Prediction of COPD- and smoking status by network-based multi-’omics data fusion analysis

Chuan-Xing Li1, Shama Naz2, Bettina Levänen1, Maria Eldh1, Susanne Gabrielsson1, David J. Erle4, Craig E. Wheelock2, C. Magnus Sköld1, Åsa M. Wheelock1

1Respiratory Medicine Unit, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet/Karolinska University Hospital, Stockholm, Sweden.
2Department of Medical Biochemistry and Biophysics, Division of Physiological Chemistry 2, Karolinska Institutet, Sweden.
3Translational Immunology Unit, Department of Medicine, Karolinska Institutet/Karolinska University Hospital, Stockholm, Sweden.
4Lung Biology Center and Division of Pulmonary and Critical Care, Department of Medicine, University of California San Francisco, USA.

Background: COPD is an umbrella diagnosis defined by obstructive lung function impairments, and is likely to be caused by a multitude of etiologies including environmental exposures, genetic predispositions and developmental factors. Molecular prediction and mechanistic modeling of COPD will therefore be essential in order to develop relevant diagnostic and treatment options for this constantly growing patient group. The availability of large-scale multi-’omics datasets and the development of computational systems medicine approaches have provided the means to elucidate global alterations in the complex etiology of COPD.

Objective: To classify smoking- and COPD disease status by network-based multi-’omics data fusion analysis. We hypothesize that bridging and integration of multi-molecular level data will provide improved power for classification of smoking and COPD diagnosis.

Methods: ‘Omic data sets from the Karolinska COSMIC study were utilized, which includes 120 clinically well characterized subjects from 4 groups: healthy never-smokers, smokers with normal lung function, and current smoker as well as ex-smoker patients with early stage COPD (GOLD stage I-II/A-B). A network-based multi-’omics data fusion analysis was performed using unsupervised feature selection by variance network-based multi-’omics data fusion and clustering of subjects by Similarity Network Fusion analysis. Performance was evaluated by comparison between the predicted cluster and the known groups by Normalized mutual information (NMI) within all possible ‘omic combinations (n=127) and feature selection parameters (n=76).

Results: We found that the prediction power was increased with the number of (complementary) ‘omic data blocks used. The best predictor for all four groups (NMI=0.85) was achieved by integration of proteome, mRNA transcriptome data from BAL cells, combined with metabolomics and exosomal miRNA profiling from BAL fluid, including 276 significant COPD-influenced features (p<0.01). The corresponding average NMI for single, double and triple ‘omic integration are 0.35, 0.44 and 0.60 respectively. Smoking is a dominant factor across all the platforms, so it is easier to distinguish current smoking from non/ex-smoking subjects even with unsupervised feature selection: By integration of proteome data from BAL cells and miRNAs transcriptome from exosomes (110 features), the NMI of current smoking vs. other subjects is close to 1.0, whereas the NMI based on the individual ‘omics blocks was 0.82 and 0.08 respectively.

Conclusions: The results indicate that there is both comprehensive and redundant information within multi-’omics datasets. The predictive power will be highly increased by integration of comprehensive multi-’omics data. Even for dominant factors, such as smoking, the predictive power will be significantly improved by multi-’omics fusion than single ‘omics. Further expansion of these methods to allow sub-phenotyping of at-risk smokers and COPD patients may lead to improved techniques for early diagnosis of COPD.

Keywords: Multi-’omics data fusion; Systems Medicine; Smoking