THE LANCET
Digital Health

Supplementary appendix 1

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Digit Health* 2021; published online Oct 19. https://doi.org/10.1016/S2589-7500(21)00180-1.

Table of Contents

Supplementary materials

Dataset details of binary classification problems

The IDaRS deep neural network model is trained for binary classification. For instance, in the case of hypermutated (HM) tumor prediction, the two classes are high mutation density (HMD) and low mutation density (LMD). For MSI pathway prediction, the two classes are microsatellite instable-High (MSI-High) (MSI) and microsatellite stable (MSS) tumors; MSI-Low are combined with MSS in line with general literature as well as another recent study.[1] For chromosomal instability (CIN) pathway, the two classes are CIN and genome stable (GS), and for CPG island methylator phenotype (CIMP) pathway, the two classes are CIMP-High (CIMP-H from hereon) and CIMP-Low (CIMP-L from hereon), whereas for BRAF and TP53 mutation prediction, the two classes are mutant (MUT) and Wild Type (WT). During the IDaRS training, MSI, CIN, CIMP-H and MUT labels are used as positive class labels and MSS, GS, CIMP-L, and WT as negative class labels.

For MSI prediction, a total of 428 WSIs (n=423, where n denotes the number of unique patients) were available with MSI labels in the TCGA-CRC-DX cohort also used in a study.[1,2] The full cohort of 502 TCGA-CRC-DX WSIs is also used for BRAF, KRAS, and TP53 prediction whereas a subset (for which the pathway labels are available) is used for the prediction of HM, CIN and CIMP pathways. **Table S1** lists the number of patients and WSIs in TCGA-CRC-DX, after excluding slides with missing metadata and those with less than ten tumor tiles, and the number of patients in different stage groupings for each pathway.

	n	n_{slides}	Median Age (y)	Gender		Stage I	Stage II	Stage III	Stage IV
				F	M				
HM/CIN	425	430	66.5	216	209	72	149	135	72
MSI	423	428	67.0	214	209	72	149	135	72
CIMP	235	239	68.0	112	123	48	92	72	27
Mutations	497	502	67.0	246	251	72	151	135	144

Table S1. Number of cases (n) and slides (n_{slides}) in TCGA-CRC-DX, and associated statistics for different stage groupings and demographic variables obtained from Liu et al.[3]

Ground truth labels

The ground truth labels of TCGA-CRC-DX for HMD/LMD, MSI/MSS, CIN/GS, and CIMP- H/L are obtained from Liu et al.[3] HMD tumors are defined as those with mutation density > 10 per megabase. MSI pathways are defined as those arising from defective DNA mismatch repair. CIN exhibit marked aneuploidy defined by a clonal deletion score (> 0.0249) and GS lack such aneuploidy. The high and low frequencies of DNA hypermethylation are used to stratify CIMP-H and CIMP-L respectively, where more than 30% of the CIMP-H tumors lack *MLH1* silencing and MSI. The Venn diagrams shown in **Fig S1** illustrate the overlap of MSI cases with other pathways (**Fig S1a**) and HM with MSI pathway (**Fig S1b**) in terms of shared number of cases.

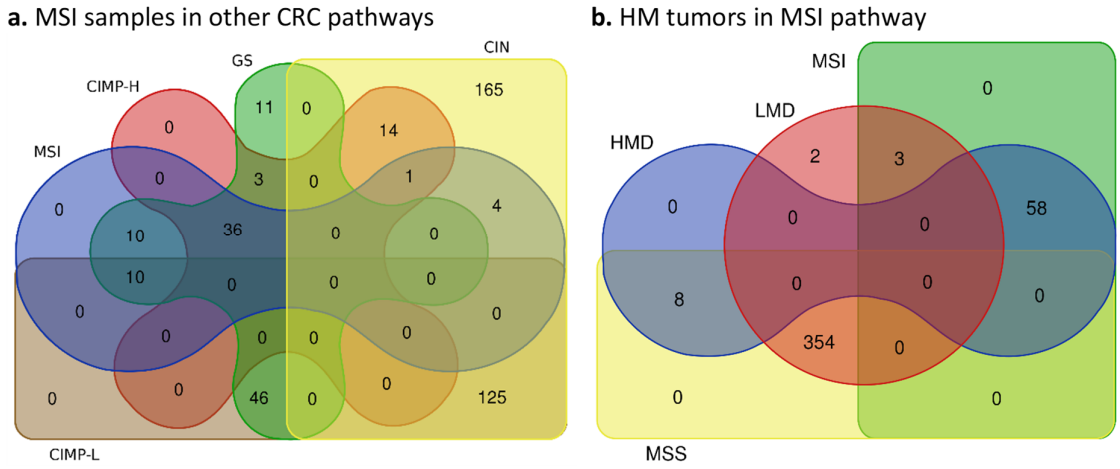


Fig S1 Number of samples shared among different molecular types of CRC. a. MSI samples in CIMP-H, GS, CIN, and CIMP-L subgroups. b. HM tumors in MSI and MSS subgroups.

TCGA cohort images and molecular labels of TCGA-CRC-DX

Whole-slide images for the TCGA cohort can be downloaded using the manifest file added as part of Appendix 2 (MS Excel sheet, GDC_manifest.2021.01.08). The molecular data of TCGA-CRC-DX cohort (obtained from

the molecular data published in Liu *et al.*[3]) is available as part of Appendix 2 (MS Excel sheet , TCGA-CRC-DX Ground Truths). TCGA-CRC-DX cohort with clinical follow-up data is publicly available at http://cancer.digitalslidearchive.net/. The process of slide selection was done blinded to all other clinicopathological variables and outcome data. The process of extracting the regions was blinded to all clinicopathological data, outcome data, and molecular data. More extensive clinical data on the subjects in this cohort are publicly available via the GDC data portal at https://portal.gdc.cancer.gov/projects/TCGA-COAD and https://portal.gdc.cancer.gov/projects/TCGA-READ. Tissue samples in this cohort were provided by multiple institutions in different countries, which are listed at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tissue-source-site-codes.

Supplementary methods

Data preprocessing

All slides are preprocessed for tissue segmentation using Otsu thresholding.[4] Then tiles are extracted from the segmented tissue region only. Following the protocol outlined in[5] and used in[1,6], tissue region in each WSI is divided into tiles of size 512×512, i.e. square tiles of 256 µm edge length at 20× magnification. Slides with less than ten tiles (three slides in total) were excluded in our experiments. For the subsequent downstream analysis through deep learning models, these tile images were resized to 1·14µm/pixel into square tiles of size 224×224, making input images of standard size for ResNet34. Before, training the IDaRS model, the color distribution of all tiles is normalized using structure preserving[7] stain normalization.

Tumor segmentation

To perform the subsequent analysis in tumor regions of the slides, a deep neural network is trained on a balanced dataset of tumor and non-tumor tiles to extract tumor tiles from TCGA slides. Total of 35,436 tiles for the dataset are taken from seven TCGA slides and two publicly available datasets.[1,8] Seventy percent of the data is used for training, fifteen percent for validation to fine tune for the best model and the remaining fifteen percent is held out to test the network performance with an unseen set of tiles. The test set contained 2,493 (47.8%) and 2,720 (52.2%) images of normal and tumor tiles, respectively. The model achieved an accuracy and F1 score of 99% for the test tiles. This model is then used to extract tumor tiles (a total of nearly 450K tiles) from the entire TCGA-CRC-DX cohort.

Neural architecture and training

Our neural network model is a fine-tuned version of ResNet34[9], pre-trained on ImageNet, for the prediction of slide labels, i.e. the status of molecular pathway and genetic mutations. Training is performed for 30 iterations with a batch size of 256. In each iteration, the top 5 tiles are carried to the next training iteration along with 45 randomly selected tiles from each slide. Thus, a training subset of only 50 image tiles from each slide was selected for each iteration.

We use the symmetric cross entropy (SCE) loss[10] during the training process to overcome the risk of training error associated with weak labels, mathematically shown as $\ell_{SCE} = \alpha\ell_{CE} + \beta\ell_{RCE}$. The values of α and β are found empirically ($\alpha = 2$ and $\beta = 3$) by trying different combinations between 0 and 5, as reported in original research.[10] The SCE loss ℓ_{SCE} is a weighted sum of standard CE ℓ_{CE} and reversed cross entropy ℓ_{RCE} losses. The SCE loss enforces a balance between learning and robustness to weakly labeled tile samples.

The deep learning model trained using IDaRS gives a probability score to each tile. This score can be considered as the likelihood of a tile belonging to the positive class in the binary classification setting. Scores of all tiles in a WSI are considered for aggregating into a WSI score and the results are reported for average probability based aggregation after comparing different aggregation schemes – which included majority vote, average top ten probabilities in a WSI, maximum of probabilities, average of probabilities, geometric mean of probabilities, median of probabilities, and average of the probabilities which are greater than the median probability in a WSI – to obtain the best aggregation scheme for each prediction problem.

The training procedure is guided by an Adam optimizer[11] with learning rate and weight decay of $0{\cdot}4\times10^{-4}$. The PyTorch deep learning library[12] was used for the implementation. A set of data augmentation including random rotation with maximum angle of 15 degrees, random horizontal and vertical flip transforms available in PyTorch applied on-the-fly. Additional data augmentations of center-crop, random resized crop, and random crop of image size 224×224 used on all input tiles (of size 256×256) to get multiple of it centered at different spatial locations in the tile. A validation set used for saving checkpoints as an early stopping criterion during the model training. All experiments performed on Nvidia DGX-2 Deep Learning System with 16× 32GB Tesla V100

Volta GPUs in a shared environment. The deep learning model is built on a single GPU equipped with dedicated RAM of 32 GB and 6 worker threads.

Draw and rank sampling

A major challenge in computational pathology lies in discovering and interpreting unknown patterns from the large amount of pixel data in WSIs, especially when only a slide-level label is provided. A typical WSI may contain 150,000×100,000 pixels or more, which is too large to feed into most deep neural networks directly. Therefore, a WSI is usually divided into square tiles (or image patches) before applying deep learning, resulting in a set of tiles with limited visual context. However, the available ground truth is often a high-level label of the WSI instead of well-informed annotations at the level of tissue or cellular regions. A WSI usually consists of tens of thousands of tiles, many with no relationship with the slide-level label, where a significant and meaningful pattern may comprise a small visual field, a tile, or a few of these. It is non-trivial to decide which tiles/patches in each WSI are to be used for training. In addition, having no specific regions annotated at the cellular level is another challenge for effective training of a model.

Conventionally, a deep neural network is trained on all tiles of the training set[1], assuming the ground truth is strongly labeled at the tile level. This is not expedient with only a slide-level label available, as there may be redundant and irrelevant tiles in the WSI resulting in a less than optimal training of the model. Besides, each WSI contributes different number of tiles, between ten to five thousand in our case, potentially introducing training bias to particular samples. The slides which produce a large number of tiles can have a greater influence on the trained model as compared to those which produce a small number of tiles. A random selection of tiles from each WSI may solve the training bias and inefficiency but may not improve the prediction accuracy. In fact, selecting tiles that are most predictive of a given WSI label in a brute force manner is a computationally intractable problem.

In order to reduce the undesirable impact of the aforementioned issues on training, the learning problem of training a deep neural network model $f(.;\theta)$ with trainable weights θ can be modeled as a weakly-supervised machine learning problem with the following empirical error formulation:

$$\min_{\theta} \frac{1}{|S|} \sum_{S_I \in S} \frac{1}{|S_I|} \sum_{x_i \in S_I} \ell\big(y_I, f(x_i;\theta)\big) \tag{1}$$

Here, S denotes the WSI training set with $S_I \in S$ representing the set of tiles $x_i \in S_I$ in the tumor region of a given WSI I with WSI-level label $y_I \in \{-1, +1\}$. This formulation allows choosing the most representative sub-images from each slide by minimizing the loss function $\ell(y_I, f(x_i;\theta))$ between the slide level label y_I and the tile-level output $f(x_i;\theta)$. However, as mentioned above, calculating the loss over all tiles in a given WSI is computationally intractable. Consequently, we propose a draw-and-rank strategy that samples a set of random tiles in a given WSI using the decision function $f(x_i;\theta)$ while maintaining the most representative tiles in an iterative manner across the training epochs. Fig1b shows the concept diagram of the proposed IDaRS method. A more detailed description of IDaRS training is given in Algorithm 1.

The proposed algorithm uses a quasi Monte-Carlo sampling method for selection of tumor tiles from each slide in each training iteration. Specifically, for a given WSI, a representative subset of tumor tiles is obtained based on a minimum loss criterion. This also helps cope with the information density and training bias because of the varied number of image tiles from each WSI. Since random tiles are chosen in a training set, which may or may not contain discriminative tiles in each iteration, each training iteration concludes by carrying the top *k* tiles from the current iteration as a part of the training set through to the next iteration. We choose top *k* tiles from each slide ranked by the maximum probability of a tile for every class label.

IDaRS Training Algorithm

In IDaRS training algorithm (next page), the set S and V represents the input and validation sets of WSIs for IDaRS, respectively, and t and I represent a given iteration and slide. The parameters T, r, *and* k represent maximum training iterations, number of randomly drawn tiles and number of top tiles obtained from each slide per iteration. For our experiments, we set T =30, r =45, and k =5, however, tunning these parameters might be desirable for different weakly-supervised WSI classification problems.

Existing state-of-the-art results

To the best of our knowledge, Kather *et al.*[1] were the first to propose a deep learning based automated prediction of MSI status from H&E images. They use the same deep learning pipeline with more data and multiple cohorts in[2] to predict MSI status from H&E images. The authors used standard fully supervised image classification by

transfer learning the ResNet18 model[9] pre-trained on ImageNet. Standard CE loss was used to train the model with Adam optimizer with learning rate and weight decay of $0{\cdot}4\times10^{-4}$.

Algorithm 1: IDaRS training algorithm.

INPUT: A WSI training set $S=\{S_1,S_2,\ldots,S_N\}$
A validation set of V WSIs
OUTPUT: The trained IDaRS deep neural network model f
PARAMETERS:
T, r, k
STEPS:
1. For each WSI I, initialize S_I^1 with $k + r$ tiles drawn from I uniformly at random
2. For t= 1 : T
   3. For I = 1 : $|S|$
      4. Update weights of f based on loss over S_I^t
      5. Sort tiles in S_I^t according to the maximum prediction score of each tile in ascending order
      6. Initialize S_I^{t+1} with top k tiles from sorted S_I^t and r randomly drawn tiles from S_I
   7. Calculate AUROC over V
   8. Save checkpoint f if validation AUROC improves

The training dataset of color normalized tiles was balanced by under sampling of the majority class (MSS). Tile scores obtained from the deep learning models are aggregated into slide score using average or majority voting. Again to the best of our knowledge, the study by Kather et al.[6] was the first and only one that reports automatic BRAF, KRAS, and TP53 prediction results, which considered mutation prediction as a fully supervised deep learning problem. To overcome the bias resulting from some WSIs having a large number of tiles, each WSI was allowed to contribute a maximum of 1,000 tiles. No color normalization method was used. The *Shufflenet*[13] model pre-trained on ImageNet was used for transfer learning, with the CE loss and an Adam optimizer with learning rate and weight decay of $0{\cdot}4\times10^{-4}$.

Data splits

In 4-fold cross-validation experiments, we used two folds for training, one fold as a validation set for keeping the best performing model and another one held-out for measuring the performance of the model on an unseen test set in a blinded way. In case there were multiple slides of the same patient, we ensure that they were grouped in the same fold. Considering the stochastic element of IDaRS, each fold is executed three times to get average AUROC and standard deviation per fold, which is further averaged over each fold to get average AUROC and standard deviation of multi-fold cross validation. For four-fold cross validation experiments, we have made the splits stratified keeping the same ratios of class imbalance in train and validation sets. **Table S*2*** shows the number of positive, negative, and total slides per fold.

Problem	***Number of slides***	**Fold1**	**Fold2**	**Fold3**	**Fold4**
HM	HMD	17	18	18	14
	LMD	90	90	91	92
	Total	107	108	109	106
MSI	MSI	15	17	16	14
	MSS	92	92	95	87
	Total	107	109	111	101
CIN	CIN	80	79	77	77
	GS	30	30	29	28
	Total	110	109	106	105
CIMP	CIMP-H	14	14	14	13
	CIMP-L	48	46	42	48
	Total	62	60	56	61
BRAF	MUT	14	14	15	16
	WT	109	112	109	113
	Total	123	126	124	129
TP53	MUT	71	76	75	72
	WT	50	52	52	54
	Total	121	128	127	126

Table S2. Cross-validation splits across the positive and negative classes

Supplementary results

Added value by IDaRS

To verify the added value of IDaRS, we have conducted ablation experiments using MSI prediction as a benchmark problem using train/test split Kather et al[1] and 4-fold cross-validation on TCGA cohort used in Kather et al[2], and results are provided in **Table S3** and **Table S4**, respectively. We have also conducted few experiments of IDaRS with varying number of maximum training iterations (T). By continuing the training of IDaRS to T=100, we did not see improvement in the test set or average cross-validation accuracies.

Experiments	AUROC
Standard image classification, Kather et al.[1]	0.77
Standard image classification with early stopping (T=30)	0.81
IDaRS, r=50 and k=0	0.82
IDaRS, r=45 and k=5, tumour tiles from Kather et al.[6]	0.84
IDaRS, r=45 and k=5, CE loss	0.88
IDaRS, r=45 and k=5, T=100, SCE loss	0.86
IDaRS, r=45 and k=5, T=50, SCE loss	0.88
IDaRS, r=45 and k=5, T=30, SCE loss	**0.90**

Table S3. Verifying added value of IDaRS. MSI prediction with train/test split used in Kather *et al*[1] is considered as a benchmark problem for these results. Experiments on IDaRS with k=0, different set of tumor tiles from Kather et al., cross entropy, symmetric cross entropy vs standard image classification with early stopping.

IDaRS (r=50, k=0)	AUROC
CE Loss, T=30	0.8497 (±0.03)
SCE Loss, T=30	0.8594 (±0.04)
CE Loss, T=100	**0.8597 (±0.03)**

Table S4. Ablation experiments with 4-fold cross-validation of IDaRS (r=50, k=0) on TCGA cohort.[2]

CRC pathway prediction in tumors grouped into different stages

Clinical staging of CRC, as in many other cancers, is used to establish the extent of disease spread to help in determining suitable treatment strategies. Knowing the status of MSI is important for the choice of adjuvant chemotherapy and immunotherapy for CRC.[14] Immunotherapy is usually recommended as the first-line treatment for early stage MSI CRCs[15] and second-line treatment for stage IV[16] MSI CRCs due to evidence for its significantly better disease-free survival and overall survival. Patients with stage III MSI-L/MSS CRC have better overall survival with adjuvant chemotherapy.[16] CIN and BRAF mutation have been associated with poor overall survival while the jury is still out for CIMP.[17–23] We consider the accurate prediction of MSI status for CRC tumors to be a strength of IDaRS based digital MSI scores.

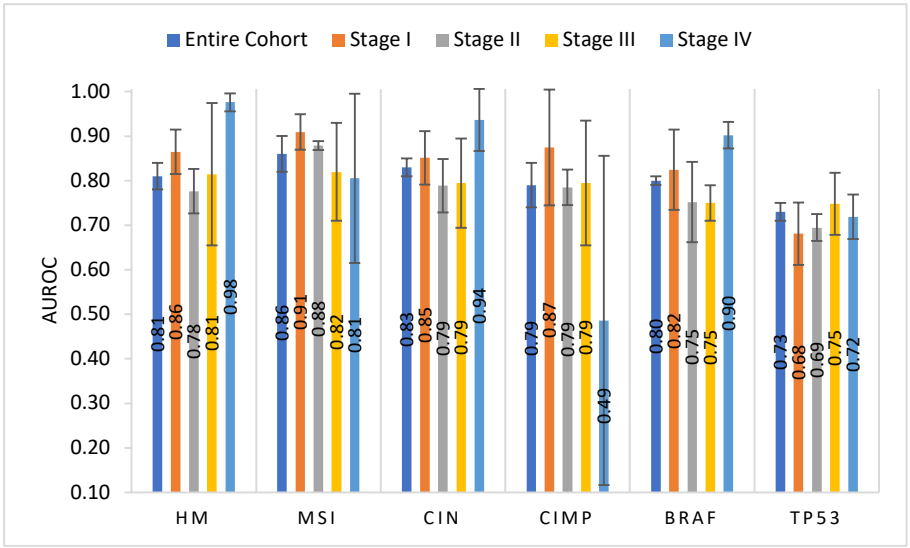


Fig S2. IDaRS predictions for pathways and mutation different stage groups. AUROC on y-axis and pathways and mutations on the x-axis. Entire cohort refers to the whole TCGA-CRC-DX cohort considering patients of all stages as one group. For each stage, AUROC is computed for a subgroup of patients belonging to the same stage while leaving out the patients belonging to other stages.

Considering importance of clinical stage in routine histopathology and CRC pathways and mutations, we divided the entire TCGA-CRC-DX cohort into four groups of patients, one group per stage, and computed AUROC for each stage group separately. We used the IDaRS scores obtained using the models trained during the cross-validation. In **Fig S2** we plot the AUROC of corresponding molecular type, pathway, and mutation prediction for four different stage groups. A stagewise breakdown of the cases is provided in **Table S1**. The prediction performance of same type, pathway, and mutation varies for different stages. Moreover, the number of cases and class ratio also varies for each pathway and mutation in each stage. Among the individual stage groups, HM, CIN, and BRAF are predicted with the highest AUROC in stage IV, whereas digital IDaRS based MSI score gives the highest accuracy for stage I and II. Prediction accuracy of the MSI status for Stage I and II groups is higher than that for Stage III and IV groups. For Stage IV cases, our method produces the highest HM, CIN, and BRAF prediction accuracy of 0·98, 0·94, and 0·90, respectively. CIMP-H and CIMP-L are much better differentiable by our method for stage I to stage III cases than for stage IV cases.

Histological feature discovery of CRC pathways

The pretrained HoVer-Net[24] model processed each of the top predictive IDaRS tiles for nucleus segmentation and classification of nuclei into one of the following five categories: neoplastic and non-neoplastic epithelial, inflammatory, mesenchymal, and necrotic cells, few examples are shown in **Fig S3**.

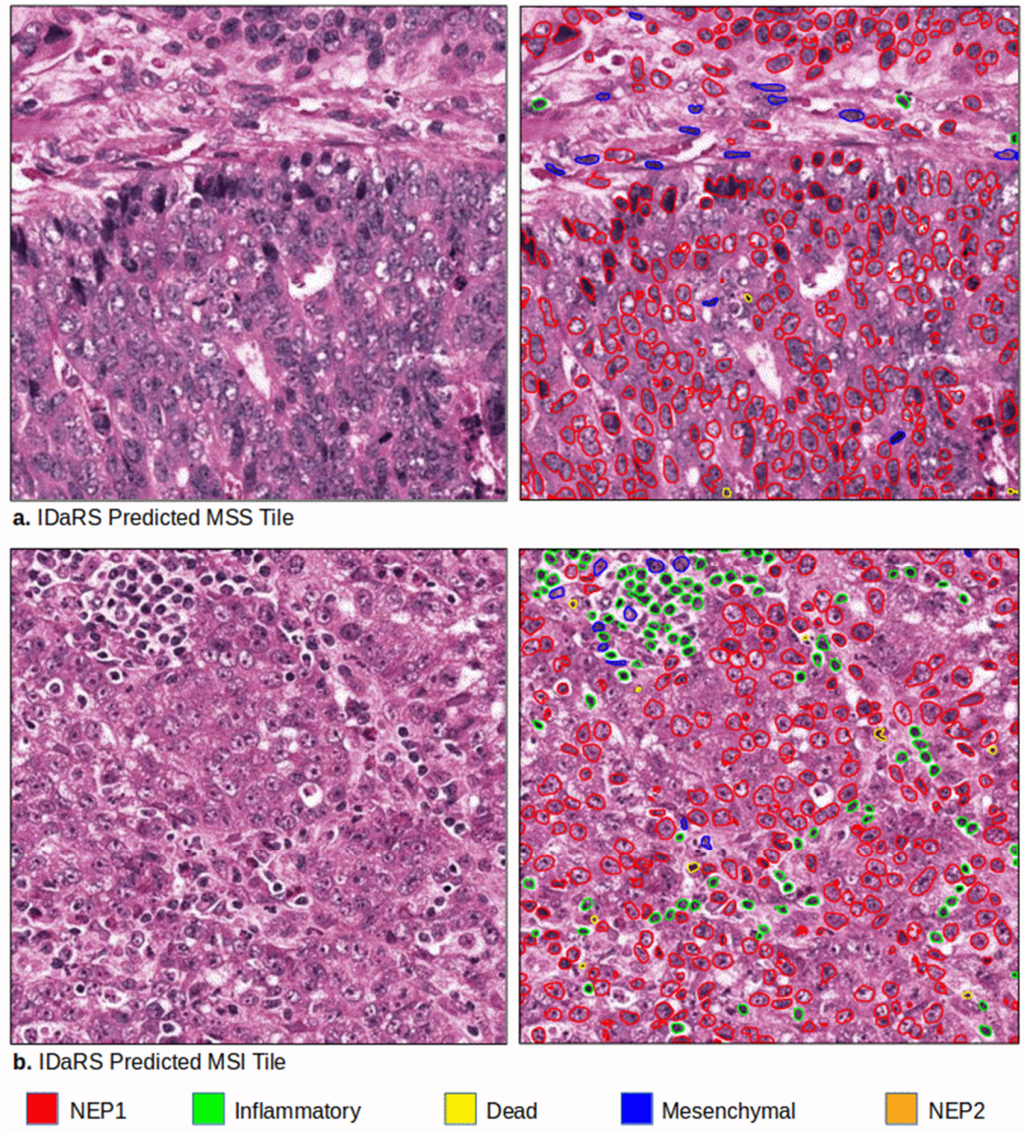


Fig S3. Nuclear segmentation and classification by HoVer-Net. Examples of **a.** top predicted MSS tile (left) with its HoVer-Net output (right) and **b.** top predicted MSI tile (left) with its HoVer-Net output (left) for tumor microenvironment analysis. Detected nuclei of different types are circled with different colors: red: NEP1, green: inflammatory, blue: mesenchymal, yellow: necrotic, and orange: NEP2.

The cell segmentation and classification results of HoVer-Net were visually examined and agreed by expert pathologists for all cell types. Non-neoplastic epithelial cells identified by the network were found to be different from those identified as neoplastic cells, but all of those were categorized as tumor cells by the experts. Neoplastic cells identified by HoVer-Net are termed as neoplastic epithelial type 1 (NEP1). Since only the tumor tiles were used, non-neoplastic epithelial identified by the network are termed as neoplastic epithelial type 2 (NEP2) cells, few examples are shown in **Fig S4** and **Fig S5**.

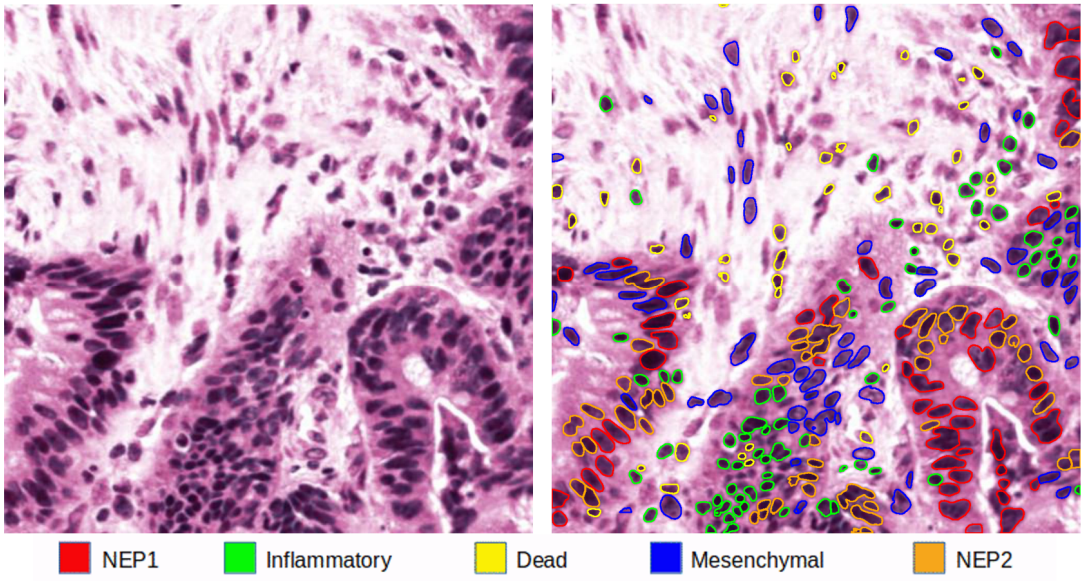


Fig S4. An example top tile of a CIN case with NEP2 type of tumor cells. Other similar tiles selected from CIN cases illustrating differential cellular composition of CIN tiles are included in a PDF file that can be downloaded from necessary additional data attached as (Tiles-CIN-Analysis.pdf).

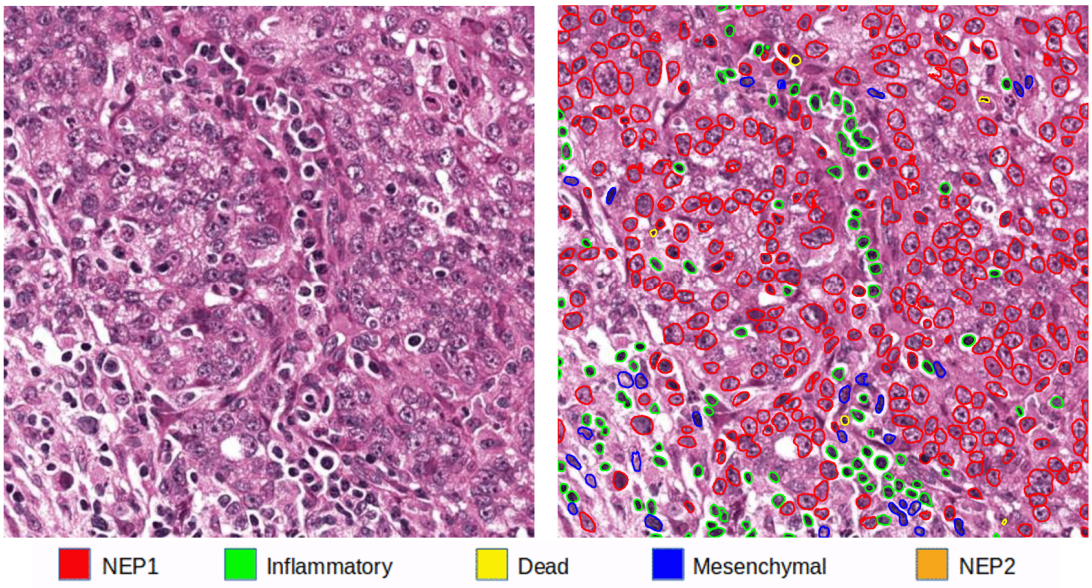


Fig S5. An example top tile of a GS case with NEP1 type of tumor cells. Other similar tiles selected from GS cases illustrating differential cellular composition of GS tiles are included in a PDF file that can be downloaded from necessary additional data attached as (Tiles-GS-Analysis.pdf).

Differential cellular composition analysis

Analysis of the cellular composition is performed on the top ten most predictive tiles in each WSI of the unseen test set (99 WSIs) as predicted by the IDaRS algorithm. For cellular composition analysis, we first employed in-house nucleus segmentation and classification method HoVer-Net[24] to localize, segment and classify different types of cells in a tile into NEP1, NEP2, inflammatory, mesenchymal, and necrosis as shown for some example images in **Fig S3**. We then used the counts of individual cell types in the tile, henceforth termed as its *cellular composition profile*, as features to differentiate between most predictive tiles from corresponding different CRC

pathways (MSI *vs* MSS, CIMP-High *vs* CIMP-Low and CIN *vs* GS) and HM tumors (HMD *vs* LMD) through four separate linear Support Vector Machine (SVM) predictors.

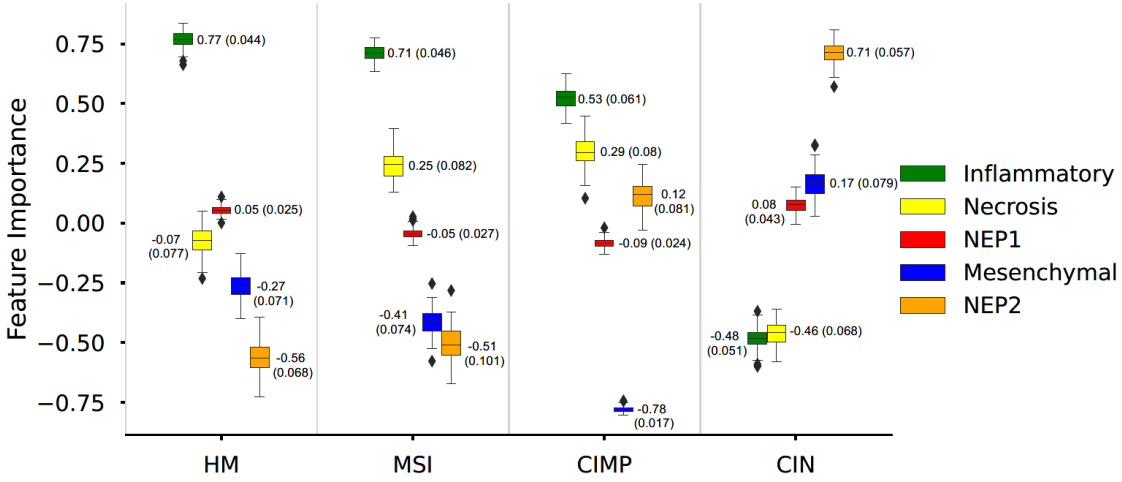


Fig S6 Differential cellular composition as histological features of CRC pathways. Box and Whisker plots of the above data. Normalized weights show the feature significance (y-axis) for HM tumors and three CRC pathways (MSI, CIMP and CIN) on x-axis. The text labels to each box show the medians (interquartile range).

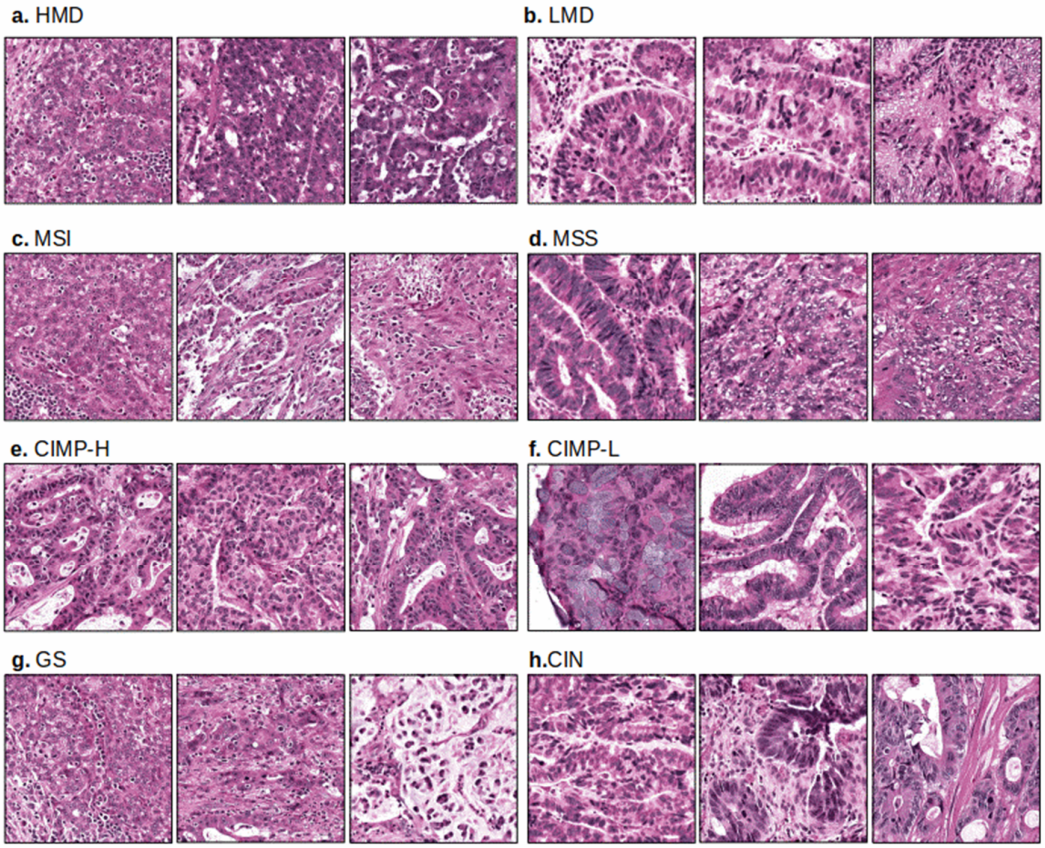


Fig S7 IDaRS discovered most representative visual fields. Examples of IDaRS predicted visual fields from different slides associated to a. HMD, b. LMD, c. MSI, d. MSS, e. CIMP-H, f. CIMP-L, g. GS, and h. CIN classes.

The weight vector of a linear SVM allows an estimation of the relative importance of different cell types in predictions of the CRC pathways.[25–27] Specifically, we analyzed the bootstrap averages of SVM weights across 100 runs over top tiles to get estimates of their relative contribution to the prediction of CRC pathways. The magnitude of different components of the resulting weight vectors as an indication of the feature importance, as shown in **Fig S6** , can be interpreted as the relative degree to which the over- or under- representation of different types of cells (NEP1, NEP2, inflammatory, mesenchymal, and dead) in the top ten tiles is predictive of

the corresponding class label (HMD vs. LMD, MSI vs. MSS, CIMP-High vs. Low and CIN vs. GS). For illustration, few examples of IDaRS predicted most representative tiles for each molecular label are shown in **Fig S7**.

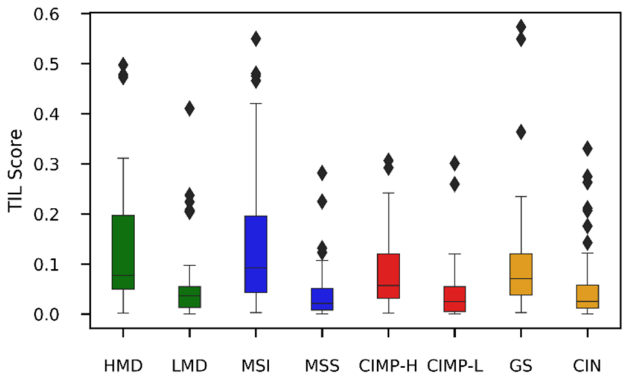


Fig S8. Quantifying TILs in HMD, MSI, CIMP-H, GS, MSS, CIMP-L and CIN. TIL abundance score (*y*-axis) and CRC pathways (*x*-axis). In IDaRS predicted top ten tiles, HMD and MSI samples got the highest TIL score. CIMP-H and GS sample also got higher TIL scores than CIMP-L and CIN.

The prognostic significance of TILs in general and its association with MSI has already been established in existing literature.[28,29] Cellular composition analysis also suggests infiltration by inflammatory cells as a key discriminating histological feature in MSI and HM tumors. Therefore, we quantified TILs of the two subclasses in each pathway and HM tumors graphically shown in **Fig S8** computed from the top ten IDaRS tiles per WSI using a TIL abundance score.[30] A paired t-test found TIL scores to be statistically significant for the MSI pathway (p<0·0024) and HM tumors (p<0·024) but not for the CIMP (p<0·09) and the CIN (p<0·157) pathways.

Overlay visualizations of prediction heatmaps

We overlay local prediction heatmaps on top of WSIs in order to further analyze the relationship of HM tumors and three CRC pathways with spatial features of the tumor microenvironment. We trained three deep learning models separately using the same patient-wise splits of training, validation and testing datasets. Each patient has multiple slide-level ground truths of being HMD/LMD, MSI/MSS, CIN/GS, and CIMP-H/CIMP-L available. A heatmap of one test set sample is shown in **Fig S9** for the visual illustration of different spatial features of the tumor microenvironment predictive in different molecular labels. A heatmap of another test set sample is shown in **Fig S10** for visual illustration.

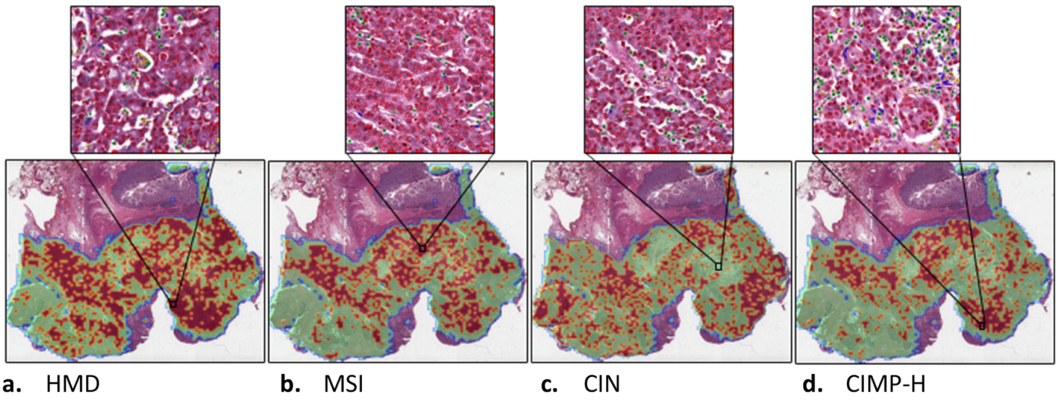


Fig S9 IDaRS tile prediction heatmaps-1. A test set slide labeled as HMD, MSI, GS and CIMP-H, WT for BRAF, TP53 and KRAS mutations. IDaRS assigned tile probabilities of being (HMD, MSI, CIN, CIMP-H) are used to generate a heatmap overlay. In the heatmap, the red color corresponds to the positive class and green to the negative class; a. Overlay heatmap of HMD prediction, HMD: Red. b. Overlay heatmap of MSI prediction, MSI: Red. c. Overlay heatmap of CIN prediction, GS: Green. d. Overlay heatmap of CIMP-H prediction, CIMP-H: Red. An IDaRS predicted top tile is also shown for each label.

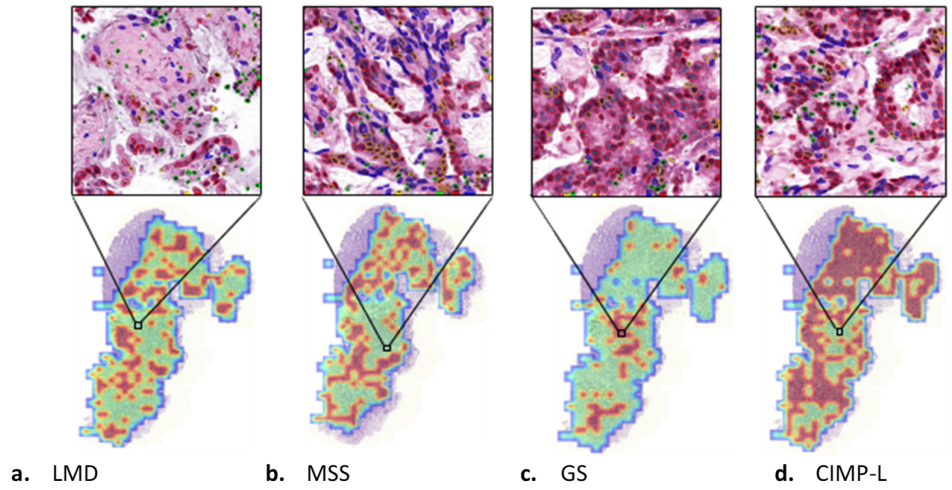


Fig S10. IDaRS tile prediction heatmaps-2. A test set slide with known KRAS mutation and wild type BRAF showing IDaRS derived overlay heat map labeled as LMD, MSS, CIN and CIMP-L. In the heatmap, the red color corresponds to the positive class and green to the negative class; a. Overlay heatmap of LMD prediction, LMD: Green. b. Overlay heatmap of MSS prediction, MSS: Green. c. overlay heatmap of CIN prediction, CIN: Red. d. overlay heatmap of CIMP-L prediction, CIMP-L: Green. An IDaRS predicted top tile is also shown for each label.

As can be observed in **Fig S9** and **Fig S10**, similar tissue regions appear to be predictive of the positive or negative status of the four CRC molecular pathways. For each pathway, both positive and negative tiles are predicted in the corresponding slide. However, tiles associated with the ground truth label are more in number in **Fig S9** whereas fewer in number in **Fig S10**. This could be linked to the difficulty of MSI prediction especially when the immunogenic response of an MSS identified cases is similar to the MSI histomorphology.[31] It may also indicate that a more sophisticated aggregation method than a simple average or majority voting may be needed to correctly classify the case presented in **Fig S10**.

Correlation analysis of the digital IDaRS based scores

We conducted an investigation into correlating the IDaRS based digital scores of HM, MSI, CIN, CIMP, BRAF and TP53 for each slide of the TCGA-CRC-DX test set used in[1] obtained by applying separately trained models on the same training set. **Fig S11Fig S11a** and **Fig S11b** show scatterplots of our IDaRS based digital scores for HM and MSI, and HM and CIMP pathways, respectively. These plots demonstrate strong positive correlation (using the Pearson correlation coefficient *r*) between the digital scores of HM and MSI (r=0·81, $p<10^{-23}$) and also between the digital scores of HM and CIMP (r=0.65, $p<10^{-12}$).

Fig S11c shows the scatterplot of our IDaRS based digital scores for MSI and CIMP, again demonstrating strong positive correlation between the two digital scores (r=0·68, $p<10^{-14}$) in line with existing literature[20,22,23] and showing that most of the MSI samples are also predicted as CIMP-H and MSS as CIMP-L. Among the single labels, MSS shows highest positive correlation of 0·96 ($p<10^{-40}$) with CIMP-L and MSI shows positive correlation of 0·49 with CIMP-H (p<0·07). Our digital scores of MSI and CIN also show strong negative correlation (r=−0·75, $p<10^{-18}$) though it did not necessarily endorse mutual exclusivity for all MSI with CIN cases. Most MSS cases and few MSI cases also got high CIN score, which is in line with previous findings about MSI and CIN (and as can also be observed in the Venn diagrams of **Fig S1**).[17,32] The IDaRS based digital scores of CIN and CIMP also show negative correlation (r=−0·67, $p<10^{-13}$), with high CIN scoring cases being scored low in CIMP and with overlapping GS cases. MSI have highest negative correlation of -0·68 (p<0·0005) with GS whereas CIMP-L has negative correlation of −0·76 with CIN ($p<10^{-12}$). IDaRS predictions also suggest MSI and CIMP are positively correlated to BRAF mutation with correlation coefficients of 0·68 ($p<10^{-14}$) and 0·93 ($p<10^{-42}$) whereas CIN has highest positive correlation of 0·60 ($p<10^{-10}$) to TP53 mutation.

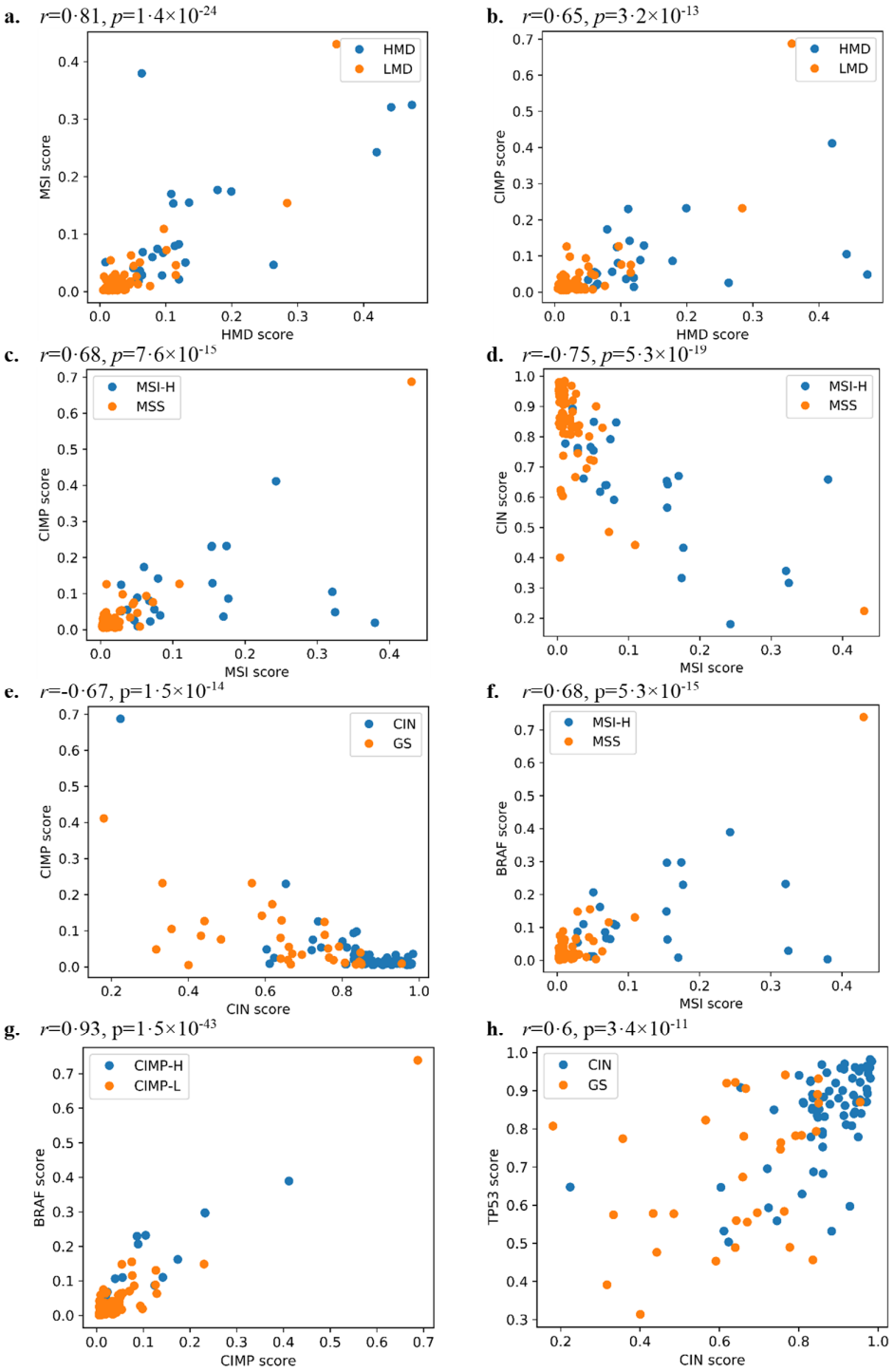


Fig S11. Correlations in pathways and mutations. Digital IDaRS scores for HM, MSI, CIN, CIMP, BRAF and TP53 of each slide of the test set obtained by applying separately trained models on the training set of the same slides: a. HMD and MSI, positively correlated. b. HMD and CIMP, positively correlated. a. b. Points are colored using HM ground truth labels. c. MSI and CIMP, positively correlated. IDaRS predicted MSI score on x-axis and CIMP score on y-axis. d. MSI and CIN, negatively correlated. MSI score on x-axis and CIN score on y-axis. c.d. Points are colored using MSI ground truth labels. e. CIMP and CIN, negatively correlated. MSI score on x-axis and CIN score on y-axis. Points are colored using CIN ground truth labels. f. MSI and BRAF, positively correlated. MSI score on x-axis and BRAF score on y-axis. Points are colored using MSI or MSS ground truth label. g. CIMP and BRAF, positively correlated. CIMP score on x-axis and BRAF score on y-axis. Points are colored using CIMP ground truth labels. h. CIN and TP53, positively correlated. CIN score on x-axis and TP53 score on y-axis. Points are colored using CIN ground truth labels.

Supplementary References

1. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med. 2019;25(7):1054–6.

2. Amelie E, Heike IG, Philip Q, et al. Clinical-grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. Gastroenterology [Internet]. 2020; Available from: http://www.sciencedirect.com/science/article/pii/S0016508520348186

3. Liu Y, Sethi NS, Hinoue T, Schneider BG, Cherniack AD, Sanchez-Vega F, et al. Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. Cancer Cell. 2018 Apr;33(4):721-735.e8.

4. Otsu N. A Threshold Selection Method from Gray-Level Histograms. N/A. 1979;9(1):62–6.

5. Muti HS, Loeffler C, Echle A, Heij LR, Buelow RD, Krause J, et al. The Aachen Protocol for Deep Learning Histopathology: A hands-on guide for data preprocessing [Internet]. 2020. Available from: https://doi.org/10.5281/zenodo.3694994

6. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer. 2020 Aug;1:789–99.

7. Vahadane A, Peng T, Sethi A, Albarqouni S, Wang L, Baust M, et al. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. IEEE Trans Med Imaging. 2016;35(8):1962–71.

8. Shaban M, Awan R, Fraz MM, Azam A, Tsang Y, Snead D, et al. Context-Aware Convolutional Neural Network for Grading of Colorectal Cancer Histology Images. IEEE Trans Med Imaging. 2020;1–1.

9. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

10. Wang Y, Ma X, Chen Z, Luo Y, Yi J, Bailey J. Symmetric Cross Entropy for Robust Learning with Noisy Labels. ArXiv. 2019;abs/1908.06112.

11. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: Bengio Y, LeCun Y, editors. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings [Internet]. 2015. Available from: http://arxiv.org/abs/1412.6980

12. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. 2019. p. 8024–35.

13. Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.

14. Kang S, Na Y, Joung SY, Lee SI, Oh SC, Min BW. The significance of microsatellite instability in colorectal cancer after controlling for clinicopathological factors: Medicine (Baltimore). 2018 Mar;97(9):e0019.

15. Chalabi M, Fanchi LF, Van den Berg JG, Beets GL, Lopez-Yurda M, Aalbers AG, et al. Neoadjuvant ipilimumab plus nivolumab in early stage colon cancer. Ann Oncol. 2018 Oct;29:viii731.

16. Sun BL. Current Microsatellite Instability Testing in Management of Colorectal Cancer. Clin Colorectal Cancer. 2020 Aug;S1533002820301043.

17. Pino MS, Chung DC. The Chromosomal Instability Pathway in Colon Cancer. Gastroenterology. 2010;138(6):2059–72.

18. Singh MP, Rai S, Pandey A, Singh NK, Srivastava S. Molecular subtypes of colorectal cancer: An emerging therapeutic opportunity for personalized medicine. Genes Dis [Internet]. 2019; Available from: http://www.sciencedirect.com/science/article/pii/S235230421930100X

19. Al-Sohaily S, Biankin A, Leong R, Kohonen-Corish M, Warusavitarne J. Molecular pathways in colorectal cancer. J Gastroenterol Hepatol. 2012;27(9):1423–31.

20. Bae JM, Kim JH, Kang GH. Molecular Subtypes of Colorectal Cancer and Their Clinicopathologic Features, With an Emphasis on the Serrated Neoplasia Pathway. Arch Pathol Lab Med. 2016;140(5):406–12.

21. Moreno V, Rebeca S-P. Altered pathways and colorectal cancer prognosis. BMC Med. 2015;13(76):1–3.

22. Shia J, Schultz N, Kuk D, Vakiani E, Middha S, Segal NH, et al. Morphological characterization of colorectal cancers in The Cancer Genome Atlas reveals distinct morphology–molecular associations: clinical and biological implications. Mod Pathol. 2017 Apr;30(4):599–609.

23. Advani SM, Advani P, DeSantis SM, Brown D, VonVille HM, Lam M, et al. Clinical, Pathological, and Molecular Characteristics of CpG Island Methylator Phenotype in Colorectal Cancer: A Systematic Review and Meta-analysis. Transl Oncol. 2018;11(5):1188–201.

24. Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal. 2019;101563.

25. Minhas F ul AA, Asif A, Arif M. CAFÉ-Map: Context Aware Feature Mapping for mining high dimensional biomedical data. Comput Biol Med. 2016;79:68–79.

26. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support Vector Machines and Kernels for Computational Biology. PLOS Comput Biol. 2008;4(10):1–10.

27. Chang Y-W, Lin C-J. Feature Ranking Using Linear SVM. In: Guyon I, Aliferis C, Cooper G, Elisseeff A, Pellet J-P, Spirtes P, et al., editors. Hong Kong: PMLR; 2008. p. 53–64. (Proceedings of Machine Learning Research; vol. 3). Available from: http://proceedings.mlr.press/v3/chang08a.html

28. Smyrk TC, Watson P, Kaul K, Lynch HT. Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma. Cancer. 2001;91(12):2417–22.

29. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing Tumor-infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method from the International Immunooncology Biomarkers Working Group: Part 1: Assessing the Host Immune Response, TILs in Invasive Breast Carcinoma and Ductal Carcinoma in Situ, Metastatic Tumor Deposits and Areas for Further Research. Adv Anat Pathol. 2017;24(5):235–51.

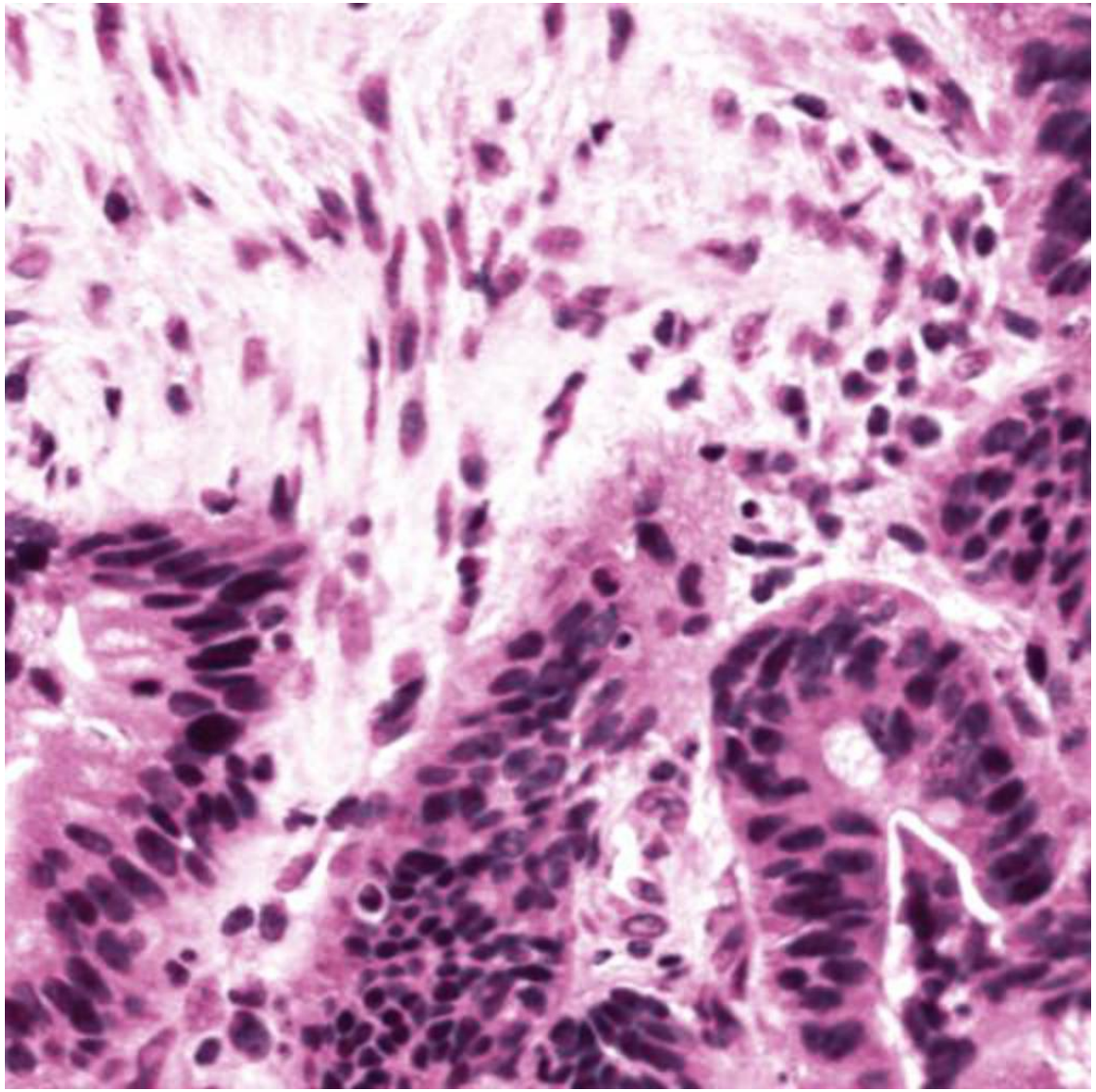
30. Shaban M, Khurram SA, Fraz MM, Alsubaie N, Masood I, Mushtaq S, et al. A Novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes Predicts Disease Free Survival in Oral Squamous Cell Carcinoma. Sci Rep. 2019;9.

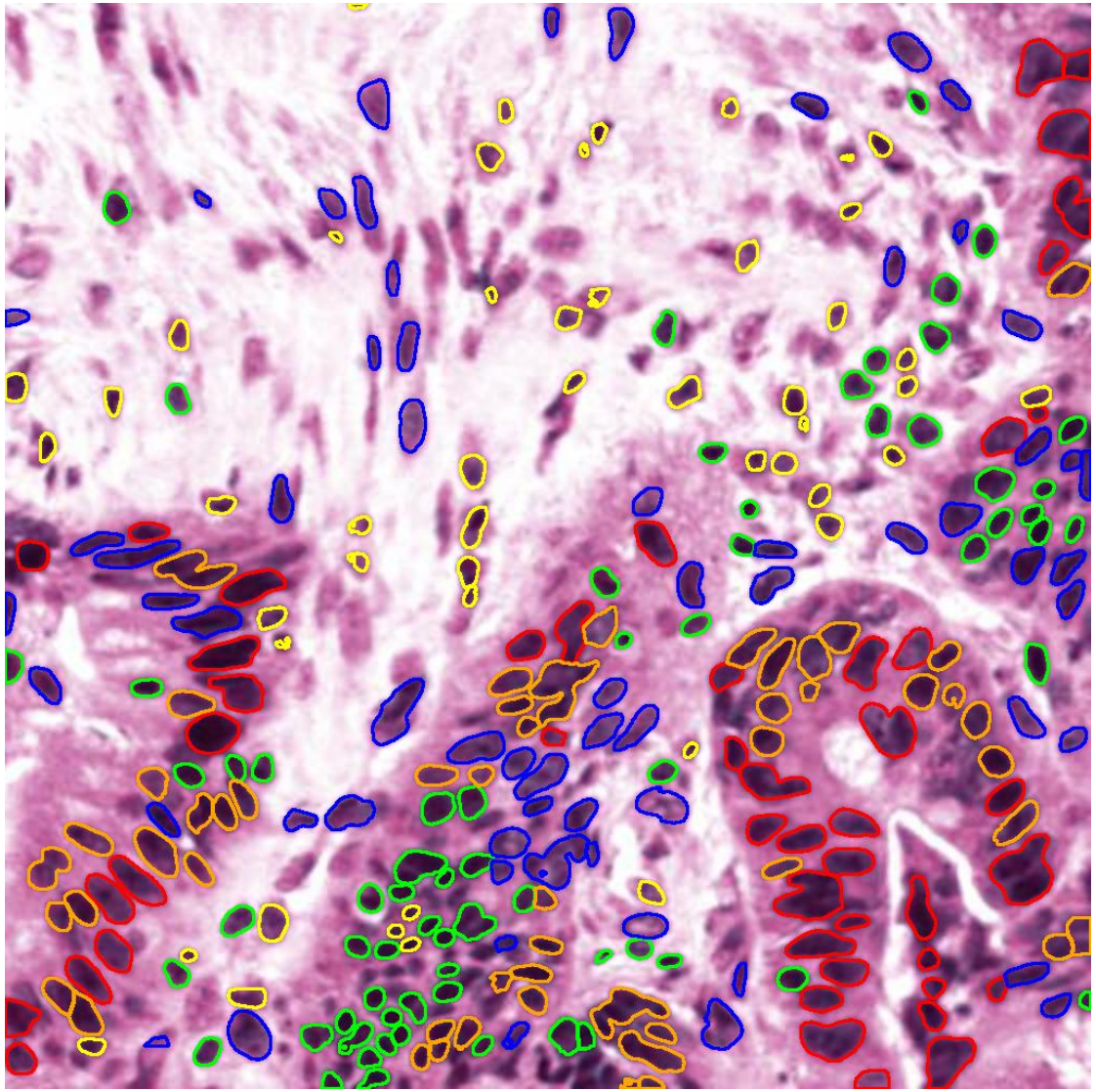
31. Saller J, Qin D, Felder S, Coppola D. Microsatellite Stable Colorectal Cancer With an Immunogenic Phenotype: Challenges in Diagnosis and Treatment. Clin Colorectal Cancer. 2020 Jun;19(2):123–31.

32. Goel A, Arnold CN, Niedzwiecki D, Chang DK, Ricciardiello L, Carethers JM, et al. Characterization of Sporadic Colon Cancer by Patterns of Genomic Instability. Cancer Res. 2003;63(7):1608–14.

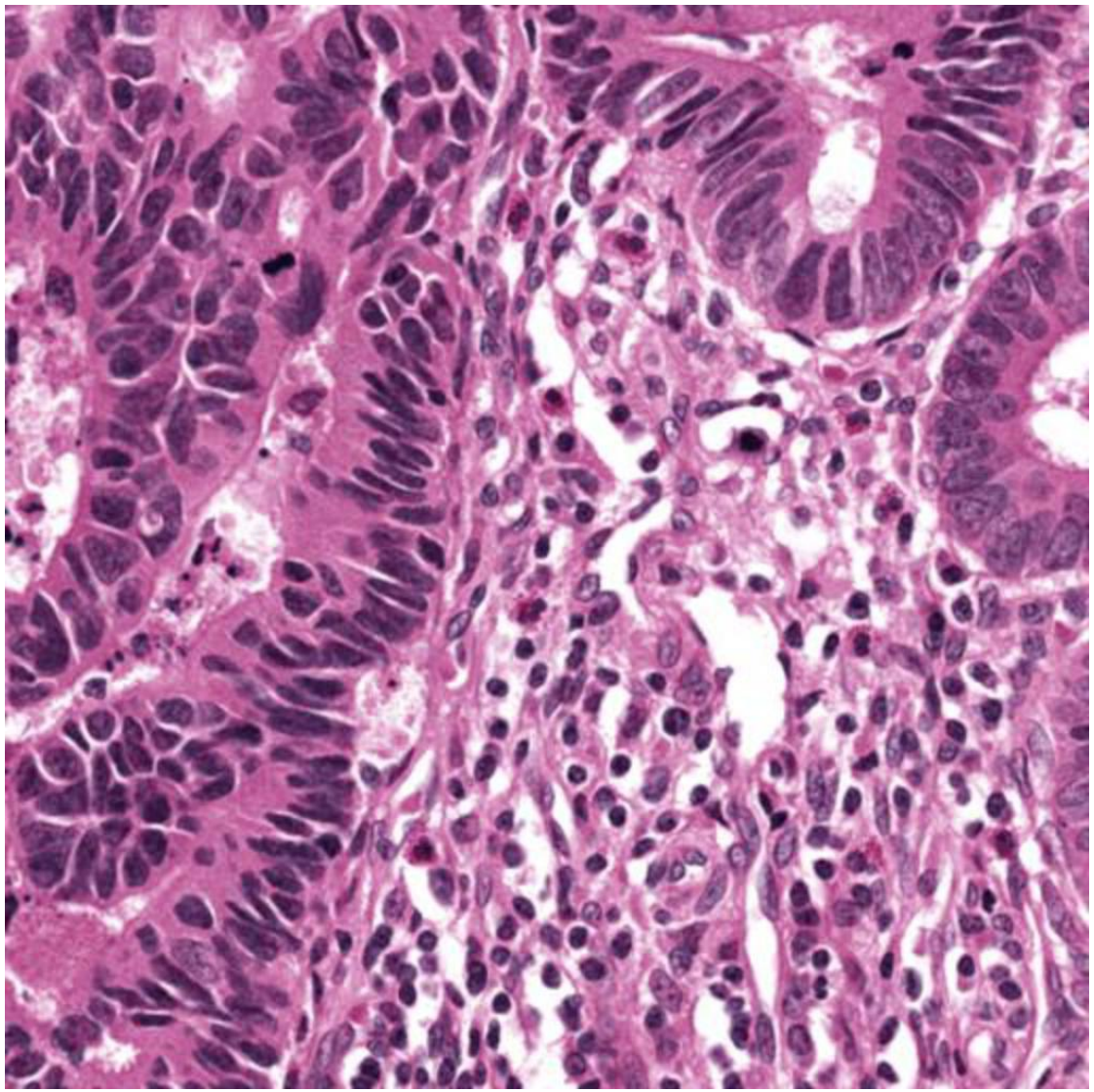
CIN cellular composition

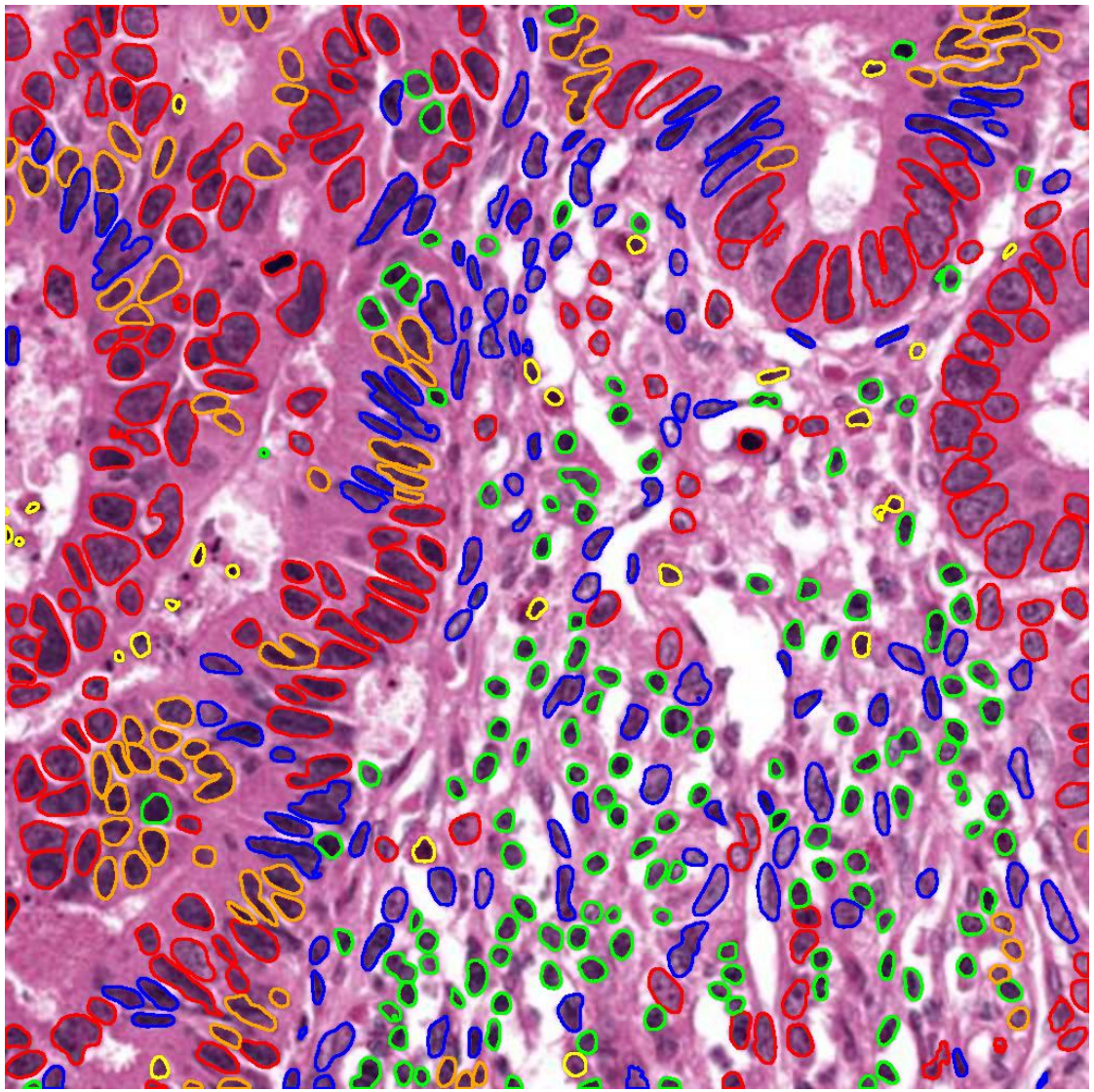
Supplementary Material



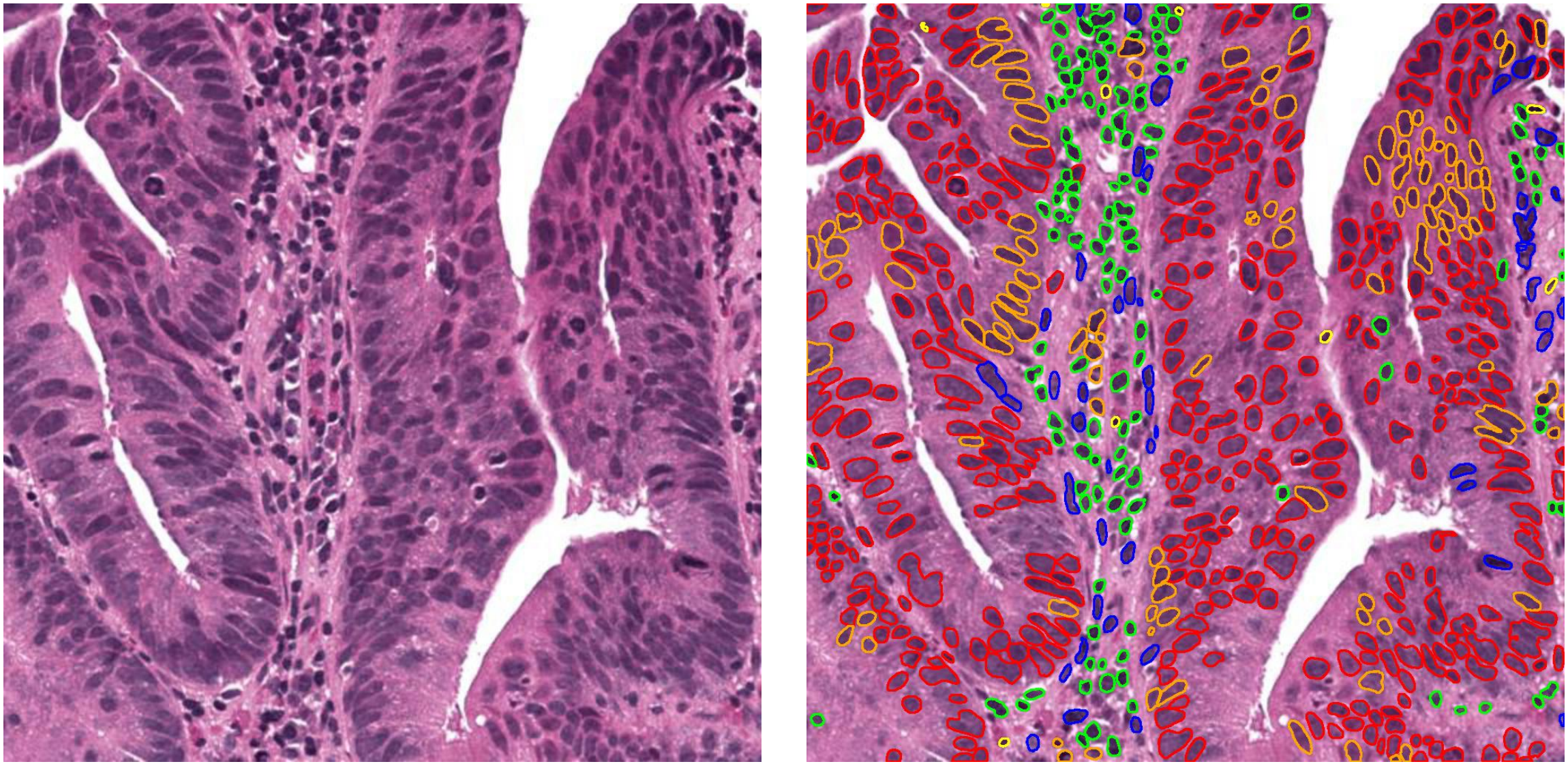


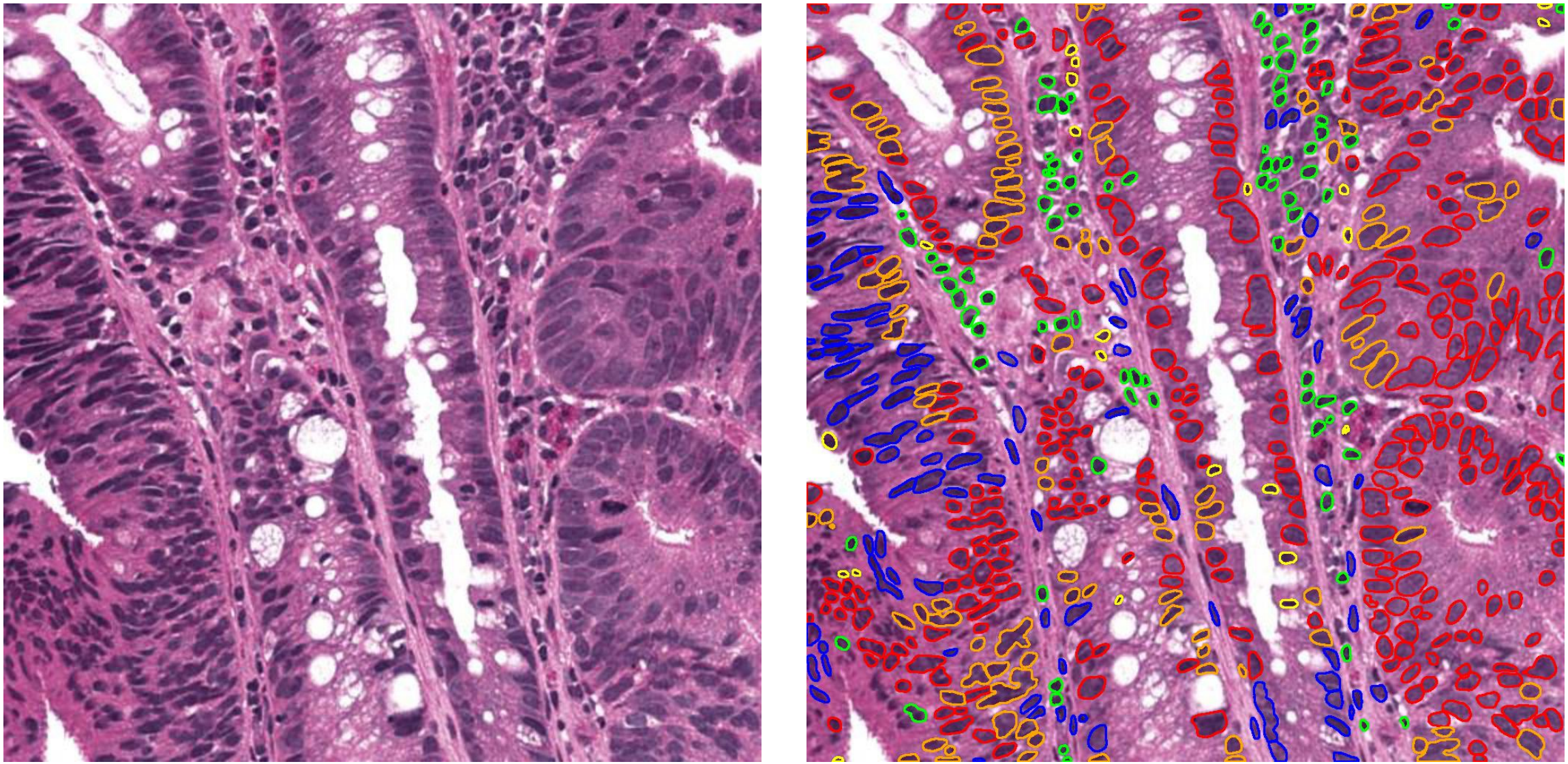
NEP1 Inflammatory Dead Mesenchymal NEP2

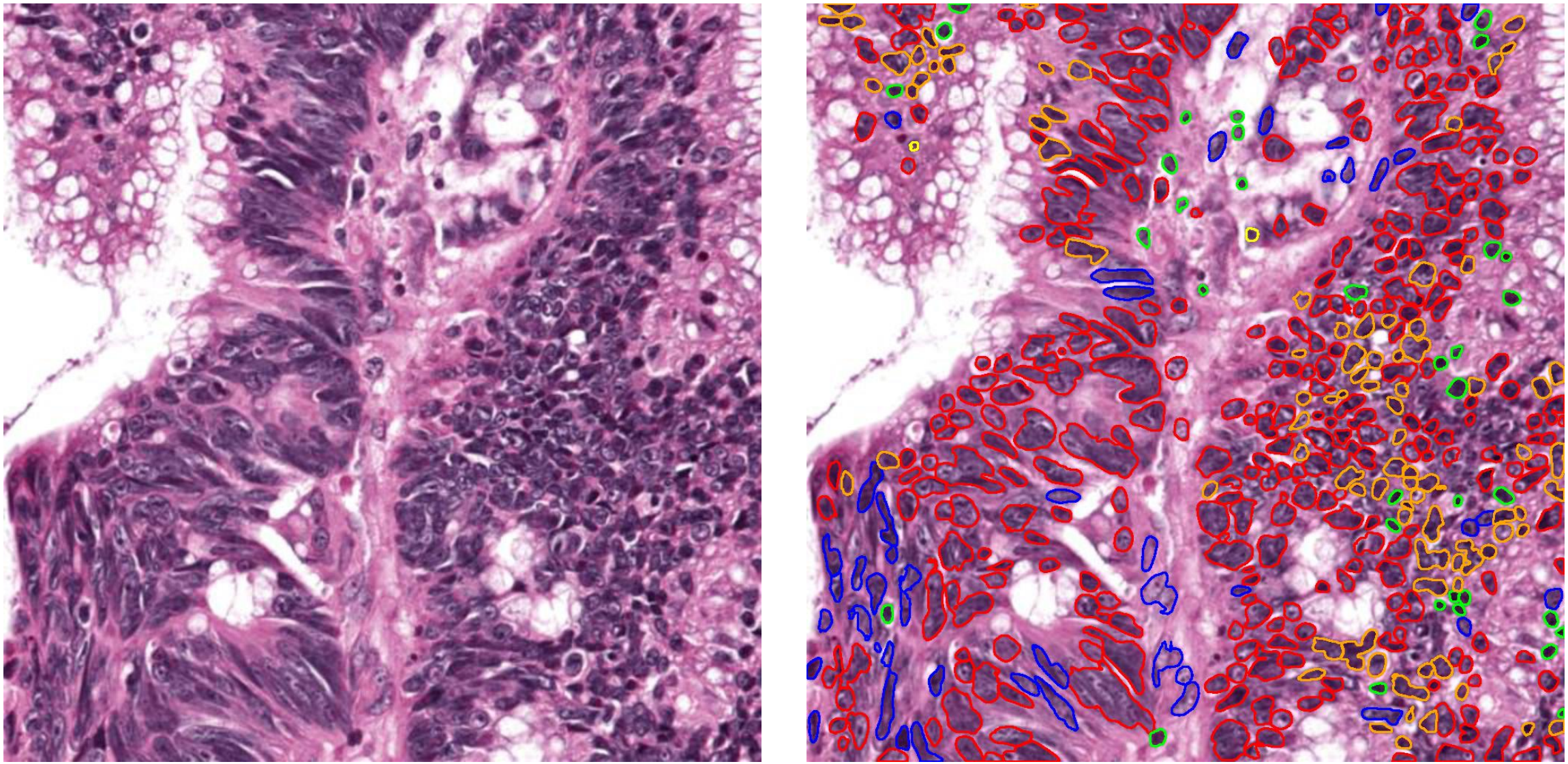


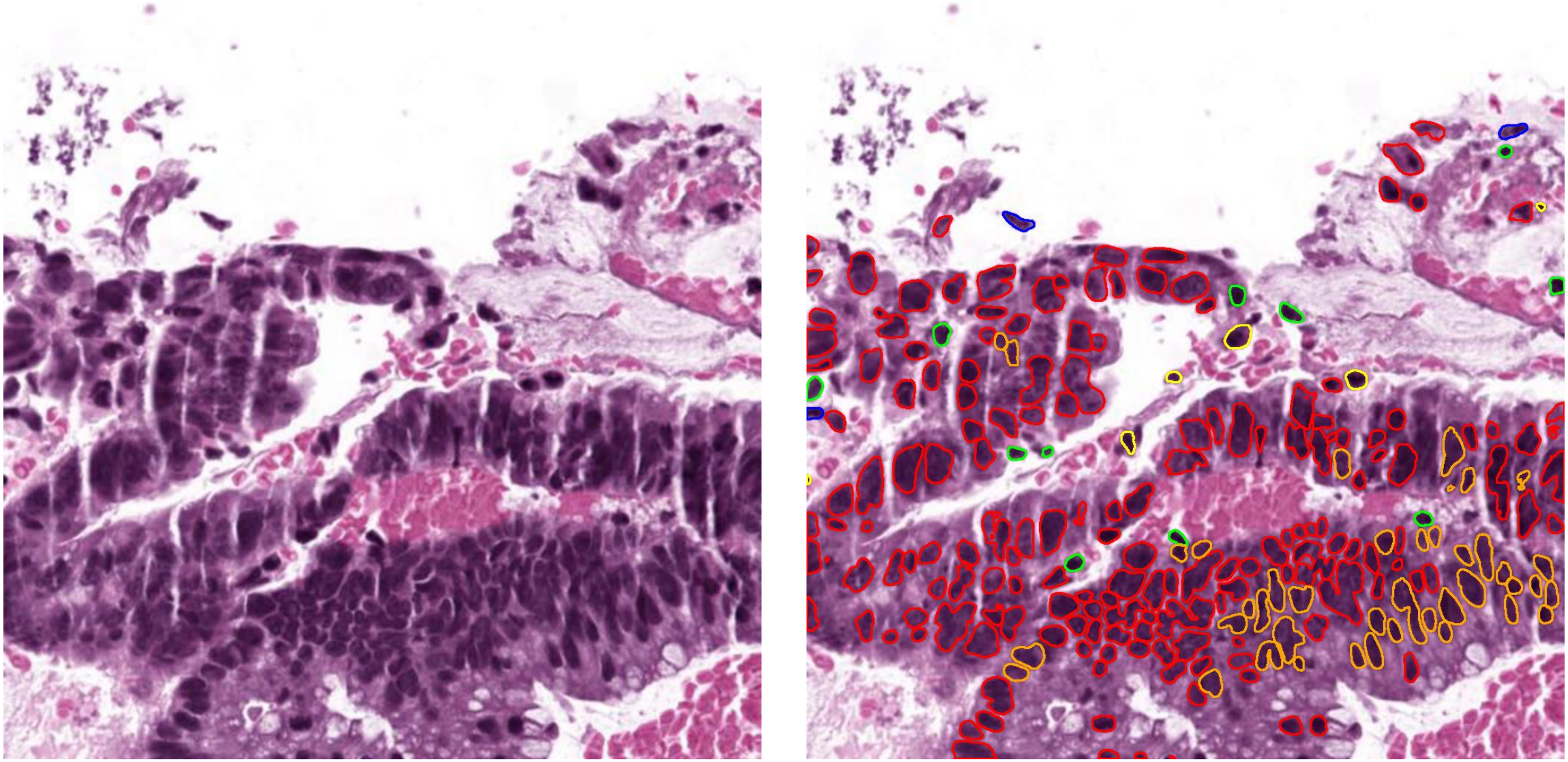


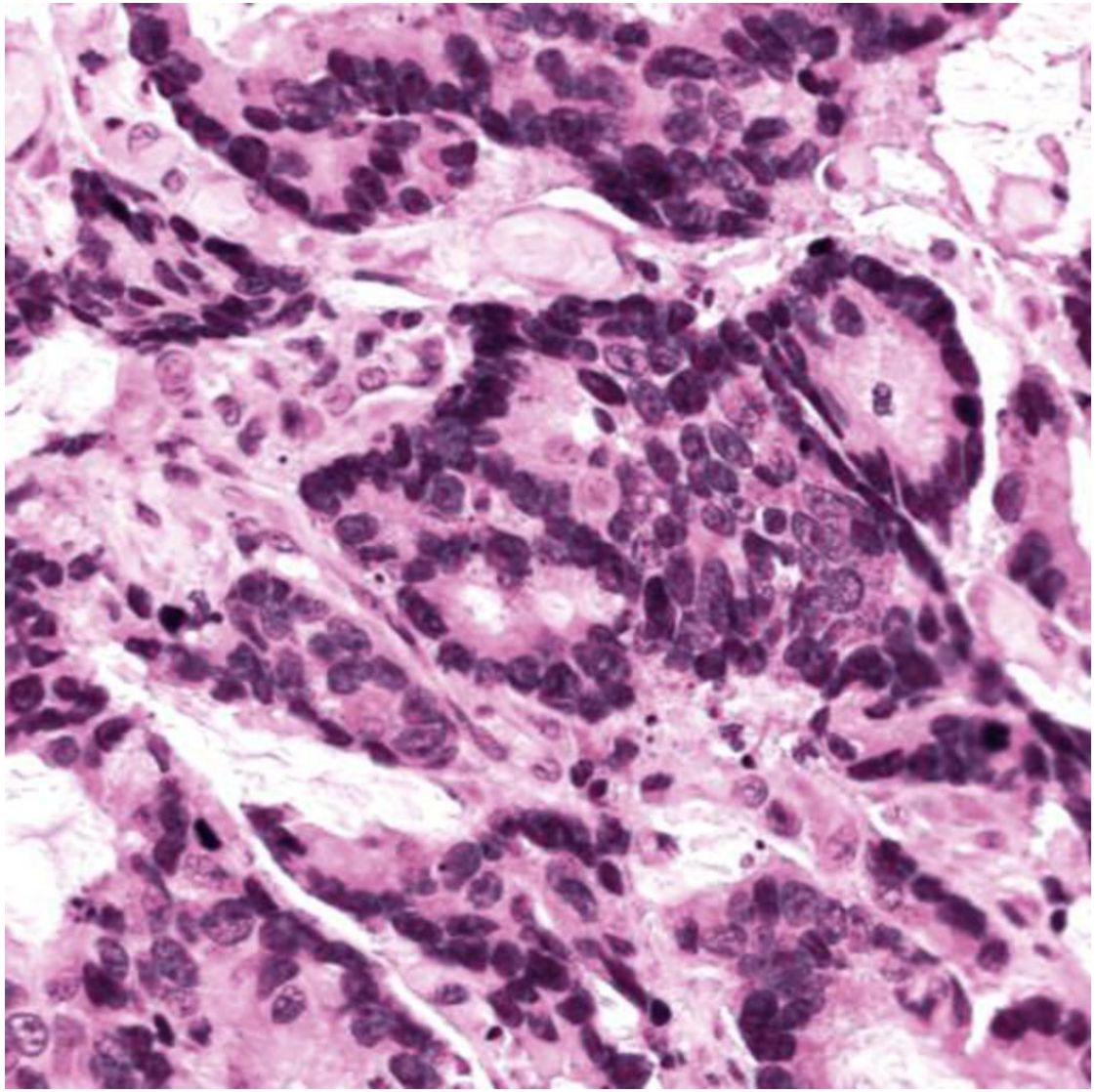
NEP1 Inflammatory Dead Mesenchymal NEP2

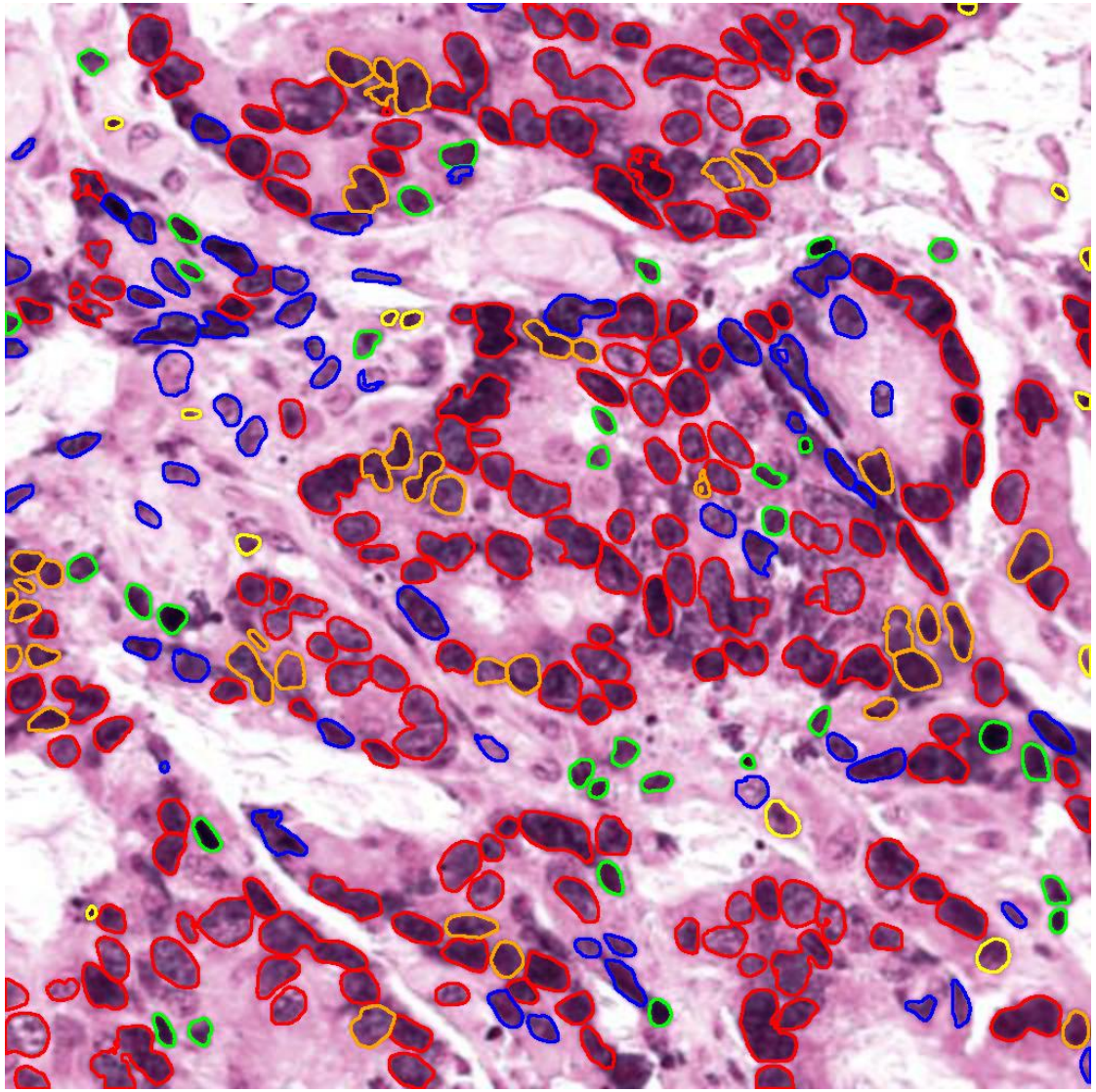

NEP1
Inflammatory
Dead
Mesenchymal
NEP2


NEP1
Inflammatory
Dead
Mesenchymal
NEP2

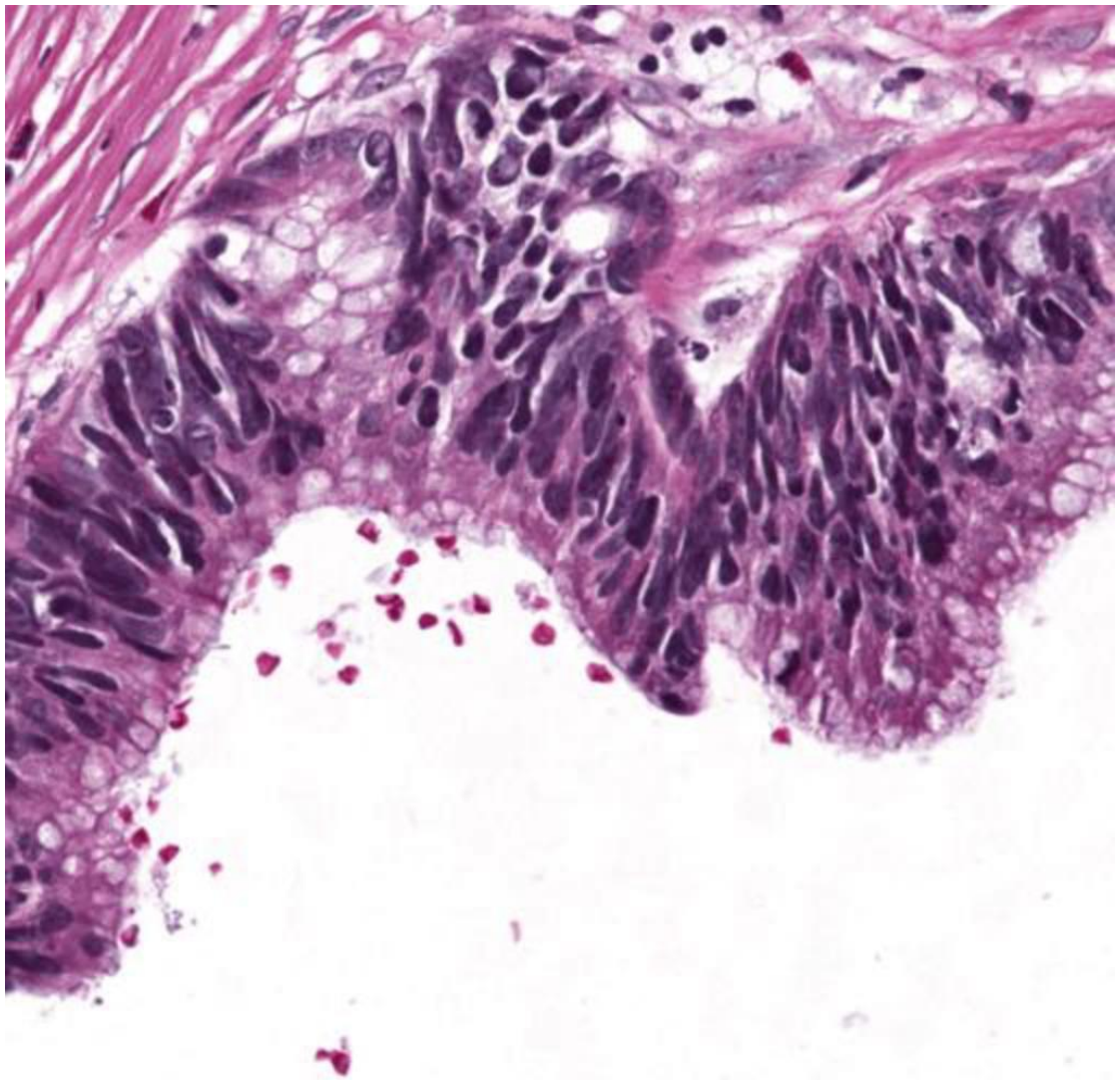

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

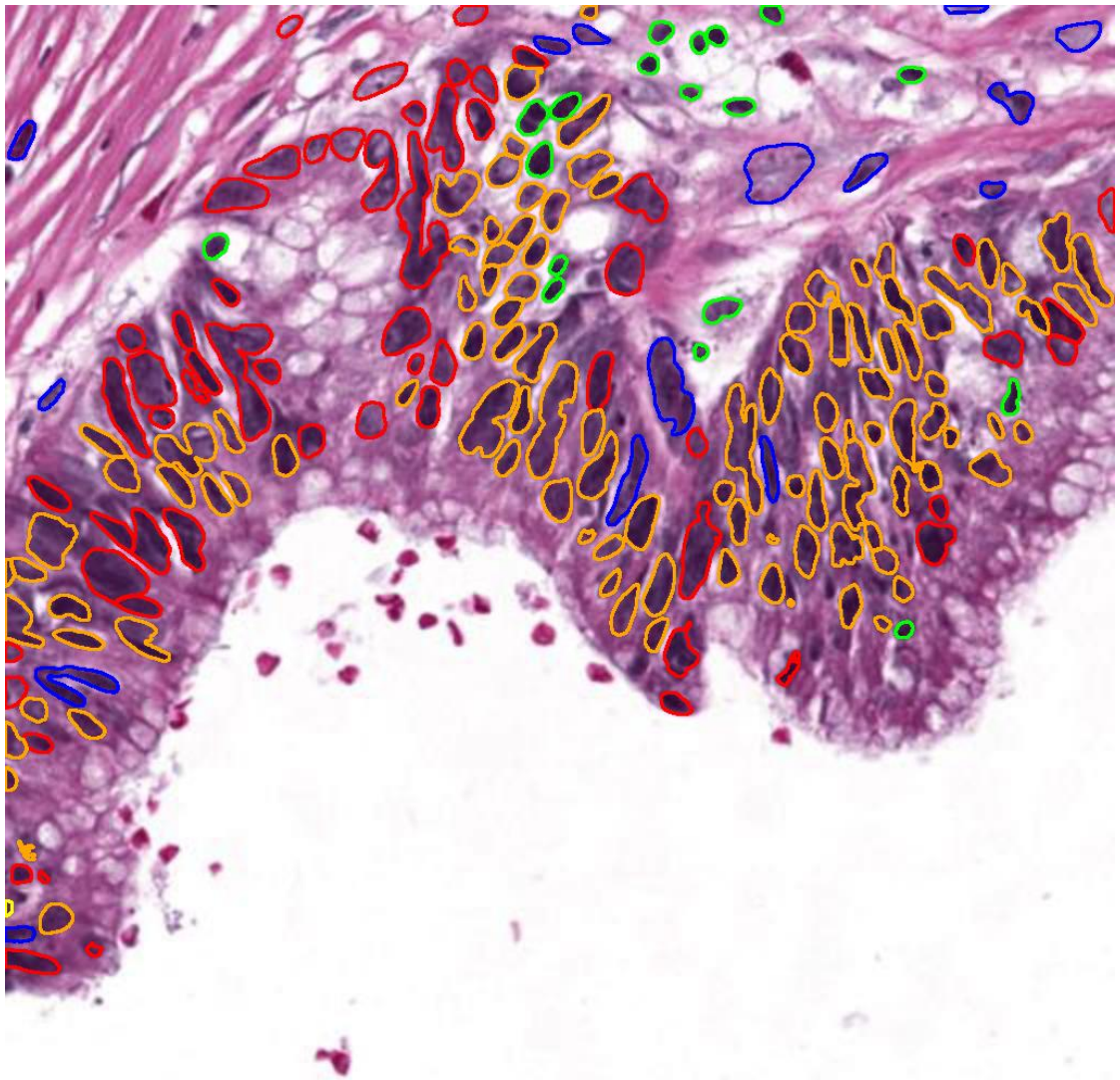

NEP1
Inflammatory
Dead
Mesenchymal
NEP2



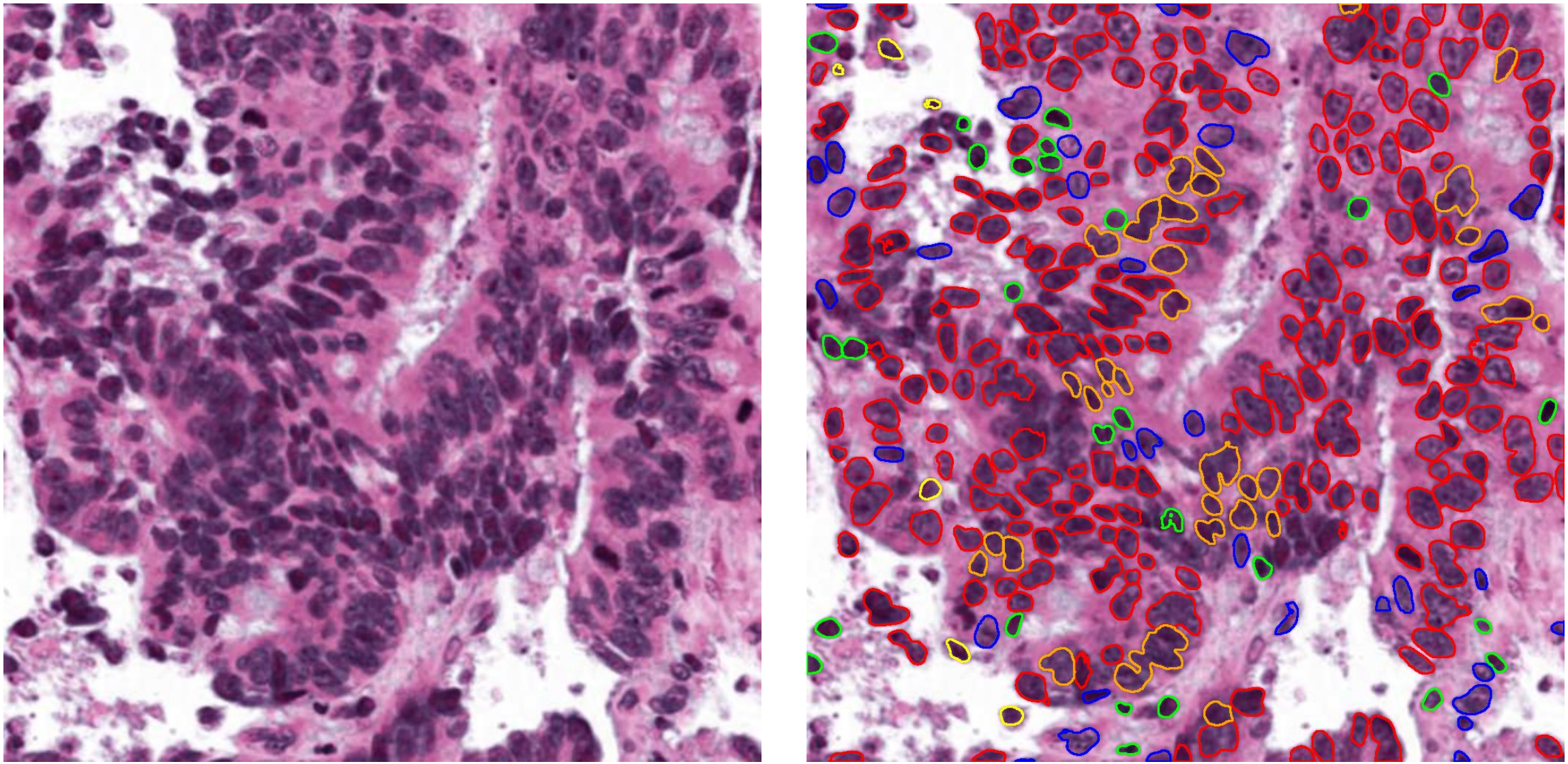


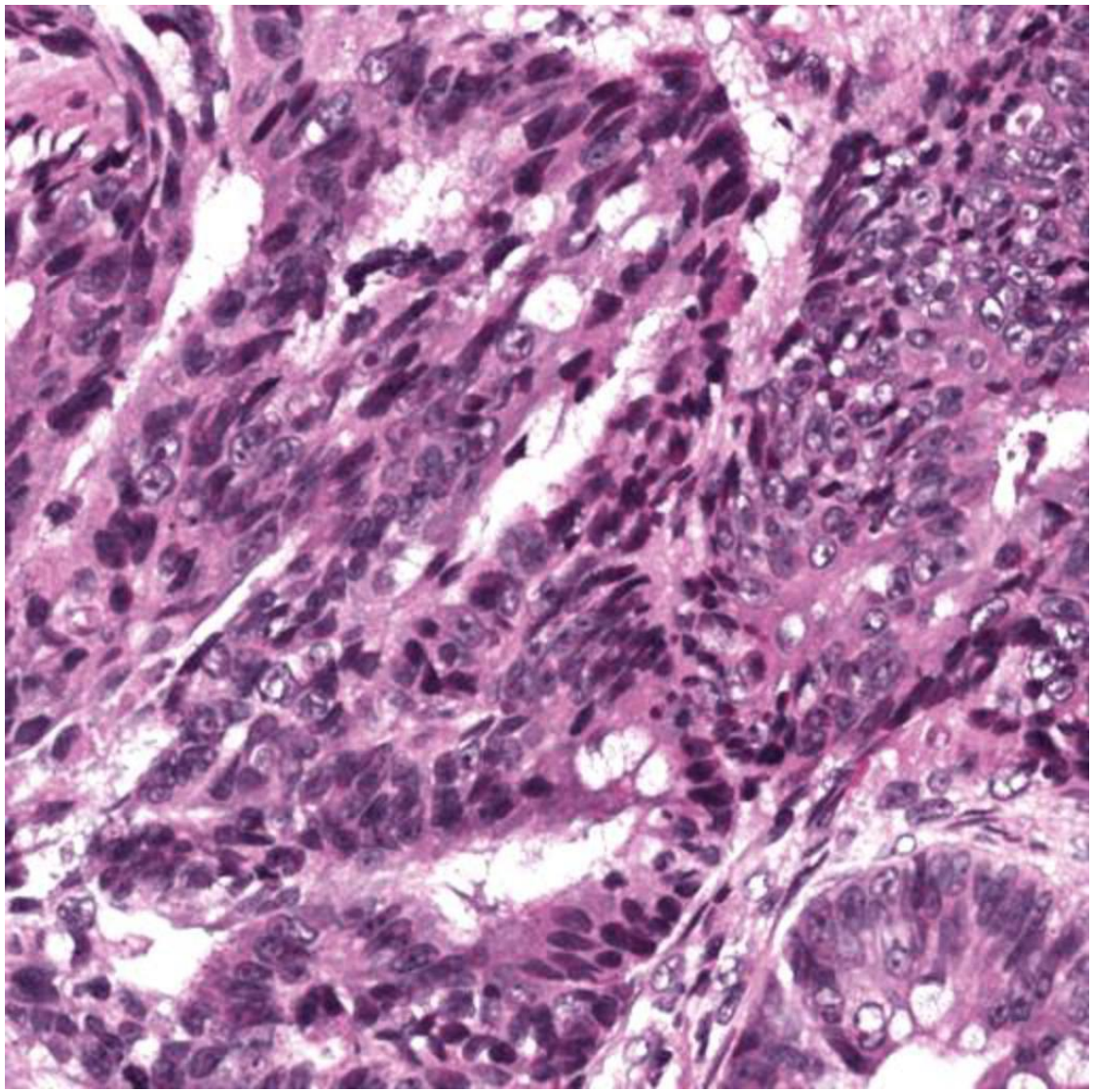
NEP1 Inflammatory Dead Mesenchymal NEP2

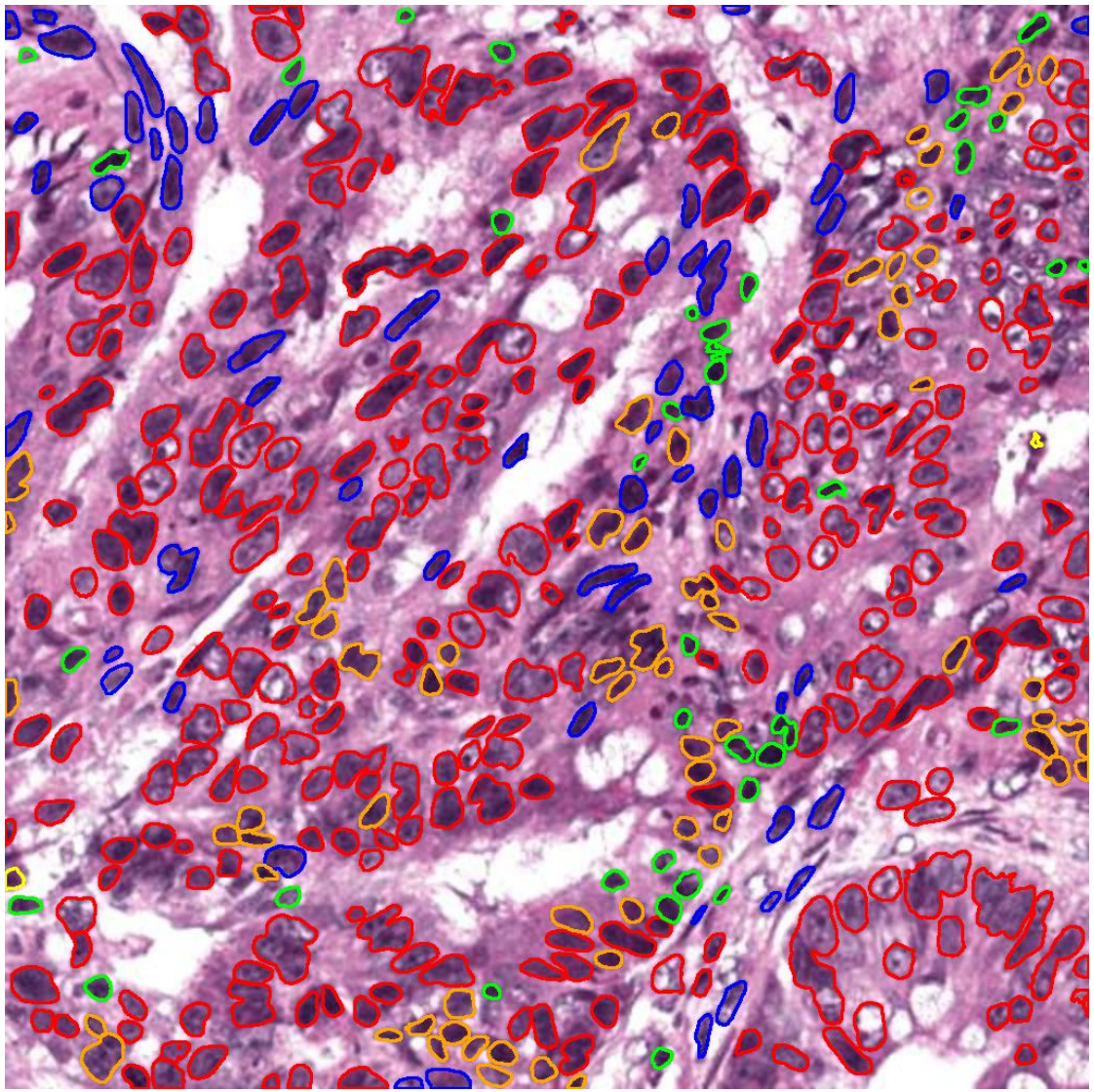




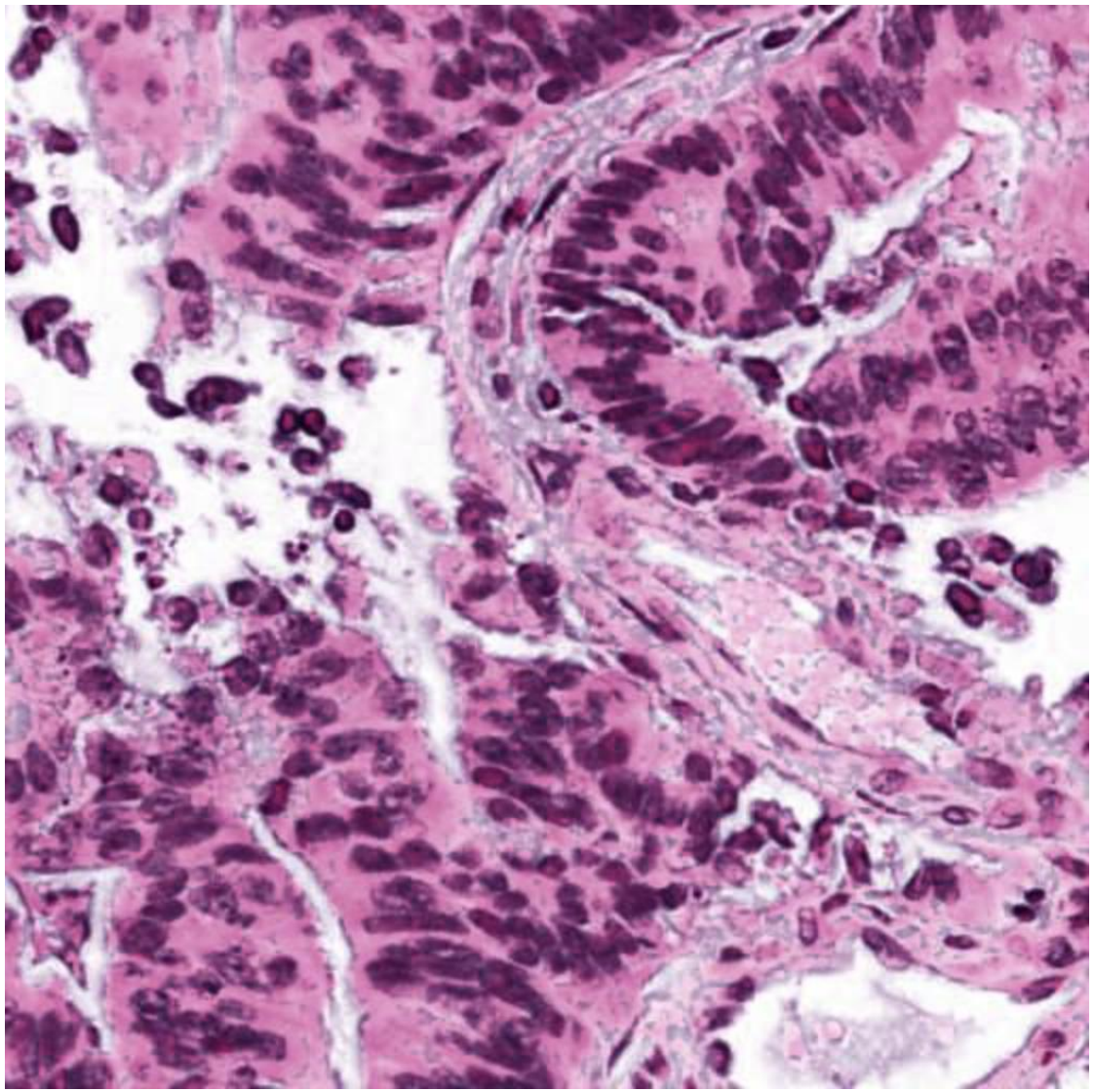
NEP1 Inflammatory Dead Mesenchymal NEP2

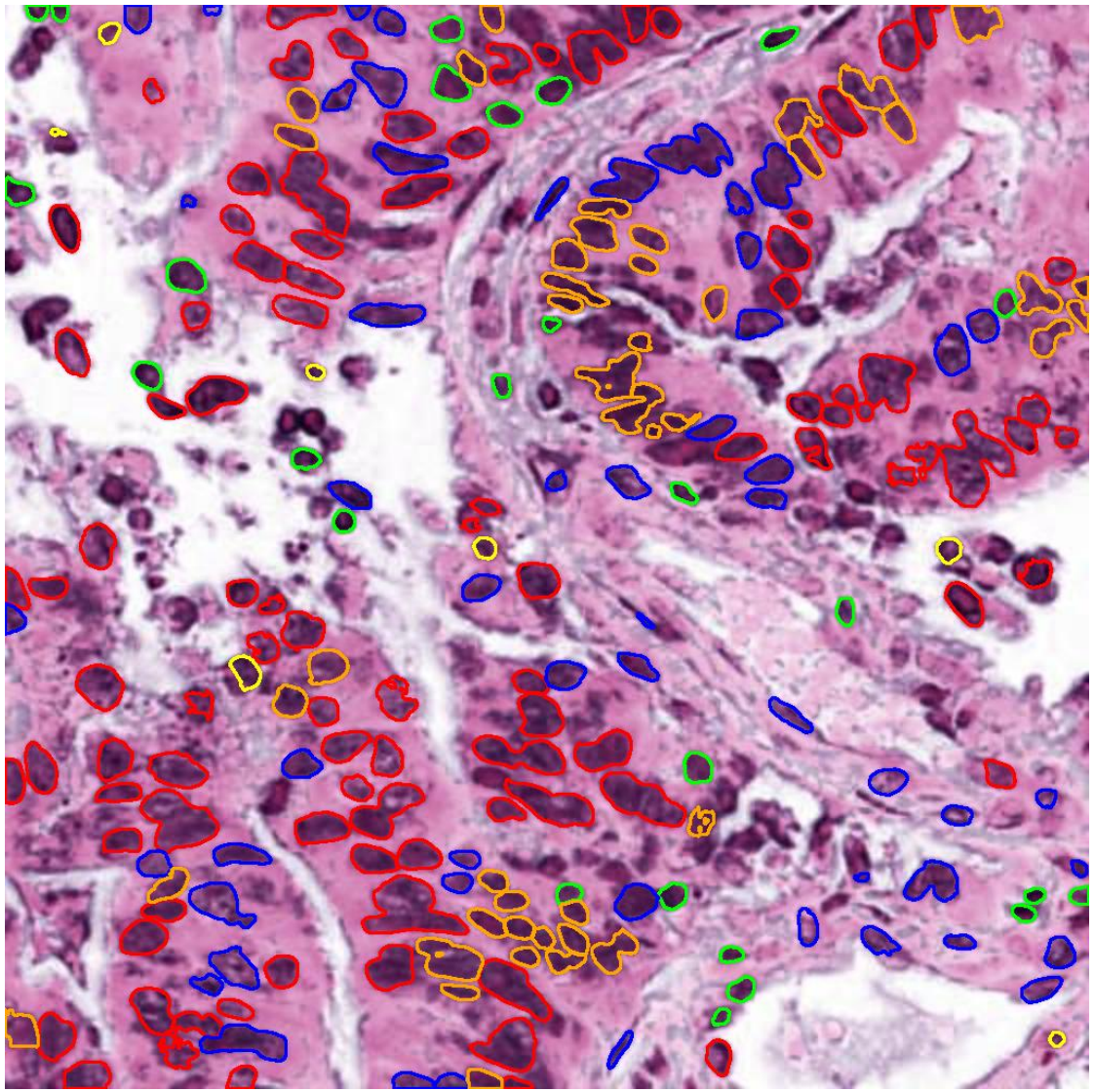

NEP1
Inflammatory
Dead
Mesenchymal
NEP2



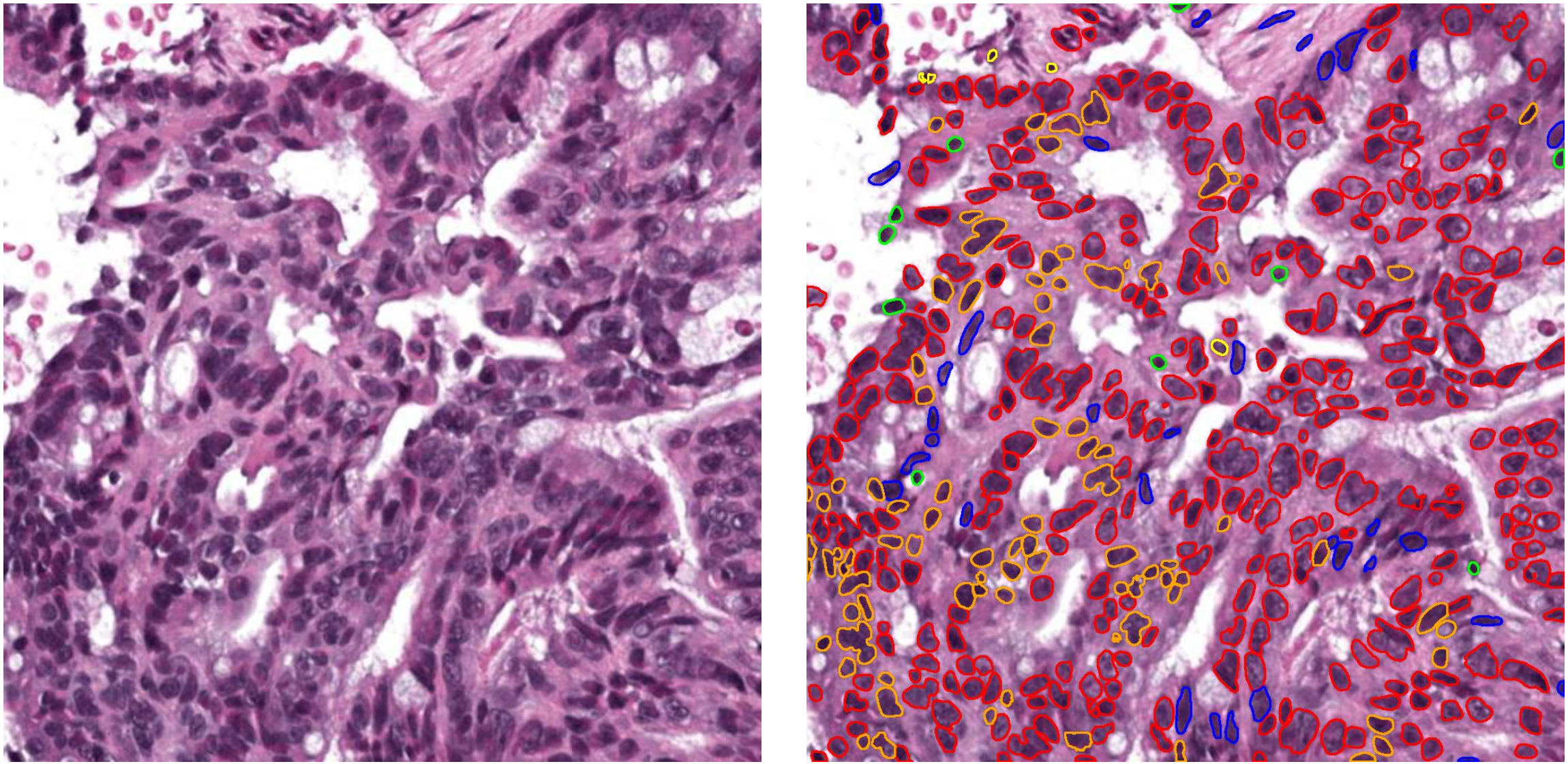


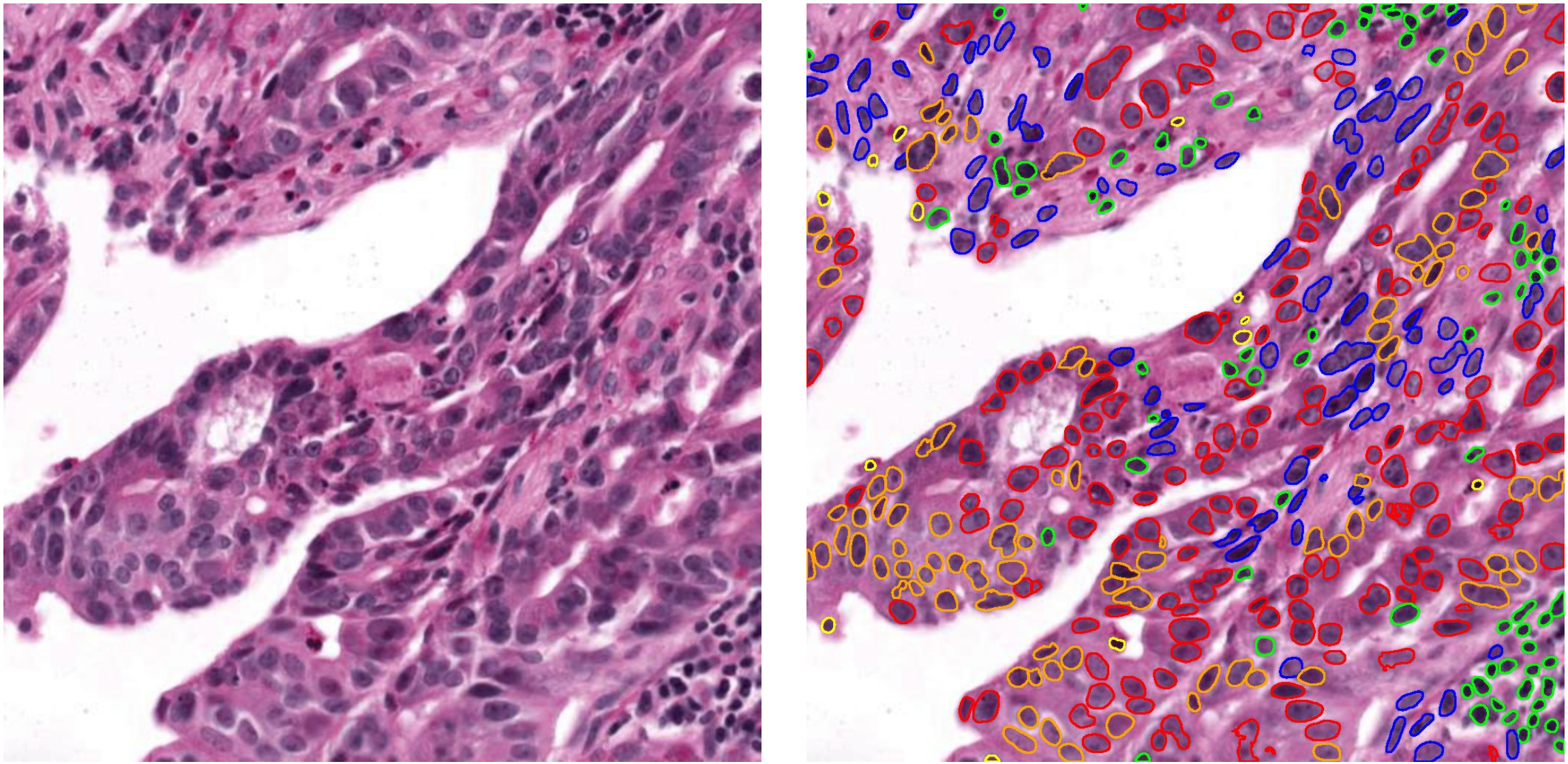
NEP1 Inflammatory Dead Mesenchymal NEP2

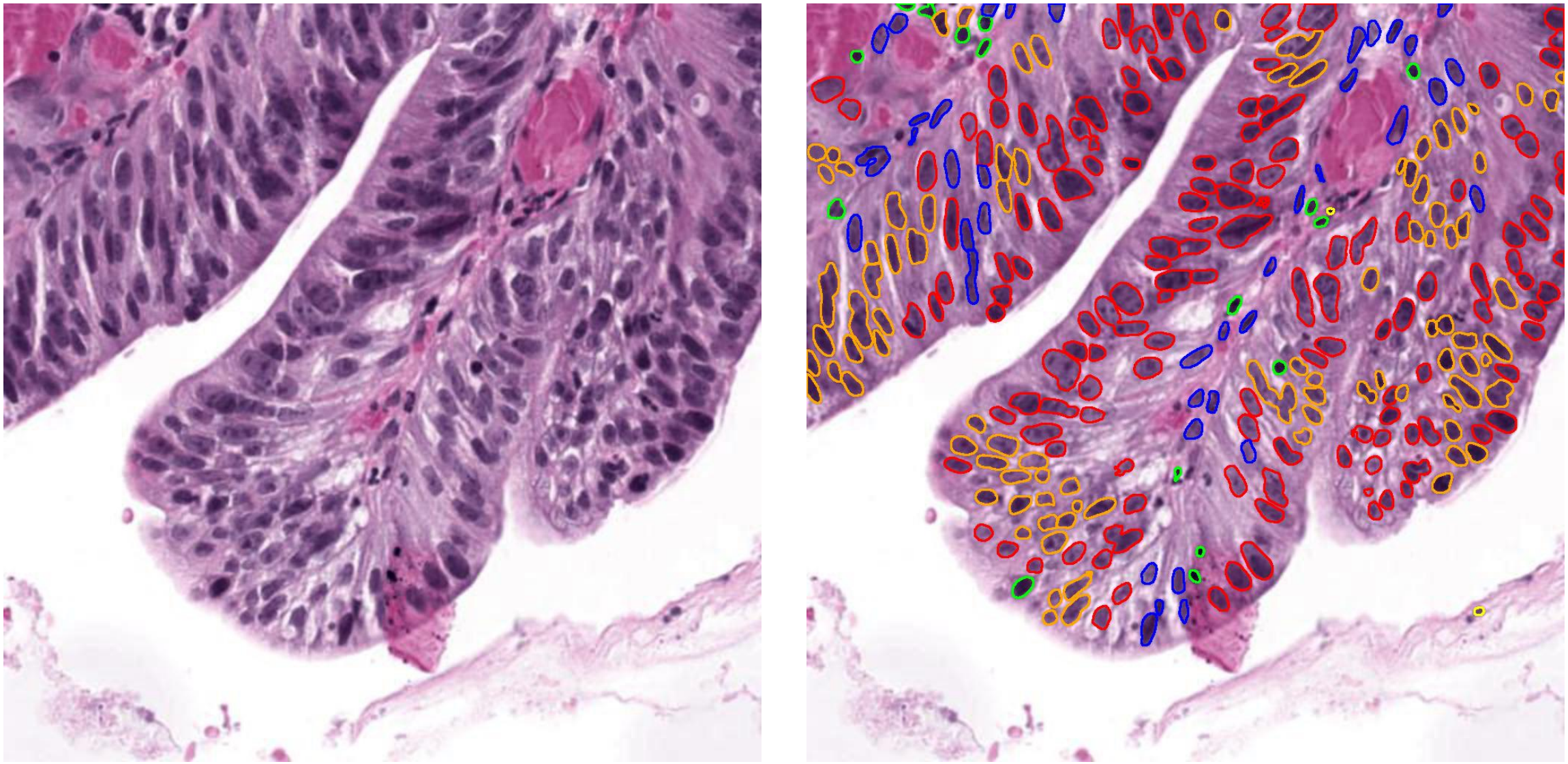


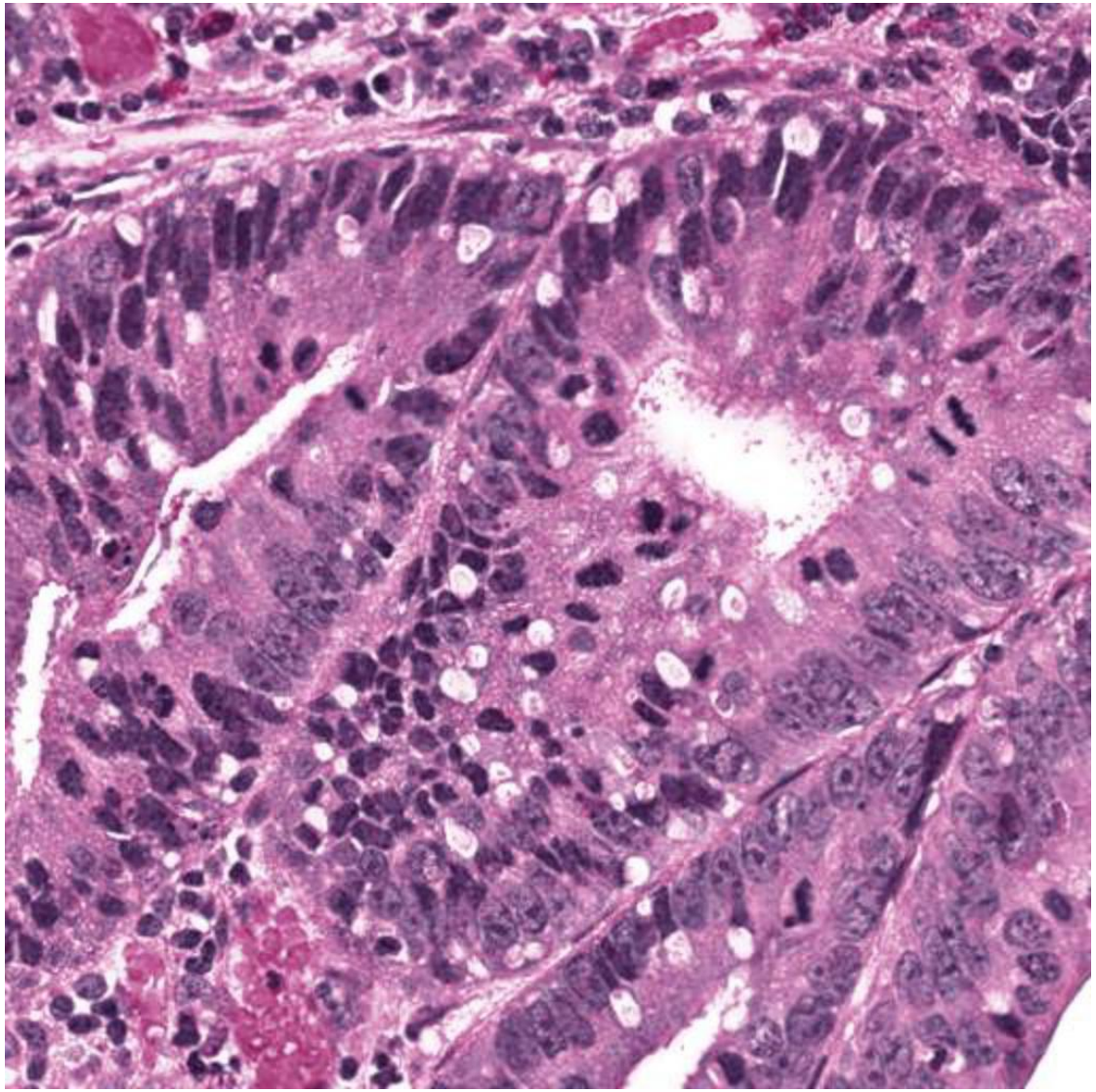


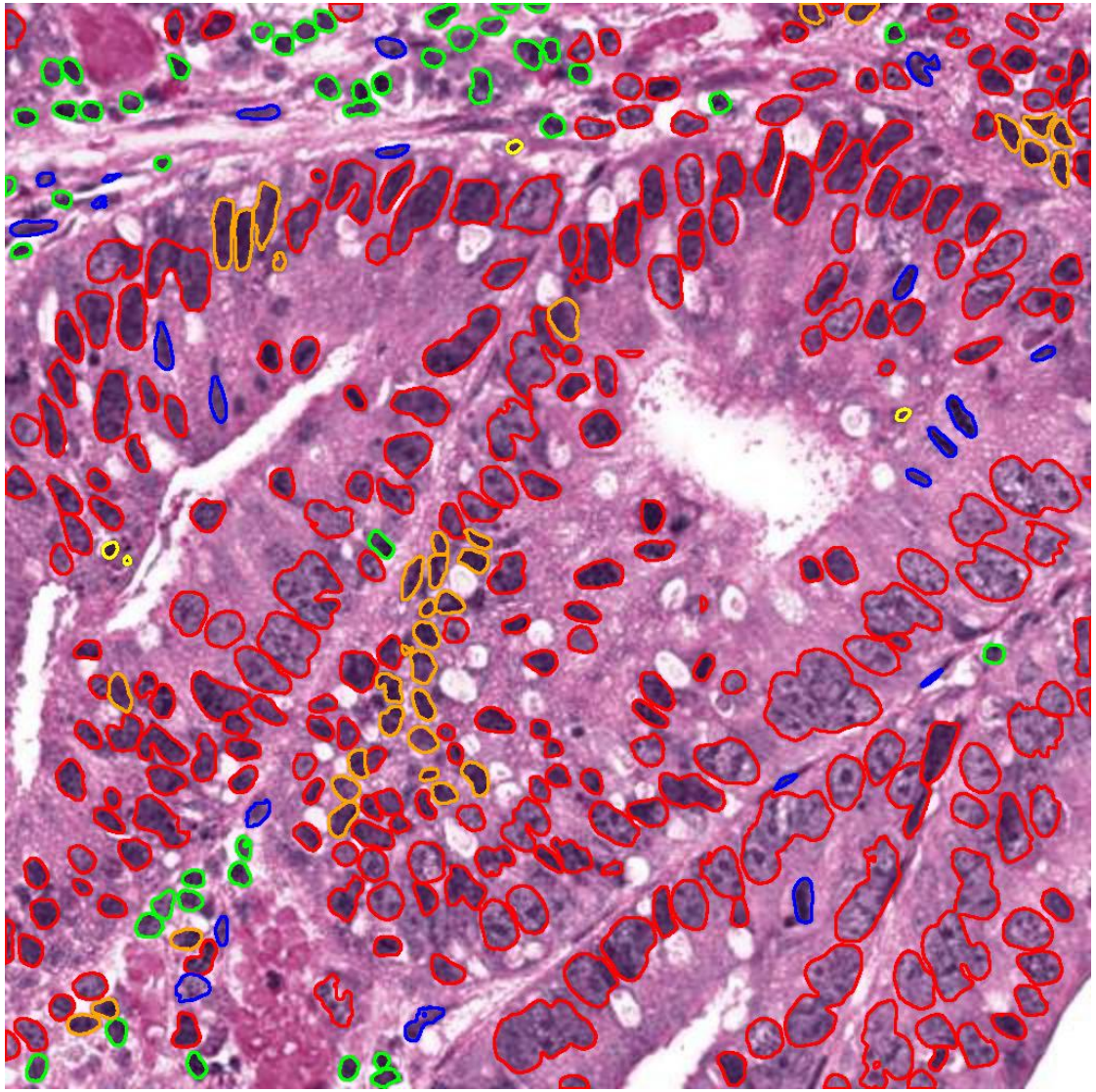
NEP1 Inflammatory Dead Mesenchymal NEP2


NEP1
Inflammatory
Dead
Mesenchymal
NEP2

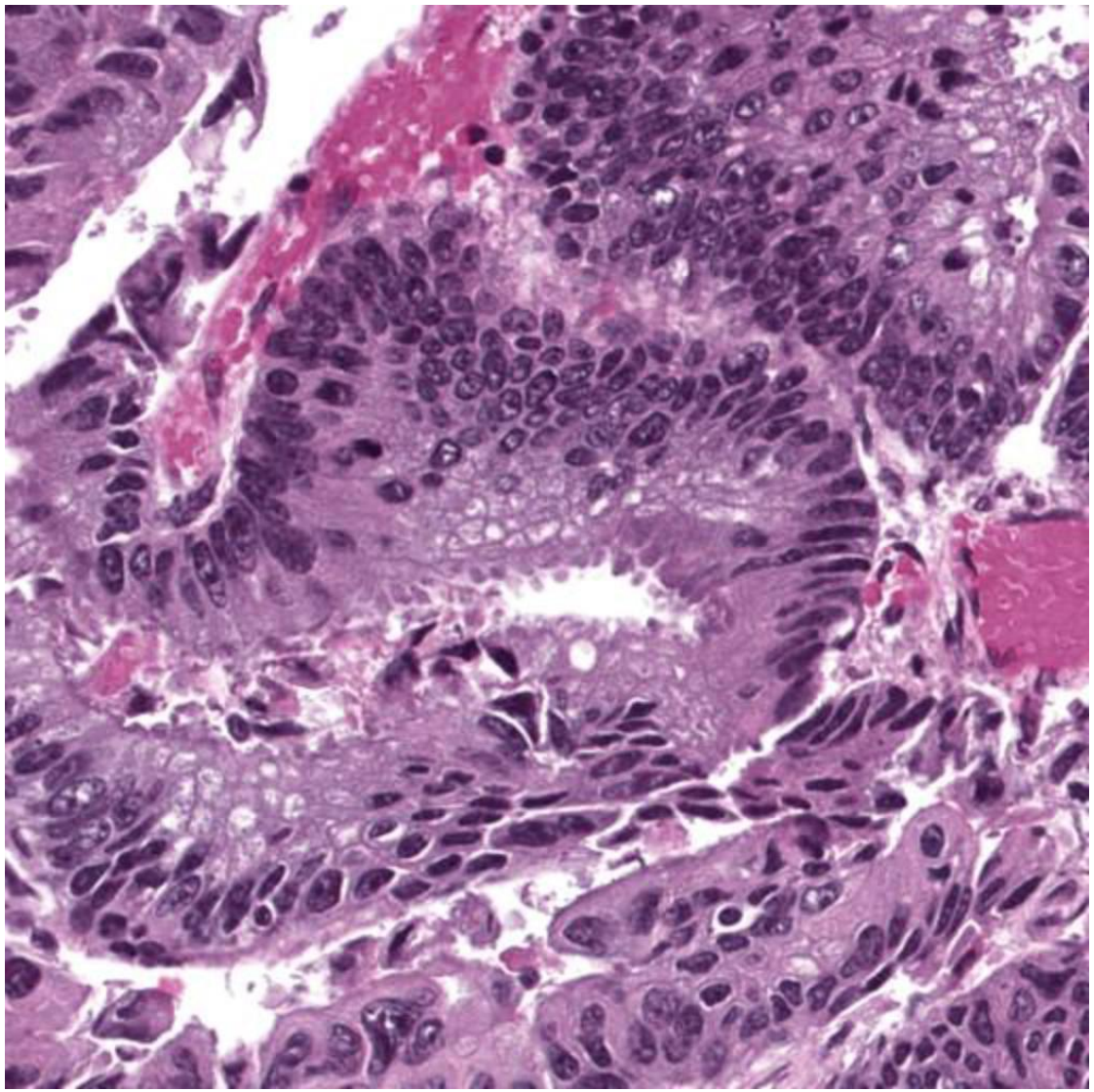

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

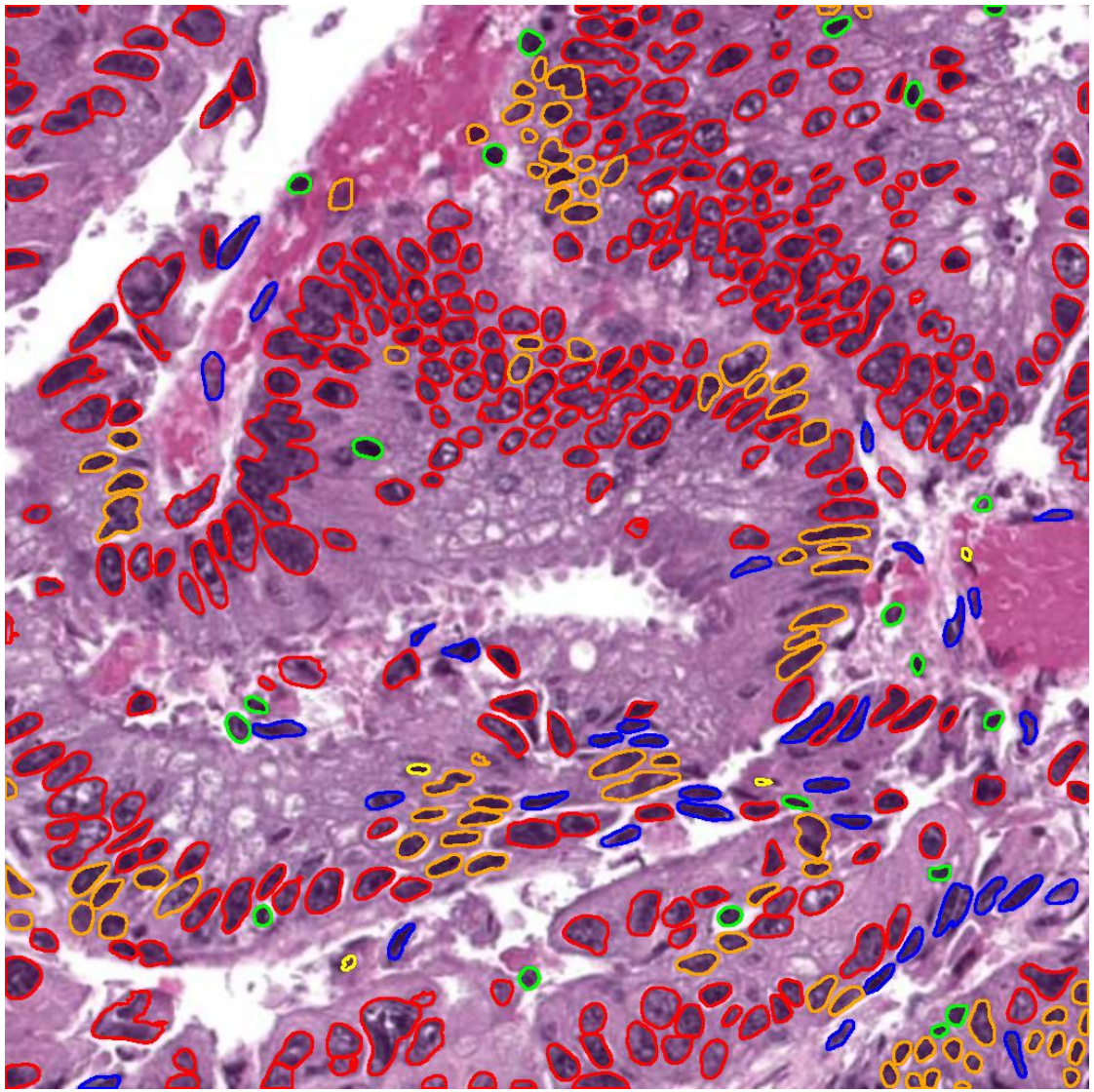

NEP1
Inflammatory
Dead
Mesenchymal
NEP2



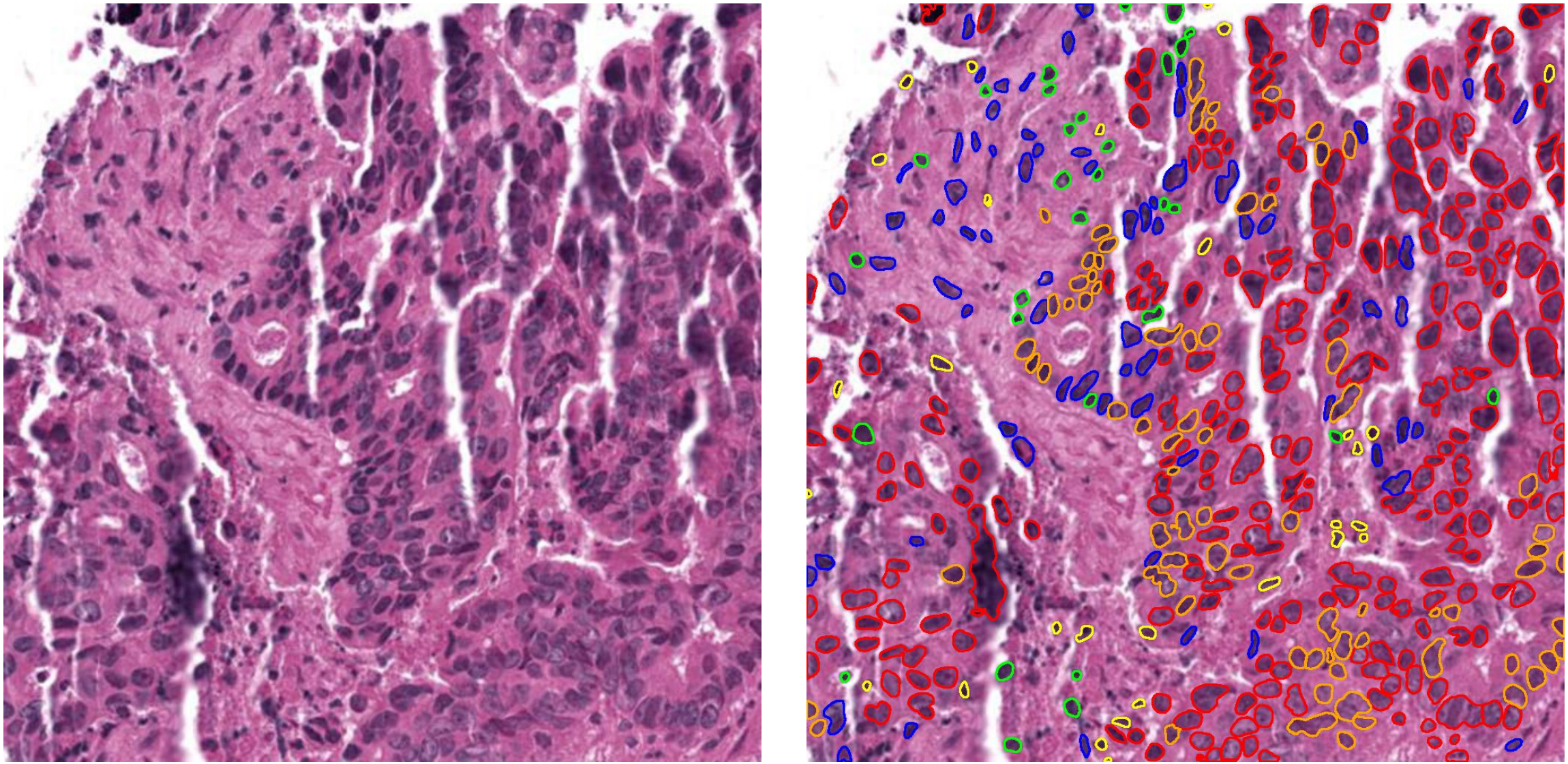


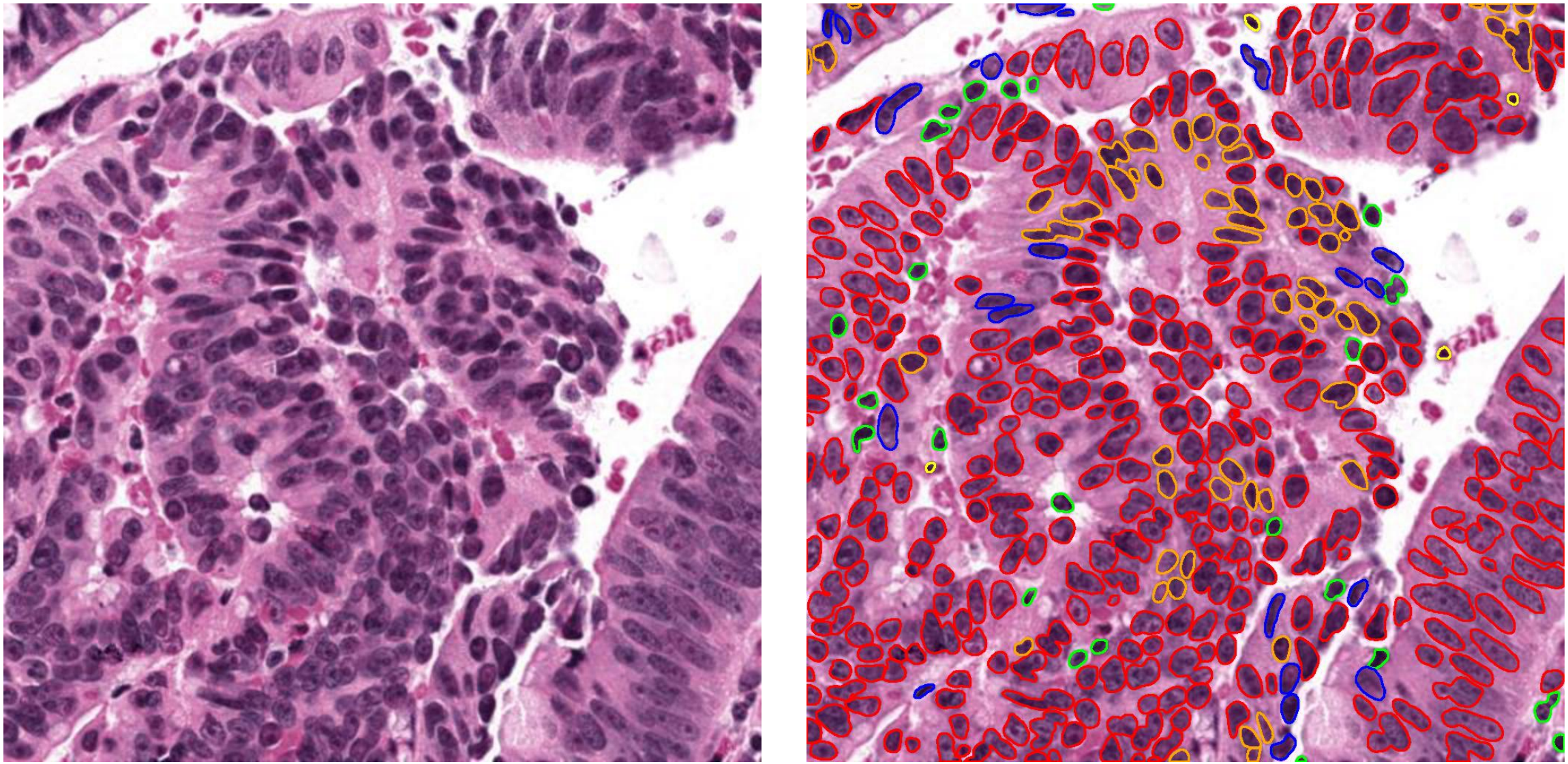
NEP1 Inflammatory Dead Mesenchymal NEP2

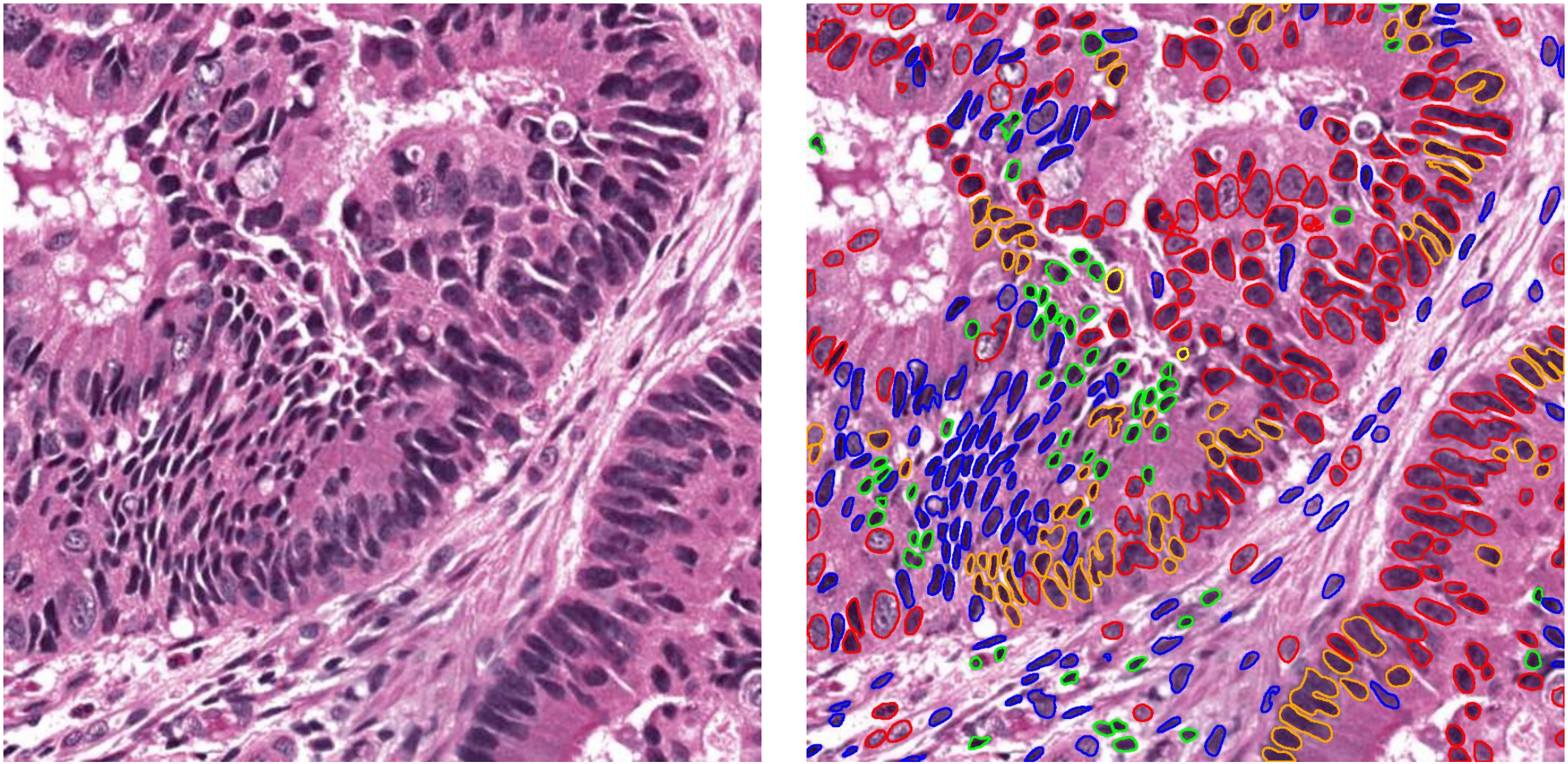


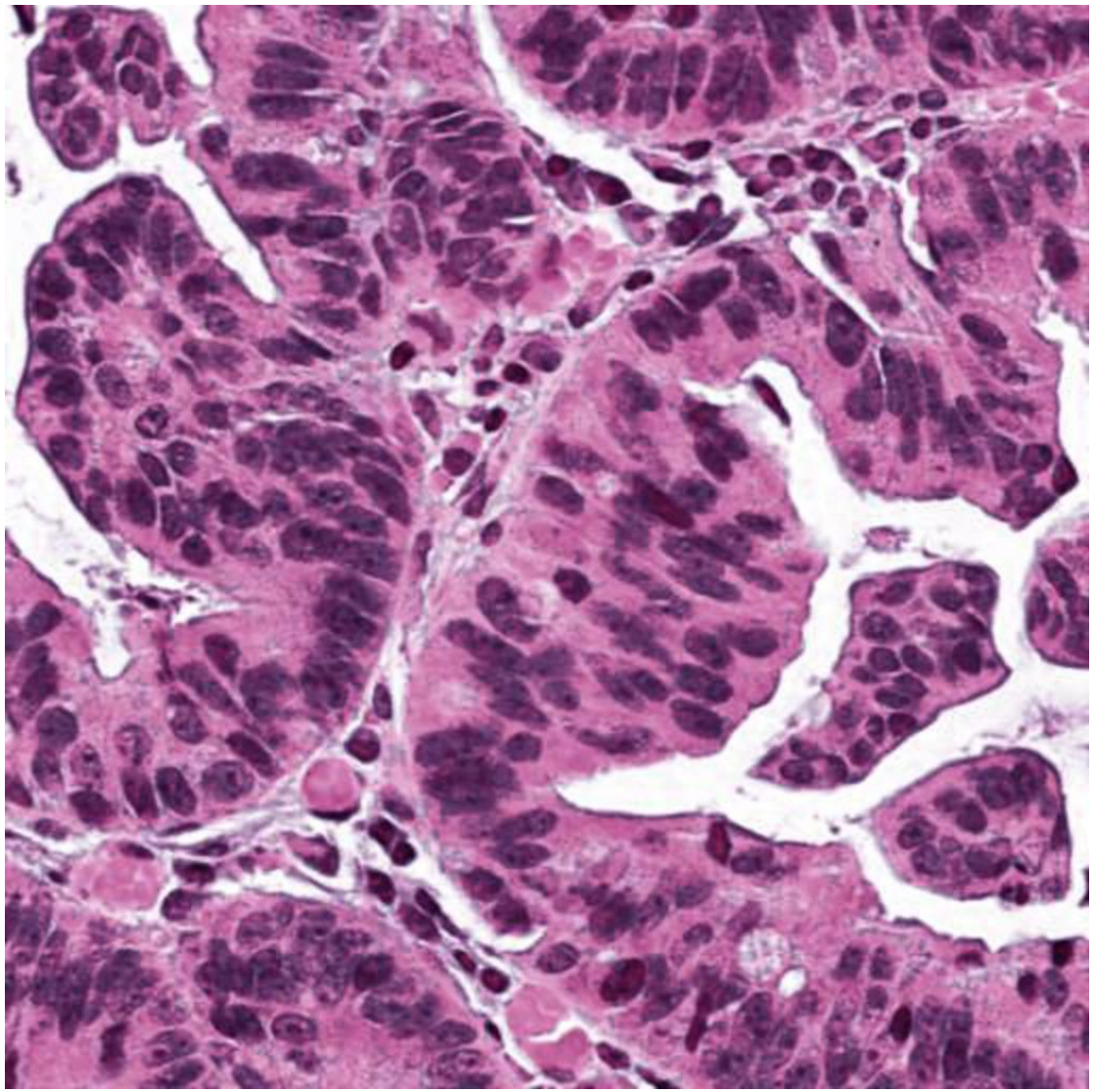


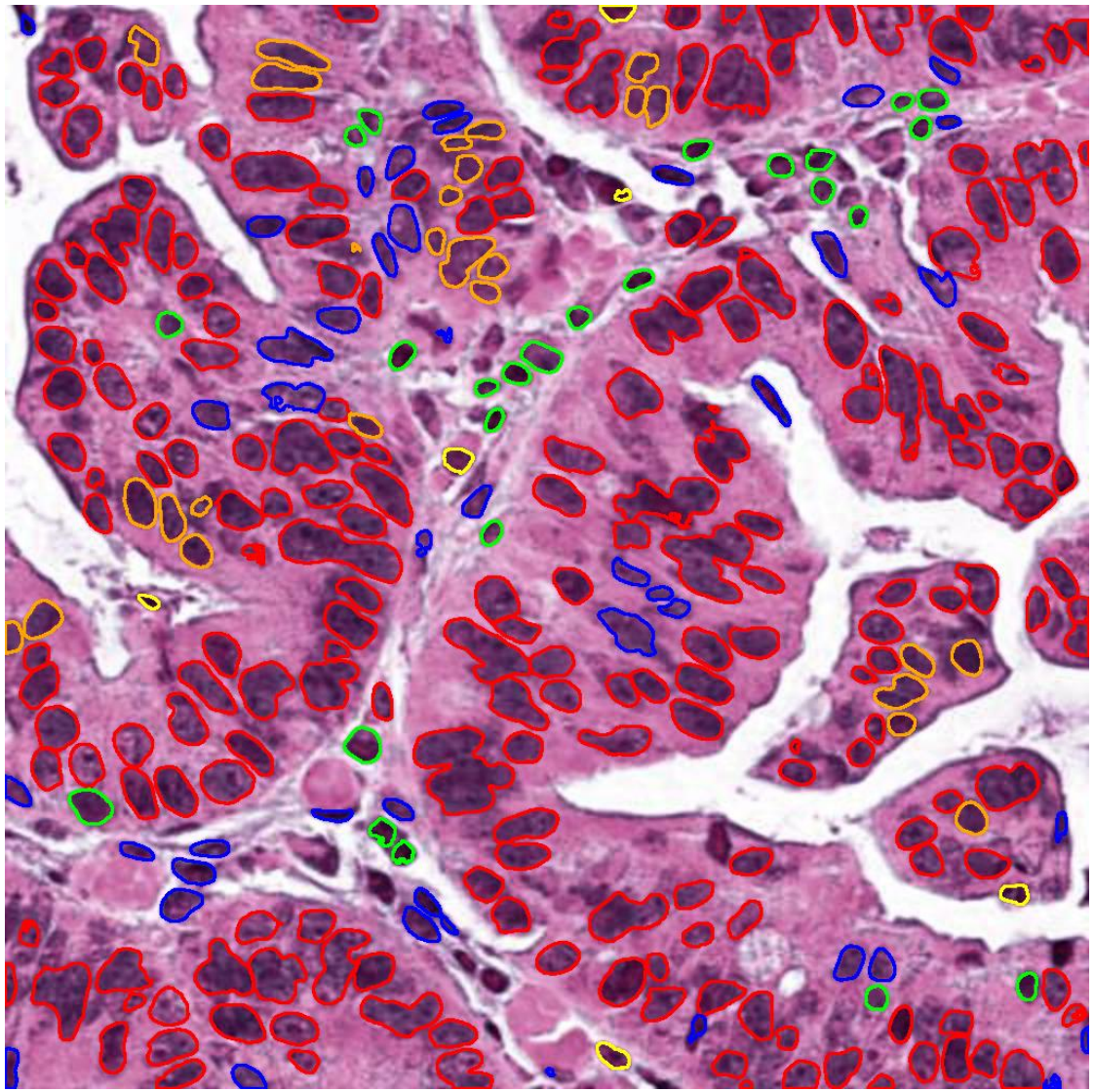
NEP1 Inflammatory Dead Mesenchymal NEP2


NEP1
Inflammatory
Dead
Mesenchymal
NEP2

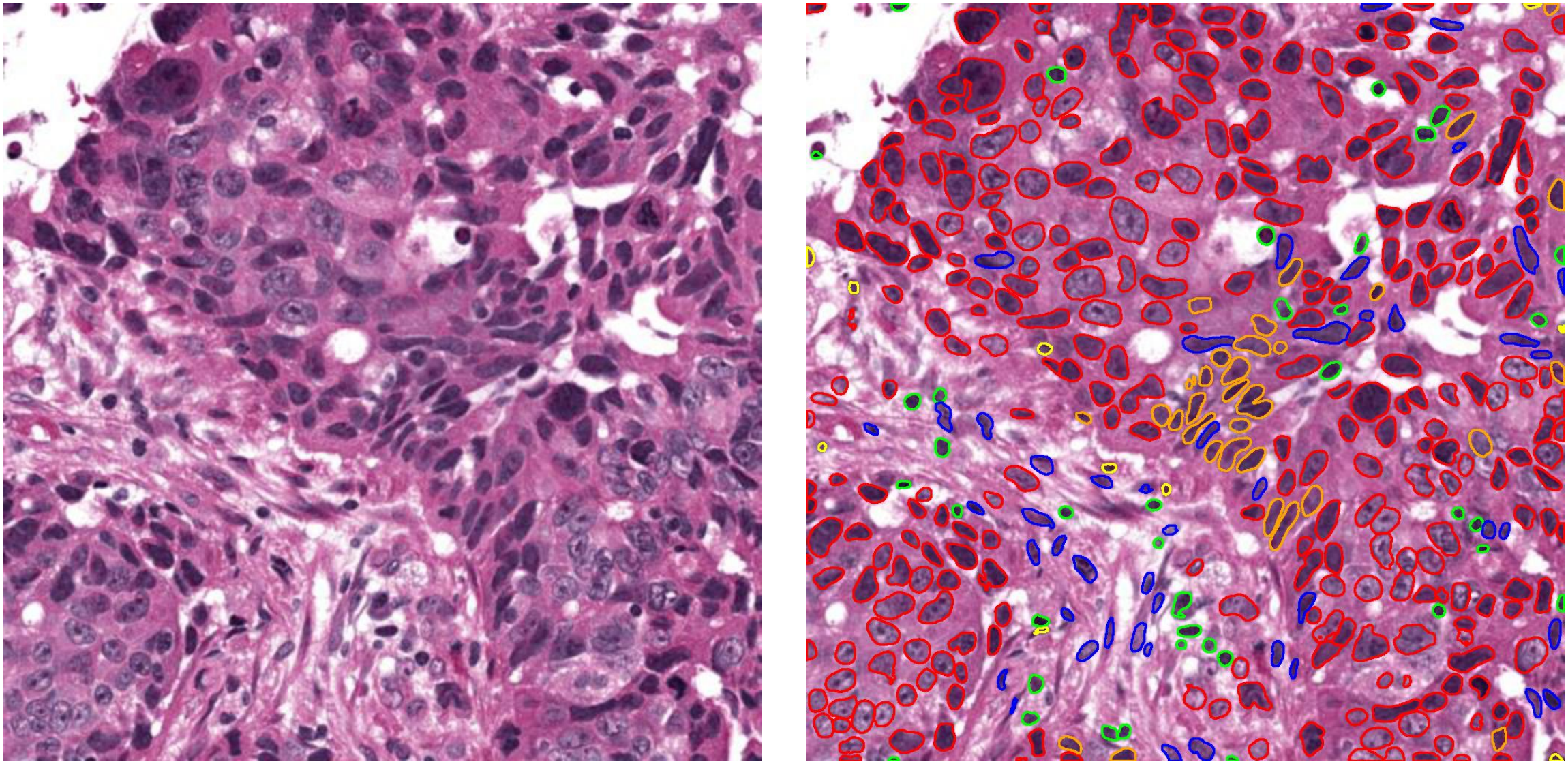

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

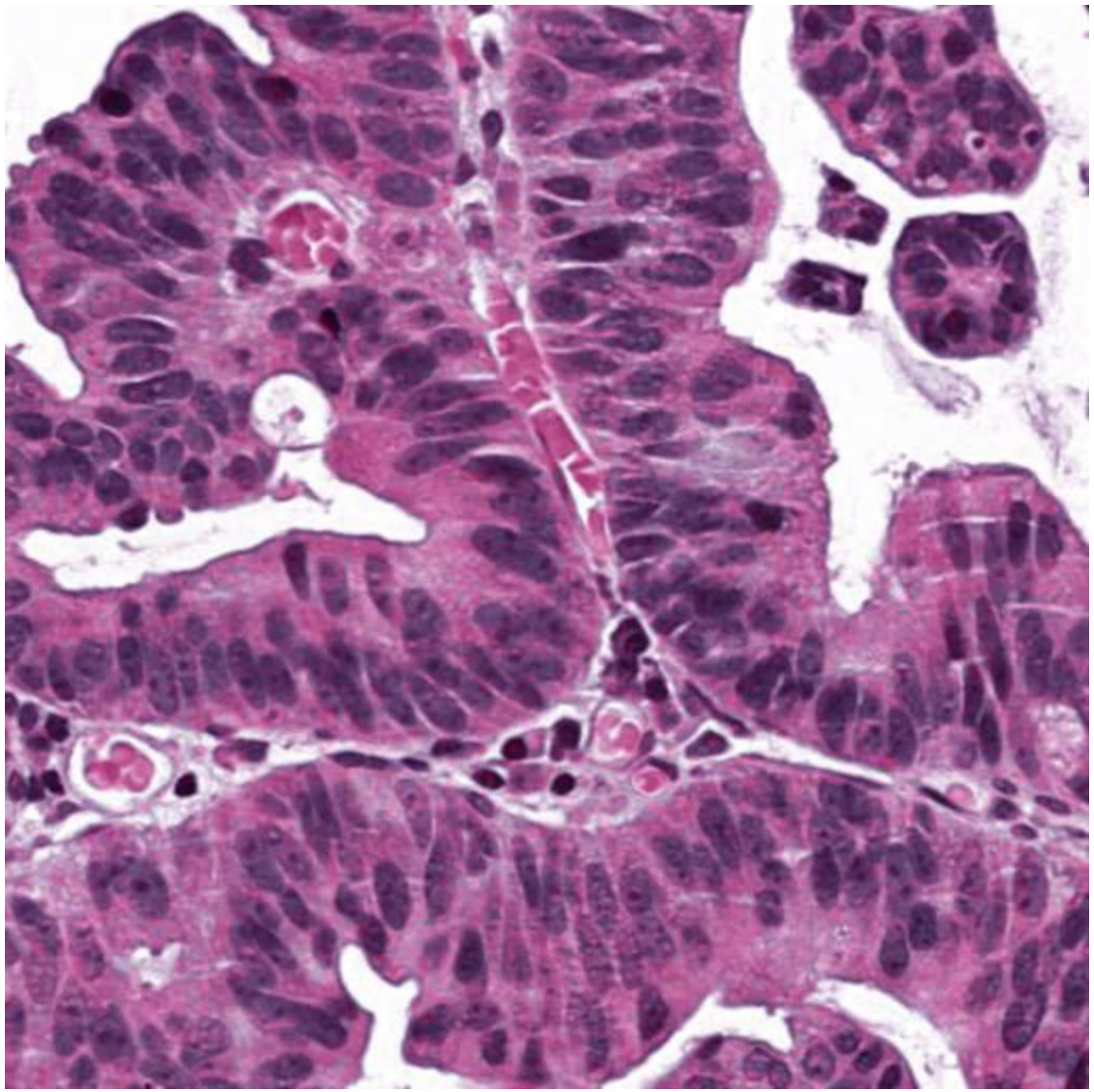

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

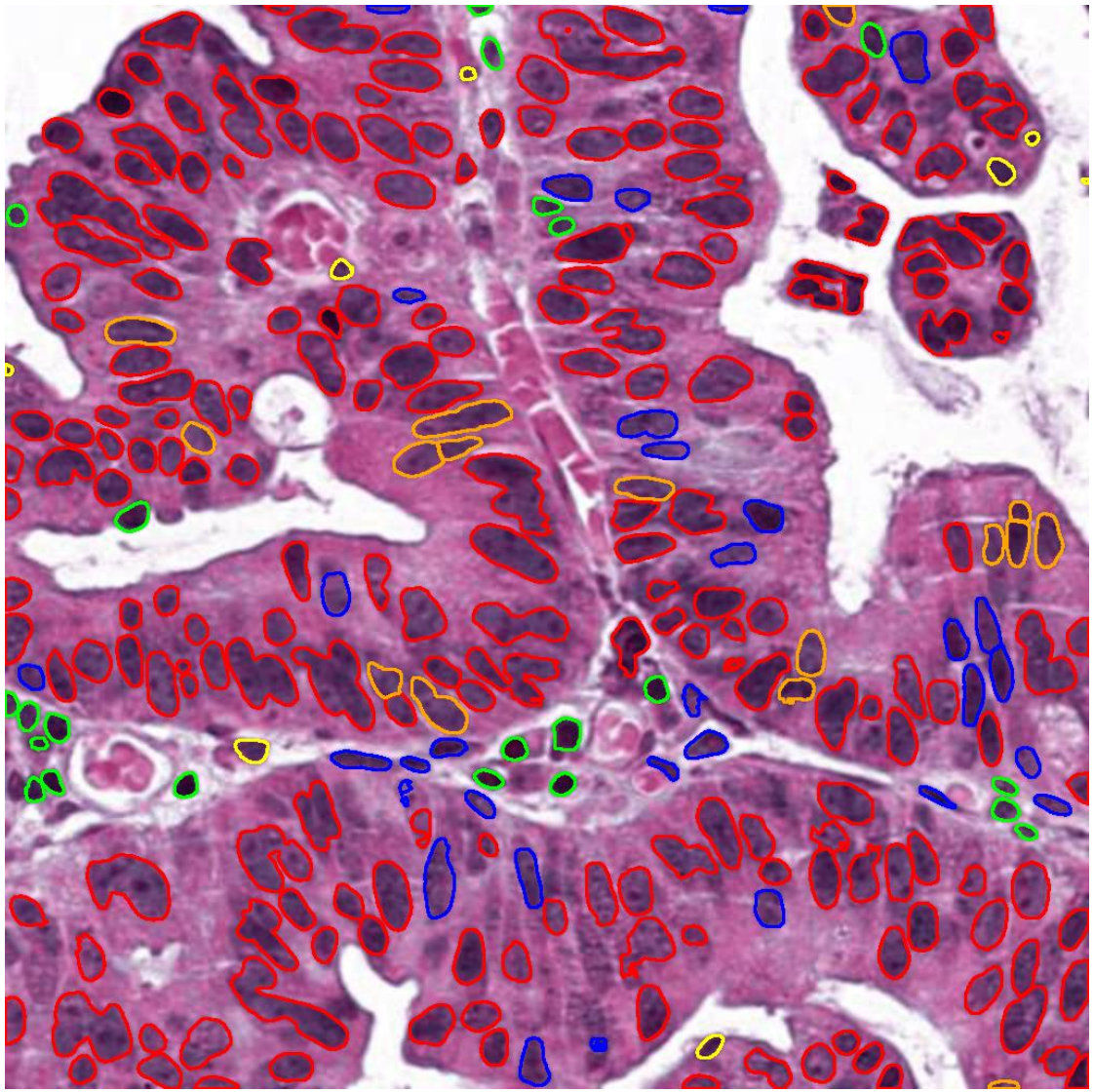




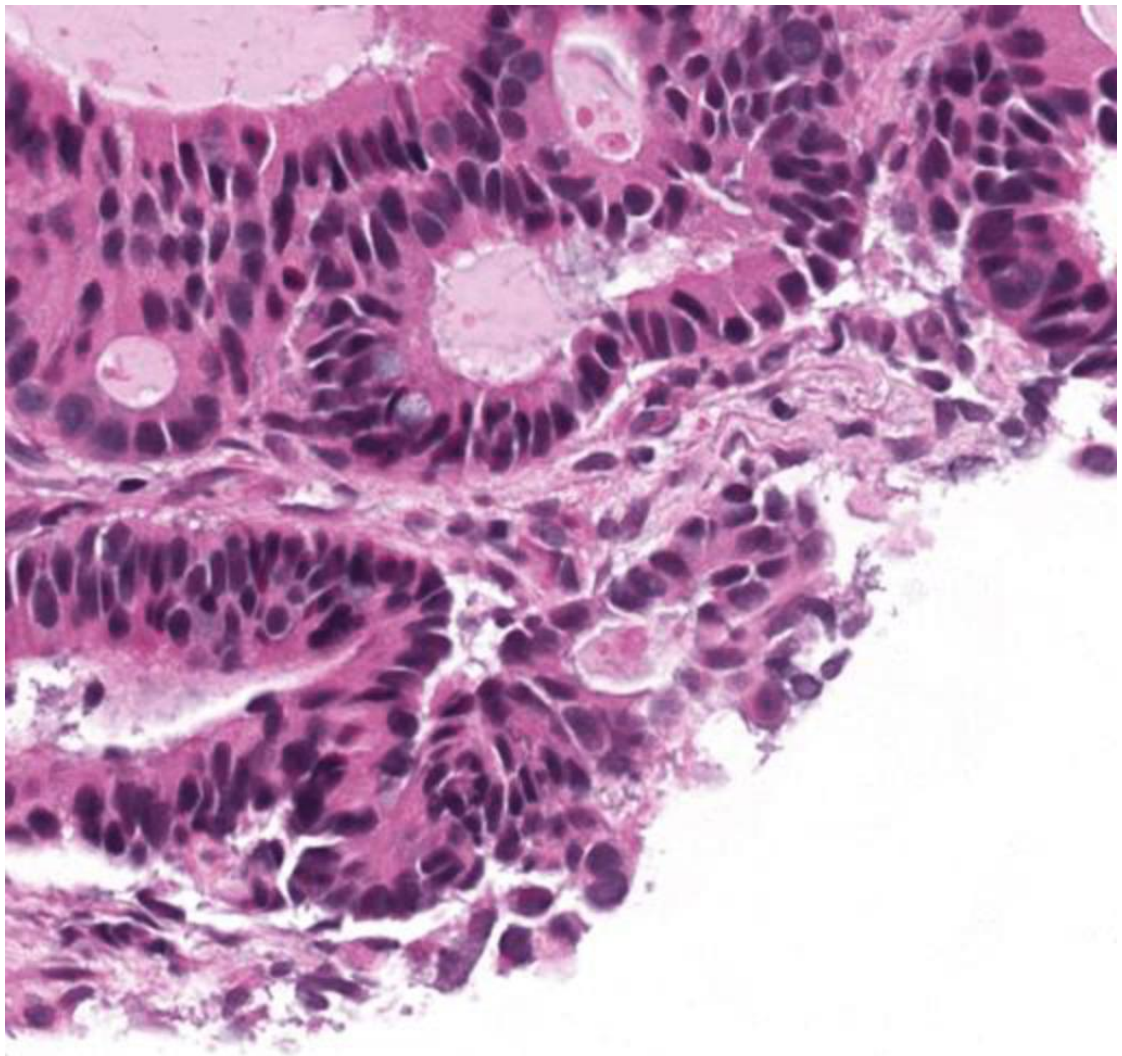
NEP1 Inflammatory Dead Mesenchymal NEP2

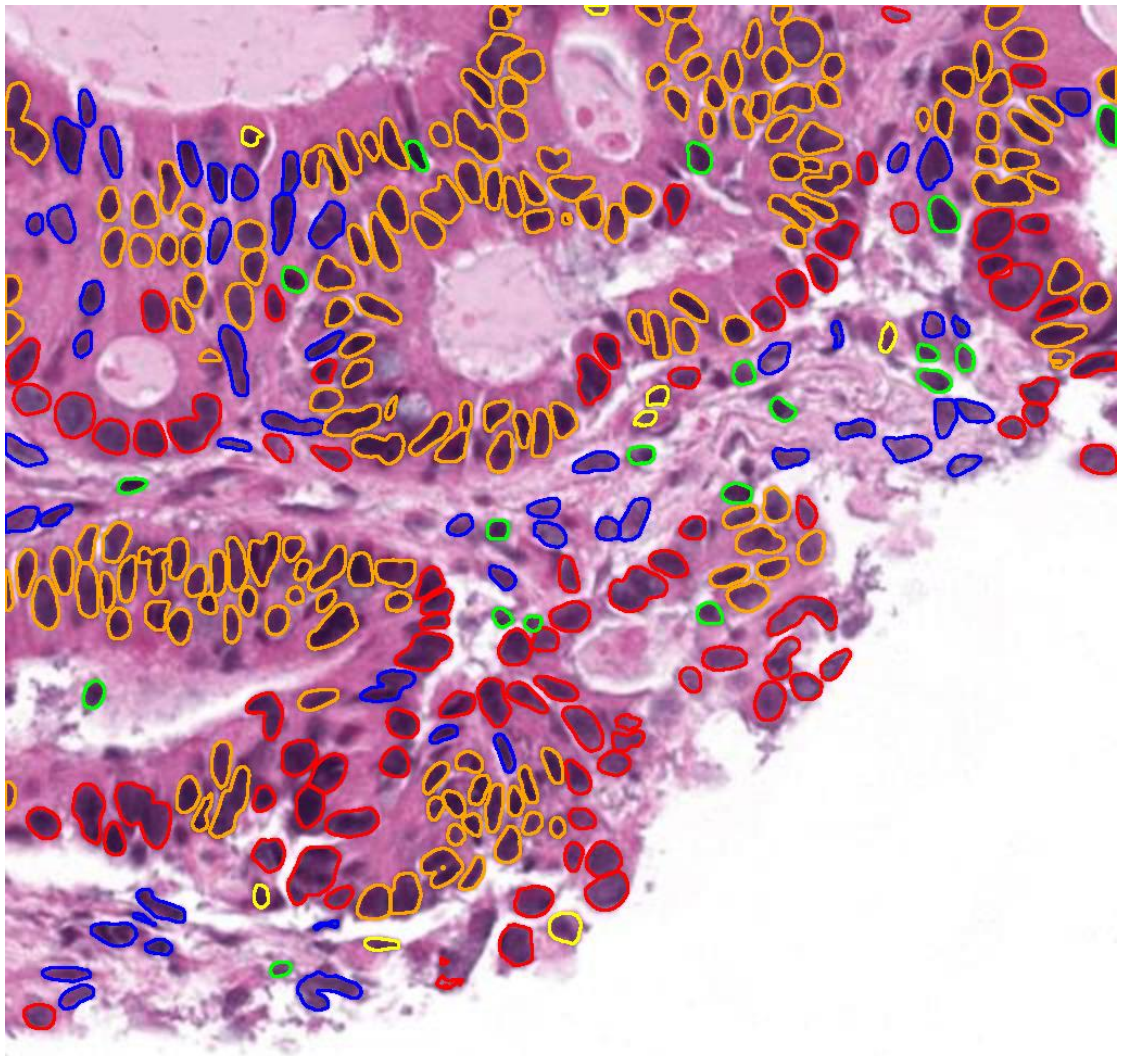

NEP1
Inflammatory
Dead
Mesenchymal
NEP2



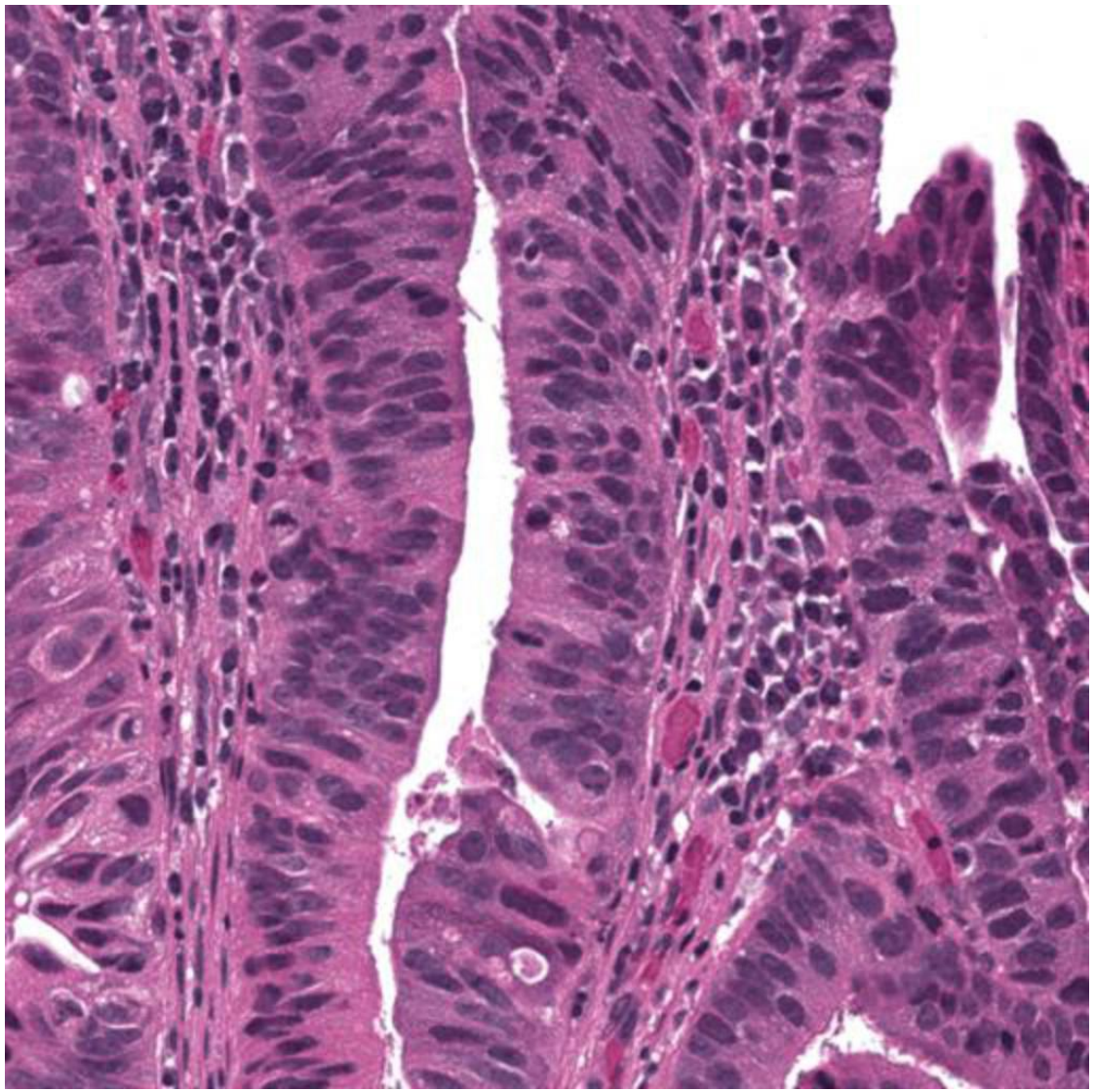


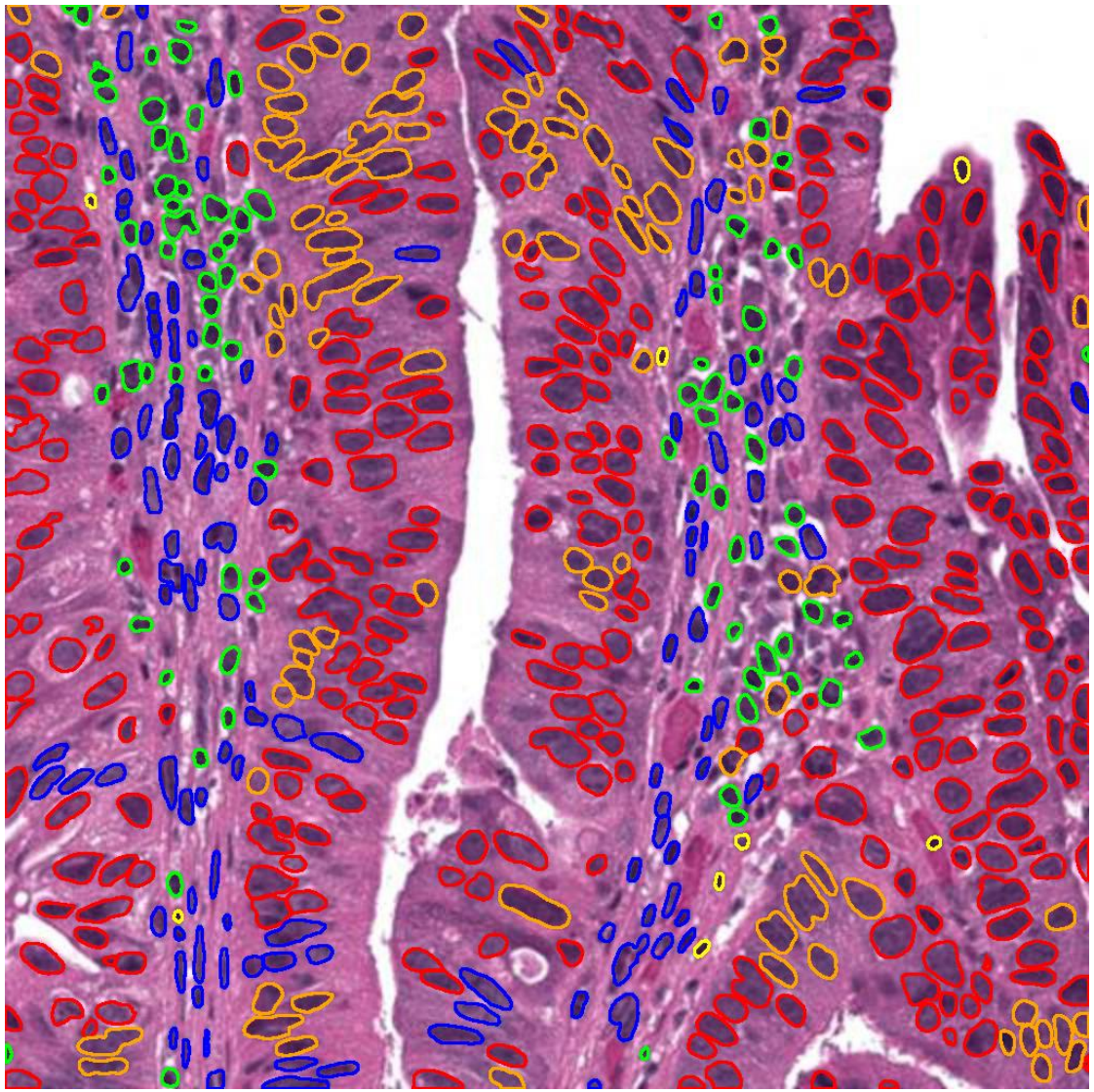
NEP1 Inflammatory Dead Mesenchymal NEP2



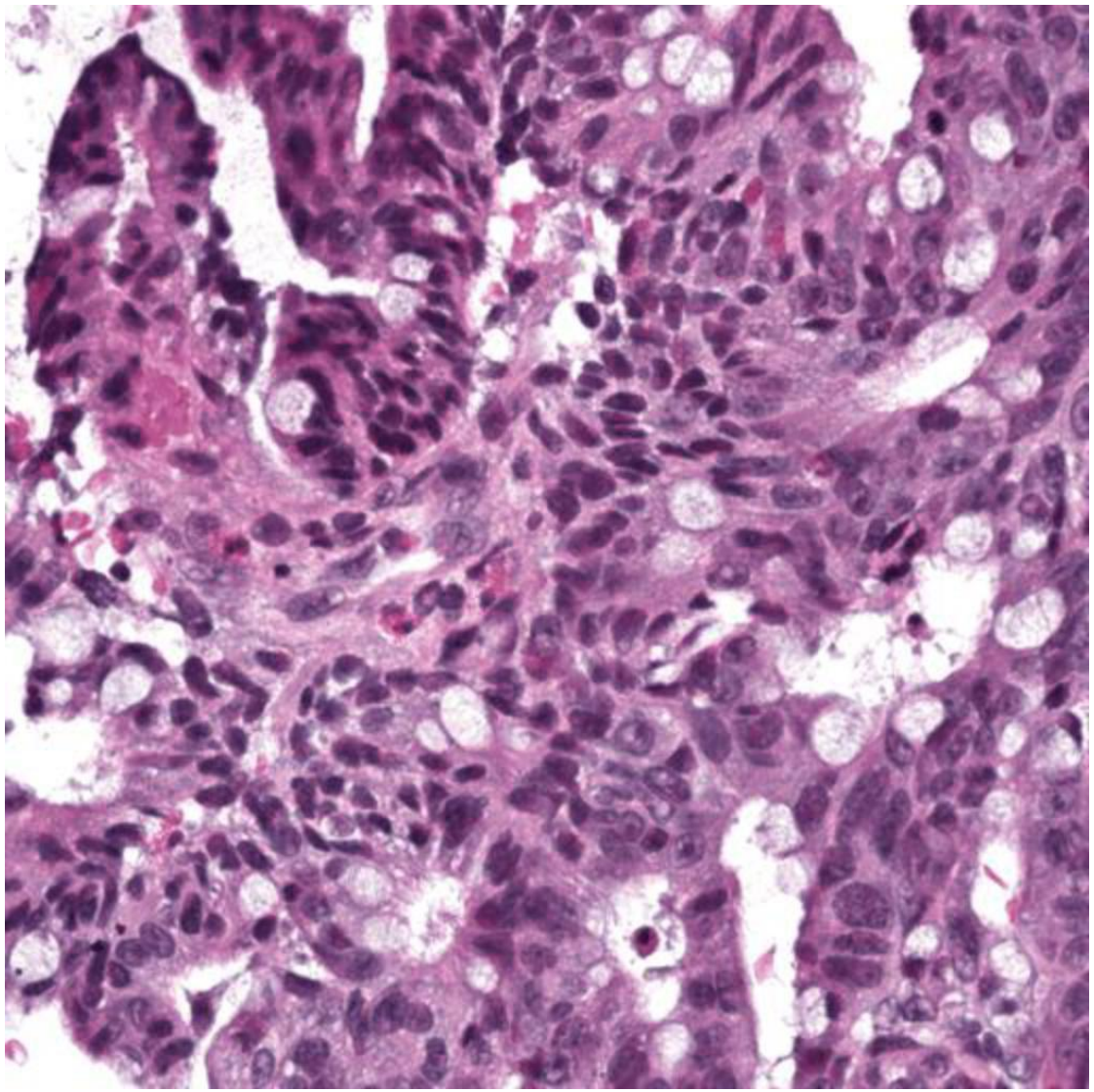


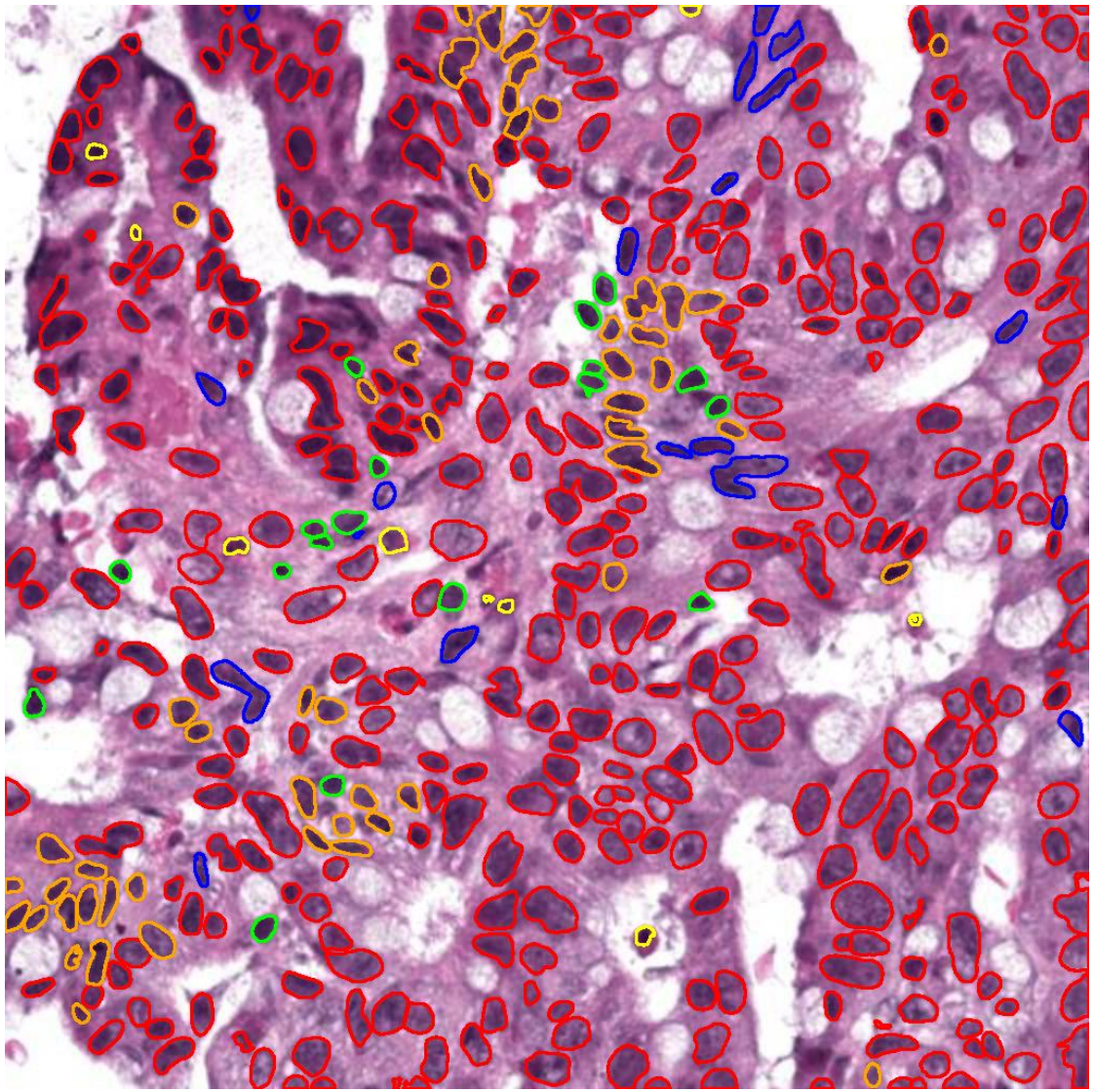
NEP1 Inflammatory Dead Mesenchymal NEP2



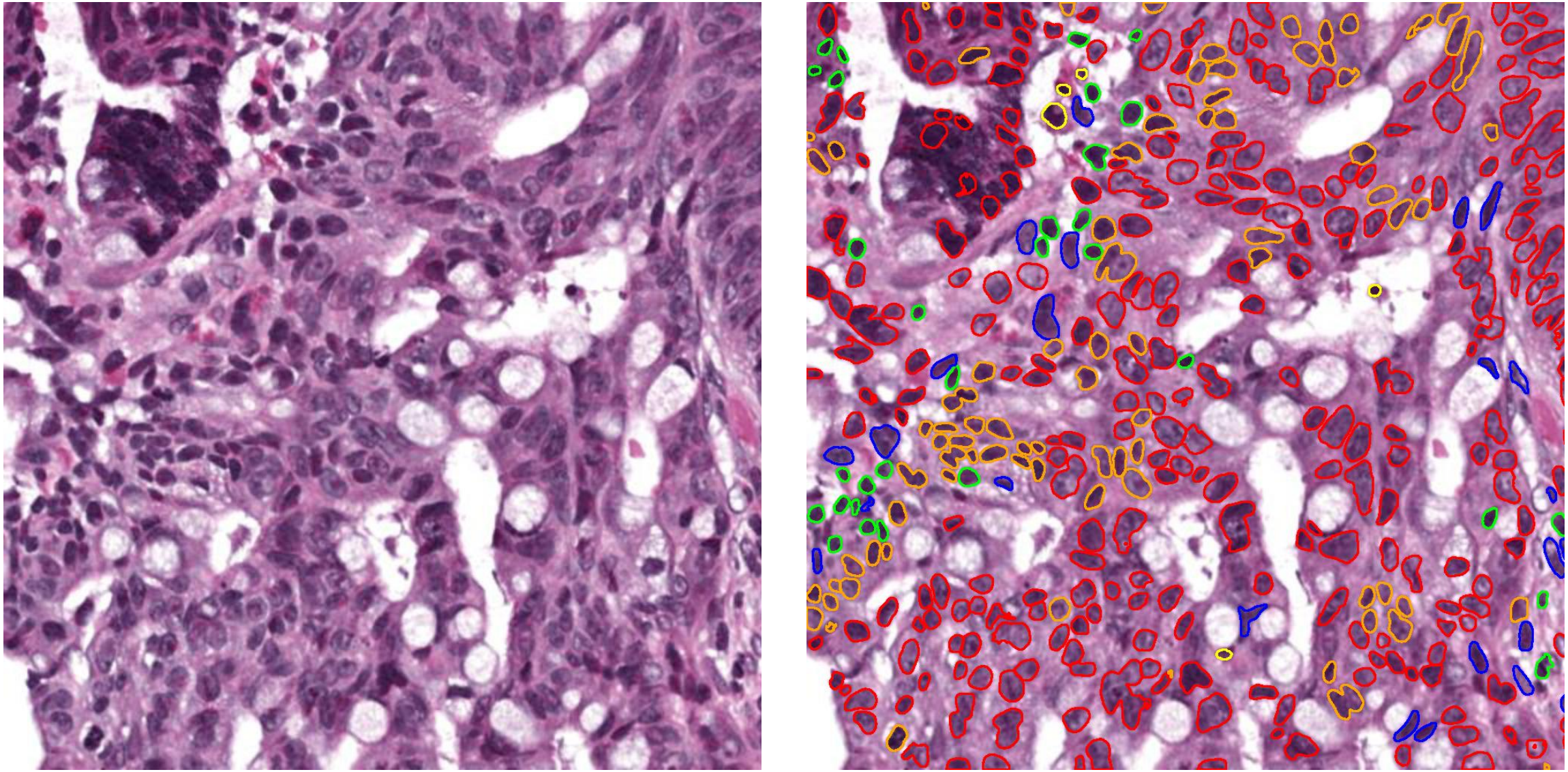


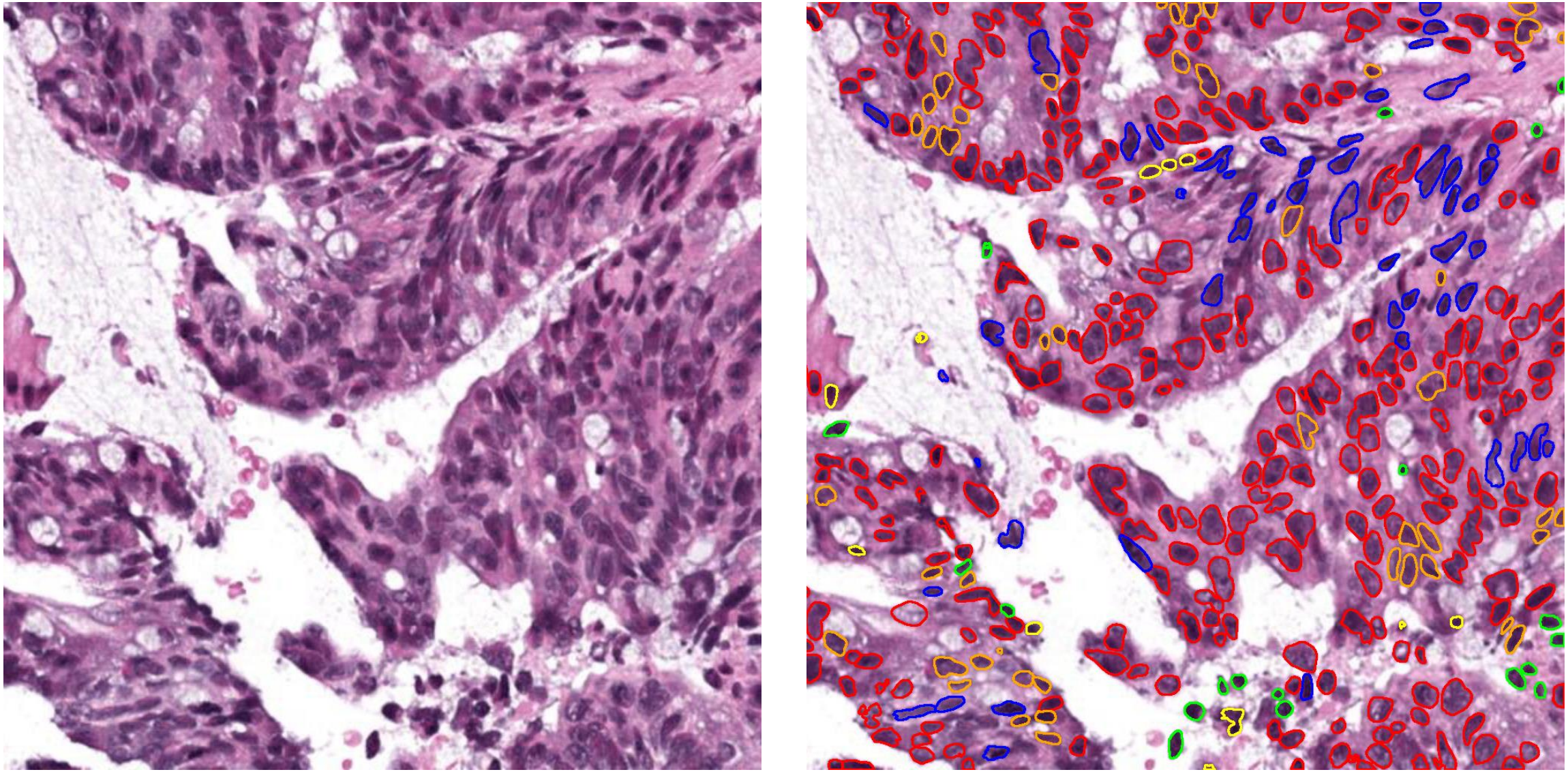
NEP1 Inflammatory Dead Mesenchymal NEP2

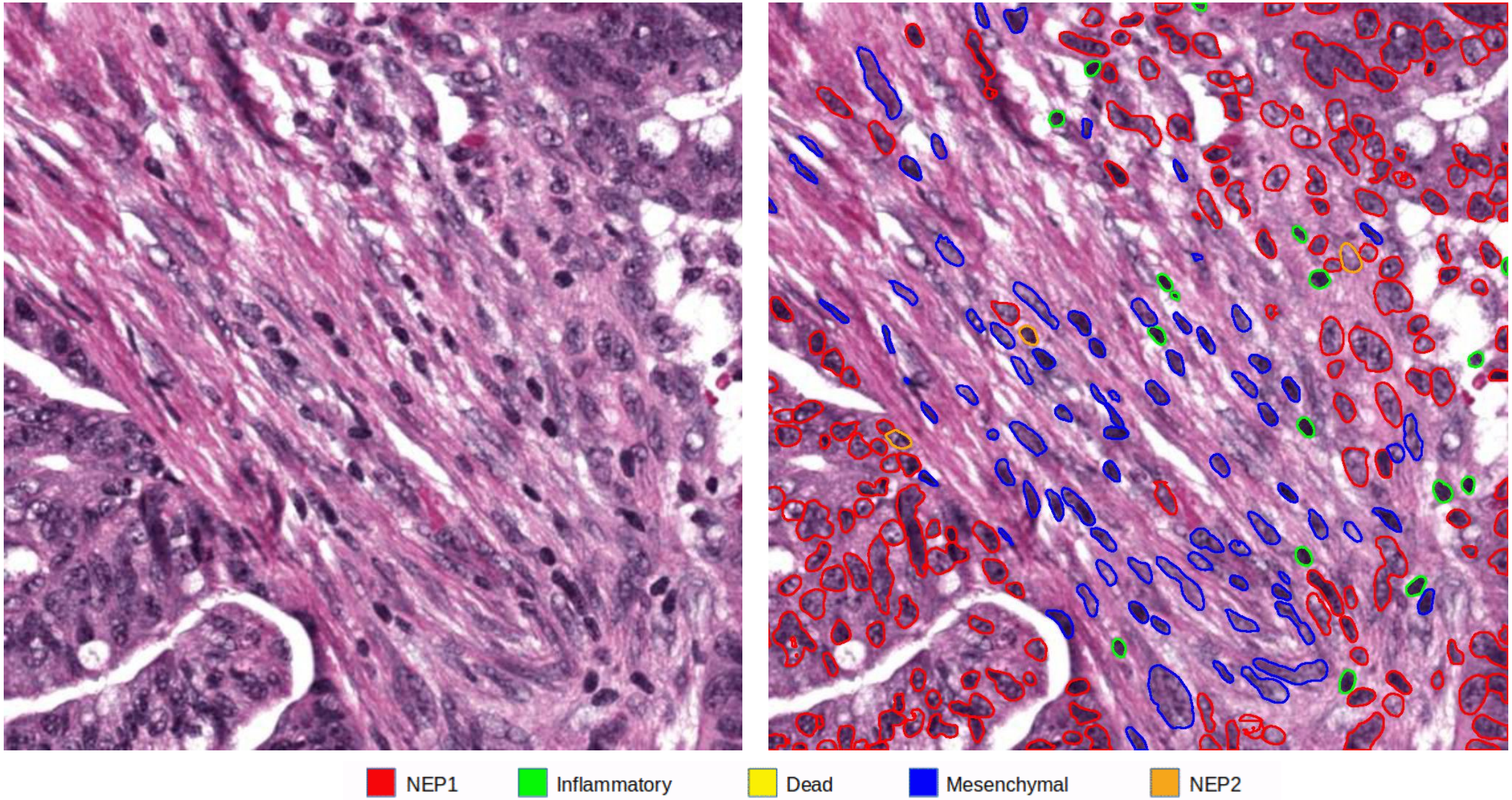


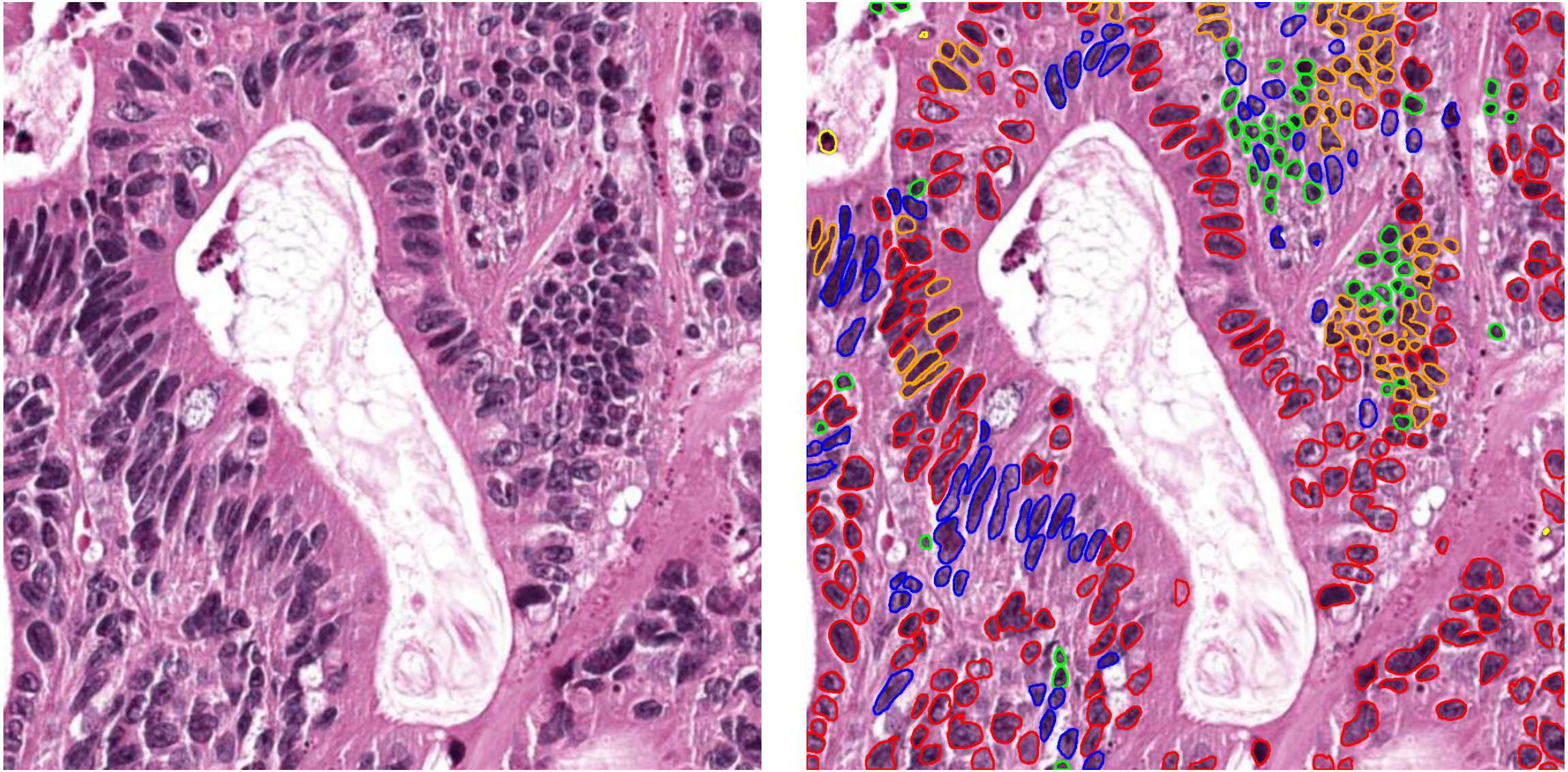


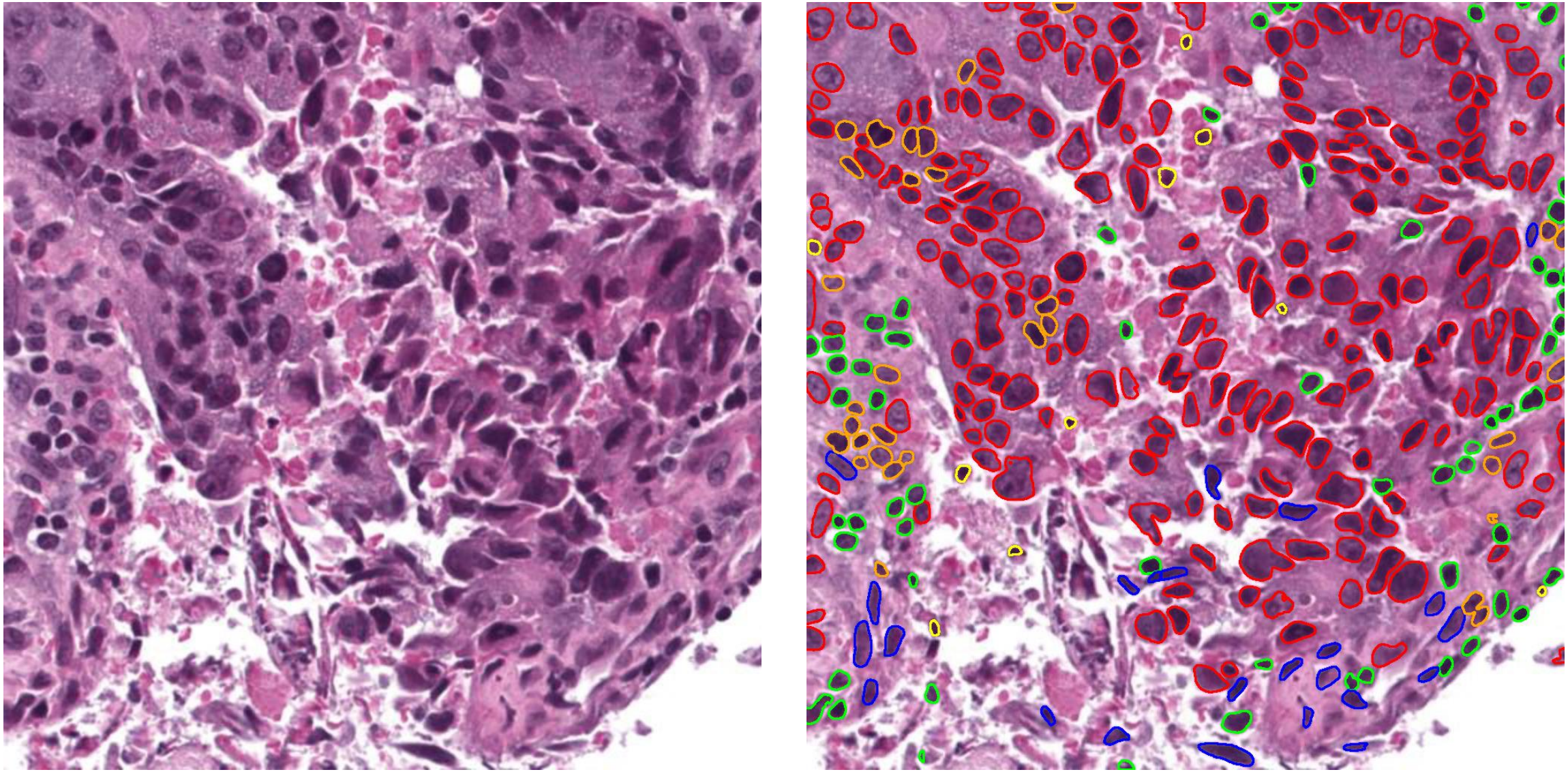
NEP1 Inflammatory Dead Mesenchymal NEP2

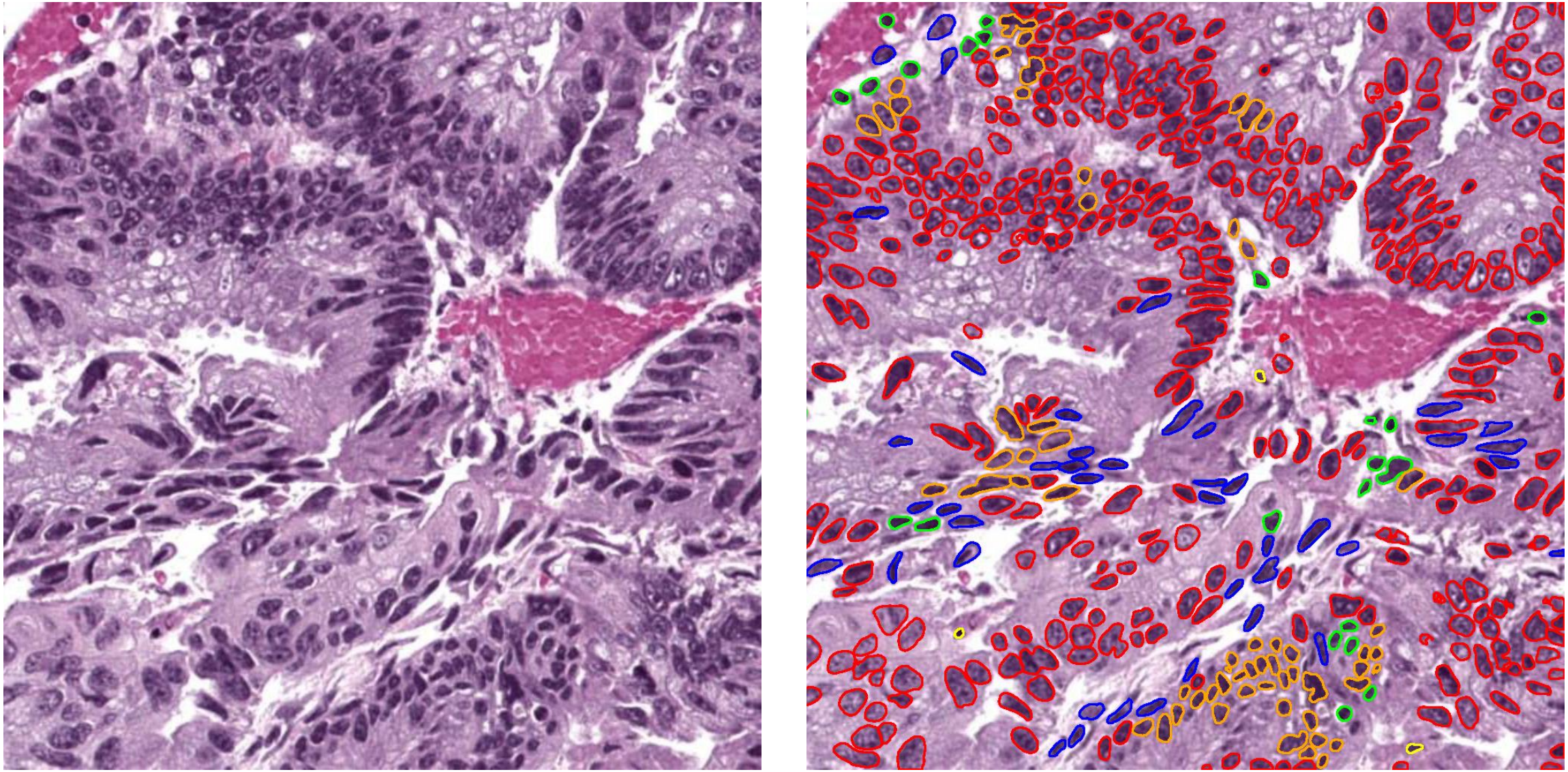

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

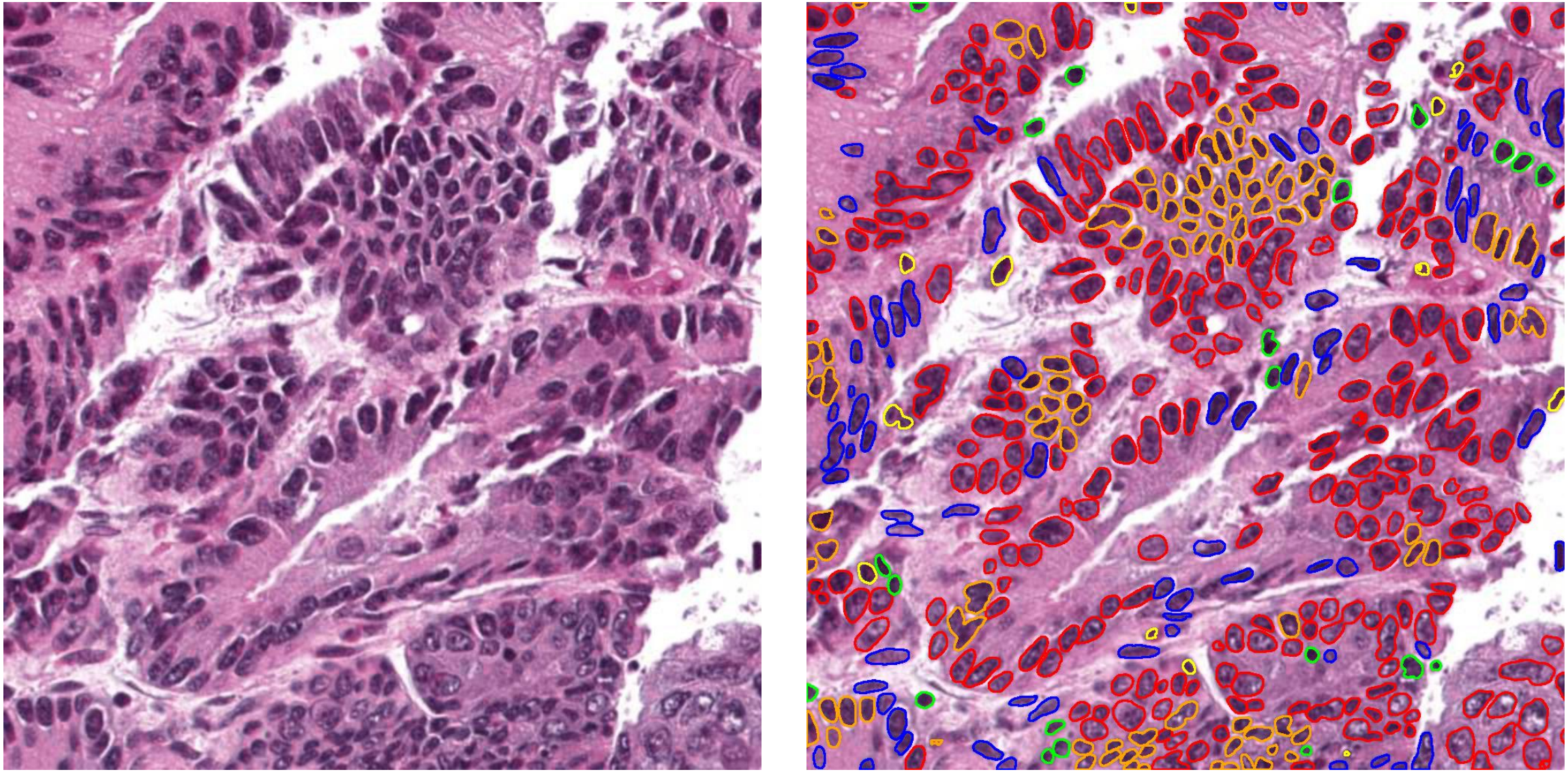

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

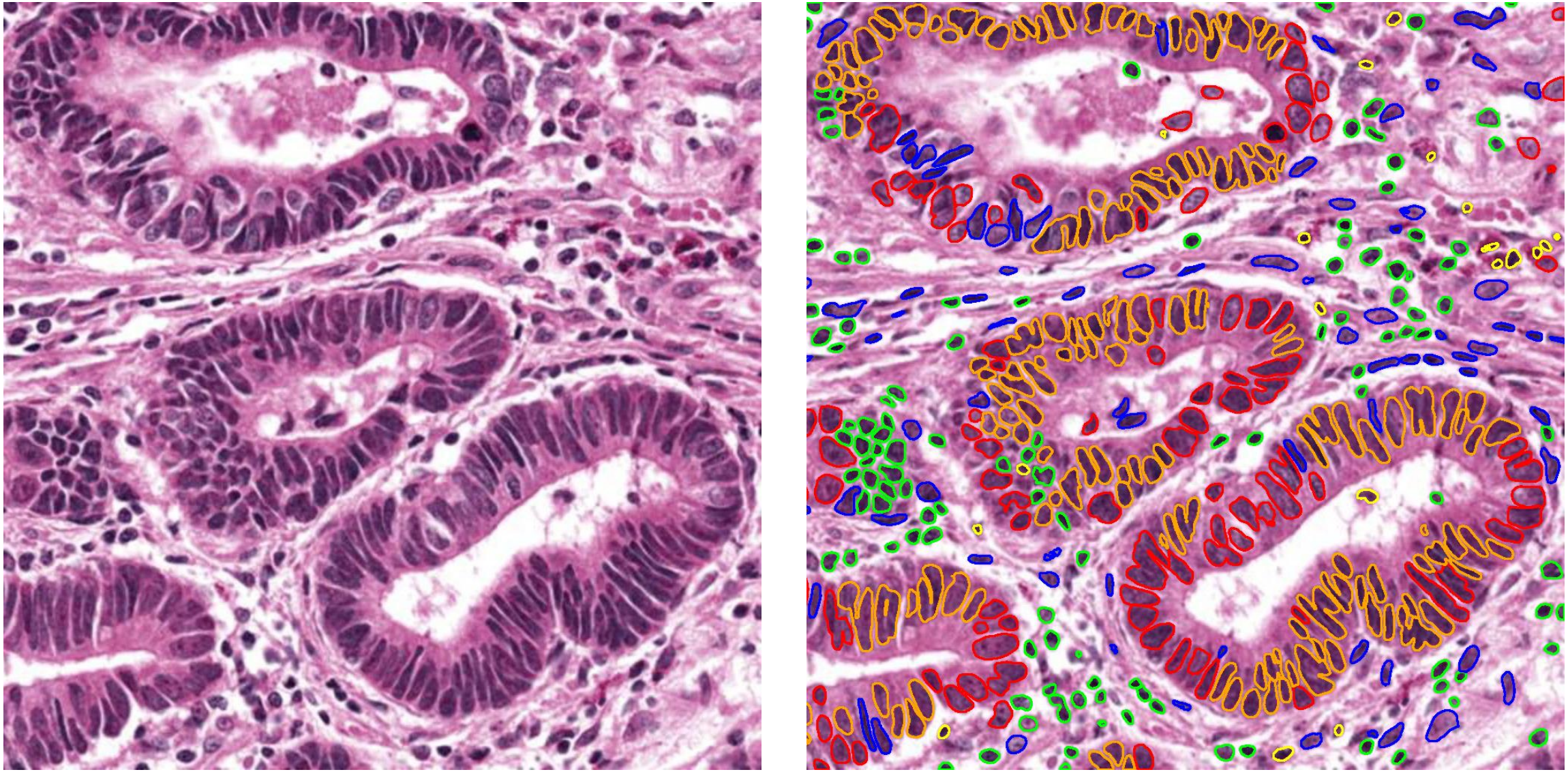

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

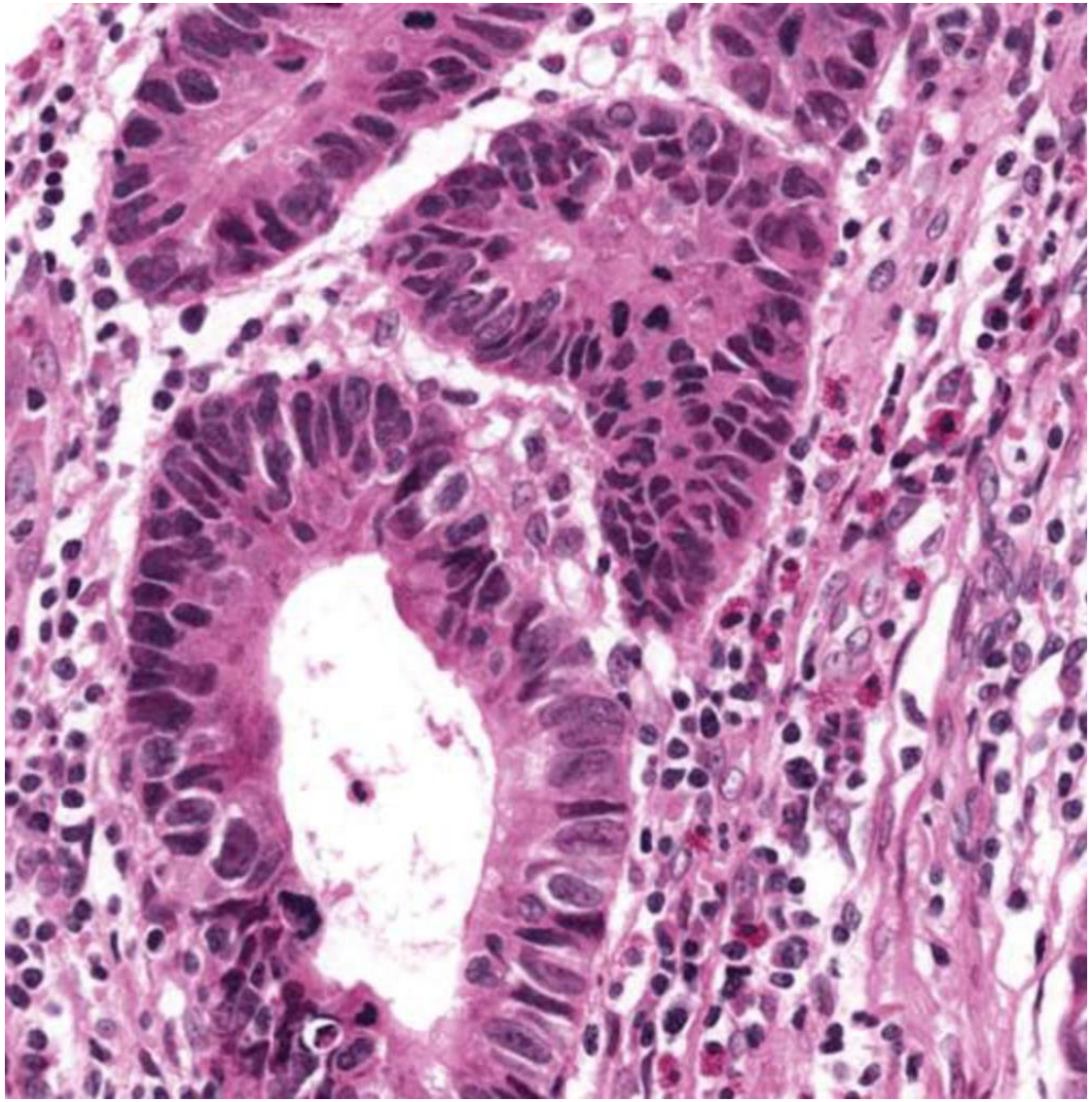

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

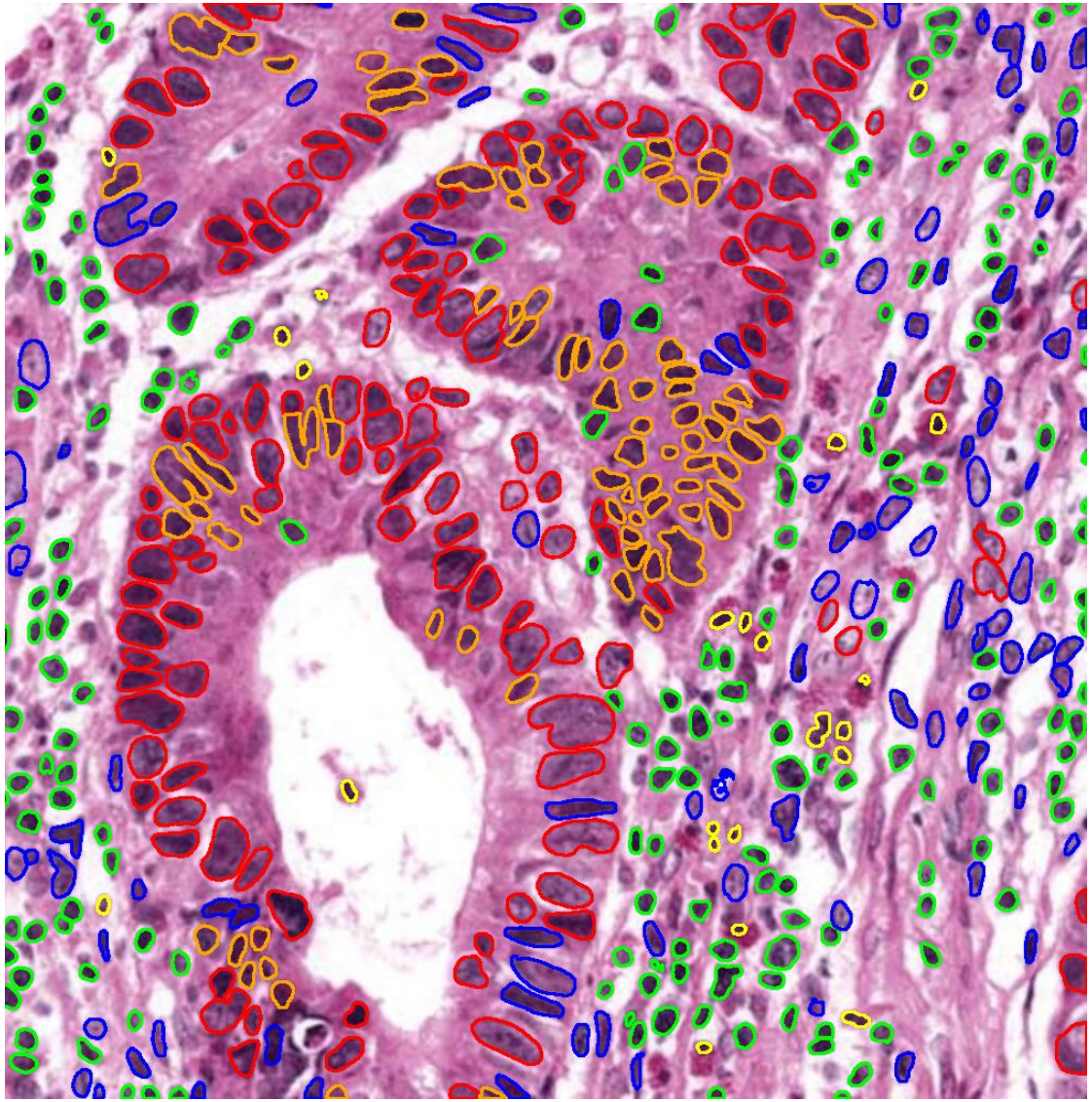

NEP1
Inflammatory
Dead
Mesenchymal
NEP2


NEP1
Inflammatory
Dead
Mesenchymal
NEP2

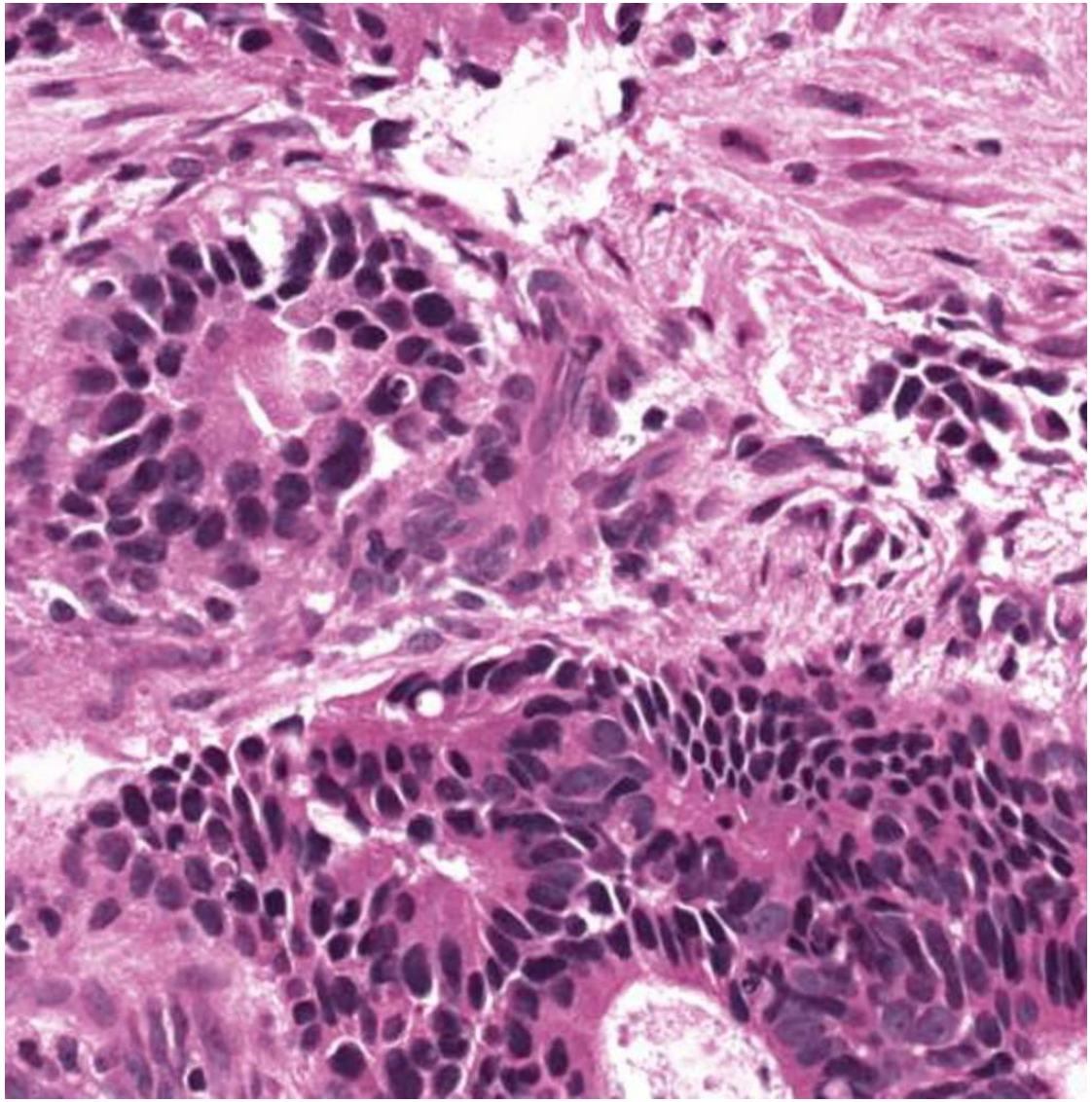

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

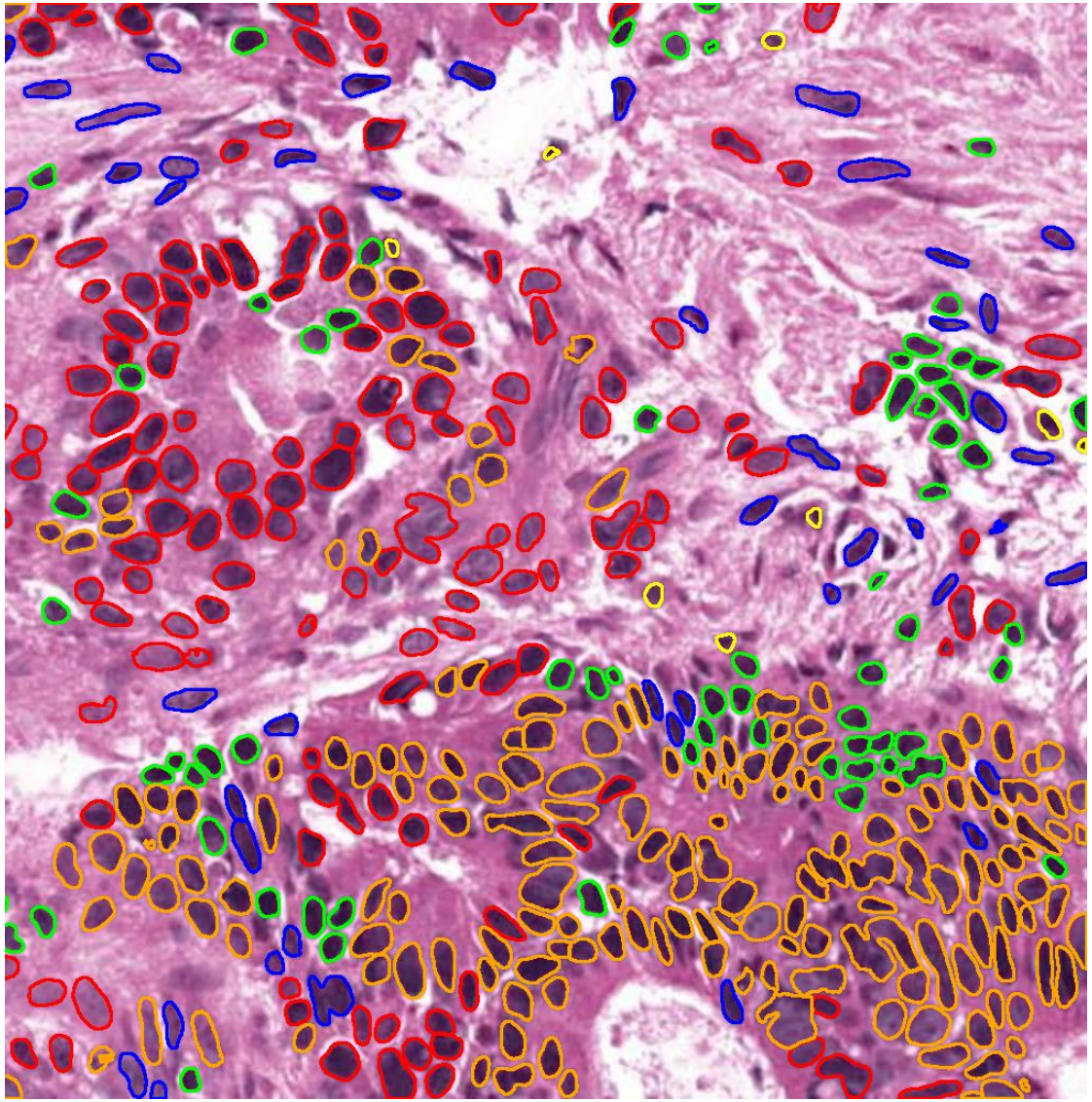

NEP1
Inflammatory
Dead
Mesenchymal
NEP2





NEP1 Inflammatory Dead Mesenchymal NEP2

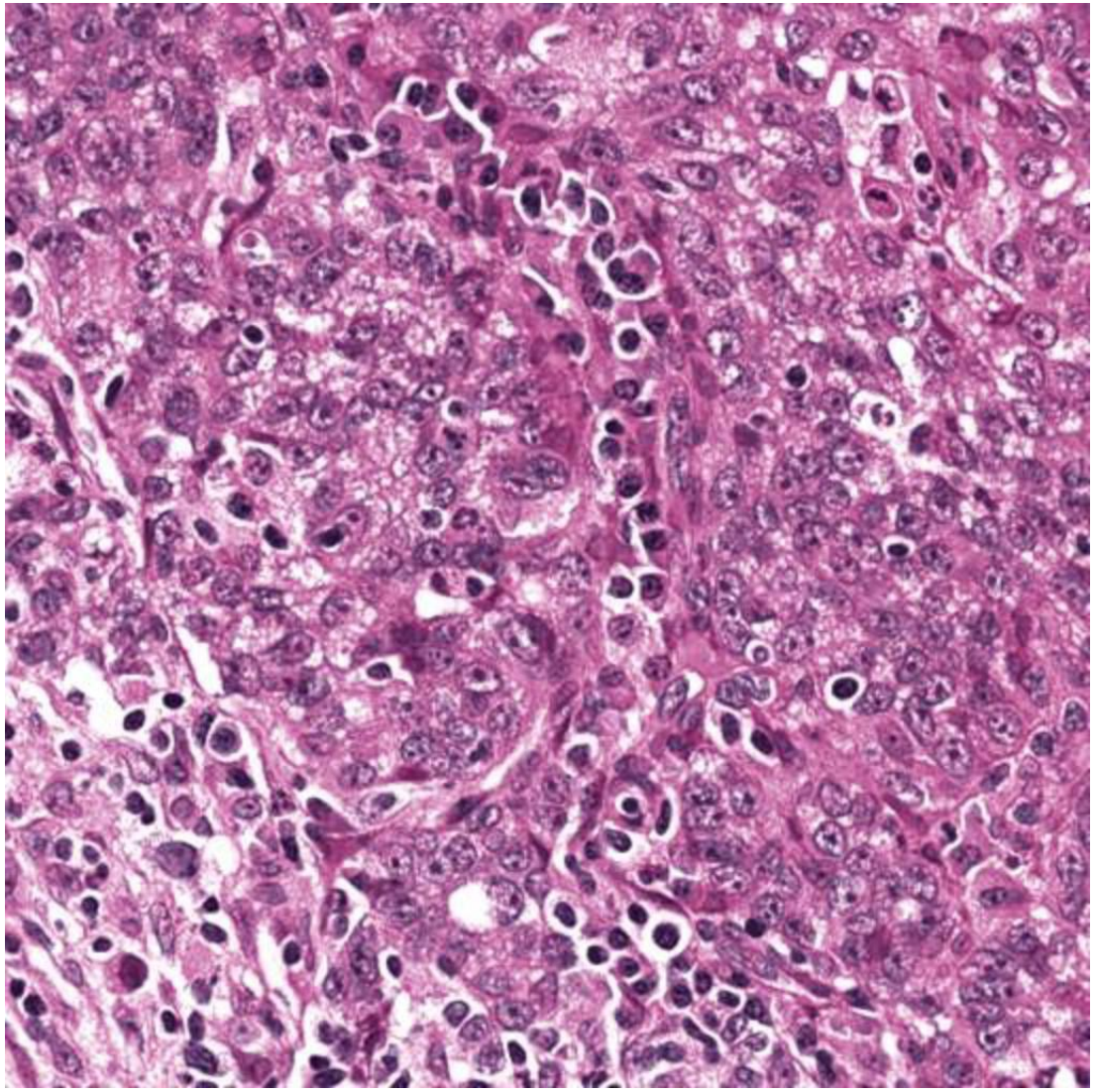


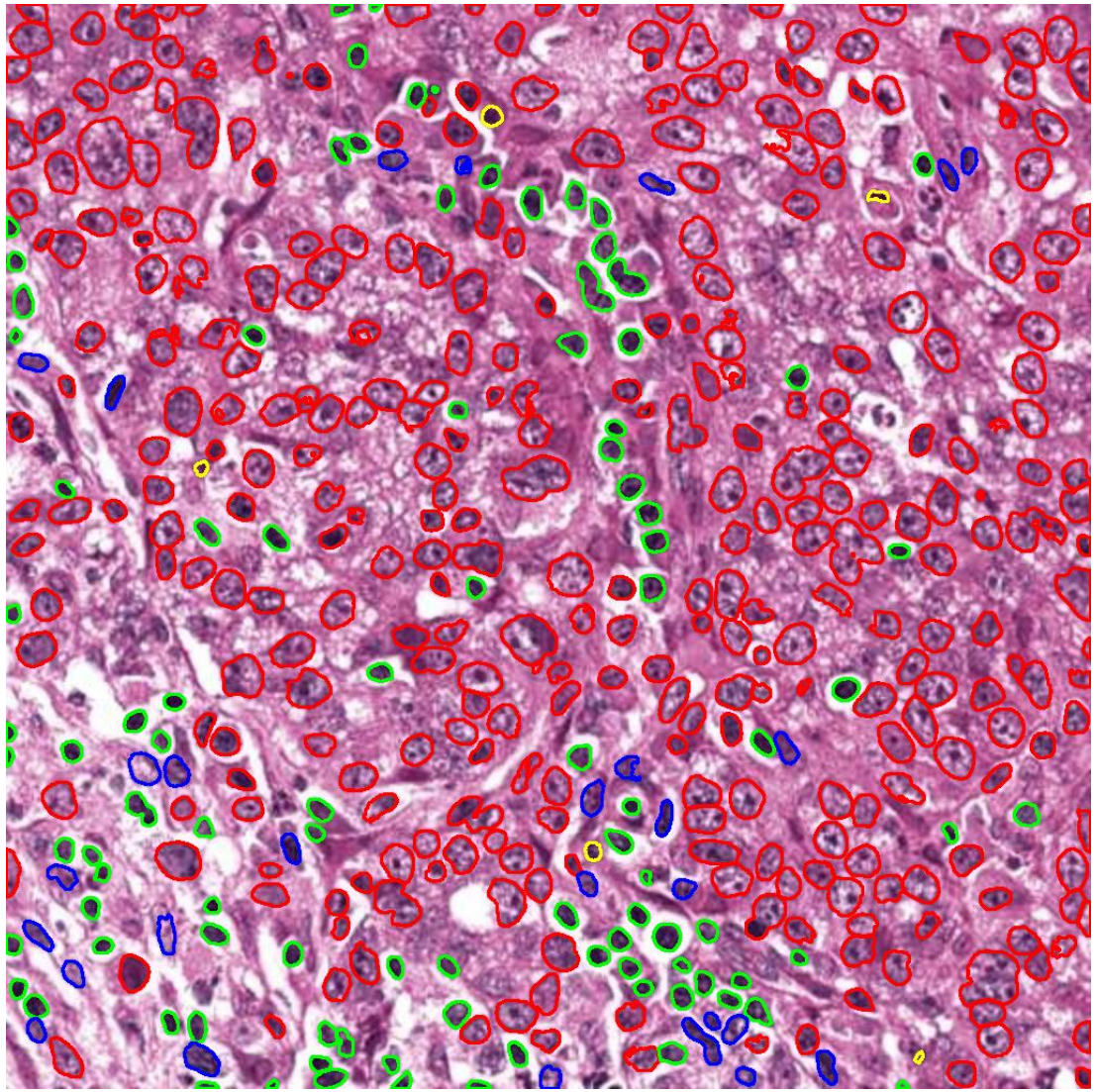


NEP1 Inflammatory Dead Mesenchymal NEP2

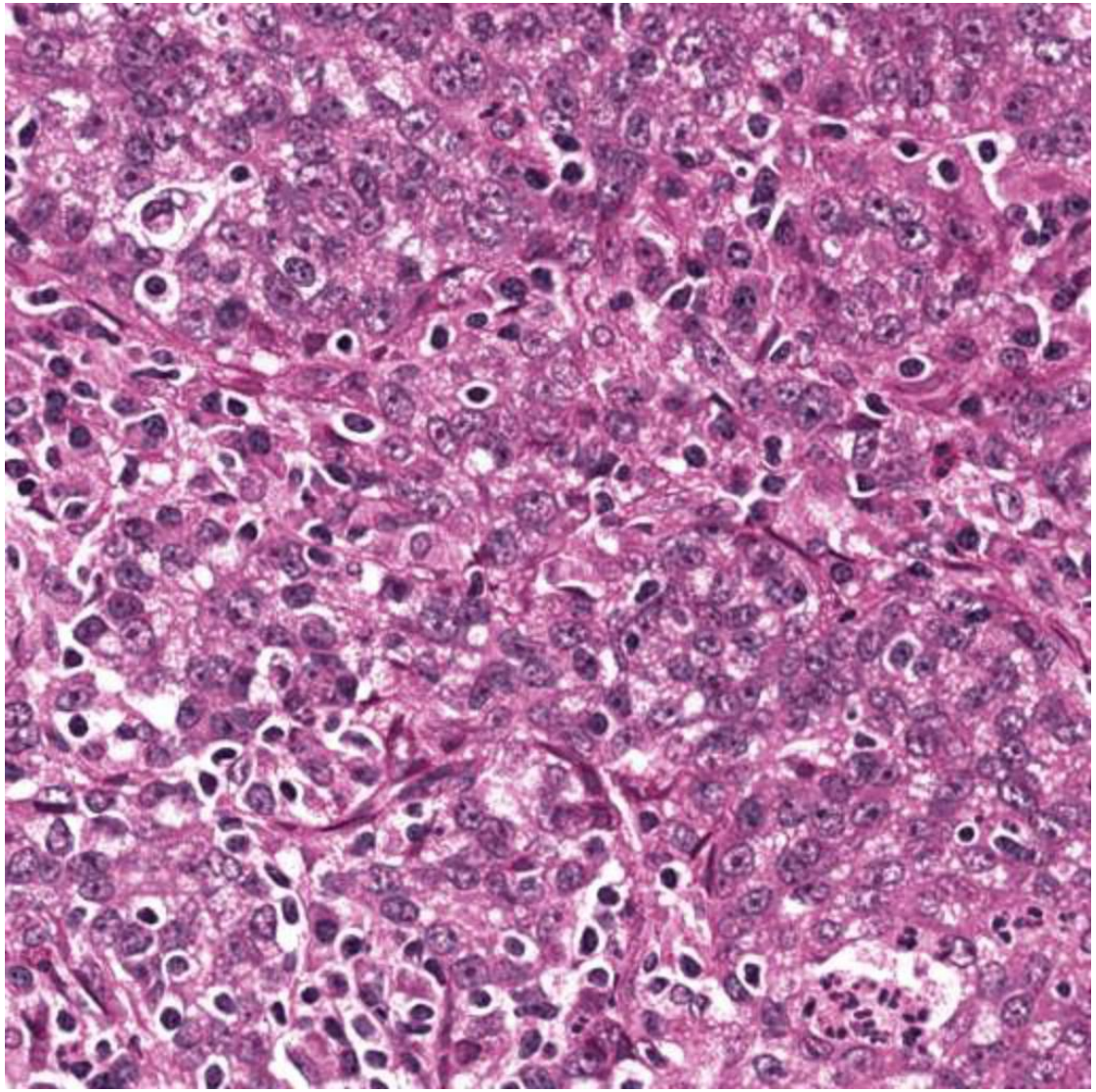
GS cellular composition

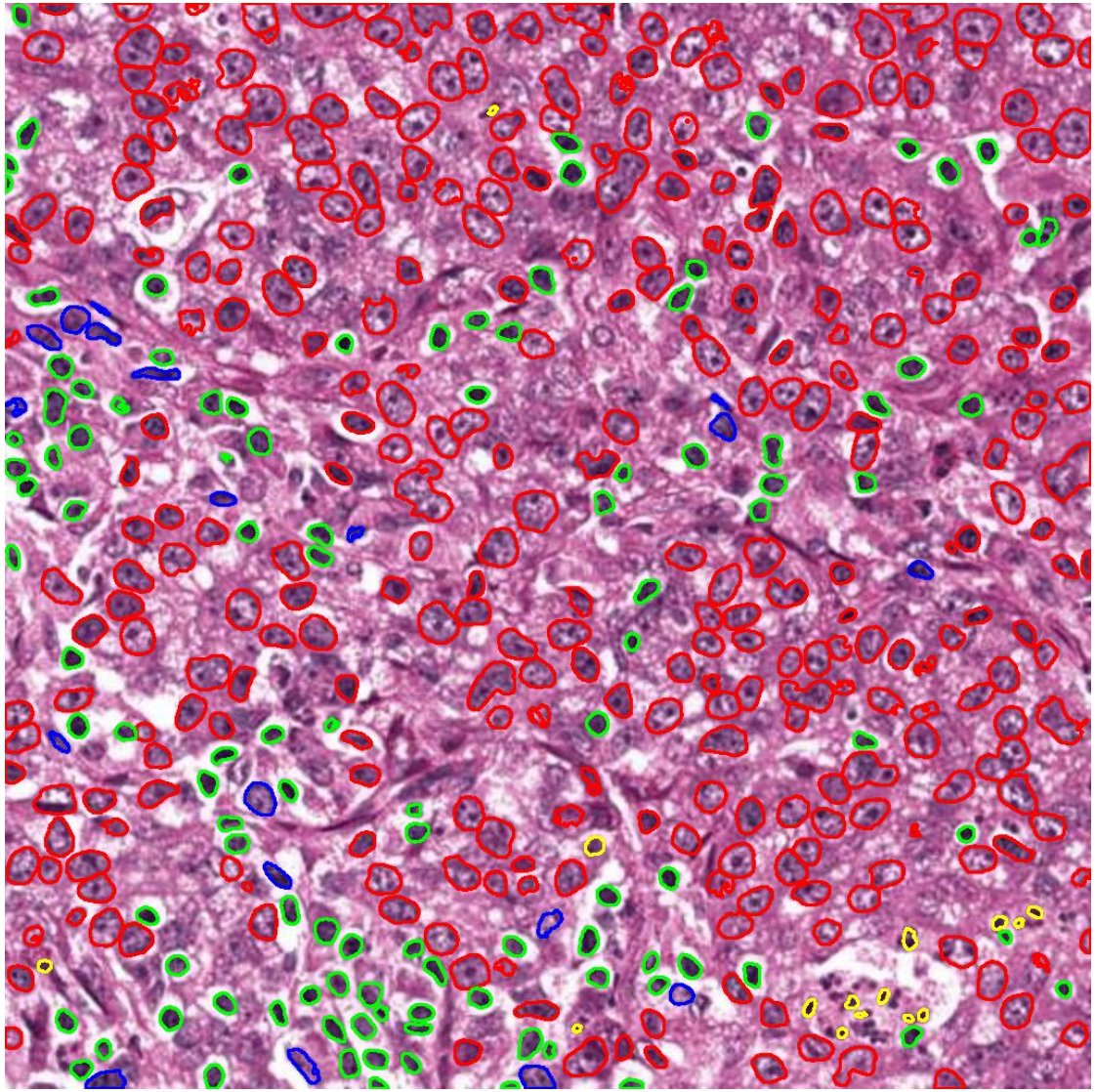
Supplementary Material



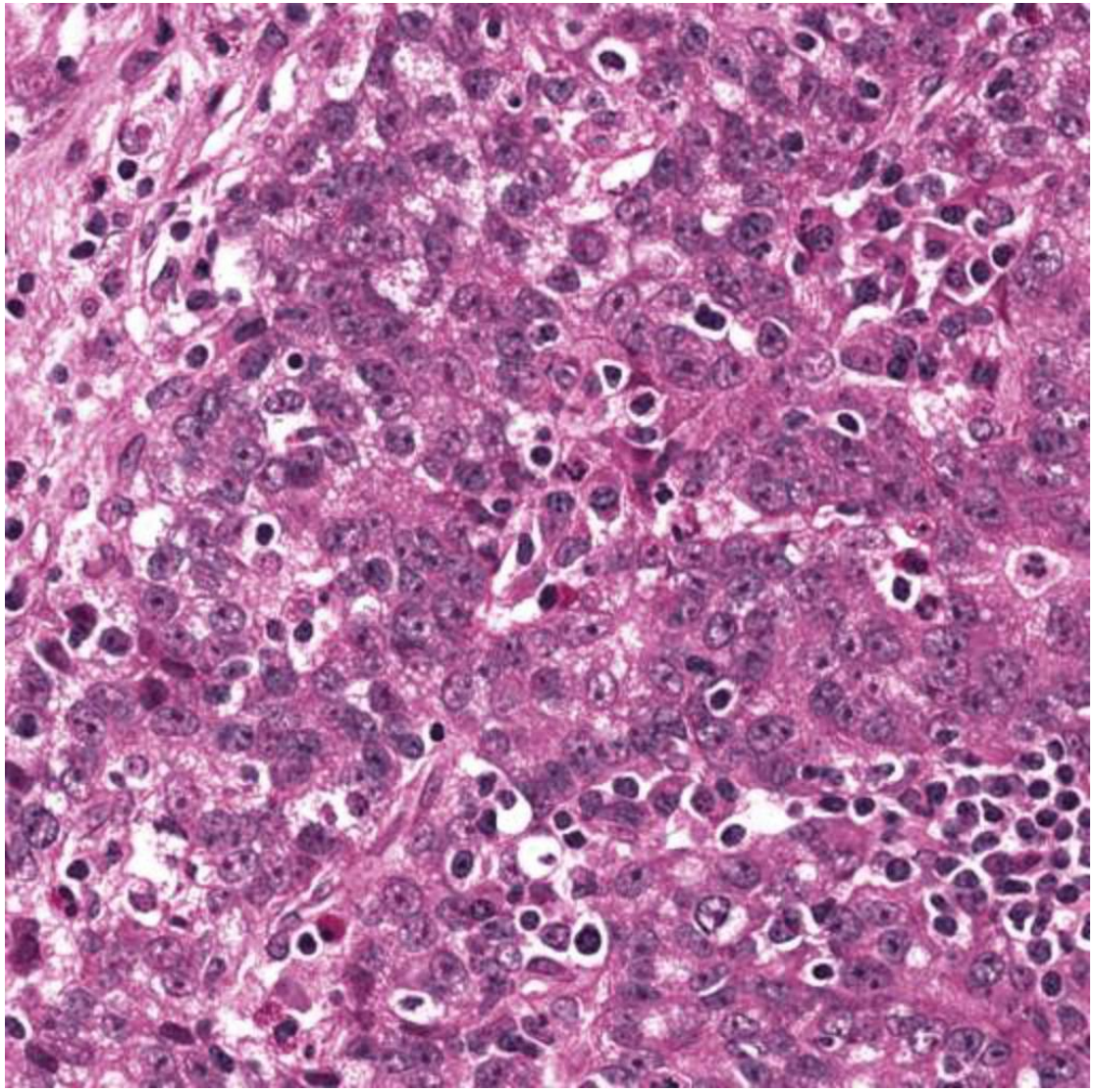


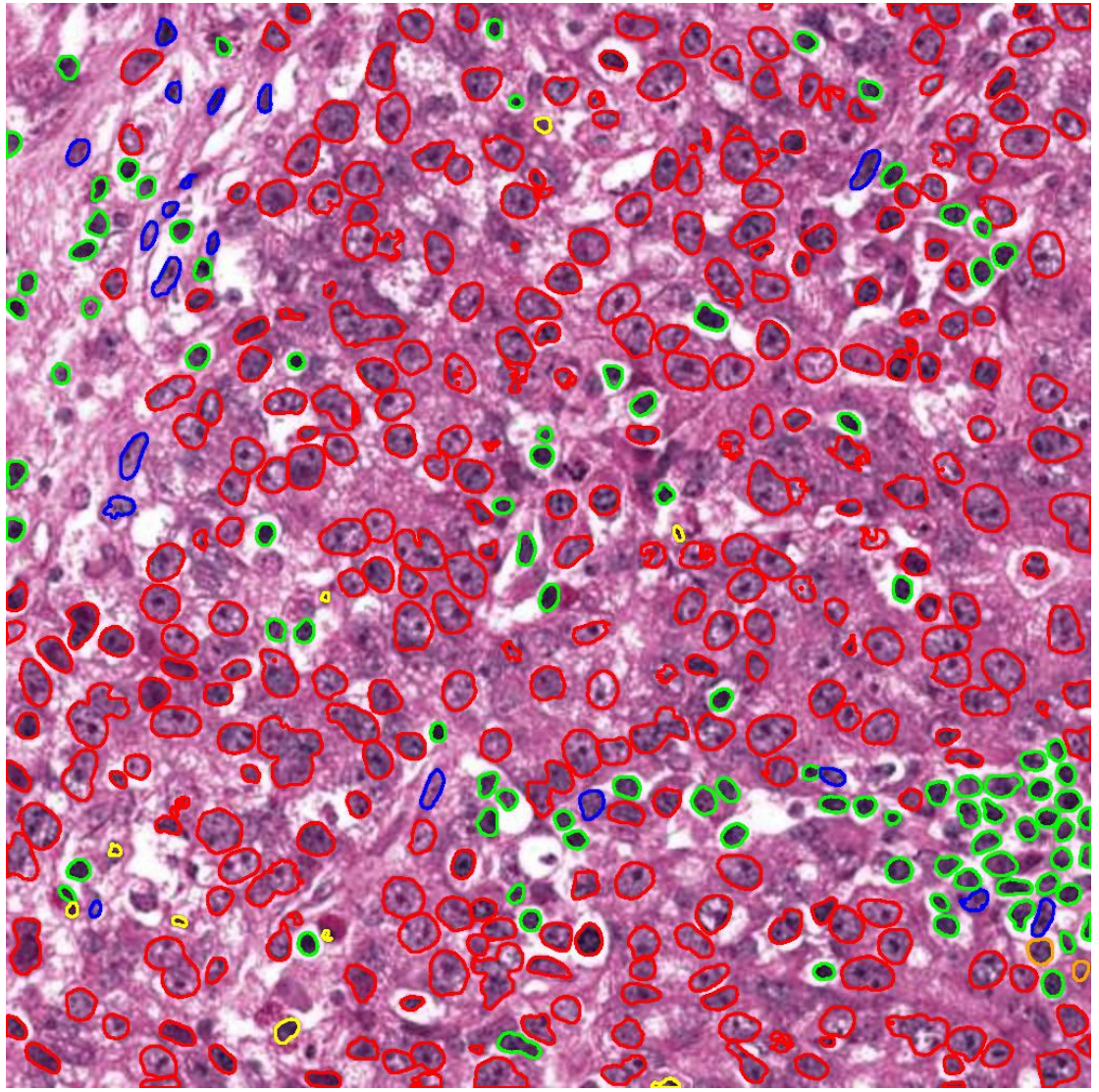
NEP1 Inflammatory Dead Mesenchymal NEP2



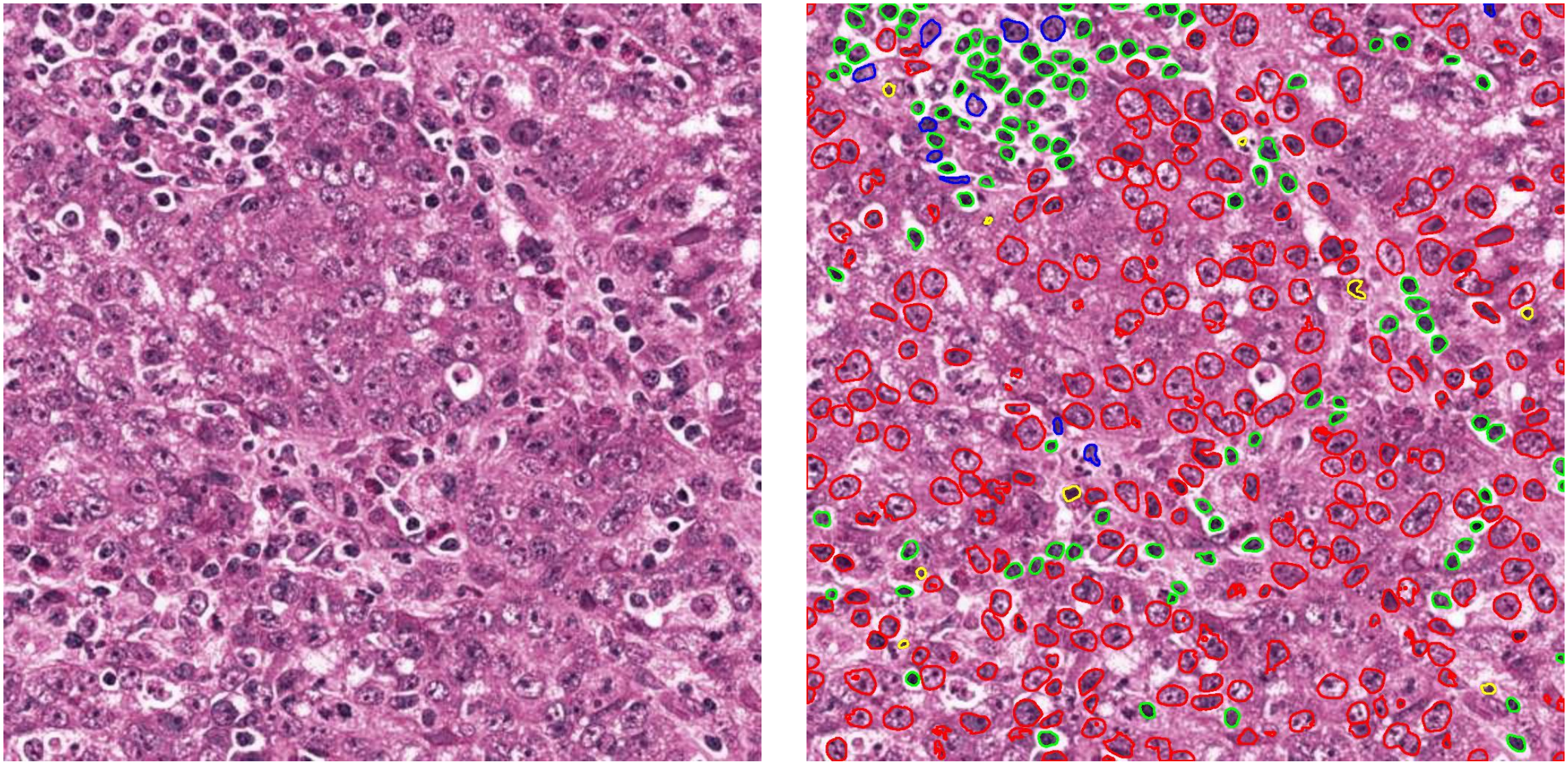


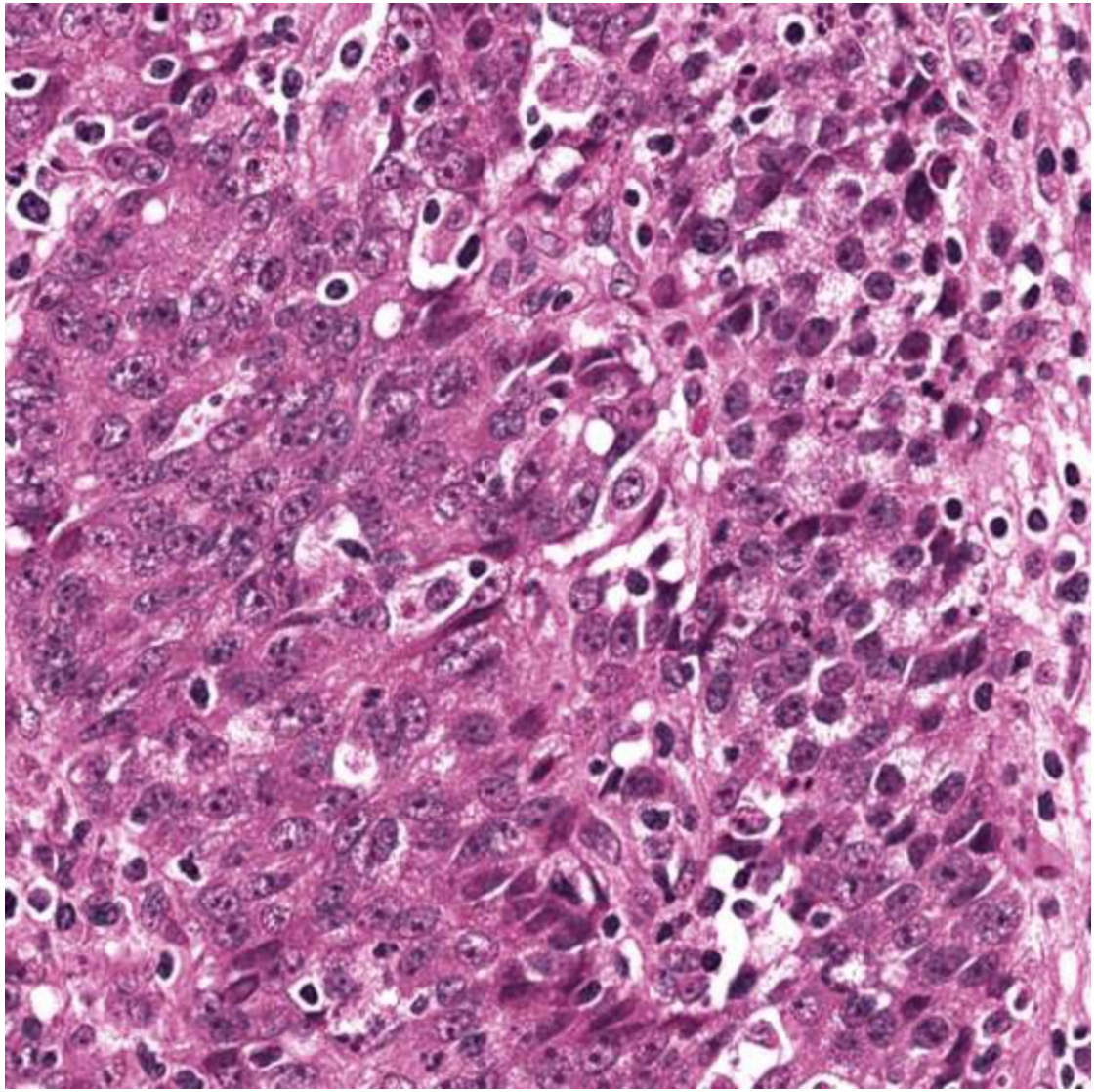
NEP1 Inflammatory Dead Mesenchymal NEP2

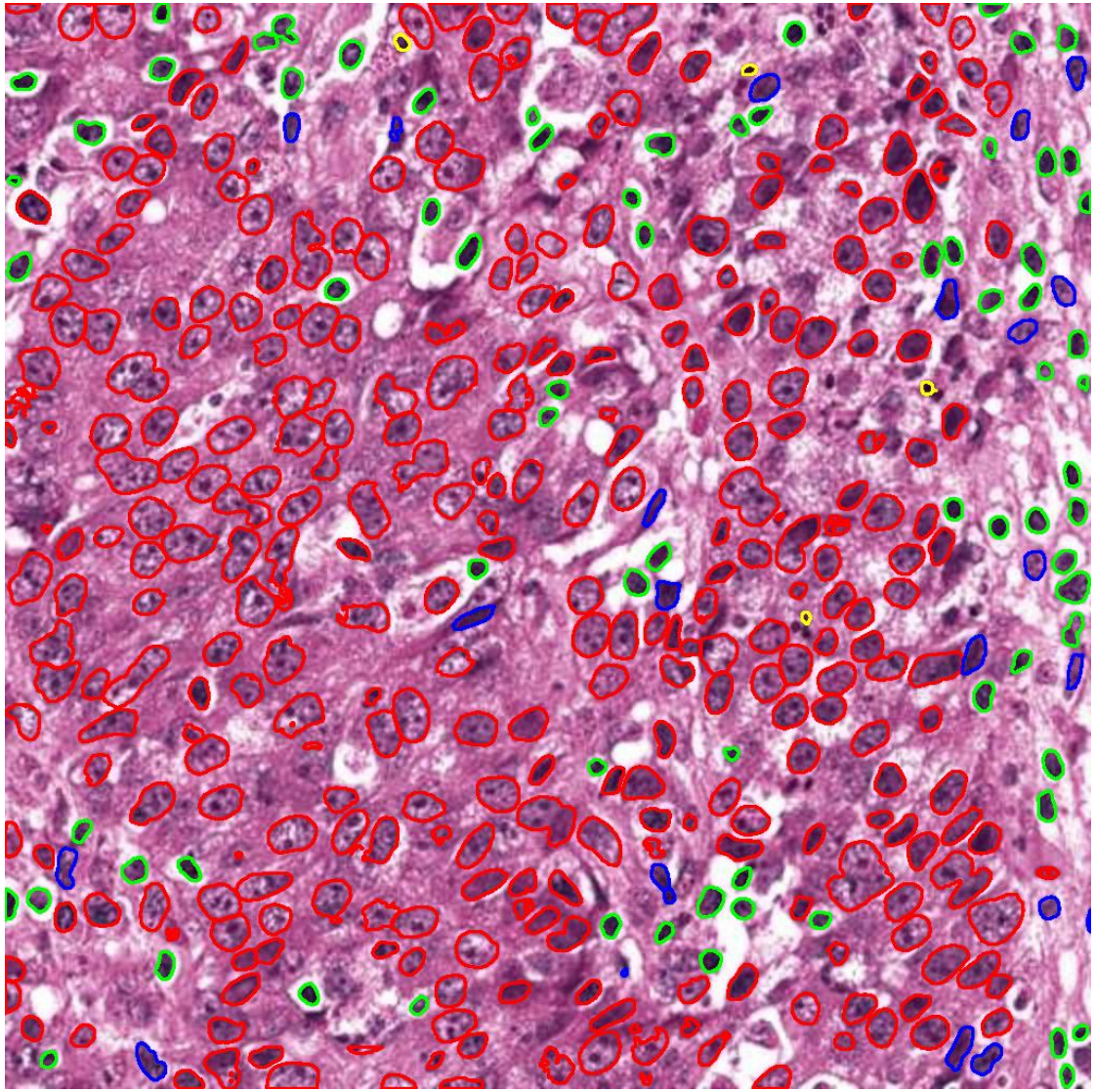




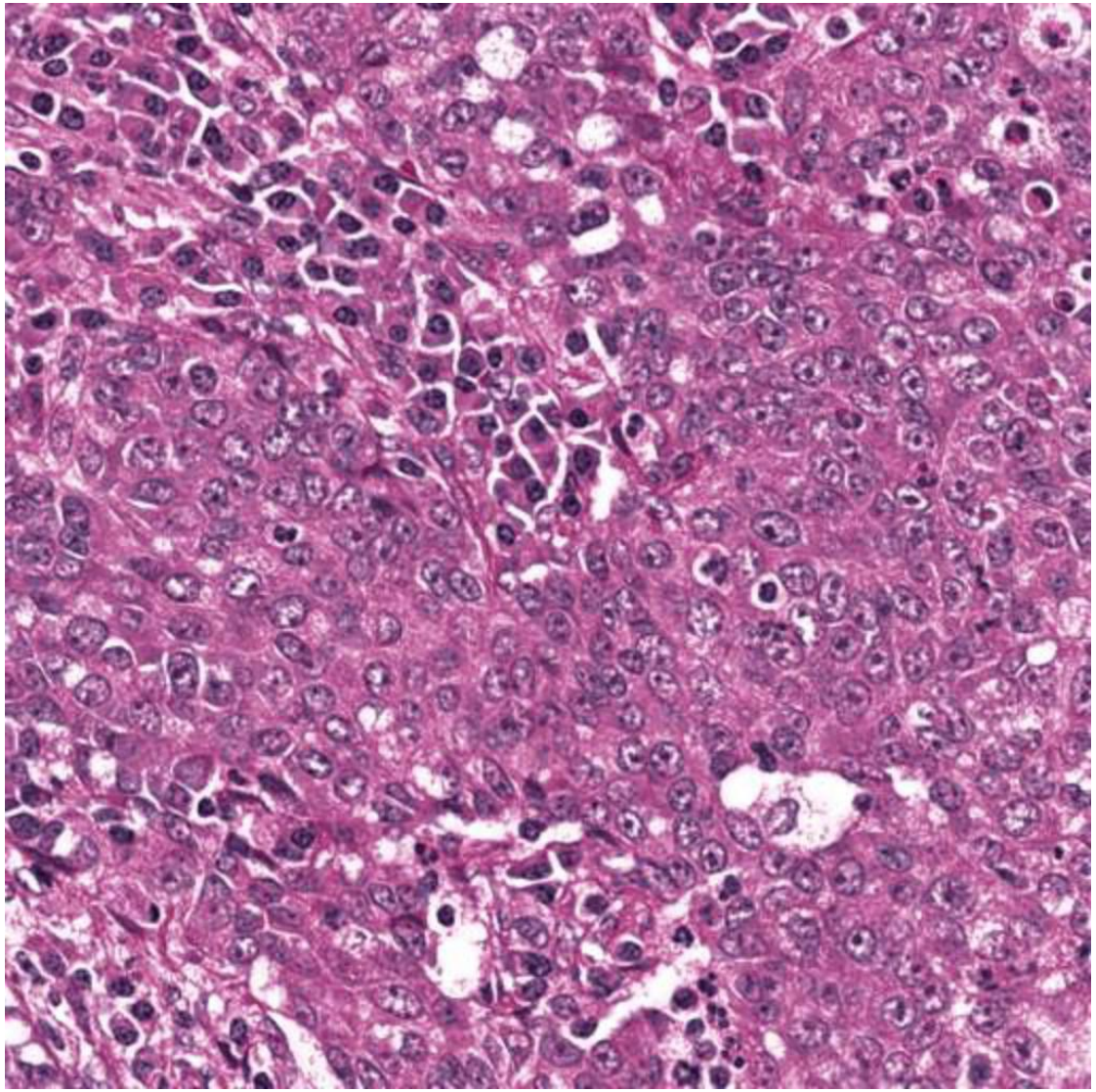
NEP1 Inflammatory Dead Mesenchymal NEP2

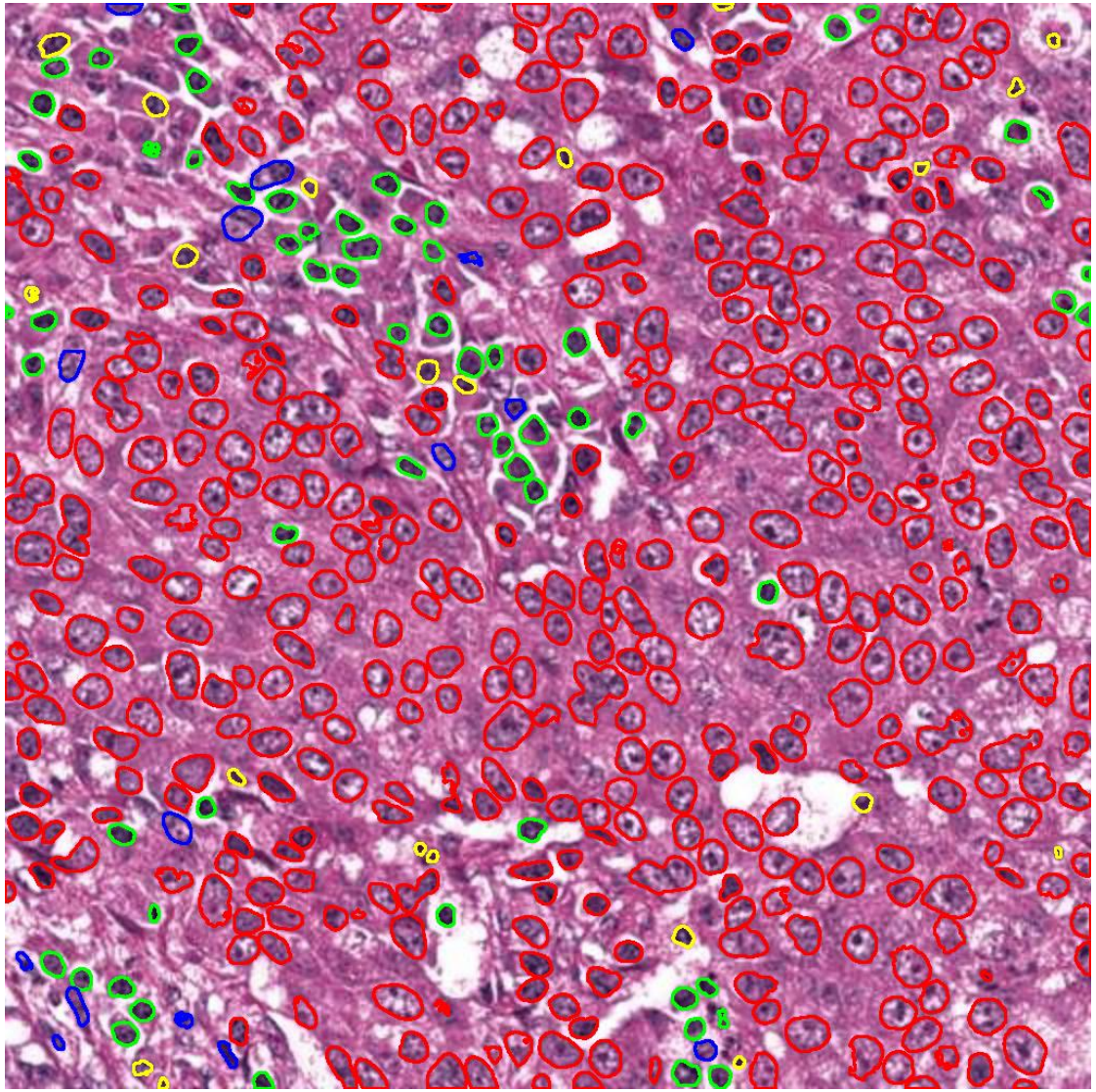

NEP1
Inflammatory
Dead
Mesenchymal
NEP2



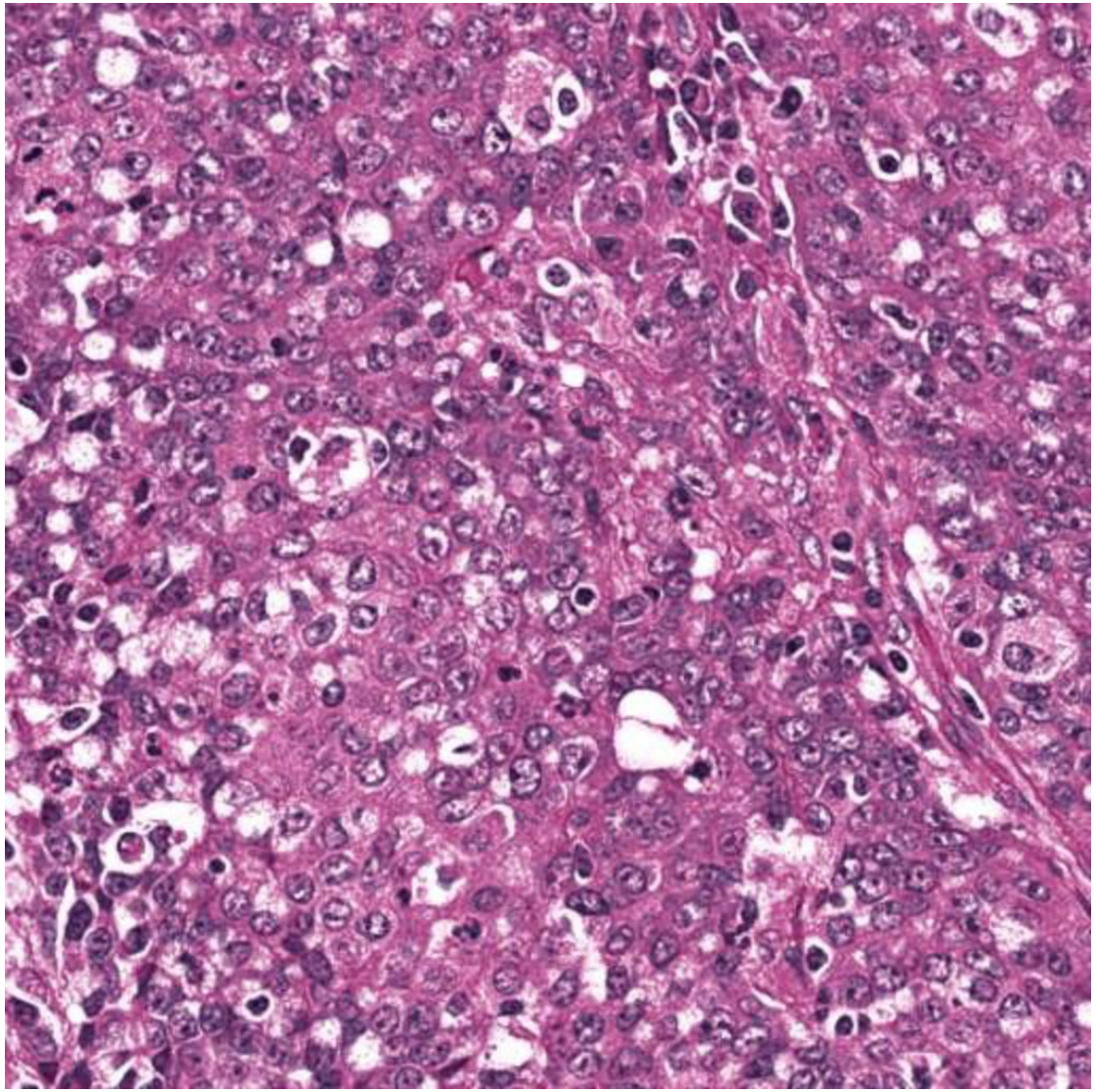


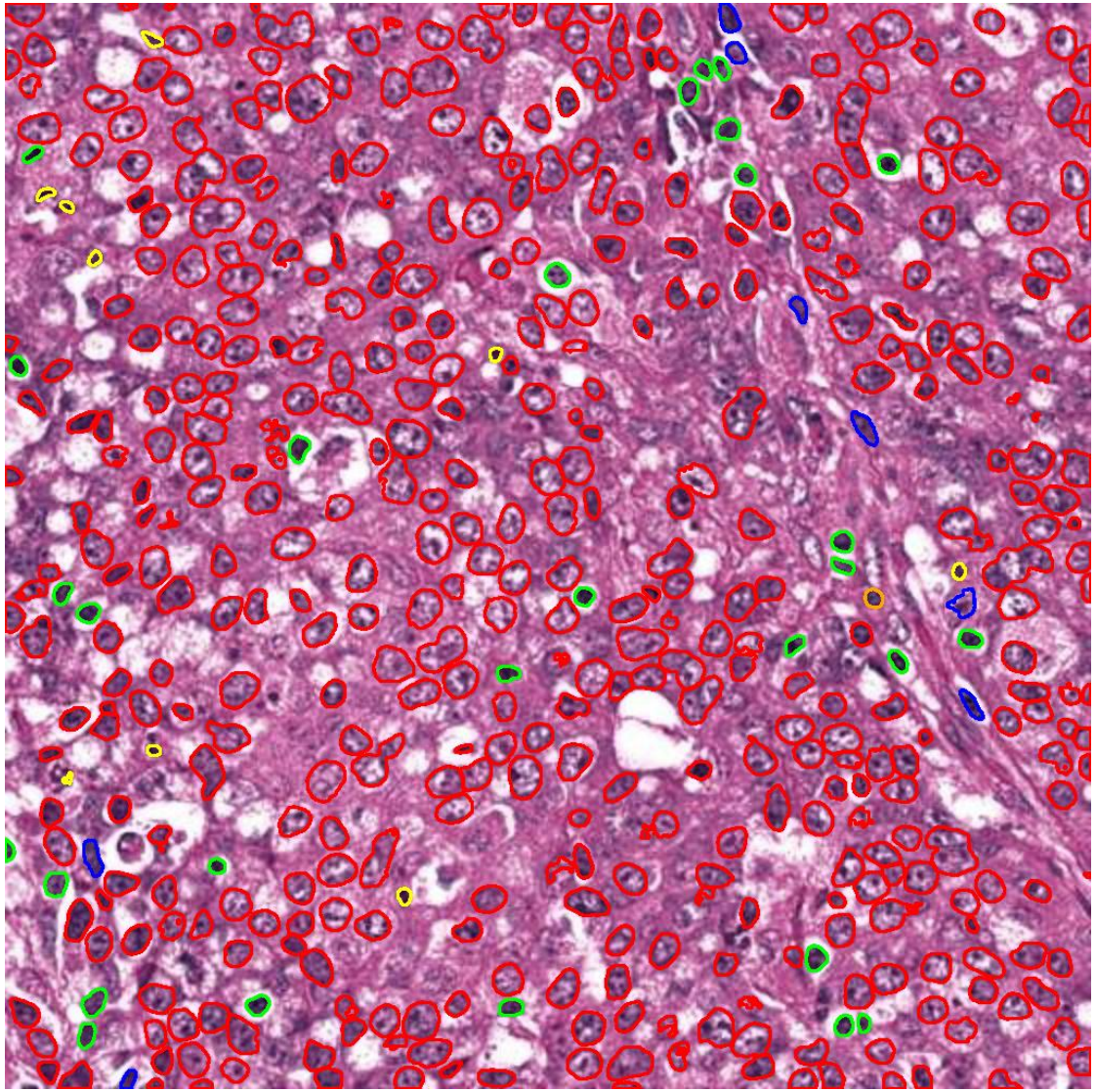
NEP1 Inflammatory Dead Mesenchymal NEP2



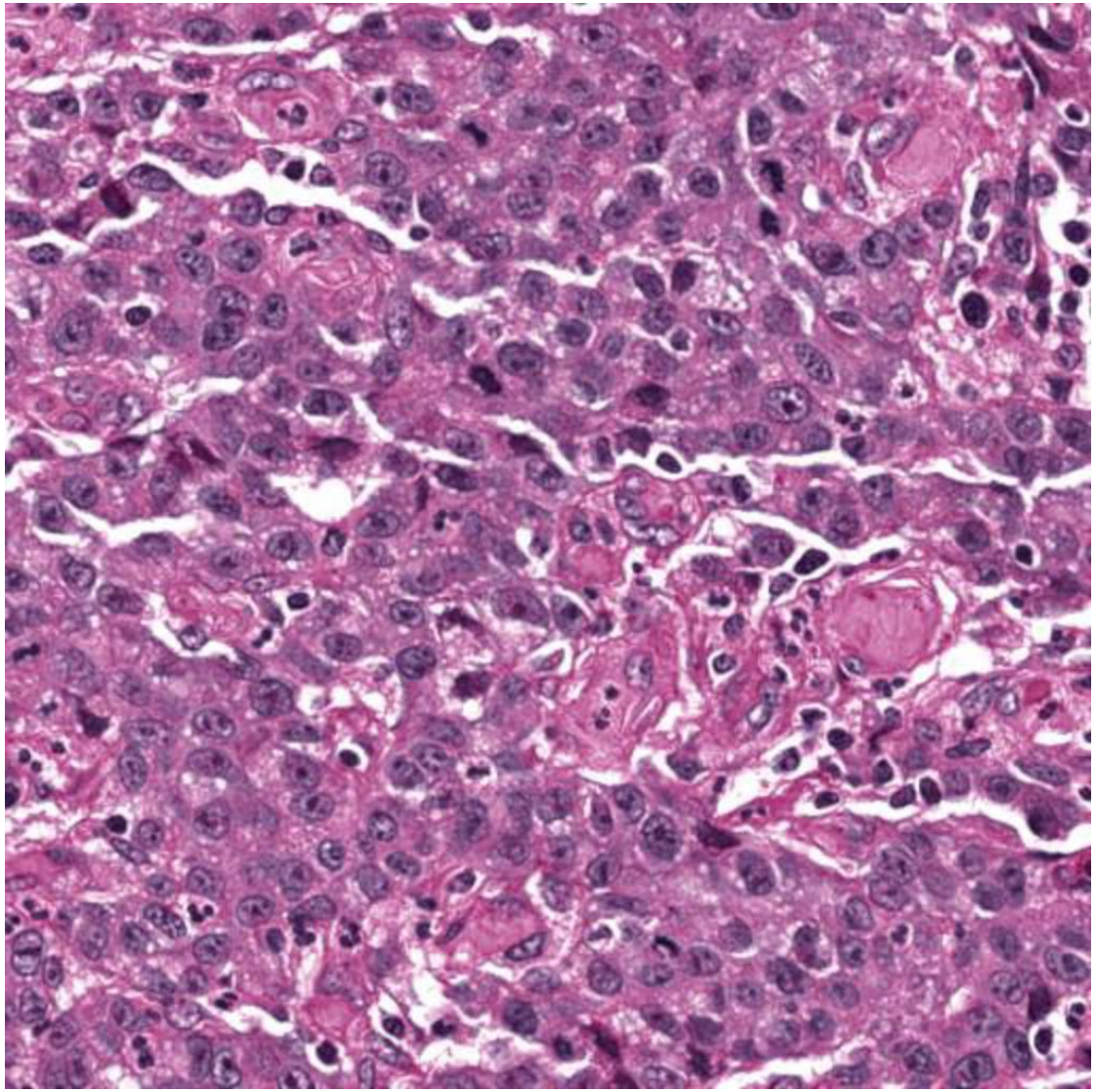


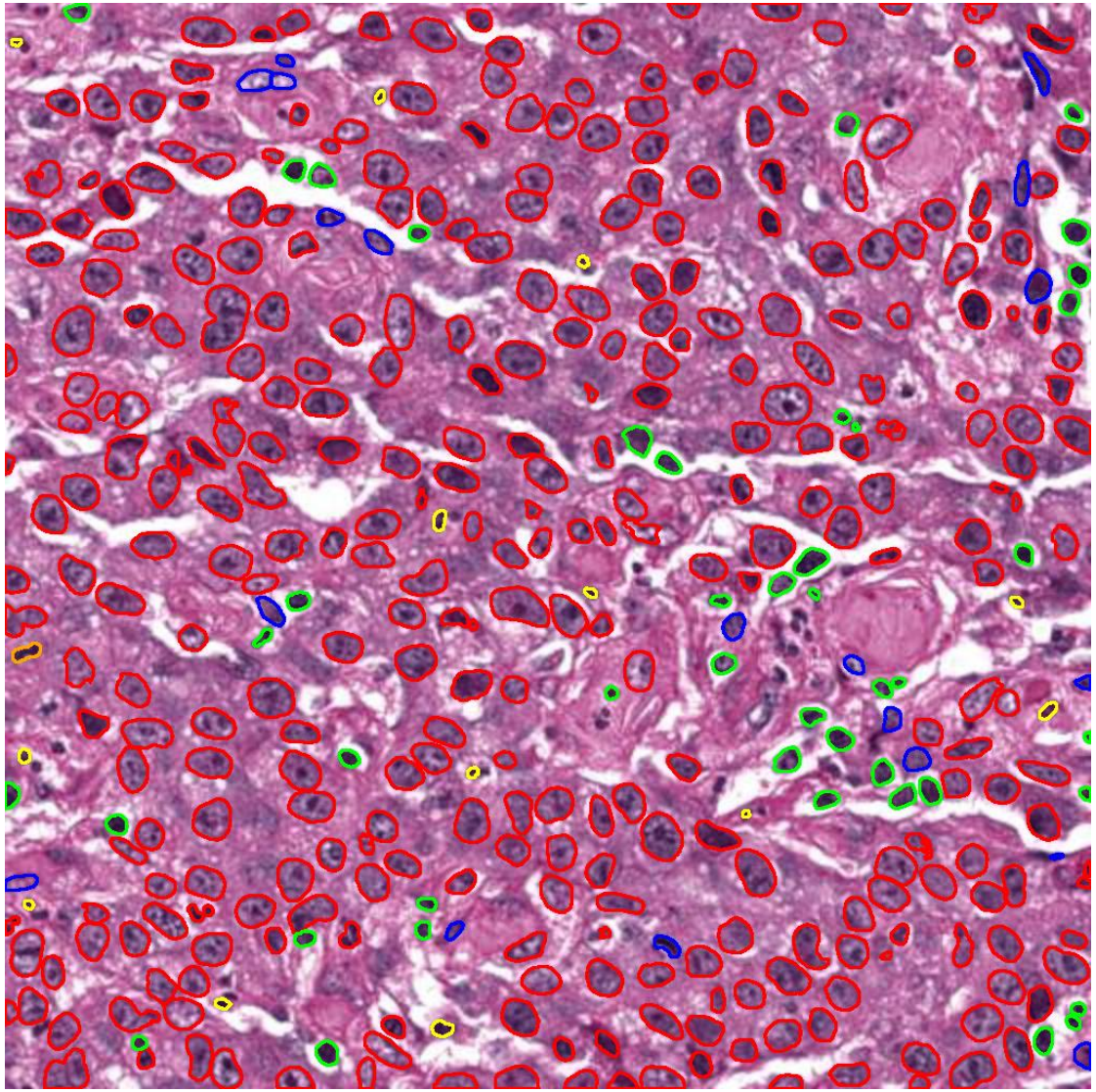
NEP1 Inflammatory Dead Mesenchymal NEP2



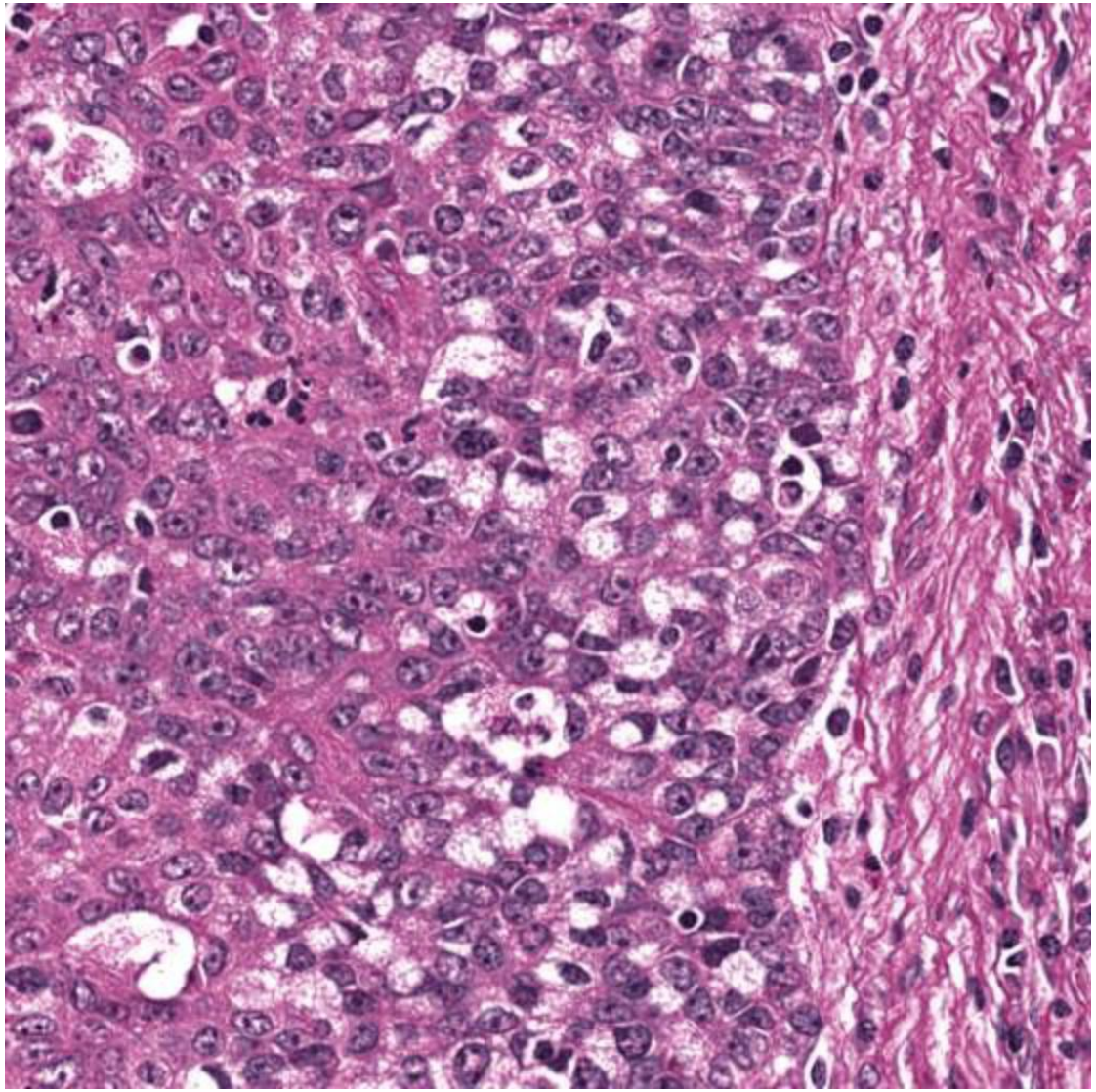


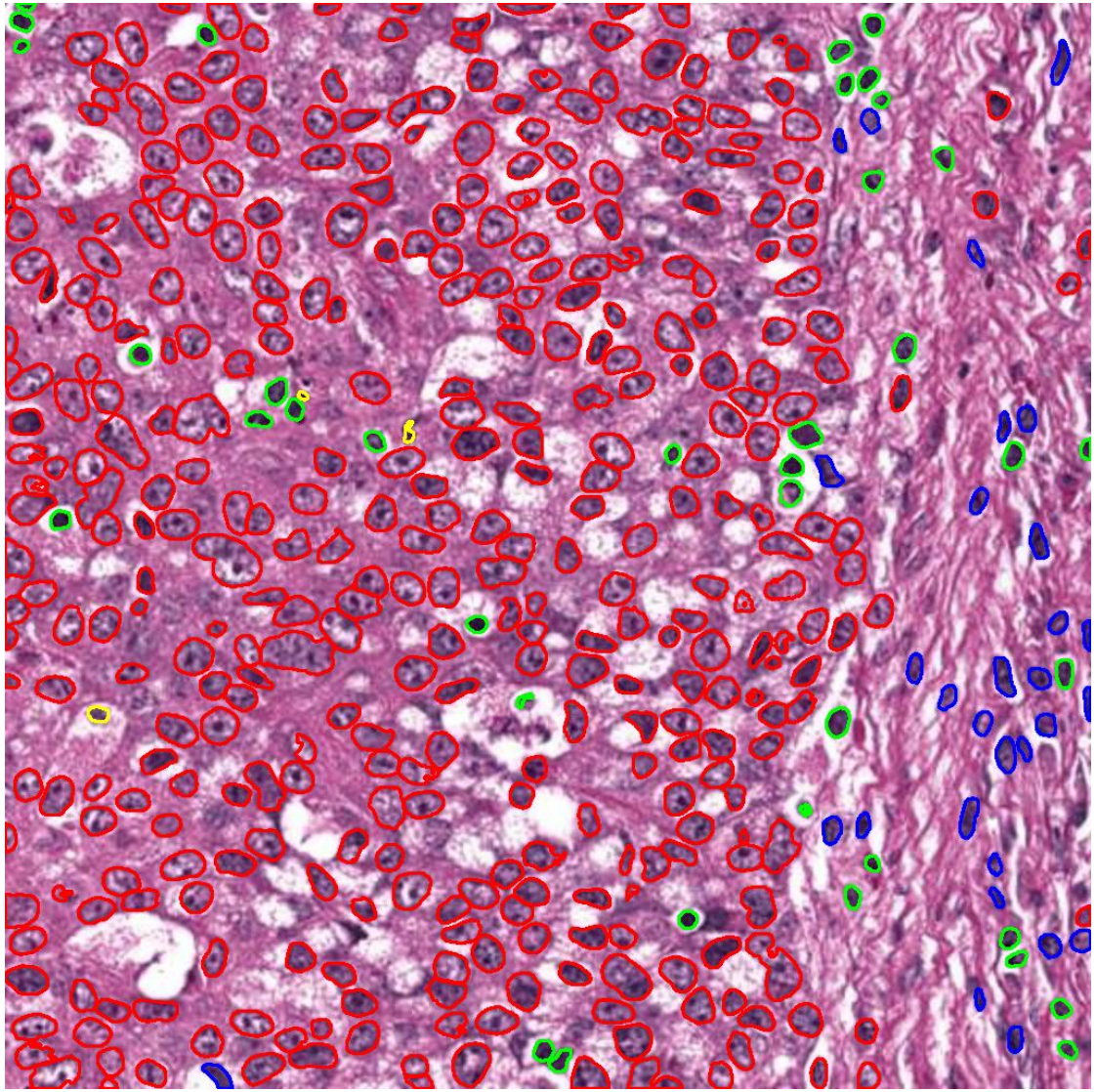
NEP1 Inflammatory Dead Mesenchymal NEP2



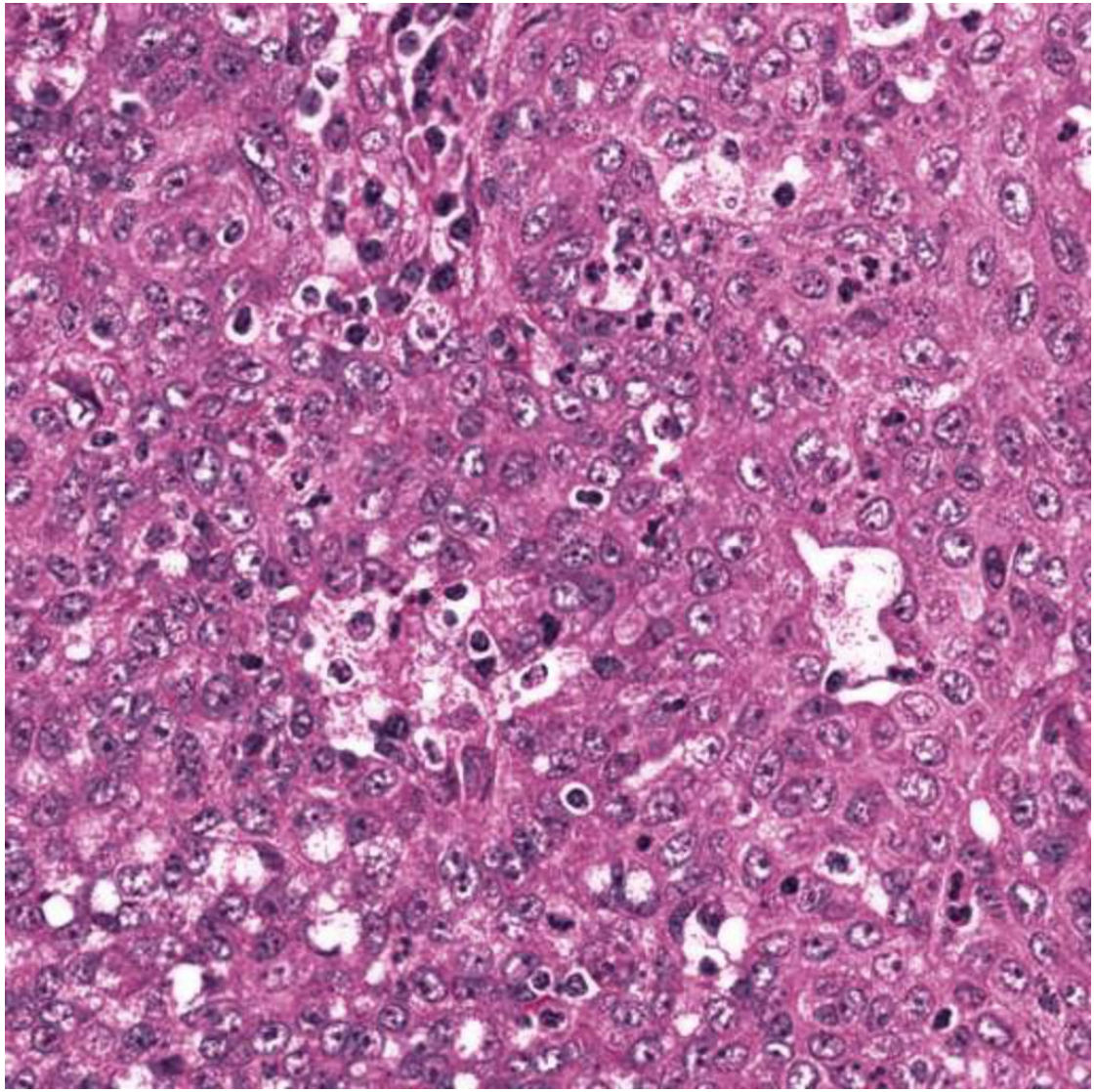


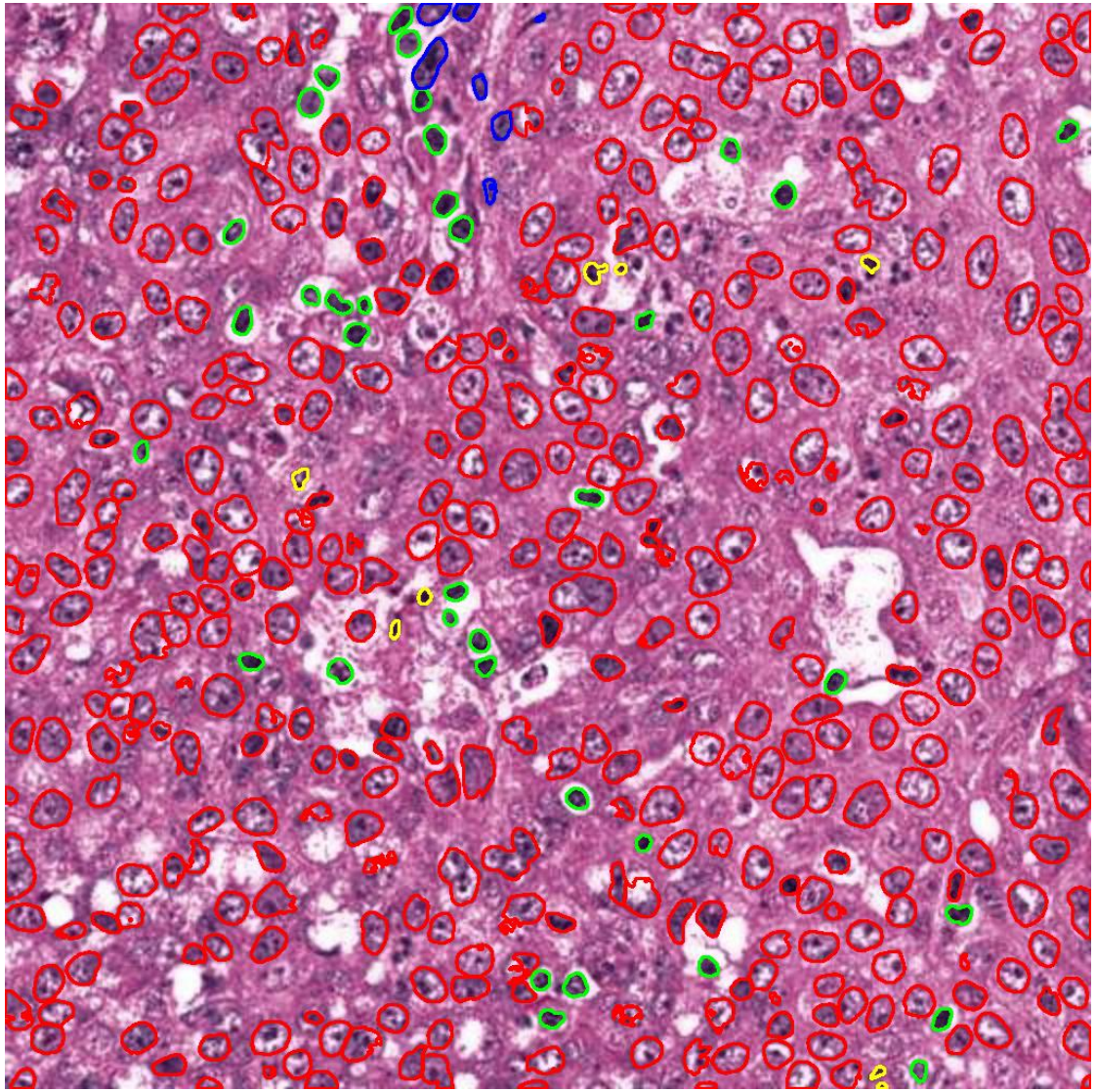
NEP1 Inflammatory Dead Mesenchymal NEP2



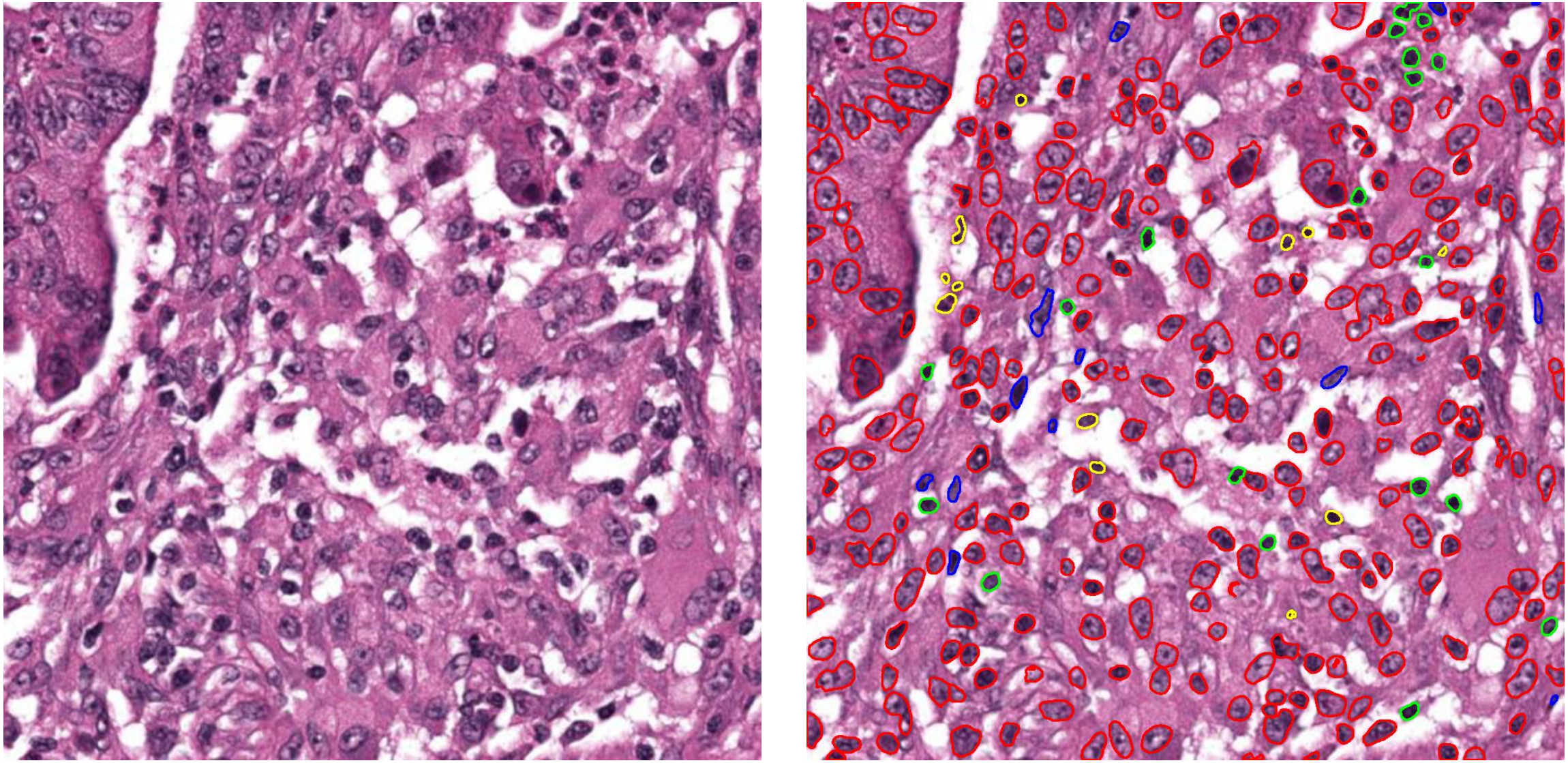


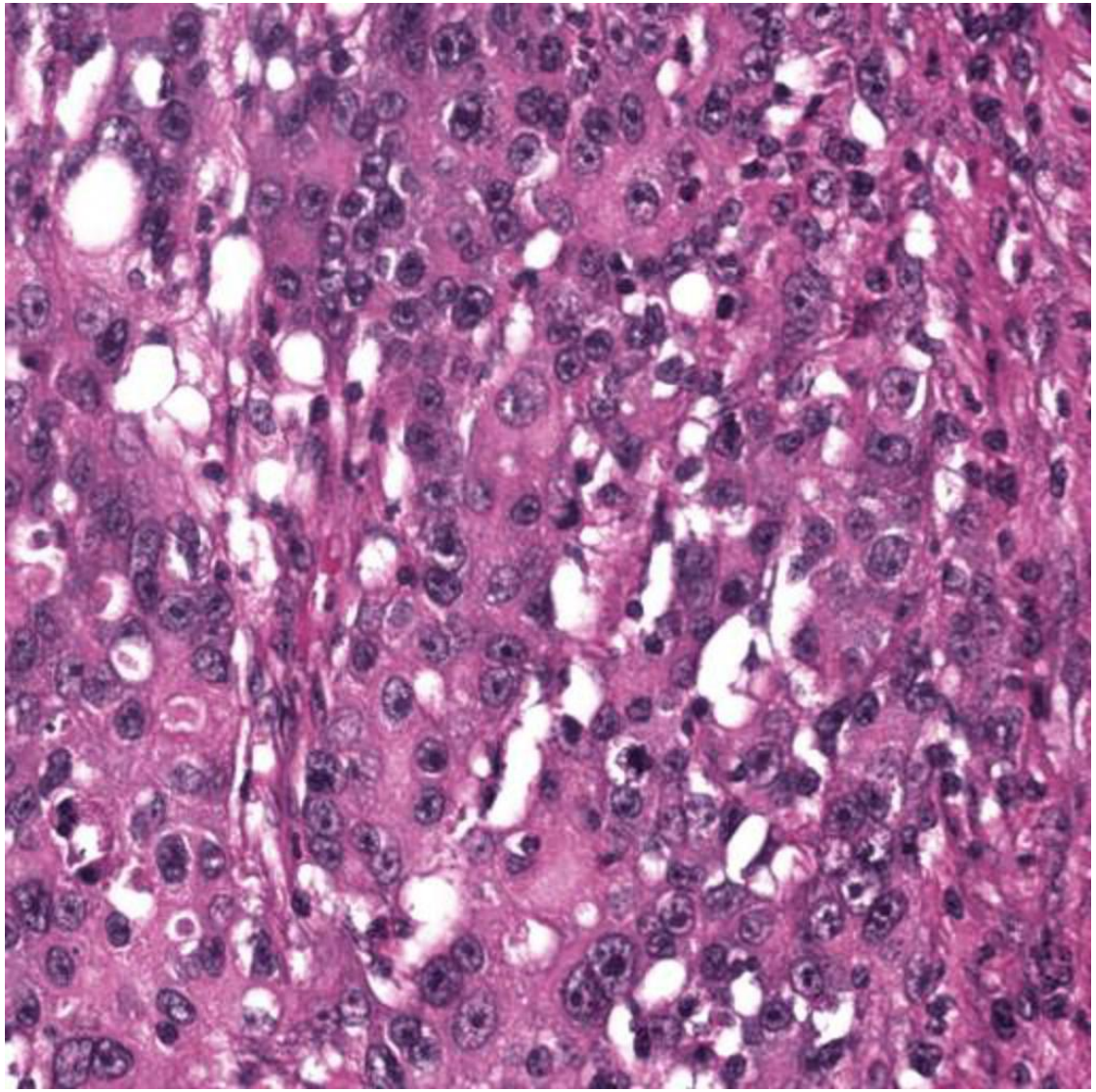
NEP1 Inflammatory Dead Mesenchymal NEP2

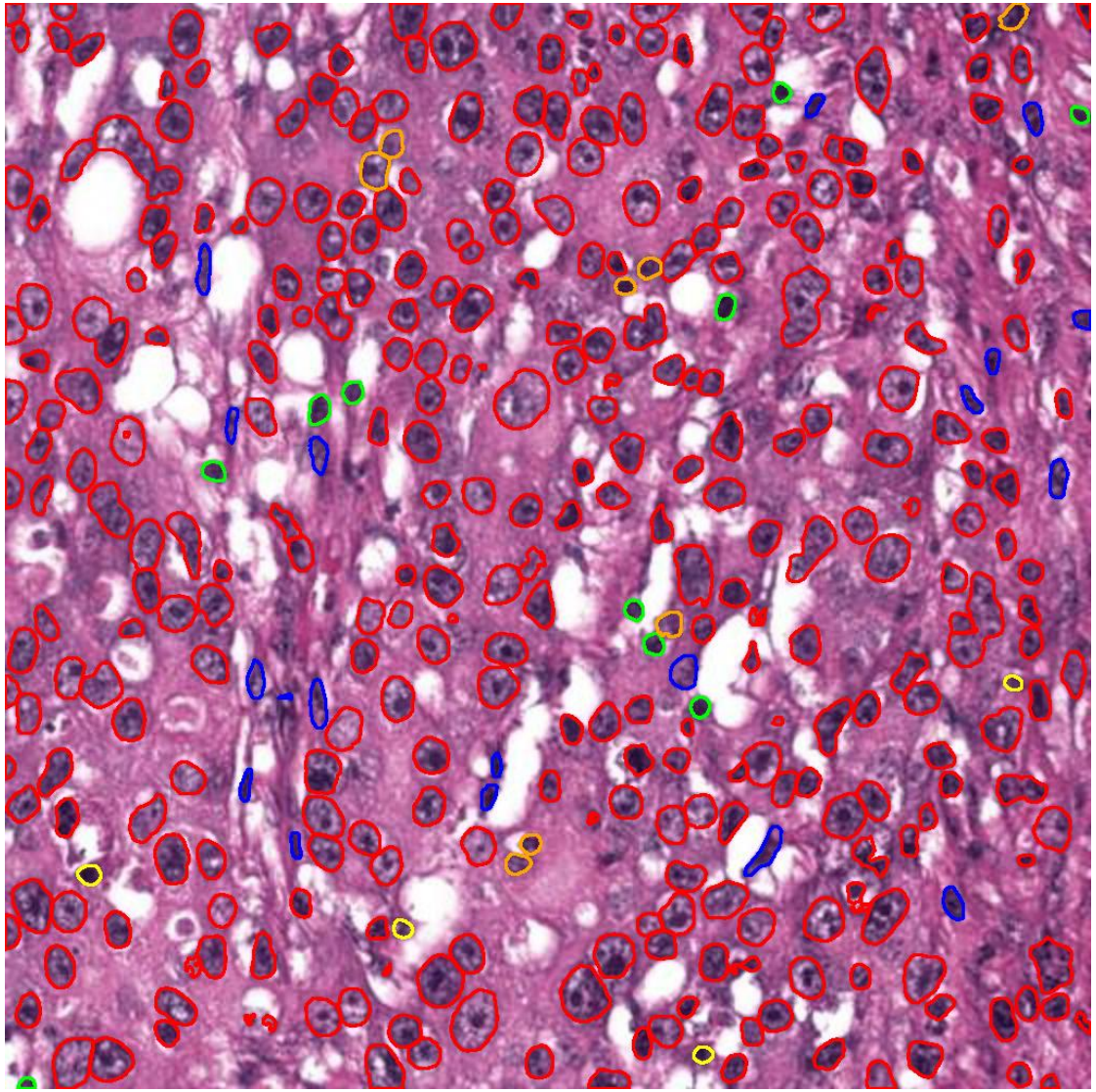




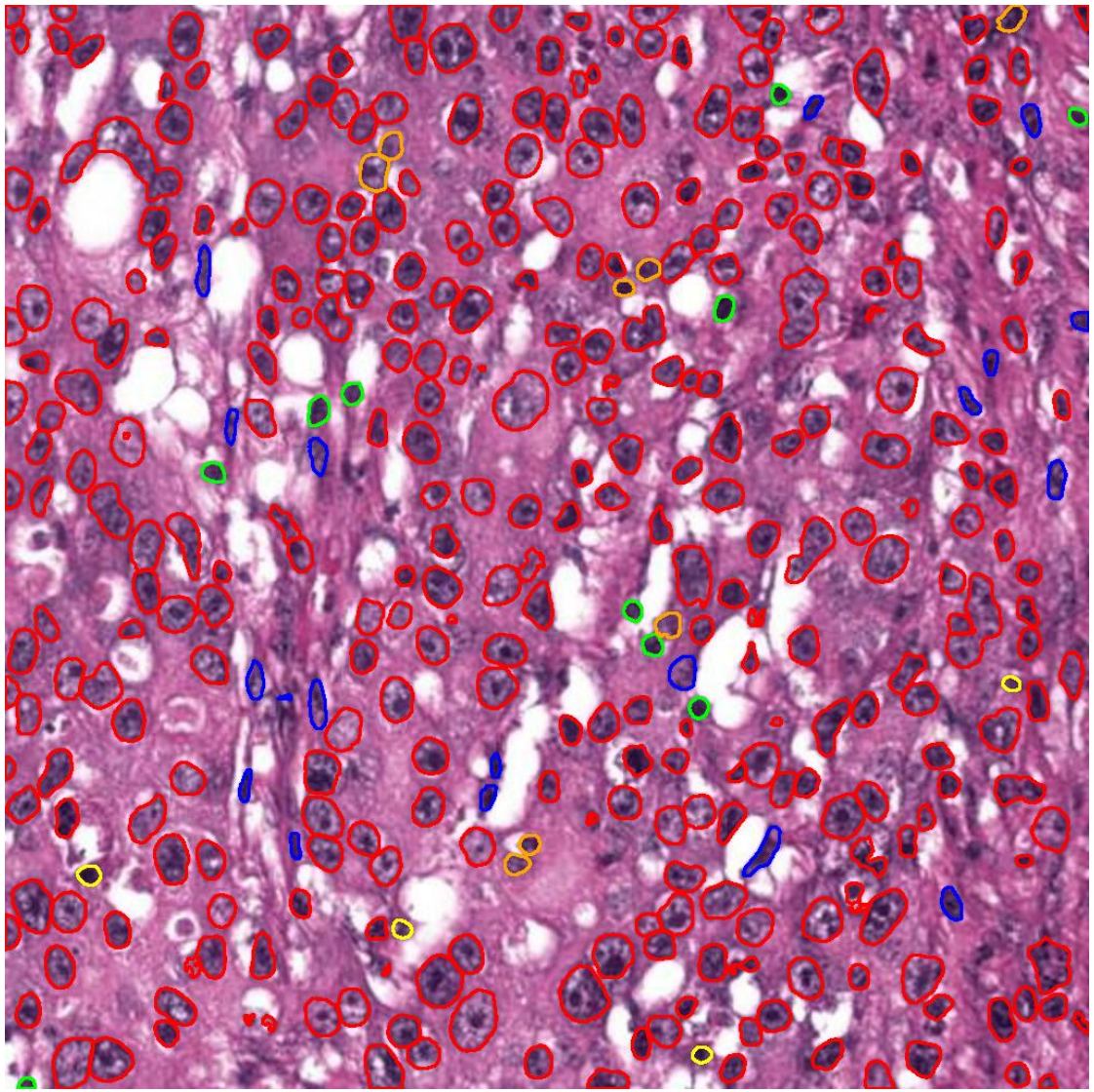
NEP1 Inflammatory Dead Mesenchymal NEP2

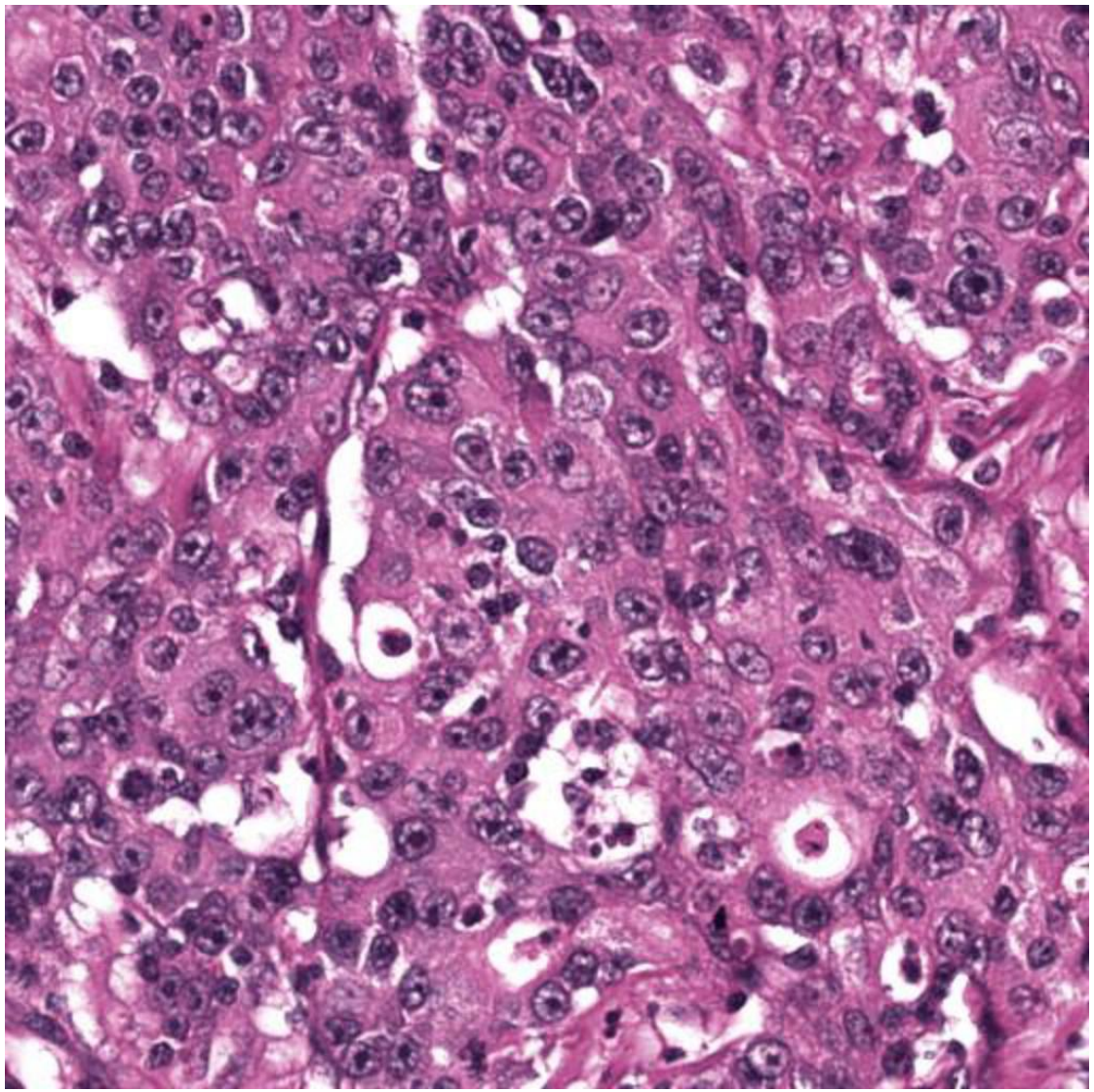

NEP1
Inflammatory
Dead
Mesenchymal
NEP2



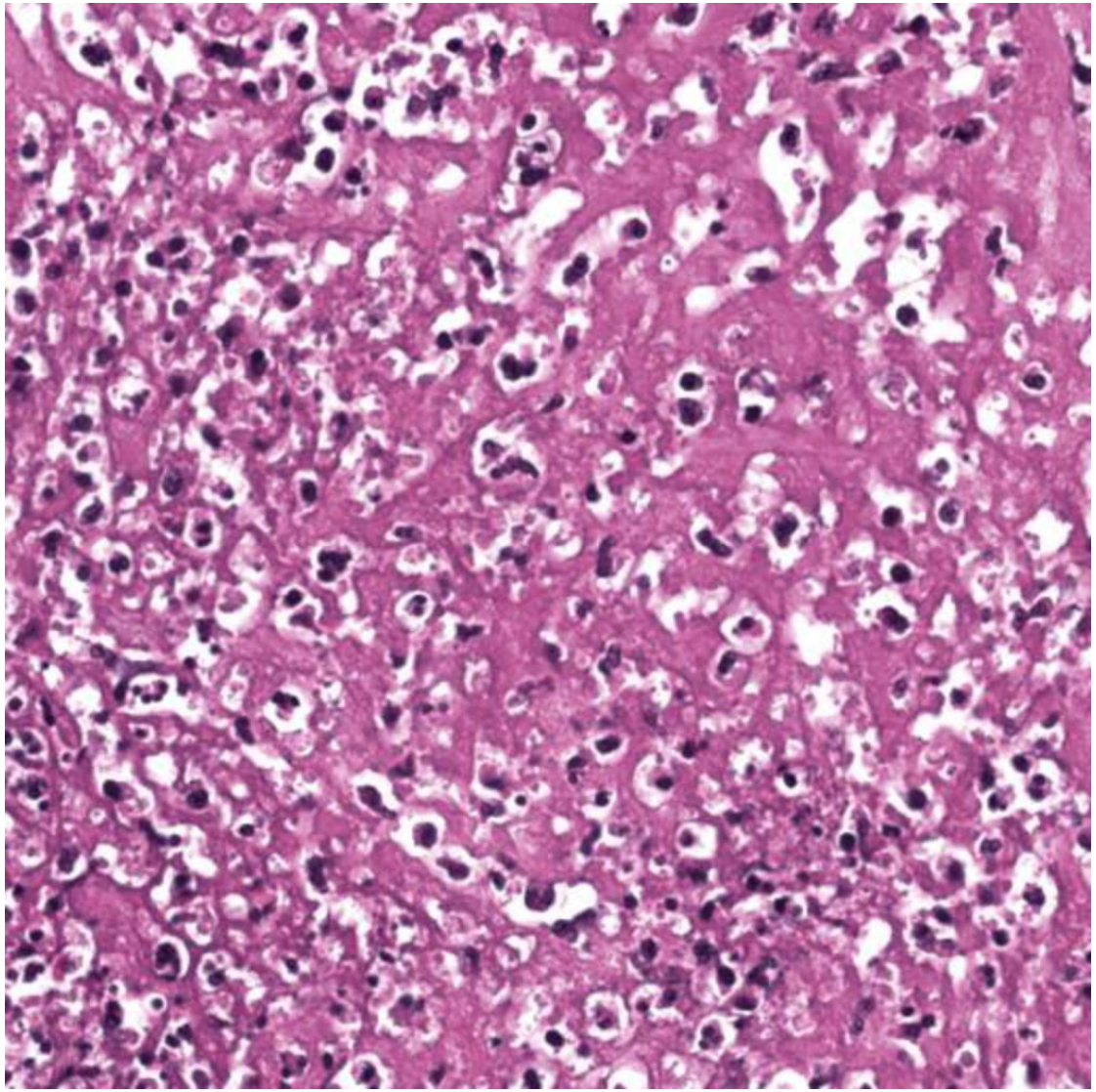


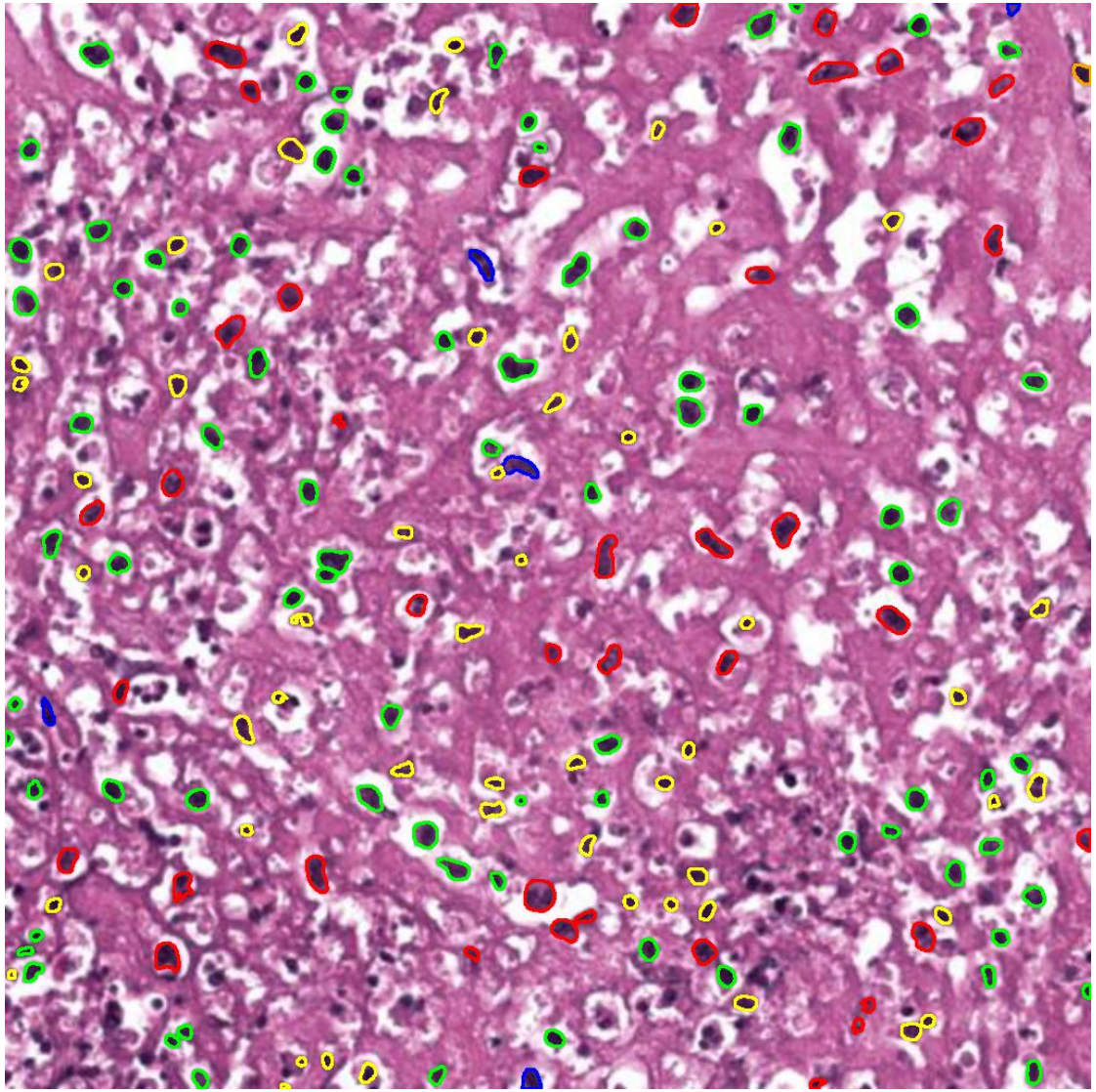
NEP1 Inflammatory Dead Mesenchymal NEP2



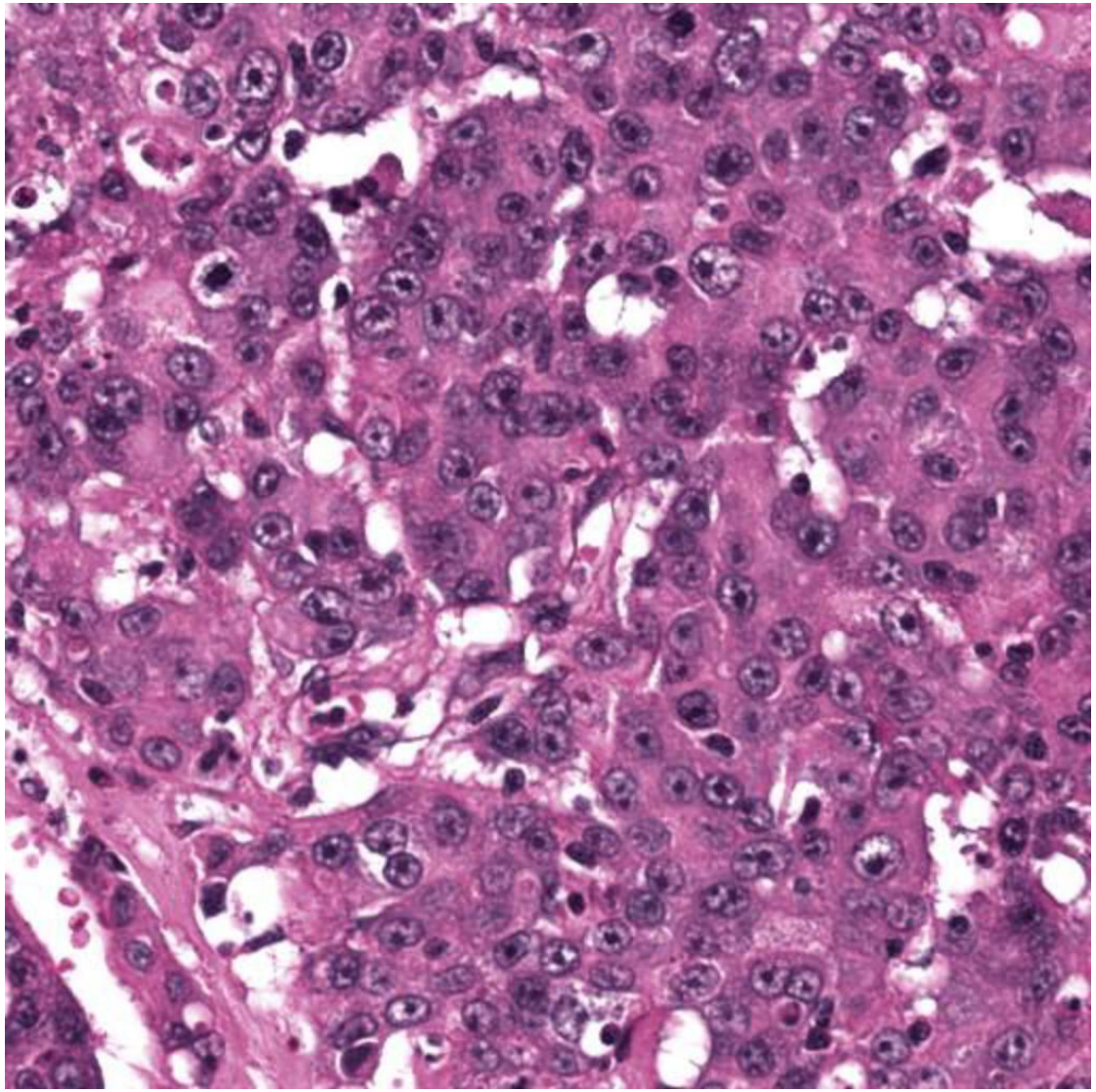


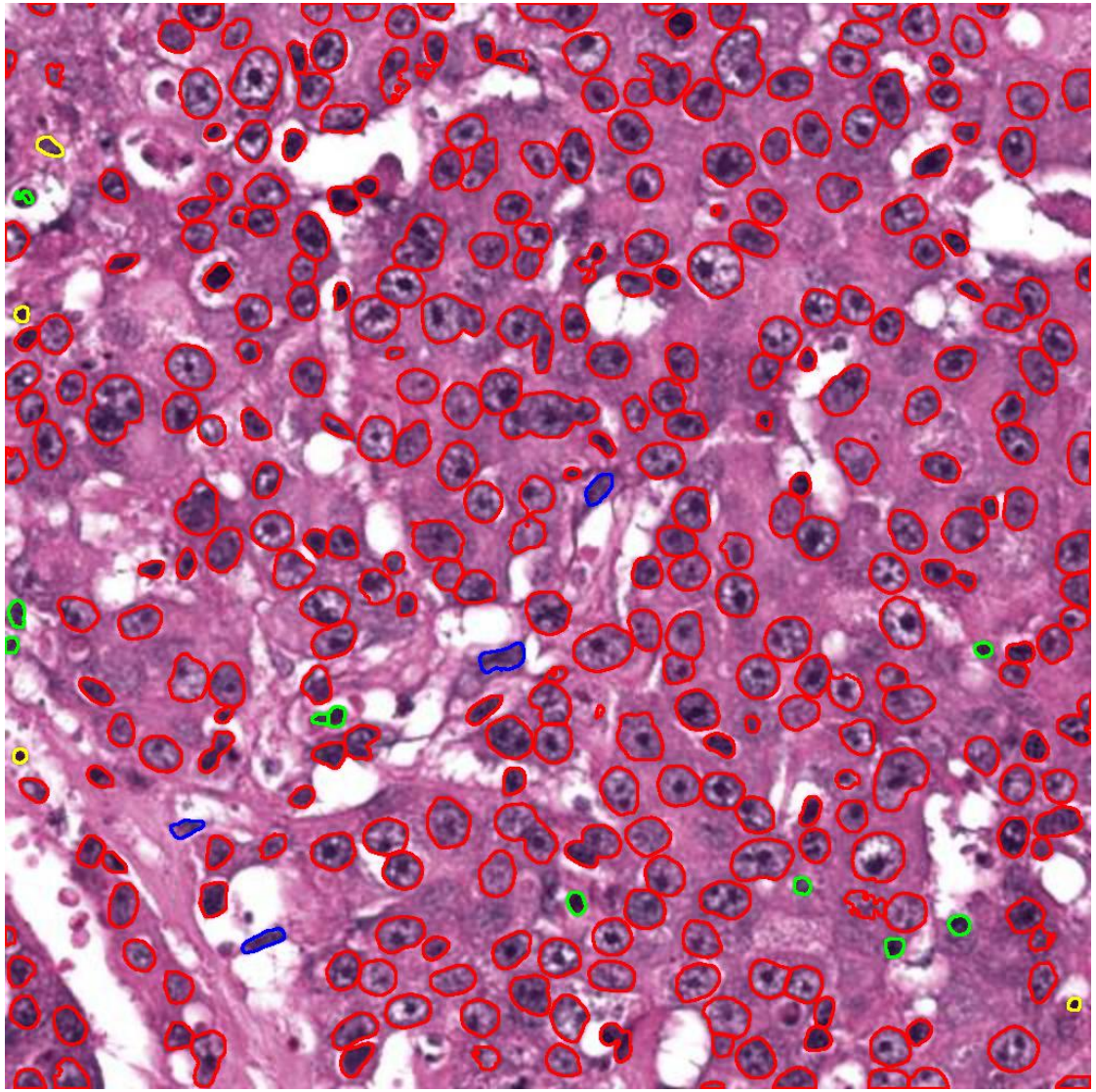
NEP1 Inflammatory Dead Mesenchymal NEP2



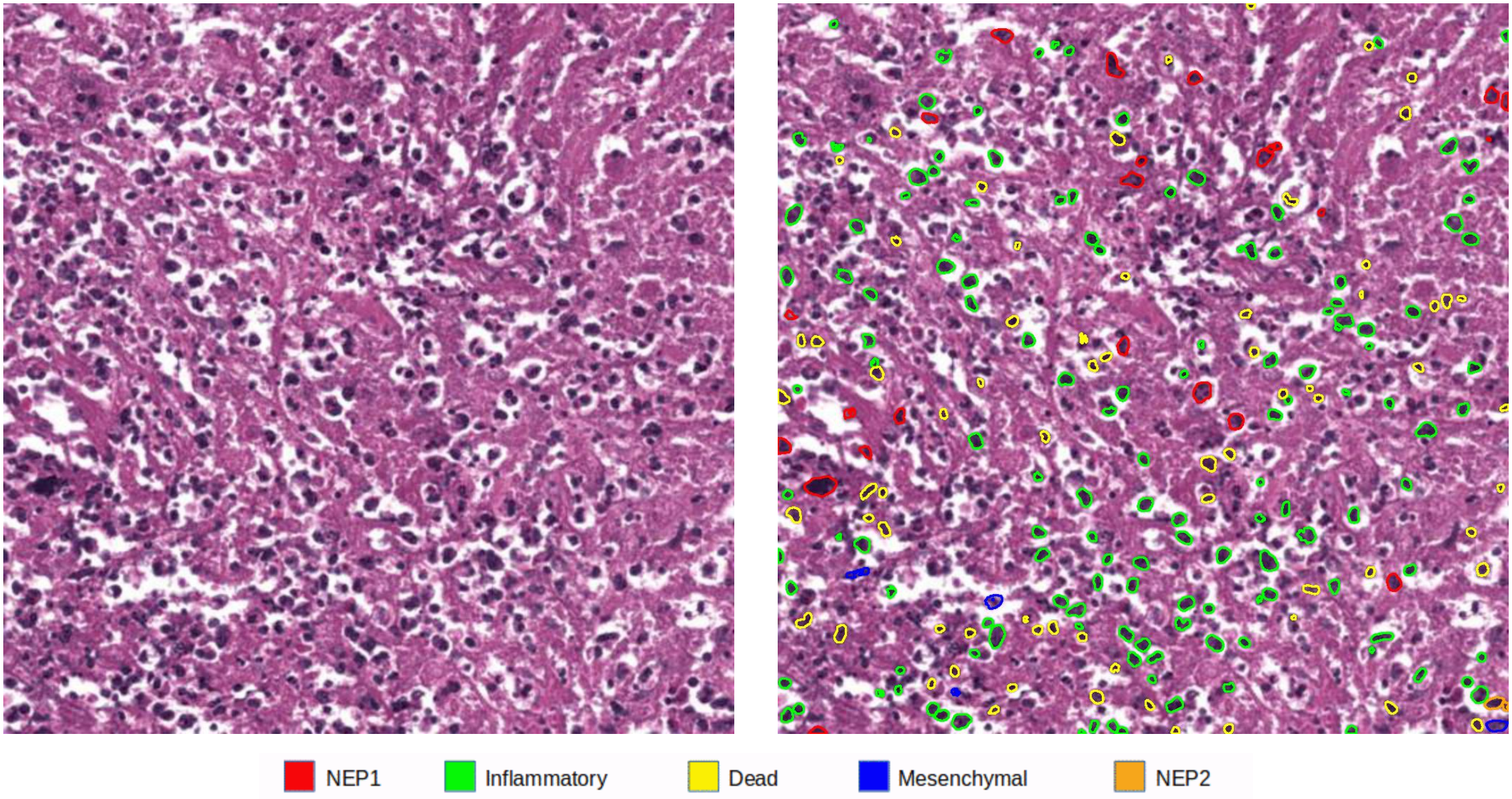


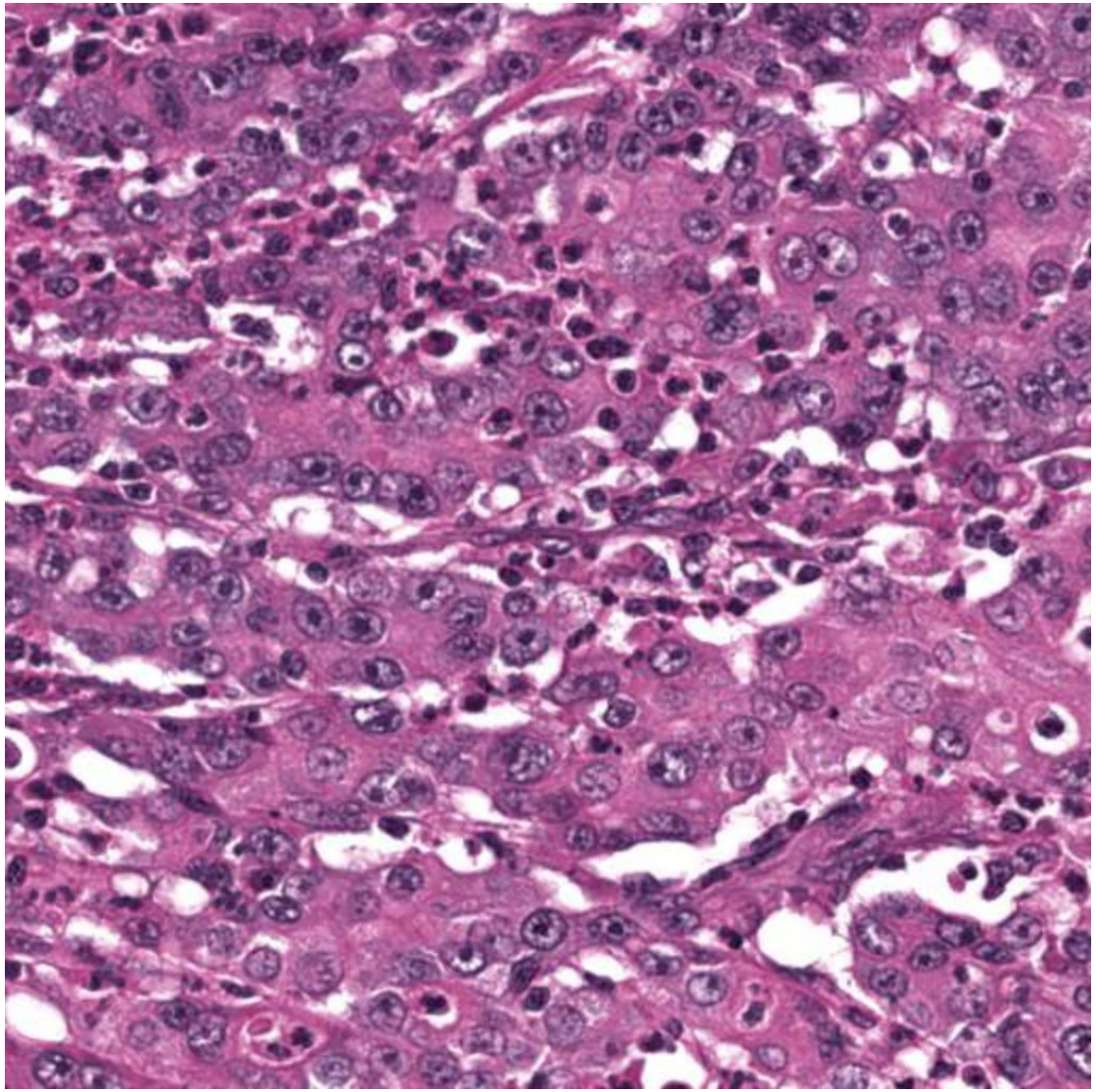
NEP1 Inflammatory Dead Mesenchymal NEP2

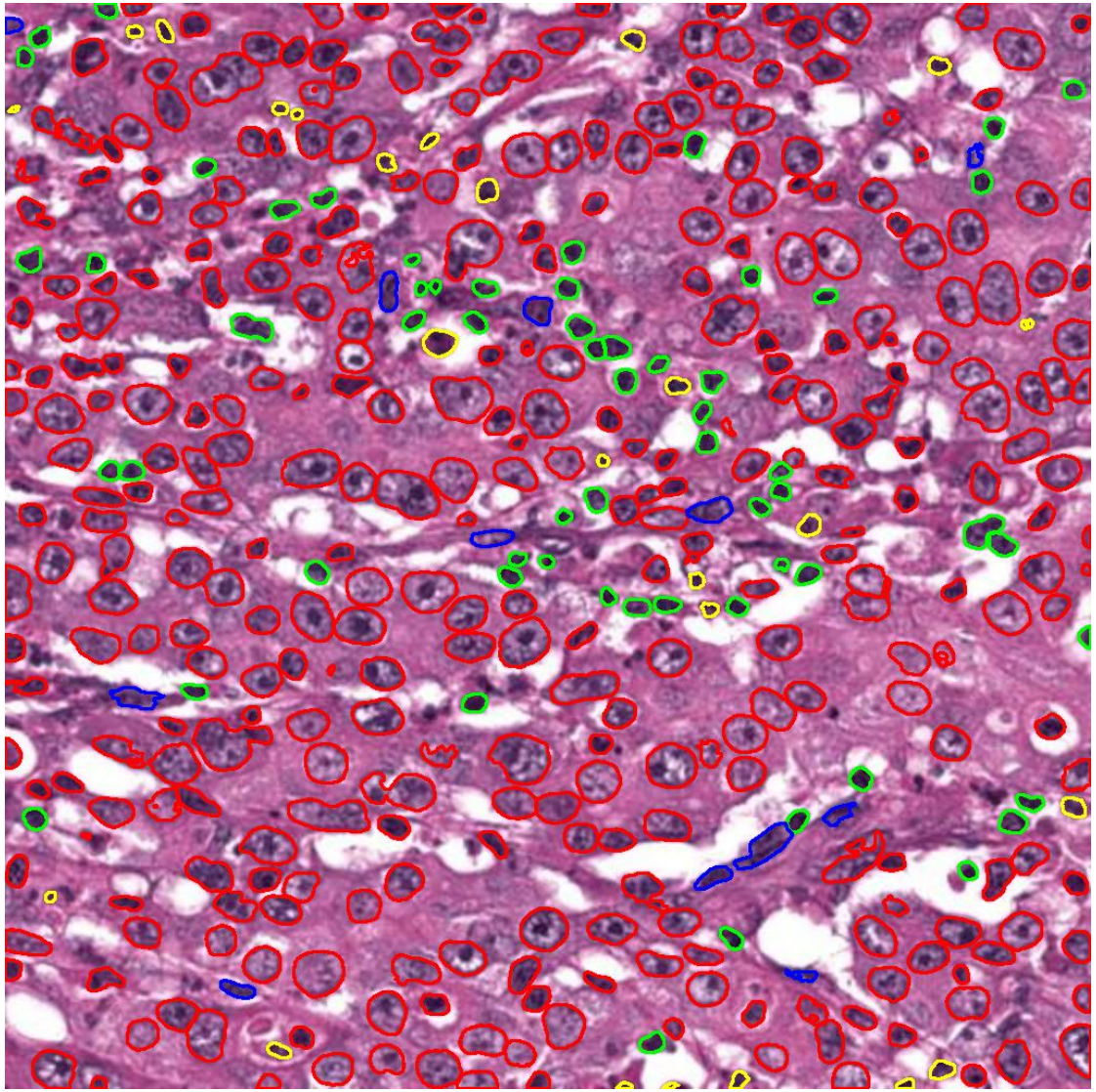




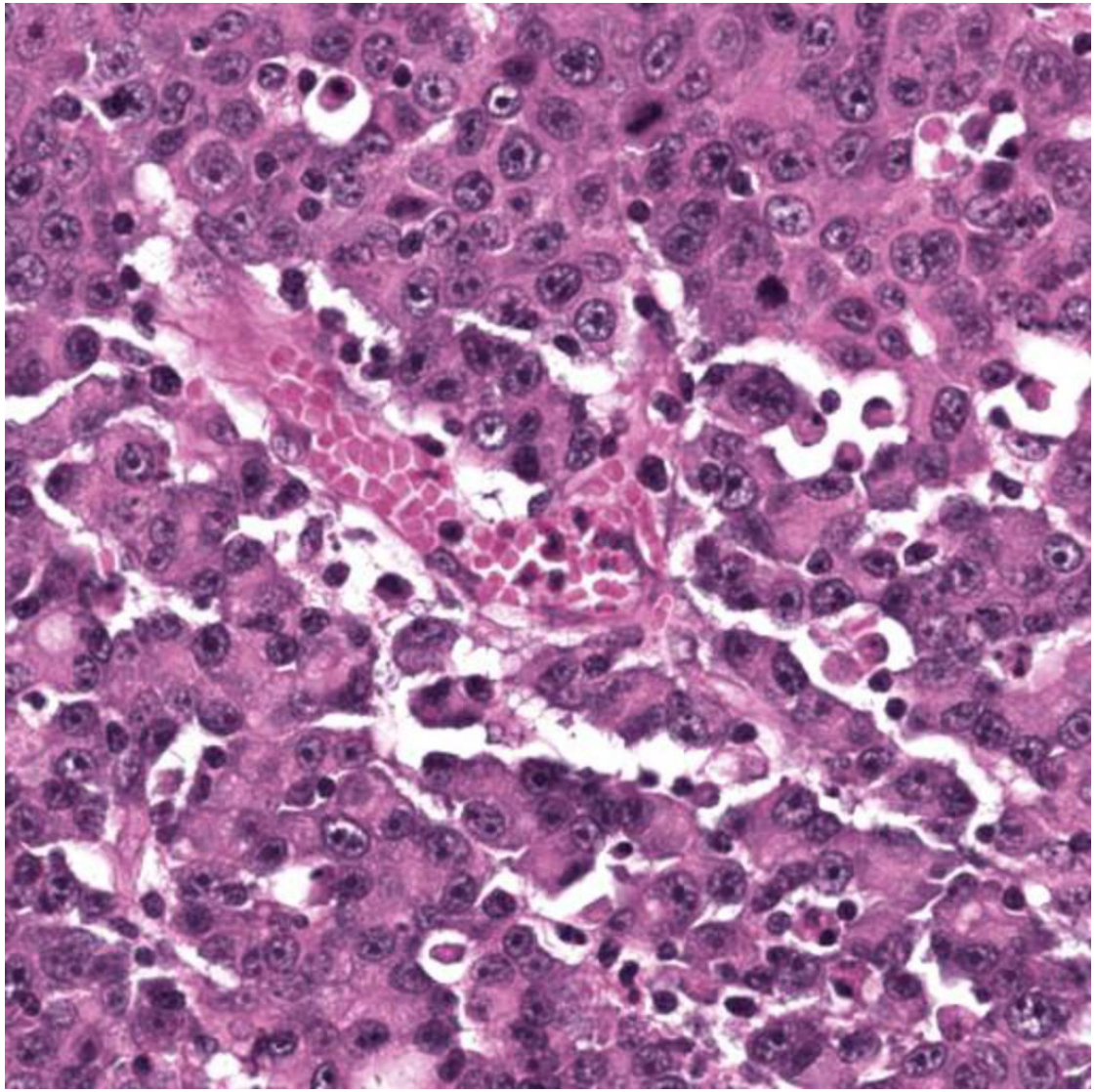
NEP1 Inflammatory Dead Mesenchymal NEP2

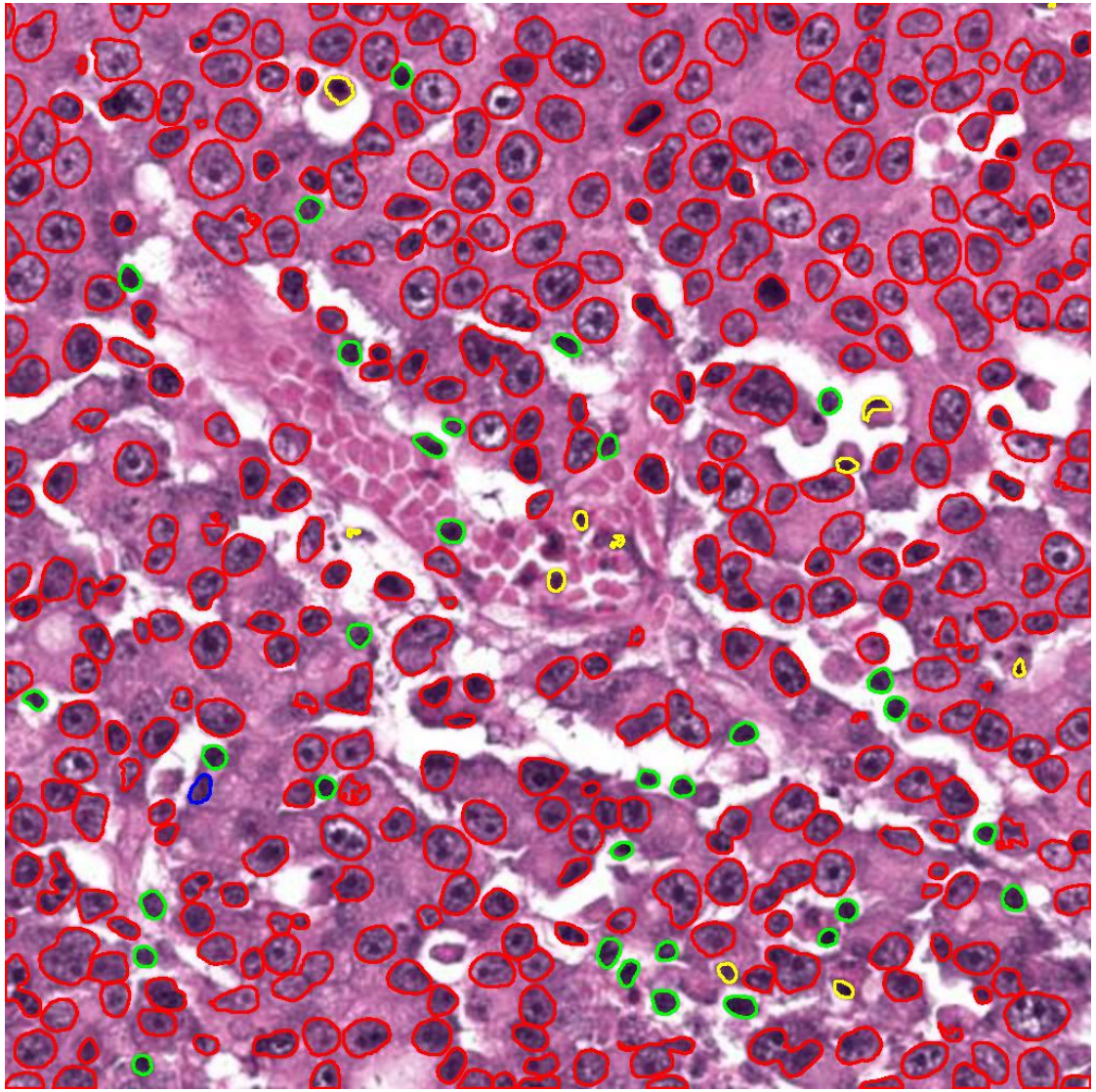

NEP1
Inflammatory
Dead
Mesenchymal
NEP2



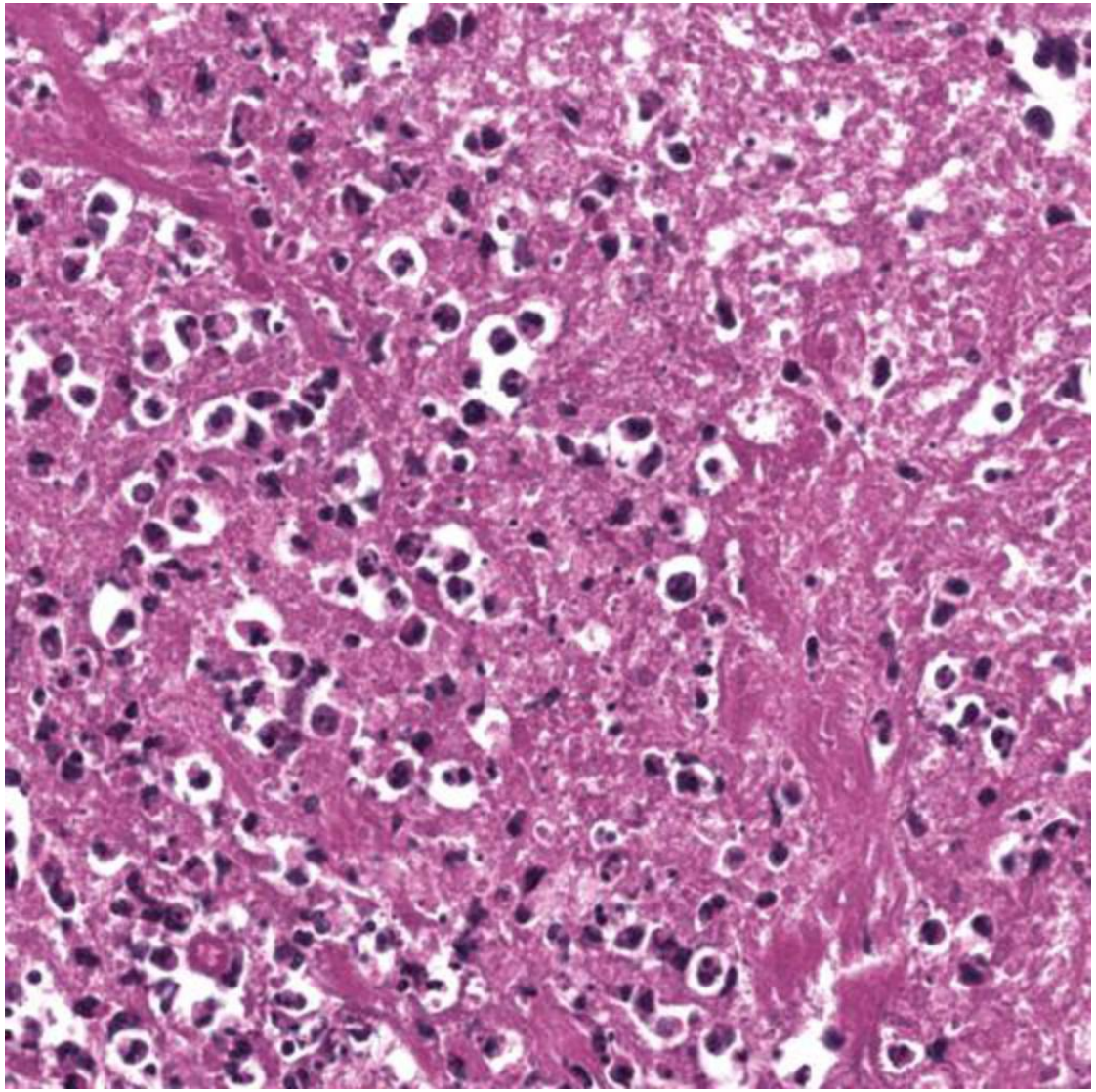


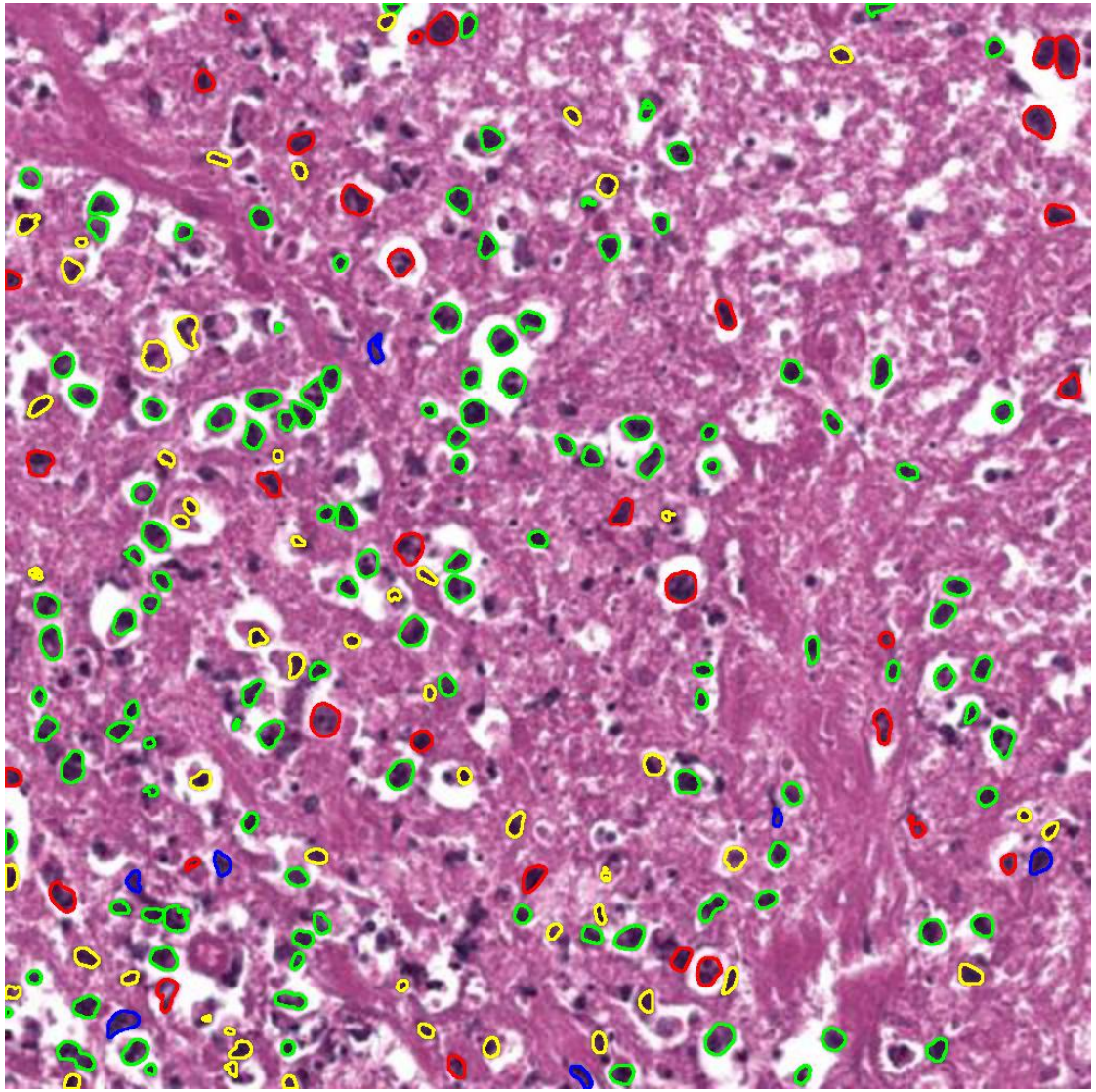
NEP1 Inflammatory Dead Mesenchymal NEP2



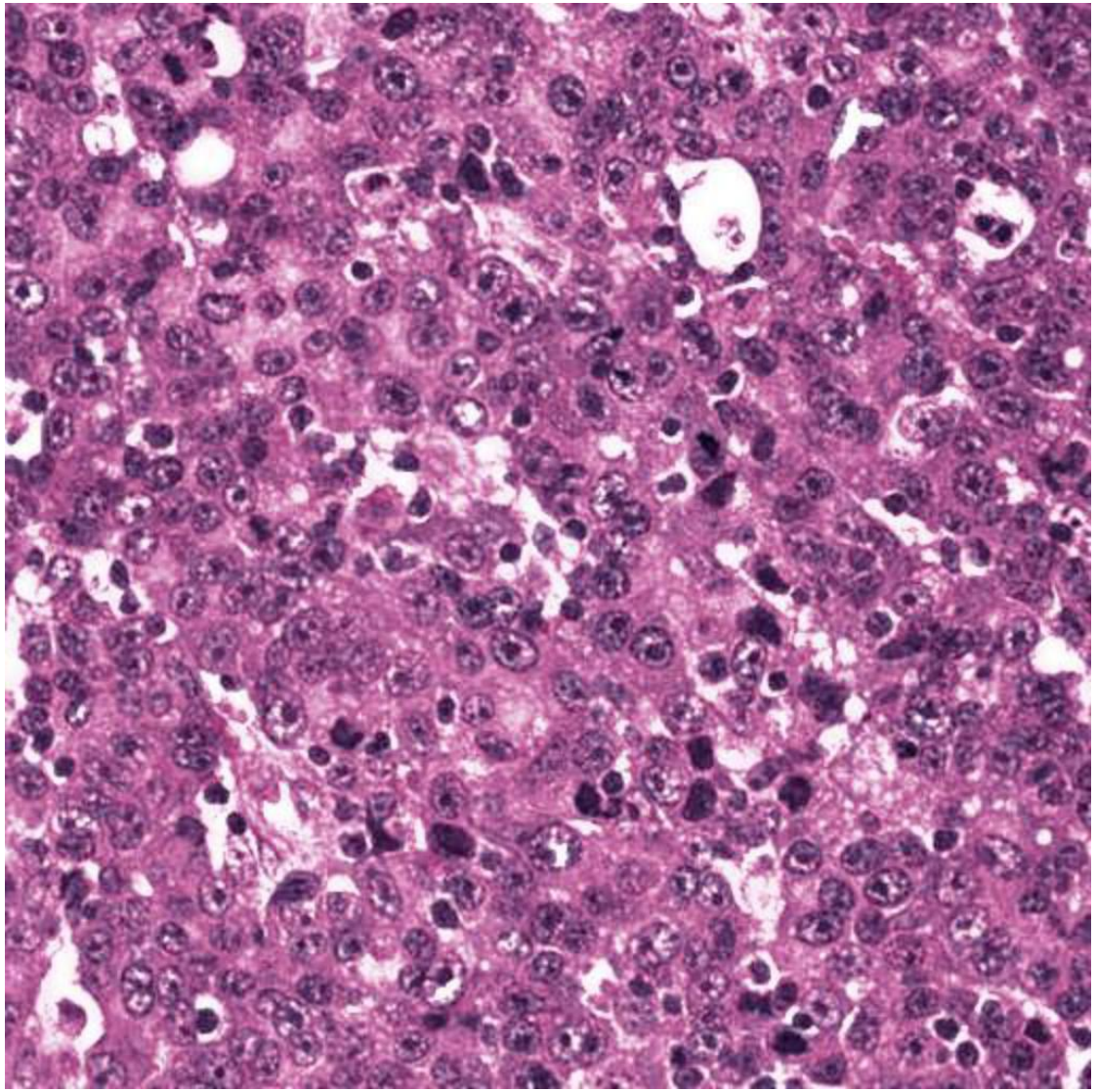


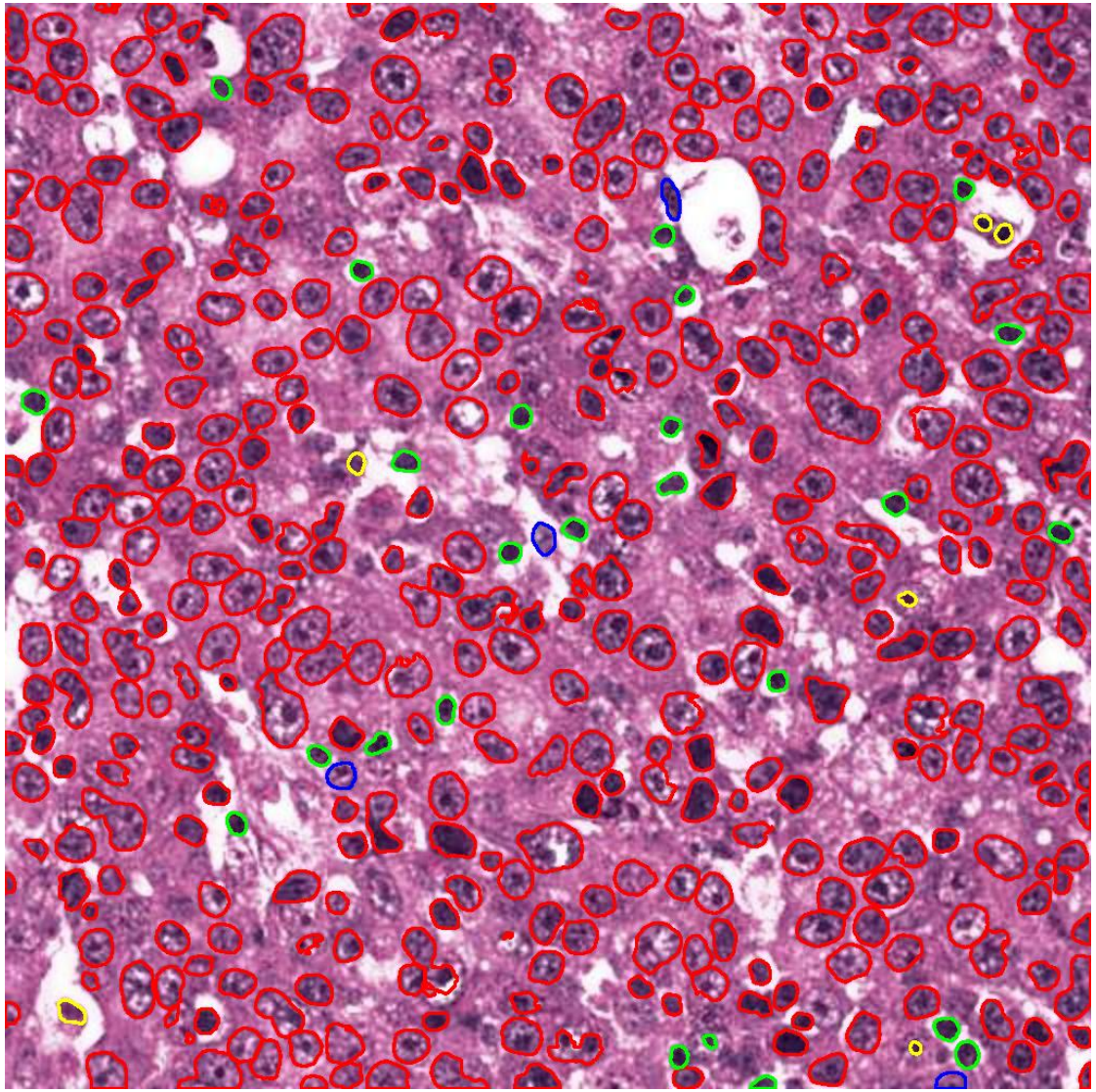
NEP1 Inflammatory Dead Mesenchymal NEP2



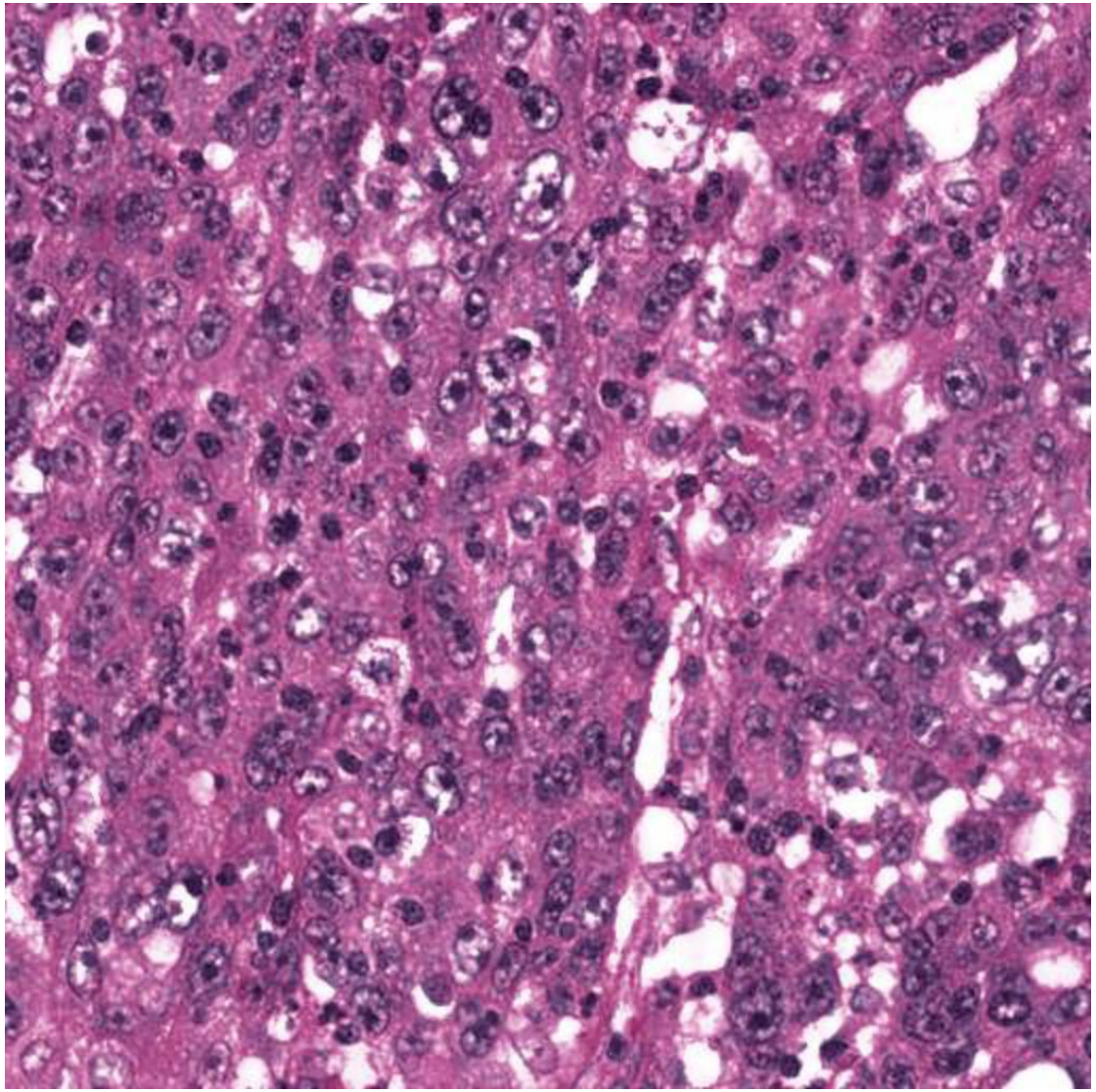


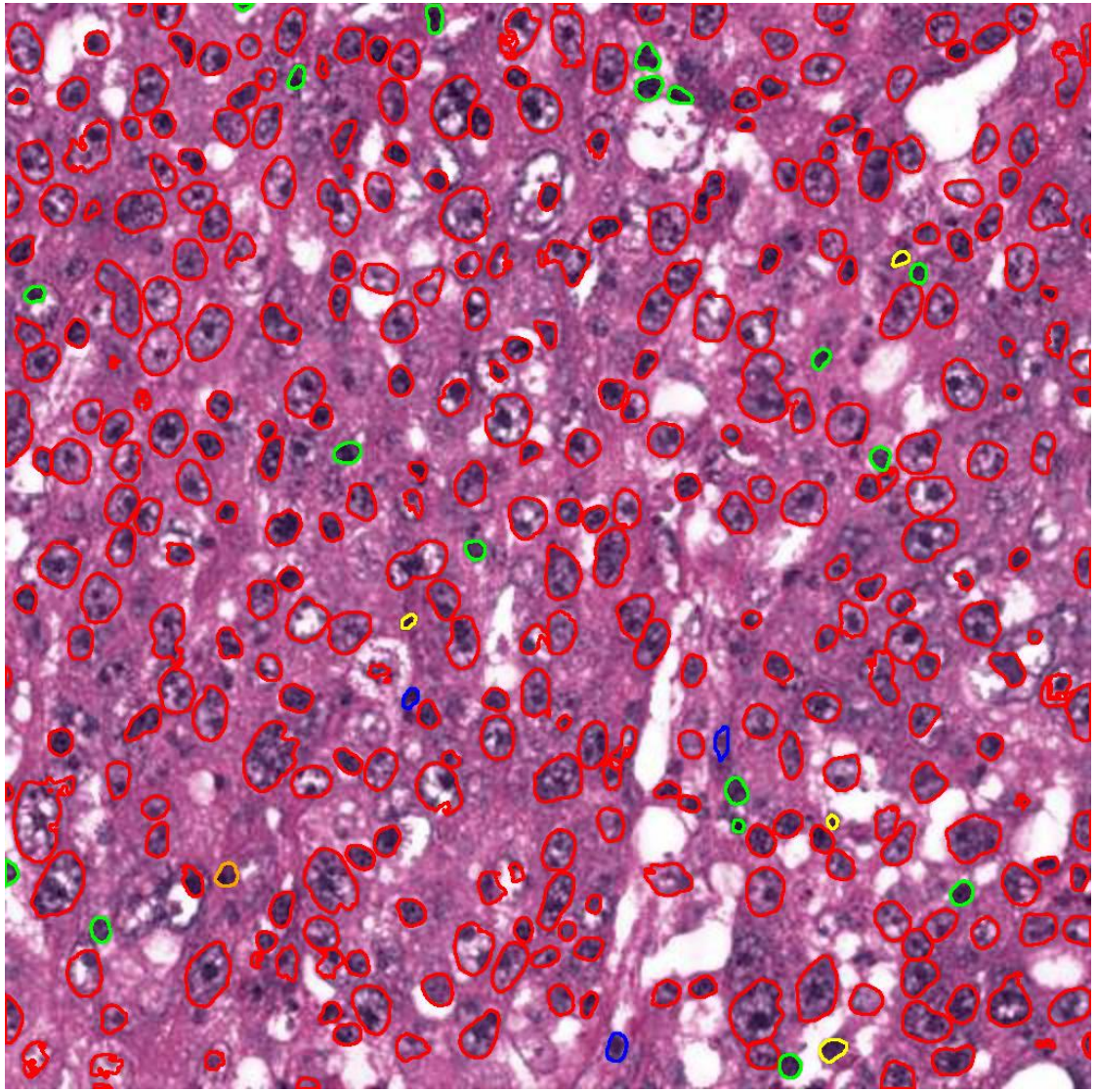
NEP1 Inflammatory Dead Mesenchymal NEP2



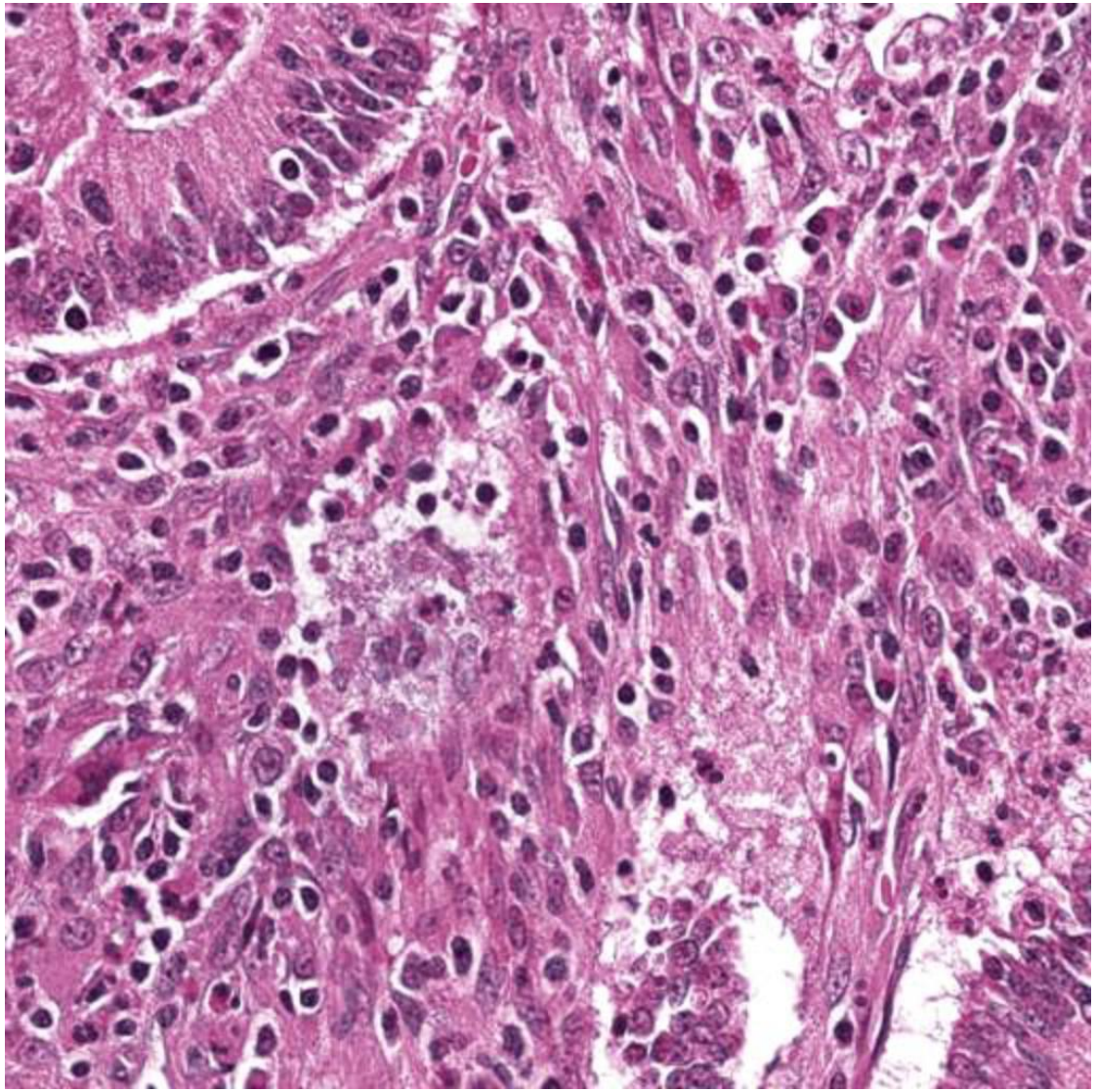


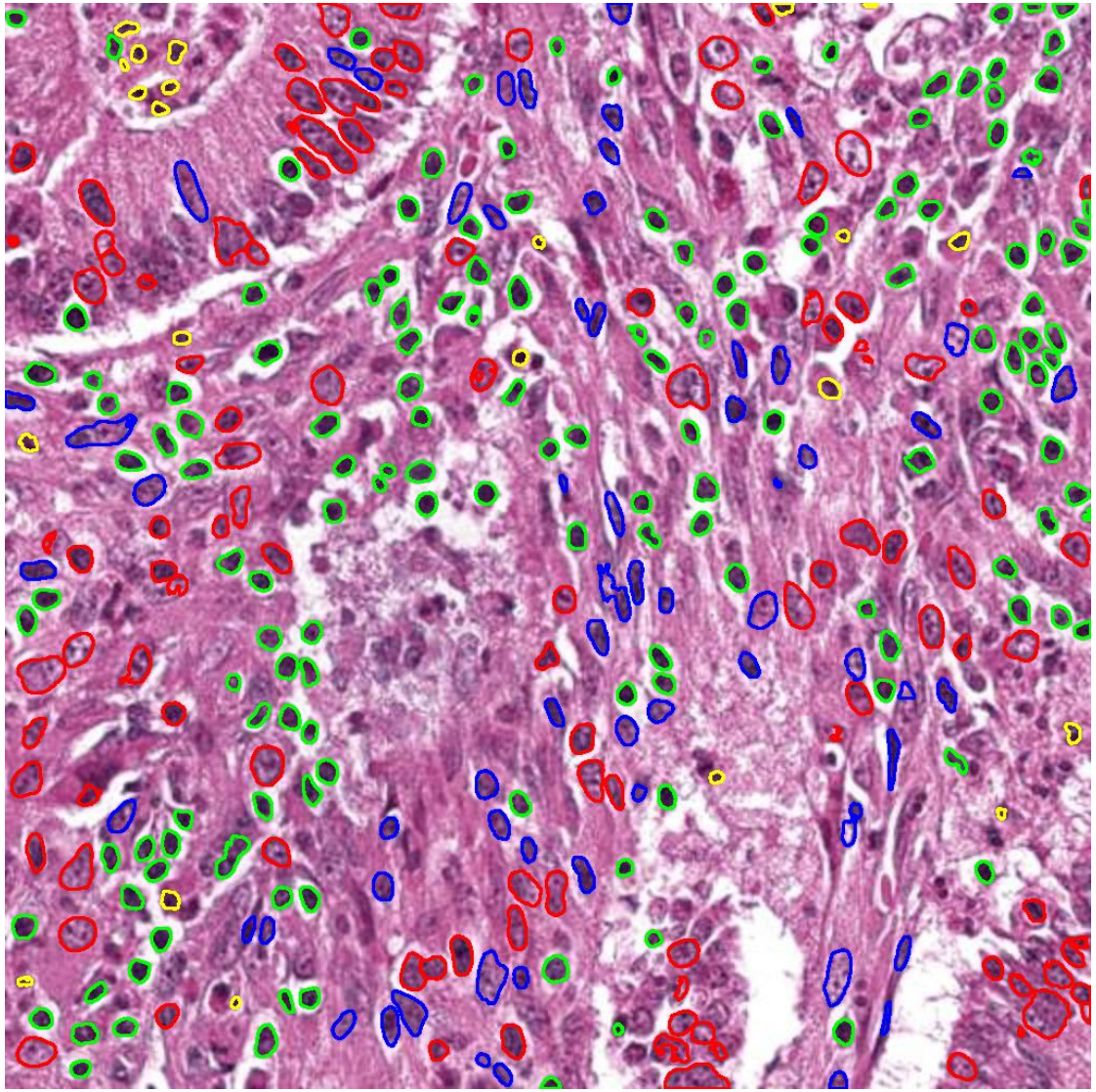
NEP1 Inflammatory Dead Mesenchymal NEP2



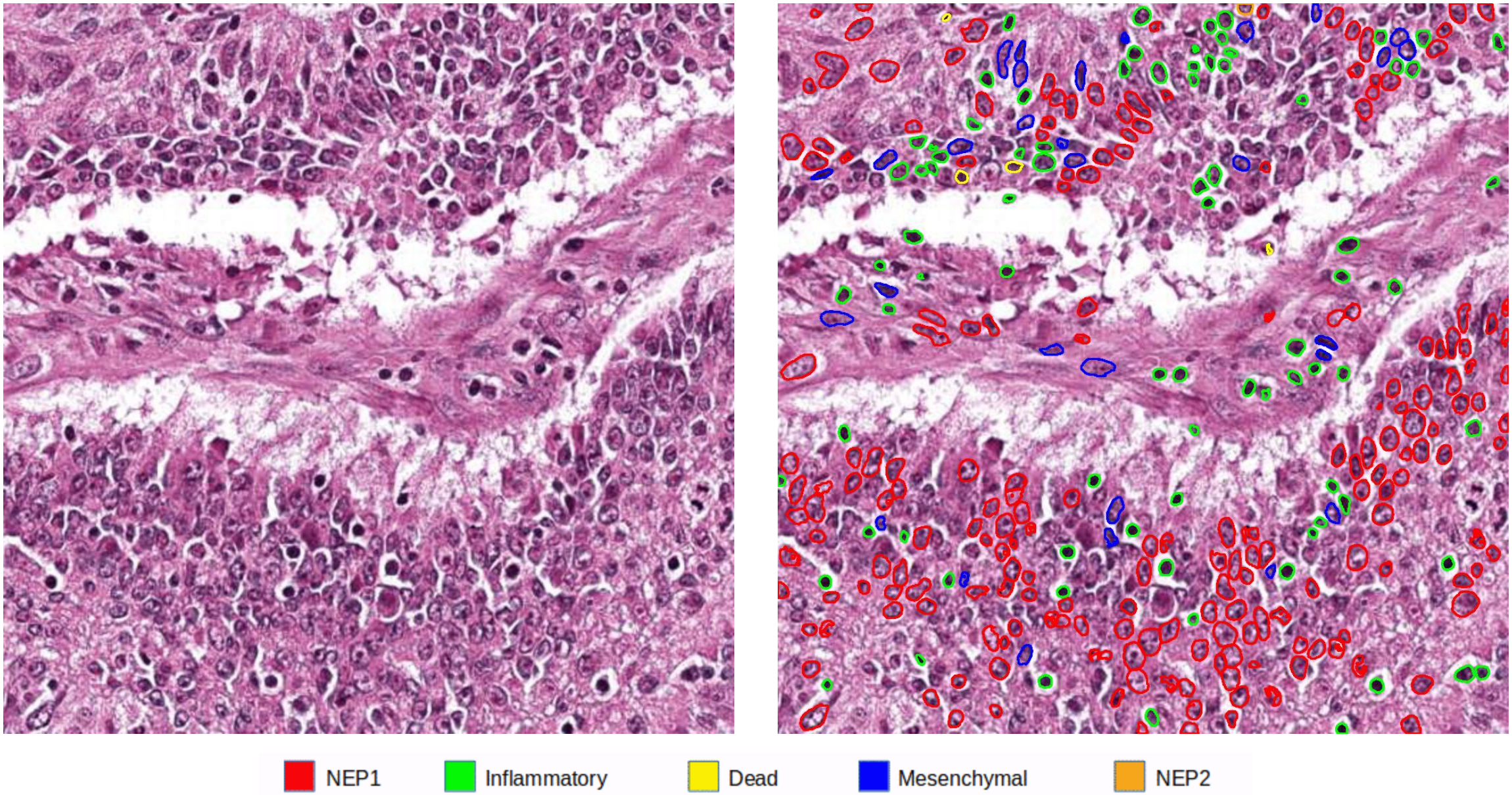


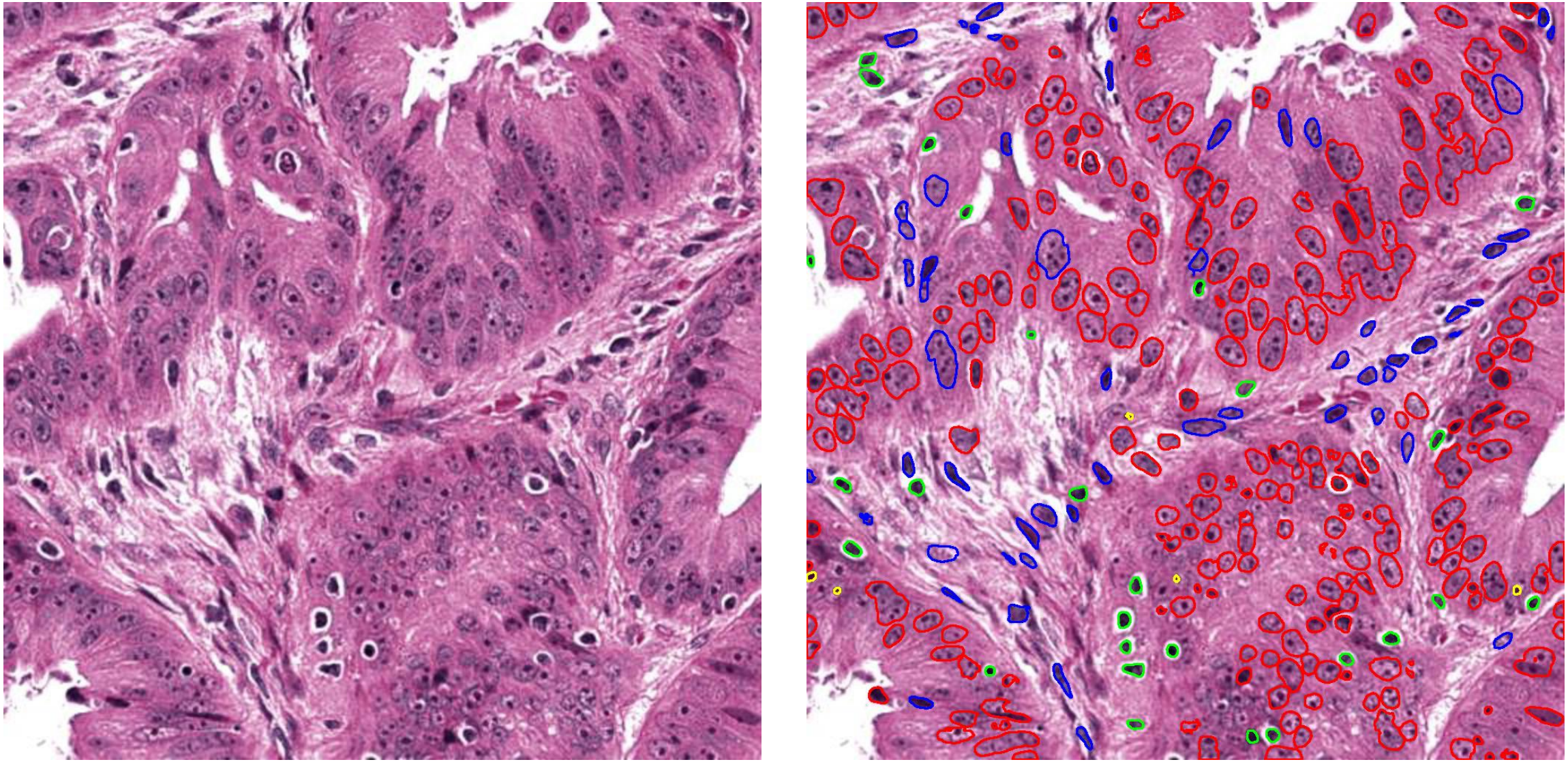
NEP1 Inflammatory Dead Mesenchymal NEP2

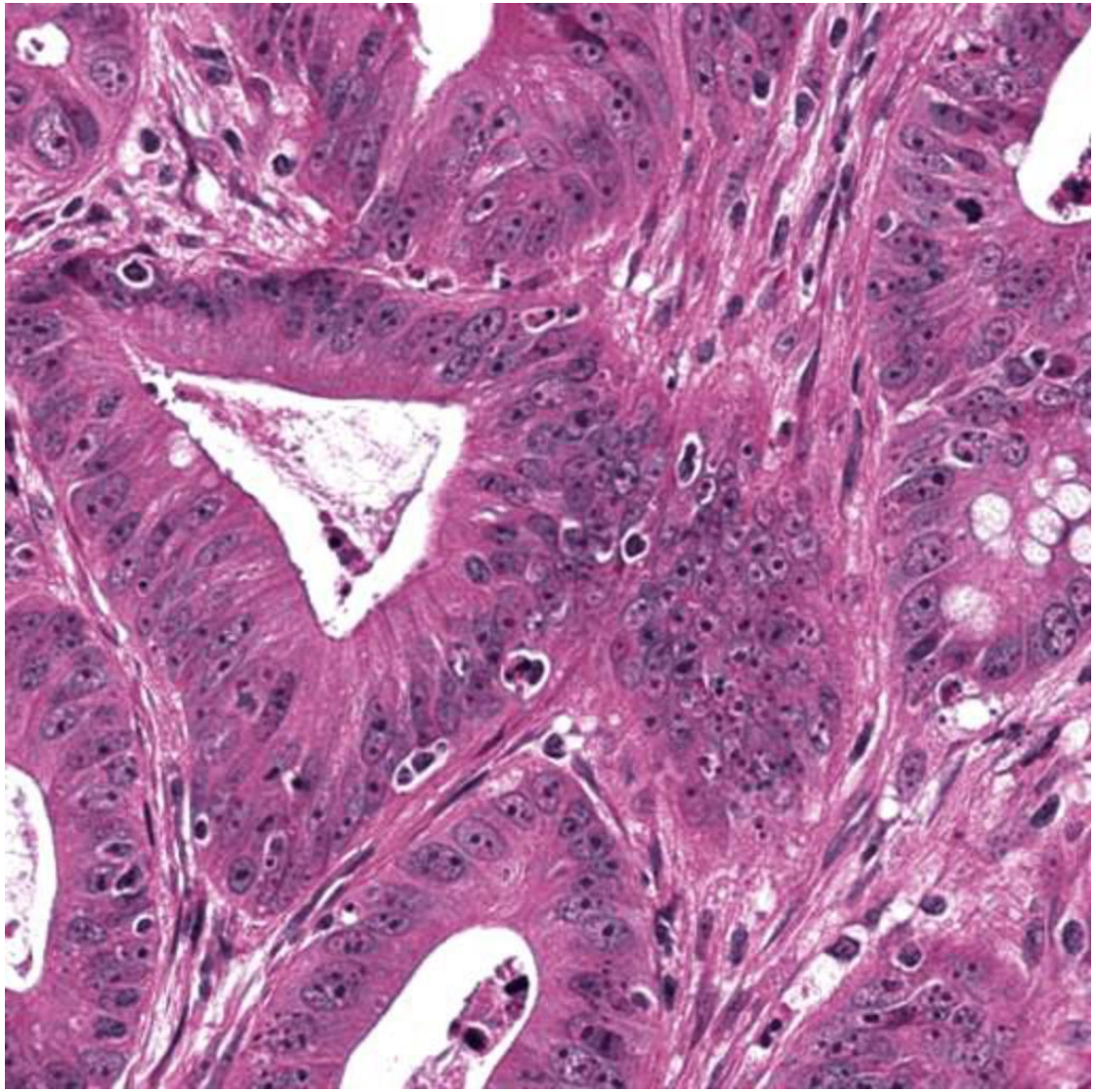


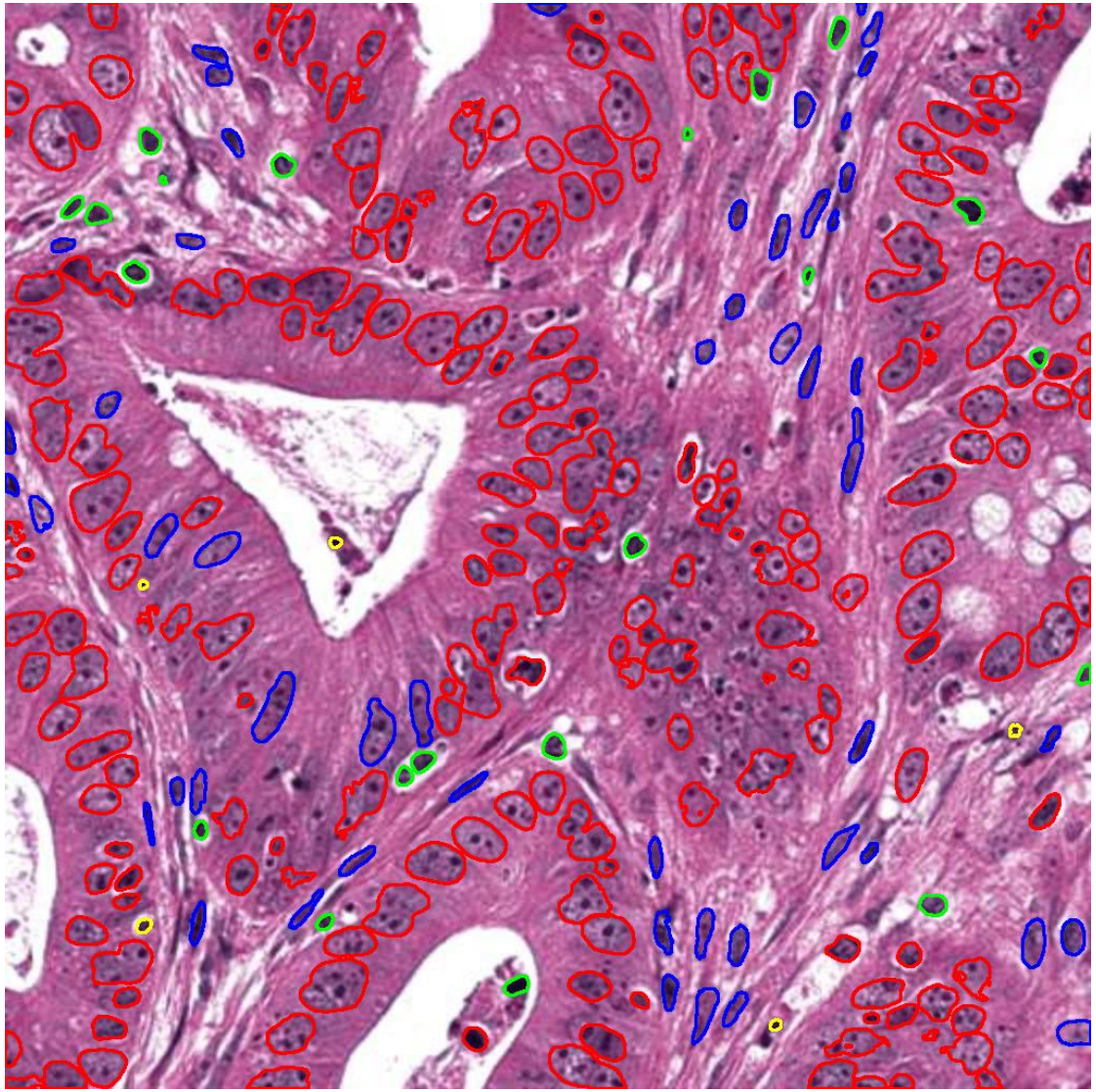


NEP1 Inflammatory Dead Mesenchymal NEP2

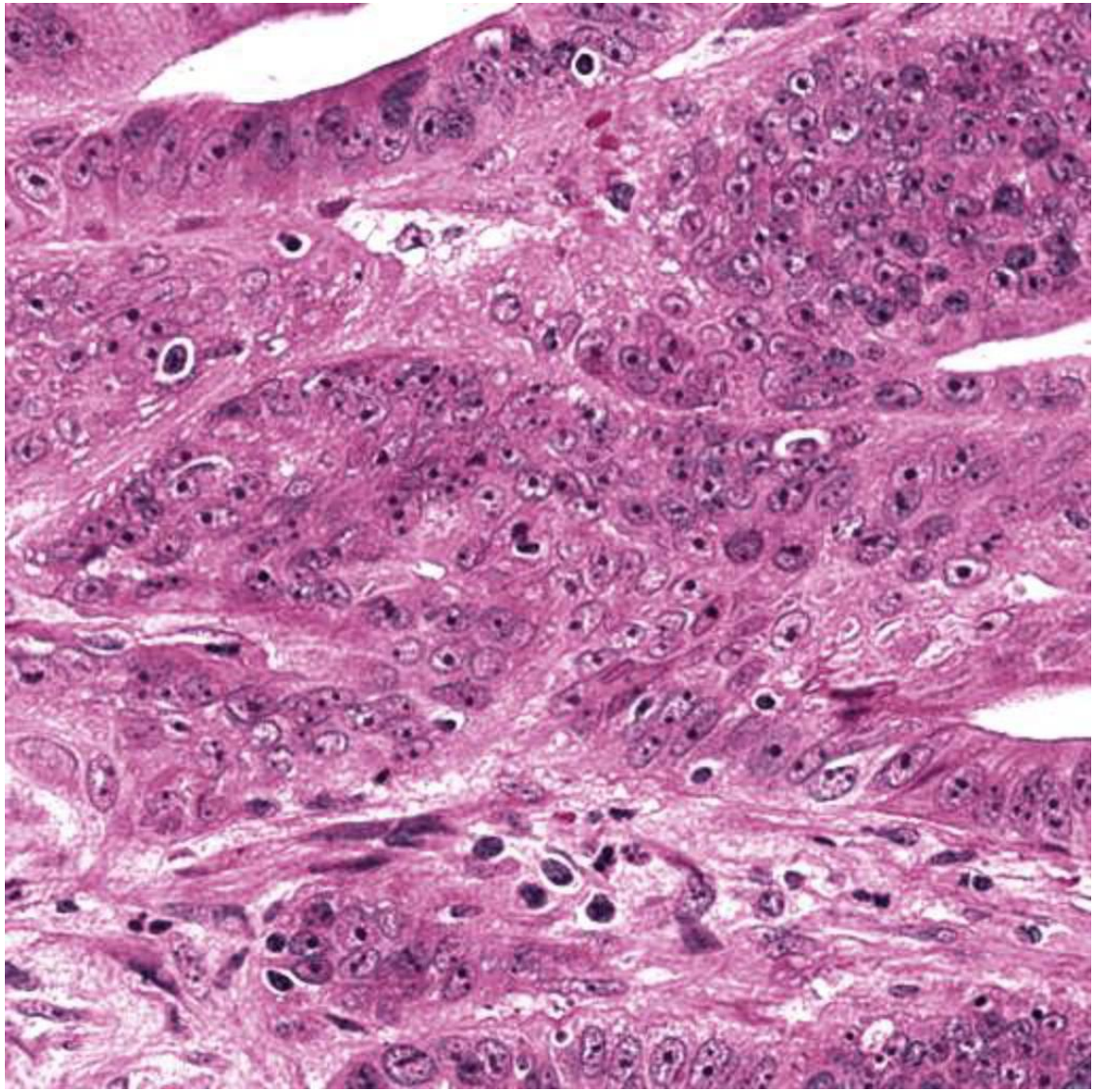

NEP1
Inflammatory
Dead
Mesenchymal
NEP2

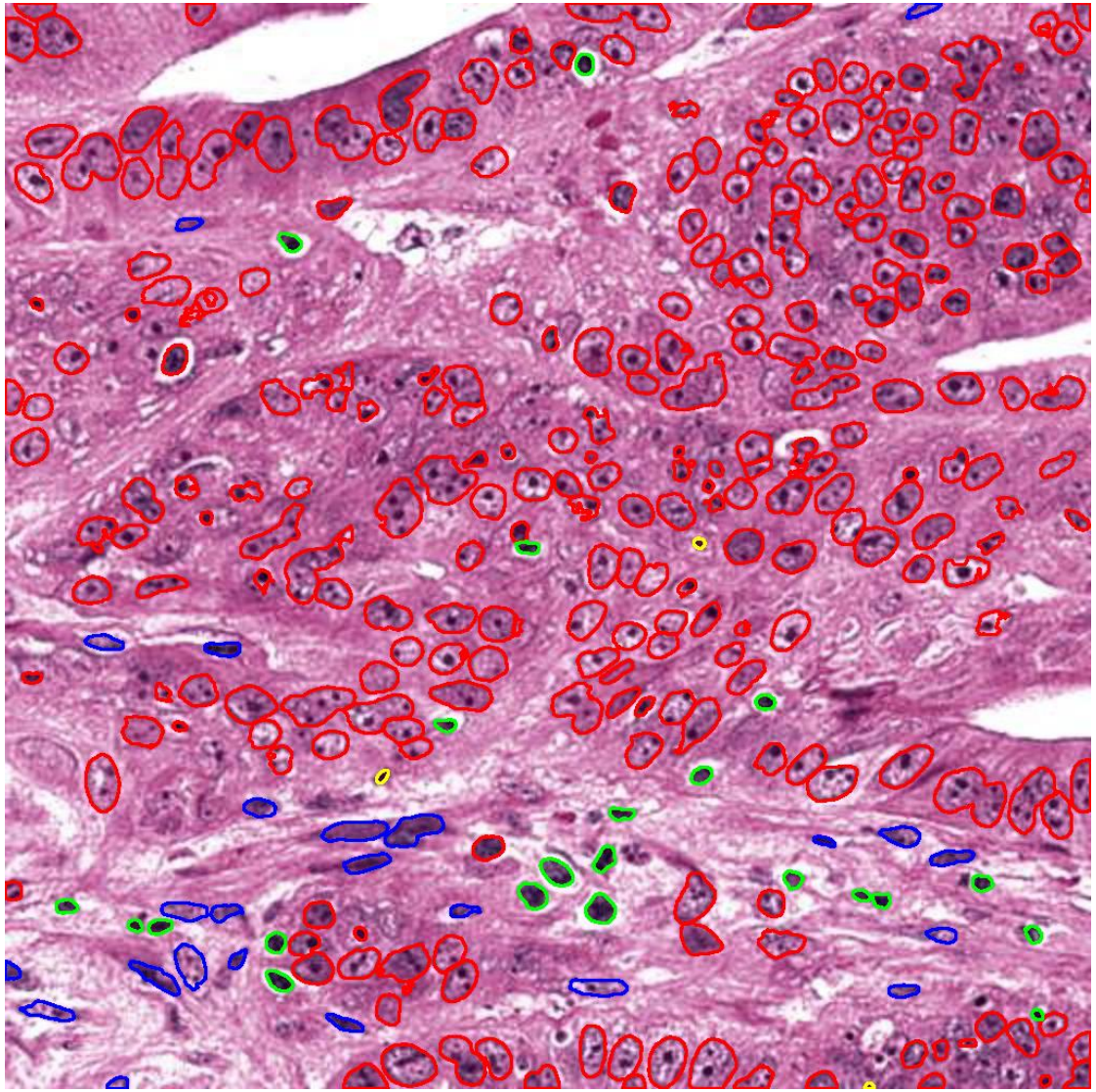

NEP1
Inflammatory
Dead
Mesenchymal
NEP2



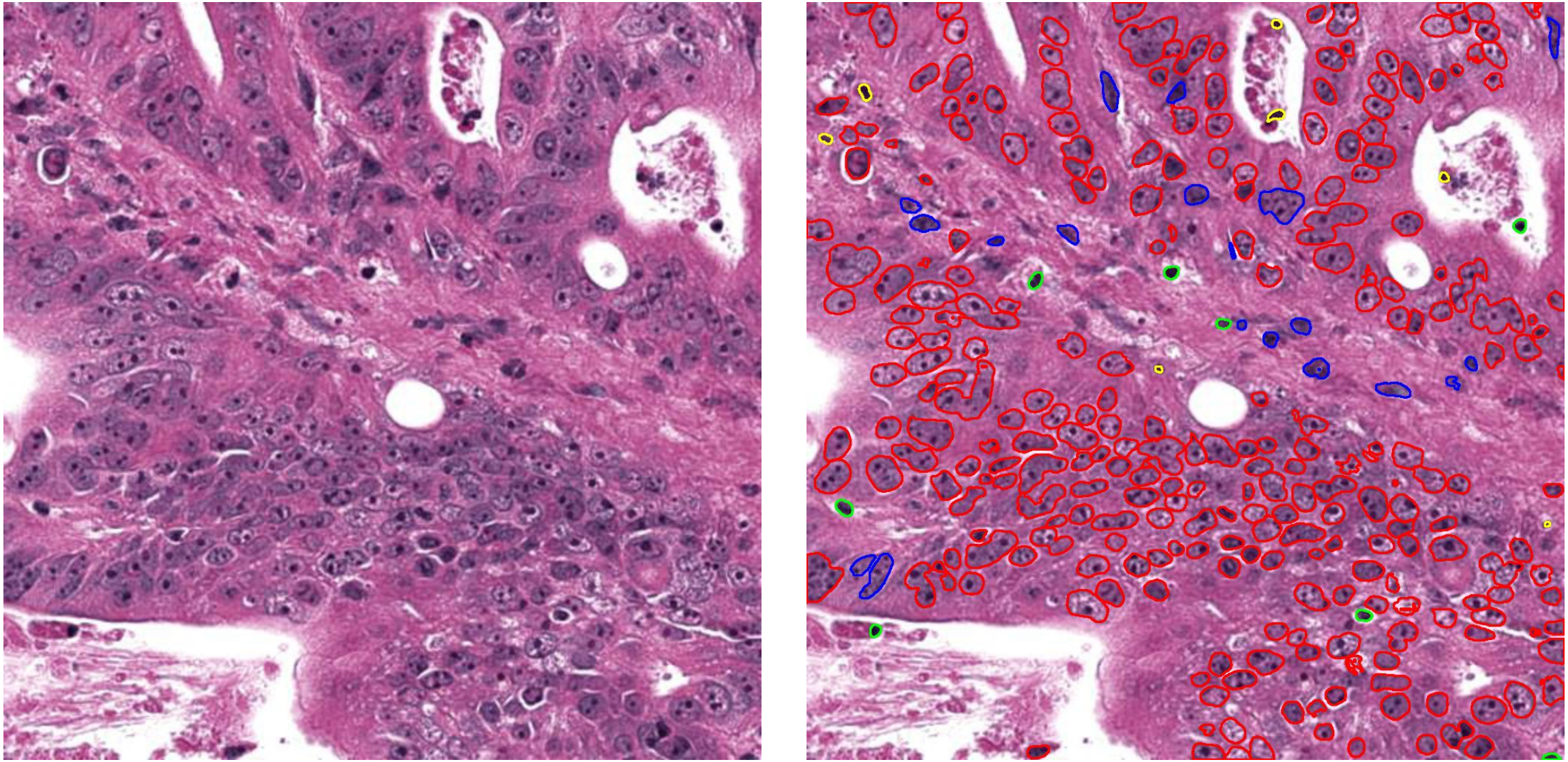


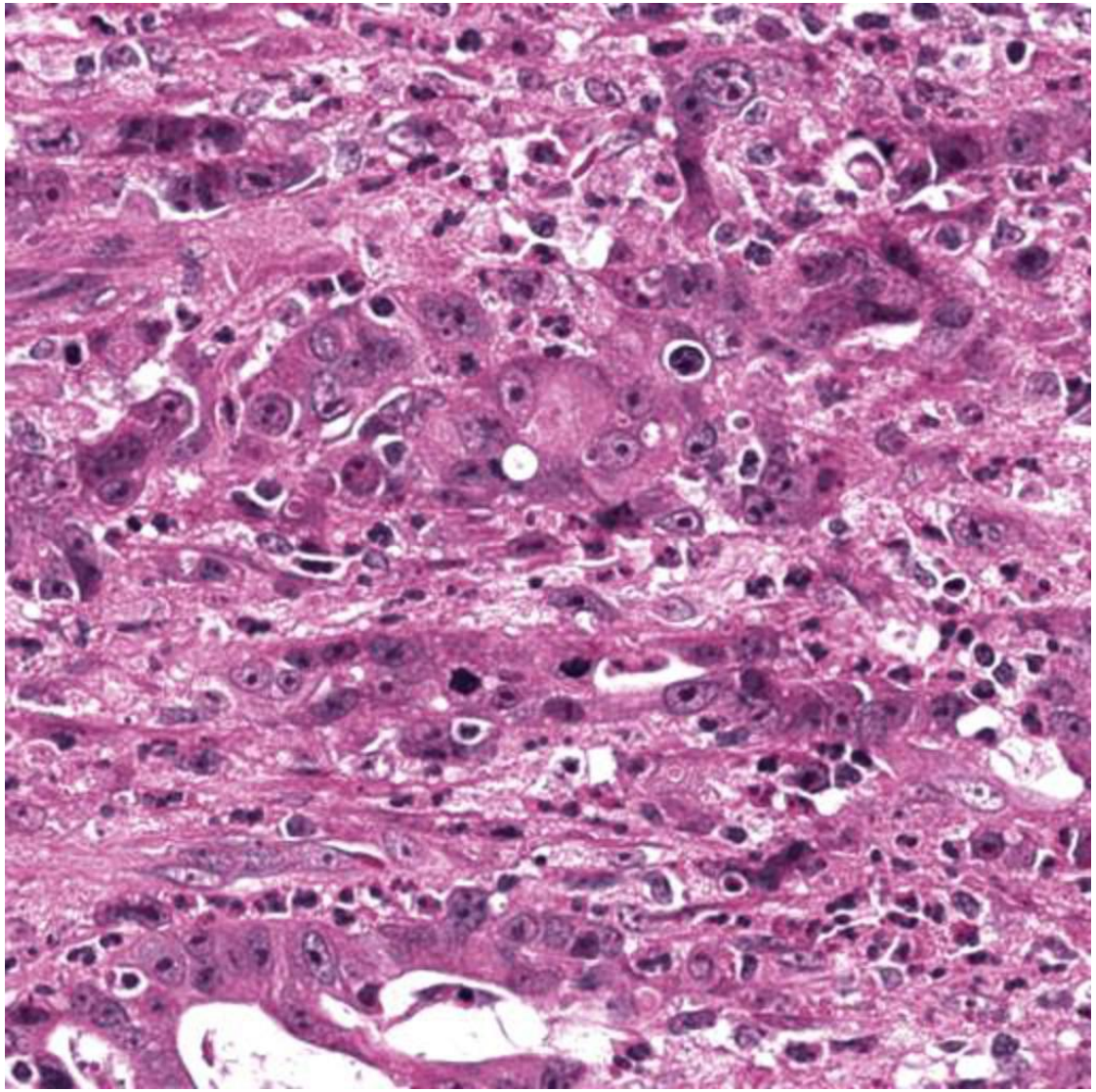
NEP1 Inflammatory Dead Mesenchymal NEP2

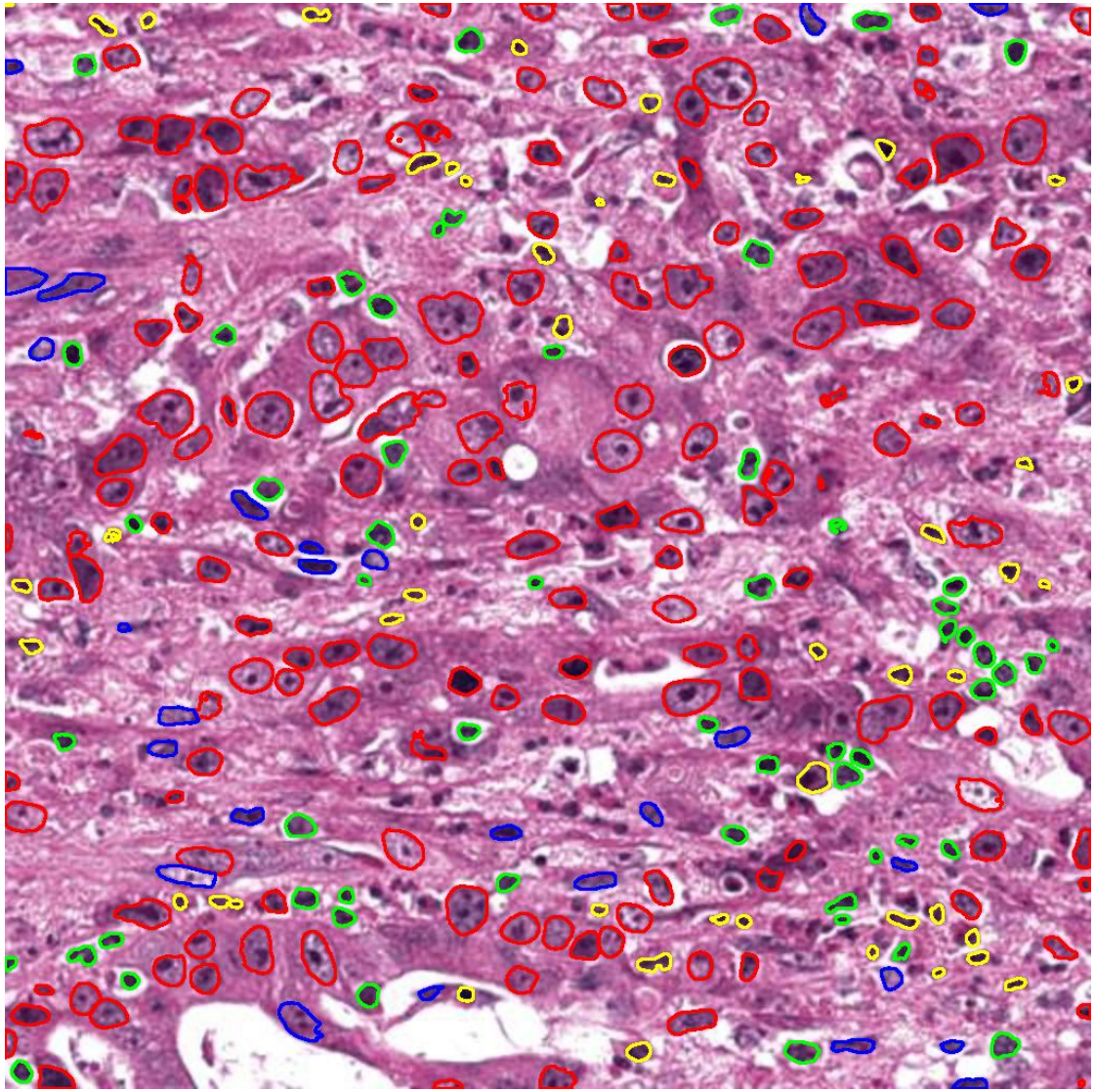




NEP1 Inflammatory Dead Mesenchymal NEP2


NEP1
Inflammatory
Dead
Mesenchymal
NEP2





NEP1 Inflammatory Dead Mesenchymal NEP2