7. Appendix

7.1. Attention Mechanism

Attention mechanisms such as self- and co-attention are permutation-equivariant feature aggregation techniques that have proliferated in natural language processing and other set-based deep learning applications [53]. Given a bag-like data structure $X \in \mathbb{R}^{M \in d_k}$ containing of d_k-dim embeddings with bag size M, the self-attention function uses X as the query Q, key K and value V matrices to learn pairwise relationships a_{ij} between embeddings $q_i \in Q, k_j \in K$, in which a_{ij} is a score that measures how much the key k_j attends to the query q_i. In computing attention scores for all $x_i \in Q$, we define the self-attention matrix $A \in \mathbb{R}^{M \times M}$ as:

$$\begin{aligned} \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) &= A \\ \text{SelfAttn}(Q, K, V) &= AV \rightarrow \widehat{V} \end{aligned} \tag{6}$$

in which the softmax operator is applied row-wise to normalize scores between 0 and 1. Using the computed attention scores A, we can update the embeddings in $\widehat{V}$ as a weighted sum of surrounding embeddings in the bag as context, in which the attention weights are defined by the pairwise similarity of embeddings Q between V. For a bag size of M, the resulting asymptotic complexity is $\mathcal{O}\left(M^2 d_k + M^2 d_k\right)$.

7.2. Survival Analysis

Preliminaries: Survival outcome prediction is an ordinal regression task that models time-to-event distributions, where the outcome of the event is not always observed (*e.g.* - right censored) [13]. In observational studies that examine overall survival in cancer patients, a censored event would result from last known patient follow-up, while an uncensored event would be observed patient death.

Following our notation in § 3.1, let X represent patient data, $t_{\text{os}} \in \mathbb{R}^+$ be overall survival time (in months), $c \in \{0, 1\}$ be right uncensorship status (death observed) in a single triplet observation in a dataset $\{X_i, t_{i,\text{os}}, c_i\}_{i=1}$. In addition, let T be a continuous random variable for overall survival time, the survival function $f_{\text{surv}}(T \geq t|X)$ measure the probability of patient X survive longer than a discrete time point t, and the hazard function $f_{\text{hazard}}(T = t \mid T \geq t, X)$ measure the probability of patient death instantaneously at t, defined as:

$$f_{\text{hazard}}(T = t) = \lim_{\partial t \to 0} \frac{P(t \leq T \leq t + \partial t | T \geq t)}{\partial t}$$

which can be used to estimate f_{surv} by integrating over of f_{hazard}. The most common method for estimating the hazard function is the Cox Proportional Hazards (CoxPH) model, in which f_{hazard} is parameterized as an exponential linear function $\lambda(t|x) = \lambda_0(t)e^{\theta X}$, where $\lambda_0(t)$ is the baseline hazard and θ are model parameters that describe how the hazard varies with features X [12]. Using deep learning, θ is the last hidden layer in a neural network, and can be optimized using Stochastic Gradient Descent with the Cox partial log-likelihood [58]:

$$\begin{aligned} l(\theta, X) &= -\sum_{i \in U} \left(X_i\theta - \log \sum_{j \subset R_i} e^{X_j\theta} \right) \\ \frac{\partial l(\theta, X)}{\partial X_i} &= \delta(i)\theta - \sum_{i,j \in R_j, U} \frac{\theta e^{X_i\theta}}{\sum_{k \in R_j} e^{X_k\theta}} \end{aligned}$$

where U is the set of uncensored patients in the mini-batch, and R_i is the set of patients in the mini-batch whose survival or last follow-up is later than $t_{\text{obs},i}$ for patient X_i. In using the Cox partial log-likelihood in supervising deep survival models, a notable limitation is that the loss function is mini-batch dependent, as the partial log-likelihood is similar to a contrastive loss that depends on computing a loss term for each sample w.r.t. to "negative targets" / "at-risk" samples in R_i. However, in training with gigapixel WSIs, training with batch sizes greater than 1 is challenging due to: 1) space complexity of large bags, 2) variable bag sizes. Moreover, the Cox partial log-likelihood places additional strong assumptions in that all samples have the same baseline hazard function.

Weak Supervision with Limited Batch Sizes: A second approach to survival prediction using deep learning is to consider discrete time intervals and model each interval using an independent output neuron. This formulation overcomes the need for large mini-batches and allows the model to be optimized using single observations during training. Specifically, given right-censored survival outcome data, we build a discrete time survival model by partitioning the continuous time scale into non-overlapping bins: $[t_0, t_1), [t_1, t_2), [t_2, t_3), [t_3, t_4)$ based on the quartiles of survival time values (in months) of uncensored patients in each TCGA cohort. The discrete event time of each patient, indexed by j, with continuous event time $T_{j,\text{cont}}$ is then defined by:

$$T_j = r \text{ if } T_{j,\text{cont}} \in [t_r, t_{r+1}) \text{ for } r \in \{0, 1, 2, 3\} \tag{7}$$

Given the discrete time ground truth label of the j^{th} patient as Y_j. For a given patient with bag-level feature $\mathbf{h}_{\text{final}\,j}$, the last layer of the network uses the sigmoid activation and models the hazard function defined as:

$$f_{\text{hazard}}(r \mid \mathbf{h}_{\text{final}\,j}) = P(T_j = r \mid T_j \geq r, \mathbf{h}_{\text{final}\,j}) \tag{8}$$

which relates to the survival function through:

$$\begin{aligned} f_{\text{surv}}(r \mid \mathbf{h}_{\text{final}\,j}) &= P(T_j > r \mid \mathbf{h}_{\text{final}\,j}) \\ &= \prod_{u=1}^{r} (1 - f_{\text{hazard}}(u \mid \mathbf{h}_{\text{final}\,j})) \end{aligned} \tag{9}$$

$\mathcal{T}_h$ Present	$\mathcal{T}_g$ Present	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
-	-	0.576 ± 0.028	0.579 ± 0.031	0.809 ± 0.030	0.537 ± 0.051	0.614 ± 0.042	0.623
✓	-	0.639 ± 0.031	0.556 ± 0.077	0.790 ± 0.013	0.597 ± 0.062	0.622 ± 0.035	0.641
-	✓	0.619 ± 0.028	0.510 ± 0.086	0.819 ± 0.020	0.572 ± 0.045	0.628 ± 0.033	0.630
✓	✓	0.624 ± 0.034	0.580 ± 0.069	0.817 ± 0.021	0.620 ± 0.032	0.622 ± 0.019	**0.653**

Table 2: Ablation study assessing the impact of MIL Transformers in MCAT.

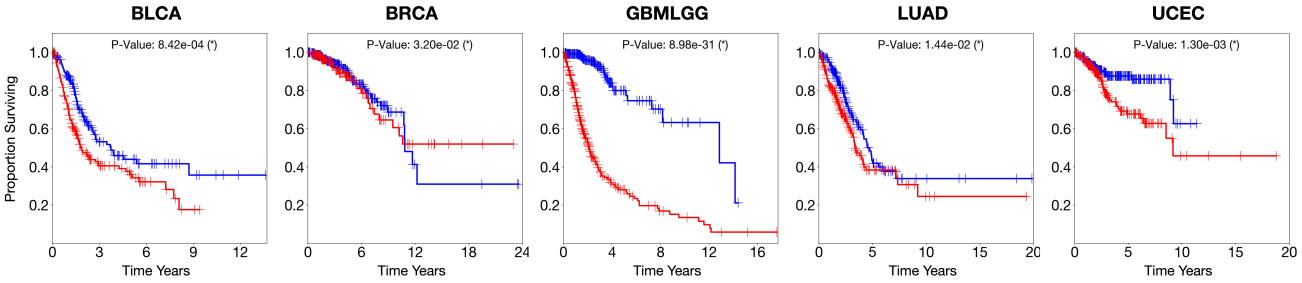

Figure 4: Kaplan-Meier Analysis showing low risk (blue) vs. high risk (red) patient stratification using predicted risk scores from MCAT w.r.t. ground truth survival time. The Logrank test was used to measure the statistical significance in comparing low vs. high risk patients as two different survival distributions.

During training, we update the model parameters using the log likelihood function for a discrete survival model [66], taking into account each patient's binary censorship status ($c_j = 1$ if the patient lived past the end of the follow-up period and $c_j = 0$ for patients who passed away during the recorded event time T_j):

$$\begin{aligned} L = -l = &-c_j \cdot \log\left(f_{\text{surv}}\left(Y_j \mid \mathbf{h}_{\text{final}\,j}\right)\right) \\ &- (1-c_j) \cdot \log\left(f_{\text{surv}}\left(Y_j - 1 \mid \mathbf{h}_{\text{final}\,j}\right)\right) \\ &- (1-c_j) \cdot \log\left(f_{\text{hazard}}\left(Y_j \mid \mathbf{h}_{\text{final}\,j}\right)\right) \end{aligned} \tag{10}$$

During training, we additionally up-weight the contribution of uncensored patient cases using a weighted sum of L and $L_{\text{uncensored}}$

$$L_{\text{surv}} = (1-\beta) \cdot L + \beta \cdot L_{\text{uncensored}} \tag{11}$$

The 2nd term of the loss function, which corresponds to uncensored patients only is computed as:

$$\begin{aligned} L_{\text{uncensored}} = &- (1-c_j) \cdot \log\left(f_{\text{surv}}\left(Y_j - 1 \mid \mathbf{h}_{\text{final}\,j}\right)\right) \\ &- (1-c_j) \cdot \log\left(f_{\text{hazard}}\left(Y_j \mid \mathbf{h}_{\text{final}\,j}\right)\right) \end{aligned} \tag{12}$$

7.3. Interpretability

To visualize multimodal interactions between WSIs and genomic features, we used a combination of attention-based and attribution-based interpretability to assess how features are used to predict risk. As discussed in § 3.2, we use the attention weights $A_{\text{coattn}} \in \mathbb{R}^{N \times M}$ computed by the GCA layer to visualize how image patches attend to each genomic embedding, where M is the bag size of the WSI and N is the number of genomic embeddings. To visualize attention maps for N different genomic embeddings, we can overlay the attention weights in each row in A_{coattn} onto the original WSI, as the Softmax operator was applied row-wise. Note that since gene attributes can belong to multiple functional categories from [32], the genomic-guided WSI embeddings $\widehat{H}_{\text{coattn}}$ may overlap with high attention attributed to similar morphological features. From visual assessment of two pathologists, we observe that computing A_{coattn} yields 2-4 unique attention maps across all cancer datasets using $N = 6$, with general observations that the oncogenesis embedding is able to localize all of the tumor regions, and the cytokine embedding localizes a mixture of tumor-adjacent stroma and immune-infiltrating tissue regions. To assess gene attribute and genomic embedding feature importance, we used Integrated Gradients (IG) [50], a gradient-based feature attribution method that attributes how prediction are made by the model with respect to the inputs. In the context of survival analysis, features with positive attribution contribute towards increasing the output value (high risk, pointing right), whereas negative attribution corresponds with decreasing the output value (low risk, pointing left). We use IG to visualize local, SHAP-like decision plots, which we perform for low risk and high risk cases for all cancer datasets in Figures 5-14.

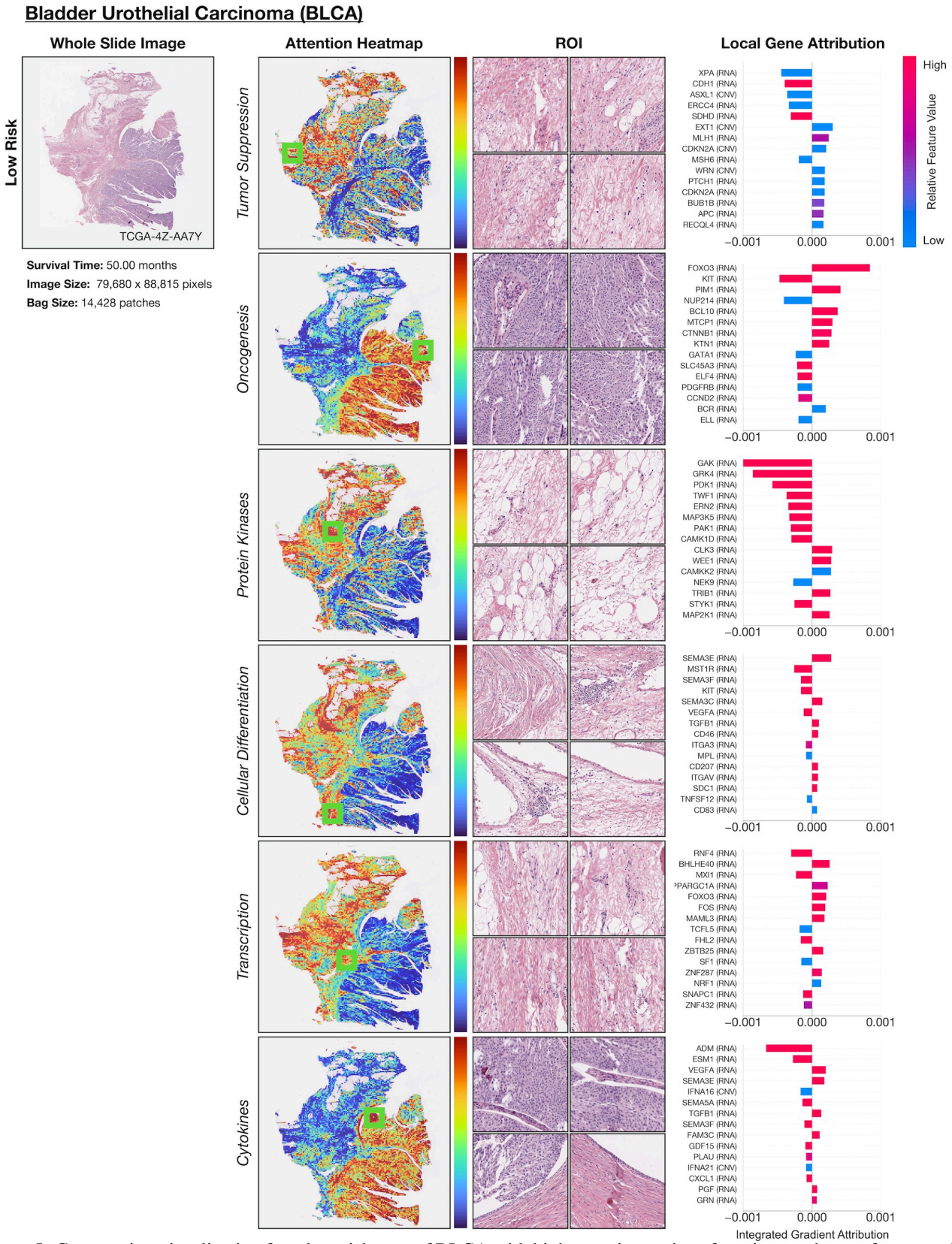

Figure 5: Co-attention visualization for a low-risk case of BLCA with high attention regions focusing on sheets of tumor cells for the Oncogenesis and Cytokines gene embeddings, and adjacent normal fibrous stroma and surrounding adipose tissue for the other gene embeddings.

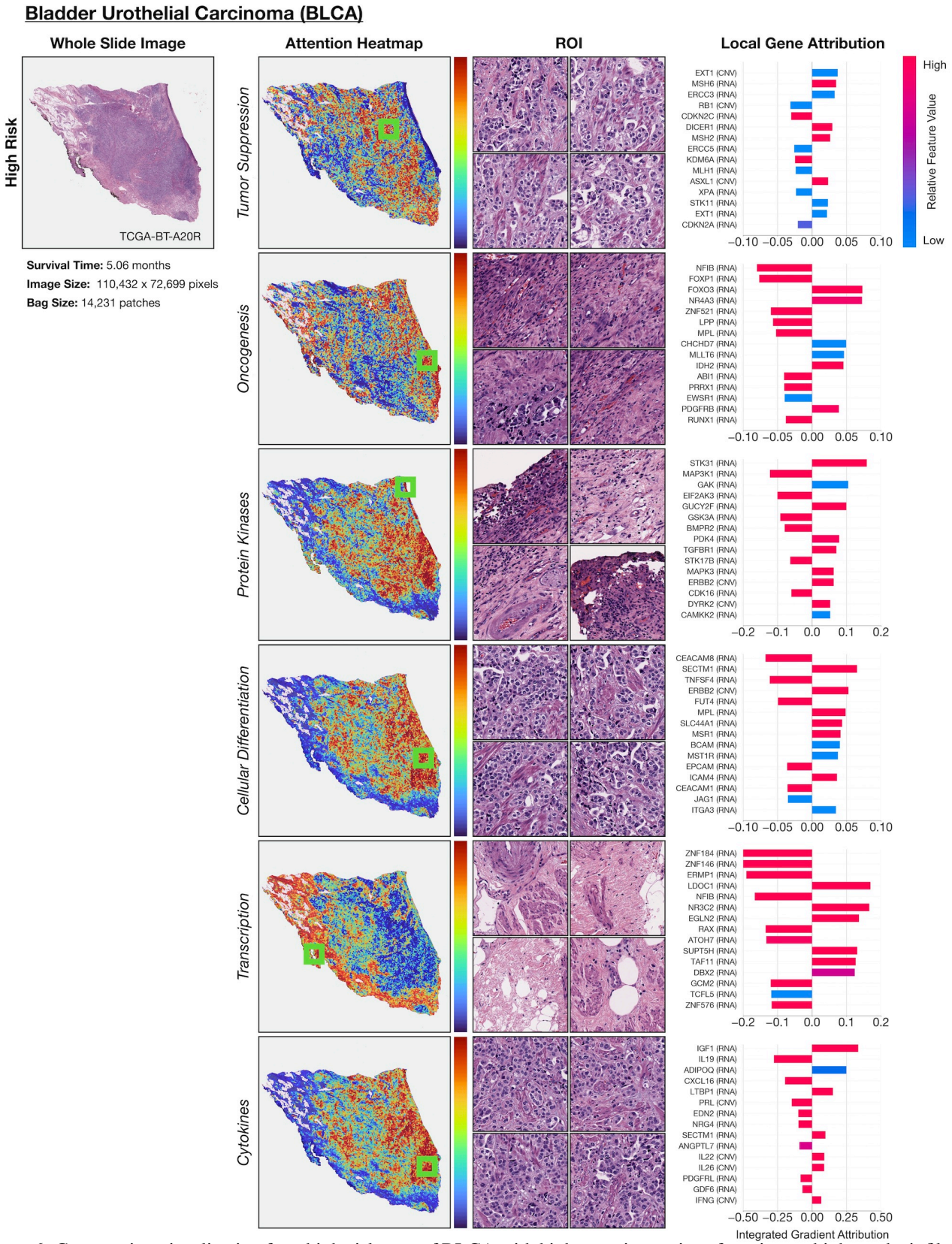

Figure 6: Co-attention visualization for a high-risk case of BLCA with high attention regions focusing on high-grade, infiltrative tumor for the Tumor Suppression, Oncogenesis, Cellular Differentiation, and Cytokines gene embeddings, and adjacent muscle, stroma, and large vessels for the other gene embeddings.

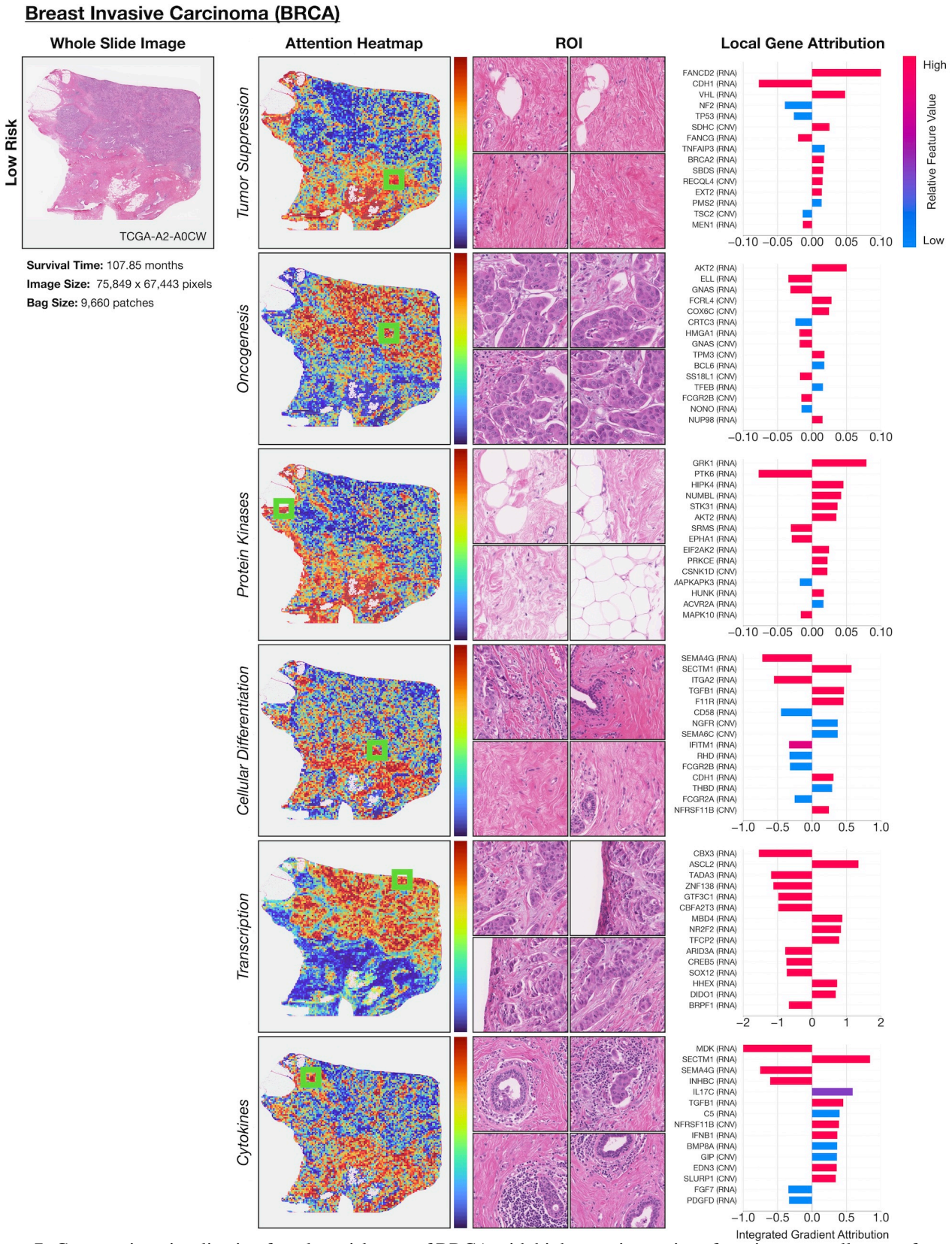

Figure 7: Co-attention visualization for a low-risk case of BRCA with high attention regions focusing on small nests of tumor cells for the Oncogenesis and Transcription gene embeddings, and adjacent dense, fibrous breast stroma, adipose tissue, and reactive background glands for the other gene embeddings.

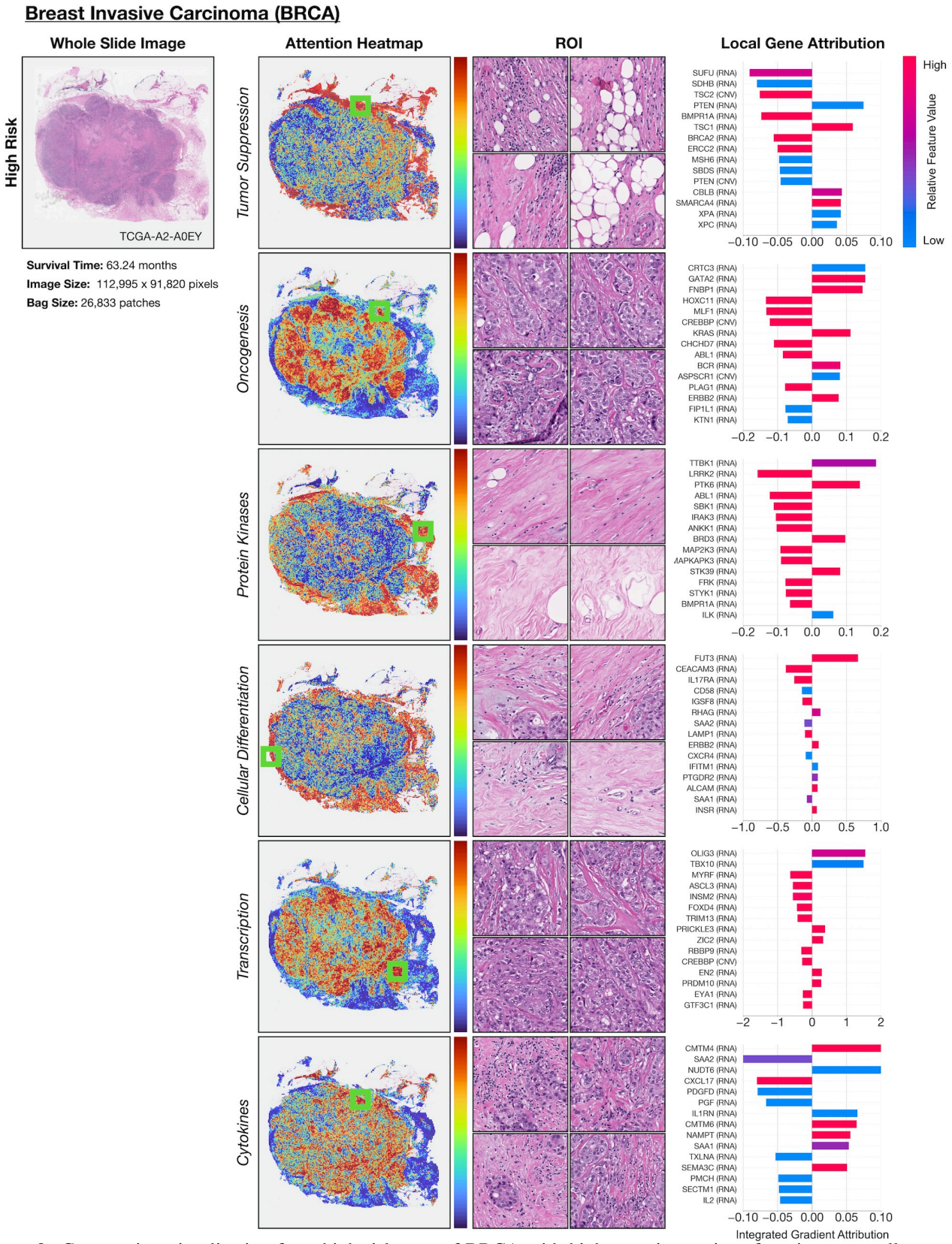

Figure 8: Co-attention visualization for a high-risk case of BRCA with high attention regions focusing on small nests of tumor cells for the Oncogenesis, Transcription, and Cytokines gene embeddings, and adjacent fibrous stroma, adipose tissue, and desmoplastic stroma for the other gene embeddings.

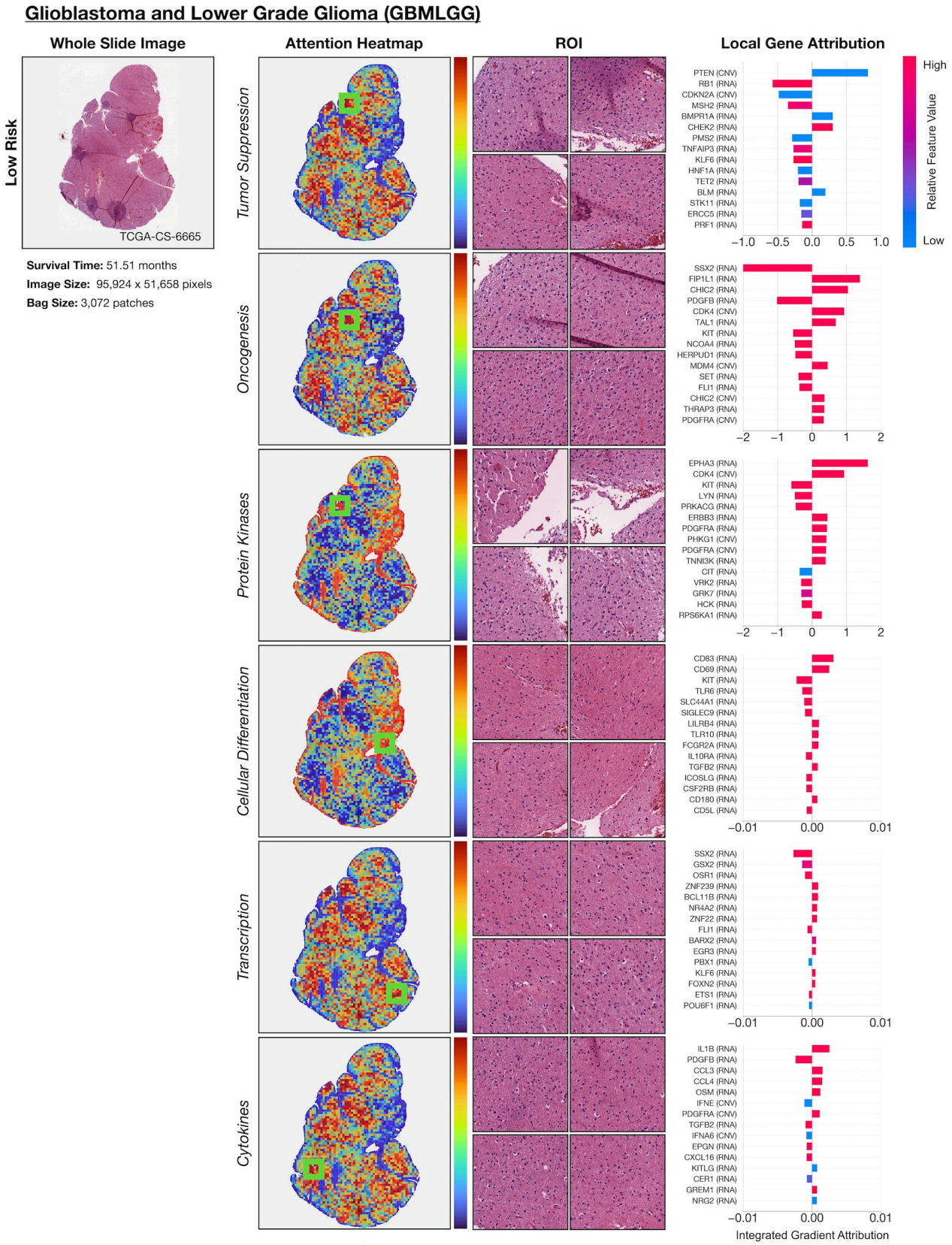

Figure 9: Co-attention visualization for a low-risk case of GBMLGG demonstrating high attention on areas of gliotic brain tissue for Tumor Suppression, Oncogenesis, Transcription, and Cytokines gene groups, and high attention on blood and relatively normocellular brain tissue for Protein Kinases and Cellular Differentiation gene embeddings.

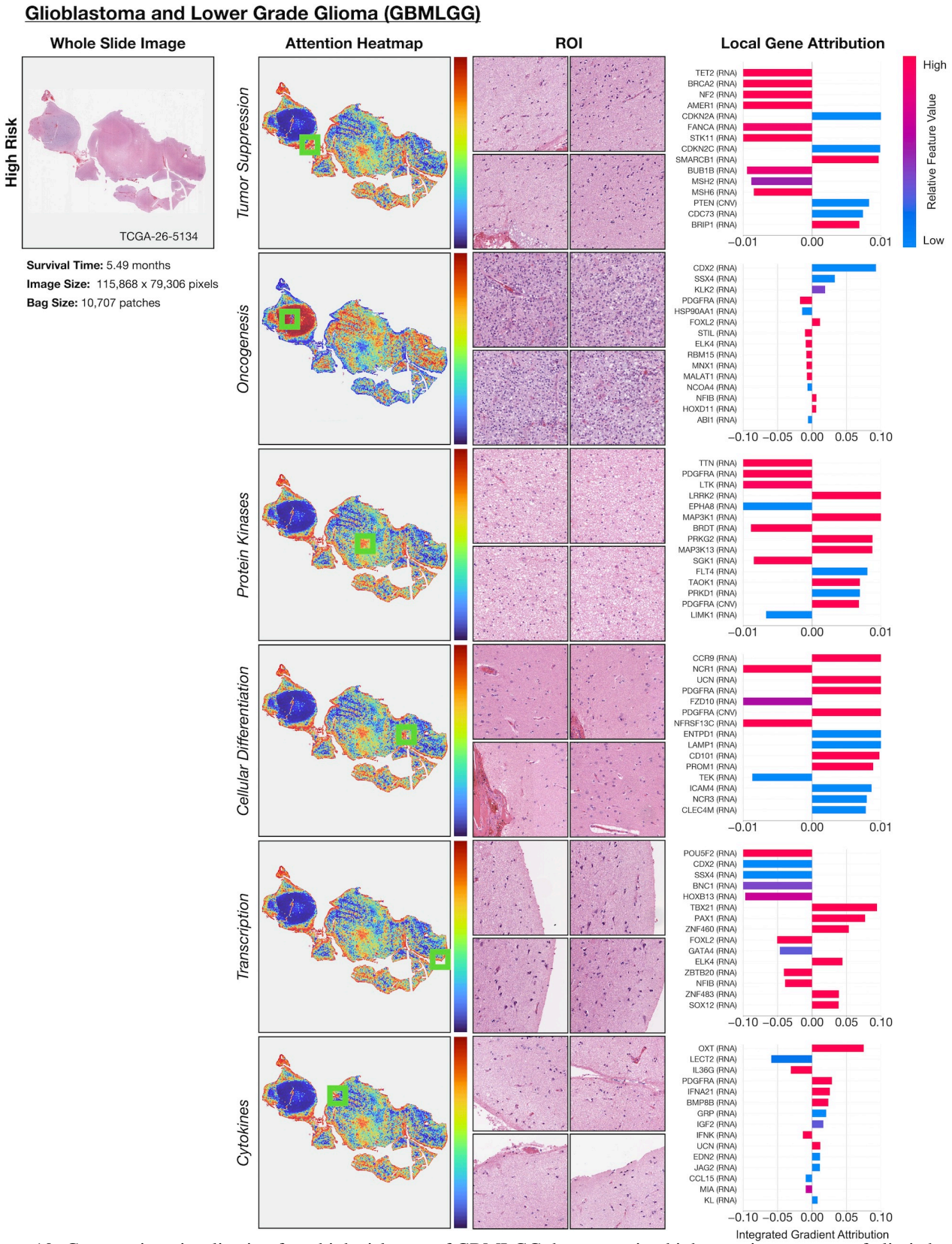

Figure 10: Co-attention visualization for a high-risk case of GBMLGG demonstrating high attention on areas of gliotic brain tissue for the Oncogenesis gene embedding and high attention on relatively normocellular brain tissue for Tumor Suppression, Protein Kinases, Cellular Differentiation, Transcription, and Cytokines gene embeddings.

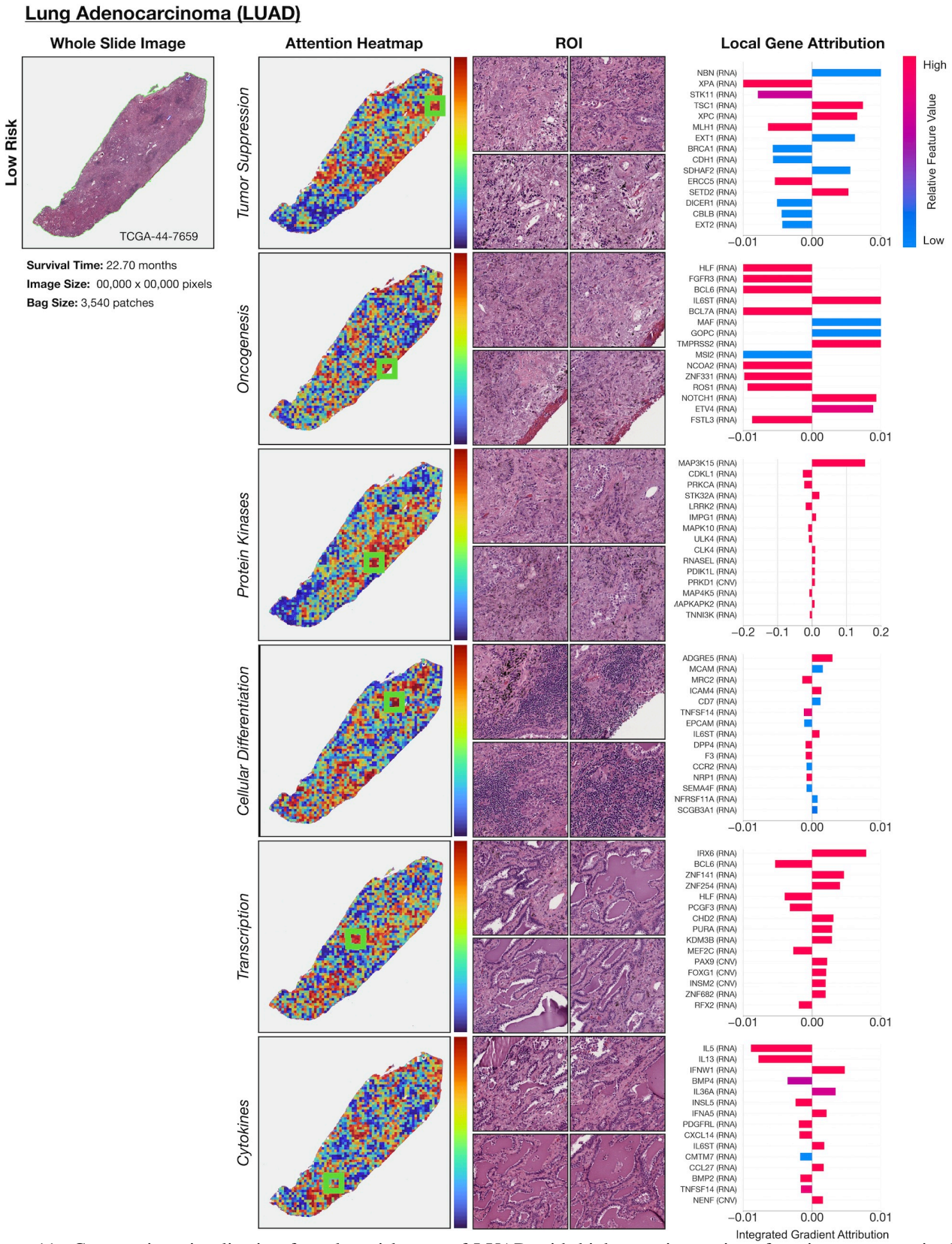

Figure 11: Co-attention visualization for a low-risk case of LUAD with high attention regions focusing on tumor in the Tumor Suppression, Oncogenes, and Protein Kinases gene groups, areas containing many lymphocytes in the Cellular Differentiation gene embeddings, and expanded airspaces containing proteinaceous fluid in the Transcription and Cytokines gene embeddings.

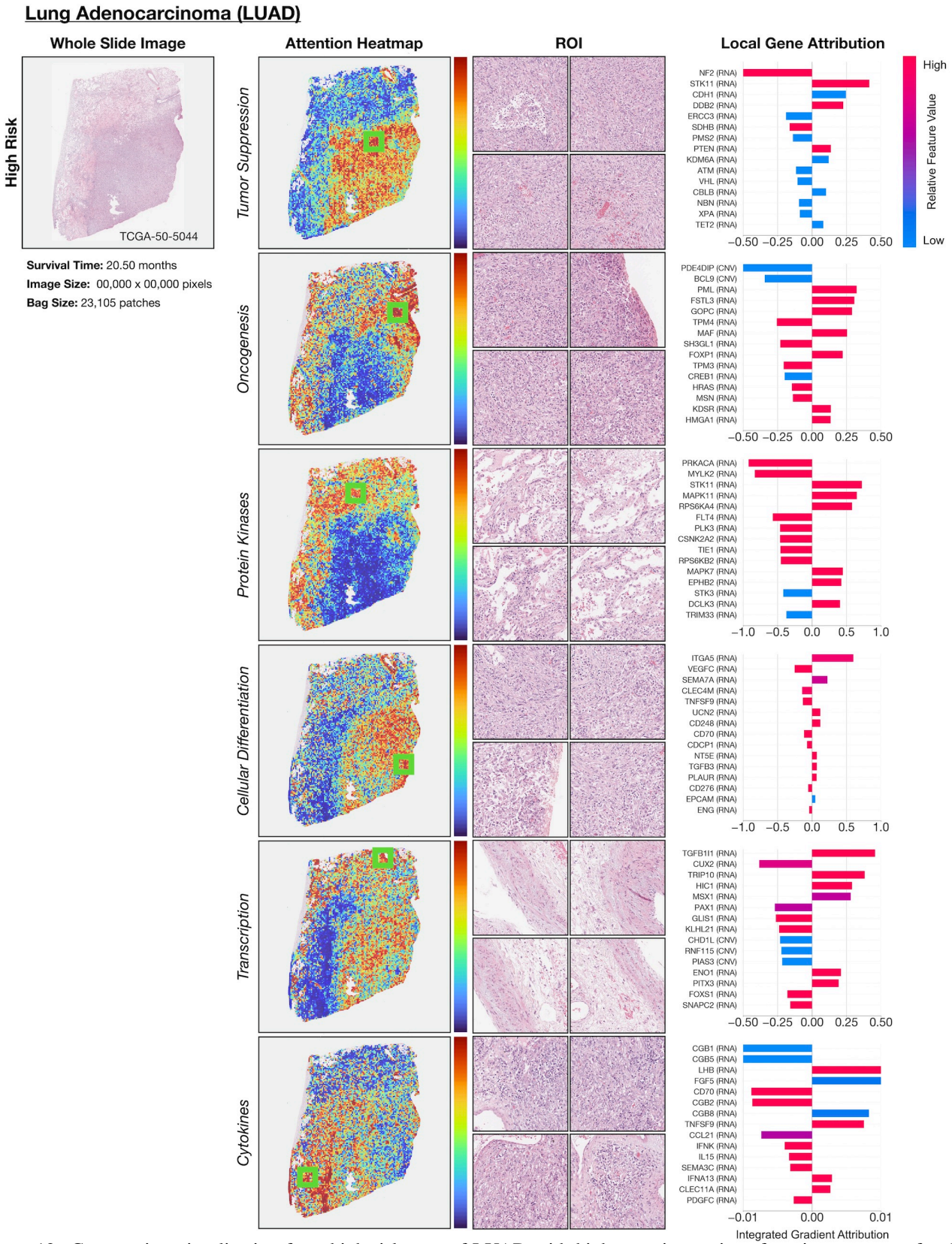

Figure 12: Co-attention visualization for a high-risk case of LUAD with high attention regions focusing on tumor for the Tumor Suppression and Cellular Differentiation gene embeddings, and adjacent normal lung, vessels, and scarred lung parenchyma for the other gene embeddings.

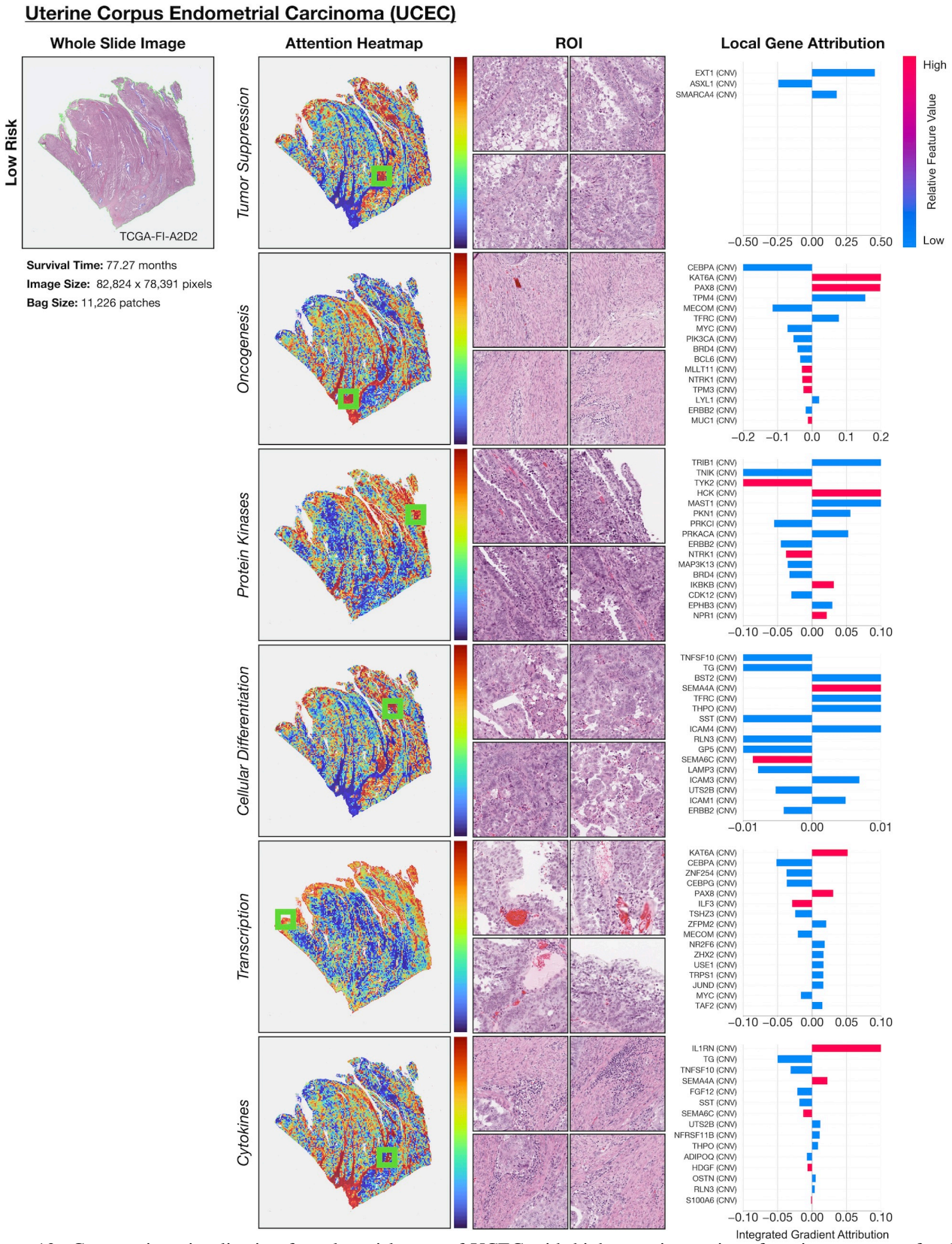

Figure 13: Co-attention visualization for a low-risk case of UCEC with high attention regions focusing on tumor for the Tumor Suppression, Protein Kinases, Cellular Differentiation, Transcription, and Cytokines gene embeddings, and adjacent background muscle and desmoplastic stroma for the other gene embeddings.

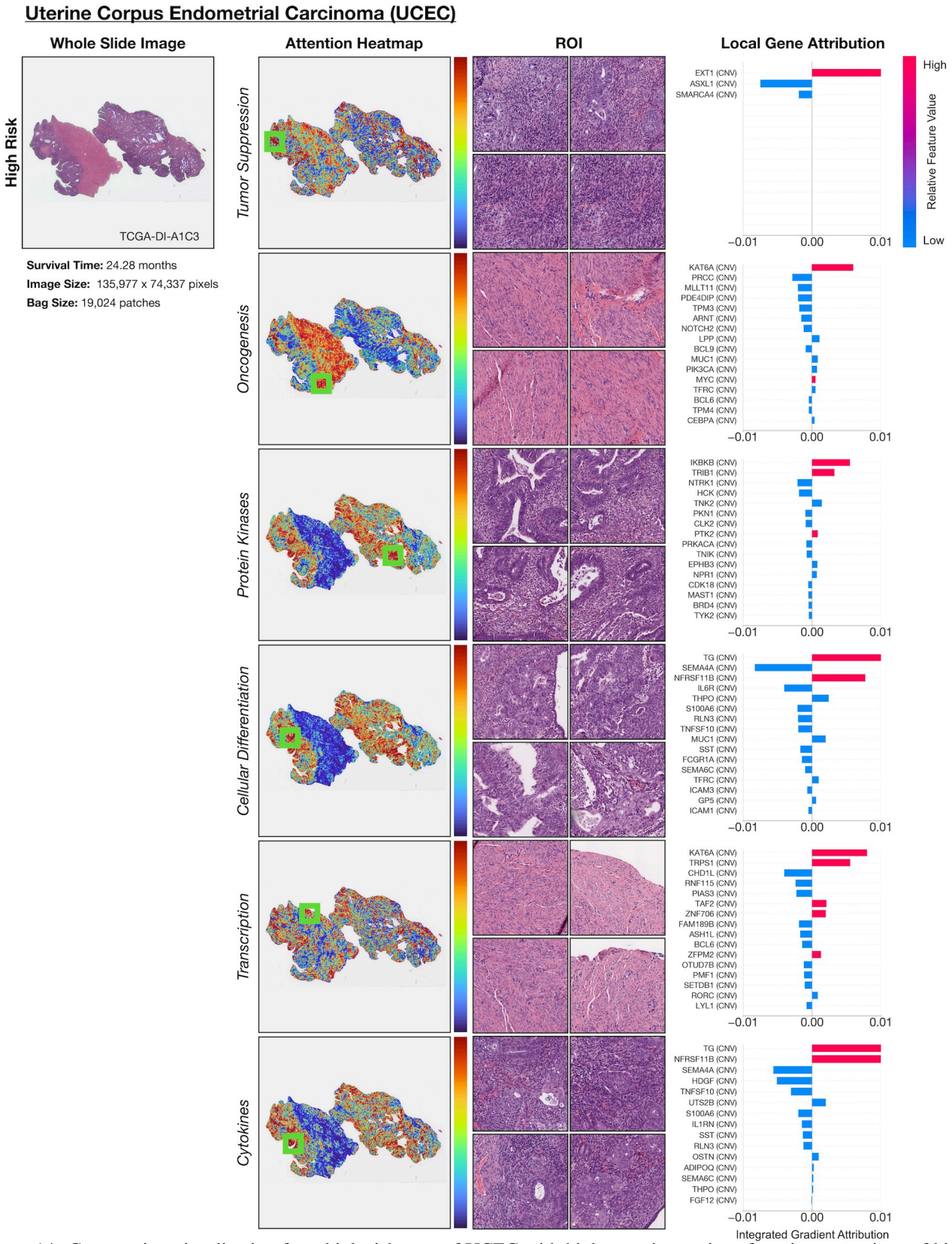

Figure 14: Co-attention visualization for a high-risk case of UCEC with high attention regions focusing on regions of high grade tumor with surrounding desmoplastic stroma for the Tumor Suppression, Protein Kinases, Cellular Differentiation, Cytokines gene embeddings, and adjacent background muscle for the other gene embeddings.