Journal of Crohn's and Colitis, 2025, **19**, jjae155
https://doi.org/10.1093/ecco-jcc/jjae155
Advance access publication 27 September 2024
Original Article

The Intestinal Mucin Isoform Landscape Reveals Region-Specific Biomarker Panels for Inflammatory Bowel Disease Patient Stratification

Wout Arras,[a,b] Tom Breugelmans,[a,b] Baptiste Oosterlinck,[a,b] Joris G. De Man,[a,b] Surbhi Malhotra-Kumar,[c] Steven Abrams,[d,e] Steven Van Laere,[f] Elisabeth Macken,[g] Michaël Somers,[g] Aranzazu Jauregui-Amezaga,[g] Benedicte Y. De Winter,[a,b,g,*] Annemieke Smet[a,b,*]

[a]Laboratory of Experimental Medicine and Pediatrics, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium
[b]Infla-Med, Centre of Excellence, University of Antwerp, Antwerp, Belgium
[c]Laboratory of Medical Microbiology, Vaccine and Infectious Disease Institute, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium
[d]Global Health Institute, Department of Family Medicine and Population Health, University of Antwerp, Antwerp, Belgium
[e]Data Science Institute, Interuniversity Institute for Biostatistics and statistical Bioinformatics, University of Hasselt, Diepenbeek, Belgium
[f]Center for Oncological Research, Integrated Personalized and Precision Oncology Network, University of Antwerp, Antwerp, Belgium
[g]Division of Gastroenterology and Hepatology, Antwerp University Hospital, Edegem, Belgium

Corresponding author: Annemieke Smet, Laboratory of Experimental Medicine and Pediatrics, Faculty of Medicine and Health Sciences, University of Antwerp, Universiteitsplein 1, 2610 Antwerp, Belgium. Tel: +32 3 265 27 11; Email: Annemieke.smet@uantwerpen.be
*Authors contributed equally.

Abstract

Background and aims: Mucosal healing is considered a key therapeutic endpoint in inflammatory bowel diseases (IBD) and comprises endoscopic improvement of inflammation without taking barrier healing into account. Mucins are critical components of the mucosal barrier function that give rise to structurally diverse isoforms. Unraveling disease-associated mucin isoforms that could act as an indication for barrier function would greatly enhance IBD management.

Methods: We present the intestinal mucin RNA isoform landscape in IBD and control patients using a targeted mucin isoform sequencing approach on a discovery cohort (n = 106). Random Forest modeling (n = 1683 samples) with external validation (n = 130 samples) identified unique mucin RNA isoform panels that accurately stratified IBD patients in multiple subpopulations based on inflammation, IBD subtype (Crohn's disease [CD], ulcerative colitis [UC]), and anatomical location of the intestinal tract (i.e. ileum, proximal colon, distal colon, and rectum).

Results: Particularly, the mucin RNA isoform panels obtained from the inflamed UC and CD distal colon showed high performance in distinguishing inflamed biopsies from their control counterparts (AUC of 93.3% and 91.1% in the training, 95.0% and 96.0% in the test, and 89.5% and 78.3% in the external validation datasets, respectively). Furthermore, the differentially expressed *MUC4* (PB.1238.363), *MUC5AC* (PB.2811.15), *MUC16* (ENST00000397910.8), and *MUC1* (ENST00000462317.5 and ENST00000620103.4) RNA isoforms frequently occurred throughout the different panels highlighting their role in IBD pathogenesis.

Conclusions: We unveiled region-specific mucin RNA isoform panels capturing the heterogeneity of the IBD patient population and showing great potential to indicate barrier function in IBD patients.

Key Words: Splice-variant; barrier dysfunction; biomarker

1. Introduction

Crohn's disease (CD) and ulcerative colitis (UC), collectively known as inflammatory bowel diseases (IBD), are chronic, progressive, and heterogeneous diseases characterized by recurring inflammation of the gastrointestinal tract in association with mucosal barrier dysfunction and gut dysbiosis. The clinical symptoms usually involve severe diarrhea, abdominal pain, fatigue, and weight loss.[1] The incidence of IBD is increasing with 6.8 million cases globally.[2] Due to the multifactorial etiology and absence of a specific therapeutic target for IBD patients, treatment relies on controlling the exacerbations/flares of the disease.[3] Numerous therapies are available, but the current knowledge does not help clinicians to choose the right therapy for the right patient, forcing them to do so empirically. As a result, approximately one-third of the patients fail to respond to the initial biologic therapy and up to 50% lose response over time.[1,4,5] According to the recent guidelines, mucosal healing is considered a key therapeutic endpoint for IBD and encompasses the restoration of the mucosal barrier and the resolution of inflammation.[6] Mucosal healing, as evaluated through endoscopy, is often combined with symptom relief and/or histological assessment of intestinal biopsies to predict long-term remission which, however, only measures the presence of inflammation without taking barrier function and thus the importance of barrier healing into account.[6,7] Biomarkers used to assess inflammation, like fecal calprotectin in stool and C-reactive protein in blood, have proven their worth in the follow-up of disease, but their predictive value for mucosal healing and therapy response is unsatisfactory.[8,9] Objective measurements at the molecular level to monitor mucosal healing are therefore lacking and identifying biomarkers that act as an indication for mucosal healing could significantly improve disease monitoring and treatment success in IBD patients.

Intestinal mucosal barrier dysfunction is generally accepted as an important contributor to these diseases.[1,10] Hence, characterization of the intestinal mucosal barrier in IBD could provide novel biomarkers, permitting an accurate and robust assessment of treatment response by monitoring mucosal barrier repairment in addition to inflammation, and thus mucosal healing at the molecular level. The general architecture of the intestinal mucosal barrier features a thick mucus layer, a single layer of epithelial cells, and the inner lamina propria hosting innate and adaptive immune cells. Mucins are the gatekeepers of the mucus barrier and are expressed either as secretory (MUC2, MUC5AC/B, MUC6, and MUC19) or as transmembrane glycoproteins (MUC1, MUC3A, MUC4, MUC12, MUC13, MUC15, MUC16, MUC17, MUC20, and MUC21). The secreted mucins, produced by goblet cells, are the major constituents of the mucus layer. Below the mucus layer, epithelial cells that do not secrete mucus, present a dense network of highly diverse transmembrane mucins which form the glycocalyx.[11,12] Besides their protective function, transmembrane mucins possess several epidermal growth factor (EGF) domains on their extracellular tail and numerous phosphorylation sites on their intracellular tail enabling them to participate in intracellular signal transduction and to play an important role in epithelial cell homeostasis.[12] Aberrant mucin expression, characterized by a reduced MUC2 secretion and overexpression of transmembrane mucins, has been described in IBD[11,13–15] as well as the active involvement of transmembrane mucins in mucosal barrier dysfunction by affecting junctional protein expression and cell polarity through JAK/STAT, SNAI1/ZEB1, and ROCK2/MAPK signaling.[13] In addition, we also showed that aberrant mucin signatures associate with IBD presentation and activity, underlining their potential use as molecular markers to aid in disease management.[16]

Furthermore, mucins are highly polymorphic and the presence of genetic alterations[17–19] can affect alternative mucin gene splicing resulting in a large repertoire of structurally diverse isoforms. While most RNA isoforms produced from the same mucin gene locus encode similar biological functions, others may alter protein function resulting in the progression towards disease.[12,19,20] Such disease-associated mucin RNA isoforms can thus act as novel biomarkers to mirror mucosal barrier function. Currently, the structural heterogeneity of mucin isoforms involved in mucosal barrier dysfunction in IBD remains unexplored. Therefore, in this study, we performed a novel targeted long-read mucin RNA isoform sequencing approach in conjunction with short-read bulk RNA sequencing to map and quantify the mucin RNA isoform landscape in the intestinal tract of IBD and control patients. Downstream bioinformatics analysis based on Random Forest classification and external validation further unveiled unique and distinct mucin RNA isoform panels that stratified the heterogeneous IBD patient population with high performance in multiple subpopulations based on inflammation, IBD subtype (i.e. CD and UC) and anatomical location in the intestinal tract.

2. Methods

2.1. Ethics approval statement

The study was conducted in accordance with the Declaration of Helsinki and approval for the study protocol was granted by the Ethics Committee of the University Hospital of Antwerp, Belgium (registration numbers B300201733423 and B3002020000162). A written informed consent was obtained from all participants prior to their enrollment in the study. Samples were registered and stored until analysis in the Biobank Antwerpen, Antwerp, Belgium (BE 71 030 031 000; BBMR-ERIC, Belgian no. access: 1, Last: April 10, 2021 [BIORESOURCE]).

2.2. Patients and sample collection

IBD patients undergoing endoscopy for medical indications (i.e. acute flare or follow-up when in remission) were recruited at the policlinic of the Department Gastroenterology and Hepatology of the Antwerp University Hospital, a tertiary hospital in Belgium from 2018 to 2023 (Supplementary Table 1). During endoscopy, ileal and colonic biopsies were collected from macroscopically inflamed and noninflamed regions of the intestinal tract. Colonic biopsies are subdivided according to their anatomical origin: proximal colon (including the cecum, ascending colon, and transverse colon), distal colon (comprising the descending colon, sigmoid colon, and rectum), and rectum. Patients without a history of IBD undergoing an endoscopy due to a positive immunological fecal occult blood test (iFOBT) showing no endoscopic inflammation, were included as control patients. Biopsies were immediately submerged in RNA later (Sigma), fresh-frozen in liquid nitrogen, and stored at −80°C until RNA extraction.

2.3. RNA isolation and quality control

Total RNA was extracted from 106 intestinal biopsies (Supplementary Table 1) using the Nucleospin RNA Plus kit (Macherey-Nagel) according to the manufacturer's instructions. The purity of the RNA was assessed through spectrophotometric analysis with the NanoDrop ND-1000 (Thermo Fisher Scientific). Concentration and RNA quality were evaluated with the Qubit Fluorometer (Qubit Broad Range RNA kit, Thermo Fisher Scientific) and 2100 Bioanalyzer (RNA 6000 Nano kit, Agilent Technologies).

2.4. Multiplex targeted isoform sequencing: library preparation and single-molecule real-time sequencing

The Pacific Biosciences (PacBio) Sequel platform was used to sequence the RNA samples extracted from the 106 intestinal biopsies randomly allocated to different batches (Figure 1A). From the total RNA, a cDNA library was generated by reverse transcription of full-length RNA transcripts with the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (New England Biolabs). Barcoded oligo-dT primers were used to allow for multiplexing (maximum 12-plex) after which hybrid capture was performed by using a custom NGS discovery pool (IDT) developed for the capture of all mucin genes present in the human reference

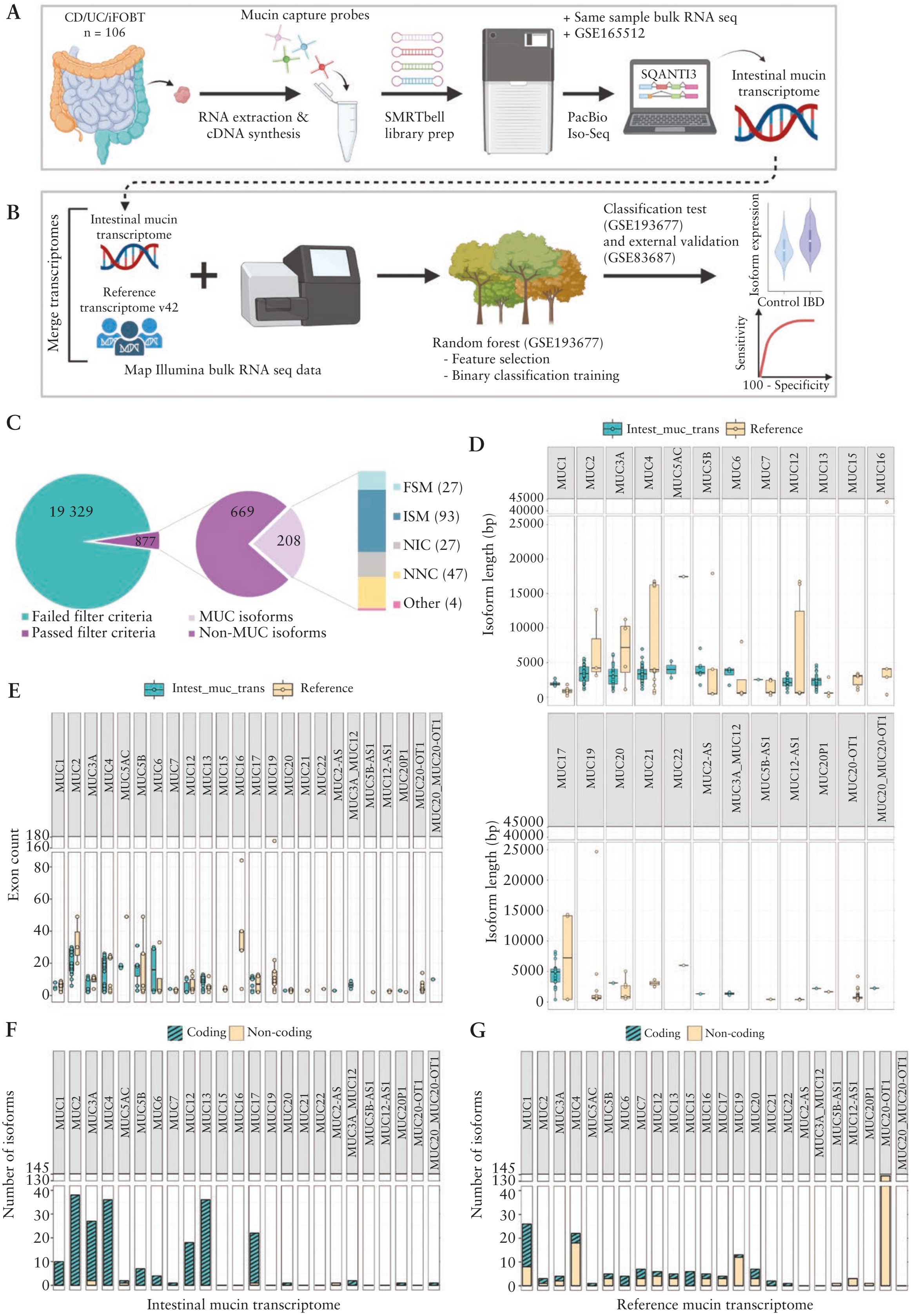


Figure 1 Overview of the methodological approach and datasets used for (A) establishing the intestinal mucin transcriptome and (B) evaluating the biomarker potential of mucin RNA isoforms. (C) Schematic overview of the filtering steps performed to obtain the intestinal mucin transcriptome. (D) Comparison of isoform length between the mucin RNA isoforms from the intestinal mucin transcriptome and the human reference transcriptome. (E) Comparison of exon count between the mucin RNA isoforms from the intestinal mucin transcriptome and the human reference transcriptome. (F) Number of isoforms in the intestinal mucin transcriptome grouped per mucin gene with accompanying coding potential. (G) Number of isoforms in the human reference transcriptome grouped per mucin gene with accompanying coding potential. In the boxplots, the horizontal line within the box represents the median, the lower hinge of the box denotes the first quartile and the upper hinge corresponds to the third quartile. Whiskers extend to the largest data point within 1.5 times the interquartile range of the upper and lower hinge. Data points beyond these ranges are plotted individually. Abbreviations: FSM, full splice match; ISM, incomplete splice match; NIC, novel in catalog; NNC, novel not in catalog. Created with BioRender.com.

genome GRCh38, i.e. *MUC1*, *MUC2*, *MUC3A*, *MUC4*, *MUC5AC*, *MUC5B*, *MUC6*, *MUC7*, *MUC12*, *MUC13*, *MUC15*, *MUC16*, *MUC17*, *MUC19*, *MUC20*, *MUC21*, and *MUC22* gene transcripts. This pool, based on the gene annotations of GRCh38, consists of high-fidelity, individually synthesized, 5′-biotinylated oligos targeting the exons and 5′ and 3′ UTR regions of the 17 mucins (Supplementary Table 2). After the removal of the probes with high risk for off-target capturing, a pool of 2032 probes (containing 1704 probes with low and 328 probes with moderate off-target risk) was generated (Figure 1A). The captured cDNA was amplified and after SMRTbell library preparation, samples were sequenced on an SMRTcell 1M v3 LR tray.

2.5. RNA-seq library preparation and Illumina sequencing

Total RNA from a subset of intestinal biopsies (i.e. 48 out of the 106 biopsies; Supplementary Table 3) was used to obtain a poly-A enriched library for bulk RNA sequencing on the Illumina NovaSeq 6000 platform (v1.5 Reagent kit) with 150 bp unstranded paired-end reads (Figure 1A).

2.6. Publicly available datasets for model building and external validation

GSE165512 (170 samples),[21] GSE83687 (134 samples),[22] and GSE193677 (2489 samples)[23] comprising Illumina short-read sequencing data of colonic and ileal biopsies of IBD and control patients were retrieved from the Gene Expression Omnibus using the sratoolkit (v3.0.0). From the GSE83687 dataset, 4 samples were removed because they originated from noninflamed regions or the biopsied location was not known. From the GSE193677 dataset, we removed 36 samples, either because they were retrieved from patients who had undergone pouch surgery or due to the absence of an endoscopic severity level in the metadata. In addition, 770 endoscopic samples were removed because they could not be unambiguously classified as inflamed or noninflamed. All Illumina short-read data were trimmed with fastP (v0.20.0) prior to the analysis.

2.7. Downstream bioinformatics analysis pipeline

2.7.1. Establishment of the intestinal mucin RNA isoform landscape

PacBio Isoseq raw reads were analyzed using the IsoSeq 3 pipeline in Linux bash (v5.0.17(1)). In short, starting from the raw sequencing reads from SMRTlink, circular consensus sequences reads were generated (ccs v6.4.0). From the resulting isoform sequences (≥Q20), primers were removed and the samples demultiplexed (lima v2.6.0). Poly(A) tails and concatemers were removed from the obtained full-length reads (isoseq refine v3.8.0) and successively, the full-length non-concatemer reads were clustered (isoseq cluster, v3.8.0). The resulting high-quality isoforms were aligned to the human genome assembly GRCh38 (pbmm2 v1.9.0) and redundant isoforms collapsed to obtain count information on unique isoforms (coverage 99% and identity 95%, isoseq collapse v3.8.0). For further classification, quality control, and rigorous filtering of the isoforms, SQANTI3 (v3.5.1)[24,25] was used in conjunction with supporting short reads (Illumina) from a subset of 48 samples, also sequenced on the PacBio platform, and from the publicly available dataset GSE165512 by using STAR (v2.7.10a)[26] and Kallisto (v0.48.0)[27] (Figure 1A). In addition to the default SQANTI3 filter, we incorporated 4 supplementary rules being that all isoforms 1) should have a count ≥ 2 in 3 or more PacBio sequenced samples, 2) have an expression higher than zero according to the Illumina sequenced samples (except when the transcript is classified as a full splice match), 3) are bite negative, and 4) map to a mucin gene. Failure to comply with one of these rules resulted in the exclusion of the isoform. The obtained intestinal mucin RNA isoform landscape was merged with the human reference transcriptome GRCh38 (Gencode release 42) by using gffcompare (v0.12.6)[28] (Figure 1B).

2.7.2. Model building based on Random Forest classification

Bulk RNA sequencing data from the GSE193677 dataset was quantified based on the transcripts present in the combined MUC transcriptome (Kallisto v0.48.0) for colon ($n = 1221$) and ileum samples ($n = 462$), separately. Reads were normalized (sleuth v.0.30.1)[29] and only mucin RNA isoforms with a minimal count of 5 in 30 or more samples were retained. The resulting count data were log2 transformed and transcripts of which the expression was highly correlated (>75%) were removed (caret v6.0.94). The GSE193677 dataset was randomly split into a training dataset containing 80% of the samples and a test dataset with the remaining 20% of the samples. Additional dataset manipulation was done with dplyr (v1.1.2). Before model training, the training data were balanced when necessary by undersampling. Using the training data, mucin RNA isoforms with high importance were identified based on a Random Forest algorithm adapted from Brieuc et al.[30] (Figure 1B). In short, for each binary comparison, importance values for every mucin RNA isoform were estimated. Based on their importance, mucin RNA isoforms were divided into different groups. The group with the lowest out-of-bag error rate was selected and used for backward purging to obtain the final mucin RNA isoform panel. To minimize the chance of overfitting, the number of isoforms used for each classification was optimized by selecting the panel with the fewest isoforms within the 2% range of the panel with the lowest out-of-bag error rate (OOB-ER; randomForest v4.7.1.1; Figure 1B).[31] GSE83687 was used as a dataset for external validation.

2.8. Statistical analysis and data visualization

Boxplots and bar charts visualizing the characteristics of the intestinal mucin transcriptome (isoform length, exon amount, coding potential, and isoform counts), heatmaps with mucin isoform expression data, and violin plots showing mucin RNA isoform expression for each panel were made by using dplyr (v1.1.2), tidyr (v1.3.0), ggplot2 (v3.4.4), ggpattern (v1.0.1), ggbreak (v0.1.2), and ComplexHeatmap (v2.16.0).[32] The performance of the different Random Forest models when applied to the training, test, and external validation data are presented as receiver operating characteristic (ROC) curves with accompanying area under the receiver operating characteristic curve (AUC) generated by using the pROC package (v1.18.4) in R Visualization of the exon-intron structure of the mucin RNA isoforms and the canonical isoform for each mucin presented was done by using transPlotR (v0.0.2)[33] and rtracklayer (v1.62.0). Tables with patient characteristics were generated with finalfit (v1.0.7), in which statistical hypothesis testing for differences between levels of categorical variables was done using Chi-square tests and the significance of differences in group-specific means of continuous variables was assessed using an omnibus ANOVA F-test. The correlation between age and mucin isoform expression was assessed using Spearman's rank correlation test (stats, v4.3.1) and visualized using corrplot (v0.92). Differential expression of mucin isoforms was done relying on the Wilcoxon rank-sum test corrected for multiple testing within each panel using the Bonferroni method. All analyses, including Random Forest classification, were carried out by using R (v4.3.1) and Rstudio (v2023.03.0 + 386).

3. Results

3.1. Discovery of the intestinal mucin RNA isoform landscape

We developed a targeted long-read mucin isoform sequencing pipeline enabling the identification of mucin RNA isoform transcripts expressed in the intestinal tract of IBD and control patients (Figure 1A). Patient demographics and clinical characteristics are shown in Supplementary Table 1. A total of 20 206 unique RNA isoforms were identified. After rigorous filtering, facilitated by the addition of short-read sequencing data from our own IBD and control subcohorts (Supplementary Table 3) and the public dataset GSE165512 (Figure 1A), 877 isoforms were retained, from which 208 isoform transcripts originated from a mucin gene (Figure 1C). These included RNA isoforms derived from *MUC1*, *MUC2*, *MUC3A*, *MUC4*, *MUC5AC*, *MUC5B*, *MUC6*, *MUC7*, *MUC12*, *MUC13*, *MUC17*, and *MUC20* (Figure 1D-1G). In addition, one RNA isoform mapped to the *MUC20* pseudogene 1 (designated as *MUC20P1*) and one to an unidentified gene on the antisense strand at the chromosomal location of *MUC2* (designated as *MUC2-AS*). Two RNA isoforms aligned to exons from both *MUC3A* and *MUC12* (designated as *MUC3A_MUC12*) and one to sequences originating from both *MUC20* and *MUC20-OT1* (designated as *MUC20_MUC20-OT1*; Figure 1D-1G). To further characterize the intestinal mucin RNA isoform landscape in our patient population, assigned as the "intestinal mucin transcriptome" (Figure 1), the isoform length, exon count, and coding potential were compared to the mucin RNA isoforms present in the human reference transcriptome. The isoform length of the intestinal mucin transcriptome varied between 385 and 8241 base pairs (bp) whereas the human reference transcriptome contained relatively large mucin RNA isoforms (Figure 1D). This was further reflected by the difference in exon count observed in both transcriptomes (Figure 1E). Furthermore, most of the mucin RNA isoforms from the intestinal mucin transcriptome are transcripts with protein-coding potential (Figure 1F), which contrasts with the mucin RNA isoforms residing in the reference mucin transcriptome showing a more proportional distribution between coding and noncoding isoforms (Figure 1G). Of note, the intestinal mucin RNA isoform landscape comprised distinctly more transcripts of *MUC2*, *MUC3A*, *MUC4*, *MUC12*, *MUC13*, and *MUC17* whereas the human reference transcriptome contained RNA isoforms of particularly the *MUC1*, *MUC7*, *MUC15*, *MUC19*, *MUC20*, and *MUC20-OT1* genes (Figure 1F-1G). All 268 mucin RNA isoforms present in the human reference transcriptome were merged with our intestinal mucin transcriptome (i.e. 208 mucin RNA isoforms) resulting in a combined mucin RNA isoform landscape of 386 unique transcripts (Figure 1B). More specifically, 118 of the 208 intestinal mucin RNA isoform transcripts were considered novel and designated with a PB number. Ninety were merged with an existing mucin RNA isoform from the human reference transcriptome and are represented with an ENST transcript ID. Furthermore, after mapping the short-read sequencing data of the GSE193677 dataset ($n = 1683$) to the combined mucin isoform landscape, 289 mucin RNA isoforms were assigned to be expressed in the colon and 261 in the ileum (Supplementary Table 4). The overall abundance of each mucin RNA isoform expressed in the colon and ileum of IBD and control patients is represented in a heatmap (Figure 2A, 2B). A relatively high abundance of *MUC2*, *MUC4*, *MUC5AC*, *MUC5B*, *MUC12*, and *MUC13* RNA isoforms was shown in the colon of IBD patients with clear differences seen between the proximal and distal colonic region and inflamed and noninflamed biopsies for *MUC12* RNA isoform expression (Figure 2A). On the contrary, *MUC12* RNA isoforms appeared to be less abundant in the terminal ileum of IBD patients, whereas a clear difference was noted in *MUC1*, *MUC4*, *MUC5AC*, *MUC5B*, *MUC6*, and *MUC12* RNA isoform expression depending on the inflammatory status in the ileum and population cohort (Figure 2B).

3.2. Mucin RNA isoform features as discriminators for general IBD and subtype classification

To determine whether mucin RNA isoforms can be used as discriminators to classify biopsies derived from IBD patients and its subtypes (i.e. UC and CD) in the presence/absence of inflammation from control patients, Random Forest-based feature selection was carried out on the combined mucin RNA isoform landscape extracting specific mucin RNA isoform panels associated with IBD, UC, or CD compared to control patients based on their OOB-ER. An equal distribution of sex between the groups for each model (Supplementary Tables 5-24) and no significant correlation between the expression of mucin isoforms, selected for the panels, and age was found (Supplementary Figure 1). The expression level of each mucin RNA isoform in the panel was then used to train, test, and validate the model for binary classification. Their discriminative performance is presented using ROC curves with accompanying AUC values and the expression levels of the mucin RNA isoforms of each respective panel are shown in violin plots (Figures 3 and 4; Supplementary Figures 2 and 3). The characteristics of the combined mucin RNA isoform landscape and patient demographics for each model are summarized in Supplementary Tables 4-24. The mucin RNA isoforms occurring throughout both the inflamed and noninflamed panels are schematically depicted as Venn diagrams in Figure 5 and Supplementary Figure 4.

Feature selection to distinguish IBD patients with colonic inflammation from control patients without inflammation in their intestinal tract unveiled a panel of 10 mucin RNA isoforms originating from 8 different mucin genes and half of them were identified as novel (Figure 3A, 3B; Supplementary Tables 4, 5). This panel performed well in the training and test dataset (AUC of 92.3% and 92.5%, respectively). Its performance decreased slightly in the external validation dataset to an AUC of 72.2% (Figure 3A). All 10 isoforms were significantly increased in the IBD cohort compared to the control patients, except for the *MUC20* (ENST00000447234.7) and *MUC12* (ENST00000473098.5) RNA isoforms which were significantly downregulated in the IBD patient group (Figure 3B).

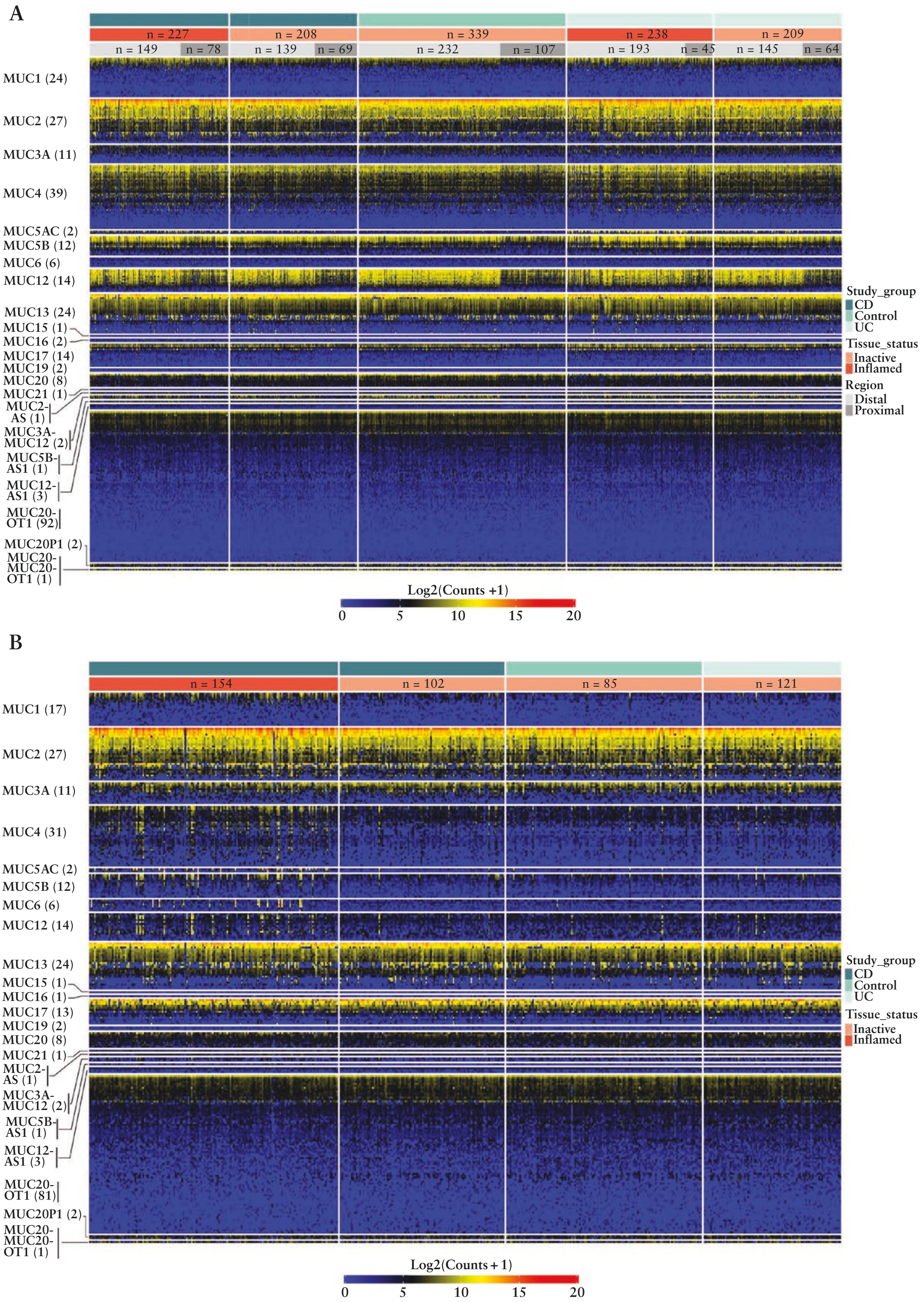


Figure 2. (A) Heatmap visualization of the mucin RNA isoform expression data in colonic (i.e. proximal and distal regions) inflamed and noninflamed biopsies of Crohn's disease and ulcerative colitis patients, and noninflamed biopsies of control patients. (B) Heatmap visualization of the mucin RNA isoform expression data in ileal biopsies originating from an inflamed and noninflamed region of Crohn's disease patients and a noninflamed region of ulcerative colitis and control patients. The number of patients (*n*) in each group is shown above each heatmap.

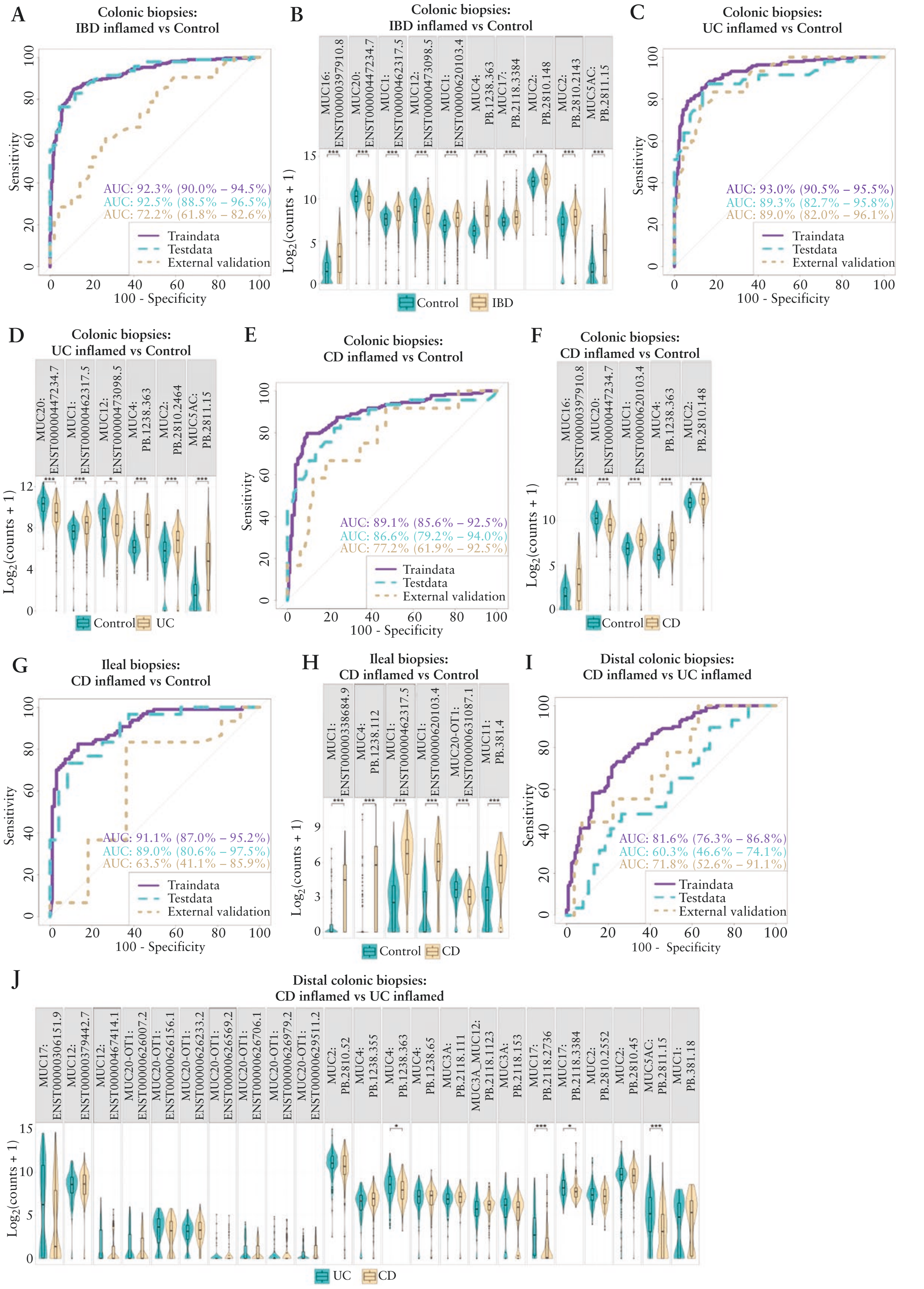


Figure 3. Mucin RNA isoform expression levels as variables associated with IBD and subtypes (UC or CD) upon inflammation. Overview of model performance for training (80% of GSE193677), test (20% of GSE193677), and external validation (GSE83687) datasets with accompanying violin plots, containing the mucin RNA isoform expression panels as predictors for a certain classification are provided. ROC curve, AUC, and violin plots for the prediction model of (A, B) inflamed colonic biopsies of IBD patients (n_{Train} = 272, n_{Test} = 93, $n_{\text{ext. val.}}$ = 42) versus noninflamed colonic biopsies of control patients (n_{Train} = 272, n_{Test} = 67, $n_{\text{ext. val.}}$ = 49), (C, D) inflamed colonic biopsies of UC patients (n_{Train} = 191, n_{Test} = 47, $n_{\text{ext. val.}}$ = 30) versus noninflamed colonic biopsies of control patients (n_{Train} = 191, n_{Test} = 67, $n_{\text{ext. val.}}$ = 49), (E, F) inflamed colonic biopsies of CD patients (n_{Train} = 182, n_{Test} = 45, $n_{\text{ext. val.}}$ = 12)

By using the same approach, we identified several panels to distinguish UC and CD patients from control patients. These included a panel of 6 mucin RNA isoforms that accurately discriminated inflamed colonic biopsies of UC patients from noninflamed colonic control biopsies (i.e. AUC of 93.0% [based on the training data], AUC of 89.3% [test data]; Figure 3C, 3D; Supplementary Tables 4, 6), a panel of 5 mucin RNA isoforms as a major determinant for inflammation in the colon of CD patients (AUC of 89.1% [training data], AUC of 86.6% [test data]; Figure 3E, 3F; Supplementary Tables 4, 7) and a panel of 6 mucin RNA isoforms that associated with inflammation in the ileum of CD patients compared to control patients (AUC of 91.1% [training data], AUC of 89.0% [test data]; Figure 3G, 3H; Supplementary Tables 4, 8). Within these panels, each mucin RNA isoform originated from a different mucin gene (Figure 3D, 3F) except for the CD ileum inflamed panel where 4 RNA isoforms originated from the *MUC1* gene (Figure 3H). Performance of the UC colon and CD colon inflamed models when applied to the external validation dataset remained remarkably high (i.e. AUC values of 89.0% and 77.2%, respectively; Figure 3C, 3E). Interestingly, the *MUC4* (PB.1238.363) and *MUC20* (ENST00000447234.7) RNA isoforms were shared by the UC colon inflamed and CD colon inflamed panels (Figures 3D, 3F, 5) whereas the *MUC1* RNA isoform (ENST00000462317.5) was common between the UC colon inflamed and CD ileum inflamed panels (Figures 3D, 3H, 5) and the *MUC1* RNA isoform (ENST00000620103.4) between the CD colon inflamed and CD ileum inflamed panels (Figures 3F, 3H, 5). Most of the mucin RNA isoforms from the three models were upregulated upon inflammation in the UC or CD group (Figure 3D, 3F, 3H). Only the *MUC12* (ENST00000473098.5; UC colon inflamed panel; Figure 3D), *MUC20* (ENST00000447234.7; UC and CD colon inflamed panel; Figure 3D, 3F), and *MUC20-OT1* (ENST00000631087.1; CD inflamed ileum panel; Figure 3H) RNA isoforms were downregulated in the presence of inflammation.

In addition, we also designed prediction models to classify IBD patients from control patients in the absence of inflammation. However, due to the absence of noninflamed biopsies in the IBD cohort of the external validation dataset, ROC curves and AUC values could only be obtained for the training and test data. Training the model to distinguish noninflamed colonic biopsies of IBD patients from noninflamed control biopsies based on a panel of 10 mucin RNA isoforms resulted in an AUC of 78.4% for the training data and 68.2% for the test data. Interestingly, *MUC5AC* (PB.2811.15) that resided in the IBD colon inflamed panel was also found to be overexpressed in the IBD noninflamed panel (Supplementary Figure 2A, 2B and Tables 4, 9; Figure 3B). In addition, the *MUC5B* (ENST00000525715.5, ENST00000527802.1, and PB.2816.52), *MUC3A_MUC12* (PB.2118.1123) and *MUC2* (PB.2810.148) RNA isoforms were also overexpressed in the noninflamed IBD group, while *MUC20-OT1* (ENST00000446521.1) was downregulated (Supplementary Figure 2B). The performance of the models distinguishing between noninflamed biopsies from CD or UC patients and control patients was similar in the test and training datasets (Supplementary Figure 2C-2F and Tables 10, 11). Only the *MUC5AC* (PB.2811.15) and *MUC20* (ENST00000447234.7) RNA isoforms of the UC colon inflamed panel (Figure 3D) were also identified in the UC colon noninflamed panel (Supplementary Figure 2D), whereas the *MUC2* (PB.2810.148) and *MUC1* (ENST00000620103.4) RNA isoforms were shared among the CD colon and ileum inflamed/noninflamed panels, respectively (Figure 3F, 3H; Supplementary Figure 2F, 2H and Tables 4, 12). Notably, the *MUC5AC* (PB.2811.15) RNA isoform was overexpressed and *MUC20* (ENST00000447234.7) RNA isoform downregulated in UC patients compared to the control group independent of inflammation (Figure 3D; Supplementary Figure 2D). In the biopsies of CD patients, the *MUC1* (ENST00000620103.4) and *MUC2* (PB.2810.148) RNA isoforms were upregulated in both the inflamed and noninflamed biopsies compared to the controls (Figure 3F, 3H; Supplementary Figure 2F, 2H). Setting up a prediction model that distinguished between colonic biopsies from CD and UC patients in the presence/absence of inflammation was shown to be difficult as reflected by the lower AUC values for the different datasets (Figure 3I; Supplementary Figure 2I and Tables 13, 14). Both panels encompassed a high number of mucin RNA isoforms (Figure 3J; Supplementary Figure 2J), particularly in the presence of inflammation, in which the majority of isoforms were not differentially expressed.

3.3. Region-specific mucin isoform panels for UC and CD patient subpopulation stratification

One of the key features that attribute to disease heterogeneity is the variable inflammatory pattern expressed in the intestinal tract of IBD patients.[1] More specifically, UC patients have a more uniform mucosal inflammation that starts in the rectum and further extends toward the distal part of the colon whereas patchy and transmural inflammatory areas are seen in the terminal ileum and throughout the colon of CD patients.[1] Here, we further investigated whether the Random Forest approach for IBD subtype classification could be improved when integrating the colonic disease location, i.e. proximal colon (encompassing cecum, ascending and transverse colon), distal colon (encompassing descending colon, sigmoid and rectum), and rectum (Figure 4; Supplementary Figures 2, 3). Feature selection to distinguish UC patients with inflammation in the distal colon from control patients unveiled a panel of 4 differentially expressed mucin RNA isoforms with similar high performance (i.e. AUC values of 93.3% (training data), 95.0% (test data) and 89.5% (external validation data); Figure 4A, 4B; Supplementary Tables 4, 15) compared to the UC colon inflamed panel (Figure 3D). Interestingly, the *MUC4* (PB.1238.363), *MUC5AC* (PB2811.15), and *MUC20*

versus noninflamed colonic biopsies of control patients (n_{Train} = 182, n_{Test} = 67, $n_{\text{ext. val.}}$ = 49), (G, H) inflamed ileal biopsies of CD patients (n_{Train} = 97, n_{Test} = 30, $n_{\text{ext. val.}}$ = 30) versus noninflamed ileal biopsies of control patients (n_{Train} = 97, n_{Test} = 24, $n_{\text{ext. val.}}$ = 11), and (I, J) inflamed distal colonic biopsies of CD patients (n_{Train} = 120, n_{Test} = 38, $n_{\text{ext. val.}}$ = 9) versus inflamed distal colonic biopsies of UC patients (n_{Train} = 120, n_{Test} = 29, $n_{\text{ext. val.}}$ = 27). Differential mucin isoform expression was assessed using the Wilcoxon rank-sum test corrected for multiple testing using the Bonferroni method. Significant differences are indicated by *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. In the boxplots, the horizontal line within the box represents the median, the lower hinge of the box denotes the first quartile and the upper hinge corresponds to the third quartile. Whiskers extend to the largest data point within 1.5 times the interquartile range of the upper and lower hinge. Data points beyond these ranges are plotted individually. Abbreviations: C, control patients; colon-dist, distal colon; ext.val, external validation.

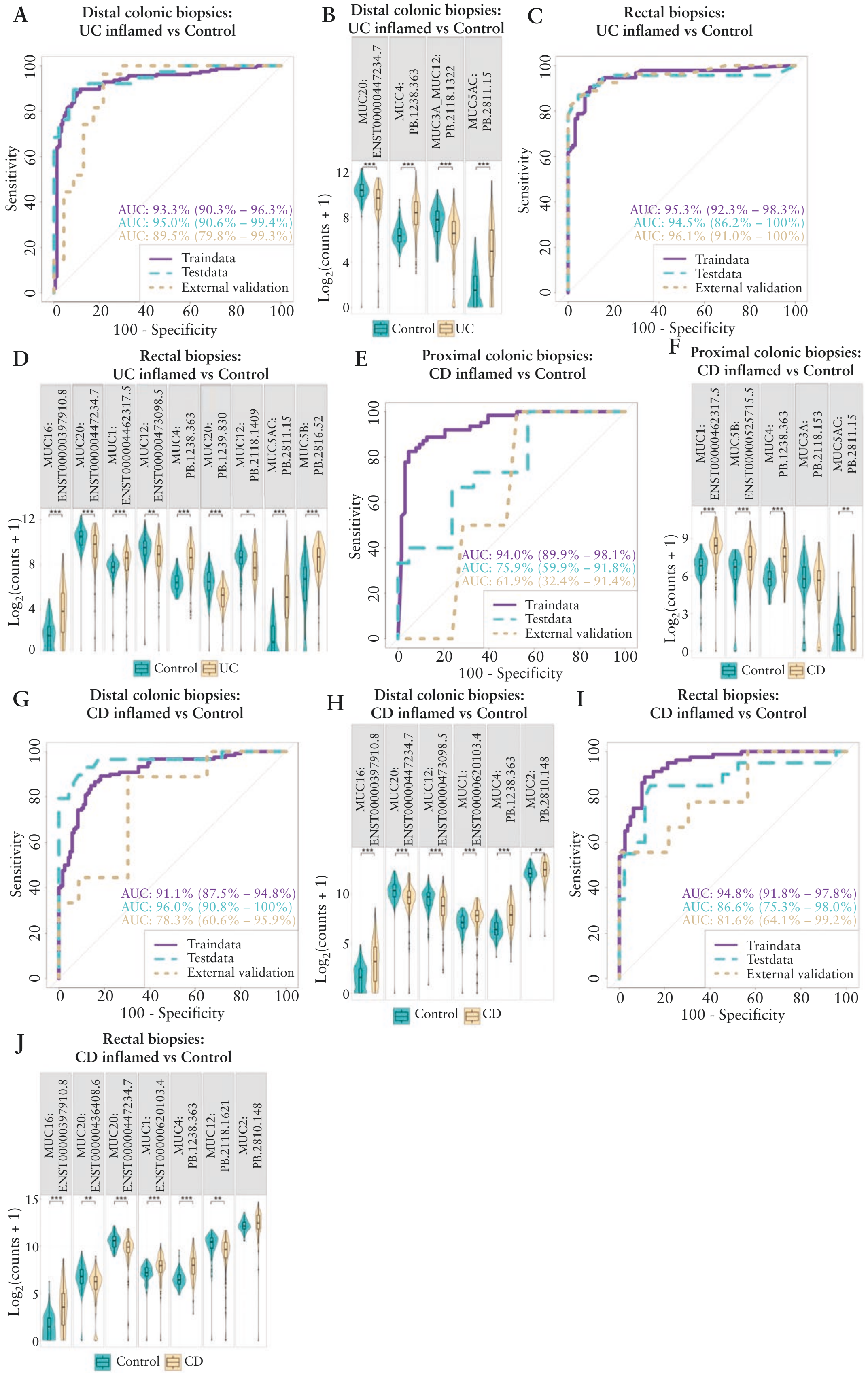


Figure 4. Mucin RNA isoform expression levels as variables associated with IBD subtypes (UC or CD) and location in the intestinal tract (i.e. proximal colon, distal colon, and rectum) upon inflammation. Overview of model performance for training (80% of GSE193677), test (20% of GSE193677), and external validation (GSE83687) datasets with accompanying violin plots, containing the mucin RNA isoform expression panels as predictors for a certain classification are provided. ROC curve, AUC, and violin plots for the prediction model of (A, B) inflamed distal colonic biopsies of UC patients (n_{Train} = 155, n_{Test} = 38, $n_{\text{ext. val.}}$ = 27) versus noninflamed biopsies of control patients (n_{Train} = 155, n_{Test} = 46, $n_{\text{ext. val.}}$ = 23), (C, D) inflamed rectal biopsies of

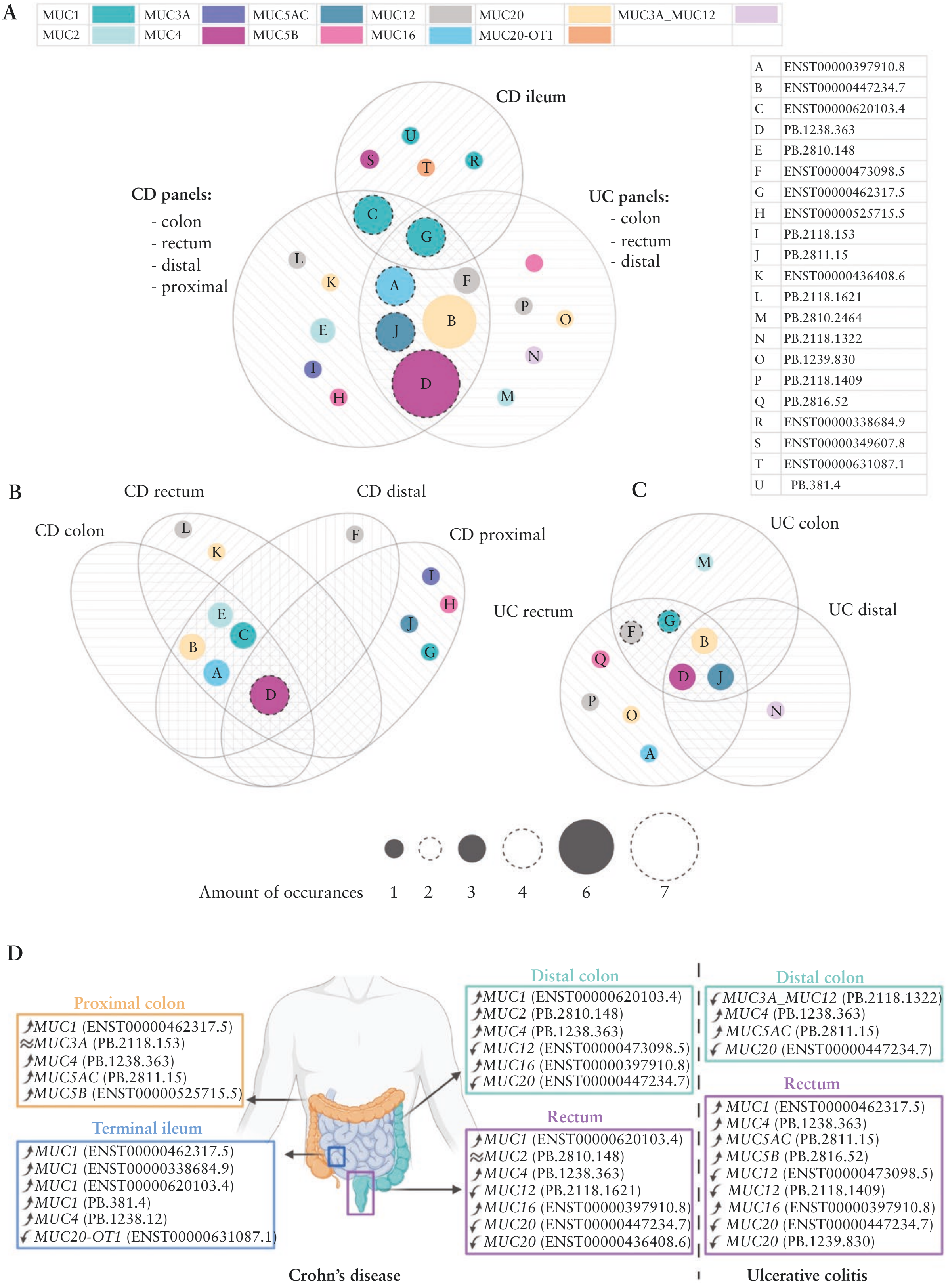
(ENST00000447234.7) RNA isoforms were common in both panels (Figures 3D, 4B, 5C). The prediction model built to discriminate UC patients with rectal inflammation from control patients revealed a panel of 9 differentially expressed mucin RNA isoforms (Figure 4C, 4D; Supplementary Tables 4, 16), of which 5 were common with the UC colon inflamed panel (Figures 3D, 4D, 5C) and 3 with the UC distal colon inflamed panel (Figures 4B, 4D, 5C). The panel for discriminating inflamed rectal biopsies of UC patients with noninflamed control biopsies also performed very well in the training, test, and external validation datasets with an AUC of 95.3%, 94.5%, and 96.1%, respectively (Figure 4C; note that for the rectal UC inflamed model, all distal colonic UC samples (including sigmoid and colon descendens) were used for external validation due to an insufficient number of rectal biopsies in this dataset). When training the Random Forest model to distinguish, inflamed proximal colon biopsies of CD patients from their noninflamed control counterpart (Supplementary Table 17), the AUC decreased to 75.9% in the test dataset and 61.9% in the external validation dataset (Figure 4E). However, the isoform panels built to discriminate inflamed distal colon or rectal biopsies of CD patients from their noninflamed control counterparts (Supplementary Tables 18, 19), demonstrated a slightly improved performance for the training (distal colon: AUC = 91.1%; rectum: AUC = 94.8%; Figure 4G, 4I), test (distal colon: AUC = 96.0%; rectum: AUC = 86.6%; Figure 4G, 4I), and external validation datasets (distal colon: AUC = 78.3%; rectum: AUC = 81.6%; Figure 4G, 4I) compared to the CD colon inflamed model (Figure 3E) (also for the rectal CD inflamed model, all distal colonic CD samples (including sigmoid and colon descendens) were used for external validation due to an insufficient number of rectal biopsies in this dataset). The isoforms in all three region-specific CD models were differentially expressed except for the *MUC3A* (PB.2118.153) RNA isoform in the CD proximal colon inflamed panel and the *MUC2* (PB.2810.148) RNA isoform in the CD rectum inflamed panel (Figure 4F, 4J). Interestingly, the CD colon distal inflamed and rectum inflamed panels encompassed all mucin RNA isoforms from the CD colon inflamed panel, in addition to 1 (*MUC12*: ENST00000473098.5) and 2 (*MUC12*: PB.2810.148; *MUC20*: ENST00000436408.6) RNA isoforms, respectively (Figures 3F, 4J, 4H, 5B; Supplementary Table 4). On the contrary, the CD proximal colon inflamed panel consisted of completely different mucin RNA isoforms ($n = 5$), except for the *MUC4* (PB.1238.363) RNA isoform, compared to the CD colon inflamed panel (Figures 4F, 5B; Supplementary Table 4).

Finally, we also investigated the performance of the region-specific UC and CD models in the absence of inflammation. Similarly, due to the absence of noninflamed biopsies in the UC and CD cohorts of the external validation dataset, ROC curves and AUC values could only be obtained for the training and test datasets (Supplementary Figure 3 and Tables 20–24). Distinguishing region-specific noninflamed biopsies in the colon of UC and CD patients compared to control patients resulted in high AUC values for the training data but performance decreased in the test data, specifically for UC and CD rectum models (Supplementary Figure 3). The UC distal colon noninflamed panel contained 3 mucin RNA isoforms (*MUC3A_MUC12* [PB.2118.1322], *MUC5AC* [PB.2811.15], and *MUC20* [ENST00000447234.7]) which also resided in the UC distal colon inflamed panel but with *MUC3A_MUC12* (PB.2118.1322) no longer being downregulated (Figure 4B; Supplementary Figure 3B), whereas all mucin RNA isoforms from the UC rectum noninflamed panel were differentially expressed and occurred in the UC rectum inflamed panel (Figure 4D; Supplementary Figure 3D). On the contrary, the CD proximal inflamed and noninflamed panels had no mucin RNA isoforms in common (Figure 4F; Supplementary Figure 3F). Of all 13 mucin RNA isoforms in the CD distal colon noninflamed panel, only the *MUC12* (ENST00000473098.5) RNA isoform was also present (and downregulated) in the CD distal colon inflamed panel (Figure 4H; Supplementary Figure 3H). Of the 23 RNA isoforms from the CD rectum noninflamed panel, only the *MUC4* (PB.1238.363) RNA isoform was present (and upregulated) in its inflamed counterpart (Figure 4J; Supplementary Figure 3J) Furthermore, the *MUC4* (PB.1238.363) and *MUC12* (ENST00000473098.5) RNA isoforms were respectively up- and downregulated in both their inflamed and noninflamed CD region-specific panels (Figure 4H, 4J; Supplementary Figure 3H, 3J).

3.4. Structural characterization of abundant mucin RNA isoforms in IBD

Several mucin RNA isoforms frequently occurred throughout the different panels highlighting their importance for further investigation. To identify structural differences between these isoforms, the exon-intron structure of all mucin RNA isoforms identified in the inflamed models is illustrated in Figure 6. More specifically, the *MUC1* RNA isoform ENST00000462317.5 appeared in the panels comparing inflamed biopsies of CD ileal, CD proximal, UC colonic, and UC rectal with (region-matched) noninflamed control biopsies (Figure 5). Interestingly, this 7 exon-long *MUC1* (ENST00000462317.5) RNA isoform lacks the first 2 exons compared to the canonical *MUC1* (ENST00000620103.4) isoform, which encodes for a part of the extracellular region and gains 1 exon near the 3′ end of the transcript (Figure 6). The canonical *MUC1* (ENST00000620103.4; Figure 6) RNA isoform also appeared in different panelsl; however, its

UC patients ($n_{\text{Train}} = 94$, $n_{\text{Test}} = 23$, $n_{\text{ext. val.}} = 27$) versus noninflamed biopsies of control patients ($n_{\text{Train}} = 94$, $n_{\text{Test}} = 44$, $n_{\text{ext. val.}} = 23$, the external validation consisted of all distal colon samples to increase the sample size), (E, F) inflamed proximal colonic biopsies of CD patients ($n_{\text{Train}} = 63$, $n_{\text{Test}} = 15$, $n_{\text{ext. val.}} = 2$) versus noninflamed proximal colonic biopsies of control patients ($n_{\text{Train}} = 63$, $n_{\text{Test}} = 21$, $n_{\text{ext. val.}} = 21$), (G, H) inflamed distal colonic biopsies of CD patients ($n_{\text{Train}} = 120$, $n_{\text{Test}} = 29$, $n_{\text{ext. val.}} = 9$) versus noninflamed distal colonic biopsies of control patients ($n_{\text{Train}} = 120$, $n_{\text{Test}} = 46$, $n_{\text{ext. val.}} = 23$), and (I, J) inflamed rectal biopsies of CD patients ($n_{\text{Train}} = 80$, $n_{\text{Test}} = 20$, $n_{\text{ext. val.}} = 9$) versus noninflamed rectal biopsies of control patients ($n_{\text{Train}} = 80$, $n_{\text{Test}} = 44$, $n_{\text{ext. val.}} = 23$; the external validation consisted of all distal colon samples to increase the sample size). Differential mucin isoform expression was determined using the Wilcoxon rank-sum test corrected for multiple testing using the Bonferroni method. Significant differences are indicated by *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. In the boxplots, the horizontal line within the box represents the median, the lower hinge of the box denotes the first quartile and the upper hinge corresponds to the third quartile. Whiskers extend to the largest data point within 1.5 times the interquartile range of the upper and lower hinge. Data points beyond these ranges are plotted individually. Abbreviations: C, control patients; colon-dist, distal colon; colon-prox, proximal colon; ext.val, external validation.



A	ENST00000397910.8
B	ENST00000447234.7
C	ENST00000620103.4
D	PB.1238.363
E	PB.2810.148
F	ENST00000473098.5
G	ENST00000462317.5
H	ENST00000525715.5
I	PB.2118.153
J	PB.2811.15
K	ENST00000436408.6
L	PB.2118.1621
M	PB.2810.2464
N	PB.2118.1322
O	PB.1239.830
P	PB.2118.1409
Q	PB.2816.52
R	ENST00000338684.9
S	ENST00000349607.8
T	ENST00000631087.1
U	PB.381.4

Figure 5. Venn diagrams representing an overview of the mucin isoforms obtained by the feature selection Random Forest throughout multiple models involving inflamed biopsies. (A) Venn diagram summarizing the mucin isoforms as predictors to classify inflamed colonic biopsies of CD or UC patients from noninflamed biopsies of control patients and inflamed ileal biopsies of CD patients from noninflamed biopsies of control patients. Isoforms present in the colon, rectum, distal, and/or proximal CD panel are combined as are the isoforms present in the colon, rectum, and/or distal UC panel. (B) Venn diagram summarizing the mucin isoforms as predictors to classify inflamed biopsies from the proximal colon, distal colon, and rectum from CD patients from their noninflamed control counterparts. (C) Venn diagram summarizing the mucin isoforms as predictors to classify inflamed biopsies from the distal colon and rectum from UC patients from their noninflamed control counterparts. The colors represent the mucin gene from which the isoform

presence was limited to the CD inflamed panels, with the exception of the CD proximal colon inflamed panel (Figure 5B), and to the CD noninflamed ileum panel (Supplementary Figure 4). In all models comparing inflamed IBD with noninflamed control colonic biopsies, the *MUC4* (PB.1238.363) RNA isoform was selected by the feature selection procedure suggesting its high discriminative value (Figure 5). In addition, *MUC4* (PB.1238.363) also appeared in the CD rectum noninflamed panel (Supplementary Figure 4). Compared to the 25 exon-long canonical *MUC4* (ENST00000463781.8) sequence, *MUC4* (PB.1238.363) comprises only 7 exons and misses a large part at its 3′ end which encodes the intracellular, transmembrane and part of the extracellular region. Since it also contained novel splice sites, it was categorized as a "novel_not_in_catalog" isoform (Figure 6; Supplementary Table 4). The canonical *MUC20* (ENST00000447234.7) and novel *MUC5AC* (PB.2811.15) RNA isoforms occurred with a high frequency in the inflamed colon of CD and UC panels, while also being present in some noninflamed panels for UC (*MUC20* [ENST00000447234.7]) or both CD and UC (*MUC5AC* [PB.2811.15]; Figure 5; Supplementary Figure 4). *MUC20* (ENST00000447234.7) is a relatively small isoform, only having 4 exons, which stands in steep contrast with many other isoforms like *MUC5AC* (PB.2811.15) having 17 exons. Identical to *MUC4* (PB.1238.363), the *MUC5AC* (PB.2811.15) RNA isoform was categorized as "novel_not_in_catalog" due to the presence of splice sites that were previously not identified for *MUC5AC*. In addition, this isoform lacks a major part at its 5′ end thereby missing the sequence that encodes for the signal peptide, several Von Willebrand factor type D domains, and the PTS domain, rich in proline, threonine, and serine.[34] Finally, like *MUC20* (ENST00000447234.7) and *MUC1* (ENST00000620103.4), the 84 exon-long *MUC16* (ENST00000397910.8) occurred in all inflamed colonic CD panels, except for the proximal colonic panels (Figure 5B), and was also present in the UC rectum inflamed panel (Figure 5C, 5D). Of note, a large number of unique mucin RNA isoforms were identified in the UC and CD noninflamed panels (Supplementary Figure 4), suggesting the Random Forest approach has difficulty in finding discriminative isoforms to classify the noninflamed IBD subgroups and the control patients from one another.

4. Discussion

Our study is the first to describe the complexity of the mucin RNA isoform landscape expressed in the terminal ileum and colon of IBD and control patients using a targeted long-read mucin isoform sequencing approach with custom-designed probes, encompassing 17 mucin genes present in the human reference genome in combination with short-read sequencing. We combined our identified intestinal mucin transcriptome with the human reference transcriptome to obtain the expression of the global mucin RNA isoform repertoire. Differences in the mucin RNA isoform landscapes between both transcriptomes can partially be explained by mucin RNA isoform expression itself, which is not limited to the intestines in the reference transcriptome but varies between tissues, sample types, and diseases.[35–39] The predicted coding potential of the intestinal-type mucin RNA isoforms was overall higher than those from the human reference transcriptome with many novel RNA isoforms discovered, especially for the *MUC2*, *MUC3A*, *MUC4*, *MUC12*, *MUC13,* and *MUC17* genes. The abundance of putatively coding isoforms in the intestinal mucin transcriptome can be explained by the oligo-dT primers used to enrich polyadenylated transcripts during the library preparation for long-read sequencing ensuring consistency with the polyadenylated enriched libraries of the short-read sequencing data.[40] Some transcripts from the intestinal mucin isoform transcriptome were also composed of fragments from two different genes, such as *MUC20_MUC20-OT1* and *MUC3A_MUC12*. Whether these fusion transcripts originate from structural rearrangements (at the DNA level), or during transcription or splicing events, requires further investigation. Nevertheless, their detection might be attributed to our targeted long-read sequencing approach which allowed sequencing of full-length mucin isoforms, even those with an expression level too low to be detected by the PacBio sequencing platform without mucin-specific enrichment.[41] The short-read sequencing data (GSE193677) from more than 1500 intestinal biopsies of IBD and control patients was then mapped to the combined mucin RNA isoform landscape to assess the expression levels of the different mucin RNA isoforms in the intestinal tract of the different study groups. The obtained heatmaps clearly highlight distinct mucin isoform signatures between ileal and colonic samples, tissue status (inflamed versus noninflamed), and patient cohorts, but also region-specific signatures were noted within the colon. These findings further add to the heterogeneity of IBD and suggest that the mucosal changes that occur upon inflammation throughout different anatomical locations of the colon are not uniform.

We subsequently assessed the predictive performance of our newly discovered mucin RNA isoform landscape expressed in the terminal ileum and colon for IBD patient stratification using Random Forest classification. The performance of each designed Random Forest model depended on the specific comparisons that were made between study groups, the anatomical location from where the biopsies originated, and the presence/absence of inflammation. Distinguishing inflamed colonic biopsies from IBD (irrespective of subtype), UC, or CD patients and inflamed ileal biopsies from CD patients from their noninflamed control counterparts unveiled mucin RNA isoform panels with high performance in both training and test data. The discriminative potential of the UC and CD colon inflamed panels in the training and test datasets was overall improved when dividing the colonic biopsies in different groups based on their location in the intestinal tract, i.e. proximal colon, distal colon, and rectum. However, the functional significance of specific mucin isoforms in relation to the gastrointestinal region warrants further investigation. Remarkably, the performance of the mucin RNA isoform panels when testing on external validation data, still remained high for the models distinguishing UC inflamed colonic biopsies from controls, and even slightly increased when the location of the intestinal tract was taken into account. Likewise,

originates, the letters refer to the specific mucin isoform and the size of the circle for each isoform indicates how many panels it occurs. (D) Graphical overview of the different mucin isoform panels obtained for each anatomical region in Crohn's disease and ulcerative colitis patients. Significant up- and downregulation when compared to mucin isoform expression in control samples is indicated with arrows in front of each isoform. Created with BioRender.com.

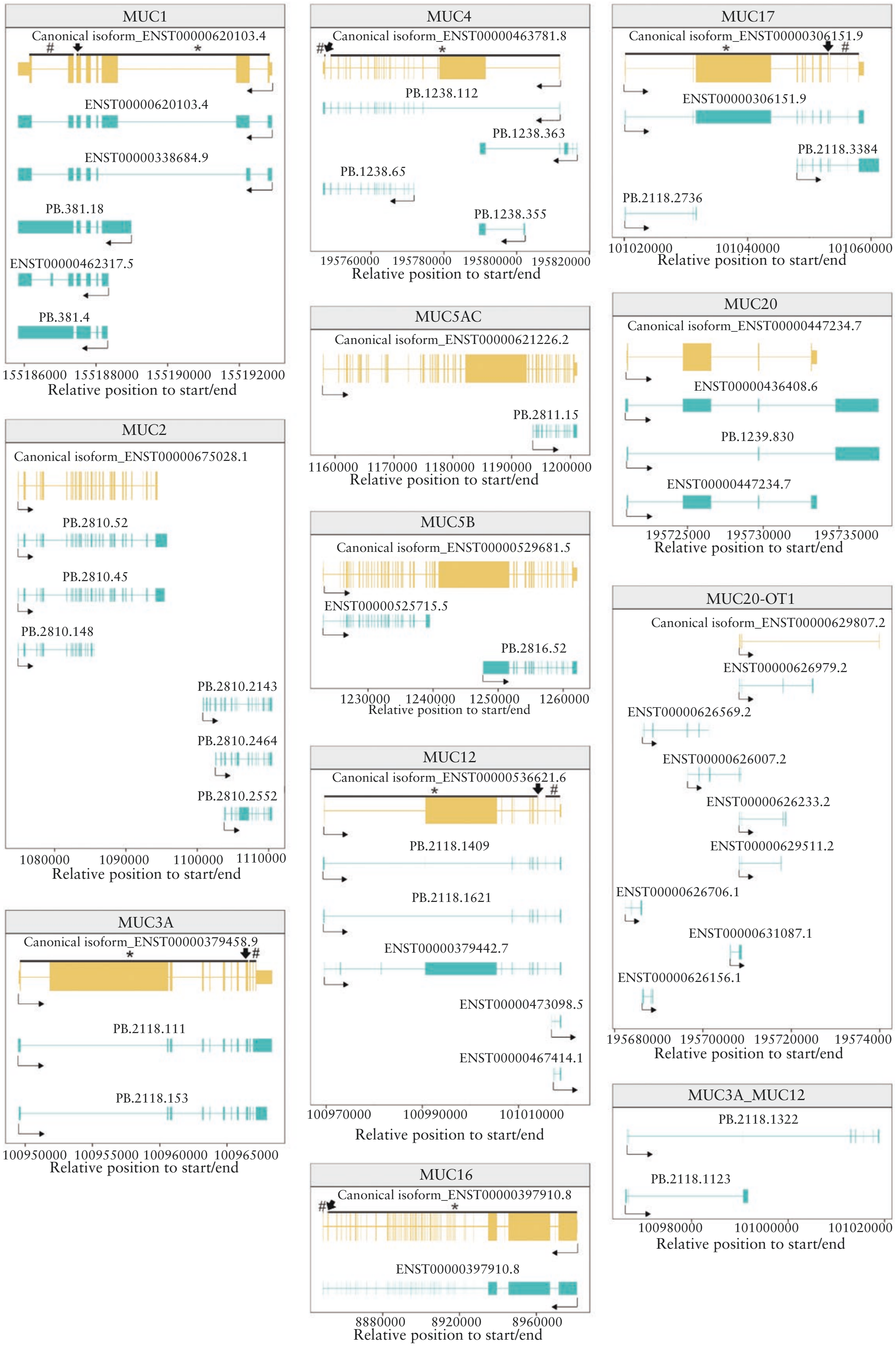


Figure 6. Schematic representation of the exon-intron structure of the isoforms encompassing the IBD, UC, CD colonic, and ileal inflamed panels. The yellow mucin isoforms are the ENSEMBLE (version 110: Jul 2023) canonical isoforms present as a reference for comparison. On each canonical isoform originating from a transmembrane mucin, the sequence coding for the cytoplasmic (#), transmembrane (↓), and extracellular (*) regions are highlighted. The arrow below each isoform indicates the direction of transcription.

the models validated on external data to discriminate inflamed distal and rectal biopsies of CD patients from controls exhibited a minor decrease in performance compared to the test and training datasets. In contrast, a more substantial decrease in AUC was observed for models trained on CD inflamed biopsies from the proximal colon and ileum. This reduction in performance might be explained by the large heterogeneity in IBD that is not fully represented in the limited amount of proximal biopsy samples present in the external validation dataset.

The model that distinguishes between inflamed colonic biopsies from CD and UC patients performed poorly which was also reflected by the large amount of selected mucin RNA isoforms. Furthermore, while the panels from all other models almost exclusively consisted of differentially expressed isoforms, the majority of mucin RNA isoforms in this panel had similar expression levels. In addition, the models trained to distinguish noninflamed biopsies of IBD patients (and subtypes) from noninflamed control biopsies also showed a decrease in performance for the training compared to the models trained to distinguish inflamed from noninflamed biopsies. The above observations, combined with the lower amount of differentially expressed mucin isoforms, suggest a closer resemblance in mucin RNA isoform expression between the inflamed colon of UC and CD patients and the noninflamed intestinal tract of IBD and control patients.

Interestingly, multiple mucin RNA isoforms also frequently occurred among the different inflamed panels suggesting that the Random Forest approach is consistent and not just capturing noise (Figure 5A–C). Furthermore, some mucin isoforms occurred in both the inflamed panel and in its noninflamed counterpart, indicating that mucin expression is still altered in IBD patients showing no macroscopic inflammation in their intestinal tract, as has also been reported recently.[16] Besides, multiple studies also identified mucin gene polymorphisms that cause structural changes in the mucin protein and have been associated with the susceptibility to IBD.[19,42,43] Therefore, further research is needed to determine whether the mucin RNA isoforms present in both the inflamed and noninflamed panels result from a genetic mutation that predisposes individuals to IBD.

Irrespective of the panel in which the isoforms occurred, *MUC4* (PB.1238.363), *MUC5AC* (PB.2811.15), *MUC16* (ENST00000397910.8), and *MUC1* (ENST00000462317.5, ENST00000620103.4) were upregulated upon inflammation, whereas the *MUC20* (ENST00000447234.7) and *MUC12* (ENST00000473098.5) RNA isoforms were downregulated. While the overexpression of the *MUC1* and *MUC4* genes in IBD is generally accepted,[11,15,44,45] literature describing the expression of the other mucin genes from which these isoforms originate is limited and sometimes ambiguous.[19,46,47] Some of the abundant mucin RNA isoforms lack several exons compared to their canonical counterpart resulting in the absence of functional domains. These included the *MUC4* (PB.1238.363) RNA isoform which lacks its cytoplasmic tail and transmembrane domain, only retaining a part of its extracellular domain. As such, *MUC4* (PB.1238.363) resembles rather a secreted than a transmembrane mucin. Das et al.[48] suggested MUC4 to be a driver of intestinal inflammation after observing an increased resistance of *MUC4*$^{-/-}$ mice to DSS-induced colitis. However, functional repercussions of *MUC4* isoforms without cytoplasmic and transmembrane domains should be further investigated, as studies suggested isoform-dependent effects in pancreatic cancer, such as complex formation with other isoforms influencing signal transduction.[20] In contrast to MUC4, MUC5AC expression was found to have a protective effect in DSS-induced colitis mice models.[46] The relatively small *MUC5AC* (PB.2811.15), comprising only 2768 bp compared to the 17 448 bp long canonical *MUC5AC* (ENST00000621226.2), is missing the sequences necessary to encode several Von Willebrand factor type D domains and the PTS domain.[12] The absence of this PTS domain reduces the extent to which the isoform can be glycosylated, thereby preventing it from executing the protective function of the canonical *MUC5AC* (ENST00000621226.2).[34] However, since *MUC5AC* (PB.2811.15) also lost its N-terminal signal sequence present in the canonical *MUC5AC*, the question remains whether this *MUC5AC* isoform gets secreted or stays intracellularly, impacting signal transduction.[34,49] In the case of *MUC16*, it was the canonical isoform ENST00000397910.8 that was found to recur in different panels. This transmembrane mucin has been associated with several cancers, in which its expression increases after stimulation with TNF-α and IFN-γ, due to an NF-κβ binding site in the proximal promoter region of *MUC16*.[50] As both TNF-α and IFN-γ are key cytokines in the pathogenesis of IBD, it is likely that they are also responsible for the upregulation of this *MUC16* RNA isoform observed in the inflamed colon of our IBD patient cohort. The only mucin RNA isoform occurring in both the inflamed ileal panel and several colonic inflamed panels for CD and UC is the *MUC1* (ENST00000462317.5) RNA isoform. Translated *MUC1* isoforms missing parts of their extracellular domain (like *MUC1* [ENST00000462317.5]), have previously been described to form protein complexes with other MUC1 isoforms resulting in an altered cellular morphology.[51] In addition, Cascio et al.[52] showed the formation of a MUC1-P65 complex that caused an upregulation of proinflammatory cytokines IL-6 and TNF-α by enhancing the NF-κβ pathway. While all tested MUC1 isoforms succeeded in inducing the NF-κβ pathway, MUC1 isoforms lacking the variable tandem repeat region (VNTR) were found to be less proinflammatory. Interestingly, the VNTR region (encoded by exon 2 of the canonical *MUC1* isoform (ENST00000620103.4)[53]) is not present in *MUC1* (ENST00000462317.5). Nonetheless, its role in the IBD pathogenesis requires further investigation.

The use of public datasets inferred some limitations in our study. The external validation included a relatively low number of samples for certain regions, necessitating the inclusion of all distal colon samples to increase the dataset size for validating the models trained on rectal samples. In addition, external validation data containing follow-up samples of IBD patients in therapy-specific trials will help to demonstrate the clinical utility of the mucin isoform panels. Unfortunately, at this moment, no such RNA sequencing data are available that have location-specific biopsy information, an acceptable cohort size, and a sufficiently large sequencing read length. Nevertheless, the strength of our study lies in the identification of distinct mucin RNA isoform panels that are associated with IBD, its subtypes (CD, UC), inflammation, and location in the intestinal tract (Figure 5D), and accurately stratify IBD patient subpopulations. Also, the diversity of mucin RNA isoform expression between the studied population cohorts highlights the large heterogeneity within IBD and the importance of the anatomical location in the intestinal tract that should be taken into account in further studies.

5. Conclusion

We unveiled the intestinal mucin isoform landscape that reflects the heterogeneity of the IBD patient population. More specifically, known and novel mucin RNA isoforms were found to be significantly up- or downregulated depending on the inflammatory status and anatomical location of the biopsy in the intestinal tract. Furthermore, the identified mucin RNA isoform inflamed panels that associate with the IBD subtype and anatomical region of the intestinal tract, show great potential to indicate epithelial barrier function at the molecular level in addition to endoscopic and histologic measures of inflammation, and thus mucosal healing in IBD patients. Therefore, adequate independent external validation in biopsies from large prospective IBD cohorts under therapy is recommended to further clinically validate the potential of these mucin RNA isoform panels to monitor barrier healing in the therapeutic management of IBD.

Funding

This work was supported by the Antwerp University Valorisation Funds (IOF-SBO number 42601 [A.S., B.Y.D.W.]; IOF-POC number 48068 [A.S., B.Y.D.W.]), the Research Foundation Flanders (FWO) grants (G031222N [A.S.], W001620N [A.S.]), FWO-Hercules Medium-Scale Research Infrastructure (#006518N, PacBio Sequel platform, S.M.-K., A.S., B.Y.D.W.), and the Flemish Government-initiated Methusalem Excellence Program (VAX-IDEA, S.M.-K.).

Conflict of Interest

W.A., T.B., B.Y.D.W., and A.S. are inventors of a patent related to mucin isoforms in diseases characterized by barrier dysfunction, including IBD, irritable bowel syndrome, gastrointestinal infections, and cancer (WO/2021/013479; EP23214719.9). All other authors declare no conflict of interest.

Acknowledgments

We express gratitude to the patients for their willingness to take part in this study. We also extend our appreciation to F. van Aert, S. Schelkens, and E. Vanhaesebrouck for their outstanding support in patient recruitment and sample collection.

Author contributions

W.A., A.S., T.B., and B.Y.D.W. conceptualized and designed the study. W.A. performed the experiments, statistical analyses, and bioinformatics computation. B.O., S.A., and S.V.L. provided expertise during data analysis. W.A. and A.S. interpreted the data and took the lead in writing the manuscript. E.M., M.S., and A.J.-A. collected patient samples and clinical data. S.M.K. provided access to the PacBio sequencing platform. W.A., A.S., T.B, B.O., J.G.D.M., S.M.K., S.A, S.V.L., E.M, M.S, A.J.-A., and B.Y.D.W. critically revised the manuscript.

Data Availability

R scripts and supporting data are available by request to the corresponding author. Public datasets were retrieved from the Gene Expression Omnibus database of the National Center for Biotechnology Information. Novel Illumina sequencing data were uploaded to the NIH database and can be accessed via the BioProject PRJNA1086386.

Supplementary Data

Supplementary data are available online at *ECCO-JCC* online.

References

1. Chang JT. Pathophysiology of inflammatory bowel diseases. *N Engl J Med* 2020;**383**:2652–64.
2. Alatab S, Sepanlou SG, Ikuta K, *et al*. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol* 2020;**5**:17–30.
3. Kim JW, Kim SY. The era of Janus kinase inhibitors for inflammatory bowel disease treatment. *Int J Mol Sci* 2021;**22**:11322. doi:10.3390/ijms222111322
4. Ben Ghezala I, Charkaoui M, Michiels C, Bardou M, Luu M. Small molecule drugs in inflammatory bowel diseases. *Pharmaceuticals* 2021;**14**:637.
5. Selkirk JV, Yan YG, Ching N, Paget K, Hargreaves R. In vitro assessment of the binding and functional responses of ozanimod and its plasma metabolites across human sphingosine 1-phosphate receptors. *Eur J Pharmacol* 2023;**941**:175442.
6. Maaser C, Sturm A, Vavricka SR, *et al*.; European Crohn's and Colitis Organisation [ECCO] and the European Society of Gastrointestinal and Abdominal Radiology [ESGAR]. ECCO-ESGAR Guideline for Diagnostic Assessment in IBD Part 1: initial diagnosis, monitoring of known IBD, detection of complications. *J Crohns Colitis* 2019;**13**:144–64.
7. Turner D, Ricciuto A, Lewis A, *et al*.; International Organization for the Study of IBD. STRIDE-II: an update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) Initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. *Gastroenterology* 2021;**160**:1570–83.
8. Atreya R, Neurath MF. Current and future targets for mucosal healing in inflammatory bowel disease. *Visc Med* 2017;**33**:82–8.
9. Falvey JD, Hoskin T, Meijer B, *et al*. Disease activity assessment in IBD: clinical indices and biomarkers fail to predict endoscopic remission. *Inflamm Bowel Dis* 2015;**21**:824–31.
10. Keita Å V, Lindqvist CM, Öst A, Magana CDL, Schoultz I, Halfvarson J. Gut barrier dysfunction—a primary defect in twins with Crohn's disease predominantly caused by genetic predisposition. *J Crohns Colitis* 2018;**12**:1200–9.
11. Breugelmans T, Van Spaendonk H, De Man JG, *et al*. In-depth study of transmembrane mucins in association with intestinal barrier dysfunction during the course of T cell transfer and DSS-induced colitis. *J Crohns Colitis* 2020;**14**:974–94.
12. Breugelmans T, Oosterlinck B, Arras W, *et al*. The role of mucins in gastrointestinal barrier function during health and disease. *Lancet Gastroenterol Hepatol* 2022;**7**:455–71.
13. Breugelmans T, Arras W, Oosterlinck B, *et al*. IL-22-activated MUC13 impacts on colonic barrier function through JAK1/STAT3, SNAI1/ZEB1 and ROCK2/MAPK signaling. *Cells* 2023;**12**:1224.
14. Senapati S, Ho SB, Sharma P, *et al*. Expression of intestinal MUC17 membrane-bound mucin in inflammatory and neoplastic diseases of the colon. *J Clin Pathol* 2010;**63**:702–7.
15. Vancamelbeke M, Vanuytsel T, Farré R, *et al*. Genetic and transcriptomic bases of intestinal epithelial barrier dysfunction in inflammatory bowel disease. *Inflamm Bowel Dis* 2017;**23**:1718–29.
16. Breugelmans T, Arras W, Boen LE, *et al*. Aberrant mucin expression profiles associate with pediatric inflammatory bowel disease presentation and activity. *Inflamm Bowel Dis* 2022;**29**:589–601.

17. Sheng YH, Lourie R, Linden SK, *et al*. The MUC13 cell-surface mucin protects against intestinal inflammation by inhibiting epithelial cell apoptosis. *Gut* 2011;**60**:1661–70.
18. Kyo K, Muto T, Nagawa H, Lathrop GM, Nakamura Y. Associations of distinct variants of the intestinal mucin gene MUC3A with ulcerative colitis and Crohn's disease. *J Hum Genet* 2001;**46**:5–20.
19. Moehle C, Ackermann N, Langmann T, *et al*. Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. *J Mol Med* 2006;**84**:1055–66.
20. Kumar S, Cruz E, Joshi S, *et al*. Genetic variants of mucins: unexplored conundrum. *Carcinogenesis* 2017;**38**:671–9.
21. Massimino L, Lamparelli LA, Houshyar Y, *et al*. The inflammatory bowel disease transcriptome and metatranscriptome meta-analysis (IBD TaMMA) framework. *Nat Comput Sci* 2021;**1**:511–5.
22. Peters LA, Perrigoue J, Mortha A, *et al*. A functional genomics predictive network model identifies regulators of inflammatory bowel disease. *Nat Genet* 2017;**49**:1437–49.
23. Argmann C, Hou R, Ungaro RC, *et al*. Biopsy and blood-based molecular biomarker of inflammation in IBD. *Gut* 2023;**72**:1271–87.
24. Tardaguila M, de la Fuente L, Marti C, *et al*. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* 2018;**28**:396–411.
25. Pardo-Palacios FJ, Arzalluz-Luque A, Kondratova L, *et al*. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. *Nat Methods* 2024;**21**:793–7.
26. Dobin A, Davis CA, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.
27. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;**34**:525–7.
28. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Res* 2020;**9**:ISCB Comm J–304.
29. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods* 2017;**14**:687–90.
30. Brieuc MSO, Waters CD, Drinan DP, Naish KA. A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Mol Ecol Resour* 2018;**18**:755–66.
31. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
32. Gu Z. Complex heatmap visualization. *IMeta* 2022;**1**:e43.
33. Jun Z. transPlotR: an elegant package to visualize gene structures. 2022. https://github.com/junjunlab/transPlotR.
34. van de Bovenkamp JHB, Hau CM, Strous GJAM., Büller HA, Dekker J, Einerhand AWC. Molecular cloning of human gastric mucin MUC5AC reveals conserved cysteine-rich D-domains and a putative leucine zipper motif. *Biochem Biophys Res Commun* 1998;**245**:853–9.
35. Thornton DJ, Rousseau K, McGuckin MA. Structure and function of the polymeric mucins in airways mucus. *Annu Rev Physiol* 2008;**70**:459–86.
36. Lakshmanan I, Ponnusamy MP, Macha MA, *et al*. Mucins in lung cancer: diagnostic, prognostic, and therapeutic implications. *J Thorac Oncol* 2015;**10**:19–27.
37. Hattrup CL, Gendler SJ. Structure and function of the cell surface (tethered) mucins. *Annu Rev Physiol* 2008;**70**:431–57.
38. Nath S, Mukherjee P. MUC1: a multifaceted oncoprotein with a key role in cancer progression. *Trends Mol Med* 2014;**20**:332–42.
39. Manne A, Kasi A, Esnakula AK, Paluri RK. Predictive value of MUC5AC signature in pancreatic ductal adenocarcinoma: a hypothesis based on preclinical evidence. *Int J Mol Sci* 2023;**24**:8087.
40. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep* 2018;**8**:4781.
41. Dorney R, Dhungel BP, Rasko JEJ, Hebbard L, Schmitz U. Recent advances in cancer fusion transcript detection. *Brief Bioinform* 2023;**24**:bbac519.
42. Kyo K, Parkes M, Takei Y, *et al*. Association of ulcerative colitis with rare VNTR alleles of the human intestinal mucin gene, MUC3. *Hum Mol Genet* 1999;**8**:307–11.
43. Visschedijk MC, Alberts R, Mucha S, *et al*.; Initiative on Crohn and Colitis. Pooled resequencing of 122 ulcerative colitis genes in a large Dutch cohort suggests population-specific associations of rare variants in MUC2. *PLoS One* 2016;**11**:e0159609.
44. Gersemann M, Becker S, Kübler I, *et al*. Differences in goblet cell differentiation between Crohn's disease and ulcerative colitis. *Differentiation* 2009;**77**:84–94.
45. Furr AE, Ranganathan S, Finn OJ. Aberrant expression of MUC1 mucin in pediatric inflammatory bowel disease. *Pediatr Dev Pathol* 2010;**13**:24–31.
46. Olli KE, Rapp C, O'Connell L, *et al*. Muc5ac expression protects the colonic barrier in experimental colitis. *Inflamm Bowel Dis* 2020;**26**:1353–67.
47. Yamamoto-Furusho JK, Ascaño-Gutiérrez I, Furuzawa-Carballeda J, Fonseca-Camarillo G. Differential expression of MUC12, MUC16, and MUC20 in patients with active and remission ulcerative colitis. *Mediators Inflamm* 2015;**2015**:659018.
48. Das S, Rachagani S, Sheinin Y, *et al*. Mice deficient in Muc4 are resistant to experimental colitis and colitis-associated colorectal cancer. *Oncogene* 2016;**35**:2645–54.
49. Liaci AM, Förster F. Take me home, protein roads: structural insights into signal peptide interactions during ER translocation. *Int J Mol Sci* 2021;**22**:11871.
50. Morgado M, Sutton MN, Simmons M, *et al*. Tumor necrosis factor-α and interferon-γ stimulate MUC16 (CA125) expression in breast, endometrial and ovarian cancers through NFκB. *Oncotarget* 2016;**7**:14871–84.
51. Baruch A, Hartmann M, Yoeli M, *et al*. The breast cancer-associated MUC1 gene generates both a receptor and its cognate binding protein1. *Cancer Res* 1999;**59**:1552–61.
52. Cascio S, Zhang L, Finn OJ. MUC1 protein expression in tumor cells regulates transcription of proinflammatory cytokines by forming a complex with nuclear factor-κB p65 and binding to cytokine promoters: importance of extracellular domain. *J Biol Chem* 2011;**286**:42248–56.
53. Zhang L, Vlad A, Milcarek C, Finn OJ. Human mucin MUC1 RNA undergoes different types of alternative splicing resulting in multiple isoforms. *Cancer Immunol Immunother* 2013;**62**:423–35.