Performant gene expression signatures predicting overall survival in resected stage I NSCLC have potential for translation in the clinic

Mathilde Bateson* [1], Charles Ferté[2], Yann Gaston-Mathé[3], David Planchard[2], 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

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.

30% of the Stage I patients recur within the 5 years after the tumor resection, leading to their death by cancer in most cases.

To improve the prediction of probability of overall survival in these patients is critical, in order to better identify the stage I patients at high risk of recurrence or death that could benefit of adjuvant therapy.

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 robust predictive signatures 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), Raponi et al (n=59), TCGA LUSC (n=35) and TCGA LUAD (n=60)

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)	Raponi (n=59)	Lusc (n=35)	Luad (n=60)
Histology	ADC	198 (100%)	22 (59%)	70 (100%)	**290 (95%)**	19 (61%)	73 (60%)	0 (0%)	0 (0%)	60 (100%)
	SCC	0 (0%)	15 (41%)	0 (0%)	**15 (5%)**	12 (39%)	49 (40%)	59 (100%)	35 (100%)	0 (0%)
Stage	1A	96 (48%)	9 (24%)	33 (47%)	**138 (45%)**	0 (0%)	112 (92%)	21 (36%)	6 (17%)	22 (37%)
	1B	102 (52%)	28 (76%)	37 (53%)	**167 (55%)**	31 (100%)	10 (8%)	38 (64%)	29 (83%)	38 (63%)
Gender	Female	99 (50%)	10 (27%)	40 (57%)	**149 (49%)**	11 (35%)	16 (13%)	22 (37%)	9 (26%)	32 (53%)
3yr survival	% os3yr=0	19%	40%	29%	**24%**	23%	29%	39%	57%	30%

Data included in the overall survival study of Stage 1 NSCLC

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

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.

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:

- Wald's P-value < 0.05 for each training dataset
- Same sign of regression coefficient in all training datasets.

We then ran a multivariate GLM model with the selected variables on a training data set named "DirHouBhat" (merge of the 3 learning datasets), and we removed all variables not verifying Wald's P-value < 0.01.

We obtained a list of 6 variables (S1) verifying P-value < 0.01 in the training set.

3. Model generation and selection

For the fitting of the model derived from the selection (S1), we used the merged dataset DirHouBhat, and then reapplied the fitted model to each dataset separately.

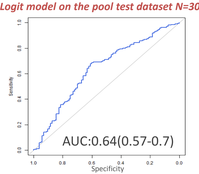
- Using the S1 variables, we trained a penalized regression model (ElasticNet) in DirHouBhat. The hyperparameter alpha was set to .1, the hyperparameter lambda was determined using a 10 fold cross validation approach. We then applied the fitted model to the validation datasets.
- ROC AUC was computed 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

RESULTS

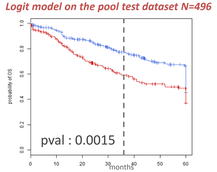
The final model comprising the following genes: HSD3B1, ING3, PDE6H, POU2F1, RARRES3 and TIMP2 had an AUC significantly > 0.5 in the 3 learning sets. It was also robust and statistically significant (p = 10-3) in the pooled test dataset.

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

Logit model on the pool test dataset N=307



Logit model on the pool test dataset N=496

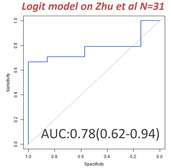


1. The model was robust in all 5 test datasets

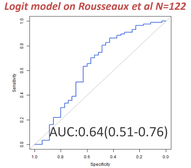
AUC > 0.5 in all 5 datasets and consistent discrimination of high vs low-risk groups in all 5 datasets

Performance (ROC-AUC) of model in the 5 independent tests datasets

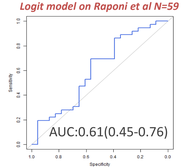
Logit model on Zhu et al N=31



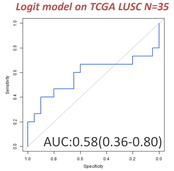
Logit model on Rousseaux et al N=122



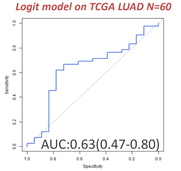
Logit model on Raponi et al N=59



Logit model on TCGA LUSC N=35

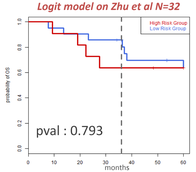


Logit model on TCGA LUAD N=60

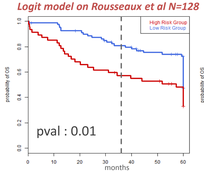


Kaplan Meier curves of overall survival for high- vs. low-risk predicted by model

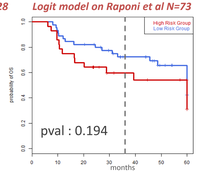
Logit model on Zhu et al N=32



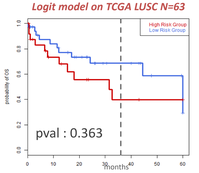
Logit model on Rousseaux et al N=128



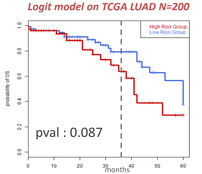
Logit model on Raponi et al N=73



Logit model on TCGA LUSC N=63



Logit model on TCGA LUAD N=200



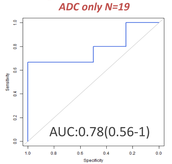
- P-value in KM curves is significant in two datasets (Rousseaux and TCGA LUAD). Small number of patients in validation cohorts impacts on statistical significance in Zhu, Raponi and Lusc
- Model performance in TCGA LUSC and Raponi et al. is lower, possibly due to different composition of population (100% SCC, whereas only 5% SCC in total learning data)

2. POU2F1 expression predicts 3-year survival in ADC patients

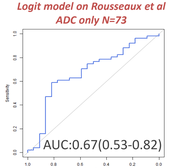
Of the 6 variables from the previous model, POU2F1 was found to be have a good predictive performance of overall survival, specifically in ADC patients.

Performance (ROC-AUC) of logit model GLM(os3yr~POU2F1) in the three independent test datasets restricted to ADC only

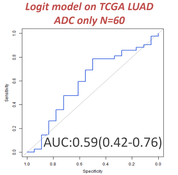
Logit model on Zhu et al ADC only N=19



Logit model on Rousseaux et al ADC only N=73

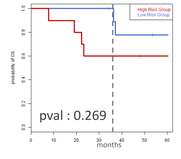


Logit model on TCGA LUAD ADC only N=60

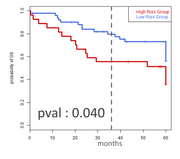


Kaplan Meier curves of overall survival for high- vs. low-risk predicted by logit model GLM(os3yr~POU2F1) in the three independent test datasets restricted to ADC only

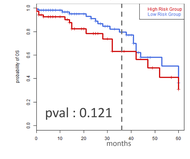
Logit model on Zhu et al ADC only N=20



Logit model on Rousseaux et al ADC only N=78



Logit model on TCGA LUAD ADC only N=200



DISCUSSION

A robust model on 5 independent test sets

Variable selection from 3 different high-dimensionality gene expression data sets produced a high-performing and robust prognosis classifier in 5 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

To our knowledge, this is the first time that a gene expression signature predicting outcome for resected stage 1 NSCLC is successfully replicated in 5 different independent datasets.

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=5) and the different nature of the test datasets (Agilent CGH arrays, Affymetrix CGH arrays, RNA-Seq data)

iii. some of the genes composing our signatures were previously associated with cancer outcome in other tumor types (e.g. POU2F1) or to the sensitivity to anticancer agents (e.g. TIMP2).

These results should be replicated in a prospective trial to assess their potential value in the clinical setting.

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

Raponi M, Zhang Y, Yu J, Chen G et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006 Aug 1;66(15):7466-72. PMID: 16885343