Descriptor

Patterns

# Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study

## Highlights

- MIL model successfully predicts a sample's tumor purity from histopathology slides
- MIL model learns to spatially resolve tumor purity from sample-level labels
- Tumor purity varies spatially within a sample
- Pathologists' region selection is vital for correct percentage tumor nuclei estimation

## Authors

Mustafa Umit Oner, Jianbin Chen, Egor Revkov, ..., Anders Jacobsen Skanderup, Wing-Kin Sung, Hwee Kuan Lee

## Correspondence

ksung@comp.nus.edu.sg

## In brief

Selecting a sample with sufficient tumor content is crucial for the proper operation of sequencing methods. This study developed a deep learning model predicting the percentage of cancer cells (tumor purity) within a tissue section from its digital histopathology slides to support pathologists in sample selection for genomic sequencing. The model successfully predicted tumor purity in eight different cancers and produced tumor purity maps showing the spatial variation within sections without requiring pixel-level annotations from pathologists during training.

Oner et al., 2022, Patterns *3*, 100399
February 11, 2022 © 2021 The Authors.
https://doi.org/10.1016/j.patter.2021.100399

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

# Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study

Mustafa Umit Oner,[1,2] Jianbin Chen,[3] Egor Revkov,[3,2] Anne James,[4] Seow Ye Heng,[4] Arife Neslihan Kaya,[3] Jacob Josiah Santiago Alvarez,[3,2] Angela Takano,[4] Xin Min Cheng,[4] Tony Kiat Hon Lim,[4] Daniel Shao Weng Tan,[5,6,3] Weiwei Zhai,[3,7,8] Anders Jacobsen Skanderup,[3,2,5] Wing-Kin Sung,[2,3,13,*] and Hwee Kuan Lee[1,2,9,10,11,12]

[1]Bioinformatics Institute, Agency for Science, Technology and Research (A*STAR), Singapore 138671, Singapore
[2]School of Computing, National University of Singapore, Singapore 117417, Singapore
[3]Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR), Singapore 138672, Singapore
[4]Department of Anatomical Pathology, Singapore General Hospital, Singapore 169608, Singapore
[5]Division of Medical Oncology, National Cancer Centre Singapore, Singapore 169610, Singapore
[6]Oncology Academic Clinical Programme, Duke-NUS Medical School, Singapore 169857, Singapore
[7]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[8]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China
[9]Singapore Eye Research Institute (SERI), Singapore 169856, Singapore
[10]Image and Pervasive Access Lab (IPAL), Singapore 138632, Singapore
[11]Rehabilitation Research Institute of Singapore, Singapore 308232, Singapore
[12]Singapore Institute for Clinical Sciences, Singapore 117609, Singapore
[13]Lead contact
*Correspondence: ksung@comp.nus.edu.sg

https://doi.org/10.1016/j.patter.2021.100399

**THE BIGGER PICTURE** Given some big data and coarse-level labels, extracting fine-level information is a demanding yet rewarding challenge in data science. This study develops a machine learning model utilizing big data and exploiting coarse-level labels to reveal fine-level details within the data. Although it can be applied to different data science tasks with enormous data and coarse labels, we applied it to a computational histopathology task with gigapixel histopathology slides and sample-level labels. Specifically, the model revealed spatial resolution of tumor purity within histopathology slides using only sample-level genomic tumor purity values during training. This can also be extended to other omics features, providing precious information about cancer biology and promising personalized, precision medicine. Such studies are of great clinical importance in discovering imaging biomarkers and better understanding the tumor microenvironment.

1 2 3 4 5 **Proof-of-Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

## SUMMARY

Tumor purity is the percentage of cancer cells within a tissue section. Pathologists estimate tumor purity to select samples for genomic analysis by manually reading hematoxylin-eosin (H&E)-stained slides, which is tedious, time consuming, and prone to inter-observer variability. Besides, pathologists' estimates do not correlate well with genomic tumor purity values, which are inferred from genomic data and accepted as accurate for downstream analysis. We developed a deep multiple instance learning model predicting tumor purity from H&E-stained digital histopathology slides. Our model successfully predicted tumor purity in eight The Cancer Genome Atlas (TCGA) cohorts and a local Singapore cohort. The predictions were highly consistent with genomic tumor purity values. Thus, our model can be utilized to select samples for genomic analysis, which will help reduce pathologists' workload and decrease inter-observer variability. Furthermore, our model provided tumor purity maps showing the spatial variation within sections. They can help better understand the tumor microenvironment.

## INTRODUCTION

High-throughput genomic analysis has become an indispensable tool for cancer research and has enabled precision oncology.[1,2] One of the crucial factors affecting the quality of genomic analysis is the proportion of cancer cells in the samples.[3] Tumors consist of a complex mixture of cells, such as cancer cells, normal epithelial cells, stromal cells, and infiltrating immune cells.[4] The proportion of cancer cells in a section can significantly influence the accuracy of not only sequencing experiments but also precision oncology. The subjective estimates of the percentage of cancer cells within a tissue section—or tumor purities—are routinely evaluated by pathologists.[5]

The tumor purity affects both high-throughput data acquisition and analysis. To detect genetic variations of a tumor sample by next-generation sequencing, the sample needs to have sufficient cancer cells.[6–8] Therefore, an accurate tumor purity estimation is of great clinical importance. A sample with low tumor purity, for example, may lead to a false-negative test result, potentially resulting in missed therapeutic opportunities.[6] Besides, the genomic analysis should incorporate the tumor purity to account for normal cell contamination, which can have confounding effects on analysis results.[5,9–14] A novel immunotherapy gene signature missed by traditional methods, for example, was discovered using a differential expression analysis incorporating tumor purity.[5] The tumor purity is also associated with clinical variables.[15–17] Low tumor purity, for instance, was associated with poor prognosis in glioma,[15] colon cancer,[16] and gastric cancer.[17] Moreover, tumor purity was a promising predictor for therapeutic response in colon cancer[16] and gastric cancer.[17]

A pathologist estimates tumor purity by reading hematoxylin and eosin (H&E)-stained histopathology slides. Essentially, the pathologist counts the percentage of tumor nuclei over a region of interest (ROI) in the slide. The tumor purity estimated in this way is referred to as percentage tumor nuclei in this study. The percentage tumor nuclei estimates are usually used for sample selection and interpretation of results in the molecular analysis. The pathologist can read any H&E-stained slide and estimate percentage tumor nuclei based on a cellular-level analysis. Thus, this approach is widely applicable, and it has a cellular-level resolution. However, counting tumor nuclei is tedious and time consuming. More importantly, there exists inter-observer variability between pathologists' estimates.[6,18]

Tumor purity can also be inferred from different types of genomic data, such as somatic copy number[19–25] and mutations,[26–31] gene expression data,[32–35] and DNA methylation data.[36–39] The tumor purity obtained from these methods will be referred to as genomic tumor purity in this study. Genomic tumor purity values are usually used in genomics analysis to mitigate confounding effects of normal cell contamination[40–42] and in correlational studies to investigate the associations between tumor purity and clinical variables.[43] Nowadays, genomic tumor purity is accepted as ‘‘accurate’’ for downstream analysis.[19,26,28,32,35] Genomic methods generally produce consistent values on different cancer datasets in The Cancer Genome Atlas (TCGA).[5] However, they do not work well for the low-tumor-content samples. Furthermore, genomic methods cannot provide information on the spatial organization of the tumor microenvironment. Hence, both genomics methods and pathologists' slide reading approach have different strengths and limitations.

Pathologists routinely estimate percentage tumor nuclei in tissue sections. However, besides previously stated challenges, pathologists' estimates do not correlate well with genomic tumor purity values.[5,13] To assist pathologists, this study develops a machine learning model that predicts the tumor purity from H&E-stained histopathology slides such that the predictions are consistent with the genomic tumor purity values. In addition to giving accurate tumor purity measurements, our model is cost-effective compared with genomics methods. It also provides information about the spatial organization of the tumor microenvironment.

Two types of machine learning models can be utilized to predict tumor purity from digital histopathology slides: patch-based models and multiple instance learning (MIL) models. The patch-based models require pathologists' pixel-level annotations showing whether each pixel is cancerous or normal. Although different studies employed this approach for tumor purity prediction,[44–49] they had limited coverage since pixel-level annotations are rarely available, expensive, and tedious. On the other hand, the MIL models do not require pixel-level annotations. Instead, they use sample-level labels, which are weak labels providing only aggregate information rather than pixel-level information. However, they can easily be collected from pathology reports, electronic health records, or different data modalities. The MIL models were successfully used in various digital pathology tasks,[50–52] whereas this is the first study using the MIL approach to predict tumor purity. This study uses sample-level genomic tumor purity values as labels during training and does not require tedious pixel-level annotations by pathologists.

We formulate predicting tumor purity of a sample from its H&E-stained histopathology slides as an MIL task (Figure 1A). The sample's top and bottom slides are cropped into many patches, and these patches are collected to form a bag. Then, the task is to predict the bag-level label of tumor purity. To achieve this task, we developed a novel MIL model with a ‘‘distribution’’ pooling filter (see experimental procedures for details).

Our MIL models successfully predicted sample-level tumor purity in different TCGA cohorts and a local Singapore cohort. The predictions were consistent with genomic tumor purity values (Figure 2). Besides, we obtained spatially resolved tumor purity maps showing the variation of tumor purity over the slides (Figures 1B and 4). We also showed that our MIL models learned discriminant features for cancerous versus normal histology (Figures 1C and 5) and classified samples into tumor versus normal almost perfectly in all cohorts (Figures 1D and 3B).

## RESULTS

In this study, there were 10 different TCGA cohorts and a local Singapore cohort. Each TCGA cohort had more than 400 patients, and the Singapore cohort had 179 lung adenocarcinoma patients, such that each patient had both histopathology slides and corresponding genomic sequencing data (Table 1, see also Tables S1 and S2). The histopathology slides in each cohort were randomly segregated at the patient level into training, validation, and test sets (Figures S1 and S2). We trained our MIL model on the training set and chose the best set of model weights based on validation set performance. Finally, we

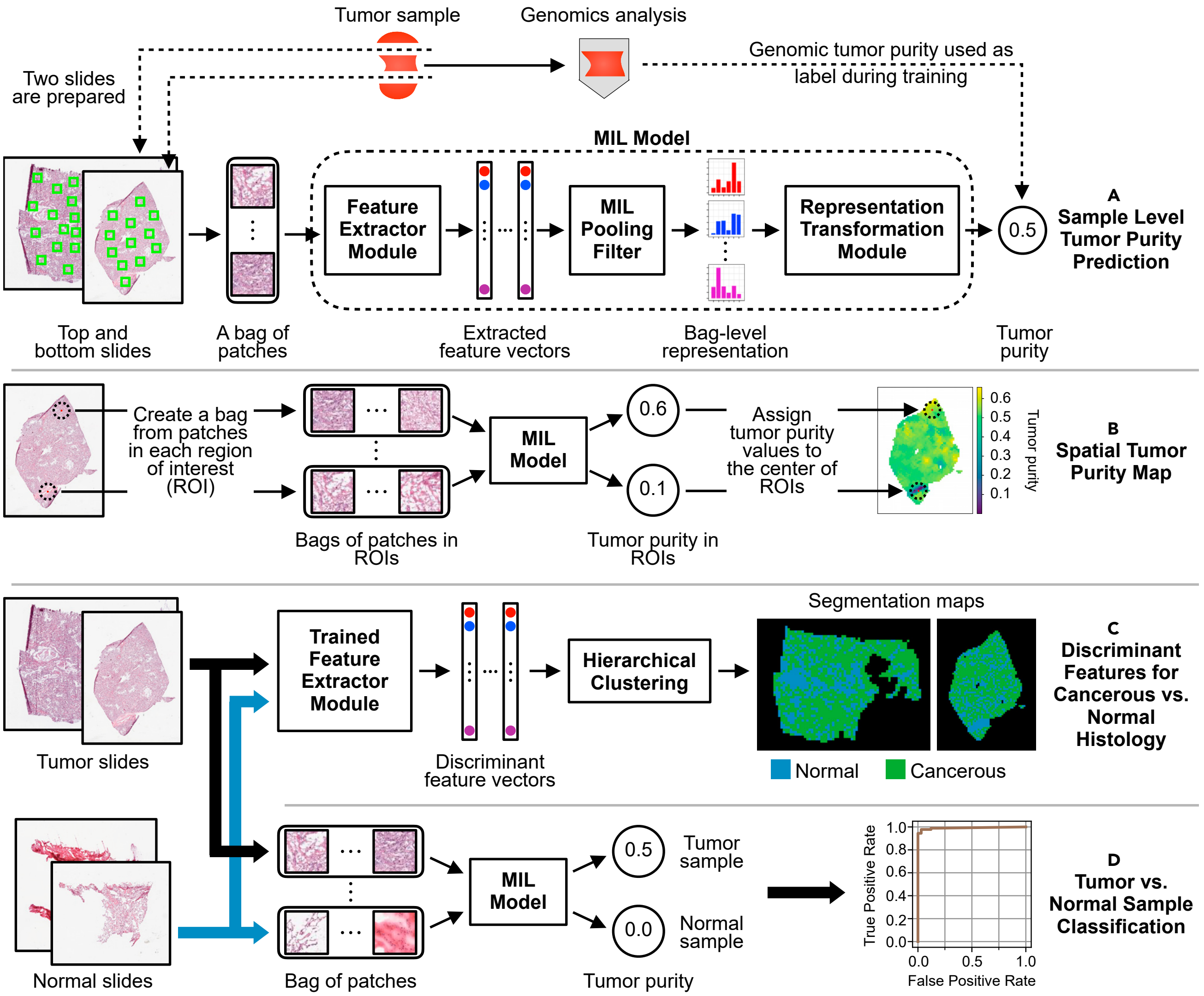


**Figure 1. A novel MIL model predicts sample-level tumor purity from H&E-stained digital histopathology slides**

(A) Our model accepts a bag of patches randomly cropped from the top and bottom slides of a sample as input and predicts the sample's tumor purity at its output. The feature extractor module extracts a feature vector for each patch inside the bag. The MIL pooling filter, namely distribution pooling, summarizes extracted features into a bag-level representation by estimating marginal feature distributions. Finally, the bag-level representation transformation module predicts the sample-level tumor purity. We use tumor purity values inferred from genomic sequencing data by ABSOLUTE[19] as ground-truth labels during training.

(B) We obtain a spatial tumor purity map for a slide by inferring tumor purity over each 1-mm$^2$ ROI within the slide in a sliding window fashion. The map shows the variation of tumor purity over the slide.

(C) Our MIL model learned discriminant features for cancerous versus normal histology from sample-level genomic tumor purity labels without requiring exhaustive annotations from pathologists. We used discriminant features to obtain cancerous versus normal segmentation maps for tumor slides. Trained feature extractor module extracts features of patches from tumor and normal slides of a patient. Then, segmentation maps are obtained by hierarchical clustering over the extracted feature vectors.

(D) Our MIL model successfully classifies samples into tumor versus normal.

evaluated the performance of our trained MIL model on the data of completely unseen patients in the hold-out test set. Each patient in the test set was like a new patient walking into the clinic.[54]

## MIL models' tumor purity predictions correlate significantly with genomic tumor purity values

Our models' performance in 10 different TCGA cohorts was evaluated by correlation analyses between genomic tumor purity values obtained from ABSOLUTE[19] and our MIL models' predictions. The performance metric was Spearman's rank correlation coefficient.

We obtained significant correlations ($p < 0.05$) in eight cohorts, namely breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), brain lower grade glioma (LGG), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma

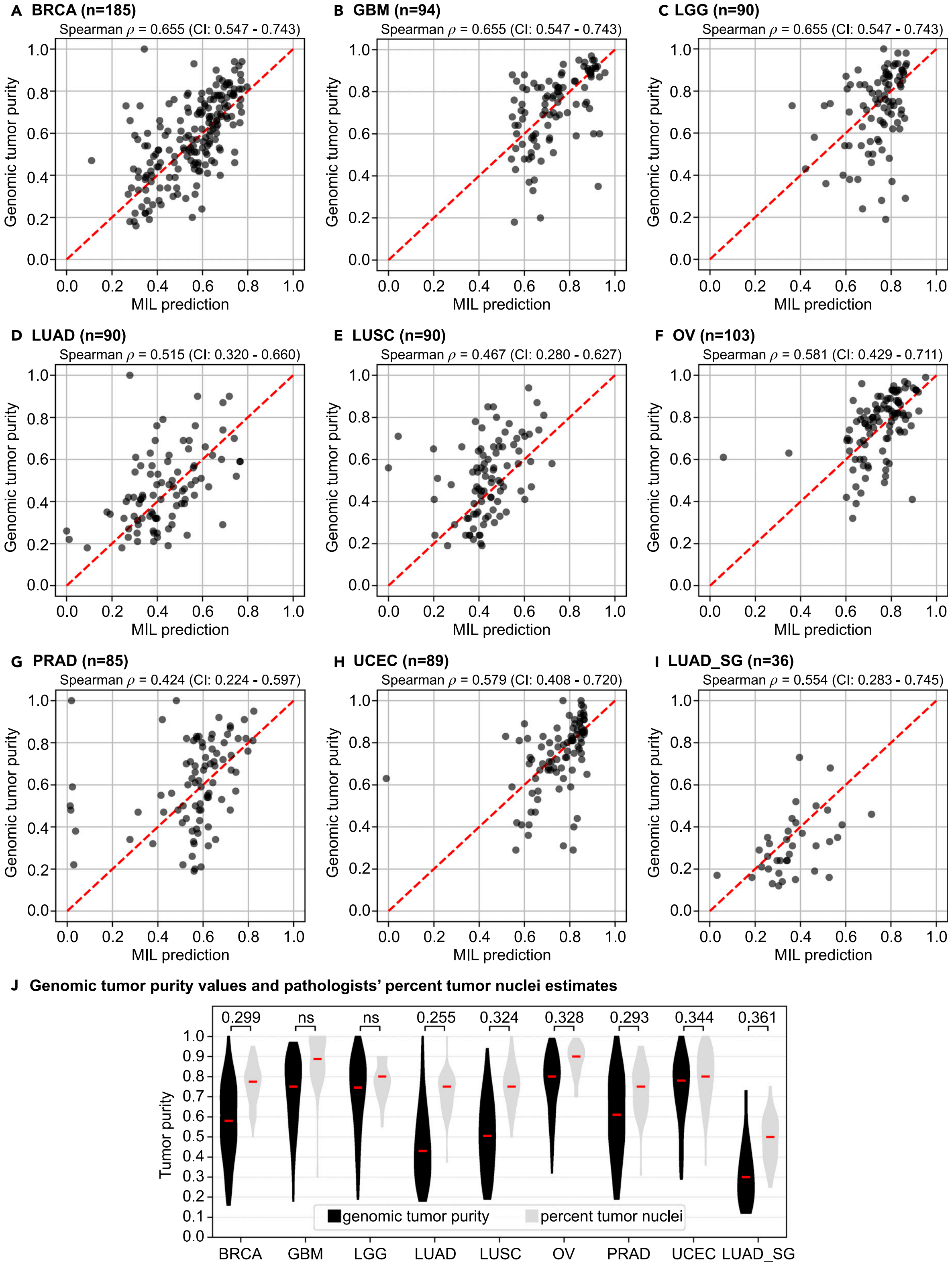


**Figure 2. The MIL model's tumor purity predictions correlate significantly with genomic tumor purity values**

(A–I) A scatterplot of genomic tumor purity versus the MIL model's prediction is given for only tumor samples in the test set of each cohort: (A) BRCA, (B) GBM, (C) LGG, (D) LUAD, (E) LUSC, (F) OV, (G) PRAD, (H) UCEC, and (I) LUAD_SG. Correlation coefficients with 95% CIs are given at the top of each plot. Note that the red dotted line in each plot shows the diagonal (i.e., y = x line). All data points would align on the diagonal line in case of zero prediction error.

*(legend continued on next page)*

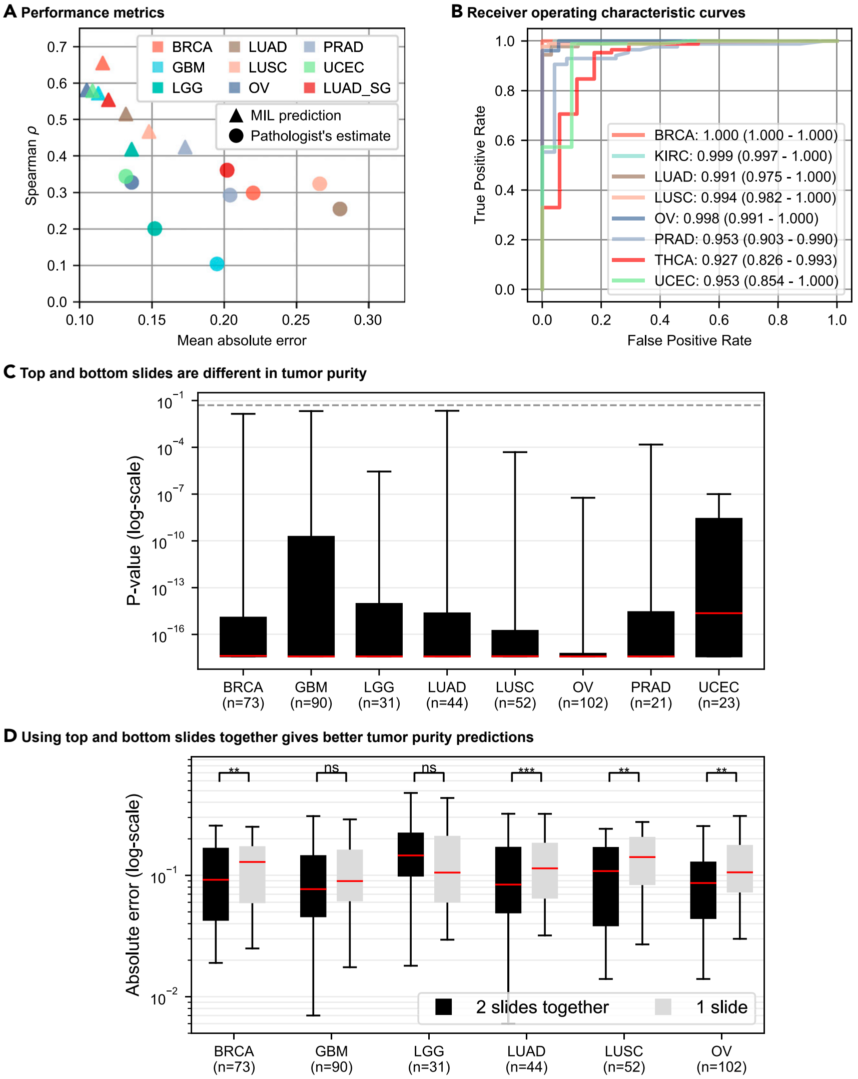


**Figure 3. Performance analysis of MIL models**

(A and B) MIL models perform better than percentage tumor nuclei estimates and successfully classify samples into tumor versus normal. (A) Spearman's correlation coefficient versus mean-absolute-error plot is given for MIL models' tumor purity predictions (represented by triangles) and pathologists' percentage tumor nuclei estimates (represented by circles) in the test sets of different cohorts (showed in different colors). MIL models' predictions achieve lower mean absolute error and higher Spearman's correlation coefficient than percentage tumor nuclei estimates. See also Tables S3 and S5. (B) ROC curve analysis over MIL models' predictions for tumor versus normal sample classification. The area under curve values with 95% CIs are given in the legend. MIL models successfully classified samples into tumor versus normal in all cohorts.

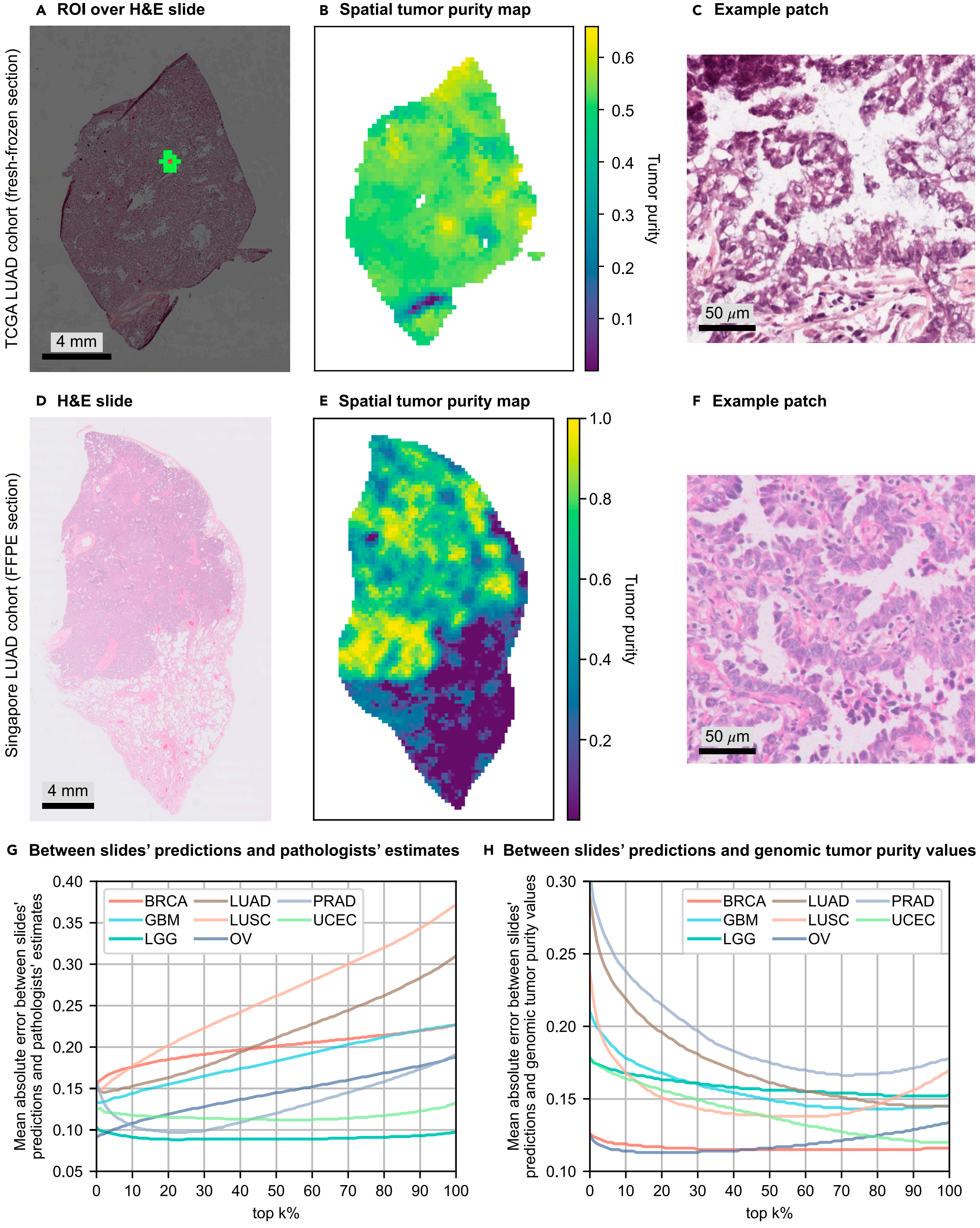
(C and D) The top and bottom slides of a tumor sample are different in tumor purity. In the test set of each cohort, for a tumor sample having top and bottom slides, we conducted two experiments. (C) The trained MIL model's predictions from the top and bottom slides of a sample are statistically compared using Wilcoxon signed-rank test.[53] Each box plot summarizes the p values obtained in a cohort. For at least 95% of the samples in each cohort, the top and bottom slides are significantly different ($p < 0.05$) in tumor purity. The dashed line shows $p = 0.05$. See also Table S6. (D) For each sample, the absolute error between genomic tumor purity value and the MIL model's prediction using both slides and the expected value of absolute errors between genomic tumor purity value and the MIL model's predictions over individual slides are calculated. Box plots summarize the absolute errors in two approaches. They are statistically compared using Wilcoxon signed-rank test,[53] and the results are presented on top of the plots such that $p > 0.05$ (ns, not significant), $^{*}p \leq 0.05$, $^{**}p \leq 0.01$, and $^{***}p \leq 0.001$. See also Table S7. Whiskers show 5th and 95th percentiles, and red lines show median values. n, number of tumor samples with two slides.

(PRAD), and uterine corpus endometrial carcinoma (UCEC) (Figures 2A–2H and Table S3). While the minimum Spearman's $\rho_{mil}$ = 0.418 ($p = 4.1 \times 10^{-5}$; 95% confidence interval [CI], 0.226–0.574) was in the LGG cohort, the maximum Spearman's $\rho_{mil}$ = 0.655 ($p = 4.6 \times 10^{-24}$; 95% CI, 0.547–0.743) was in the BRCA cohort. We compared our MIL models'' predictions with tumor purity values obtained from ESTIMATE[32] as well and observed similar performance (Table S4).

We repeated the correlation analyses between genomic tumor purity values and pathologists' percentage tumor nuclei estimates (Figure 2J and Table S3). While the minimum Spearman's $\rho_{path}$ = 0.240 ($p = 2.7 \times 10^{-2}$; 95% CI, 0.009–0.446) was in the thyroid carcinoma (THCA) cohort, the maximum Spearman's $\rho_{path}$ = 0.344 ($p = 9.8 \times 10^{-4}$; 95% CI, 0.139–0.531) was in the UCEC cohort. There was no significant correlation in the GBM and LGG cohorts. Hence, the minimum correlation with MIL predictions ($\rho_{mil}$ = 0.418 in the LGG cohort) was higher than the maximum correlation with pathologists' percentage tumor nuclei estimates ($\rho_{path}$ = 0.344 in the UCEC cohort). This implies that MIL predictions are more consistent with genomic tumor purity values than the pathologists' percentage tumor nuclei estimates.

Moreover, we conducted statistical tests on correlation coefficients to compare our MIL models' predictions and pathologists' percentage tumor nuclei estimates. We used the Fisher's z transformation-based method of Meng et al.[55] Two methods were compared only when there was a significant correlation for both methods in a cohort (Table S3). MIL predictions were significantly better than pathologists' estimates in all cohorts except LUSC and PRAD. For these cohorts, two methods

(J) Violin plots summarize genomic tumor purity values and pathologists' percentage tumor nuclei estimates in the test set of each cohort. Correlation coefficients are given at the top. Red lines show median values. ns, not significant; n, the number of tumor samples.
See also Tables S3 and S5.



*(legend on next page)*

performed on par ($p_{comp} = 1.7 \times 10^{-1} > 0.05$ for the LUSC and $p_{comp} = 2.0 \times 10^{-1} > 0.05$ for the PRAD) in the test sets.

### MIL models' predictions have lower mean absolute error than percentage tumor nuclei estimates

Apart from Spearman's correlation coefficients, we also checked the mean absolute errors between genomic tumor purity values and MIL models' predictions, and between genomic tumor purity values and pathologists' percentage tumor nuclei estimates (Table S5).

In the analyses of MIL predictions, the minimum and maximum mean-absolute-error values of $\mu_{e_{mil}} = 0.105$ (standard deviation $\sigma_{e_{mil}} = 0.091$) and $\mu_{e_{mil}} = 0.173$ ($\sigma_{e_{mil}} = 0.154$) were obtained in the OV cohort and the PRAD cohort, respectively. On the other hand, in the analyses of pathologists' percentage tumor nuclei estimates, the minimum and maximum mean-absolute-error values of $\mu_{e_{path}} = 0.132$ ($\sigma_{e_{path}} = 0.124$) and $\mu_{e_{path}} = 0.280$ ($\sigma_{e_{path}} = 0.151$) were obtained in the UCEC cohort and the LUAD cohort, respectively. In all cohorts, pathologists' estimates were generally higher than genomic tumor purity values (Figure 2J).

Similar to our comparison in correlation analyses, we compared two methods based on absolute errors in the test sets of different cohorts. We used the Wilcoxon signed-rank test[53] on absolute error values for tumor samples in the test sets (Table S5). Absolute error values in MIL predictions were significantly lower than those in pathologists' percentage tumor nuclei estimates in all cohorts except the LGG cohort. Two methods performed similarly ($p_{comp} = 5.4 \times 10^{-2} > 0.05$) in the test set of the LGG cohort.

Figure 3A summarizes correlation and absolute error analyses. We observed that MIL predictions had lower mean absolute error and higher Spearman's correlation coefficient than pathologists' percentage tumor nuclei estimates.

### MIL model predicts tumor purity from H&E-stained slides of FFPE sections in the Singapore cohort

Our MIL models successfully predicted tumor purity from H&E-stained digital histopathology slides of fresh-frozen sections in different TCGA cohorts. Besides, we evaluated their performance on slides of formalin-fixed paraffin-embedded (FFPE) sections in a local Singapore cohort consisting of 179 lung adenocarcinoma patients (see Note S1 for details). Similar to TCGA cohorts, we segregated data at the patient level (Table 1 and Figure S3).

We used transfer learning and initialized the model with the weights of the MIL model trained on the TCGA LUAD cohort. Then, we froze the weights of all layers in the network except the first convolutional layer in the feature extractor module (Figure 1A). This helped the network adapt the first layer weights to learn the tissue morphology in FFPE sections, which were different from fresh-frozen sections (Figure S4). Note that, while the FFPE method preserves morphology better and is the routine in histopathology, the fresh-frozen method preserves nucleic acids better and is preferred for molecular analysis.[56]

Similar to the performance in the TCGA LUAD cohort, we obtained a Spearman's $\rho_{mil} = 0.554$ ($p = 4.6 \times 10^{-4}$; 95% CI, 0.283–0.745) and the mean absolute error of $\mu_{e_{mil}} = 0.120$ ($\sigma_{e_{mil}} = 0.091$) in the test set of the Singapore LUAD (LUAD_SG) cohort (Figures 2I and 3A). There were substantial differences between the TCGA and LUAD_SG cohorts, such as tissue preservation method (fresh-frozen versus FFPE) and ancestry of patients (European versus East Asian). However, our MIL model successfully predicted tumor purity from slides of FFPE sections using transfer learning with minimal training only in the first convolutional layer of the feature extractor module. The results suggested that our MIL models learned robust features for tumor purity prediction tasks at the higher levels of the network. We also checked the performance of the TCGA LUAD model directly on the LUAD_SG cohort used as an external validation set (Figure S5). Nevertheless, we did not get a significant correlation ($\rho_{mil} = 0.141$; $p = 0.06 > 0.05$), which highlighted the necessity of adapting the weights of the first layer in feature extractor to FFPE slides using transfer learning.

For pathologists' estimates, we obtained a Spearman's $\rho_{path} = 0.361$ ($p = 3.0 \times 10^{-2}$; 95% CI, 0.029–0.644) and the mean absolute error of $\mu_{e_{path}} = 0.202$ ($\sigma_{e_{path}} = 0.105$) in the test set of the LUAD_SG cohort (Figure 2J). We statistically compared the MIL model's predictions and percentage tumor nuclei estimates. While the difference was not significant ($p_{comp,\rho} = 2.3 \times 10^{-1} > 0.05$) in terms of correlation coefficient, it was significant ($p_{comp,abs} = 7.3 \times 10^{-4} < 0.05$) in terms of absolute error.

### Tumor purity varies spatially within a sample: Top and bottom slides of a sample are different in tumor purity

Intra-tumor heterogeneity is a well-known phenomenon in solid cancers.[57–61] It results in therapeutic failure and drug resistance.[62] We checked whether it is observable from tumor purity predictions of the trained MIL model on the top and bottom slides of a sample. For each slide of a tumor sample with both top and bottom slides in a cohort, 100 bags are created from the slide's patches and predictions are obtained from the trained MIL model. Then, the predictions of two slides are statistically compared using the Wilcoxon signed-rank test.[53]

Figure 3C shows the box plot of p values obtained from the statistical tests in each cohort's test set. There is a significant difference between the MIL predictions on the top and bottom slides of the same tumor sample. In all cohorts, at least 75% of

**Figure 4. Incorrect size and selection of ROI might cause overestimation in percentage tumor nuclei estimates**

(A–H) For a slide of a fresh-frozen section in the TCGA LUAD cohort, (A) shows the ROI centered on a patch and consisting of 16 closest patches to that particular patch (≈ 1 $mm^2$ at the specimen level). Tumor purity corresponding to the patch is predicted over the ROI. (B) The tumor purity map for all patches within the slide. Similarly, (D) shows a slide of a FFPE section in the LUAD_SG cohort, and (E) shows its corresponding tumor purity map. (C) and (F) show example patches cropped from cancerous regions in the slides shown in (A) and (D), respectively. (G and H) To investigate the effect of the size and selection of ROI on pathologists' percentage tumor nuclei estimates, we conducted error analyses over the slides' tumor purity values by gradually extending the ROI. We calculated the slide's tumor purity as the average of top-k% of the patches with the highest scores ($k = 0, \cdots, 100$) in the tumor purity map ($k = 0$: the patch with the highest tumor purity). In different cohorts, we plotted mean absolute error versus top-k% of the patches for error analyses between slides' predictions and pathologists' percentage tumor nuclei estimates in (G) and slides' predictions and genomic tumor purity values in (H). See also Figures S4 and S6–S10.

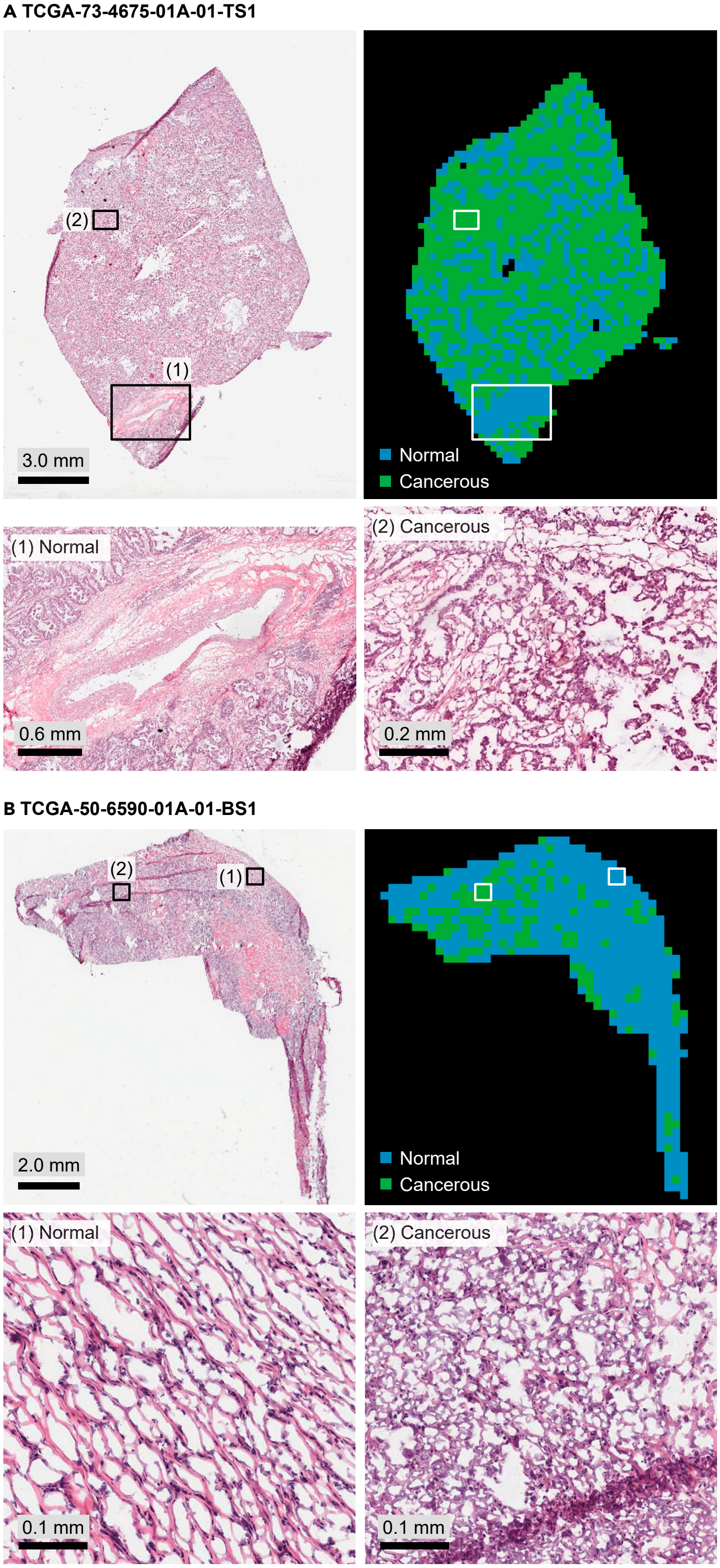


**Figure 5. Cancerous versus normal segmentation maps obtained by performing a clustering over the features extracted by the trained MIL model's feature extractor module are consistent with LUAD histopathology**

(A and B) We show H&E-stained slides, color-coded segmentation maps, and example zoom-in areas for two slides in the test set of LUAD cohort: (A) TCGA-73-4675-01A-01-TS1, and (B) TCGA-50-6590-01A-01-BS1. See also supplemental experimental procedures for details.

**Table 1. The TCGA and Singapore cohorts**

| | Tumor samples | | | Normal samples | | |
|---|---|---|---|---|---|---|
| Cohorts | Train | Validation | Test | Train | Validation | Test |
| BRCA | 559 | 185 | 185 | 76 | 27 | 30 |
| GBM | 285 | 95 | 94 | 0 | 0 | 0 |
| KIRC | 261 | 85 | 89 | 220 | 71 | 73 |
| LGG | 273 | 91 | 90 | 0 | 0 | 0 |
| LUAD | 266 | 90 | 90 | 101 | 37 | 33 |
| LUSC | 273 | 90 | 90 | 132 | 41 | 47 |
| OV | 310 | 103 | 103 | 53 | 13 | 18 |
| PRAD | 258 | 85 | 85 | 72 | 15 | 24 |
| THCA | 258 | 85 | 85 | 48 | 18 | 17 |
| UCEC | 270 | 90 | 89 | 18 | 4 | 10 |
| LUAD_SG | 107 | 36 | 36 | 0 | 0 | 0 |

In each cohort, a patient has only one tumor sample and one matching normal sample, if available. The numbers of tumor and matching normal samples in training, validation, and test sets are presented for each cohort. The data are segregated at the patient level. See also Tables S1 and S2, Figures S1–S3, and Note S1.

samples have p value $p < 1.0 \times 10^{-8}$ and at least 95% of samples have p value $p < 0.05$. Hence, we conclude that there is a variation in tumor purity between the top and bottom sections of a tumor sample; i.e., tumor purity varies spatially within the sample.

The degree of spatial variation in tumor purity is different for different cancer types (Table S6). The UCEC, LGG, and GBM cohorts had the lowest mean absolute differences ($\mu_{dabs}$) between top and bottom slides' predictions ($\mu_{dabs} \leq 0.090$); i.e., they were the most spatially homogeneous cancers among all cohorts. On the other hand, the PRAD cohort had the highest mean absolute difference ($\mu_{dabs} = 0.144$); i.e., it was the most spatially heterogeneous cancer in tumor purity.

## Predicting a sample's tumor purity using both top and bottom slides is better than using only one slide

We checked if there is a significant difference between predicting a sample's tumor purity by using both slides (top and bottom) and using only one slide. For a tumor sample with two slides in a cohort, let $p_{smpl}$ be genomic tumor purity value of the sample; $\widehat{p}_{smpl}$ be tumor purity prediction obtained from trained MIL model by using both of the slides together; $\widehat{p}_{sld1}$ and $\widehat{p}_{sld2}$ be tumor purity predictions obtained from trained MIL model for individual slides. We compared the absolute error of sample-level prediction $e_{smpl} = | \widehat{p}_{smpl} - p_{smpl}|$ and the expected value of absolute errors of slide-level predictions $e_{sld} = 0.5 * (| \widehat{p}_{sld1} - p_{smpl}| + | \widehat{p}_{sld2} - p_{smpl}|)$. We used the Wilcoxon signed-rank test[53] on the difference of $e_{smpl} - e_{sld}$ (Table S7). Note that the PRAD ($n = 21$) and UCEC ($n = 23$) cohorts were excluded from this study due to few samples with two slides.

In the test sets of BRCA, LUAD, LUSC, and OV cohorts, using both slides for tumor purity prediction gave better results in terms of absolute error (Figure 3D). However, in the test sets of GBM and LGG cohorts, there was no significant difference using both slides or one slide alone. Indeed, this is not surprising since they had the lowest mean absolute differences between the slides' predictions (Table S6); i.e., the most spatially homogeneous tumors. In fact, when both slides are the same, sample-level prediction and slide predictions would be the same.

We conclude that predicting a sample's tumor purity using both the top and bottom slides together is better than using only one of them whenever possible.

## Spatial tumor purity map analysis reveals the probable cause of pathologists' high percentage tumor nuclei estimates

Pathologists' percentage tumor nuclei estimates were generally higher than genomic tumor purity values for all TCGA cohorts in our analysis (Figure 2J and see also Figures S1 and S2). Previous studies also stated that,[5,13] but the reasons remain unclear. We hypothesized that incorrect size and selection of ROI might be the cause. We obtained tumor purity maps by our trained MIL models in different TCGA cohorts and conducted error analysis over them to test our hypothesis.

We followed the same procedure as in Smits et al.[6] to simulate pathologists' percentage tumor nuclei estimation. Tumor purity is predicted over an ROI of 1 mm × 1 mm around each patch in a slide, which corresponds to 16 patches at 20× zoom level (each patch is around 256 μm × 256 μm at the specimen level) (Figure 4A). Then, the predicted value is assigned to the patch in the tumor purity map (Figure 4B). We also obtained tumor purity maps for slides in the Singapore cohort (Figures 4D, 4E, and S6–S10).

We observed that a tumor purity map shows variation within the slide, which implies that ROI selection is crucial in pathologists' percentage tumor nuclei estimation. Since tumor purity was higher in pathologists' percentage tumor nuclei estimates, we investigated whether pathologists might have selected high tumor content regions over the slides for percentage tumor nuclei estimation. The highest prediction in a slide's tumor purity map was used as the slide's tumor purity value. Then, error analyses were conducted over the slides' tumor purity values compared with pathologists' percentage tumor nuclei estimates and genomic tumor purity values. The error analyses were repeated by gradually extending the ROI such that a slide's tumor purity was calculated as the average of top-k% of the patches with the highest scores ($k = 0, \cdots, 100$) in the slide's tumor purity map.

We discovered that the mean absolute error between the slides' predictions and pathologists' percentage tumor nuclei estimates increases as we extend the ROI to cover the lower tumor purity regions (Figure 4G). These observations suggested that pathologists may tend to select high-tumor-content regions to estimate percentage tumor nuclei. The LGG and UCEC cohorts may look exceptional with almost constant mean-absolute-error plots. However, this is expected since these two cohorts' samples have high genomic tumor purity values (Figure S1), so the variation within the slides is very low. The PRAD cohort's plot also has a different pattern than the others. It has an initial decrease and an increase in the later stages, emphasizing the importance of the ROI size. The pathologists may need to analyze a bigger ROI depending on the morphology of the tissue origin to reach a certain nuclei count while estimating percentage tumor nuclei. The PRAD may be one of them due to the glandular structure of the prostate.

Furthermore, as the ROI grows, the mean absolute error between the slides' predictions and genomic tumor purity values decreases (Figure 4H). Indeed, this is expected since our MIL models converge to their original performance of prediction over the whole slide (Figure 2). It is even more evident in the LUSC and OV cohorts. The error decreases initially but increases later since our MIL models underestimated the tumor purity compared with genomic tumor purity values in these cohorts.

### MIL model learns discriminant features for cancerous versus normal tissue histology

We explored the capability of our MIL model's feature extractor on learning discriminant features for cancerous versus normal tissue histology while being trained on sample-level genomic tumor purity labels. For each patient having both tumor and matching normal samples, features of patches cropped over the slides of the tumor and normal samples were extracted using the trained feature extractor module of the MIL model. Then, slide-level cancerous versus normal segmentation maps were obtained by performing a hierarchical clustering over the extracted feature vectors (Figure 1C and see supplemental experimental procedures for details). The resolution of segmentation was at the patch level, and each patch was around 256 μm × 256 μm at the specimen level.

In the test set of the LUAD cohort, there were 33 patients both with tumor and matching normal samples. We constructed slide-level segmentation maps for these patients (Figure 5). We observed that segmentation maps were consistent with the LUAD histopathology during the qualitative assessment of the segmentation maps. While healthy tissue components, like blood vessels, stroma regions, and normal tissue structures, were labeled normal, regions invaded by neoplastic cells were labeled cancerous. Hence, we qualitatively validated that our MIL model learned discriminant features for cancerous versus normal tissue histology in LUAD from sample-level genomic tumor purity labels without requiring pixel-level annotations from pathologists.

### MIL model successfully classifies samples into tumor versus normal

A good tumor purity predictor should be able to discriminate between tumor and normal. We checked our MIL model's performance in the tumor versus normal sample classification task. Tumor purity predictions for all samples in the test set of each cohort were obtained and a receiver operating characteristic (ROC) curve analysis was conducted. Then, the area under the ROC curve (AUC) was calculated and a 95% CI was constructed using the percentile bootstrap method.[63] Note that GBM and LGG cohorts were excluded from analysis since there were no normal slides in these cohorts.

Our MIL models successfully discriminated tumor samples from normal samples in all cohorts with AUC values greater than or equal to 0.927 (Figure 3B). We got the minimum and maximum AUC values of 0.927 (95% CI, 0.826–0.993) and 1.000 (95% CI, 1.000–1.000) on the test sets of THCA and BRCA cohorts, respectively. Note that, although we did not get a strong correlation between genomic tumor purity values and MIL predictions in the test sets of kidney renal clear cell carcinoma (KIRC) and THCA cohorts, our models successfully classified samples into tumor versus normal in these cohorts.

Furthermore, we obtained an AUC value of 0.991 (95% CI, 0.975–1.000) on the test set of the LUAD cohort. Our model outperformed the classical image processing and machine-learning-based method of Yu et al.[64] (AUC, 0.85) and the DNA plasma-based method of Sozzi et al.[65] (AUC, 0.94). Besides, our model performed on par with the deep learning model of Coudray et al.[66] (AUC, 0.993), which was trained on tumor versus normal classification, and the deep learning model of Fu et al.[67] (AUC, 0.977 with 95% CI, 0.976–0.978), which was fine-tuned on pathologists' percentage tumor nuclei estimates in a transfer learning setup. However, there is one concern about the dataset preparation methods of Coudray et al.[66] and Fu et al.[67] They obtained the datasets by segregating data either at slide level[66] or at patch level.[67] These data segregation methods might lead to a severe data leakage problem, and the models' performance might be illusory.

## DISCUSSION

Accurate tumor purity estimation is crucial for high-throughput genomic analysis. It is routinely estimated by pathologists; however, pathologists' estimates suffer from inter-observer variability and do not correlate well with genomic tumor purity values. Besides, percentage tumor nuclei estimation by pathologists is tedious and time consuming. To overcome these challenges, we developed a novel MIL model with a distribution pooling filter. It predicted tumor purity from H&E-stained histopathology slides of fresh-frozen and FFPE sections in different TCGA cohorts and a Singapore cohort, respectively. The predictions were consistent with genomic tumor purity values, and they outperformed pathologists' percentage tumor nuclei estimates in the TCGA cohorts.

Hence, our MIL models can be utilized for sample selection for high-throughput genomic analysis, which will help reduce pathologists' workload and decrease inter-observer variability. Moreover, spatially resolved tumor purity maps obtained using our MIL models can substantially contribute to a better understanding of the tumor microenvironment. Lastly, our models' predictions can be used as prognostic biomarkers to stratify patients.

### MIL model can pre-screen slides for genomic analysis

The current workflow for sample selection for genomic analysis includes screening slides of 8–12 sections, choosing the most appropriate slide, and, possibly, marking out a high tumor content region on the slide for macrodissection before extraction. This adds a heavy burden to pathologists' workload. To help pathologists, our MIL model can pre-screen the slides and suggest the best slide (with the highest predicted tumor purity) for high-throughput sequencing. Moreover, it can propose high-tumor-content regions over the slide for macrodissection via spatial tumor purity map, which is remarkably important for low-purity samples (for example, in lung cancers).

Furthermore, our MIL model's predictions can be used as a quality control metric to decide if a section has enough tumor content for sequencing or if a section requires deeper sequencing. This can avoid wasting the limited amount of tissue (especially in biopsy samples) in failed sequencing attempts.

### Genomic tumor purity values and pathologists' percentage tumor nuclei estimates are complementary

While genomic tumor purity values have recently been recognized as accurate for downstream genomic analysis after sequencing,[35] sequencing is still subject to sample selection based on pathologists' percentage tumor nuclei estimates. On the other hand, pathologists' slide reading method is inherently limited since it requires cellular-level analysis. It can give reliable results over a selected ROI, but it may not be applicable for sample-level tumor purity prediction, ideally requiring the analysis of gigapixel digital histopathology slides. Therefore, we used sample-level genomic tumor purity values as labels for training our MIL models. Now, we can use our MIL models to support pathologists for sample selection for molecular analysis by prescreening slides and proposing ROIs for further assessment.

### Spatially resolved tumor purity maps can enhance spatial omics

We obtained tumor purity maps showing the variation of tumor purity in slides using our trained MIL models (Figures 4B and 4E). They can potentially help understand the interaction of cancer cells with other tissue components (like normal epithelial, stromal, and immune cells) in the tumor microenvironment, which is a key player in tumor formation and primary determinant of therapeutic response.[68,69] Furthermore, they can enhance spatial-omics technologies.[70–73]

### Weak tumor purity labels innately necessitated an MIL approach

Previous studies based on patch-based models worked on few cancer types with relatively few patients (like 10 patients[46] or 64 patients[47,48]) since they required pixel-level annotations (rarely available). However, using genomic tumor purity values as sample-level weak labels enabled us to conduct a pan-cancer study on 10 different TCGA cohorts, where each cohort had more than 400 patients. On the other hand, unlike pixel-level annotations providing whether each cell is cancerous or normal, the genomic tumor purity of a sample tells us only the proportion of cancer cells within the sample. Therefore, training a machine learning model using weak tumor purity labels innately necessitated an MIL approach, where a sample was represented as a bag of patches from the sample's slides, and the sample's genomic tumor purity value was used as the bag's label.

### The sources of error in MIL predictions

Our MIL models successfully predicted tumor purity (Figure 2). However, they slightly deviated from the genomic tumor purity values. There may be different sources of prediction errors. While some of them can be eliminated, some are inevitable.

First, we have fewer patients in our datasets (300 patients per training set) than traditional deep learning datasets containing millions of independent samples.[74] Considering the complexity of cancer, our MIL models effectively captured features that distinguish cancerous versus normal. We also expect that the performance will improve with the increasing number of patients. Indeed, we obtained the best performance in our largest cohort of BRCA (559 patients in the training set).

Second, our MIL model uses histopathology slides from the top and bottom sections of the tumor portion. We have already shown the variation in tumor purity between the top and bottom sections of the tumor samples. Thus, for samples with only one slide, the prediction error is expected to be higher.

Third, we checked if necrosis regions inside the slides affect our MIL models' performances. For each cohort (except the LGG cohort, in which all samples have percentage necrosis of 0), Spearman's correlation coefficients between absolute errors in MIL predictions and percentage necrosis values are calculated in the test set (Table S8). There is no significant correlation ($p > 0.05$) in any cohorts except the LUSC cohort, in which we observe a low correlation of 0.253 ($p = 1.6 \times 10^{-2}$; 95% CI, 0.062, 0.432). Overall, it seems that our models can handle necrosis regions well.

Last, our model's predictions are based on morphology in H&E-stained histopathology slides. However, genomic tumor purity values were based on DNA data, and all the effects of genetic changes (so the genomic tumor purity changes) may not be observable from the slides due to the selective dyeing characteristics of H&E staining. Even some genetic changes may not manifest in morphology.[75]

### Why do MIL predictions perform better than percentage tumor nuclei estimates?

Compared with the percentage tumor nuclei estimates by pathologists, our MIL models' predictions gave a higher correlation and lower mean absolute error with genomic tumor purity values (Figure 3A). One of the primary reasons for this superiority is that the MIL models were trained directly on genomic tumor purity values, which enabled the MIL models to learn associated features.

Another reason might be that pathologists concentrate more on tumor cells than infiltrating normal cells within the tumor, which may result in missed normal tissue components. Moreover, cancer cells are usually enlarged. They occupy more space than normal tissue components, stromal cells, and infiltrating lymphocytes, which may create an implication of high tumor content.[18] Pathologists may fail to incorporate this effect in their estimates correctly and may overestimate percentage tumor nuclei. Indeed, this was the case in the cohorts we analyzed (Figures 2J, S1, and S2).

Finally, while our MIL models predict tumor purity over the whole slide, pathologists estimate the percentage tumor nuclei by analyzing some selected ROI over the slide. Therefore, the size and selection of the ROI might cause the overestimation in pathologists' percentage tumor nuclei estimates (Figure 4G).

### Limitations and future work

Our MIL models, by design, apply to any tumor sample with H&E-stained histopathology slides. We tested them on tumor samples with a broad range of tumor purity values. However, testing them on samples with percentage tumor nuclei lower than TCGA threshold would strengthen the applicability of our MIL models. It is reserved for future work.

We evaluated our MIL models on hold-out test sets to simulate real-world clinical workflow and obtained successful results. Besides, our analysis on the LUAD_SG cohort using transfer learning with minimal training for domain adaptation showed that our MIL models learned robust features for tumor purity prediction tasks. However, we could not validate our models on external cohorts due to differences between fresh-frozen and FFPE tissue preservation methods, which might further consolidate their robustness.

We qualitatively validated our spatially resolved tumor purity maps. Quantitative validation of them using spatial-omics technologies is reserved for future work, which requires recruitment of a prospective cohort, conducting spatial-omics and image analysis, and evaluating purity maps obtained from the MIL model against spatial-omics.

Lastly, our MIL models are deep learning based, and deep learning algorithms perform better with more data. Training of the models with larger cohorts would help to improve the model performance by better capturing patient-to-patient variations.

## EXPERIMENTAL PROCEDURES

### Resource availability

#### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Wing-Kin Sung (ksung@comp.nus.edu.sg).

#### Materials availability

This study did not generate new unique reagents.

#### Data and code availability

- All TCGA datasets are publicly available. Manifest files are provided with original code to download H&E-stained digital histopathology slides using GDC Data Transfer Tool. Genomic tumor purity values were downloaded from https://gdc.cancer.gov/about-data/publications/pancanatlas under filename TCGA_mastercalls.abs_ tables_JSedit.fixed.txt. They are also given in Data S1. For the Singapore cohort, genomic tumor purity values and representative histological images are publicly available from OncoSG (https://src.gisapps.org/OncoSG/) under dataset Lung Adenocarcinoma (GIS, 2019).
- All original code has been deposited at Zenodo under the https://doi.org/10.5281/zenodo.5606981 and is publicly available as of the date of publication. The repository provides a detailed step-by-step explanation, from downloading H&E-stained digital histopathology slides to obtaining spatially resolved tumor purity maps (SRTPMs).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

### Datasets

We downloaded H&E-stained histopathology slides of fresh-frozen sections in 10 different cohorts in TCGA (Table 1). We selected these cohorts since they have more than 400 patients with both histopathology slides and corresponding genomic sequencing data in TCGA. Each patient had a tumor sample, and some patients also had matching normal samples.

In TCGA, primary tumor samples and matching normal samples (adjacent non-neoplastic solid tissue or blood) were collected at the Tissue Source Sites (TSSs) from patients who had received no prior treatment (chemotherapy or radiotherapy) for their disease. Collected samples were frozen and shipped overnight to the Biospecimen Core Resource (BCR) for TCGA while maintaining a temperature less than $-180^\circ$C.[76]

At the BCR, each frozen sample was cut into portions.[77] Then, two glass slides (sometimes only one) were prepared by cutting sections 4–6 μm thick from the top and bottom of a portion[78] and staining with H&E.[79] Based on the information from the BCR (via personal communication), these slides were scanned at 40× magnification using an Aperio XT slide scanner. A board-certified pathologist reviewed the slides. Upon passing pathology review, the remaining portion without any tumor enrichment was sent for genomic analysis.[80] In other words, the (top and bottom) slides and the portion sent for genomic analysis were immediate neighbors.

During review (personal communication), a pathologist estimated (1) percentage tumor nuclei and percentage of all other nuclei, which add up to 100%; and (2) percentage cellular tumor, percentage normal, percentage stroma, and percentage necrosis, which add up to 100%. Percentage tumor nuclei in each slide of a frozen tumor section was estimated by evaluating at least 10 specimen fields (excluding necrosis regions) via the digital slide viewer. Tumor portions with percentage tumor nuclei of $\geq$60% and percentage necrosis of $\leq$20% were accepted to the study and sent for genomic analysis.[76] Besides, pathologists confirmed from slides of frozen normal sections that the adjacent normal tissues (if available) were free of tumor cells.

We also collected H&E-stained histopathology slides of FFPE sections in an East Asian cohort consisting of 179 lung adenocarcinoma patients in Singapore. In the Singapore cohort, only one slide was prepared for each tumor sample from the top section of the tissue used for sequencing, and there were no normal samples. All the slides of FFPE sections were prepared, stained, and scanned at 40× magnification using The Philips IntelliSite Pathology Solution (Koninklijke Philips, The Netherlands) in the same laboratory in Singapore.

In each cohort, we randomly segregated the data at the patient level (i.e., slides from the same patient should be in the same set) into training (60%), validation (20%), and test (20%) sets (Table 1), which had similar tumor purity distributions (Figures S1 and S2). Note that segregating data at the patient level is crucial to prevent data leakage while training machine learning models.[54] The training set was used to train the machine learning model, the validation set was used to choose the best model, and the test set was held out as unseen data for evaluation of the best model. The list of patients and slides in each set are given in Document S2.

### MIL model

Our novel MIL model consists of three modules: feature extractor module, MIL pooling filter, and bag-level representation transformation module (Figure 1A). We use neural networks to implement the feature extractor module and the bag-level representation transformation module to parameterize the learning process fully (see supplemental experimental procedures for details). We use our novel distribution pooling filter as the MIL pooling filter. It is more expressive than standard pooling filters (like mean and maximum pooling) regarding the amount of information captured while obtaining bag-level representations.[81] Given a bag of patches, the feature extractor module extracts a feature vector for each patch inside the bag. Then, thanks to its superiority, the distribution pooling filter obtains a strong bag-level representation by estimating the marginal distributions of the extracted features. Finally, the bag-level representation transformation module predicts tumor purity. This system of neural network modules is trained end-to-end using samples' genomic tumor purity values as labels.

### Training of MIL models

To prepare machine learning datasets, tissue regions inside histopathology slides were detected by applying OTSU thresholding. Over the tissue regions, non-overlapping 512 × 512 patches at 20× zoom level (specimen-level pixel size, 0.5 μm × 0.5 μm) were cropped.

During training, we used a bag of patches cropped from a sample's top and bottom slides as the input and the sample's tumor purity value obtained from genomic sequencing data by ABSOLUTE[19] as the ground-truth label (Figure S1). At each epoch, one bag per sample is created on the fly by randomly selecting 200 patches from the sample's patches. We also used the matching normal samples whenever available to enable our model to capture the information related to normal tissue histology. We assigned a tumor purity value of 0.0 to a matching normal sample as the ground-truth label. Note that there were no normal samples in the Singapore cohort.

We initialized the models' weights randomly and trained them end-to-end using the ADAM optimizer with a learning rate of 0.0001 and L2 regularization on the weights with a weight decay of 0.0005. The batch size was 1. We used absolute error as the loss function and employed early stopping based on loss in the validation set to avoid overfitting.

### Predicting tumor purity of a sample

We created 100 bags for each sample in the test set and obtained tumor purity predictions from the trained model. Each bag was created by randomly selecting 200 patches from the available patches cropped from the sample's (top and bottom) slides. We used the average of 100 predictions as the sample's tumor purity prediction during performance evaluation.

### Statistical analysis

We obtained 95% CIs for Spearman's rank correlation coefficients and area under the ROC curves using the percentile bootstrap method.[63] To compare the performance of two methods (our MIL models' predictions and pathologists' percentage tumor nuclei estimates), we used the Fisher's z transformation-based method of Meng et al.[55] on Spearman's rank correlation coefficients and Wilcoxon signed-rank test[53] on absolute error values.

All statistical tests were two sided and statistical significance was considered when $p < 0.05$. We used scipy.stats (v1.4.1) python library for statistical tests.

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100399.

## ACKNOWLEDGMENTS

This work is partly supported by the Biomedical Research Council of the Agency for Science, Technology and Research, Singapore, and the National University of Singapore, Singapore. We thank Jay Bowen (Principal Investigator at the BCR) for his detailed explanations about how the slides are prepared, stained, and scanned at the BCR for TCGA and providing detailed information about the pathology review process of the slides at the BCR for sample selection for downstream analyses. We thank Valerie Yang for constructive comments on the manuscript.

## AUTHOR CONTRIBUTIONS

M.U.O., W.K.S., and H.K.L. conducted the machine learning study. J.C., E.R., A.N.K., J.J.S.A., W.Z., and A.J.S. performed genomic analysis. D.S.W.T. designed the clinical study of the Singapore cohort. A.T. and X.M.C. selected the cases in the Singapore cohort and conducted the histopathology review. A.J. retrieved the slides and estimated percentage tumor nuclei values. S.Y.H. scanned the slides and exported digital slide images. T.K.H.L. reviewed the slides and contributed to the methods discussion and manuscript review.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 3, 2021
Revised: September 7, 2021
Accepted: November 3, 2021
Published: February: 11, 2022

## REFERENCES

1. Schuster, S.C. (2008). Next-generation sequencing transforms today's biology. Nat. Methods *5*, 16–18.
2. Xuan, J., Yu, Y., Qing, T., Guo, L., and Shi, L. (2013). Next-generation sequencing in the clinic: promises and challenges. Cancer Lett. *340*, 284–295.
3. Jennings, L.J., Arcila, M.E., Corless, C., Kamel-Reid, S., Lubin, I.M., Pfeifer, J., Temple-Smolkin, R.L., Voelkerding, K.V., and Nikiforova, M.N. (2017). Guidelines for validation of next-generation sequencing–based oncology panels: a joint consensus recommendation of the Association for Molecular Pathology and College of American Pathologists. J. Mol. Diagn. *19*, 341–365.
4. Whiteside, T. (2008). The tumor microenvironment and its role in promoting tumor growth. Oncogene *27*, 5904–5912.
5. Aran, D., Sirota, M., and Butte, A.J. (2015). Systematic pan-cancer analysis of tumour purity. Nat. Commun. *6*, 1–12.
6. Smits, A.J., Kummer, J.A., De Bruin, P.C., Bol, M., Van Den Tweel, J.G., Seldenrijk, K.A., Willems, S.M., Offerhaus, G.J.A., De Weger, et al. (2014). The estimation of tumor cell percentage for molecular testing by pathologists is not accurate. Mod. Pathol. *27*, 168–174.
7. Kim, J., Park, W.-Y., Kim, N.K., Jang, S.J., Chun, S.-M., Sung, C.-O., Choi, J., Ko, Y.-H., Choi, Y.-L., Shim, H.S., et al. (2017). Good laboratory standards for clinical next-generation sequencing cancer panel tests. J. Pathol. Transl. Med. *51*, 191.
8. Patel, N.M., Jo, H., Eberhard, D.A., Yin, X., Hayward, M.C., Stein, M.K., Hayes, D.N., and Grilley-Olson, J.E. (2019). Improved tumor purity metrics in next-generation sequencing for clinical practice: the integrated interpretation of neoplastic cellularity and sequencing results (IINCaSe) approach. Appl. Immunohistochem. Mol. Morphol. *27*, 764–772.
9. Elloumi, F., Hu, Z., Li, Y., Parker, J.S., Gulley, M.L., Amos, K.D., and Troester, M.A. (2011). Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. BMC Med. Genomics *4*, 1–12.
10. Isella, C., Terrasi, A., Bellomo, S.E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C., et al. (2015). Stromal contribution to the colorectal cancer transcriptome. Nat. Genet. *47*, 312–319.
11. Zhang, W., Feng, H., Wu, H., and Zheng, X. (2017). Accounting for tumor purity improves cancer subtype classification from DNA methylation data. Bioinformatics *33*, 2651–2657.
12. Rhee, J.-K., Jung, Y.C., Kim, K.R., Yoo, J., Kim, J., Lee, Y.-J., Ko, Y.H., Lee, H.H., Cho, B.C., and Kim, T.-M. (2018). Impact of tumor purity on immune gene expression and clustering analyses across multiple cancer types. Cancer Immunol. Res. *6*, 87–97.
13. Haider, S., Tyekucheva, S., Prandi, D., Fox, N.S., Ahn, J., Xu, A.W., Pantazi, A., Park, P.J., Laird, P.W., Sander, C., et al. (2020). Systematic assessment of tumor purity and its clinical implications. JCO Precis. Oncol. *4*, 995–1005.
14. Cheng, J., He, J., Wang, S., Zhao, Z., Yan, H., Guan, Q., Li, J., Guo, Z., and Ao, L. (2020). Biased influences of low tumor purity on mutation detection in cancer. Front. Mol. Biosci. *7*, 533196.
15. Zhang, C., Cheng, W., Ren, X., Wang, Z., Liu, X., Li, G., Han, S., Jiang, T., and Wu, A. (2017). Tumor purity as an underlying key factor in glioma. Clin. Cancer Res. *23*, 6279–6291.
16. Mao, Y., Feng, Q., Zheng, P., Yang, L., Liu, T., Xu, Y., Zhu, D., Chang, W., Ji, M., Ren, L., et al. (2018). Low tumor purity is associated with poor prognosis, heavy mutation burden, and intense immune phenotype in colon cancer. Cancer Manag. Res. *10*, 3569.
17. Gong, Z., Zhang, J., and Guo, W. (2020). Tumor purity as a prognosis and immunotherapy relevant feature in gastric cancer. Cancer Med. *9*, 9052–9063.
18. Mikubo, M., Seto, K., Kitamura, A., Nakaguro, M., Hattori, Y., Maeda, N., Miyazaki, T., Watanabe, K., Murakami, H., Tsukamoto, T., et al. (2020). Calculating the tumor nuclei content for comprehensive cancer panel testing. J. Thorac. Oncol. *15*, 130–137.
19. Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. *30*, 413–421.
20. Oesper, L., Mahmoody, A., and Raphael, B.J. (2013). Theta: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. *14*, R80.
21. Chen, H., Bell, J.M., Zavala, N.A., Ji, H.P., and Zhang, N.R. (2015). Allele-specific copy number profiling by next-generation DNA sequencing. Nucleic Acids Res. *43*, e23.
22. Yu, G., Zhang, B., Bova, G.S., Xu, J., Shih, I.-M., and Wang, Y. (2011). BACOM: in silico detection of genomic deletion types and correction of normal cell contamination in copy number data. Bioinformatics *27*, 1473–1480.

23. Zhang, B., Hou, X., Yuan, X., Shih, I.-M., Zhang, Z., Clarke, R., Wang, R.R., Fu, Y., Madhavan, S., Wang, Y., et al. (2014). AISAIC: a software suite for accurate identification of significant aberrations in cancers. Bioinformatics *30*, 431–433.

24. Yuan, X., Bai, J., Zhang, J., Yang, L., Duan, J., Li, Y., and Gao, M. (2018). CONDEL: detecting copy number variation and genotyping deletion zygosity from single tumor samples using sequence data. IEEE/ACM Trans. Comput. Biol. Bioinform. *17*, 1141–1153.

25. Yuan, X., Yu, J., Xi, J., Yang, L., Shang, J., Li, Z., and Duan, J. (2019). CNV_IFTV: an isolation forest and total variation-based detection of CNVs from short-read sequencing data. IEEE/ACM Trans. Comput. Biol. Bioinform. *18*, 539–549.

26. Van Loo, P., Nordgard, S.H., Lingjærde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., et al. (2010). Allele-specific copy number analysis of tumors. Proc. Natl. Acad. Sci. U S A *107*, 16910–16915.

27. Andor, N., Harness, J.V., Mueller, S., Mewes, H.W., and Petritsch, C. (2014). EXPANDS: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics *30*, 50–60.

28. Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C. (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann. Oncol. *26*, 64–70.

29. Su, X., Zhang, L., Zhang, J., Meric-Bernstam, F., and Weinstein, J.N. (2012). PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics *28*, 2265–2266.

30. Larson, N.B., and Fridley, B.L. (2013). PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. Bioinformatics *29*, 1888–1889.

31. Yuan, X., Li, Z., Zhao, H., Bai, J., and Zhang, J. (2020). Accurate inference of tumor purity and absolute copy numbers from high-throughput sequencing data. Front. Genet. *11*, 458.

32. Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres- Garcia, W., Treviño, V., Shen, H., Laird, P.W., Levine, D.A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nat. Commun. *4*, 1–11.

33. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D.E., and Gfeller, D. (2017). Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. eLife *6*, e26476.

34. Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat. Biotechnol. *37*, 773–782.

35. Li, Y., Umbach, D.M., Bingham, A., Li, Q.-J., Zhuang, Y., and Li, L. (2019). Putative biomarkers for predicting tumor sample purity based on gene expression data. BMC Genomics *20*, 1–12.

36. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., and Kelsey, K.T. (2012). DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics *13*, 86.

37. Zheng, X., Zhao, Q., Wu, H.-J., Li, W., Wang, H., Meyer, C.A., Qin, Q.A., Xu, H., Zang, C., Jiang, P., et al. (2014). MethylPurify: tumor purity deconvolution and differential methylation detection from single tumor DNA methylomes. Genome Biol. *15*, 1–13.

38. Zhang, N., Wu, H.-J., Zhang, W., Wang, J., Wu, H., and Zheng, X. (2015). Predicting tumor purity from methylation microarray data. Bioinformatics *31*, 3401–3405.

39. Zheng, X., Zhang, N., Wu, H.-J., and Wu, H. (2017). Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. Genome Biol. *18*, 1–14.

40. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.-Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

41. Akbani, R., Ng, P.K.S., Werner, H.M., Shahmoradgoli, M., Zhang, F., Ju, Z., Liu, W., Yang, J.-Y., Yoshihara, K., Li, J., et al. (2014). A pan-cancer proteomic perspective on the cancer genome atlas. Nat. Commun. *5*, 1–15.

42. Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

43. Chen, J., Yang, H., Teo, A.S.M., Amer, L.B., Sherbaf, F.G., Tan, C.Q., Alvarez, J.J.S., Lu, B., Lim, J.Q., Takano, A., et al. (2020). Genomic landscape of lung adenocarcinoma in East Asians. Nat. Genet. *52*, 177–186.

44. Viray, H., Coulter, M., Li, K., Lane, K., Madan, A., Mitchell, K., Schalper, K., Hoyt, C., and Rimm, D.L. (2014). Automated objective determination of percentage of malignant nuclei for mutation testing. Appl. Immunohistochem. Mol. Morphol. *22*, 363.

45. Hamilton, P.W., Wang, Y., Boyd, C., James, J.A., Loughrey, M.B., Hougton, J.P., Boyle, D.P., Kelly, P., Maxwell, P., McCleary, D., et al. (2015). Automated tumor analysis for molecular profiling in lung cancer. Oncotarget *6*, 27938.

46. Azimi, V., Chang, Y.H., Thibault, G., Smith, J., Tsujikawa, T., Kukull, B., Jensen, B., Corless, C., Margolin, A., and Gray, J.W. (2017). Breast cancer histopathology image analysis pipeline for tumor purity estimation. In 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) (IEEE), pp. 1137–1140.

47. Pei, Z., Cao, S., Lu, L., and Chen, W. (2019). Direct cellularity estimation on breast cancer histopathology images using transfer learning. Comput. Math. Methods Med. *2019*, 3041250.

48. Rakhlin, A., Tiulpin, A., Shvets, A.A., Kalinin, A.A., Iglovikov, V.I., and Nikolenko, S.. (2019). Breast tumor cellularity assessment using deep neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0.

49. Greene, C., O'Doherty, E., Abdullahi Sidi, F., Bingham, V., Fisher, N.C., Humphries, M.P., Craig, S.G., Harewood, L., McQuaid, S., Lewis, C., et al. (2021). The potential of digital image analysis to determine tumor cell content in biobanked formalin-fixed, paraffin-embedded tissue samples. Biopreserv. Biobank. *19*, 324–331.

50. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. *25*, 1301–1309.

51. Tomita, N., Abdollahi, B., Wei, J., Ren, B., Suriawinata, A., and Hassanpour, S. (2019). Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw. open *2*, e1914645.

52. Oner, M.U., Lee, H.K., and Sung, W.-K.. (2020). Weakly supervised clustering by exploiting unique class count. In International Conference on Learning Representations.

53. Wilcoxon, F. (1992). Individual comparisons by ranking methods. In Breakthroughs in Statistics (Springer), pp. 196–202.

54. Oner, M.U., Cheng, Y.-C., Lee, H.K., and Sung, W.-K. (2020). Training machine learning models on patient level data segregation is crucial in practical clinical applications. medRxiv. https://doi.org/10.1101/2020.04.23.20076406.

55. Meng, X.-L., Rosenthal, R., and Rubin, D.B. (1992). Comparing correlated correlation coefficients. Psychol. Bull. *111*, 172.

56. Spencer, D.H., Sehn, J.K., Abel, H.J., Watson, M.A., Pfeifer, J.D., and Duncavage, E.J. (2013). Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. J. Mol. Diagn. *15*, 623–633.

57. Jamal-Hanjani, M., Wilson, G.A., McGranahan, N., Birkbak, N.J., Watkins, T.B., Veeriah, S., Shafi, S., Johnson, D.H., Mitter, R., Rosenthal, R., et al. (2017). Tracking the evolution of non–small-cell lung cancer. N. Engl. J. Med. *376*, 2109–2121.

58. Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M., Papaemmanuil, E., Brewer, D.S., Kallio, H.M., Högnäs, G., Annala, M.,

et al. (2015). The evolutionary history of lethal metastatic prostate cancer. Nature *520*, 353.

59. Gerlinger, M., Horswell, S., Larkin, J., Rowan, A.J., Salm, M.P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C.R., et al. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat. Genet. *46*, 225.

60. Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., et al. (2015). A big bang model of human colorectal tumor growth. Nat. Genet. *47*, 209.

61. Zhai, W., Lim, T.K.-H., Zhang, T., Phang, S.-T., Tiang, Z., Guan, P., Ng, M.-H., Lim, J.Q., Yao, F., Li, Z., et al. (2017). The spatial organization of intra-tumour heterogeneity and evolutionary trajectories of metastases in hepatocellular carcinoma. Nat. Commun. *8*, 4565.

62. McGranahan, N., and Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell *27*, 15–26.

63. Efron, B. (1992). Bootstrap methods: another look at the jackknife. In Breakthroughs in Statistics (Springer), pp. 569–593.

64. Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., and Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. *7*, 1–10.

65. Sozzi, G., Conte, D., Leon, M., Cirincione, R., Roz, L., Ratcliffe, C., Roz, E., Cirenei, N., Bellomi, M., Pelosi, G., et al. (2003). Quantification of free circulating DNA as a diagnostic marker in lung cancer. J. Clin. Oncol. *21*, 3902–3908.

66. Coudray, N., Ocampo, P.S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., Moreira, A.L., Razavian, N., and Tsirigos, A. (2018). Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat. Med. *24*, 1559–1567.

67. Fu, Y., Jung, A.W., Torne, R.V., Gonzalez, S., Vöhringer, H., Shmatko, A., Yates, L.R., Jimenez-Linan, M., Moore, L., and Gerstung, M. (2020). Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nat. Cancer *1*, 1–11. https://doi.org/10.1038/s43018-020-0085-8.

68. Hanahan, D., and Coussens, L.M. (2012). Accessories to the crime: functions of cells recruited to the tumor microenvironment. Cancer Cell *21*, 309–322.

69. Junttila, M.R., and de Sauvage, F.J. (2013). Influence of tumour microenvironment heterogeneity on therapeutic response. Nature *501*, 346.

70. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*. https://doi.org/10.1126/science.aaa6090.

71. Svensson, V., Teichmann, S.A., and Stegle, O. (2018). SpatialDE: identification of spatially variable genes. Nat. Methods *15*, 343–346.

72. Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A., Dulac, C., et al. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science *362*, eaau5324.

73. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., III, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.

74. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In 2009 IEEE conference on computer vision and pattern recognition (Ieee), pp. 248–255.

75. Caiado, F., Silva-Santos, B., and Norell, H. (2016). Intra-tumour heterogeneity–going beyond genetics. FEBS J. *283*, 2245–2258.

76. Lazar, A.J., McLellan, M.D., Bailey, M.H., Miller, C.A., Appelbaum, E.L., Cordes, M.G., Fronick, C.C., Fulton, L.A., Fulton, R.S., Mardis, E.R., et al. (2017). Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. Cell *171*, 950–965.

77. (2020). L020 standard operating procedure (SOP) for sectioning and portioning frozen tissue samples in logistics. https://brd.nci.nih.gov/brd/sop/show/1447.

78. (2020). H006 standard operating procedure (SOP) for sectioning frozen tissue samples in histology. https://brd.nci.nih.gov/brd/sop/show/1442.

79. (2020). H001 standard operating procedure (SOP) for hematoxylin and eosin (H&E) staining coverslipping. https://brd.nci.nih.gov/brd/sop/show/1421.

80. (2020). L016 standard operating procedure (SOP) for preparing frozen tissue for molecular analysis. https://brd.nci.nih.gov/brd/sop/show/1445.

81. Oner, M.U., Kye-Jet, J.M.S., Lee, H.K., and Sung, W.-K. (2020). Studying the effect of MIL pooling filters on MIL tasks. arXiv, arXiv:2006.01561 https://arxiv.org/abs/2006.01561.

Patterns, Volume *3*

## Supplemental information

# Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study

**Mustafa Umit Oner, Jianbin Chen, Egor Revkov, Anne James, Seow Ye Heng, Arife Neslihan Kaya, Jacob Josiah Santiago Alvarez, Angela Takano, Xin Min Cheng, Tony Kiat Hon Lim, Daniel Shao Weng Tan, Weiwei Zhai, Anders Jacobsen Skanderup, Wing-Kin Sung, and Hwee Kuan Lee**

# SUPPLEMENTAL ITEMS

**Table S1: The number of samples, slides, and patches in each TCGA cohort.** Each patient has only one tumor sample and one normal sample if available. Note that "tumor slide" and "normal slides" refer to the slides of tumor samples and normal samples, respectively. Similarly, "tumor patches" and "normal patches" refer to patches cropped over "tumor slides" and "normal slides", respectively. Related to Table 1.

| | # samples | | | # slides | | | # patches | | |
|---|---|---|---|---|---|---|---|---|---|
| | normal | tumor | total | normal | tumor | total | normal | tumor | total |
| BRCA | 133 | 929 | 1,062 | 312 | 1,280 | 1,592 | 84,196 | 710,446 | 794,642 |
| GBM | 0 | 474 | 474 | 0 | 917 | 917 | 0 | 618,649 | 618,649 |
| KIRC | 364 | 435 | 799 | 454 | 841 | 1295 | 466,883 | 655,625 | 1,122,508 |
| LGG | 0 | 454 | 454 | 0 | 625 | 625 | 0 | 347,065 | 347,065 |
| LUAD | 171 | 446 | 617 | 200 | 694 | 894 | 108,876 | 490,401 | 599,277 |
| LUSC | 220 | 453 | 673 | 333 | 714 | 1,047 | 166,181 | 544,778 | 710,959 |
| OV | 84 | 516 | 600 | 142 | 1,031 | 1,173 | 72,385 | 1,122,620 | 1,195,005 |
| PRAD | 111 | 428 | 539 | 111 | 535 | 646 | 75,798 | 338,120 | 413,918 |
| THCA | 83 | 428 | 511 | 83 | 443 | 526 | 30,234 | 199,275 | 229,509 |
| UCEC | 32 | 449 | 481 | 34 | 589 | 623 | 17,359 | 314,624 | 331,983 |

**Table S2: The number of samples in different genomic tumor purity and percent tumor nuclei groups (<10% , 10-25%, 25-50%, and ≥50%).** Related to Table 1.

| | genomic tumor purity | | | | percent tumor nuclei | | | |
|---|---|---|---|---|---|---|---|---|
| | <10% | 10-25% | 25-50% | ≥50% | <10% | 10-25% | 25-50% | ≥50% |
| BRCA | 1 | 44 | 247 | 637 | 0 | 0 | 10 | 919 |
| GBM | 0 | 6 | 43 | 425 | 1 | 0 | 4 | 469 |
| KIRC | 0 | 7 | 158 | 270 | 0 | 0 | 0 | 435 |
| LGG | 0 | 10 | 54 | 390 | 0 | 0 | 3 | 451 |
| LUAD | 1 | 40 | 225 | 180 | 0 | 0 | 5 | 441 |
| LUSC | 0 | 40 | 188 | 225 | 0 | 0 | 5 | 448 |
| OV | 0 | 1 | 28 | 487 | 0 | 0 | 0 | 516 |
| PRAD | 0 | 24 | 117 | 287 | 0 | 0 | 7 | 421 |
| THCA | 0 | 3 | 50 | 375 | 0 | 0 | 0 | 428 |
| UCEC | 0 | 6 | 43 | 400 | 0 | 0 | 6 | 443 |

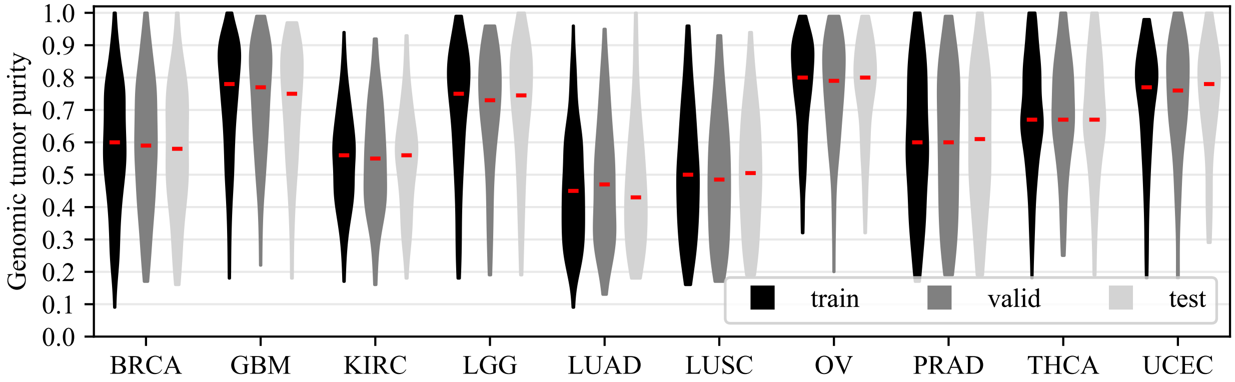


**Figure S1: Violin plots of genomic tumor purity values (obtained using ABSOLUTE[1]) in the training, validation, and test sets of each TCGA cohort.** Related to Table 1.

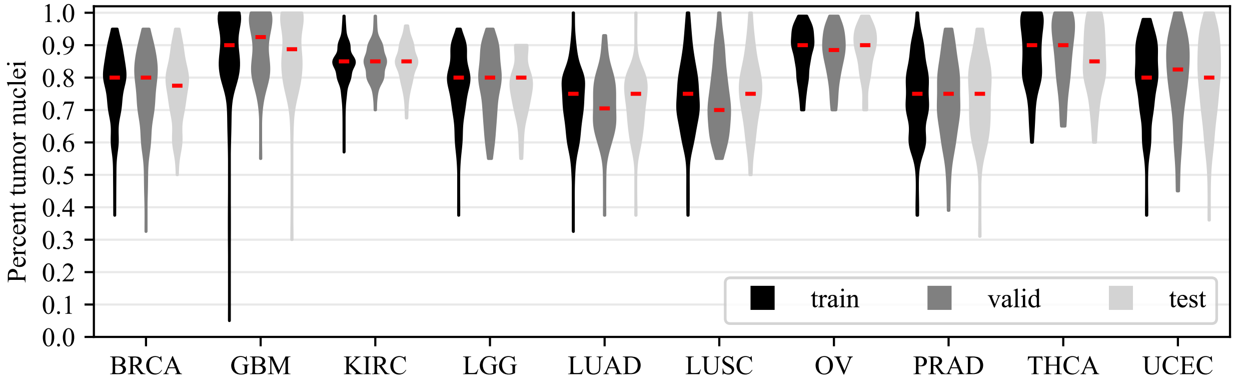


**Figure S2: Violin plots of percent tumor nuclei values (collected from TCGA data portal) in each TCGA cohort's training, validation, and test sets.** Related to Table 1.

**Table S3: Comparison of methods based on Spearman's correlation coefficients in the test sets of different cohorts.** Spearman's correlation coefficients between genomic tumor purity values and MIL predictions ($\rho_{mil}$) and genomic tumor purity values and pathologists' percent tumor nuclei estimates ($\rho_{path}$) in the test sets of different cohorts are calculated for only the tumor samples. Then, they are compared using the method in Meng et al.[2]. The number of tumor samples (n), Spearman's correlation coefficients together with calculated p-values ($P_{\rho_{mil}}$ and $P_{\rho_{path}}$) and 95% confidence intervals ($CI_{\rho_{mil}}$ and $CI_{\rho_{path}}$), and calculated p-values in statistical tests ($P_{comp}$) are presented. Note that if the calculated correlation in any method is not significant (i.e., $P_{\rho_{mil}} > 5.0e-02$ or $P_{\rho_{path}} > 5.0e-02$), the statistical test is not conducted. It is indicated by 'x'. The best methods are highlighted in bold. Related to Figure 2 and Figure 3A.

| | | MIL prediction | | | Pathologist's estimate | | | Comparison |
|---|---|---|---|---|---|---|---|---|
| | n | $\rho_{mil}$ | $P_{\rho_{mil}}$ | $CI_{\rho_{mil}}$ | $\rho_{path}$ | $P_{\rho_{path}}$ | $CI_{\rho_{path}}$ | $P_{comp}$ |
| BRCA | 185 | **0.655** | **4.6e-24** | **0.547 - 0.743** | 0.299 | 3.6e-05 | 0.162 - 0.429 | **1.4e-07** |
| GBM | 94 | 0.572 | 1.7e-09 | 0.389 - 0.721 | 0.104 | 3.2e-01 | -0.102 - 0.309 | x |
| LGG | 90 | 0.418 | 4.1e-05 | 0.226 - 0.574 | 0.201 | 5.7e-02 | -0.029 - 0.392 | x |
| LUAD | 90 | **0.515** | **2.1e-07** | **0.320 - 0.660** | 0.255 | 1.5e-02 | 0.036 - 0.448 | **1.2e-02** |
| LUSC | 90 | 0.467 | 3.5e-06 | 0.280 - 0.627 | 0.324 | 1.8e-03 | 0.118 - 0.503 | 1.7e-01 |
| OV | 103 | **0.581** | **1.3e-10** | **0.429 - 0.711** | 0.328 | 7.1e-04 | 0.132 - 0.518 | **9.4e-03** |
| PRAD | 85 | 0.424 | 5.3e-05 | 0.224 - 0.597 | 0.293 | 6.5e-03 | 0.074 - 0.504 | 2.0e-01 |
| UCEC | 89 | **0.579** | **2.7e-09** | **0.408 - 0.720** | 0.344 | 9.8e-04 | 0.139 - 0.531 | **2.6e-02** |

**Table S4: Spearman's correlation coefficients.** Spearman's correlation coefficients between (i) genomic tumor purity values from ABSOLUTE[1] (ABS) and MIL predictions (MIL), (ii) genomic tumor purity values from ESTIMATE[3] (EST) and MIL predictions, and (iii) genomic tumor purity values from ABSOLUTE and genomic tumor purity values from ESTIMATE are calculated for the tumor samples having corresponding values in the test sets. The number of tumor samples (n), correlation coefficients ($\rho$) together with calculated p-values ($P$) and 95% confidence intervals ($CI$) are presented.

| | | ABS vs. MIL | | | EST vs. MIL | | | EST vs ABS | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | n | $\rho$ | $P$ | $CI$ | $\rho$ | $P$ | $CI$ | $\rho$ | $P$ | $CI$ |
| BRCA | 186 | 0.655 | 4.6e-24 | 0.547 - 0.743 | 0.519 | 4.0e-14 | 0.401 - 0.615 | 0.611 | 2.4e-20 | 0.496 - 0.709 |
| GBM | 22 | 0.610 | 3.3e-03 | 0.162 - 0.882 | 0.528 | 1.4e-02 | 0.112 - 0.821 | 0.732 | 1.6e-04 | 0.439 - 0.898 |
| LGG | 91 | 0.418 | 4.1e-05 | 0.226 - 0.574 | 0.139 | 1.9e-01 | -0.076 - 0.333 | 0.352 | 6.6e-04 | 0.142 - 0.531 |
| LUAD | 91 | 0.515 | 2.1e-07 | 0.320 - 0.660 | 0.546 | 2.5e-08 | 0.391 - 0.674 | 0.645 | 6.7e-12 | 0.468 - 0.779 |
| LUSC | 88 | 0.447 | 1.4e-05 | 0.264 - 0.611 | 0.350 | 8.9e-04 | 0.150 - 0.524 | 0.628 | 7.5e-11 | 0.466 - 0.752 |
| OV | 52 | 0.596 | 3.9e-06 | 0.360 - 0.768 | 0.579 | 8.5e-06 | 0.323 - 0.763 | 0.708 | 6.2e-09 | 0.532 - 0.824 |
| PRAD | 86 | 0.424 | 5.3e-05 | 0.224 - 0.597 | 0.319 | 3.0e-03 | 0.109 - 0.496 | 0.447 | 1.8e-05 | 0.241 - 0.634 |
| UCEC | 40 | 0.574 | 1.3e-04 | 0.284 - 0.788 | 0.400 | 1.2e-02 | 0.057 - 0.695 | 0.580 | 1.1e-04 | 0.291 - 0.789 |

**Table S5: Comparison of methods based on absolute errors in the test sets of different cohorts.** Absolute errors between genomic tumor purity values and MIL predictions ($e_{mil}$) and genomic tumor purity values and pathologists' percent tumor nuclei estimates ($e_{path}$) in the test sets of different cohorts are calculated for only the tumor samples. Then, they are compared using the Wilcoxon signed-rank test[4]. The number of tumor samples (n), mean absolute errors ($\mu_{e_{mil}}$ and $\mu_{e_{path}}$) together with standard deviations ($\sigma_{e_{mil}}$ and $\sigma_{e_{path}}$), median absolute errors ($m_{e_{mil}}$ and $m_{e_{path}}$) together with interquartile ranges ($IQR_{e_{mil}}$ and $IQR_{e_{path}}$), and calculated p-values in the statistical tests ($P_{comp}$) are presented. The best methods are highlighted in bold. Related to Figure 2 and Figure 3A.

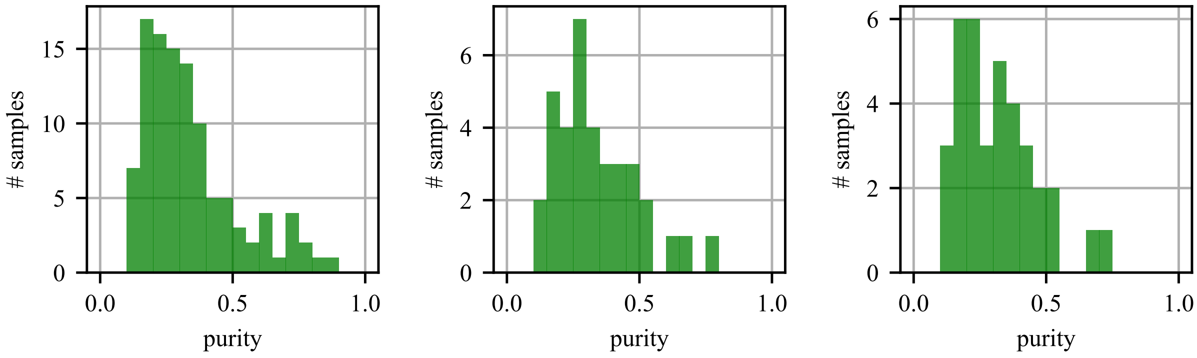
| | | MIL prediction | | | | Pathologist's estimate | | | | Comp. |
|---|---|---|---|---|---|---|---|---|---|---|
| | n | $\mu_{e_{mil}}$ | $\sigma_{e_{mil}}$ | $m_{e_{mil}}$ | $IQR_{e_{mil}}$ | $\mu_{e_{path}}$ | $\sigma_{e_{path}}$ | $m_{e_{path}}$ | $IQR_{e_{path}}$ | $P_{comp}$ |
| BRCA | 185 | **0.116** | **0.097** | **0.104** | **0.043 - 0.159** | 0.220 | 0.147 | 0.200 | 0.105 - 0.310 | **2.5e-13** |
| GBM | 94 | **0.113** | **0.106** | **0.074** | **0.046 - 0.142** | 0.195 | 0.158 | 0.145 | 0.080 - 0.260 | **2.1e-07** |
| LGG | 90 | 0.136 | 0.119 | 0.105 | 0.052 - 0.188 | 0.152 | 0.122 | 0.130 | 0.060 - 0.200 | 5.4e-02 |
| LUAD | 90 | **0.132** | **0.109** | **0.112** | **0.060 - 0.175** | 0.280 | 0.151 | 0.275 | 0.170 - 0.395 | **3.9e-09** |
| LUSC | 90 | **0.148** | **0.122** | **0.125** | **0.054 - 0.196** | 0.266 | 0.150 | 0.250 | 0.140 - 0.375 | **5.8e-06** |
| OV | 103 | **0.105** | **0.091** | **0.086** | **0.043 - 0.127** | 0.136 | 0.126 | 0.110 | 0.030 - 0.190 | **1.6e-02** |
| PRAD | 85 | **0.173** | **0.154** | **0.130** | **0.068 - 0.240** | 0.204 | 0.141 | 0.180 | 0.090 - 0.285 | **1.4e-02** |
| UCEC | 89 | **0.109** | **0.120** | **0.072** | **0.027 - 0.142** | 0.132 | 0.124 | 0.100 | 0.040 - 0.170 | **1.4e-02** |

## Note S1: Singapore Cohort

Singapore cohort consists of 179 lung adenocarcinoma patients having East Asian ancestry. Each patient has one tumor sample, and one slide is prepared from each tumor sample (except one sample in the training set). The slides are prepared from formalin-fixed paraffin-embedded sections (FFPE). On the contrary to FFPE sections in the Singapore cohort, slides in the TCGA cohorts are prepared from fresh-frozen sections. These two tissue preservation methods are quite different from each other. While the FFPE method preserves morphology better and is the routine in histopathology, the fresh-frozen method preserves nucleic acids better and is preferred for molecular analysis[5]. The number of samples, slides and patches in the training, validation and test sets of the Singapore cohort are presented below.

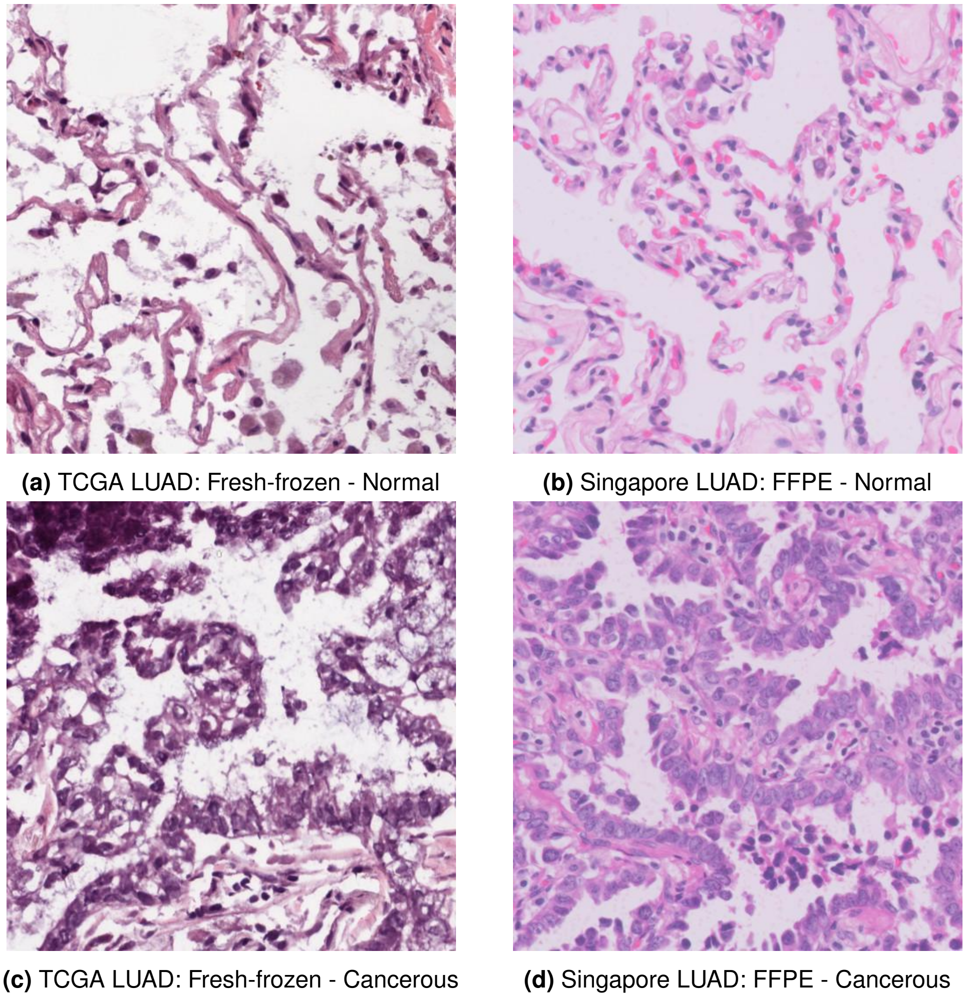
**Singapore cohort: the number of samples, slides, and patches. Note that each patient has only one tumor sample. Related to Table 1.**

| | # samples | | | # slides | | | # patches | | |
|---|---|---|---|---|---|---|---|---|---|
| dataset | normal | tumor | total | normal | tumor | total | normal | tumor | total |
| training | 0 | 107 | 107 | 0 | 108 | 108 | 0 | 525,961 | 525,961 |
| validation | 0 | 36 | 36 | 0 | 36 | 36 | 0 | 190,971 | 190,971 |
| test | 0 | 36 | 36 | 0 | 36 | 36 | 0 | 182,383 | 182,383 |
| all | 0 | 179 | 179 | 0 | 180 | 180 | 0 | 899,315 | 899,315 |



**(a)** Training **(b)** Validation **(c)** Test

**Figure S3: Singapore cohort: genomic tumor purity histograms for (a) training, (b) validation, and (c) test sets. Related to Table 1.**



**(a)** TCGA LUAD: Fresh-frozen - Normal

**(b)** Singapore LUAD: FFPE - Normal

**(c)** TCGA LUAD: Fresh-frozen - Cancerous

**(d)** Singapore LUAD: FFPE - Cancerous

**Figure S4: Example patches cropped from slides of fresh-frozen and formalin-fixed paraffin-embedded (FFPE) sections.** (**a, c**) A normal patch and a cancerous patch cropped from slides of fresh-frozen sections in the TCGA LUAD cohort. (**b, d**) A normal patch and a cancerous patch cropped from slides of FFPE sections in the Singapore LUAD cohort. Related to Figure 4.

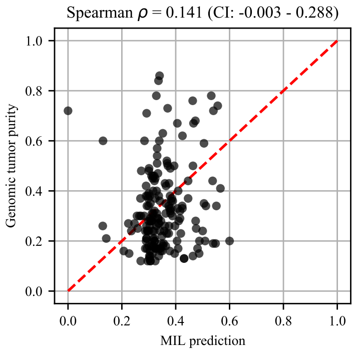


**Figure S5: External validation on Singapore cohort.** We checked the performance of the TCGA LUAD model directly on the Singapore LUAD cohort (with n=179 tumor samples) used as an external validation set. Scatter plot of genomic tumor purity vs. MIL model prediction. Diagonal red dotted line shows the y=x line.

**Table S6: Statistics of the absolute difference between the predictions of a tumor sample's top and bottom slides.** In the test set of each cohort, for a tumor sample with two slides, the absolute difference ($d_{abs}$) between the tumor purity predictions of the slides is calculated. Then, the number of tumor samples with two slides (n), the mean absolute difference ($\mu_{d_{abs}}$), the standard deviation of the absolute difference ($\sigma_{d_{abs}}$), the median absolute difference ($m_{d_{abs}}$), and the interquartile range ($IQR_{d_{abs}}$) are presented. Related to Figure 3C.

| | n | $\mu_{d_{abs}}$ | $\sigma_{d_{abs}}$ | $m_{d_{abs}}$ | $IQR_{d_{abs}}$ |
|---|---|---|---|---|---|
| BRCA | 73 | 0.101 | 0.106 | 0.063 | 0.031 - 0.115 |
| GBM | 90 | 0.090 | 0.083 | 0.068 | 0.016 - 0.141 |
| LGG | 31 | 0.086 | 0.089 | 0.054 | 0.023 - 0.139 |
| LUAD | 44 | 0.100 | 0.110 | 0.059 | 0.023 - 0.125 |
| LUSC | 52 | 0.106 | 0.123 | 0.062 | 0.030 - 0.144 |
| OV | 102 | 0.125 | 0.156 | 0.080 | 0.032 - 0.150 |
| PRAD | 21 | 0.144 | 0.189 | 0.086 | 0.027 - 0.134 |
| UCEC | 23 | 0.063 | 0.056 | 0.042 | 0.021 - 0.089 |

**Table S7: Comparing the absolute errors of sample-level predictions and the expected value of the absolute errors of slide-level predictions in the test sets of different cohorts.** In the test set of each cohort, for a tumor sample with two slides, the absolute errors between genomic tumor purity values and sample-level MIL predictions ($e_{smpl}$) and the expected value of absolute errors between genomic tumor purity values and slide-level MIL predictions ($e_{sld}$) are calculated. Then, the number of samples with two slides (n), the mean absolute errors ($\mu_{e_{smpl}}$ and $\mu_{e_{sld}}$) together with standard deviations ($\sigma_{e_{smpl}}$ and $\sigma_{e_{sld}}$), the median absolute errors ($m_{e_{smpl}}$ and $m_{e_{sld}}$) together with interquartile ranges ($IQR_{e_{smpl}}$ and $IQR_{e_{sld}}$), and the calculated p-values in the statistical tests ($P_{comp}$) are presented. Note that the PRAD (n=21) and UCEC (n=23) cohorts were excluded from this study due to few samples with two slides. The best methods are highlighted in bold. Related to Figure 3D.

| | | Sample level | | | | Slide level | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | n | $\mu_{e_{smpl}}$ | $\sigma_{e_{smpl}}$ | $m_{e_{smpl}}$ | $IQR_{e_{smpl}}$ | $\mu_{e_{sld}}$ | $\sigma_{e_{sld}}$ | $m_{e_{sld}}$ | $IQR_{e_{sld}}$ | $P_{comp}$ |
| BRCA | 73 | **0.114** | **0.082** | **0.092** | **0.043 - 0.166** | 0.126 | 0.073 | 0.129 | 0.060 - 0.171 | **2.8e-03** |
| GBM | 90 | 0.115 | 0.107 | 0.076 | 0.046 - 0.145 | 0.118 | 0.096 | 0.089 | 0.062 - 0.161 | 7.1e-01 |
| LGG | 31 | 0.178 | 0.149 | 0.146 | 0.100 - 0.218 | 0.168 | 0.152 | 0.106 | 0.067 - 0.198 | 5.6e-01 |
| LUAD | 44 | **0.118** | **0.102** | **0.084** | **0.050 - 0.168** | 0.138 | 0.102 | 0.121 | 0.067 - 0.181 | **3.7e-04** |
| LUSC | 52 | **0.124** | **0.092** | **0.109** | **0.039 - 0.168** | 0.150 | 0.096 | 0.143 | 0.085 - 0.201 | **1.7e-03** |
| OV | 102 | **0.106** | **0.091** | **0.086** | **0.043 - 0.128** | 0.135 | 0.100 | 0.105 | 0.073 - 0.176 | **5.0e-03** |

**Table S8: Spearman's correlation coefficients between absolute errors in MIL predictions and percent necrosis values ($\rho$) are calculated in the test set of each cohort.** The number of samples (n), correlation coefficients together with calculated p-values (P) and 95% confidence intervals (95% CI) are presented for tumor samples only. There is no significant correlation (P>0.05) in any cohorts except LUSC, in which the correlation is 0.253 (P=1.6e-02 < 0.05). The LGG cohort is excluded from analysis since all samples have percent necrosis of 0.

| | n | $\rho$ | 95% CI | P |
|---|---|---|---|---|
| BRCA | 185 | 0.089 | -0.054, 0.236 | 2.3e-01 |
| GBM | 94 | -0.040 | -0.232, 0.150 | 7.0e-01 |
| LUAD | 90 | 0.034 | -0.187, 0.267 | 7.5e-01 |
| LUSC | 90 | 0.253 | 0.062, 0.432 | 1.6e-02 |
| OV | 103 | 0.044 | -0.157, 0.236 | 6.6e-01 |
| PRAD | 85 | -0.050 | -0.262, 0.170 | 6.5e-01 |
| UCEC | 89 | -0.023 | -0.230, 0.187 | 8.3e-01 |

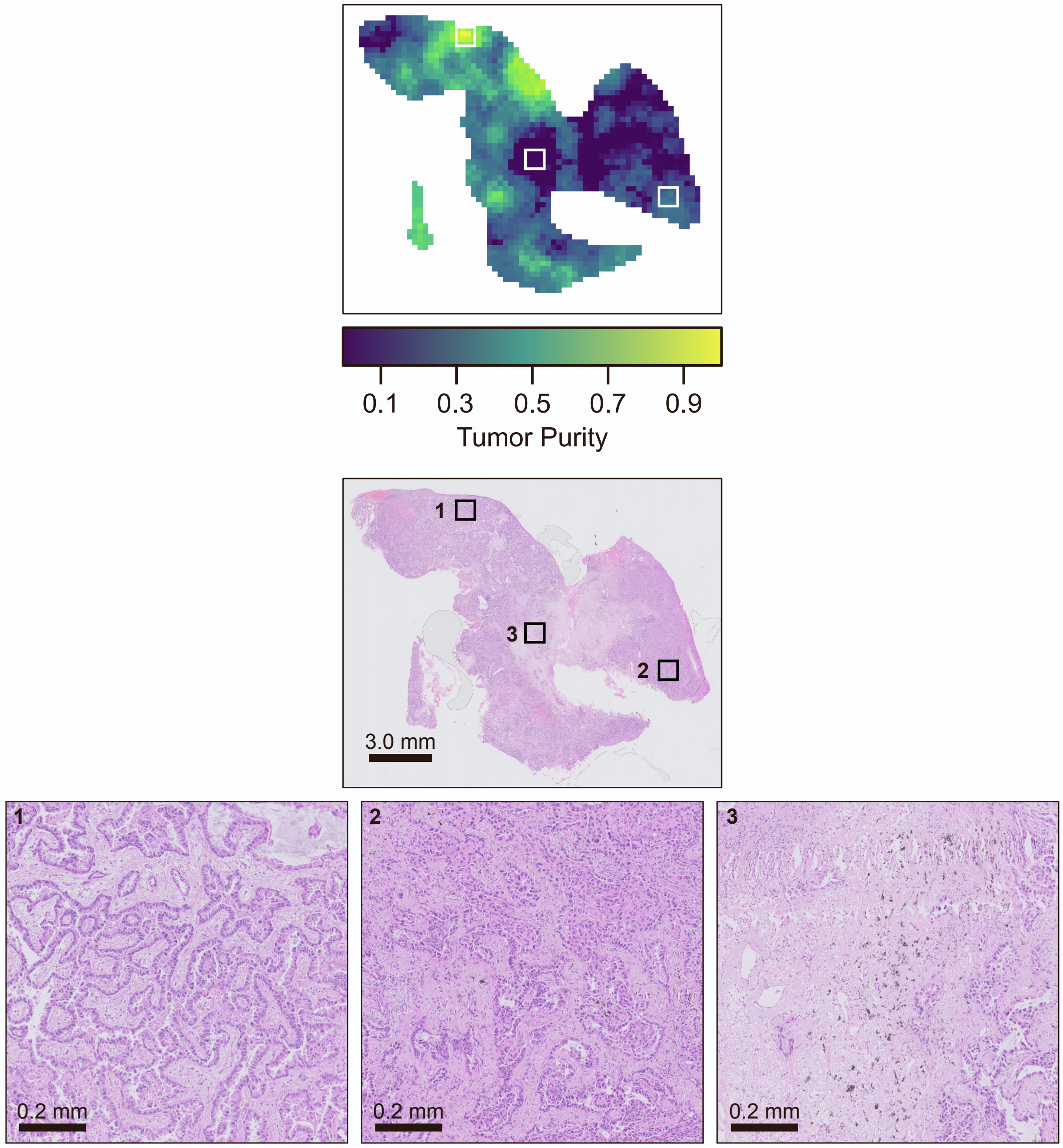


**Figure S6: Tumor purity map for A186 in the Singapore Cohort.** Genomic tumor purity was 0.340 and our MIL model predicted tumor purity as 0.339, so the absolute error was 0.001. Related to Figure 4.

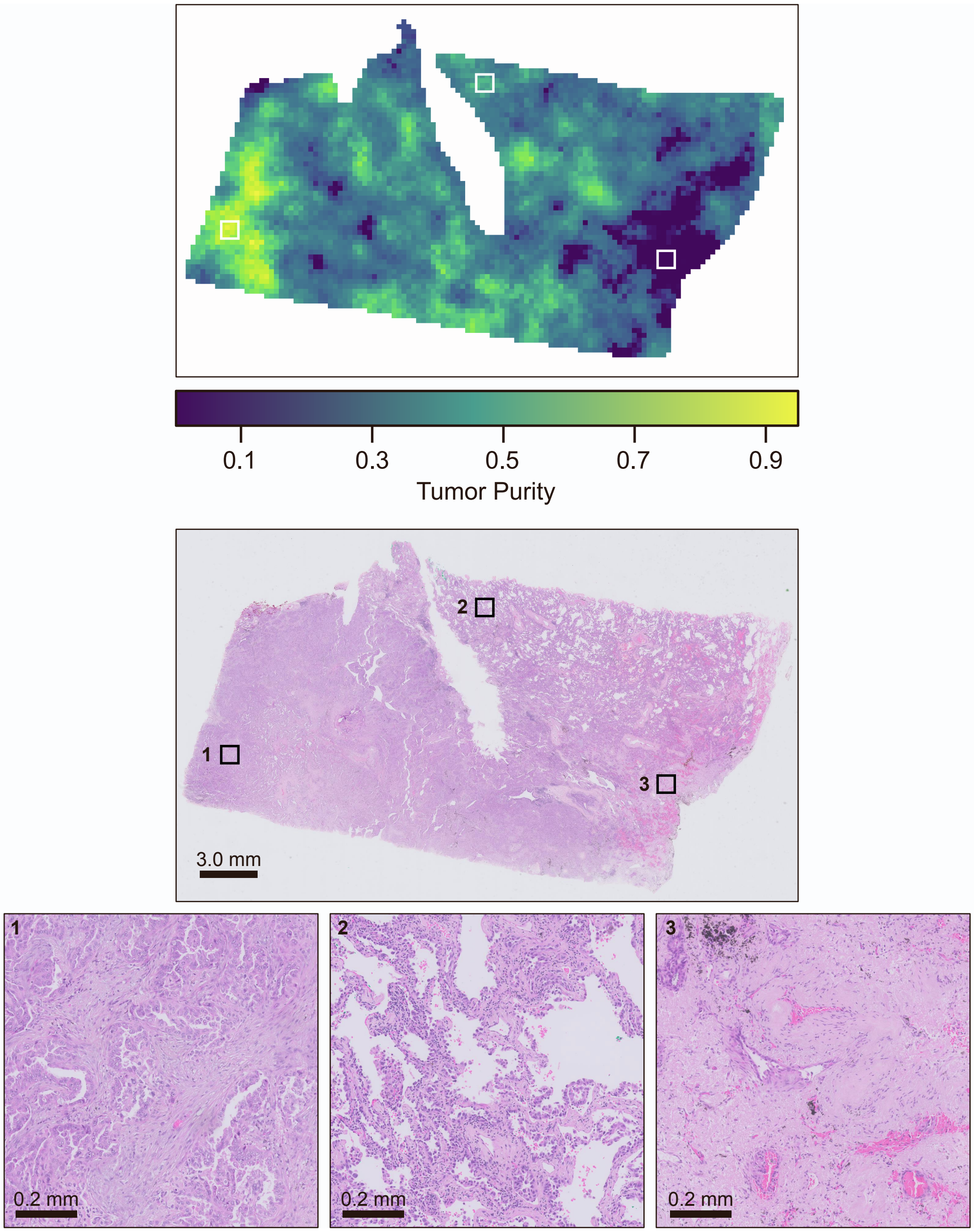


**Figure S7: Tumor purity map for A537 in the Singapore Cohort.** Genomic tumor purity was 0.420 and our MIL model predicted tumor purity as 0.380, so the absolute error was 0.04. Related to Figure 4.

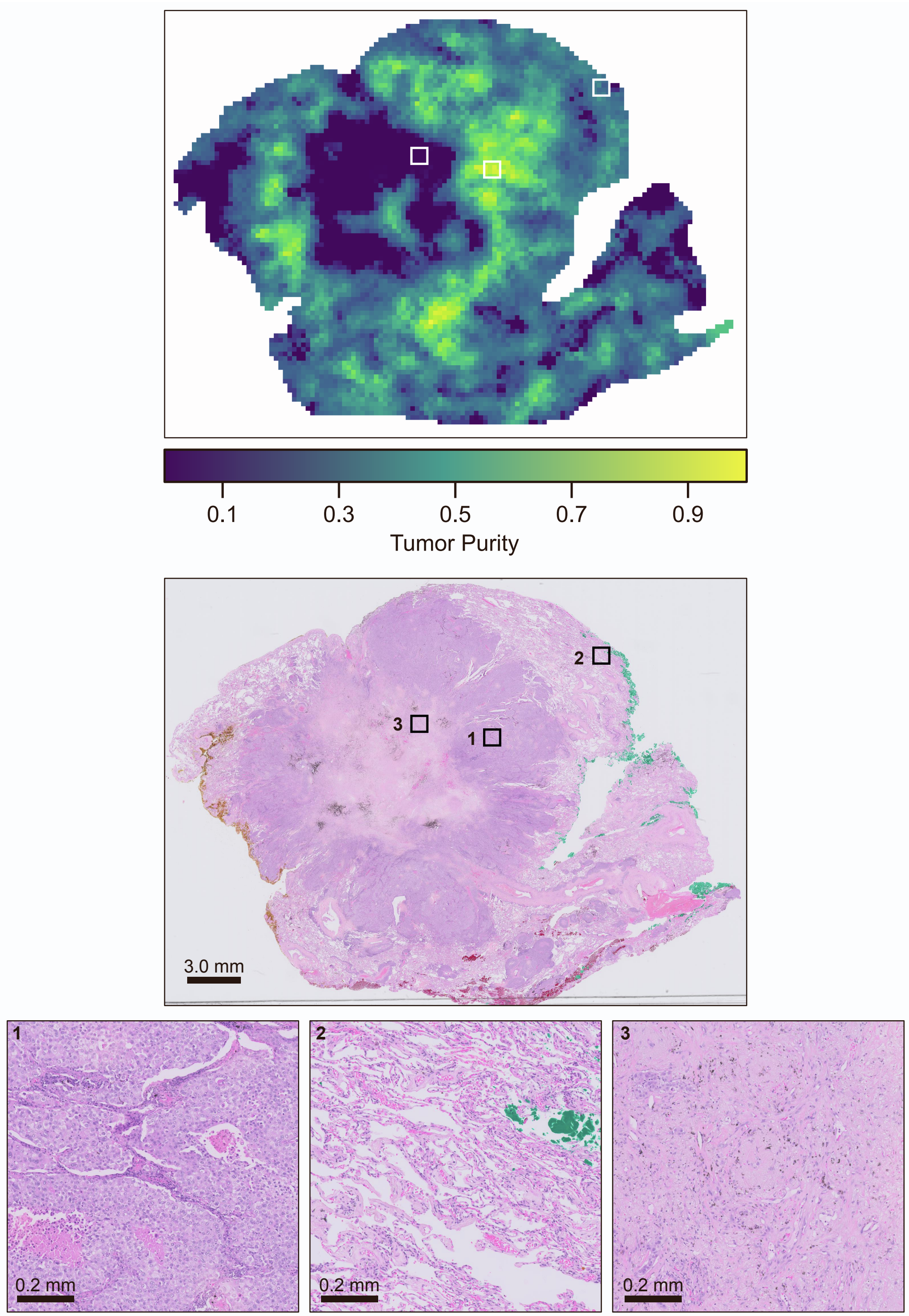


**Figure S8: Tumor purity map for A143 in the Singapore Cohort.** Genomic tumor purity was 0.240 and our MIL model predicted tumor purity as 0.339, so the absolute error was 0.099. Related to Figure 4.

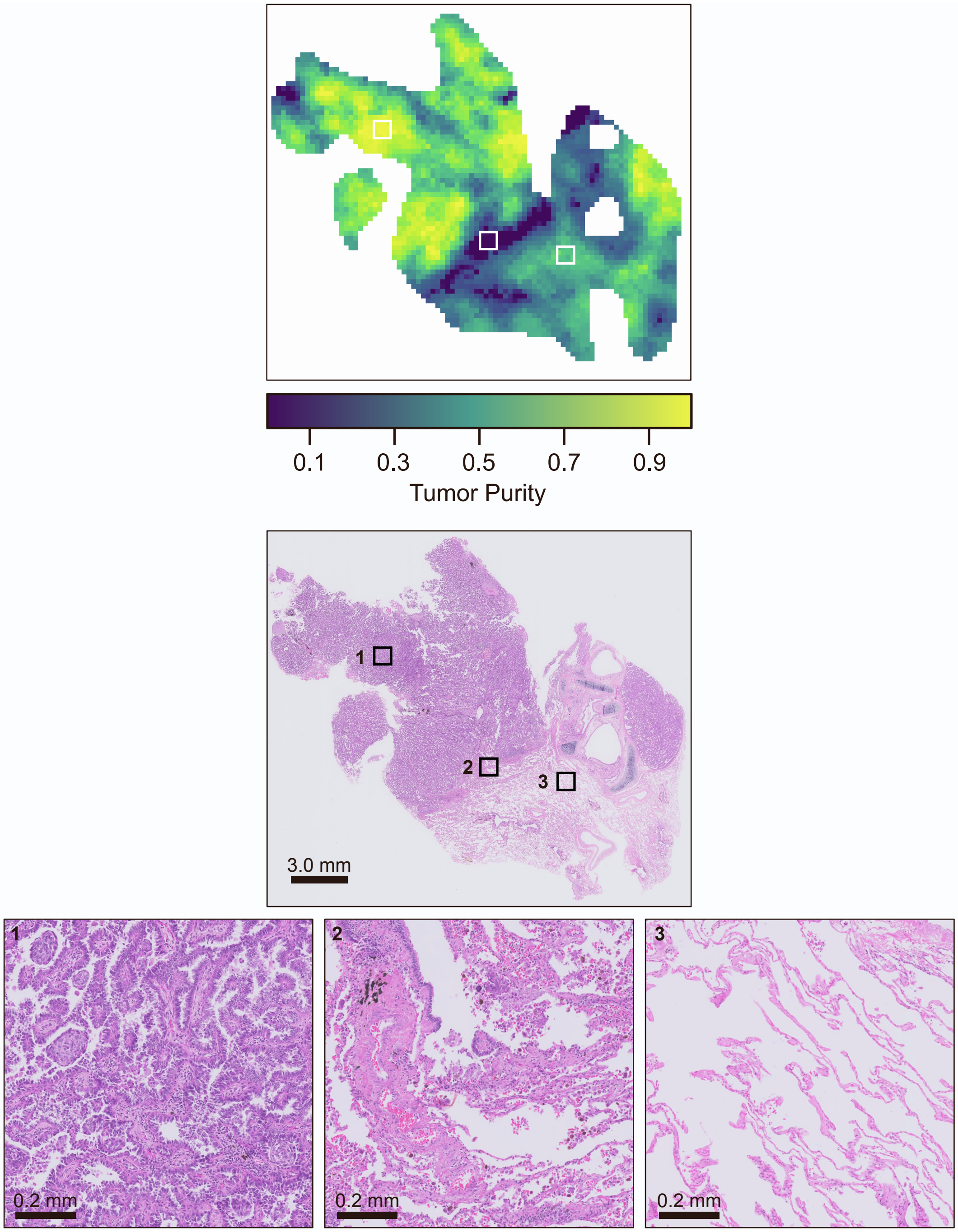


**Figure S9: Tumor purity map for A219 in the Singapore Cohort.** Genomic tumor purity was 0.410 and our MIL model predicted tumor purity as 0.584, so the absolute error was 0.174. Related to Figure 4.

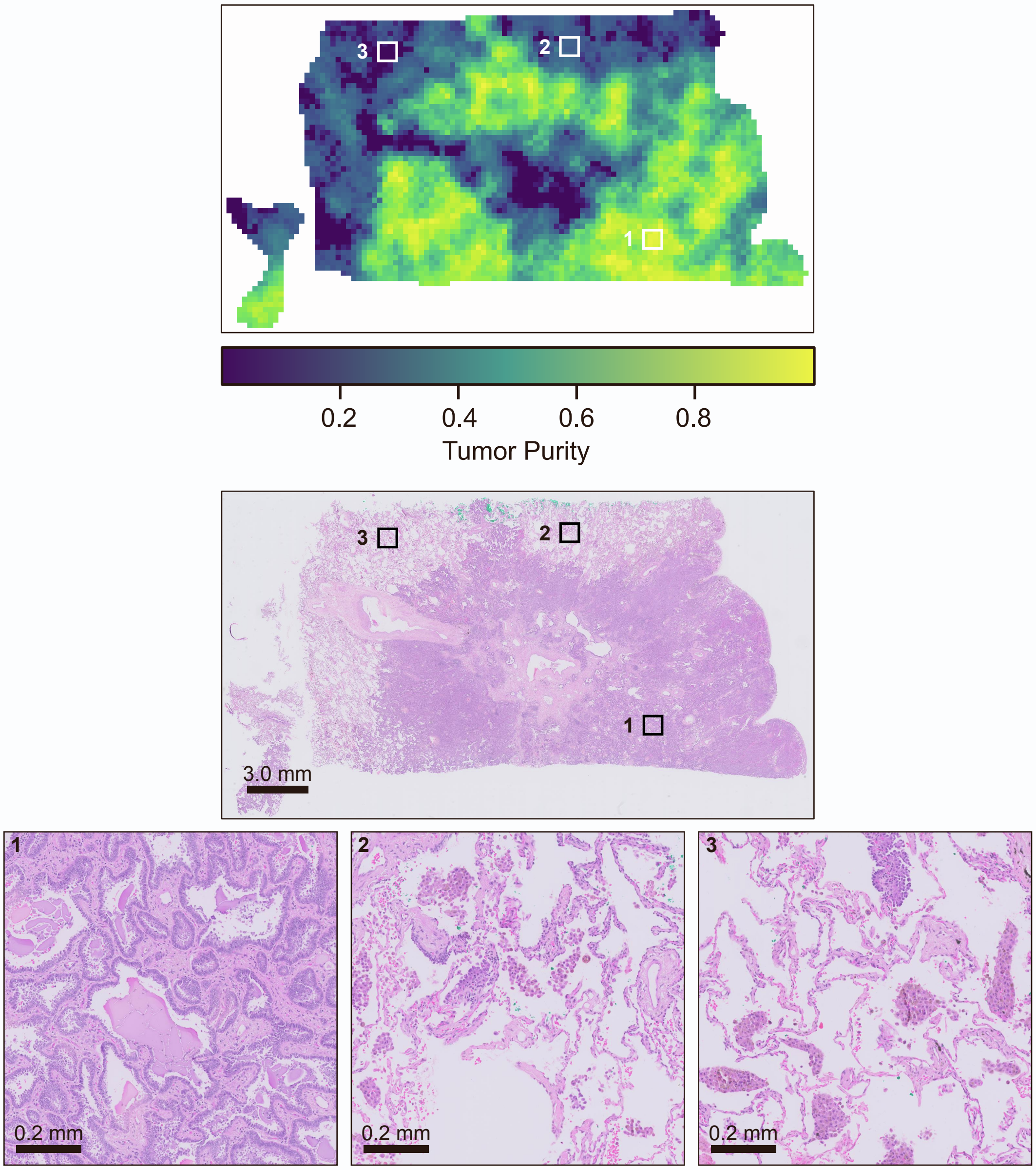


**Figure S10: Tumor purity map for A126 in the Singapore Cohort.** Genomic tumor purity was 0.160 and our MIL model predicted tumor purity as 0.527, so the absolute error was 0.367. Related to Figure 4.

# SUPPLEMENTAL EXPERIMENTAL PROCEDURES

## MIL Framework

### Problem formulation and notation

Let $\mathcal{D}$ be a MIL dataset such that for each $(X, Y) \in \mathcal{D}$, $X = \{x_1, x_2, \cdots, x_N\} \subseteq \mathcal{I}$ and $Y \in \mathcal{Y}$, where $\mathcal{I}$ is the instance space, and $\mathcal{Y}$ is the bag label space. Note that we fix the number of instances in a bag to N for clarity of notation, yet our formulation is also valid for bags with the variable number of instances.

Given any pair $(X, Y) \in \mathcal{D}$, our objective is to predict bag label $Y$ for a given bag of instances $X$. Here, a bag label $Y$ is the genomic tumor purity of a sample, and a bag $X$ is a collection of cropped patches over the sample's slides. Let $\hat{Y}$ be the predicted bag label of $X$. To obtain $\hat{Y}$, we designed a novel MIL framework consisting of three stages.

The first stage is a *feature extractor* module $\theta_{\text{feature}} : \mathcal{I} \rightarrow \mathcal{F}$, where $\mathcal{F}$ is the feature space. For each $x_i \in X$, the *feature extractor* module takes $x_i$ as input, extracts $J$ features and outputs a feature vector: $\boldsymbol{f}_{x_i} = \theta_{\text{feature}}(x_i) = [f^1_{x_i}, f^2_{x_i}, \cdots, f^J_{x_i}] \in \mathcal{F}$, where $f^j_{x_i} \in \mathbb{R}$ is the $j^{th}$ feature value and $\mathcal{F} = \mathbb{R}^J$. Let $\boldsymbol{F}_X = [\boldsymbol{f}_{x_1}, \boldsymbol{f}_{x_2}, \cdots, \boldsymbol{f}_{x_N}] \in \mathbb{R}^{JN}$ be feature matrix constructed from extracted feature vectors such that $i^{th}$ column corresponds to $\boldsymbol{f}_{x_i}$.

The second stage is a *MIL pooling filter* module $\theta_{\text{filter}} : \mathbb{R}^{JN} \rightarrow \mathcal{H}$, where $\mathcal{H}$ is the bag-level representation space. The *MIL pooling filter* module takes the feature matrix $\boldsymbol{F}_X$ as input and aggregates the extracted feature vectors to obtain a bag-level representation: $\boldsymbol{h}_X = \theta_{\text{filter}}(\boldsymbol{F}_X) \in \mathcal{H}$.

The last stage is a *bag-level representation transformation* module $\theta_{\text{transform}} : \mathcal{H} \rightarrow \mathcal{Y}$. It transforms the bag-level representation into the predicted bag label: $\hat{Y} = \theta_{\text{transform}}(\boldsymbol{h}_X)$.

We use neural networks to implement $\theta_{\text{feature}}$ and $\theta_{\text{transform}}$ so that we can fully parameterize the learning process. For $\theta_{\text{filter}}$, we use our novel 'distribution' pooling filter. This system of neural networks is end-to-end trainable.

### Distribution Pooling Filter

Our previous study[6] defined the family of distribution-based pooling filters as: Given a feature matrix $\boldsymbol{F}_X = [f^j_{x_i} | f^j_{x_i} \in \mathbb{R},\ i = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, J]$ obtained from a bag $X = \{x_1, x_2, \cdots, x_N\}$, its bag level representation is obtained by estimating a marginal distribution over each extracted feature. Let $\tilde{p}^j_X : \mathbb{R} \rightarrow \mathbb{R}^+ \cup \{0\}$ be the estimated marginal distribution obtained over $j^{th}$ extracted feature and $\tilde{p}^j_X \in \mathbb{P}$ where $\mathbb{P}$ is the set of all possible marginal distributions. $\tilde{p}^j_X$ is calculated by using kernel density estimation[7], which employs a Gaussian kernel with standard deviation $\sigma$, as shown in the Eq. 1. Each instance has two attention based weights, feature weight $\alpha_i$ and kernel weight $\beta_i$, obtained from neural network modules. Hence, the bag level representation $\boldsymbol{h}_X = [\tilde{p}^j_X\ |\tilde{p}^j_X \in \mathbb{P}, j = 1, 2, \cdots, J] \in \mathcal{H}$ where $\mathcal{H} = \mathbb{P}^J$. Note that the estimated marginal distributions are uniformly binned during training neural network models for computational purposes.

$$\tilde{p}^j_X(v) = \sum_{i=1}^{N} \beta_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(v - \alpha_i f^j_{x_i}\right)^2}\ \forall_{j=1,2,\cdots,J} \tag{1}$$

Our previous study formally proved that the distribution-based pooling filters are more expressive than the point estimate-based counterparts (like max and mean pooling) regarding the amount of information captured while obtaining bag-level representations[6]. Then, we empirically showed that models with distribution-based pooling filters perform equal or better than that with point estimate-based pooling filters on distinct real-world MIL tasks.

In this study, we used standard deviation of $\sigma = 0.05$ and the estimated marginal distributions were uniformly binned into 21 bins. Note that attention weights in 'distribution' pooling were fixed to $\alpha_i = 1\ \forall_i$ and $\beta_i = \frac{1}{N}\ \forall_i$ where $N$ is the number of instances per bag.

### Neural network architectures and hyper-parameters

We used a ResNet18[8] model as the *feature extractor* module and a three-layer multi-layer-perceptron as the *bag-level representation transformation* module.

During the training of the models, we prepared bags on the go. A bag was created by randomly sampling 200 patches (instances) from all available patches previously cropped over a sample's slides. The patch size was $512 \times 512$. Data augmentation (random cropping with a size of $299 \times 299$ and random horizontal/vertical flipping) was also applied on the patches. We extracted 128 features for each instance inside the bag.

The architecture and list of hyper-parameters used in MIL models are given below.

**Neural network architecture and list of hyper-parameters used in the MIL models.**

| | |
|---|---|
| Architecture | input - $299 \times 299 \times 3$<br>ResNet18 (128 nodes in the last fc layer)<br>'distribution' pooling<br>Dropout(0.5)<br>fc-384 + ReLU<br>Dropout(0.5)<br>fc-192 + ReLU<br>Dropout(0.5)<br>fc-1 *(regression)* |
| patch size | $512 \times 512$ |
| random crop size | $299 \times 299$ |
| # instances per bag ($N$) | 200 |
| # features ($J$) | 128 |
| # bins in 'distribution' filters | 21 |
| $\sigma$ in Gaussian kernel | 0.05 |
| Optimizer | ADAM |
| Learning rate | $1e-4$ |
| $L2$ regularization weight decay | 0.0005 |
| batch size | 1 |

## Segmentation of Histopathology Slides in The TCGA LUAD Cohort

In the TCGA LUAD cohort, for each patient with a matching normal sample, we used the trained feature extractor module of our MIL model to extract features of patches cropped over the slides of the tumor and normal samples of the patient. Then, we clustered the patches by using hierarchical clustering over the extracted feature vectors. We determined the distance threshold in hierarchical clustering such that there were 4 clusters among the patches from slides of the normal sample. This made our clustering approach robust against patient-to-patient variations. Indeed, this was why we decided to use both tumor and normal samples of the patient. In other words, instead of determining a global distance threshold for all patients, we calculated patient-specific distance threshold values to capture inter-patient variations.

Each cluster can be assigned one of two labels: cancerous or normal. Ideally, a cluster with a cancerous label can contain patches only from slides of the tumor sample. On the other hand, a cluster with a normal label can contain patches from slides of both the tumor and the normal samples since the tumor sample may also contain normal tissue components. As a post-processing step, we analyzed normal clusters. If the number of patches from slides of the normal sample in a normal cluster was less than 10%, we split this cluster into two such that patches from slides of the tumor sample were assigned to a new cancerous cluster. Finally, we created segmentation masks for slides of the tumor sample by using cluster labels assigned to the patches.

## SUPPLEMENTAL REFERENCES

[1] Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., et al. (2012). Absolute quantification of somatic dna alterations in human cancer. Nature biotechnology, 30(5):413–421.

[2] Meng, X.-L., Rosenthal, R., and Rubin, D. B. (1992). Comparing correlated correlation coefficients. Psychological bulletin, 111(1):172.

[3] Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Treviño, V., Shen, H., Laird, P. W., Levine, D. A., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. Nature communications, 4(1):1–11.

[4] Wilcoxon, F. (1992). Individual comparisons by ranking methods. In Breakthroughs in statistics, pages 196–202. Springer.

[5] Spencer, D. H., Sehn, J. K., Abel, H. J., Watson, M. A., Pfeifer, J. D., and Duncavage, E. J. (2013). Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. The Journal of molecular diagnostics, 15(5):623–633.

[6] Oner, M. U., Kye-Jet, J. M. S., Lee, H. K., and Sung, W.-K. (2020). Studying the effect of mil pooling filters on mil tasks. arXiv preprint arXiv:2006.01561.

[7] Parzen, E. (1962). On estimation of a probability density function and mode. The annals of mathematical statistics, 33(3):1065–1076.

[8] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.