Automated multi-scale computational pathotyping (AMSCP) of inflamed synovial tissue

Richard D. Bell[1,2,*], Matthew Brendel[2,*], Maxwell Konnaris[1], Justin Xiang[3], Miguel Otero[1,2], Mark A. Fontana[1,2], Accelerating Medicines Partnership Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP RA/SLE) Consortium, Edward DiCarlo[1], Jennifer Anolik[4], Laura Donlin[1], Dana Orange[1,5], H. Mark Kenney[4], Edward M. Schwarz[4†], Lionel B Ivashkiv[1,2†] and Fei Wang[2†]

*, † Authors contributed equally

Affiliations: [1]Hospital for Special Surgery, [2]Weill Cornell Medical College, [3]Horace Greely High School, [4]University of Rochester Medical Center, and [5]The Rockefeller University

Corresponding Author:
Richard D. Bell, PhD
515 E 71st New York, NY
bellr@hss.edu

Abstract

Rheumatoid arthritis (RA) is a complex immune-mediated inflammatory disorder in which patients suffer from inflammatory-erosive arthritis. Recent advances on histopathology heterogeneity of RA pannus tissue revealed three distinct phenotypes based on cellular composition (pauci-immune, diffuse and lymphoid), suggesting distinct etiologies that warrant specific targeted therapy. Thus, cost-effective alternatives to clinical pathology phenotyping are needed for research and disparate healthcare. To this end, we developed an automated multi-scale computational pathotyping (AMSCP) pipeline with two distinct components that can be

leveraged together or independently: 1) segmentation of different tissue types to characterize tissue-level changes, and 2) cell type classification within each tissue compartment that assesses change across disease states. Initial training and validation were completed on 264 knee histology sections from mice with TNF-transgenic (n=233) and injected zymosan induced (n=32) inflammatory arthritis. Peak tissue segmentation performance with a frequency weighted mean intersection over union was 0.94 ± 0.01 and peak cell classification F1 was 0.83 ± 0.12. We then leveraged these models and adapted them to analyze RA pannus tissue clinically phenotyped as pauci-immune (n=5), diffuse (n=28) and lymphoid (n=27), achieving peak cell classification performance with F1 score of 0.81 ± 0.06. Regression analysis demonstrated a highly significant correlation between AMSCP of lymphocyte counts and average Krenn Inflammation Score (rho = 0.88; p<0.0001). While a simple threshold of 1.1% of plasma cells demonstrated the phenotyping potential of our automated approach vs. a clinical pathologist with a sensitivity and specificity of 0.81 and 0.73. Taken together, we find AMSCP to be a valuable cost-effective method for research. Follow-up studies to assess its clinical utility are warranted.

Introduction

Disease pathotyping with histopathology, the discovery of disease subtypes using target organ histology, is a critical step in understanding etiology and response to therapy in heterogeneous diseases, like rheumatoid arthritis (RA). Our understanding of RA, which is a chronic, inflammatory joint disease, has greatly benefited from histopathology subtyping because the disease has distinct and disparate etiologies with largely stable pathotypes [1, 2] that show differential response to therapy [3-7]. However, the process of pathotyping a patient can be resource intensive involving both basic and immune-stains requiring a high level of expertise by pathologists to interpret tissue and cellular histologic features, and prone to inter-

and intra-observer variations [8, 9]. More cost effective and efficient procedures need to be developed in order incorporate these types of data into a precision medicine decision making process.

Recent work describing RA pathotypes uncover three distinct synovial pathotypes 1) cellular dense, lymphocyte rich (lymphoid), 2) myeloid rich with few lymphocytes (diffuse/myeloid), and 3) fibroblast rich (pauci-immune); which are identifiable through distinct cellular and tissue level changes within synovial joint biopsies [10-14]. These pathotypes also correlate with antibody positivity (i.e. anti-citrullinated peptide antibodies) with the lymphoid type enriched in antibody positive patients whereas both the diffuse/myeloid and pauci-immune types have equal contributions of both antibody positive and negative patients. This aligns with preclinical models that rely on antibody dependent (e.g. Collagen Induced Arthritis and Serum Induced Arthritis) and independent mechanism (e.g. humanized TNF Transgenic and Zymosan Induced Arthritis) to study disease and are phenotypically similar to these pathotypes [15]. Thue, tools that allow us to study both murine and human pathology would improve our overall understanding of this heterogeneous disease.

Computational tools to study histological changes have been shown to augment pathologist workflows and allow for the identification of disease specific patterns [16]. In particular, machine learning, and specifically deep learning, is a data-driven framework that has recently had success in the automated analysis of musculoskeletal imaging data [17]. Additionally, computational tools can automate and provide a more holistic analysis a wide variety of histopathologic tissue and cell-level changes to enable a more detailed understanding of disease subtypes. However, automated and comprehensive tool to study both tissue and cell type specific changes in arthritis that can quantify therapeutic or clinically meaningful differences has not yet been developed [17].

In this work, we developed an automated multi-scale computational pathotyping (AMSCP) model to analyze tissue and cell-level changes during the progression of inflammatory

arthritis. This model can pathotype both mouse and human inflammatory arthritis in therapeutic intervention studies and clinical meaningful scenarios. We leveraged innovative transfer- and active- learning techniques to improve model performance and workload efficiency. Our modeling framework consists of two distinct components that can be utilized together or independently, (1) a method to segment different tissue types to characterize tissue-level changes and (2) a method to classify cell types within each tissue compartment to study how these change across disease states. We utilized two mouse models of inflammatory arthritis to train and validate our computational models with subsequent implementation on additional datasets to measure therapeutic efficacy, known biologic differences and discover novel pathologic changes. Then, we utilize a human synovial biopsy data set from the Accelerating Medicines Partnership Rheumatoid Arthritis (AMP-RA) to demonstrate our model's utility in a clinical setting by classifying lymphoid pathotypes from diffuse/myeloid pathotype and identifying cell types associated with the pauci-immune pathotype while preserving spatial cell-level data.

Methods

Dataset Description

All mouse work was approve by the University Committee on Animal Resources at the University of Rochester Medical Center and the Institutional Animal Care and Use Committee at the Hospital for Special Surgery. Whole slide images (WSIs) of sagittal mouse knee sections were taken from two different mouse models of inflammatory arthritis and the accompanying controls for segmentation experiments (**Supplemental Figure 1A**). Batch A consisted of male and female TNF Transgenic mice (TNF-Tg, n=47) and wild-type littermates (WT, n = 15) used in a previous publication [18, 19]. Batch B consisted of male and female knees that received intra-articular injections of 180ug of Zymosan to induce Zymosan Induced Arthritis (ZIA, n = 24) and control contralateral limbs (Control, n = 8) [20] that were euthanized on Day 7 after injection. Different batches were used to test model generalizability across different biological

mechanisms of arthritis development, differences in H&E staining protocols and slide scanner used to digitally capture slides at 40x magnification (Batch A: VS120 Olympus, 0.173 µm per pixel; Batch B: CS2 Aperio Leica, 0.253 µm per pixel).

To further test model generalizability, 2 different holdout datasets were used to validate the model: 1) H&E-stained sagittal knee WSIs from Bell et al [18, 19] and 2) Orange G-H&E stained sagittal knee WSIs from Kenney et al [21]. Slides from Bell et al were ensured to not be used in the initial model training, internal validation or testing. These included slides from 6-month-old male TNF-Tg (n=33 slides from 14 knees) and WT littermates (n=43 slides from 17 knees), and slides from 9.5 month old male TNF-Tg mice (Anti-TNF: n=24 slides from 8 knees; Placebo: n= 29 slides from 10 knees) and WT littermates (Placebo: n=42 slides from 15 knees) either treated with a 6 week course of Anti-TNF antibodies or Placebo control. To generate downstream measurements of tissues of interest, a region of interest (ROI) was drawn from the tibial growth plate to femoral growth plate including the anterior and posterior extra articular tissue.

WSIs of H&E-stained human synovial biopsies were collected from the Accelerating Medicines Partnership Rheumatoid Arthritis (AMP-RA) Phase II consortium [22]. In short, synovial tissue biopsies were acquired from RA patients at 13 different clinical sites in the United States and 2 from the United Kingdom from October 2016 to February 2020. The study was performed in accordance with protocols approved by the institutional review board at each site. The tissue was paraffin embedded, stained with H&E and imaged on a VS120 Olympus. Board certified pathologists in the AMP-RA consortium classified each case as either lymphoid (n = 27), diffuse/myeloid (n=28) or pauci-immune (n=5), and provided the Krenn inflammation score for each case [23].

Semantic Segmentation Annotation and Preprocessing

Manual annotations were performed within QuPath [24] to assign labels for WSIs. To test model performance across tissue types at different granularity, multiple different class structures were tested (**Supplemental Figure 2**). Eleven different classes were manually annotated, including synovium, muscle and tendon, growth plate, bone marrow, cortical bone, trabecular bone, articular cartilage, meniscus, fat, bone marrow fat, and histology artifact (i.e., out of focus). A seven-class, nine-class, and ten-class segmentation task was generated by merging histologically similar tissues, such as merging cartilage and meniscus into the same class (**Supplemental Figure 2**). Overall, we estimate about 250 hours were spent annotating the 11 tissue classes on 94 WSIs.

Due to the gigapixel nature of WSIs, the entire slide cannot be fed directly into a deep learning model. Previous work has shown that WSIs can be broken into patches, in this case 512 pixel x 512 pixel, to perform downstream learning tasks [25]. For semantic segmentation, a custom QuPath script was used to export patches at a 4x downsample while filtering out regions of the scanned slide that lacked annotations or without tissue. Additionally, images were normalized to mean 0 and standard deviation 1 by sampling a subset of patches to get mean and standard deviation RGB statistics.

Semantic Segmentation Models and Training Strategy

Two different training strategies were employed to test model performances (**Supplemental Figure 1B**). First, a mixed training strategy was used, in which slides from Batch A and Batch B were mixed and randomly assigned to a training, validation, and test split, using 70%, 15%, and 15% for each split, respectively (3 splits total). In this training strategy, stratification was performed by batch and disease type to ensure even splitting into each set, and each patch was split on the slide level to prevent data leakage across splits. Next a single batch training strategy was tested to assess how site-specific differences in histology slides may

impact model performance [26]. In this training strategy, Batch A was used for the training and validation sets, and Batch B was used as the held-out test set.

Data augmentation strategies were also tested to assess their impact across the different training strategies (**Supplemental Figure 3**). We had three different levels of augmentation tested, (1) None, (2) Low, (3) Medium, and (4) High. The python package imgaug[27] as used to implement augmentation. Augmentation was applied in the following way: (1) None had zero augmentation, (2) Low had 2-4 augmentation process applied 25% of the time, (3) Medium had 3-7 augmentations applied 50% of the time and (4) High augmentation had 5-11 augmentation process applied 50% of the time during training. Augmentation was randomly selected from 11 different types of augmentations including, horizontal flip (p = 0.5), coarse dropout or pixel dropout from 0.2x the original image resolution (p=0.1), one of three different rotation types at 90°, 180°, and 270°, additive gaussian noise sampled from a normal distribution with mean 0 and variance 0.2*255, blur using gaussian kernel with sigma of 1.5, hue modification using addition (-30,10) and saturation modification using multiplication (0.5,1.5) and linear contrast (0.5,2), brightness adjustment both add(-30,30) and multiply (0.5,1.5), and color change adjustment (3000,8000).

To compare segmentation performance using both conventional machine learning and deep learning methods, we tested two different model architectures. A Random Forest (RF) model implemented with OpenCV [28] in QuPath and an U-Net++ [29] deep learning model implemented in PyTorch [30]. To train the RF, we first applied a Simple Linear Iterative Clustering algorithm (σ = 5, spacing = 20 µm, Max Iterations = 1, Regularization = 0.01) in which each over segmented area was considered a super-pixel. We then performed color and shape feature extraction of each super-pixel, and calculated the average features of super-pixels within 40 and 80 µms. These features were then used in the RF and the tissue class of each super-pixel was used to supervise the model. After the RF was trained, its performance was assessed fully in QuPath. For our deep learning pipeline, we utilized the

segmentation_models_pytorch [31] package for the UNET++ implementation with an ImageNet pre-trained EfficientNet-B5 backbone for the encoder [32]. We utilized the combo loss, which was the arithmetic average of the dice loss and binary cross entropy (BCE) loss during model training [33]. A learning rate scheduler was used to train the model using a step size of 2 and a gamma of 0.5. The learning rate started at 0.025 using the stochastic gradient descent (SGD) optimizer with momentum (0.9) and regularization of $1x10^{-4}$ and was trained for 10 epochs.

When jointly training across 10 different classes, this model provided no prediction for the meniscus class resulting in a mIOU of 0 ± 0 (Supplemental Figure 6), likely due to insufficient training examples (Supplemental Figure 2; 0.9% frequency overall). As the meniscus was important for downstream biological analyses, we developed a strategy to improve predictions for this class. A second U-Net++ model was fine-tuned from the nine-class model (i.e. the model that has cartilage and meniscus merged into one class) by changing the prediction (final) layer specifically to predict between cartilage and meniscus. We then re-trained the full model only on images and Ground Truth annotations (GTs) that contained either cartilage or meniscus within the mixed training set.

Previous work had suggested that U-Net based semantic segmentation can have image boundary level artifacts [34]. Therefore, we assessed how including patch level overlap for prediction can improve model performance. We included experiments using no overlap, 50% overlap and 66% overlap between them with images from both batches (**Supplemental Figure 4**). To analyze the results, we loop through all the predictions (*N*) for the entire WSI and calculate a majority vote for each pixel after thresholding to removes low confidence predictions (pixel value of 75). *N* can be variable depending upon the region, for example if it is on border, but typically is between 4-9.

Semantic Segmentation Model Evaluation, Inference on External Validation Set and Statistical Evaluation

206 To evaluate the semantic segmentation model performance for the validation and test
207 set slides with ground-truth labels, we calculated the mean Intersection over Union (mIOU) and
208 frequency-weighted mIOU to prevent very rare classes from drastically impacting overall model
209 performance [35]. All model hyperparameter optimization was performed on the validation set,
210 and once the above parameters were chosen the models were used to infer segmentation on
211 the test set and the mIOUs were calculated.
212 We used the optimal settings from the training/validation process to evaluate the model
213 performance on the held-out datasets. Pearson's correlation was used to compare the hand
214 drawn synovial tissue area reported in Bell et al [18] with model classified synovial tissue area.
215 The fine-tuned 10-class model was used to infer tissue segmentation on the held-out data [21,
216 36]. Specifically, the UNET++ 9 class model was used to predict tissues classes on each patch,
217 and if the nine-class model had a prediction output for the combined cartilage and meniscus
218 class on a patch, then the patch was passed through the fine-tuned 2 class model to assign to
219 either the meniscus or cartilage class. Predictions were merged by only allowing the fined tuned
220 predictions to be within the predictions from the combined Cartilage-Meniscus class. Once the
221 inference was complete, a ROI was drawn on each slide from femoral growth plate to tibial
222 growth plate around the joint to restrict the downstream analysis to the joint space, subchondral
223 bone, and synovial adjacent tissue. Tissue area was calculated and a One-Way ANOVA with
224 Tukey's post-hoc adjustment was used to detect significant differences.
225
226 *Cell Type Classification Framework and Preprocessing*
227 For cell type classification, a combination of transfer learning and active learning was
228 used to identify several different cell types that exist within the joint tissue. Cell type
229 classification can be broken into a two-step process, (1) segmentation and (2) classification. For
230 cell segmentation, transfer learning was used by leveraging a deep learning model, HoVer-Net
231 [37], pretrained on the PanNuke dataset [38], to extract nuclei regions. Image patches (1024 x

1024 pixels) at 40x magnification were given as inputs, and ROI contours of nuclei were obtained to perform feature extraction upon. These nuclei with their features and labels (detailed below) were then leveraged in a gradient boosted decision tree (GBDT) model to classify cells.

The input for the classifier was derived from features extracted from each ROI generated by HoVer-Net. Specifically, every nuclei from the json file output of HoVer-Net was converted into a ROI (.roi) file to be read into FIJI/ImageJ [39] for feature extraction using a custom script. Detailed workflow for the ImageJ/FIJI analysis is described as follows. Each image was split using the built-in color deconvolution [40] algorithm in FIJI into hematoxylin and eosin color channels. For each channel the nuclei were measured for several different parameters, including morphological quantities (area, perimeter, circularity, feret's diameter, feret angle, aspect ratio, roundness, and solidity) and staining color quantities (mean, mode, min, max, standard deviation, skewness, median, and kurtosis), and the nuclei ROI was enlarged to calculate cell specific cytoplasmic H&E color information for each cell. Characteristics of neighboring cells at several different distance ranges were used to include tissue level context into cell type classification [41]. A final total of 854 features were extracted for the downstream analysis.

Known healthy and pathologic cell types that contribute to inflammatory arthritis were then annotated on both mouse and human tissues (**Supplemental Figures 8** and **10**). The mouse proof-of-concept classification task consisted of bone-embedded cells, vessel cells, adipo-stromal cells (both adipose and stromal cells within fatty tissue), synovial fibroblasts (healthy and pathologic), chondrocytes, lymphocytes, and other synovial lining cells (healthy and pathologic) as detailed in **Supplemental Figures 8**; annotated by subject-matter experts familiar with histologic analysis of these cell types. These cells were annotated on six healthy, eight mild disease and five severely diseased TNF-Tg knee sections. For human samples, a clinically meaningful set of cell types were labeled by a senior pathologist, following a standard cell type hierarchy (**Supplemental Figure 10**). These included stromal/connective tissue cells,

synovial lining cells, synovial fibroblasts, vascular endothelial cells, tissue macrophages/histocytes, lymphocytes, and plasma cells. These cells were labeled on five lymphoid, five diffuse and three pauci-immune cases. Nuclei were then mapped to manually annotated nuclei by checking if a nuclei's centroid (as determined by Hover-Net) was within an annotation mask.

Cell Type Classification Model and Active Learning

A total of 4,712 cells were annotated for mouse cell type classification from seven different classes. Cells were labeled from a total of 19 different slides. A GBDT model was trained for cell type classification, using 5-fold cross validation to select the best models. To evaluate model performance, cell types were predicted for two different biological settings, (1) to identify cell composition changes across different disease severities on the remaining cells (>300,000) on the 19 slides, and (2) identify differences between male and female mice in the context of disease progression in a held-out dataset [18]. Finally, average synovial inflammatory infiltrate scores and average pannus invasion scores were correlated with lymphocyte and synovial lining cell counts respectively (Spearman's Correlation).

Human annotation was the time-consuming step for the cell type classification pipeline. Therefore, we applied the active learning strategy to improve the annotation efficiency for cell annotation of human samples. To develop this strategy, we tested a proof-of-concept active learning strategy using labeled data from the mouse H&E slides (**Supplemental Figure 9**). Active learning is an iterative process that consists of three main steps, (1) annotation, (2) model training, and (3) sample selection for further annotations. Its goal is to select the samples that can lead to the largest model performance improvement when adding to the training data after annotation. To validate the strategy, 100 different rounds of 5-fold cross-validation were performed. Average F1-scores were reported for each class and a macro-F1 score was additionally reported. 25 runs of 5-fold cross validation were removed due to cells from a single

class not being present in both the training and testing sets. For the training dataset for each split, 5% of cells were first randomly selected as the first set of cells selected as being labeled annotated. Subsequently, the GBDT classifier was trained using this randomly selected data. Several different metrics for determining cells for annotation and subsequent model finetuning, including smallest margin uncertainty [42], least confidence uncertainty [43] and entropy-based uncertainty [42] were assessed. The top 5% of cells were added to the training dataset and the cycle of model training and evaluation and new cell annotations continued until the entire training dataset was used. A random selection of cells after shuffling was also tested to compare model performance to the various active learning strategies. The package modal [44] was leveraged in our implementation. Mean and 95% confidence intervals are reports for each subset across the 75 different runs of 5-fold cross validation.

Cell Classification Model Evaluations

Confusion matrices were generated for model prediction along with F1-scores calculated as $2 \cdot \frac{precision \cdot recall}{precision + recall}$, where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$, where TP, FP, FN stand for true positives, false positives, false negatives. Models were tested using known cell types within specific tissue types to evaluate the model qualitatively.

Analysis of Human Synovial Biopsy Slides

Active learning was then leveraged for human cell type classification using H&E-stained slides of human synovial biopsy tissue of RA patients from the AMP consortium as described above. Multiple rounds of cell type labeling were performed with the assistance of active learning, to obtain a total of 2,639 cells grouped in seven different cell types, detailed in Supplemental Figure 9 (stromal/connective tissue cells n=597, synovial lining cells n= 309, synovial fibroblasts n = 189, vascular endothelial cells n = 486, Tissue Macrophages/Histocytes

n = 201, lymphocytes n = 826, and plasma cells n = 310). A cell type classification model using GBDT with five-fold cross validation was trained on this dataset and inferences from the best performing model applied to all cells on the slides within this patient cohort. Summary cell type quantification (total cell counts and percent of total) was then assessed for each patient. Two analyses were performed using the derived cell types from the cell classification model. First cell type proportions were analyzed across different pathotypes. Specifically, statistical significance testing using lymphocyte, plasma cell, and fibroblast slide proportions were evaluated across pathotypes. Second, pathologist-derived, and clinically relevant Krenn inflammation scores were correlated with Spearman's correlation to the counts of the lymphocytes for each slide to assess disease severity across human samples.

Visualization of Data

Uniform Manifold Approximation and Projection [45] visualization was used for feature representations between batches and cell type framework features. Masks for each class were reimported into QuPath [24] for visualization purposes.

Dataset and Code Availability

All original data and analysis scripts will be provided upon reasonable request.

Results

Deep Learning Segmentation Can Identify Major Tissue Types within Mouse Knee Histology and Measure Therapeutic Response

Several model training choices, including patch overlap, training strategy, and use of different amounts of augmentation during training, were empirically derived from initial experiments to inform the final training of the deep learning segmentation model (UNET++) [34, 46]. First, we tested if 0%, 50% or 66% patch overlap was more performative and determined

that 66% overlap performed the best (0% Overlap fwIOU: 0.72 ± 0.04; 50% Overlap fwIOU: 0.93 ± 0.02; 66% Overlap fwIOU: 0.95 ± 0.01; **Supplemental Figure 4A**). Qualitatively, there were less tiling edge artifacts in the 66% overlap vs 50% overlap results (50% - Arrows **Supplemental Figure 4B** vs **Figure 1**). Second, a mixed training strategy was shown to overcome the large staining batch effect (**Supplemental Figure 5A**) commonly seen in histology datasets with all levels of augmentation equally performative (**Supplemental Figure 5B**). However, if the data is restricted to a single batch and data augmentation is introduced, model performance for the single batch training strategy becomes comparable to model performance using the mixed strategy (High Augmentation UNET++: 0.81± 0.02 vs 0.80 ± 0.06 mIOU for mixed and single batch respectively), demonstrating the need for augmentation to generalize across batches with this level of heterogeneity in staining and other imaging variations (**Supplemental Figure 5B and C**). Thus, we choose to employ a mixed training strategy, with 66% patch overlap and with high augmentation to optimize model performance and generalizability.

Once the training strategy was established, model performance was benchmarked across segmentation tasks at multiple tissue granularities and compared with a standard RF model built-in to QuPath (**Figure 1**). As expected, as the number of different tissue types increases, model performance decreases across both the RF model and UNET++ model, and the DL model outperform the RF at all levels. Interestingly, the magnitude of change is smaller for the UNET++ model compared to the RF model. When testing the UNET++ model, using the ten-class granularity, model performance drops from 0.88 ± 0.06 mIOU for the cartilage and meniscus class to 0.83 ± 0.05 mIOU for the cartilage class and 0.0 ± 0.0 mIOU for the meniscus class indicating a complete loss of meniscus identification (**Supplemental Figure 6**). Because defining the amount of cartilage and meniscus is an incredibility important pathologic readout in joint diseases (e.g. pannus invasion at end stage arthritis), we developed a finely tuned two class model and placed it sequentially after the 9-class model. Predictions from both were

incorporated during majority voting process and performance jumps for this fine-tuned resulting ten-class model from 0.72 ± 0.01 mIOU to 0.82 ± 0.02 mIOU (**Figure 1A and B**). Specifically, performance for the cartilage and meniscus classes were 0.9 ± 0.04 and 0.92 ± 0.05, respectively. We additionally observe that the worst performing classes are the artifact class and bone marrow fat class, the two most infrequent classes, suggesting that fine tuning may work to improve performance for one or both (**Supplemental Figure 6)**. Thus, for all subsequent work we either used the 9-class model (termed Original model) for ease of computation or the Fine-Tuned 10 class model for meniscus segmentation because these provided the best predictive performance while encompassing the most amount of tissues.

After training the model on the mouse cohorts as described, the model was then externally validated (i.e., completely independent of the training, testing and validation process above) in two ways. 1) We first validated our Fine-Tuned model directly on previously published data by comparing hand drawn histomorphometry outlines of the synovial tissue on slides (n=9) in the Test set from the TNF-Tg cohort (Batch A) published in Bell et al 2019 [18]. There was a significant positive correlation between the DL segmentation area and the hand drawn area (r^2=0.96, **Figure 2A**) which demonstrates the accuracy of our method with hand drawn histomorphometry. 2) We then validated our model in a real-world setting by collecting 171 slides (64 knees, 2-to-3 histologic levels per knee) from 6-to-8-month-old TNF-Tg mice treated with Anti-TNF therapy for 6 weeks and slides from TNF-Tg and WT naïve and placebo-treated mice used as controls [19, 47] (**Figure 2B**). Male TNF-Tg mice display a robust inflammatory arthritis with synovial hyperplasia and pannus invasion of the distal femur and femoral articular cartilage starting around 3 months and reaching end-stage at 8 months of age, and our model measures this robust response well (**Figure 2C**, WT vs TNF-Tg, p<0.0001). Anti-TNF therapy is known to reduce synovitis yet does not alter trabecular bone loss in mice with established disease (>6 months old) [48, 49]. Our DL segmentation appropriately modeled these well-established structural changes autonomously (**Figure 2 C&D**), and it uncovered that cartilage

387 degradation is also moderately reduced when Anti-TNF therapy is provided at 8 months of age
388 for 6 weeks, which is an expected but novel result. Interestingly, trabecular bone area is already
389 decreased in 6-month-old TNF-Tg mice compared to WT counterparts while cartilage area is
390 not, suggesting that trabecular bone loss occurs before cartilage loss. Additionally, multiple
391 other tissue structures can be studied simultaneously (**Supplemental Figure 7**) showing the
392 versatility of studying H&E segmentation models to assess tissue structural changes within the
393 context of mouse disease models. Taken together, these analyses suggests that our model
394 accurately detects relevant and meaningful biologic treatment effects with the potential to
395 discover novel structural changes. It is important to note that some of these slides were stained
396 with a variation of traditional H&E, H&E-Orange-G (see *Methods*) demonstrating that our
397 training strategy choices produced a model that is robust even when introducing stain
398 variations.
399
400 *Tissue-Specific and Arthritis Effector Cell Types can be Identified with ML in TNF-Tg Mice*
401 We annotated a total of 4,712 cells across three stages of murine inflammatory arthritis
402 in the TNF-Tg mice (n=6 healthy, n=8 mild disease, and n=5 severe, **Supplemental Figure 8**)
403 to build a cell type classification model that could recognize cells from various disease stages.
404 We developed novel custom feature extraction methods, and the fidelity of these methods was
405 demonstrated in 2D UMAP space, as most cell types are clearly separable (**Figure 3A**). We
406 next built a GBDT classification model using stratified 5-fold cross validation training and tested
407 our models' predictions using three methods. First, we calculated the average (± SD of folds) F1
408 of each cell class in the test sets (**Figure 3B**). Each cell class demonstrated a good F1,
409 between 0.61 for vessel cells and 0.92 for adipo-stromal cells with synovial associated cell
410 classifications among the best (synovial lining cells = 0.87 ± 0.03; synovial fibroblasts = 0.79 ±
411 0.05; and synovial lymphocytes = 0.91 ± 0.05). To validate our classification, we used two
412 strategies that utilized tissue and disease context. Tissue segmentation using the Original

model was performed on these training slides and the remaining >300,000 cells cell-type classification was inferred. Tissues with known homogenous cell types, fat tissue and cartilage/meniscus, were investigated and cell types among these tissues were plotted. Within these adipose tissue and cartilage/meniscus, the most predicted cell type were adipo-stromal cells and chondrocytes, respectively (**Figure 3C**). Next, to assess predictions in the context of disease, we utilized the synovial tissue predictions only (as determined by the tissue segmentation model **Figure 1**) and stratified by disease severity. As shown by **Figure 3D**, synovial-specific increases in synovial fibroblasts, synovial lining cells and lymphocytes are seen with increasing disease severity. These results suggest that our cell type model can produce high quality predictions that are sensitive to disease stage.

Given the promising intra-test set performance and tissue- and disease-state specificity of the cell type modeling, we aimed to further validate our model with a larger data-set with more biologic variation. To do this, we utilize the previously described sexually dimorphic synovial pathology in TNF-Tg mice [18] and collected slides from 3-to-5.5-month-old WT and TNF-Tg-male and female mice. Confirming our previous observation using traditional histologic scoring, we found a significant increase of lymphocytes in female TNF-Tg synovium at 3 months of age with concomitant significant increase in synovial lining cells (**Figure 4A** and **B**). Interestingly, sexual dimorphism was not observed when assessing synovial fibroblasts, which is a novel finding (**Figure 4C**). Finally, we correlated our computationally derived lymphocyte and synovial lining cell counts with the synovial inflammatory score and pannus score, respectively (**Figure 4D** and **E**), and found good correlation between the parameters. This relationship also held for total cell counts (**Figure 4F**).

To reduce the annotation time, we explored various active learning approaches retrospectively in the murine cell type data set. Model performance using all the active learning strategies with 45% of the total training size was comparable to using the complete dataset with random sampling (0.8188 (0.8157-0.8219) vs 0.8213 (0.8184-0.8243), F1 ± SD). Additionally,

with 45% of the data, the model performance using active learning was higher than using a randomly sampled set of examples (0.8188 (0.8157-0.8219) vs 0.8082(0.8053-0.8111) F1 ± SD) (**Supplemental Figure 10**). Additionally, the mean 5-fold CV macro F1-score was most different at the 10%-25% of annotated data range, indicating that active learning can drastically improve model performance with fewer examples.

Cell Type Modeling on Human RA Synovial Biopsies Predicts Pathotypes and Correlates with Clinical Outcomes

After validating the active learning strategy in murine tissue, we next applied this approach to generate cell type annotations on human synovial tissue sections, aiming to reduce the overall histopathological evaluation time for our pathologist. We collected a small subset of initial annotations, predicted cell types on new cells and calculated the entropy-based uncertainty, ranking the most uncertain cells for future annotation by the pathologist. After multiple rounds with active learning, a total of 2,341 cells were annotated (**Supplemental Figure 11**). Using a gradient boosted decision tree with 5-fold cross validation we achieved model performance ranging from 0.74-0.88 F1-scores for all classes (**Figure 5A**). An example confusion matrix from the best performing fold demonstrates that undifferentiated stromal-connective cells are confused with vascular endothelial cells at the highest frequency and the plasma cells and lymphocytes are also often confused with each other (**Figure 5B**). Example prediction overlays demonstrate the high quality that our model produces for all our given cell types (**Figure 5C**).

To validate model performance, cell types were quantified for the entire synovial biopsy dataset with known across the three known pathotypes. Importantly, our cell type prediction a consistent with the previously described cellular distribution within each pathotype [10, 11, 14, 50]. Specifically, synovial fibroblasts are primarily found in the pauci-immune pathotype and some diffuse cases; while lymphocytes and plasma cells are found primarily in the lymphoid

pathotype (**Figure 6A**). Interestingly, if a simple threshold is placed at 1.1% plasma cell to classify lymphoid (≥1.1%) or diffuse (<1.1%) cases, 19 out of 24 lymphoid cases and 22 out of 29 diffuse cases are correctly identified. Further, lymphocyte counts strongly correlates with the Krenn inflammation score (Spearman's rho = 0.88, **Figure 6B**). Example images of the three pathotypes and their respective predictions demonstrate the cellular distributions of these pathotypes (**Figure 6C**).

Discussion

Here, we demonstrate that multi-scale modeling of synovial histopathology can pathotype RA and inflammatory arthritis in clinically meaningful settings, such as treatment response. Our model will reduce the analytical bottleneck associated with histopathology assessment in both the clinical and preclinical settings allowing quicker times to intervention or hypothesis resolution. Furthermore, it will reduce the amount of accessory immunostaining required for pathotyping by using H&E stains to infer cell types, like lymphocytes and plasma cells, which would otherwise need an immunohistochemical (IHC) stain to confirm specific cell types. This could be very impactful as the diffuse pathotype (plasma cell poor with enrichment in myeloid cells) show improved response to tocilizumab (anti-IL-6R antibody) [50] while the fibroblast rich pauci-immune pathotype shows inadequate response to anti-TNF therapy [11]. While we did not have enough specimens to build a model to classify the pauci-immune pathotype in the AMP-RA data set, we were able to classify a diffuse vs lymphoid pathotype utilizing a simple threshold on plasma cell percent with a sensitivity and specificity of 0.81 and 0.73, respectively.

Computational approaches to understanding tissue and cellular information from histology slides have greatly improved in recent years. Specifically, tools to segment or classify of malignant tissues and cells from biopsy specimens [51], and cell type classification on

cytology blood smears [52] have seen the largest amount of development with many applications acquiring FDA approval [53]. These tools have dramatically increase the throughput of clinical histologic analysis by, for example, red flagging potential malignant cases with >0.95 AUC for further review [54], classifying and counting cells on cytology smears or slides for detection of malignant cells and other pathologies [37, 52]. However, there is a dearth of computational pathology tools outside of cancer or outside of diagnostic settings and within the space of musculoskeletal pathologies [55]. Our work represents the first set of comprehensive tissue and cellular analysis tools for both pre-clinical and clinical phenotyping (in this case pathotyping) in inflammatory arthritis. In adjacent work, Pati and colleagues utilized hierarchical graph convolutional neural networks to integrate information from both the cellular and tissue levels on 2048px x 1536px sized images (0.42 uM^2 pixels) [56]. This model was largely effective at detecting cancerous images with an F1 of 84.9% ± 0.8%, however did not perform well for non-cancerous, pre-cancerous or normal images with F1s of 56.6 ± 1.7%, 66.1 ± 3.7%, and 66.2 ± 1.7%, respectively. In other work, HoVer-Net is a state-of-the-art nucleus segmentation and cell type classification deep learning model for H&E-stained tissues [37]. However, the model was originally trained to identify malignant cells from other stromal (epithelium, fibroblast, muscle or endothelium) or inflammatory cells and only on small image patches. In synovial biopsies from RA patients' further delineation of inflammatory cells is critical to pathotyping. Thus, we utilized the nuclear segmentation portion of the model and transferred this knowledge into our custom feature extraction pipeline to build a model to classify discrete immune cells relevant in RA synovial pathology.

Some computational approaches have already been successfully applied to H&E slides of synovial biopsies of RA patients to quantify cellular changes, such as nuclei density and its association with clinical inflammatory measures [57], and simple counting of CD3+ T-cells or CD68+ macrophages on IHC [58-60]. Further, pathologist scores of specific cells types have been associated with quantitative inflammatory gene expression changes in the RA synovium

[61]. Our approach aimed to incorporate the important cell types from this previous work while producing models that only require H&E-stained tissues. However, RA has been shown to be a complex polygenic autoimmune disorder with various environmental risk factors contributing to multiple etiologies [12, 62, 63]. As a consequence of this complexity, many RA patients are refractory to existing approved therapies [64-67]. This highlights the need for a personalized medicine approach to improve clinical trial design and treatment allocation, as done in recent trials like the R4RA [6] and PEAC [68] that utilize ultrasound guided synovial biopsies and pathotype evaluation to stratify patients. Our models may be able to improve the workflow of these types of clinical trials and reduce the overall cost.

One major benefit of our model is the fact that we used H&E-stained tissues. These tissues are routinely and easily collected and represent a large proportion of historical datasets allowing for larger retrospective studies. While some of the computational pathology tools utilize H&E-stained slides, many have utilized special stains geared towards tissue- or cell type-specific classification [55]. This limits the overall throughput and utility of such pipelines by adding additional steps and costs that may be prohibitive. In addition, using a common stain facilitates transfer learning applications to other musculoskeletal pathologies, like osteoarthritis, bone fracture or disc degeneration, with different etiologies but similar tissue involvement and histopathology requirements.

Histopathology scoring that utilize Likert scales [69] is the gold standard analysis method of histology. For RA, various types of assessment, including Krenn lining and inflammation scores, rely on a consensus grading system to account for the challenge in analyzing H&E-stained tissue [70]. However, many scoring mechanisms quantify high level pathologic features (e.g. percent of area effected) or composite measures, without the level of detail included in our pipeline. The added benefit of quantifying cellular abundance and discrete tissue margins is that it can increase our ability to detect and predict the sensitivity to biological or therapeutic differences over Likert scale-based histology scores. For example, in our

previous work we used a Likert scale to quantify synovial inflammatory infiltrate in which a score of 3 corresponded to ">30 inflammatory cells thick" while in our current work we use our computational approaches to count the specific number of lymphocytes that are present in the synovium. This is evident when the effect size (Glasse's dela) is calculated for Female vs Male, TNF-Tg, 3-month-old of the scores vs the cell counts is compared (Histologic Scoring: 2.79 vs Lymphocyte Counts: 74.82) suggesting our cell classification model is orders of magnitude more sensitive to biologic differences than histologic scoring. These differences in quantification method represents ~25 fold increase in measured effect size at 3 months between female and male TNF-Tg mice (Figure 5A) using the computational approach, suggesting a greatly enhance the ability to detect very small differences in therapeutic or other interventions to modulate the pathology. In addition to more granular quantification of pathology these analyses are also more efficient. For example, to generate the annotations to build our segmentation model we spent 200+ hours drawing annotations on the 94 slides. However, to infer the tissue segment on the 174 slides in Figure 2 the model took ~30 hours of hands-off compute time with only 2-3 hours of labor to visualize the results, representing a ~120-fold increase in efficiency.

Lastly, utilizing computational tools that quantify multiple tissues and cell types improves the ability to find novel phenomena. For example, while our primary focus is on the synovial pathology in inflammatory arthritis, having a model that measures cartilage, meniscus, and bone pathology provides a more comprehensive picture of disease. This allows easier detection of off-target or unexpected therapeutic effects with a singular methodology. However, reliance solely on computational modeling maybe increase false positives and expert-level quality control is advised for high impact results.

In conclusion, we have developed a set of models that can characterize tissue and cellular pathology in pre-clinical and clinical inflammatory arthritis settings. These models can be leveraged to better understand disease mechanisms in pre-clinical models and be used in a precision medicine pipeline to improve patients' health.

Acknowledgments

Accelerating Medicines Partnership Program: Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP RA/SLE) Network includes:

Jennifer Albrecht[1], William Apruzzese[2], Brendan F. Boyce[1], David L. Boyle[3], Michael B. Brenner[2], S. Louis Bridges Jr.[10], Christopher D. Buckley[4], Jane H. Buckner[5], Vivian P. Bykerk[1,3], James Dolan[2], Thomas M. Eisenhaure[6], Andrew Filer[4], Gary S. Firestein[3], Chamith Y. Fonseka[2,6], Ellen M. Gravallese[7], Peter K. Gregersen[8], Joel M. Guthridge[9], Maria Gutierrez-Arcelus[2,6], Nir Hacohen[6], V. Michael Holers[7], Laura B. Hughes[10], Eddie A. James[5], Judith A. James[9], A. Helena Jonsson[2], Josh Keegan[2], Stephen Kelly[11], James A. Lederer[2], Yvonne C. Lee[12], David J. Lieb[6], Arthur M. Mandelin II[12], Mandy J. McGeachy[13], Michael A. McNamara[1,3], Joseph R. Mears[2,6], Nida Meednu[1], Fumitaka Mizoguchi[2,14], Larry Moreland[13], Jennifer P. Nguyen[2], Akiko Noma[6], Chad Nusbaum[6], Harris Perlman[12], Javier Rangel-Moreno[1], Christopher T. Ritchlin[1], William H. Robinson[24], Mina Rohani-Pichavant[24], Cristina Rozo[3], Karen Salomon-Escoto[7], Jennifer Seifert[7], Anupamaa Seshadri[2], Kamil Slowikowski[2,6], Danielle Sutherby[6], Darren Tabechian[1], Jason D. Turner[4], Paul J. Utz[15], Gerald F. M. Watts[2], Kevin Wei[2], Costantino Pitzalis[16], Deepak A. Rao[2], Michael B. Brenner[2], Soumya Raychaudhuri[2]

[1]University of Rochester Medical Center, Rochester, NY, USA.

[2]Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.

[3]University of California, San Diego, La Jolla, CA, USA.

[4]University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK.

[5]Benaroya Research Institute at Virginia Mason, Seattle, WA, USA.

[6]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[7]University of Massachusetts Medical School, Worcester, MA, USA.

[8]Feinstein Institute for Medical Research, Northwell Health, Manhasset, New York, NY, USA.

[9]Oklahoma Medical Research Foundation, Oklahoma City, OK, USA.

[10]University of Alabama at Birmingham, Birmingham, AL, USA.

[11]Barts Health NHS Trust, London, UK.

[12]Northwestern University Feinberg School of Medicine, Chicago, IL, USA.

[13]University of Pittsburgh School of Medicine, Pittsburgh, PA, USA.

[14]Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan.

[15]Stanford University School of Medicine, Palo Alto, CA, USA.

[16]Centre for Experimental Medicine & Rheumatology, William Harvey Research Institute, Queen Mary University of London; London, UK

Figure Legends

Figure 1. A fine-tuned 10 class model can segment relevant tissue in inflammatory arthritis. Mean Intersection over union (mIOU) and class frequency weighted mIOU statistics from the RF and DL segmentation models are presented in A, and demonstrate that UNET++ 9 Class model that was then fine tuned to segment the meniscus and cartilage performs the best. Representative images of H&E (B) image, with Ground Truth (GT, C), RF (D), UNET++ (E) and the Fine-Tuned model's prediction overlays demonstrate the successful segmentation and mistakes. Specifically, the area where the synovial pannus tissue reaches the cartilage (B^{ii} - E^{ii}), the RF model (D^{ii}) underperforms while the UNET++ (E^{ii}) are almost perfect. A common mistake within all models is misclassifying the trabecular bone (B^{iii} - E^{iii}) for other tissues, with the RF (D^{iii}) model misclasses for cartilage, muscle and tendon, and cortical bone; while UNET++ (E^{iii}) misclass only as cortical bone. When investigating the overall misclassification of

cortical bone for trabecular bone it occurs in <5% of all tissue area. Importantly, the Fine-Tuned model (F) successfully segments the Cartilage and Meniscus from each other (F^{ii} and F^{iii}).

Figure 2. The fine tuned UNET++ model measures treatment response in the TNF-Tg with Anti-TNF therapy. Inferred test set synovial area correlated excellently with previously published hand drawn histomorphometry data (A). We then developed an independent cohort of slides from previously published work with a biologically relevant therapeutic intervention to test if our model is sensitive to detect these well described therapeutic response (B). As predicted from previous work, TNF-Tg have increased synovial area and lower trabecular bone area, with Anti-TNF therapy decreasing synovial area but does not alter trabecular area (C and D). Interestingly, we are able to detect the loss of cartilage area at end stage disease compared to younger TNF-Tg mice, and the therapeutic intervention protects from this loss (C, Right Panel).

Figure 3. Cell type classification model successfully identifies important cell types in inflammatory arthritis. Utilizing Principal Component Analysis to reduce features and Uniform Manifold Approximation and Projection (UMAP) to project in 3D, we see good separation of Lymphocytes, Bone-Embedded Cells, Chondrocytes, and Adipo-stromal Cells (A). The gradient boosted decision tree performs well with an overall F1 of 0.83 ± 0.12 (M ± SD) and the class specific F1 is shown in (B). We then predicted the cell class (>75% predicted probability) of all other cells identified on the slides (~300,000). To provide quality control, we looked at the predicted cells within the Fat tissue and Cartilage/Meniscus and our model appropriately classifies the majority cells within these tissue as fat cells or chondrocytes, respectively (C). We next looked at the cells within the synovium and found that our model classifies fibroblast, lymphocytes and other synovial lining cells well, and their distribution meets the expectation in healthy mild and severely disease knees (D).

644 **Figure 4. Cell type modeling recapitulates the sexual dimorphism of TNF-Tg inflammatory**
645 **arthritis** Using this model to perform cell type predictions in a sexually dimorphic model of
646 inflammatory arthritis, the computational pathology model describes previously demonstrated
647 difference in inflammatory lymphoid infiltrates (here as Lymphocytes, **A**) and synovial
648 hyperplasia (**B**, Synovial Lining cells). Interestingly, there is no difference in fibroblast cell (**C**)
649 counts between TNF-Tg male and female mice. Importantly, we found a significant correlation
650 between lymphocyte counts and the inflammatory infiltrate score (**D**); synovial lining cell counts
651 and pannus invasion score (**E**); and between the total cell count and total cell area (as originally
652 measured, **F**).
653
654 **Figure 5. Cell type modeling correctly classifies synovial stromal and immune cells in RA**
655 **synovial biopsies.** The overall average F1 from the five-folds was 0.81±0.06, with all individual
656 cell types performing well with greater than 0.75 F1 scores (A). Interrogating the confusion
657 matrix (B) from the best performing model, we see that the most common mistakes are that
658 plasma cell and lymphocytes are mistaken for each other and all the stromal cell populations
659 (Stromal/Connective cells, Vascular Endothelial cells and Synovial Lining cells) can be
660 confused. Representative images of cell class predictions with the original H&E on the top row
661 and the cell type predictions overlays on the bottom row (C). F : Fibroblast, L : Lymphoid, M/H :
662 Macrophage/Histocyte, PC : Plasma Cell, S/C : Stromal/Connective Cell, SLC : Synovial Lining
663 Cell; VEC : Vascular Endothelial Cell.
664
665 **Figure 6. Cell type modeling associates with clinical pathotypes and correlates with**
666 **clinical outcomes measures.** Cell type predictions were made on 60 RA biopsy specimens
667 (n=5 Pauci-Immune, n=27 Diffuse, n=26 Lymphoid) and plots of the Synovial Fibroblasts,
668 Lymphocytes, and Plasma Cells percent of total cells demonstrate the known clinical differences
669 among these pathotypes (**A**). We found a significant correlation between lymphocyte counts and

the Krenn Inflammation score (**B**). Representative H&E and prediction overlay images (**D**) demonstrate the cellular diversity between the Pauci-immune, Diffuse, and Lymphoid pathotypes of RA.

Supplemental Figure Legends

Supplemental Figure 1. Training, Testing and Validation with batch effect study design. A) Description of Batch A from the TNF-Tg cohort and Batch B from the ZIA cohort. B) Description of the two training strategies, Mixed Training and Single Batch training.

Supplemental Figure 2. Tissue Hierarchy Annotation Strategy. Description of the hierarchical tissue classification scheme and which tissues are among each of the classes. The overall frequency of each class is also reported on the right.

Supplemental Figure 3. Examples of Image Augmentation. The original image is provided in A with Gaussian and Speckle Noise (B), Drop Out (C), Rotation (D), Hue and Saturation with Drop Out(E), and Color Temperature adjustment (F).

Supplemental Figure 4. Using highly overlapping tiles with a per pixel majority vote produces the best performing model. mIOU and fwIOU statistics of the three tiling mechanisms are show in A (*data is also presented in in Figure 1). Representative prediction overlays of the no overlap tiling strategy which shows very poor performance, and the 50% overlap tiling strategy which has a prediction artifact caused by poor predictions at the edges of the tiles (Arrows, B).

Supplemental Figure 5. A mixed training strategy with high augmentation builds a high performing segmentation model. Staining batch effect is a major issue with histology and we

696 intentionally collected images in two different batches despite both being stained with H&E. The
697 difference in RGB distributions between the two batches (A) is shown in 3D RGB space. The
698 mean and frequency weighted IOU scores of all the mixed training models are similar and
699 outperform the single batch training (B). During single batch training, high augmentation preform
700 that best. The SLIC-RF model is provided as reference and underperforms compared to all DL
701 models. Representative images of both batches with the H&E, GT, RF, No Augmentation and
702 High Augmentation overlays are presented in C. Single batch high augmentation model
703 qualitatively produces more segmentation artifacts (arrows).
704
705 **Supplemental Figure 6. Tables of class specific mIOU's across all models.** The 7 class, 9
706 class, 10 class and 11 class model performance with each class is presented (A-C).
707
708 **Supplemental Figure 7. The UNET++ Fine Tuned 10 class model phenotypes the TNF-Tg**
709 **with Anti-TNF therapy.** The remaining 5 tissue class predictions do not show any meaningful
710 differences between the Anit-TNF and control samples.
711
712 **Supplemental Figure 8. Mouse Cell Type Classification scheme.** We selected 7 basic cell
713 types to classify and annotated on healthy (n=6), mild disease (n = 8) and severely diseased
714 (n=5) TNF-Tg slides. We annotated bone-embedded cells (A, n=312), vessel cells (B, n = 378),
715 adipo-stromal cells (C, n = 506), fibroblasts (healthy, D; diseased, E; n = 749), chondrocytes
716 (articular, F; growth plate, G; n = 625), lymphocytes (H, n = 467), and all other synovial lining
717 cells (healthy, I; diseased, J; n = 1675).
718
719 **Supplemental Figure 9. Active learning can reduce the number of training samples.** Three
720 independent active learning statistics were evaluated (Entropy, Blue; Margin, Red: Uncertainty,
721 Purple) to select cells and compared against randomly selecting cells (Green). Using the any of

the active learning uncertainty calculations to select new training data to add outperforms random sampling overall (A, green curve is lower between 5-60%). Bone-embedded cells, Synovial Fibroblasts, synovial fats cells and synovial vessel cells all demonstrate improved performance with active learning while lymphocytes, synovial lining cells and growth plate chondrocytes do not (B-H). Interestingly, maximal training occurred with only ~30% of the total training data set, suggesting this mechanism could be used to generate new annotations on a novel dataset. Note the different scale on D.

Supplemental Figure 10. Human Cell type Classification scheme for Synovial Biopsy Specimens. We identified 13 cases from the Accelerating Medicine Partnership – Rheumatoid Arthritis (AMP-RA) Phase 2 cohort that were distributed between the three RA synovial pathotypes (Lymphoid, n = 5; Diffuse, n = 5; Pauci-Immune, n = 3). A board-certified pathologist annotated 7 cell types among these 13 cases: stromal/connective tissue cells (A, n=541), synovial lining cells (B, n= 314), fibroblasts (C, n = 180), vascular endothelial cells (D, n = 349), Tissue Macrophages/Histocytes (E, n = 176), lymphocytes (F, Green Arrows; n = 555), and plasma cells (F, Blue Arrows; n = 226). In order to obtain these annotations, we used the uncertainty sampling active learning strategy as presented in Supp Figure 9 to select cells for our pathologist to annotate. The table describes the general hierarchy of cell labels, for example, if the pathologist could not further distinguish a stromal cell to be either a vascular endothelial cell or fibroblast, then the cell would be labeled as a stromal/connective cell**.**

References

1. Firestein, G.S., *The disease formerly known as rheumatoid arthritis.* Arthritis Res Ther, 2014. **16**(3): p. 114.

2. Lewis, M.J., et al., *Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes.* Cell Rep, 2019. **28**(9): p. 2455-2470 e5.

3. Wang, J., et al., *Synovial Inflammatory Pathways Characterize Anti-TNF-Responsive Rheumatoid Arthritis Patients.* Arthritis Rheumatol, 2022. **74**(12): p. 1916-1927.

4. Rivellese, F., et al., *Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial.* Nat Med, 2022. **28**(6): p. 1256-1268.

5. Micheroli, R., et al., *Role of synovial fibroblast subsets across synovial pathotypes in rheumatoid arthritis: a deconvolution analysis.* RMD Open, 2022. **8**(1).

6. Humby, F., et al., *Rituximab versus tocilizumab in anti-TNF inadequate responder patients with rheumatoid arthritis (R4RA): 16-week outcomes of a stratified, biopsy-driven, multicentre, open-label, phase 4 randomised controlled trial.* Lancet, 2021. **397**(10271): p. 305-317.

7. Nerviani, A., et al., *A Pauci-Immune Synovial Pathotype Predicts Inadequate Response to TNFalpha-Blockade in Rheumatoid Arthritis Patients.* Front Immunol, 2020. **11**: p. 845.

8. Elmore, J.G., et al., *Diagnostic concordance among pathologists interpreting breast biopsy specimens.* Jama, 2015. **313**(11): p. 1122-1132.

9. Elmore, J.G., et al., *Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study.* bmj, 2017. **357**.

10. Pitzalis, C., S. Kelly, and F. Humby, *New learnings on the pathophysiology of RA from synovial biopsies.* Curr Opin Rheumatol, 2013. **25**(3): p. 334-44.

11. Nerviani, A., et al., *A Pauci-Immune Synovial Pathotype Predicts Inadequate Response to TNFα-Blockade in Rheumatoid Arthritis Patients.* Front Immunol, 2020. **11**: p. 845.

12. Rivellese, F., et al., *Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial.* Nature medicine, 2022. **28**(6): p. 1256-1268.

13. Humby, F., et al., *Synovial cellular and molecular signatures stratify clinical response to csDMARD therapy and predict radiographic progression in early rheumatoid arthritis patients.* Ann Rheum Dis, 2019. **78**(6): p. 761-772.

14. Manzo, A., et al., *Histopathology of the synovial tissue: perspectives for biomarker development in chronic inflammatory arthritides.* Reumatismo, 2018. **70**(3): p. 121-132.

15. Chang, M.H. and P.A. Nigrovic, *Antibody-dependent and -independent mechanisms of inflammatory arthritis.* JCI Insight, 2019. **4**(5).

16. Raciti, P., et al., *Clinical Validation of Artificial Intelligence–Augmented Pathology Diagnosis Demonstrates Significant Gains in Diagnostic Accuracy in Prostate Cancer Detection.* Archives of Pathology & Laboratory Medicine, 2022.

17. Konnaris, M.A., et al., *Computational pathology for musculoskeletal conditions using machine learning: advances, trends, and challenges.* Arthritis Research & Therapy, 2022. **24**(1): p. 1-15.

18. Bell, R.D., et al., *Selective sexual dimorphisms in musculoskeletal and cardiopulmonary pathologic manifestations and mortality incidence in the tumor necrosis factor–transgenic mouse model of rheumatoid arthritis.* Arthritis & Rheumatology, 2019. **71**(9): p. 1512-1523.

19. Bell, R.D., et al., *iNOS dependent and independent phases of lymph node expansion in mice with TNF-induced inflammatory-erosive arthritis.* Arthritis Res Ther, 2019. **21**(1): p. 240.

20. Frasnelli, M.E., et al., *TLR2 modulates inflammation in zymosan-induced arthritis in mice.* Arthritis Res Ther, 2005. **7**(2): p. 1-10.

21. Kenney, H.M., et al., *Persistent popliteal lymphatic muscle cell coverage defects despite amelioration of arthritis and recovery of popliteal lymphatic vessel function in TNF-Tg mice following anti-TNF therapy.* Scientific Reports, 2022. **12**(1): p. 12751.

22. Zhang, F., et al., *Cellular deconstruction of inflamed synovium defines diverse inflammatory phenotypes in rheumatoid arthritis.* BioRxiv, 2022: p. 2022.02. 25.481990.

23. Krenn, V., et al., *Grading of chronic synovitis--a histopathological grading system for molecular and diagnostic pathology.* Pathol Res Pract, 2002. **198**(5): p. 317-25.

24. Bankhead, P., et al., *QuPath: Open source software for digital pathology image analysis.* Scientific reports, 2017. **7**(1): p. 1-7.

25. Campanella, G., et al., *Clinical-grade computational pathology using weakly supervised deep learning on whole slide images.* Nature medicine, 2019. **25**(8): p. 1301-1309.

26. Howard, F.M., et al., *The impact of site-specific digital histology signatures on deep learning model accuracy and bias.* Nature communications, 2021. **12**(1): p. 1-13.

27. *imgaug: a library for image augmentation in machine learning experiments.* 2023 5/21/23].

28. Bradski, G. *The OpenCV Library*. 2000 5/21/23]; Available from: https://opencv.org/.

29. Zhou, Z., et al. *Unet++: A nested u-net architecture for medical image segmentation*. in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. 2018. Springer.

30. Paszke, A., et al., *Pytorch: An imperative style, high-performance deep learning library.* Advances in neural information processing systems, 2019. **32**.

31. Iakubovskii, P. *Segmentation Models Pytorch*. 2019; Available from: https://github.com/qubvel/segmentation_models.pytorch.

32. Tan, M. and Q. Le. *Efficientnet: Rethinking model scaling for convolutional neural networks*. in *International conference on machine learning*. 2019. PMLR.

836 33. Taghanaki, S.A., et al., *Combo loss: Handling input and output imbalance in*
837 *multi-organ segmentation.* Computerized Medical Imaging and Graphics, 2019.
838 **75**: p. 24-33.
839 34. Chan, L., et al. *Histosegnet: Semantic segmentation of histological tissue type in*
840 *whole slide images*. in *Proceedings of the IEEE/CVF International Conference on*
841 *Computer Vision*. 2019.
842 35. Thoma, M., *A survey of semantic segmentation.* arXiv preprint arXiv:1602.06541,
843 2016.
844 36. Stokbro, K., et al., *Does Mandible-First Sequencing Increase Maxillary Surgical*
845 *Accuracy in Bimaxillary Procedures?* J Oral Maxillofac Surg, 2019. **77**(9): p.
846 1882-1893.
847 37. Graham, S., et al., *Hover-Net: Simultaneous segmentation and classification of*
848 *nuclei in multi-tissue histology images.* Med Image Anal, 2019. **58**: p. 101563.
849 38. Gamper, J., et al., *Pannuke dataset extension, insights and baselines.* arXiv
850 preprint arXiv:2003.10778, 2020.
851 39. Schindelin, J., et al., *Fiji: an open-source platform for biological-image analysis.*
852 Nature methods, 2012. **9**(7): p. 676-682.
853 40. Ruifrok, A.C. and D.A. Johnston, *Quantification of histochemical staining by color*
854 *deconvolution.* Analytical and quantitative cytology and histology, 2001. **23**(4): p.
855 291-299.
856 41. Wang, X., et al., *Spatial interplay patterns of cancer nuclei and tumor-infiltrating*
857 *lymphocytes (TILs) predict clinical benefit for immune checkpoint inhibitors.*
858 Science Advances, 2022. **8**(22): p. eabn3966.

42. Joshi, A.J., F. Porikli, and N. Papanikolopoulos. *Multi-class active learning for image classification*. in *2009 ieee conference on computer vision and pattern recognition*. 2009. IEEE.

43. Lewis, D.D. and J. Catlett, *Heterogeneous uncertainty sampling for supervised learning*, in *Machine learning proceedings 1994*. 1994, Elsevier. p. 148-156.

44. Tivadar Danka, P.H., *modAL: A modular active learning framework for Pythonmodular active learning framework for {P}ython}.* arXiv.

45. L. McInnes, J.H., N Saul, L GroBberger, *UMAP: Uniform Manifold Approximation and Projection.* The Journal of Open Source Software, 2018. **3**(29): p. 861.

46. Tellez, D., et al., *Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology.* Medical image analysis, 2019. **58**: p. 101544.

47. Kenney, H.M., et al., *Persistent popliteal lymphatic muscle cell coverage defects despite amelioration of arthritis and recovery of popliteal lymphatic vessel function in TNF-Tg mice following anti-TNF therapy.* Sci Rep, 2022. **12**(1): p. 12751.

48. Yi, X., et al., *TNF-Polarized Macrophages Produce Insulin-like 6 Peptide to Stimulate Bone Formation in Rheumatoid Arthritis in Mice.* J Bone Miner Res, 2021. **36**(12): p. 2426-2439.

49. Shealy, D.J., et al., *Anti-TNF-alpha antibody allows healing of joint damage in polyarthritic transgenic mice.* Arthritis Res, 2002. **4**(5): p. R7.

50. Rivellese, F., et al., *Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial.* Nat Med, 2022. **28**(6): p. 1256-1268.

51. Cifci, D., et al., *AI in Computational Pathology of Cancer: Improving Diagnostic Workflows and Clinical Outcomes?* Annual Review of Cancer Biology, 2023. **7**(1): p. undefined-undefined.

52. Jiang, H., et al., *Deep learning for computational cytology: A survey.* Med Image Anal, 2023. **84**: p. 102691.

53. Muehlematter, U.J., P. Daniore, and K.N. Vokinger, *Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): a comparative analysis.* Lancet Digit Health, 2021. **3**(3): p. e195-e203.

54. Campanella, G., et al., *Clinical-grade computational pathology using weakly supervised deep learning on whole slide images.* Nat Med, 2019. **25**(8): p. 1301-1309.

55. Konnaris, M.A., et al., *Computational pathology for musculoskeletal conditions using machine learning: advances, trends, and challenges.* Arthritis Res Ther, 2022. **24**(1): p. 68.

56. Pati, P., et al., *Hierarchical graph representations in digital pathology.* Med Image Anal, 2022. **75**: p. 102264.

57. Guan, S., et al., *Rheumatoid Arthritis Synovial Inflammation Quantification Using Computer Vision.* ACR Open Rheumatology, 2022. **4**(4): p. 322-331.

58. Kraan, M.C., et al., *Quantification of the cell infiltrate in synovial tissue by digital image analysis.* Rheumatology (Oxford), 2000. **39**(1): p. 43-9.

59. Haringman, J.J., et al., *Synovial tissue macrophages: a sensitive biomarker for response to treatment in patients with rheumatoid arthritis.* Ann Rheum Dis, 2005. **64**(6): p. 834-8.

60. Rooney, T., et al., *Microscopic measurement of inflammation in synovial tissue: inter-observer agreement for manual quantitative, semiquantitative and computerised digital image analysis.* Ann Rheum Dis, 2007. **66**(12): p. 1656-60.

61. Orange, D.E., et al., *Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histologic features and RNA sequencing data.* Arthritis & Rheumatology, 2018. **70**(5): p. 690-701.

62. Firestein, G.S., *The disease formerly known as rheumatoid arthritis.* Arthritis research & therapy, 2014. **16**: p. 1-3.

63. Mizoguchi, F., et al., *Functionally distinct disease-associated fibroblast subsets in rheumatoid arthritis.* Nature communications, 2018. **9**(1): p. 789.

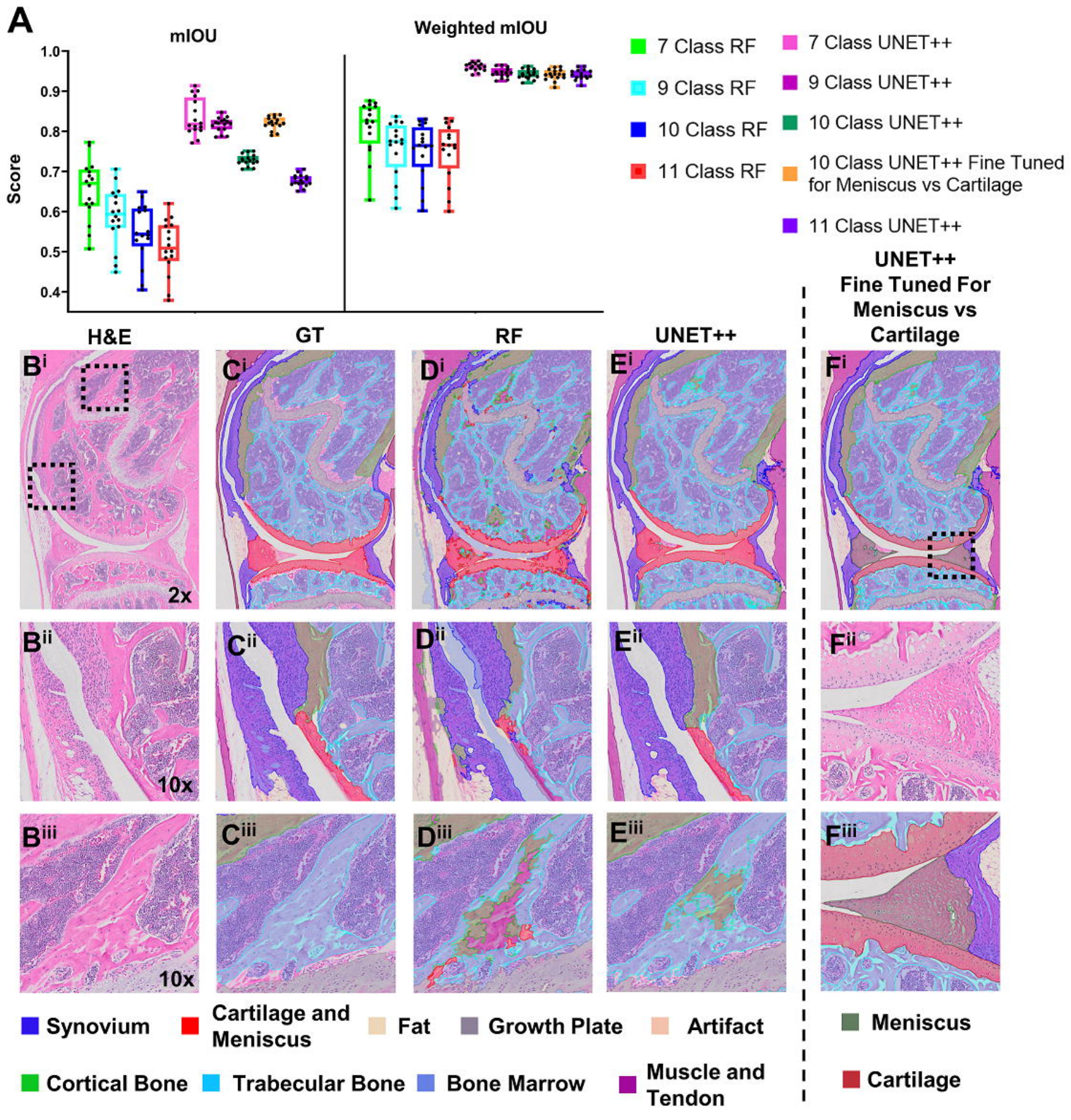
64. Buch, M.H., *Defining refractory rheumatoid arthritis.* Annals of the rheumatic diseases, 2018. **77**(7): p. 966-969.

65. Smolen, J.S., et al., *Rheumatoid arthritis.* Nat Rev Dis Primers, 2018. **4**: p. 18001.

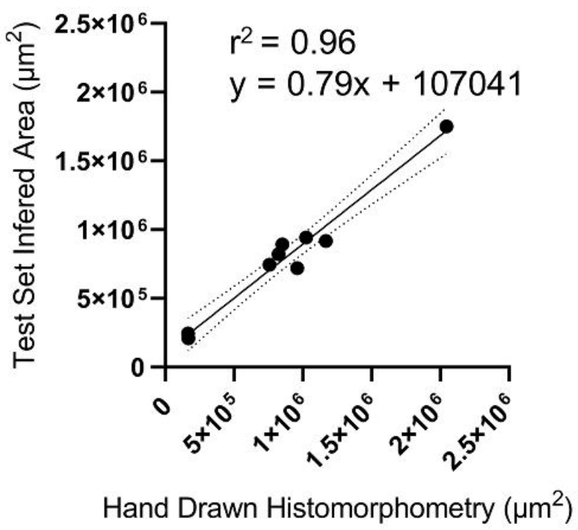
66. Fraenkel, L., et al., *2021 American College of Rheumatology guideline for the treatment of rheumatoid arthritis.* Arthritis & Rheumatology, 2021. **73**(7): p. 1108-1123.

67. Donlin, L.T., *Inching closer to precision treatment for rheumatoid arthritis.* Nature Medicine, 2022. **28**(6): p. 1129-1131.

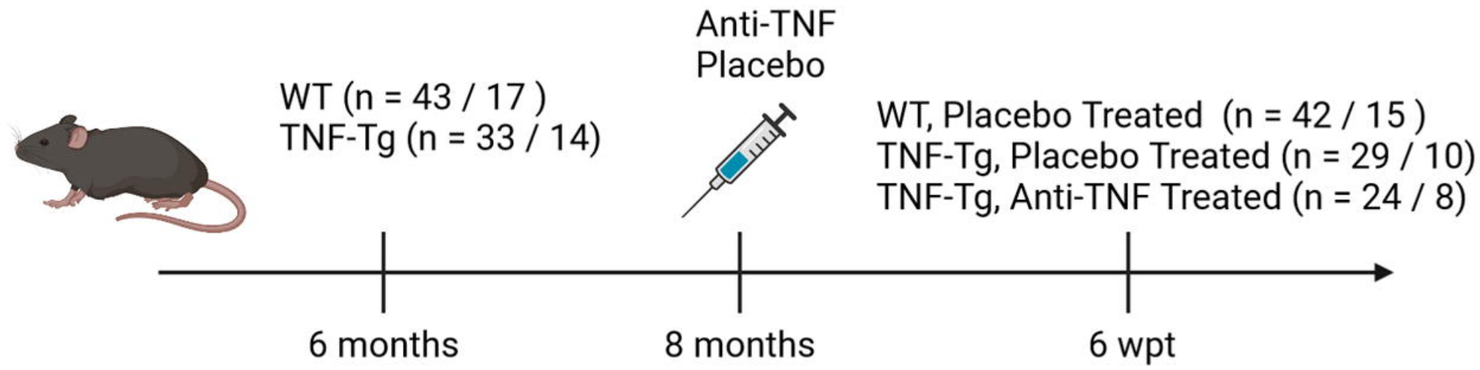
925 68. Lliso-Ribera, G., et al., *Synovial tissue signatures enhance clinical classification*
926 *and prognostic/treatment response algorithms in early inflammatory arthritis and*
927 *predict requirement for subsequent biological therapy: results from the*
928 *pathobiology of early arthritis cohort (PEAC).* Ann Rheum Dis, 2019. **78**(12): p.
929 1642-1652.
930 69. Likert, R., *A technique for the measurement of attitudes.* Archives of Psychology,
931 1932. **22 140**: p. 55-55.
932 70. Zhang, F., et al., *Defining inflammatory cell states in rheumatoid arthritis joint*
933 *synovial tissues by integrating single-cell transcriptomics and mass cytometry.*
934 Nature immunology, 2019. **20**(7): p. 928-942.
935


A
mIOU
Weighted mIOU
Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
7 Class RF
9 Class RF
10 Class RF
11 Class RF
7 Class UNET++
9 Class UNET++
10 Class UNET++
10 Class UNET++ Fine Tuned for Meniscus vs Cartilage
11 Class UNET++
H&E
GT
RF
UNET++
UNET++ Fine Tuned For Meniscus vs Cartilage
Bi
Ci
Di
Ei
Fi
2x
Bii
Cii
Dii
Eii
Fii
10x
Biii
Ciii
Diii
Eiii
Fiii
10x
Synovium
Cartilage and Meniscus
Fat
Growth Plate
Artifact
Cortical Bone
Trabecular Bone
Bone Marrow
Muscle and Tendon
Meniscus
Cartilage

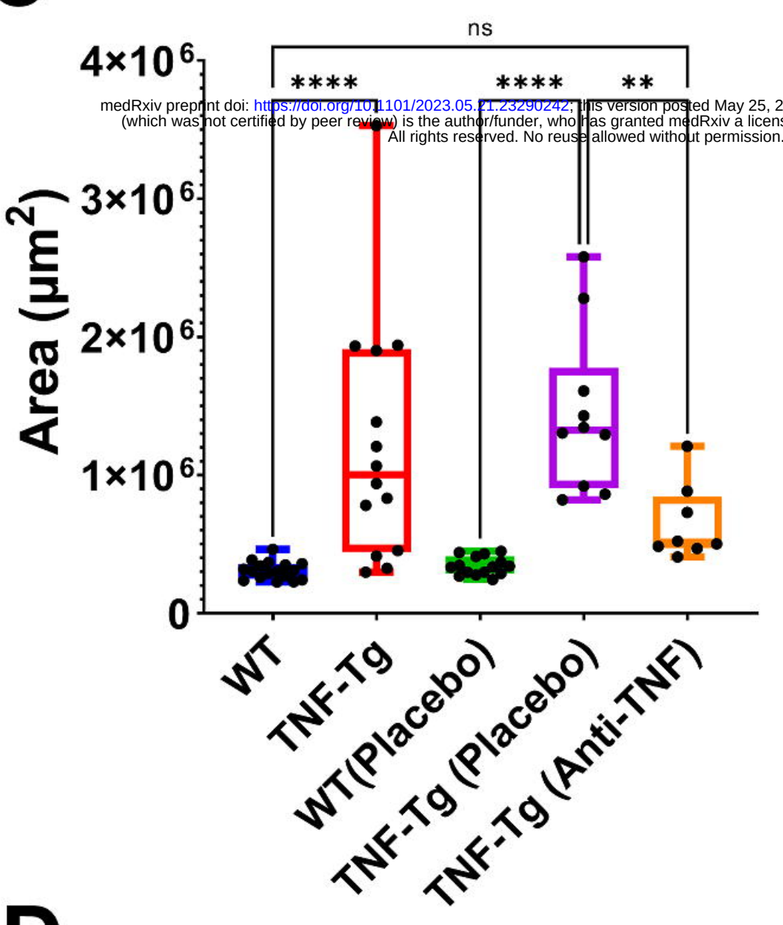
A

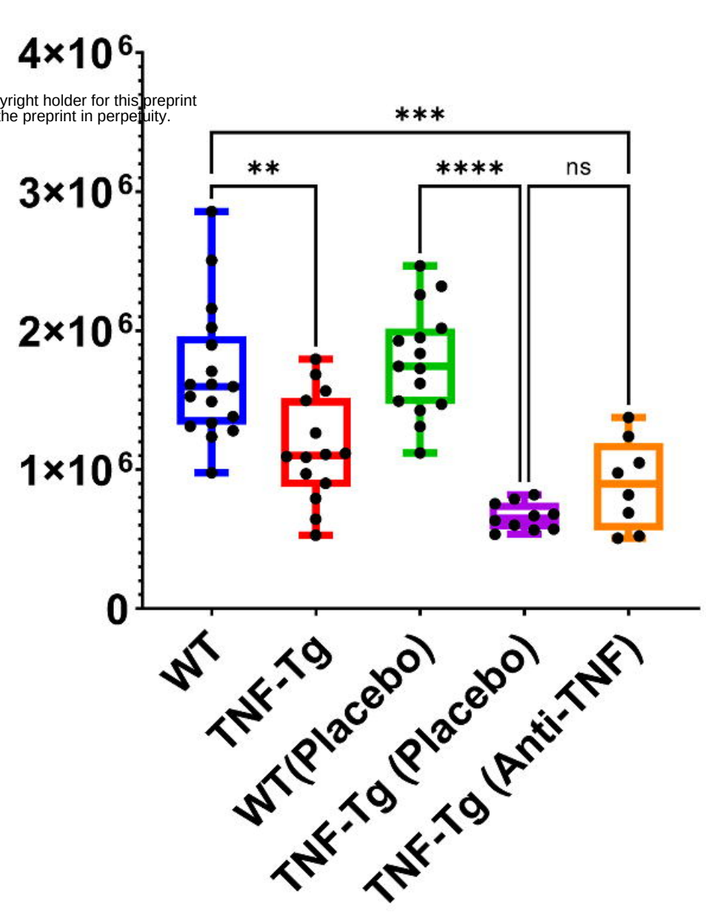


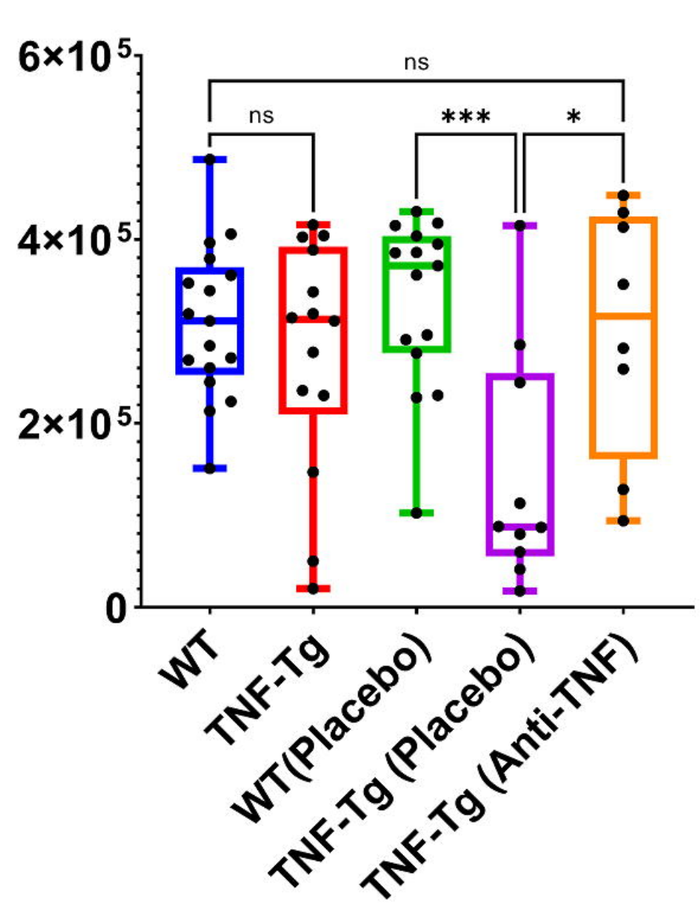
B



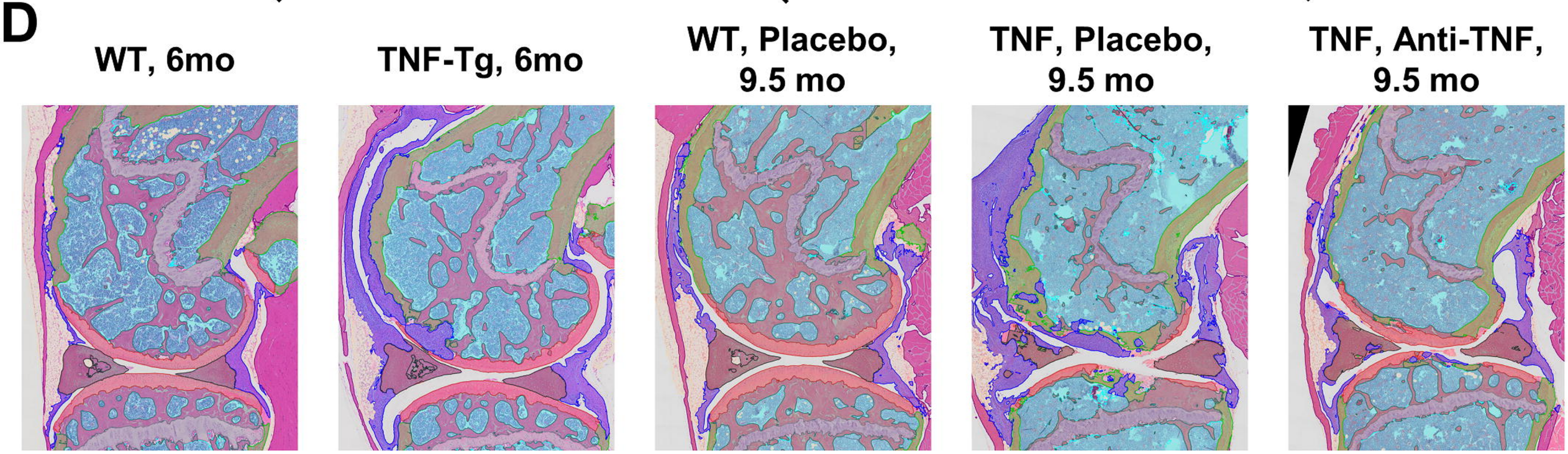
C



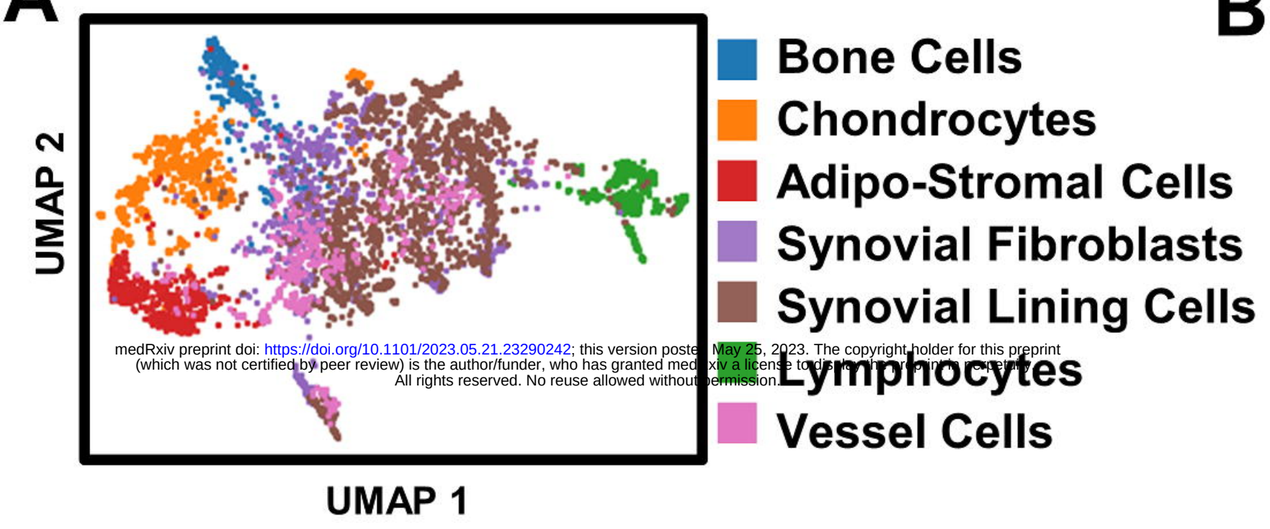




D



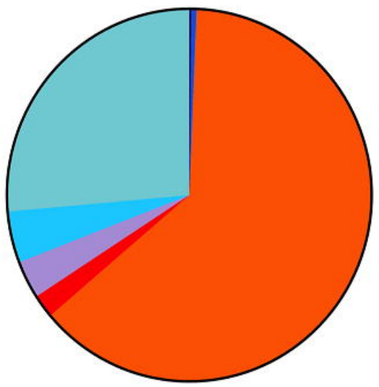
A



B

Class	F1 (Mean ± SD of folds)
Bone-Embedded Cells	0.87 ± 0.03
Chondrocytes	0.89 ± 0.06
Lymphocytes	0.91± 0.05
Adipo-Stromal Cell	0.92 ± 0.04
Fibroblasts	0.79 ± 0.05
Synovial Lining Cells	0.87 ± 0.03
Vessel Cells	0.61 ± 0.16

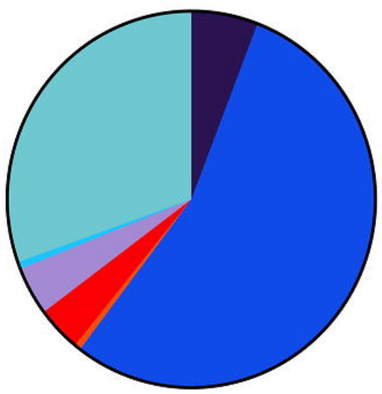
C



Predicted Cells in Fat Tissue

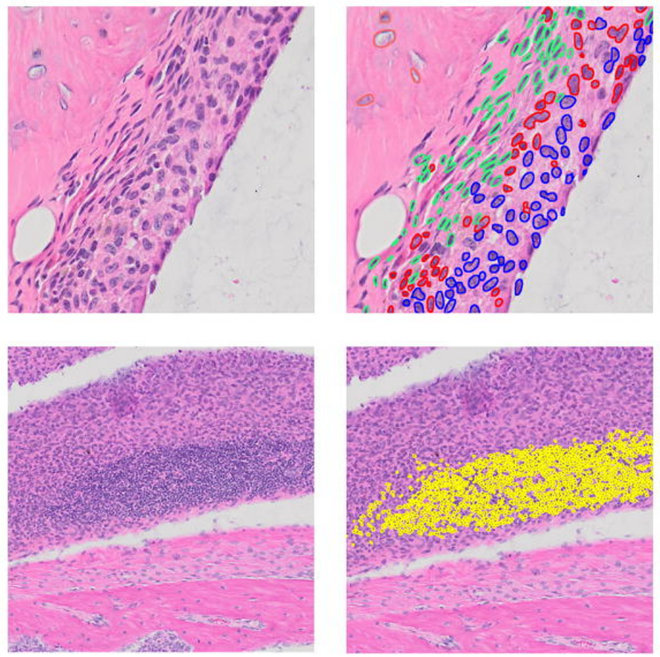
- 63.16% Fat Cells
- 26.41% No Label
- 4.46% Synovial Vessel Cells
- 3.29% Synovial Lining Cells
- 2.04% Synovial Fibroblasts
- 0.41% Chondrocytes
- 0.20% Bone Cell
- 0.03% Lymphocytes

Predicted Cells in Cartilage and Meniscus Tissue

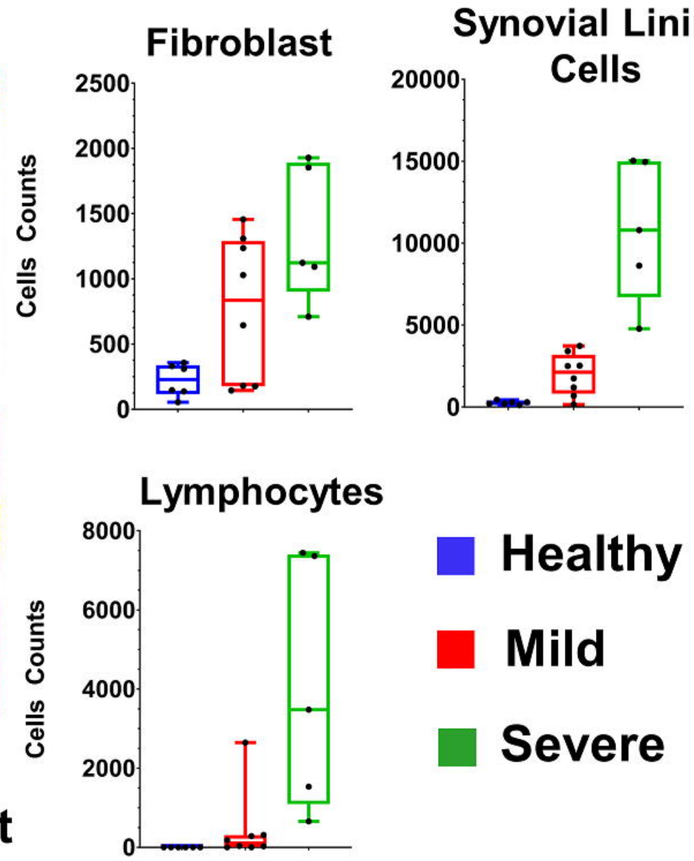


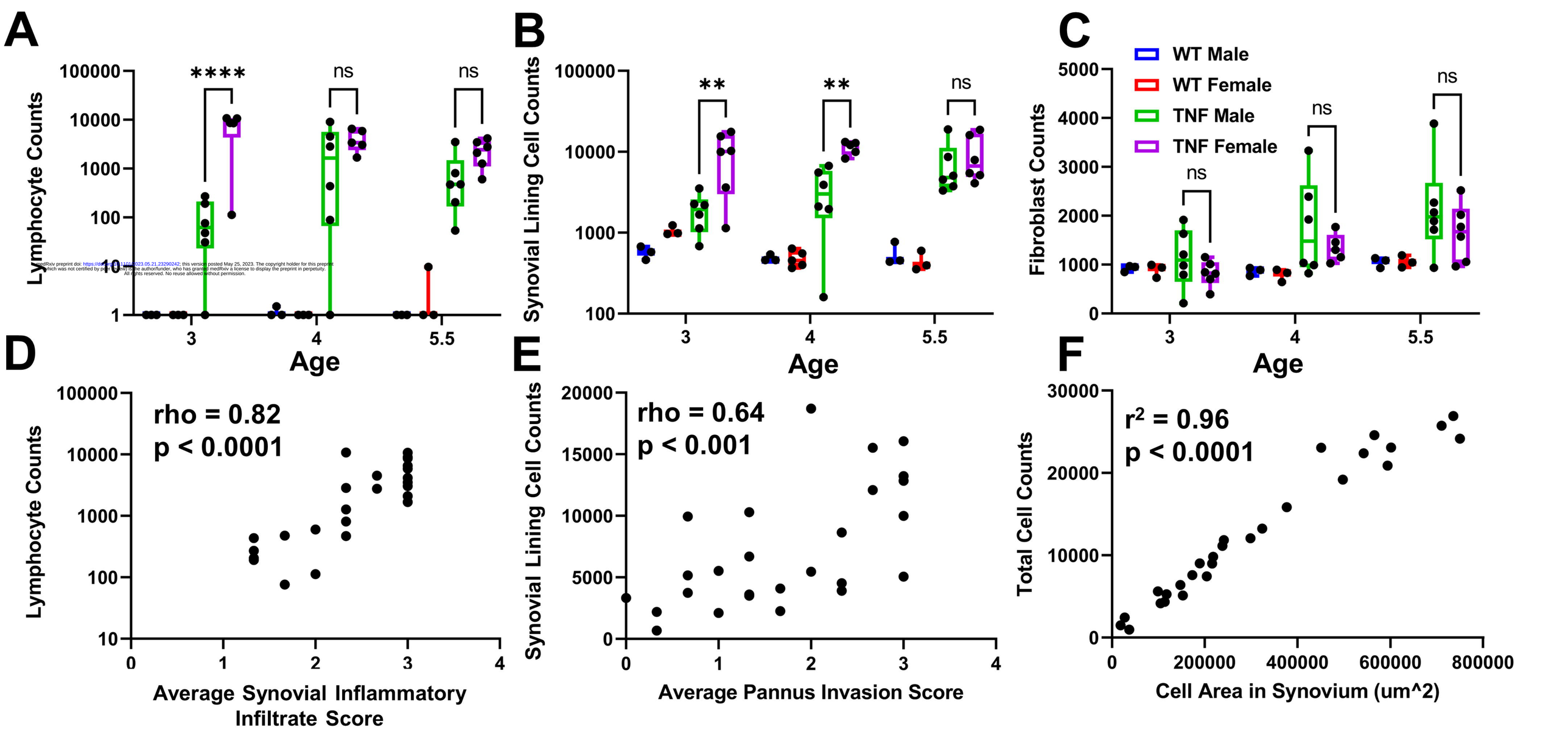
- 54.48% Chondrocytes
- 30.42% No Label
- 5.85% Bone Cell
- 3.91% Synovial Fibroblasts
- 4.18% Synovial Lining Cells
- 0.61% Synovial Vessel Cells
- 0.55% Fat Cells

D



Unlabeled Cells
Fibroblast
Lymphocytes
Synovial Lining Cells

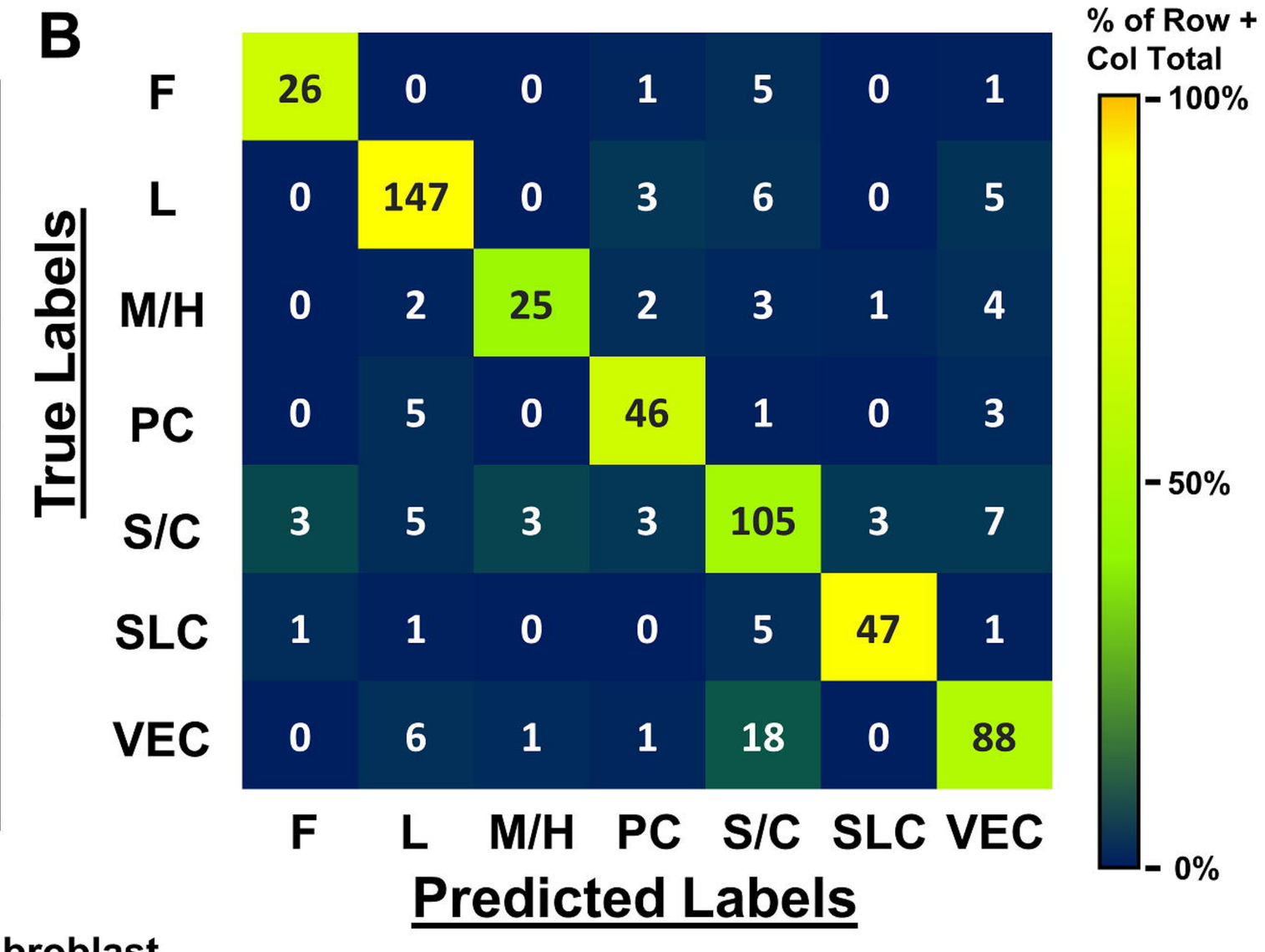




A

Class	F1 (Mean ± SD of folds)
Fibroblasts	0.77 ± 0.04
Lymphocytes	0.87 ± 0.01
Macrophages/ Histiocytes	0.82 ± 0.05
Plasma Cell	0.74 ± 0.06
Stromal/ Connective	0.75 ± 0.02
Synovial Lining Cell	0.88 ± 0.03
Vascular Endothelial Cell	0.79 ± 0.02

B



C