Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies

Boyd, Mette; Thodberg, Malte; Vitezic, Morana; Bornholdt, Jette; Vitting-Seerup, Kristoffer; Chen, Yun; Coskun, Mehmet; Li, Yuan; Lo, Bobby Zhao Sheng; Klausen, Pia

*Published in:* Nature Communications

*Link to article, DOI:* 10.1038/s41467-018-03766-z

*Publication date:* 2018

*Document Version*  
Publisher's PDF, also known as Version of record

*Link back to DTU Orbit*

*Citation (APA):*  
Characterization of the enhancer and promoter landscape of inflammatory bowel disease from human colon biopsies

Mette Boyd1,2, Malte Thodberg1,2, Morana Vitezic1,2, Jette Bornholdt1,2, Kristoffer Vitting-Seerup1,2, Yun Chen1,2, Mehmet Coskun1,2, Yuan Li3, Bobby Zhao Sheng Lo1,2,3, Pia Klausen4, Pawel Jan Schweiger2, Anders Gorm Pedersen5, Nicolas Rapin2,6,7, Kerstin Skovgaard8, Katja Dahlgaard9, Robin Andersson1, Thilde Bagger Terkelsen1,2, Berit Lilje1,2, Jesper Thorvald Troelsen9, Andreas Munk Petersen10,11, Kim Bak Jensen2,7, Ismail Gögenur12, Peter Thielsen3, Jakob Benedict Seidelin3, Ole Haagen Nielsen3, Jacob Tveiten Bjerrum3 & Albin Sandelin1,2

Inflammatory bowel disease (IBD) is a chronic intestinal disorder, with two main types: Crohn’s disease (CD) and ulcerative colitis (UC), whose molecular pathology is not well understood. The majority of IBD-associated SNPs are located in non-coding regions and are hard to characterize since regulatory regions in IBD are not known. Here we profile transcription start sites (TSSs) and enhancers in the descending colon of 94 IBD patients and controls. IBD-upregulated promoters and enhancers are highly enriched for IBD-associated SNPs and are bound by the same transcription factors. IBD-specific TSSs are associated to genes with roles in both inflammatory cascades and gut epithelia while TSSs distinguishing UC and CD are associated to gut epithelia functions. We find that as few as 35 TSSs can distinguish active CD, UC, and controls with 85% accuracy in an independent cohort. Our data constitute a foundation for understanding the molecular pathology, gene regulation, and genetics of IBD.

1Department of Biology, University of Copenhagen, 2200 Copenhagen N, Denmark. 2Biotech Research and Innovation Centre, University of Copenhagen, 2200 Copenhagen N, Denmark. 3Department of Gastroenterology, Medical Section, Herlev Hospital, University of Copenhagen, 2730 Herlev, Denmark. 4Department of Gastroenterology, Surgical Section, Herlev Hospital, 2730 Herlev, Denmark. 5DTU Bioinformatics, Technical University of Denmark, 2800 Lyngby, Denmark. 6The Finsen Laboratory, Rigshospitalet, University of Copenhagen, 2200 Copenhagen N, Denmark. 7Novo Nordisk Foundation Center for Stem Cell Biology, DanStem, University of Copenhagen, 2200 Copenhagen N, Denmark. 8Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800 Lyngby, Denmark. 9Department of Science and Environment (INM), Roskilde University, 4000 Roskilde, Denmark. 10Hvidovre Hospital, Gastrounit Medical Division, University of Copenhagen, 2650 Hvidovre, Denmark. 11Hvidovre Hospital, Department of Clinical Microbiology, University of Copenhagen, 2650 Hvidovre, Denmark. 12Centre for Surgical Science, Department of Surgery, Zealand University Hospital, 4600 Koege, Denmark. These authors contributed equally: Mette Boyd, Malte Thodberg, Morana Vitezic, and Jette Bornholdt. Correspondence and requests for materials should be addressed to J.T.B. (email: jacob.wium.bjerrum@regionh.dk) or to A.S. (email: albin@binf.ku.dk).
Inflammatory bowel disease (IBD) is an umbrella term for a range of chronic idiopathic disorders, of which Crohn’s disease (CD) and ulcerative colitis (UC) constitute the two major entities, both with an increasing incidence and prevalence worldwide. With an estimated 2.5–3 × 10^8 patients in Europe alone, UC is characterized by mucosal inflammation of the colon, whereas CD may affect all layers of the intestine throughout the gastrointestinal tract. The distinction between CD and UC is critical for correct medication and especially surgery, yet the diagnosis is challenging. In a previous meta-analysis of 22,038 IBD patients, it was impossible to distinguish UC or CD in 13% of cases. This is a major problem for patients who have failed medical therapy and are facing colectomy. Genome-wide association studies (GWAS) have identified >200 loci containing IBD-associated variants, but IBD-associated single-nucleotide polymorphisms (SNPs) can only explain 15–20% of the disease phenotype. Moreover, ~70% of IBD-associated SNPs are non-coding and might affect gene regulation. However, a comprehensive map of active promoters and enhancers is lacking for IBD patients.

Genome-wide 5′-RNA sequencing of capped RNAs (Cap Analysis of Gene Expression, CAGE) can detect transcription start sites (TSSs) and thereby promoter regions. Distal enhancer regions can also be detected by CAGE, because active enhancers transcribe enhancer RNAs (eRNAs). eRNA expression is a powerful proxy for cell-specific enhancer activity and CAGE-identified enhancers are two to three times more likely to validate in vitro than non-transcribed enhancers detected by chromatin-based methods. An advantage of CAGE is that it can be easily applied on small biological samples, such as colonic biopsies that are routinely taken when diagnosing IBD.

Here we present CAGE analysis on biopsies from the descending colon from 94 IBD patients and controls. These data enabled annotation of IBD-regulated enhancers and TSSs, and characterization of IBD-associated SNPs in such regions. Furthermore, we define a small subset of TSSs that allow for accurate classification between UC, CD, and control subjects.

Results

The TSS expression landscape of IBD. We recruited 94 subjects undergoing lower endoscopy as part of routine visits, diagnosed as UC, CD, or control based on ref. 13. UC patients were graded by the Mayo score: a score ≤2 with endoscopic sub-score of 0 (no macroscopic inflammation) as quiescent UC and >2 as active UC. CD patients were graded by the Harvey–Bradshaw score: a score <5 as quiescent CD and ≥5 as active CD. For controls, an endoscopy was performed due to gastrointestinal symptoms but all clinical investigations returned normal. We extracted RNA from pinch biopsies from the descending colon of 94 subjects (cohort 1): 25 active UC (UCa), 20 active CD (CDa), 17 UC patients in remission (UCi), 3 CD patients in remission (CDi), and 29 control subjects (Ctrl) (Fig. 1a, Supplementary Table 1 and Supplementary Data 1). For UCa and CDa, biopsies taken from visually inflamed segments were confirmed by histology (Supplementary Figure 1). The choice of focusing on macroscopically inflamed rather than not visibly inflamed tissue was based on practical and statistical considerations: (i) in clinical diagnosis of IBD, histological examination is made on biopsies obtained from grossly inflamed intestine, together with endoscopic findings, medical history, laboratory, and imaging procedures. (ii) We wanted to characterize the most affected colonic regions of the disease to maximize comparability between biopsies, since non-macroscopically inflamed tissue might still show inflammation at the molecular level.

For each subject, we prepared a CAGE library (Supplementary Data 2). We defined 48,593 expressed TSSs based on CAGE tag clusters (Supplementary Data 3). Most CAGE-defined TSSs were within gene loci: 22.6% (11,013) corresponded to annotated main TSSs from GENCODE models, 12.8% (6,229) to annotated alternative TSSs and 40.7% (19,787) TSSs were putative novel alternative TSSs within known gene loci, although the majority of these were lowly expressed (Fig. 1b). These may reflect spurious transcription initiation, TSSs expressed in rare cells or noise, but were not classified as bidirectional transcribed enhancers in our analysis (see below). The ST6GAL1 gene, exemplifies the common occurrence of known and novel alternative TSSs in the same gene. The annotated main ST6GAL1 TSS was detected but did not change between groups, while one novel alternative TSS was upregulated in IBD, and another was downregulated (Fig. 1c).

CAGE TSSs principal component analysis (PCA) showed a separation between active IBD (UCa, CDa) and other groups (UCi, CDi, Ctrl), and a weaker separation between CDa and UCa (Fig. 2a). Surprisingly, UCa and CDi were not readily distinguishable from Ctrl (Supplementary Note 1). Hence, we excluded UCa and CDi from the systematic analysis below.

In order to identify differentially expressed TSSs, we used the generalized linear model framework in edgeR. As observed above, CDa and UCa samples were characterized by a shared response compared to Ctrl (PC1 in Fig. 2a), and secondly by differences between CDa and UCa (PC2 in Fig. 2a). To capture this, we defined sets of significantly up- or downregulated TSSs identified in both CDa and UCa vs. Ctrl (IBDup and IBDdown). To identify TSSs distinguishing CDa and UCa, we identified TSSs significantly upregulated in CDa vs. UCa, defined as CDspec, and TSSs significantly upregulated in UCa vs. CDa, defined as Uspec. The differential expression analysis recapitulated the PCA results: the number of TSSs within IBDspec (4736) and IBDdown (2536) was much larger than in CDspec (337) and Uspec (71) (Fig. 2b, Supplementary Data 4). The inclusion of additional patient data in the analysis, i.e. gender and previous medication, did not affect the number of differentially expressed TSSs substantially, indicating these effects are small compared to the CDa/UCa/Ctrl diagnosis (Supplementary Figure 2a). We also identified genes where one or more TSSs responded differently compared to the other TSSs inside the same gene (as exemplified in Fig. 1c): 2068 genes showed this pattern in IBDup/IBDdown and 82 in CDspec/UCspec (Supplementary Data 5).

Because most functional annotations are on gene rather than TSS level, we defined differentially expressed genes by summing the contribution of each TSS within each GENCODE gene model and repeating the differential expression analysis above (Supplementary Data 6, Supplementary Fig. 2b). Gene Ontology (GO) analysis (Fig. 2c, Supplementary Data 7) of IBDup genes showed a strong overrepresentation of GO terms related to inflammatory response and cytokines. IBDup genes were also enriched for colon-specific processes previously related to IBD pathogenesis, for example remodeling of the extracellular matrix (FDR = 2.66e−09) and antibacterial peptide secretion (FDR = 1.02e−14). Because small networks of biologically linked genes are challenging to identify through GO analysis, we supplemented this analysis with STRING network analysis. This identified smaller sets of upregulated genes with known functions in gut epithelia barrier integrity, including gap junctions (connexins AQP5, GJA4,5) and tight junctions (claudins CLDN1, 2, 10, 14, and 18) (Supplementary Fig. 2c). Interestingly, the full set of CLDN genes could distinguish IBD from Ctrl subjects, because a subset was part of the IBDdown set (Supplementary Fig. 2d–e). More generally, the IBDdown set was enriched for terms related to xenobiotic response and drug/steroid processing (Fig. 2c): STRING analysis showed that proteins corresponding to IBDdown genes were involved in cell cycle and growth, solute and membrane transport and maintenance of fluid balance in the...
Fig. 1 Defining the TSS landscape of IBD. 

**a** Overview of data set. Pinch biopsies from the descending colon were taken from 94 human subjects, classified into active ulcerative colitis (UCa), active Crohn’s disease (CDa), UC and CD patients in remission (UCi, CDi) and controls (Ctrl: subjects screened for IBD where all subsequent investigations turned out normal). For each biopsy, a CAGE library was produced, resulting in the detection of TSSs and enhancer regions. Schematics show the typical inflammatory patterns in the intestinal system, the approximate location of biopsy sampling and number of subjects in each group. 

**b** Detection and annotation of gene TSSs. Top panel shows an example gene with CAGE-defined TSSs, which are annotated as main, alternative or novel TSSs defined by their overlap with GENCODE gene annotation as indicated in callouts. CAGE-defined TSSs not falling into any of the categories were defined as novel intergenic TSSs. Left bottom panel shows the number of detected TSSs in each category (colors correspond to callouts in top panel), split by CAGE expression strength measured as tags per million (TPM). Right bottom panel shows the expression distribution of each category of TSSs as boxplots. 

**c** Genome-browser example of the detection of annotated and novel TSSs in the ST6GAL1 gene. From top to bottom, the browser plot shows the genomic location investigated, RefSeq gene annotation (exons are denoted as boxes, green indicate forward strand transcription). Below, CAGE TPM expression on the forward strand is shown as average across subjects (green bars) and for individuals (pink heat map, each row is one subject, columns are widened 5× for readability), split by subject group. Annotated and novel TSSs, annotated as in b are highlighted. Note that the first novel alternative TSS is upregulated in CDa and UCa vs. remaining groups, while the last novel alternative TSS has the opposite pattern (block arrows indicate TSSs and their overall strength in each subject group for these two TSSs). Conversely, the annotated TSSs are detected but not substantially changing between groups.
intestine (Supplementary Fig. 2f), with a strong sub-network of genes involved in steroid hormone biosynthesis.

The CDspec set was enriched for terms related to steroid, lipid, and lipoprotein metabolism (Fig. 2c). Because of the limited number of genes in the UCspec set, only a handful of GO terms were overrepresented, including several terms linked to gap junctions. These results largely mirrored those identified by RNA-seq in ileal pediatric CD20. This study identified a CD-vs-UC upregulation of apolipoprotein genes, which was also present in our set (Supplementary Fig. 2g). Thus, it is clear that while CDa and UCa share a strong inflammatory component, we show that genes that differentiate CDa and UCa are not primarily related to...
immune response but rather to cellular functions associated to gut and gut epithelia.

Differential expression analysis will favor genes showing low variance within groups. However, gene expression variance across patients with the same diagnosis is relevant for patient stratification and precision medicine. We found that five antibacterial peptide genes, defensins DEFA5 and DEFA6 (normally expressed in small intestine and Paneth cells) and c-lectins REG1A, REG1B, and REG3A (normally expressed in pancreas and small intestine) were highly upregulated in IBD and at the same time showed extreme variance across UC and CD subjects (Fig. 2d). The upregulation of these genes was highly correlated to cytokine induction (Supplementary Fig. 2h). The large variance might indicate patient-specific responses to inflammatory factors, disease chronicity or cell metabolism, and again highlights the role of epithelia-specific genes IBD response.

Colonic biopsies consist of multiple cell types, whose composition may change in IBD. We assessed this in two ways. First, we took advantage of the FANTOM5 TSS atlas, where individual TSSs were annotated as being preferentially expressed in one or more cell types sampled across the human body based on CAGE expression.21 With these data, we could estimate overall cell type enrichments for each differentially expressed TSS group across all FANTOM5 primary cell types (Fig. 2e, Supplementary Data 8). IBDup TSSs were enriched in immune-related cells, while IBDdown TSSs were enriched in epithelial cells. Cell-type enrichment patterns of CD_ppec and UC_ppec TSSs were different: CD_ppec TSSs were mostly enriched for immune-related cells (including basophils, monocytes, and eosinophils) and intestinal epithelial cells, while the UC_ppec TSSs were mostly enriched in epithelial cells and mesenchymal cells. Although T-helper cells have previously suggested to define the difference between CDa and UCa (reviewed in ref. 22), they were not enriched in our analysis, consistent with previous results23.

Second, because FANTOM5 data only included cells from healthy tissue, we wanted to also assess whether induction of an inflammatory response induced IBD signature genes in immune cells and/or epithelial cells. To address this, we used primary cultures of colonic epithelia organoids and monocytes, which were stimulated with TNF (also known as TNFα), one of the most prominent inflammatory molecules associated with flaring disease in IBD patients.25 Cells were stimulated with TNF for 4 and 24 h (Fig. 2f, Supplementary Fig. 3a, Supplementary Data 9) and the expression of 35 TSSs (established in the classification section below) was measured by qPCR. The same TSSs were measured by qPCR in gut biopsies from an independent cohort with 18 CDa, 37 UCa and 46 Ctrl subjects (cohort 2). We found that out of a total of 21 TSSs that were upregulated in UCa or CDa vs. control, 66% (14/21) and 81% (17/21) were also upregulated in TNF-stimulated organoids and monocytes, respectively, at any time point; 47% (10) of TSS were upregulated in both stimulated organoids and monocytes (Fig. 2f and Supplementary Fig. 3b, c). It is thus unlikely that the IBD expression change is solely attributed to immune cells; epithelial cells are likely part of shaping the transcriptional changes associated with UC and CD.

Thus, in agreement with the GO analysis, these cell type enrichment results indicated that the difference between Ctrl, CDa, and UCa does not only lie in immune cells and their associated general immune response pathways, but also in cells and genes associated with gut epithelia.

The enhancer landscape of inflammatory bowel disease. Gene and TSS-focused analysis as above can give insights into gene expression but not their regulation by distal enhancers. As established previously, bidirectional CAGE peaks accurately predict enhancer locations and activity13. Using the same method as in ref. 13, we predicted a permissive set of 37,231 enhancers, and a strict subset of 10,670 enhancers where ≥8 samples had detectable enhancer expression (the latter set was used in the rest of the analyses; Supplementary Data 10, 11 and Supplementary Fig. 4a). The strict enhancer set overlapped ENCODE DNase hypersensitive sites (DHSs)26 in 90% of cases (P < 2e−16, by sampling), and these enhancers were significantly more conserved across mammals than non-transcribed DHSs from gut tissue27 and randomly selected genomic regions (P < 5.2e−6, two-sided Mann–Whitney U test; Fig. 3a). These enhancers were also strongly enriched for ENCODE transcription factor (TF) ChIP-seq peaks from diverse cells (Fig. 3b). Similarly, they were highly enriched for H3K27ac and H3K4me1 ChIP-seq signals from rectal and colonic mucosa, and to a lesser degree other intestinal regions27 (Fig. 3c and Supplementary Fig. 4b). FAIRE-seq signals from colonic CD and control patients28,29 were also enriched (Supplementary Fig. 4c). However, the regions had less H3K27ac and H3K4me1 ChIP-seq enrichment in immune cells and non-gut tissues (Fig. 3c and Supplementary Fig. 4b). Thus, the enhancer candidates had the hallmarks of enhancer regions in general, but seemed most strongly used in gut tissue.

PCA of enhancer expression showed separation between inflamed samples (UCa and CDa) and Ctrl, but only minor separation between UCa and CDa (Fig. 3d). We used EBSeq to
define four differentially expressed enhancer groups, analogous to the TSS groups defined above: shared CDa and UCa up/downregulated vs. Ctrl (IBD_up and IBD_down), upregulated in CDa vs. UCa (CDspec) and vice versa (UCspec) (Supplementary Data 12 and Supplementary Fig. 4d). In agreement with the PCA, the IBD_up and IBD_down enhancer sets were much larger than CDspec and UCspec sets (Fig. 3e).

Next, we used HOMER 31 to identify shared predicted TF binding sites around IBD_up and IBD_down enhancers (±300 bp around enhancer midpoints; CDspec and UCspec sets were too small for meaningful analysis). The same analysis was used for promoter regions of corresponding TSS groups (defined as −500/+100 bp around TSSs) (Fig. 3f). IBD_up promoters and IBD_down enhancers were enriched for sites linked to TFs previously associated with inflammation, including ETS1, IRF1, STAT4, and RELA. IBD_down promoters and enhancers site enrichments were also highly similar: in particular, HNF4A and CDX2 sites were over-represented, consistent with their important roles in...
 differentiation of gut epithelial cells. Decreased CDX2 and HNF4A expression were previously observed in IBD, and indeed, site enrichments generally correlated with the expression of the associated TF (Fig. 3). We observed similar enrichment patterns when analyzing enhancer and promoter regions for overlapping ENCODE TF ChIP-seq peaks (Supplementary Fig. 5). Thus, IBD-induced enhancers and promoters share similar TF binding pattern enrichments, which are distinct from those of IBD-repressed enhancers and promoters.

To understand enhancer regulation, enhancers must be linked to target gene TSSs. As shown previously, enhancer-TSS pairs can be predicted based on CAGE co-expression across samples. Following the same approach, we identified 21,502 enhancer-TSS pairs within 500 kb with positive and significant Pearson correlation (correlation test, FDR < 0.05) (Supplementary Data 13). Of these pairs, 8507 (39.6%) were between a differentially expressed TSS and an enhancer, regardless of its differential expression status. Around half of IBDup and IBDdown TSSs (54.7% and 47.9%, respectively) were linked with at least one enhancer, compared to smaller fractions for CDspec, UCspec (40.9% and 18.3%) and remaining TSSs (Fig. 3g). Enhancer-linked TSSs in IBDup and IBDdown groups were commonly linked with more than one enhancer: in particular, IBDup TSSs linked to >3 enhancers were as frequently observed as TSSs linked to a single enhancer (Fig. 3g).

The NOD2 locus, a cluster of chemokine genes (CXCL1-3, CXCL5-6, and CXCL8), and two receptors for these chemokines (CXCR1 and CXCR2), represent examples of enhancers linked with neighboring TSSs of genes associated with IBD pathogenesis (Fig. 4a–c). The examples show putative distal regulators for key IBD genes, and highlighted a potential regulatory SNP that may affect the chemokine response pathway. Due to the common occurrence of enhancers linked to cytokine-related genes, we annotated the cytokine–cytokine receptor interaction pathway with linked enhancers and alternative TSSs (Supplementary Fig. 6).

Characterization of IBD-upregulated enhancer clusters. Previously, large regions having enhancer-like chromatin features—so-called “super”—or “stretch”—enhancers—were identified as drivers of central biological processes. By eye, clusters of CAGE-defined enhancers were evident, for instance in the region surrounding CEBPB, a key inflammatory response TF (Fig. 5a). We defined CAGE-based enhancer clusters by requiring >2 consecutive enhancer midpoints spaced ≤15 kb from each other, resulting in a set of 693 enhancer clusters (Supplementary Data 14), where 67% overlapped enhancer clusters defined by ChIP-seq in immune and gut cells. CAGE-defined enhancer regions within clusters often co-occurred with ENCODE TF ChIP-seq peaks (Fig. 5a).

The number of CAGE-defined enhancers within a cluster correlated with the average fraction of IBDup but not IBDdown enhancers in the same cluster (Fig. 5b). The same trend was observed for predicted enhancer targets: TSSs linked to at least one enhancer within a cluster had a higher IBD vs. Ctrl fold change if the enhancer cluster contained >6 enhancers compared to TSSs linked to clusters with fewer enhancers (P < 2.2e−16, two-sided t-test) (Fig. 5c and Supplementary Fig. 7a). Consistent with these observations, if an IBDup TSS was linked to an enhancer in a cluster with >6 enhancers, it was typically linked with all or most enhancers within the cluster (Supplementary Fig. 7b). Altogether, this suggests that enhancer clusters with many members and singleton enhancers have distinct functions in IBD.

To further explore the relation between the number of enhancers within a cluster and their regulatory function in IBD, we compared (i) IBDup TSS-linked enhancers from enhancer clusters with >6 members with (ii) IBDup TSS-linked singleton enhancers (not belonging to any cluster). Singleton enhancers had a greater overlap with ENCODE ChIP-seq peaks corresponding to TFs associated with inflammation (including RELA, SP1, STAT factors, and IRF1) (Fig. 5d), while enhancer clusters were overrepresented for ChIP-seq peaks for chromatin remodelers and insulators, including SMARCA4 and CTCF. As SMARCA4 has a function in activation of repressed genomic regions and CTCF is an insulator, these larger enhancer clusters might reflect large-scale changes in chromatin structure induced by the inflammatory response, while smaller enhancer clusters may be driven by the binding of one or a few TFs.

SNP overrepresentation in enhancers and promoters. As discussed above, a majority of IBD-associated SNPs are located in intergenic regions, and may impact gene regulation. The UC-associated SNP overlapping an IBDup enhancer located between cytokine receptors CXCR1 and CXCR2 (Fig. 4c) exemplifies an SNP with potential regulatory function. As we did not have access to the genotypes of our subjects, we investigated the overlap between publically available IBD-associated SNPs and enhancer and promoter regions using two complementary approaches: GWAS SNP enrichment and partitioned heritability of IBD.

First, we obtained all SNPs significantly associated with a disease/trait from the GWAS catalog. For each GWAS SNP, we merged the region covered by other SNPs in linkage disequilibrium (LD, R2 > 0.75 and within 500 kb) into a single block, using...
the 1000 genomes CEU reference population. If LD blocks overlapped, they were merged into an LD clump.

For every GWAS disease, we assessed the overlap between LD clumps with our identified promoter (−500/+100 bp around TSS peaks) and enhancer regions (±300 bp around enhancer midpoints) by using empirical Bayes to shrink fraction of overlapping LD clumps toward the average of all GWAS diseases. Only IBD, CD and UC LD clumps had a high degree of overlap with both the promoter and enhancer sets (Fig. 6a and Supplementary Fig. 8a). Next, we analyzed whether differentially expressed TSSs and enhancers were enriched for LD clump overlaps. IBD up promoters and enhancers were enriched for LD clumps associated with several immune-related diseases, but only IBD, CD, and UC LD clumps were enriched in both promoters and enhancers.
that were up- or downregulated between subject groups, but cannot assess whether expression data for a single biopsy contain enough information to accurately classify the subject as active UC, active CD, or control. This is a relevant question as the diagnosis between active UC and CD is often clinically difficult, and expression data might be useful to increase overall diagnosis accuracy. We reasoned that CAGE-based analysis is not realistic in a clinical setting; instead, we sought to define a small set of CAGE-based biomarkers for active UC, active CD and controls that were amendable to robust targeted methods such as qPCR.

To achieve this, we used successive steps of biomarker selection and testing (Fig. 7a). First, we analyzed the CAGE data from the 25 UCa, 20 CDa, and 29 Ctrl subjects in cohort 1 with an ensemble of statistical and manual curation methods to extract 274 TSSs/enhancers distinguishing UCa, CDa, and Ctrl groups (Supplementary Data 15). A Random Forrest (RF) classification framework trained on CAGE data for these 274 biomarkers could predict UCa, CDa, and Ctrl diagnosis with an overall accuracy of 95%, assessed by five-fold cross-validation (Fig. 7b, left and middle panels). This high accuracy could be retained when reducing the number of biomarkers (Fig. 7b, right panel). We were able to design and validate qPCR primers corresponding to 161 of these biomarkers (Supplementary Data 15, 16), and analyzed their expression using microfluidic qPCR on the same biopsies as above (cohort 1: 161 primer pairs analyzed in biopsies from 74 subjects) (Supplementary Data 17). An RF trained on the microfluidic qPCR expression data gave an overall accuracy of 84% assessed by five-fold cross-validation (Fig. 7c, left and middle panels). Thus, accurate classification of IBD based on these biomarkers was achievable across experimental methods. Similar accuracy levels were achieved using fewer biomarkers: we retained comparable accuracy using 30–40 biomarkers (Fig. 7c, right panel). Encouraged by this, we reduced the number of biomarkers to a final set of 35. In this final selection no enhancers were retained, consistent with their lower RNA abundance compared to gene TSS.

To test whether the classification power of our biomarkers generalized beyond cohort 1, we enrolled a second independent validation cohort (cohort 2: 18 CDa, 46 Ctrl, 37 UCa; Supplementary Table 1), which included biopsies acquired at two different hospitals by different physicians, and measured the expression of same 35 biomarkers using microfluidic qPCR (Supplementary Data 18). We then trained a RF on microfluidic qPCR data from the 35 biomarkers from cohort 1 and predicted the diagnosis of the subjects in cohort 2. We achieved an overall accuracy of 85% in cohort 2 (Fig. 7d, left and middle panels), strongly suggesting that our selected biomarkers generalize to the larger population. Although the CD sensitivity was lower when using qPCR than with CAGE (see Discussion), our results were in all cases substantially better than expected by chance, as estimated by training on randomly shuffled cohort 1 labels and then
Fig. 5 Characterization of enhancer clusters in IBD. a Example of an enhancer cluster in the CEBPB locus. Genome browser screenshot of the CEBPB locus, organized as in a but also showing chromatin-defined enhancer clusters from dbSUPER36, and a CAGE derived enhancer cluster located ~150 kb downstream of CEBPB. Because CEBPB has two nearby alternative TSSs with similar activity, most enhancers are linked to both. Lower panel shows a zoom-in of the enhancer cluster where ENCODE transcription factor ChIP-seq peaks are displayed: each black line corresponds to one ChIP-seq peak. b Relation between enhancer IBD-up/downregulation and number of enhancers within an enhancer cluster. Bar plot shows the fraction of enhancers that are significantly downregulated (IBD_{down}, gray) or upregulated (IBD_{up}, white) in UCa and C Da vs. Ctrl, grouped by the number of enhancers within an enhancer cluster. Enrichments not part of clusters are included for comparisons (singleton enhancers). The expected overlap by chance for each bar is indicated as dotted lines, with 95% confidence intervals. c Relation between the number of enhancers within an enhancer cluster and IBD upregulation of linked TSSs. Boxplots show the distribution of IBD vs. Ctrl log_{2} fold changes of TSSs linked to singleton enhancers or enhancer clusters as in b. TSSs are grouped by how many enhancers the linked enhancer cluster contains. d Overrepresentation of ENCODE TF ChIP-seq peaks in singleton enhancers vs. enhancer clusters with >6 members linked to IBD-upregulated TSSs. X-axis shows the log_{2} fold change in ENCODE ChIP-seq peak over-representation in single enhancers vs. enhancers within enhancer clusters having >6 enhancers, where 0 indicates no difference between sets. Only enhancers linked to IBD_{up} TSSs are analyzed. Y-axis shows the associated over-representation P-value. Each dot corresponds to one type of ChIP-seq peak, colored by whether they are annotated as inflammation-associated factors (purple), SMARC- or CTCF factors (orange), or other factors (gray). Factors of interest are highlighted.
predicting the labels of cohort 2 (Fig. 7d, right panel and Supplementary Figure 9a, b). To explore whether CD sensitivity could be improved further, we employed the gradient boosting decision tree method implemented in XGBoost41, training on the same 35 biomarkers cohort 1 and predicted the diagnosis of the patients in cohort 2. This resulted in an increase in CDa classification sensitivity by 13% points with no overall classification performance decrease (Supplementary Figure 9c). Thus, our initial results are not specific to machine learning method employed and may be improved by the use of more complex techniques. In summary, we have shown that a small set of qPCR primers, selected from CAGE data, could distinguish control, active UC, and active CD biopsies.

**Discussion**

Here we have profiled the enhancer and TSS landscape of biopsies taken in the descending colon of admitted IBD patients and controls. Such biopsies are identical to biopsies used in current...
diagnosis methods. This has the advantage that samples are highly clinically relevant, and that transcriptional changes in the disease states measured will be close to the in vivo reality. Thus, from a diagnostic perspective, such tissue samples are more relevant than selected primary cells isolated from tissue. As with all tissue samples, expression changes may be partially due to changes in cell composition. We showed that aside from a shared increase of immune cell-linked TSSs, active CD and UC seem to have different composition of cells, where UC-specific gene expression was more strongly linked to epithelial cells.

CAGE data interpreted on gene level identified similar pathway and GO term enrichments as previous RNA-seq studies. A recent study identified two distinct CD populations based on RNA expression data: we could not replicate this finding but the CAGE data did show the active CD population to be much more transcriptionally diverse than active UC or controls, as discussed below.

This is a large-scale analysis of enhancer activities across a total of 94 CD, UC, and control subjects: ref. 28 investigated two smaller CD populations (N = 10 and 9) using FAIRE-seq, a method identifying accessible DNA, while H3K27ac ChIP-seq has been performed in gut tissue for even fewer individuals, and not in an IBD context. With our data, we could establish that promoters and enhancers that are upregulated in IBD share the same DNA-binding patterns enrichments, and that enhancer clusters upregulated in IBD are distinct from corresponding singleton enhancers.

The enhancer component of IBD reported here gives new possibilities for understanding the disease and its genetics. We show a clear overrepresentation of IBD-associated SNPs in both IBD-upregulated enhancer and promoter regions, and these regions had the largest IBD heritability enrichment compared to a large set of other genomic regions. Thus, the IBD TSS and promoters and enhancers that are upregulated in IBD share the same DNA-binding patterns enrichments, and that enhancer clusters upregulated in IBD are distinct from corresponding singleton enhancers.

Aside from adding to the functional and genetic understanding of IBD, CAGE data can be used to distinguish between active UC, CD, and control subjects. This accuracy could be retained using as few as 35 TSSs, quantified by microfluidic qPCR and validated in an independent cohort. Although the method is not at usable
in clinical settings in its current form, it may provide an inroad in parallel biopsies taken within an inch of the 1st biopsy. The biopsies were immediately placed in RNA later® Stabilization Solution (Ambion©, Life Technologies) and kept at 4 °C for 24 h before long-term storage at −80 °C. The RNA was extracted using PureLink® RNA Mini Kit (Ambion©, Life Technologies) and quantified by means of the ND 1000 spectrophotometer (Thermo Scientific). The purity was measured as the A260/280 ratio and was between 1.9 and 2.1. RNA quality (integrity) was determined using the 2100 Bioanalyzer Instrument (Agilent Technologies) with the Agilent RNA 6000 Pico Kit (Agilent Technologies) and the RNA was purified using the RNeasy Mini Kit as recommended by manufacturer with lysis buffer containing 1% 2-mercaptoethanol (Sigma) and on-column DNase digestion. The average yield was 700 ng, with RIN values between 8 and 9.2. TSS expression analysis of 36 TSS targets (as defined below) was done using the microfluidics Fluidigm platform in parallel with biopsy qPCR, as described below.

Monocyte isolation and stimulation.

Blood was freshly collected from IBD patients with informed written consent. First, freshly drawn blood was diluted at a ratio of 1:1 with PBS. Peripheral blood mononuclear cells (PBMCs) were isolated from diluted blood by Ficoll-Paque density gradient centrifugation according to the manufacturer’s instructions (GE Healthcare, Uppsala, Sweden). Harvested PBMCs were used for monocytes isolation. CD14+ monocytes were isolated from PBMCs by negative immunomagnetic bead separation using Monocytes Isolation Kit II (Miltenyi Biotec, CA, USA). After monocyte isolation, 1 × 10^6 monocytes were plated in each well of 24-well plates (TPP, Trasadingen, Switzerland) in 1 ml of RPMI-1640 medium containing 10% FBS (Sigma-Aldrich), 50 μg/ml penicillin, 50 μg/ml streptomycin, and 0.5 mg/ml gentamycin) at 37 °C in a cell culture incubator. Cells were cultured overnight and stimulated with 10 ng/ml TNFa the next day and harvested at indicated time. Control cells were not treated with TNFa but otherwise treated in the same way. Cells were pelleted, snap frozen on dry ice and stored at −80 °C until use. RNA was purified using Qiagen RNeasy Mini Kit as recommended by manufacturer with lysis buffer containing 1% 2-mercaptoethanol (Sigma) and on-column DNase digestion. The average yield was 700 ng, with RIN values between 8 and 9.2. TSS expression analysis of 36 TSS targets (as defined below) was done using the microfluidics Fluidigm platform in parallel with biopsy qPCR, as described below.

Ethnicity.

The IBD sample collection and study was approved by the Scientific Ethics Committee of the Capital Region of Denmark, filed under the ethical permits no. M 36/10, M 5/09, M 8/09-057, 43691/H-6-2014-051, and 53619/H-6-2014-051, covering both biopsy collection and the application of data to next-generation sequencing approaches. All individuals provided written informed consent to participate in this study. In this context, participation meant that five extra biopsies were obtained in addition to the number of biopsies required on medical grounds, extending the examination time by 2-3 min. Ctrl subjects were a subset of the above set, but were assigned “Ctrl” when all clinical and paraclinical investigations subsequently indicated no inflammation. Participants were informed both orally and in writing in compliance with the Declaration of Helsinki and the guidelines of the Danish National Scientific Ethics Committee. The gut epithelia organoid and monocyte studies were approved by the local bioethical committee (ethical approval numbers 1302159 and 51899, respectively). All samples were obtained from patients who provided informed consent before surgery (for gut organoids) or blood sampling (monocytes).

Tissue isolation and RNA extraction.

Mucosal pinch biopsies (weights on average 15 mg) were obtained from the descending colon using endoscopic forceps. We choose the descending colon in order to avoid intersegmental variation. The endoscopic diagnosis of active or inactive disease was confirmed by histopathology conducted on parallel biopsies taken within an inch of the 1st biopsy. The biopsies were immediately placed in RNA later® Stabilization Solution (Ambion©, Life Technologies) and kept at 4 °C for 24 h before long-term storage at −80 °C. The RNA was extracted using PureLink® RNA Mini Kit (Ambion©, Life Technologies) and quantified by means of the ND 1000 spectrophotometer (Thermo Scientific). The purity was measured as the A260/280 ratio and was between 1.9 and 2.1. RNA quality (integrity) was determined using the 2100 Bioanalyzer Instrument (Agilent Technologies) with the Agilent RNA 6000 Pico Kit (Agilent Technologies) and the RNA was purified using the RNeasy Mini Kit as recommended by manufacturer with lysis buffer containing 1% 2-mercaptoethanol (Sigma) and on-column DNase digestion. The average yield was 700 ng, with RIN values between 8 and 9.2. TSS expression analysis of 36 TSS targets (as defined below) was done using the microfluidics Fluidigm platform in parallel with biopsy qPCR, as described below.

Ethnicity.

The IBD sample collection and study was approved by the Scientific Ethics Committee of the Capital Region of Denmark, filed under the ethical permits no. M 36/10, M 5/09, M 8/09-057, 43691/H-6-2014-051, and 53619/H-6-2014-051, covering both biopsy collection and the application of data to next-generation sequencing approaches. All individuals provided written informed consent to participate in this study. In this context, participation meant that five extra biopsies were obtained in addition to the number of biopsies required on medical grounds, extending the examination time by 2-3 min. Ctrl subjects were a subset of the above set, but were assigned “Ctrl” when all clinical and paraclinical investigations subsequently indicated no inflammation. Participants were informed both orally and in writing in compliance with the Declaration of Helsinki and the guidelines of the Danish National Scientific Ethics Committee. The gut epithelia organoid and monocyte studies were approved by the local bioethical committee (ethical approval numbers 1302159 and 51899, respectively). All samples were obtained from patients who provided informed consent before surgery (for gut organoids) or blood sampling (monocytes).
isolated using Trizol (Invitrogen) in combination with PureLink minicolumns (Life Technologies). The PurEsq minikit was used as recommended by manufacturer including the on-column DNase step. The average yield was 5540 ng with a RIN between 6.8 and 9.8. TSS expression analysis of 36 TSS targets (as defined below) was done using the microfluidics Fluidigm platform in parallel with biopsy qPCR, as described below.

**CAGE library preparation, sequencing, mapping, and processing.** CAGE libraries were prepared according to ref. 11 with an input of 1500 ng of total RNA as starting material. Libraries were prepared otherwise the same as described in ref. 11. The following four barcodes were used: no. 2 (CTT), no. 3 (GAT), no. 6 (ACG), and no. 8 (ATC). All used primers and adaptors were purchased from Integrated DNA Technologies (IDT). A HiSeq2000 instrument from Illumina at the National High-throughput DNA sequencing Centre, University of Copenhagen was used. Phi-X spike-ins were added for each lane. Reads were matched to their originating samples if barcodes matched identically. Linker sequences were removed and reads were filtered, requiring at least 30% in the bases using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) with the length of 25 bp. CAGE tags were mapped using Bowtie2 (version 0.12.7) to the hg19 assembly using v = 2 and otherwise standard settings but allowing for multiple alignments. Subsequently, only uniquely mapping reads were retained. Reads that mapped to chrM were discarded. Supplementary Data 2 shows mapping statistics. In short, for each library on average 17.3 million reads mapped, of these, on average 74% (S.D.± 2.8) mapped to a unique location.

**Evolutionary conservation analysis of enhancer regions.** Enhancer regions from the strict set were aligned by their midpoint. Enhancers whose ±2000 bp flanks from the midpoints overlapped the empirical enhancer set at least eight times (empirical enhancer set) were included as described in the main text, only CDA, UCA, and Ctrl samples were included in this analysis. The four major batches were included as blocking factors (Supplementary Data 1). The following

**Visualization of ChIP signals at predicted enhancer regions.** ChIP heatmaps for transcription factor peaks were derived from Chip-Seq peaks in the ENCODE consortium (version 3, including 161 TFS from 91 cell types). Number of peaks were aggregated for coverage per bp and used as TFBS signals in the heatmap. H3K27ac and H4K4me1 ChIP-Seq signals used for differently arranged heatmaps were downloaded from the Roadmap Epigenetics Consortium. Heat maps were ordered by the increasing width of enhancers from the strict set and centered at enhancer midpoint. Chip-Seq signals falling in ±200 bp of the enhancers were visualized in 401 bins as averages. Binned values were then broken into percentiles and assigned to a color gradient ranging from red (high Chip-Seq signals) to blue (low Chip-Seq signals).

**Definition of enhancer clusters.** Consecutive enhancers from the strict set were chained together if their distance was no longer than 15 kb, producing enhancer clusters with varying number of enhancer members. An enhancer cluster was defined as linked with a TSS if at least one of its enhancer members had a significant and positive correlation with the TSS, as defined above.

**Exploratory differential expression analysis data.** We defined an expression matrix consisting of CAGE-defined TSSs (as defined above) and their summed number of CAGE tags for each subject. Raw counts were normalized to log-TPM values and subjected to a variance stabilizing transformation using the variance Stabilizing Transformation function from the DESeq2 package (with blind = TRUE) to improve linear modeling of the data by dampening the mean-variance trend inherent in count-type data. Principal Component Analysis (PCA) was used to explore major expression patterns in the dataset using the pcaNorm function from base R (with scale = TRUE and center = TRUE). This revealed four major groups which were not related to any specific part of the study, but was corrupted with the date, on which the individual samples were prepared (Supplementary Fig. 11). This effect may be due to change of reagent batches e.g. the biotin hydrazine and antarctic phosphatase which were replaced between batches. To remove these batch effects for plotting purposes, we used the ComBat function (default parameters) from the sva package, after first performing a variance stabilizing transformation with blind = FALSE (Supplementary Fig. 11). All samples and all five experimental groups (CDa, UCa, CDi, UCi, and Ctrl) were used for the batch effect correction for the initial PCA plot (Figs. 2a and 3d). Only CDa, UCa, and Ctrl were used for batch-corrected gene-level expression plots for the later parts of the main text (Fig. 3f, Supplementary Fig. 2d, e, h and Supplementary Data 3).

**Evolutionary conservation analysis of enhancer regions.** Enhancer regions from the strict set were aligned by their midpoint. Enhancers whose ±2000 bp flanks from the midpoints overlapped the empirical enhancer set were discarded. As described in the main text, only CDA, UCA, and Ctrl samples were included in this analysis. The four major batches were included as blocking factors (Supplementary Data 1). The following...
contrasts were used: (i) IBd: average of CDa and UCa different from Ctrl and (ii) CDvsUC: CD different from UC. Resulting \( P \)-values were corrected for multiple testing using a step-down Bonferroni method (producing FDR values). To obtain the final sets of differentially expressed TSSs (or genes) used in the main analysis, we filtered TSSs (or genes) with \( \log_2 \) fold change \( > 1 \) and \( FDR < 0.05 \). The IBd\textsubscript{UP}/IBd\textsubscript{DOWN} SETS correspond to TSSs (or genes) with a positive/negative \( \log_2 \) fold change in contrast \( i \) and the CD\textsubscript{SPEC}/UC\textsubscript{SPEC} sets correspond to TSSs (or genes) with a positive/negative \( \log_2 \) fold change in contrast \( ii \), where a positive \( \log_2 \) fold change indicates a higher expression in CDa compared to UCa (and vice versa). The diffSpliceDGE function was used to identify differential TSS usage within annotated genes. The Smies method was used to aggregate TSS \( P \)-values to gene level. To decide whether including additional covariates would improve modeling of the data, we added effects of various medical labels, such as medication, smoking, gender, etc. to the design matrix described above, and used limma-voom to test for differential expression in a similar manner (due to the higher speed of limma compared to edgeR). Most extra covariates did not have any significantly differentially expressed genes, except minor effects for gender and SABA treatment. Since these were far smaller than the effects of either inflammation or batch, we chose not to include them in the main analysis shown in the main text (Supplementary Fig. 2a).

### Differential expression analysis of enhancer regions

Because edgeR analysis has limited power for globally locally expressed entities like enhancer RNAs, we used an alternative approach to assess differential expression between conditions for these data. The enhancer RNA data were normalized using the median-based method implemented in the DESeq package \( \text{default}^{35} \). Instead of normalizing directly on the counts of enhancer RNA, we used the counts from TSSs from the same samples. These signals are typically orders of magnitude higher and the measurement error is therefore associated with much smaller uncertainties. The R-package EBSeq (version 1.10.0) was used for the identification of differentially expressed enhancers based on the normalized data\(^{30} \). EBSeq assumes that any given data row (enhancer count data for CDa, UCa, and Ctrl samples, respectively) can follow one of a number of possible expression patterns, corresponding to different plausible hypotheses about that specific enhancer. For the three conditions considered here, there are a total of five possible patterns

**Pattern 1:** Enhancers expressed at the same level in CDa, UCa, and Ctrl (non-differential expression).

**Pattern 2:** Enhancers with the same expression level in CDa and UCa, and a different expression level in Ctrl (general IBd enhancers).

**Pattern 3:** Enhancers having similar expression levels in CDa and Ctrl, while being expressed differently in UCa (enhancers specific for UCa).

**Pattern 4:** Enhancers having similar expression levels in UCa and Ctrl, while being expressed differently in CDa (enhancers specific for CDa).

**Pattern 5:** Enhancers expressed differently in all three conditions.

For each enhancer in the dataset, EBSeq computes the posterior probabilities that it follows each of these five patterns. EBSeq also produces posterior estimates of the fold change in expression levels between each pair of the three conditions for any given enhancer. Another feature of the EBSeq method is that it uses hierarchical (or multi-level) modeling: the expression level (read count) of each individual enhancer is assumed to follow negative binomial distributions whose \( r \) and \( q \) parameters are inferred from the data. For any given enhancer, these parameters may be identical across conditions (non-differential expression) or different across conditions (differential expression). Different enhancers each have their own \( r \) and \( q \) parameters (assumed to be random), but all \( r \) and \( q \) parameters are also estimated from data. This multi-level modeling approach results in “borrowing of strength” (inference from data on one enhancer will help inform inference for other enhancers), and “shrinkage” (parameter estimates for enhancers with little data will be pulled toward the mean). Since the majority of all enhancers are not differentially expressed the approach is conservative (it requires a substantial amount of data to pull log-fold estimates away from zero) and thereby helps avoiding issues with multiple testing.

Specifically, the analysis was performed using the EBMultiTest function in the EBSeq package. The set to 100 iterations. Convergence was checked by inspecting plots of the hyper-parameters \( \alpha \text{Beta} \) and \( \text{Beta} \) and the mixture parameter \( p \) and ensuring their values had settled at stable values at the end of the iterations. The function QP was used to generate QP-plots confirming that the beta prior used in the analysis was appropriate (data not shown). We used a posterior probability cutoff of 0.9 for deciding when to label an enhancer as differentially expressed, corresponding to a false discovery rate of 0.1. Labeled enhancers were then recombined with the classification of differential expression made for TSSs as follows (\( P = \) posterior probability):

- **IBd\textsubscript{UP}:** \( P(\text{Pattern} 2) \geq 0.9 \) and (CDa and UCa positive log fold change vs. Ctrl)
- **IBd\textsubscript{DOWN}:** \( P(\text{Pattern} 2) \geq 0.9 \) and (CDa and UCa negative log fold change vs. Ctrl)
- **CD\textsubscript{SPEC}:** \( P(\text{Pattern} 4) > 0.9 \)
- **U\textsubscript{SPEC}:** \( P(\text{Pattern} 3) > 0.9 \)

There is no simple way of accounting for batch effects in the software used here, and we therefore performed the analysis independently for the two batches. Results reported in this paper are for the larger batch (58 subjects), but we found very similar results for the smaller batch (Supplementary Fig. 4d). Analyses were performed both on the larger (permissive) enhancer set, and (the smaller) restricted enhancer set. On the whole, enhancers identified as being differentially expressed in the strict data set were a subset of those identified based on analysis of the permissive data set (data not shown).

### Gene ontology and protein interaction analysis

For differential expression analysis performed at the gene level, we performed enrichment tests for GO-terms using the gProfiler \( \text{package}^{20} \) using the gProfiler function. Default settings were used, except the background set (i.e., “universe”) was set to all expressed genes. The results of the enrichment analysis on each of the four sets can be found in Supplementary Data 7. Additionally, we also queried each of our gene-level differential expression sets against the STRING database of protein-protein interaction networks (using default settings, but removing proteins with no connections to any other proteins).

### FANTOM5 cell specificity enrichment analysis

For TSSs that were differentially expressed, we performed enrichment tests for FANTOM5 cell specificity of TSSs. We downloaded the table of cell type specificity of CAGE tag clusters from \( \text{http://fantom.gsc.riken.jp/5/datasets/phase1.1.extra/Sample_ontology_enrichment_of_CAGE_peaks/} \). For each TSS, we noted whether it was differentially expressed and if it overlapped (based on genomic coordinates) with a FANTOM5 cell-specific tag cluster. We used the \text{table} function from \text{R} to create a \( 2 \times 2 \) contingency table for differential expression statistics and cell specificity, and tested for independence between rows and columns of this table using a Fisher’s exact test (via \text{R}’s \text{fisher.test}). \( P \)-values for each cell category were corrected for multiple testing using the Benjamin–Hochberg method. The complete table of statistics for all cell type terms can be found in Supplementary Data 8.

### Transcription factor binding site enrichment analysis

Promoter regions were defined as follows: for each TSS defined as differentially expressed in the respective sets (IBd\textsubscript{UP}, IBd\textsubscript{DOWN}, UC\textsubscript{SPEC}, CD\textsubscript{SPEC}), on the whole, enhancers identified as differentially expressed in the TSSs permissive enhancer set. On the whole, enhancers identified as differentially expressed in our analysis (see above) were manually paired with CAGE-defined transcription factor expression (ComBat-corrected gene-level expression).

### ENCODE transcription factor ChIP peaks enrichment

ChIP-Seq peaks were downloaded from the ENCODE consortium (161 TFs from 91 cell types; see above section for heatmap visualization for details). We analyzed the enrichment of each set of TF peaks in a given set of enhancer regions (e.g., IBd\textsubscript{UP} enhancers) by constructing a contingency table as defined in Supplementary Table 2. Enrichment of a given TF in the region of interest was defined as \( \frac{A}{(A+B+C)} \) (terms defined in Supplementary Table 2) and significance of enrichment was tested by Fisher’s exact test for each TF based on the above contingency table. For the analysis in Supplementary Fig. 5 the regions of interest were IBd\textsubscript{UP} or IBd\textsubscript{DOWN} enhancers and the background regions were non-differentially expressed enhancers. In Fig. 3d, we only focused on enhancers that were associated with IBd\textsubscript{UP} TSSs by expression correlation as defined previously. Among these enhancers, we defined the regions of interest as the enhancers from large enhancer clusters (>6 enhancers in a cluster) and the background regions as singleton enhancers (not part of any cluster).

### Selection of biomarkers

In order to create an initial candidate list for machine learning-based selection and classification, we reasoned that any single prioritization method has its own disadvantages. Therefore, we performed an ensemble approach where we integrated the results of multiple analysis methods each aimed to extract TSSs with high power to distinguish the subject groups. As our main goal was classification by machine learning, we will refer to these TSS regions as ‘Fea- ture biomarkers’ in this section (instead of biomarkers as in the main text). To assess the shared inflammatory response in UCa and CDa vs. Ctrl was strong, we divided our analysis into an IB set of features (shared CdA and UCa up- or downregulation vs. Ctrl), and a set of features corresponding to features differentially expressed between UCa and CDa.

The ensemble analysis consisted of the following components:

i. edger\(^{38} \). In a similar fashion to the quasi-likelihood implementation explained above, standard edger \( \text{(fitGLM and glmLRT)} \) was used to test for effects of CDa vs. Ctrl, UCa vs. Ctrl, CDa vs. UCa, and CDa\&UCa vs. Ctrl (shared

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03766-z | www.nature.com/naturecommunications
inflammatory response), while controlling for batch effects. Extracted TSSs had a Benjamini–Hochberg FDR < 0.05 and a log₂ fold change ≥ 1, for a total of 1000 sets.

ii. Partial Least Squares Discriminant Analysis (PLS-DA): The pLSA function from the R-package DiscriminR (https://cran.r-project.org/package=DiscrimInR) was used to perform PLSDA with two components on the variance stabilized expression values from above, using all amplified independent qPCR gene patients, in three group categories: CD+, CD-, UCa+ UCa- and Ctrl. The 500 TSSs with the highest and lowest loadings for the component separating Ctrl from the remaining samples were extracted for a total of 1000 TSSs.

iii. ComBat54 followed by PCA: ComBat corrected values were used for PCA analysis. We identified the component that contributed the most to the difference between the two states and sampled 1000 TSSs from each side with the highest rotation values. We focused on the PCAs that gave the most explanations for each of the comparisons (PCA1 for shared IBD-specific signal and PCA2 for CDspec and UCapec), producing two lists for IBd (from each side one list) and one list for CDspec and UCapec (from each side one list).

iv. ComBat followed by limma55: Combat normalized data (as described above) was used for differential expression analysis using the limma package and the lmFit function. TSSs that passed the threshold of log fold change ≤−1 and FDR < 0.05 were taken into account making two lists for IBd (over-represented in UCa vs. Ctrl and overrepresented in CDa vs. Ctrl) and one list for CDa vs. UCa.

v. Random Forest analysis. We utilized the inherent ability for Random Forests (RFs) to rank features by their importance for classification accuracy. To do this, we used the randomForest function from the randomForest R package (https://CRAN.R-project.org/package=RandomForest) on Combat-normalized TC expression data (as described above in the “Exploratory data analysis” section). To circumvent the “Large p, small N” problem with our data (74 samples, 48,593 TCS), which makes difficult to find the most predictive features, we used an iterative feature selection approach where we in each step remove a subset of the least important features. More specifically: in each iteration we first evenly divided all the data into subsets each containing max 500 features. For each of these data subsets a RF was trained to classify the patient labels based only on the data in that particular subset. Next, for each of the data subsets, we removed the 5% of features that had the lowest classification power (lowest MeanDecreaseAccuracy values) resulting in a list of features that is analyzed in the next iteration of the procedure. To make sure an important feature was not lost due to a challenging subset of data, in each iteration we randomly divided all our data into subsets 10 times. In other words, features had several chances of showing its importance. This iterative procedure was terminated when less than 200 features were left. Each RF was made with default parameters except specifying that 501 trees per forest should be generated and specifying the number of variables randomly sampled as candidates at each split by setting the mtry parameter to sqrt(48593) (that is 6969 TCS). The high-confidence data set was obtained by requiring that (i) the tag cluster was known_alternative_cds, known_non_alternative_cds, intergenic, “known_alternative_tss”, “known_non_alternative_tss”, or “tss” (ii) the width of the TC was in the range of 2–100 nucleotides (both included), (iii) the TC was expressed more than 100 TPM in at least 3 subjects.

As a summary, the ensemble produced 10 candidate sets for separating CDa and UCa from Ctrl (3 edgeR, 4 ComBat, 2 RandomForest, 1 PLSDA) and 5 candidate sets for separating CDa and UCa from non-UCa (edgeR, limma, 2 RandomForests, and PCA).

Next, we overlapped the lists from each comparison to find a consensus set of classification candidates. For the IBd vs. Ctrl comparison, we selected features that appeared in at least 9 out of 10 lists, resulting in 63 features. From the UC/CD set we selected the 169 features that appeared in at least 2 out of 5 lists. The reasoning behind this unbalanced selection was to have a higher number of features for the more difficult task of distinguishing CDa from UCa, compared to the easier task of distinguishing Ctrl from UCa and CDa.

The two lists were merged and a short list of expert-curated known and novel biomarkers (21 TSSs and 10 enhancers) was added. The combined list was then further pruned by manual curation, taking into account overall expression strength and genomic loci complexity. In total, the initial list comprised 263 features, which was used for initial classification using CAGE data (see below).

**Microfluidic qPCR analysis.** For qPCR analysis, the 263 features identified above were assessed for qPCR primer targeting suitability. 181 candidates were taken into the primer design stage, where genomic complexity and primer design feasibility was evaluated manually.

Primers were designed using Primer3 (v0.4.0)56 aiming for an amplicon length of 70–110 bp and a primer melting temperature of 60 °C calculated by Breslauer thermodynamics. Primers were designed to be intron-spanning when possible, and each designed primer-set was analyzed with the UCSC browser in-silico PCR tool in order to ensure a unique region would be amplified. Primers were synthesized by Integrated DNA Technologies (IDT). Primer sequences and conversion data are included in the Supplementary Data. Primer efficiencies and dynamic ranges were acquired from standard curves constructed from several separate dilution series of pooled cDNA: with the dilutions 1, 1.25, 1, 100, 1,500, and 1,250. Melting curves of amplicons were measured to ensure primer specificity. 161 primer pairs (features) successfully passed our quality control and were used for analysis of cohort 1 on the Fluidigm platform.

The CDNA synthesis and preamplification for cohort 1 was performed as described previously57. Total RNA (500 ng RNA, from biopsies) was converted into cDNA using QuantiTect Reverse Transcription kit (Qiagen), using a mix of random and oligo-dT primers, as per the manufacturer’s instructions. We used 2 separate cDNA reactions for each DNA sample. Preamplification was performed using TaqMan Prep Master Mix (Applied Biosystems). A 500 μl primer mix (200 nM) combining all primers to be used on the Fluidigm plate was prepared by pooling 5 μl of all primer pairs (20 μM) and filling up the remaining volume with low EDTA TE-buffer (VWR International). TaqMan Prep Master Mix (3 μl) was mixed with 2.5 μl of the 200 nM primer mix, 2.5 μl diluted cDNA, and incubated at 95 °C for 10 min, followed by 15–21 cycles of 95 °C for 15 s and 60 °C for 4 min (the number of cycles depended on the expression measured by CAGE). 16 U of Exonuclease I (New England Biolabs) was added to the preamplified cDNA, and incubated 30 min at 37 °C, 80 °C for 15 min. The preamplified cDNA was diluted with low EDTA TE-buffer (VWR International) either 1:5 if the cDNA was preamplified for 21 cycles or 1:10 if the cDNA was preamplified for 19 or 15 cycles.

RNA expression was analyzed by real-time qRT-PCR in the microfluidics system BioMark™ 96.96 Dynamic Array (Fluidigm) following the procedure described previously57. The following cycle parameter were used: Thermal Mix with 2 min at 50 °C, 30 min at 70 °C, 10 min at 25 °C, UNG and Hot start with 2 min at 50 °C, 10 min at 95 °C, followed by 35 cycles with denaturing for 15 s at 95 °C and annealing/elongation for 1 min at 60 °C. Melting curves were generated after each run to confirm a single PCR product (from 60 °C to 95 °C, increasing 1 °C per 3 s). We performed reactions in duplicates. Non-template controls were used to indicate problems with sample contaminations or non-specific amplification. To assess potential DNA contamination, non-reverse transcriptase controls were used.

We used the Fluidigm Real-Time PCR Analysis software 3.0.2 (Fluidigm) to acquire Cq values. These values were exported to GenEx (MultiD) for data pre-processing including normalization to reference genes, interplate calibration, individual PCR efficiency correction for each primer assay, and averaging of cDNA technical replicates. Using GeNorm50 and NormFinder51, we identified the most stable reference genes of the 10 reference genes included on the plates. This was done separately for all plates run with the same number of preamplification cycles. The used reference genes for each preamplification batch are specified in Supplementary Data 16. By normalizing to these reference genes, ΔCq values were calculated and used in the down-stream analysis.

**Feature selection for cohort 2.** The 161 features were reduced to 36 features by ranking the features for classification power. Briefly, a RF was trained to predict subject groups, both using all groups and one-vs.-rest specification (i.e. UCa vs. non-UCa) using the randomForest R package.56 The optimal value for the mtry parameter was found using 100 five-fold cross-validations using the caret package (https://cran.r-project.org/web/packages/caret/index.html) train function and tune = 1000. Feature importance for both classification tasks were extracted using the importance function with type = 2 (mean decrease in node impurity or gini index). Primer pairs with a high importance in both complete and one-vs.-rest classification were selected as the most predictive. In addition to these RF predictive features, six gene-based features were selected to validate pathology-based findings in the first cohort.

cDNA synthesis and preamplification for cohort 2 was performed as described above for cohort 1 with an adjustment of the number of pre-amplification cycles. In this cohort CDNA was either preamplified 15 or 20 cycles, depending on expression level of the primer targets. The preamplified cDNA was diluted with low EDTA TE-buffer for 15 International volumes if the cDNA was preamplified for 20 cycles, or 1:10 if the cDNA was preamplified for 15 cycles. RNA expression was analyzed by real-time qRT-PCR in the microfluidics system BioMark™ 192.24 Dynamic Array (Fluidigm) following the same protocol as used for the cohort 1 runs, with volumes adjusted to the 192.24 format, as described by manufacturer. Expression data (ΔCq values) were acquired and treated similar to cohort 1 above.

**qPCR analysis of monocyte and organoid samples.** cDNA synthesis and preamplification was performed as described for cohort 1 and 2, but with an input of 500 ng (organoids) or 440 ng (RNA in microliters) of RNA in microliters due to limited amounts of RNA. RNA expression was analyzed by real-time qRT-PCR in the microfluidics system BioMark™ 192.24 Dynamic Array (Fluidigm) as described
for cohort 2. Monocyte samples and organoid samples were placed on two independent plates.

**Batch correction and expression normalization of qPCR data.** For each primer and all ΔCq values from cohort 1 and cohort 2 were batch corrected using the removalBatchEffect function from limma, giving the subject/group labels (CDa, UCa, Ctrl) as covariates and cohort and hospital of origin as batch effects and specifying the ‘robust’ method should be used for the linear models. To make mononuclear and monocyte data comparable with the batch-corrected data a batch correction factor, obtained as the mean of the batch correction factors for the two hospitals, was applied to the organoid and monocyte data. Afterwards the batch corrected ΔCq values were calculated in a two-step process: (1) \( x = 2^{\Delta Cq} \). (2) \( \Delta Cq = \log_2 (x/\text{min}(x)) \).

**Preprocessing of microfluidic qPCR data.** For classification analysis of cohorts 1 and 2 (see below), only Ctrl, UCa, and CDa subjects were considered. We removed one subject and one primer pair from the analysis which both had > 10% "NA" values from the ΔCq Fluidigm data (Supplementary Data 18), making the final feature count 35. The remaining missing values \( (N = 16, 0.26\% \text{ of total}) \) were imputed via bagged regression trees using the preprocess function from the "caret" R package (https://CRAN.R-project.org/package=caret) specifying methods to the logImpute function. Experimental practice and reporting were performed according to MIQE guidelines.

**Comparison of organoid, biopsy, and monocyte qPCR data.** Only measurable ΔCq Fluidigm data were analyzed (values labeled "missing" or "infinite" were discarded). For each qPCR primer we performed the following quantifications:

i. Inflammatory response in UCa or CDa as log2(ΔCq_CDa-UCa/UCa) (average of ΔCq_α_anti inflammatory)

ii. 4 or 24 h TNF response in gut epithelia organoid as log2(ΔCq_α_organoid_TNF_4h/24h) (average(ΔCq_α_all_organoid_TNF_4h/24h))

iii. 4 or 24 h TNF response in monocyte as log2(ΔCq_α_monocyte_TNF_4h/24h) (average(ΔCq_α_all_monocyte_TNF_4h/24h))

Median of responses across replicates or subject group (CDa, UCa, monocyte TNF 4 h, monocyte TNF 24 h, organoid TNF 4 h and organoid TNF 24 h) was then used to indicate the IBD or TNFα response of the group. Specifically, if the median IBD or TNFα response of a qPCR primer in a sample group was >0, it was labeled upregulated in the sample group, and vice versa. qPCR primers with missing median in any of the sample groups were discarded, resulting in analysis of 30 qPCR primers in Fig. 2f and Supplementary Fig. 3c.

**Real-time quantitative PCR.** For specific enhancer RNA expression validations (Fig. 4), we used the Applied Biosystems® QuantStudio™ 6 Flex Real-Time PCR System, with the 384-well plate. This platform allows for higher reaction volumes compared to the Fluidigm system. The cDNA preparations were done as described for cohort 1. All samples were run in triplicates in 5 μl reaction volumes using SYBR® Master Mix (Life Technologies) as described by manufacturer, with a primer concentration of 0.5 μM, and under the following conditions: 50 °C for 2 min, 95 °C for 2 min, 45 cycles of 95 °C for 15 s, 58 °C for 1 min. All results were normalized to the reference gene PIAS4.

**Prediction of subject diagnosis using random forests.** For all three classification analyses in Fig. 7, we used a two-step RF classification approach with the randomForest package (https://CRAN.R-project.org/package=randomForest). In this approach, the first step was a RF trained to distinguish IBD (CDa and UCa) from Ctrl and the second step was a RF trained to distinguish CDa from UCa. The two-step approach was chosen over an all-in-one classification analysis of cohorts 1 and used for prediction in cohort 2. For each of the 101 iterations, the overall accuracy and CDa, UCa, and Ctrl-group-specific accuracy, sensitivity and specificity were calculated. For each of these, we reported the average and 95% confidence interval across the 101 iterations. The CDa, UCa and Ctrl-group-specific performance measures were calculated by for each label recasting the problem in a binary setting (e.g. for the “Ctrl” label converting both CDa and UCa to “non-Ctrl”) and then assessing accuracy, sensitivity, and specificity. The overall accuracy was assessed on all three categories.

To assess the observed vs. expected performance in classification, we trained 1001 RF models on randomly shuffled labels (again only based on cohort 1 data). These models were then used to predict the patient labels of cohort 2 and these predictions were compared to the true labels. Performance measures were calculated as described above.

To analyze the relation between number of features in the model and classification accuracy, for a selected number of features referred to as \( i \), we randomly selected \( i \) features from the total set of features and performed five-fold cross-validation with the same parameters as identified by the grid search above. This procedure was repeated 11 times for each selected \( i \). The average accuracy and 95% confidence interval of these repeated analyses were reported for each \( i \).

**Prediction of subject diagnosis using XGBoost.** We used the XGBoost framework to perform the same classification as done with Random Forests in Fig. 7c (training on cohort 1 qPCR data and evaluating on the independent cohort 2 qPCR data). For additional power, each sample was scored using a database of curated pathways (wikipathways.org). Briefly, for each patient, we computed pathway expression scores, based on the sum of overlapping genes in each pathway and the qPCR data. From the 569 original pathways in the database, 16 returned a non-null score. These scores were then used to provide 16 supplemental features, on top of the 35 original ones, for the training of the classifier. After parameter optimization on the training data (max_depth, min_child_weight and subsample), the final classifier was used on the cohort 2.

**Analysis of GWAS data.** A recent version of the GWAS catalog was downloaded (2016-12-15) and lifted to hg19 using the gwasucr R-package (http://bioconductor.org/packages/gwasucr/). To calculate linkage disequilibrium (LD), we used genotype data from whole-genome sequencing of Europeans. This was obtained from the VCF files from the 1000 Genomes EUR population release 20130502. We use only diallelic SNPs with <5% missing genotypes that were polymorphic in the CEU population. SNPs from the GWAS catalog were merged with 1000G SNPs by genomic position, and PLINK was used to calculate LD between all SNPs within 500 kb (using the function plink --r2 -ld-window-kb 500 -ld-window 99999 -ld-window-r2 0.5). Pairs of SNPs were considered to be in LD if \( R^2 > 0.75 \). LD clumps were found by expanding all SNPs with their associated LD-SNPs, and then reducing the set by merging overlapping regions.

To assess LD clump enrichment, Empirical Bayes shrinkage was performed with the ebbr R-package (https://github.com/dg/two/ebbr) package using the ebbr fit prior and augment functions (with default arguments). Prior means were calculated as \( \alpha + \alpha (\text{alpha} + \beta) \) and added to the plot of shrunken proportions. LD-clump (one-sided) enrichment was calculated using the kegg function from the limma R-package, but using user-supplied sets: universe set was set to all TSSs/ enhancers, de the respectively differentially expressed sets and gene/pathway set to all pairs of overlapping LD clumps for a GWAS disease. No trend based on an additional covariate was used.

Partitioned heritability analysis was performed using the ldsc tool from (https://github.com/bulik/lolsc). The two analyses below were done based on recommended practice by the creators of the tool: https://github.com/bulik/lolsc/wiki/Partitioned-Heritability. LD-scores were calculated for TSSs, enhancers, and IBD_0, and IBD_1, sets (including both promoter (~500/-100 bp around TSS peaks) and enhancer (~300 bp around enhancer midpoints)). IBD SNP summary statistics were obtained from the International IBD Genomics Consortium, https://www.ibdgenetics.org/downloads.html (downloaded 2016-12-22). SNP summary statistics were processed using wmerge_sumstats.py (with settings -N-nas 128828 -N-con 21770 for sample sizes (as described in ref. 7). Base model enrichment statistics for this set of GWAS summary statistics can be seen in Supplementary Fig. 8b. All four extra categories were added to the baseline model for Fig. 6d. Additionally, each new annotation was separately added to the baseline model in sequence, similar to the cell-type group analysis from the original paper. Proportions, enrichments and coefficients were obtained from each model and compared (Supplementary Fig. 8c).

**Data availability.** CAGE data from this study has been deposited in GEO database under accession number GSE95437. Microfluidics qPCR data are available in Supplementary Data 9, 17, 18.
Acknowledgements

Work in the A.S. laboratory was supported by grants from the Lundbeck Foundation, the Novo Nordisk Foundation, Innovation Fund Denmark, the Danish Council for Independent Research and Elixir Denmark. M.V. was supported by a Marie Skłodowska-Curie Individual Fellowship (grant 659505). O.H.N. was supported by Axel Møller Fonden, Holger og Ruth Hesses Mindefond, P.A. Messerschmidt og Hustrus Foundation, and the Foundation of Aase and Ejnar Danielsen. R.A. was supported by grants from the Danish Council for Independent Research and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant 638273). K.B.J. was supported by the Lundbeck Foundation and Danish Cancer Society. We thank Anders Albrechtsen for help with statistical genetics analyses, and Lene Buhl Riis for help with histology.

Author contributions

Additional information
Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-03766-z.

Competing interests: The authors A.S., M.B., J.B., M.V., M.T., K.V.S., O.H.N. and J.B. have filed a patent for the method for classifying CD/UC vs. control based on qPCR, based on this study. The remaining authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

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

© The Author(s) 2018