The mutational profile and clonal landscape of the inflammatory bowel disease affected colon

Sigurgeir Olafsson[1], Rebecca E. McIntyre[1], Tim Coorens[1], Tim Butler[1], Philip Robinson[1], Henry Lee-Six[1], Mathijs A. Sanders[1,2], Kenneth Arestang[3], Claire Dawson[3], Monika Tripathi[4], Konstantina Strongili[3], Yvette Hooks[1], Michael R. Stratton[1], Miles Parkes[3], Inigo Martincorena[1],Tim Raine[3], Peter J. Campbell[1*], Carl A. Anderson[1*]

1. Wellcome Sanger Institute, Hinxton, UK

2. Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands

3. Department of Gastroenterology, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK.

4. Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK.

***Address for correspondence and material requests:**

Dr Peter J. Campbell, Cancer Genome Project, Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom.

Telephone: +44 (0) 1223 834244.

e-mail: pc8@sanger.ac.uk

Dr Carl A Anderson, Genomics of inflammation and immunity, Human Genetics, Wellcome Sanger Institute, Hinxton CB10 1SA, United Kingdom.

Telephone: +44 (0) 1223 892371.

e-mail: carl.anderson@sanger.ac.uk

Summary paragraph

Inflammatory bowel disease (IBD) is a chronic inflammatory disease associated with increased risk of gastrointestinal cancers[1–3] but our understanding of the effects of IBD on the mutational profile and clonal structure of the colon is limited. Here, we isolated and whole-genome sequenced 370 colonic crypts from 45 IBD patients, and compared these to 413 crypts from 41 non-IBD controls. We estimated the base substitution rate of affected colonic epithelial cells to be doubled after IBD onset. This change was primarily driven by acceleration of mutational processes ubiquitously observed in normal colon, and we did not detect an IBD-specific mutational process. In contrast to the normal colon, where clonal expansions outside the confines of the crypt are rare, we observed widespread millimeter-scale clonal expansions. We also found that non-synonymous mutations in *ARID1A, PIGR and ZC3H12A,* and genes in the interleukin 17 and Toll-like receptor pathways, were under positive selection in colonic crypts from IBD patients. With the exception of *ARID1A*, these genes and pathways have not been previously associated with cancer risk. Our results provide new insights into the consequences of chronic intestinal inflammation on the

mutational profile and clonal structure of colonic epithelia and point to potential therapeutic targets for IBD.

Introduction

Inflammatory bowel disease (IBD) is a debilitating disease characterized by repeated flares of intestinal inflammation. The two major subtypes of IBD, Crohn's disease (CD) and ulcerative colitis (UC), are distinguished by the location, continuity and nature of the inflammatory lesions. UC affects only the large intestine, spreading continuously from the distal to proximal colon, whereas CD most commonly affects the small and large intestine, and is characterized by discontinuous patches of inflammation. IBD patients have a 1.7-fold increased risk of developing gastrointestinal cancers compared to the general population, with cancer risk being associated with the duration, extent and severity of disease[1–3]. As a result, IBD patients require regular endoscopic screening and may undergo prophylactic colectomy to mitigate this risk[1,2].

That somatic mutations contribute to the development of cancer is well established, but their patterns, burden and functional consequences in diseases other than cancer have not been extensively studied. Methodological developments have now enabled the analysis of polyclonal somatic tissues, allowing characterization of somatic mutations in normal tissues such as skin[4], oesophagus[5,6], endometrium[7,8] and colon[9,10]. In the setting of non-neoplastic diseases, chronic liver disease has had the most attention, with studies showing that compared to healthy liver, hepatic cirrhosis is associated with acquisition of new mutational processes, increased mutation burden and larger clonal expansions[11–13].

Colonic epithelium is well suited to the study of somatic mutations on account of its clonal structure. It is organized into millions of colonic crypts, finger-like invaginations composed of approximately 2000 cells[14] each that extend into the lamina propria below. At the base of each crypt resides a small number of stem cells undergoing continuous self-renewal through stochastic cell divisions[15,16]. As a result, the progeny of a single stem cell iteratively sweep the entire niche and the epithelial cells that line the crypt are the progeny of this single clone. Active inflammatory bowel disease disrupts these normal stem cell dynamics - the epithelial lining is damaged, the organised crypt structure is ablated and the barrier between lumen and mucosa is disrupted.

We hypothesised that the recurrent cycles of inflammation, ulceration and regeneration seen in inflammatory bowel disease could impact the mutational and clonal structure of intestinal epithelial cells. To test these hypotheses we isolated and whole-genome sequenced 370 colonic crypts from 45 IBD patients with varying degrees of current and previous colonic inflammation, and compared the mutation burden, clonal structure, mutagen exposure and driver mutation landscape to colonic crypts from healthy donors[9].

Results

IBD doubles the mutation rate of colonic epithelial cells

We used laser capture microdissection (LCM) to isolate 370 colonic crypts from endoscopic biopsies taken from 28 UC patients and 17 CD patients (Supplementary Table 1, Extended Data Figure 1). Biopsies were annotated as never-, previously- or actively inflamed at

the time of sampling (Methods). The dissected crypts were whole-genome sequenced to a mean depth of 17.7, allowing us to call somatic substitutions, small insertions and deletions (indels) and larger deletions and duplications (Methods, Supplementary Tables 2, 3 and 4).

To assess if IBD is associated with a difference in the mutation burden of the colonic epithelium, we combined our data with data from 413 crypts sequenced as part of our recent study of somatic mutations in normal colon[9] (hereafter referred to as the control data). We fitted linear mixed-effects models that enabled us to estimate the independent effects of age, disease duration and biopsy location on mutation burden, while controlling for the within-patient and within-biopsy correlations inherent in our sampling strategy (Methods). We estimated the effect of IBD to be an additional 49 substitutions per crypt per year of disease duration (26-71 95% CI, $P=4.9x10^{-5}$, linear mixed effects models and likelihood ratio - Figure 1A). These mutations are in addition to the 41 (29-52, 95% CI) substitutions we estimated are accumulated on average per year of life, suggesting that mutation rates are roughly doubled in regions of the colon affected by IBD. We found greater between-patient variance in the IBD cohort compared with the controls (SD=867 and 585 mutations for cases and controls, respectively. $P=6.4x10^{-7}$, likelihood ratio test). This was expected as patients vary in their disease severity and response to treatment. We similarly estimate an increase in the indel burden of seven indels per crypt per year of IBD (5 - 9 95% CI, $P=3.1x10^{-9}$ - Figure 1B) in addition to the estimated 1 (0.3-1.8 95% CI) indel per year that is accumulated per year of life.

We called large somatic deletions and duplications across all samples in our cohort. The burden of structural variants (SVs) was modest in both datasets (Figure 1C) but we observed a significant disease effect of 0.075 (0.033 - 0.012 95% CI, $P=4.7x10^{-4}$, Poisson regression) SVs, or one SV per crypt per 13.3 years of disease, on average. We additionally found increased rates

of deletions at the fragile site chr16p13.2, but not at two other fragile sites, chr16q23.1 and chr3p14.2 (P = 0.0083, 0.29, 0.58, respectively. Poisson regression. Extended Data Figure 2). We found no significant difference in the burden of substitutions, indels or SVs between UC and CD patients (P=0.18, 0.10 and 0.43, respectively).

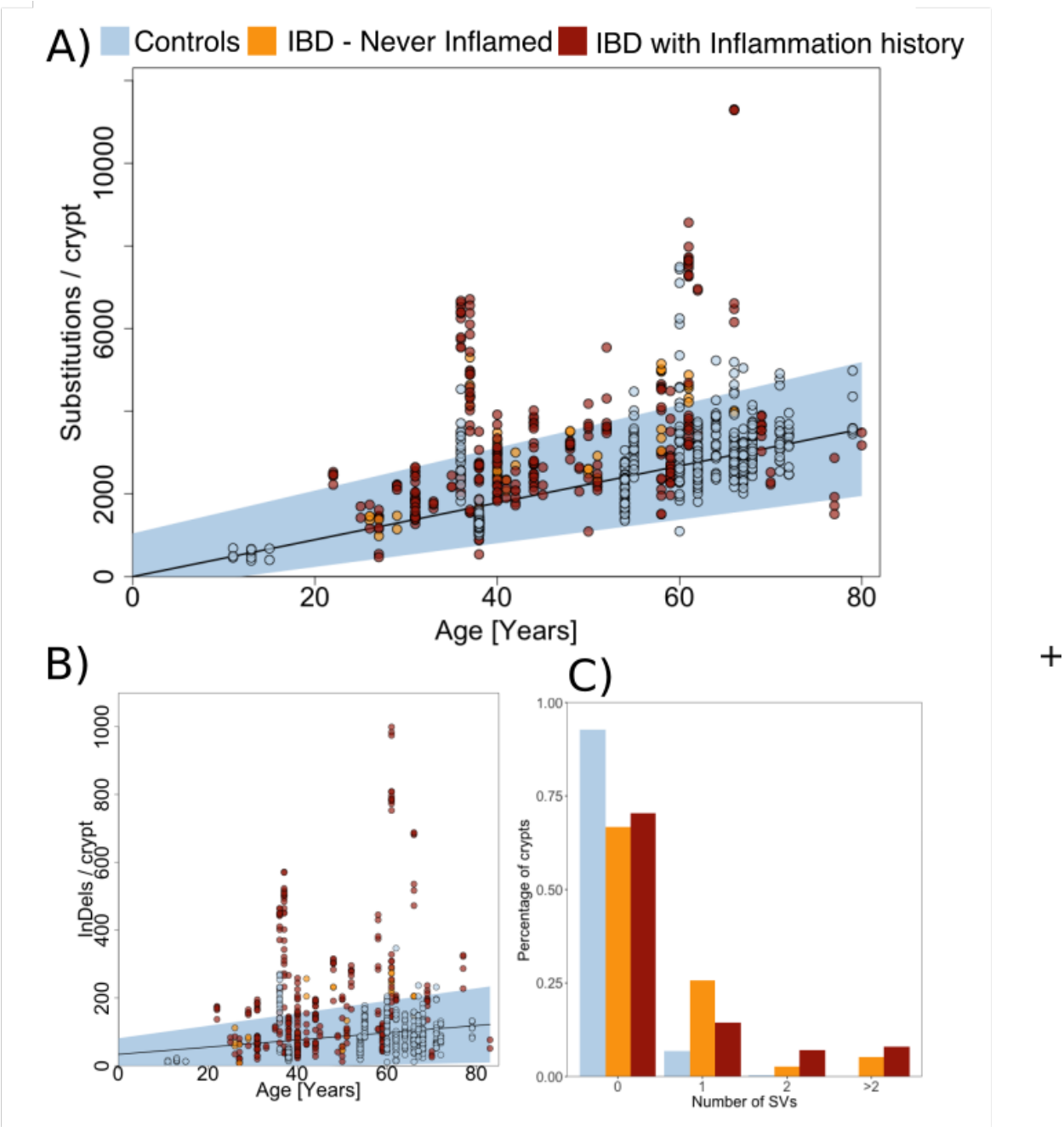


Figure 1. Mutation burden in the IBD colon. A) Substitution burden as a function of age. Each point represents a colonic crypt and is coloured by disease status. The line shows the effect of age on mutation burden as estimated by fitting a linear mixed effects model correcting for sampling location and the within-biopsy and within-patient correlation structure. The blue shaded area represents the 95% confidence interval. B) Same as A) but for indels. C) The fraction of crypts found to carry structural variants in IBD and controls.

IBD accelerates normal mutagenic processes

The somatic mutations found in the cells of a colonic crypt reflect the mutational processes that have acted on the stem cells and their progenitors since conception. Distinct mutational processes each leave a characteristic pattern, a mutational signature, within the genome, distinguished by the specific base changes and their local sequence context[17,18]. We extracted single base substitution (SBS) mutational signatures for both IBD and control crypts (Methods), and did not identify a mutational signature specific to IBD, CD or UC (Figure 2A, Supplementary Table 2). Rather, we found that approximately 80% of the increase in mutation burden is explained by an increased burden of mutational signatures 1, 5 and 18, as defined by Alexandrov et al., which are also found ubiquitously in normal colon[9,10] (an increase of 12 (6-18 95% CI), 22 (12-33 95% CI) and 6 (3-8 95% CI) mutations per crypt per year of disease, respectively. $P=1.1x10^{-4}$, $2.8x10^{-5}$ and $5.4x10^{-5}$, linear mixed effects models, likelihood ratio - Figure 2B). Signatures 1 and 5 are clock-like and thought to be associated with cell proliferation, while signature 18 has been linked with reactive oxygen species[18].

The remaining 20% of the increase in mutation burden is a consequence of rarer mutational processes and treatment. For example, we found 77 crypts with over 150 mutations

attributed to purine treatment in a subset of seven IBD patients, five of whom have a documented history of such treatment. However, the number of mutations attributed to purine was not associated with purine therapy duration, and some patients showed large mutation burdens despite brief, or indeed no, documented clinical exposure (Figure 2C).

We observed that two signatures previously discovered in the normal colon[9], SBSA and SBSB, were also present in the context of IBD. As in normal colon, these mutational processes were primarily active early in life (Extended Data Figures 3 and 4). We also found signatures 2 and 13, which are associated with APOBEC activity, and signatures 17a and 17b, which are of unknown aetiology, to be active in a small number of crypts with high mutation burdens. Finally, we found signature 35, associated with platinum compound therapy, in one patient with a history of platinum treatment for squamous cell carcinoma of the tongue. We did not find a difference in the number of mutations attributed to signatures 1, 5 or 18 between UC and CD patients (P=0.23, P=0.32 and P=0.25, respectively).

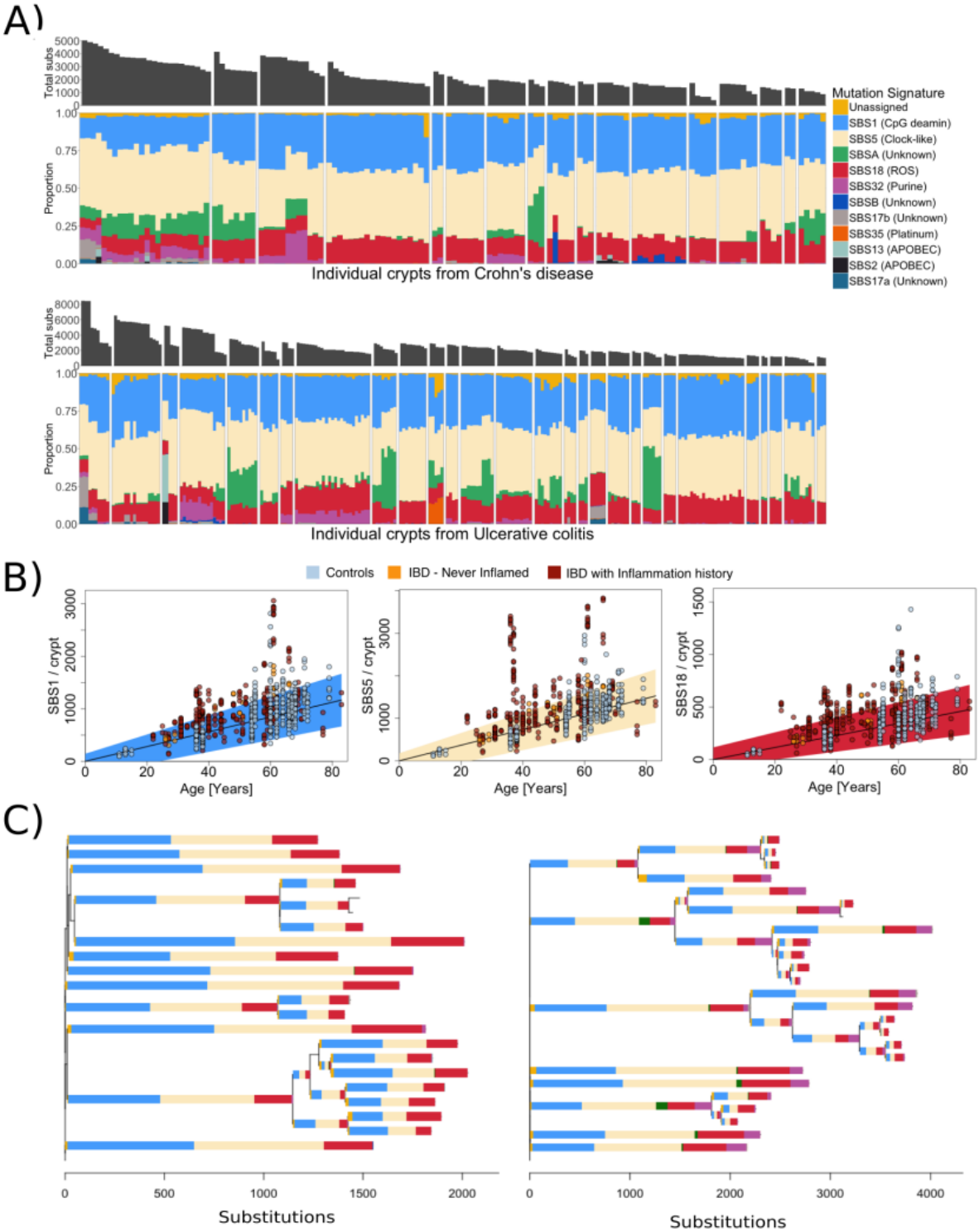


Figure 2. Mutational signatures in the IBD colon. A) A stacked bar-plot showing the relative contribution of each signature in each crypt. Crypts are grouped by patient and ordered by

mutation burden, and crypts from Crohn's disease and Ulcerative colitis patients are shown separately. Signature nomenclature is the same as in Alexandrov et al (2019). B) Burden of mutational signatures 1, 5 and 18 with age. The shaded area represents the 95% confidence interval around the age estimate. C) Phylogenic trees of two patients with widespread ulcerative colitis. The colours of the branches reflect the relative contribution of each mutational signature extracted for those branches as in A). The patient on the left has received azathioprine treatment for 10 years but shows no SBS32 burden (purple). In contrast, the patient on the right received azathioprine for 2 weeks and mercaptopurine for 2 weeks and had significant adverse reactions to both drugs. SBS32 is found in most crypts from this patient. All crypts are from inflamed biopsies. SBS: Substitution signature.

IBD creates a patchwork of millimeter-scale clones

Colonic crypts divide by a process called crypt fission, whereby a crypt bifurcates at the base and branching elongates in a zip-like manner towards the lumen. This process is relatively rare in the normal colon, wherein each crypt fissions on average only once every 27 years[9,19]. Compared to normal colon, we found much larger clonal expansions in IBD patients, evident of numerous crypt fission events occurring late in molecular time. We observed several examples of individual clones spanning entire 2-3 mm endoscopic biopsies (Figure 3, Extended Data Figures 3 and 4). However, when we biopsied the same inflamed or previously inflamed region more than once, we did not observe any clones that stretched between biopsies that were taken a few millimeters apart (11 patients; Extended Data Figures 3 and 4). Several biopsies contained more than one clone. Taken together, these observations suggest that crypts in the IBD colon are often clonal on the scale of millimeters, but rarely centimeters, and that IBD-affected regions are

generally not dominated by a single major clone, but are more accurately viewed as an oligoclonal patchwork of clones that often grow considerably larger than in healthy colon. We hypothesize that these are the result of rapid crypt fission events following inflammation-driven bottlenecks. Clones stretching large distances have been described in IBD[20] and field cancerization, the wide-spread expansion of clones carrying pathogenic mutations, is thought to precede the development of neoplastic lesions in IBD[21]. Mutations in *TP53, APC* and *KRAS* are thought to play a key role in the initiation of these events[20,22], but no coding mutations in these genes were found in our dataset, suggesting that clones start expanding even before these mutations are acquired.

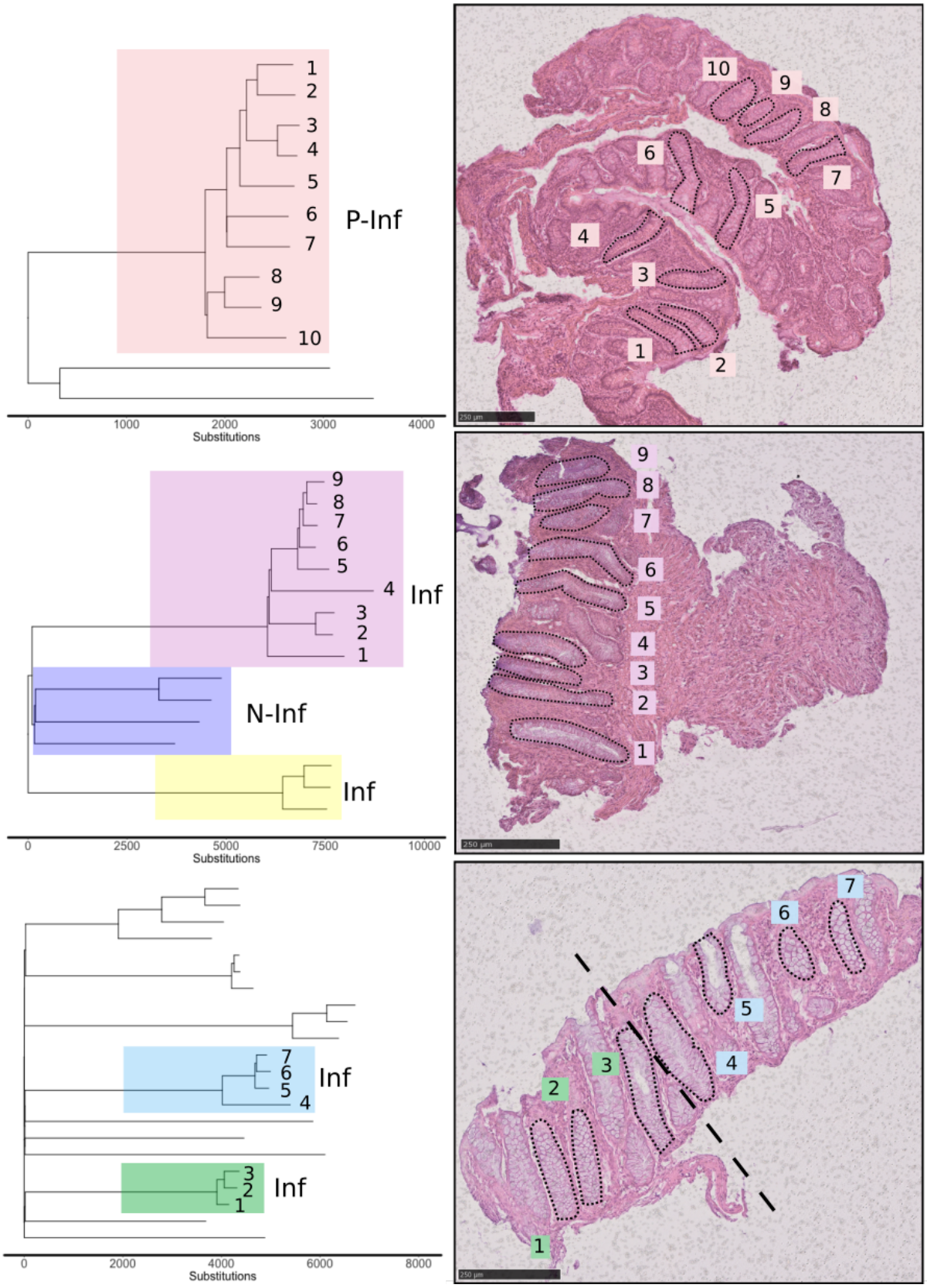

1
2
3
4
5
6
7
8
9
10
P-Inf
0
1000
2000
3000
4000
Substitutions
9
8
7
6
5
4
3
2
1
Inf
N-Inf
Inf
0
2500
5000
7500
10000
Substitutions
7
6
5
4
Inf
3
2
1
Inf
0
2000
4000
6000
8000
Substitutions
10
9
8
7
6
5
4
3
2
1
9
8
7
6
5
4
3
2
1
7
6
5
4
3
2
1

Figure 3. Examples of clonal expansions in three IBD patients. (Top) A phylogenic tree of crypts sampled from a 38 year old patient with a 21 year history of ulcerative colitis. The accompanying biopsy image shows the crypts from the shaded area - a clone stretching for >2 mm of tissue. (Middle) A phylogenic tree of crypts sampled from a 61 year old patient with a 27 year history of ulcerative colitis. The clones highlighted in purple and yellow come from biopsies taken millimeters apart. The accompanying biopsy image shows the crypts from the purple clone. (Bottom) A phylogenic tree of crypts sampled from a 37 year old patient with a 25 year history of Crohn's disease affecting the colon. A biopsy overlaps two clones (in blue and green). Inf: Inflamed. P-Inf: Previously inflamed. N-Inf: Never inflamed.

IBD drives positive selection in colonic stem cells

The recurrent cycles of inflammation and remission which characterise IBD could create an environment in which clones containing advantageous mutations may selectively spread in the mucosa. This advantage may manifest either through faster cell division and elevated crypt fission rate or through increased resistance to inflammation. To search for evidence of selection in the IBD colon, we applied the dN/dScv method[23] (Methods). Briefly, dN/dScv uses the observed number of synonymous mutations to estimate background mutation rates, searching for genes where the number of non-synonymous mutations is significantly higher than this background expectation, after accounting for local sequence context and gene-to-gene variation in mutation rates.

The application of dNdScv revealed three genes, *ARID1A, PIGR* and *ZC3H12A*, to be under positive selection in the IBD colon (Figure 4A). *ARID1A* is a well-established tumour suppressor, both in sporadic- and colitis-associated colorectal cancer[23,24], and also plays a role in

maintaining the intestinal stem cell niche[25]. Heterozygous truncating mutations in *ARID1A* preceded several clonal expansions in our dataset and some patients carried distinct *ARID1A* mutations in different crypts (Figure 4B).

To our knowledge, mutations in the other two genes, *PIGR* and *ZC3H12A*, have not been described in cancer. *ZC3H12A* encodes an RNAse, Regnase-1, which potentiates the mechanistic target of rapamycin kinase (mTOR) and purine metabolism in epithelial cells[26]. Mice with intestinal epithelial cell-specific knock-out mutations of *Zc3h12a* are more resistant to dextran sulfate sodium-induced epithelial injury[26] - a common murine model of IBD. We hypothesize that mutations in *ZC3H12A* protect human colonic epithelial cells from the damaging effects of inflammation and facilitate rapid healing of the mucosa without predisposing to cancer. This would suggest that therapeutic modulation of *ZC3H12A* may represent a potential treatment strategy in IBD.

PIGR encodes the poly-immunoglobulin (Ig) receptor, which transfers polymeric Igs produced by plasma cells in the mucosal wall across the epithelium to be secreted into the intestinal lumen[27]. *Pigr* knock-out mice exhibit decreased epithelial barrier integrity and increased susceptibility to mucosal infections and penetration of commensal bacteria into tissues[28].

We next carried out a pathway-level dN/dS analysis, searching for enrichment of missense and truncating variants across 15 gene sets that were defined *a priori* because of their relevance in either colorectal carcinogenesis or IBD pathology (Figure 4C, Supplementary Table 5, Extended Data Figure 5, Methods). We observed a nominally significant enrichment in genes associated with colorectal cancer (q=0.043) and cancer of any type (q=0.043), but these were driven by the mutations in *ARID1A* described above. As previously stated, we did not find a

single mutation in *APC*, *KRAS* or *TP53*, the genes most commonly mutated in both sporadic- and colitis-associated colorectal cancer[23,24]. We did observe two large-scale deletions that overlap known tumour suppressors, *PIK3R1* and *CUX1*, and are likely drivers. Both precede clonal expansions in our dataset (Figure 4B and Extended Data Figure 6).

Interestingly, the pathway-level dNdS also revealed a significant enrichment of truncating mutations in the interleukin-17 (IL17) signaling pathway (q=0.0044) and Toll-like receptor (TLR) cascades (q=0.0072) (Figure 4C). Resident microbiota interact with the epithelium to stimulate *Pigr* expression through TLR binding of microbe-associated molecular patterns[27,29]. Intestinal epithelial-cell specific knock-out of components of the IL17 pathway in mice also results in commensal dysbiosis through down-regulation of *Pigr* and other genes[30]. Although we cannot rule out the possibility that mutations in *PIGR* and the TLR and IL17 pathways confer upon cells some survival advantage in the context of inflammation, together these results suggest that somatic mutations may initiate, maintain or perpetuate IBD pathogenesis through disruption of microbe-epithelial homeostasis. Further study is required to test this hypothesis.

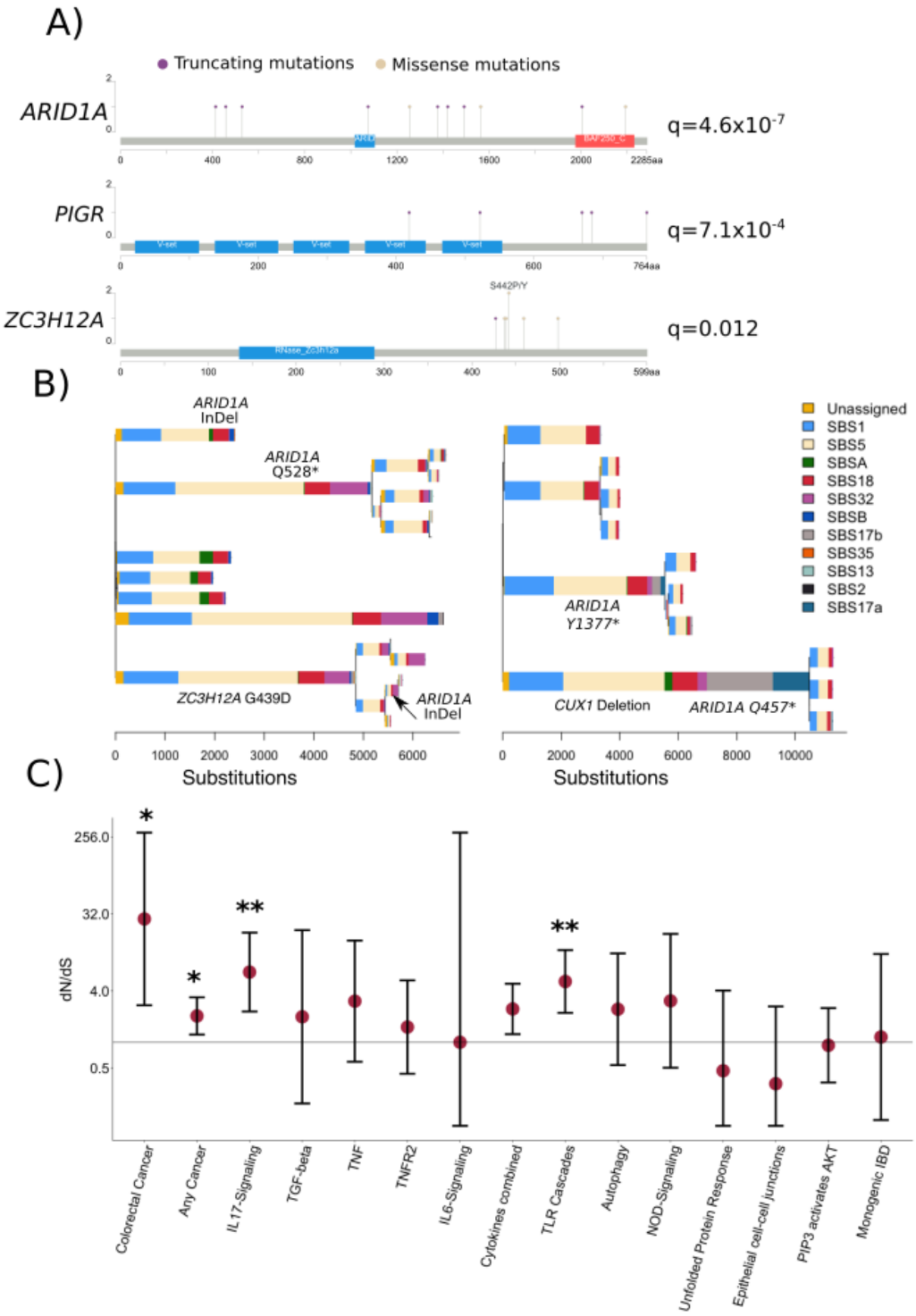


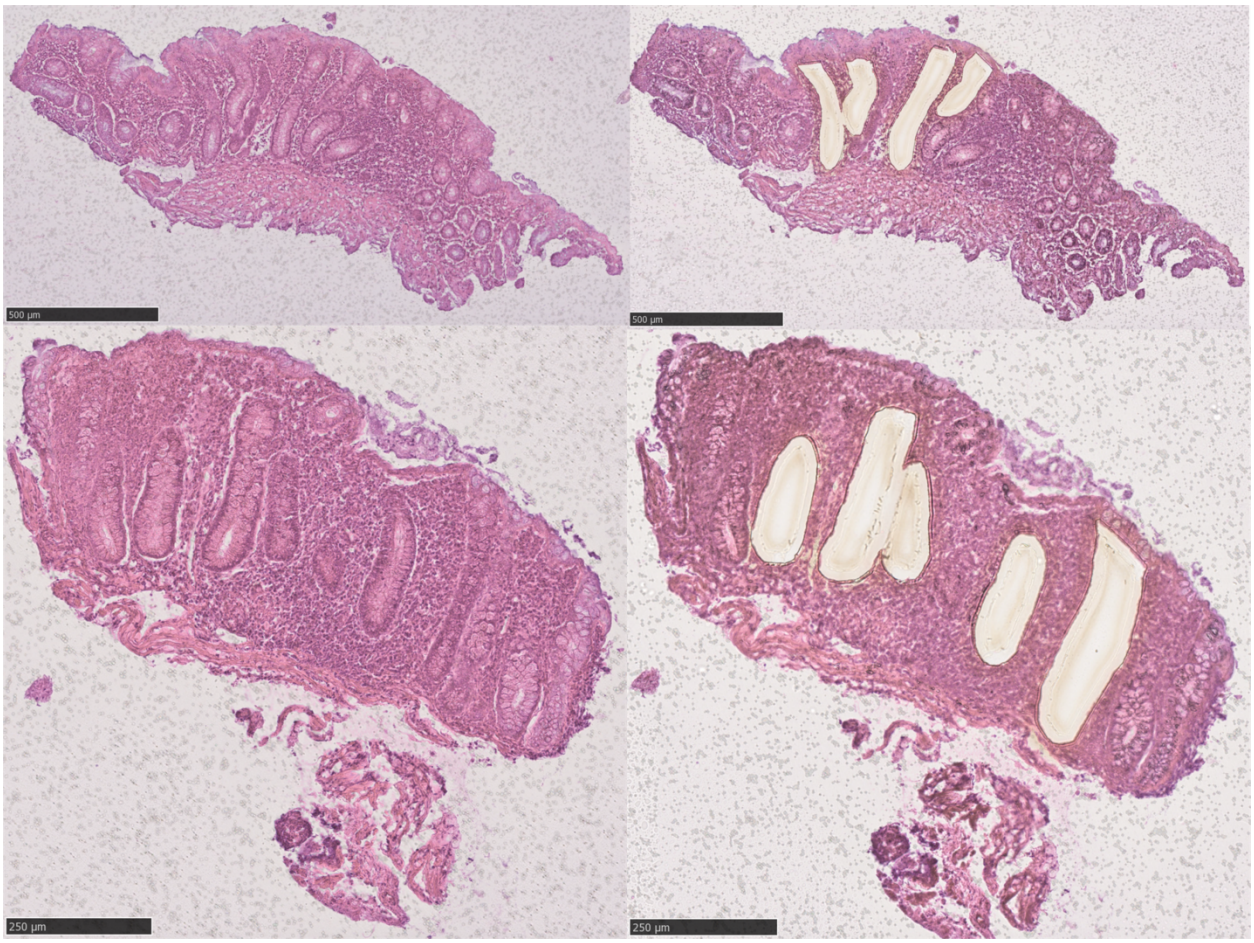
Figure 4. Driver mutations and positive selection in IBD. A) A lollipop plot showing the location of mutations found in *ARID1A, PIGR* and *ZC3H12A* - genes found to be under positive selection in the IBD colon. B) Phylogenic trees of two patients showing the location of driver

mutations on the phylogenic tree. C) Pathway-level dN/dS ratios for truncating mutations in known cancer genes and cellular pathways important in IBD pathogenesis. Error bars represent 95% confidence intervals. *q<0.05. **q<0.005, Benjamini-Hochberg.

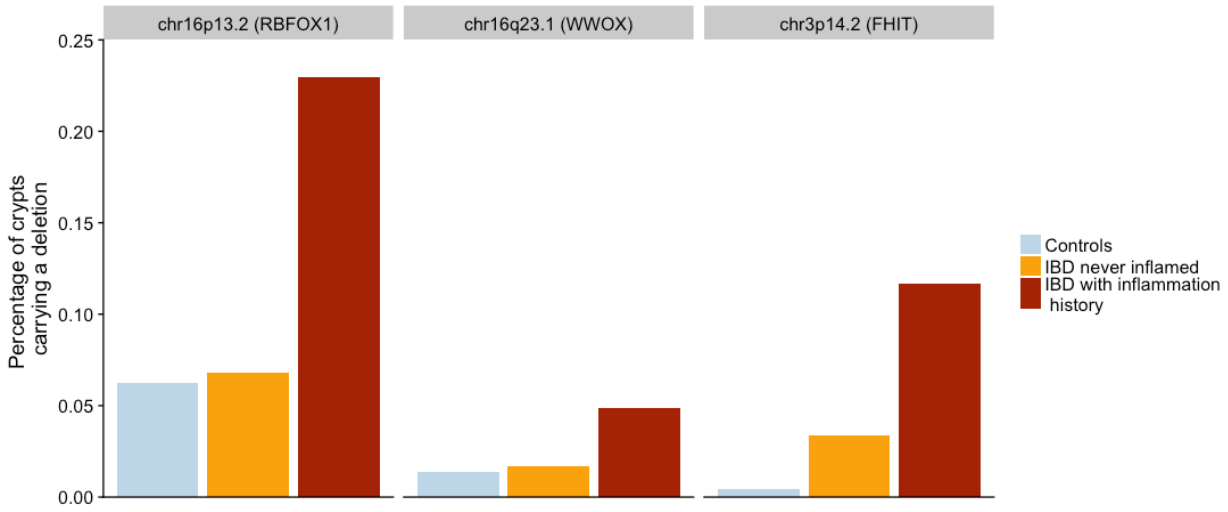
Conclusions

We have characterized the somatic mutation landscape of the IBD affected colon. We show that the mutation burden is doubled after disease onset and the increase is mostly driven by acceleration of common mutational processes. We show that millimeter scale clonal expansions are common in IBD, even in the absence of *TP53*, *KRAS* or *APC* mutations. Truncating mutations in *ARID1A* commonly drive clonal expansions and may be early events in neoplasm development in IBD patients. An excess of mutations was also found in two genes with roles in maintaining epithelial homeostasis, *PIGR* and *ZC3H12A,* but which we believe don't drive carcinogenesis and two pathways, IL17 and TLR signaling, with established roles in IBD pathogenesis. These results further our understanding of the early steps of neoplasm development under inflammatory conditions and suggest that studying somatic mutations in complex diseases may reveal novel drug targets and causal mechanisms.

Extended Figures and figure legends



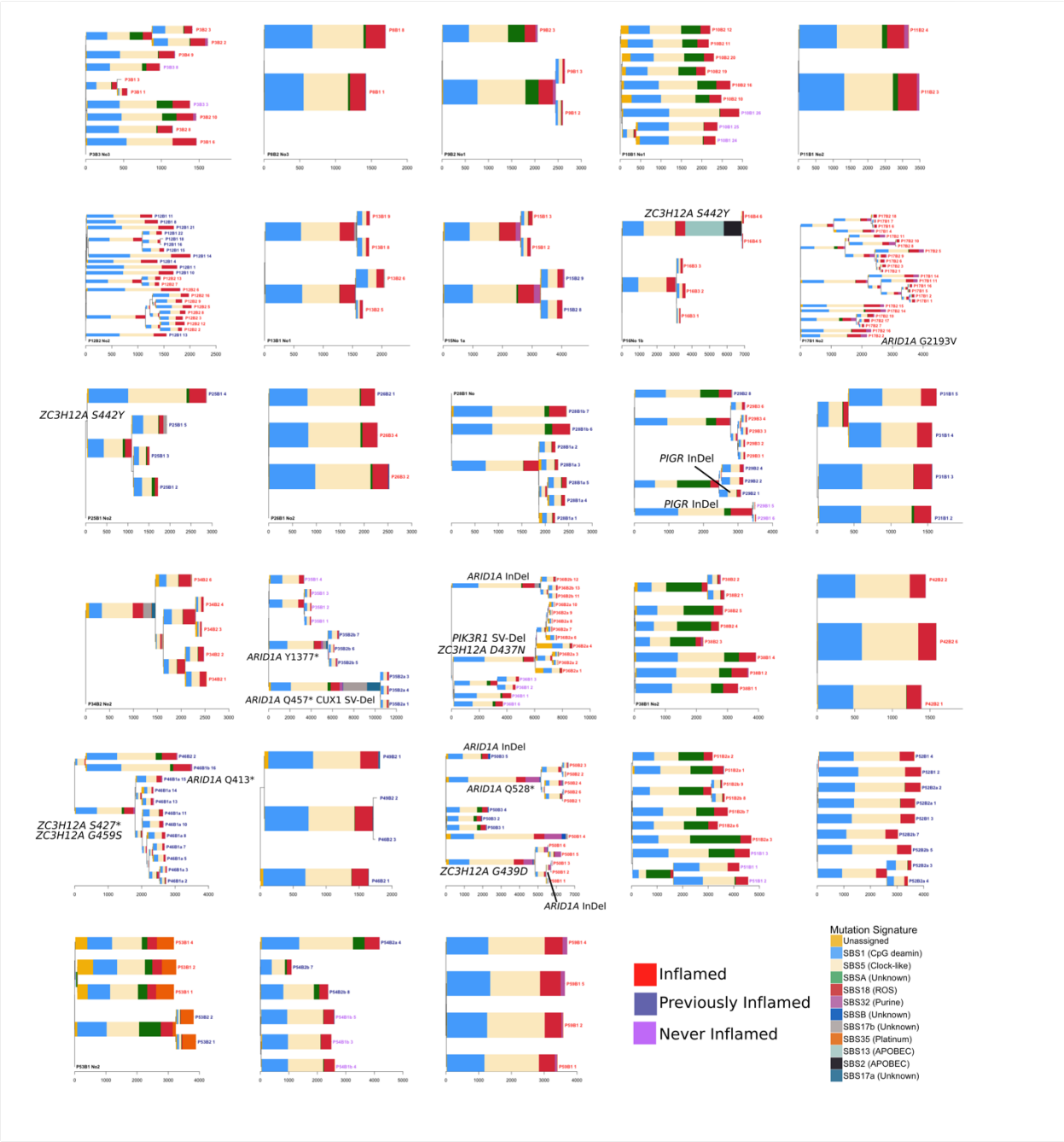
Extended Data Figure 1: Laser capture microdissection of crypts from two biopsies. The left side shows representative images of sections of endoscopic biopsies from colonic tissue before crypt dissection. The right side shows the same sections after crypt dissection.



Extended Data Figure 2: Fraction of crypts carrying deletions at three fragile sites of the genome. Only chr16p13.2 reaches statistical significance (P=0.0083).

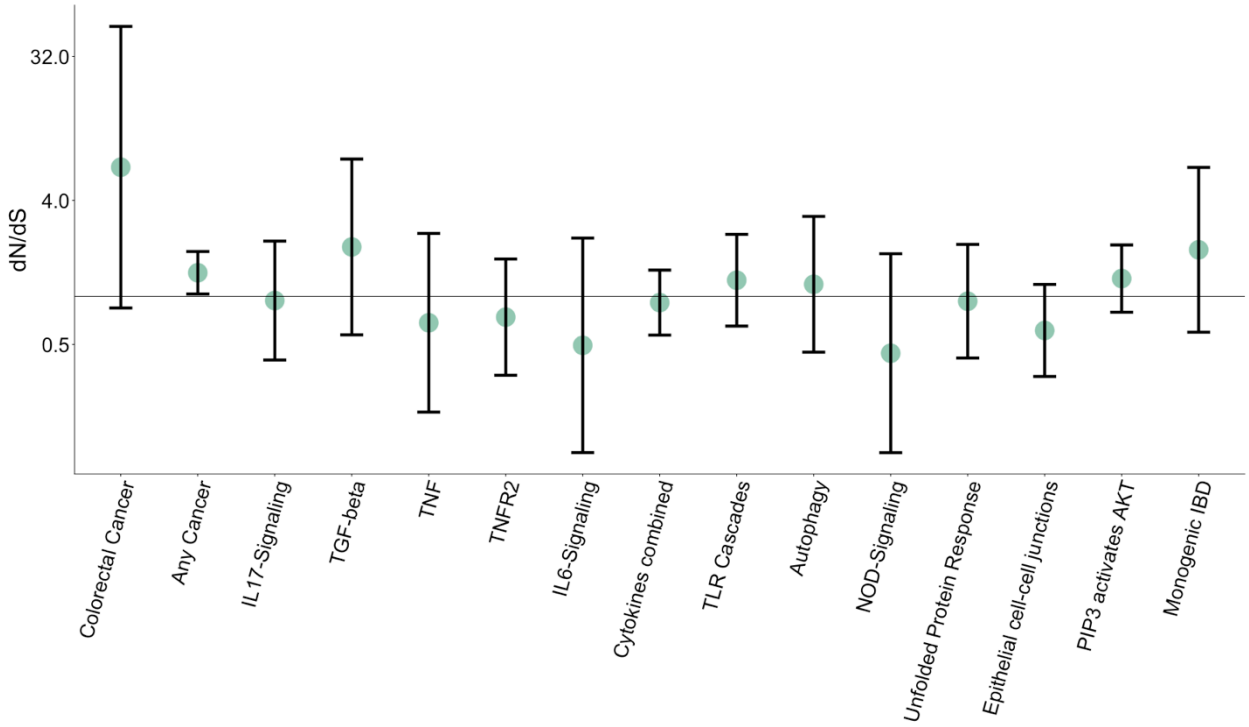
Extended Data Figure 3: Phylogenic trees for all Crohn's disease patients. Mutational signatures are overlaid on the trees and likely driver mutations are mapped to the branch in which they occur. Crypts are labelled on the form PXBY_Z where PX is the patient number, BY the biopsy number (with a,b and c denoting biopsies taken a few millimeters apart from the same

site) and Z is the crypt number. The colour of the labels indicates whether a crypt comes from an inflamed, previously inflamed or never inflamed site.

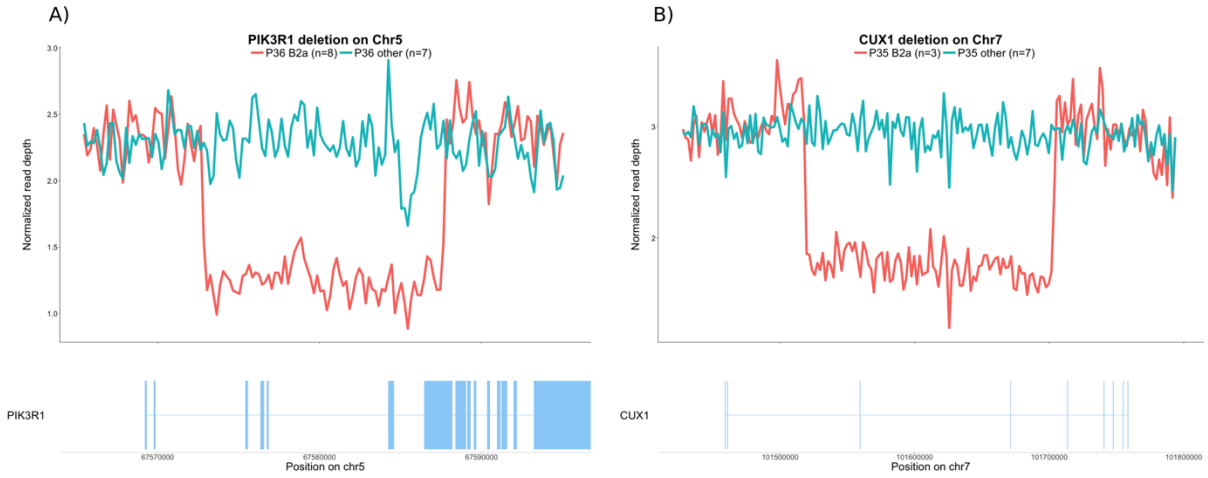


Extended Data Figure 4: Phylogenic trees for all ulcerative colitis patients. Mutational signatures are overlaid on the trees and likely driver mutations are mapped to the branch in

which they occur. Crypts are labelled on the form PXBY_Z where PX is the patient number, BY the biopsy number (with a,b and c denoting biopsies taken a few millimeters apart from the same site) and Z is the crypt number. The colour of the labels indicates whether a crypt comes from an inflamed, previously inflamed or never inflamed site.



Extended Data Figure 5: Pathway-level dN/dS ratios for missense mutations in known cancer genes and cellular pathways important in IBD pathogenesis.



Extended Data Figure 6: Structural variants of probable driver status. The figure compares normalized read depths of crypts called as carriers/non carriers. A) A deletion covering five exons of *PIK3R1* found to precede a clonal expansion in biopsy 2a of patient 36 (Figure 3 of the main text, middle panel, purple clone). B) A deletion covering three exons of *CUX1* and found to precede a clonal expansion in biopsy 2a of patient 35.

Data availability

All sequencing data will be made available via the European Genome Phenome Archive prior to publication (https://ega-archive.org/) and accession codes provided. Histology images are available from the authors upon request.

Code availability

All code supporting this study can be found at https://github.com/Solafsson/somaticIBD.

References

1. Beaugerie, L. & Itzkowitz, S. H. Cancers Complicating Inflammatory Bowel Disease. *N. Engl. J. Med.* **372**, 1441–1452 (2015).
2. Adami, H.-O. *et al.* The continuing uncertainty about cancer risk in inflammatory bowel disease. *Gut* **65**, 889–893 (2016).
3. Lutgens, M. W. M. D. *et al.* Declining risk of colorectal cancer in inflammatory bowel disease: an updated meta-analysis of population-based cohort studies. *Inflamm. Bowel Dis.* **19**, 789–799 (2013).
4. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
5. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
6. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
7. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *bioRxiv* 505685 (2018).
8. Suda, K. *et al.* Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep.* **24**, 1777–1789 (2018).
9. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
10. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during

life. *Nature* **538**, 260–264 (2016).

11. Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
12. Kim, S. K. *et al.* Comprehensive analysis of genetic aberrations linked to tumorigenesis in regenerative nodules of liver cirrhosis. *J. Gastroenterol.* **54**, 628–640 (2019).
13. Zhu, M. *et al.* Somatic Mutations Increase Hepatic Clonal Fitness and Regeneration in Chronic Liver Disease. *Cell* **177**, 608–621.e12 (2019).
14. Potten, C. S., Kellett, M., Roberts, S. A., Rew, D. A. & Wilson, G. D. Measurement of in vivo proliferation in human colorectal mucosa using bromodeoxyuridine. *Gut* **33**, 71–78 (1992).
15. Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822–825 (2010).
16. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
17. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
18. Alexandrov, L. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv* 322859 (2018).
19. Nicholson, A. M. *et al.* Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell* **22**, 909–918.e8 (2018).
20. Galandiuk, S. *et al.* Field Cancerization in the Intestinal Epithelium of Patients With Crohn's Ileocolitis. *Gastroenterology* **142**, 855–864.e8 (2012).
21. Choi, C.-H. R., Bakir, I. A., Hart, A. L. & Graham, T. A. Clonal evolution of colorectal

cancer in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **14**, 218–229 (2017).

22. Leedham, S. J. *et al.* Clonality, founder mutations, and field cancerization in human ulcerative colitis-associated neoplasia. *Gastroenterology* **136**, 542–50.e6 (2009).
23. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
24. Baker, A.-M. *et al.* Evolutionary history of human colitis-associated colorectal cancer. *Gut* gutjnl–2018–316191 (2018).
25. Hiramatsu, Y. *et al.* Arid1a is essential for intestinal stem cells through Sox9 regulation. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 1704–1713 (2019).
26. Nagahama, Y. *et al.* Regnase-1 controls colon epithelial regeneration via regulation of mTOR and purine metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 11036–11041 (2018).
27. Johansen, F.-E. & Kaetzel, C. S. Regulation of the polymeric immunoglobulin receptor and IgA transport: new advances in environmental factors that stimulate pIgR expression and its role in mucosal immunity. *Mucosal Immunol.* **4**, 598–602 (2011).
28. Johansen, F. E. *et al.* Absence of epithelial immunoglobulin A transport, with increased mucosal leakiness, in polymeric immunoglobulin receptor/secretory component-deficient mice. *J. Exp. Med.* **190**, 915–922 (1999).
29. Frantz, A. L. *et al.* Targeted deletion of MyD88 in intestinal epithelial cells results in compromised antibacterial immunity associated with downregulation of polymeric immunoglobulin receptor, mucin-2, and antibacterial peptides. *Mucosal Immunol.* **5**, 501–512 (2012).
30. Kumar, P. *et al.* Intestinal Interleukin-17 Receptor Signaling Mediates Reciprocal Control of the Gut Microbiota and Autoimmune Inflammation. *Immunity* **44**, 659–671 (2016).

Acknowledgements

This work was supported by the Wellcome Trust grant [206194]. We thank the staff of Wellcome Sanger Institute's Sequencing, Sample Management and Informatics facilities for their contribution to the study. We thank Paul Scott for his help with the laser capture microdissection and Dr. Doug Winton for discussion on the design of the study and interpretation of its results. Finally, we give our dearest thanks to the participants of this study, who consented to having an invasive and sometimes painful colonoscopy procedure extended in order to donate samples for this study.

Author contributions

SO, RM, MP, IM, TR, PJC and CAA designed the study. KA, CD, KS, MP and TR were involved in cohort ascertainment, phenotypic characterization and recruitment. SO, RM and YH embedded and sectioned biopsies and SO and RM also imaged, microdissected, and lysed colonic crypts with contributions from HLS and TB. MT reviewed biopsy histology. SO, TC, PR, TB, HLS and MS contributed to statistical analyses, mutation calling and filtering. IM, PJC and CAA oversaw statistical analyses. SO, MRS, IM, TR, PJC and CAA interpreted the findings. SO wrote the first draft of the manuscript and all authors contributed to its final version. PJC and CAA supervised the study.

Competing interest

CAA is a paid consultant for Genomics plc and Celgene. All other authors declare no competing interests.

Methods

Human tissue attainment and processing

Colonic pinch-biopsies were donated by IBD patients undergoing regular surveillance of their disease at Addenbrooke's hospital, Cambridge (Supplementary Table 1). All samples were obtained with informed consent of the donor and the study was approved by the National Health Service (NHS) Research Ethics Committee (Cambridge South, REC ID 17/EE/0338) and by the Wellcome Trust Sanger Institute Human Materials and Data Management Committee (approval number 17/113). We have complied with all relevant ethical regulations.

All donors are of white-European ancestry. The time between clinical diagnosis and date of biopsy was used to define the disease duration of a given individual. We added six months to this number for all patients because symptoms often precede diagnosis by several months and to avoid setting the disease duration to zero for patients who donated samples at the time of diagnosis. Time of purine treatment was estimated by consulting electronic health records from NHS databases. Biopsies were annotated as never, previously or actively inflamed using all available clinical data and NHS histopathology archives. The biopsy images (or an image of a second biopsy from the same site of the colon) were reviewed by a histopathologist. None of the patients had colorectal cancer or adenoma.

Biopsies from patients 1-26 were embedded in optimal cutting temperature (OCT) compound and sectioned, stained and fixed as previously described[1]. Subsequent biopsies were embedded in paraffin because this better preserved the morphology of the tissue. Biopsies were sectioned (10-20 µm), fixed to 4 µm PEN membrane slides (11600288, Leica) and stained with hematoxylin and eosin. Crypts were dissected using laser capture microdissection microscopy (LMD7000, Leica) and lyzed using ARCTURUS PicoPure DNA extraction kit (Applied Biosystems) according to the manufacturer's instructions. DNA libraries were prepared as previously described[1].

Whole genome sequencing

Samples from patient 1 through 19 (Supplementary Tables 1 and 2) were sequenced on Illumina XTEN® machines as previously described[1]. The remaining samples were sequenced on Illumina Htp NovaSeq 6000® machines using 150bp, paired end reads. Reads were aligned to the human reference genome (NCBI build37) using BWA-MEM.

Mutation calling and filtering

Substitutions

Base substitution calling was carried out in three steps: Discovery, filtering and genotyping. Mutations were first called using the Cancer Variants through Expectation Maximisation (CaVEMan) algorithm[2]. CaveMan uses a Bayesian classifier, incorporating base quality, read position, read orientation and more, to derive a posterior probability of all possible

genotypes at every candidate site. Out of concern about a field cancerization effect, patients 1 through 26, and patients from which only a few crypts were sequenced, were analysed using a matched normal sample dissected from non-epithelial tissue from one of the biopsies. As it became apparent that clones did not stretch between biopsies, we stopped sequencing non-epithelial tissue control samples from patients if crypts were dissected from multiple biopsies.

The substitution calls were next filtered, as previously[1] described, to remove mapping artefacts, common single nucleotide polymorphisms and calls associated with the formation of cruciform DNA structures during library preparation. When matched normal samples were unavailable for the calling (see above), a large number of rarer germline variants remained post filtering. All sites where a somatic mutation was called in any crypt from a given patient were subsequently genotyped in all other samples from that patient by constructing read pileups and counting the number of mutant and wild-type reads. Only reads with a mapping quality of 30 or higher, and bases with a base quality of 30 or higher, were counted.

We next performed an exact binomial test to remove germline variants. True heterozygous germline variants should be present at a variant allele frequency (VAF) of 0.5 in all samples from an individual. Across all samples from a given individual, we aggregated variant and read counts at sites where a single nucleotide variant was called in at least one sample. We then used a one-sided exact binomial test to distinguish germline variants from somatic variants. The null hypothesis was that germline variants were drawn from a binomial distribution with a probability of success of 0.5, or 0.95 for the sex chromosomes in men. The alternative hypothesis was that these variants were drawn from distributions with a lower probability of success. The resulting p-values were corrected for multiple testing using the Benjamini-Hochberg method. A variant was classified as somatic if $q < 10^{-3}$, or $q < 10^{-2}$ if fewer than five crypts had been

dissected for the patient. For variants classified as somatic, we fitted a beta-binomial distribution to the number of variant supporting reads and total number of reads across crypts from the same patient. For every somatic variant, we determined the maximum likelihood overdispersion parameter (ρ) in a grid-based way (ranging the value of ρ from 10^{-6} to $10^{-0.05}$). A low overdispersion captures artefactual variants because they appear to be randomly distributed across samples and can be modelled as being drawn from a binomial distribution. In contrast, true somatic variants will be present at at a VAF close to 0.5 in some, but not all crypt genomes, and are thus best represented by a beta-binomial with a high overdispersion. To distinguish artefacts from true variants, we used ρ=0.1 as a threshold, below which variants were considered artefacts. The code for this filtering approach is an adaptation of the Shearwater variant caller[3]. Finally, we filtered out variants that were supported by fewer than three reads or where the sequencing depth was less than five.

Indels

Short deletions and insertions were called using the Pindel algorithm[4]. We applied the same restrictions on median VAF and read counts as for substitutions, and germline indel calls were filtered using the same binomial filters as described above.

Structural variants

Structural variants were called using the BRASS algorithm (https://github.com/cancerit/BRASS) as previously described[1]. Calls were filtered using AnnotateBRASS (https://github.com/MathijsSanders/AnnotateBRASS) as previously described[5]. When a matched normal sample was not available for a patient, we used a clonally unrelated sample from the same individual to filter germline variants. All variants passing filters were

manually reviewed in a genome browser. For discovery of deletions at fragile sites of the genome, we manually reviewed the three regions in all the genomes.

Constructing phylogenic trees

We used the MPBoot software[6] to create a phylogenic tree for each patient. MPBoot uses ultrafast bootstrap approximation to generate a maximum parsimony consensus tree. We assigned mutations to branches using a maximum likelihood approach, removing mutations which didn't adhere to the tree structure ($P<0.01$, maximum likelihood estimation).

Mutation rate comparisons between IBD patients and controls.

Any test for a difference in mutation burden between cohorts must take into account all factors, biological and technical, which correlate with disease and/or affect mutation calling sensitivity. For our comparison of IBD and normal, we fitted a linear mixed effects model taking the following factors into account:

1. Age is the most important predictor of mutation burden and the age distribution of the two cohorts is different. We include a fixed effect for age in our model to account for this.
2. Mutation burden differs for different sectors of the colon[1]. The IBD cohort is enriched with samples from the left side, as this is the area predominantly affected in UC patients. We include a fixed effect for location within the colon to account for this.
3. Observations are non-independent. We include in the model random effects for patient and for biopsy, with the random effect for biopsy nested within that for the patient.

4. Most embryonic mutations will be removed by our filters as germline so at birth the mutation count is near zero. We therefore do not include a random intercept in our model but constrain the intercept to zero. The biological interpretation of this is that there are no somatic mutations present at birth.
5. The between-patient variance is likely greater in the IBD cohort as patients vary in the duration, extent and severity of their disease. The within-patient variance is also likely greater in the IBD cohort as biopsies taken from different sites of the colon vary in their disease exposure, number and duration of flares etc. To model this, we construct a general positive-definite variance-covariance matrix for the random effects of patient and biopsy by cohort.
6. Any difference in the clonality of the colon between IBD patients and controls will affect the relative sensitivity to detect somatic mutations. To account for this, we adjusted the branch lengths of the phylogenic trees and used the adjusted mutation counts as the response variable in our models. The adjustment was carried out as follows. Mutations with low variant allele frequencies (VAFs) will be missed at low coverage. Therefore, for each crypt, we first fitted a truncated binomial distribution to the VAF distribution of the crypt to estimate the true underlying median VAF (this is different from 0.5 because recent mutations may not yet have been fixed in the stem cell niche, and because of contamination of lymphocytes and other cells from the lamina propria, which do not carry the same somatic mutations as the epithelial cells). We next simulated 100,000 mutation call attempts by drawing the coverage of each call from a Poisson distribution, with the lambda set as the median coverage of the sample, and multiplying that with the median VAF estimate from the truncated binomial. The resulting value represents the

number of reads that carry the mutated allele. We calculated sensitivity for the sample, S_s, as the fraction of draws that resulted in four or more mutant reads, which is the number required by CaVEMan to call a mutation. The sensitivity of a branch with n daughter crypts, S_b, was then calculated as:

$$S_b = 1 - (1-S_{s1})*...*(1-S_{si})*...*(1-S_{sn})$$

The adjusted mutation count is thus the observed mutation count divided by the sensitivity of the branch. In this way, the mutation count of clones formed of stand-alone crypts is augmented more than that of branches with multiple daughter crypts. Even after these steps, a small but significant effect of coverage remained (beta=29, $P=4.7x10^{-7}$).

We compared this model with one which includes disease duration as a fixed effect using a likelihood ratio test. The disease durations for never inflamed regions of the colons of IBD patients was set to zero. We report the effects of the model without coverage as we believe that is the most accurate effect estimate but note that this makes little difference to our model (see accompanying code for full details of this analysis).

As comparatively few SVs are found in our dataset, we used Poisson regression within a generalized linear mixed effects framework to test for differences in SV number between cases and controls. We included the same random and fixed effects described above for base substitutions and compared models with and without disease duration using a likelihood ratio test. The above statistical tests are two-sided as are all statistical tests performed in this manuscript .

Mutational signature extraction and analyses

Mutational signatures were extracted using a hierarchical Dirichlet process. This has the advantage of allowing simultaneous fitting to existing signatures and discovery of new signatures. We pooled the control and the IBD data and extracted signatures as previously described[1], treating each branch of a phylogenic tree with more than 50 mutations as a sample, and using signatures reported in colorectal cancer as priors. We also included signature 32, which is attributed to azathioprine therapy[7], in the priors.

Selection analyses

To search for mutations under positive selection, we used the dNdScv method[8]. We included never inflamed samples from the IBD cohort in the analysis as some uncertainty existed regarding the annotation of a handful of never-inflamed biopsies and we estimated that our analysis would suffer more from potential exclusion of drivers than from inclusion of more neutral mutations. We used the Benjamini-Hochberg method to correct for multiple testing.

To look for enrichment of mutations in pathways we defined 15 gene-sets (Supplementary Table 5) as follows. We included all genes found to be under selection in colorectal cancer[8] and a list of 369 genes associated with any cancer compiled for our previous work[8]. We also chose a set of cellular pathways known to be important in IBD pathogenesis and epithelial homeostasis. The Reactome database was used to define the pathways[9], see Supplementary table 5 for accession dates and Reactome IDs of pathways. We chose the cytokine pathways TNF-Signaling, TNFR2, IL6, TGFb and IL17 for testing. We also defined a combined list of cytokines which included all of the above as well as IFNg, IL10, IL20, IL23, IL28, and IL36. We also decided to test other pathways shown by us and others through genome-wide association studies to be important in IBD pathogenesis[10,11]. These were Toll-like receptor

cascades, NOD-signaling, autophagy, unfolded protein response and epithelial cell-cell junctions. We included the PIP3/AKT signaling pathways as it is downstream of many of the pathways defined above and we had discovered two deletions affecting this pathway before performing the analysis. Finally, we defined a list of genes known to cause early-onset, monogenic forms of IBD. Many of the genes defined in the literature affect myeloid cell development and cause severe immunodeficiencies[12,13]. We restricted our analysis to the union of monogenic-IBD genes which either are specifically thought to affect epithelial cells or were members of any of the pathways above.

We extracted global dN/dS values for missense and truncating variants separately and used the Benjamini-Hochberg method to correct for multiple testing.

Method References

1. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
2. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
3. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).
4. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
5. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium.

bioRxiv 505685 (2018).

6. Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.* **18**, 11 (2018).
7. Alexandrov, L. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv* 322859 (2018).
8. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
9. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
10. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
11. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
12. Uhlig, H. H. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. *Gut* **62**, 1795–1805 (2013).
13. Uhlig, H. H. *et al.* The Diagnostic Approach to Monogenic Very Early Onset Inflammatory Bowel Disease. *Gastroenterology* **147**, 990–1007.e3 (2014).