

User-friendly and machine learning-empowered platform for classification of NSCLC based on RNA-seq profiling

Francois Collin¹[™], M. Niemira², A. Krętowski^{2,3}, J. Nikliński⁴, M. Kwaśniewski¹ Medical University of Bialystok, Poland: ¹Centre for Bioinformatics and Data Analysis, ²Center of Clinical Research, ³Department of Endocrinology, Diabetology and Internal Medicine, ⁴Department of Clinical Molecular Biology *Irancois.collin@umb.edu.pl*

Introduction

RNA-seq is a standard method for transcriptome profiling of cancer. Studies attempted to identify limited number of marker genes for diagnostic/prognostic of Non-Small Cell Lung Cancer NSCLC.

We propose to:

- use the complete transcriptome profile for high prediction accuracy.

Materials and Methods

Learning dataset:

123 patients, tumour/non-tumour samples, histo-pathological subtypes (MOBIT project, Niklinski 2017)

Machine learning:

Hissi -- *

- answer specific questions using transcriptome.

Results

Reactive modeling: the model adapts to the provided genes.

Best predictive results for whole transcriptome gene set. Good performance for lower number of genes.

Predictions used for:

Inpu

(gen

- diagnostic based on whole available gene set.
- hypothesis test (e.g. robustness of gene signature)
- hypothesis generation (e.g.identification of important genes)

Random-forest model (Breiman 2001)

Model testing: Sensitivity/specificity and ROC curves

Render: User friendly web platform (R + shiny + flexdashboard)





	[MoBIT: data valorisation / Pi	ototype v0.6] Presentation (1) Lung Cancer (2) SCC (3) Adenocarcinoma (supp) Prognostic markers (1/2) (supp) Prognostic ma	arkers (2/2) (supp) Refs
		About the diag.	Advise
t data es/count	Dev. Info Francois COLLIN, Ph.D 09.04.2019 - 17h37 Running on non-lung-tissue Spies produce inconsistent	The AI algorithm learned:	To increase the sensitivity of your diagnostic model, you may reduce the diagnosis threshold in the settings panel.
		 how to discriminate positive Lung Cancer tumour (LC+) and negative Lung Cancer (LC-, <i>i.e.</i> normal tissue versus tumoral sample). based on the expression profile of the available genes in the MoBIT data pool (annotation v37). growing 500 classification trees (settings panel). and performance was assessed comparing actual and estimated diagnosis. 	An increase sensitivity: • will avoid a negative • will in unit





A simple use of complex data:

+ Phenotyping based on transcriptome analysis.

+ Easy comparison/external validation.

Perspectives for:

+ personalised medicine, subtypes based on gene expression
+ prognostic (depends on data availability on later stages of follow-up)

Machine learning and artificial intelligence can **increase usability of complex genomic data**, necessary step toward wider implementation of personalised medicine.

References

Niklinski J et al. "Systematic biobanking, novel imaging techniques, and advanced molecular analysis for precise tumor diagnosis and therapy: The Polish MOBIT project." Advances in medical sciences 62.2 (2017): 405-413.

Breiman, L. "Random forests." Machine learning 45.1 (2001): 5-32.

 $\sim\sim$ No conflicts of interest to declare. $\sim\sim$



In the provide the second s