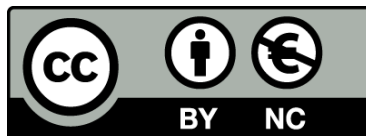




UNIVERSITAT<sub>DE</sub>  
BARCELONA

# Molecular epidemiology study on genetically regulated gene expression in the colonic mucosa and its role in disease susceptibility

Virginia Díez Obrero



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial 4.0. Espanya de Creative Commons**.

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial 4.0. España de Creative Commons**.

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial 4.0. Spain License**.



# **Molecular epidemiology study on genetically regulated gene expression in the colonic mucosa and its role in disease susceptibility**

Memòria de tesi doctoral presentada per

**Virginia Díez Obrero**

per optar al grau de Doctora per la Universitat de Barcelona

**Dirigida pels Drs. Víctor Raúl Moreno Aguado i Robert Carreras Torres,**  
de la Facultat de Medicina i Ciències de la Salut de la Universitat de Barcelona (UB), el Programa ONCOBELL de L'Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), el Programa d'Analítica de Dades en Oncologia de L'Institut Català d'Oncologia (ICO) i el Centre d'investigació Biomèdica en Red d'Epidemiologia i Salut Pública (CIBERESP)

Programa de Doctorat en Medicina i Recerca Translacional

Facultat de Medicina i Ciències de la Salut

Universitat de Barcelona

Setembre 2021





El Dr. Víctor Raúl Moreno Aguado i el Dr. Robert Carreras Torres, de la Facultat de Medicina i Ciències de la Salut de la Universitat de Barcelona (UB), el Programa ONCOBELL de L'Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), el Programa d'Analítica de Dades en Oncologia de L'Institut Català d'Oncologia (ICO) i el Centre d'investigació Biomèdica en Red d'Epidemiologia i Salut Pública (CIBERESP)

CERTIFIQUEN:

que la Tesi realitzada per la estudiant de doctorat Virginia Díez Obrero sota la seva supervisió, titulada “Molecular epidemiology study on genetically regulated gene expression in the colonic mucosa and its role in disease susceptibility”, és adequada per ser defensada públicament.

Barcelona, a 2 de Setembre de 2021.

Dr. Víctor Raúl Moreno  
Aguado  
Director

Dr. Robert Carreras  
Torres  
Co-director

Virginia Díez  
Obrero  
Estudiant de doctorat









Ítaca t'ha donat el bell viatge.  
Sense ella no hauries pas sortit cap a fer-lo.  
Res més no té que et pugui ja donar.  
I si la trobes pobra, no és que Ítaca t'hagi enganyat.  
Savi com bé t'has fet, amb tanta experiència,  
ja hauràs pogut comprendre què volen dir les Ítaques.

**Extracte d'“Ítaca” (1911)**

**Konstantinos P. Kavafis**

Ítaca te ha dado un viaje hermoso.  
Sin ella no te habrías puesto en marcha.  
Pero no tiene ya más que ofrecerte.  
Y aunque la encuentres pobre, Ítaca no te ha engañado.  
Así, sabio como te has vuelto, con tanta experiencia,  
ya habrás comprendido el significado de las Ítacas.

**Extracto de “Ítaca” (1911)**

**Konstantinos P. Kavafis**

Ithaka gave you the marvelous journey.  
Without her you wouldn't have set out.  
She has nothing left to give you now.  
And if you find her poor, Ithaka won't have fooled you.  
Wise as you will have become, so full of experience,  
you'll have understood by then what these Ithakas mean.

**Excerpt from “Ithaka” (1911)**

**Konstantinos P. Kavafis**



## ACKNOWLEDGEMENTS

I have been very lucky to share these last four years with very nice, supportive, and talented people who have helped me to become a scientist and develop this Thesis. I am very grateful to all of them.

Dr. Victor Moreno gave me the opportunity to join his team and participate in top projects. He has trusted me, has given me independence to propose analyses, and has always been considerate and willing to listen and help.

Dr. Robert Carreras has been a catalyst in my learning process. He has provided helpful support and honest advice. I admire his talent for mentoring.

It has been a pleasure to work with current and former members of the Oncology Data Analytics Program. Ferran Moratalla is a wonderful colleague and makes delicious limoncello. I would like to thank Dr. Rebeca Sanz, Anna Díez, Dr. Mireia Obón, and Dr. Gemma Ibáñez for their help and encouragement. Special thanks to Ania Alay, who has always been supportive and *recursive*.

In addition, I am very happy to have been working in the GECCO consortium, in such an international and collaborative environment. The leadership of Dr. Ulrike Peters is inspiring and admirable. I have learnt a lot from the colleagues working in the FIGI and GxE working groups, especially from Dr. Graham Casey's team.

Also, I would like to thank Dr. James McKay for hosting me at his group at IARC. Although it was an unusual stay during pandemic times, it was a very enriching experience, and it was nice to meet inspiring scientists such as Dr. Aida Ferreiro.

Finally, I would like to thank the hundreds of anonymous individuals who altruistically donated their samples and shared their genetic and clinical information for research. Without their trust and commitment, the studies presented in this Thesis would not have been possible. This Thesis is dedicated to all of them.

Thank you!



## **FUNDING**

This Thesis has been developed thanks to the “Programa de Formació de Professorat Universitari” (FPU) fellowship [FPU16/00599] awarded to Virginia Díez Obrero by the Spanish Ministry of Education, Culture and Sport. The stay at the International Agency for Research on Cancer (IARC) was supported by the European Molecular Biology Organization (EMBO) Short-Term fellowship [number 8870].

The work carried out in this Thesis was funded by the National Institutes of Health [R01 CA204279, R01 CA143237 and R01 CA201407]; the Agency for Management of University and Research Grants (AGAUR) of the Catalan Government [2017SGR723]; the Instituto de Salud Carlos III, co-funded by FEDER funds, a way to build Europe [PI14-00613, and PI17-00092]; the Spanish Association Against Cancer (AECC) Scientific Foundation [GCTRA18022MORE]; and the Centro de investigación biomédica en red. Epidemiología y salud pública (CIBERESP) [CB07/02/2005].





## TABLE OF CONTENTS

|   |    |
|---|----|
| LIST OF FIGURES   | 15 |
| GLOSSARY  | 17 |
| LIST OF THE ARTICLES THAT COMPRISE THE THESIS   | 21 |
| THESIS SUMMARY  | 23 |
| RESUM   | 27 |
| 1. INTRODUCTION   | 33 |
| 1.1. Gene expression: a key molecular process involved in disease etiology                  | 33 |
| 1.1.1. Regulatory processes of gene expression  | 33 |
| 1.1.1.1. Mechanisms of mRNA splicing  | 33 |
| 1.1.1.2. Genetic regulation of gene expression  | 36 |
| 1.1.2. Approaches to link genetically regulated gene expression with disease susceptibility | 39 |
| 1.1.2.1. Functional annotation  | 40 |
| 1.1.2.2. Colocalization   | 40 |
| 1.1.2.3. Transcriptome-wide association   | 42 |
| 1.2. The human colon in health and disease  | 44 |
| 1.2.1. Anatomy and main functions   | 44 |
| 1.2.2. Normal colon gene expression   | 48 |
| 1.2.3. Common complex diseases affecting the colon  | 49 |
| 1.2.3.1. Colorectal cancer  | 50 |
| 1.2.3.2. Inflammatory bowel disease   | 53 |
| 2. HYPOTHESES   | 57 |
| 3. OBJECTIVES   | 59 |

|   |     |
|---|-----|
| 4. MATERIALS AND METHODS AND RESULTS                                      | 61  |
| 4.1. BarcUVa-Seq normal colon e/sQTLs.                                    | 61  |
| 4.2. The Colon Transcriptome Explorer (CoTrEx) 2.0.                       | 80  |
| 4.3. Transcription-Wide Association Study for Inflammatory Bowel Disease. | 86  |
| 5. DISCUSSION   | 99  |
| 5.1. Discussion of Objective 1  | 99  |
| 5.2. Discussion of Objective 2  | 105 |
| 5.3. Discussion of Objective 3  | 108 |
| 5.4. Global discussion  | 112 |
| 6. CONCLUSIONS  | 117 |
| 7. REFERENCES   | 119 |

## LIST OF FIGURES

|   |     |
|---|-----|
| Figure 1. Overview of alternative splicing (AS).  | 34  |
| Figure 2. Constitutive and alternative splicing events.   | 36  |
| Figure 3. Overview of eQTL mapping.   | 37  |
| Figure 4. e/sQTL effect patterns across human tissues.  | 38  |
| Figure 5. Screenshot of the GTEx Transcript Browser.  | 39  |
| Figure 6. Overview of colocalization.   | 41  |
| Figure 7. Overview of the transcription-wide association study (TWAS) approach.                       | 43  |
| Figure 8. The human colon anatomy.  | 46  |
| Figure 9. Schematic of a colonic crypt.   | 47  |
| Figure 10. Screenshot of the Colonomics gene expression browser.                                      | 49  |
| Figure 11. Overview of colorectal cancer (CRC) development.   | 51  |
| Figure 12. Manhattan plot of the TWAS results for CRC.  | 52  |
| Figure 13. Dysregulation of immune response in inflammatory bowel diseases (IBD).                     | 54  |
| Figure 14. Molecular pathways driving IBD.  | 56  |
| Figure 15. The gut-brain axis and cell type composition of the colon mucosa.                          | 102 |
| Figure 16. Linkage, causality, and pleiotropy effects on colocalization.                              | 113 |
| Figure 17. Approaches for translating disease associated risk <i>loci</i> into targeted therapeutics. | 115 |



## GLOSSARY

**Alternative splicing:** Alternative splicing (AS) is the molecular process by which a gene can derive into multiple mRNA transcript isoforms that originate from the same *locus*. There have been characterized different mechanisms of AS, also called AS events.

**BarcUVa-Seq:** The University of Barcelona and University of Virginia genotyping and RNA sequencing project is a collaboration between researchers in the University of Barcelona and the University of Virginia. In this project, there were collected colon tissue biopsies, blood samples and epidemiological information from up to 485 healthy adults.

**e/sQTLs:** Expression and splicing quantitative trait loci (e/sQTLs) refer to single nucleotide polymorphisms (SNPs) statistically associated with gene expression and alternative splicing profiles, respectively.

**Fine-mapping:** Statistical fine-mapping is an approach to identify the causal gene(s) or variant(s) involved in a trait of interest and its variability. It assigns a probability for each gene or variant in a given *locus* to be the cause underlying an association signal.

**Gene expression:** Gene expression is the process by which a DNA sequence is converted into the encoded molecule (mainly a protein or a non-coding RNA). The level of expression of a given gene is indicative of the level of activity of this gene in the cell.

**Genetic susceptibility:** Increased likelihood of developing a particular disease or trait due to the presence of germline genetic variants. Also called genetic/inherited predisposition.

**Genetically regulated gene expression:** Gene expression which levels are influenced in a given direction by the alleles of genetic variants, such as SNPs.

**Genotype:** Combination of possible alleles present at a specific *locus*, given by the pair of chromosomes.

**GTEx:** The genotype-tissue expression project (GTEx) represents the largest atlas of human tissue gene expression to date. It includes samples from up to 49 different tissues and cell types that were collected from 838 post-mortem donors.

**GWAS:** Genome-wide association studies (GWAS) are case-control epidemiologic studies that survey the entire genome for measuring association between single nucleotide polymorphisms (SNPs) and phenotypes, such as disease status.

**Linkage disequilibrium:** Phenomenon according to which a nonrandom association exists between alleles at different *loci*, which appear associated in a population more often than by chance.

**mRNA:** Messenger RNA (mRNA) is the direct product of the process of gene expression and a key intermediate molecule between a given DNA sequence and the protein that it encodes.

**Pleiotropy:** the phenomenon by which one genetic variant appears associated with multiple phenotypes.

**RNA-Seq:** RNA Sequencing (RNA-Seq) is a high throughput technique for assessing the transcriptome of a given biological sample. It produces sequencing reads to be mapped to a reference genome/transcriptome. It allows quantifying the levels of

RNA and assessing global gene expression as well as the expression of alternatively spliced transcripts.

**SNP:** A single nucleotide polymorphism (SNP) is a type of germline genetic variation affecting one single DNA nucleotide, *i.e.* adenine (A), thymine (T), cytosine (C), or guanine (G). A common SNP occurs when the frequency of the variant is present in at least 1% of genetic sequences in a particular population.

**SNP-based heritability:** The SNP-based heritability ( $h^2_{\text{SNP}}$ ) can be defined as the proportion of phenotypic variance explained by a set of SNPs. These SNPs could be those included on a genotyping array, those sequenced from whole-genome/exome sequencing or imputed from reference SNP imputation panels.

**Transcript:** A transcript is a RNA sequence generated from DNA during the process of transcription. There are different types of transcripts, which can be broadly classified according to their potential to be translated into proteins.

**Transcriptome:** The transcriptome encompasses the entire collection of RNA molecules expressed from the genome. The human transcriptome contains 63,568 RNA sequences (according to the annotations provided by GENCODE release 29). These include 22,705 (around 36%) protein-coding genes.

**TWAS:** Transcriptome-wide association studies (TWAS) are case-control studies that test for the association between genetically predicted gene expression and a complex disease or trait of interest. They identify genes whose imputed expression in a particular tissue/cell of interest is up or downregulated in cases in comparison with controls. They do not profile gene expression, rather, they impute it from genotype data.





## LIST OF THE ARTICLES THAT COMPRISE THE THESIS

This Thesis is in the form of a collection of published articles. It comprises three main objectives, and three articles (one published, one accepted for publication and one prepared for submission). Journal metrics indicated below are according to the Journal Citation Reports (JCR) 2021.

### Objective 1

To provide reference transcriptome-wide gene expression and alternative splicing profiles of colon mucosal biopsies from healthy adults, as well as their differences across colon location and corresponding e/sQTLs. Also, to identify complex traits and diseases whose SNP-based heritability is enriched in the e/sQTLs identified, and propose candidate susceptibility genes for these phenotypes.

**Díez-Obrero V**, Dampier CH, Moratalla-Navarro F, Devall M, Plummer SJ, Díez-Villanueva A, *et al.* Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci. *Cell Mol Gastroenterol Hepatol.* 2021 Feb 16;12(1):181–97.

Journal impact factor: 9.225; Quartile: 1; Subject category: Gastroenterology & Hepatology.

### Objective 2

To develop a web resource to explore population-based normal colon transcriptome profiles, e/sQTLs, gene expression prediction models, as well as to annotate SNPs with eQTLs.

**Díez-Obrero V**, Moratalla-Navarro F, Dampier CH, Devall M, Carreras-Torres R, Casey G, *et al.* The Colon Transcriptome Explorer (CoTrEx) 2.0, a reference resource for exploring population-based normal colon gene expression. Prepared for submission.

### **Objective 3**

To propose candidate genes whose genetically regulated gene expression is associated with IBD, including genes in specific colon subsites that are expression markers of colon cell types; with emphasis on gene expression markers of colon cell types, and gene enrichment in IBD therapy-related molecular pathways. Also, to identify candidate susceptibility genes specific for the epithelial, immune/blood, mesenchymal and neural tissue categories.

**Díez-Obrero V**, Moratalla-Navarro F, Ibáñez-Sanz G, Guardiola J, Rodríguez-Moranta F, Obón-Santacana M, *et al.* Transcriptome-wide association study for inflammatory bowel disease reveals novel candidate susceptibility genes in specific colon subsites and tissue categories. *J Crohns Colitis* [Internet]. 2021 Jul 21; Available from: <http://dx.doi.org/10.1093/ecco-jcc/jjab131>.

Journal impact factor: 9.071; Quartile: 1; Subject category: Gastroenterology & Hepatology.

## THESIS SUMMARY

### Title

Molecular epidemiology study on genetically regulated gene expression in the colonic mucosa and its role in disease susceptibility.

### Introduction

Gene expression is a key molecular process that is tightly regulated by a wide repertoire of molecular mechanisms. Most genes undergo alternative splicing (AS), a process that generates different transcript isoforms from a single locus. Genetic variants (*i.e.* SNPs) that regulate gene expression and AS are called expression and splicing quantitative trait *loci* (eQTLs and sQTLs), respectively, and can operate in a tissue-specific manner.

Gene expression can be involved in disease etiology, and the analyses of genetically regulated gene expression can provide evidence of a causal relation. The statistical colocalization between eQTLs and risk SNPs allows linking gene expression in a specific tissue or cell type with disease risk. In addition, the eQTL summary statistics can be used to predict gene expression from genotype data. These gene expression prediction models are used in transcriptome-wide association studies (TWAS) to associate differential levels of predicted gene expression with disease status.

In the case of colon tissue, there is a lack of an adequate representation of the transcriptome of the epithelial mucosa from biopsies of living individuals, which has limited the profiling of gene expression and its genetic regulation across the colon. Two major common chronic diseases affecting the colon are colorectal cancer (CRC) and inflammatory bowel disease (IBD). Genome-wide association studies (GWAS) have identified a total of 141 and 240 genetic risk variants related to them, respectively. Although some functional evidence has been provided, the mechanisms underlying genetic susceptibility for these and other colon-related diseases are not yet fully understood.

## **Hypotheses**

Colon gene expression and AS profiles derived from RNA sequencing of mucosal biopsies can provide good estimates of the colon tissue transcriptome. They may vary across the colon anatomy, and can be controlled by genetic variants (*i.e.* SNPs). In addition, an interactive web-based resource can facilitate researchers a quick and centralized exploration of colon gene expression-related data. Finally, genetically regulated gene expression in the colon can play a role in the susceptibility to complex traits and diseases, such as CRC or IBD, and candidate susceptibility genes for these diseases could be nominated.

## **Objectives**

1. To provide reference transcriptome-wide gene expression and alternative splicing profiles of colon mucosal biopsies from healthy adults, as well as their differences across colon location and corresponding e/sQTLs. Also, to identify complex traits and diseases whose SNP-based heritability is enriched in the e/sQTLs identified, and propose candidate susceptibility genes for these phenotypes.
2. To develop a web resource to explore normal colon transcriptomic profiles, e/sQTLs, gene expression prediction models, as well as to annotate SNPs with colon eQTLs.
3. To propose candidate genes whose genetically regulated gene expression is associated with IBD, including genes in specific colon subsites that are expression markers of colon cell types, and genes that are enriched in relevant molecular pathways for IBD, such as therapy-related pathways. Also, to identify candidate susceptibility genes specific for the epithelial, immune/blood, mesenchymal and neural tissue categories.

## **Methods**

We included 445 individuals with tissue RNA-Seq and genome-wide genotyping data. RNA-Seq reads were aligned to the reference transcriptome with STAR. Gene expression was quantified with RSEM. AS events were profiled with SUPPA2 and LeafCutter. Genotyping was performed with the Illumina OncoArray BeadChip. Allelic dosages from about 40,000,000 SNPs were obtained after imputation with the Haplotype Reference Consortium panel. Differential gene expression was carried out with the edgeR R package. e/sQTLs were mapped with FastQTL. Models were adjusted for sex, age, colon location, sequencing batch, probabilistic estimation of expression residuals (PEER) factors, and genetic ancestry. Heritability enrichment and colocalization analyses were performed with LD-Score and fastENLOC software, respectively, using GWAS summary statistics data.

The web application was developed with the R package Shiny.

We generated gene expression prediction models using elastic net regression. Models were compiled for a total of 62 tissues and cell types, including the BarcUVa-Seq, GTEx and CEDAR datasets. We performed single and multi-tissue TWAS following the S-PrediXcan and S-MultiXcan approaches, respectively. We used publicly available IBD, Crohn's disease (CD) and ulcerative colitis (UC) GWAS summary statistics from about 60,000 individuals. For the gene enrichment analysis, signaling and regulatory pathways from the Pathway Interaction Database were used, and enrichment values were measured by hypergeometric tests.

## **Main results**

We generated the BarcUVa-Seq dataset, which included colon biopsy gene expression and genome-wide genotypes from 445 healthy people. We described the gene expression and alternative splicing differences across colon subsites. We identified 11,739 eQTLs and 1,125 sQTLs. We found that part of the SNP-based heritability of diseases affecting colon tissue, such as CRC and IBD, but also of

diseases affecting other tissues, such as psychiatric conditions, can be partly explained by the identified QTLs. We provided candidate susceptibility genes for these phenotypes.

The Colon Transcriptome Explorer (CoTrEx) 2.0 was hosted online at <https://barcuvasseq.org/cotrex/>. It is based on BarcUVa-Seq and GTEx colon data and features plots, tables, and customization options for exploring gene and transcript expression profiles, e/sQTLs, summary statistics of gene expression prediction models, and regulatory and coexpression networks.

Finally, we identified 136, 116 and 88 novel candidate susceptibility genes for IBD, CD and UC, respectively, expressed across 62 tissues and cell types. We provided 39 novel genes whose expression in the colon is associated with IBD status, including expression markers for specific colon cell types. Additionally, we identified 78 novel susceptibility genes whose expression was associated with IBD exclusively in immune (N=19), epithelial (N=25), mesenchymal (N=22) and neural (N=12) tissue categories. The associated genes were involved in relevant molecular pathways, including pathways related to the immune system, and pathways related to known IBD therapeutics, such as tumor necrosis factor signaling.

## **Conclusions**

We provided a large characterization of gene expression and AS, and their genetic regulation, across colon subsites. The findings add biological insight into complex traits and diseases influenced by transcriptomic changes in the colonic mucosa.

We provided the Colon Transcriptome Explorer 2.0 including large population-based normal colon gene expression resources.

We proposed novel genes whose genetically regulated gene expression across tissues and cell types is associated with IBD status. These genes might be prioritized in further functional studies.

## RESUM

### Títol

Molecular epidemiology study on genetically regulated gene expression in the colonic mucosa and its role in disease susceptibility.

### Introducció

L'expressió gènica és un procés molecular clau, estretament regulat per un ampli repertori de mecanismes moleculars. La majoria dels gens es sotmeten a un *splicing* alternatiu (AS), un procés que genera diferents isoformes a partir d'un sol *locus*. Les variants genètiques (*i.e.* SNPs) que regulen l'expressió gènica i l'AS es denominen *loci* de trets quantitius d'expressió i *splicing* (eQTL i sQTL), respectivament, i mostren especificitat de teixit.

L'expressió gènica pot estar implicada en l'etiologia de les malalties, i les anàlisis de l'expressió gènica regulada genèticament poden proporcionar evidències d'una relació causal. La colocalització estadística entre SNPs de risc i els eQTL al llarg del genoma permet vincular l'expressió gènica en un tipus específic de teixit o cèl·lula amb el risc de malaltia. A més, els eQTLs poden explicar part de l'heretabilitat basada en SNPs de malalties complexes comunes. A més, les estadístiques d'associació resumides d'eQTLs es poden utilitzar per predir l'expressió gènica a partir de dades de genotips. Aquests models d'expressió gènica predits genèticament es poden utilitzar en estudis d'associació a tot el transcriptoma (TWAS) en els quals es comparen els nivells d'expressió gènica predits en el teixit d'interès entre casos i controls.

En el cas del teixit del còlon, manca una representació adequada del transcriptoma de la mucosa epitelial a partir de biòpsies d'individus vius; cosa que ha limitat el perfilat de l'expressió gènica, i la seva regulació genètica, a tot el còlon. Pel que fa a les malalties cròniques habituals que afecten el còlon, els estudis d'associació al llarg del genoma (GWAS) han identificat 141 i 240 variants genètiques de risc



relacionades amb el càncer colorectal (CCR) i la malaltia inflamatòria intestinal (MII), respectivament. Tot i que s'han observat algunes evidències funcionals, els mecanismes subjacents a la susceptibilitat genètica per aquestes i altres malalties relacionades amb el còlon, encara no es comprenen en la seva totalitat.

### **Hipòtesis**

L'expressió dels gens i els perfils de *splicing* alternatiu derivats de la seqüenciació de l'ARN de biòpsies de mucosa de còlon poden proporcionar bones estimacions del transcriptoma del teixit del còlon. Aquests poden variar segons l'anatomia del còlon i poden ser controlats per variants genètiques (*i.e.* SNP). A més, un recurs web interactiu pot facilitar als investigadors un accés i exploració ràpida i centralitzada de dades relacionades amb l'expressió gènica del còlon. Finalment, l'expressió gènica regulada genèticament al còlon pot jugar un paper en la susceptibilitat genètica a trets i malalties complexes, com ara CCR o MII, i es poden designar gens de susceptibilitat per a aquestes malalties.

### **Objectius**

1. Proporcionar perfils de referència a nivell de transcriptoma complet, tant d'expressió genètica com de *splicing* alternatiu, de biòpsies de mucosa del còlon d'adults sans, així com les seves diferències entre distintes localitzacions al còlon, i els seus e/sQTLs corresponents. També, identificar trets i malalties complexes la heretabilitat basada en SNPs de les quals estiga enriquida per els e/sQTLs identificats, i proposar gens de susceptibilitat per a aquests fenotips.
2. Desenvolupar un recurs web per explorar els perfils transcriptòmics del còlon sa, els seus e/sQTLs, i els seus models de predicció d'expressió genètica, així com per anotar SNPs de risc amb eQTLs de còlon.

3. Proposar gens l'expressió gènica regulada genèticament dels quals s'associï a MII, incloent gens a llocs específics del còlon que siguin marcadors d'expressió de cèl·lules, i gens que siguin enriquits a rutes moleculars rellevants per la MII, com rutes relacionades amb teràpies per MII. També, identificar gens de susceptibilitat específics per les categories de teixits epitelial, immune, mesenquimal i neural.

## **Mètodes**

Es van incloure 445 individus amb dades de genotips i de RNA-Seq de teixit. Les seqüències de RNA es van alinear amb el transcriptoma de referència emprant STAR. L'expressió gènica es va quantificar amb RSEM. Es van quantificar els events de *splicing* alternatiu amb SUPPA2 i LeafCutter. El genotipat al llarg del genoma es va realitzar amb l'Illumina OncoArray BeadChip. Es van obtenir les dosis al·lèliques d'uns 40.000.000 SNPs després de la seua imputació amb el panell Haplotype Reference Consortium. L'expressió gènica diferencial es va dur a terme amb el paquet estadístic de R edgeR. Els e/sQTLs es van computar amb FastQTL. Els models es van ajustar per sexe, edat, localització del còlon, lot de seqüenciació, estimació probabilística dels factors residuals d'expressió (PEER) i ascendència genètica. Les anàlisis de colocalització i proporció de la varianza de l'heretabilitat es van realitzar amb els programes fastENLOC i LD-Score, respectivament, utilitzant dades d'estadístiques resumides de GWAS.

L'aplicació web es va desenvolupar amb el paquet R Shiny.

Els models de predicció d'expressió genètica es van generar mitjançant regressió de tipus xarxa elàstica. En general, vam predir l'expressió de gens en 62 teixits i tipus de cèl·lules sanguínies utilitzant dades dels projectes BarcUVa-Seq, GTEx i CEDAR. Els TWAS per teixit i combinant varis teixits es van realitzar seguint els protocols S-PrediXcan i S-MultiXcan, respectivament. Per aquests anàlisis, es van incloure dades d'estadístiques resumides de GWAS públiques de MII, colitis ulcerosa i malaltia de

Crohn d'uns 60.000 individus. Per als anàlisis d'enriquiment de gens es van utilitzar les rutes moleculars de senyalització i reguladores de la Pathway Interaction Database, i els valors d'enriquiment es van calcular amb tests hipergeomètrics.

### **Resultats principals**

Es va generar el conjunt de dades "BarcUVa-Seq", que inclou l'expressió gènica de biòpsies de còlon i genotips al llarg del genoma de 445 persones sanes. Vam descriure les diferències a nivell d'expressió gènica i de *splicing* alternatiu entre distintes localitzacions al llarg del còlon. Vam identificar 1,739 eQTLs i 1,125 sQTLs. Vam trobar que una proporció considerable de l'heretabilitat basada en SNPs de malalties que afecten el còlon es pot explicar pels QTLs identificats, com ara el càncer colorectal i la MII, però també de malalties que afecten altres teixits, com les afeccions psiquiàtriques. També, vam proporcionar gens de susceptibilitat per a aquests fenotips.

Es va desenvolupar el "Colon Transcriptome Explorer" (CoTrEx) i la seva versió actualitzada 2.0. CoTrEx està disponible a <https://barcuvaseq.org/cotrex/>. Aquesta eina web es basa en dades de còlon dels projectes "BarcUVa-Seq" i "GTEx" i presenta gràfics, taules i opcions interactives per explorar perfils d'expressió de gens i trànscrips, e/sQTLs, models de predicció d'expressió genètica i xarxes reguladores i de coexpressió.

Finalment, es van identificar 136, 116 and 88 nous gens de susceptibilitat per MII, malaltia de Crohn i colitis ulcerosa, respectivament. Es va proporcionar 39 nous gens de susceptibilitat l'expressió dels quals al còlon s'associa amb MII. Aquests gens inclouen marcadors d'expressió per a tipus específics de cèl·lules al còlon. Per altra banda, en la metaanàlisi de tots els resultats, vam trobar 186 nous gens de susceptibilitat. A més, es van identificar 78 nous gens de susceptibilitat l'expressió dels quals s'associa amb MII exclusivament en teixits immunes (N=19), epitelials (N=25), mesenquimals (N=22) i neuronals (N=12). Els gens associats participen en

vies moleculars rellevants, incloses vies relacionades amb teràpies conegudes de la MII, com la senyalització del factor de necrosi tumoral.

### **Conclusions**

Es va proporcionar una exhaustiva caracterització de l'expressió gènica i el *splicing* alternatiu al llarg del còlon. Els resultats amplien els coneixements sobre trets i malalties complexes influenciades per canvis transcriptòmics al còlon.

Es va desenvolupar el *Colon Transcriptome Explorer 2.0* incloent dades d'expressió genètica al còlon.

Es va proposar al voltant de dos-cents gens nous l'expressió gènica regulada genèticament dels quals, a una sèrie de teixits i tipus de cèl·lules, està associada a MII. Aquests gens s'haurien de prioritzar a estudis funcionals posteriors.



## **1. INTRODUCTION**

### **1.1. Gene expression: a key molecular process involved in disease etiology**

Gene expression represents a key complex molecular process that acts as an intermediary between the DNA and the functional molecules that carry out cellular functions. It is inferred from the number of transcripts quantified in a given cell/tissue at a given time, and reflects the amount of activity of the measured genes. Gene expression is dynamic and tightly regulated. While there are more than 20,000 protein coding genes, not all are expressed in every cell, and there is a notable heterogeneity in gene expression across tissues (1). In addition, gene expression levels constitute a significant source of phenotypic diversity among individuals within populations and it can play a key role in disease susceptibility (2).

#### **1.1.1. Regulatory processes of gene expression**

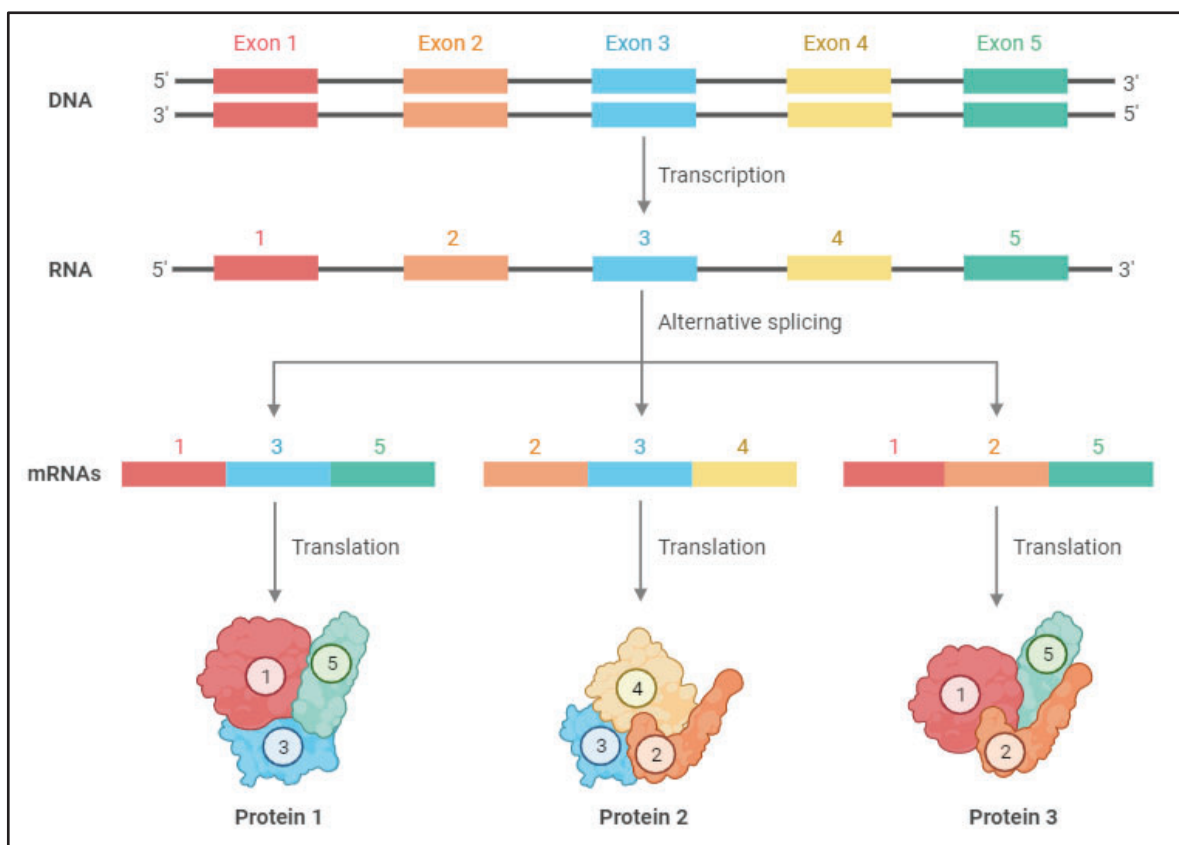
Characterizing the regulatory architecture of transcriptome-wide gene expression is a key challenge. Regulation occurs at different levels that can be broadly classified into regulation of transcription initiation, regulation of messenger RNA (mRNA) processing, and post-transcriptional regulation. Each regulatory level involves a wide variety of interrelated complex molecular processes. First, the regulation of transcription initiation includes mechanisms related to chromatin accessibility, alternative transcription start sites or transcription factor binding. Secondly, regulation of mRNA processing includes mechanisms regarding mRNA splicing, RNA editing, nonsense-mediated decay and regulation by microRNA. Finally, post-transcriptional regulation includes processes such as post-translational RNA modification and translation (3).

##### **1.1.1.1. Mechanisms of mRNA splicing**

The mRNA splicing is an essential gene expression regulatory mechanism whose primary function is the removal of non-coding introns. This process is carried out by

the spliceosome, a large ribonucleoprotein complex whose core components are highly conserved (3). The spliceosome requires the recognition of the splice sites, which are essential nucleotides that aid in the recognition of exons. Splice sites can be constitutive or alternative, depending on whether they are always (constitutive) or only sometimes (alternative) recognized and spliced into the mature mRNA (4).

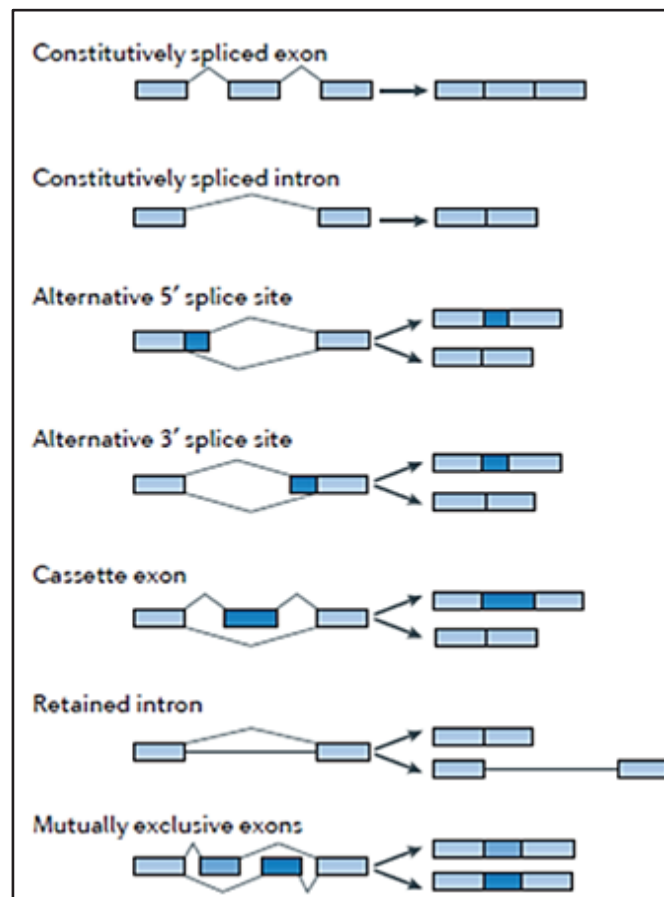
Alternative splicing (AS) is the process by which multiple mRNA transcript isoforms are generated due to the different selection of splice sites, which will ultimately result in different polypeptides that while originating from the same *locus* can be notably different (3). See **Figure 1**. Approximately 94% of human multi-exonic genes undergo AS, and 86% of these have a minor transcript isoform frequency of 15% or higher (5). The selection of a particular transcript isoform is often performed during the early stages of splice site recognition and spliceosome assembly (7).



**Figure 1. Overview of alternative splicing (AS).** General schematic of AS, which is shown as an intermediate step between DNA transcription and protein translation. Reprinted from “Gene Splicing”, by BioRender.com (2021). Retrieved from <https://app.biorender.com/biorender-templates>.

AS events refer to the different mechanisms by which a gene can be alternatively spliced, generating variability in the exonic structure of mature mRNAs. The exon skipping event is the most common in humans, where a given exon is present in some transcripts but not others. Another common event occurring in nearly 75% of multi-exon genes is the intron retention event, where introns are spliced out in some transcripts but not others (6). Other highly relevant, well characterized AS events are alternative 5' (donor) and 3' (acceptor) splice sites, where the 5'/3' splice sites differ among transcripts, and mutually exclusive exons, where two or more exons are retained (7). Also, it is important to note that a given transcript isoform can be the outcome of multiple simultaneous AS events (5). A scheme including the AS events mentioned above is shown in **Figure 2**. Other AS events include alternative first or last exons, where the first/last exon varies among transcripts, and alternative 5' or 3' UTR, where the 5'/3' untranslated region of a transcript varies among transcripts. Events that are not included in the groups stated above are less characterized and often classified as complex AS events (7).





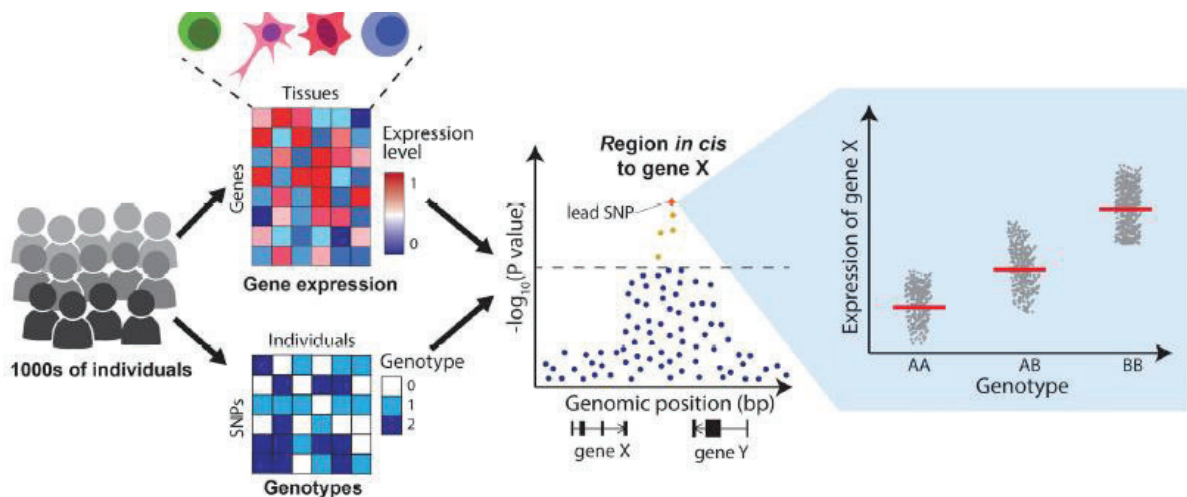
**Figure 2. Constitutive and alternative splicing events.** Scheme showing constitutive splicing events as well as the following AS events: alternative donor (5') and acceptor (3') splice sites, cassette exon (*i.e.* exon skipping), intron retention, and mutually exclusive exons. Light blue: constitutive sequence that always is included in the mature mRNA; mid/darker-blue: alternative sequence that can be either included or excluded in the mature mRNA. Adapted from Dvinge et al. (4).

### 1.1.1.2. Genetic regulation of gene expression

A significant proportion of gene expression variation is heritable (2). Genetic variants regulating gene expression can be local or distant, according to their relative position to the gene they regulate. Local variants are commonly defined to be located within 1 Mb of the transcription start site (TSS) of the regulated gene (8). An example of local regulation is the presence of a genetic variant within the gene promoter, which can decrease the binding affinity of the RNA polymerase, and consequently affects the expression levels of the target gene. On the other hand, an example of distal regulation is a genetic variant affecting the binding of regulatory

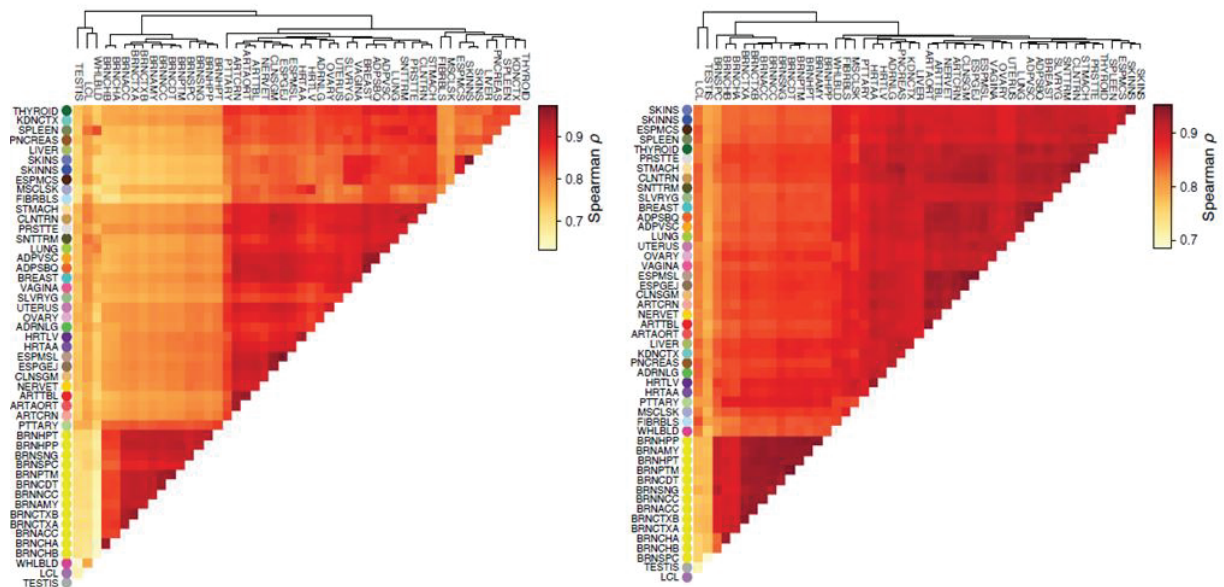
elements in sequences located in other chromosomes different to the chromosome where lies the gene of interest (2).

The genetic variants regulating gene expression and splicing are referred to as expression and splicing quantitative trait *loci* (e/sQTL), respectively. Their mapping consists of deriving statistical associations between the alleles of genotyped single nucleotide polymorphisms (SNPs) and gene expression and AS, respectively, in a particular tissue or cell-type. If association exists in a given population, individuals with different genotypes would show different average values for the studied gene/AS event (9) (see eQTL mapping schematic in **Figure 3**). e/sQTL mapping has been possible thanks to advances in high-throughput sequencing technologies, which have allowed measuring quantitatively the gene expression and AS in a genome-wide manner, and also due to the implementation of sequencing projects involving a large number of individuals. In this sense, the Genotype-Tissue Expression (GTEx) project represents the largest atlas of human tissue gene expression to date, including samples from up to 49 different tissues and cell types collected from 838 post-mortem donors (8).



**Figure 3. Overview of eQTL mapping.** eQTL mapping requires sequencing and genotyping of a large number of individuals in the tissues or cell types of interest. In this framework, statistical associations between SNP genotypes and gene expression levels of nearby genes are obtained. Adapted from Cano-Gamez et al. (10).

The generation of e/sQTL catalogs has enabled a detailed investigation of the genetic architecture of transcriptional variation in human tissues across different populations, sexes, age ranges and cellular conditions, among other variables. In general, gene expression and AS heterogeneity across tissues is consistent with the heterogeneity of e/sQTLs effects on them (8). There is a high degree of tissue similarity in terms of their regulation by e/sQTLs. In **Figure 4** it is depicted a hierarchical clustering of tissues according to their e/sQTL effects, where the brain regions (in yellow) form a separate cluster. Also, testis, lymphoblastoid cell lines and whole blood are less related to other tissues. Splicing measures are more tissue specific than gene expression, but genetic effects on splicing tend to be highly shared, which is consistent with pairwise tissue-sharing patterns, as shown in **Figure 4**.



**Figure 4. e/sQTL effect patterns across human tissues.** Tissue clustering with pairwise Spearman correlation of eQTLs (left) and sQTLs (right) effect sizes. Adapted from the GTEx consortium (8).

To explore gene expression and splicing patterns across tissues as well as e/sQTLs, there have been developed different publicly-accessible interactive online resources, such as the GTEx Transcript Browser (11). In this particular resource, the expression metrics of the transcript isoforms of a gene of interest can be retrieved,

and the gene expression-based tissue-relatedness patterns can be visualized. See an example of a gene shown in the GTEx Transcript Browser in **Figure 5**.



**Figure 5. Screenshot of the GTEx Transcript Browser.** The transcript expression levels of the gene ACTN1 across tissues are visualized in a heatmap. It includes hierarchical clusters grouping similar tissues according to the expression of the transcripts across tissues. From (11).

### 1.1.2. Approaches to link genetically regulated gene expression with disease susceptibility

Characterizing the functional impact of human genetic variation and its influence in diseases is a main challenge in current biology. This is crucial in the case of common complex diseases, in which a modest fraction of the estimated heritability can be explained by multiple genetic variants (12). Genome-wide association studies (GWAS) have reported strong genotype-phenotype associations, but the functional role of these GWAS-identified genetic variants is not fully ascertained, specially for non-exonic SNPs, which are likely to influence processes such as transcriptional regulation, noncoding RNA function or epigenetic regulation (13).

Gene expression and AS play an important role in mediating genetic susceptibility to disease, as evidenced by the significant enrichment of eQTLs and sQTLs in the SNP-based heritability estimation of complex traits and diseases (14). These SNP

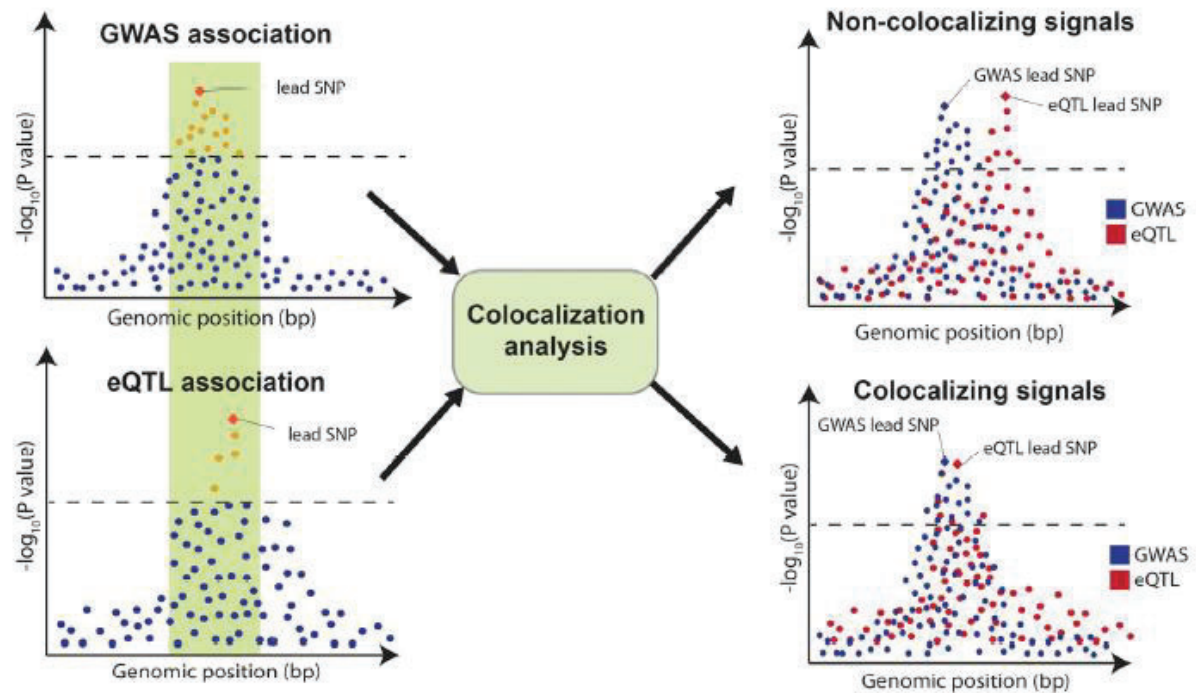
enrichments prioritize disease-relevant tissues and cell types, where transcriptional regulation may play an important role in the development of traits/diseases (10). Strategies for prioritizing genes at GWAS *loci* include functional annotation, colocalization, and transcriptome-wide association studies (TWAS).

#### **1.1.2.1. Functional annotation**

The most straightforward approach for nominating candidate effector genes in GWAS-identified risk *loci* is to perform functional annotation (15). Functional annotation consists of overlapping risk SNPs with the SNPs participating in e/sQTLs (*i.e.* e/sSNPs). If GWAS-identified SNPs and a e/sSNPs lie at the same *locus*, or are correlated due to linkage disequilibrium (LD), it could be inferred that the risk SNP confers susceptibility to disease through the modulation of the gene expression or AS feature of the target gene, with a specific direction and effect as indicated by the e/sQTL statistics. A GWAS signal overlapping an eQTL signal is indicative of potential functional relevance of the risk SNP through the modulation of gene expression. This procedure is especially useful in *loci* where there are multiple genes near a GWAS signal (10).

#### **1.1.2.2. Colocalization**

About half of identified common genetic variants are estimated to have a role in the expression of at least one gene, which may cause false positive eQTL annotations that appear due to chance, driven by LD patterns (16). To tackle this limitation, the colocalization approach applies a formal statistical test that takes LD into account for the identification of statistical signals in a genetic *locus* that colocalize, *i.e.* share the same causal variant for both GWAS and e/sQTLs signals (10). See a schematic of colocalization in **Figure 6**. Numerous colocalization methods exist, with different assumptions and statistical approaches, including the popular COLOC (17) and ENLOC methods, which estimate the posterior probability that one (or more) causal variants are shared between two traits (*e.g.* between a GWAS SNP and an eSNP).



**Figure 6. Overview of colocalization.** Colocalization compares the GWAS and eQTL associations while modelling the LD patterns of the locus, to find if both signals are driven by the same causal variants. Adapted from Cano-Gamez et al. (10).

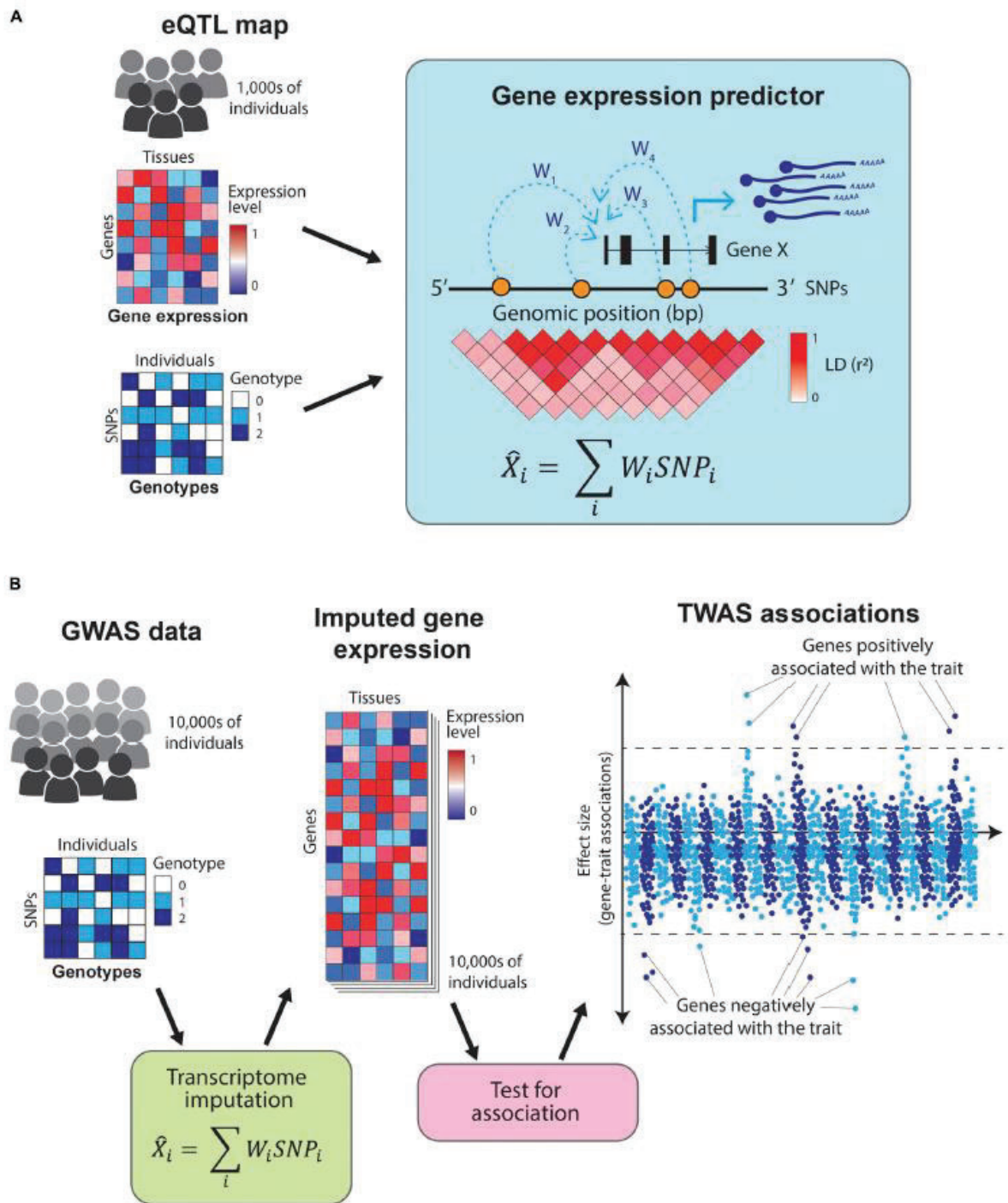
The COLOC approach (17) assumes that there is at most one causal variant per trait (*e.g.* per GWAS signal) and applies a Bayesian framework to compute the odds of colocalization in a *locus* compared with the absence of association. According to this approach, colocalization appears in a *locus* when the probability of association between the GWAS risk SNP and the eSNP due to a single colocalized SNP is higher than this association probability due to two independent colocalizing signals (10). On the other hand, the ENLOC (18) approach combines colocalization with SNP-enrichment under the reasoning that as the majority of GWAS-identified risk SNPs are enriched in tissue-specific eQTLs, then most overlaps between the two traits will be driven by true colocalizations in the relevant tissue/cell type. Therefore, this method applies a Bayesian model that weighs the probability estimations by the tissue-specific SNP enrichment estimations.



### 1.1.2.3. Transcriptome-wide association

Despite significant increases in GWAS sample sizes during the last years, they did not reach the plateau to capture the complete landscape of genetic variants that contribute to complex traits and diseases. This is notable in the case of variants with small effect sizes, which detection requires very large GWAS sample sizes. A novel strategy that permits finding new risk *loci* related to gene expression is the TWAS approach (10).

A TWAS tests for association between gene expression and a complex trait/disease status comparing predicted (*i.e.* imputed) gene expression levels among cases and controls. More in detail, the prediction of the gene expression by statistical procedures can be performed thanks to the use of reference imputation panels, which consists of statistical associations between SNPs and observed gene expression levels obtained from sequencing projects in target tissues of interest. Applying these gene expression prediction models to GWAS-derived data (which is often publicly available for many complex traits and diseases) allows the imputation and comparison of predicted gene expression levels between cases and controls (19). This way, TWAS approach avoids sequencing the mRNA of multiple tissues and cell types from hundreds of thousands cases and controls, which would be unfeasible, especially for non-easily accessible tissues. See an overview of the TWAS approach in **Figure 7**. Additionally, the indirect TWAS testing approach using germline genetics data provides evidence of directional causality from gene expression to disease risk (19).



**Figure 7. Overview of the transcription-wide association study (TWAS) approach. (A)** TWAS uses SNP-gene association results in a tissue of interest (“eQTL map”) to train predictors, *i.e.* gene expression prediction models, which estimate the SNP weights ( $w_i$ ) on expression levels of nearby genes, accounting for LD. **(B)** Predictors are used to impute gene expression levels in individuals included in a GWAS study, whose genotype data is available. Finally, the imputed gene expression values are tested for association with the trait/disease, resulting in a set of genes whose expression positively or negatively influences the trait. Reprinted from Cano-Gamez *et al.* (10).



Different statistical approaches have been developed to implement TWAS. Some of them require individual-level genotypes, which are often not publicly accessible. Others, such as the Summary-PrediXcan (S-PrediXcan) approach (20), allows using GWAS summary statistics (*i.e.* association statistics between SNPs and a given trait/disease), facilitating the application of TWAS for many complex traits and diseases. Additional improvements of these methods, such as the S-MultiXcan approach, allow meta-analysing TWAS results across tissues (21), providing more powerful estimates given the notable sharing in genetic regulation of gene expression across tissues. The development of improved TWAS approaches that provide more accurate estimates is an ongoing active area of research, and new methodologies and model implementations are continuously being published.

Finally, the limitations of the TWAS approach and its comparison with other methods have been assessed (22). Of note, a source of possible false positive results provided by the TWAS approach is the correlation of predicted expression between genes within a *locus* due to correlation between multiple eSNPs and the GWAS hit SNPs. To mitigate the biases that appear due to co-regulation between genes, the fine-mapping of causal gene sets (FOCUS) approach was developed, which directly models predicted expression correlations and uses this information to assign genes posterior probabilities of causality (23). This fine-mapping approach is of special relevance in *loci* where multiple TWAS signals appear.

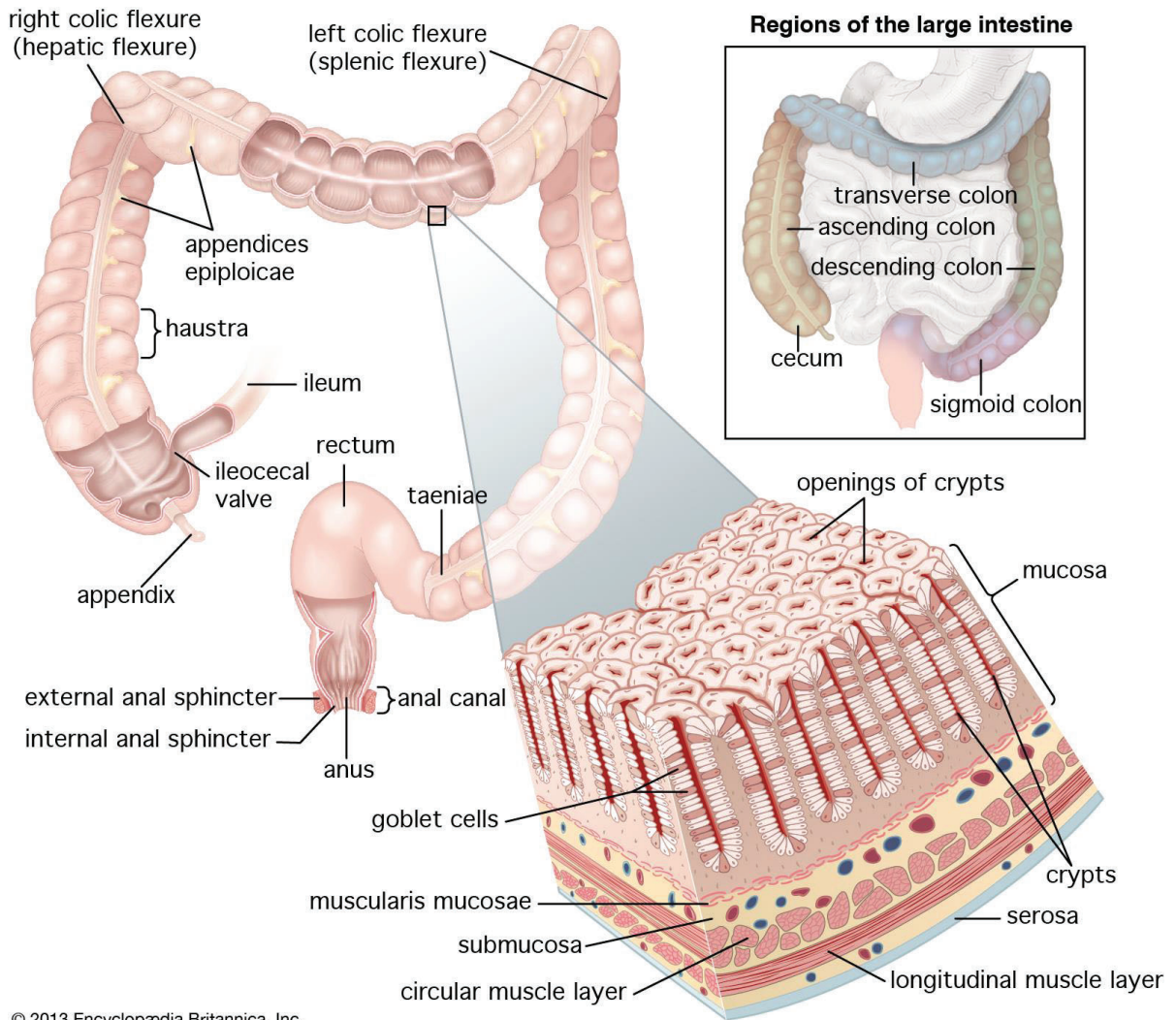
## **1.2. The human colon in health and disease**

### **1.2.1. Anatomy and main functions**

The colon, also known as large intestine or large bowel, is an organ of the digestive system located in the lower gastrointestinal tract. The main function of the colon is to reabsorb water, some nutrients and electrolytes from partially digested food, generating a solid waste (*i.e.* stool) (24). Its approximate length is 1.50 meters and

is composed of four regions, *i.e.* the cecum and ascending colon, the transverse colon, the descending colon, and the sigmoid colon (25) (see a colon schematic in **Figure 8**). Anatomically the colon can be also divided into a two-region model comprising the right and left colon. The right colon consists of the cecum, ascending colon, hepatic flexure and the right half of the transverse colon; and the left colon consists of the left half of the transverse colon, splenic flexure, descending colon, and sigmoid colon. Additional categories of colon anatomy include the proximal colon, which refers to the cecum, ascending and transverse colon; and the distal colon, which includes the descending and the sigmoid colon (26).

The wall of the colon is composed of the following tissue layers: mucosa (including superficial mucosa, also known as epithelium, lamina propria and muscularis mucosa layers), submucosa, muscularis propria (including circular and longitudinal muscular layers), subserosa and serosa (see **Figure 8**).

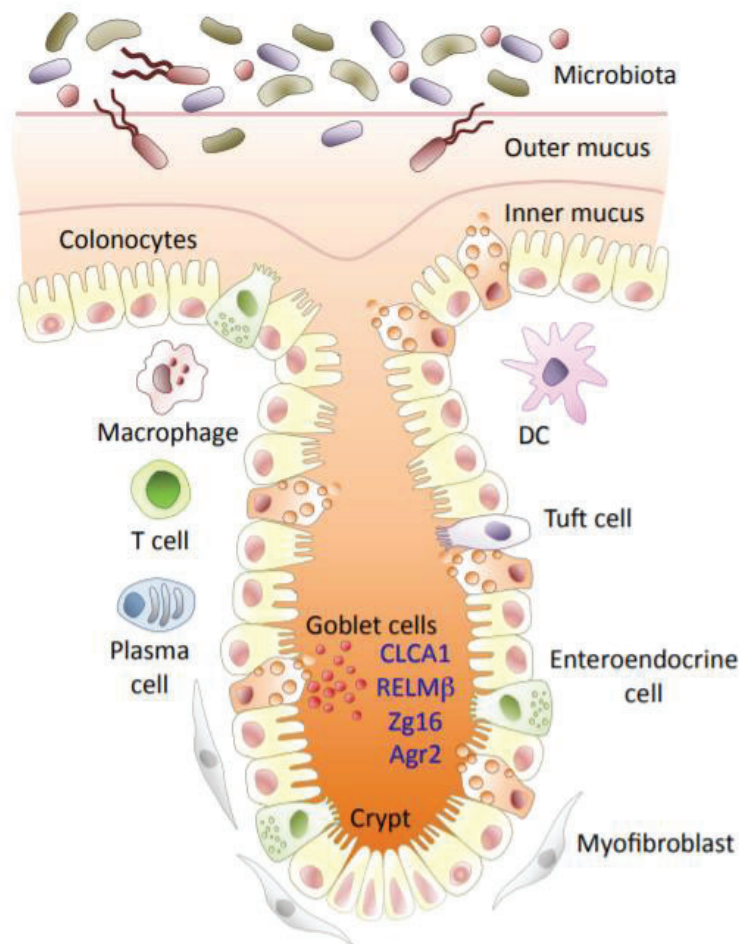


© 2013 Encyclopædia Britannica, Inc.

**Figure 8. The human colon anatomy.** This figure illustrates the main anatomical parts of the human colon, i.e. cecum, ascending, traverse, descending and sigmoid colon; along with a cross-section depicting each layer of the colonic wall. Reprinted from Encyclopedia Britannica (27).

The colon mucosa mainly consists of a single layer of epithelial cells, *i.e.* colonocytes, which are joined together by tight and adherens junctions and form a contiguous and selectively permeable membrane, which is crucial to ensure that the contents of the intestinal lumen are not drained (28). At the base of the mucosa lie the colonic crypts, which are invaginations that greatly increase the total surface area of the colonic epithelium, augmenting its potential to absorb water. Along with colonocytes, goblet cells are the major cell types of the colonic crypts. Goblet cells produce mucins, such as MUC2 mucin, and other large glycosylated proteins, which compose the outer and inner mucus layers (29) (see **Figure 9**). The mucus is a key

protective barrier, as it represents the first line of defense against bacteria (30). Additional cell types that compose the colonic mucosa include proliferative cells (*i.e.* stem cells), immune-related cells, such as Tuft cells, and enteroendocrine cells (29) (see Figure 9).



**Figure 9. Schematic of a colonic crypt.** The major cell types that compose the colonic mucosa are depicted, as well as other key components such as the mucus and the microbiota. DC, dendritic cell. Reprinted from Allaire et al. (29).

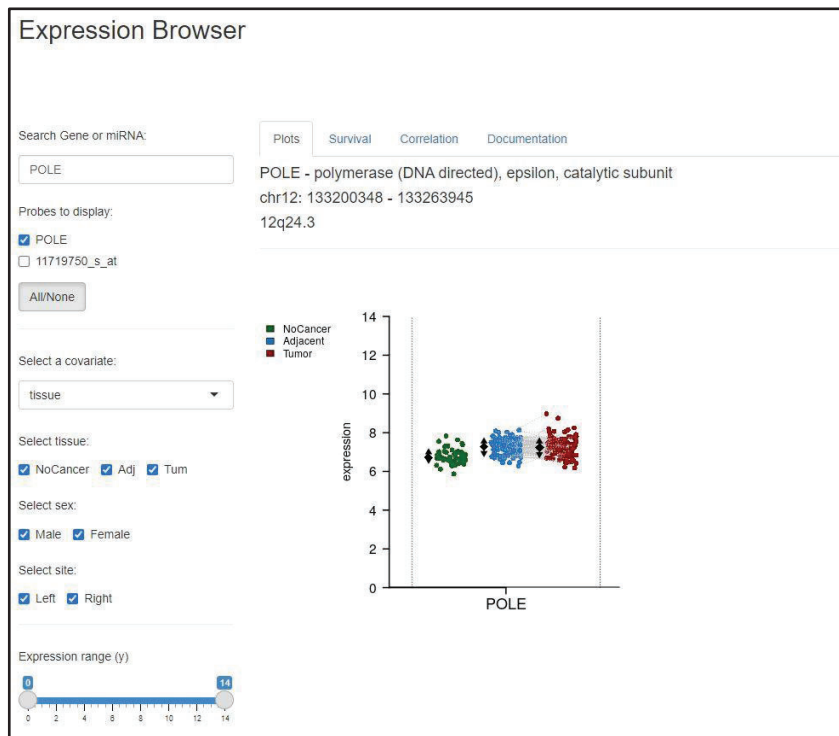
The colon mucosa is a key coordinator of mucosal immunity, and acts as a barrier to the external environment, integrating a variety of signals, including those from metabolites, microbiota and immune system (29). This selectively permeable barrier is dynamic and adapts to environmental perturbations, being able to respond appropriately to pathogens while remaining tolerant to innocuous agents like microbial metabolites and nutrients from food. Different interrelated molecular

mechanisms coordinate to maintain a homeostatic balance, which if disturbed can cause serious disease (31).

### **1.2.2. Normal colon gene expression**

Normal colon refers to a non-diseased non-neoplastic colon, with absence of macroscopic lesions, such as polyps, observed at colonoscopy (32). The characterization of gene expression variability across anatomical locations of normal colon epithelium from biopsies of healthy individuals has initially been addressed using expression microarrays technology (33,34). Specifically, a total of 154 genes were identified to be differentially expressed between the proximal and distal subsites, following a gradient of expression along the colon (33). Among them it outstands the family of homeobox genes, which were overexpressed in the proximal colon with respect to the distal colon. These genes encode transcription factors essential for controlling cell growth and differentiation, suggesting different regenerative processes of the epithelial cells according to colon location (34). Of note, expression microarrays do not provide estimates of AS events.

The Colonomics project collected normal colon biopsies from 50 healthy individuals, along with 100 normal colon biopsies adjacent/paired to colorectal tumors, and reported 29,073 eQTLs (35). In addition, gene expression in normal colon was compared with that from tumors. The Colonomics Expression Browser was developed to facilitate the access to this data as well as for exploring and visualizing colon gene expression levels (36). A screenshot of the browser is shown in **Figure 10**. Also, an eQTL Browser with this data is implemented, which includes plots and filtering and customization options (37).



**Figure 10. Screenshot of the Colonomics gene expression browser.** It shows the expression levels of the gene *POLE*. From (36).

On the other hand, the last version of the GTEx project (8) provided RNA Sequencing (RNA-Seq)-based gene expression profiles of transverse (N=368) and sigmoid (N=318) colon, based on post-mortem donors, which cause of death was different from colon-related diseases. The sample collection of these data was not homogeneous between colon locations. In the case of transverse colon, it is based on tissue from the entire colonic wall, and, in the case of sigmoid, it lacks the mucosa layer, and it is mainly represented by muscularis mucosa (8). Therefore, the corresponding gene expression profiles are not comparable between colon locations, as reported elsewhere (38,39). This project identified a total of 11,687 and 10,550 eQTLs as well as a total of 3,459 and 3,269 sQTLs for transverse and sigmoid colon, respectively (8).

### 1.2.3. Common complex diseases affecting the colon

Maintaining intestinal homeostasis is crucial for the normal functioning of the colon. The dysregulation of epithelial homeostasis, by factors such as infection, chemicals,



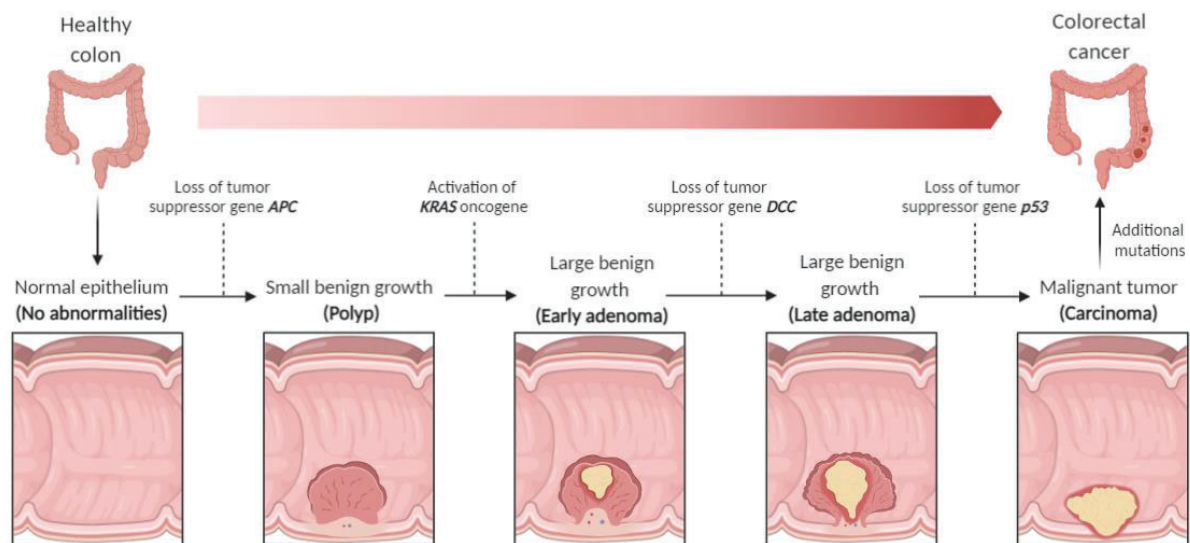
or genetics, can lead to chronic diseases. Two major colon-related diseases causing an important public health burden in developed countries are colorectal cancer (CRC) and inflammatory bowel disease (IBD) (40).

### 1.2.3.1. Colorectal cancer

CRC is the third most commonly diagnosed cancer worldwide and the second leading cause of cancer death globally (41). Its etiology is heterogeneous and differs by anatomical location and subtype. Approximately 60-65% of CRC cases are sporadic, *i.e.* occur in individuals without CRC family history or inherited risk-increasing genetic mutations. CRC is largely attributable to modifiable environmental risk factors, such as obesity, physical inactivity, nutritionally poor diets and smoking habit, which makes it more prevalent in westernized countries (42).

There have been described different CRC carcinogenic pathways, including the adenoma-carcinoma, the serrated and the inflammatory pathways. The most common is the adenoma-carcinoma pathway, in which there is a progressive accumulation of (epi)genetic alterations that drive the transformation of normal cells to a polyp, to an early adenoma, to a late adenoma and, finally, to a carcinoma. This process is depicted in **Figure 11**. Crucial genetic alterations are inactivating mutations in the tumor suppressor gene *APC*, which overactivates the Wnt/ $\beta$ -catenin signaling pathway, that provokes cell proliferation. Subsequently the oncogene *KRAS* acquires mutations, promoting the growth of the adenoma. Also, the inactivation of the tumour suppressor gene *TP53* contributes to the progression to CRC (42) (see **Figure 11**). Next, the serrated pathway is highlighted by the progression from normal cells to hyperplastic polyp, to sessile serrated adenoma and to CRC. It is characterised by mutations of the oncogene *BRAF*, the activation of the MAPK pathway, and CpG island methylator phenotype (CIMP) positivity (42). Finally, the inflammation-associated carcinogenic pathway remains the less frequent and appears particularly in ulcerative colitis patients. Precursor genetic

alterations are not prevalent in this pathway, and mutations in *APC* and *TP53* occur early and late in the development of carcinogenesis, respectively (42).



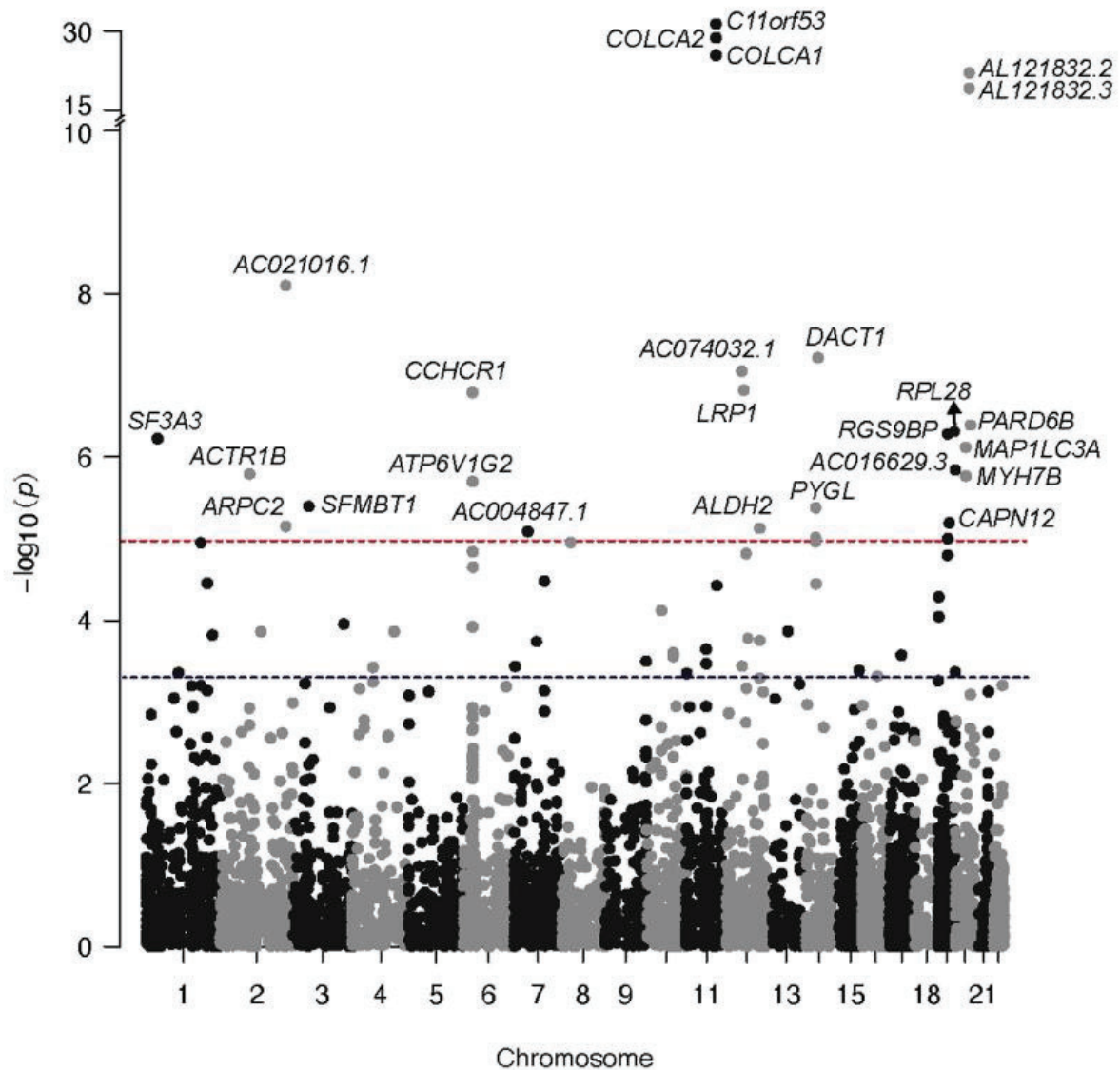
**Figure 11. Overview of colorectal cancer (CRC) development.** Schematic depicting the key stages from healthy colon to CRC of the adenoma-carcinoma pathway. Adapted from “The Multi-Hit Model of Colorectal Cancer”, by BioRender.com (2021). Retrieved from <https://app.biorender.com/biorender-templates>.

The germline genetic architecture of CRC has been addressed in GWAS, mostly based on subjects of European descent (43,44). CRC has a strong heritable basis, with an estimated SNP-based heritability of 29% (95% confidence interval [95% CI]: 24%-35%) (44). Specifically, a total of 141 independent SNPs, genome-wide distributed, have been proposed to affect CRC risk (45). Most variants lie in non-coding genomic regions influencing gene regulation and are enriched in active regulatory regions identified in colon tissue, such as enhancers (44). Also, substantial genetic susceptibility heterogeneity has been defined between proximal and distal colorectal tumors, specifically, 48 risk *loci* show tumor-location specificity (46).

Candidate genes have been proposed at several GWAS-identified risk *loci*, mainly based on functional annotation and colocalization with colon and blood eQTLs, and other lines of evidence that incorporate epigenomics data (e.g. chromatin-chromatin interaction and DNA accessibility data). In addition, a TWAS was



implemented for CRC, including the prediction of gene expression in the transverse colon of a total of 125,478 subjects (58,131 CRC cases). This study identified 25 genes associated with CRC, which were located both within and outside GWAS-identified CRC risk *loci* (47). The symbols and the distribution of these genes by chromosome are depicted in a Manhattan plot in **Figure 12**.



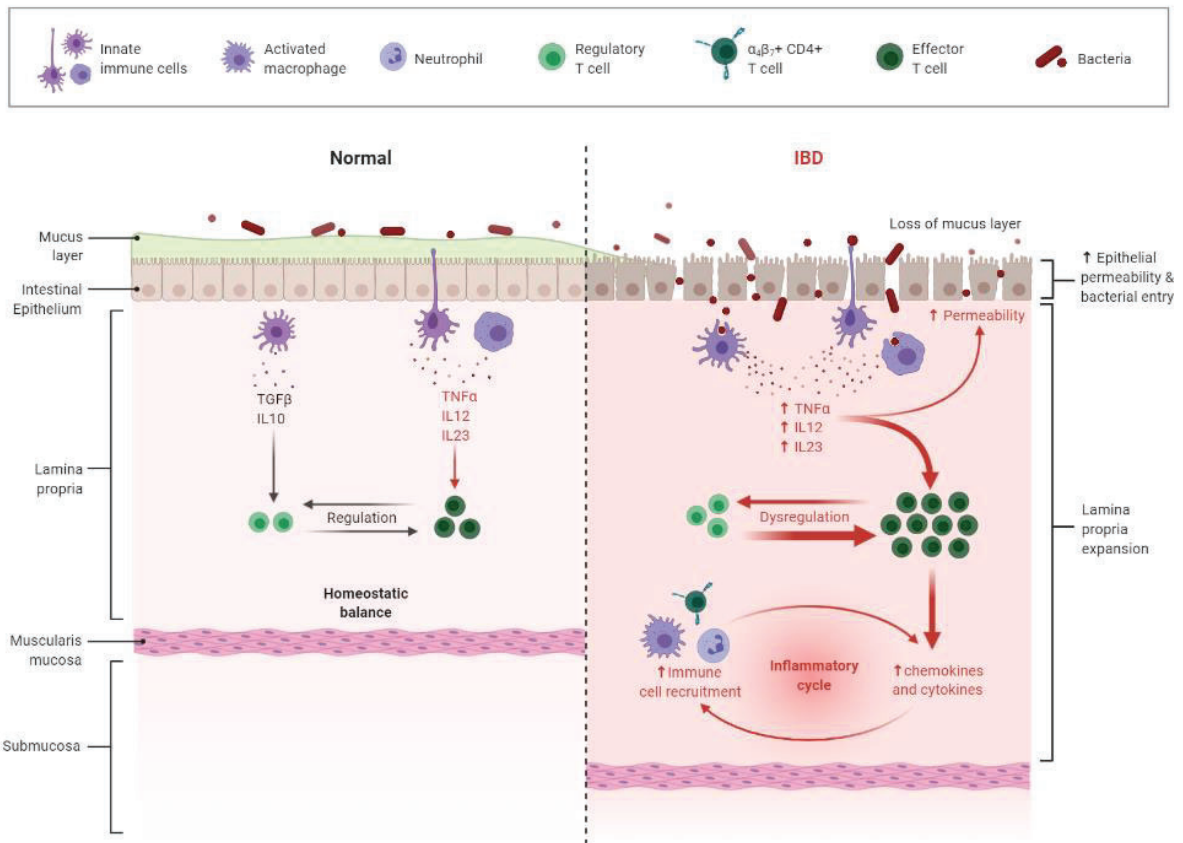
**Figure 12. Manhattan plot of the TWAS results for CRC.** The blue and red lines represent a false discovery rate (FDR)-corrected significance level of  $P < 6.6 \times 10^{-4}$  and a Bonferroni corrected threshold of  $P < 9.1 \times 10^{-6}$ , respectively. Reprinted from Guo *et al.* (47).

### 1.2.3.2. Inflammatory bowel disease

IBD is a chronic inflammatory disease that affects the gastrointestinal tract. Its prevalence has been rising, becoming a global disease with a rapidly increasing incidence in newly westernized societies. Environmental risk factors include antibiotic usage and smoking (48).

IBD is characterized by a dysregulated immune response, whose exacerbated effect causes serious damage of the intestinal epithelium. Although the complete landscape of molecular mechanisms that drive disease pathogenesis is not fully elucidated, molecular pathways that maintain the mucosal immunity homeostasis have been described as key pathways (49). Indeed, current IBD therapeutic agents are mostly limited to block the mediators of inflammation (50).

Many Inflammatory molecules have been implicated in the IBD pathogenesis (51). Important players in the dysregulation of immune response in IBD are the Tumor Necrosis Factor (TNF) and the interleukins (IL), such as IL-12 and IL-23, whose overexpression contributes to the dysfunction of the adaptive immune system and the increased permeability of the intestinal mucosa. This mucosal impairment facilitates the infiltration of bacteria to deeper layers of the intestinal wall, causing the overactivation of inflammatory processes mediated by innate immune cells, such as neutrophils and activated macrophages. This process, in turn, causes the further release of pro-inflammatory cytokines, driving a cycle of inflammation and intestinal damage that ultimately leads to epithelial cell death (51) (see **Figure 13**).



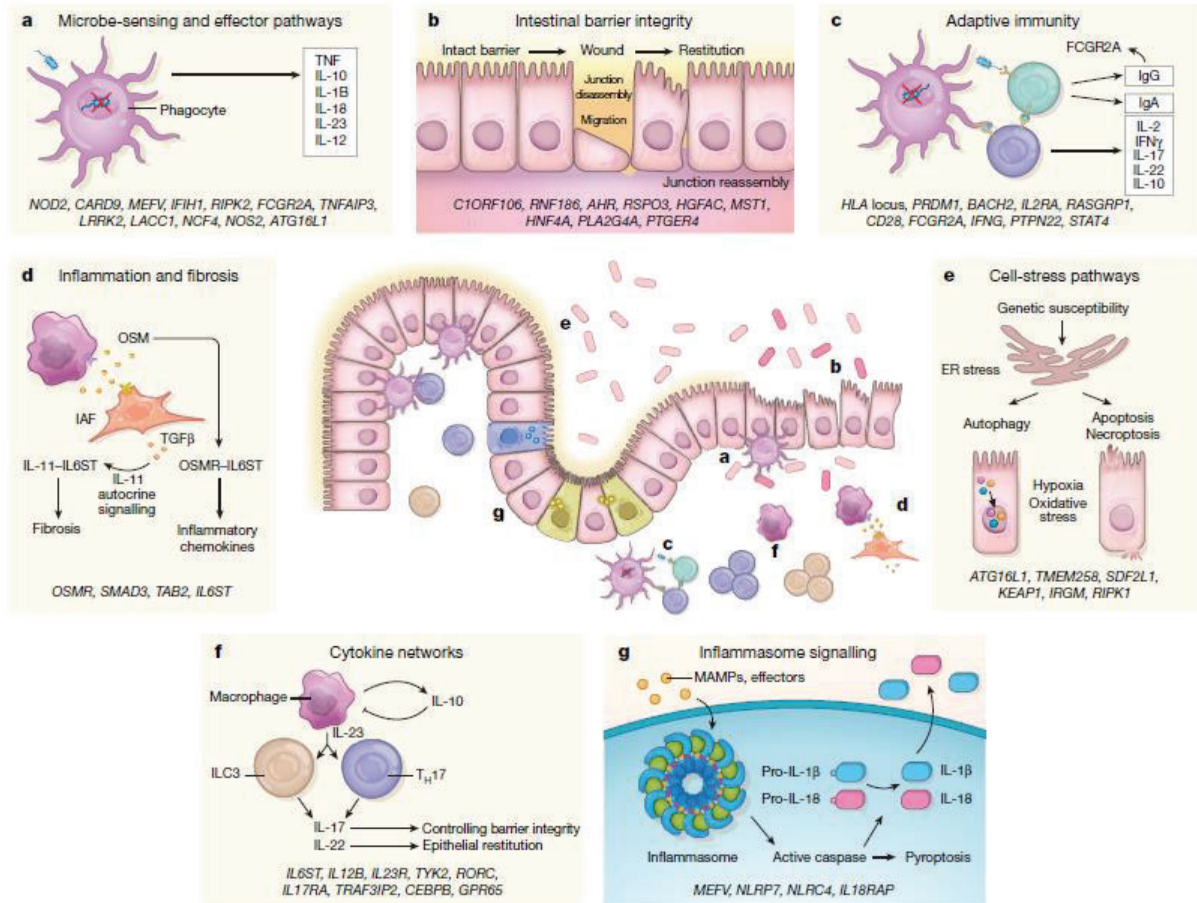
**Figure 13. Dysregulation of immune response in inflammatory bowel diseases (IBD).** Key features of immune dysregulation that occur in IBD pathogenesis are depicted. Reprinted from “Immune Response in IBD”, by BioRender.com (2021). Retrieved from <https://app.biorender.com/biorender-templates>.

IBD encompasses two similar disease subtypes, which are Crohn’s disease (CD) and ulcerative colitis (UC). Different characteristics and clinical manifestations differentiate these two subtypes, being CD more heterogeneous (40). For example, CD can comprise multiple separate areas of inflammation, and can damage all layers of the intestinal wall, forming deep perforations. In contrast, UC forms a continuous patch of inflammation, and damages the innermost lining of the intestinal wall (52). Regarding the disease localization, while UC is restricted to the colon and rectum, CD can affect any part of the gastrointestinal tract. Specifically, CD mainly affects the terminal ileum of the small intestine, and it affects the colon in only 25% of cases. Different features distinguish ileal from colonic CD, including pathophysiological and genetic factors. For example, there is higher neutrophil

activity in colonic than ileal CD. Understanding these disease location differences can translate into more individualised therapies (53).

Genetic risk factors have been identified to be associated with IBD by GWAS that include nearly 60 thousand subjects, including 25 thousand IBD patients (54). These studies have identified a total of 241 independent risk SNPs. The estimation of the SNP-based heritability for IBD is 14% (95% CI: 12%-16%), 20% for CD (95% CI: 16%-24%), and 13% for UC (95% CI: 10%-15%) (14,54). Putative effector genes associated with these variants have been proposed by eQTL evidence (54), as well as by gene expression network-based approaches in intestinal and immune cell types relevant for IBD (55,56). Many genetically regulated genes have been mapped to key molecular pathways that drive IBD, including cell-stress and integrity intestinal barrier-related pathways, in addition to immune related pathways (see **Figure 14**) (49). However, the complete picture of the molecular mechanisms of IBD susceptibility and the implicated genes are not completely ascertained.

Finally, although there are some TWAS that include IBD (57–59), it has not been conducted a TWAS that focuses entirely on IBD and its two main subtypes, and that leverages the full potential of the datasets currently available from relevant tissues and immune cell types.



**Figure 14. Molecular pathways driving IBD.** This figure shows the key pathways that drive IBD and the risk genes that have been mapped to each of them (indicated at the bottom of each panel). Reprinted from Graham DB *et al.* (49).

## 2. HYPOTHESES

Colon gene expression and AS profiles derived from RNA sequencing of mucosal biopsies may provide good estimates of the colon tissue transcriptome. The generation of these new data from more than 400 healthy living individuals may represent a large reference dataset of normal colon, given the sample size, specimen collection procedures and sequencing technology used.

Gene expression and AS profiles may vary across the colon anatomy. The sample collection procedures of these newly generated data, in contrast to the currently available population-based colon gene expression data, may provide comparable profiles across the colon.

Colon gene expression and AS can be controlled by cis-regulatory processes involving common germline genetic variation (*i.e.* SNPs) physically located closely to the corresponding gene. The association parameters between SNP genotypes and gene expression levels can be estimated.

Genetically regulated gene expression in the colon may play a role in the genetic susceptibility to complex traits and diseases, including not only those directly affecting the colon, such as CRC or IBD, but also others indirectly related with it, such as those affected by molecular processes taking place in the framework of the gut-brain axis. In silico approaches such as QTL-heritability enrichment and colocalization could be employed to identify diseases in which the colon physiology is playing a role and to nominate candidate susceptibility genes for these diseases, respectively.

An interactive web-based resource may facilitate researchers a quick and centralized access and exploration of population-based colon gene expression-related data.

Gene expression across tissues and cell subtypes can be predicted from genotype data in a large cohort of IBD cases and controls. Different levels of predicted gene

expression between IBD cases and controls can be measured and may reveal new candidate IBD susceptibility genes. These genes may be tissue-specific and different according to colon location and tissue category. Also, they may participate in pathways related to approved IBD therapies, such as tumor necrosis factor signaling. Also, candidate genes may not only provide functional insight in the molecular processes underlying disease pathogenesis, but also may guide further research on new targeted therapeutics.



### 3. OBJECTIVES

1. To provide reference profiles for transcriptome-wide gene expression and alternative splicing of colon mucosal biopsies from healthy adults.
  - 1.1. To describe the differences of these profiles across colon anatomic subsites (ascending, transverse, and descending colon).
  - 1.2. To provide the associations of these profiles with SNPs (*i.e.* e/sQTLs).
  - 1.3. To identify complex traits and diseases whose SNP-based heritability is partly explained by the identified e/sQTLs, and propose candidate susceptibility genes for these phenotypes.
2. To develop a web resource to explore population-based normal colon transcriptomic profiles, e/sQTLs, gene expression prediction models, as well as to annotate SNPs with colon eQTLs.
3. To propose candidate genes whose genetically regulated gene expression in specific tissues and cell subtypes is associated with inflammatory bowel disease, Crohn's disease, and ulcerative colitis status, separately.
  - 3.1. To identify candidate susceptibility genes in specific colon subsites, with emphasis on gene expression markers of specific cell types.
  - 3.2. To identify regulatory and signaling molecular pathways in which the candidate susceptibility genes are enriched, including IBD therapy-related pathways.
  - 3.3. To find candidate susceptibility genes specific for the epithelial, immune/blood, mesenchymal and neural tissue categories.





## **4. MATERIALS AND METHODS AND RESULTS**

### **4.1. BarcUVa-Seq normal colon e/sQTLs.**

The first objective of the Thesis was “to provide reference profiles for transcriptome-wide gene expression and alternative splicing of colon mucosal biopsies from healthy adults, as well as their differences across colon location and corresponding e/sQTLs. Also, to identify complex traits and diseases whose SNP-based heritability is enriched in the identified e/sQTLs, and propose candidate susceptibility genes for these phenotypes”.

To address this objective, we developed the article entitled “Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci”.

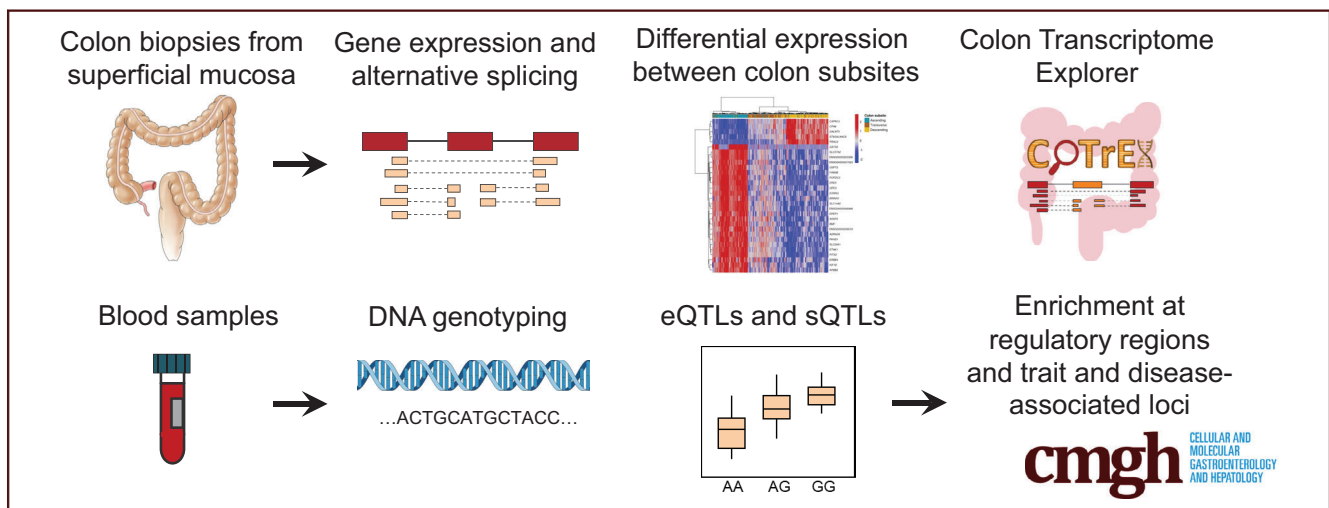
## ORIGINAL RESEARCH

## Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci



Virginia Díez-Obrero,<sup>1,2,3,4</sup> Christopher H. Dampier,<sup>5,6,7</sup> Ferran Moratalla-Navarro,<sup>1,3,4</sup> Matthew Devall,<sup>5,6</sup> Sarah J. Plummer,<sup>5,6</sup> Anna Díez-Villanueva,<sup>1,2,3</sup> Ulrike Peters,<sup>8,9</sup> Stephanie Bien,<sup>8,9</sup> Jeroen R. Huyghe,<sup>8,9</sup> Anshul Kundaje,<sup>10</sup> Gemma Ibáñez-Sanz,<sup>1,2,3,11</sup> Elisabeth Guinó,<sup>1,2,3</sup> Mireia Obón-Santacana,<sup>1,2,3</sup> Robert Carreras-Torres,<sup>1,2,3</sup> Graham Casey,<sup>5,6</sup> and Víctor Moreno<sup>1,2,3,4</sup>

<sup>1</sup>Oncology Data Analytics Program, Catalan Institute of Oncology, L'Hospitalet de Llobregat, Barcelona; <sup>2</sup>Colorectal Cancer Group, Molecular Mechanisms and Experimental Therapy in Oncology (ONCOBELL) Program, Bellvitge Biomedical Research Institute; <sup>3</sup>Consortium for Biomedical Research in Epidemiology and Public Health, Madrid, Spain; <sup>4</sup>Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain; <sup>5</sup>Center for Public Health Genomics, University of Virginia; <sup>6</sup>Department of Public Health Sciences, University of Virginia; <sup>7</sup>Department of Surgery, University of Virginia, Charlottesville, Virginia; <sup>8</sup>Epidemiology Department, University of Washington, Seattle, Washington; <sup>9</sup>Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington; <sup>10</sup>Department of Genetics, Stanford University, Stanford, California; <sup>11</sup>Gastroenterology Department, Bellvitge University Hospital, L'Hospitalet de Llobregat, Barcelona, Spain



## SUMMARY

We profiled gene expression and alternative splicing of non-neoplastic colon from biopsy specimens from 445 healthy individuals. We showed that single-nucleotide polymorphisms associated with these profiles are enriched in disease-associated loci, including colorectal cancer and inflammatory bowel disease.

**BACKGROUND & AIMS:** The association of genetic variation with tissue-specific gene expression and alternative splicing guides functional characterization of complex trait-associated loci and may suggest novel genes implicated in disease. Here, our aims were as follows: (1) to generate reference profiles of colon mucosa gene expression and alternative splicing and compare them across colon subsites (ascending, transverse, and descending), (2) to identify expression and splicing

quantitative trait loci (QTLs), (3) to find traits for which identified QTLs contribute to single-nucleotide polymorphism (SNP)-based heritability, (4) to propose candidate effector genes, and (5) to provide a web-based visualization resource.

**METHODS:** We collected colonic mucosal biopsy specimens from 485 healthy adults and performed bulk RNA sequencing. We performed genome-wide SNP genotyping from blood leukocytes. Statistical approaches and bioinformatics software were used for QTL identification and downstream analyses.

**RESULTS:** We provided a complete quantification of gene expression and alternative splicing across colon subsites and described their differences. We identified thousands of expression and splicing QTLs and defined their enrichment at genome-wide regulatory regions. We found that part of the SNP-based heritability of diseases affecting colon tissue, such as colorectal cancer and inflammatory bowel disease, but also of diseases affecting other tissues, such as psychiatric conditions,

can be explained by the identified QTLs. We provided candidate effector genes for multiple phenotypes. Finally, we provided the Colon Transcriptome Explorer web application.

**CONCLUSIONS:** We provide a large characterization of gene expression and splicing across colon subsites. Our findings provide greater etiologic insight into complex traits and diseases influenced by transcriptomic changes in colon tissue. (*Cell Mol Gastroenterol Hepatol* 2021;12:181–197; <https://doi.org/10.1016/j.jcmgh.2021.02.003>)

**Keywords:** Gene Expression; Alternative Splicing; QTLs; Colon.

Transcriptome-wide gene expression profiles of normal colon tissue have been assessed in population-based studies, using data sets with a range of different characteristics, including variable colon anatomic subsites, collection methods, sample sizes, sequencing technologies, and data processing methods.<sup>1–8</sup> A large public transcriptome data set for non-neoplastic colon tissue from the Genotype-Tissue Expression (GTEx) project included samples collected from the transverse and sigmoid colon of post-mortem subjects and included both mucosa and muscularis propria.<sup>8</sup> In most studies, the transcriptome is assessed in terms of gene expression, however, a comprehensive characterization of alternative splicing (AS) has not been performed in normal colon epithelial tissue derived from living individuals.

AS is a post-transcriptional regulatory mechanism by which multiple messenger RNA transcripts are produced from a single locus, enabling enlargement of cellular functions.<sup>9</sup> More than 90% of human genes have the potential to undergo AS.<sup>10</sup> Common AS patterns include exon skipping, alternative 5' and 3' splice sites, mutually exclusive exons, intron retention, and alternative first or last exons.<sup>11</sup> Based on these predefined patterns and transcript expression levels, different AS events and their relative abundances can be identified for a given gene.<sup>12</sup> In addition, by measuring alternative excision of introns, novel and more complex alternative splicing events can be identified.<sup>13</sup> AS has been assessed in multiple tissue types across several large cohorts, including healthy<sup>8</sup> and pathologic tissues,<sup>14–16</sup> allowing the association of particular AS events with phenotypes such as age<sup>17</sup> and cancer type.<sup>14–16</sup> In colon tissue, AS events have been measured in adenocarcinomas and paired adjacent normal tissue and have been associated with colorectal cancer (CRC) anatomic location<sup>18</sup> and prognosis.<sup>18–20</sup>

Single-nucleotide polymorphisms (SNPs) have been associated with gene expression (ie, expression quantitative trait loci [eQTLs]) and AS (sQTLs), and increasingly are identified in studies of both normal<sup>8,21–25</sup> and malignant tissues.<sup>26</sup> Such associations can indicate the functional effects of SNPs at genetic risk loci, help prioritize SNPs and genes for functional assays, serve as prognostic biomarkers, and suggest disease mechanisms.<sup>10,26,27</sup> In the case of normal colon tissue, eQTL data sets have been generated,<sup>1–8</sup> but there is no information about sQTLs derived from living individuals.

In this study, we analyzed a novel RNA sequencing (RNA-Seq) data set of normal colon tissue biopsy specimens including colon anatomic subsites not investigated previously (ascending, transverse, and descending). Our data set is representative of the transcriptome of colon epithelial cells of living subjects because all biopsy specimens were collected from mucosa at colonoscopy. This characteristic makes it optimal for investigating the normal physiology across the colon, and it is relevant not only for studying the etiologic aspects of diseases affecting this tissue, such as CRC, but also for diseases affecting other tissues, such as those that imply epithelial–neuronal communication<sup>28</sup> and those affected by perturbations of intestinal permeability.<sup>29</sup>


The aims of this study were as follows: (1) to provide a reference transcriptomic data set for normal colon epithelium by profiling gene expression and AS, (2) to identify SNPs associated with variation in gene expression and AS (ie, QTLs), (3) to list traits for which identified QTLs contribute to SNP-based heritability, (4) to prioritize candidate effector genes, and (5) to provide a web-based resource to visualize the expression profiles and QTLs.

## Results

The University of Barcelona and University of Virginia genotyping and RNA Sequencing Project: A Novel Reference Data Set for Colon Tissue Transcriptome Analysis

The University of Barcelona and University of Virginia genotyping and RNA sequencing project (BarcUVA-Seq) cross-sectional study included 485 adult volunteers found to have an endoscopically healthy colon (ie, a normal colon without polyps or other lesions) from whom we collected superficial colon biopsy specimens and blood samples. Bulk RNA was isolated from biopsy samples and sequenced in several batches. Subjects were genotyped using the Illumina (San Diego, CA) OncoArray 500K beadchip,<sup>30</sup> and genome-wide SNPs were imputed. After filtering the data to select for individuals with high-quality RNA-Seq and genotype samples (see the Materials and Methods section), we included data from 445 individuals, among whom 283 were female (64%). Biopsy specimens were obtained from sites along the ascending (n = 138; 31%), transverse (n = 143; 32%), and descending (n = 164; 37%) colon (Table 1). We profiled gene expression and alternative splicing and identified cis-acting eQTLs and sQTLs (see the Materials and Methods section).

**Abbreviations used in this paper:** AS, alternative splicing; BarcUVA-Seq, University of Barcelona and University of Virginia genotyping and RNA sequencing project; CoTrEx, colon transcriptome explorer; CRC, colorectal cancer; eGene, eQTL gene; eQTL, expression quantitative trait locus; eSNP, eQTL SNP; FDR, false-discovery rate; FWER, family-wise error rate; GTEx, Genotype-Tissue Expression project; GWAS, genome-wide association study; LD, linkage disequilibrium; MAF, minor allele frequency; PSI, percent splicing index; RBP, RNA-binding proteins; RNA-Seq, RNA sequencing; sGene, sQTL gene; SNP, single-nucleotide polymorphism; sSNP, sQTL SNP; sQTL, splicing quantitative trait locus; TSS, transcription start site.

 Most current article

© 2021 The Authors. Published by Elsevier Inc. on behalf of the AGA Institute. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

2352-345X

<https://doi.org/10.1016/j.jcmgh.2021.02.003>

**Table 1.** BarcUVa-Seq Data Set Descriptive

|   |               |
|---|---------------|
| Total individuals, N  | 445           |
| Sex, n (%)  |               |
| Female  | 283 (63.6)    |
| Male  | 162 (36.4)    |
| Age, y, means $\pm$ SD  | 60 $\pm$ 7.44 |
| Colon anatomic location overall<br>and stratified by sex, n (%) |               |
| Ascending (right)   | 138 (31.0)    |
| Female  | 86 (62.3)     |
| Male  | 52 (37.7)     |
| Transverse  | 143 (32.1)    |
| Female  | 90 (62.9)     |
| Male  | 53 (37.1)     |
| Descending (left)   | 164 (36.9)    |
| Female  | 107 (65.2)    |
| Male  | 57 (34.8)     |

### Gene Expression and Alternative Splicing

Expression was analyzed based on GENCODE (E;BL-EBI, Hinxton, UK) release 19 annotations.<sup>31</sup> After filtering out features with low or no expression, 21,281 genes and 104,769 transcripts remained (see the Materials and Methods section). Gene and transcript abundances of interest can be visualized online (see the Colon Transcriptome Explorer [CoTrEx] section). We considered 13,243 AS events in 6178 genes after applying filters (see AS events annotations in [Supplementary Table 1](#)). We categorized AS events as follows: alternative first exons (30%), exon skipping (24%), alternative 3' splice-site (12%), alternative 5' splice-site (12%), intron retention (10%), alternative last exons (10%), and mutually exclusive exons (1%) ([Figure 1](#), [Table 2](#)). Most genes had AS events from 1 or 2 categories, and few had AS events from up to 6 categories. In addition, as a complementary AS metric, we computed the abundances of 269,586 alternatively excised introns that were grouped in 73,313 clusters. Some introns (23%) were novel and 77% were annotated in 15,912 genes. We filtered introns by low expression or low complexity and considered only 42,808 intron clusters annotated in 8953 genes for sQTL analysis (see the Materials and Methods section).

### Transcriptomic Profiles Differ Between Colon Subsites

We aimed to identify genes and splicing features that were expressed differentially across colon subsites, situating the transverse colon as an intermediate phenotype (see the Materials and Methods section). Overall, 4430 genes were expressed differentially between ascending, transverse, and descending subsites (family-wise error rate [FWER],  $\leq 0.05$ ), with absolute log fold changes of up to 3.7 ([Figure 2A](#)). Hierarchical clustering of the top 30 genes with the smallest FWER showed the transverse colon clustered with descending colon ([Figure 2B](#)). Full differential gene expression results are listed in [Supplementary Table 2](#). Next, we tested whether genes expressed differentially across subsites were enriched for features in a wide array of curated gene sets, signatures, functional pathways, and

ontologies. We found enrichment in a gene set associated with normal colon tissue transformation into adenoma, in pathways involved in drug metabolism, and in other biological processes such as antimicrobial humoral response. Full enrichment results are listed in [Supplementary Table 3](#). For splicing, we found 236 genes with different relative abundances of AS events (false-discovery rate [FDR],  $\leq 0.05$ ) ([Supplementary Tables 4 and 5](#)) and 280 genes with different relative abundances of excised introns between the ascending and descending colon (FDR,  $\leq 0.05$ ) ([Supplementary Table 6](#)).

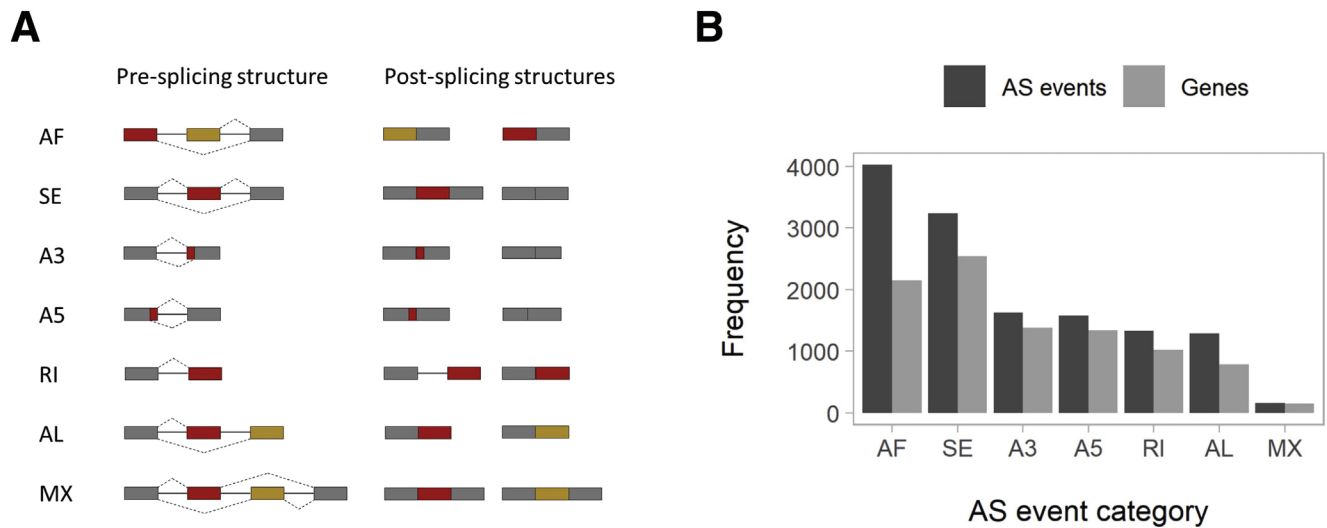
### Identification of eQTLs and sQTLs

We identified 11,739 eQTLs (Q value  $\leq 0.05$ ) including 11,427 unique SNPs (eSNPs) associated with the expression of 11,739 genes (eGenes) ([Supplementary Table 7](#)). Most eSNPs were associated with a single eGene, but we found eSNPs associated with up to 6 eGenes. Neither the location of the eSNPs relative to the gene transcription start site (TSS) nor the allele frequency were associated with the eSNP effect ([Figure 3](#)). eQTLs can be explored on the CoTrEx web application (see the Colon Transcriptome Explorer section). Full eQTL summary statistics are publicly available (see the Data availability statement). In addition, we performed eQTL interaction analysis for colon subsites (ascending vs descending) and found 26 eQTLs with a Q value of 0.05 or less ([Supplementary Table 8](#)). The eQTL rs6684275-*RIMKLA* showed an inverse association in the ascending colon compared with the descending colon ([Figure 4](#)).

Next, we mapped 1125 sQTLs (Q value  $\leq 0.05$ ) including 1122 unique SNPs (sSNPs) associated with 1125 genes (sGenes) ([Supplementary Table 9](#)). The proportions of AS categories among SNP-associated AS events were similar to those found for total AS events ([Table 2](#)). Although we found 82% of sGenes among eGenes, only 8% of sGenes shared the same genetic variants with eGenes (6%) or harbored variants in high linkage disequilibrium ( $LD R^2 > 0.8$ ) with eSNPs (2%) ([Figure 5A](#)). In addition, we identified an additional set of 1062 sQTLs (Q value  $\leq 0.05$ ) of 1058 sSNPs associated with clusters of excised introns in 1062 genes ([Supplementary Table 10](#)) and observed that 40% of these sGenes were in common with sGenes associated with AS events. sQTLs can be explored on the CoTrEx web application (see Colon Transcriptome Explorer section), and full summary statistics are publicly available (see Data availability statement).

### Replication and Meta-Analysis With GTEx

We performed replication and meta-analyses using data from the GTEx project v8.<sup>8</sup> For replication analysis, we used samples from the sigmoid and transverse colon ( $n = 318$  and  $n = 368$ , respectively). For the replication of eQTLs, we downloaded the list of GTEx eQTLs (see the Materials and Methods section). For the replication of sQTLs we used GTEx transcript expression data for computing AS events as well as SNPs for computing sQTLs using the same approach considered for BarcUVa-Seq data ([Supplementary Tables 11](#)



**Figure 1. Alternative splicing events.** (A) Scheme of gene and alternatively spliced transcripts structure in 7 AS categories: alternative first exons (AF), exon skipping (SE), alternative 3' splice-site (A3), alternative 5' splice-site (A5), intron retention (RI), alternative last exons (AL), and mutually exclusive exons (MX). Constitutive exons (ie, those maintained in all processed transcripts after splicing) are shown in gray. Exons in red or gold alternatively are present in processed transcripts after splicing. *Dashed line* indicates different splicing processing for a gene. (B) Frequency of AS events and genes by AS category. One gene can be processed according to different AS categories.

and 12). We explored the  $P$  value distributions between BarcUVa-Seq and GTEx colon data sets and computed the  $\pi_1$  statistic<sup>32</sup> (Figure 6). For eQTLs, a higher replication value ( $\pi_1 = 0.76$ ) was obtained for GTEx transverse colon than for sigmoid colon ( $\pi_1 = 0.56$ ). For sQTLs the same replication statistic was obtained for both GTEx colon tissue data sets ( $\pi_1 = 0.67$ ).

We performed a meta-analysis of BarcUVa-Seq eQTLs with the full GTEx v8 data set ( $n = 49$  tissues) using a multivariate adaptive shrinkage approach.<sup>33</sup> Hierarchical clustering of pairwise correlations on the resulting effect sizes showed that BarcUVa-Seq eQTLs from colonic mucosa clustered with GTEx eQTLs from transverse colon and terminal ileum (Figure 7A). The correlations between BarcUVa-Seq eQTL effect sizes and all GTEx tissues showed that transverse colon, terminal ileum, stomach, minor salivary

gland, and kidney cortex are the GTEx tissues with highest correlation ( $\rho > 0.7$ ) (Figure 7B).

### Annotation and Functional Enrichment Analyses

We observed eSNPs and sSNPs distributed in patterns similar to each other across the following genomic regions: introns, intergenic regions, upstream and downstream gene regions, 3' and 5' untranslated regions and splice regions (including donor and acceptor variants). Intronic variants were the most common from both types of SNPs. Intergenic and upstream regions harbored higher proportions of eSNPs than sSNPs, and splice and untranslated regions harbored higher proportions of sSNPs than eSNPs (Figure 5B). Functional consequences also were assessed: most SNPs were not classified, but a small proportion of SNPs were

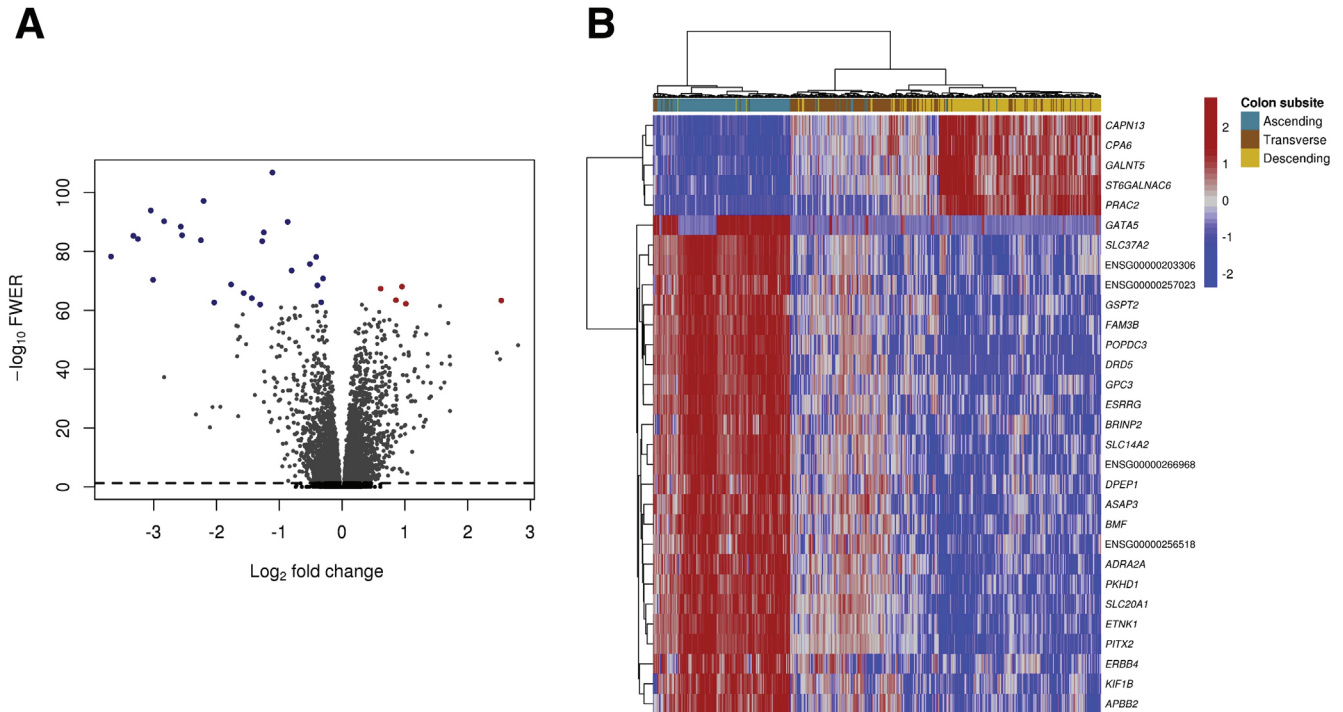
**Table 2.** Description of AS Events and Genes by AS category

| Event category | Total AS events, n (%) | Total genes, n (%) | AS events associated with sSNPs, n (%) |
|----------------|------------------------|--------------------|--|
| SE             | 3235 (24.43)           | 2542 (41.20)       | 316 (28.1)                             |
| AF             | 4023 (30.38)           | 2146 (34.78)       | 253 (22.5)                             |
| A3             | 1627 (12.29)           | 1378 (22.33)       | 140 (12.4)                             |
| A5             | 1579 (11.92)           | 1344 (21.78)       | 148 (13.2)                             |
| RI             | 1327 (10.02)           | 1022 (16.56)       | 126 (11.2)                             |
| AL             | 1292 (9.76)            | 785 (12.72)        | 259 (11.5)                             |
| MX             | 160 (1.21)             | 148 (2.40)         | 12 (1.1)                               |
| Overall        | 13,243 (100.00)        | 6170 (100.00)      | 1125 (100.0)                           |

NOTE. A given gene can have AS events from up to 6 categories.

AF, alternative first exons; AL, alternative last exon; A3, alternative 3' splice-site; A5, alternative 5' splice-site; RI, intron retention; MX, mutually exclusive exons; SE, exon skipping.



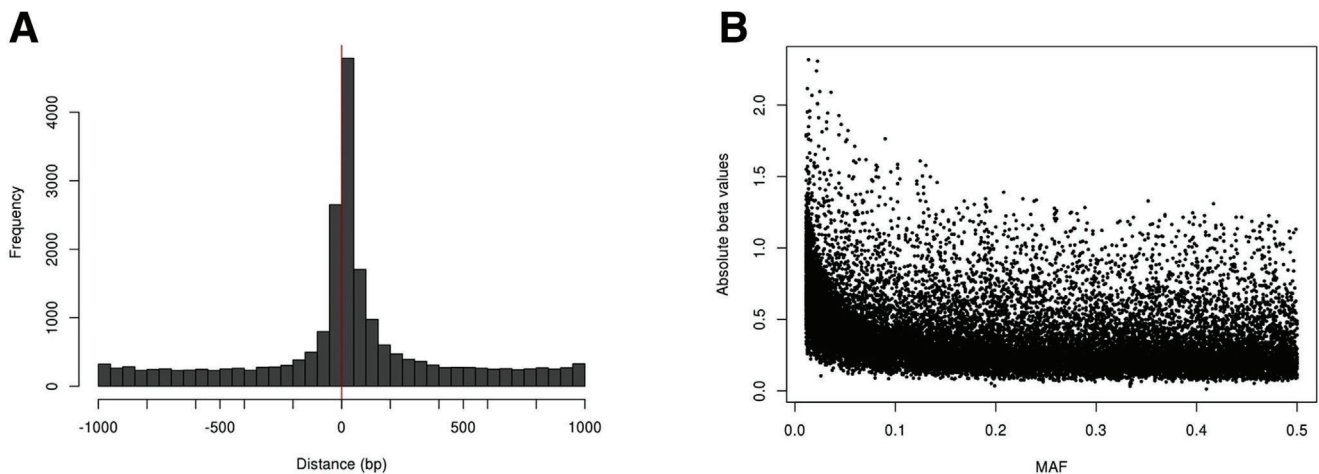


**Figure 2. Differential gene expression profiles across colon anatomic subsites.** (A) Volcano plot showing the distribution of gene log fold changes and statistical significance. Points above the horizontal dashed line represent genes considered significantly differentially expressed ( $\text{FWER} \leq 0.05$ ). Points in red and blue color represent genes over (red) and underexpressed (blue) following a consistent trend from ascending to descending colon (ie, overexpressed in transverse relative to ascending colon and overexpressed in descending relative to transverse). (B) Heatmap showing the expression profiles of the top 30 differentially expressed genes across colon subsites ranked by FWER-adjusted  $P$  values. Hierarchical clustering shows the similarity between genes (rows) and samples (columns) based on Euclidean distances.

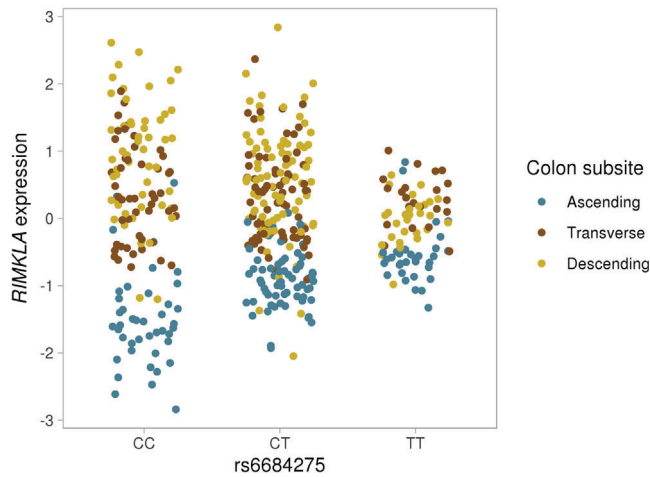
classified as nonsense, start loss, frameshift, canonical splice site, missense, or synonymous variants (Supplementary Table 13).

Next, we performed enrichment analysis at regulatory regions (open chromatin regions, active enhancers, super-enhancers, and transcription factor binding sites) using data derived from colon cell lines as well as from normal and

cancerous colon tissue. We found significant enrichment ( $P$  value  $\leq .05$ ) in all types of regulatory regions for both eSNPs and sSNPs. In addition, we looked for enrichment in target sites distributed across the genome of 170 RNA-binding proteins (RBPs). The top 20 RBPs with the lowest  $P$  values for eSNP enrichment are included in Figure 8A. Of those RBPs, 15 also were among the top 20 RBPs most



**Figure 3. eQTLs features.** (A) Distribution of distances between eSNPs location and corresponding eGenes TSS. (B) Distribution of absolute beta values (slope associated with the nominal  $P$  value of association) of eQTLs and eSNPs minor allele frequencies (MAF). These variables were not correlated ( $r = 0.14$ ).



**Figure 4. Example of eQTLs interacting with colon subsite.** Distribution of expression level (inverse normal transformed trimmed means of M values) of *RIMKLA* by rs6684275 genotype and colon subsite.

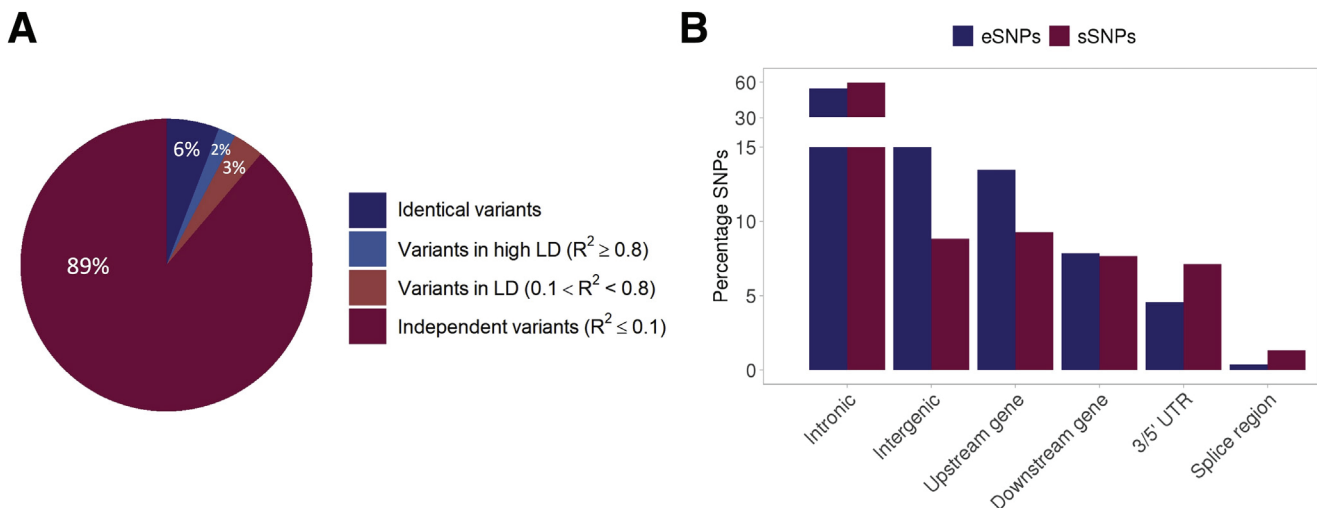
enriched for sSNPs. In both cases, the heterogeneous nuclear ribonucleoprotein C was the RBP with the most significant enrichment. The RBPs with highest enrichment values for sSNPs are included in Figure 8B. We observed sSNPs enriched at binding sites of spliceosome constituents such as the splicing factor U2 small nuclear RNA auxiliary factor 1. Full enrichment results are listed in Supplementary Table 14.

### Phenotype Heritability Enrichment and Colocalization Analyses

To quantify the ability of BarcUVa-Seq QTLs to explain a phenotype's genetic risk loci, we analyzed eSNPs/sSNPs in the context of their potential contribution to total SNP-based

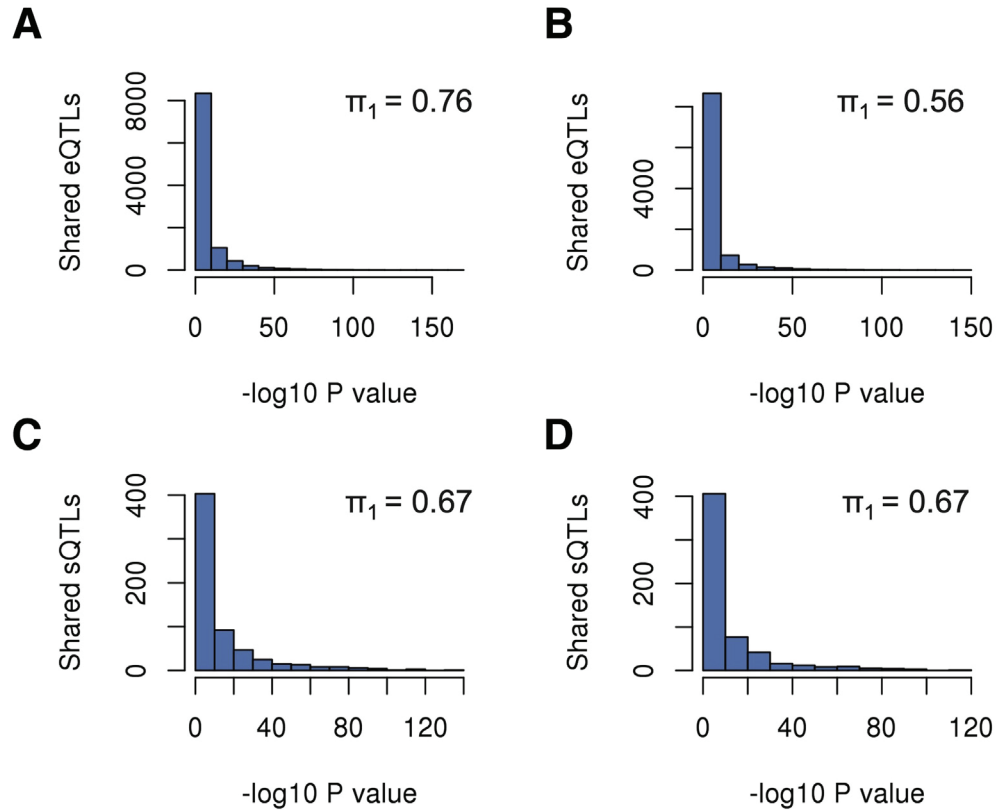
heritability estimates of multiple complex traits. SNP-based heritability is the heritability of traits captured by SNPs in a SNP array in the context of a genome-wide association study (GWAS). We performed SNP-based heritability enrichment tests in 63 complex diseases and traits that we considered a priori to influence or be influenced by colon homeostasis. We observed that eSNPs were enriched in the SNP-based heritability estimation of 20 diseases or traits after Bonferroni adjustment ( $P$  value  $\leq 8 \times 10^{-4}$ ) and 31 diseases or traits at an unadjusted  $P$  value  $\leq .01$ . SNP-heritability enrichments for 33 traits and diseases are included in Figure 9A, and full results are listed in Supplementary Table 15. BarcUVa-Seq eSNPs explained 17% of the total SNP-based heritability of CRC ( $P$  value =  $9 \times 10^{-8}$ ), which accounts for 10% of the phenotype (based on a recent GWAS study<sup>34</sup>). Interestingly, eSNPs also were enriched in the SNP-based heritability estimation of psychiatric-neuronal disease, such as schizophrenia, bipolar disorder, and multisite chronic pain. BarcUVa-Seq sSNPs were enriched in the SNP-based heritability estimation of 10 diseases and traits at a  $P$  value  $\leq .01$ , but no enrichments were statistically significant after Bonferroni adjustment (Figure 9B shows 33 representative traits or diseases, Supplementary Table 15 has the full list of results). BarcUVa-Seq sSNPs explained 3% of the total SNP heritability of ulcerative colitis ( $P$  value = .02), which accounts for 13% of the phenotype (Figure 9B).

Subsequently, to nominate candidate genes at GWAS-identified genetic risk loci, we performed colocalization analyses for the complex traits and diseases that passed Bonferroni correction for SNP-based heritability analysis for BarcUVa-Seq eSNPs. The regional colocalization probability is used as a proxy for the gene's causality, that is, to quantify the probability that an eQTL and a GWAS signal share the same causal variant.<sup>35</sup> In the case of CRC, we identified 13 genes with regional colocalization probability greater than

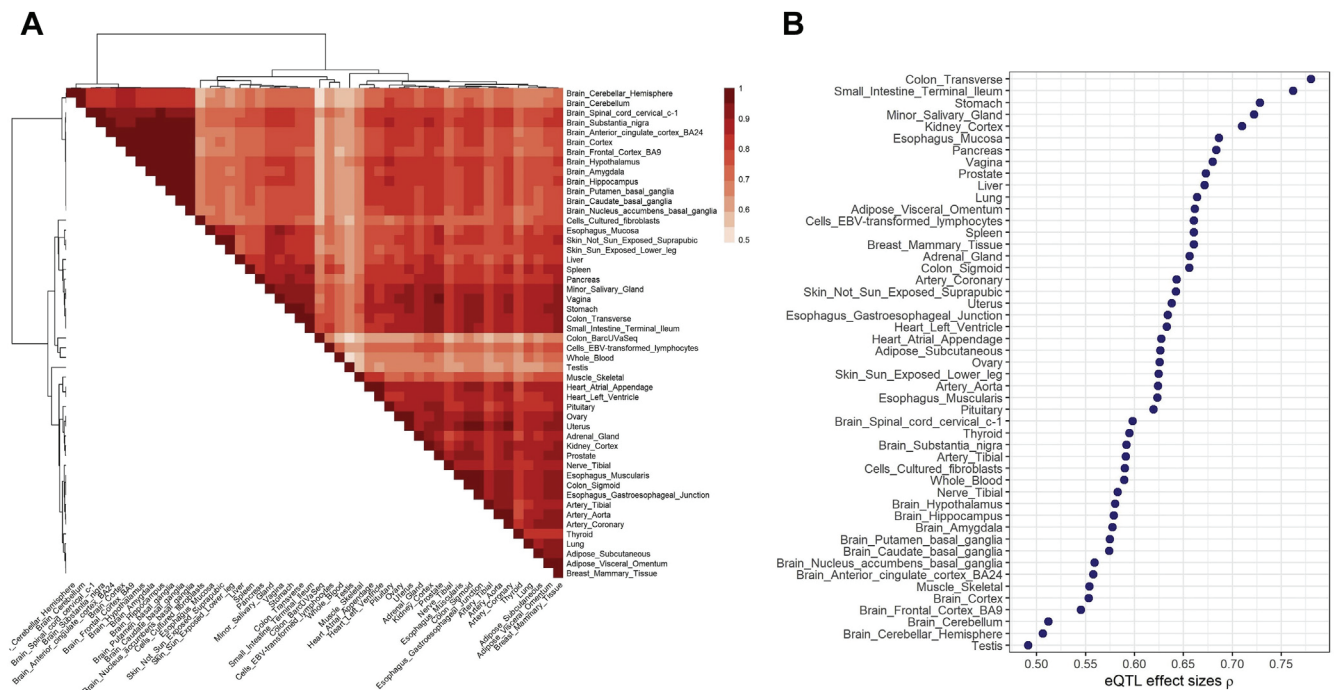


**Figure 5. Colocalization among sSNPs and eSNPs and genomic region annotation.** (A) Percentages of colocalization patterns among sSNPs and eSNPs in common genes according to measures of LD  $R^2$ . (B) Percentages of eSNPs and sSNPs at specific genomic regions, note that the plot is gapped between 15% and 30% and rescaled between 30% and 60% to show the differences in the categories with the lowest representation. UTR, untranslated region.

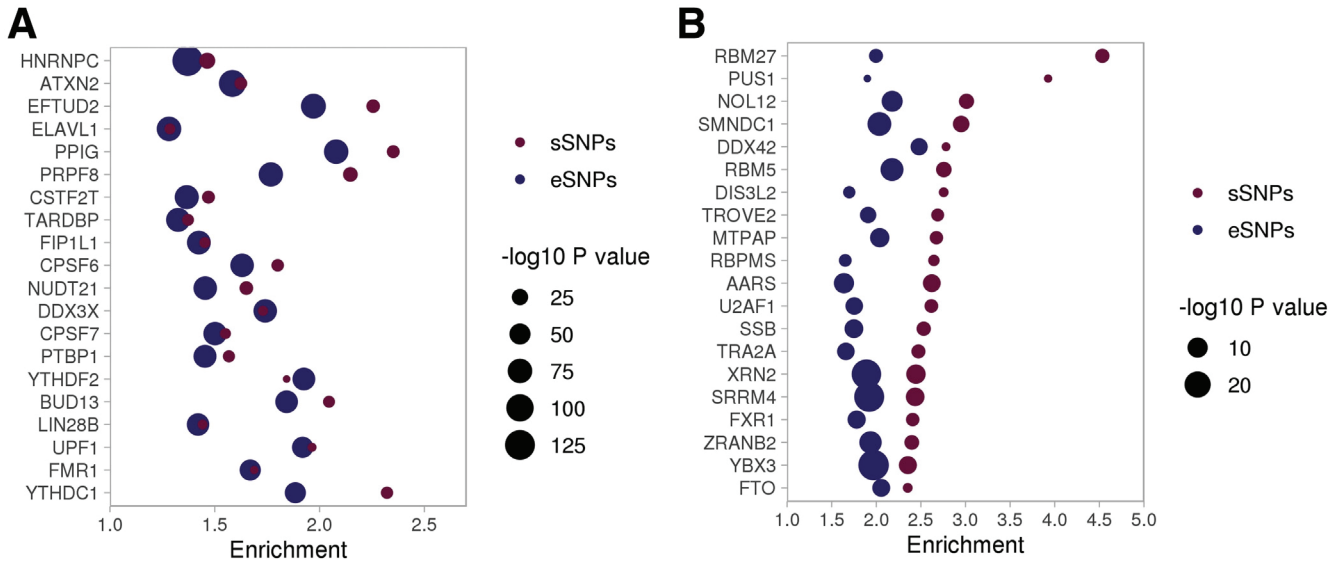




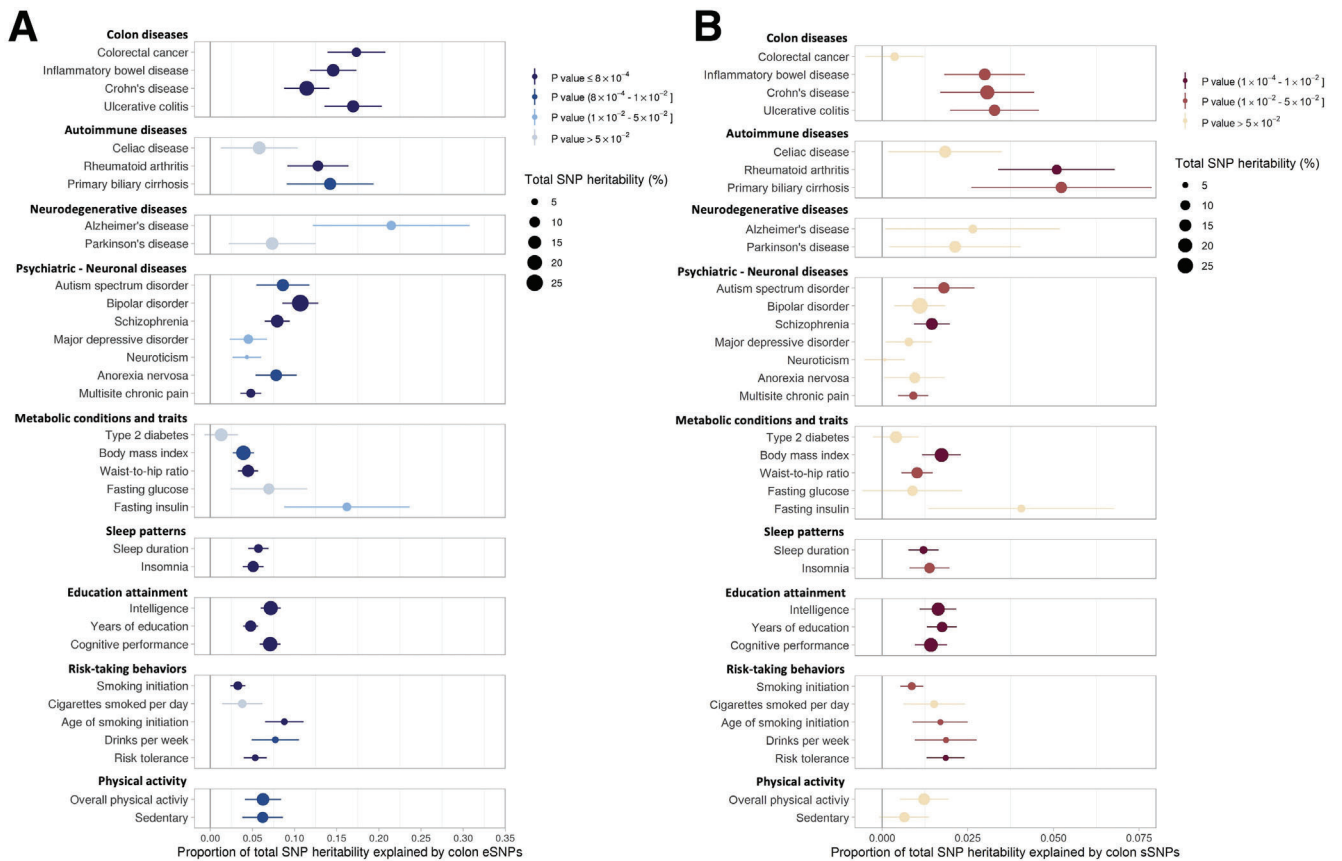
**Figure 6. Replication analysis of eQTLs/sQTLs with GTEx v8 colon data.** The value of the  $\pi_1$  statistic is shown. The distribution of  $P$  values is shown for (A) transverse colon eQTLs, (B) sigmoid colon eQTLs, (C) transverse colon sQTLs, and (D) sigmoid colon sQTLs.



**Figure 7. Meta-analysis with GTEx v8 tissues.** (A) Clustering of BarcUVa-Seq and GTEx v8 tissues based on pairwise Spearman correlation of eQTL effect sizes derived from mashr meta-analysis. We only considered significant ( $FDR \leq 0.05$ ) and active (local false sign rate [LFSR]  $\leq 0.05$ ) eQTLs. (B) Spearman correlation of eQTL effect sizes between BarcUVa-Seq and GTEx v8 tissues. eQTL effect sizes were derived from mashr meta-analysis. We only considered significant ( $FDR \leq 0.05$ ) and active (LFSR  $\leq 0.05$ ) eQTLs.



**Figure 8. Enrichment of eSNPs/sSNPs in binding sites across the genome of RBPs.** (A) The top 20 RBP with the lowest enrichment *P* values for eSNPs. (B) The top 20 RBPs with the highest enrichment values for sSNPs (*P* value < .05).



**Figure 9. BarCUVa-Seq QTL enrichment results for total SNP heritability of 33 complex traits and diseases related to colon tissue.** (A) Proportion of total SNP heritability explained by eSNPs is shown on the x axis, along with error bars. The size of the points indicates the percentage of the total SNP heritability out of the total heritability of the phenotype. (B) Proportion of total SNP heritability explained by sSNPs is shown on the x axis, along with error bars. The size of the points indicates the percentage of the total SNP heritability out of the total heritability of the phenotype.

0.9, including known risk genes such as *COLCA1* and *COLCA2*,<sup>6</sup> as well as other less-well-described genes such as *ANKRD36*. In the case of inflammatory bowel disease, we identified 6 genes with a regional colocalization probability greater than 0.9, such as *IRF8* and *RGS14* (Figure 10). Full results are available in the Supplementary Data.

### Colon Transcriptome Explorer

Gene and transcript abundances for the BarcUVa-Seq data set, as well as eQTLs/sQTLs, have been loaded into the web-based visualization resource CoTrEx. This tool facilitates searches for genes and transcripts of interest for their visualization in customizable plots, such as a strip chart, heatmap, and principal component analysis (PCA) plots. The interactive application includes different options for filtering and coloring the data by covariates. Figure 11 shows an example in the Expression tab. CoTrEx is freely available online at <http://barcuvaseq.org/cotrex>.

## Discussion

In the present study we analyzed a large data set (BarcUVa-Seq) comprising germline SNPs and transcriptome profiles from mucosal biopsy specimens of ascending, transverse, and descending colon collected from 445 healthy living individuals. Differential expression patterns were identified across colon subsites. We profiled 11,739 eQTLs comprising 11,427 unique SNPs associated with the expression of 11,739 genes. In addition, we identified 13,243 AS events from 7 distinct AS categories and identified 1125 AS events in 1125 genes associated with 1122 unique SNPs (sQTLs). These eQTLs/sQTLs frequently were intronic and enriched in regulatory regions. We showed how these are useful for annotation of GWAS-identified risk loci and prioritization of candidate effector genes. Moreover, we replicated and meta-analyzed our QTLs with GTEx v8 data. Finally, we built an interactive web resource to explore the expression profiles and QTLs of the BarcUVa-Seq data set.

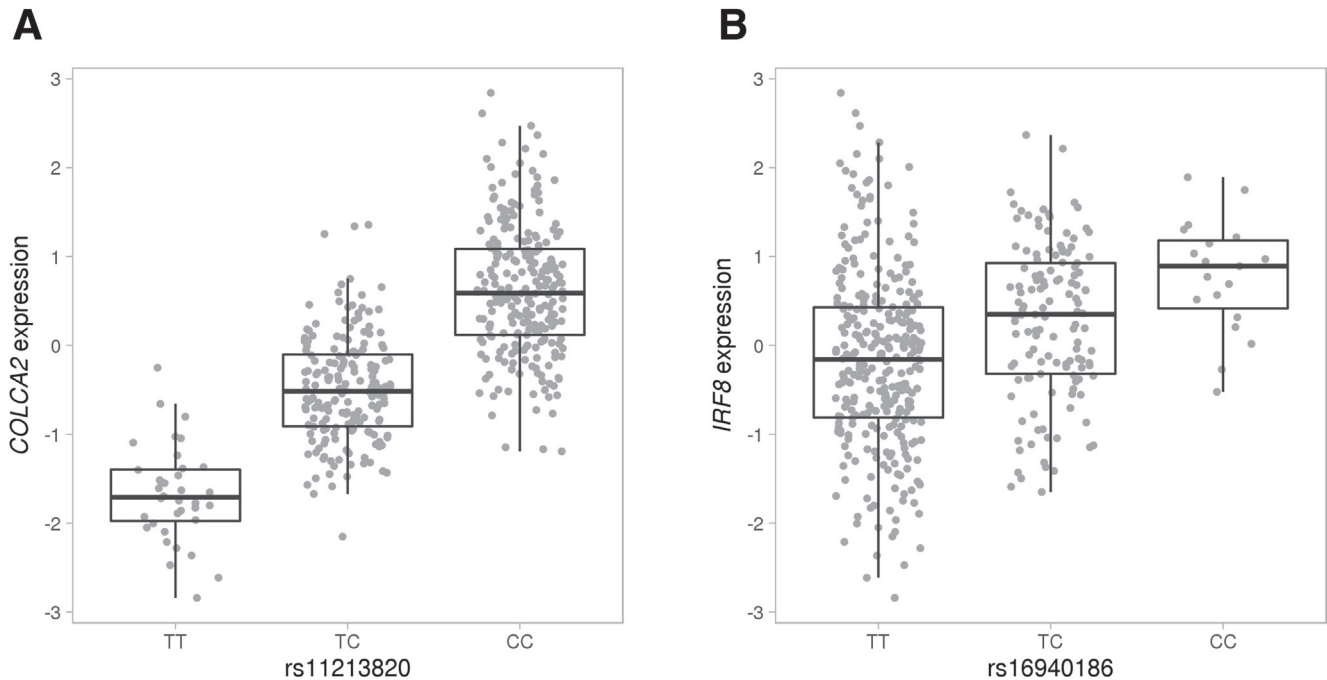
In contrast to BarcUVa-Seq, the GTEx project provided RNA-Seq data on sigmoid and transverse colon tissue from post-mortem subjects and extracted RNA from full-thickness and muscularis-only sections.<sup>8,36</sup> Our novel BarcUVa-Seq data set overcomes some of the limitations of the GTEx colon data sets. BarcUVa-Seq samples were collected as superficial mucosal biopsy specimens in living subjects undergoing colonoscopy, which provide an optimal representation of the normal physiology of the colon epithelium. Moreover, they included subsites of the large intestine not assessed previously. Together with the enrichment of colon epithelial cells in superficial biopsy specimens, inclusion of ascending, transverse, and descending colon samples make BarcUVa-Seq a unique colon transcriptome data set.

Next-generation RNA-Seq data provide estimates of AS. Although long-read sequencing technologies can provide transcriptomic profiles with full-length isoform information, such technologies have lower base-level fidelity and are less feasible in large population-based studies at their current

cost.<sup>11</sup> In this study we used 2 complementary methods to provide a comprehensive profile of AS. The frequencies of genes with specific AS patterns that we identified in colon tissue are similar to those described in other tissues, where genes with exon skipping events were the most frequent.<sup>17</sup> Predicting AS events helps generate hypotheses about specific molecular mechanisms involved in post-transcriptional modifications. In contrast to profiling individual transcripts to characterize the transcriptome, AS events group transcripts with similar structure. However, the profiles of annotated AS events are sensitive to the choice of transcript annotations,<sup>11</sup> and other measures of AS, such as clusters of excised introns, complement the characterization of AS events.<sup>13</sup>

Regarding colon location, transcriptomic differences between subsites in normal colon have been described previously,<sup>37</sup> including gene expression differences in genes from the cytochrome P450 family. In addition, different AS events have been identified between CRC tumors located in the ascending and descending colon.<sup>38</sup> Indeed, tumor distribution across the colon has been associated with differential mutation and immune profiles, prognosis, and treatment response.<sup>39,40</sup> In this study, we identified a subset of genes expressed differentially between colon subsites that are involved in molecular pathways related to lipid, xenobiotic, and drug metabolism, and a subset of genes involved in antimicrobial response. We observed that the gene expression profile of transverse colon tissue was more similar to the descending than to the ascending colon, which was unexpected based on embryologic origin and adult blood supply. Differential gene expression across the colon may reflect differences in cell type composition because we find gene markers of different cell types of the colon epithelium shown by single-cell RNA-Seq studies.<sup>41-43</sup> For instance, using our data, we confirmed that goblet cell markers defined elsewhere,<sup>41</sup> such as *MUC2* and *TFF3*, are overexpressed in descending colon (Supplementary Table 2), which supports previous findings that have shown that goblet cell content increases caudally from duodenum to distal colon.<sup>44</sup> Differential expression also may be influenced by differential exposure owing to variability in luminal content along the length of the colon, including microbial communities.<sup>43</sup>

We identified eQTLs and sQTLs assumed to participate in the transcriptional regulation of colon epithelium via cis mechanisms. These had strong replication in the transverse colon from GTEx v8 and were more similar to tissues with a high proportion of mucosa (eg, terminal ileum, stomach, and salivary gland) than others from GTEx v8, showing the robustness of BarcUVa-Seq data. The lower replication value in sigmoid colon may be owing to the higher proportion of muscularis in this tissue.<sup>8,36</sup> We found fewer sGenes than eGenes, partly because the number of genes that showed splicing variability was lower than genes with expression variability. In addition, we had lower power to detect expression for transcripts than for genes at our depth of coverage. We found similar distributions of eSNPs/sSNPs around gene TSSs, as well as across estimated effect sizes, genomic locations, and functional consequences. We



**Figure 10. The top eQTLs of the genes with the highest regional colocalization probability for CRC and inflammatory bowel disease.** (A) Expression level (inverse normal transformed trimmed means of M values [TMMs]) of *COLCA2* by genotype of the eSNP rs11213820. (B) Expression level (inverse normal transformed TMMs) of *IRF8* by genotype of the eSNP rs16940186.

observed a high proportion of sGenes among eGenes, as reported elsewhere.<sup>24,25</sup> Although they can colocalize, eQTLs and sQTLs usually are independent.<sup>27</sup> sQTLs add information to eQTLs as they associate SNPs with changes in relative use of specific sets of transcripts sharing a common structure and post-transcriptional mechanism.

In this study, we showed that regulation of gene expression and AS is associated with tissue-specific epigenetic variations, including chromatin remodeling and histone modifications.<sup>45</sup> The dysregulation of these features has been associated with initiation and progression of diseases such as CRC.<sup>45,46</sup> We showed that normal colon eSNPs/sSNPs are present at many important regulatory regions marked by epigenetic signatures, such as open chromatin and proximal enhancers of both normal and malignant colon tissue. In addition, we identified specific RBPs and transcription factors as potential regulators of AS in normal colon.

We provide a comprehensive profile of AS for normal tissue along colon subsites in living subjects. We described differential gene expression and splicing between the ascending and descending normal colon, which involved genes of immune response and drug metabolism. We expanded the number of colon QTLs and assessed eQTL interaction with colon subsites. In addition, we observed that colon eQTLs/sQTLs contributed to the SNP-based heritability of brain-related traits and disease, supporting a model of epithelial-neuronal communication along the gut-brain axis.<sup>28</sup> Thus, our QTL catalog may be of potential interest for researchers investigating traits and diseases

that do not primarily affect the colon, but other organs. It is important to note that these results could reflect a common regulation of expression between tissues. In addition, colocalization alludes to potential molecular mechanisms associated with risk loci, but may not prove to be directly causal.

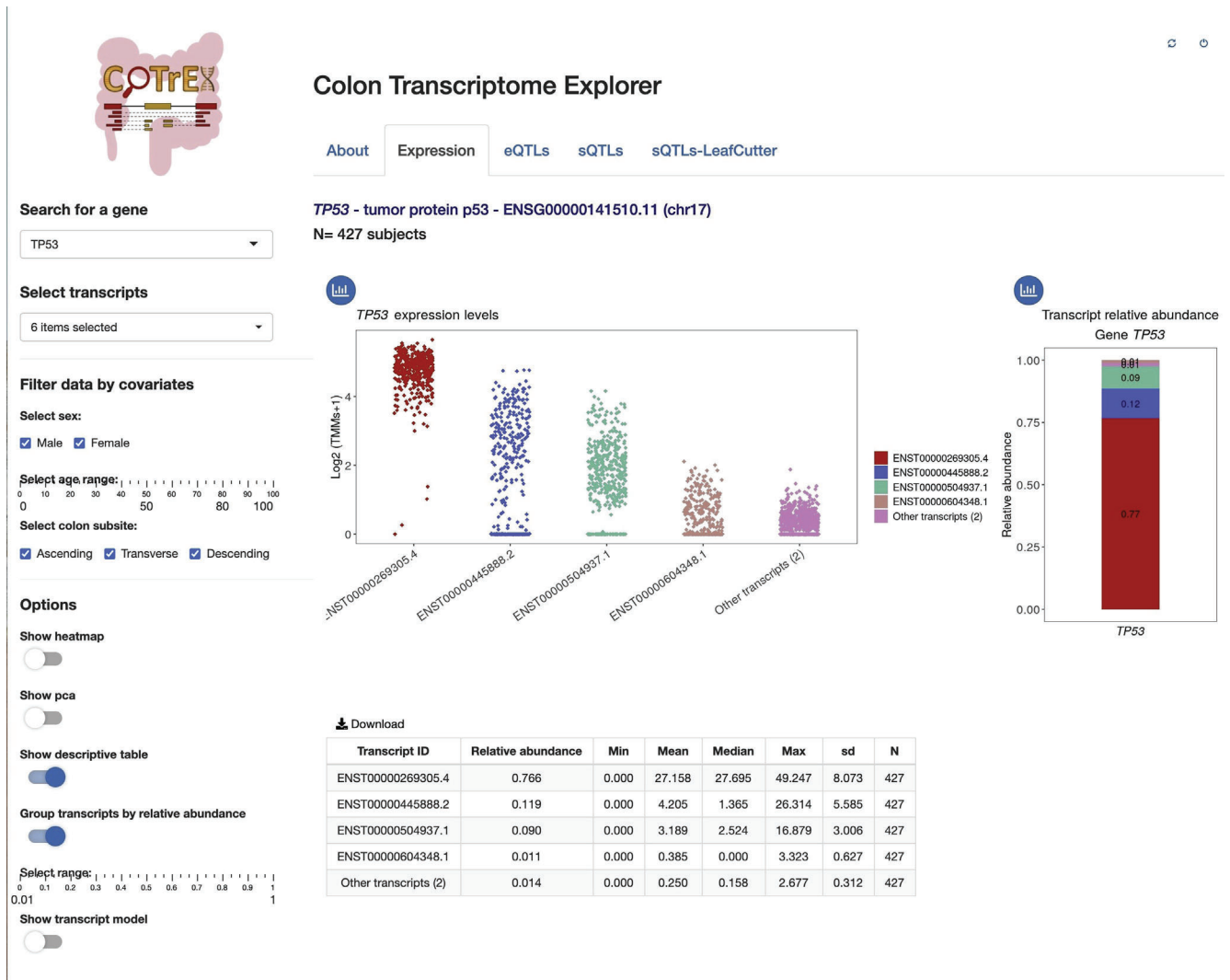
Overall, our findings provide evidence of the regulation of gene expression and alternative splicing in the colon as potential underlying mechanisms of genetic risk loci and should serve as a rich resource for the research community.

## Methods

### Sample Collection

Subjects included in the study ( $n = 445$ ; 64% females) had a mean age of 60 years, were almost all of European ancestry, and received an indication for colonoscopy after a positive fecal immunochemical test result (hemoglobin level,  $>20$  mg Hb/g) or by direct referral by their medical doctor. Subjects had no lesions at colonoscopy and no history of polyps or CRC. Non-neoplastic colon mucosa biopsy specimens were obtained endoscopically from the ascending ( $n = 138$ ; 31%), transverse ( $n = 143$ ; 32%), and descending ( $n = 164$ ; 37%) colon (Table 1). Peripheral blood samples also were collected. Informed consent was obtained from all participants. The corresponding study protocol was approved by the Bellvitge University Hospital Ethics Committee (PR073/11 and PR286/15) and followed national and international directives on ethics and data protection. More information about the BarcUVa-Seq project





**Figure 11. Overview of the expression tab of CoTrEx.** As an example, the transcript expression values and relative abundances of the *TP53* gene are shown, along with different display options.

can be accessed online at <https://barcuvaseq.org>. All authors had access to the study data and reviewed and approved the final manuscript.

### RNA-Seq Library Preparation and Sequencing

RNA was extracted from frozen tissue using the mirVana kit (Thermo Fisher Scientific, Waltham, MA) after homogenization using the Minilys bead mill (Bertin Instruments, Montigny le Bretonneux, France). The RNA was DNase treated and concentrated using the RNA Clean and Concentrator-5 kit (Zymo Research, Irvine, CA). Quantification of total RNA was executed using a Qubit Fluorometer (Invitrogen, Waltham, MA). An Agilent (Santa Clara, CA) 2100 Bioanalyzer or TapeStation was used to assess quality. For library preparation, the Illumina TruSeq Stranded Total RNA Library Prep Gold kit was used. Libraries were tagged with unique adapter indexes. Final libraries were validated on the Agilent 2100 Bioanalyzer, quantified via quantitative

polymerase chain reaction, pooled at equimolar ratios, diluted, denatured, and loaded onto an Illumina HiSeq 2500 (high-output mode), for batches 1–7, or a NovaSeq 6000, for batch 8, instruments using a paired-end flowcell.

### RNA-Seq Data Processing

Low-quality bases, sequencing adapters, and ribosomal RNA of raw sequences were trimmed from RNA-Seq reads using BBTools suite (Joint Genome Institute, Berkeley, CA).<sup>47</sup> FastQC (Babraham Bioinformatics, Cambridge, UK)<sup>48</sup> was used for quality control. Trimmed reads were aligned against human transcriptome using the Genome Reference Consortium human reference 37 assembly (GRCh37/hg19) with the Spliced Transcripts Alignment to a Reference (STAR, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY) software in 2-pass mode<sup>49</sup> using GENCODE (EMBL-EBI, Hinxton, UK) release 19 annotations, which include a total of 57,952 genes and 196,667 transcripts.<sup>31</sup> We only included

samples with a depth of coverage greater than 10 million mappable paired-end reads, a multimapping rate lower than 15%, and a unique mapping rate greater than 80%. The mean library size was 32M (SD, 8.5M). Gene and transcript expression were quantified with RSEM (University of Wisconsin-Madison, Madison, WI).<sup>50</sup> Genes and transcripts with fewer than 6 and 3 counts, respectively, in less than 10% of the samples were considered not expressed and filtered out. Trimmed mean of M values were computed from counts to correct for library size and RNA composition.

### Genotype Data Processing

Genotyping of approximately 400,000 SNPs was performed with the Illumina OncoArray BeadChip.<sup>30</sup> We only included samples with a genotyping rate greater than 95%. The following aspects also were assessed before imputation: duplication and relatedness greater than 0.8, missing rate per SNP greater than 0.1, missing rate per sample greater than 0.1, sex concordance (genetic and reported sex), heterozygosity: means  $\pm$  4 SD and Hardy-Weinberg disequilibrium  $P$  value less than  $1 \times 10^{-4}$ . We obtained allelic dosages from 39,117,105 and 1,228,035 SNPs for autosomes and chromosome X, respectively, using SHAPEIT (University of Oxford, Oxford, UK)<sup>51</sup> for phasing and Minimac 3 (University of Michigan, Ann Arbor, MI)<sup>53</sup> for imputation with The Haplotype Reference Consortium panel on the Michigan Imputation Server.<sup>52</sup> SNPs with an imputation quality of  $R^2$  less than 0.7 or minor allele frequency (MAF) less than 1% were excluded, resulting in 6,804,675 and 183,788 SNPs for autosomes and chromosome X, respectively. Allelic dosages were used for subsequent QTL analyses. SNP IDs were annotated using dbSNP version 142.<sup>53</sup> Principal components of genetic data were obtained with PLINK 1.9 (Complete Genomics, Mountain View, CA).<sup>54</sup> We checked that both genotype and RNA-Seq samples had been labeled correctly and belonged to the same individual using Picard Tools CheckFingerprint (Broad Institute, Cambridge, MA).

### Alternative Splicing Profiling

For quantifying AS, we used 2 complementary methods that provide the relative abundance (ie, percent splicing index [PSI]) of specific AS features. Seven types of AS events were determined based on GENCODE version 19 annotations with SUPPA2 (Catalan Institution for Research and Advanced Studies, Barcelona, Spain).<sup>12</sup> In this case, the PSI reflects the proportion of transcripts of a given gene showing a specific AS event (ie, inclusion transcripts) of the total transcripts of the gene.<sup>11</sup> This metric was calculated with SUPPA2 for each AS event by dividing the expression levels of the inclusion transcripts by the total expression levels of all transcripts of the gene. We kept AS events in which the median PSI for all samples was between 0.05 and 0.95 (see AS events annotations in [Supplementary Table 1](#)). As a complementary approach, we used LeafCutter (Stanford University, Stanford, CA)<sup>13</sup> following the analysis procedure described elsewhere<sup>8</sup> to compute the relative abundance of alternatively excised introns.

### Differential Gene Expression and Splicing Analysis

Differential gene expression analysis was performed using a quasi-likelihood F-test implemented in the R package edgeR (Garvan Institute of Medical Research, Parkville, Australia).<sup>55</sup> Ward's minimum variance method with Euclidean distances was used for hierarchical clustering. For differential splicing analysis, normalized PSI values of AS events were fitted in a linear model adjusted for sex, age, and sequencing batch using the R package limma (University of Melbourne, Parkville, Australia).<sup>56</sup> The function *diffSplice* was used to perform an F test to find the differences between AS event log-fold-changes of a gene and yield a single gene-level  $P$  value. T tests for individual AS events also were performed with *diffSplice*. Differential use of excised introns was performed with LeafCutter,<sup>13</sup> adjusting for sex, age, and sequencing batch. Functional enrichment analysis was performed with FUMA *gen2func* (University of Amsterdam, Amsterdam, The Netherlands)<sup>57</sup> using differentially expressed genes with FWER of 0.05 or less. FWER values were estimated for correcting for multiple testing using a Bonferroni correction.

### eQTL/sQTL Mapping

We mapped QTLs within 1 Mb of the TSSs for given genes and assumed QTLs influenced expression of nearby genes via cis mechanisms. For QTL identification we used FastQTL (University of Geneva Medical School, Geneva, Switzerland) version 2.0.<sup>58</sup> We applied an inverse normal transformation on gene trimmed means of M values and PSI values, which mitigates the effect of outliers and normalizes the expression distribution across samples. We adjusted the models for age, sex, sequencing batch, tissue anatomic location, genetic ancestry (2 principal components), and probabilistic estimation of expression residuals factors,<sup>59</sup> which capture the effects of unknown confounding variables. We chose the number of probabilistic estimation of expression residuals factors that maximized the discovery of eGenes/sGenes. FDR (Storey and Tibshirani procedure) was computed with R package *qvalue* (Princeton University, Princeton, NJ).<sup>60</sup> For colon subsite eQTL interaction analysis we used the FastQTL version 2.0 interaction mode.<sup>57</sup>

### Replication and Meta-Analysis With GTEx Data

For replication analysis, we estimated  $\pi_1$ <sup>33</sup> with the R package *qvalue*.<sup>60</sup> This statistic reflects the proportion of true positives among BarcUVa-Seq QTLs that also were detected by the corresponding QTL analysis in GTEx v8. Following a common approach described elsewhere,<sup>8</sup> we only included associations involving the SNP with the lowest  $P$  value for each gene to avoid including many SNPs in LD. For meta-analysis, full GTEx v8 eQTL summary statistics ( $n = 49$  tissues) were downloaded from the Google Cloud Platform (Mountain View, CA) under *gtex*-resources. We used a multivariate adaptive shrinkage approach using the R package *mashr* (University of Chicago, Chicago, IL)<sup>33</sup> following the same analytic pipeline

described elsewhere.<sup>8</sup> Effect size estimates and local false sign rate output by mashr were used as metrics of QTL magnitude and activity, respectively. A local false sign rate less than 0.05 was used as a threshold for significant QTL activity.

### Annotation and Functional Enrichment Analysis

For the annotation of genomic regions and classification of variants according to their functional consequence we used the ENSEMBL Variant Effect Predictor (EMBL-EBI, Hinxton, UK).<sup>61</sup> We used the *-pick* flag to extract a single annotation per variant following an ordered set of criteria to prioritize annotations. For functional enrichment analysis in regulatory regions distributed across the genome (Supplementary Table 14), we compiled a list of publicly available regions relevant for colon tissue from different studies (ie, active enhancers,<sup>46</sup> variant enhancer loci,<sup>46</sup> open chromatin sites,<sup>34,46</sup> superenhancers,<sup>62</sup> and transcription factor binding sites<sup>63</sup>). Regions from multiple samples of the same assay type were joined. In addition, we downloaded RNA binding protein sites, including splicing factor binding sites, from CLIPdb (Tsinghua University, Beijing, China).<sup>64</sup> We used GREGOR (University of Michigan, Ann Arbor, MI),<sup>65</sup> which defines enrichment (fold change) as the ratio between the number of observed vs expected SNPs overlapping the regulatory regions. This approach accounts for the number of LD proxies, gene proximity, and MAF.

### Phenotype Heritability Enrichment and Colocalization Analyses

For the SNP-based heritability enrichment analysis (partitioned heritability analysis) of eSNPs/sSNPs among disease-/trait-associated loci, we applied linkage disequilibrium score regression using the software LD Score (Broad Institute of MIT, Cambridge, MA)<sup>66</sup> with baselineLD model. A list with the GWAS summary statistics used for this analysis and related information can be found in Supplementary Table 15. Total SNP heritability for the tested phenotypes was estimated in observed scale for continuous traits and in liability scale for binary traits, using LD score regression from a total of 1,217,312 SNPs with a MAF greater than 0.05 in HapMap phase 3 populations (NHGRI, Bethesda, MD).<sup>66</sup> Under the null hypothesis of all SNPs contributing equally to the total SNP-based heritability, we would expect that the 1122 sSNPs and 11,427 eSNPs identified in this study explain approximately 0.09% and 0.94%, respectively, of estimated total SNP heritability. Population prevalence and lifetime risk in the case of CRC was curated from the literature. For colocalization we used the fastENLOC (University of Michigan)<sup>35</sup> approach. We computed Z-score-derived posterior inclusion probabilities for GWAS summary statistics with TORUS (University of Michigan)<sup>67</sup> and assigned LD blocks to each locus using the references defined elsewhere.<sup>68</sup> We performed multi-SNP fine-mapping analysis of eQTLs with DAP-G (University of Michigan).<sup>69</sup>

### Web Application

The web-based visualization resource CoTrEx was developed with the RStudio platform Shiny (Boston, MA)<sup>70</sup> using open-source software.

### Data Availability

The RNA-Seq and SNP data that support the findings of this study as well as the sample covariates are available from the European Genome-phenome Archive under accession number EGAS00001004891. Complete summary statistics (including all FastQTL nominal pass results) for all QTLs identified in this study are available from the Digital Repository of the University of Barcelona at <http://hdl.handle.net/2445/172697>. Top-QTLs per gene are available in Supplementary Tables 7, 9, 10, 11, and 13.

### References

1. Momozawa Y, Dmitrieva J, Théâtre E, Deffontaine V, Rahmouni S, Charlotheaux B, Crins F, Docampo E, Elansary M, Gori A-S, Lecut C, Mariman R, Mni M, Oury C, Altukhov I, Alexeev D, Aulchenko Y, Amininejad L, Bouma G, Hoentjen F, Löwenberg M, Oldenburg B, Pierik MJ, Vander Meulen-de Jong AE, Janneke van der Woude C, Visschedijk MC, International IBD Genetics Consortium, Lathrop M, Hugot J-P, Weersma RK, De Vos M, Franchimont D, Vermeire S, Kubo M, Louis E, Georges M. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat Commun* 2018;9:2427.
2. Closa A, Cordero D, Sanz-Pamplona R, Solé X, Crous-Bou M, Paré-Brunet L, Berenguer A, Guino E, Lopez-Doriga A, Guardiola J, Biondo S, Salazar R, Moreno V. Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. *Carcinogenesis* 2014;35:2039–2046.
3. Moreno V, Alonso MH, Closa A, Vallés X, Diez-Villanueva A, Valle L, Castellví-Bel S, Sanz-Pamplona R, Lopez-Doriga A, Cordero D, Solé X. Colon-specific eQTL analysis to inform on functional SNPs. *Br J Cancer* 2018; 119:971–977.
4. Singh T, Levine AP, Smith PJ, Smith AM, Segal AW, Barrett JC. Characterization of expression quantitative trait loci in the human colon. *Inflamm Bowel Dis* 2015; 21:251–256.
5. Hulusi I, Gamazon ER, Skol AD, Xicola RM, Llor X, Onel K, Ellis NA, Kupfer SS. Enrichment of inflammatory bowel disease and colorectal cancer risk variants in colon expression quantitative trait loci. *BMC Genomics* 2015; 16:138.
6. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, Farrington S, Svinti V, Palles C, Orlando G, Sud A, Holroyd A, Penegar S, Theodoratou E, Vaughan-Shaw P, Campbell H, Zgaga L, Hayward C, Campbell A, Harris S, Deary IJ, Starr J, Gatcombe L, Pinna M, Briggs S, Martin L, Jaeger E, Sharma-Oates A, East J, Leedham S, Arnold R, Johnstone E, Wang H, Kerr D, Kerr R, Maughan T, Kaplan R, Al-Tassan N, Palin K,



- Hänninen UA, Cajuso T, Tanskanen T, Kondelin J, Kaasinen E, Sarin A-P, Eriksson JG, Rissanen H, Knekt P, Pukkala E, Jousilahti P, Salomaa V, Ripatti S, Palotie A, Renkonen-Sinisalo L, Lepistö A, Böhm J, Mecklin J-P, Buchanan DD, Win A-K, Hopper J, Jenkins ME, Lindor NM, Newcomb PA, Gallinger S, Duggan D, Casey G, Hoffmann P, Nöthen MM, Jöckel K-H, Easton DF, Pharoah PDP, Peto J, Canzian F, Swerdlow A, Eeles RA, Kote-Jarai Z, Muir K, Pashayan N, PRACTICAL Consortium, Harkin A, Allan K, McQueen J, Paul J, Iveson T, Saunders M, Butterbach K, Chang-Claude J, Hoffmeister M, Brenner H, Kirac I, Matošević P, Hofer P, Brezina S, Gsur A, Cheadle JP, Aaltonen LA, Tomlinson I, Houlston RS, Dunlop MG. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun* 2019;10:2154.
7. Ongen H, Andersen CL, Bramsen JB, Oster B, Rasmussen MH, Ferreira PG, Sandoval J, Vidal E, Whiffin N, Planchon A, Padioleau I, Bielser D, Romano L, Tomlinson I, Houlston RS, Esteller M, Orntoft TF, Dermitzakis ET. Putative cis-regulatory drivers in colorectal cancer. *Nature* 2014;512:87–90.
  8. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;369:1318–1330.
  9. Di C, Syafrizayanti, Zhang Q, Chen Y, Wang Y, Zhang X, Liu Y, Sun C, Zhang H, Hoheisel JD. Function, clinical application, and strategies of Pre-mRNA splicing in cancer. *Cell Death Differ* 2018;26:1181–1194.
  10. Manning KS, Cooper TA. The roles of RNA processing in translating genotype to phenotype. *Nat Rev Mol Cell Biol* 2017;18:102–114.
  11. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The expanding landscape of alternative splicing variation in human populations. *Am J Hum Genet* 2018;102:11–26.
  12. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, Eyraas E. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol* 2018;19:40.
  13. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK. Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* 2018;50:151–158.
  14. Ryan M, Wong WC, Brown R, Akbani R, Su X, Broom B, Melott J, Weinstein J. TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. *Nucleic Acids Res* 2016;44:D1018–D1022.
  15. Kahles A, Lehmann K-V, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, Stegle O, Kohlbacher O, Sander C, Cancer Genome Atlas Research Network, Ratsch G. Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* 2018;34:211–224.e6.
  16. Climente-González H, Porta-Pardo E, Godzik A, Eyraas E. The functional impact of alternative splicing in cancer. *Cell Rep* 2017;20:2215–2226.
  17. Wang K, Wu D, Zhang H, Das A, Basu M, Malin J, Cao K, Hannenhalli S. Comprehensive map of age-associated splicing changes across human tissues and their contributions to age-associated diseases. *Sci Rep* 2018;8:10929.
  18. Huang X, Liu J, Mo X, Liu H, Wei C, Huang L, Chen J, Tian C, Meng Y, Wu G, Xie W, P C FJ, Liu Z, Tang W. Systematic profiling of alternative splicing events and splicing factors in left- and right-sided colon cancer. *Aging* 2019;11:8270–8293.
  19. Xiong Y, Deng Y, Wang K, Zhou H, Zheng X, Si L, Fu Z. Profiles of alternative splicing in colorectal cancer and their clinical significance: A study based on large-scale sequencing data. *EBioMedicine* 2018;36:183–195.
  20. Zong Z, Li H, Yi C, Ying H, Zhu Z, Wang H. Genome-wide profiling of prognostic alternative splicing signature in colorectal cancer. *Front Oncol* 2018;8:537.
  21. Takata A, Matsumoto N, Kato T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nat Commun* 2017;8:14519.
  22. Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, Johnson AD, Levy D, O'Donnell CJ. Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet* 2015;47:345–352.
  23. Li YI, Wong G, Humphrey J, Raj T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nat Commun* 2019;10:994.
  24. Rotival M, Quach H, Quintana-Murci L. Defining the genetic and evolutionary architecture of alternative splicing in response to infection. *Nat Commun* 2019;10:1671.
  25. Walker RL, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, de la Torre-Ubieta L, Pasaniuc B, Stein JL, Geschwind DH. Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell* 2019;179:750–771.e22.
  26. Tian J, Wang Z, Mei S, Yang N, Yang Y, Ke J, Zhu Y, Gong Y, Zou D, Peng X, Wang X, Wan H, Zhong R, Chang J, Gong J, Han L, Miao X. CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Res* 2019;47:D909–D916.
  27. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK. RNA splicing is a primary link between genetic variation and disease. *Science* 2016;352:600–604.
  28. Najjar SA, Davis BM, Albers KM. Epithelial-neuronal communication in the colon: implications for visceral pain. *Trends Neurosci* 2020;43:170–181.
  29. Camilleri M. Leaky gut: mechanisms, measurement and clinical implications in humans. *Gut* 2019;68:1516–1526.
  30. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, Casey G, Hunter DJ, Sellers TA, Gruber SB, Dunning AM, Michailidou K, Fachal L, Doheny K, Spurdle AB, Li Y, Xiao X, Romm J, Pugh E, Coetzee GA, Hazelett DJ, Bojesen SE, Caga-Anan C, Haiman CA, Kamal A, Luccarini C, Tessier D, Vincent D, Bacot F, Van Den Berg DJ, Nelson S, Demetriades S, Goldgar DE, Couch FJ, Forman JL, Giles GG, Conti DV, Bickeböller H, Risch A, Waldenberger M, Brüske-Hohlfeld I, Hicks BD, Ling H, McGuffog L, Lee A, Kuchenbaecker K, Soucy P, Manj J, Cunningham JM, Butterbach K, Kote-Jarai Z, Kraft P, FitzGerald L, Lindström S, Adams M, McKay JD, Phelan CM, Benlloch S, Kelemen LE, Brennan P,



- Riggan M, O'Mara TA, Shen H, Shi Y, Thompson DJ, Goodman MT, Nielsen SF, Berchuck A, Laboissiere S, Schmit SL, Shelford T, Edlund CK, Taylor JA, Field JK, Park SK, Offit K, Thomassen M, Schmutzler R, Ottini L, Hung RJ, Marchini J, Amin AI, Olama A, Peters U, Eeles RA, Seldin MF, Gillanders E, Seminara D, Antoniou AC, Pharoah PDP, Chenevix-Trench G, Chanock SJ, Simard J, Easton DF. The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev* 2017;26:126–135.
31. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigó R, Hubbard TJ. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res* 2012;22:1760–1774.
  32. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 2003;100:9440–9445.
  33. Urbut SM, Wang G, Carbonetto P, Stephens M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* 2019;51:187–195.
  34. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, Conti DV, Qu C, Jeon J, Edlund CK, Greenside P, Wainberg M, Schumacher FR, Smith JD, Levine DM, Nelson SC, Sinnott-Armstrong NA, Albanes D, Alonso MH, Anderson K, Arnau-Collell C, Arndt V, Bamia C, Banbury BL, Baron JA, Berndt SI, Bézieau S, Bishop DT, Boehm J, Boeing H, Brenner H, Brezina S, Buch S, Buchanan DD, Burnett-Hartman A, Butterbach K, Caan BJ, Campbell PT, Carlson CS, Castellví-Bel S, Chan AT, Chang-Claude J, Chanock SJ, Chirlaque M-D, Cho SH, Connolly CM, Cross AJ, Cuk K, Curtis KR, de la Chapelle A, Doheny KF, Duggan D, Easton DF, Elias SG, Elliott F, English DR, Feskens EJM, Figueiredo JC, Fischer R, FitzGerald LM, Forman D, Gala M, Gallinger S, Gauderman WJ, Giles GG, Gillanders E, Gong J, Goodman PJ, Grady WM, Grove JS, Gsur A, Gunter MJ, Haile RW, Hampe J, Hampel H, Harlid S, Hayes RB, Hofer P, Hoffmeister M, Hopper JL, Hsu W-L, Huang W-Y, Hudson TJ, Hunter DJ, Ibañez-Sanz G, Idos GE, Ingersoll R, Jackson RD, Jacobs EJ, Jenkins MA, Joshi AD, Joshi CE, Keku TO, Key TJ, Kim HR, Kobayashi E, Kolonel LN, Kooperberg C, Kühn T, Küry S, Kweon S-S, Larsson SC, Laurie CA, Le Marchand L, Leal SM, Lee SC, Lejbkovicz F, Lemire M, Li CI, Li L, Lieb W, Lin Y, Lindblom A, Lindor NM, Ling H, Louie TL, Männistö S, Markowitz SD, Martín V, Masala G, McNeil CE, Melas M, Milne RL, Moreno L, Murphy N, Myte R, Naccarati A, Newcomb PA, Offit K, Ogino S, Onland-Moret NC, Pardini B, Parfrey PS, Pearlman R, Perduca V, Pharoah PDP, Pinchev M, Platz EA, Prentice RL, Pugh E, Raskin L, Rennert G, Rennert HS, Riboli E, Rodríguez-Barranco M, Romm J, Sakoda LC, Schafmayer C, Schoen RE, Seminara D, Shah M, Shelford T, Shin M-H, Shulman K, Sieri S, Slattery ML, Southey MC, Stadler ZK, Stegmaier C, Su Y-R, Tangen CM, Thibodeau SN, Thomas DC, Thomas SS, Toland AE, Trichopoulou A, Ulrich CM, Van Den Berg DJ, van Duijnhoven FJB, Van Guelpen B, van Kranen H, Vijai J, Visvanathan K, Vodicka P, Vodickova L, Vymetalkova V, Weigl K, Weinstein SJ, White E, Win AK, Wolf CR, Wolk A, Woods MO, Wu AH, Zaidi SH, Zanke BW, Zhang Q, Zheng W, Scacheri PC, Potter JD, Bassik MC, Kundaje A, Casey G, Moreno V, Abecasis GR, Nickerson DA, Gruber SB, Hsu L, Peters U. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* 2019;51:76–87.
  35. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet* 2017;13:e1006646.
  36. Breschi A, Muñoz-Aguirre M, Wucher V, Davis CA, Garrido-Martín D, Djebali S, Gillis J, Pervouchine DD, Vlasova A, Dobin A, Zaleski C, Drenkow J, Danyko C, Scavelli A, Reverter F, Snyder MP, Gingeras TR, Guigó R. A limited set of transcriptional programs define major cell types. *Genome Res* 2020;30:1047–1059.
  37. Glebov OK, Rodriguez LM, Nakahara K, Jenkins J, Cliatt J, Humbyrd C-J, DeNobile J, Soballe P, Simon R, Wright G, Lynch P, Patterson S, Lynch H, Gallinger S, Buchbinder A, Gordon G, Hawk E, Kirsch IR. Distinguishing right from left colon by the pattern of gene expression. *Cancer Epidemiol Biomarkers Prev* 2003;12:755–762.
  38. Puccini A, Marshall JL, Salem ME. Molecular variances between right- and left-sided colon cancers. *Curr Colorectal Cancer Rep* 2018;14:152–158.
  39. Zhang L, Zhao Y, Dai Y, Cheng J-N, Gong Z, Feng Y, Sun C, Jia Q, Zhu B. Immune landscape of colorectal cancer tumor microenvironment from different primary tumor location. *Front Immunol* 2018;9:1578.
  40. Stintzing S, Tejpar S, Gibbs P, Thiebach L, Lenz H-J. Understanding the role of primary tumour localisation in colorectal cancer treatment and outcomes. *Eur J Cancer* 2017;84:69–80.
  41. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, Herbst RH, Rogel N, Slyper M, Waldman J, Sud M, Andrews E, Velonias G, Haber AL, Jagadeesh K, Vickovic S, Yao J, Stevens C, Dionne D, Nguyen LT, Villani A-C, Hofree M, Creasey EA, Huang H, Rozenblatt-Rosen O, Garber JJ, Khalili H, Desch AN, Daly MJ, Ananthakrishnan AN, Shalek AK, Xavier RJ, Regev A. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 2019;178:714–730.e22.
  42. Parikh K, Antanaviciute A, Fawcner-Corbett D, Jagielowicz M, Aulicino A, Lagerholm C, Davis S, Kinchen J, Chen HH, Alham NK, Ashley N, Johnson E, Hublitz P, Bao L, Lukomska J, Andev RS, Björklund E, Kessler BM, Fischer R, Goldin R, Koohy H, Simmons A. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* 2019;567:49–55.

43. James KR, Gomes T, Elmentaite R, Kumar N, Gulliver EL, King HW, Stares MD, Bareham BR, Ferdinand JR, Petrova VN, Polański K, Forster SC, Jarvis LB, Suchanek O, Howlett S, James LK, Jones JL, Meyer KB, Clatworthy MR, Saeb-Parsy K, Lawley TD, Teichmann SA. Distinct microbial and immune niches of the human colon. *Nat Immunol* 2020;21:343–353.
44. Kim YS, Ho SB. Intestinal goblet cells and mucins in health and disease: recent insights and progress. *Curr Gastroenterol Rep* 2010;12:319–330.
45. Amirkhah R, Naderi-Meshkin H, Shah JS, Dunne PD, Schmitz U. The intricate interplay between epigenetic events, alternative splicing and noncoding RNA deregulation in colorectal cancer. *Cells* 2019;8:929.
46. Cohen AJ, Saiakhova A, Corradin O, Luppino JM, Lovrenert K, Bartels CF, Morrow JJ, Mack SC, Dhillon G, Beard L, Myeroff L, Kalady MF, Willis J, Bradner JE, Keri RA, Berger NA, Pruett-Miller SM, Markowitz SD, Scacheri PC. Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat Commun* 2017;8:14400.
47. Bushnell B. BBtools. BBMap short read aligner, and other bioinformatic tools. Available from: [sourceforge.net/projects/bbmap](https://sourceforge.net/projects/bbmap). Accessed December 2019.
48. Andrews S. FastQC: a quality control tool for high throughput sequence data 2010. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>. Accessed December 2019.
49. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21.
50. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;12:323.
51. O’Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, McQuillan R, Fraser RM, Campbell H, Polasek O, Asiki G, Ekoru K, Hayward C, Wright AF, Vitart V, Navarro P, Zagury J-F, Wilson JF, Toniolo D, Gasparini P, Soranzo N, Sandhu MS, Marchini J. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet* 2014;10:e1004234.
52. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, Schlessinger D, Stambolian D, Loh P-R, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, Fuchsberger C. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48:1284–1287.
53. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2015;43:D6–D17.
54. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.
55. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–140.
56. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43:e47.
57. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;8:1826.
58. Ongen H, Buil A, Brown AA, Dermitzakis ET, Delaneau O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* 2016;32:1479–1485.
59. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* 2012;7:500–507.
60. John D, Storey Andrew J; Bass ADDR. qvalue. Q-value estimation for false discovery rate control, 2018. Available from: <http://github.com/jdstorey/qvalue>. Accessed November 2020.
61. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. *Genome Biol* 2016;17:122.
62. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA, Young RA. Super-enhancers in the control of cell identity and disease. *Cell* 2013;155:934–947.
63. The ENCODE Project Consortium. A user’s guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011;9:e1001046.
64. Yang Y-CT, Di C, Hu B, Zhou M, Liu Y, Song N, Li Y, Umetsu J, Lu ZJ. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* 2015;16:51.
65. Schmidt EM, Zhang J, Zhou W, Chen J, Mohlke KL, Chen YE, Willer CJ. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* 2015;31:2601–2606.
66. Bulik-Sullivan BK, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015;47:291–295.
67. Wen X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann Appl Stat* 2016;10:1619–1638.
68. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* 2016;32:283–285.
69. Wen X, Lee Y, Luca F, Pique-Regi R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am J Hum Genet* 2016;98:1114–1129.
70. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. Shiny: web application framework for R. 2018. Available from: <https://CRAN.R-project.org/package=shiny>. Accessed November 2020.

---

Received July 2, 2020. Accepted February 8, 2021.

**Correspondence**

Address correspondence to: Graham Casey, PhD, Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia,

MSB Room 3238, PO Box 800717, Charlottesville, Virginia 22908-0717. e-mail: [gc8r@virginia.edu](mailto:gc8r@virginia.edu); or Victor Moreno, MD, Catalan Institute of Oncology, Oncology Data Analytics Program, Hospital Duran i Reynals, Gran Via de L'Hospitalet, 199-203, 08908 L'Hospitalet de Llobregat, Barcelona, Spain. e-mail: [v.moreno@iconcologia.net](mailto:v.moreno@iconcologia.net).

#### Acknowledgments

The authors thank the "Centres de Recerca de Catalunya" (CERCA) Program, Generalitat de Catalunya for institutional support. The authors particularly acknowledge the patients participating in this study, the endoscopy units from the Bellvitge University Hospital and the Viladecans Hospital, as well as Carmen Atencia, Judith Rocamora, Susana Lopez, Gemma Aiza, and the Biobank, Bellvitge University Hospital, Catalan Institute of Oncology Bellvitge Biomedical Research Institute (HUB-ICO-IDIBELL) (PT17/0015/0024) for their collaboration. RNA-Seq was provided by the Genomics Core Facility of the Case Western Reserve University (CWRU) School of Medicine's Genetics and Genome Sciences Department as well as the Northwest Genomics Center at the University of Washington. Colon artwork in the CoTrEx logo is designed by Smashicons from Flaticon (Málaga, Spain).

#### CRediT Authorship Contributions

Virginia Díez-Obrero (Data curation: Lead; Formal analysis: Lead; Software: Lead; Visualization: Lead; Writing – original draft: Lead)

Christopher H Dampier (Data curation: Lead; Formal analysis: Equal; Writing – original draft: Lead; Writing – review & editing: Lead)

Ferran Moratalla-Navarro (Data curation: Equal; Formal analysis: Equal; Software: Equal; Writing – review & editing: Equal)

Matthew Devall (Data curation: Equal; Formal analysis: Equal; Writing – review & editing: Equal)

Sarah J Plummer (Data curation: Equal; Resources: Lead; Writing – review & editing: Equal)

Anna Díez-Villanueva (Formal analysis: Equal; Software: Equal; Writing – review & editing: Equal)

Ulrike Peters (Funding acquisition: Equal; Resources: Equal; Supervision: Equal; Writing – review & editing: Equal)

Stephanie Bien (Supervision: Equal; Writing – review & editing: Equal)

Jeroen R Huyghe (Supervision: Equal; Writing – review & editing: Equal)

Anshul Kundaje (Supervision: Equal)

Gemma Ibáñez-Sanz (Resources: Lead; Writing – review & editing: Equal)  
Elisabeth Guinó (Data curation: Lead)

Mireia Obón-Santacana (Data curation: Equal; Writing – review & editing: Equal)

Robert Carreras-Torres (Conceptualization: Equal; Software: Equal; Supervision: Equal; Writing – original draft: Lead; Writing – review & editing: Lead)

Graham Casey (Conceptualization: Lead; Funding acquisition: Lead; Resources: Lead; Supervision: Lead; Writing – review & editing: Equal)

Victor Moreno (Conceptualization: Lead; Funding acquisition: Lead; Resources: Lead; Software: Equal; Supervision: Lead; Writing – review & editing: Equal)

#### Conflicts of interest

The authors disclose no conflicts.

#### Funding

Supported by the Agency for Management of University and Research Grants of the Catalan Government grants 2017SGR723; the Instituto de Salud Carlos III, co-funded by European Regional Development Fund (FEDER) funds "A Way to Build Europe" grants PI14-00613, PI17-00092; the Spanish Association Against Cancer Scientific Foundation grant GCTRA18022MORE; Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP) CB07/02/2005; and the National Institutes of Health grants R01CA204279, R01CA143237, and R01CA201407. Also supported by EU H2020 - Marie Skłodowska-Curie (MSC) grant 796216 (R.C.T.); a postdoctoral fellowship through the "Fundación Científica de la Asociación Española Contra el Cáncer (AECC)" (M.O.S.); the National Institutes of Health training grant T32 5T32CA163177-07 (C.H.D.); and the Ministerio de Universidades through predoctoral fellowship number FPU16/00599 for the "Formación del Profesorado Universitario" (V.D.O.). A sample collection of this work was supported by the Xarxa de Bancs de Tumors de Catalunya (XBTC) sponsored by Pla Director d'Oncologia de Catalunya, "Plataforma Biobancos PT13/0010/0013," and the Biobank of the Catalan Institute of Oncology (ICOBIOBANC), sponsored by the Catalan Institute of Oncology. This work was supported in part by National Institutes of Health/ National Cancer Institute grants CA143237 and CA204279 (G.C.).

**Supplementary Materials** are provided online, available at:

[https://www.cmghjournal.org/article/S2352-345X\(21\)00036-9/fulltext#supplementaryMaterial](https://www.cmghjournal.org/article/S2352-345X(21)00036-9/fulltext#supplementaryMaterial)

- Supplementary Tables (large)

<https://www.cmghjournal.org/cms/10.1016/j.jcmgh.2021.02.003/attachment/26aa43ac-2917-4c50-9fc5-abdca66dcb5f/mmc1.xlsx>

- Supplementary Data (large)

<https://www.cmghjournal.org/cms/10.1016/j.jcmgh.2021.02.003/attachment/8eb502ca-a08f-4e1e-8b14-7cd49390fb75/mmc2.xlsx>

#### **4.2. The Colon Transcriptome Explorer (CoTrEx) 2.0.**

The second objective of the Thesis was “to develop a web resource to explore population-based normal colon transcriptome profiles, e/sQTLs, gene expression prediction models, as well as to annotate SNPs with eQTLs”.

To address this objective, we developed the article entitled “The Colon Transcriptome Explorer (CoTrEx) 2.0, a reference resource for exploring population-based normal colon gene expression”.

# The Colon Transcriptome Explorer (CoTrEx) 2.0: a reference web-based resource for exploring population-based normal colon gene expression

Virginia Díez-Obrero<sup>1,2,3,4</sup>, Ferran Moratalla-Navarro<sup>1,3,4</sup>, Christopher Heaton Dampier<sup>5,6</sup>, Matthew Devall<sup>5,6</sup>, Robert Carreras-Torres<sup>1,2,3</sup>, Graham Casey<sup>5,6</sup> and Victor Moreno<sup>1,2,3,4,\*</sup>

<sup>1</sup> Oncology Data Analytics Program, Catalan Institute of Oncology (ICO). L'Hospitalet de Llobregat, Barcelona, Spain.

<sup>2</sup> Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL). L'Hospitalet de Llobregat, Barcelona, Spain.

<sup>3</sup> Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Spain.

<sup>4</sup> Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain.

<sup>5</sup> Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA.

<sup>6</sup> Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA.

\* Correspondence: v.moreno@iconcologia.net; Tel.: +34 932 607 434

**Abstract:** Gene expression data is key for the functional annotation of single nucleotide polymorphisms (SNPs) identified in genome-wide association studies (GWAS). Expression and splicing quantitative trait loci (e/sQTLs) in normal colon tissue, such as those from the University of Barcelona and University of Virginia RNA sequencing project (BarcUVa-Seq) and the Genotype-Tissue Expression project (GTEx), are required to gain biological insight of colon-related diseases risk loci. Moreover, transcriptome-wide association studies (TWAS) rely on reference gene expression imputation panels in the tissue of interest to nominate susceptibility genes. Also, it is of high interest to study the relationships between genes in a network framework. For facilitating these analyses, we have updated and expanded the scope of the Colon Transcriptome Explorer (CoTrEx) to the version 2.0. This web-based resource provides exhaustive visualization and analysis of transcriptome-wide gene expression profiles of normal colon tissue from BarcUVa-Seq and GTEx. In addition to the integration of new datasets, CoTrEx 2.0 provides additional e/sQTLs sets, as well as gene expression prediction models and regulatory and co-expression networks. It is freely available at <https://barcuvaseq.org/cotrex/>. Overall, it is of high interest for researchers aiming to investigate the genetic susceptibility to colon-related complex traits and diseases.

**Keywords:** RNA-Seq; bioinformatics; web application; gene expression; alternative splicing; visualization; molecular epidemiology

## 1. Introduction

Datasets of both blood DNA genotyping and RNA sequencing (RNA-Seq) of biopsy samples from a large number of healthy individuals are valuable resources for studies in molecular epidemiology. For example, they provide expression and splicing quantitative trait loci (e/sQTLs) for the annotation of genome-wide association studies (GWAS)-identified risk single nucleotide polymorphisms (SNPs) and gene expression prediction models for transcriptome-wide association studies (TWAS). In this sense, the University of Barcelona and University of Virginia genotyping and sequencing project (BarcUVa-Seq) provided gene expression and alternative splicing profiles of normal (i.e. non-neoplastic, without lesions) colon biopsies from ascending (N=138), transverse (N=143) and descending (N=164) subsites. The expression profiles and their association statistics with



germline genetic variants, i.e. e/sQTLs, were recently reported and included in the initial version of the Colon Transcriptome Explorer (CoTrEx) [1]. Additionally, the Genotype-Tissue Expression (GTEx) project provided normal colon e/sQTLs from transverse (N=368) and sigmoid (N=318) colon samples from corpses [2]. Although the gene expression and related information is provided as supplementary material or deposited in public online repositories, it is often difficult and time consuming for researchers to access the data and analyze and visualize their gene of interest, especially for non-bioinformaticians.

In this article we present the CoTrEx 2.0, an interactive web resource that facilitates the exhaustive visualization and analysis of normal colon gene expression and alternative splicing data from BarcUVa-Seq and GTEx projects. This version, in addition to incorporating GTEx colon datasets and new customization options, provides additional e/sQTL sets, a SNP annotation tool, prediction models statistics for gene expression imputation, and regulatory and gene co-expression networks.

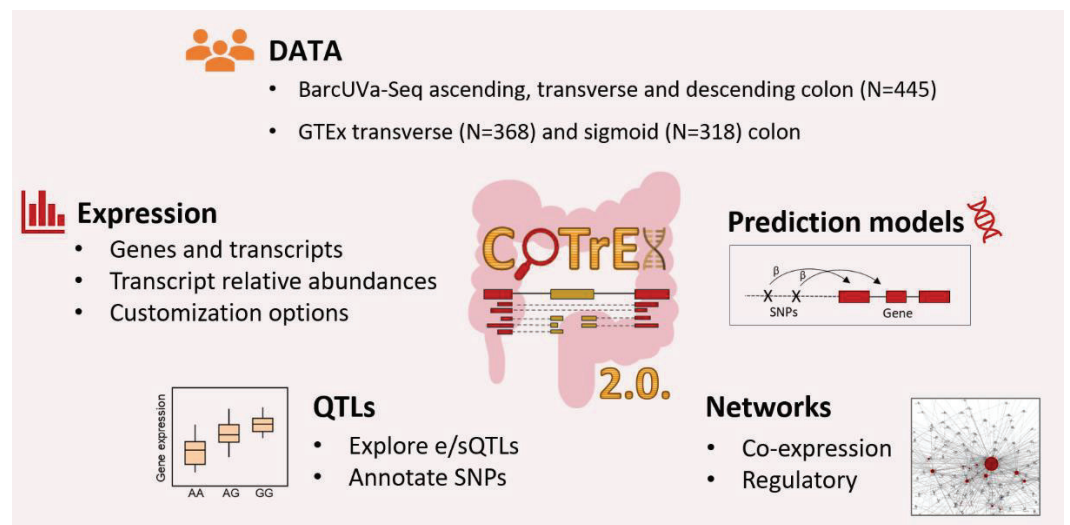
## 2. Description of CoTrEx 2.0

CoTrEx 2.0 is a web-based resource that includes normal colon gene expression data from BarcUVa-Seq and GTEx projects (see schema in Figure 1). Its main components are divided in the “Expression”, “QTLs”, “Prediction models”, and “Networks” tabs.

On the “Expression” tab, users can search for a gene of interest, select one or more associated transcripts and visualize their expression in multiple ways. On the left panel, the following options are available: i) filter the input data by sex, age and colon anatomic location, ii) select specific visualization features (e.g. heatmap, PCA plot), and iii) group transcripts by relative abundance according to a selected expression threshold (i.e. if 0.05 is selected, the lowest expressed 5% of transcripts is grouped in a single category labeled “Other transcripts”). On the main panel, a customizable stripchart and a barplot are displayed. For example, points in the stripchart can be colored by covariates of interest, and transcript expression can be hidden to show only the expression of selected genes. Annotation by covariate is also available for heatmaps and PCA plots.

On the “QTLs” tab, users can explore lists of significant colon e/sQTLs, including summary statistics and customizable plots showing the distribution of gene expression/percent splicing index by SNP genotype. Users can also search for association statistics for SNPs of interest by selecting the “Annotate SNPs” option. The “Prediction models” tab includes elastic net-based gene expression prediction models for the entire colon and by colon subsite. Descriptive statistics of the prediction models and the SNP weights can be obtained for a gene of interest.

On the “Networks” tab, by selecting the “Regulatory network” option, users can explore gene interactions between TFs and regulated target genes in a network. Arrows are directed from TFs to target genes (either TFs or non-TFs). It is possible to explore first and second order step neighbors by selecting the corresponding option. Descriptive and topological network parameters are provided in tables, including the mutual information (MI) values for each interaction, which indicate the strength of an interaction. The weighted correlation network analysis (WGCNA) approach [3] was used for exploring patterns of correlated gene expression in a gene co-expression network framework. This method makes groups of highly interconnected genes called modules. A total of 20 modules with a mean of 777 highly correlated genes per module were defined, each of them labelled with a color name. The gene-module assignments can be downloaded, and hierarchical clusters of all modules can be explored.



**Figure 1.** CoTrEx 2.0 schematic.

### 3. Discussion

We have updated and expanded the scope of CoTrEx to the newest version 2.0, including new data and functionalities. This version includes gene expression and alternative splicing-related data from the GTEx v8 transverse and sigmoid colon. In this version, the genes and transcripts visualized on the Expression tab can be filtered or colored according to the individuals' age and sex. Also, the expression statistics associated with the selected samples can be retrieved. Transcripts can be grouped by relative abundance and hierarchical clustering can be observed in a heatmap. These features are not provided by the GTEx Transcript Browser [4]. In addition, we provide a SNP annotation tool on the QTLs tab where users can provide a list of SNPs of interest to explore associations with genes located up to 1Mb of distance. In contrast, the GTEx eQTL Calculator [5] requires that the users provide the gene ID in addition to the SNP of interest. This is not convenient in cases where the user wants to explore SNP-gene associations of all genes nearby a SNP of interest. Also, the gene expression prediction models can be downloaded from the Prediction models tab, which are useful for investigators interested in performing TWAS and nominating candidate susceptibility genes for a phenotype of interest. A list of complex traits and diseases for which the gene expression prediction models provided in CoTrEx 2.0 are relevant for TWAS is provided elsewhere [1]. Future developments of CoTrEx 2.0 would include additional QTL sets generated, such as regulatory QTLs, associated with changes in interactions between genes.

In conclusion, CoTrEx 2.0 facilitates a quick and centralized access to explore and analyze the most up to date reference gene expression and splicing profiles for non-neoplastic human colon tissue, and their associations with germline genetic variants, which facilitates the understanding of the transcriptomic basis of this tissue. Finally, the CoTrEx 2.0 is a valuable resource for researchers interested in annotating risk loci identified in colon-related GWAS, in performing TWAS for colon-related diseases, and in unraveling the mechanisms underlying inherited susceptibility to colon-related diseases.

### 4. Materials and Methods

CoTrEx 2.0 was built with the R platform Shiny [6]. Gene and transcript expression counts and e/sQTLs of GTEx v8 sigmoid and transverse colon were obtained from the database of Genotypes and Phenotypes (dbGaP) at <http://www.ncbi.nlm.nih.gov/gap> through dbGaP accession number phs000424.v8.p2. Genes with at least 6 counts in more than 20% of the samples were provided. Expression counts were transformed to trimmed mean of M-values (TMMs). Gene expression prediction models of GTEx v8 were obtained



from elsewhere [1,7] (see data availability statement). Gene expression prediction models of BarcUVa-Seq were generated for the whole sample size and for subsets of the data according to the anatomic location where the biopsies were collected (ascending, transverse and descending colon). The elastic net-based models were generated following the PredictDB pipeline, which was the one used for GTEx v8 data [7]. Following this pipeline, we considered significant gene models those with a predictive performance  $P < 0.05$  and  $R^2 > 0.1$ . Gene expression data was adjusted for sex, sequencing batch, probabilistic estimation of expression residuals [PEER] factors [8] and genetic ancestry (2 principal components).

The BC3net R package [9] was used to generate weighted directed gene regulatory networks between 2,195 transcription factors (TFs) and 8,785 target genes. TFs were chosen according to three GO annotations: GO:0045449 “regulation of transcription”, GO:0001071 “Nucleic acid binding transcription factor activity”, and GO:0140110 “transcription regulator activity”. A total of 1,000 bootstraps were run to get a robust final network. Finally, the weighted correlation network analysis (WGCNA) was performed with the WGCNA R package [3]. A soft thresholding of 6 was selected to approximate to scale free topology.

**Author Contributions:** Conceptualization, V.M. and V.D.; methodology, V.D, F.M., R.C.; software, V.D.; data curation, V.D., F.M., C.D., M.D.,R.C; writing—original draft preparation, V.D.; writing—review and editing, V.D., F.M., C.D., M.D., R.C., G.C., V.M.; visualization, V.D.; supervision, V.M.; funding acquisition, V.M., G.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Institutes of Health, grant numbers R01 CA204279, R01 CA143237 and R01 CA201407; the Agency for Management of University and Research Grants (AGAUR) of the Catalan Government, grant number 2017SGR723; the Instituto de Salud Carlos III, co-funded by FEDER funds –a way to build Europe, grant numbers PI14-00613, PI17-00092; the Spanish Association Against Cancer (AECC) Scientific Foundation, grant number GCTRA18022MORE; and the Centro de investigación biomédica en red. Epidemiología y salud pública (CIBERESP), grant number CB07/02/2005. RCT received funding through the EU H2020 – MSC, grant number 796216; CHD received funding through the National Institutes of Health, grant number T32 5T32CA163177-07; VDO received founding through the Spanish “Ministerio de Educación, Cultura y Deporte”, grant number FPU16/00599.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** GTEx v8 sigmoid and transverse colon data were obtained from the database of Genotypes and Phenotypes (dbGaP) at <http://www.ncbi.nlm.nih.gov/gap> through dbGaP accession number phs000424.v8.p2. GTEx v8 gene expression prediction models were obtained from Zenodo, at <https://dx.doi.org/10.5281/zenodo.3519321>.

**Acknowledgments:** We thank the CERCA Program, Generalitat de Catalunya, for institutional support.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Díez-Obrero, V.; Dampier, C.H.; Moratalla-Navarro, F.; Devall, M.; Plummer, S.J.; Díez-Villanueva, A.; Peters, U.; Bien, S.; Huyghe, J.R.; Kundaje, A.; et al. Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci. *Cell Mol Gastroenterol Hepatol* **2021**, doi:10.1016/j.jcmgh.2021.02.003.
- GTEx Consortium The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* **2020**, *369*, 1318–1330.
- Langfelder, P.; Horvath, S. WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* **2008**, *9*.
- GTEx Transcript Browser Available online: <https://gtexportal.org/home/transcriptPage> (accessed on 17 May 2021).
- GTEx eQTL Calculator Available online: <https://gtexportal.org/home/testyourown> (accessed on 17 May 2021).
- Chang, W.; Cheng, J.; Allaire, J.J.; Sievert, C.; Schloerke, B.; Xie, Y.; Allen, J.; McPherson, J.; Dipert, A.; Borges, B. Shiny: Web Application Framework for R; 2021.
- Barbeira, A.N.; Liang, Y.; Bonazzola, R.; Wang, G.; Wheeler, H.E.; Melia, O.J.; Aguet, F.; Ardlie, K.G.; Wen, X.; Im, H.K.; et al. Fine-Mapping and QTL Tissue-Sharing Information Improve Causal Gene Identification and Transcriptome Prediction Performance. *bioRxiv* 2020:2020.03.19.997213. Doi: 10.1101/2020.03.19.997213.

- 
8. Stegle, O.; Parts, L.; Piipari, M.; Winn, J.; Durbin, R. Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses. *Nat. Protoc.* **2012**, *7*, 500–507. 185  
186
  9. de Matos Simoes, R.; Emmert-Streib, F. Bagging Statistical Network Inference from Large-Scale Gene Expression Data. *PLoS One* **2012**, *7*, e33624. 187  
188  
189

### **4.3. Transcription-Wide Association Study for Inflammatory Bowel Disease.**

The third objective of this Thesis was “to propose candidate genes whose genetically regulated gene expression is associated with IBD, including genes in specific colon subsites; with emphasis on gene expression markers of colon cell types, and gene enrichment in IBD therapy-related molecular pathways. Also, identify candidate susceptibility genes specific for the epithelial, immune/blood, mesenchymal and neural tissue categories”.

To address this objective we developed the article entitled “Transcriptome-wide association study for inflammatory bowel disease reveals novel candidate susceptibility genes in specific colon subsites and tissue categories”.



Original Article

# Transcriptome-Wide Association Study for Inflammatory Bowel Disease Reveals Novel Candidate Susceptibility Genes in Specific Colon Subsites and Tissue Categories

Virginia Díez-Obrero,<sup>a,b,c,d</sup> Ferran Moratalla-Navarro,<sup>a,c,d</sup>  
Gemma Ibáñez-Sanz,<sup>a,b,c,e</sup> Jordi Guardiola,<sup>e,○</sup>  
Francisco Rodríguez-Moranta,<sup>e</sup> Mireia Obón-Santacana,<sup>a,b,c</sup>  
Anna Díez-Villanueva,<sup>a,b,c</sup> Christopher Heaton Dampier,<sup>f,g</sup>  
Matthew Devall,<sup>f,g,○</sup> Robert Carreras-Torres,<sup>a,b,c</sup> Graham Casey,<sup>f,g</sup>  
Victor Moreno<sup>a,b,c,d,○</sup>

<sup>a</sup>Oncology Data Analytics Program, Catalan Institute of Oncology (ICO), L'Hospitalet de Llobregat, Barcelona, Spain  
<sup>b</sup>ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain  
<sup>c</sup>Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain <sup>d</sup>Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain <sup>e</sup>Gastroenterology Department, Bellvitge University Hospital, L'Hospitalet de Llobregat, Spain <sup>f</sup>Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA <sup>g</sup>Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA

Corresponding author: Dr Victor Moreno, Catalan Institute of Oncology, Oncology Data Analytics Program, Hospital Duran i Reynals, Gran Via de l'Hospitalet, 199–203, 08908 L'Hospitalet de Llobregat (Barcelona) Spain. Tel: +34 932 607 434; Email: [v.moreno@iconcologia.net](mailto:v.moreno@iconcologia.net)

## Abstract

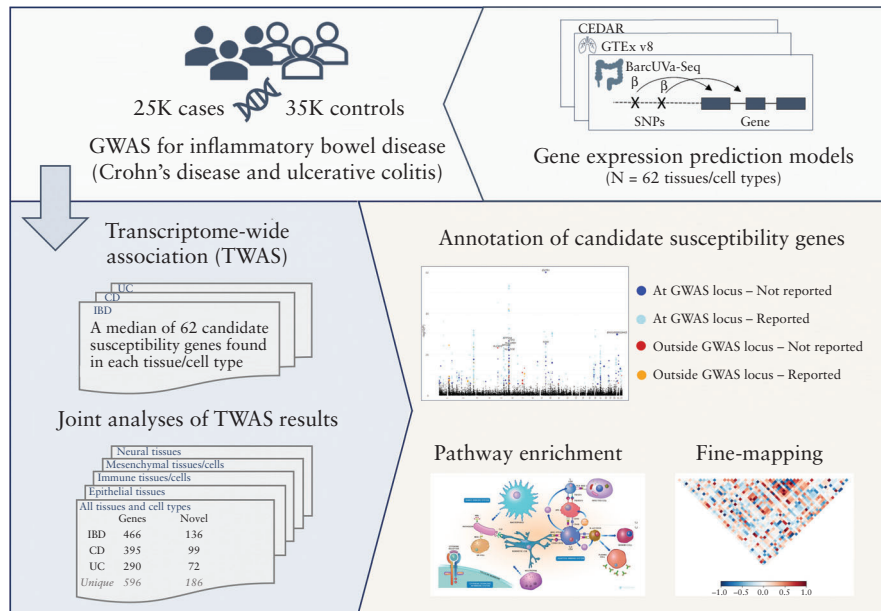
**Background and Aims:** Genome-wide association studies [GWAS] for inflammatory bowel disease [IBD] have identified 240 risk variants. However, the benefit of understanding the genetic architecture of IBD remains to be exploited. Transcriptome-wide association studies [TWAS] associate gene expression with genetic susceptibility to disease, providing functional insight into risk loci. In this study, we integrate relevant datasets for IBD and perform a TWAS to nominate novel genes implicated in IBD genetic susceptibility.

**Methods:** We applied elastic net regression to generate gene expression prediction models for the University of Barcelona and University of Virginia RNA sequencing project [BarcUVa-Seq] and correlated expression and disease association research [CEDAR] datasets. Together with Genotype-Tissue Expression project [GTEx] data, and GWAS results from about 60 000 individuals, we employed Summary-PrediXcan and Summary-MultiXcan for single and joint analyses of TWAS results, respectively.

**Results:** BarcUVa-Seq TWAS revealed 39 novel genes whose expression in the colon is associated with IBD genetic susceptibility. They included expression markers for specific colon cell types. TWAS meta-analysis including all tissues/cell types provided 186 novel candidate susceptibility genes. Additionally, we identified 78 novel susceptibility genes whose expression is associated with IBD exclusively in immune ( $N = 19$ ), epithelial ( $N = 25$ ), mesenchymal ( $N = 22$ ) and neural

( $N = 12$ ) tissue categories. Associated genes were involved in relevant molecular pathways, including pathways related to known IBD therapeutics, such as tumour necrosis factor signalling. **Conclusion:** These findings provide insight into tissue-specific molecular processes underlying IBD genetic susceptibility. Associated genes could be candidate targets for new therapeutics and should be prioritized in functional studies.

## Graphical Abstract



**Key Words:** Transcriptome-wide association study; genetic susceptibility; gene expression

## 1. Introduction

Inflammatory bowel disease [IBD] is a chronic inflammatory disorder of the gastrointestinal tract that encompasses two main disease subtypes, namely Crohn's disease [CD] and ulcerative colitis [UC]. IBD is caused by immune dysregulation and aberrant inflammatory responses to gut microbiota that result in tissue damage. Clinical manifestations of CD are more heterogeneous than those of UC. UC is restricted to the large intestine, whereas CD can affect any part of the gastrointestinal tract and involves the colon in only 25% of cases.<sup>1,2</sup>

Germline genetic variants have been associated with IBD susceptibility. The largest genome-wide association study [GWAS] for IBD identified 240 independent risk single nucleotide polymorphisms [SNPs].<sup>3</sup> Some of them have been functionally characterized and found to affect established mechanisms of IBD pathogenesis, including impaired autophagy, interleukin [IL]-17/IL-23 axis/type 3 innate lymphoid cells, and failure to suppress aberrant immune responses.<sup>4</sup> GWAS SNPs have also been associated with genes whose encoded proteins participate in pathways targeted by approved IBD therapies such as infliximab and adalimumab, which are monoclonal antibodies that modulate tumour necrosis factor [TNF] signalling.<sup>3</sup> Despite these successes the mechanisms by which GWAS-identified genetic variants, especially non-coding variants, confer susceptibility are not yet fully understood.<sup>2</sup>

The hypothesis that risk SNPs modify expression of nearby genes and influence development of disease is supported by recent work

in multiple tissues.<sup>2</sup> One recently published study<sup>5</sup> presented a large gene expression dataset from normal colon tissue and showed strong evidence that genetically regulated gene expression in the colon is involved in IBD genetic susceptibility. Another study<sup>6</sup> related genetic risk variants to gene expression in circulating immune cells to identify IBD susceptibility genes.

Sequencing RNA from multiple tissues/cell types of thousands of subjects with and without IBD to associate gene expression with disease is costly and not feasible for some tissues. In addition, this approach cannot distinguish whether altered gene expression is a cause rather than a consequence of disease. A solution to these limitations is provided by the transcriptome-wide association study [TWAS] statistical approach, which permits prediction of gene expression from genetic data, and thereby enables imputation of gene expression for subjects included in GWAS. TWAS uses reference imputation panels (i.e. predictive models generated from population-based germline genotype and tissue-specific gene expression data) to associate genetically regulated gene expression with traits and diseases. The TWAS approach provides biological context for interpreting disease risk loci by nominating candidate susceptibility genes not only at GWAS risk regions but also at other potential regions that current GWAS have not been powered to detect.<sup>7</sup>

Previous TWAS for IBD<sup>8-10</sup> reported candidate susceptibility genes based on prior versions of the Genotype-Tissue Expression project [GTEx],<sup>11</sup> and the only study that analysed the latest version (v8) of GTEx<sup>9</sup> did not provide results for CD. The University

of Barcelona and University of Virginia RNA sequencing project [BarcUVA-Seq]<sup>5</sup> recently provided a gene expression dataset of colon biopsies across colon subsites. Another source of relevant data for autoimmune diseases is the array-derived correlated expression and disease association research [CEDAR]<sup>6</sup> dataset, which includes gene expression data of circulating immune cell types. To the best of our knowledge, no published study has utilized these reference panels to perform a joint TWAS (i.e. a TWAS meta-analysis that combines TWAS results of individual tissues for increasing the statistical power to find associations) across tissues to strengthen the evidence for genes involved in IBD susceptibility.

In this study, we perform an integrative TWAS analysis to identify novel candidate susceptibility genes whose expression influences IBD pathogenesis. This study includes reference datasets of tissues and blood cell types relevant to IBD and leverages GWAS results for IBD from a dataset including about 25 000 cases and 35 000 controls.<sup>3</sup> We nominate genes whose expression in more than 60 tissues and cell types, including specific colon anatomical subsites (ascending, transverse and descending colon), is associated with IBD and its subtypes (CD and UC). Finally, we assess associations specific for epithelial, immune, mesenchymal and neural tissue categories.

## 2. Materials and Methods

### 2.1. GWAS summary statistics

We downloaded publicly available IBD, UC and CD GWAS summary statistics from a large study including about 60 000 subjects.<sup>3</sup> We performed liftover of SNP coordinates to the GRCh38 genome reference using Crossmap.<sup>12</sup> Reference SNP cluster IDs [rsIDs] were annotated according to dbSNP v151 to match IDs from reference panels.

### 2.2. BarcUVA-Seq data processing

BarcUVA-Seq data<sup>5</sup> include genome-wide genotypes and gene expression from ascending ( $n = 138$ ), transverse ( $n = 143$ ) and descending ( $n = 164$ ) colon. Expression data were processed as described elsewhere.<sup>5</sup> The GENCODE v26 gene model<sup>13</sup> was used to facilitate integration with GTEx v8 data. Genotypes were imputed with the TOPMed (version r2) reference panel on the Michigan Imputation Server.<sup>14</sup> SNPs were filtered by minor allele frequency [MAF] 0.01 and imputation quality (i.e.  $R^2$ ) 0.8. For each panel, we assessed population heterogeneity using 2318 ancestry-informative marker SNPs with the plink *pca* method.<sup>15</sup>

### 2.3. CEDAR data processing

CEDAR data<sup>6</sup> were obtained from the Array Express repository under accession numbers E-MTAB-6666 and E-MTAB-6667 for genotypes and expression data, respectively. The data include gene expression from terminal ileum, transverse colon, rectum, platelets, CD15<sup>+</sup> granulocytes, CD19<sup>+</sup> B lymphocytes, CD8<sup>+</sup> T lymphocytes, CD4<sup>+</sup> T lymphocytes and CD14<sup>+</sup> monocytes. Corresponding sample sizes are provided in [Supplementary Table 1](#). Expression arrays were processed with the *iluminaio* R package.<sup>16</sup> Expression variability between samples was assessed with graphical visualization of expression values in box plots to ensure that no extreme outliers appeared in the dataset. Quantile normalization was performed. Gene annotation was harmonized to GENCODE v26 annotations<sup>13</sup> to facilitate integration with GTEx v8 data. Genotypes were imputed with the Haplotype Reference Consortium panel on the Michigan Imputation Server,<sup>14</sup> and lifted over to the GRCh38 genome reference with

Crossmap.<sup>12</sup> We filtered SNPs by MAF 0.01 and imputation quality (i.e.  $R^2$ ) 0.8. For each panel, we assessed population heterogeneity using 2318 ancestry-informative marker SNPs with the plink *pca* method.<sup>15</sup>

### 2.4. Gene expression prediction models

We downloaded GTEx v8 elastic net regularized regression-based imputation panels ( $N = 49$  tissues/cell types) from PredictDB.<sup>11,17</sup> We generated gene expression prediction models using gastrointestinal tissue and blood cell gene expression data from BarcUVA-Seq (ascending, transverse and descending and ‘any’ colon, where ‘any’ includes all three subsites) and CEDAR (terminal ileum, transverse colon, rectum, platelets, CD15<sup>+</sup> granulocytes, CD19<sup>+</sup> B lymphocytes, CD8<sup>+</sup> T lymphocytes, CD4<sup>+</sup> T lymphocytes and CD14<sup>+</sup> monocytes) datasets, using elastic net regularized regression. CEDAR gene expression was adjusted by sex, age and sequencing batch. BarcUVA-Seq gene expression was adjusted for sex, sequencing batch, probabilistic estimation of expression residuals [PEER] factors<sup>18</sup> and genetic ancestry (two principal components). To be consistent with the PredictDB pipeline followed by the GTEx team for generating the GTEx v8 models,<sup>11,17</sup> we considered significant gene models as those with a predictive performance  $p < 0.05$  and  $R^2 > 0.1$ . Summary statistics and SNP weights of BarcUVA-Seq prediction models were loaded into the Colon Transcriptome Explorer [CoTrEx] 2.0 web resource.<sup>19</sup> Altogether, we compiled a total of 62 reference imputation panels of expression prediction models with a median of 4848 significant genes per panel (ranging from 1003 to 10 013 genes) ([Supplementary Table 1](#)). As expected, the number of significant prediction models increased with the sample size of the imputation panels.

### 2.5. Transcriptome-wide association analyses

The TWAS approach, in a first step, predicts gene expression from genotype data of subjects from whom gene expression has not been measured. This is achieved thanks to tissue-specific gene expression prediction models, i.e. reference imputation panels (see previous subsection ‘Gene expression prediction models’). Next, the inferred gene expression is tested for association with a particular phenotype (e.g. IBD). The Summary-PrediXcan (S-PrediXcan) method<sup>20</sup> used in this study combines the last two steps into one, and therefore does not need individual-level genotype data; instead, it uses the summary parameters of the statistical association between SNPs and the phenotype of interest, commonly referred to as ‘summary statistics’ (see 2.1. Methods sub-section on the summary statistics we used). Along with GWAS summary statistics, it uses the SNP expression weights to impute the expression of a given gene; and uses the variance and covariances of the included SNPs to correct for linkage disequilibrium (LD) biases.<sup>20</sup> Specifically, S-PrediXcan computes a Z-score (Wald statistic) as a measure of the association between predicted gene expression and a phenotype. The main analytical expression used is as follows:

$$Z_g \approx \sum_{l \in \text{Model}_g} \omega_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{se(\hat{\beta}_l)}$$

where  $\omega_{lg}$  is the weight of SNP  $l$  in the prediction of the expression of gene  $g$ ;  $\hat{\beta}_l$  is the GWAS coefficients for SNP  $l$ ;  $se(\hat{\beta}_l)$  is the standard error of  $\hat{\beta}_l$ ;  $\hat{\sigma}_l$  is the estimated variance of SNP  $l$ , and  $\hat{\sigma}_g$  is the estimated variance of the predicted expression of gene  $g$ .<sup>20</sup> We considered as significant those genes that passed Bonferroni correction (0.05/total number of genes).



On the other hand, we used Summary-MultiXcan (S-MultiXcan)<sup>21</sup> for the joint analysis of TWAS results across multiple tissues. Briefly, MultiXcan consists of fitting a linear regression of the phenotype on predicted expression from multiple tissue models jointly.<sup>21</sup> On a similar basis to the S-PrediXcan approach explained above, the MultiXcan framework was extended to be used with GWAS summary statistics. Specifically, S-MultiXcan combines single-tissue S-PrediXcan results, along with LD information from a reference panel for the estimation of their joint effect across tissues on the phenotype.<sup>21</sup> We considered as significant only those genes that passed Bonferroni correction and that had a  $p$ -value  $\leq 10^{-4}$  in the panel with lowest  $p$ , as advised elsewhere,<sup>21</sup> to minimize errors due to LD mismatches.

The categorization of expression panels as epithelial, immune/blood, mesenchymal and neural categories was based on their histological origin (for CEDAR and BarcUVa-Seq datasets), and on their classification above the third quartile of the corresponding categories established by Breschi *et al.*<sup>22</sup> (for GTEx v8 datasets).

## 2.6. Gene annotation

Gene symbols were annotated according to the HUGO Gene Nomenclature Committee.<sup>23</sup> Genes were annotated as novel if they did not appear in the GWAS catalogue genes for IBD,<sup>24</sup> were not indicated in large GWAS previously published elsewhere,<sup>3,25</sup> or did not appear in the TWAS-hub resource for IBD.<sup>8</sup> Genes were annotated at GWAS loci if their transcription start sites were within 1 Mb of any of the top 240 SNPs identified by IBD GWAS.<sup>3</sup> In the case of significant genes predicted using BarcUVa-Seq colon panels, we annotated the cells for which genes were expression markers according to a study by Smillie *et al.* that characterized the colon transcriptome at single-cell resolution.<sup>26</sup>

## 2.7. Fine-mapping

In the context of a TWAS, the fine-mapping approach aims to prioritize candidate genes with higher likelihood of being causal for the association. This is especially important for TWAS-associated loci with multiple genes, where the correlation of expression between genes tend to be high and which might bias the results, in a similar manner as LD does with GWAS-identified SNPs. To address this topic, probabilistic fine-mapping was performed using the fine-mapping of causal gene sets [FOCUS] approach.<sup>27</sup> FOCUS provides fine-mapping at each of the TWAS-identified loci by integrating GWAS summary statistic data, the SNP expression weights for each tissue and LD-related statistics among all SNPs in each locus. Specifically, it applies a probabilistic framework to assign to every gene in a given TWAS-associated locus a posterior probability [PIP] that indicates the likelihood of a given gene to explain the observed TWAS association signal.<sup>27</sup> We used the FOCUS software with default parameters and provided FOCUS with genes passing Bonferroni correction in TWAS analyses, and considered as probably causal those genes included in a credible set with a nominal confidence of 90% and with a PIP > 0.5.

## 2.8. Pathway enrichment analysis

We included signalling and regulatory pathways from the Pathway Interaction Database<sup>28</sup> in pathway enrichment analysis. Enrichment was measured by hypergeometric tests. We only reported pathways that had an enrichment  $q$  value <0.05 [false discovery rate computed with the Benjamini–Hochberg method].

## 2.9. Data availability statement

The data underlying this article were derived from sources in the public domain. IBD, CD and UC summary statistics are available

at [ftp://ftp.sanger.ac.uk/pub/project/humgen/summary\\_statistics/human/2016-11-07/](ftp://ftp.sanger.ac.uk/pub/project/humgen/summary_statistics/human/2016-11-07/); GTEx v8-derived gene expression prediction models are available in Zenodo, at <https://dx.doi.org/10.5281/zenodo.3519321>; BarcUVa-Seq-derived prediction models are available in the Colon Transcriptome Explorer version 2.0, at <https://barcuvaseq.org/cotrex/>; and CEDAR data were obtained from the Array Express repository, under accession numbers E-MTAB-6666 and E-MTAB-6667 for genotypes and expression data, respectively.

## 3. Results

### 3.1. Transcriptome-wide associations

We evaluated associations between genetically regulated gene expression and IBD, CD and UC status, separately for each tissue/blood cell type. In the TWAS for IBD we found significant association for a median of 62 genes per tissue/cell type [ranging from ten to 124]. As expected, the number of significant associations increased with the sample size of the imputation panel. Also, we found fewer associated genes in the tissues/blood cell types of the CEDAR dataset, which was based on expression arrays to profile gene expression, than in the BarcUVa-Seq and GTEx datasets, which were based on RNA-seq. We found CD4<sup>+</sup>, CD14<sup>+</sup> and CD19<sup>+</sup> cells, rectum tissue, and BarcUVa-Seq transverse colon tissue among the tissues/cell types with the highest percentage of genes significantly associated with IBD. A summary of TWAS results for the three IBD phenotypes is provided in [Supplementary Table 2](#). Complete TWAS results in all tissues and cell types for IBD, CD and UC phenotypes are provided in [Supplementary Data 1](#).

### 3.2. BarcUVa-Seq colon TWAS

TWAS results generated with BarcUVa-Seq-derived panels [ascending, transverse, descending and any colon] are summarized in [Table 1](#) and shown in [Figure 1](#). We found 124 unique candidate susceptibility genes, including 39 that were novel (i.e. not reported in other large association studies see Methods). Among the 81 and 57 genes associated with CD and UC, respectively, we found 26 shared genes, and 55 and 31 genes specific for each disease subtype, respectively. CD-specific genes included Liver Enriched Antimicrobial Peptide 2 [*LEAP2*], and Ubiquitin D [*UBD*], both novel and specific for descending colon. UC-specific genes included Tripartite Motif Containing 31 [*TRIM31*], which was specific to ascending colon, and Abhydrolase Domain Containing 11 [*ABHD11*], which was specific to descending colon. We provide complete annotated results for BarcUVa-Seq candidate susceptibility genes in [Supplementary Data 2](#).

To identify cell types within the colon likely to mediate genetic susceptibility to IBD, we intersected lists of candidate susceptibility genes from BarcUVa-Seq TWAS with lists of expression marker genes of specific cell types derived from colon single cell RNA sequencing [scRNA-Seq] profiles.<sup>26</sup> We found 33 candidate susceptibility genes were markers for a total of 28 cell types across colon subsites [[Figure 2A](#)] and IBD phenotypes [[Figure 2B](#)]. Cell types were categorized into epithelial, fibroblast, endothelial, myeloid, T cell and B cell types [see Methods for annotation details]. The candidate susceptibility genes identified in ascending and transverse colon TWAS and in the TWAS for UC were more frequently markers of specific cell types than susceptibility genes identified in descending colon TWAS and in the TWAS for CD, respectively [see [Figure 2](#)]. Among these findings, we found ten novel candidate susceptibility cell marker genes, which are described in [Table 2](#). These included two fibroblast markers, three markers of myeloid cell types [e.g. inflammatory monocyte], four markers of epithelial cell types [such as M cell, goblet cell and enterocyte] and two

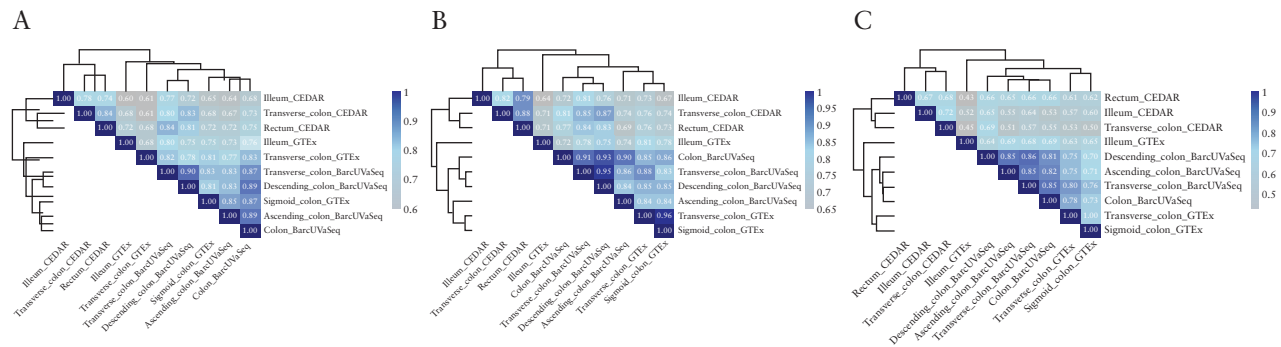




**Table 2.** Summary of genes whose genetically regulated expression in the colon is associated with IBD. Subset of ten genes reported as expression markers of cell types in the colon and that have not previously been reported as IBD susceptibility genes

| Phenotype | Colon subsite                     | Locus            | Gene symbol                 | Gene name   | GWAS SNP [p value]                             | Cell type   | Cell category            | Pathway   | Z score        | TWAS <i>p</i>        |
|-----------|-----------------------------------|------------------|-----------------------------|---|--|---|--------------------------|---|----------------|----------------------|
| IBD       | Transverse                        | 1q21.3           | <i>CTSS</i>                 | Cathepsin S   | rs17800987 [1.07E-16]                          | Inflammatory monocytes  | Myeloid                  | Trafficking and processing of endosomal Toll-like receptors | -4.73          | 2.27E-06             |
| IBD<br>UC | Colon<br>Colon                    | 1p36.12          | <i>WNT4</i>                 | Wnt Family Member 4                                 | rs12568930 [1.00E-17]<br>rs34920465 [9.01E-16] | WNT2B+, Fos-lo 2 cells  | Fibroblasts              | WNT ligand biogenesis and trafficking                       | -6.71<br>-5.99 | 2.01E-11<br>2.08E-09 |
| CD        | Colon                             | 5p13.1           | <i>C7</i>                   | Complement C7                                       | rs6451494 [8.258E-56]                          | RSPO3+ cells, WNT2B+, Fos-lo 2  | Fibroblasts              | Terminal pathway of complement                              | 4.75           | 2.04E-06             |
| CD<br>UC  | Descending<br>Ascending           | 6p22.1<br>6p22.1 | <i>UBD</i><br><i>TRIM31</i> | Ubiquitin D Tripartite Motif Containing 31          | rs11859512 [4.76E-13]<br>rs2270191 [1.139E-22] | Microfold cells<br>Goblet cells, enterocytes, immature enterocytes  | Epithelial<br>Epithelial | —<br>Interferon gamma signalling                            | 4.33<br>-4.70  | 1.47E-05<br>2.58E-06 |
| IBD<br>UC | Colon<br>Colon                    | 6p21.32          | <i>HLA-DOB</i>              | Major Histocompatibility Complex, Class II, DO Beta | rs6927022 [5.00E-133]<br>rs9271176 [4.20E-91]  | Cycling B cells, germinal centre [GC] cells, follicular cells; dendritic cells  | B cells; myeloid         | MHC class II antigen presentation                           | -4.96<br>-5.52 | 7.19E-07<br>3.35E-08 |
| IBD       | Transverse<br>Descending<br>Colon | 6p21.32          | <i>PSMB9</i>                | Proteasome 20S Subunit Beta 9                       | rs6927022 [5.00E-133]                          | CD8+ lamina propria [LP] cells  | T cells                  | Proteasome  | -4.50<br>-5.64 | 6.90E-06<br>1.74E-08 |
| CD        | Descending<br>Colon               | 7q11.23          | <i>CLDN4</i>                | Claudin 4   | rs185605448 [4.84E-04]                         | Transit-amplifying [TA] cells, immature goblet, immature enterocytes, stem cells, secretory TA, enterocytes, goblet, Best4+ enterocytes, enterocyte progenitors | Epithelial               | Tight junction  | -4.82<br>-4.46 | 1.71E-07<br>8.27E-06 |
| UC        | Colon                             | 7q11.23          | <i>CLDN4</i>                | Claudin 4   | rs11981405 [1.77E-07]                          | Transit-amplifying [TA] cells, immature goblet, immature enterocytes, stem cells, secretory TA, enterocytes, goblet, Best4+ enterocytes, enterocyte progenitors | Epithelial               | Tight junction  | 4.96           | 6.98E-07             |
| IBD       | Colon                             | 9q34.3           | <i>C8G</i>                  | Complement C8 Gamma Chain                           | rs10781499 [4.00E-56]                          | Enterocytes   | Epithelial               | Terminal pathway of complement                              | -5.11          | 3.27E-07             |
| IBD       | Colon                             | 14q13.2          | <i>NFKB1A</i>               | NFκB Inhibitor Alpha                                | rs2384352 [3.12E-13]                           | Tregs, innate lymphoid cells [ILCs]; CD69+, Mast  | T cells; myeloid         | TNF signalling  | 4.85           | 1.22E-06             |

GWAS SNP refers to the SNP identified by GWAS<sup>1</sup> with lowest *p* value among those located up to 1 Mb from the TSS of the associated gene. The pathway with the lowest enrichment *q* value [computed by a hypergeometric test] is indicated. If the enrichment *q* value is  $\geq 0.05$  the pathway is not provided. REACTOME and KEGG pathway sources were used. Gene symbols in bold refer to genes that were reported as differentially expressed between patient-derived and healthy colon biopsies.<sup>30</sup> Positive Z scores indicate that higher gene expression is associated with higher IBD risk.



**Figure 3.** Replication of TWAS results in lower intestinal tissues. Correlation of the predicted effect of gene expression across tissues identified in TWAS for [A] IBD, [B] CD and [C] UC. Hierarchical clustering of tissues is shown. Correlation values are indicated by the colour scale.

**Table 3.** Summary of significant candidate susceptibility genes identified in the joint analyses of TWAS results across all tissues/cell types

| Disease subtype        | Genes | Genes at GWAS loci | Novel genes | Fine-mapped genes |
|------------------------|-------|--------------------|-------------|-------------------|
| IBD                    | 466   | 388                | 136         | 50                |
| CD                     | 395   | 32                 | 116         | 44                |
| UC                     | 290   | 27                 | 88          | 31                |
| <i>Unique elements</i> | 596   | 440                | 186         | 47                |

Genes: significant genes passing Bonferroni correction and with lowest individual  $p \leq 1E-4$ ; GWAS loci: within 1 Mb of any top SNP found at corresponding GWAS; Novel: not reported in other large genome-wide association studies [see Methods]; Fine-mapped: genes included in fine-mapping credible sets and with >50% probability of being causal in their given signal.

Finally, to test for consistency of results across datasets from lower intestinal tissues, we correlated the predicted effect of the TWAS associations between tissues [see Figure 3]. As expected, we found the lowest correlations for the CEDAR dataset, which might be due to the technology used for assessing gene expression [arrays] in contrast to RNA-Seq, used by GTEx and BarcUva-Seq. We found high correlations between BarcUva-Seq-derived effects [all sites] and GTEx transverse colon-derived effects [ $r \geq 0.75$  in all three IBD phenotypes]. A lower correlation of results with GTEx sigmoid colon might be due to the higher component of muscularis tissue than of epithelial tissue present in the samples of this dataset.<sup>22</sup>

### 3.3. Joint analyses of TWAS results

To gain more power for discovery we performed a meta-analysis of all TWAS results obtained separately for IBD, CD and UC [summarized in Table 3]. In these joint analyses, we combined the TWAS results of all tissues/cell types [see Methods] and found 466, 395 and 290 significant genes for IBD, CD and UC risk, respectively, comprising 596 unique candidate susceptibility genes. These findings included 186 novel genes (i.e. not reported in other large association studies [see Methods]). Overall, we found candidate susceptibility genes nearby (i.e. with the gene Transcription Start Site [TSS] within 1 Mb of) 106 of the 240 top SNPs reported in IBD GWAS.<sup>3</sup> The Manhattan plot for IBD TWAS is shown in Figure 4. The most significant association is with Endosome Associated Trafficking Regulator 1 [ENTR1] [ $p = 8.27 \times 10^{-61}$ ]. We found 85 unique signalling and regulatory pathways significantly enriched in significantly associated

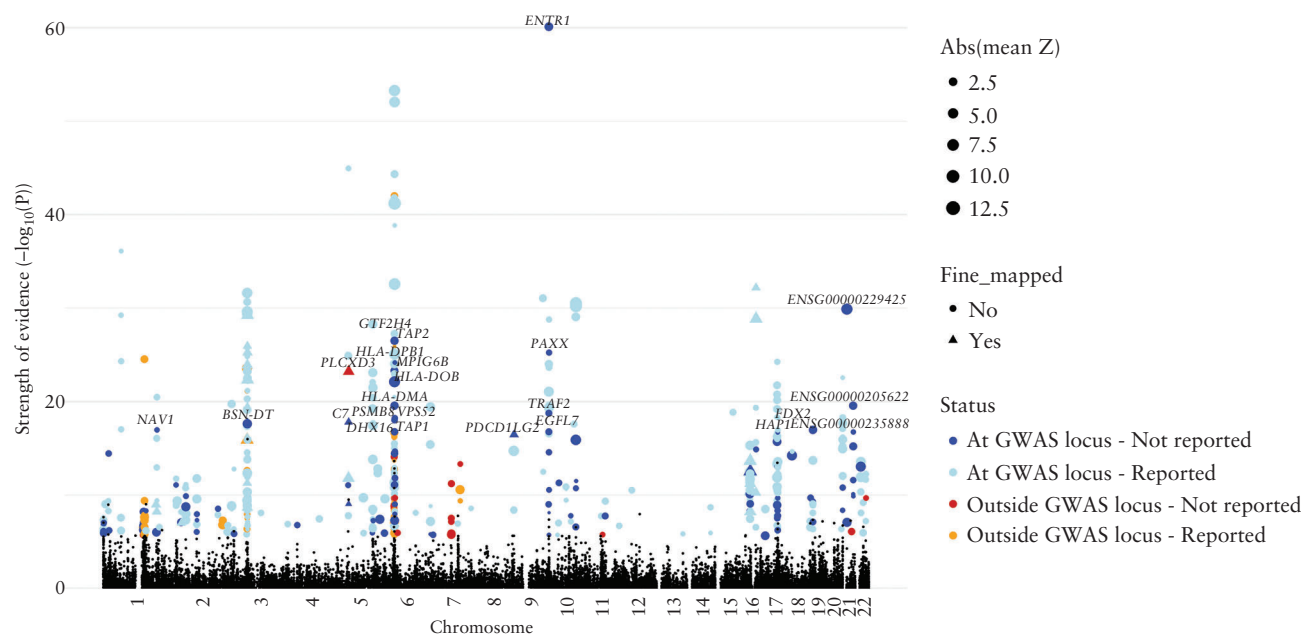
genes [Supplementary Table 4]. Among these pathways we found IL-12, IL-23, integrins and TNF-related pathways, which have high therapeutic relevance for IBD. Novel genes in these therapeutic pathways are summarized in Table 4.

We next performed fine-mapping of significantly associated genes to prioritize those with high probability of explaining the association signal in loci where multiple genes were identified. We found 50, 44 and 31 fine-mapped genes for IBD, CD and UC, respectively, comprising a total of 47 unique genes. Of these, we identified six novel genes [Supplementary Table 5], including five protein coding genes and one long non-coding RNA gene. These include genes that participate in the complement immune response.

### 3.4. Category-specific joint analyses

Next, to identify genes that participate in IBD susceptibility-related molecular mechanisms exclusively in specific tissue categories, we performed joint analyses combining different sets of TWAS results [see Methods] into epithelial [ $n = 18$ ], immune/blood [ $n = 14$ ], mesenchymal [ $n = 11$ ] and neural [ $n = 15$ ] categories based on histological and transcriptional characteristics [details in Methods and Supplementary Table 1]. Most significant genes found in these analyses had been previously identified in the joint analyses with all TWAS results [described in the previous section], but these tissue-category stratified joint analyses reported 93, 101 and 66 additional significant genes for IBD, CD and UC, respectively. Some of these genes were specific to IBD subtype and tissue category [see Table 5]. For example, we identified 26 genes specific to CD and the immune/blood category, which represented 11.9% of the total significant genes found in that analysis. In contrast, we found eight genes [5.2%] specific to UC and the immune/blood category [Table 5; Supplementary Figure 1]. A total of 78 category-specific genes (immune [ $N = 19$ ], epithelial [ $N = 25$ ], mesenchymal [ $N = 22$ ] and neural [ $N = 12$ ]) were not previously described by other studies [description given in Supplementary Table 6]. For example, we found that Aph-1 Homolog A, Gamma-Secretase Subunit [APH1A] underexpression in neural tissues was associated with IBD [ $p = 2.41E-06$ ]. This gene participates in presenilin action in Notch and Wnt signalling, and in syndecan-3-mediated signalling events, among others.<sup>28</sup>

In addition, we investigated gene pathway enrichment among category-specific IBD-associated genes. We identified 31 additional significantly enriched pathways not found in the pathway analysis of genes from the main TWAS meta-analysis described in the previous section [results in Supplementary Table 4]. Full results for all TWAS joint analyses [main and category-specific meta-analyses of TWAS results] are provided in Supplementary Data 3.



**Figure 4.** Manhattan plot of TWAS joint analysis for IBD. Each point represents a gene. Genes significantly associated are coloured. Novel genes with  $p < 1E-16$  are labelled.

**Table 4.** Summary of novel genes involved in signalling pathways of high therapeutic relevance for IBD

| Gene symbol    | Gene name  | Locus   | GWAS SNP              | TWAS P   | Mean Z | Pathway                                  | Drug name[s]           |
|----------------|--|---------|-----------------------|----------|--------|--|------------------------|
| <i>HLA-A</i>   | Major Histocompatibility Complex, Class I, A     | 6p22.1  | rs10826797 [3.99E-13] | 1.05E-06 | 1.72   | IL12-mediated signalling events          | Ustekinumab            |
| <i>MAP4K4</i>  | Mitogen-Activated Protein Kinase Kinase Kinase 4 | 2q11.2  | rs13001325 [2.51E-23] | 8.82E-07 | -1.09  | TNF receptor signalling pathway          | Infliximab, adalimumab |
| <i>TRAF2</i>   | TNF Receptor Associated Factor 2                 | 9q34.3  | rs10781499 [4.00E-56] | 1.78E-19 | -2.44  | TNF receptor signalling pathway          | Infliximab, adalimumab |
| <i>COL11A2</i> | Collagen Type XI Alpha 2 Chain                   | 6p21.32 | rs6927022 [5.00E-133] | 4.94E-06 | 1.05   | Beta1 integrin cell surface interactions | Vedolizumab            |

GWAS SNP refers to the SNP identified by GWAS<sup>3</sup> with lowest  $p$  value among those located up to 1 Mb from the TSS of the associated gene. Pathway refers to a signalling pathway in which the gene is significantly enriched [ $q$  value  $< 0.05$ ] and which is related to the drug indicated.

#### 4. Discussion

In this study, we identified candidate genes that may modulate the inherited risk of IBD and could eventually be exploited as novel therapeutic targets. We integrated transcriptomic and genetic information to predict gene expression in 59 957 genotyped subjects, including 25 042 with IBD, and discovered new associations between gene expression and IBD status. Additional insight into colon subsite-specific mechanisms was provided by site-specific expression prediction models trained on the recently published BarcUVA-Seq expression quantitative trait locus dataset. To gain new insights into the mechanisms underlying IBD, we performed a large, multi-dataset TWAS,<sup>21</sup> including predictive models for colon epithelium-enriched tissues and blood cell types of high relevance for IBD.

There are notable advantages of TWAS over other traditionally used approaches [such as GWAS and differential gene expression analysis] for nominating candidate genes that participate in disease pathogenesis. On the one hand, GWAS just identify risk SNPs and, except some obvious cases where an SNP lies in coding regions, this approach does not provide the candidate downstream functional

effects of the SNP on the phenotype/disease. On the other hand, differential gene expression analysis using observational rather than predicted gene expression measures does not provide causal inference. In this sense, the genetic variants that regulate gene expression are not affected by the disease, and therefore the direction of the effect, from gene expression to the disease, and not the opposite, can be made for the TWAS-identified genes.

TWAS based on expression models trained on BarcUVA-Seq ascending, transverse and descending colon allowed comparison of IBD-associated genetically regulated gene expression across different colon subsites. We identified susceptibility genes specific for colon subsites and IBD subtype. The strongest association signal [ $p = 1.33 \times 10^{-104}$ ] involved Phosphodiesterase 4B [*PDE4B*], whose expression was associated with IBD only in ascending colon. *PDE4B* is a candidate therapeutic target for paediatric-onset IBD,<sup>30</sup> and expression of *PDE4B* was associated with UC in patient-derived colon biopsies.<sup>29</sup> The TSS of *PDE4B* is located over 1 Mb from any top GWAS SNP, and the gene has not previously been associated with IBD susceptibility. Another novel gene involved in genetic

**Table 5.** Summary of TWAS joint analyses combining results of specific tissues/cell types

| Phenotype       | Analysis     | Tissues/cell types | Significant genes [not found in joint analysis of all TWAS] | Category-specific gene [% of significant genes] | Novel category-specific genes |
|-----------------|--------------|--------------------|---|---|-------------------------------|
| IBD             | All          | 61                 | 466   | —   | —                             |
|                 | Immune/blood | 14                 | 239 [32]  | 19 [8.0%]                                       | 6                             |
|                 | Epithelial   | 18                 | 271 [33]  | 25 [9.2%]                                       | 9                             |
|                 | Mesenchymal  | 11                 | 252 [33]  | 23 [9.1%]                                       | 13                            |
|                 | Neural       | 15                 | 230 [15]  | 9 [3.9%]  | 3                             |
| CD              | All          | 61                 | 395   | —   | —                             |
|                 | Immune/blood | 14                 | 218 [38]  | 26 [11.9%]                                      | 11                            |
|                 | Epithelial   | 18                 | 231 [31]  | 23 [10.0%]                                      | 11                            |
|                 | Mesenchymal  | 11                 | 220 [31]  | 16 [6.8%]                                       | 10                            |
|                 | Neural       | 15                 | 200 [25]  | 17 [8.5%]                                       | 8                             |
| UC              | All          | 61                 | 290   | -   | -                             |
|                 | Immune/blood | 14                 | 154 [17]  | 8 [5.2%]  | 2                             |
|                 | Epithelial   | 18                 | 175 [25]  | 17 [9.7%]                                       | 8                             |
|                 | Mesenchymal  | 11                 | 154 [19]  | 12 [7.8%]                                       | 5                             |
|                 | Neural       | 15                 | 151 [21]  | 16 [10.6%]                                      | 2                             |
| Unique elements | —            | —                  | —   | —   | 78                            |

susceptibility to IBD is *UBD*, whose expression in descending colon was significantly associated with CD status. *UBD* is an expression marker for the microfold [M] cell, a type of colon epithelial cell associated with colon inflammation in UC-derived colon biopsies.<sup>26</sup> *UBD* has also been reported to be upregulated in patient-derived colon biopsies<sup>29</sup> and may be a target for anti-TNF- $\alpha$  treatment.<sup>31</sup> As these examples demonstrate, the candidate susceptibility genes identified in this study could be promising therapeutic targets for IBD treatment. Expression levels of two other genes, *TRIM31* and Claudin 4 [*CLDN4*], were associated with IBD status for the first time. *TRIM31* is an expression marker of colon goblet and enterocyte cells and was significant only in the ascending colon TWAS for UC and in the joint analysis of TWAS results of epithelial tissues, suggesting tissue type specificity. *TRIM31* downregulation has been linked to bacterial invasion.<sup>32</sup> *CLDN4* is involved in the control of colon epithelial barrier function, including the maintenance of tight junction integrity. These examples highlight new directions for emerging treatment approaches.

BarcUVa-Seq TWAS revealed 39 novel IBD candidate susceptibility genes, including expression markers of 28 cell types found in the colon.<sup>26</sup> This finding allowed us to link IBD risk SNPs to colon-specific cell types that may affect genetic susceptibility. The risk SNP rs12568930 at 1p36.12 was associated with WNT2B+ Fos-lo 2 cells [a subtype of colon inflammatory fibroblasts] through the expression of Wnt Family Member 4 [*WNT4*]. The mechanisms of intestinal fibrosis in IBD are poorly understood, which impedes the development of anti-fibrotic therapies.<sup>33</sup>

Next, we meta-analysed TWAS results in joint analyses that combine single-tissue results to increase the statistical power to identify associations [see Methods], given the shared patterns of genetically regulated expression across human tissues. The advantages of this integrative approach have been described elsewhere.<sup>21</sup> Joint analysis of all TWAS results showed 596 genes whose genetically regulated expression might be involved in IBD genetic susceptibility, including 186 genes that were not previously reported in other large association studies [Table 3]. Our meta-analysis highlighted *ENTR1* as an important susceptibility gene. This gene encodes a protein involved in presentation of TNF receptors on the cell surface, and the modulation of TNF-induced apoptosis.<sup>34</sup> We also reported other novel genes encoding proteins that play important roles in TNF signalling. For

example, TNF Receptor Associated Factor 2 [*TRAF2*],<sup>35</sup> involved in TNF signalling, may be targeted by anti-TNF IBD therapeutics.

Finally, we performed joint analyses of single-tissue TWAS results by histological category to identify associations specific to particular tissue types, which may point to specific molecular mechanisms underlying IBD genetic risk and may give insight into potential targeted therapies. These category-specific analyses identified additional susceptibility genes, allowed us to link risk SNPs to specific tissue types, and provided insight into tissue-type specific mechanisms, as revealed by pathway enrichment analysis.

An important limitation of the TWAS approach is the possibility of spurious correlation between IBD causal SNPs and SNPs regulating gene expression of nearby genes, which could drive non-causal associations, as reported elsewhere.<sup>36</sup> This affects especially the human major histocompatibility complex [MHC] region, which features high LD between SNPs and includes several immune-related genes, such as human leukocyte antigen [HLA] genes. Indeed, the mean of significantly associated genes per locus in IBD joint analysis was three genes, whereas 6p21.33 [MHC-related] and 3p21.31 were associated with 52 and 56 genes, respectively. The high number of associations motivated fine-mapping of these loci. Our fine-mapping approach modelled correlation among significant signals and assigned a probability to explain the observed association signal for every gene in a given locus at a nominal confidence of 90%.<sup>27</sup> The number of significant signals per locus was reduced after considering only fine-mapped genes. In particular, the 6p21.33 locus retained two probable causal genes out of 52 significantly associated genes.

Among significantly associated genes with strong evidence for causality after the fine mapping of other loci, we found six genes not previously reported. These included the Programmed Cell Death 1 Ligand 2 [*PDCD1LG2*] gene at 9p24.1, which has been linked to immunosuppression by inhibition of T-cell proliferation<sup>37</sup> as well as the Complement C6 and C7 genes [*C6*, *C7*] at 5p13.1, which are also involved in immunoregulatory processes. In addition, we found a long non-coding RNA [*Lnc-ATXN2L-1*], a type of molecule that remains understudied and is considered a promising topic of research.<sup>38</sup>

In comparison with other published TWAS studies for IBD,<sup>8-10</sup> this study provided more robust statistical associations. This is due to the use of a large number of gene expression datasets, including

some of high relevance to IBD, not previously included in other TWAS [i.e. BarcUVa-Seq, CEDAR], meta-analyses including TWAS of many tissues/cell types, and fine-mapping of significant association signals. Many of the significant associations we observed have been identified by other large association studies, including GWAS<sup>3,24</sup> and TWAS for IBD,<sup>8</sup> and other studies based on patient biopsy samples<sup>46</sup> [see Table 3], but our analysis still discovered novel associations.

Our results may guide other investigators to prioritize potential genes of interest for further functional studies. Indeed, the candidate genes we proposed would require extensive validation in an experimental setting, through, for example, the use of engineered organoid models, or CRISPR screens, which was beyond the scope of this study. We supported the robustness of our results by strong statistical significance and by showing overlap with genes described by other high-impact studies. Also, associated genes were enriched in relevant pathways for IBD, mostly immune-related, which might be potential therapeutic targets.

## Funding

This work was supported by the National Institutes of Health [R01 CA204279, R01 CA143237 and R01 CA201407 to G.C.]; the Agency for Management of University and Research Grants [AGAUR] of the Catalan Government [2017SGR723 to V.M.]; the Instituto de Salud Carlos III, co-funded by Instituto de Salud Carlos III funds – a way to build Europe [PI14-00613, and PI17-00092 to V.M.]; the Spanish Association Against Cancer [AECC] Scientific Foundation [GCTRA18022MORE to V.M.]; and the Centro de investigación biomédica en red. Epidemiología y salud pública [CIBERESP] [CB07/02/2005 to V.M.]. R.C.T. received funding through the Marie Skłodowska-Curie actions - Horizon 2020 - European Union grant no. 796216; M.O.S. received a postdoctoral fellowship through the 'Fundación Científica de la Asociación Española Contra el Cáncer [AECC]'; C.H.D. received funding through the National Institutes of Health training grant T32 5T32CA163177-07 [PI: Craig Slingluff, MD]; V.D.O. received a predoctoral fellowship through the Ministerio de Educación, Cultura y Deporte - Gobierno de España FPU16/00599.

## Conflict of Interest

The authors disclose no relevant conflicts of interest.

## Author Contributions

V.D.O.: Data collection and curation, statistical analysis, analysis and interpretation of data, draft writing. F.M.N.: Statistical analysis, analysis and interpretation of data, draft review and editing. G.I.S.: Analysis and interpretation of data, draft review and editing. J.G., F.R.M., M.O.S., C.H.D., M.D.: Analysis and interpretation of data, draft review and editing. A.D.V.: Statistical analysis, analysis and interpretation of data, draft review and editing. R.C.T.: Study conception and design, supervision, data collection and curation, analysis and interpretation of data, draft review and editing. G.C.: Funding acquisition, analysis and interpretation of data, draft review and editing. V.M.: Funding acquisition, study conception and design, analysis and interpretation of data, draft review and editing, supervision.

## Acknowledgments

We thank the CERCA Program, Generalitat de Catalunya, for institutional support.

## Supplementary Data

Supplementary data are available at ECCO-JCC online.

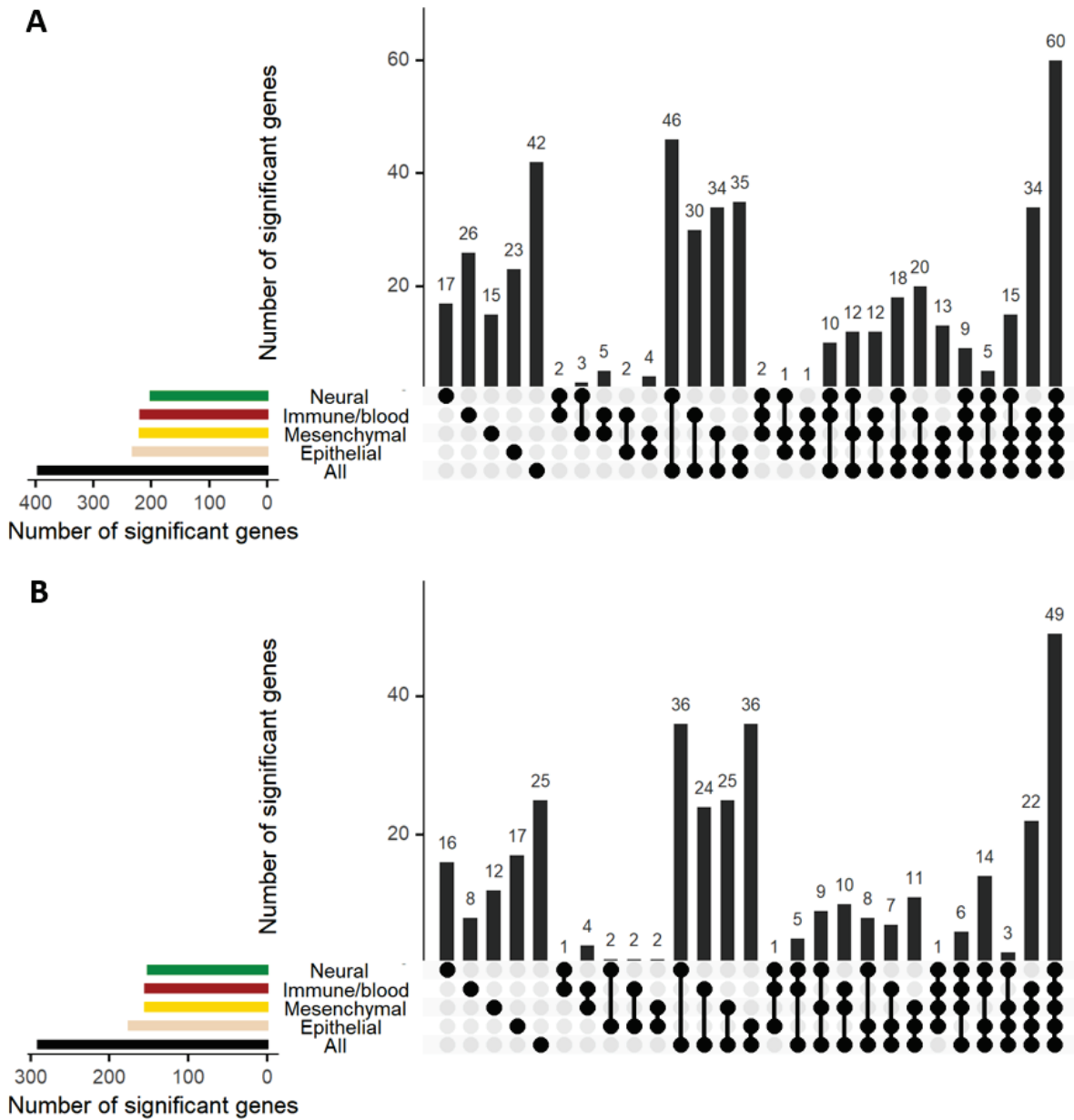
## References

- Furey TS, Sethupathy P, Sheikh SZ. Redefining the IBDs using genome-scale molecular phenotyping. *Nat Rev Gastroenterol Hepatol* 2019;16:296–311.
- Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* 2020;578:527–39.
- de Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* 2017;49:256–61.
- Verstockt B, Smith KG, Lee JC. Genome-wide association studies in Crohn's disease: past, present and future. *Clin Transl Immunology* 2018;7:e1001.
- Díez-Obrero V, Dampier CH, Moratalla-Navarro F, et al. Genetic effects on transcriptome profiles in colon epithelium provide functional insights for genetic risk loci. *Cell Mol Gastroenterol Hepatol* 2021;12:181–97.
- Momozawa Y, Dmitrieva J, Théâtre E, et al.; International IBD Genetics Consortium. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat Commun* 2018;9:2427.
- Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016;48:245–52.
- Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasianic B. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am J Hum Genet* 2017;100:473–87.
- Barbeira AN, Bonazzola R, Gamazon ER, et al.; GTEx GWAS Working Group; GTEx Consortium. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol* 2021;22:49.
- Dai Y, Pei G, Zhao Z, Jia P. A convergent study of genetic variants associated with Crohn's disease: evidence from GWAS, gene expression, methylation, eQTL and TWAS. *Front Genet* 2019. Doi: 10.3389/fgene.2019.00318.
- GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;369:1318–30.
- Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* 2014;30:1006–7.
- Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 2019;47:D766–73.
- Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48:1284–7.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.
- Smith ML, Baggerly KA, Bengtsson H, Ritchie ME, Hansen KD. illuminaio: an open source IDAT parsing tool for Illumina microarrays. *F1000Res* 2013;2:264.
- Barbeira AN, Melia OJ, Liang Y, Bonazzola R, Wang G, Wheeler HE, et al. Fine-mapping and QTL tissue-sharing information improves the reliability of causal gene identification. *Genet Epidemiol* 2020. Doi: 10.1002/gepi.22346.
- Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* 2012;7:500–7.
- CoTrEx 2.0. *The Colon Transcriptome Explorer Version 2.0*. <https://barcuvaseq.org/cotrex/>. Accessed May 14, 2021.
- Barbeira AN, Dickinson SP, Bonazzola R, et al.; GTEx Consortium. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* 2018;9:1825.
- Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet* 2019;15:e1007889.
- Breschi A, Muñoz-Agüirre M, Wucher V, et al. A limited set of transcriptional programs define major cell types. *Genome Res* 2020;30:1047–59.



23. Bruford EA, Braschi B, Denny P, Jones TEM, Seal RL, Tweedie S. Guidelines for human gene nomenclature. *Nat Genet* 2020;**52**:754–8.
24. Buniello A, MacArthur JAL, Cerezo M, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;**47**:D1005–12.
25. Huang H, Fang M, Jostins L, *et al.*; International Inflammatory Bowel Disease Genetics Consortium. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* 2017;**547**:173–8.
26. Smillie CS, Biton M, Ordovas-Montanes J, *et al.* Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 2019;**178**:714–30.e22.
27. Mancuso N, Freund MK, Johnson R, *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet* 2019;**51**:675–82.
28. Schaefer CF, Anthony K, Krupa S, *et al.* PID: the pathway interaction database. *Nucleic Acids Res* 2009;**37**:D674–9.
29. Taman H, Fenton CG, Hensel IV, Anderssen E, Florholmen J, Paulssen RH. Transcriptomic landscape of treatment—naïve ulcerative colitis. *J Crohns Colitis* 2018;**12**:327–36.
30. Huang B, Chen Z, Geng L, *et al.* Mucosal profiling of pediatric-onset colitis and IBD reveals common pathogenics and therapeutic pathways. *Cell* 2019;**179**:1160–76.e24.
31. Kawamoto A, Nagata S, Anzai S, *et al.* Ubiquitin D is upregulated by synergy of notch signalling and TNF- $\alpha$  in the inflamed intestinal epithelia of IBD patients. *J Crohns Colitis* 2019;**13**:495–509.
32. Ra EA, Lee TA, Won Kim S, *et al.* TRIM31 promotes Atg5/Atg7-independent autophagy in intestinal cells. *Nat Commun* 2016;**7**:11726.
33. Mao R, Rimola J, Chen MH, Rieder F. Intestinal fibrosis: the Achilles heel of inflammatory bowel diseases? *J Dig Dis* 2020;**21**:306–7.
34. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
35. Jin J, Xiao Y, Hu H, *et al.* Proinflammatory TLR signalling is regulated by a TRAF2-dependent proteolysis mechanism in macrophages. *Nat Commun* 2015;**6**:1–12.
36. Wainberg M, Sinnott-Armstrong N, Mancuso N, *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* 2019;**51**:592–9.
37. Solinas C, Aiello M, Rozali E, Lambertini M, Willard-Gallo K, Migliori E. Programmed cell death-ligand 2: a neglected but important target in the immune response to cancer? *Transl Oncol* 2020;**13**:100811.
38. Lin L, Zhou G, Chen P, *et al.* Which long noncoding RNAs and circular RNAs contribute to inflammatory bowel disease? *Cell Death Dis* 2020;**11**:456.

Supplementary Materials



**Supplementary Figure 1.** Distribution of the number of significant genes identified in different TWAS joint analyses and their intersections, for (A) CD and (B) UC.

**Supplementary Tables** are provided online, available at

<https://academic.oup.com/ecco-jcc/advance-article-abstract/doi/10.1093/ecco-jcc/jjab131/6324884> (large tables).

## **5. DISCUSSION**

This section includes an individual discussion for each of the objectives of the Thesis, achieved in the studies provided in section 4. Finally, a global discussion integrating all objectives is included.

### **5.1. Discussion of Objective 1**

The first objective of this Thesis was to provide reference profiles for transcriptome-wide gene expression and alternative splicing of colon mucosal biopsies from healthy adults, as well as their differences across colon location and corresponding e/sQTLs; also, to identify complex traits and diseases whose SNP-based heritability is enriched in the identified e/sQTLs, and propose candidate susceptibility genes for these phenotypes. To achieve this objective, a comprehensive study was carried out and published in the Cellular and Molecular Gastroenterology and Hepatology (CMGH) journal as an original article entitled “Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci”.

In this study, we generated a new dataset from 445 healthy individuals consisting of gene expression bulk RNA-Seq data and germline genotypes. We compared gene expression and AS profiles across ascending, transverse and descending colon subsites. We provided e/sQTLs, performed replication and meta-analysis with GTEx data, and assessed their enrichment in genome-wide regulatory regions and in the SNP-based heritability of common complex traits and diseases. Finally, we provided candidate susceptibility genes for 20 complex traits/diseases by colocalization analysis, whose expression in the colon contributes to their risk.

It is important to highlight that generating a good-quality dataset for molecular epidemiological research is challenging. Recruiting hundreds of healthy individuals to donate blood and colon tissue samples requires generous volunteers willing to altruistically contribute to research. Also, to design and implement a sample collection protocol is not straightforward and implies the coordinated work of many



professionals, such as gastroenterologists, biobank managers, laboratory technicians, staff from sequencing facilities and data analysts. The whole circuit must be properly orchestrated, and each step is important to achieve good quality samples that are homogeneous across individuals. Despite the challenges of this process, we successfully recruited nearly five hundred people and obtained good quality data, for both genotype and RNA-Seq data from 445 individuals. We named this project “University of Barcelona and University of Virginia RNA sequencing project” (BarcUVa-Seq), a name that also reflects the complexity that implies the coordination of research teams from different international institutions.

The characterization of the transcriptome from normal colon tissue made by the GTEx project presented some limitations that made it incomplete and partially inadequate. Samples were collected heterogeneously for transverse and sigmoid colon from postmortem donors. These included not only the mucosa but also deeper layers of the colonic wall, *e.g.* the sigmoid tissue was enriched in muscular tissue (8). In contrast, BarcUVa-Seq samples were representative from the colonic mucosa of healthy individuals and were collected homogeneously from ascending, transverse and descending colon locations. This characteristic, in contrast to GTEx data, allowed us to report gene expression and AS differences across ascending, transverse and descending colon locations.

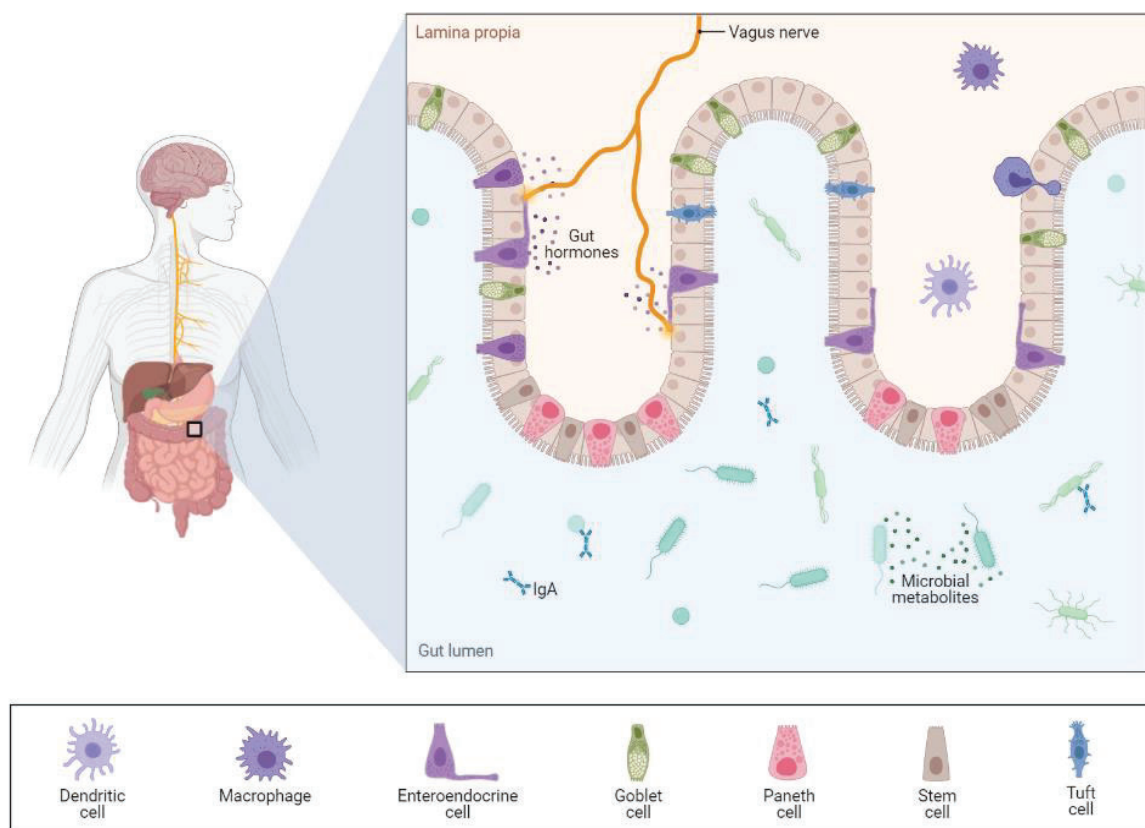
Regarding the novel differential expression results, we found more than four thousand genes whose expression linearly increases/decreases across the colon track. This implies a great advance in the knowledge of this topic and a better-defined picture of the colon transcriptome, in comparison to what was known from other studies, which reported only around one hundred genes differentially expressed between proximal and distal colon (33). Our results go in line with the fact that there are different molecular environments across the colon. For example, along the colon there is a progressive increase in pH, and different cell-type composition, and microbiome and related metabolites abundances (60). In addition,

we found that transverse colon gene expression profiles were more similar to those from descending than to those from ascending colon. This does not go in line with the embryological origin and blood supply of the colon, shared between ascending and transverse colon (61). This finding points to the need for further molecular analysis involving gene expression across colon subsites that give insights to the molecular processes driving these differences.

Then, we generated reference catalogs of normal colon e/sQTLs. The relationship between these two types of QTLs have been assessed in other studies (62,63). Although the molecular mechanisms underlying the genetic regulation of both gene expression and splicing are different, different types of QTLs can colocalize. In our study, we identified that most (~90%) e/sQTLs are independent of each other, suggesting different regulatory mechanisms implied, as reported by similar studies. We hypothesized that mechanisms exclusive to sQTL imply SNPs overlapping splice sites, and this was reflected in our results by a higher representation of sSNPs in these sites, as well as by sSNPs being significantly enriched in the binding sites of splice factors. Also, similarly to other studies (8,62), we found e/sQTLs in non-coding regions, enriched in regulatory sites such as enhancers. Our moderated replication of e/sQTLs with those e/sQTLs from GTEx sigmoid colon tissue reflected the different cell type composition of the samples between datasets. Indeed, we followed the GTEx analytical procedures (8) to analyze RNA-Seq data and compute e/sQTLs to ensure that lower replication estimates are due to biological rather than analytical factors.

Moreover, we identified complex traits and diseases whose SNP-based heritability is enriched by the identified e/sQTLs. As expected, we found strong evidence for colon-related diseases such as CRC and IBD, but also for complex traits/diseases not directly affecting the colon, such as behavioral traits and psychiatric diseases. This finding corroborated our hypothesis of finding diseases affected by molecular processes taking place in the framework of the gut-brain axis (depicted in **Figure**

15). A key component of the gut-brain axis is the vagus nerve, which is proposed to be responsible for transferring signaling molecules between the colon and the brain (64). These molecules include hormones that are produced by enteroendocrine cells in the colonic mucosa, as well as molecules produced in response to signaling processes derived from the communication with the microbiota (see **Figure 15**). Overall, this finding suggests that genetic regulation of colon gene expression plays a role not only in the colon, but in the systemic physiology.



**Figure 15. The gut-brain axis and cell type composition of the colon mucosa.** The main components of the gut-brain axis, as well as different cell types that are part of the colon mucosa are shown. Reprinted from “Gut-Brain Axis”, by BioRender.com (2021). Retrieved from <https://app.biorender.com/biorender-templates>.

In addition, in the study we proposed candidate susceptibility genes for the traits/diseases that showed strong evidence of having its SNP-based heritability influenced by eQTLs. For this purpose, we used colocalization analysis and reported as candidate genes those that showed high probability of colocalization with GWAS-identified risk SNPs. In the case of CRC we leveraged the data provided in a large

GWAS study, including about 35,000 CRC cases (43), and identified a total of 32 candidate susceptibility genes, including *LAMC1*, *TRIM28* and *SMAD9*, which participate in integrin cell surface, p53 and BMP receptor signaling pathways, respectively (65).

We are aware of the limitations of our study. One aspect is that we provide estimates of AS events based on short-read RNA-Seq data, which are not adequate nor intended to provide good resolution profiles of AS features. In contrast, improved technologies such as long-read and high-coverage RNA-Seq (66) would be better suited for this analysis, but due to the costs of this technology, it is not feasible yet to be scaled to the level of an epidemiological study including hundreds of individuals. Also, a wide variety of computational methods have been developed to quantify AS (63), such as the two complementary approaches that we used in our study (67,68). In addition, recent studies provided an improved statistical approach that showed improved estimates of AS and sQTLs across tissues of the GTEx dataset (69).

On the other hand, our transcriptome profiling is based on bulk RNA-Seq, which consists of the sequencing of RNA from a mixture of cell types that compose the colon mucosal tissue (70). Gene expression differences measured in bulk tissue transcriptomes may reflect changes in cellular composition rather than changes in the expression of genes in individual cells. One approach that could be implemented to tackle this limitation is computational deconvolution of bulk RNA-Seq, which provides an enrichment score of specific cell types and tissues. This score can be either included as a covariate for adjustment in the eQTL mapping, or as an outcome, to generate cell-type associated eQTLs (71). Another solution of this limitation is provided by the single-cell (sc) RNA-Seq sequencing technology, which significantly improves the resolution and could be employed to derive cell-type colon specific e/sQTLs. This technology applied to eQTL mapping is a promising path

of research in the field, as it can provide cell-specific molecular mechanisms of disease susceptibility (72).

Finally, it is important to remark that the uniqueness of BarcUVa-Seq data supports the high relevance of the results we obtained from its analysis as well as its potential utility to be used in further analysis. In this sense, we expect that this data represents a rich and valuable resource for the scientific community interested in investigating the human colon gene expression. Potential examples of use are exemplified by already published studies (73,74) that compare gene expression and AS profiles between normal colon and diseased colon derived samples, such as inflamed or neoplastic colon tissue.

## 5.2. Discussion of Objective 2

The second objective of this Thesis was to develop a web resource to explore population-based normal colon transcriptome profiles, e/sQTLs, gene expression prediction models, as well as to annotate SNPs with eQTLs. To achieve this objective, the Colon Transcriptome Explorer (CoTrEx) was developed, updated to the 2.0 version, and hosted online at <https://barcuvasq.org/cotrex/> to be publicly accessible. This work is entitled “The Colon Transcriptome Explorer (CoTrEx) 2.0, a reference resource for exploring population-based normal colon gene expression” and it is prepared for submission.

In this study, we provided an online interactive application that provides RNA-Seq-based gene expression data from the BarcUVa-Seq and GTEx colon tissues. It provides four main functionalities summarized in the Expression, QTLs, Prediction models and Networks tabs. Briefly, the Expression tab provides custom visualization of gene and transcript expression levels, as well as their related summary expression statistics. The QTLs tab provides e/sQTLs catalogs, its visualization, and a QTL annotation tool to annotate SNPs of interest. The Prediction models tab provides elastic-net based genetic gene expression prediction models, including summary statistics and SNP weights for each gene. Finally, the Networks tab provides relationships between genes based on regulatory and gene co-expression networks.

The relevance and potential usefulness of CoTrEx 2.0 are notable. It might be useful for researchers investigating 1) the transcriptomic basis of the colon, 2) the genetic regulatory processes affecting gene expression and AS in the colon, 3) the functional relevance of SNPs identified in GWAS, and 4) the molecular processes underlying susceptibility to colon-related traits/diseases. In relation to this last point, CoTrEx 2.0 provides genetic gene expression prediction models, a key input data to perform TWAS and nominate candidate effector genes associated with susceptibility to complex traits/diseases.

CoTrEx 2.0 presents many advantages in comparison with other similar resources. It is user-friendly and provides a quick and centralized access to highly requested gene expression-related data. Also, it includes the most updated version and largest data publicly available for colon tissue, *i.e.* from BarcUVa-Seq and GTEx version 8. These aspects make it a reference resource and differentiates it from other resources providing similar data and utilities. For example, the GTEx Transcript Browser (11) does not provide a custom visualization of transcript abundances. Also, it lacks the option for transcript grouping and data filtering, as well as additional visualization parameters that can be set in CoTrEx 2.0. Similarly, other GTEx online resources such as the GTEx eQTL Dashboard (75) or the GTEx eQTL Calculator (76) are not as comprehensive as CoTrEx 2.0 in terms of visualization capabilities. For example, a notable disadvantage of the GTEx eQTL Calculator is that it requires that users provide the gene ID in addition to the SNP of interest. This is not convenient in cases where the user wants to explore all SNP-gene associations of all genes nearby a SNP of interest.

Of note, although the original version of the CoTrEx was presented in the CMGH paper describing the BarcUVa-Seq dataset, it was substantially improved afterwards, which motivated us to describe it as an independent publication. The main features implemented in CoTrEx 2.0 include the incorporation of GTEx colon gene expression data, the incorporation of additional e/sQTL sets and the SNP annotation tool into the QTLs tab, the development of the Networks tab, as well as extra customization options implemented throughout the application.

Finally, future developments of CoTrEx would include additional features. For example, the incorporation of more QTL sets, such as regulatory QTLs (rQTLs) and eQTLs interacting with specific exposures. Also, a multi-gene query option would provide the visualization of the expression of multiple genes in annotated heatmaps, reflecting expression patterns by a covariate of interest such as colon location. In addition, the implementation of formal statistical tests to analyze the



data could be added in further releases of the resource, facilitating researchers to perform differential gene expression analysis by covariates of interest.

### 5.3. Discussion of Objective 3

The third objective of this Thesis was to propose candidate genes whose genetically regulated gene expression is associated with IBD, including genes in specific colon subsites that are expression markers of colon cell types, and genes that are enriched in relevant molecular pathways for IBD, such as therapy-related ones. Also, to identify candidate susceptibility genes specific for the epithelial, immune/blood, mesenchymal and neural tissue categories. To achieve this objective, a comprehensive study was carried out and accepted for publication in the Journal of Crohn's & Colitis (JCC) as an original article entitled “Transcriptome-wide association study for inflammatory bowel disease reveals novel candidate susceptibility genes in specific colon subsites and tissue categories”.

In this study, we imputed gene expression across a large set of tissues and cell types in a cohort of about 60,000 subjects, including around 25,000 IBD cases (54), and performed a comprehensive TWAS to nominate candidate susceptibility genes for IBD, CD and UC, respectively. Also, we combined TWAS results of histologically similar tissues and suggested susceptibility genes that could act in a tissue/cell-type specific manner. In addition, in the case of colon tissue, we identified candidate genes that could potentially point to colon subsite and cell-type specific molecular mechanisms of IBD susceptibility. This was achieved by using information derived from single-cell RNA-Seq data (56) to indicate the cell types whose expression markers overlap with the candidate susceptibility genes identified in the TWAS. Moreover, to further increase the association evidence of the proposed candidate genes, we carried out statistical fine-mapping (23) and indicated those with strongest evidence.

We identified genes that participate in key molecular pathways for IBD pathogenesis (49), including genes that maintain the intestinal barrier integrity, genes involved in the innate and adaptive immune system, genes related to interactions with the microbiome, and genes acting in other key pathways such as

autophagy and fibrosis. This highlights the relevance of known pathways driving IBD and expands the number of genes involved in their dysregulation during IBD pathogenesis. Some of these genes and pathways have a higher relevance in particular tissues. To highlight this aspect, in the study we specified the candidate genes found at each tissue and cell type, as well as by tissue-type category. For example, we found stronger associations in the colon for genes participating in the maintenance of epithelial tight junctions. Also, we found genes specifically in neural tissues that could be related to neuroimmune mechanisms. In this sense, this study would support the role of the enteric nervous system in the complex interplay of molecular pathways driving IBD, as described elsewhere (77).

Moreover, we found genes that participated in signaling and regulatory molecular pathways targeted by commonly used IBD therapeutics, such as monoclonal antibodies that modulate the immune response. There was previous evidence that GWAS-identified risk SNPs were related to therapeutically relevant pathways, such as tumor necrosis factor (TNF) signaling, and interleukin and autophagy-related pathways (54). In our study we expanded the knowledge on candidate susceptibility genes involved in these processes, including genes such as *HLA-A*, *MAP4K4*, *TRAF2* and *COL11A2*, which might be potential targets or modulators of therapeutic agents.

To provide further evidence of association, we compared our results with those based on observational data. We overlapped the identified candidate susceptibility genes for UC in the colon with genes differentially expressed between colon biopsies of UC patients and controls (78). We found a slightly moderate replication (around 20%) and a high concordance of direction of the effect. This low rate of concordance between predicted-based vs observational-based differential expression results was expected. This could be explained as observational data is based on smallest sample sizes ( $n=15$  in this case) and therefore has lower statistical power; also, gene expression dysregulation processes caused by IBD status can

confound and bias the results, also known as reverse causation. In contrast, TWAS-based candidate susceptibility genes are not biased by the effect of disease, as they are predicted from blood DNA genotypes (which cannot be altered by disease), and therefore, the directionality of the effect can be established, from susceptibility genes to disease, but not vice versa. In this sense, it is also notable that the imputation panels in which TWAS is based should be from healthy normal tissue, instead of inflamed tissue from patients, which is often used and might bias the results.

In comparison with previous TWAS for IBD (57–59), we carried out a large, detailed, and complete study; including the most up to date genetic gene expression prediction models applied in a public large GWAS study. We included novel tissues/cells (55) not previously assessed in this framework and of high relevance for IBD, such as tissue across colon subsites and immune blood cells. Our approach provided not only novelty and comprehensiveness, but also more robust results because of an increased statistical power due to sample size and meta-analysis. Importantly, most of our results replicated with those provided in other large-scale genome/transcriptome-wide studies, supporting their robustness. To facilitate the comparison with other studies as well as the integration of our results in further studies, we highlighted the novel results and provided the complete association and statistical parameters for all genes and tissues as supplemental data.

We are aware that the comparison of TWAS results across tissues and cell types to nominate potentially more relevant genes is challenging. This is partially because associations between genetically regulated gene expression and IBD risk were strongly dependent on the sample size of the used datasets, showing a positive correlation between the number of identified genes and the sample size of the reference prediction panels. Therefore, conscious of this limitation, we compared TWAS results across tissues according to the correlation of the association effect sizes of significant genes, and we found higher homogeneity across tissues and cell

types within each disease subtype (CD and UC) than between them. Understanding the cell-specific role of disease-associated variants is crucial and an active area of research (59). Also, we found a higher number of significant genes for CD than for UC, which might be related to the higher estimated SNP-based heritability observed for CD than for UC.

Finally, regarding the implication of the results, these can be of high relevance for other researchers interested in 1) investigating the molecular processes driving IBD and their differences according to disease subtype (CD and UC) and tissue/cell type, 2) identifying potential candidate therapeutic targets, and 3) the development of predictive risk models based on predicted gene expression.

#### 5.4. Global discussion

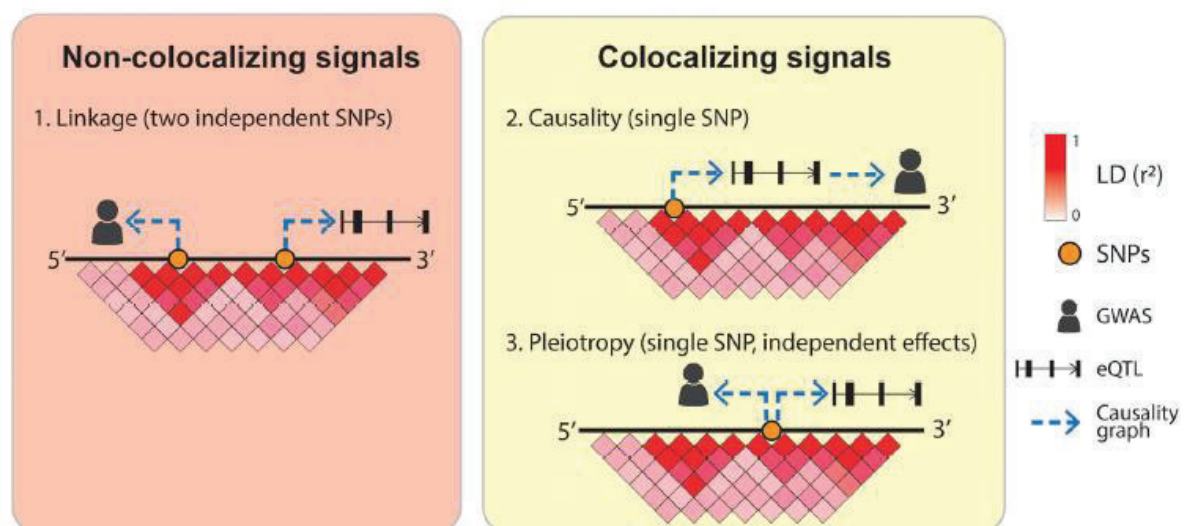
Exponential advances in the development of sequencing technologies during the last two decades have provided a feasible generation of large omics datasets. These advances, along with those in the fields of bioinformatics and biostatistics providing robust inferences and interpretations of these datasets, have promoted a paradigm shift in the study of life sciences and have provided advances in Medicine (79). More in detail, large scale genome-wide association analyses such as GWAS, and its functional interpretation and translation to targeted therapies, is a field that is gaining momentum (80).

In this up-to-date framework, we have carried out studies that give insights into the regulatory effects on colon gene expression and its role in providing susceptibility to colon-related diseases such as CRC and IBD. In addition, we used bioinformatic tools to develop a web-based interactive resource to help other investigators to easily benefit from the large colon gene expression-related data we and other researchers provide.

Altogether, we lay the foundations for important discoveries coming from the exploitation of the benefits provided by our investigations. For example, the genetic gene expression prediction models provided in CoTrEx 2.0 might be used to perform TWAS for a wide range of complex traits and diseases, which, similarly to what we evidenced in our TWAS for IBD, might point to strong candidate susceptibility genes in relevant signaling pathways and cell types. One promising application of our colon eQTL reference data and imputation panels is their use to give more insight into the biology underlying autoimmune diseases, such as rheumatoid or celiac disease. These diseases are influenced by altered permeability of the colon mucosa, thus nominating related genes can be useful for specific drug design.

We are aware that the architecture of transcriptional variation is complex, and that there are some limitations inherent to the statistical approaches used to link genetically predicted gene expression with disease. In this regard, to address LD-

related issues we performed colocalization rather than functional annotation; and to address the issue of correlation between genes we used fine mapping. Although these approaches mitigate LD linkage, they are subjected to biases due to other genetic phenomena such as epistasis, genotype environment interaction, and pleiotropy. For example, overlapping eQTL and GWAS signals can be linked by LD, or can colocalize through any of these two scenarios: 1) causality, *i.e.* a single-causal SNP affecting the trait by modulating the expression of a gene; or 2) pleiotropy, *i.e.* a single-causal SNP with independent effects on trait and gene expression (10) (see **Figure 16**). Identifying a causal rather than a pleiotropic effect is an expanding area of research, and there have been developing tools based on Mendelian randomization approaches that address this topic (81).



**Figure 16. Linkage, causality, and pleiotropy effects on colocalization.** Colocalization addresses linkage but can be driven either by true causality or by pleiotropy. Adapted from (10).

Large scale genome-wide association analyses have resulted in great advances in the field, including the understanding of disease biology and the development of targeted therapies. Despite hints of success, many doubts have been recently raised about the utility of GWAS (82,83) under the reasoning that *everything points to nothing*. As GWAS studies continuously provide additional variants associated with disease, in the end, every DNA region active in a tissue would be involved in a

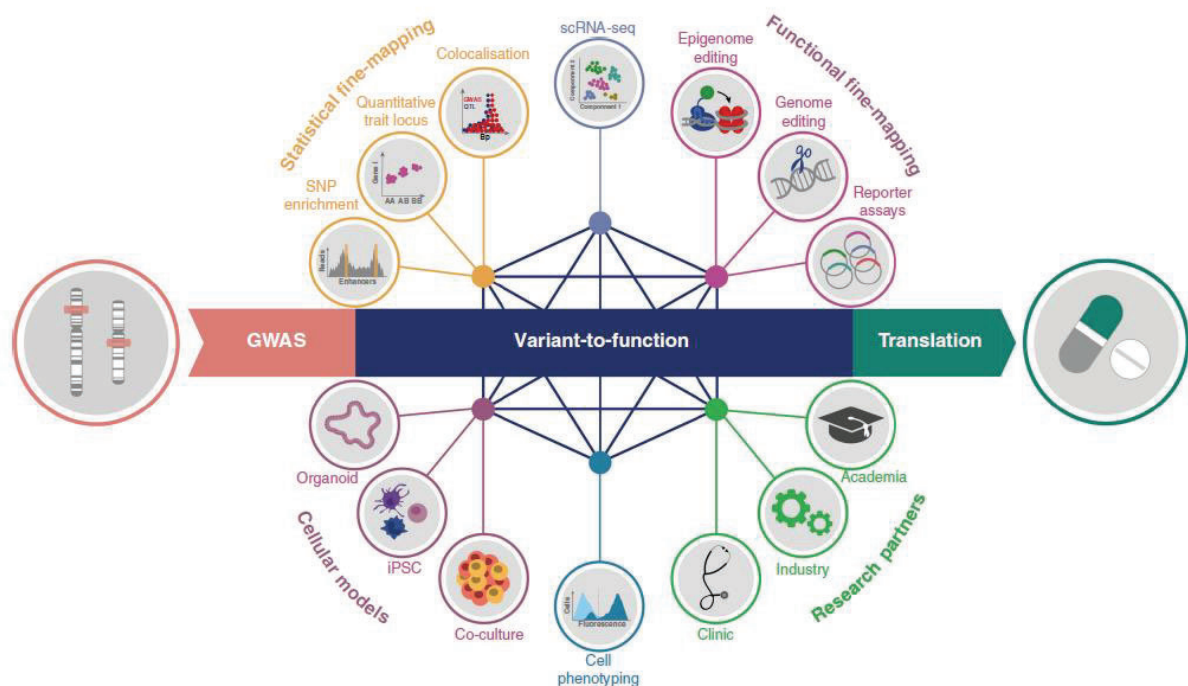


common disease, and all expressed genes in a tissue can be implicated. Therefore, scientists raising this concern claim that it will become indefensible that there's a simple biological interpretation for each gene associated with disease. They articulate that instead of increasing sample size and the efforts to carry out larger GWAS studies, research might be focused on understanding biochemical networks and the connections between the molecules participating in them (82). However, both strategies are complementary and will provide a broader understanding of disease etiology.

Then, it is important to remark that undesired research practices that bias and override findings are extended and often overlooked (84). A concerning topic that stands out is the lack of reproducibility. For example, a study pointed out that most (more than 70%) researchers have failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own (85). Addressing the aspects that contribute to these research practices is crucial, as improving the reliability of research will increase the credibility of the published scientific literature and accelerate discoveries (84). Specifically, in the field of bioinformatics, some aspects must be considered. For example, analyses can be run with different algorithms and statistical methods that are equally valid. The selection of the one that either has the friendliest user interface or gives the most interesting results might bias the research findings. Also, analyses may require the subjective selection of *a priori* parameters, whose different selection will likely generate different results. Moreover, some analyses can become deprecated because of advances in the databases/reference sets. Therefore, the selection of the correct genome build and the adequate database is key, as well as investigating their limitations, including their lack of corrections or updates to annotations (79).

Finally, regarding the future directions of this field, the coordinated interplay of researchers from different disciplines will result in the successful identification of disease-associated *loci*, as well as its translation to meaningful discoveries that

provide clinical solutions, such as drugs and risk prediction tools (80). A summary of the key players that will drive these advances is summarized in **Figure 17**. Functional fine-mapping approaches, in addition to those that we used in our studies, such as clustered regularly interspaced short palindromic repeats (CRISPR) screens and massively parallel reporter assays (MPRA) may serve as *in vitro* validation tools of the candidate susceptibility genes. Another essential component driving advances in the field are scRNA-Seq-derived approaches, which are still in their first stages of development, but represent a promising technology. Also, advances in approaches to provide adequate cellular models represent another hot topic that is rapidly expanding, including organ-on-a-chip, and engineered organoid cultures (see **Figure 17**).



**Figure 17. Approaches for translating disease associated risk *loci* into targeted therapeutics.** Making sense of GWAS-identified risk *loci* would require the orchestrated conjunction of key cross-disciplinary areas of development. Thanks to these combined efforts, significant advancements will take place for the translation of knowledge to novel clinical solutions. Reprinted from Lichou & Trynka (80).

Overall, our results might aid other researchers to generate novel hypotheses that guide future investigations on the molecular basis of complex traits and diseases.

Specially for those affected by gene expression changes in the colon. Also, the work included in this Thesis will pave the way for future developments on risk prediction approaches and targeted therapies in IBD and CRC.

## 6. CONCLUSIONS

1. We have generated reference profiles of gene expression and alternative splicing of normal colon tissue based on a sample size of 445 healthy individuals.
2. We have found 4,430 genes differentially expressed across ascending, transverse and descending colon subsites.
3. We have reported 11,739 eQTLs and 1,125 sQTLs in normal colon tissue. About 50% of the SNPs involved in these QTLs were intronic. Also, they were enriched in regulatory regions, suggesting additional functional relevance in the colon.
4. We have identified 20 complex traits and diseases whose SNP-based heritability estimation is significantly enriched in the eQTLs identified, and we have proposed candidate susceptibility genes for these phenotypes.
5. We have provided insight into the genes and molecular processes underlying disease susceptibility. These genes should be prioritized in functional studies and could be targets for new therapeutics.
6. The Colon Transcriptome Explorer 2.0, an interactive web-based resource, was built and hosted online at <https://barcuvasseq.org/cotrex/>. This application is of interest for visualizing gene and transcript expression levels in the colon, as well as exploring SNP-expression associations and annotating SNPs with colon eQTLs.
7. We identified 136, 116 and 88 novel candidate susceptibility genes for IBD, CD and UC, respectively. We described in detail the novel genes that were identified in the colon (N=39) as well as those identified in immune (N=19), epithelial (N=25), mesenchymal (N=22) and neural (N=12) tissue categories.

8. The candidate genes we proposed for IBD participate in regulatory and signaling pathways mostly related to the immune system, as well as in other key pathways, such as the maintenance of the colon mucosa integrity and pathways related to IBD therapeutics.

## 7. REFERENCES

1. Adams J. Transcriptome: connecting the genome to gene function. *Nature Education*; 2008. (1(1):195).
2. Skelly DA, Ronald J, Akey JM. Inherited Variation in Gene Expression [Internet]. Vol. 10, *Annual Review of Genomics and Human Genetics*. 2009. p. 313–32. Available from: <http://dx.doi.org/10.1146/annurev-genom-082908-150121>
3. Holste D, Ohler U. Strategies for identifying RNA splicing regulatory motifs and predicting alternative splicing events. *PLoS Comput Biol*. 2008 Jan;4(1):e21.
4. Dvinge H, Kim E, Abdel-Wahab O, Bradley RK. RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer*. 2016 Jul;16(7):413–30.
5. Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes [Internet]. Vol. 456, *Nature*. 2008. p. 470–6. Available from: <http://dx.doi.org/10.1038/nature07509>
6. Scotti MM, Swanson MS. RNA mis-splicing in disease [Internet]. Vol. 17, *Nature Reviews Genetics*. 2016. p. 19–32. Available from: <http://dx.doi.org/10.1038/nrg.2015.3>
7. Sammeth M, Foissac S, Guigó R. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol*. 2008 Aug 8;4(8):e1000147.
8. Consortium TG, The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues [Internet]. Vol. 369, *Science*. 2020. p. 1318–30. Available from: <http://dx.doi.org/10.1126/science.aaz1776>

9. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet.* 2009 Aug;10(8):565–77.
10. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet.* 2020 May 13;11:424.
11. The GTEx Consortium. GTEx Transcript Browser [Internet]. 2021 [cited 2021 Jul 25]. Available from: <https://gtexportal.org/home/transcriptPage>
12. Hemminki K, Försti A, Houlston R, Bermejo JL. Searching for the missing heritability of complex diseases. *Hum Mutat.* 2011 Feb;32(2):259–62.
13. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet.* 2013 Nov 7;93(5):779–97.
14. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015 Mar;47(3):291–5.
15. Broekema RV, Bakker OB, Jonkers IH. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.* 2020 Jan;10(1):190221.
16. Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet.* 2019 May;51(5):768–9.
17. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 2014 May;10(5):e1004383.



18. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet.* 2017 Mar;13(3):e1006646.
19. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet.* 2016 Mar;48(3):245–52.
20. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun.* 2018 May 8;9(1):1825.
21. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* 2019 Jan;15(1):e1007889.
22. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet.* 2019 Apr;51(4):592–9.
23. Mancuso N, Freund MK, Johnson R, Shi H, Kichaev G, Gusev A, et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet.* 2019 Apr;51(4):675–82.
24. National Cancer Institute. NCI Dictionary of Cancer Terms [Internet]. 2021 [cited 2021 Jul 25]. Available from: <https://www.cancer.gov/publications/dictionaries/cancer-terms>
25. Azzouz LL, Sharma S. Physiology, Large Intestine. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2020.
26. U. S. National Institutes of Health, National Cancer Institute. SEER Training

Modules, Anatomy of Colon and Rectum [Internet]. 2021 [cited 2021 Jul 25]. Available from: <https://training.seer.cancer.gov/colorectal/anatomy/>

27. Britannica, The Editors of Encyclopaedia. Large intestine [Internet]. Encyclopædia Britannica. 2020 [cited 2021 Jul 25]. Available from: <https://www.britannica.com/science/large-intestine#/media/1/330544/68639>
28. Chelakkot C, Ghim J, Ryu SH. Mechanisms regulating intestinal barrier integrity and its pathological implications. *Exp Mol Med*. 2018 Aug 16;50(8):1–9.
29. Allaire JM, Crowley SM, Law HT, Chang S-Y, Ko H-J, Vallance BA. The Intestinal Epithelium: Central Coordinator of Mucosal Immunity. *Trends Immunol*. 2018 Sep;39(9):677–96.
30. Johansson MEV, Sjövall H, Hansson GC. The gastrointestinal mucus system in health and disease. *Nat Rev Gastroenterol Hepatol*. 2013 Jun;10(6):352–61.
31. Funk MC, Zhou J, Boutros M. Ageing, metabolism and the intestine [Internet]. Vol. 21, EMBO reports. 2020. Available from: <http://dx.doi.org/10.15252/embr.202050047>
32. Moreels TG, Dewit O. Normal Endoscopic Appearance of the Colon and the Terminal Ileum [Internet]. *Colitis*. 2018. p. 31–6. Available from: [http://dx.doi.org/10.1007/978-3-319-89503-1\\_4](http://dx.doi.org/10.1007/978-3-319-89503-1_4)
33. LaPointe LC, Dunne R, Brown GS, Worthley DL, Molloy PL, Wattchow D, et al. Map of differential transcript expression in the normal human large intestine. *Physiol Genomics*. 2008 Mar 14;33(1):50–64.
34. Sanz-Pamplona R, Cordero D, Berenguer A, Lejbkowitz F, Rennert H, Salazar R, et al. Gene expression differences between colon and rectum tumors. *Clin*

- Cancer Res. 2011 Dec 1;17(23):7303–12.
35. Moreno V, Alonso MH, Closa A, Vallés X, Diez-Villanueva A, Valle L, et al. Colon-specific eQTL analysis to inform on functional SNPs. *Br J Cancer*. 2018 Oct;119(8):971–7.
  36. The Colonomics Team. The Colonomics Expression Browser [Internet]. 2016 [cited 2021 Jul 25]. Available from: <https://www.colonomics.org/expression-browser/>
  37. The Colonomics Team. The Colonomics eQTL Browser [Internet]. 2016 [cited 2021 Jul 25]. Available from: <https://www.colonomics.org/eqtlbrowser/>
  38. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes [Internet]. Vol. 8, *Nature Communications*. 2017. Available from: <http://dx.doi.org/10.1038/s41467-017-01027-z>
  39. Breschi A, Muñoz-Aguirre M, Wucher V, Davis CA, Garrido-Martín D, Djebali S, et al. A limited set of transcriptional programs define major cell types. *Genome Res*. 2020 Jul;30(7):1047–59.
  40. U.S. National Library of Medicine. Colonic Diseases [Internet]. 2021 [cited 2021 Jul 25]. Available from: <https://medlineplus.gov/colonicdiseases.html>
  41. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018 Nov;68(6):394–424.
  42. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat Rev Gastroenterol Hepatol*. 2019 Dec;16(12):713–32.

43. Huyghe JR, Bien SA, Harrison TA, Kang HM, Chen S, Schmit SL, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet.* 2019 Jan;51(1):76–87.
44. Law PJ, Timofeeva M, Fernandez-Rozadilla C, Broderick P, Studd J, Fernandez-Tajes J, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun.* 2019 May 14;10(1):2154.
45. Thomas M, Sakoda LC, Hoffmeister M, Rosenthal EA, Lee JK, van Duijnhoven FJB, et al. Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. *Am J Hum Genet.* 2020 Sep 3;107(3):432–44.
46. Huyghe JR, Harrison TA, Bien SA, Hampel H, Figueiredo JC, Schmit SL, et al. Genetic architectures of proximal and distal colorectal cancer are partly distinct. *Gut.* 2021 Jul;70(7):1325–34.
47. Guo X, Lin W, Wen W, Huyghe J, Bien S, Cai Q, et al. Identifying Novel Susceptibility Genes for Colorectal Cancer Risk From a Transcriptome-Wide Association Study of 125,478 Subjects. *Gastroenterology.* 2021 Mar;160(4):1164–78.e6.
48. Windsor JW, Kaplan GG. Evolving Epidemiology of IBD. *Curr Gastroenterol Rep.* 2019 Jul 23;21(8):40.
49. Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature.* 2020 Feb;578(7796):527–39.
50. Furey TS, Sethupathy P, Sheikh SZ. Redefining the IBDs using genome-scale molecular phenotyping. *Nat Rev Gastroenterol Hepatol.* 2019 May;16(5):296–311.
51. Park JH, Peyrin-Biroulet L, Eisenhut M, Shin JI. IBD immunopathogenesis: A comprehensive review of inflammatory molecules. *Autoimmun Rev.* 2017

Apr;16(4):416–26.

52. Eisenstein M. Biology: A slow-motion epidemic. *Nature*. 2016 Dec 21;540(7634):S98–9.
53. Pierre N, Salée C, Vieujean S, Bequet E, Merli A-M, Siegmund B, et al. Review article: distinctions between ileal and colonic Crohn’s disease: from physiology to pathology. *Aliment Pharmacol Ther* [Internet]. 2021 Jul 23; Available from: <http://dx.doi.org/10.1111/apt.16536>
54. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet*. 2017 Feb;49(2):256–61.
55. Momozawa Y, Dmitrieva J, Théâtre E, Deffontaine V, Rahmouni S, Charloreaux B, et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat Commun*. 2018 Jun 21;9(1):2427.
56. Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell*. 2019 Jul 25;178(3):714–30.e22.
57. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am J Hum Genet*. 2017 Mar 2;100(3):473–87.
58. Barbeira AN, Bonazzola R, Gamazon ER, Liang Y, Park Y, Kim-Hellmuth S, et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol*. 2021 Jan 26;22(1):49.
59. Dai Y, Pei G, Zhao Z, Jia P. A Convergent Study of Genetic Variants Associated

With Crohn's Disease: Evidence From GWAS, Gene Expression, Methylation, eQTL and TWAS. *Front Genet.* 2019 Apr 9;10:318.

60. Zheng D, Liwinski T, Elinav E. Interaction between microbiota and immunity in health and disease. *Cell Res.* 2020 Jun;30(6):492–506.
61. Rubian FA, Al Rubian F, Keijzer R. Normal Embryology, Anatomy, and Physiology of the Gastrointestinal Tract [Internet]. *Pearls and Tricks in Pediatric Surgery.* 2021. p. 147–53. Available from: [http://dx.doi.org/10.1007/978-3-030-51067-1\\_21](http://dx.doi.org/10.1007/978-3-030-51067-1_21)
62. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. *Science.* 2016 Apr 29;352(6285):600–4.
63. Park E, Pan Z, Zhang Z, Lin L, Xing Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet.* 2018 Jan 4;102(1):11–26.
64. Underwood E. Newly detailed nerve links between brain and other organs shape thoughts, memories, and feelings [Internet]. *Science News.* 2021 [cited 2021 Jul 25]. Available from: [https://www.sciencemag.org/news/2021/06/newly-detailed-nerve-links-between-brain-and-other-organs-shape-thoughts-memories-and?utm\\_campaign=SciMag](https://www.sciencemag.org/news/2021/06/newly-detailed-nerve-links-between-brain-and-other-organs-shape-thoughts-memories-and?utm_campaign=SciMag)
65. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. *Nucleic Acids Res.* 2009 Jan;37(Database issue):D674–9.
66. Oikonomopoulos S, Bayega A, Fahiminiya S, Djambazian H, Berube P, Ragoussis J. Methodologies for Transcript Profiling Using Long-Read

Technologies. *Front Genet.* 2020 Jul 7;11:606.

67. Trincado JL, Entizne JC, Hysenaj G, Singh B, Skalic M, Elliott DJ, et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* 2018 Mar 23;19(1):40.
68. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet.* 2018 Jan;50(1):151–8.
69. Garrido-Martín D, Borsari B, Calvo M, Reverter F, Guigó R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat Commun.* 2021 Feb 1;12(1):727.
70. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016 Jan 26;17:13.
71. Donovan MKR, D'Antonio-Chronowska A, D'Antonio M, Frazer KA. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat Commun.* 2020 Feb 19;11(1):955.
72. van der Wijst M, de Vries DH, Groot HE, Trynka G, Hon CC, Bonder MJ, et al. The single-cell eQTLGen consortium. *Elife* [Internet]. 2020 Mar 9;9. Available from: <http://dx.doi.org/10.7554/eLife.52155>
73. Dampier CH, Devall M, Jennelle LT, Díez-Obrero V, Plummer SJ, Moreno V, et al. Oncogenic Features in Histologically Normal Mucosa: Novel Insights Into Field Effect From a Mega-Analysis of Colorectal Transcriptomes. *Clin Transl Gastroenterol.* 2020 Jul;11(7):e00210.
74. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. The Functional Impact of Alternative Splicing in Cancer [Internet]. Vol. 20, *Cell Reports*. 2017. p.



- 2215–26. Available from: <http://dx.doi.org/10.1016/j.celrep.2017.08.012>
75. The GTEx Consortium. The GTEx eQTL Dashboard [Internet]. 2021 [cited 2021 Jul 25]. Available from:  
<https://www.gtexportal.org/home/eqtlDashboardPage>
76. The GTEx Consortium. The GTEx eQTL Calculator [Internet]. 2021 [cited 2021 Jul 25]. Available from: <https://www.gtexportal.org/home/testyourown>
77. Margolis KG, Gershon MD. Enteric Neuronal Regulation of Intestinal Inflammation. *Trends Neurosci*. 2016 Sep;39(9):614–24.
78. Taman H, Fenton CG, Hensel IV, Anderssen E, Florholmen J, Paulssen RH. Transcriptomic Landscape of Treatment-Naïve Ulcerative Colitis. *J Crohns Colitis*. 2018 Feb 28;12(3):327–36.
79. Manzoni C, Kia DA, Vandrovцова J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Brief Bioinform*. 2018 Mar 1;19(2):286–302.
80. Lichou F, Trynka G. Functional studies of GWAS variants are gaining momentum. *Nat Commun*. 2020 Dec 8;11(1):6283.
81. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. 2016 May;48(5):481–7.
82. Callaway E. New concerns raised over value of genome-wide disease studies [Internet]. Vol. 546, *Nature*. 2017. p. 463–463. Available from:  
<http://dx.doi.org/10.1038/nature.2017.22152>
83. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From

Polygenic to Omnigenic. *Cell*. 2017 Jun 15;169(7):1177–86.

84. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, et al. A manifesto for reproducible science. *Nat Hum Behav*. 2017 Jan 10;1:0021.
85. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature*. 2016 May 26;533(7604):452–4.