Conserved transcriptomic profile between mouse and human colitis allows unsupervised patient stratification

Paulo Czarnewski, Sara M. Parigi, Chiara Sorini, Oscar E. Diaz, Srustidhar Das, Nicola Gagliani & Eduardo J. Villablanca

Clinical manifestations and response to therapies in ulcerative colitis (UC) are heterogeneous, yet patient classification criteria for tailored therapies are currently lacking. Here, we present an unsupervised molecular classification of UC patients, concordant with response to therapy in independent retrospective cohorts. We show that classical clustering of UC patient tissue transcriptomic data sets does not identify clinically relevant profiles, likely due to associated covariates. To overcome this, we compare cross-sectional human data sets with a newly generated longitudinal transcriptome profile of murine DSS-induced colitis. We show that the majority of colitis risk-associated gene expression peaks during the inflammatory rather than the recovery phase. Moreover, we achieve UC patient clustering into two distinct transcriptomic profiles, differing in neutrophil-related gene activation. Notably, 87% of patients in UC1 cluster are unresponsive to two most widely used biological therapies. These results demonstrate that cross-species comparison enables stratification of patients undistinguishable by other molecular approaches.

1 Immunology and Allergy Unit, Department of Medicine, Solna, Karolinska Institute and University Hospital, 17176 Stockholm, Sweden. 2 Department of Medicine and Department of General, Visceral and Thoracic Surgery, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany. 3 These authors jointly supervised this work: Nicola Gagliani, Eduardo J. Villablanca. Correspondence and requests for materials should be addressed to E.J.V. (email: eduardo.villablanca@ki.se)
Ulcerative colitis (UC) is a type of inflammatory bowel disease (IBD) that is mostly restricted to the colon and is characterized by changes in the mucosal architecture, epithelial function, increase in immune cell infiltration, and an elevated concentration of inflammatory cytokines. Symptoms include diarrhea, abdominal pain, rectal bleeding, lack of appetite, and fatigue, all of which significantly affect the patient’s quality of life. UC is a heterogeneous disease, presenting diverse macroscopic features, symptoms, grades of inflammation, and colonic affected areas1,2.

Although there is no definitive cure for UC, there are biological therapies available, which target the inflammatory response during UC by means of inhibiting proinflammatory cytokines or by blocking immune cell migration3. Among these, the most frequently used biological therapies in UC patients block tumor necrosis factor (TNF) with anti-TNF antibodies (such as infliximab, IFX)4 or leukocyte migration (such as vedolizumab, VDZ)5,6. However, about 35% and 50% of patients poorly achieve clinical response to IFX and VDZ, respectively. Patients that do not respond develop adverse effects, most notably increased risk of infections, thus requiring continuous medical monitoring and ultimate surgical intervention7,8.

In an attempt to identify genes/pathways as a potential novel therapeutic target, genome-wide association studies (GWAS) have identified more than 200 polymorphisms associated with a higher susceptibility to IBD9,10. Moreover, an unbiased stratification of UC subtypes has not been reported at the molecular and functional levels, to the best of our knowledge. Here, using transcriptomic data from a well-characterized experimental model of colitis, we identify conserved genes between mouse and UC patients. As a result, we offer insights into IBD-risk gene kinetics and to molecularly stratify UC patients in an unsupervised manner.

**Results**

**Human UC is highly variable at the transcriptome level.** In order to stratify UC patients into molecular profiles, we combined four publicly available human UC cohort data sets (n = 102 patients), in which transcriptomic microarrays of the total colonic biopsies were performed12–15 (Table 1 and Supplementary Fig. 1). We ranked genes using the top 100 most variable genes and further tested whether transcriptomic profiles exist (Fig. 1a). Analysis by visual assessment of cluster tendency (VAT)16 indicated that biopsies presented high inter-sample dissimilarities (Fig. 1b), suggesting a poor overall tendency to form consistent clusters. Dimensionality reduction analysis by t-SNE using the top highly variable genes also indicates the formation of a single group with no apparent subdivisions (Fig. 1c). Then, we further statistically tested whether multi-cluster substructures were present in the data set, since most clustering algorithms define transcriptomic profiles even on random noise17–19. However, bootstrapping analysis using the Hartigan’s Dip test18,20 presented a low cluster substructure trend (p > 0.9), regardless of the gene-ranking metrics used (Fig. 1d). Independently of the clustering tendency results, we forced patient subdivision using hierarchical clustering and tested for cluster stability using bootstrapping17,18,21. In line with previous results, formed clusters were highly unstable using the list of highly variable genes (AU ≈ 0%) (Fig. 1e). These results indicate that without prior knowledge of patient subdivision, standard gene-ranking strategies do not allow clustering of UC patients into consistent molecularly distinct profiles.

**Time-series analysis during colon inflammation and repair.** One cause of such inter-patient variability can be attributed to the sampling procedure, which contributes largely to the total data variance and masks real biological differences22,23. To overcome the total data variance, we sought to identify the genes that contribute to inflammation in an independent and unsupervised manner. To this end, we focused the analysis on a list of evolutionarily conserved genes that best discriminate the nuances of inflammation in a well-characterized colitis mouse model24.

To identify these evolutionarily conserved genes, we first elucidated through an unbiased manner which genes and pathways are differentially regulated during mouse colonic inflammation, followed by a tissue regeneration phase. In particular, we took advantage of the widely used dextran sodium sulfate (DSS)-induced model of colitis. This model is one of the few characterized by a phase of damage, followed by a phase of regeneration. Therefore, this model gave the possibility to identify also sets of genes essential in the regeneration phase, a key step toward the resolution of the inflammation. In short, mice were exposed to DSS in the drinking water for 7 days, and then allowed to recover for the following 7 days. During this period, we collected colonic tissue samples every second day to then be analyzed by RNA sequencing (RNA-seq), histology, and flow cytometry (Fig. 2a and Supplementary Fig. 2). First, we confirmed that 7 days of DSS exposure resulted in continuous body weight loss and acute disease severity, until day 10 to then initiate the recovery phase (Supplementary Fig. 2a–b). Histological analysis confirmed epithelial damage, such as desquamation of the epithelial layer on day 6 (Supplementary Fig. 2c), while labeling proliferating cells within crypts (Ki67 staining) indicated a disrupted crypt architecture by day 6 and restoration by day 14 (Supplementary Fig. 2c). Loss of the epithelial cells (CD45−EpCAM+) by day 7–10 and restoration by day 14 was further confirmed by flow cytometry (Supplementary Fig. 2d). To test whether the epithelial barrier integrity was restored by day 14, we gavaged FITC-dextran and measured its concentration in the serum. We detected higher FITC-dextran concentrations on day 7, which indicate barrier disruption, whereas basal levels were detected by day 14, indicating restoration of the barrier integrity.

| Data set ID | Total | Responders | Nonresponders | Ref. |
|-------------|-------|------------|--------------|------|
| Infliximab: |       |            |              |      |
| GSE12251    | 23    | 11         | 12           | 13   |
| GSE73661    | 23    | 15         | 8            | 15   |
| GSE23597    | 32    | 7          | 25           | 14   |
| GSE16879    | 24    | 16         | 8            | 12   |
| Sum         | 102   | 49         | 53           |      |
| Vedolizumab:|       |            |              |      |
| GSE73661    | 37    | 23         | 14           | 15   |
| Pediatric UC.| | | | |
| GSE109142   | 206   | 105        | 101          | 33   |

Data sets used for the classification of ulcerative colitis molecular profiles. Only the number of patients used for analysis are shown (inflamed mucosa before receiving any therapy).
**Fig. 1** Human ulcerative colitis transcriptional signature does not cluster patients into molecular profiles. a Schematic representation of the strategy used for patient group identification, in which four publicly available data sets were combined. Gene ranking was done using the most variable genes in the human data set, which were used for clustering analysis. b Sample dissimilarity heatmaps for visual analysis of clustering tendency (VAT), comparing the human data set using the top 100 variable genes. c t-SNE plot using the top 100 variable genes in the human data set. Each point represents a patient sample. d Hartigan’s Dip test for clustering tendency using all genes in the data set, the top 100 variable genes, the top 100 highly dispersed genes, or the top 100 leading genes in the principal components. e Bootstrapping analysis of hierarchical clustering, comparing the human data set using the top 100 variable genes in the human data set. Numbers in orange indicate the approximately unbiased (AU) p-value, shown as a percentage. AU closer to zero indicates a cluster with low stability. Boxplot represents the median (center line, second quantile), first, and third quartiles (box) and whiskers extending 1.5 times the interquartile range (IQR).

**Fig. 2** Unbiased characterization of the DSS colitis reveals conserved inflammatory signature between mice and humans. a Schematic illustration of the experimental design. Mice received DSS in drinking water for 7 days, after which the treatment was replaced with water. Samples were collected at indicated time points. b PCA on differentially expressed gene counts. Samples were color-coded according to their respective day of collection (from gray to orange to blue). The percentage of variance explained by the respective principal component is indicated in parenthesis. c Clustered heatmap of all differentially expressed genes (left). The mean expression of each gene module is shown (right). Functional annotation of genes in each cluster was done based on Gene Ontology (GO) enrichment. Only the top three enriched processes are shown, sorted by p-value. d Venn diagram comparing the list of DEGs in treatment-naive UC and the DEGs identified in mouse DSS colitis (upper). Among the 650 genes shared among those lists (in red), the number and percentage of genes were found in each module identified in our mouse data set (lower). Modules highlighted in bold are the ones enriched for inflammatory terms in (c). e GO and KEGG enrichment analysis out of the 650 shared genes identified in (d), sorted by p-value.
Inflammatory pathways conserved between mice and humans. Having characterized genes and pathways that are associated with intestinal inflammation and tissue repair during experimental colitis, we investigated whether such pathways are conserved in humans. To this end, we compared the list of DEGs from the mouse experimental colitis with the recently published list of DEGs found in newly diagnosed treatment-naïve ulcerative colitis patients. This is a cohort containing human RNA-seq data, where they report DEGs between UC patients versus healthy controls. We found that among the 4045 mouse DEGs, 650 genes were also found among the list of DEG obtained comparing UC patients versus healthy controls. Moreover, 26 genes of the 75 IBD-risk genes found in our data set were also found among the list of DEGs from human UC. GO and KEGG enrichment analysis revealed that those 650 genes were enriched for inflammatory pathways related to neutrophil degranulation and chemotaxis, as well as cytokine and inflammatory signaling pathways. These results showed that most of the genes/pathways conserved between experimental mouse colitis and human UC are associated with inflammatory responses.

Temporal classification of IBD-risk genes. To understand the temporal expression of the genes associated with the identified IBD polymorphisms (candidate IBD-risk genes), we checked the expression of genes associated with UC or CD identified by single-variant fine-mapping resolution into the list of DEGs from the mouse data set. Out of the 233 reported candidate IBD-risk genes, 40 genes presented very low or undetectable counts in the mouse data set. Among them, we found that 75 genes were found among the DEGs in our mouse data set (Supplementary Fig. 3a and Supplementary Data set 7). In order to provide temporal information regarding the expression of IBD-risk genes during inflammation and repair, we utilized the mouse transcriptional landscape to map at which time point homolog IBD-risk genes were up- or downregulated. Out of the 75 genes shared between mouse DEGs and IBD-risk genes, 45 (60%) were mapped to modules m1, m3, and m4, which represent the acute phase of inflammation (Supplementary Fig. 5c and Supplementary Data set 7). Among them, we found Card9, Il12b, Stat3, Stat4, and Cd40, which have been reported to exert functions during the acute phase of intestinal inflammation.
By contrast, _Fut1_, _Sult1a1_, _Hes5_, and _Tnfsf15_ were mapped to modules m8, m7, and m2, which are downregulated during acute inflammation, while _Rasip_, _Nmi5_, and _Rtel1_ matched with module 6, which is associated with genes that are upregulated during the recovery phase after acute inflammation (Supplementary Fig. 5c). These data thus provided temporal information on when IBD-risk genes are differentially expressed during damage and repair, providing useful insights into their potential roles during inflammation and recovery.

**Conserved genes distinguish two UC transcriptomic profiles.** Having identified genes that contribute to inflammatory pathways that are conserved between mice and humans, we next used those genes to assess whether UC patients can be subdivided into distinct transcriptomic profiles (Table 1, Fig. 3a). To this end, we selected the top 100 leading genes in PC1 and PC2 from the mouse colitis data set and identified the respective human homologs (Fig. 3a). We found that 57 genes were shared between mice and humans. Of these, only 17 genes were found among the 100 most variable genes of the human data set (Supplementary Fig. 6), which might explain why patient classification using highly variable genes was not possible.

Therefore, we performed an unsupervised analysis of the human data set, using 57 homolog genes (Fig. 3a). Of note, VAT analysis using these 57 homolog genes indicated the distinction into two major patient transcriptomic profiles (Fig. 3b), which also resulted in reduced Hartigan’s unimodality test (p < 0.001, Fig. 3c). This indicates that by using mouse most variable genes as opposed to the sole top human variable ones, it is possible to obtain a higher clustering tendency of the UC patient data. To test whether using the mouse homologs also impacted on cluster stability, we performed a bootstrapping analysis. This time, clustering using the top mouse homolog genes resulted in reduced Hartigan’s dip test p-value, shown as percentage. AU closer to zero indicates a cluster with low stability. (Fig. 3d), compared with using the top human highly variable genes (AU = 0%) (Fig. 1e). Hierarchical agglomerative clustering using the mouse homolog genes thus defined two UC transcriptomic profiles, namely UC1 and UC2, comprising 60 and 42 patients, respectively (Fig. 3e). The UC1 transcriptomic profile is defined as patients presenting the higher average expression of the inflammatory genes compared with UC2 (Fig. 3f). We also observed that neither UC1 nor UC2 profiles were discriminated by the overall macroscopic tissue disease severity (Fig. 3g), suggesting that although these two UC profiles are indistinguishable based on the histological Mayo score, they are transcriptionally distinct.

**Key genes allow the distinction between UC1 and UC2 profiles.** In order to characterize UC1 and UC2 beyond conserved genes, we performed differential expression analysis using all genes present in the human data set. We were able to identify 205 highly differentially expressed genes, among which 187 were upregulated in UC1 and 18 were upregulated in UC2 (Fig. 4a). Detailed tables with information on all DEGs comparing UC1 and UC2 are available for exploration (Supplementary Data set 8 and Supplementary Fig. 7a). Among those, cytokines (_TNF_, _IL11_), enzymes (_NOX1_, _MMP3_, and _CYP26B1_), calcium-binding proteins (_S100A8_, _S100A9_), chemokines (_TREM1_, _CXCL8_), and other
proteins related to the inflammatory response (NR3C2, BCL2A1, PARM1, and TNFSF13B) were clearly able to discriminate UC1 from UC2 (Fig. 4b) and Supplementary Fig. 7). Enrichment analysis for cell types, GO, and KEGG pathways revealed that genes highly expressed in UC1 (187) were associated with terms related to neutrophil, neutrophil degranulation, and cytokine–cytokine receptor interaction, respectively (Supplementary Fig. 7b). Venn diagram of the top enriched terms revealed that many overlapping genes are shared among these pathways (Fig. 4c), suggesting that UC1 patients present a distinct transcriptional signature enriched in neutrophil activity and cytokine signaling compared with UC2 patients.

We trained a logistic regression classifier, using each of the DEGs between UC1 and UC2 to identify key genes that could be further used in the clinics for distinction of UC1 and UC2. Genes were tested and scored individually, using the area under the curve (AUC) as a combined measure of sensitivity and specificity (Fig. 4d). We observed that genes such as TREM1 (AUC = 99%), CYP26B1 (AUC = 97%), and CXCL8 (AUC = 97%) were among the top markers to distinguish UC1 from UC2. Other genes such as WNT5A, BCL2A1, CSAR1, MMP1, MMP3, and IL11 also presented AUC scores above 90% and also represented good candidates for UC1 and UC2 distinction in clinical practice.

UC1 and UC2 respond differently to biological therapies. While we stratified UC patients into two molecularly distinct profiles, it was unclear whether UC1 and UC2 show different treatment responses to biological therapies. To address this, we used the patient-specific treatment response obtained 4–8 weeks after the biopsy was taken and treatment with IFX started (Table 1). Interestingly, we observed that on average, 70% of the patients belonging to the UC2 transcriptomic profile responded to infliximab therapy (Fig. 5a), in contrast to <10% of the patients classified as UC1, regardless of the data set analyzed (Fig. 5a).

To extend the applicability of our findings, we made use of another set of UC patients receiving vedolizumab and repeated the same procedure as before (Table 1). Transcriptomic data from UC patients were analyzed using the most relevant genes identified in our mouse colitis model and then clustered as described above to reveal UC1 and UC2. Between them, UC1 presented a higher expression of the conserved inflammatory genes (Fig. 5b). We observed that about 60% of the patients belonging to the UC2 transcriptomic profile responded to VDZ, in comparison with about 13% of the patients belonging to the UC1 transcriptomic profile (Fig. 5c). Taken together, the data indicate that patients belonging to the UC2 transcriptomic profile, which presents a higher percentage of response, respond to either IFX or VDZ treatment. Importantly, our approach actually allows a more accurate identification of those patients with UC1, in which 87% of the patients are refractory to both IFX and VDZ.

Next, we tested the robustness of our cross-species unbiased approach to classify UC1 and UC2 patients, using a data set in which a distinct IBD onset (pediatric) and methodology to obtain gene expression profiles (RNA-seq) was used33. Hence, using RNA-seq data from 206 pediatric UC patients (GSE109142)33 (Table 1), we were able to identify two major transcriptomic profiles, in which the enrichment of inflammatory genes distinguished the UC1 from the UC2 transcriptomic profile (Fig. 5d). Differentially expressed genes between UC1 and UC2 identified using microarray data (e.g., TREM1) were also observed in analysis from RNA-seq data (Fig. 5e). Moreover, we observed that 107 of the DEGs were shared between microarray and RNA-seq data sets (Supplementary Fig. 8a and Supplementary Data set 9), among which the key genes used to differentiate UC1 from UC2 were found (see Fig. 4). We further compared the p-values and the expression trends of those 107 DEGs found in both data sets. We observed that those genes are highly significant, regardless of the data set used (Supplementary Fig. 8b), and that 100% of them have the exact same expression
results indicate that the UC1 and UC2 identification (Supplementary Fig. 9a). However, pediatric UC1–calprotectin levels between UC1 and UC2 from the pediatric profile regardless of the methodology used to obtain gene expression IFX, while <37% of patients classified as UC1 (Fig. 5f) responded to IFX therapy in each group. 

Discussion
A systematic study demonstrated that biopsy sampling was the major source of inter-patient variability. Therefore, such technical variations can mask real biological differences, even though UC is known to present a high level of variability in macroscopic and endoscopic scoring among patients. To solve this, we limited the analysis to the relevant genes for inflammation, including the phases of tissue repair and regeneration. By using the key DEGs obtained by a mouse model of colitis, we were able to “ignore” genes that were highly variable between patients (e.g., as a result of technical variation), and focus only on those that contribute to inflammation. This allowed us to temporally assign IBD-risk genes and molecularly subclassify UC patients into two profiles: one of these characterized by genes involved in neutrophil recruitment, activation, and degranulation, and by low response to biologicals.

Different experimental models to study mucosal immune processes associated with the pathogenesis of UC are available. Among them, the DSS-induced colitis model is broadly used, due to its simplicity and applicability with different therapeutic drugs. Early studies characterized the temporal changes by qPCR for a handful of inflammatory markers, but how noninflammatory (i.e., repair-related) genes fluctuate over time during tissue repair was unknown. Others had previously performed a kinetic microarray analysis only during the acute inflammation phase of DSS (from days 0 to 6), but whether those genes continue to be expressed during tissue repair remained unclear. Moreover, although the DSS-induced colitis model has been extensively used for the study of UC, an open reference for gene expression during intestinal inflammation and tissue repair was still missing. Here, we used a time-series transcriptional characterization of colitis, which allowed us to identify which of the genes contribute to most of the nuances of inflammation over time. In addition, this paper provides an open data source that can be further investigated by others with different questions. As an example, we provided a temporal assignment of IBD-risk genes that might offer insight into their potential functions. Finally, our data show that the DSS mouse model is a relevant model for studying certain aspects of human UC.

Fig. 5 UC1 and UC2 differ on their response to IFX and VDZ therapy. a Individual patient response to IFX therapy in each group and the percentage of patients responding to IFX in each cohort. b t-SNE plot using the top variable genes identified from the mouse data set. Average expression of mouse homolog genes used to subdivide patients (left), where dark blue color indicates a higher average expression. Unsupervised hierarchical agglomerative clustering was used to automatically define patient subdivision (right). Dashed line delimits UC1 (triangle) and UC2 (circles) patients. c Individual patient response to VDZ therapy in each group and the percentage of patients responding to VDZ in the cohort (right). d t-SNE for data set GSE109 containing 206 pediatric UC patients plot, using the top variable genes identified from the mouse data set. Each point represents a patient sample. The average expression of mouse homolog genes used (left) to subdivide patients into UC1 or UC2 (right). Unsupervised hierarchical agglomerative clustering was used to automatically define patient subdivision (right). Dashed line delimits UC1 (triangle) and UC2 (circles) patients. e t-SNE overlay of the expression level of selected DEGs between UC1 and UC2, showing inter-patient variation in the pediatric UC cohort. f Individual patient response to IFX therapy in each group and the percentage of patients responding to IFX in each cohort.
Previous studies identified the molecular differences between responder and non-responder IBD patients\(^\text{13}\). These studies were purposely biased by an a priori knowledge of the responder and non-responder IBD samples. In contrast, we successfully classified the patients using a completely unsupervised approach, and therefore, we have potentially identified genes that go beyond the responsiveness to the therapy, by describing the molecular signature of the identified profiles. We were able to do this by using the key DEGs found in the mouse model of colitis, by “debiasing” the human analysis, by “ignoring” genes that were highly variable between patients, and by focusing only on those genes that contribute to inflammation. Consequently, we identified two transcriptomic profiles of ulcerative colitis (UC1 and UC2) based on all adult and pediatric patients.

While per definition, both UC1 and UC2 subpopulations are considered inflamed, only UC1 patients present a higher expression of genes associated with neutrophil degranulation and cytokine signaling, and only 10% of these patients responded to biological therapies. Similar to our results, others have shown that IL-6, IL11, IL13RA, STC1, and PTGS2 were downregulated in patients responsive to IFX\(^\text{13}\) (namely UC2 in our study). Another recent report showed that the gene OSM is upregulated in IBD patients compared with healthy controls and is predictive of anti-TNF responsiveness\(^\text{39}\). However, we did not find OSM as differentially expressed between UC1 and UC2 patients. For VDZ, however, a signature for prediction of response to therapy was still missing\(^\text{15}\). While UC1 and UC2 likely define two transient inflammatory states of the colon, the causes of origin and development of such a conserved signature remain unknown.

The identification of UC2, which is characterized by responsiveness to both IFX and VDZ, may have direct implications in the clinical setting. For example, it indicates that UC2 patients would benefit from a treatment with IFX only, since IFX therapy has a higher response rate\(^\text{6}\) and is more cost-effective compared with VDZ\(^\text{40}\). On the other hand, identification of non-responsiveness to both IFX and VDZ in the UC1 patient transcriptomic profile, suggests that another line of therapy should be applied. For example, we observed that the B-cell activation factor (TNFSF13B, protein BAFF) was to be found upregulated in UC1 patients. This suggests a potential role of B cells in UC1. Moreover, B cells are known to enhance inflammatory responses by cytokine secretion, such as TNF and IL-6\(^\text{41}\), which are also upregulated in UC1 patients. B-cell depletion using an anti-CD20 antibody in a small cohort showed a trend in reducing inflammation, although nonsignificant\(^\text{42}\). However, it remains possible that B-cell depletion might affect only UC1 patients, but not UC2. Similarly, we also observed that UC1 patients have a higher expression of genes involved in the JAK/STAT signaling pathway (PTP4A3, SOCS3) and cytokine signaling (IL6 and IL1B), suggesting a potential role of other therapies for this transcriptomic profile, such as canakinumab (anti-IL-6 mAb), siltuximab (anti-IL-1β mAb), JMS-053 (PTP4A3 inhibitor), and others might apply.

In summary, we have performed an unbiased characterization of the inflammatory and tissue repair processes, using a mouse colitis model, providing a useful resource for understanding colonic inflammation. Many of the genes identified in mice were also detected in human UC patients, thus allowing us to explore the temporal expression of IBD-risk genes during the course of inflammation and gain useful insights into their potential function. Furthermore, they allowed us to identify for the first time two clinically relevant molecular ulcerative colitis profiles (UC1 and UC2) in an unsupervised manner in both adult and pediatric UC patients. Thus, our methodology identified two molecularly distinct UC profiles and serves as a proof of concept for the use of transcriptomic data from highly controlled mice experiments to perform unsupervised and biologically driven analysis of highly variable human data sets.

### Methods

#### Mouse and induction of DSS colitis.

Animal experiments were done following institutional guidelines of the Stockholm Regional Ethics Committee under approved ethical permit number N89/15. Female 8–12-week-old C57BL/6J mice were obtained from Scanbur (Charles River, Germany) and housed in environmentally enriched ventilated cages under specific pathogen-free conditions (SPF) at the Astrid Fagaras laboratory (AFL, Karolinska Institutet) under 12-h light cycle and ad libitum access to drinking water and rat and mouse diet (RM100; Special Diet Services). For induction of colitis, 2.5% w/v dextran sulfate sodium (DSS; Assaymetrics) was supplemented in drinking water and given to mice for 7 consecutive days, with a change on day 3. After the treatment was ceased, mice returned to receive standard water. Mice were monitored everyday for alterations in body weight, disease activity index (DAI)\(^\text{43}\). Mice were anesthetized with isoflurane and killed for blood and tissue sampling.

#### Mouse gene expression by mRNA sequencing.

Colon samples were stored in RNAlater (Ambion) at −80°C until further use. Colon samples were homogenized using a bead-beating system (Precellys) for total RNA purification, using RNAasy kit (Qiagen) following the manufacturers’ recommendations. RNA purity and quantity were measured by NanoDrop spectrophotometer (ThermoFisher). All samples were screened for RNA integrity check and presented RIN values above 8 on 2100 Bioanalyzer instrument (Agilent). Samples were submitted to Novogene for library preparation, using TruSeq Stranded mRNA Library Prep Kit (poly-A selection) and sequencing using HiSeq-2500 platform (Illumina). Samples were aligned to single-end 50 bp sequence in the murine genome\(^\text{44}\), aligned using hisat\(^\text{2}\), trimming using Trimmmomatic\(^\text{45}\), and further proceeded for abundance estimation using Kallisto\(^\text{46}\), further data analysis was done in R programming language (Rstudio). Genes with an absolute read count <5 in at least three samples were considered with low expression and filtered out. Differences in tissue cell composition that could affect transcriptional pools were balanced by means of removing unwanted variation, based on negative control genes, using the RUVG function implemented in RUVseq package\(^\text{45}\). Analysis revealed that library sizes strongly correlated with several known intestinal housekeeping genes, such as Hprt (r = 0.87) and Gapdh (r = 0.85), but not Actb (r = 0.88). Moreover, genes such as Cd63 (0.94), Trappc (r = 0.97), and Cd63 (0.97) and Slc5a3 (r = 0.96) correlated even more strongly to the library sizes, indicating potentially novel housekeeping genes during colonic inflammation. Negative controls genes were thus defined as genes with a positive Pearson correlation above 0.9 to their respective sample library sizes. Estimated unwanted variation vectors were then used as covariates for calculation of differentially expressed genes (DEGs) using EdgeR package\(^\text{47}\). EdgeR is specialized in performing time-series differential expression by means of generalized linear model (glm) function\(^\text{27}\), where time points were parsed as independent factors in the contrast matrix, thus allowing detection of differentially expressed genes at any given time point. Genes were considered differentially expressed when the overall false discovery rate (FDR) <0.01 and at least one time point had fold change >1.5. DEGs identified in this manner were used for dimensionality reduction by principal component analysis (PCA), from which gene variance contribution to the total variation can be calculated. Identification of gene modules was done based on smoothed temporal expression curves\(^\text{48}\). Briefly, genewise log-fold changes were smoothed using spline curves and further grouped into modules by using inverse Pearson correlation as the distance for hierarchical agglomerative clustering with Ward’s method (“ward. D2”). Functional gene annotation was performed on each gene module individually, using the Gene Ontology (GO, Biological Process\(^\text{2017}\)) and the Kyoto Encyclopedia of Genes and Genomes (KEGG\(^\text{2016}\)) libraries with enrichR package\(^\text{49}\).

#### UC and IBD-risk gene mapping to the murine RNA-seq data set.

To identify which of the genes are shared between mouse and human ulcerative colitis, we compared the list of DEGs identified by the DSS data set and the list of genes identified by Taman et al\(^\text{26}\). Mapping of IBD-risk genes was done, using the list of IBD-risk genes identified in mice as genes with a positive Pearson correlation above 0.9 to their respective sample library sizes. Estimated unwanted variation vectors were then used as covariates for calculation of differentially expressed genes (DEGs) using EdgeR package\(^\text{47}\). EdgeR is specialized in performing time-series differential expression by means of generalized linear model (glm) function\(^\text{27}\), where time points were parsed as independent factors in the contrast matrix, thus allowing detection of differentially expressed genes at any given time point. Genes were considered differentially expressed when the overall false discovery rate (FDR) <0.01 and at least one time point had fold change >1.5. DEGs identified in this manner were used for dimensionality reduction by principal component analysis (PCA), from which gene variance contribution to the total variation can be calculated. Identification of gene modules was done based on smoothed temporal expression curves\(^\text{48}\). Briefly, genewise log-fold changes were smoothed using spline curves and further grouped into modules by using inverse Pearson correlation as the distance for hierarchical agglomerative clustering with Ward’s method (“ward. D2”). Functional gene annotation was performed on each gene module individually, using the Gene Ontology (GO, Biological Process\(^\text{2017}\)) and the Kyoto Encyclopedia of Genes and Genomes (KEGG\(^\text{2016}\)) libraries with enrichR package\(^\text{49}\).

#### Classification of adult UC molecular subtypes.

To investigate whether the molecular differences observed in the mouse model could also be found in humans, we made use of four human microarray data sets from GSE122513, GSE23359\(^\text{2}\), GSE23359\(^\text{2}\), and GSE58079\(^\text{4}\). Combining these datasets resulted in a gene expression and metadata of 447 patients, containing information such as disease type (UC or CD), Mayo macroscopic score, the therapy given, when the sample was collected, and the response to infliximab (IFX) or to vedolizumab (VDZ). Across all data sets, patients were considered in remission if presenting a Mayo score of 2 or 3 (out of 6). Similarly, patients were considered to respond to therapy, when the respective Mayo score reduced to 0 or 1, between 4 and 8 weeks of treatment with IFX or between 6 and 52 weeks of treatment with VDZ. For this
study, we included only patients with UC before receiving any therapy (either IFX or VDZ), comprising a total transcriptional profile of 143 patients, of which 102 received IFX and 41 for VDZ.

Probes with log2 fluorescence count lower than 6 in at least 10 samples were excluded from the analysis. Batches between the data set were observed and corrected using the ComBat solution for 24 h, and embedded in paraffin. Five-micrometer sections were stained with H&E, Ki67 (1:100, Cat# M3-14520, Thermo Scientific) staining was performed according to a previously published protocol57. A pathologist accessed the tissue pathological score in a blinded manner and scored the sections as previously described58.

### FITC-dextran assay

Assessment of the epithelial barrier integrity by FITC dextran was done as previously described59. Mice were gavaged with 10 mg/mL FITC-dextran (Molecular Probes, Eugene, OR) at 37°C 0–5 h before sacrifice. The following day, the opposite side of the abdomen was gavaged with 5 mg/mL FITC-dextran. Mice were sacrificed at 3 h post gavage, the abdomen was opened, and both the jejunum and ileum were removed. The jejunum was flushed with PBS and the ileum was dried and weighed. The FITC-dextran content was calculated by interpolation to 12-dilution FITC-dextran standard curve.

### Statistical analyses

Statistical analyses were performed using Prism Software 6.0 (GraphPad). Two-sample comparisons were compared using a two-tailed Student’s t test. ANOVA with Dunnett’s post hoc was used for calculation of significance at multiple time points relative to the control (day 0). Noncontinuous data were compared using a nonparametric Mann–Whitney U test. The results were considered significant when p < 0.05.

### Reporting summary

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

### Data availability

All the raw data generated in this study were deposited at the Gene Expression Omnibus under accession number GSE131032.

### Code availability

Codes used in this paper are available on Github (https://github.com/czarnewski/uu_classification).

Received: 9 January 2019 Accepted: 31 May 2019

Published online: 28 June 2019

---

### References

1. Magro, P. et al. Third European evidence-based consensus on diagnosis and management of ulcerative colitis. Part I: Definition, diagnosis, extra-intestinal manifestations, pregnancy, cancer surveillance, surgery, and fecal incontinence disorders. J. Crohns Colitis 11, 649–670 (2017).

2. Satsangi, J., Silverberg, M. S., Vermeire, S. & Colombel, J. F. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. Gut 55, 749–753 (2006).

3. Danese, S. New therapeutic approaches for inflammatory bowel disease: from the bench to the bedside. Gut 61, 918–932 (2012).

4. Rutgeerts, P. et al. Infliximab for induction and maintenance therapy for ulcerative colitis. New Engl. J. Med. 353, 2462–2476 (2005).

5. Feagan, B. G. et al. Vedolizumab as induction and maintenance therapy for ulcerative colitis. N. Engl. J. Med. 369, 699–710 (2013).
6. Paramsothy, S., Rosenstein, A. K., Mehandru, S. & Colombel, J. F. The current state of the art for biological therapies and new small molecules in inflammatory bowel disease. *Mucosal Immunol.* 11, 1558–1570 (2018).

7. Ordas, I., Eckmann, L., Talamin, M., Baumgart, D. C. & Sandborn, W. J. Ulcerative colitis. *Lancet* 380, 1606–1619 (2012).

8. Harbord, M. et al. Third European evidence-based consensus on diagnosis and management of ulcerative colitis. Part 2: Current management. *J. Crohns. Colitis.* 21, 1–36 (2008).

9. Graham, D. B. & Xavier, R. J. From genetics of inflammatory bowel disease toward mechanistic insights. *Trends Immunol.* 34, 371–378 (2013).

10. Huang, H. et al. Fine-mapping in inflammatory bowel disease loci to single-variant resolution. *Nature 547*, 173–178 (2017).

11. D’Haens, G. R. et al. The London Position Statement of the World Congress of Gastroenterology on Biological Therapy for IBD with the European Crohn’s and Colitis Organization: when to start, when to stop, which drug to choose, and how to predict response? *Am. J. Gastroenterol.* 106, 199–212 (2011).

12. Arijs, I. et al. Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. *PloS ONE* 4, e7984 (2009).

13. Arijs, I. et al. Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis. *Gut* 58, 1612–1619 (2009).

14. Toedter, G. et al. Gene expression profiling and response signatures associated with differential responses to infliximab treatment in ulcerative colitis. *Am. J. Gastroenterol.* 106, 199–212 (2011).

15. Arijs, I. et al. Effect of vedolizumab (anti-a4β7-integrin) therapy on histological healing and mucosal gene expression in patients with UC. *Gut* 67, 43–52 (2016).

16. Bezdek, J. C. & Hathaway, R. J. VAT: a Tool for Visual Assessment of Clustering Tendency. *Proc. 2002 Int. Joint Conference on Neural Networks, Colorado* 3, 2225–2230 (2002).

17. Ronan, T., Qi, Z. & Naege, K. M. Avoiding common pitfalls when clustering biological data. *Sci. Signal.* 9, re6 (2016).

18. D’Haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* 23, 1499–1501 (2005).

19. Adolfsen, A., Ackerman, M. & Brownstain, N. C. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognit.* 88, 13–26 (2019).

20. Hartigan, J. A. & Hartigan, P. M. The dip test of unimodality. *Ann. Stat.* 13, 70–84 (1985).

21. Suzuki, R. & Shimodaira, H. Pcvlusl: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542 (2006).

22. Bakay, M. et al. Sources of variability and effect of experimental approach on expression profiling data interpretation. * BMC Bioinforma.* 3, 4 (2002).

23. McIntryre, L. M. et al. RNA-seq: technical variability and sampling. *BMC Genom.* 12, 293 (2011).

24. Eichele, D. D. & Kharbanda, K. K. Dextran sodium sulfate colitis murine model: An inducible tool for advancing our understanding of inflammatory bowel diseases pathogenesis. *World J. Gastroenterol.* 23, 6016–6027 (2017).

25. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of DNA-seq quantification. *BMC Bioinforma.* 34, 525–527 (2019).

26. Riso, D., Ngi, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902 (2014).

27. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Biometrics* 26, 139–149 (2010).

28. Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S. & Simon, I. Continuous representations of time-series gene expression data. *J. Comput. Biol.* 10, 341–356 (2003).

29. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W92–W97 (2016).

30. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883 (2012).

31. Bartenhagen, C., Klein, H. U., Ruckert, C., Jiang, X. & Dugas, M. Comparative study of unsupervised dimension reduction techniques for the visualization of gene expression data. *BMC Genom.* 6016, 567–576 (2009).

32. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, Article 3 (2004).

33. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE* 12, e0190152 (2017).

34. Parigi, S. M. et al. Fli3 ligand expands bona fide innate lymphoid cell precursors in vivo. *Sci. Rep.* 8, 154 (2018).

35. Das, S. et al. Mice deficient in Mc4r are resistant to experimental colitis and colitis-associated colorectal cancer. *Oncogene* 35, 2645–2654 (2016).

36. Erben, U. et al. A guide to histomorphological evaluation of intestinal inflammation in mouse models. *Int. J. Clin. Exp. Pathol.* 7, 4557–4576 (2014).

37. Brandl, K. et al. Enhanced sensitivity to DSS colitis caused by a hypomorphic Mb1p1 mutation disrupting the ATF6-driven unfolded protein response. *Proc. Natl Acad. Sci. USA* 106, 3300–3305 (2009).

Acknowledgements

We would like to thank Stefan Bonn, Samuel Huber, and Charlotte Hedlin for critical reading and suggestions on the paper. We thank Elaine Hussey for editorial assistance.

EJF was supported by grants from the Swedish Research Council VR grant K2015-68X-22765-01-6, Formas grant nr. FR-20160005, and Wallenberg Academy Fellow (WAF) program.

Author contributions

P.C. and E.J.V. conceived the idea and wrote the paper. P.C. performed bioinformatics analysis and created schematic illustrations. N.G. and E.J.V. provided reagents and guidance. P.C., S.M.P., S.D. and O.E.D. performed the experiments. P.C., N.G. and E.J.V. analyzed and interpreted the data. All authors contributed to paper writing.
Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-10769-x.

Competing interests: The authors declare no competing interests.

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

Peer review information: Nature Communications would like to thank Lee Denson and other anonymous reviewers for their contributions to the peer review of this work. Peer review reports are available.

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