cancers

MDPI

Article

Deep Learning-Based Inference of Mutational Signature Exposures from Histology Across Tumour Types

Byeong Chang Sim [1,2,3], JunSik Kim [1], Jiyeon Park [1], YunJi Kim [1], JungHyun Park [1], Taehoon Ko [1,2], Ji-Won Chun [1,2] and Tae-Min Kim [1,2,3,4,*]

[1] CMC Institute for Basic Medical Science, The Catholic Medical Center, The Catholic University of Korea, Seoul 06591, Republic of Korea; 202553049@catholic.ac.kr (B.C.S.); 22501836@cmcnu.or.kr (J.K.); pjy0088@catholic.ac.kr (J.P.); yunjikim@cmcnu.or.kr (Y.K.); 22602175@cmcnu.or.kr (J.P.); thko@catholic.ac.kr (T.K.); jwchun@catholic.ac.kr (J.-W.C.)
[2] Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
[3] Graduate School, The Catholic University of Korea, Seoul 06591, Republic of Korea
[4] Cancer Research Institute, College of Medicine, The Catholic University of Korea, Seoul 06591, Republic of Korea
* Correspondence: tmkim@catholic.ac.kr; Tel.: +82-2-3147-8426

Simple Summary

Mutational signatures record the processes that shape cancer genomes, including DNA repair defects and exposure to mutagens. We investigated whether routine pathology images could infer the relative activity of these signatures across seven cancer types. Models were developed using 2741 patients and evaluated in a separate cohort of 580 patients. Fourteen cancer–signature combinations showed nominal positive associations in both cohorts. Multiple-comparison-adjusted values are also reported for the complete panel. Additional analyses examined how the choice of slides affected the results. We also assessed links between image-derived scores and genomic features related to DNA repair and instability. The models produced relative scores associated with specific mutational signatures in both cohorts. Further studies are needed to interpret the tissue features underlying these associations and assess their clinical value.

Abstract

Background/Objectives: Mutational signatures reflect the processes that shape cancer genomes. We investigated whether continuous mutational-signature exposures could be inferred from routine histology across tumour types. Methods: We developed whole-slide image regression models for 30 single-base substitution (SBS) signatures in 2741 The Cancer Genome Atlas (TCGA) patients across seven tumour types and evaluated them in 580 Clinical Proteomic Tumor Analysis Consortium (CPTAC) patients. RetCCL mean-feature ridge regression and RetCCL/CTransPath attention-based multiple-instance learning were combined by fixed unweighted averaging. Final ensemble outputs were standardised SBS scores. Patient-disjoint evaluation, full-panel false discovery rate (FDR) adjustment and tumour-labelled slide sensitivity were performed. Results: Among 185 pairs with correlations available in both cohorts, 73 showed positive correlations in both cohorts. Fourteen had nominal positive-direction $p < 0.05$ in both, including breast cancer–SBS3 (TCGA r = 0.289, CPTAC r = 0.215) and lung adenocarcinoma–SBS4 (r = 0.128 and 0.263). Cohort-specific FDR-adjusted values are provided for all 210 pairs. All 14 associations remained positive after restricting inputs to tumour-labelled slides. The SBS3 score correlated with continuous homologous recombination deficiency (r = 0.332). Exploratory TCGA SBS3, SBS6 and SBS10 scores yielded areas under the receiver operating characteristic curve

Check for updates

Academic Editor: Vasyl Vasko

Received: 6 October 2026
Revised: 23 October 2026
Accepted: 23 October 2026
Published: 25 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

of 0.693, 0.619 and 0.628 for homologous recombination deficiency-high, microsatellite instability and qualifying *POLE*/*POLD1* mutations. External MSI and polymerase AUCs were 0.573 and 0.680, respectively. Conclusions: Histology-derived scores were associated with selected genomic SBS exposures in both cohorts. Further studies are needed to identify the tissue features underlying these associations and assess their clinical relevance.

Keywords: computational pathology; whole-slide image; mutational signature; single-base substitution; multiple-instance learning; genomic instability

1. Introduction

Cancer genomes accumulate somatic mutations through diverse endogenous and exogenous mutational processes, including spontaneous DNA damage, replication errors, defective DNA repair, and exposure to environmental mutagens. These processes leave characteristic sequence patterns in cancer genomes, serving as molecular fossils of genomic processes that have operated throughout tumour evolution. Computational deconvolution of these sequence context patterns has resolved them into a repertoire of mutational signatures [1]. For example, the single-nucleotide variant catalogue of a tumour can be decomposed into single-base substitution (SBS) signatures, with the exposure of each signature quantifying its contribution to the overall mutational catalogue. A number of SBS signatures have been linked to specific biological processes, e.g., SBS1 is a clock-like, age-associated signature. SBS2 and SBS13 reflect APOBEC cytidine deaminase activity. SBS3 is associated with homologous recombination deficiency (HRD) [2]. SBS6 is associated with mismatch repair (MMR) deficiency and microsatellite instability (MSI) [3]. Unlike individual driver mutations, mutational signatures integrate the cumulative consequences of endogenous and exogenous processes that have shaped a tumour genome.

Histopathological examination of surgical resections and tissue biopsies has long been central to cancer diagnosis and clinical evaluation. Computational pathology has further shown that routine haematoxylin and eosin (H&E) morphology contains information about driver mutations, molecular subtypes, gene expression programmes, MSI, HRD and other genomic phenotypes [4–8]. More focused studies have directly predicted MSI or HRD from whole-slide images (WSIs), including pan-cancer and externally validated settings [5,9–12]. In these studies, the molecular phenotype itself is generally used as the image-level prediction target. Mutational signatures have also been used as genomic readouts of upstream repair defects. Valieris et al. used Signature 3 and Signature 20 activities to define HRD and MMR-deficient labels for histology-based classification rather than attempting to predict the continuous signature activities themselves [13]. Together, these studies establish links between histology and clinically interpretable consequences of mutational processes. They provide a basis for asking whether histology also contains information about the quantitative activities of the underlying signatures.

A preprint describing the Fuji multimodal pathology foundation model has included binary mutational-signature inference among its training tasks [14], providing a further example of molecular supervision based on mutational signatures. However, systematic prediction of continuous SBS exposures across multiple tumour types, with evaluation of the same tumour-type signature targets in an independent cohort, remains largely uncharacterised. This task is challenging because SBS exposures are continuous, frequently zero-inflated, and strongly dependent on cancer type. Attention-based multiple-instance learning (MIL) enables weakly supervised WSI analysis by aggregating information across thousands of tiles without requiring tile-level annotations [15]. Encoders pretrained on

histopathology images can transform individual tiles into informative morphological representations [16,17], while regression-based WSI approaches have established the feasibility of predicting continuous molecular phenotypes [12]. We therefore evaluated whether continuous SBS1–SBS30 exposures could be inferred from tumour histology across seven tumour types, using cancer-specific models trained in TCGA and evaluated in tumour-type-matched CPTAC cohorts. Continuous exposures were the regression targets. The ensemble outputs were standardised SBS scores evaluated against those targets. In exploratory analyses, we additionally asked whether histology-derived SBS3, SBS6 and SBS10 scores retained information about their biologically related genomic features, namely, HRD, MSI, and *POLE*/*POLD1* proofreading-domain mutations, without training the image models directly on those downstream labels.

2. Materials and Methods

2.1. Study Cohorts

H&E whole-slide images and genomic data were assembled from TCGA and the NCI Genomic Data Commons for breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), and uterine corpus endometrial carcinoma (UCEC) [18,19]. Tumour-type-matched CPTAC cohorts were obtained from the CPTAC resource, including the endometrial carcinoma study [20,21]. Analysis-specific sample sizes varied according to the availability of genomic targets and model predictions. WSI and genomic data were matched by patient identifier within each cohort. Patient characteristics, specimen annotations and the patient selection flow are summarised in Supplementary Data S1.

2.2. Mutation Signature Evaluation

Exome-level somatic mutation data were used for TCGA and CPTAC. Single-nucleotide variants were used to derive SBS signatures. Proportional SBS1–SBS30 exposures were estimated with deconstructSigs [22] using the built-in signatures.cosmic legacy reference, default trinucleotide normalisation (tri.counts.method = "default"), more than 50 observed substitutions, and a signature cutoff of 0.06. Default normalisation and a 0.06 contribution cutoff have been used with COSMIC 30-signature references in a previous whole-exome sequencing study [23]. deconstructSigs was applied separately to each eligible genomic tumour sample. When multiple eligible genomic samples mapped to the same patient, proportional weights were averaged to obtain a patient-level target. Exposure proportions supplied the molecular training targets. Final image-derived outputs were standardised scores. Target distributions, zero fractions and cohort-level substitution count and eligibility summaries are provided in Supplementary Data S3. Genomic samples and patients are reported separately. To assess dependence on target estimation, the same image-derived scores were compared with saved Sigfit (v1.2.0) legacy-SBS1–SBS30 estimates and SigProfilerAssignment estimates using COSMIC v2, v3.5 and v3.6 (software v1.1.5 for COSMIC v2 and v3.6; v1.1.4 for COSMIC v3.5). Each original-versus-alternative comparison used the same matched patients. Only matching signature names were compared. Split signatures were not combined. Source-specific sample coverage and comparison results are provided in Supplementary Data S3 (signature_comparison).

2.3. Image Processing and Feature Extraction

Slides were processed with the End2End and Marugoto regression workflow [12]. WSIs were resampled to 1.143 μm per pixel, and non-overlapping 224 × 224-pixel tiles, representing 256 × 256 μm, were extracted after Canny-edge tissue filtering [24]. Stain

appearance was normalised with the Macenko method [25]. Frozen RetCCL ResNet-50 (2048-dimensional) and CTransPath (768-dimensional) features [16,17] were coordinate-aligned. Separate learned 128-dimensional branches were concatenated and fused into a 256-dimensional tile representation. For patients with multiple eligible slides, aligned tiles from all slides were concatenated before bag sampling with a fixed patient-specific seed. The standardised analysis used patient-level RetCCL mean features and multi-output gated-attention MIL bags of 1024 and 2048 tiles. Encoder weights were not updated during downstream model fitting. Tile quality control (QC) scores combined occupancy, tissue fraction, edge density and focus. The formula, thresholds, component definitions and fixed averaging rule are provided in Supplementary Data S1. Software details and recorded environment versions are provided in Supplementary Data S1 and the public code repository.

Image inputs included all matched eligible slides. No pixel-level tumour-region mask or Tumor_Segment_Acceptable restriction was applied. TCGA inputs comprised 5135 primary-tumour-labelled slides (1800 diagnostic FFPE and 3335 frozen tissue sections). CPTAC inputs comprised 2239 slides (1509 tumour-labelled, 638 normal-labelled and 92 with unspecified tissue). Embedding was recorded as FFPE for 1571 slides, OCT for 262 and unavailable in the matched extract for 406. Among 580 CPTAC patients, 451 had at least one normal-labelled slide, and 10 had no tumour-labelled slide. Labels describe the submitted slide, not the tumour fraction of every tile. Cohort-specific slide type and specimen preparation summaries are provided in Supplementary Data S3.

2.4. WSI-Based SBS Prediction

The analysis evaluated all 210 combinations of seven tumour types and SBS1–SBS30. A separate cancer-specific model jointly predicted all 30 proportional exposures. TCGA evaluation used patient-disjoint shuffled five-fold cross-validation with three independent repetitions. Within each outer-training partition, 20% of patients were reserved for model selection. At each 1024- and 2048-tile scale, we compared random sampling without replacement, with equal selection probability for each tile, against a quality hybrid scheme. In the latter, half of each bag comprised the tiles with the highest QC scores. The remaining half was sampled at random without replacement from the other tiles. Both schemes were evaluated with or without fusion of the target attention representation with the mean and standard deviation of encoded tiles within the sampled bag. The selected variant in each scale, cancer, seed and fold maximised the stored inner-validation median Pearson correlation and produced all 30 outputs in that context. MIL optimisation used AdamW (learning rate 10^{-4}, weight decay 10^{-4}, gradient-norm clipping 5.0), at most 25 epochs and patience 8. The loss was Hellinger composition loss plus 0.25 times weighted signature-presence binary cross-entropy, 0.5 times square-root exposure mean-squared error and 0.10 times macro Pearson loss. Checkpoint scoring included SBS outputs with at least three non-zero inner-validation targets and non-zero target variance. The principal component analysis (PCA)-ridge component used RetCCL patient mean features from the inner-training patients, training-fitted standardisation, 16 principal components and ridge alpha 100. It was fitted to square-root exposures, with non-negative predictions squared back to the exposure scale. The PCA-ridge, 1024-tile MIL and 2048-tile MIL outputs were each standardised using the corresponding inner-validation mean and population standard deviation, then combined by a fixed unweighted 1:1:1 arithmetic mean. The resulting outputs were standardised SBS scores, not calibrated exposure proportions. TCGA out-of-fold scores were averaged across the three repetition-specific predictions for each patient. CPTAC predictions were averaged across the 15 TCGA-trained models. Model

specifications are in Supplementary Data S1. Full performance results are in Supplementary Data S3.

2.5. Exploratory Associations of SBS Scores with Genomic Features

Exploratory genomic-feature associations used TCGA out-of-fold primary SBS3, SBS6 and SBS10 scores for BRCA HRD, UCEC MSI and UCEC polymerase status, respectively. All patients with the corresponding SBS score and required clinical label were included. Genomic-feature labels were used only for evaluation and were not used to fit, select or recalibrate the signature scores. The continuous HRD score was the sum of loss of heterozygosity, telomeric allelic imbalance and large-scale state transition scores [26], with HRD-high defined as $\geq$42 [27]. The UCEC MSI status was defined using MANTIS [28] scores from the published TCGA pan-cancer analysis [29]. For the present study, scores > 0.6 were classified as positive and scores < 0.4 as negative. Scores from 0.4 to 0.6, inclusive, and missing values were excluded. This yielded 365 of the 422 primary UCEC patients. The originally specified polymerase endpoint was a composite. A patient was positive if at least one of the exact literature-informed qualifying proofreading-domain missense substitutions in *POLE* OR *POLD1* listed in Supplementary Data S2 was identified, and negative if no listed substitution was identified. Double-positive patients were counted once. Negative status denotes absence of a qualifying recorded substitution rather than complete *POLE*/*POLD1* wild type. The 11 *POLE* substitutions match the somatic endometrial carcinoma list in León-Castillo et al., Table 3 [30]. The seven *POLD1* substitutions are a study-defined selection supported by the functional, somatic signature and familial cancer literature [3,31,32]. This research selection is distinguished from germline clinical pathogenicity classification.

External analyses assessed the primary CPTAC UCEC SBS6 and SBS10 scores against source-published MSI-H versus MSS status and the original *POLE* OR *POLD1* composite, respectively. The same variant list and patient-level OR rule in Supplementary Data S2 were applied to the source-publication mutation event annotations (cBioPortal ucec_cptac_2020_mutations, queried using the complete 95-sample ucec_cptac_2020_sequenced list [21]). All 82 evaluated patients were present in that sequenced sample list. Compatible CPTAC BRCA genomic-scar HRD scores or HRD-high labels (score $\geq$ 42) were not obtained from the CPTAC sources examined. MANTIS sensitivity compared the original TCGA < 0.4/>0.6 rule with a >0.4-positive/$\leq$0.4-negative rule using all available continuous scores [28]. This changes the classification rule and evaluable patient set together, while retaining the image scores.

2.6. Statistical Analysis

Pearson correlation between each standardised image-derived score and its observed exposure was the primary association measure, with patient-level Fisher-z 95% confidence intervals. Positive-direction tests (H1: $\rho > 0$) were retained for all 210 tumour type–signature pairs. Benjamini–Hochberg adjustment was applied separately to all 210 tests in each cohort [33]. An FDR threshold of 0.20 was used for this exploratory analysis to reduce the risk of overlooking potentially relevant associations. The original nominal $p < 0.05$ criterion in both cohorts was retained to identify recurring positive associations. Full effects, confidence intervals and raw/adjusted *p*-values are provided in Supplementary Data S3. Positive–positive, negative–negative and discordant pairs were reported separately. Spearman correlations used the same predictions. Sparse targets were interpreted with their distributions and non-zero counts.

Genomic-feature AUCs were calculated directly from the specified signature scores. AUC confidence intervals used 10,000 class-stratified patient-bootstrap replicates and percentile limits. Continuous SBS3–HRD associations used two-sided Pearson tests, and

binary comparisons used two-sided Mann–Whitney U tests. Unadjusted two-sided *p*-values are reported for these exploratory associations.

2.7. Image-Input Sensitivity Analysis

All 210 combinations of seven CPTAC cancer types and 30 SBS targets were assessed using tile features selected by source-slide annotation. Tile embeddings from normal-labelled and unspecified source slides were excluded, and those from explicitly tumour-labelled source slides were retained. Retained tile features were pooled per patient and sampled into the original MIL bags. RetCCL patient-mean features for PCA–ridge were also recomputed from retained tiles. The genomic targets and component averaging were retained. Across the seven cancer types, both input conditions used the same 570 retained patients, with 2221 all-input slides versus 1509 tumour-labelled slides. Ten patients without a tumour-labelled input were excluded only from this supplementary assessment. The original 580-patient assessment remained primary. Complete paired correlations, confidence intervals, raw *p*-values and restricted-input cohort-adjusted values are supplied in Supplementary Data S3. Figure S3 displays the original 14 nominal examples. Specimen groups and normal-slide inclusion were compared descriptively using existing all-slide scores, with patient counts reported. Source-slide annotations do not identify the tumour at the tile level.

2.8. Attention Visualisation

For each of UCEC–SBS13, UCEC–SBS3, BRCA–SBS3 and LUAD–SBS4 in TCGA and CPTAC, an illustrative patient was selected using the median positive genomic exposure among patients with a tumour-labelled slide. Original MIL weights were displayed at sampled coordinates. The H&E overview, overlay and higher-/lower-weight tiles show how one MIL component weights the sampled regions during aggregation. Each display reports the patient genomic exposure and standardised ensemble score. Colour indicates the within-patient attention rank, and unselected regions are uncoloured. Expert pathological interpretation was not performed.

3. Results

3.1. Study Population and Analysis Framework

The WSI–genomic analysis comprised 2741 TCGA and 580 CPTAC patients across seven tumour types—BRCA, HNSC, KIRC, LUAD, LUSC, OV and UCEC—with analysis-specific sample sizes determined by the availability of both patient-level image predictions and deconstructSigs exposures (Figure 1). The TCGA prediction cohorts included 530 BRCA, 368 HNSC, 258 KIRC, 406 LUAD, 418 LUSC, 339 OV and 422 UCEC cases. The corresponding CPTAC cohorts comprised 77, 104, 83, 86, 103, 45 and 82 cases, respectively. Each tumour type contributed patient-level predictions for all 30 SBS targets, allowing the same signature landscape to be compared between the TCGA and CPTAC cohorts.

Each TCGA patient received an out-of-fold prediction generated by models trained without that patient. The corresponding TCGA-trained models were then applied to CPTAC without cohort-specific refitting.



Figure 1. Study cohorts and workflow for histology-based SBS analysis. (**a**) H&E whole-slide images underwent Canny-based tissue filtering, Macenko stain normalisation and non-overlapping 224 × 224-pixel tessellation at 1.143 µm per pixel before frozen RetCCL and CTransPath feature extraction. (**b**) For the 210-pair analysis, each gated-attention MIL component formed a patient representation by attention-weighted tile aggregation. Its SBS outputs and the PCA ridge outputs were standardised using inner-validation values and combined by a fixed unweighted 1:1:1 arithmetic mean. (**c**) Patient-disjoint TCGA cross-validation, model selection within the training data, fixed unweighted model averaging and CPTAC prediction. Primary TCGA out-of-fold SBS3, SBS6 and SBS10 scores were assessed against corresponding genomic features. (**d**) Matched WSI–genomic patient counts by cancer and cohort. Arrows indicate the direction of processing and prediction. Colours distinguish workflow components in panels (**a**–**c**) and cancer types in panel d, as indicated by the labels and legend. Abbreviations: single-base substitution (SBS), haematoxylin and eosin (H&E), whole-slide image (WSI), multiple-instance learning (MIL), principal component analysis (PCA), quality control (QC), The Cancer Genome Atlas (TCGA), Clinical Proteomic Tumor Analysis Consortium (CPTAC), homologous recombination deficiency (HRD), microsatellite instability (MSI), breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC).

3.2. Associations of Histology-Derived SBS Scores with Observed Exposures

Across the 210 tumour type–SBS combinations, Pearson correlations could be calculated in both cohorts for 185 pairs (73 positive–positive, 46 negative–negative and 66 discordant). The 119 direction-concordant pairs are a descriptive total. Negative–negative

associations are not counted as successful positive-direction predictions. Thirty-five TCGA pairs and 34 CPTAC pairs had nominal positive-direction $p < 0.05$. Fourteen met this criterion in both cohorts, namely, BRCA–SBS2, BRCA–SBS3, BRCA–SBS13, HNSC–SBS19, KIRC–SBS1, LUAD–SBS4, LUAD–SBS24, LUSC–SBS1, LUSC–SBS4, LUSC–SBS23, UCEC–SBS3, UCEC–SBS6, UCEC–SBS10 and UCEC–SBS13 (Figures 2 and 3). At an exploratory FDR threshold of 0.20, nine of these 14 associations remained significant in both cohorts. Complete two-cohort effects and confidence intervals are provided in Figure S1, with raw and adjusted *p*-values in Supplementary Data S3.

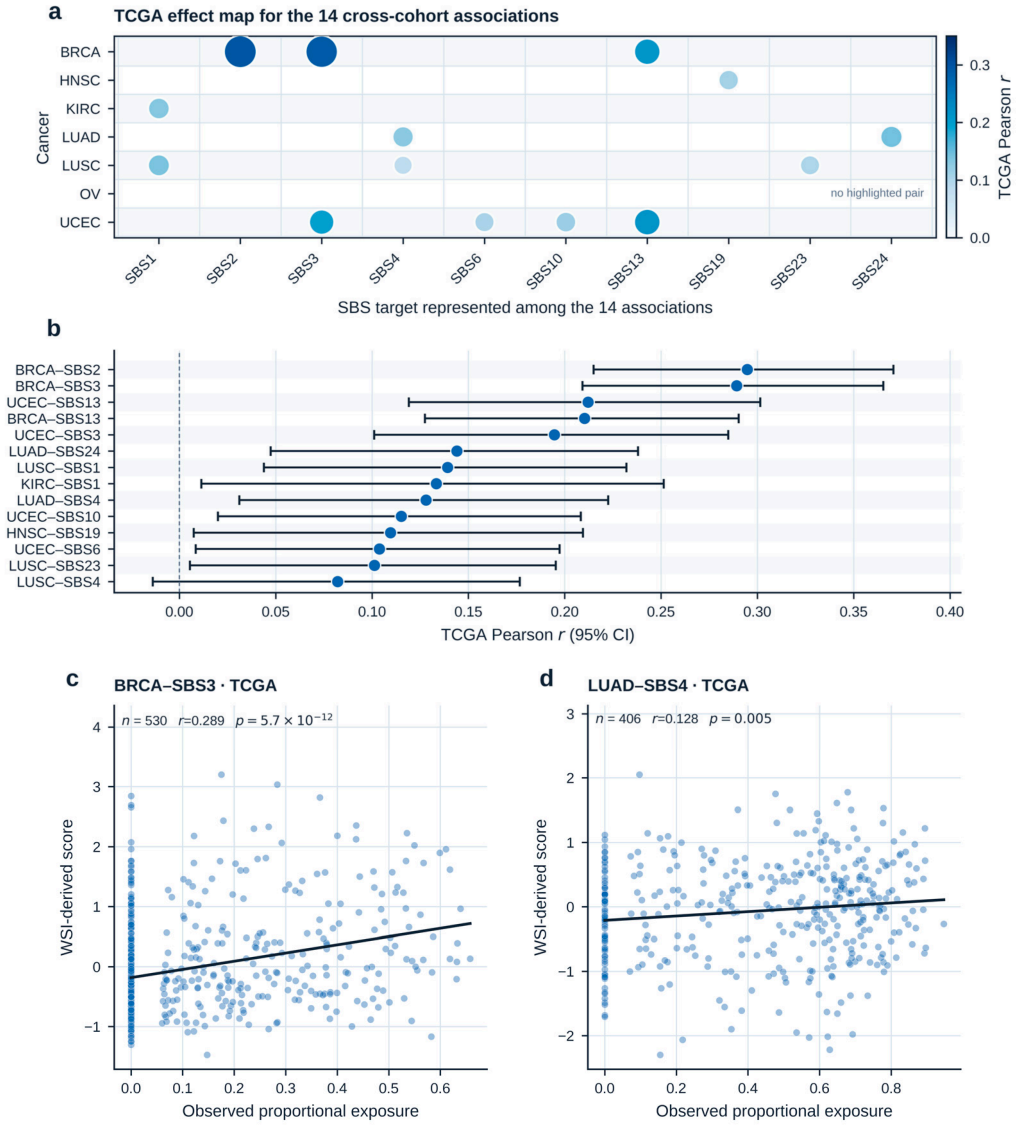


Figure 2. TCGA associations in the 210-pair SBS analysis. (**a**) Bubble map of the 14 nominally recurring cancer–SBS pairs that showed positive-direction $p < 0.05$ in both TCGA and CPTAC. Bubble area and blue intensity encode the TCGA Pearson r. OV is retained as an empty row because no OV pair met the cross-cohort criterion. The complete 210-pair TCGA landscape is shown in Supplementary Figure S1. (**b**) TCGA effects with Fisher-z 95% confidence intervals for the same 14 pairs. (**c**) BRCA–SBS3 ($n = 530$, r = 0.289, $p = 5.68 \times 10^{-12}$). (**d**) LUAD–SBS4 ($n = 406$, r = 0.128, $p = 0.00492$). Blue points in panels c and d represent individual TCGA patients; solid lines show least-squares fits. The dashed vertical line in panel b marks r = 0.

https://doi.org/10.3390/cancers18213446

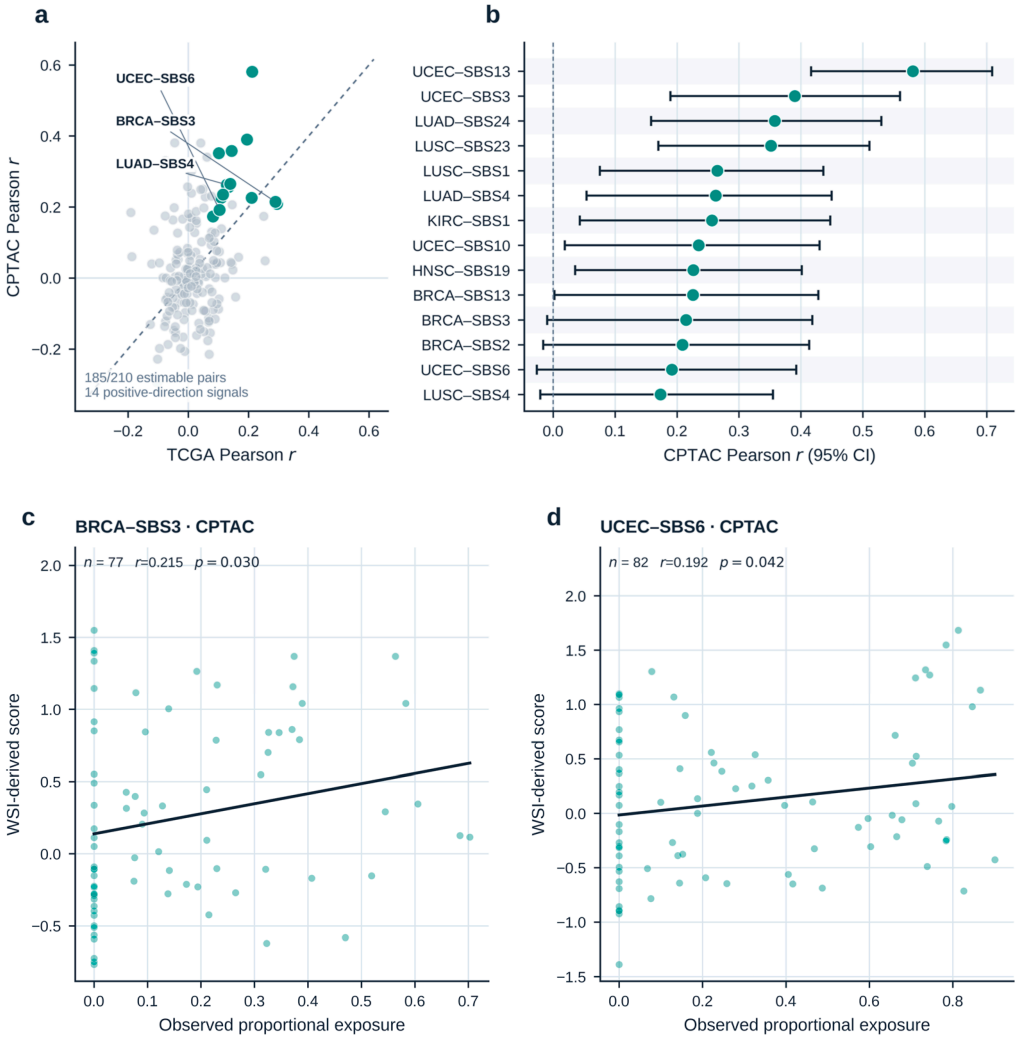


Figure 3. External validation of the SBS landscape in CPTAC. (**a**) TCGA and CPTAC Pearson effects for the 185 cancer–SBS pairs with correlations available in both cohorts. Teal symbols identify the 14 nominally recurring pairs that showed positive-direction $p < 0.05$ in both cohorts; grey symbols denote the remaining estimable pairs. The dashed diagonal marks equal correlations in the two cohorts. (**b**) CPTAC effects and Fisher-z 95% confidence intervals for these 14 pairs; the dashed vertical line marks r = 0. (**c**) BRCA–SBS3 (n = 77, r = 0.215, p = 0.0305). (**d**) UCEC–SBS6 (n = 82, r = 0.192, p = 0.0422). Teal points in panels c and d represent individual CPTAC patients; solid lines show least-squares fits.

Several of these nominally recurring pairs corresponded to mutational processes with well-established aetiologies. HRD-associated BRCA–SBS3 showed positive correlations in both cohorts (TCGA r = 0.289, CPTAC r = 0.215). Smoking-associated SBS4 showed positive correlations in both cohorts for lung cancers, including LUAD (TCGA r = 0.128, CPTAC r = 0.263) and LUSC. APOBEC-associated signatures were represented by BRCA–SBS2, BRCA–SBS13 and UCEC–SBS13, while MSI-associated SBS6 was nominally positive in both UCEC cohorts (TCGA r = 0.104, CPTAC r = 0.192). Positive associations in both cohorts occurred across several mutational processes and tumour types, with generally modest effect sizes.

The Pearson findings were also examined using Spearman rank correlations. Across pairs with available correlations, Pearson and Spearman effect estimates were closely aligned, with correlations between the two effect-size measures of 0.924 in TCGA and 0.907 in CPTAC (Supplementary Figure S3a,b). Sign agreement between Pearson and Spearman estimates was observed for 174 of 200 TCGA pairs and 166 of 187 CPTAC pairs.

Using the unchanged image-derived scores, Sigfit and COSMIC v2 SigProfilerAssignment each retained positive correlations for all 14 original pairs in TCGA and 13 of 14 in CPTAC. With COSMIC v3.5 and v3.6, 12 of 13 comparable original pairs remained positive in TCGA and nine of 12 in CPTAC. Across the full set of comparable pairs, each alternative produced both increases and decreases in correlation. These comparisons used matched patients within each source and did not retrain the image models (Supplementary Data S3, signature_comparison).

3.3. Image-Input Sensitivity

The complete 210-pair input assessment compares the same 570 CPTAC patients using tile features from 2221 all-input source slides versus 1509 tumour-labelled slides. Four BRCA and six OV patients lack an explicitly tumour-labelled input in the matched metadata and are excluded only from this comparison. The primary analysis retained all 580 patients. All 14 original nominal pairs remained positively correlated after tumour-labelled slide restriction. Seven correlations increased and seven decreased. Among these 14 pairs, nominal CPTAC $p < 0.05$ was observed for 12 using all inputs and 10 after restriction. BRCA–SBS3 changed from r = 0.211 to 0.267, LUAD–SBS4 from 0.263 to 0.266 and UCEC–SBS13 from 0.581 to 0.584. UCEC–SBS3 and LUSC–SBS4 weakened from 0.390 to 0.344 and 0.173 to 0.072, respectively. Figure S3 summarises these pairs. Supplementary Data S3 reports input counts and all 210 paired correlations with confidence intervals. Observed correlations were higher in the FFPE-only group for some associations and in the frozen/OCT-only group for others. Performance differences by normal-slide inclusion could not be assessed reliably because comparison groups were absent or small. Group-specific results and sample sizes are provided in Supplementary Data S3.

3.4. Exploratory Associations with Genomic Features

The exploratory genomic-feature analyses used primary TCGA out-of-fold SBS3, SBS6 and SBS10 scores. In TCGA BRCA, the WSI-predicted SBS3 score correlated with the continuous HRD score ($n = 423$, r = 0.332, 95% CI 0.244–0.414, $p = 2.47 \times 10^{-12}$, Figure 4). The same score discriminated HRD-high status, defined by an HRD score ≥ 42, with an AUC of 0.693 (95% CI 0.638–0.745, 132 positive patients among 423, $p = 2.09 \times 10^{-10}$).

In TCGA UCEC, the primary SBS6 score discriminated MANTIS-defined MSI status with an AUC of 0.619 (95% CI 0.555–0.681, 102 positive patients among 365, $p = 4.34 \times 10^{-4}$, Figure 4 and Supplementary Data S3).

The WSI-predicted SBS10 score similarly discriminated UCEC tumours carrying at least one of the originally listed qualifying *POLE* OR *POLD1* proofreading-domain substitutions with an AUC of 0.628 (95% CI 0.537–0.716, 47 positive patients among 422, $p = 0.00409$, Figure 4 and Supplementary Data S2). The mutation-positive group contained 44 *POLE*-positive and four *POLD1*-positive patients, including one patient positive for both genes. Double-positive patients were counted once, giving 47 qualifying-variant-positive and 375 qualifying-list-negative patients. No *POLD1*-specific performance inference was made.

Using the published MANTIS 0.4 threshold in TCGA increased the evaluated set from 365 to 409 of 422 UCEC patients and positive labels from 102 to 146. Thirteen patients lacked a continuous MANTIS score. Using the same SBS6 scores, the AUC was 0.592 (95% CI 0.536–0.649), compared with 0.619 under the original intermediate-score exclusion. In CPTAC UCEC, the primary SBS6 score yielded AUC = 0.573 (95% CI 0.429–0.709, 25 MSI-H patients among 82, two-sided $p = 0.295$). The same *POLE* OR *POLD1* list and rule identified six positives and 76 negatives. Five qualifying positive patients carried *POLE* p.P286R, and one carried *POLE* p.V411L. No qualifying *POLD1*-positive case was identified. The unchanged SBS10 score yielded AUC = 0.680 (95% CI 0.474–0.862, $p = 0.150$). The AUC

estimates refer to the prespecified composite endpoint. *POLD1*-specific performance could not be evaluated. These endpoint analyses were exploratory (Figure S4 and Supplementary Data S3). External evaluation of the original BRCA genomic-scar HRD endpoint was not completed because equivalent labels were not obtained from the sources examined.

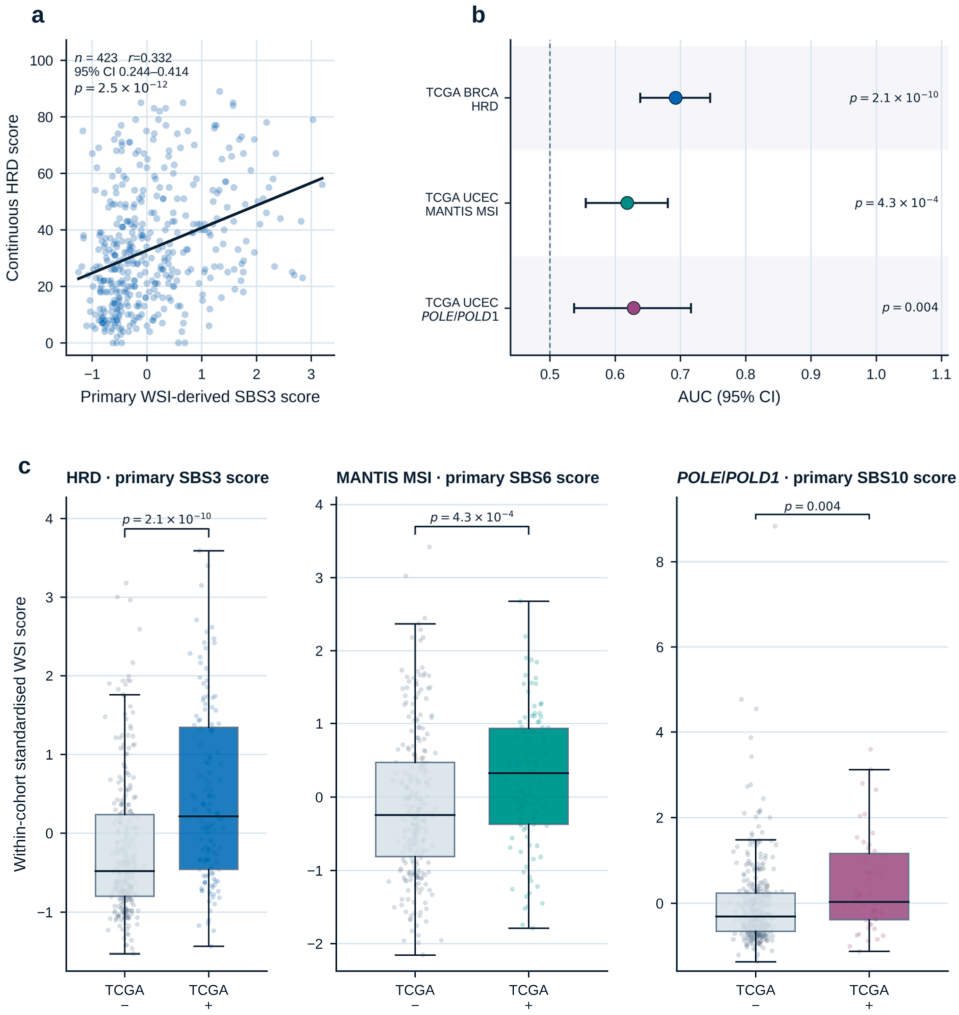


Figure 4. Exploratory associations of WSI-derived SBS scores with genomic features in TCGA. (**a**) SBS3 score versus continuous HRD (n = 423, r = 0.332, 95% CI 0.244–0.414, $p = 2.47 \times 10^{-12}$). (**b**) AUCs with bootstrap 95% confidence intervals for HRD-high using SBS3 (AUC = 0.693), MANTIS-defined MSI using the primary SBS6 score (AUC = 0.619) and the presence of the original qualifying *POLE* OR *POLD1* proofreading-domain substitutions using SBS10 (AUC = 0.628). The dashed vertical line in panel b marks AUC = 0.5. Blue, teal and purple denote HRD, MSI and the *POLE*/*POLD1* composite, respectively. (**c**) Score distributions for the same direct associations; grey indicates negative cases and the corresponding endpoint colour indicates positive cases. Scores in panel c were z-standardised within each clinical association cohort for display only. All association metrics were calculated from the out-of-fold scores. The p-value in panel a is from a two-sided Pearson correlation test. p-values for binary-group comparisons are from two-sided Mann–Whitney U tests. In panel c, centre lines indicate medians, boxes show the interquartile range (IQR), whiskers extend to the most extreme observations within 1.5 times the IQR, and points represent individual patients.

4. Discussion

This study evaluated histology-based inference of continuous SBS exposures across seven cancer types. The final ensemble outputs are standardised scores associated with genomic exposures, rather than calibrated exposure proportions. Fourteen cancer–signature pairs showed positive correlations with nominal $p < 0.05$ in both patient-disjoint TCGA eval-

uation and independent CPTAC patients, spanning six cancer types and several established mutational processes. Their recurrence in two patient populations supports a relationship between histology and these genomic exposures. Requiring nominal significance in both cohorts does not itself control the FDR across the screened associations. The cohort-specific adjusted *p*-values account for the multiple comparisons in this exploratory screen.

Previous computational pathology studies have established associations between H&E morphology and individual driver mutations, MSI, molecular subtypes and broad multi-omic biomarkers in single-cancer and pan-cancer settings [4–8,34,35]. Studies targeting HRD or MSI have further demonstrated attention-MIL regression or classification approaches and external validation for these clinically relevant phenotypes [9–12].

Valieris et al. used SBS3 and SBS20 activities to define HRD and MMR-deficient labels for WSI classification [13]. The Fuji multimodal pathology foundation model included binary mutational-signature inference during molecularly supervised pretraining [14]. Here, individual SBS exposures were regression targets in a fixed seven-cancer by 30-signature analysis, and the resulting models were assessed in tumour-type-matched CPTAC cohorts without refitting. The cross-cohort candidates include signatures related to homologous recombination deficiency (SBS3) [2], tobacco mutagenesis (SBS4), APOBEC activity (SBS2/SBS13) and mismatch repair deficiency (SBS6) [1,3]. These associations suggest links between mutational history and tissue appearance, although their cellular basis remains to be determined.

Our whole-tissue approach follows that of El Nahhas et al. [12], who used TCGA and CPTAC slides without prior tumour region segmentation. Valieris et al. [13] used background-filtered WSI tiles and subsequently identified tumour and non-tumour tissue among selected regions. All 14 original associations retained positive correlations after removing normal-labelled and unspecified slide inputs. These results indicate that normal-labelled or unspecified source inputs can influence individual correlations but do not fully account for the observed positive associations.

The genomic-feature analyses examined whether scores trained on SBS exposures retained information about related genomic states. The SBS3 association with HRD supports this interpretation in TCGA. In CPTAC, the SBS10 AUC was 0.680 for the predefined *POLE*/*POLD1* composite, while the SBS6 MSI AUC was 0.573. These external estimates had wide confidence intervals, particularly for the six polymerase-positive patients, whose qualifying variants were all in *POLE*. The analyses therefore remain exploratory and do not support *POLD1*-specific conclusions. Compatible genomic-scar HRD labels were unavailable for the matched CPTAC BRCA patients.

The input sensitivity analysis could not isolate the effect of normal tissue because normal-labelled and unspecified slides were excluded together. Slide labels do not identify tumour regions within individual tiles. Specimen-type comparisons used existing all-slide scores and were limited by small groups and incomplete annotations. The influence of normal-slide inclusion could not be assessed reliably because comparison groups were absent or small. Acquisition and site-specific histological characteristics can also affect molecular prediction [36]. Expert pathological review by pathologists with expertise across multiple tumour types is warranted to strengthen the biological interpretation of the attention examples. Such review could not be arranged, so the tissue types and possible artefacts in these examples remain unverified.

deconstructSigs was evaluated using simulated signature mixtures and TCGA data [22]. Although individual associations changed with alternative fitting approaches and reference catalogues, no alternative produced a consistent improvement across the evaluated cancer–signature pairs. We therefore retained the original deconstructSigs-based SBS1–SBS30 framework for consistency across cancer types and cohorts. Exome mutation

counts, similarity between signatures, sparse exposures and the 0.06 contribution cutoff can affect the estimated targets.

5. Conclusions

Cancer-specific deep learning models inferred relative mutational signature activity from histology, with selected score–exposure associations observed in both TCGA and CPTAC. The complete seven-cancer, 30-signature analysis; adjustment for multiple testing; and input sensitivity analyses describe the strength and uncertainty of these associations. Further comparisons of signature estimates, expert pathological review and assessment in additional independent cohorts are needed to establish their biological and clinical relevance.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers18213446/s1, Figure S1: Signature associations across cancer types; Figure S2: Target distributions, sparsity and catalogue counts; Figure S3: Robustness of score–exposure associations; Figure S4: Exploratory genomic-feature associations and MANTIS sensitivity; Figure S5: Paired TCGA/CPTAC attention examples. Supplementary Data S1 provides model settings, cohort characteristics and selection details. Supplementary Data S2 provides endpoint definitions, qualifying variants and the TCGA clinical scores and results used in Figure 4. Supplementary Data S3 provides complete association statistics, input sensitivity, clinical endpoint comparisons, target diagnostics, slide composition, specimen strata and alternative estimation comparisons. References [18–21,26,37,38] are also cited in the supplementary materials.

Author Contributions: Conceptualisation and writing—review and editing, T.-M.K.; methodology, software, formal analysis, visualisation and data curation, B.C.S.; validation, J.K., Y.K., J.P. (Jiyeon Park) and J.P. (JungHyun Park); methodology review, T.K. and J.-W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by grants from the National Research Foundation of Korea (NRF) (grant numbers RS-2024-00454948, RS-2025-02273009 and RS-2026-25504267) and by the Catholic Education Foundation through the Basic Medical Science Facilitation Program at the Catholic Medical Center of The Catholic University of Korea.

Institutional Review Board Statement: This study used de-identified secondary data from TCGA and CPTAC under the data use conditions of the originating repositories. No participants were recruited and no additional human samples were collected for this analysis.

Informed Consent Statement: Additional informed consent was not required for this secondary analysis of de-identified TCGA and CPTAC data made available for research use.

Data Availability Statement: TCGA data are available through the NCI Genomic Data Commons (https://portal.gdc.cancer.gov/). CPTAC images are available through The Cancer Imaging Archive (https://www.cancerimagingarchive.net/) and mutation data through doi:10.5281/zenodo.8394329. Derived results and source identifiers are supplied in Supplementary Data S1–S3. Controlled-access source data remain subject to repository requirements. TCGA clinical data were obtained from the Pan-Cancer Clinical Data Resource [37], and slide annotations were interpreted using the published pathology description [38]. The analysis code, software versions and environment specifications, preprocessing details, fold assignments, manifests and permissible derived results have been uploaded to a public GitHub repository (https://github.com/byeongchangsim/cancers-sbs (accessed on 22 October 2026); code snapshot: commit 4a6149661259).

Acknowledgments: We thank the participants and investigators of TCGA and CPTAC for making the data available for research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alexandrov, L.B.; Stratton, M.R. Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **2014**, *24*, 52–60. [CrossRef] [PubMed]
2. Polak, P.; Kim, J.; Braunstein, L.Z.; Karlic, R.; Haradhavala, N.J.; Tiao, G.; Rosebrock, D.; Livitz, D.; Kübler, K.; Mouw, K.W.; et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* **2017**, *49*, 1476–1486. [CrossRef] [PubMed]
3. Haradhvala, N.J.; Kim, J.; Maruvka, Y.E.; Polak, P.; Rosebrock, D.; Livitz, D.; Hess, J.M.; Leshchiner, I.; Kamburov, A.; Mouw, K.W.; et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **2018**, *9*, 1746. [CrossRef] [PubMed]
4. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **2018**, *24*, 1559–1567. [CrossRef] [PubMed]
5. Kather, J.N.; Pearson, A.T.; Halama, N.; Jäger, D.; Krause, J.; Loosen, S.H.; Marx, A.; Boor, P.; Tacke, F.; Neumann, U.P.; et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **2019**, *25*, 1054–1056. [CrossRef] [PubMed]
6. Jin, D.; Liang, S.; Shmatko, A.; Arnold, A.; Horst, D.; Grünewald, T.G.P.; Gerstung, M.; Bai, X. Teacher-student collaborated multiple instance learning for pan-cancer PDL1 expression prediction from histopathology slides. *Nat. Commun.* **2024**, *15*, 3063. [CrossRef] [PubMed]
7. Fu, Y.; Jung, A.W.; Torne, R.V.; Gonzalez, S.; Vöhringer, H.; Shmatko, A.; Yates, L.R.; Jimenez-Linan, M.; Moore, L.; Gerstung, M. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **2020**, *1*, 800–810. [CrossRef] [PubMed]
8. Arslan, S.; Schmidt, J.; Bass, C.; Mehrotra, D.; Geraldes, A.; Singhal, S.; Hense, J.; Li, X.; Raharja-Liu, P.; Maiques, O.; et al. A systematic pan-cancer study on deep learning-based prediction of multi-omic biomarkers from routine pathology images. *Commun. Med.* **2024**, *4*, 48. [CrossRef] [PubMed]
9. Loeffler, C.M.L.; El Nahhas, O.S.M.; Muti, H.S.; Carrero, Z.I.; Seibel, T.; van Treeck, M.; Cifci, D.; Gustav, M.; Bretz, K.; Gaisa, N.T.; et al. Prediction of homologous recombination deficiency from routine histology with attention-based multiple instance learning in nine different tumor types. *BMC Biol.* **2024**, *22*, 225. [CrossRef] [PubMed]
10. Bergstrom, E.N.; Abbasi, A.; Díaz-Gay, M.; Galland, L.; Ladoire, S.; Lippman, S.M.; Alexandrov, L.B. Deep learning artificial intelligence predicts homologous recombination deficiency and platinum response from histologic slides. *J. Clin. Oncol.* **2024**, *42*, 3550–3560. [CrossRef] [PubMed]
11. Wang, C.W.; Muzakky, H.; Firdi, N.P.; Liu, T.C.; Lai, P.J.; Wang, Y.C.; Yu, M.H.; Chao, T.K. Deep learning to assess microsatellite instability directly from histopathological whole slide images in endometrial cancer. *npj Digit. Med.* **2024**, *7*, 143. [CrossRef] [PubMed]
12. El Nahhas, O.S.M.; Loeffler, C.M.L.; Carrero, Z.I.; van Treeck, M.; Kolbinger, F.R.; Hewitt, K.J.; Muti, H.S.; Graziani, M.; Zeng, Q.; Calderaro, J.; et al. Regression-based Deep-Learning predicts molecular biomarkers from pathology slides. *Nat. Commun.* **2024**, *15*, 1253. [CrossRef] [PubMed]
13. Valieris, R.; Amaro, L.; Osório, C.A.B.T.; Bueno, A.P.; Rosales Mitrowsky, R.A.; Carraro, D.M.; Nunes, D.N.; Dias-Neto, E.; Silva, I.T.D. Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer. *Cancers* **2020**, *12*, 3687. [CrossRef] [PubMed]
14. Li, Q.; Sang, J.; Xiao, Y.; Zhang, P.; Chen, W.; Tran, T.N.; Li, Z.; Li, S.; Zhang, T.; Wang, G. Genome-scale molecular profiling from routine histopathology with a multimodal pathology foundation model. *Res. Sq.* **2026**, rs.3.rs.9949149. [CrossRef] [PubMed]
15. Ilse, M.; Tomczak, J.M.; Welling, M. Attention-based deep multiple instance learning. *Proc. Mach. Learn. Res.* **2018**, *80*, 2127–2136. Available online: https://proceedings.mlr.press/v80/ilse18a.html (accessed on 22 October 2026).
16. Wang, X.; Du, Y.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Yang, W.; Huang, J.; Han, X. RetCCL: Clustering-guided contrastive learning for whole-slide image retrieval. *Med. Image Anal.* **2023**, *83*, 102645. [CrossRef] [PubMed]
17. Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Yang, W.; Huang, J.; Han, X. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **2022**, *81*, 102559. [CrossRef] [PubMed]
18. Cancer Genome Atlas Research Network; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef] [PubMed]
19. Zhang, Z.; Hernandez, K.; Savage, J.; Li, S.; Miller, D.; Agrawal, S.; Ortuno, F.; Staudt, L.M.; Heath, A.; Grossman, R.L. Uniform genomic data analysis in the NCI Genomic Data Commons. *Nat. Commun.* **2021**, *12*, 1226. [CrossRef] [PubMed]
20. Li, Y.; Dou, Y.; Da Veiga Leprevost, F.; Geffen, Y.; Calinawan, A.P.; Aguet, F.; Akiyama, Y.; Anand, S.; Birger, C.; Cao, S.; et al. Proteogenomic data and resources for pan-cancer analysis. *Cancer Cell* **2023**, *41*, 1397–1406. [CrossRef] [PubMed]

21. Dou, Y.; Kawaler, E.A.; Cui Zhou, D.; Gritsenko, M.A.; Huang, C.; Blumenberg, L.; Karpova, A.; Petyuk, V.A.; Savage, S.R.; Satpathy, S.; et al. Proteogenomic characterization of endometrial carcinoma. *Cell* **2020**, *180*, 729–748.e26. [CrossRef] [PubMed]
22. Rosenthal, R.; McGranahan, N.; Herrero, J.; Taylor, B.S.; Swanton, C. DeconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **2016**, *17*, 31. [CrossRef] [PubMed]
23. Murphy, C.; Muscat, A.; Ashley, D.; Mukaro, V.; West, L.; Liao, Y.; Chisanga, D.; Shi, W.; Collins, I.; Baron-Hay, S.; et al. Tailored NEOadjuvant epirubicin, cyclophosphamide and Nanoparticle Albumin-Bound paclitaxel for breast cancer: The phase II NEONAB trial—Clinical outcomes and molecular determinants of response. *PLoS ONE* **2019**, *14*, e0210891. [CrossRef] [PubMed]
24. Canny, J. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *8*, 679–698. [CrossRef]
25. Macenko, M.; Niethammer, M.; Marron, J.S.; Borland, D.; Woosley, J.T.; Guan, X.; Schmitt, C.; Thomas, N.E. A method for normalizing histology slides for quantitative analysis. In Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 28 June–1 July 2009; pp. 1107–1110. [CrossRef]
26. Knijnenburg, T.A.; Wang, L.; Zimmermann, M.T.; Chambwe, N.; Gao, G.F.; Cherniack, A.D.; Fan, H.; Shen, H.; Way, G.P.; Greene, C.S.; et al. Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. *Cell Rep.* **2018**, *23*, 239–254.e6. [CrossRef] [PubMed]
27. Telli, M.L.; Timms, K.M.; Reid, J.; Hennessy, B.; Mills, G.B.; Jensen, K.C.; Szallasi, Z.; Barry, W.T.; Winer, E.P.; Tung, N.M.; et al. Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. *Clin. Cancer Res.* **2016**, *22*, 3764–3773. [CrossRef] [PubMed]
28. Kautto, E.A.; Bonneville, R.; Miya, J.; Yu, L.; Krook, M.A.; Reeser, J.W.; Roychowdhury, S. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget* **2017**, *8*, 7452–7463. [CrossRef] [PubMed]
29. Bonneville, R.; Krook, M.A.; Kautto, E.A.; Miya, J.; Wing, M.R.; Chen, H.Z.; Reeser, J.W.; Yu, L.; Roychowdhury, S. Landscape of microsatellite instability across 39 cancer types. *JCO Precis. Oncol.* **2017**, *2017*, PO.17.00073. [CrossRef] [PubMed]
30. León-Castillo, A.; Britton, H.; McConechy, M.K.; McAlpine, J.N.; Nout, R.; Kommoss, S.; Brucker, S.Y.; Carlson, J.W.; Epstein, E.; Rau, T.T.; et al. Interpretation of somatic *POLE* mutations in endometrial carcinoma. *J. Pathol.* **2020**, *250*, 323–335. [CrossRef] [PubMed]
31. Mur, P.; Viana-Errasti, J.; García-Mulero, S.; Magraner-Pardo, L.; Muñoz, I.G.; Pons, T.; Capellá, G.; Pineda, M.; Feliubadaló, L.; Valle, L. Recommendations for the classification of germline variants in the exonuclease domain of *POLE* and *POLD1*. *Genome Med.* **2023**, *15*, 85. [CrossRef] [PubMed]
32. Bellido, F.; Pineda, M.; Aiza, G.; Valdés-Mas, R.; Navarro, M.; Puente, D.A.; Pons, T.; González, S.; Iglesias, S.; Darder, E.; et al. *POLE* and *POLD1* mutations in 529 kindred with familial colorectal cancer and/or polyposis: Review of reported cases and recommendations for genetic testing and surveillance. *Genet. Med.* **2016**, *18*, 325–332. [CrossRef] [PubMed]
33. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300. [CrossRef]
34. Cifci, D.; Foersch, S.; Kather, J.N. Artificial intelligence to identify genetic alterations in conventional histopathology. *J. Pathol.* **2022**, *257*, 430–444. [CrossRef] [PubMed]
35. Schrammen, P.L.; Ghaffari Laleh, N.; Echle, A.; Truhn, D.; Schulz, V.; Brinker, T.J.; Brenner, H.; Chang-Claude, J.; Alwers, E.; Brobeil, A.; et al. Weakly supervised annotation-free cancer detection and prediction of genotype in routine histopathology. *J. Pathol.* **2022**, *256*, 50–60. [CrossRef] [PubMed]
36. Howard, F.M.; Dolezal, J.; Kochanny, S.; Schulte, J.; Chen, H.; Heij, L.; Huo, D.; Nanda, R.; Olopade, O.I.; Kather, J.N.; et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat. Commun.* **2021**, *12*, 4423. [CrossRef] [PubMed]
37. Liu, J.; Lichtenberg, T.; Hoadley, K.A.; Poisson, L.M.; Lazar, A.J.; Cherniack, A.D.; Kovatich, A.J.; Benz, C.C.; Levine, D.A.; Lee, A.V.; et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **2018**, *173*, 400–416.e11. [CrossRef] [PubMed]
38. Cooper, L.A.D.; Demicco, E.G.; Saltz, J.H.; Powell, R.T.; Rao, A.; Lazar, A.J. PanCancer insights from The Cancer Genome Atlas: The pathologist's perspective. *J. Pathol.* **2018**, *244*, 512–524. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.