Deep Learning Identifies FGFR Alterations from H&E Whole Slide Images in Bladder Cancer

Poster 119

Josh Och[1], Bolesław L Osinski[1], Kshitij Ingale[1], Sun Hae Hong[1], Caleb Willis[1], Rohan P Joshi[1], Nike Beaubier[1], and Martin C Stumpe[1]

[1]Tempus Labs, Inc., Chicago, IL // **Disclosure:** All authors are employees of Tempus Labs, a for-profit company // **Correspondence:** josh.och@tempus.com

INTRODUCTION

Several targeted therapies for FGFR alterations in bladder cancer are either currently in clinical trials or already FDA-approved. FGFR alterations — including activating single nucleotide variants (SNVs) and fusions — are common in bladder cancer and detectable via next-generation sequencing (NGS). The ability to rapidly screen patients based on routine pathology would help prioritize patients for NGS testing. Here, we developed a model using H&E whole slide images (WSIs) to predict FGFR alterations using real-world data.

METHODS

WSIs and ground truth labels pertaining to FGFR mutational status (obtained by DNA-seq) were collected from primary and metastatic bladder cancer specimens (**n=3,652, Table 1**). Positive labels, denoted FGFR+, were defined as those harboring a pathogenic SNV or fusion of FGFR, as confirmed by a molecular pathologist (n=577, including 556 FGFR3, and 21 either FGFR1, FGFR2 or FGFR4). Model development was performed as follows: i) a custom attention-based convolutional neural network with ResNet-18 backbone was trained to predict FGFR status from each WSI in the training set (60%), ii) hyperparameters were selected using an optimization set (20%) and iii) performance was reported on a test set of data (20%). Training, optimization, and testing was performed in 5-fold cross-validation (CV). Cohorts were stratified to maintain a similar distribution of tissue sites and scanner types across each fold. Finally, to assess generalizability, the same 5 model folds were evaluated on a set of TCGA Bladder Cancer diagnostic slides (n=383, including 52 FGFR+).

Covariate	Value (sample size, % FGFR+)	p-value*
Tissue Site	Bladder (2072, 15.4%), Other (1580, 16.3%)	0.082
Scanner Make	Philips (1823, 17.6%), Leica (1829, 14.0%)	0.003
Tumor Grade	High Grade (2028, 16.0%), Low Grade (103, 32.0%), Unknown (1521, 14.4%)	3.6e-06

Table 1. For select covariates, the number of samples in each group and the FGFR positivity rate for that group.
*p-value (chi-square test) indicates the likelihood of this breakdown occurring assuming that the given covariate has no relationship with FGFR positivity.

SUMMARY

- An attention-based deep multiple instance learning model trained on **H&E whole-slide images** is capable of **predicting FGFR SNVs and fusions** in bladder cancer.
- Model performance was similar across tissue sites, scanner types, and tumor grades, and it **generalized** to an external TCGA cohort.

RESULTS

Illustration of deep learning pipeline for predicting FGFR status from WSIs

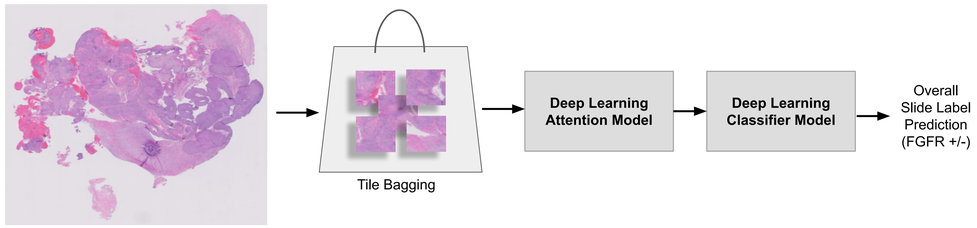

Figure 1. Whole slide images (WSIs) are broken up into tiles. Tile data are grouped by slide-level classes and passed to two deep learning modules to create a prediction of FGFR status. The model weights are iteratively updated until the area under the receiver operating characteristic curve (AUROC) for the validation set no longer improves.

Tumor tiles were more often used by the model to make correct FGFR +/- predictions than stromal areas or tiles with low tissue content

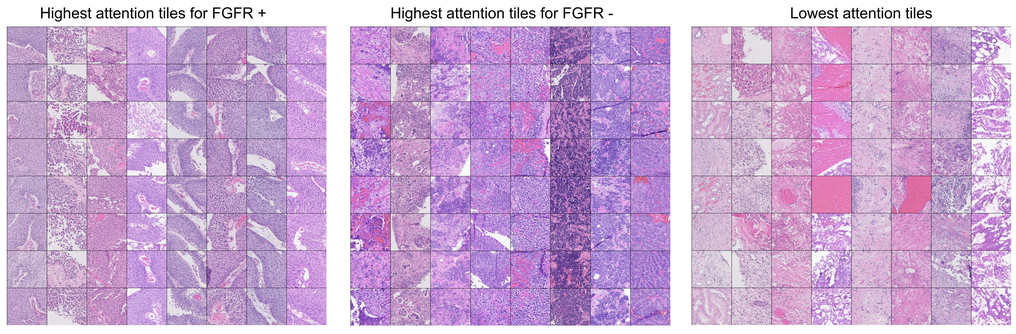

Figure 2. Top 8 highest attention tiles for WSIs correctly predicted as FGFR+ (left) and FGFR- (middle). Right: Top 8 lowest attention tiles. Each column represents a different WSI.

Acknowledgements: We thank the Tempus Scientific Communications team for poster review and Kunal Nagpal for insightful discussions.

FGFR predictions are reliable and generalizable

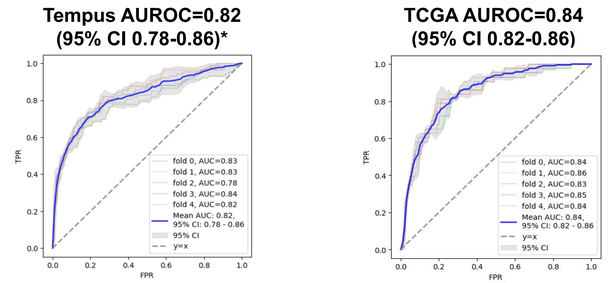

Figure 3. ROC curves for models evaluated on Tempus (left) and TCGA (right) test sets. Light-colored lines are ROC curves for each of the 5 folds, the bold line shows the mean across folds, and the shaded area shows a 95% confidence interval.
*Tempus AUROC significantly outperforms a linear model trained only on clinical and confounding features. Baseline linear model AUROC=0.60 (95% CI 0.56-0.65)
Baseline Model Features: Stage, Grade, Tissue site, Scanner, Procedure type, Tumor %, Race

Model performance is consistent across tissue sites, scanner types and tumor grades

Subgroup	FGFR- Samples	FGFR+ Samples	AUROC (95% CI)
Bladder Site	1752	320	0.85 (0.79-0.91)
Other Tissue Site	1323	257	0.78 (0.67-0.89)
Philips Scanner	1502	321	0.81 (0.75-0.87)
Leica Scanner	1573	256	0.82 (0.73-0.92)
High Grade	1704	324	0.80 (0.76-0.85)
Low Grade	70	33	0.94 (0.86-1.00)
Unknown Grade	1301	220	0.79 (0.72-0.86)

Table 2. AUROC for our model on each sample subset. 95% CI is computed using the 5 folds in our study and assuming a normal distribution.