

Inferring Molecular Signatures in Colorectal Cancer Directly from Routine Whole-Slide Images

Author: *Piotr Keller¹

1. Predictive Systems in Biomedicine (PRISM) Lab,
Department of Computer Science, University of
Warwick, Coventry, UK
*Correspondence to Piotr.Keller@warwick.ac.uk

Disclosure: The author has declared no conflicts of interest.

Keywords: Colorectal cancer (CRC),
computational pathology, graph neural network,
molecular signatures.

Citation: Oncol AMJ. 2026;3[1]:101-102.
<https://doi.org/10.33590/oncolamj/IU9TN11N>

BACKGROUND AND AIMS

Molecular profiling has advanced colorectal cancer (CRC) research but remains limited in the clinic due to cost and tissue constraints. In contrast, hematoxylin and eosin (H&E) whole-slide images (WSI) are routinely available. Recent deep learning studies show that certain molecular signatures can be inferred from morphology. However, prediction of patient-level gene set activity remains underexplored despite their pathobiological significance as coordinated drivers of cancer risk and their interpretable definitions. Furthermore, gene set-level signatures are likely to be more robust and biologically meaningful than single-gene signals because they capture coordinated transcriptional programs.

MATERIALS AND METHODS

The author and their team developed SPARROW, a custom graph neural network trained to predict over 200 molecular signatures simultaneously from WSI graph representations of tumor, stroma, lymphocytic, and mucosal regions in primary

untreated resection specimens. SPARROW was trained on The Cancer Genome Atlas Program (TCGA; n=585) and externally validated in the PETACC3 trial dataset (n=1,160). It predicts pathway enrichment scores, point mutations, and clinically actionable CRC subtypes.

RESULTS

As shown in Table 1, SPARROW accurately predicts key molecular features of CRC directly from routine H&E slides and generalizes robustly to the independent PETACC3 trial dataset. The strongest concordance was observed for biologically and clinically relevant programs, including intrinsic consensus molecular subtype 3 (iCMS3) genes, fetal enteric progenitor pathway genes, a 22-gene YAP/TAZ transcriptional target signature, and epithelial-specific high-risk gene set (epiHR) activity. SPARROW also showed good performance for clinically used classifications, including Consensus Molecular Subtypes (CMS) and intrinsic CMS (iCMS) subtypes, interferon phenotype, and *BRAF* mutation status. Importantly, image-derived molecular scores were predictive of relapse-free survival, showing their prognostic value for patient risk stratification.

CONCLUSION

SPARROW demonstrates that some key molecular features of CRC, including transcriptomic subtypes and genomic alterations, are robustly encoded in and learnable from routine H&E histology. This enables histology to serve as a cost-effective and rapid surrogate for predicting key molecular signatures and actionable molecular subtypes while also allowing

Table 1: Performance of SPARROW for molecular signature prediction.

Molecular signature	TCGA (4-fold CV)	PETACC3 (external)	Relapse-free survival analysis (PETACC3)
iCMS3 (ρ)	0.64±0.05	0.66	HR: 1.2*; 95% CI: 1.04–1.38
Fetal mustata (ρ)	0.47±0.06	0.61	HR: 1.74*; 95% CI: 1.45–2.08
YAP_22 (ρ)	0.41±0.07	0.49	HR: 2.54*; 95% CI: 2.03–3.18
EpiHR (ρ)	0.40±0.08	0.45	HR: 1.59*; 95% CI: 1.31–1.92
iCMS2 versus iCMS3 (AUC)	0.90±0.02	0.92	HR: 1.12; 95% CI: 1.00–1.26
CMS (4 groups; AUC)	0.84±0.01	0.84	HR: 6.3* (CMS4 versus other); 95% CI: 3.92–10.01
Interferon (high/low; AUC)	0.82±0.08	0.88	HR: 1.14; 95% CI: 0.86–1.05
BRAF mutation (AUC)	0.82±0.07	0.80	HR: 2.12*; 95% CI: 1.13–4.29

*p-value <0.05 based on log-rank test using median pathway score.

Performance is reported as Spearman correlation coefficient (ρ) for continuous molecular signatures and AUC for discrete molecular phenotypes. Results are shown for internal validation using four-fold cross-validation in the TCGA cohort and external validation in the PETACC3 trial cohort. Prognostic relevance of image-derived molecular scores was assessed using relapse-free survival analysis in PETACC3, reported as HRs with 95% CIs. Asterisks indicate statistical significance (p<0.05) based on log-rank testing using the median predicted pathway score as the threshold.

AUC: area under the receiver operating characteristic curve; CMS: Consensus Molecular Subtypes; CV: cross-validation; EpiHR: epithelial-specific high-risk gene set; HR: hazard ratio; iCSM: intrinsic CMS; TCGA: The Cancer Genome Atlas Program.

Reference

1. Keller P et al. Inferring molecular signatures in colorectal cancer directly from routine whole-slide images. J Clin Oncol. 2026;44(Suppl 16):3522.

mining of spatially localized image signatures associated with these transcriptional programs. Future work will involve a detailed analysis of the morphological correlates of these signatures, their association with patient treatment response, and multimodal integration.¹