1 **Conserved transcriptomic profile between mouse and human colitis**

2 **allows temporal dynamic visualization of IBD-risk genes and**

3 **unsupervised patient stratification**

4

5 Paulo Czarnewski[1], Sara M. Parigi[1], Chiara Sorini[1], Oscar E. Diaz[1], Srustidhar Das[1], Nicola

6 Gagliani[1,2,§], Eduardo J. Villablanca[1,§,*]

7

8 [1] Immunology and Allergy Unit, department of Medicine, Solna, Karolinska Institute and

9 University Hospital, 17176 Stockholm, Sweden

10 [2] I. Department of General, Visceral and Thoracic Surgery / Department of Medicine,

11 University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany.

12

13 [§] These authors jointly supervised this work.

14 * Corresponding author

15 Eduardo J. Villablanca, PhD

16 Immunology and Allergy Unit, L2:04,

17 Karolinska Institutet,

18 Stockholm, SE17176

19 E-mail: eduardo.villablanca@ki.se (EJV)

Abstract

Despite the fact that ulcerative colitis (UC) patients show heterogeneous clinical manifestation and diverse response to biological therapies, all UC patients are classified as one group. Therefore, there is a lack of tailored therapies. In order to design these, an unsupervised molecular re-classification of UC patients is evoked. Classical clustering approaches based on tissue transcriptomic data were not able to classify UC patients into subgroups, likely due to associated covariates. In addition, while genome wide association studies (GWAS) have identified potential new target genes, their temporal dynamic revealing the optimal therapeutic window of time remains to be elucidated. To overcome the limitations, we generated time-series transcriptome data from a mouse model of colitis, which was then cross-compared with human datasets. This allowed us to visualize IBD-risk gene expression kinetics and reveal that the expression of the majority of IBD-risk genes peak during the inflammatory phase, and not the recovery phase. Moreover, by restricting the analysis to the most differentially expressed genes shared between mouse and human, we were able to cluster UC patients into two subgroups, termed UC1 and UC2. We found that UC1 patients expressed higher copy of genes involved in neutrophil recruitment, activation and degranulation compared to UC2. Of note, we found that over 87% of UC1 patients failed to respond to two of the most widely-used biological therapies for UC.

This study serves as a proof of concept that cross-species comparison of gene expression profiles enables the temporal annotation of disease-associated gene expression and the stratification of patients as of yet considered molecularly undistinguishable.

Introduction

Ulcerative colitis (UC) is a type of inflammatory bowel diseases (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 patient's quality of life. UC is recognized as a heterogeneous disease, presenting diverse macroscopic features, symptoms, grads of inflammation and colonic affected areas [1,2].

Although there is no definitive cure for UC, there are biological therapies available which target the inflammatory response during UC by means of inhibiting pro-inflammatory cytokines or by blocking immune cell migration [3]. 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% [4,6] and 50% [5,6] 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 intervention [7,8].

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 higher susceptibility to IBD [9,10]. However, the function and temporal expression of IBD risk genes during experimental colitis are yet to be elucidated [9,10].

Furthermore, while there is an obvious clinical heterogeneity among UC patients as seen for example by the location affected (i.e. distal colitis, left-sided and pancolitis, and responder and non-responder) and the extent of the severity, initial treatment for these patient subgroups is identical and modified only if the patients have not responded [6,8]. Biomarkers that could distinguish the different entities of the UC spectrum are currently lacking and they are required in order to achieve the highly needed stratification of UC patients into molecularly functional subgroups [8,11]. Moreover, an unbiased stratification of UC subtypes

69   has never been accomplished at the molecular and functional levels. Here, using

70   transcriptomic data from a well-characterized experimental model of colitis we were able to

71   identify conserved genes between mouse and UC patients. As a result, we were able to gain

72   insights into IBD-risk gene kinetics and to molecularly stratify UC patients in an unsupervised

73   manner.

74 **Results**

75 **Human UC is highly variable at the transcriptome level**

76 In order to molecularly stratify UC patients into subgroups, we combined 4 publicly available

77 human UC cohort datasets (n=102 patients), in which transcriptomic microarrays of total

78 colonic biopsies was performed [12-15] (**Table1** and **Fig S1**). We ranked genes using the top

79 100 most variable genes and further tested whether molecular subgroups exist (**Fig 1a**).

80 Analysis by visual assessment of cluster tendency (VAT)[16] indicated that biopsies presented

81 high inter-sample dissimilarities (**Fig1b**), suggesting a poor overall tendency to form

82 consistent clusters. Dimensionality reduction analysis by tSNE using the top highly variable

83 genes also indicates the formation of a single group with no apparent subdivisions (**Fig 1c**).

84 Then, we further statistically tested whether multi-cluster substructures were present in the

85 dataset, since most clustering algorithms define subgroups even on random noise [17-19].

86 However, bootstrapping analysis using the Hartigan's Dip test [19,20] presented a low cluster

87 substructure trend ($p > 0.9$), regardless of the gene ranking metrics used (**Fig 1d**).

88 Independently of the clustering tendency results, we forced patient subdivision using

89 hierarchical clustering and tested for cluster stability using bootstrapping [17,18,21]. In line with

90 previous results, formed clusters were highly unstable using the list of highly variable genes

91 (AU ≈ 0%) (**Fig 1e**). These results indicate that without prior knowledge of patient subdivision,

92 standard gene ranking strategies do not allow clustering of UC patients into molecularly

93 distinct subgroups.

94

95 **Time-series reveals processes underlying colon inflammation and repair**

96 One cause of such inter-patient variability can be attributed to the sampling procedure, which

97 contributes largely to the total data variance and masks real biological differences [22,23]. To

98 overcome the total data variance, we sought to identify the genes that contribute to

99 inflammation in an independent and unsupervised manner. To this end, we focused the

100 analysis on a list of evolutionarily conserved genes that best discriminate the nuances of

101 inflammation in a well-characterized colitis mouse model [24].

102 To identify these evolutionarily conserved genes, we first elucidated through an unbiased

103 manner which genes and pathways are differentially regulated during mouse colonic

104 inflammation followed by a tissue regeneration phase. In particular, we took advantage of the

105 widely used dextran sodium sulfate (DSS)-induced model of colitis. This model is one of the

106 few characterized by a phase of damage followed by a phase of regeneration. Therefore, this

107 model gave the possibility to identify also sets of genes essential in the regeneration phase, a

108 key step towards the resolution of the inflammation. In short, mice were exposed to DSS in

109 the drinking water for 7 days, then allowed to recover for the following 7 days. During this

110 period, we collected colonic tissue samples every second day to then be analyzed by RNA

111 sequencing (RNA-seq), histology and flow cytometry (**Fig 2a and Fig S2**). First, we confirmed

112 that 7 days of DSS exposure resulted in continuous body weight loss and acute disease

113 severity until day 10 to then initiate the recovery phase (**Fig S2a-b**). Histological analysis

114 confirmed epithelial damage, such as desquamation of the epithelial layer on day 6 (**Fig S2c**),

115 while labeling proliferating cells within crypts (Ki67 staining) indicated disrupted crypt

116 architecture by day 6 and restoration by d14 (**Fig S2c**). Loss of the epithelial cells

117 (CD45^{neg}EpCAM^{+}) by day 7-10 and restoration by day 14 was further confirmed by flow

118 cytometry (**Fig S2d**). To test whether the epithelial barrier integrity was restored by day 14,

119 we gavaged FITC-dextran and measured its concentration in the serum. We detected higher

120 FITC-dextran concentrations on day 7, which indicates barrier disruption, whereas basal

121 levels were detected by day 14 indicating restoration of the barrier integrity (**Fig S2e**). Thus,

122 on the basis of this characterization we will refer to d6-d10 and d12-d14 as acute phase and

123 recovery phase, respectively.

124 Next, we performed a RNA-seq analysis from colonic samples throughout the experiment and

125 computed differentially expressed genes (DEGs) taking the complete kinetics of expression

126 into consideration for p-value estimation using EdgeR [25] (see Methods). A detailed list of all

127 genes found differentially expressed is available for further exploration (**Table S1**). Principal

128 component analysis (PCA) on DEGs revealed that samples displayed a sequential temporal

129 path in PCA space, starting on day 0, passing through day 7 (acute) and ultimately reaching

130 day 14 (recovery) (**Fig 2b**). Of note, samples from day 14 did not reach the same gene

131 expression profile compared to day 0, suggesting that complete molecular restoration was not

132 reached by day 14. We observed that over 70% of the variance among the differentially

133 expressed transcripts is retained in the first 5 principal components (PCs) (**Fig S3a**), and that

134 each principal component corresponds to a unique expression kinetic through the time course

135 of DSS-colitis (**Fig S3b**). For instance, the variance explained by PC1 peaked at the acute

136 phase and returned to almost normal levels on day 14 (recovery), capturing most of the

137 variance related to inflammatory genes that peaked from days 7 to 10, such as *Ly6g*, *Reg3b*,

138 *Reg3g*, *S100a8*, *S100a9*, *Mmp3*, *Mmp8*, *Mmp10* (**Fig S3b and c**). On the other hand, the

139 variance explained by PC2 peaked on day 4 during DSS administration, to return close to

140 normal by day 7, thus, capturing most of the variance related to genes expressed during

141 initiation of inflammation, such as *Mcpt1*, *Mcpt2*, *Mmp3*, *Mmp10*, *Il11*, *Scnn1g* and *Best2* (**Fig**

142 **S3b and c**). These results indicate that several of the genes modulated between days 4-10

143 are related to inflammation and together contribute the most to the variance in the dataset.

144

145 By using hierarchical clustering on the spline smoothed gene expression of DEGs, we were

146 able to classify the gene expression into 9 modules (**Fig 2c**). For further exploration,

147 expression values for all genes in each module are available (**Table S1)**. Three gene modules

148 (m2, m7 and m8) were down regulated during the acute and recovery phases of DSS-induced

149 inflammation, with lowest peak on days 6, 10 and 12, respectively. GO and KEGG enrichment

150 analysis suggest that these modules represent genes mainly involved with epithelial cell

151 functions, such as PPAR signaling (*Acsl1*, *Fabp1*), small molecule metabolism (*Sult1a1*,

152 *Sult1b1*) and fat digestion and absorption (*Paqr8*, *Clps*, *Pla2g3*) (**Fig 2c and Fig S4a**).

153 On the other hand, six modules (m9, m3, m1, m4, m6 and m5) were up-regulated over the

154 early, acute and recovery phases of DSS-induced inflammation, peaking on days 2, 6, 7, 10,

155 12 and 14, respectively. Among those, processes such as cytokine signaling (*Il11*, *Il12b*, *Il6*,

156 *Il1b*), leukocyte migration (*Sell*, *Ccr1*, *Ccr2*, *Cxcl2*, *Cxcr3*), neutrophil degranulation (*Ly6g*,

157　　*Itgam*, *Itgax*, *Cd300a*), matrix remodeling (*Mmp3*, *Mmp7*, *Mmp10*), response to

158　　lipopolysaccharide (*Saa3*, *Nox2*) as well as several inflammatory signaling pathways (*Stat3*,

159　　*Jak3*, *Nfkbia*, *Smad4*, *Birc3*) were enriched, suggesting the interplay of several immune cells

160　　and pathways as a cause/trigger of inflammation, especially during the acute phase (**Fig 2c**

161　　**and Fig S4b**). Moreover, modules m9 and m5 presented two degrees of bimodal expression

162　　pattern, peaking at day 2-4 (early phase), with slight down-regulation between days 7-10 and

163　　a second peak on days 12-14 (recovery phase). Genes in those modules were associated

164　　mainly with cell cycle (*Ttk*, *Cdc7*, *Cdc20*, *Cdc25c*, *Ccna2*, *Ccnb1*, *Ccnb2*) and cholesterol

165　　biosynthetic pathways (*Acat2*, *Sqle*, *Mvd*, *Hmgcs1*), respectively (**Fig 2c and Fig S4b**). Many

166　　other genes and GO/KEGG pathways not shown here are fully accessible for exploration of

167　　individual genes and their clusters (**Table S1, S2 and S3**). Taken together, time-series

168　　transcriptomic characterization of mouse colonic inflammation identifies distinct gene

169　　expression kinetics associated with epithelial and immune cell related pathways during the

170　　course of colitis.

171

172　　**Inflammatory pathways are the most conserved between mouse and human colitis**

173　　Having characterized genes and pathways that are associated with intestinal inflammation

174　　and tissue repair during experimental colitis, we investigated whether such pathways are

175　　conserved in humans. To this end, we compared the list of DEGs from the mouse

176　　experimental colitis with the recently published list of DEGs found in newly diagnosed

177　　treatment-naïve ulcerative colitis patients [26]. This is a cohort containing human RNA-seq data,

178　　where they report DEGs between UC patients versus healthy controls. We found that among

179　　the 4045 mouse DEG, 650 genes were also found among the list of DEG obtained comparing

180　　UC patients versus healthy controls (**Fig 2d and TableS4**). Out of the 650 genes shared

181　　between mouse and humans, 53.9% were identified in the inflammatory modules m1 (28.2%),

182　　m3 (14.2%) and m4 (11.5%) (**Fig 2d**). This suggests that acute inflammatory genes in m1,

183　　m3 and m4 are conserved between DSS-induced colitis and UC. GO and KEGG enrichment

184　　analysis revealed that those 650 genes were enriched for inflammatory pathways related to

185 neutrophil degranulation and chemotaxis, as well as cytokine and inflammatory signaling

186 pathways (**Fig 2e** and **TableS5**). These results showed that most of the genes/pathways

187 conserved between experimental mouse colitis and human UC are associated with

188 inflammatory responses.

189

190 **Forward translation from mouse to human UC patients allows the temporal**

191 **classification of the IBD risk genes**

192 To understand the temporal expression of the genes associated with the identified IBD

193 polymorphisms (candidate IBD risk genes) [9], we checked the expression of genes associated

194 with UC or CD identified by single variant fine-mapping resolution [10] into the list of DEGs from

195 the mouse dataset. Out of the 233 reported candidate IBD risk genes, 40 genes presented

196 very low or undetectable counts in the mouse dataset (i.e., *IL23R*, *SULT1A2*, *ERAP2*,

197 *MUC19*), 118 were detected but did not have their expression altered through the

198 development of inflammation (i.e., *TNFRSF14*, *ATG16L1*, *GPR35*, *TNFSF8*) and 75 were

199 found among the DEGs in our mouse dataset (**Fig S5a** and **Table S6**). Among these, many

200 IBD-risk genes with already known functions during mouse colitis were found (e.g. *IFNG*,

201 *GPR65*, *ITGAL*, *CCL7*, *STAT3*, *FUT1*, *CD40*, *SULT1A1*, *MUC1*, *CARD9*, *IL12B*, *IRF1*, *CD5*),

202 being specifically present in gene modules related to inflammation m1, m3 and m4. Moreover,

203 26 genes of the 75 IBD risk genes found in our dataset are shared between UC and CD (i.e.

204 *CARD9*, *SULT1A1*, *STAT3*, *GPR65*, *IL12B*), while 10 and 39 were restricted to UC or CD,

205 respectively (**Fig S5b** and **Table S7**). In order to provide temporal information regarding the

206 expression of IBD-risk genes during inflammation and repair, we utilized the mouse

207 transcriptional landscape to map at which time point homolog IBD risk-genes were up- or

208 down-regulated. Out of the 75 genes shared between mouse DEGs and IBD risk genes, 45

209 (60%) were mapped to modules m1, m3 and m4, which represent the acute phase of

210 inflammation (**Fig S5c** and **Table S7**). Among them we found *Card9, Ifng, Il12b, Stat3, Stat4,*

211 *Cd40,* which have been reported to exert functions during the acute phase of intestinal

212 inflammation [27-32]. By contrast, *Fut1, Sult1a1, Hes5 and Tnfsf15* were mapped to modules

213   m8, m7 and m2, which are down-regulated during acute inflammation, while *Rasip, Ntn5 and*

214   *Rtel1* matched with module 6 which is associated with genes that are up-regulated during the

215   recovery phase after acute inflammation (**Fig S5c**). These data thus provided temporal

216   information on when IBD risk genes are differentially expressed during damage and tissue

217   repair, providing useful insights into their potential roles during inflammation and recovery.

218

219   **Key conserved inflammatory genes distinguish two human ulcerative colitis**

220   **subgroups**

221   Having identified genes that contribute to inflammatory pathways that are conserved between

222   mice and humans, we next used those genes to assess whether UC patients can be

223   subdivided into subgroups (**Table1**, **Fig 3a**). To this end, we selected the top 100 leading

224   genes in PC1 and PC2 from the mouse colitis dataset and identified the respective human

225   homologs (**Fig 3a**). We found that 57 genes were shared between mice and humans. Of

226   these, only 17 genes were found among the 100 most variable genes of the human dataset

227   (**Fig S6**), which might explains why patient classification using highly variable genes was not

228   possible.

229   Therefore, we performed an unsupervised analysis of the human dataset using the 57

230   homolog genes (**Fig 3a**). Of note, VAT analysis using these 57 homolog genes indicated the

231   distinction into 2 major patient subgroups (**Fig 3b**), which also resulted in reduced Hartigan's

232   unimodality test ($p < 0.001$, **Fig 3c**). This indicates that by using mouse most variable genes

233   as opposed to the sole top human variable ones, it is possible to obtain higher clustering

234   tendency of the UC patient data. To test whether using the mouse homologs also impacted

235   on cluster stability, we performed a bootstrapping analysis. This time, clustering using the top

236   mouse homolog genes resulted in clusters with higher stability (AU ≈ 80%) (**Fig 3d**),

237   compared to using the top human highly variable genes (AU ≈ 0%) (**Fig 1e**). Hierarchical

238   agglomerative clustering using the mouse homolog genes thus defined 2 UC subtypes,

239   namely UC1 and UC2, comprising 60 and 42 patients, respectively (**Fig 3e**). The UC1

240   subgroup is defined as patients presenting the higher average expression of the inflammatory

241   genes compared to UC2 (**Fig 3f**). We also observed that neither UC1 nor UC2 subtypes were

242   discriminated by the overall macroscopic disease severity (**Fig 3g**), suggesting that although

243   these two UC subtypes are indistinguishable based on Mayo score, they are transcriptionally

244   distinct.

245

246   **UC1 and UC2 are transcriptionally distinct**

247   In order to characterize UC1 and UC2 beyond conserved genes, we performed differential

248   expression analysis using all genes present in the human dataset. We were able to identify

249   205 highly differentially expressed genes, among which 187 were up-regulated in UC1 and 18

250   were up-regulated in UC2 (**Fig 4a**). Detailed tables with information on all DEGs comparing

251   UC1 and UC2 are available for exploration (**Table S8 and Fig S7a**). Among those, cytokines

252   (*TNF*, *IL11*), enzymes (*NOX1*, *MMP3*, *CYP26B1*), calcium-binding proteins (*S100A8*,

253   *S100A9*), chemokines (*TREM1*, *CXCL8*) and other proteins related to the inflammatory

254   response (*NR3C2, BCL2A1, PARM1, TNFSF13B*) were clearly able to discriminate UC1 from

255   UC2 (**Fig 4b and Fig S7**). Enrichment analysis for cell types, GO, and KEGG pathways

256   revealed that genes highly expressed in UC1 (187) were associated with terms related to

257   neutrophil, neutrophil degranulation and cytokine-cytokine receptor interaction, respectively

258   (**Fig S7b**). Venn diagram of the top enriched terms revealed many overlapping genes are

259   shared among these pathways (**Fig 4c**), suggesting that UC1 patients present a distinct

260   transcriptional signature enriched in neutrophil activity and cytokine signaling compared to

261   UC2 patients.

262   We trained a logistic regression classifier using each of the DEGs between UC1 and UC2 to

263   identify key genes that could be further used in the clinics for distinction of UC1 and UC2.

264   Genes were tested and scored individually using the area under the curve (AUC) as a

265   combined measure of sensitivity and specificity (**Fig 4d**). We observed that genes such as

266   *TREM1* (AUC=99%), *CYP26B1* (AUC=97%) and *CXCL8* (AUC=97%) were among the top

267   markers to distinguish UC1 from UC2. Other genes such as *WNT5*, *BCL2A1*, *C5AR1*, *MMP1*,

268   *MMP3* and *IL11* also presented AUC scores above 90% and also represented good

269   candidates for UC1 and UC2 distinction in clinical practice.

270

271   **UC1 and UC2 respond differently to biological therapies**

272   While we stratified UC patients into two molecularly distinct subgroups, it was unclear whether

273   UC1 and UC2 show different treatment responses to biological therapies. To address this, we

274   used the patient-specific treatment response obtained 4 to 8 weeks after the biopsy was

275   taken and treatment with IFX started (**Table 1**). Interestingly, we observed that on average,

276   70% of the patients belonging to the subtype UC2 responded to infliximab therapy (**Fig 5a**) in

277   contrast to less than 10% of the patients classified as UC1, regardless of the dataset

278   analyzed (**Fig 5a**).

279   To extend the applicability of our findings, we made use of another set of UC patients

280   receiving vedolizumab and repeated the same procedure as before (**Table 1**). Transcriptomic

281   data from UC patients were analyzed using the most relevant genes identified in our mouse

282   colitis model and then clustered as described above to reveal UC1 and UC2. Between them,

283   UC1 presented a higher expression of the conserved inflammatory genes (**Fig 5b**). We

284   observed that about 60% of the patients belonging to the UC2 subgroup responded to VDZ, in

285   comparison to about 13% of the patients belonging to the UC1 subgroup (**Fig 5c**). Taken

286   together, the data indicates that patients belonging to the UC2 subgroup, which present a

287   higher percentage of response, respond to either IFX or VDZ treatment. Importantly, our

288   approach actually allows a more accurate identification of those patients with UC1, in which

289   87% of the patients are refractory to both IFX and VDZ.

290

291   **Discussion**

292   A systematic study demonstrated that biopsy sampling was the major source of inter-patient

293   variability[22]. Therefore, such technical variations can mask real biological differences, even

294   though UC is known to present a high level of variability in macroscopic and endoscopic

295   scoring among patients[1,2,8]. To solve this, we limited the analysis to the relevant genes for

296   inflammation including the phases of tissue repair and regeneration. By using the key DEGs

297   obtained by a mouse model of colitis, we were able to "ignore" genes that were highly variable

298   between patients (e.g. as a result of technical variation), and focus only on those that

299   contribute to inflammation. This allowed us to temporally classify IBD risk genes and

300   molecularly sub-classify UC patients into two subgroups; one of these characterized by genes

301   involved in neutrophil recruitment, activation and degranulation, and by low response to

302   biologicals.

303   Different experimental models to study mucosal immune processes associated with the

304   pathogenesis of UC are available [33,34]. Among them, the DSS-induced colitis model is broadly

305   used due to its simplicity and applicability with different therapeutic drugs [35]. Early studies

306   characterized the temporal changes by qPCR for a handful of inflammatory markers [36], but

307   how non-inflammatory (i.e. repair-related) genes fluctuate over time during tissue repair was

308   unknown. Others had previously performed a kinetic microarray analysis only during the acute

309   inflammation phase of DSS (from days 0 to 6)[37], but whether those genes continue to be

310   expressed during tissue repair remained unclear. Moreover, although the DSS-induced colitis

311   model has been extensively used for the study of UC, an open reference for gene expression

312   during intestinal inflammation and tissue repair was still missing. Here, we used a time-series

313   transcriptional characterization of colitis, which allowed us to identify which genes contribute

314   to most of the nuances of inflammation over time. In addition, this manuscript provides an

315   open data source that can be further investigated by others with different questions. As an

316   example, we provided a temporal assignment of IBD risk genes that might offer insight into

317   their potential functions. Finally, our data show that the DSS mouse model is a relevant model

318   for studying certain aspects of human UC.

319   Previous studies identified the molecular differences among responder and non-responder

320   IBD patients [13]. These studies were purposely biased by an *a priori* knowledge of the

321   responder and non-responder IBD samples. In contrast, we successfully classified the

322   patients using a completely unsupervised approach and therefore, we have potentially

323   identified genes that go beyond the responsiveness to the therapy by describing the

324 molecular signature of the identified subgroups. We were able to do this by using the key

325 DEGs found in the mouse model of colitis, by "debiasing" the human analysis, by "ignoring"

326 genes that were highly variable between patients, and by focusing only on those genes that

327 contribute to inflammation. Consequently, we identified two subpopulations of UC patients

328 (UC1 and UC2).

329 While per definition both UC1 and UC2 subpopulations are considered inflamed, only UC1

330 patients present higher expression of genes associated with neutrophil degranulation and

331 cytokine signaling, and only 10% of these patients responded to biological therapies. Similar

332 to our results, others have shown that IL6, IL11, IL13RA, STC1 and PTGS2 were down-

333 regulated in patients responsive to IFX [13] (namely UC2 in our study). Another recent report

334 showed that the gene OSM is up-regulated in IBD patients compared to healthy controls and

335 is predictive of anti-TNF responsiveness [38]. However, we did not find OSM as differentially

336 expressed between UC1 and UC2 patients. For VDZ, however, a signature for prediction of

337 response to therapy was still missing [15].

338 The identification of UC2 which is characterized by responsiveness to both IFX and VDZ may

339 have direct implications in the clinical setting. For example, it indicates that UC2 patients

340 would benefit from a treatment with IFX only, since IFX therapy has a higher response rate [6]

341 and is more cost-effective compared to VDZ [39]. On the other hand, identification of non-

342 responsiveness to both IFX and VDZ in the UC1 patient subgroup, suggests that another line

343 of therapy should be applied. For example, we observed that the B cell activation factor

344 (*TNFSF13B*, protein BAFF) was to be found up-regulated in UC1 patients. This suggests a

345 potential role of B cells in UC1. Moreover, B cells are known to enhance inflammatory

346 responses by cytokine secretion such as TNF and IL-6 [40], which are also up-regulated in UC1

347 patients. B cell depletion using anti-CD20 antibody in a small cohort showed a trend in

348 reducing inflammation, although non-significant [41]. However, it remains possible that B cell

349 depletion might affect only UC1 patients, but not UC2. Similarly, we also observed that UC1

350 patients have a higher expression of genes involved in the JAK/STAT signaling pathway

351 (*PTP4A3*, *SOCS3*) and cytokine signaling (*IL6* and *IL1B*), suggesting a potential role of other

352   therapies for this subgroup, such as canakinumab (anti IL-6 mAb), siltuximab (anti IL-1β

353   mAb), JMS-053 (PTP4A3 inhibitor) and others might apply.

354   In summary, we have performed an unbiased characterization of the inflammatory and tissue

355   repair processes using a mouse colitis model, providing a useful resource for understanding

356   colonic inflammation. Many of the genes identified in mice were also detected in human UC

357   patients, thus allowing us to explore the temporal expression of IBD risk genes during the

358   course of inflammation and gain useful insights into their potential function. Furthermore, they

359   allowed us to identify for the first time two clinically relevant molecular ulcerative colitis

360   subsets (UC1 and UC2) in an unsupervised manner. Thus, our methodology identified two

361   molecularly distinct UC subgroups and will serve as a proof of concept for the use of

362   transcriptomic data from highly controlled mice experiments to perform unsupervised and

363   biologically-driven analysis of highly variable human datasets.

364

Methods

366   All methods used in this paper are described in the Online Methods linked to this manuscript.

367

Acknowledgements

373

Author Contributions

375   PC and EJV conceived the idea and wrote the paper. PC performed bioinformatics analysis

376   and schematic illustrations. NG and EJV provided reagents and guidance. PC, SMP, SD, CS

377   and OED performed the experiments. PC, NG and EJV analyzed and interpreted the data. All

378   authors contributed to manuscript writing.

References

1   Magro, F. et al. Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders. Journal of Crohn's & colitis 11, 649-670, doi:10.1093/ecco-jcc/jjx008 (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, doi:10.1136/gut.2005.082909 (2006).

3   Danese, S. New therapies for inflammatory bowel disease: from the bench to the bedside. Gut 61, 918-932, doi:10.1136/gutjnl-2011-300904 (2012).

4   Rutgeerts, P. et al. Infliximab for induction and maintenance therapy for ulcerative colitis. The New England journal of medicine 353, 2462-2476, doi:10.1056/NEJMoa050516 (2005).

5   Feagan, B. G. et al. Vedolizumab as induction and maintenance therapy for ulcerative colitis. The New England journal of medicine 369, 699-710, doi:10.1056/NEJMoa1215734 (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 immunology, doi:10.1038/s41385-018-0050-3 (2018).

7   Ordas, I., Eckmann, L., Talamini, M., Baumgart, D. C. & Sandborn, W. J. Ulcerative colitis. Lancet 380, 1606-1619, doi:10.1016/S0140-6736(12)60150-0 (2012).

8   Harbord, M. et al. Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 2: Current Management. Journal of Crohn's & colitis 11, 769-784, doi:10.1093/ecco-jcc/jjx009 (2017).

9   Graham, D. B. & Xavier, R. J. From genetics of inflammatory bowel disease towards mechanistic insights. Trends in immunology 34, 371-378, doi:10.1016/j.it.2013.04.001 (2013).

10   Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 547, 173-178, doi:10.1038/nature22969 (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? The American journal of gastroenterology 106, 199-212; quiz 213, doi:10.1038/ajg.2010.392 (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, doi:10.1371/journal.pone.0007984 (2009).

13   Arijs, I. et al. Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis. Gut 58, 1612-1619, doi:10.1136/gut.2009.178665 (2009).

14   Toedter, G. et al. Gene expression profiling and response signatures associated with differential responses to infliximab treatment in ulcerative colitis. The American journal of gastroenterology 106, 1272-1280, doi:10.1038/ajg.2011.83 (2011).

15   Arijs, I. et al. Effect of vedolizumab (anti-alpha4beta7-integrin) therapy on histological healing and mucosal gene expression in patients with UC. Gut 67, 43-52, doi:10.1136/gutjnl-2016-312293 (2016).

16   Bezdek, J. C. & Hathaway, R. J. VAT: a tool for visual assessment of (cluster) tendency. 2225-2230, doi:10.1109/ijcnn.2002.1007487 (2002).

427   17   Ronan, T., Qi, Z. & Naegle, K. M. Avoiding common pitfalls when clustering biological
428   data. *Science signaling* **9**, re6, doi:10.1126/scisignal.aad1932 (2016).

429   18   D'Haeseleer, P. How does gene expression clustering work? *Nature biotechnology*
430   **23**, 1499-1501, doi:10.1038/nbt1205-1499 (2005).

431   19   Adolfson, A. A., M.; Brownstain, N. C. *To Cluster, or Not to Cluster: How to Answer*
432   *the Question* (2017).

433   20   Hartigan, J. A. & Hartigan, P. M. The Dip Test of Unimodality. *The Annals of Statistics*
434   **13**, 70-84, doi:10.1214/aos/1176346577 (1985).

435   21   Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in
436   hierarchical clustering. *Bioinformatics* **22**, 1540-1542,
437   doi:10.1093/bioinformatics/btl117 (2006).

438   22   Bakay, M. *et al.* Sources of variability and effect of experimental approach on
439   expression profiling data interpretation. *BMC bioinformatics* **3**, 4 (2002).

440   23   McIntyre, L. M. *et al.* RNA-seq: technical variability and sampling. *BMC genomics* **12**,
441   293, doi:10.1186/1471-2164-12-293 (2011).

442   24   Eichele, D. D. & Kharbanda, K. K. Dextran sodium sulfate colitis murine model: An
443   indispensable tool for advancing our understanding of inflammatory bowel diseases
444   pathogenesis. *World journal of gastroenterology* **23**, 6016-6029,
445   doi:10.3748/wjg.v23.i33.6016 (2017).

446   25   McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of
447   multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids*
448   *research* **40**, 4288-4297, doi:10.1093/nar/gks042 (2012).

449   26   Taman, H. *et al.* Transcriptomic Landscape of Treatment-Naive Ulcerative Colitis.
450   *Journal of Crohn's & colitis* **12**, 327-336, doi:10.1093/ecco-jcc/jjx139 (2018).

451   27   Lamas, B. *et al.* CARD9 impacts colitis by altering gut microbiota metabolism of
452   tryptophan into aryl hydrocarbon receptor ligands. *Nature medicine* **22**, 598-605,
453   doi:10.1038/nm.4102 (2016).

454   28   Ito, R. *et al.* Interferon-gamma is causatively involved in experimental inflammatory
455   bowel disease in mice. *Clinical and experimental immunology* **146**, 330-338,
456   doi:10.1111/j.1365-2249.2006.03214.x (2006).

457   29   Kobayashi, T. *et al.* NFIL3-deficient mice develop microbiota-dependent, IL-12/23-
458   driven spontaneous colitis. *Journal of immunology* **192**, 1918-1927,
459   doi:10.4049/jimmunol.1301819 (2014).

460   30   Xu, J. *et al.* Stat4 is critical for the balance between Th17 cells and regulatory T cells
461   in colitis. *Journal of immunology* **186**, 6597-6606, doi:10.4049/jimmunol.1004074
462   (2011).

463   31   Liu, J. *et al.* Chronic inflammation up-regulates P-gp in peripheral mononuclear blood
464   cells via the STAT3/Nf-kappab pathway in 2,4,6-trinitrobenzene sulfonic acid-induced
465   colitis mice. *Scientific reports* **5**, 13558, doi:10.1038/srep13558 (2015).

466   32   Borcherding, F. *et al.* The CD40-CD40L pathway contributes to the proinflammatory
467   function of intestinal epithelial cells in inflammatory bowel disease. *The American*
468   *journal of pathology* **176**, 1816-1827, doi:10.2353/ajpath.2010.090461 (2010).

469   33   Kiesler, P., Fuss, I. J. & Strober, W. Experimental Models of Inflammatory Bowel
470   Diseases. *Cellular and molecular gastroenterology and hepatology* **1**, 154-170,
471   doi:10.1016/j.jcmgh.2015.01.006 (2015).

472   34   Wirtz, S. *et al.* Chemically induced mouse models of acute and chronic intestinal
473   inflammation. *Nature protocols* **12**, 1295-1309, doi:10.1038/nprot.2017.044 (2017).

474   35   Melgar, S. *et al.* Validation of murine dextran sulfate sodium-induced colitis using four
475   therapeutic agents for human inflammatory bowel disease. *International*
476   *immunopharmacology* **8**, 836-844, doi:10.1016/j.intimp.2008.01.036 (2008).

477   36   Yan, Y. *et al.* Temporal and spatial analysis of clinical and molecular parameters in
478   dextran sodium sulfate induced colitis. *PloS one* **4**, e6073,
479   doi:10.1371/journal.pone.0006073 (2009).

480   37   Fang, K. *et al.* Temporal genomewide expression profiling of DSS colitis reveals
481   novel inflammatory and angiogenesis genes similar to ulcerative colitis. *Physiological*
482   *genomics* **43**, 43-56, doi:10.1152/physiolgenomics.00138.2010 (2011).

483   38   West, N. R. *et al.* Oncostatin M drives intestinal inflammation and predicts response
484   to tumor necrosis factor-neutralizing therapy in patients with inflammatory bowel
485   disease. *Nature medicine* **23**, 579-589, doi:10.1038/nm.4307 (2017).

486   39   Yokomizo, L., Limketkai, B. & Park, K. T. Cost-effectiveness of adalimumab,
487   infliximab or vedolizumab as first-line biological therapy in moderate-to-severe
488   ulcerative colitis. *BMJ open gastroenterology* **3**, e000093, doi:10.1136/bmjgast-2016-
489   000093 (2016).

490   40   Shen, P. & Fillatreau, S. Antibody-independent functions of B cells: a focus on
491   cytokines. *Nature reviews. Immunology* **15**, 441-451, doi:10.1038/nri3857 (2015).

492   41   Leiper, K. *et al.* Randomised placebo-controlled trial of rituximab (anti-CD20) in active
493   ulcerative colitis. *Gut* **60**, 1520-1526, doi:10.1136/gut.2010.225482 (2011).

494

495

496

497 **Figure legends**

498 **Figure 1. Human ulcerative colitis transcriptional profiles cluster patients into**

499 **molecular subgroups.** (**a**) Schematic representation of the strategy used for patient group

500 identification, in which four publicly available datasets were combined. Gene ranking was

501 done using the most variable genes in the human dataset, which were used for clustering

502 analysis. (**b**) Sample dissimilarity heatmaps for visual analysis of clustering tendency (VAT),

503 comparing the human dataset using the top 100 variable genes. (**c**) tSNE plot using the top

504 100 variable genes in the human dataset. Each point represents a patient sample. (**d**)

505 Hartigan's Dip test for clustering tendency using all genes in the dataset, the top 100 variable

506 genes, the top 100 highly dispersed genes or the top 100 leading genes in the principal

507 components. (**e**) Bootstrapping analysis of hierarchical clustering, comparing the human

508 dataset using the top 100 variable genes in the human dataset. Numbers in orange indicate

509 the approximately unbiased (AU) p-value, shown as a percentage. AU closer to zero indicates

510 a cluster with low stability.

511

512 **Figure 2. Unbiased characterization of the DSS colitis reveals conserved inflammatory**

513 **signature between mice and humans.** (**a**) Schematic illustration of the experimental design.

514 Mice received DSS in drinking water for 7 days, after which the treatment was replaced with

515 water. Samples were collected at indicated time points. (**b**) PCA on differentially expressed

516 gene counts. Samples were color-coded according to their respective day of collection (from

517 grey to orange to blue). The percentage of variance explained by the respective principal

518 component indicated in parenthesis. (**c**) Clustered heatmap of all differentially expressed

519 genes (left). The mean expression of each gene module is shown (right). Functional

520 annotation of genes in each cluster was done based on Gene Ontology (GO) enrichment.

521 Only the top 3 enriched processes are shown, sorted by P-value. (**d**) Venn diagram

522 comparing the list of DEGs in treatment-naive UC and the DEGs identified in mouse DSS

523 colitis (upper). Among the 650 genes shared among those lists (in red), the number and

524 percentage of genes found in each module identified in our mouse dataset (lower). Modules

525 highlighted in bold are the ones enriched for inflammatory terms in **Fig 2c**. (**e**) GO and KEGG

526 enrichment analysis out of the 650 shared genes identified in (**d**), sorted by P-value.

527

528 **Figure 3. Conserved inflammatory gene signature distinguishes two UC subgroups.** (**a**)

529 Schematic representation of the strategy used for patient group identification. Four publicly

530 available datasets were combined. Gene ranking was done using the most variable genes

531 identified mouse dataset that had a homolog in humans. (**b**) Sample dissimilarity heatmaps

532 for visual analysis of clustering tendency (VAT), comparing the human dataset using the top

533 mouse gene homologs. (**c**) Hartigan's Dip test for clustering tendency comparing the analysis

534 using top 100 variable genes and the top mouse gene homologs. (**d**) Bootstrapping analysis

535 of hierarchical clustering, comparing the human dataset using the top mouse gene homologs.

536 Numbers in orange indicate the approximately unbiased (AU) p-value, shown as percentage.

537 AU closer to zero indicates a cluster with low stability. (**e**) tSNE plot using the top variable

538 genes identified from the mouse dataset. Each point represents a patient sample. tSNE plot

539 showing the separation of 2 patient subgroups (left). Unsupervised hierarchical agglomerative

540 clustering was used to automatically define patient subdivision (center). Dashed line delimits

541 UC1 (triangle) and UC2 (circles) patients. (**f**) Average expression of mouse homolog genes

542 used to subdivide patients (right), where dark blue colour indicates higher average

543 expression. Dashed line delimits UC1 (triangle) and UC2 (circles) patients. (**g**) Assessment of

544 Mayo clinical subscore in patients from UC1 and UC2. Mann-Whitney test was used for

545 comparison.

546

547 **Figure 4. UC1 subgroup is enriched for the inflammatory signature.** (**a**) Heatmap of

548 DEGs between UC1 and UC2 patients including all genes in the human dataset. Only the

549 selected genes are shown, grouped by functional categories and respective to the expression

550 level. (**b**) tSNE overlay of the expression level of selected DEGs between UC1 and UC2,

551 showing inter-patient variation. (**c**) Venn diagram of the top GO, KEGG and cell enriched

552 terms identified from the DEGs between UC1 and UC2. (**d**) Top 20 genes ranked by area

553 under the curve (AUC) for specificity and sensitivity to distinguish UC1 from UC2, among the

554 list of DEGs (left). Classification was carried out using logistic regression. The fitted values of

555 prediction are shown for selected genes (right).

556

557 **Figure 5. UC1 and UC2 differ on their repose to IFX and VDZ therapy.** (**a**) Individual

558 patient response to IFX therapy in each group and the percentage of patients responding to

559 IFX in each cohort. (**b**) tSNE plot using the top variable genes identified from the mouse

560 dataset. Average expression of mouse homolog genes used to subdivide patients (left),

561 where dark blue colour indicates higher average expression. Unsupervised hierarchical

562 agglomerative clustering was used to automatically define patient subdivision (right). Dashed

563 line delimits UC1 (triangle) and UC2 (circles) patients. (**c**) Individual patient response to VDZ

564 therapy in each group and the percentage of patients responding to VDZ in the cohort (right).

565

566 ## Supplementary figure legends

567 **Figure S1. Normalization of publicly available ulcerative colitis datasets.** (**a**)

568 Multidimensional scaling plots before and after batch effect correction using ComBat. (**b**)

569 Relative log expression plots comparing samples from the different datasets before and after

570 adjusting for batches using ComBat.

571

572 **Figure S2. Macroscopic alterations in mice during DSS-induced colitis.** (**a**) Body weight

573 change over the time course of colitis. *$P < 0.05$; two-way ANOVA. (**b**) Disease activity index

574 score (DAI) over time (in arbitrary units, A.U.). *$P < 0.05$; two-way ANOVA. (**c**) Representative

575 histological section of the colonic tissue at indicated time points. H&E (upper) and

576 immunohistochemistry staining for Ki-67 (bottom) are depicted. One representative figure out

577 of three experiments. Scale bar 50 μm.

578 (**d**) Flow cytometry data showing colonic epithelial cell ($EpCAM^+CD45^-$) frequencies during the

579 course of the experiment. Dot plots are representative of three experiments. The graph on the

580   right shows epithelial cell absolute numbers during the course of the experiment. *p < 0.05;

581   two-way ANOVA. (**e**) Quantification of intestinal permeability by FITC-dextran assay. Mice

582   were gavaged with 10 mg/mL of FITC-dextran and sacrificed 4 hours later for quantification of

583   fluorescence in the serum. *$p < 0.05$; two-way ANOVA. Error bars represent SEM.

584

585   **Figure S3. Identification of top leading genes and that drive overall differences in gene**

586   **expression during DSS colitis.** (**a**) Percentage of variance explained by each principal

587   component (see Fig 2b). (**b**) Overall fluctuations in the first 6 PCs over the time course of DSS

588   colitis. Note that the overall variance captured by PC6 is close to 0 and therefore not used in

589   further analysis. (**c**) Ranking of the top 20 leading genes that contribute to the variance in

590   each of the first 5 PCs.

591

592   **Figure S4. List of the top DEGs per module.** (**a**) Down-regulated gene modules m8, m7

593   and m2 (see Fig 2c). (**b**) Up-regulated gene modules m9, m1, m3, m4, m5 and m6 (see Fig

594   2c).

595

596   **Figure S5. Mouse colitis and human UC share inflammatory pathways and IBD risk**

597   **genes.** (**a**) Venn diagram comparing IBD risk genes and the list of DEGs in the mouse

598   dataset (upper). 75 genes are shared between these lists (in red). The number and

599   percentage out of the 75 IBD-risk genes presented in each mouse module is shown (below).

600   Modules highlighted in bold are the ones enriched for inflammatory terms in **Fig 2c**. (**b**) Venn

601   diagram for the genes in the list of DEGs in the mouse dataset, genes associated with UC

602   and/or to CD. Among those, 26 are shared between UC and CD (in red). (**c**) Expression level

603   of IBD risk gene mouse homologs during the DSS colitis.

604

605   **Figure S6. List of highly variable genes in humans and mouse colitis.** (**a**) Top 100 genes

606   sorted by high variance in the human dataset. Genes highlighted in red are also present

607   among the top list of homolog genes identified in the mouse colitis dataset. (**b**) Top list of

608   homolog genes identified in the mouse colitis dataset, sorted by variance on the human

609   dataset.

610

611   **Figure S7. List of DEGs between UC1 and UC2 and enrichment analysis**. (**a**) List of the

612   top 32 up-regulated and 16 down-regulated DEGs between UC1 and UC2. (**b**) Cell, GO and

613   KEGG enrichment analysis for the genes up-regulated in UC1 compared to UC2.

614

615   ## Tables

Table 1. Publicly available human datasets used for classification of ulcerative colitis subtypes. Only the number of patients used for analysis are shown (inflamed mucosa before receiving any therapy).

Dataset ID	Total	Resp	Non-resp	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]

616

Online Methods

21

22

1.1. Mice and induction of DSS colitis

23

24 Female 8-12 weeks old C57BL/6J mice were obtained from ScanBur (Charles River,

25 Germany) and housed in environmentally enriched ventilated cages under specific pathogen

26 free conditions (SPF) at Astrid Fagræus laboratory (AFL, Karolinska Institutet) under 12h light

27 cycle and receiving water and ration *ad libitum* (RM1(P), Special Diet Services). For induction

28 of colitis, 2.5% w/v dextran sulfate sodium (DSS; Affymetrics) was supplemented in drinking

29 water and given to mice for 7 consecutive days, with a change on day 3. After the treatment

30 was ceased, mice returned to receive standard water. Mice were monitored everyday for

31 alterations in body weight, disease activity index (DAI) [1]. Mice were anesthetized with

32 isofluorane and sacrificed for blood and tissue sampling. Animal experiments were done

33 following institutional guidelines of the Stockholm Regional Ethics Committee under approved

34 ethical permit number N89/15.

35

1.2. Mouse gene expression by mRNA sequencing

36

37 Colon samples were stored in RNAlater (Ambion) at -80˚C until further use. Colonic samples

38 were homogenized using bead-beating system (Precellys) for total RNA purification using

39 RNAeasy kit (Qiagen) following manufacturers recommendations. RNA purity and quantity

40 was measured by NanoDrop spectrophotometer (ThermoFisher). All samples were screened

41 for RNA integrity check and presented RIN values above 8 on 2100 Bioanalyzer instrument

42 (Agilent). Samples were submitted to Novogene for library preparation using TruSeq Stranded

43 mRNA Library Prep Kit (poly-A selection) and sequencing using HiSeq-2500 platform

44 (Illumina). Samples were sequenced using single-end 50bp sequencing[2], aiming an coverage

45 of 20M reads. Read quality was inspected using MultiQC[3], trimmed with Trimmimatic[4] and

46 further proceeded for abundance estimation using Kallisto[5].

47   Further data analysis was done in R programming language (Rstudio). Genes with absolute

48   read count less than 5 in at least 3 samples were considered with low expression and filtered

49   out. Differences in tissue cell composition that could affect transcriptional pools were

50   balanced by means of removing unwanted variation based on negative control genes using

51   the RUVg function implemented in RUVseq package[6]. Analysis revealed that library sizes

52   strongly correlated with several known intestinal housekeeping genes, such as *Hprt* (r=0.87)

53   and *Gapdh* (r=0.85), but not *Actb* (r=0.68). Moreover, genes such as *Cd63* (0.94), *Trappc*

54   (r=0.97), and *Cpped1* (0.97) and *Slc25a3* (r=0.96) correlated even more strongly to the library

55   sizes, indicating potentially novel housekeeping genes during colonic inflammation. Negative

56   controls genes were thus defined as genes with positive Pearson correlation above 0.9 to

57   their respective sample library sizes. Estimated unwanted variation vectors were then used as

58   covariates for calculation of differentially expressed genes (DEGs) using EdgeR package[7].

59   EdgeR is specialized in performing time-series differential expression by means of

60   generalized linear model (glm) function[8], where time points were parsed as independent

61   factors in the contrast matrix, thus allowing detection of differentially expressed genes at any

62   given time point. Genes were considered differentially expressed when the overall false

63   discovery rate (FDR) < 0.01 and at least one time-point had fold change > 1.5. DEGs

64   identified in this manner were used for dimensionality reduction by principal component

65   analysis (PCA), from which gene-wise contribution to the total variation can be calculated.

66   Identification of gene modules was done based on smoothed temporal expression curves[9].

67   Briefly, gene-wise log fold changes were smoothed using spline curves and further grouped

68   into modules by using Pearson correlation as distance for hierarchical agglomerative

69   clustering with Ward's method ("ward.D2"). Functional gene annotation was performed on

70   each gene module individually using the Gene Ontology (GO_Biological_Process_2017) and

71   the Kyoto Encyclopedia of Genes and Genomes (KEGG_2016) libraries with enrichR

72   package[10].

73

74   **1.3. Mapping treatment-naïve ulcerative colitis and IBD risk genes to the murine**

75   **RNA-seq dataset**

76   To identify which genes are shared between mouse and human ulcerative colitis, we

77   compared the list of DEGs identified by in the DSS dataset and the list of genes identified by

78   Taman et al.[11]. Mapping of IBD risk genes was done using the list of IBD risk genes identified

79   by fine-mapping at the single loci resolution [12]. Identification of enriched GO processes and

80   KEGG pathways was done using enrichR[10].

81

82   **1.4. Classification of ulcerative colitis molecular subtypes using genes in mouse**

83   **principal components**

84   To investigate whether the nuances of inflammation observed in the mouse model could also

85   be found in humans, we made use of four human microarray datasets from GSE12251[13],

86   GSE73661[14], GSE23597[15] and GSE16879[16]. Combined, these datasets contain gene

87   expression and metadata of 447 patients, containing information such as disease type (UC or

88   CD), Mayo macroscopic score, the therapy given, when the sample was collected and the

89   response to infliximab (IFX) or to vedolizumab (VDZ). Across all datasets, patients were

90   considered inflamed if presenting a Mayo score of 2 or 3 (out of 3). Similarly, patient were

91   considered to respond to therapy when it respective Mayo score reduced to 0 or 1, between

92   4-8 weeks of treatment with IFX or between 6-52 weeks of treatment with VDZ. For this study,

93   we included only patients with UC before receiving any therapy (either IFX or VDZ),

94   comprising a total transcriptional profiles of 143 patients, of which 102 received IFX and 41 for

95   VDZ. The list of samples used in this study is supplied as metadata table (**Table S9**).

96   Probes with log2 fluorescence count lower than 6 in at least 10 samples were excluded from

97   the analysis. Batches between dataset were observed and corrected using the ComBat

98   function in SVA package[17]. Selection of genes for further exploration was done by different

99   approaches: 1) using all genes; 2) using only the top 100 highly variable genes; 3) using the

100   genes with top 100 high dispersion; 3) The gene with high loading in principal component 1

101   and; 4) The gene with high loading in principal component 2.

102   We determined whether clustering patterns exist by 4 independent methods: 1) By

103   dimensionality reduction using tSNE. Since data originated from biopsies are known to

104   present high variability across patients[18], dimensionality reduction and visualization was done

105   using t-Stochastic neighbor embedding (t-SNE). Because of it's nonlinear characteristics, t-

106   SNE becomes less sensitive to noise and outperform PCA[19] to discriminate biopsies based

107   on shared expression patterns, rather than their absolute expression values.; 2) By visual

108   assessment of clustering tendency (VAT) using dissimilarity matrices[20]; 3) By using the

109   Hartigan's dip test[21,22], which tests whether the gene distribution are different to an unimodal

110   distribution. Values close to 1 indicate that the data is unlikely to present cluster

111   substructures. We performed bootstrapping 100 times on 90% of the samples to calculate

112   Hartigan's dip test p-value. The comparison between bootstrapping with human highly

113   variable genes and mouse PCs (see below) was done using paired Mann-Whitney test; 4) By

114   dividing patients into subgroups using hierarchical agglomerative clustering. Cluster stability

115   was determined by bootstrapping 300 times on 90% or the samples, resulting in the

116   approximate unbiased (AU) statistics[23]. Clusters with AU closer to 100 present higher

117   stability.

118   Instead of using the top variable genes as above, we alternatively used the top genes

119   identified in the mouse RNA-seq DSS colitis dataset (see above). To this end, the top 100

120   genes identified in PC1 and PC2 were selected for identification of the respective human

121   homologs. Together, 175 genes were found in top genes in both PC1 and PC2 and from

122   these, 148 genes had a homolog in humans. In total, 57 homolog genes were found between

123   our mouse PCs and the human dataset. Dimensionality reduction was performed with tSNE.

124   Assessment of clustering tendency was done as described above. Agglomerative clustering

125   on the Euclidean distance using complete linkage was used to discriminate patient subgroups

126   UC1 and UC2. For the matter of definition used in this study, patients that present higher

127   mean expression of the 57 mouse-human homologs were classified as UC1, while those with

128   low expression were classified as UC2. Differences in expression between UC1 and UC2

129   were calculated using eBayes method in limma package[24]. Probes with fold changes above

130   1.5 and FDR lower than 0.001 were considered significantly differentially expressed.

131   Identification of enriched GO, KEGG and cell types was done using enrichR[10].

132   To identify which genes can discern UC1 from UC2, we trained a logistic regression classifier

133   for each gene individually and comparing to the UC1 and UC2 classification mentioned

134   above. The sensibility and sensitivity of the prediction was summarized using the area under

135   the curve (AUC) method. Genes with AUC values closer to 1 (100%) have a better accuracy

136   to distinguish UC1 and UC2 patients.

137

138   **1.5. Lamina propria cell isolation for analysis by flow cytometry**

139   Cell isolation from the colonic tissue was performed as previously described [26] with

140   modifications. Briefly, tissues were open longitudinally, cut into 1cm pieces and washed with

141   PBS. The epithelial cell fraction was obtained by incubating the tissue with Buffer-A (PBS, 5%

142   FCS, 5 mM EDTA) at 37°C for 20 minutes under agitation at 600 rpm. The supernatant was

143   collected and kept on ice while the remaining tissue was washed 2 times with PBS. Tissue

144   were digested with collagenase solution containing 0.15 mg/ml Liberase TL (Roche) and

145   0.1 mg/ml DNase I (Roche) in HBSS and incubated at 37°C for 60 minutes under agitation at

146   1200 rpm. The digested and the epithelial cell fraction were mixed, filtered through a 100 um

147   cell strainer, pelleted by centrifugation at 1750 rpm and re-suspended in Buffer-A. Cell

148   suspensions were blocked with Fc-blocking solution (1:1000, eBioscience) and stained with

149   the antibody mix (1:200), both at 4°for 15 minutes. The following antibodies were purchased

150   from BD Biosciences: CD45.2 (104), CD3 (500A2), CD90.2 (53-2.1), EPCAM (G8.8), CD11b

151   (M1/70), CD11c (N418), Ly6G (1A8), B220 (RA3-6B2) and CD64 (54-5/7.1). The following

152   antibodies were purchased from eBiosciences: CD103 (2E7) and Ly6C (HK1.4). Counting

153   beads (Spherotech) and DAPI (1:400, Sigma) were added to each sample to allow absolute

154   cell quantification and exclusion of dead cells. Data acquisition was done using 5-laser LSR

155   Fortessa flow cytometer (BD Biosciences) and analysis was carried out with FlowJo software

156   (TreeStar).

157

158   **1.6. Histological analyses**

159   The colonic tissue was rinsed and flushed with PBS and gently squeezed out to remove non-

160   adherent bacteria, fixed in 4% formaldehyde solution for 24 h and embedded in paraffin. 5 um

161   sections were stained with H&E. Ki67 (1:100, Cat# MA5-14520, Thermo Scientific) staining

162   was performed according to previously published protocol {26364605}. A pathologist

163   accessed the tissue pathological score in a blind manner and score the sections as previously

164   described[27].

165

166   **1.7. FITC-dextran assay**

167   Assessment of epithelial barrier integrity was done as previously described. Mice were

168   gavaged with 10 mg/mL FITC-dextran (Sigma) at different time points of DSS colitis, but on

169   the same day of sacrifice. Four hours later, mice were killed and the blood collected for

170   analysis. Sera were diluted 1:1 v/v in PBS and added to a 96-well plate for fluorescent-based

171   assays (Invitrogen) and were quantified on a fluorescent plate reader using a 535/587nm

172   ex/em filter. FITC-dextran concentration was calculated by interpolation to 12-dilution FITC-

173   dextran standard curve.

174

175   **1.8. Statistical analyses**

176   Statistical analyses were performed using Prism Software 6.0 (GraphPad). Two-sample

177   comparisons were compared using two-tailed Student's *t*-test. ANOVA with Dunnett's *post-

178   hoc* was used for calculation of significance at multiple time points relative to the control (day

179   0). Non-continuous data was compared using non-parametric Mann-Whitney U test. Results

180   were considered significant when $p < 0.05$.

181

182   **1.9. Data availability**

183   All the raw data generated in this study will be deposited in a suitable database (i.e., Gene

184   Expression Omnibus) upon acceptance of this manuscript.

185

186

187   **References**

188   1   Kim, J. J., Shajib, M. S., Manocha, M. M. & Khan, W. I. Investigating intestinal inflammation
189   in DSS-induced model of IBD. *Journal of visualized experiments : JoVE*, doi:10.3791/3678
190   (2012).

191   2   Chhangawala, S., Rudy, G., Mason, C. E. & Rosenfeld, J. A. The impact of read length on
192   quantification of differentially expressed genes and splice junction detection. *Genome biology*
193   **16**, 131, doi:10.1186/s13059-015-0697-y (2015).

194   3   Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for
195   multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048,
196   doi:10.1093/bioinformatics/btw354 (2016).

197   4   Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
198   sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

199   5   Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
200   quantification. *Nature biotechnology* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

201   6   Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor
202   analysis of control genes or samples. *Nature biotechnology* **32**, 896-902, doi:10.1038/nbt.2931
203   (2014).

204   7   Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
205   differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140,
206   doi:10.1093/bioinformatics/btp616 (2010).

207   8   McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor
208   RNA-Seq experiments with respect to biological variation. *Nucleic acids research* **40**, 4288-
209   4297, doi:10.1093/nar/gks042 (2012).

210   9   Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S. & Simon, I. Continuous
211   representations of time-series gene expression data. *Journal of computational biology : a*
212   *journal of computational molecular cell biology* **10**, 341-356,
213   doi:10.1089/10665270360688057 (2003).

214   10   Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server
215   2016 update. *Nucleic acids research* **44**, W90-97, doi:10.1093/nar/gkw377 (2016).

216   11   Taman, H. *et al.* Transcriptomic Landscape of Treatment-Naive Ulcerative Colitis. *Journal of*
217   *Crohn's & colitis* **12**, 327-336, doi:10.1093/ecco-jcc/jjx139 (2018).

218   12   Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution.
219   *Nature* **547**, 173-178, doi:10.1038/nature22969 (2017).

220   13   Arijs, I. *et al.* Mucosal gene signatures to predict response to infliximab in patients with
221   ulcerative colitis. *Gut* **58**, 1612-1619, doi:10.1136/gut.2009.178665 (2009).

222   14   Arijs, I. *et al.* Effect of vedolizumab (anti-alpha4beta7-integrin) therapy on histological
223   healing and mucosal gene expression in patients with UC. *Gut* **67**, 43-52, doi:10.1136/gutjnl-
224   2016-312293 (2016).

225   15   Arijs, I. *et al.* Mucosal gene expression of antimicrobial peptides in inflammatory bowel
226   disease before and after first infliximab treatment. *PloS one* **4**, e7984,
227   doi:10.1371/journal.pone.0007984 (2009).

228   16   Toedter, G. *et al.* Gene expression profiling and response signatures associated with
229   differential responses to infliximab treatment in ulcerative colitis. *The American journal of*
230   *gastroenterology* **106**, 1272-1280, doi:10.1038/ajg.2011.83 (2011).

231   17   Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for
232   removing batch effects and other unwanted variation in high-throughput experiments.
233   *Bioinformatics* **28**, 882-883, doi:10.1093/bioinformatics/bts034 (2012).

234   18   Bakay, M. *et al.* Sources of variability and effect of experimental approach on expression
235   profiling data interpretation. *BMC bioinformatics* **3**, 4 (2002).

236   19   Bartenhagen, C., Klein, H. U., Ruckert, C., Jiang, X. & Dugas, M. Comparative study of
237   unsupervised dimension reduction techniques for the visualization of microarray gene
238   expression data. *BMC bioinformatics* **11**, 567, doi:10.1186/1471-2105-11-567 (2010).

239   20   Bezdek, J. C. & Hathaway, R. J. VAT: a tool for visual assessment of (cluster) tendency.
240   2225-2230, doi:10.1109/ijcnn.2002.1007487 (2002).

241   21   Adolfson, A. A., M.; Brownstain, N. C. *To Cluster, or Not to Cluster: How to Answer the*
242   *Question* (2017).

243   22   Hartigan, J. A. & Hartigan, P. M. The Dip Test of Unimodality. *The Annals of Statistics* **13**,
244   70-84, doi:10.1214/aos/1176346577 (1985).

245   23   Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in
246   hierarchical clustering. *Bioinformatics* **22**, 1540-1542, doi:10.1093/bioinformatics/btl117
247   (2006).

248   24   Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression
249   in microarray experiments. *Statistical applications in genetics and molecular biology* **3**,
250   Article3, doi:10.2202/1544-6115.1027 (2004).

251   25   Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time
252   quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408,
253   doi:10.1006/meth.2001.1262 (2001).

254   26   Parigi, S. M. *et al.* Flt3 ligand expands bona fide innate lymphoid cell precursors in vivo.
255   *Scientific reports* **8**, 154, doi:10.1038/s41598-017-18283-0 (2018).

256   27   Erben, U. *et al.* A guide to histomorphological evaluation of intestinal inflammation in mouse
257   models. *International journal of clinical and experimental pathology* **7**, 4557-4576 (2014).
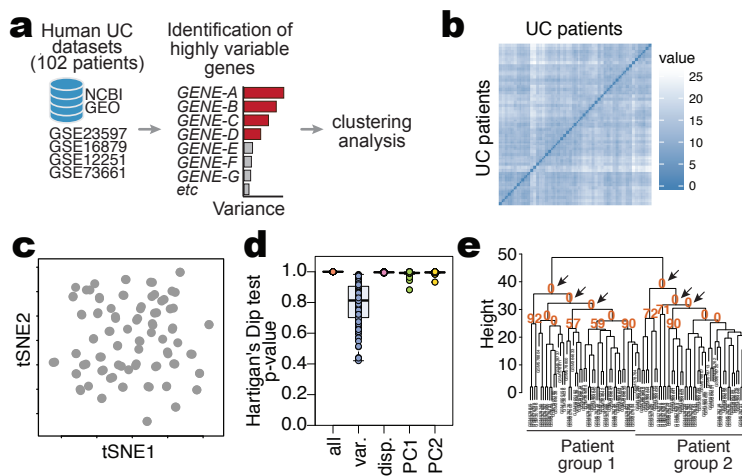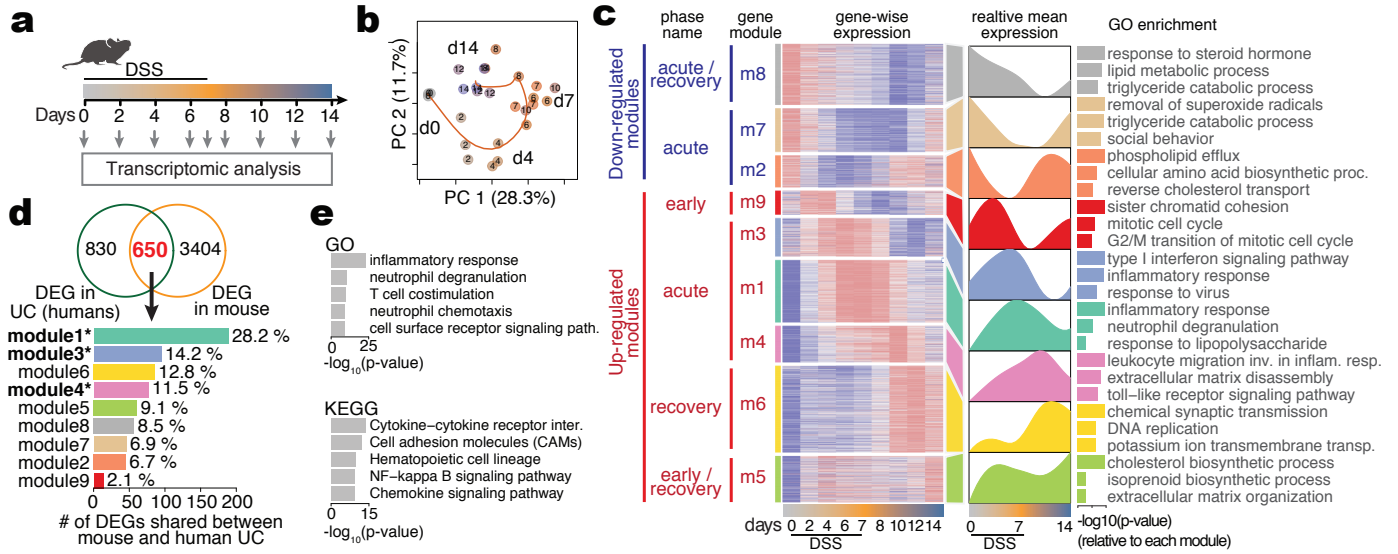
258

Figure 1

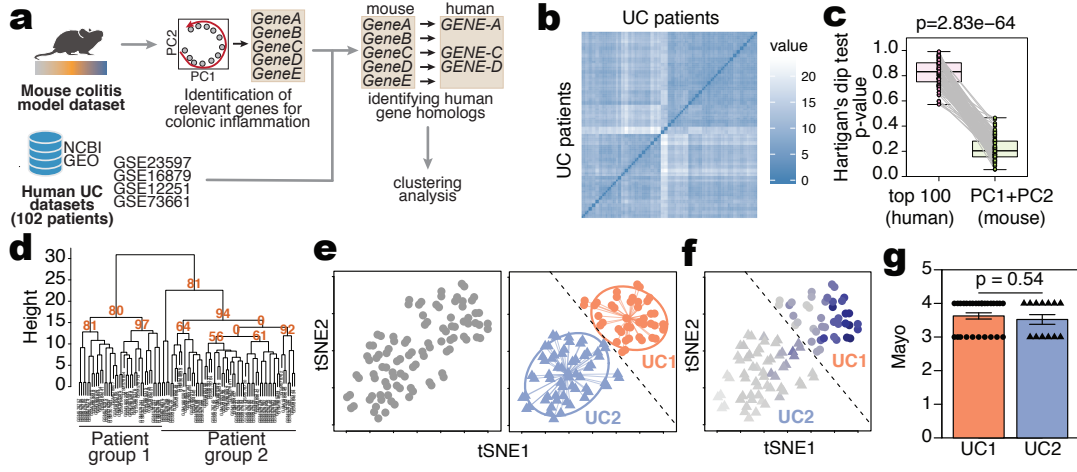**a** Human UC datasets (102 patients)

NCBI
GEO

GSE23597
GSE16879
GSE12251
GSE73661

Identification of highly variable genes

GENE-A
GENE-B
GENE-C
GENE-D
GENE-E
GENE-F
GENE-G
etc

Variance

→ clustering analysis

b UC patients

UC patients

value
25
20
15
10
5
0

c

tSNE2

tSNE1

d

Hartigan's Dip test p-value

1.0
0.8
0.6
0.4
0.2
0.0

all var. disp. PC1 PC2

e

Height

50
40
30
20
10
0

Patient group 1   Patient group 2

Figure 2

Figure 3

a

Mouse colitis model dataset

Identification of relevant genes for colonic inflammation

mouse	→	human
GeneA	→	GENE-A
GeneB	→	
GeneC	→	GENE-C
GeneD	→	GENE-D
GeneE	→	

NCBI GEO

Human UC datasets (102 patients)

GSE23597
GSE16879
GSE12251
GSE73661

identifying human gene homologs

clustering analysis

b UC patients

c p=2.83e−64

top 100 (human)   PC1+PC2 (mouse)

Hartigan's dip test p-value

d

Height

Patient group 1   Patient group 2

e tSNE2 / tSNE1   UC1   UC2

f tSNE2 / tSNE1   UC1   UC2

g p = 0.54   Mayo   UC1   UC2

Figure 4

Figure 5

a Response to IFX

Responders

tSNE2 / tSNE1

UC1 / UC2

% response to IFX

p = 0.039

■ GSE12251
● GSE16879
▲ GSE23597
◆ GSE73661

UC1 UC2

b Top genes in PC1 and PC2 (mouse homologs)

tSNE2 / tSNE1

UC1 / UC2

c Response to VDZ

Responders

tSNE2 / tSNE1

UC1 / UC2

% response to VDZ

◆ GSE73661

UC1 UC2

Figure S1

Figure S2

a Body weight change

H$_2$O
DSS

% body weight

DSS | water

days after the start of DSS treatment

b DAI score

A.U.

DSS
H$_2$O

DSS | water

days after the start of DSS treatment

c

d0 | d6 | d14

H&E

Ki-67

d

d0 | d7 | d10 | d14

22.0 | 23.7 | 16.2 | 21.0

EpCAM

CD45

# of EPCAM+ cells (million)
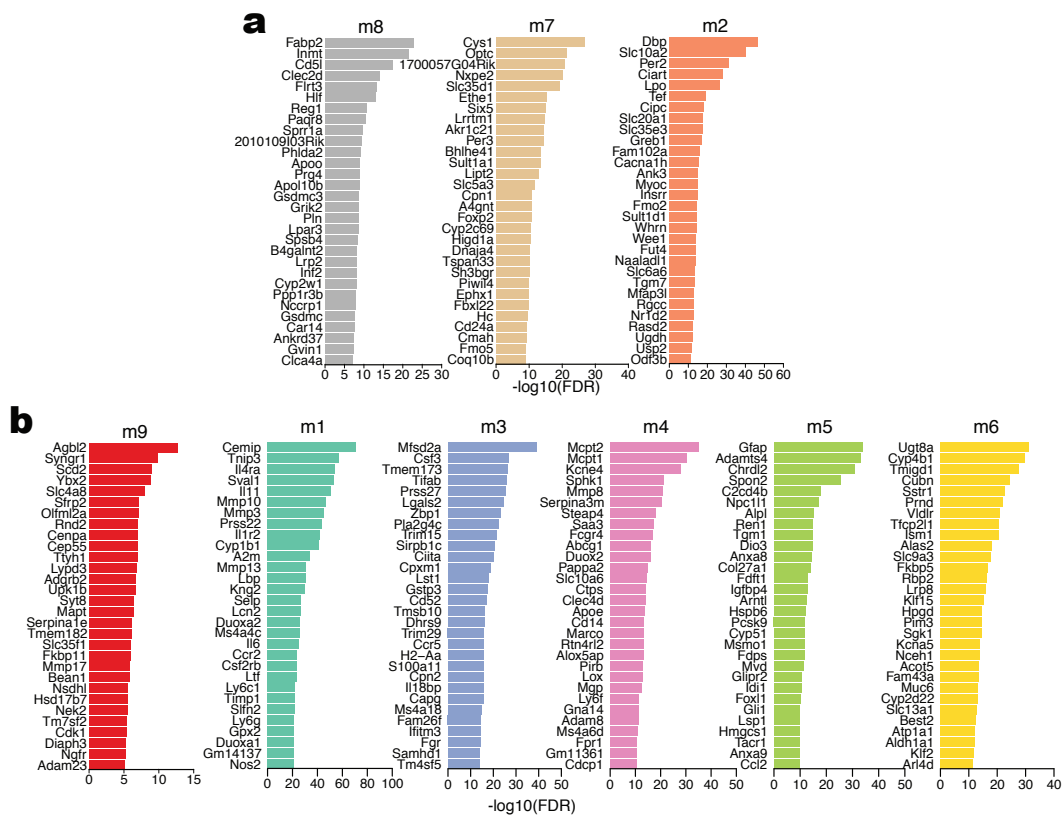
d0 | d7 | d10 | d14

e

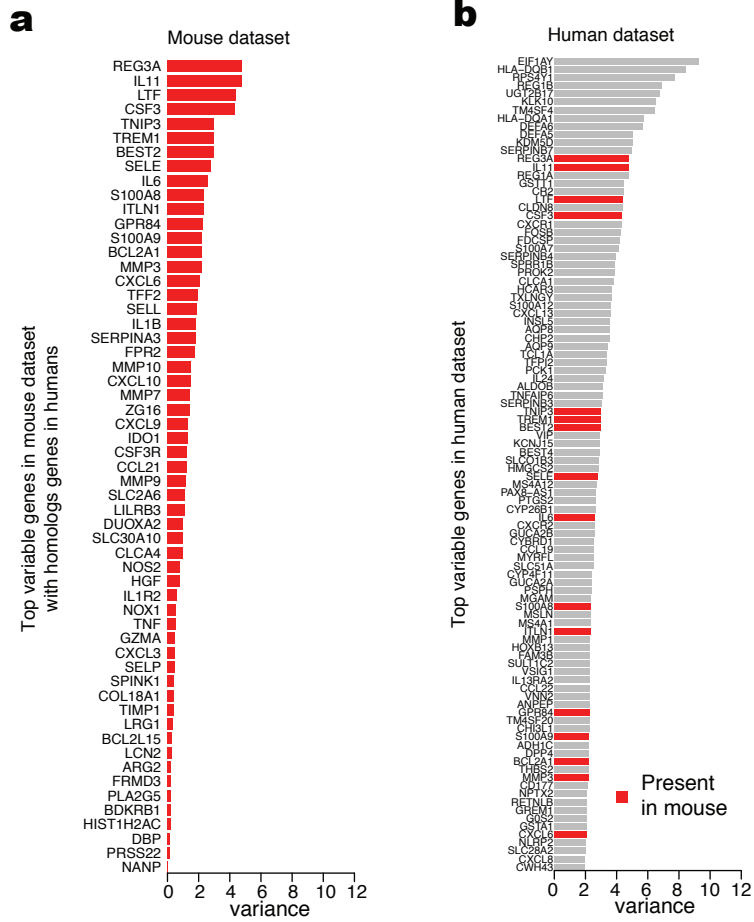FITC-dextran (µg/mL)

d0 | d7 | d14

Figure S3

Figure S4

Figure S5

Figure S6

Figure S7