Supplemental Information

Statistical models to characterize colon tumor stiffness heterogeneity through representative atom force microscopy maps

Gauthier Gadouas[1,2,3,#], Guillaume Tosato[1,2,3,4,#], Luca Costa[3,5], Alain Thierry[1,2,3], Patrice Ravel[1,2,3], Thibault Mazard[1,2,3], Pierre-Emmanuel Milhiet[3,5], Christine Bénistant[3,5], Evelyne Lopez-Crapez[1,2,3], Jacques Colinge[1,2,3,*]

1 Institut de Recherche en Cancérologie de Montpellier (IRCM), Inserm U1194, Montpellier, France

2 Institut régional du Cancer de Montpellier (ICM), Montpellier, France

3 Université de Montpellier, Montpellier, France

4 Current address: Centre de Recherche des Cordeliers, Inserm, Université Paris Cité, Sorbonne Université, Paris, France

5 Centre de Biologie Structurale (CBS), Inserm, CNRS, Montpellier, France

# Equal contributions

* Correspondence: Jacques Colinge, IRCM, 208 avenue des Apothicaires, 34298 Montpellier, France, M jacques.colinge@umontpellier.fr, T +33 (0)411 283118, F +33 (0)467 613787

Supplementary Methods

The topographical features we used from Surfalize were the following:

Developed Interfacial Area Ratio

$$S_{\mathrm{dr}} = \frac{1}{A}\left[\iint_A \left(\sqrt{1 + \left(\frac{\partial z(xy,)}{\partial x}\right)^2 + \left(\frac{\partial z(x,y)}{\partial y}\right)^2} - 1\right) \mathrm{d}x\mathrm{d}y\right],$$

where:

- A is the total area of the surface,
- $z(x, y)$ is the height function that defines the surface.

S_{dr} is a measure of a surface's roughness. It represents the ratio of the additional surface area contributed by the texture to the projected area. Essentially, S_{dr} quantifies how much the actual surface area exceeds the nominal flat area.

Root Mean Square Gradient

$$S_{\mathrm{dq}} = \frac{1}{A}\left[\iint_A \sqrt{\left(\frac{\partial z(xy,)}{\partial x}\right)^2 + \left(\frac{\partial z(x,y)}{\partial y}\right)^2}\, \mathrm{d}x\mathrm{d}y\right].$$

S_{dq} quantifies the average rate of change of the surface height. It is calculated as the root mean square of the local slopes across the surface, providing a measure of the surface's overall steepness. It is useful for understanding the degree of surface roughness and how rapidly the surface height changes over a given area.

Reduced Peak Height

The Reduced Peak Height measures the height of the peaks on a surface relative to a reference mean plane, focusing on the most prominent features. It is defined as:

$$S_{\mathrm{pk}} = \frac{1}{A_r}\int_{A_1}(z - z_1)\, dA,$$

where:

- A_r is the area of the roughness profile,
- A_1 is the area containing the highest peaks,
- z_1 is the height threshold defining the peaks.

Autocorrelation Length

Autocorrelation Length is a measure of the distance over which surface height variations are correlated. It indicates the average lateral distance over which the surface roughness features repeat.

$$S_{\text{al}} = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R(x,y)dxdy}{R(0,0)},$$

where:

- $R(x,y)$ is the autocorrelation function of the surface heights,
- $R(0,0)$ is the value of the autocorrelation function at the origin.

Skewness

Skewness is a measure of the asymmetry of the surface height distribution. It indicates whether the surface has more peaks or valleys. A positive S_{sk} suggests a surface with more peaks, while a negative S_{sk} suggests more valleys.

$$S_{\text{sk}} = \frac{1}{A_r\sigma^3}\int_{A_r}(z-\acute{z})^3\,dA,$$

where:

- A_r is the area of the roughness profile,
- z is the height,
- $\acute{z}$ is the mean height,
- σ is the standard deviation of the height.

Kurtosis

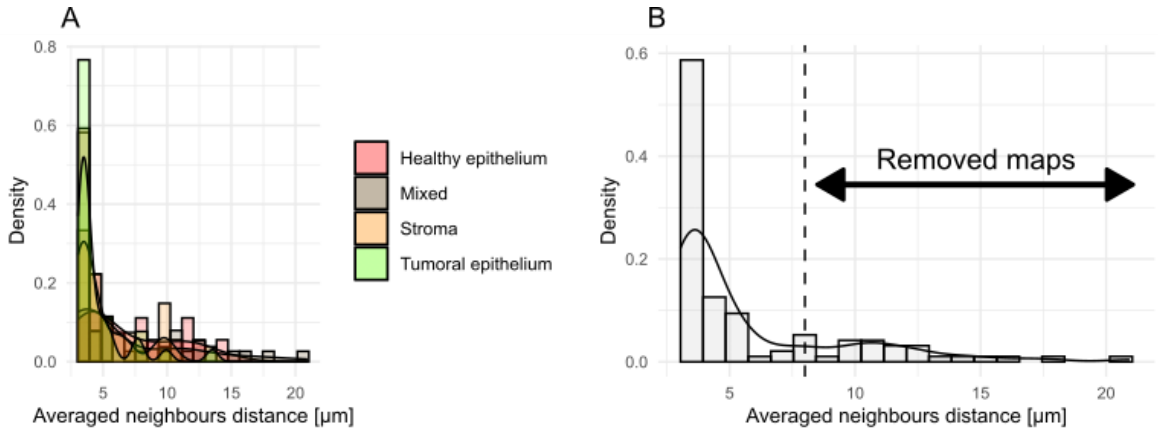
Kurtosis is a measure of the sharpness or peakedness of the surface height distribution. High kurtosis indicates a surface with sharp peaks and deep valleys, while low kurtosis indicates a flatter surface.

$$S_{\text{ku}} = \frac{1}{A_r\sigma^4}\int_{A_r}(z-\acute{z})^4\,dA.$$

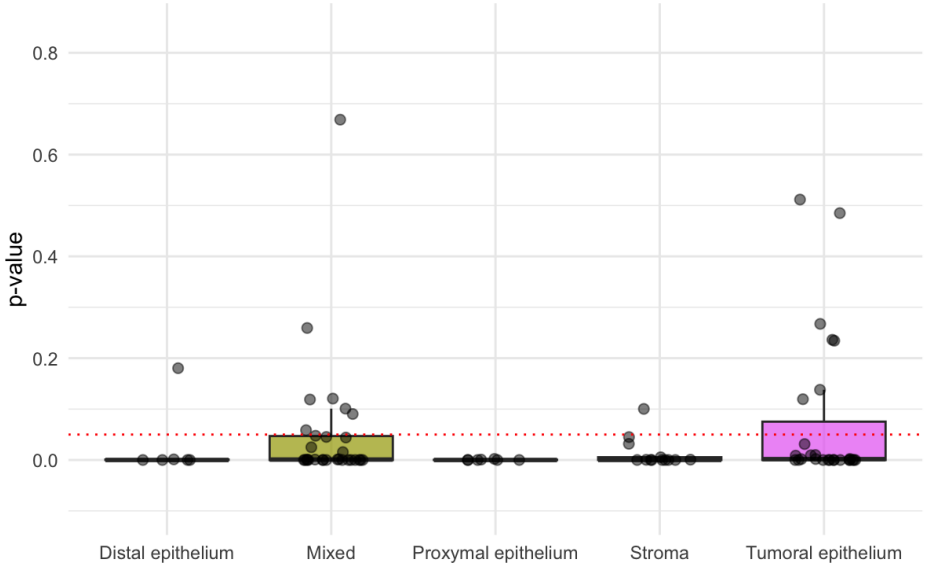
Supplementary Tables and Figures

Supplementary Table 1. Number of healthy colon areas provided by each tumor.

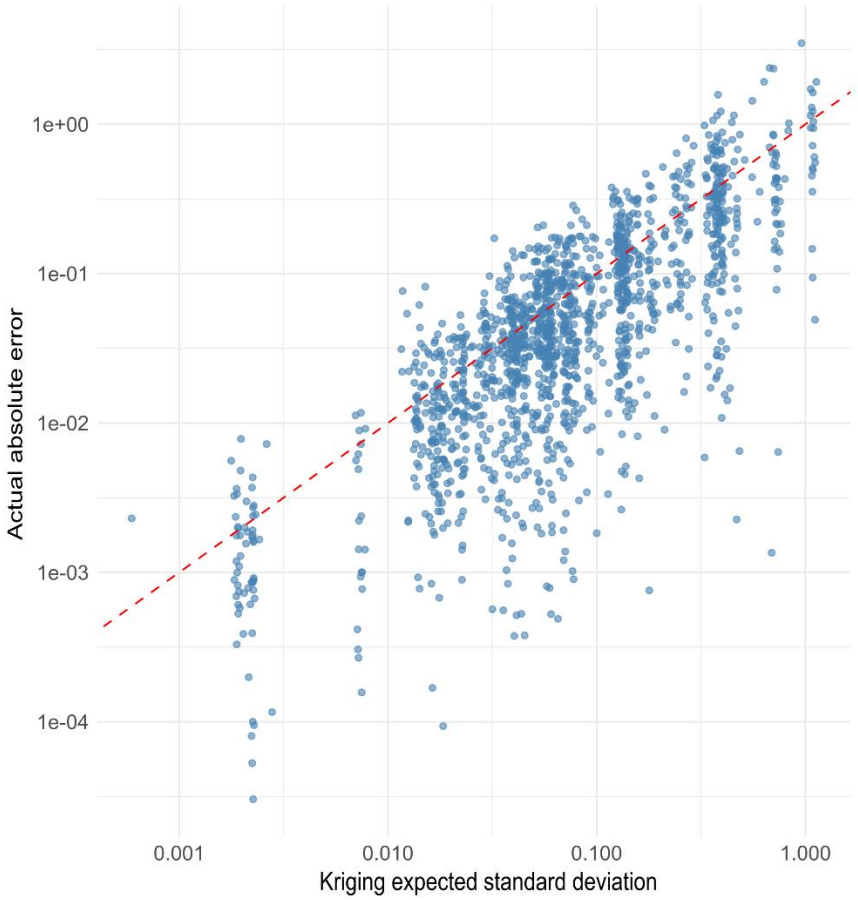
	P1	P3	P4	P5	P6	P7	P14	P19
Distal	1	1	2	2				
Proximal					2	2	1	2



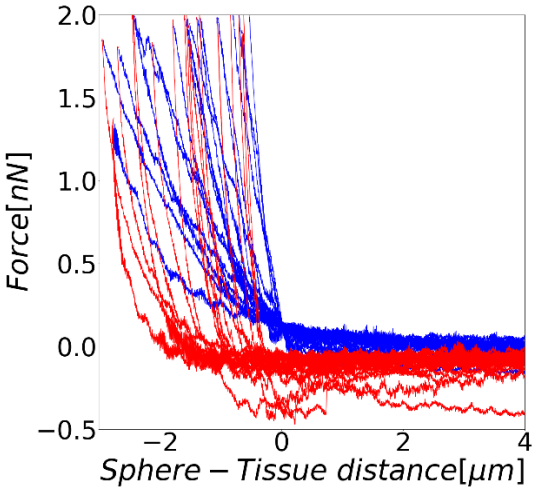
Supplementary Figure 1. Distance index distribution. (A) separated by region types. (B) Pooled.



Supplementary Figure 2. Geary's C p-values per tissue types. A p-value > 0.05 indicates no significant spatial correlation.

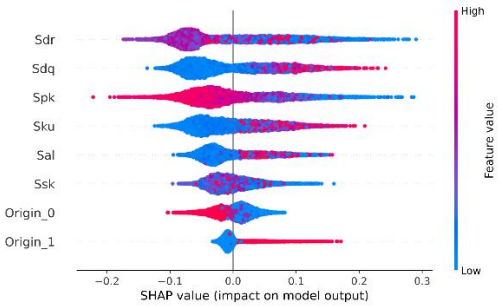


Supplementary Figure 3. Kriging interpolator expected standard deviation *versus* actual error computed by Leave-One-Out Cross-Validation (LOOCV). We note that the expected standard deviation is a conservative (typically larger) estimate of the actual error.

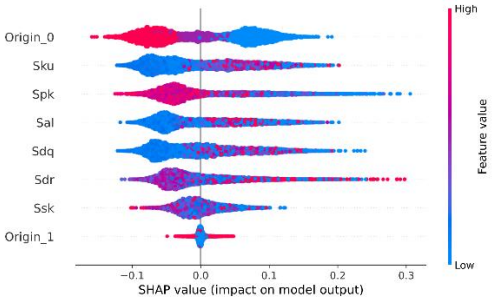


Supplementary Figure 4. Representative AFM force–distance curves. Typical approach (blue) and retract (red) curves obtained during AFM nanoindentation of the tissue surface. The horizontal axis reports the distance between the probe sphere and the tissue surface, while the vertical axis shows the measured interaction force.

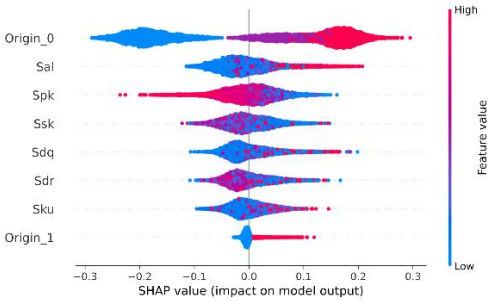
(a) Stage



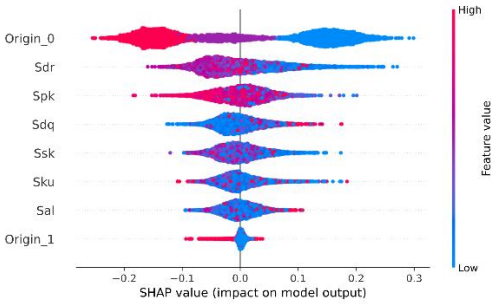
(b) Tumor localisation



(c) KRAS mutational status



(d) Presence of vascular emboli



Supplementary Figure 5. SHAP summary plots of AFM-derived topographical features predicting clinical variables. This figure presents global feature-importance analyses obtained from the random-forest models. For each clinical variable—(a) Stage, (b) tumor localisation, (c) *KRAS* mutation status, (d) presence of vascular emboli —a distinct random-forest classifier was trained using 5-fold stratified cross-validation repeated 100 times with different random seeds, ensuring balanced class representation and reducing the risk of overfitting. At each iteration the model was fitted on the training folds and evaluated on the held-out test fold. SHAP values were computed with the TreeExplainer implementation of the Python *SHAP* package. For binary classification tasks, SHAP values were calculated with respect to the positive class (label = 1); a positive SHAP value indicates that the corresponding topographical feature increases the predicted probability of the positive

class, whereas a negative value indicates the opposite effect. Across all cross-validation iterations, SHAP values from the held-out test folds were concatenated so that the importance measures reflect only out-of-sample predictions. Each panel shows a SHAP summary plot, where features (AFM-derived topographical parameters) are ranked by their mean absolute SHAP value and the distribution of SHAP values across all patients and all CV runs is displayed.