Deep learning-based approach for the characterization and quantification of histopathology in mouse models of colitis.

Short Title: Deep learning-based quantification of histopathology in mouse models of colitis.

Soma Kobayashi[1], Jason Shieh[2], Ainara Ruiz de Sabando[3], Julie Kim[2], Yang Liu[2], Sui Y. Zee[4], Prateek Prasanna[1], Agnieszka B. Bialkowska[2], Joel H. Saltz[1,4], and Vincent W. Yang[*,1,2,5].

[1]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, United States.

[2]Department of Medicine, Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, United States.

[3]Department of Medical Genetics, Complejo Hospitalario de Navarra, Pamplona, Navarra, Spain.

[4]Department of Pathology, Renaissance School of Medicine at Stony Brook University, Stony Brook, NY, United States.

[5]Department of Physiology and Biophysics, Stony Brook University, Stony Brook, NY, United States.

*Corresponding Author: Vincent.Yang@stonybrookmedicine.edu, Tel.: (631) 444-2066.

Abstract

Inflammatory bowel disease (IBD) is a chronic immune-mediated disease of the gastrointestinal tract. While therapies exist, response can be limited within the patient population. As such, researchers have studied mouse models of colitis to further understand its pathogenesis and identify new treatment targets. Although bench methods like flow cytometry and RNA-sequencing can characterize immune responses with single-cell resolution, whole murine colon specimens are processed at once. Given the simultaneous presence of colonic regions that are involved or uninvolved with abnormal histology, processing whole colons may lead to a loss of spatial context. Detecting these regions in hematoxylin and eosin (H&E)-stained colonic tissues offers the downstream potential of quantifying immune populations in areas with and without disease involvement by immunohistochemistry on serially sectioned slides. This could provide a complementary, spatially-aware approach to further characterize populations identified by other methods. However, detection of such regions requires expert interpretation by pathologists and is a tedious process that may be difficult to perform consistently across experiments. To this end, we have trained a deep learning model to detect 'Involved' and 'Uninvolved' regions from H&E-stained colonic slides across controls and three mouse models of colitis – the dextran sodium sulfate (DSS) chemical induction model, the recently established intestinal epithelium-specific, inducible *Klf5*$^{\Delta IND}$ (*Villin-CreER*^{T2}*;Klf5*$^{fl/fl}$) genetic model, and one that combines both induction methods. The trained classifier allows for extraction of 'Involved' colonic regions across mice to cluster and identify histological classes. Here, we show that quantification of 'Involved' and 'Uninvolved' image patch classes in swiss rolls of colonic specimens can be utilized to train a linear determinant analysis classifier to distinguish between mouse models. Such an approach has the potential for revealing histological links and improving synergy between various colitis mouse model studies to identify new therapeutic targets and pathophysiological mechanisms.

43 **Introduction**

44 Inflammatory bowel disease (IBD) is a state of chronic intestinal inflammation and is comprised
45 of two major subtypes, Crohn's disease (CD) and ulcerative colitis (UC). IBD afflicts approximately
46 1.6 million Americans, and as many as 70,000 new cases are diagnosed each year (1). In addition
47 to intestinal symptoms such as abdominal pain and diarrhea, patients' quality of life can be
48 severely affected by extraintestinal manifestations such as arthritis, ankylosing spondylitis,
49 erythema nodosum, pyoderma gangrenosum, iritis, uveitis, and primary sclerosing cholangitis (2).
50 Furthermore, CD has been linked to increased risk for cancer, pulmonary, gastrointestinal, genital,
51 and urinary tract diseases (1), and UC is associated with an increased risk for colorectal cancer
52 (3). Classically, CD and UC have different patterns of intestinal involvement. CD pathology is
53 discontinuous and can affect any area along the gastrointestinal tract, while UC has a continuous
54 distribution that is limited to the colon (4). Both diseases can have uninvolved intestinal regions.
55 In CD, affected areas are termed 'skip lesions' as they are interspersed with uninvolved regions,
56 and in UC, disease often does not involve the whole large intestine.

57 Mouse models of colitis have been heavily studied to understand disease pathogenesis
58 and identify new treatment targets. One of the most used, in part due to the relatively simple
59 method of induction and replicability, has been the dextran sodium sulfate (DSS) chemical
60 induction model. DSS is provided to mice dissolved in drinking water for seven days to cause
61 acute colitis (5). Groups have utilized methods such as flow cytometry and single-cell RNA
62 sequencing to characterize the immune responses in these mice (6, 7). However, these methods
63 require the processing of whole murine colons at once. As DSS-treated mice have been reported
64 to have a predominantly distal injury pattern (8, 9), such an approach may not be sufficient to fully
65 capture intracolonic heterogeneity. In addition to this injury pattern, Kolachala et al. reported
66 different cytokine profiles between the proximal and distal colon (8). There is thus a likely benefit
67 to resolving differences between this distal niche from the typically uninvolved proximal colon in
68 DSS-treated mice. Additionally, our group has recently established the *Klf5*$^{\Delta IND}$ colitis mouse

69 model (10, 11). We have also observed the simultaneous presence of colonic regions that are
70 involved and uninvolved with disease in these mice. Therefore, we were motivated to develop an
71 automated method to histologically identify these areas across these mouse models.

72 Convolutional neural networks (CNNs) have shown much promise in biomedical image
73 classification tasks due to their ability to associate visual patterns with image labels (12-14).
74 Provided that many histopathological labels are defined in part by visual patterns that pathologists
75 have consistently seen for certain clinical diseases, CNNs are well-suited for such tasks. In
76 practice, histological slides are digitized into gigapixel resolution whole slide images (WSIs). Due
77 to the size of these WSIs, they are broken up into smaller, equally-sized image patches to which
78 CNNs are applied. In this study, we sought to address whether CNNs could be incorporated in a
79 pipeline to characterize intracolonic heterogeneity within mouse models. Although preclinical
80 animal models are frequently utilized to identify novel protein and genomic targets, use of
81 histology is often limited to confirming presence or lack of certain markers without a focus on the
82 morphological and cellular patterns. As with the 'villous' or 'adenomatous' polyps, histological
83 patterns are clinically significant and a manifestation of cellular and sub-cellular
84 pathophysiological mechanisms. There is therefore a likely value to characterizing histological
85 patterns in mouse models and tying their presence to molecular characterizations. Adaptation of
86 histological phenotypes to routinely collected clinical specimens has the potential to transfer
87 knowledge gained from animal models for more spatially-granular characterizations in a non-
88 invasive manner.

89 A CNN pipeline to detect intracolonic heterogeneity could provide a histological readout
90 to aid in the characterization of colitis mouse models and the evaluation of treatment efficacies.
91 While this could be accomplished via pathologist's interpretation, doing so for many samples is
92 tedious and may be difficult to perform in a consistent manner across experiments and over time.
93 Additionally, reliable identification of such regions is the first step towards correlating
94 immunohistochemically identified cell populations to regions with and without abnormal histology.

95 This could provide a spatial context-aware characterization of immune responses. Lastly, as we
96 show in this study, a classifier trained for this task can extract diseased image regions across a
97 cohort of mice. Involved region extraction allows for clustering approaches to identify histological
98 classes, the variable presence of which can be utilized to distinguish between mouse models with
99 an overall F1 score of 95.75%.
100

101 **Results**

102 **Archived Mouse Cohort Background**

103 To train our models, we utilized archived, formalin-fixed paraffin-embedded (FFPE) swiss-
104 rolled colons (15). These were collected from three colitis mouse models and appropriate controls.
105 The colitis model treatment schedules are detailed in S1A Fig. The first mouse model is the
106 recently established *Klf5*$^{\Delta IND}$ (*Villin-CreER*^{T2}*;Klf5*$^{fl/fl}$) genetic model, wherein inducible intestinal
107 epithelium-specific knockout of *Klf5* upon five days of intraperitoneal (IP) tamoxifen (TAM)
108 injections disrupts epithelial barrier function and causes colitis (5T-*Klf5*$^{\Delta IND}$) (10, 11). For this study,
109 we utilized female mice due to the higher rate of *Klf5* knockout relative to males (11). As TAM is
110 dissolved in corn oil (CO), we also performed five days of IP CO injections in *Klf5*$^{\Delta IND}$ mice as a
111 control (5C-*Klf5*$^{\Delta IND}$). Additionally, to provide controls for the genotypes and the TAM injections,
112 we have also collected various combinations of controls with five days of IP CO and five days of
113 IP TAM injections across *Klf5*$^{\Delta IND}$, *Klf5*$^{\Delta IND/+}$ (*Villin-CreER*^{T2}*;Klf5*$^{fl/+}$), and *Klf5*^{WT} mice (*Villin-*
114 *CreER*^{T2}*;Klf5*$^{+/+}$) (S1 Table).

115 The second mouse model is the chemical dextran sodium sulfate (DSS) induction model.
116 To control for the injections in our 5T-*Klf5*$^{\Delta IND}$ model, *Klf5*$^{\Delta IND/+}$ mice received five days of IP CO
117 injections, which were then followed by seven days of 2.5% DSS in drinking water. We have seen
118 that the histology mirrors that in non-injected *Klf5*^{WT} mice treated with DSS (S1B Fig). For the third
119 model, we performed a combined induction. As homozygous 5T-*Klf5*$^{\Delta IND}$ mice exhibit increased

120 mortality in an 18-day period after induction (11), we injected heterozygous *Klf5*$^{\Delta IND/+}$ mice with
121 TAM for five days IP, then followed with seven days of 2.5% DSS in drinking water (5T-*Klf5*$^{\Delta IND/+}$
122 + DSS). Although we have yet to fully characterize the combined induction model, we wanted to
123 examine the histology in these mice and evaluate whether our computational pipeline could still
124 distinguish the single-type induction models even in the presence of these mice in the cohort. In
125 total, we have 48 mice in the archived mouse cohort, and sample numbers are available in S1
126 Table.

127 Of note, the colons extracted from these colitis models contain areas that are involved or
128 uninvolved with disease. Manually selected regions from representative whole slide images
129 (WSIs) are shown in Fig 1A. Swiss rolls allow for the visualization of a whole mouse colon on a
130 single glass side. In these preparations, the center is the proximal end, and the colon tissue can
131 be traced distally towards the outer portion of the swiss roll. Although not utilized in this study, we
132 have also provided WSIs of additional controls and comparisons (5C- *Klf5*$^{\Delta IND/+}$ and 5T- *Klf5*$^{\Delta IND/+}$)
133 for the combined induction model (S1C Fig). Clinical scoring metrics combining weight loss, stool
134 consistency, and fecal blood (16) show significant increases for 5T-*Klf5*$^{\Delta IND/+}$ + DSS mice relative
135 to 5C- *Klf5*$^{\Delta IND/+}$ + DSS and control mice (S1D Fig). Combined induction mice trend lower on
136 histological scoring relative to 5C- *Klf5*$^{\Delta IND/+}$ + DSS mice (S1E Fig), mainly due to decreased extent
137 of ulceration along the colon.

138

139 **Figure 1. Colitis Mouse Models Exhibit Regional Heterogeneity.** A) Representative whole
140 slide images (WSIs) of swiss-rolled colons from *Klf5*^{WT} (*Villin-CreER*^{T2}*;Klf5*$^{+/+}$) control and colitis
141 mouse models. Manually selected regions that are 'Involved' (red) and 'Uninvolved' (green) are
142 shown for colitis mouse model samples. 8, 11, and 16 show areas with crypt dropout. 7 and 12
143 are examples of crypt dilation. 15 is an example of distorted glands. B) Reported patterns of injury
144 in literature for colitis mouse models.

145

While the regional heterogeneity in our colitis models is visually apparent (Fig 1A), identification of these regions requires a pathologist's inspection. This can be difficult at times to perform objectively and especially over many samples. Pathologist inter- and intra-observer variability has been reported in various contexts, such as lymph node counting (17), atypical ductal hyperplasia diagnosis (18), and follicular thyroid carcinoma diagnosis (19). An automated pipeline to identify murine colonic regions involved and uninvolved with disease would provide benefits in objective reproducibility and speed. Furthermore, reliable, automated identification of these areas may open the door for more spatially motivated, histological characterizations. A goal of this pipeline is to capture the reported proximal and distal injury patterns in the 5T-*Klf5*$^{\Delta IND}$ (11) and DSS models (8, 9), respectively (Fig 1B). Following detection, involved image regions across mice can be extracted to cluster and reveal pathologies specific to mouse models. To this end, we have trained an 'Involved' versus 'Uninvolved' region classification model that generates predictions over mucosal and submucosal regions of swiss-rolled murine colonic tissues.

'Involved' versus 'Uninvolved' Classifier Training

For our classifier, we used the ResNet-34 (RN-34) architecture (20). ResNet is considered a landmark architecture for the introduction of skip connections. Deep networks introduce a high number of non-linear mathematical operations through the addition of many layers, which can complicate the gradient descent operations that allow the neural network to "learn". Gradient explosion and vanishing can occur when calculated gradients are too large or small and will impede learning (21). Skip connections allow gradient descent to skip layers of the network where this occurs, allowing for continued training. This has allowed for a deep, 34-layer ResNet that can extract even higher dimensional features from images, a powerful quality for H&E-patch classification. In this study, we trained RN-34 models that were pretrained on ImageNet, a large dataset with natural images comprising 1000 classes (22).

171 Our classifier was trained in a two-phase approach (S2A-B Fig). All WSIs across our
172 archived mouse cohort were downsampled by a factor of 8 and tiled into equally sized 224x224
173 pixel patches. The initial phase trained a classifier with ground truth patch labels corresponding
174 to mouse colitis status (S2A Fig). As such, all patches from colitis mice were labeled 'Colitis' and
175 those from control mice were labeled 'Control'. Since this labeling process does not account for
176 the intracolonic heterogeneity observed in our colitis models (Fig 1A), we performed further
177 feature extraction to extract the high dimensional patch representations learned during the first
178 phase of training (S2B Fig). K-means clustering on these representations across all archived
179 mouse cohort samples identified five patch classes (S2C-D Fig). Three of these classes ('Crypts',
180 'Lightly Packed', and 'Rosettes') were visually determined and validated with a pathologist to be
181 uninvolved, while two of these classes ('Mixed Pathology', 'Distorted Glands') were involved with
182 abnormal histology.

183 In addition to the visual evaluations, we calculated the proportion of each mouse's patches
184 that corresponded to each of these classes. Two of the three qualitatively uninvolved classes
185 were significantly enriched in our control mice, while both of our involved classes were
186 significantly enriched in colitis mice (S2D Fig). The lack of significance for the 'Rosettes' class is
187 attributed to the higher proportion of patches of this class in DSS-treated mice relative to the other
188 colitis models (S2E Fig). This is likely due to the predominantly distal injury pattern in DSS-treated
189 mice (8, 9), as rosette structures are more prevalent in the proximal colon where abnormal
190 histology is less common. Since mucosal folds are found in proximal murine colon (23, 24), the
191 increased prevalence of rosette structures is due to sectioning at these regions. In the second
192 phase of training (S2B Fig), these k-means classes were utilized as patch-level ground truth labels
193 based on their qualitative descriptions and quantitative enrichment in colitis or control mice.

194 In both phases of training, 70% of total patches were used to train the model, 10% as a
195 mid-training performance validation set, and 20% as a held-out test set. Our final model exhibited
196 an overall F1 score of 90.1% when predictions were compared to our method-generated ground

truth labels on the held-out test set (Fig 2A). This F1 score was an improvement from the overall F1 score of 77.6% for the model from the first phase of training with colitis status ground truth labels (S2F Fig). The first phase model exhibited a relatively lower F1 score of 68.87% when attempting to predict patches from colitis mice. This performance is likely explained by the presence of uninvolved patches extracted from our colitis mice labeled as 'Colitis' and motivated our second phase of model training. Additionally, when applied to 171 patches with 'Involved' and 'Uninvolved' ground truth labels provided by a pathologist, our classifier exhibited an overall F1 score of 81.41% and again showed improvement from the model from the first phase of training (S2G Fig).

Figure 2. H&E Model Performs 'Involved' versus 'Uninvolved' Classifications on Swiss-Rolled Colons. A) Confusion matrix showing final model predictions on patches from test set mice. From the 48 mice in our archived cohort (S1 Table), 34 mice (14281 patches, ~70% of total) were used for model training, 5 mice were used for mid-training validation (2192 patches, ~10% of total), and 9 mice (4171 patches, ~19%) were used as a final independent test set. B) Original input H&E-stained swiss rolls are shown with corresponding 'Involved' (red) versus 'Uninvolved' (green) classifier overlays for test set mice from archived mouse cohort. C) Original input H&E-stained swiss rolls and corresponding overlays for mice from prospectively treated mouse cohort. For both B) and C), red and green overlay colors correspond to regions with at least 50% model prediction confidence.

Small Patch Classifier Training for Preprocessing

We initially used simple thresholding of pixel intensities across each patch's red, green, and blue color channels to filter out patches with too much background and muscle. However, we observed that this approach was not robust and trained an additional RN-34 classifier pretrained on ImageNet on smaller 32x32 pixel patches to classify between 'Background', 'Muscle', 'Tissue',

and ‘Submucosa’ classes (S3 Fig). From 8 mice (2 each of control and of the three mouse models), we extracted 100 patches for each of the ‘Background’, ‘Muscle’, ‘Tissue’, and ‘Submucosa’ classes (S3A Fig). Our model was trained with these ground truth labels and applied to a test set of patches from 4 different mice as training (1 each of control and of the 3 mouse models, 100 patches for each class). This model achieved an overall F1 score of 93.5% (S3B Fig). We also generated qualitative overlays of our small patch classifier’s prediction on test set mice (S3C Fig). Additionally, we opted to implement this small patch classifier within our patch filtering process. All patches with >65% of area corresponding to regions predicted to be ‘Background’ or ‘Muscle’ are filtered out during the patch extraction process (S4 Fig).

‘Involved’ versus ‘Uninvolved’ Overlay Generation

To provide visual outputs for classifications, we generated overlays for our held-out test set mice (Fig 2B). Areas with at least 50% confidence of ‘Involved’ and ‘Uninvolved’ predictions are overlayed onto the input H&E WSIs. Overlapping patches were extracted from all downsized WSIs. This was done by taking initially extracted patch locations and iterating 20 pixels in each direction (up/down/left/right). Overlapping patch extraction led to an increase of ~450 to ~200,000 patches per mouse after repeating filtering to eliminate patches with too much muscle or background. Our classifier was applied to each overlapping patch, and ‘Involved’ and ‘Uninvolved’ prediction confidences were averaged at every pixel in the WSI. All regions predicted by our Small Patch Classifier to be ‘Background’ and ‘Muscle’ were eliminated from final overlays.

Prospective Mouse Cohort

To test the robustness of our ‘Involved’ versus ‘Uninvolved’ classifier, we treated and collected H&E-stained, FFPE swiss-rolled colons from additional mice. Specifically, this cohort allowed us to assess whether our approaches would work regardless of who treated mice, collected swiss rolls, and performed H&E staining. To complement the 48 mice in our archived

249 mouse cohort, we treated 24 additional mice. This consisted of eight controls (no injections, put
250 on normal drinking water), eight *Klf5*$^{\Delta IND/+}$ DSS-treated mice with no injections, five 5T-*Klf5*$^{\Delta IND}$,
251 and three 5T-*Klf5*$^{\Delta IND/+}$ + DSS mice (S1 Table). We elected to utilize non-injected DSS-treated
252 mice to confirm that our model properly captures the histology in these mice even without five
253 days of corn oil injections preceding the DSS. We applied our approach to these prospective
254 samples and successfully generated 'Involved' versus 'Uninvolved' overlays (Fig 2C).
255
256 **Patch Class Discovery**
257 To build upon the binary classifications generated by our final model, we aimed to define
258 more specific image patch classes. Analogous to our second phase of model training (S2B Fig),
259 we utilized our final 'Involved' versus 'Uninvolved' classifier to extract high dimensional
260 representations for patches in our dataset. Specifically, two datasets (one with all 'Involved' patch
261 representations and one with all 'Uninvolved' patch representations) were collected. Given the
262 high number of patches, principal component analysis (PCA)-based dimensionality reduction was
263 then performed on each set of patch representations before k-means clustering. Clustering on
264 'Uninvolved' patches from our archived mouse cohort identified three classes: 'Crypts' – test-tube-
265 like perspective of crypt structures, 'Lightly Packed' – more space between crypts often
266 accompanied by immune cell nuclei, and 'Rosettes' – Ring-like cross-section perspective of crypts
267 (Fig 3A-B). Of note, the discovered 'Uninvolved' patch classes matched those utilized as ground
268 truth labeling during our second phase of classifier training (Fig S2C).
269
270 **Figure 3. Discovered 'Uninvolved' Patch Classes are Enriched in Control Mice and Show**
271 **Shifts in Prevalence Across Colitis Mouse Models.** A) 'Uninvolved' patch classes with
272 qualitative class labels. Patches most representative of class labels are shown. B) One-way
273 ANOVA shows that these 'Uninvolved' patch classes are found in differing portions across colitis
274 mouse models and controls. *$p<0.05$, **$p<0.01$, ***$p<0.001$, ****$p<0.0001$.

275 While the 'Crypts' class was significantly enriched in control mice relative to all the colitis
276 mouse models, this was not true for the other two classes. Specifically, the 'Rosettes' class was
277 significantly enriched in control mice relative to 5T-*Klf5*$^{\Delta IND}$ and 5T-*Klf5*$^{\Delta IND/+}$ + DSS mice, but not
278 in 5C-*Klf5*$^{\Delta IND/+}$ + DSS mice (Fig 3B). In addition, the 'Rosettes' class was significantly enriched in
279 the 5C-*Klf5*$^{\Delta IND/+}$ + DSS mice relative to the other two colitis models. This is again likely due to the
280 predominantly distal injury pattern in DSS-treated mice (8, 9). This injury pattern is also depicted
281 in our 'Involved' versus 'Uninvolved' overlays (Fig 2B-C). Our DSS-treated mice show more
282 'Uninvolved' regions in the rosettes-enriched proximal colon (center of swiss rolls) where disease
283 is less common. In the 5T-*Klf5*$^{\Delta IND/+}$ + DSS mice, however, DSS-treatment no longer causes a
284 distal predominant injury pattern (Fig 2B-C). While the mechanism of DSS-induced colitis has not
285 been completely elucidated, the effects are believed to be dependent on tissue penetration of the
286 chemical leading to disruption of the intestinal epithelial monolayer and barrier integrity (25). DSS
287 has a variable molecular weight from 5 to 1400 kDa, and administration of forms of 500 kDa and
288 higher do not induce colitis (26). As such, the shift in the distal injury pattern of 5C-*Klf5*$^{\Delta IND/+}$ + DSS
289 mice to the proximal one in 5T-*Klf5*$^{\Delta IND/+}$ + DSS mice may be a function of increased proximal
290 tissue permeability from TAM injection-induced heterozygous knockout of KLF5 that promotes
291 greater tissue penetrance of DSS.

292 The 'Lightly Packed' class was significantly enriched in control mice relative to 5C-
293 *Klf5*$^{\Delta IND/+}$ + DSS mice but not the other two colitis mouse models. This likely relates to the
294 increased prevalence of abnormal proximal histology in the 5T-*Klf5*$^{\Delta IND}$ and 5T-*Klf5*$^{\Delta IND/+}$ + DSS
295 models relative to 5C- *Klf5*$^{\Delta IND/+}$ + DSS mice (11) (Fig 2B-C). However, the 'Crypts', which are
296 found more distally, significantly decreased in all colitis mouse models relative to controls. Thus,
297 an additional explanation is that the 'Lightly Packed' regions in these mice sit on the decision
298 border, representing an accumulation of immune cells without sufficient levels of abnormal
299 histology to garner an 'Involved' prediction. While our 5T-*Klf5*$^{\Delta IND/+}$ + DSS model exhibited a
300 significantly higher clinical score than our 5C-*Klf5*$^{\Delta IND/+}$ + DSS mice (S1D Fig), they trended lower

301 in histological scoring (S1E Fig). This may reflect the limitation of histological scoring schemes
302 that rely upon simple detection and counting of pre-defined pathological findings without
303 accounting for lower grade abnormalities in other regions. Thus, in our 5T-*Klf5*$^{\Delta IND}$ and 5T-
304 *Klf5*$^{\Delta IND/+}$ + DSS mice, the presence of 'Lightly Packed' patches may indicate an immune response
305 that is present but either protective or insufficient to generate overt abnormalities. A future
306 direction is thus to leverage this property of our pipeline to consider these borderline regions in
307 the grading of swiss rolls.

308 The 'Mixed Pathology' k-means class identified during our second phase of model training
309 (S2B-D Fig) motivated us to apply this approach to 'Involved' patches from across the archived
310 mouse cohort to identify specific types of abnormal histology. K-means clustering on these
311 'Involved' patches identified four pathologic patch classes (Fig 4A-B). By visual inspection and
312 validation by a pathologist, these were classified as 'Inflammatory' (Milder, more heterogeneous
313 histological phenotype with an influx of immune cell nuclei), 'Crypt Dropout' (loss of epithelium
314 and displacement by stroma and immune cells), 'Crypt Dilation' (expansion of crypt lumen spaces),
315 and 'Distorted Glands' (distortion of crypt structures). Of note, the 'Distorted Glands' class
316 persisted from the ground truth classes used in our second round of model training (S2C-D Fig),
317 but the 'Mixed Pathologies' class was replaced by the other classes. Notably, immune cell
318 infiltration, crypt dilation, and gland distortion are all histological features that have been observed
319 in human IBD (27-29). Furthermore, these 'Involved' patch classes were found to be present in
320 significantly different proportions across our mouse models and controls (Fig 4B). Specifically,
321 5C-*Klf5*$^{\Delta IND/+}$ + DSS mice were enriched in the 'Inflammatory' and 'Crypt Dropout' classes, 5T-
322 *Klf5*$^{\Delta IND}$ mice in the 'Inflammatory' and 'Crypt Dilation' classes, and 5T-*Klf5*$^{\Delta IND/+}$ + DSS mice in
323 the 'Crypt Dropout' and 'Distorted Glands' classes (Fig 4B).

324

325 **Figure 4. Discovered 'Involved' K-Means Patch Classes Can Be Utilized by Machine**
326 **Learning Classifier to Predict Mouse Model.** A) 'Involved' patch classes with qualitative class

327 labels. Patches most representative of class labels are shown. B) One-way ANOVA shows that
328 these 'Involved' patch classes are found in differing portions across colitis mouse models and
329 controls. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. C) Stacked bar plot of total 'Uninvolved'
330 and 'Involved' k-means patch class proportions across archived mouse cohort. Error bars show
331 mean with standard deviation. D) Confusion matrix for linear determinant analysis (LDA) classifier
332 mouse model prediction on prospective mouse cohort. LDA model was trained on the original 48
333 mice cohort of archived tissues.
334
335 Visualizing patch class proportions by stacked bar plots helps capture and summarize
336 these findings (Fig 4C). While all three 'Uninvolved' patch classes are present in control mice,
337 'Crypts' patches fall significantly for all three colitis mouse models, while 'Lightly Packed' patches
338 persist. However, the presence of 'Rosettes' are maintained in our 5C-*Klf5*$^{\Delta IND/+}$ + DSS mice. The
339 'Inflammatory' and 'Crypt Dilation' patch classes are most prevalent in our 5T-*Klf5*$^{\Delta IND}$ mice. Given
340 the shorter treatment course of this colitis mouse model (5 days), these patch classes may
341 indicate a more acute finding. On the other hand, the 'Crypt Dropout' class may be less acute as
342 prevalence is higher in our longer treatment course 5C-*Klf5*$^{\Delta IND/+}$ + DSS (7 days) and 5T-*Klf5*$^{\Delta IND/+}$
343 + DSS mice (12 days). Similarly, our 5T-*Klf5*$^{\Delta IND/+}$ + DSS mice, which have 12 days of total colitis
344 induction (five days of TAM injections, seven days of DSS, S1 Table), are the only model with
345 appreciable presence of the distorted glands class. Distorted glands are one of the histological
346 markers of chronic inflammation in human IBD (30) and exemplifies how this type of pipeline can
347 histologically categorize colitis mouse models by which facets of human disease they recapitulate.
348 In addition, partnering this approach with immunohistochemical staining can shed further light on
349 the relative contributions of length of colitis induction and immune cell presence in causing various
350 types of histological abnormalities.
351
352 **Prediction of Mouse Model from H&E Inputs**

353 Finally, we sought to address whether the variable presence of 'Uninvolved' patches and 'Involved'
354 patch classes could be utilized to predict mouse model in our prospective mouse cohort. We thus
355 formed the pipeline detailed in S5A Fig to utilize 'Uninvolved' patch and 'Involved' k-means patch
356 class proportions to train a linear determinant analysis (LDA) classifier to predict mouse model.

357 The LDA classifier was trained on the archived mouse cohort. When applied to the
358 prospective mouse cohort via the inference pipeline in S5B Fig, this LDA classifier predicted which
359 mouse model each swiss roll came from with an overall F1 score of 95.75% (Fig 4D). Of note, the
360 stacked bar plots of k-means patch classes in the prospective cohort (S5C Fig) mirrored what we
361 observed in the archived mouse cohort (Fig 4C). As one of the 5T-*Klf5*$^{\Delta IND}$ mice was classified as
362 control, we examined the input H&E and found healthier appearing histology relative to properly
363 classified 5T-*Klf5*$^{\Delta IND}$ mice. This was apparent even from low magnification with decreased luminal
364 space areas and more tightly packed crypts (S6 Fig). Although some 'Involved' areas were
365 present in the 5T-*Klf5*$^{\Delta IND}$ mouse classified as control, crypts were more intact in these regions
366 and accompanied with goblet cell presence. Goblet cells secrete mucus to provide a protective
367 lining in the intestine, and decreased goblet cells are observed in CD and UC (31). As such, this
368 pipeline can histologically identify outlier mice with weaker colitis induction and phenotypes.

369 **Discussion**

370 This study sought to address the simultaneous presence of colonic regions that are
371 histologically involved and uninvolved with abnormal histology within colitis mouse models. While
372 these regions can be identified via careful inspection by a pathologist, doing so objectively over
373 many samples is difficult and highly time-consuming. Deep learning models learn to associate
374 class labels with implicit patterns in data and are well-suited for this task. A computational method
375 to quantify the presence of abnormal histology can also serve as a histological readout for
376 experiments performed on mice. One example would be to confirm the effect of treatments

377 histologically as an accompaniment to clinical metrics like body weight, fecal blood, or stool
378 consistency.

379 We have thus presented an 'Involved' versus 'Uninvolved' classifier for colons of DSS-
380 treated, 5T-*Klf5*$^{\Delta IND}$, 5T-*Klf5*$^{\Delta IND/+}$ + DSS, and control mice. Specifically, we have utilized a two-
381 phase training approach. The first phase uses mouse colitis status as patch ground truth labels,
382 while the second phase incorporates feature extraction and clustering to generate improved,
383 patch-level ground truth labeling. Importantly, our final model shows improved agreement with a
384 pathologist relative to the model trained in just the first phase of the approach.

385 In a related study, Bédard et al. published a proof of concept application of convolutional
386 neural networks for the microscopic scoring of acute inflammation in DSS colitis (32). They utilized
387 a commercial artificial intelligence platform to detect muscle, normal mucosa, and acutely
388 inflamed mucosa in H&E-stained murine colon WSIs. The authors calculated the ratio of acutely
389 inflamed to total mucosa to assess the level of disease. Identifying these regions is an important
390 step towards the microscopic grading of colitis samples. We have similarly trained our Small Patch
391 Classifier to detect 'Background', 'Muscle', 'Mucosa', and 'Submucosa' classes from 32x32 pixel
392 patches. In our work, the Small Patch Classifier was implemented to improve our patch filtering
393 process and to focus our 'Involved' versus 'Uninvolved' classifier overlays onto the 'Mucosa' and
394 'Submucosa' regions. A future direction is to explore whether the Small Patch Classifier outputs
395 can improve our 'Involved' versus 'Uninvolved' model performance.

396 In an alternate approach, Rogers et al. attempted to segment colitis lesions in H&E-stained
397 WSIs across the DSS, CD45RBHi, and IL-10$^{-/-}$ mouse models to grade disease severity (33). The
398 authors experienced segmentation challenges due to variabilities in crypt morphologies and opted
399 to generate a workflow based on CD3 immunohistochemical staining instead. Morphological
400 variabilities likely increase more with disease relative to healthy histology. Here, we chose to train
401 a patch-based 'Involved' versus 'Uninvolved' classifier. Including the task to classify 'Uninvolved'
402 regions that should be more homogenous in morphology may help to separate out the 'Involved'

403 regions. An additional benefit is that this has allowed us to perform clustering on just the 'Involved'
404 patches to identify specific types of abnormal histology. Clustering on 'Uninvolved' patches also
405 generated classes of normal histology. As with Rogers et al., we have also incorporated multiple
406 mouse models of colitis in our study. This has allowed for comparisons showing quantitative
407 enrichment of pathologic classes in the different mouse models.

408 Specifically, we applied feature extraction, PCA-based dimensionality reduction and k-
409 means clustering on 'Uninvolved' patches to define 'Crypt', 'Rosette', and 'Lightly Packed' classes.
410 Patches of the 'Crypt' class were significantly decreased in all colitis mouse models. The 'Rosette'
411 and 'Lightly Packed' classes were present in variable proportions across mouse models, mirroring
412 shifts in proximal and distal predominant patterning of injury. As the 'Lightly Packed' classes in
413 colitis mice may represent regions of immune infiltration without overt abnormal histology, we plan
414 to assess the capacity of our model to incorporate these regions in the grading of colonic histology.

415 From 'Involved' patches, we identified four classes of abnormal histology – 'Inflammatory,
416 'Crypt Dropout', 'Crypt Dilation', and 'Distorted Glands'. DSS-treated mice were enriched in the
417 'Inflammatory' and 'Crypt Dropout' classes, 5T-*Klf5*$^{\Delta IND}$ mice in the 'Inflammatory' and 'Crypt
418 Dilation' classes, and 5T-*Klf5*$^{\Delta IND/+}$ + DSS in the 'Crypt Dropout' and 'Distorted Glands' classes.
419 The proportion of patches per swiss roll classified as 'Uninvolved' and the four 'Involved' classes
420 for each mouse was sufficient to train an LDA classifier to predict mouse model. Increased
421 prevalence of the 'Inflammatory' and 'Crypt Dilation' classes in the shorter time course 5T-*Klf5*$^{\Delta IND}$
422 (5 days) and DSS-treated (7 days) mice may indicate these are more acute findings. Increases
423 of 'Distorted Glands' in the combined induction model (12 days) and of 'Crypt Dropout' in the DSS-
424 treated mice and combined induction model may reflect the more chronic nature of these
425 histological classes. To the best of our knowledge, this is the first study to explore quantitation of
426 pathologies across mouse models to invite further assessment of histology beyond qualitative
427 descriptions. Additionally, this may help validate the capacity of mouse models to capture specific
428 aspects of human disease. Detection of specific histologic abnormalities will likely provide finer

429 grained characterizations for improved mappings to human IBD heterogeneity. To allow others to
430 utilize the code, we have made available our prospective mouse cohort images and the full
431 inference pipeline from WSI scaling and patch extraction to LDA inference.

432 Our pipeline thus essentially examines swiss rolls, detects and categorizes histology, then
433 predicts mouse model. As the 5T-*Klf5*$^{\Delta IND}$ mice have a Th17-mediated immune response (11),
434 while the DSS model has been characterized as Th1-, Th17-, and innate immunity-mediated (6,
435 34, 35), we have confirmed that the variable immune responses in these models are accompanied
436 by differences in histological pathologies. The next step is to delineate the respective contributions
437 of various immune populations to presence of specific pathologies.

438 Accordingly, we believe this pipeline is the start for a spatially motivated characterization
439 of immune populations by histologic localization to 'Involved' and 'Uninvolved' regions. While
440 methods like flow cytometry and single-cell sequencing can characterize immune responses with
441 single-cell resolution (6, 7), whole tissues must be processed at once and spatial context is
442 sacrificed. As a complement to these methods, a future direction is to map immune populations
443 stained on serially sectioned slides to this regional heterogeneity as another approach to
444 assessing immune cell functionality and subtypes. Residence of immune populations in areas
445 with and without disease may provide another angle towards characterization. Since many studies
446 assessing functionality rely on knockouts or expensive neutralizing antibody experiments (11, 36-
447 40) , such a pipeline may offer a cheaper, quicker alternative.

448 One limitation of this method is that current application is restricted to the mouse models
449 included in this study. However, we believe this is an important first step to establish and promote
450 the potential of computer vision methodologies at the bench, and specifically, within the context
451 of colitis mouse models. Another future direction is therefore to amplify the capacity of this method
452 by incorporating other colitis models within training, such as the 2,4,6-Trinitrobenzenesulfonic
453 acid (TNBS), IL-10 knockout, adoptive cell transfer, and oxazolone models (41). Doing so would

allow for the application of our method as a histological readout over a wider range of colitis mouse models.

More importantly, 'Involved' regions could then be extracted over an even more heterogeneous collection of colitis phenotypes. As these colitis models across literature differ in induction and flavor of immune responses, such an approach would open the door for synergy of knowledge gained from different mouse models. One possibility is that immune response characterizations, molecular approaches, and mechanistic studies focusing on pathophysiology could be linked to similar and dissimilar presence of types of histological findings across different murine colitis phenotypes. Notably, protein and genomic targets are often identified in preclinical animal models then evaluated for clinical value in patient specimens. Histological patterns identified in these animal models are driven by these same protein and genomic targets. Consequently, a future goal is to evaluate the clinical value and transferability of such histological phenotypes identified in animal models. This approach, therefore, has the potential to one day promote the discovery of novel therapeutic targets and pathways in IBD by serving as link between studies. As such, we believe that the integration of computational and computer vision approaches offers significant and exciting potential in bringing together the depth of human knowledge that has been gained from across colitis mouse models and in encouraging further collaboration across groups.

Materials and Methods

Mice

All studies involving mice were approved by the Stony Brook University Institutional Animal Care and Use Committee (IACUC; protocol #354918 and 245094). All mice carried an inducible Cre recombinase gene under the *Villin* promoter. These mice carried either two wild-type alleles of *Klf5* (*Villin-CreER*^{T2}*;Klf5*$^{+/+}$) or were heterozygous or homozygous for an additional *Klf5* allele

479 flanked by loxP sites (heterozygous: *Villin-CreER*^{T2}*;Klf5*$^{\Delta IND/+}$, homozygous: *Villin-CreER*^{T2}*;*
480 *Klf5*$^{\Delta IND/\Delta IND}$).

481 Treatment schedules are detailed in S1A Fig and all treatment conditions are available in
482 S1 Table. For all experiments, mice were eight to ten weeks old and female, as we have seen
483 increased Cre recombination efficiency in these mice relative to males (11). Intraperitoneal (IP)
484 injection of tamoxifen at 1mg/day dissolved in corn oil for 5 days was performed as previously
485 described (11). Control mice received 5 days of IP corn oil injections. For dextran sodium sulfate
486 (DSS) treatment, mice in the archived cohort received 2.5% DSS in drinking water. In the
487 prospective cohort, DSS-treated mice received 3% DSS, as we observed this was the optimal
488 concentration for experiments at the Stony Brook facilities. Combined induction mice received
489 2.5% DSS following 5 days of IP injection of tamoxifen at 1mg/day dissolved in corn oil.

490

491 **Whole Slide Image (WSI) Generation**

492 FFPE swiss-rolled colons were sectioned onto glass slides and stained by hematoxylin and eosin
493 (H&E) (15). These glass slides were then scanned and digitized at 40X magnification (0.17
494 μM/pixel) to a .vsi format by the Olympus VS120 Digital Virtual Slide System (VS120-L100-W).
495 The files were converted to a tiff format for further downstream use.

496

497 **H&E 'Involved' versus 'Uninvolved' Classifier Training**

498 For this study, we trained the ResNet-34 (RN-34) neural network, which is considered a
499 landmark architecture for its introduction of skip connections (20). Briefly, skip connections allow
500 the model to 'skip' parts of its architecture where training would typically be impeded due to
501 gradient explosion or vanishing. We trained a RN-34 model that was pretrained on ImageNet, a
502 large dataset with natural images comprising 1000 classes (22).

503 An overview schematic of our two-phase training approach is available in S2A-B Fig. The
504 initial step is to downsize our WSIs by a factor of 8 and extract 224x224 pixel patches. Patches

505 containing too little tissue are filtered out. The first phase, outlined in S2A Fig, utilizes mouse-
506 level colitis induction status as ground truth to label extracted patches. As such, while our colitis
507 mice show colonic presence of regions that are involved and uninvolved with disease (Fig 1A),
508 this is not addressed here. Instead, all patches from mice with colitis induction are provided a
509 'Colitis' label, while all patches from control mice are provided a 'Control' label. Once labeled,
510 70% of patches were used to train the model, 10% of patches were used assess to performance
511 mid-training, and 20% of patches were used as a held-out test set.

512 With the goal of generating more granular, patch-level ground truth labeling, we pursued
513 the approach detailed in S2B Fig. We used our "Before k-means" as a feature extractor to convert
514 all image patches in the archived mouse cohort dataset to high-dimensional 512-length numerical
515 representations learned during the first phase of training in S2A Fig. K-means clustering on these
516 representations generated patch classes (S2C Fig).

517 We then utilized these k-means clusters to generate new patch-level ground truth labels.
518 As our control mice did not receive any colitis induction, we provided 'Uninvolved' ground truth
519 labels to all patches from these mice. For our mice with colitis induction, we referred to the k-
520 means clustering results. With this new ground truth labeling, we trained another RN-34 model
521 pretrained on ImageNet using a 70/10/20% training/validation/test set split.

522

523 **Small Patch Classifier Training for Preprocessing**

524 From our archived mouse cohort dataset WSIs, we extracted 32x32 pixel patches and trained an
525 additional RN-34 classifier pretrained on ImageNet (S3 Fig). Our Small Patch Classifier predicts
526 these patches as one of the 'Background', 'Muscle', 'Tissue', and 'Submucosa' classes.

527

528 **Tissue Map Generation and Patch Filtering**

529 For every H&E-stained, swiss-rolled colon WSI, we first extract 32x32 pixel patches. The Small
530 Patch Classifier is applied to each to generate a tissue map for every input H&E WSI (S4A Fig).

531 The tissue map is of the same dimension as the input H&E where white areas represent 'Tissue'
532 and 'Submucosa' classifications, while the black areas represent 'Muscle' and 'Background'. For
533 each subsequently extracted 224x224 pixel H&E patch, the corresponding area is extracted from
534 the accompanying tissue map. Each patch is then thresholded to evaluate whether enough tissue
535 or submucosa is present. Practically, we implement a 65% threshold. Patches with more than
536 65% of 'Background' or 'Muscle' regions are filtered out, while those with less are kept (S4B Fig).
537

538 **Overlap Patch Extraction and Overlay Generation**

539 To reduce classification dependence for what patch happened to be extracted from a region in a
540 WSI, we extracted overlapping patches. Beginning with existing patch locations, we iterated 20
541 pixels 10 times in each direction (up/down/left/right) and extracted new patches. We then again
542 filtered out patches containing too little tissue. This led to an increase of ~450 to ~200,000 patches
543 per mouse.

544 To generate overlays, our 'Involved' versus 'Uninvolved' classifier was applied to all
545 patches, including overlapping patches. At every pixel in the original H&E, 'Involved' and
546 'Uninvolved' prediction confidences were averaged. As such, a greater degree of spatial context
547 is incorporated for final classifications at each pixel. Red areas correspond to pixels with >50%
548 confidence of 'Involved' predictions, while green areas correspond to pixels with >50% confidence
549 of 'Uninvolved' predictions. Furthermore, all regions in tissue maps corresponding to 'Muscle' and
550 'Background' were eliminated from the overlay. This generated final outputs with overlays only
551 over mucosal and submucosal areas (Fig 2B-C).
552

553 **Discovery of Patch Classes**

554 Overview schematic is shown in S5A Fig. From across our archived mouse cohort, we extracted
555 all patches, including overlapping patches, classified as 'Involved'. We utilized our final 'Involved'
556 versus 'Uninvolved' classifier as a feature extractor then performed PCA-based dimensionality

557 reduction to 250 principal components (PCs) that captured 95% of variability. We then performed
558 k-means clustering to identify pathologic patch classes. This process was also repeated for all
559 patches classified as 'Uninvolved' in our archived mouse cohort dataset to identify non-pathologic
560 patch classes. PCA-based dimensionality reduction for 'Uninvolved' patches required 255 PCs to
561 account for 95% of variability.
562

563 **Machine Learning Classifier to Predict Mouse Model**

564 Per mouse proportions of 'Uninvolved' patches and the 'Involved' k-means patch classes were
565 generated for every sample from the archived mouse cohort and utilized to train a linear
566 determinant analysis (LDA) classifier to predict the mouse model (Fig S5A). For inference, the
567 final schematic is shown in Fig S5B. Overlapping patches are extracted from our prospective
568 mouse cohort samples and filtered according to tissue maps generated by the small patch
569 classifier. Our 'Involved' versus 'Uninvolved' classifier is applied, and 'Involved' patches undergo
570 a further round of PCA-based dimensionality reduction that is first fit to archived mouse cohort
571 data. Our k-means model, which is also trained on our archived mouse cohort, then categorizes
572 'Involved' patches into one of the discovered classes. Finally, the per mouse proportions of
573 'Uninvolved' patches and 'Involved' k-means patch classes are generated for each mouse and
574 fed into the LDA classifier to infer the mouse model.
575

576 **Acknowledgements**

577 The authors wish to acknowledge the Stony Brook Research Histology core for their expert
578 assistance with paraffin embedding, sectioning, and performing H&E staining on our colon
579 samples. The authors also wish to acknowledge the Stony Brook Division of Laboratory Animal
580 Resources for their expert assistance with caring, housing, and maintaining mice used to generate

581 results included in this manuscript. This work is support by grants from the National Institutes of
582 Health: DK052230 to V.W.Y. and CA205109/CA225021 to J.H.S.

583 **References**

584 1. The Facts About Inflammatory Bowel Disease Crohn's & Colitis Foundation of
585 America2014 [Available from:
586 https://www.crohnscolitisfoundation.org/sites/default/files/2019-
587 02/Updated%20IBD%20Factbook.pdf.
588 2. Bernstein CN, Wajda A, Blanchard JF. The clustering of other chronic inflammatory
589 diseases in inflammatory bowel disease: a population-based study. Gastroenterology.
590 2005;129(3):827-36.
591 3. Lakatos PL, Lakatos L. Risk for colorectal cancer in ulcerative colitis: changes, causes
592 and management strategies. World J Gastroenterol. 2008;14(25):3937-47.
593 4. Thoreson R, Cullen JJ. Pathophysiology of inflammatory bowel disease: an overview.
594 Surg Clin North Am. 2007;87(3):575-85.
595 5. Solomon L, Mansor S, Mallon P, Donnelly E, Hoper M, Loughrey M, et al. The dextran
596 sulphate sodium (DSS) model of colitis: an overview. Comparative clinical pathology.
597 2010;19(3):235-9.
598 6. Nunes NS, Kim S, Sundby M, Chandran P, Burks SR, Paz AH, et al. Temporal clinical,
599 proteomic, histological and cellular immune responses of dextran sulfate sodium-induced
600 acute colitis. World J Gastroenterol. 2018;24(38):4341-55.
601 7. Ho YT, Shimbo T, Wijaya E, Kitayama T, Takaki S, Ikegami K, et al. Longitudinal Single-
602 Cell Transcriptomics Reveals a Role for Serpina3n-Mediated Resolution of Inflammation
603 in a Mouse Colitis Model. Cell Mol Gastroenterol Hepatol. 2021;12(2):547-66.
604 8. Yan Y, Kolachala V, Dalmasso G, Nguyen H, Laroui H, Sitaraman SV, et al. Temporal and
605 spatial analysis of clinical and molecular parameters in dextran sodium sulfate induced
606 colitis. PLoS One. 2009;4(6):e6073.
607 9. Whittem CG, Williams AD, Williams CS. Murine Colitis modeling using Dextran Sulfate
608 Sodium (DSS). J Vis Exp. 2010(35).

609 10. Nandan MO, Ghaleb AM, Liu Y, Bialkowska AB, McConnell BB, Shroyer KR, et al.
610 Inducible intestine-specific deletion of Kruppel-like factor 5 is characterized by a
611 regenerative response in adult mouse colon. Dev Biol. 2014;387(2):191-202.
612 11. Shieh J, Chu TH, Liu Y, Kim J, Ruiz de Sabando A, Kobayashi S, et al. KLF5 protects the
613 intestinal epithelium against Th17 immune response in a murine colitis model. JCI Insight.
614 2022;7(7).
615 12. Yadav SS, Jadhav SM. Deep convolutional neural network based medical image
616 classification for disease diagnosis. Journal of Big Data. 2019;6(1):113.
617 13. Kobayashi S, Saltz JH, Yang VW. State of machine and deep learning in histopathological
618 applications in digestive diseases. World J Gastroenterol. 2021;27(20):2545-75.
619 14. Cai L, Gao J, Zhao D. A review of the application of deep learning in medical image
620 classification and segmentation. Ann Transl Med. 2020;8(11):713.
621 15. Bialkowska A, Ghaleb AM, Nandan MO, Yang VW. Improved Swiss-rolling Technique for
622 Intestinal Tissue Preparation for Immunohistochemical and Immunofluorescent Analyses.
623 J Vis Exp. 2016(113):e54161.
624 16. Cooper HS, Murthy S, Shah R, Sedergran D. Clinicopathologic study of dextran sulfate
625 sodium experimental murine colitis. Laboratory investigation; a journal of technical
626 methods and pathology. 1993;69(2):238-49.
627 17. Parkash V, Bifulco C, Feinn R, Concato J, Jain D. To count and how to count, that is the
628 question: interobserver and intraobserver variability among pathologists in lymph node
629 counting. American journal of clinical pathology. 2010;134(1):42-9.
630 18. Jain RK, Mehta R, Dimitrov R, Larsson LG, Musto PM, Hodges KB, et al. Atypical ductal
631 hyperplasia: interobserver and intraobserver variability. Mod Pathol. 2011;24(7):917-23.
632 19. Franc B, de la Salmoniere P, Lange F, Hoang C, Louvel A, de Roquancourt A, et al.
633 Interobserver and intraobserver reproducibility in the histopathology of follicular thyroid
634 carcinoma. Hum Pathol. 2003;34(11):1092-100.

635 20. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv e-
636 prints [Internet]. 2015 December 01, 2015:[arXiv:1512.03385 p.]. Available from:
637 https://ui.adsabs.harvard.edu/abs/2015arXiv151203385H.
638 21. Hanin B. Which neural net architectures give rise to exploding and vanishing gradients?
639 arXiv preprint arXiv:180103744. 2018.
640 22. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale
641 visual recognition challenge. International journal of computer vision. 2015;115(3):211-52.
642 23. Nava GM, Friedrichsen HJ, Stappenbeck TS. Spatial organization of intestinal microbiota
643 in the mouse ascending colon. ISME J. 2011;5(4):627-38.
644 24. Kamphuis JB, Mercier-Bonin M, Eutamène H, Theodorou V. Mucus organisation is shaped
645 by colonic content; a new view. Scientific reports. 2017;7(1):1-13.
646 25. Eichele DD, Kharbanda KK. Dextran sodium sulfate colitis murine model: An
647 indispensable tool for advancing our understanding of inflammatory bowel diseases
648 pathogenesis. World journal of gastroenterology. 2017;23(33):6016.
649 26. Perse M, Cerar A. Dextran sodium sulphate colitis mouse model: traps and tricks. J
650 Biomed Biotechnol. 2012;2012:718617.
651 27. Lee H, Westerhoff M, Shen B, Liu X. Clinical Aspects of Idiopathic Inflammatory Bowel
652 Disease: A Review for Pathologists. Arch Pathol Lab Med. 2016;140(5):413-28.
653 28. Kellermann L, Riis LB. A close view on histopathological changes in inflammatory bowel
654 disease, a narrative review. Digestive Medicine Research. 2021;4.
655 29. Magro F, Langner C, Driessen A, Ensari A, Geboes K, Mantzaris GJ, et al. European
656 consensus on the histopathology of inflammatory bowel disease. J Crohns Colitis.
657 2013;7(10):827-51.
658 30. Tanaka M, Riddell RH, Saito H, Soma Y, Hidaka H, Kudo H. Morphologic criteria
659 applicable to biopsy specimens for effective distinction of inflammatory bowel disease from

660 other forms of colitis and of Crohn's disease from ulcerative colitis. Scand J Gastroenterol.
661 1999;34(1):55-67.
662 31. Gersemann M, Becker S, Kübler I, Koslowski M, Wang G, Herrlinger KR, et al. Differences
663 in goblet cell differentiation between Crohn's disease and ulcerative colitis. Differentiation.
664 2009;77(1):84-94.
665 32. Bédard A, Westerling-Bui T, Zuraw A. Proof of concept for a deep learning algorithm for
666 identification and quantification of key microscopic features in the murine model of DSS-
667 induced colitis. Toxicologic Pathology. 2021;49(4):897-904.
668 33. Rogers R, Eastham-Anderson J, DeVoss J, Lesch J, Yan D, Xu M, et al. Image analysis-
669 based approaches for scoring mouse models of colitis. Veterinary pathology.
670 2016;53(1):200-10.
671 34. Eichele DD, Kharbanda KK. Dextran sodium sulfate colitis murine model: An
672 indispensable tool for advancing our understanding of inflammatory bowel diseases
673 pathogenesis. World J Gastroenterol. 2017;23(33):6016-29.
674 35. Chassaing B, Aitken JD, Malleshappa M, Vijay-Kumar M. Dextran sulfate sodium (DSS)-
675 induced colitis in mice. Curr Protoc Immunol. 2014;104:Unit 15 25.
676 36. Kojouharoff G, Hans W, Obermeier F, MA¨ NNEL D, Andus T, SCHO¨ LMERICH J, et al.
677 Neutralization of tumour necrosis factor (TNF) but not of IL-1 reduces inflammation in
678 chronic dextran sulphate sodium-induced colitis in mice. Clinical & Experimental
679 Immunology. 1997;107(2):353-8.
680 37. Ogawa A, Andoh A, Araki Y, Bamba T, Fujiyama Y. Neutralization of interleukin-17
681 aggravates dextran sulfate sodium-induced colitis in mice. Clinical immunology.
682 2004;110(1):55-62.
683 38. Siegmund B, Fantuzzi G, Rieder F, Gamboni-Robertson F, Lehr H-A, Hartmann G, et al.
684 Neutralization of interleukin-18 reduces severity in murine colitis and intestinal IFN-γ and

685 TNF-α production. American Journal of Physiology-Regulatory, Integrative and
686 Comparative Physiology. 2001;281(4):R1264-R73.
687 39. Scheinin T, Butler DM, Salway F, Scallon B, Feldmann M. Validation of the interleukin-10
688 knockout mouse model of colitis: antitumour necrosis factor-antibodies suppress the
689 progression of colitis. Clinical & Experimental Immunology. 2003;133(1):38-43.
690 40. Ranganathan P, Jayakumar C, Manicassamy S, Ramesh G. CXCR2 knockout mice are
691 protected against DSS-colitis-induced acute kidney injury and inflammation. American
692 Journal of Physiology-Renal Physiology. 2013;305(10):F1422-F7.
693 41. Kiesler P, Fuss IJ, Strober W. Experimental Models of Inflammatory Bowel Diseases. Cell
694 Mol Gastroenterol Hepatol. 2015;1(2):154-70.
695
696

697 **Supporting Information Captions**

698

699 **Supplemental Table 1. Mouse Cohort Sample Numbers Per Genotype and Treatment.**

700

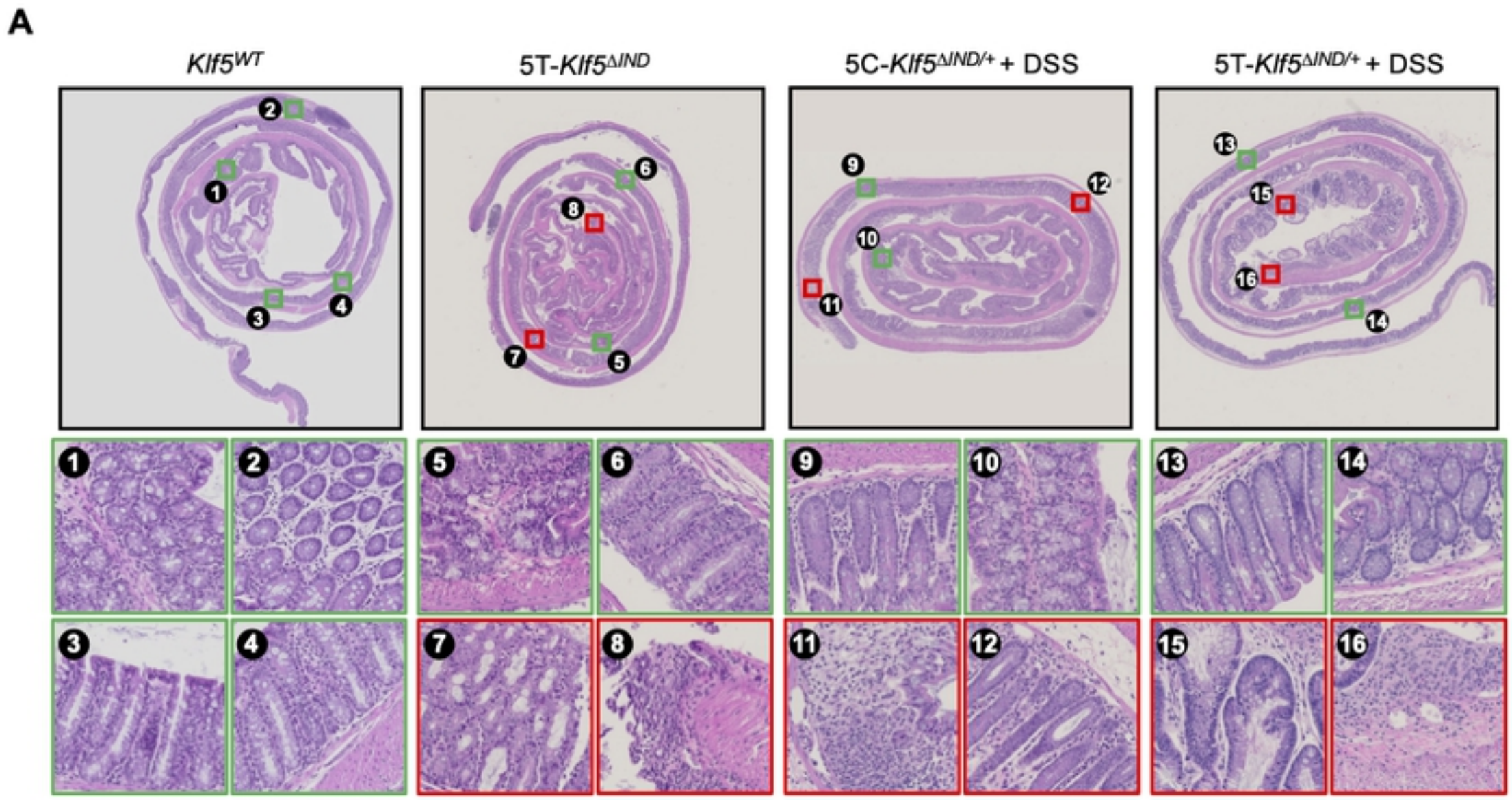
701 **Supplemental Figure 1. Colitis Mouse Models.** A) Treatment schedules for each mouse model.
702 DSS model in prospective cohort has no injections to confirm abnormal pathology is recognized
703 independent of corn oil. Additionally, 3.0% DSS was used in prospective mice, as this was
704 observed to be the optimal concentration at the Stony Brook facilities. B) Swiss roll of *Klf5*^{WT}
705 (*Villin-CreER*^{T2}*;Klf5*$^{+/+}$) mouse treated with DSS and no TAM or CO injections. C) Though not
706 utilized in this study, swiss rolls of corresponding controls (5T-*Klf5*$^{\Delta IND/+}$ and 5C-*Klf5*$^{\Delta IND/+}$) are
707 shown for combined colitis model. D) Clinical scores combining weight loss, stool consistency,
708 and fecal blood according to Cooper et al. (16) for combined colitis model, 5C-*Klf5*$^{\Delta IND/+}$ + DSS,
709 and control mice with student's t-test showing significant differences. E) Histological scores
710 according to Cooper et al. (16). One-way ANOVA *p<0.05, **p<0.01, ***p<0.001

711

712 **Supplemental Figure 2. Second Phase of Training with 'Involved' and 'Uninvolved' K-**
713 **Means Patch Cluster Labels Improves Model Performance.** A) Overview schematic showing
714 initial phase of RN-34 model training that uses mouse colitis status as patch ground truth labels.
715 Thus, intracolonic heterogeneity in colitis mice is not addressed at this round of patch labeling. B)
716 Second phase of H&E model training that uses trained RN-34 model from A) as a feature extractor
717 for patches in dataset. K-means clustering is applied on extracted features to generate patch
718 classes. Student t-test is utilized to assess whether patch classes are significantly enriched in
719 proportion between colitis and control mice. In this second phase of training, all patches from
720 control mice are labeled as 'Uninvolved', as no colitis induction was performed. For colitis mice,
721 patches are given 'Uninvolved' or 'Involved' ground truth labels according to k-means predictions.
722 C) K-means patch classes identified during second phase of training in B) used for ground truth

723 labeling. D) Student's t-tests show patch classes are found in differing portions between mice with
724 colitis induction and controls. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001. E) Rosettes are in
725 DSS-treated mice relative to other colitis models. One-way ANOVA shows a statistically
726 significant difference between groups. F) Independent test set output confusion matrix for before
727 k-means model. G) 200 patches from 4 mice (1 control, 1 of each colitis mouse model) were
728 extracted and labeled as 'Uninvolved' or 'Involved' by a pathologist. 29/200 patches were
729 discarded for not enough spatial context to provide a label. Inference using models trained in A)
730 and B) show that the k-means patch labeling approach increased prediction agreement with
731 pathologist-generated labels.
732
733 **Supplemental Figure 3. Small Patch Classifier Training for Preprocessing.** A) Example
734 32x32 pixel patches for the 'Background', 'Tissue', 'Muscle', and 'Submucosa' classes used to
735 train the small patch RN-34 classifier. B) Trained small patch RN-34 model confusion matrix
736 outputs for independent test set of 4 mice, each with 100 patches of each class (1600 total
737 patches). C) Example overlay of small patch RN-34 classifier on test set mouse.
738
739 **Supplemental Figure 4. Tissue Map Generation and Filtering Process.** A) Small patch RN-
740 34 classifier from S4 Fig is applied to all 32x32 pixel patches extracted from a WSI. A tissue map
741 is generated where 'Background' and 'Muscle' are black, while 'Tissue' and 'Submucosa' are
742 white. The yellow arrow indicates a portion of muscle that is assigned to the Background/Muscle
743 class on the corresponding tissue map. B) Example 224x224 pixel H&E patches with
744 corresponding tissue map patches and filtering decisions. For each extracted patch, the decision
745 to keep or filter out the patch is made based on whether there is more than 65% (filter) or less
746 than 65% (keep) of unwanted Background/Muscle area on the corresponding tissue map patch.
747

748 **Supplemental Figure 5. Linear Determinant Analysis Classifier Training and Inference**
749 **Schematics.** A) Overview schematic. All patches, including overlapping patches, classified as
750 'Involved' by our model undergo feature extraction by our trained 'Involved' versus 'Uninvolved'
751 classifier. Subsequent PCA-based dimensionality reduction and k-means clustering identify 4
752 'Involved' patch classes shown in Fig 4A. Per mice proportions of 'Uninvolved' patches and
753 'Involved' class patches are used to train an LDA classifier to predict mouse models. B) Overview
754 schematic for inference pipeline. Overlapping patches are extracted and patches containing too
755 much background or muscle are filtered out via tissue map process in S5 Fig. H&E classifier is
756 applied to each patch to predict 'Involved' or 'Uninvolved'. 'Involved' patches undergo further RN-
757 34 feature extraction and PCA-based dimensionality reduction. K-means model trained on training
758 cohort then classifies 'Involved' patches into one of the four 'Involved' patch classes. Patch
759 proportion of 'Uninvolved' and 'Involved' k-means classes for each mouse are generated and fed
760 into the LDA classifier for inference of mouse model. C) Stacked bar plot of total 'Uninvolved' and
761 'Involved' k-means patch class proportions across prospective mouse cohort. Error bars show
762 mean with standard deviation.
763
764 **Supplemental Figure 6. 5T-*Klf5*$^{\Delta IND}$ Swiss Roll Predicted as Control has Healthier Appearing**
765 **Histology Than 5T-*Klf5*$^{\Delta IND}$ Swiss Roll Predicted as 5T-*Klf5*$^{\Delta IND}$.** Compared to the properly
766 predicted 5T-*Klf5*$^{\Delta IND}$ swiss rolled colon (left), the colon predicted as control has fewer pathologies
767 and likely represents a mouse with weak induction of colitis. Yellow arrows indicate presence of
768 healthy goblet cells. Red arrows indicate loss of goblet cells. Orange arrows indicate crypt loss.
769



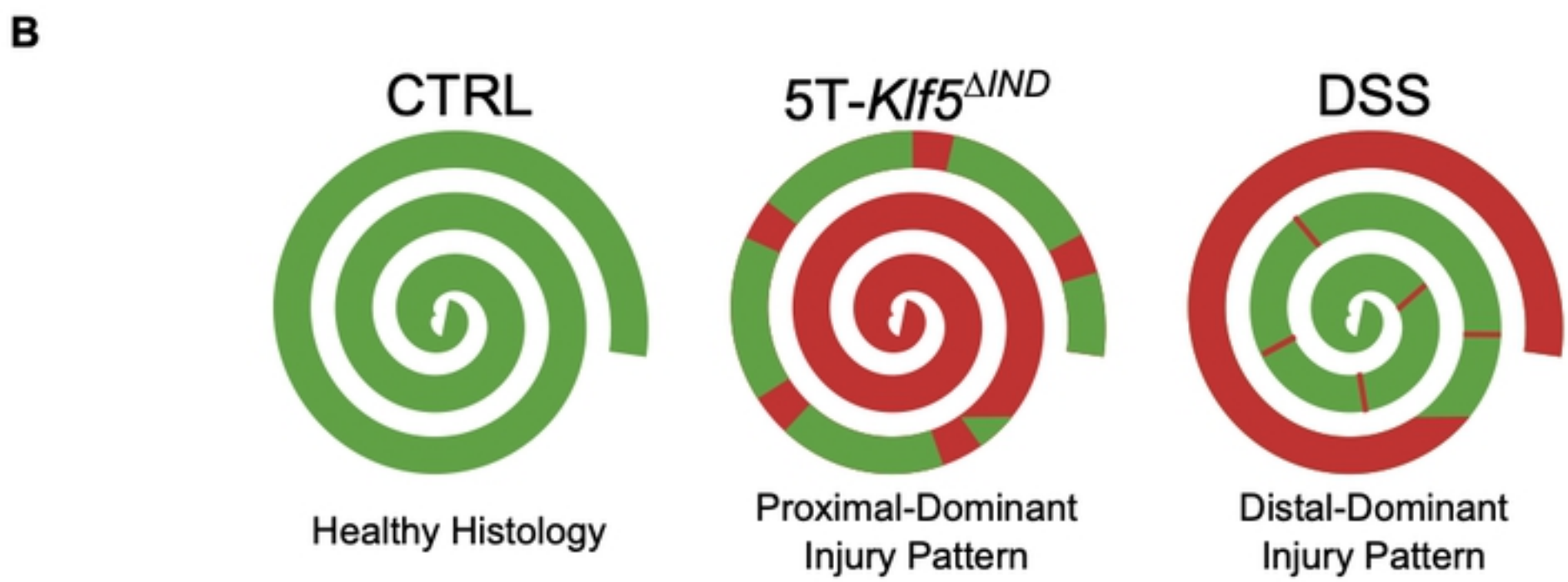


Figure 1

Figure 1

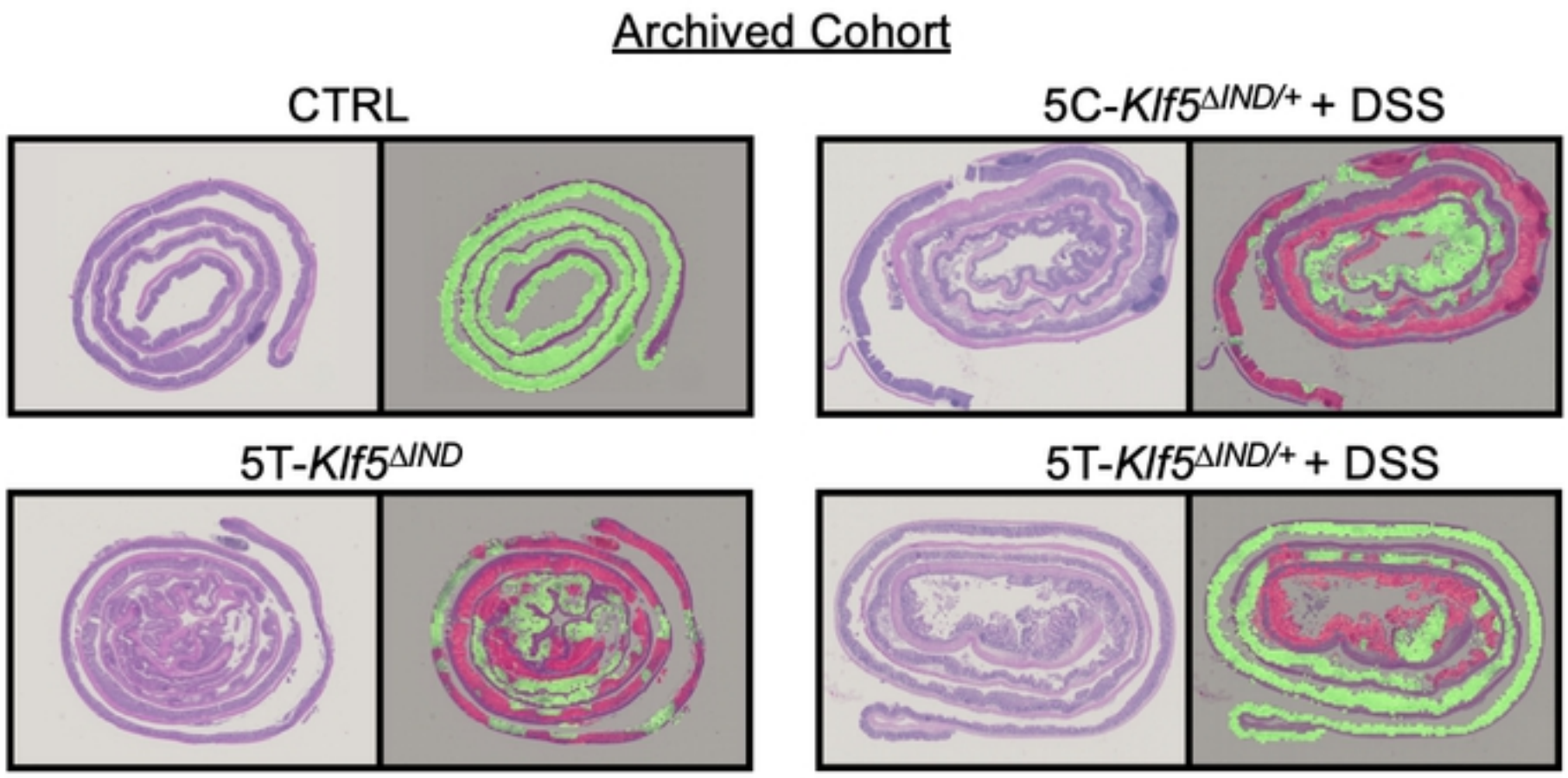
A

Ground Truth	RN-34 Prediction: Uninv.	RN-34 Prediction: Involved
Uninv.	3221	69
Involved	191	690

Overall F1 Score: 90.1%

	F1 Score	Precision	Recall
Uninv.	96.12%	*94.40%*	*97.90%*
Involved	84.15%	*90.91%*	*78.32%*

B



C

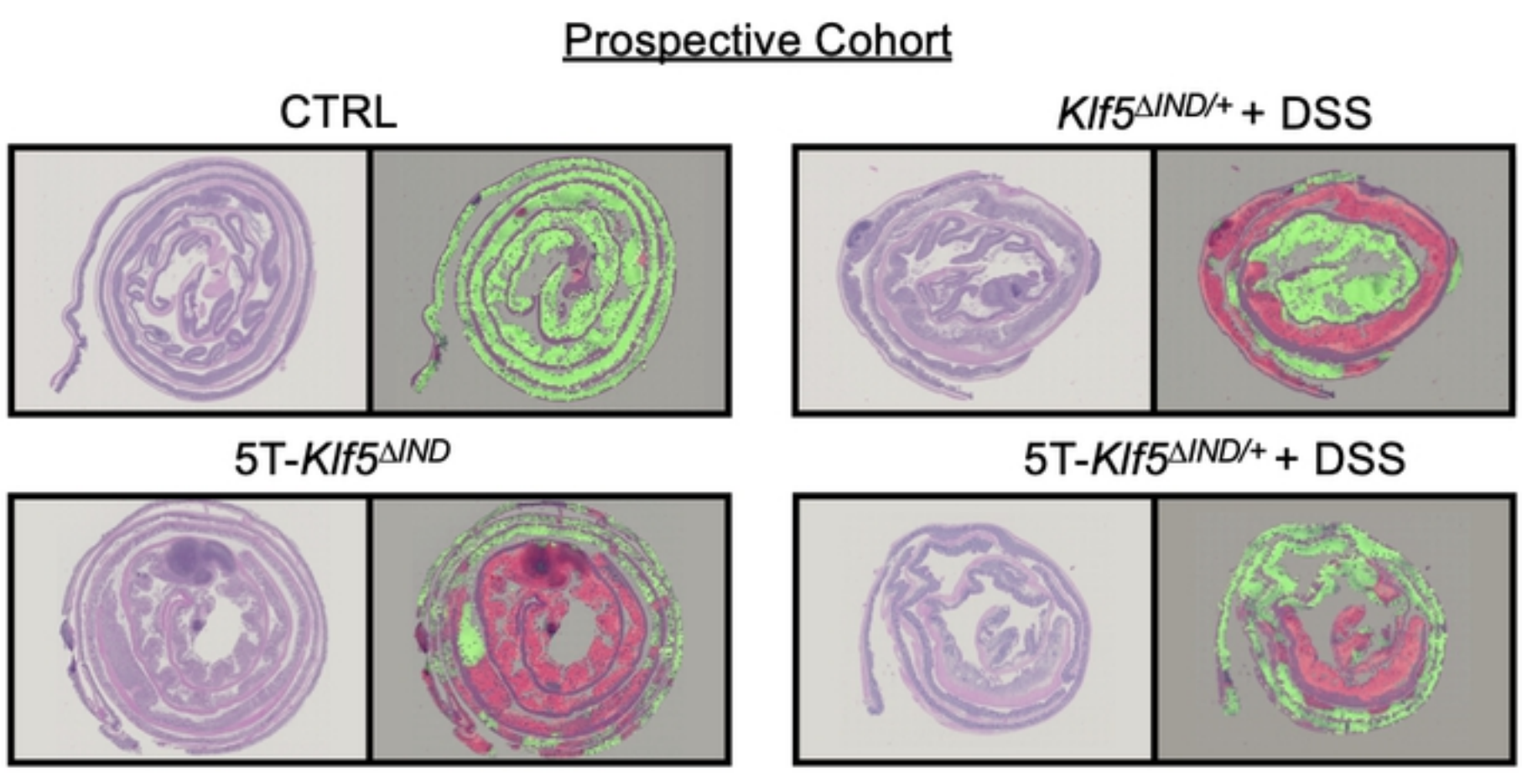


Figure 2

Figure 2

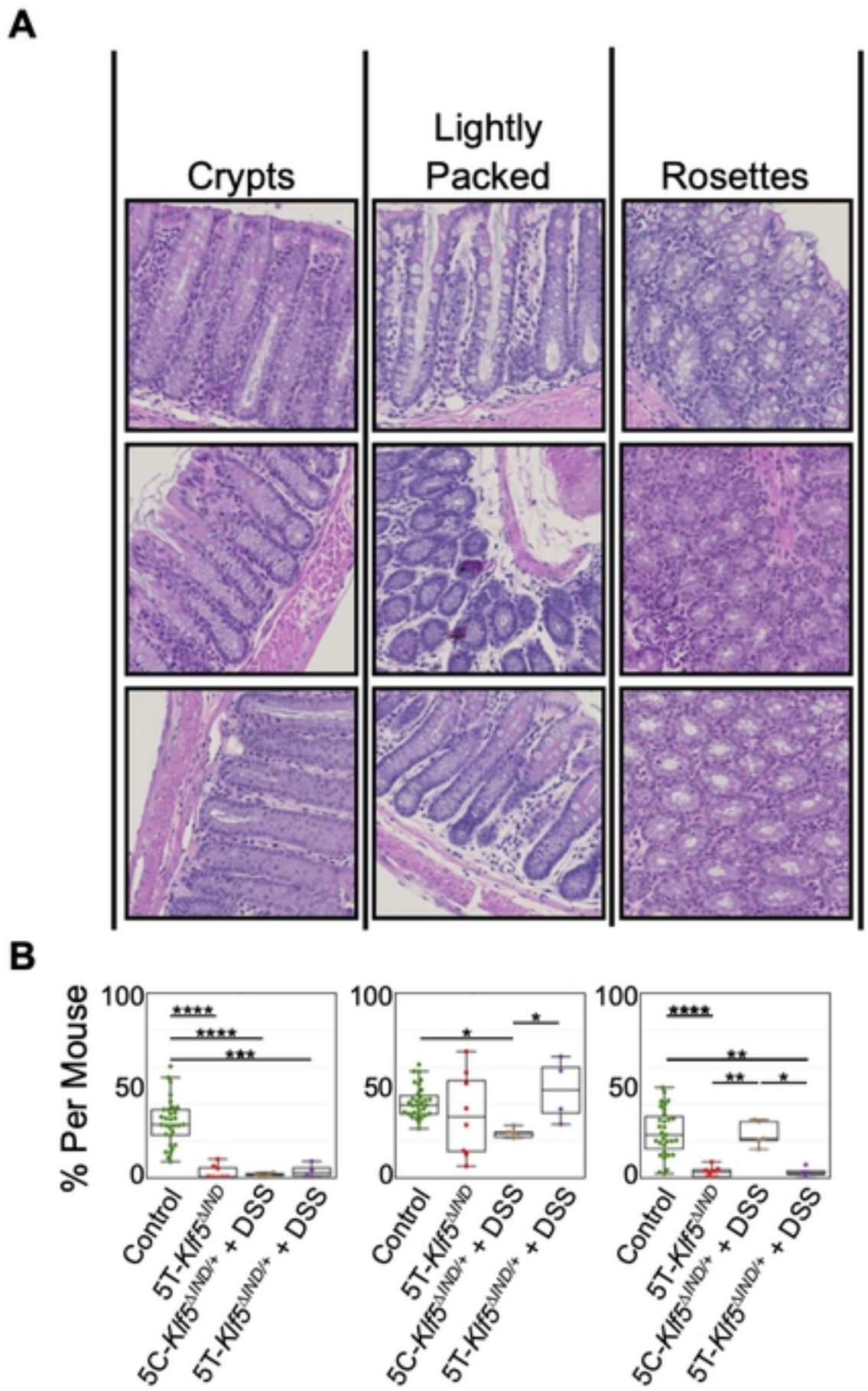
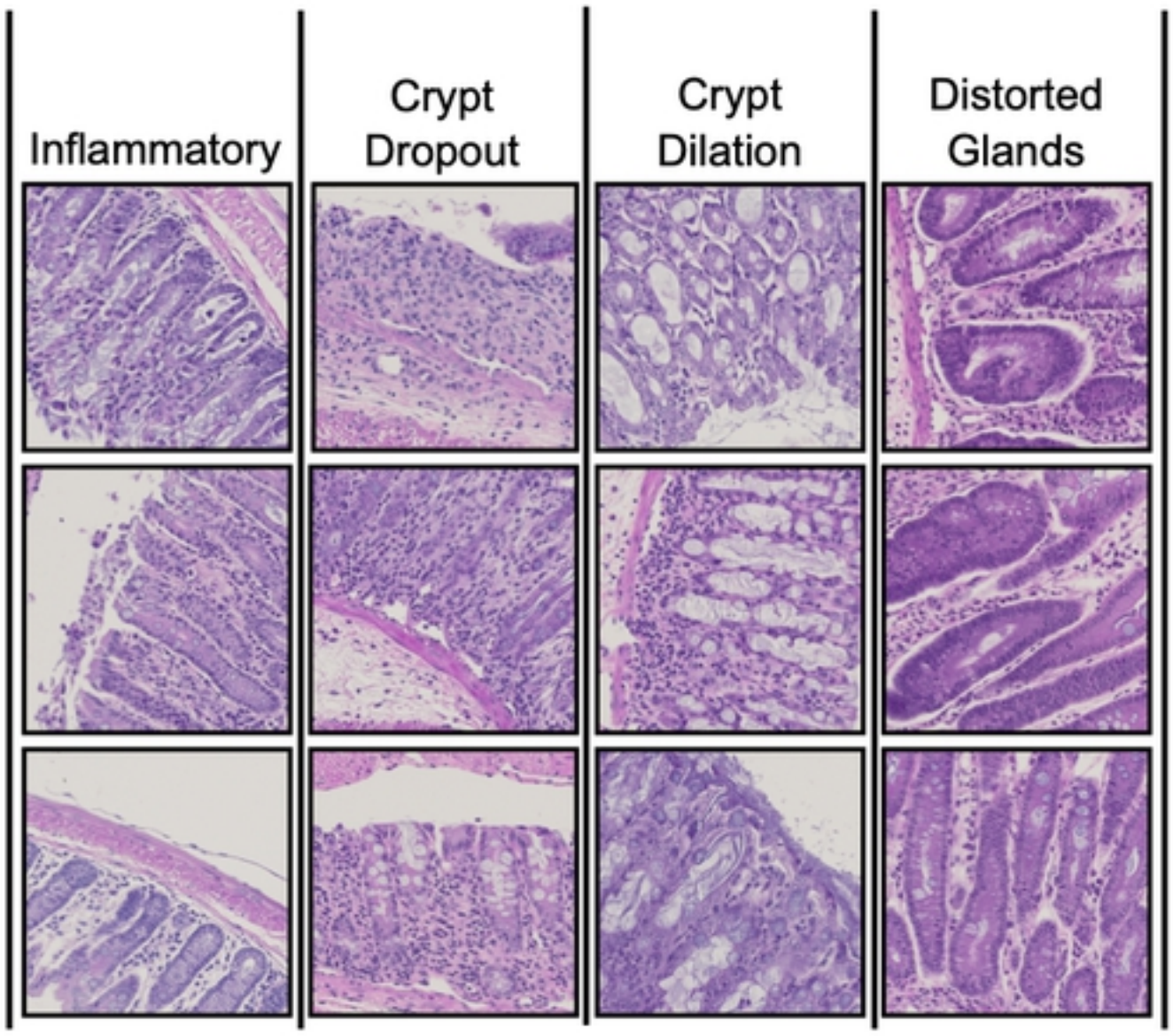


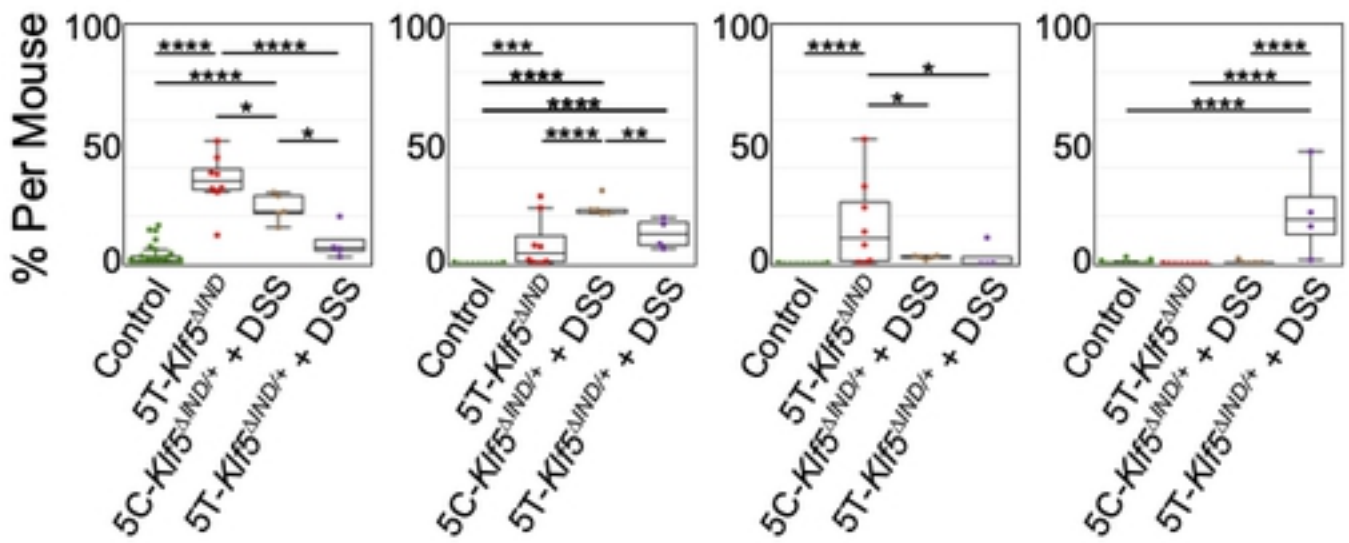
Figure 3

Figure 3

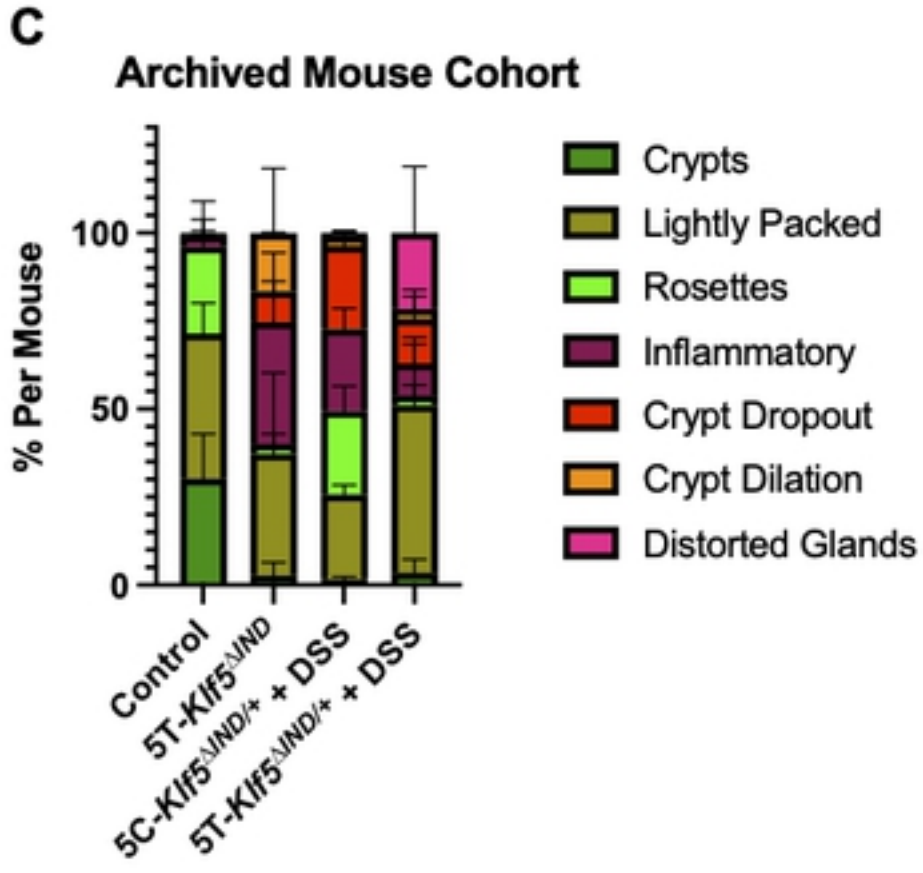
A



B



C



D

LDA Prediction

Ground Truth	CTRL	DSS	5T-*Klf5*$^{\Delta IND}$	5T-*Klf5*$^{\Delta IND/+}$ + DSS
CTRL	8	0	0	0
DSS	0	8	0	0
5T-*Klf5*$^{\Delta IND}$	1	0	4	0
5T-*Klf5*$^{\Delta IND/+}$ + DSS	0	0	0	3

Overall F1 Score: 95.75%

Overall Accuracy: 95.83%

Figure 4

Figure 4