A performant gene expression signature robustly predicts the outcome of patients diagnosed with stage 1 lung adenocarcinoma

Mathilde Bateson* [1], Charles Ferté[2], Yann Gaston-Mathé[3], Justin Guinney[4], Benjamin Besse[2], Jean-Pierre Armand[2], Jean-Charles Soria[2]

(*): corresponding author (mathilde.bateson@institut-hypercube.org) (1): Institut HyperCube; (2): Gustave Roussy; (3): YGM Consult SAS; (4): Sage Bionetworks

BACKGROUND

Although the management of metastatic lung cancer has been profoundly modified by the identification of actionable molecular traits, decision making for early-stage lung cancer still relies on the tumor stage (TNM) only.

Stage 1 patients are considered of good prognosis and usually do not receive adjuvant therapy. However, about 30% of those patients do recur within 5 years and eventually die from cancer. The better identification of these high-risk patients who could benefit of adjuvant therapy remains challenging.

The recent availability of high-dimensional molecular data (gene expression data) for lung cancer patients, simultaneously to the development of novel data mining methods are expected to dramatically improve such predictive challenges.

In this work, we aimed to increase the statistical power and model robustness by using in combination several publically available data sets to discover a robust predictive signature of outcome.

METHODS

1. Datasets and data preparation

Stringent Inclusion criteria :
- Publically available gene expression datasets
- Stage IA-IB Lung adenocarcinoma or squamous cells patients
- Surgical resection (R0)
- No adjuvant therapy (CT, RT)
- Overall survival data (36M+ follow-up)

Eight datasets were used :

- Three training sets: Directors Challenge (n=198); Hou et al (n=37); Bhattacharjee et al (n=70)
- Five test sets: Zhu et al (JBR.10 trial) (n=31), Rousseaux et al (n=122), TCGA LUAD (n=60), Raponi et al (n=59) and TCGA LUSC (n=35)

Gene expression data were normalized using RMA method and rescaled. The set of genes expression variables was restricted to the 8492 gene expression variables present in all datasets. Clinical covariates were Histology, Age and Sex. The output variable was the 3-year survival (yes/no): "os3yr".

Data included in the 3-year-survival study : distribution of important clinical variables

		DIR (n=198)	Hou (n=37)	Bhat (n=70)	All Training (n=305)	Zhu (n=31)	Rous (n=122)	TCGA LUAD (n=60)	TCGA LUSC (n=35)	Raponi (n=59)
Histology	ADC	198 (100%)	22 (59%)	70 (100%)	**290 (95%)**	19 (61%)	73 (60%)	60 (100%)	0 (0%)	0 (0%)
	SCC	0 (0%)	15 (41%)	0 (0%)	**15 (5%)**	12 (39%)	49 (40%)	0 (0%)	35 (100%)	59 (100%)
Stage	1A	96 (48%)	9 (24%)	33 (47%)	**138 (45%)**	0 (0%)	112 (92%)	22 (37%)	6 (17%)	21 (36%)
	1B	102 (52%)	28 (76%)	37 (53%)	**167 (55%)**	31 (100%)	10 (8%)	38 (63%)	29 (83%)	38 (64%)
Gender	Female	99 (50%)	10 (27%)	40 (57%)	**149 (49%)**	11 (35%)	16 (13%)	32 (53%)	9 (26%)	22 (37%)
Prob of 3yr survival (%)		81%	60%	71%	**76%**	77%	71%	70%	43%	61%

Data included in the overall survival study of Stage 1 Lung adenocarcinoma

	DIR (n=217)	Hou (n=40)	Bhat (n=70)	All Training (n=327)	Zhu (n=32)	Rous (n=128)	TCGA LUAD (n=200)	TCGA LUSC (n=63)	Raponi (n=73)
Median month to last contact or death	56	54	49	**53**	68	62	11	16	35

2. Variable selection methods

We applied a weighted logistic regression model to each learning set (Dir, Hou, Bhat): GLM(os3yr ~ var i), for each 8492 variables (i.e. genes).

With the underlying assumption that if a variable has a true biological impact on 3-year survival (os3yr), it should be visible and consistent in each 3 of the training datasets, we then selected the variables which satisfied two conditions:

- Condition 1 (C1): Wald's P-value < 0.05 for each training dataset
- Condition 2 (C2): Same sign of regression coefficient in all training datasets.

Additionally a weighted Cox regression model was applied to each learning set (Dir, Hou, Bhat), for each of the 8492 features. The S0' variables were then selected when they satisfied
- Condition 1' (C1'): Wald's P-value < 0.1 for each training dataset and (C2).

To retain the most influential variables in a multivariate setting, a regression model (Logit) using only the S0 + S0' variables was run on the merged dataset including the three training sets (DirHouBhat) 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) "DirHouBhat" and then reapplied the fitted model to each dataset separately.

- Using the S1 variables, we trained a regression model (Logit) in DirHouBhat. We then applied the fitted model to the validation datasets.
- **ROC AUC** was computed on os3yr excluding 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

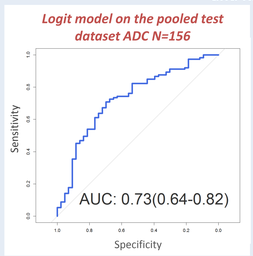
4. Model interpretation

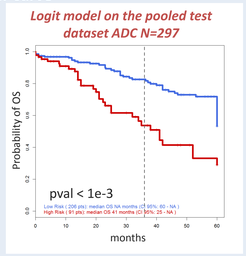
We used lasso penalized linear models on the TCGA LUAD cohort to identify the mutations and the copy number aberrations (amplification / deletion) associated with the scores predicted by our model.

RESULTS

The final model comprising the following genes: **FOSL2, HSD3B1, ING3, PDE6H, POU2F1, RARRES3** and **TIMP2** had an **AUC > 0.70** in the 3 learning sets. It was also robust and statistically significant in the independent **(p < 1e-3)** and the pooled test datasets (p =< 0.03) restricted to adenocarcinomas (ADC).

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

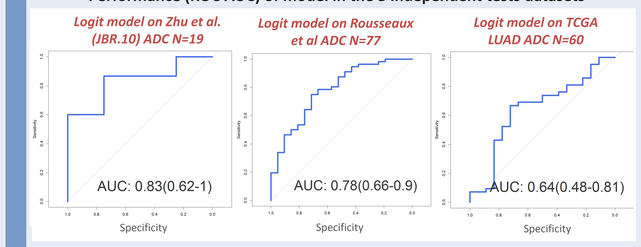




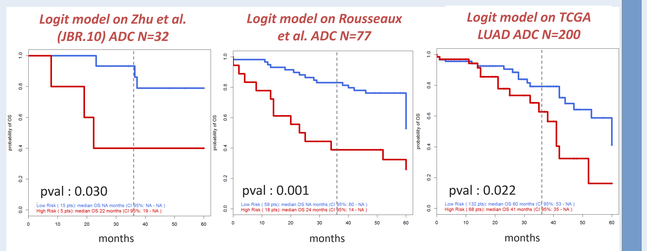
1. The model was robust in all 3 test datasets restricted to ADC

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

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



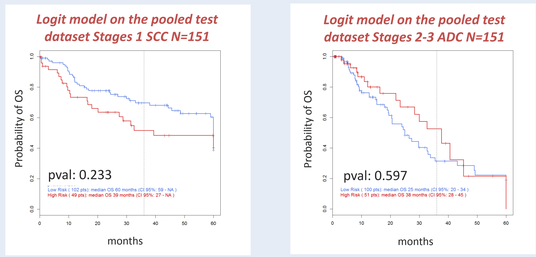
Kaplan Meier curves of overall survival for high- vs. low-risk predicted by model



2. The model is specific to stage 1 ADC

Our model was generated on stage 1 lung ADC and is specific this group : it is robust but not statistically significant in the pooled test datasets restricted to stage 1 squamous cells carcinomas (SCC). It performed poorly when applied to the independent and the pooled test datasets restricted to stage 2-3 ADC.

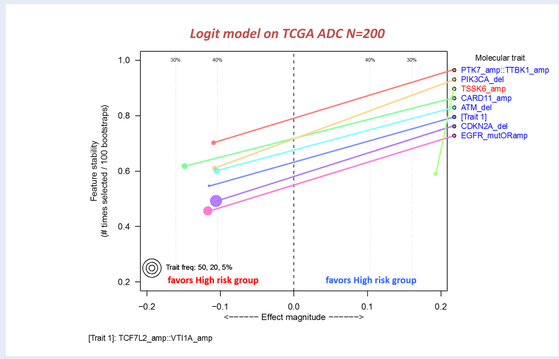
Performance (Kaplan Meier Curve) of model in the pooled test dataset restricted to stage 1 SCC and in the pooled test dataset restricted to stage 2-3 ADC



3. The model is biologically relevant

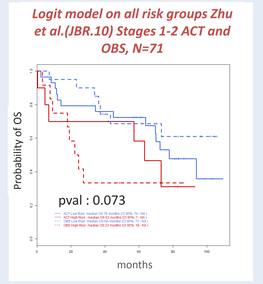
The mutations and the copy number aberrations associated with high risk profiles predicted by the signature are coherent and consistent with the literature: EGFR mutation or amplification, CDKN2A deletion, PTK7 amplification.

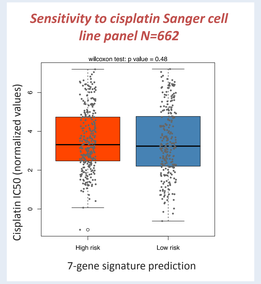
Mutation and copy number aberrations associated with the gene expression signature



4. No conclusive prediction of response to adjuvant cisplatin based chemotherapy

We replicated our model on the JBR.10 data restricted to ADC, where 32 patients were in the observation (OBS) group, while 39 received ACT. We included stages 1-2 of ADC NSCLC to increase statistical power.





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.

DISCUSSION

A robust model on 3 independent test sets

Variable selection from 3 different high-dimensionality gene expression data sets produced a high-performing and robust prognosis classifier in 3 independent test datasets. This robust approach relies on the identification of gene expression patterns that are similarly associated to prognosis in several independent datasets, to limit the risk of false discovery. We believe this approach may hold promises for future discoveries of meaningful and robust classifiers in high-dimensionality data sets.

Clinical utility : a simple and performant signature specific to stage 1 ADC

We developed a highly performant gene expression signature predicting overall survival, which is specific to stage 1 lung ADC patients and yields biological relevance.

Importantly, a number of facts strongly strengthen the translation potential of our predictors:

i. the stringent criteria we used to select our training and test datasets (known pathological stage, no adjuvant chemotherapy or radiotherapy, follow up, R0 resection, follow-up > 36 months)
ii. the number (n=3) and the different nature of the test datasets (Agilent CGH arrays, Affymetrix CGH arrays, RNA-Seq data)
iii. The mutations and the copy number aberrations associated with the signature were coherent with the literature

Our signature should be replicated in a larger trial to assess their potential value for predicting the response of early-stage lung cancer to adjuvant chemotherapy.

These promising results have the potential to change the decision making at bedside for stage 1 lung ADC patients and to improve the stratification of patients in future clinical trials.

REFERENCES

Ferté C, Trister AD, Huang E, Bot BM, Guinney J, Commo F, Sieberts S, André F, Besse B, Soria JC, Friend SH. Impact of bioinformatic procedures in the development and translation of high-throughput molecular classifiers in oncology. Clin Cancer Res. 2013 Aug 15;19(16):4315-25.

Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma, Shedden K, Taylor JMG, Enkemann SA, Tsao M-S, Yeatman TJ, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site,blinded validation study. Nat Med.2008;14:822-7.

Zhu C-Q, Ding K, Strumpf D, Weir B a, Meyerson M, Pennell N, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. J Clin Oncol. 2010;28:4417–24.

Hou J, Aerts J, Den Hamer B, Van Ijcken W, Den Bakker M, Riegman P, et al. Gene expression based classification of non-small cell lung carcinomas and survival prediction. PloS One. 2010;5:e10312.

Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, Moro-Sibilot D, Brichon PY, Lantuejoul S, Hainaut P, Laffaire J, de Reyniès A, Beer DG, Timsit JF, Brambilla C, Brambilla E, Khochbin S. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med. 2013 May 22;5(186):186ra66.

Bhattacharjee A1, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001 Nov 20;98(24):13790-5