To retain the most influential variables in a multivariate setting, a regression model (Logit) using only the 50 & 120’ variable was run on the merged dataset including the three training sets (DivTrottoli/Sumi) and only the variables verifying P-value 0.05 in the multivariate model were retained in the S1 list. We obtained a list of 7 variables (S1).

3. Model generation and selection

For the fitting of the model derived from the selection [S1], we used the merged dataset restricted to stage 1 adenocarcinoma (ADC) “DivTrottoli” and then reapplied the fitted model to each dataset separately.

- Using the S2 variables, we trained a regression model (Logit) in DivTrottoli. We then applied the fitted model to the validation datasets.
- ROCR AUC was computed on oyxcluiring patients with less than 36 month follow-up.
- Kaplan Meier curves were built and log-rank test p-value were computed to assess the model predictive performance to discriminate high- and low-risk groups using all available patient OS data. High- and low-risk were defined with a cut-off of returned probabilities of 0.5.

4. Model interpretation

We used lasso penalized linear models on the TCGA LUAD cohort to identify the mutations and the copy number aberrations (amplification / deletion) associated with the scores predicted by our model.

The final model comprising the following genes: FOIL2, HS3DL1, INGS, PDE4H, POU2F1, RARRES2 AND TIMP2 had an AUC > 0.70 in the 3 training sets. It was also robust and statistically significant in the independent (P < 1e-10) and the pooled test datasets (P < 0.03) restricted to adenocarcinomas (ADC).

Model performance in the pooled test dataset measured by ROCR AUC and Kaplan Meier curve

1. The model was robust in all 3 test datasets restricted to ADC

In all 3 datasets: AUC > 0.64 and consistent discrimination of high vs low risk groups. For pval < 0.03.

Performance (ROC AUC) of model in the 3 independent tests datasets

2. The model is specific to stage 1 ADC

Our model was generated on stage 1 lung ADC and is specific to this group: it is robust but not statistically significant in the pooled test datasets restricted to stage 2 ADC (p-value 0.03).

Performance (Kaplan Meier Curve) of model in the pooled test dataset restricted to stage 1 SCC and in the pooled test dataset restricted to stage 2 ADC

3. The model is biologically relevant

The mutations and the copy number aberrations associated with high risk profiles predicted by the signature are coherent and consistent with the literature: EGFR mutation or amplification, CDKN2A deletion, PTK7 amplification. No conclusive prediction of response to adjuvant cisplatin based chemotherapy

No conclusive prediction of response to adjuvant cisplatin based chemotherapy could be achieved on the JBR.10 data, possibly due to small population; no association with sensitivity to cisplatin was found on the Sanger cell line panel.