



# Integration of multi-omics datasets enables molecular classification of COPD

Chuan-Xing Li<sup>1</sup>, Craig E. Wheelock<sup>2</sup>, C. Magnus Sköld<sup>1,3</sup> and Åsa M. Wheelock<sup>1</sup>

## Affiliations:

<sup>1</sup>Respiratory Medicine Unit, Dept of Medicine and Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden.

<sup>2</sup>Integrative Molecular Phenotyping Laboratory, Division of Physiological Chemistry II, Dept of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden.

<sup>3</sup>Lung-Allergy Clinic, Karolinska University Hospital, Stockholm, Sweden.

## Correspondence:

Åsa M. Wheelock, Pulmonomics Group, Karolinska University Hospital J7:30, SE-171 76 Stockholm, Sweden.  
E-mail: asa.wheelock@ki.se

 @ERSpublications  
**Multi-omics integration drastically improves unsupervised molecular prediction of COPD; 100% accuracy, subgroups n=6**  
<http://ow.ly/EeiY30iWB2f>

**Cite this article as:** Li C-X, Wheelock CE, Sköld CM, *et al.* Integration of multi-omics datasets enables molecular classification of COPD. *Eur Respir J* 2018; 51: 1701930 [<https://doi.org/10.1183/13993003.01930-2017>].

**ABSTRACT** Chronic obstructive pulmonary disease (COPD) is an umbrella diagnosis caused by a multitude of underlying mechanisms, and molecular sub-phenotyping is needed to develop molecular diagnostic/prognostic tools and efficacious treatments.

The objective of these studies was to investigate whether multi-omics integration improves the accuracy of molecular classification of COPD in small cohorts.

Nine omics data blocks (comprising mRNA, micro RNA, proteomes and metabolomes) collected from several anatomical locations from 52 female subjects were integrated by similarity network fusion (SNF). Multi-omics integration significantly improved the accuracy of group classification of COPD patients from healthy never-smokers and from smokers with normal spirometry, reducing required group sizes from n=30 to n=6 at 95% power. Seven different combinations of four to seven omics platforms achieved >95% accuracy.

For the first time, a quantitative relationship between multi-omics data integration and accuracy of data-driven classification power has been demonstrated across nine omics data blocks. Integrating five to seven omics data blocks enabled 100% correct classification of COPD diagnosis with groups as small as n=6 individuals, despite strong confounding effects of current smoking. These results can serve as guidelines for the design of future systems-based multi-omics investigations, with indications that integrating five to six data blocks from several molecular levels and anatomical locations suffices to facilitate unsupervised molecular classification in small cohorts.

---

This article has supplementary material available from [erj.ersjournals.com](http://erj.ersjournals.com)

The Karolinska COSMIC cohort is registered at [ClinicalTrials.gov](http://ClinicalTrials.gov) with identifier NCT02627872.

Received: Sept 22 2017 | Accepted after revision: March 08 2018

Copyright ©ERS 2018

## Introduction

Chronic obstructive pulmonary disease (COPD) is an umbrella diagnosis currently defined by spirometry and symptoms alone. However, COPD has a multitude of aetiologies, including environmental exposures, genetic predisposition and developmental factors. Genome-wide association studies indicate that polygenic variance can explain a portion of the variance of both forced expiratory volume in 1 s (FEV<sub>1</sub>) and FEV<sub>1</sub>/forced vital capacity (FVC) [1], with enrichment of both developmental and inflammatory pathways involved in the regulation of lung function [2]. However, a very large number of biological pathways appear to be controlling lung function development and decline, with genetic variants having a limited effect on phenotype [3]. It is clear that multiple modulating mediators in the downstream molecular cascade, originating from several different anatomical compartments, are involved in the chronic inflammation and structural changes characteristic of COPD.

Owing to the large number of COPD sub-phenotypes giving rise to similar clinical characteristics, molecular sub-phenotyping of COPD represents an essential first step in the identification and classification of these subgroups, before diagnostic/prognostic tools and treatment options can be established for the respective patient subgroup. Large-scale profiling, *i.e.* genomics, proteomics, lipidomics, metabolomics and breathomics, provide the means to elucidate global alterations in complex inflammatory diseases such as COPD. However, because dysregulation at different molecular levels and anatomical locations can dominate in different disease subgroups, multi-omics integration may be necessary to facilitate the diagnosis and understanding of disease mechanisms involved in the underlying COPD disease subgroups [4–8]. Several recent studies integrating two or three omics data blocks have indicated that integrating data from multiple molecular levels improves the identification of biomarkers [9, 10], sub-phenotype prediction [11, 12] and mechanistic understanding [13] of COPD. While these specific examples provide convincing evidence of the advantages of omics integration, no quantitative evaluation of the gain in statistical power beyond dual and triple omics integration has yet been published.

Herein we present a quantitative evaluation of the integration of nine multi-molecular-level omics data blocks (genotyping, mRNA, microRNA (miRNA), proteomes and metabolomes) collected from multiple anatomical locations (airway epithelium, lung resident immune cells, airway exudates, exosomes and serum) from the Karolinska COSMIC cohort. The primary objective was to evaluate if network-based integration of three to seven omics data blocks improves the statistical power and accuracy of group classification in small cohorts, as compared to single or dual omics investigations.

As a framework for the omics integrations, we used similarity network fusion (SNF) [14], which represents a new class of integration methods [15]. The majority of network integration methods are designed to cluster variables (*e.g.* genes, mRNA, proteins) to identify biological interactions or mechanisms between known groups. By contrast, SNF is a subjects-based method that uses the similarity of the overall molecular profile between study subjects to cluster them into previously unknown groups. SNF represents a unique network approach in that a similarity network is first created for every single-omics data block, then for each pair of omics data blocks, *etc.*, in an iterative fashion until all profiles are represented. Most importantly, this process is performed in an unsupervised manner, meaning that no information regarding disease status or other group classification is included. This is an essential aspect in the context of COPD classification, because the underlying assumption is that clinical characterisation as it stands does not provide sufficient resolution to facilitate COPD sub-phenotyping [16].

The current study presents a concerted workflow designed to maximise the acquired molecular information while simultaneously minimising the number of individuals required to achieve robust statistical power. This approach provides a viable strategy for performing systems medicine-based studies in small cohorts, representing an important advancement in the field that will facilitate the design and execution of investigations conducting molecular sub-phenotyping of respiratory diseases.

## Materials and methods

For detailed descriptions of the methods, see the supplementary material.

### Clinical cohort

This study used omics data blocks from the Karolinska COSMIC cohort (ClinicalTrials.gov ID: NCT02627872), a three-group cross-sectional study [17–25] with age-matched (45–65 years) and sex-matched groups of healthy never-smokers (“Healthy”), smokers with normal lung function (“Smokers”) and COPD patients (“COPD”; Global Initiative for Chronic Obstructive Lung Disease (GOLD) stage I–II/A–B; FEV<sub>1</sub> 51–97%; FEV<sub>1</sub>/FVC <70) (table E1). Peripheral blood, bronchoalveolar lavage (BAL) and bronchial epithelial cells (BEC) were collected as previously described [17, 19]. Participants had no history of allergy or asthma, did not use inhaled or oral corticosteroids and had had no exacerbations for >3 months prior to study inclusion. Current-smokers were matched in terms of

smoking history (>10 pack years) and current smoking habits (>10 cigarettes per day over the past 6 months). Current smoking status and abstinence for >8 h prior to BAL was verified through exhaled carbon monoxide [26]. The study was approved by the Stockholm Regional Ethical Board (Case No. 2006/959–31/1) and participants provided their informed written consent.

### Omics data blocks

Nine omics data blocks (figure 1) from 52 female subjects (20 Healthy, 20 Smokers, 12 COPD) were used: mRNA from BAL cells collected by microarray [22]; miRNA from BAL cells and from exosomes from BAL fluid (BALF) collected by microarray [22, 27]; difference gel electrophoresis (DIGE) proteomics from BAL cells [17]; shotgun proteomics data from BAL cells collected by isobaric tags for relative and absolute quantitation (iTRAQ) mass spectrometry (MS) [25, 28]; shotgun proteomics data from BEC collected by means of tandem mass tag (TMT)-MS [29]; eicosanoid profiling data from serum and BALF [21]; and metabolomics data from serum [30]. For details regarding data collection platforms and data preprocessing, see previous publications and supplementary material. The missing data matrix is provided in figure E1. The motivation for including only the female subjects (n=52) was based on maximal coverage across omics platforms for each subject.

### Similarity network fusion

Network-based multi-omics data fusion analysis and subject-based clustering were performed using the R-package SNFtool ([cran.r-project.org/web/packages/SNFtool](http://cran.r-project.org/web/packages/SNFtool)) [14]. In brief, a distance matrix was calculated for each subject using each single-omics dataset, followed by the construction of similarity graphs for each single-omics dataset. In essence, each omics data block is thereby reduced to an affinity matrix, in which the number of predictors depends on the number of study subjects in the analysis, not the number of variables (*i.e.* mRNA, proteins *etc.*). Therefore, the vastly different numbers of variables between omics platforms (*e.g.* mRNA data with 40 000 variables *versus* eicosanoid data with 100 variables) does not influence the SNF analysis in the same way it would in other types of analysis workflows. The affinity matrices from the various omics platforms are then fused into a single “fused similarity matrix” representing the similarity of each subject in relation to the other study subjects. The input predictor is thus a similarity matrix from each omics data block, and not the full matrix of the original variables, thereby making it possible to integrate disparate types of data with these differing numbers of variables. Fused subject similarity graphs were constructed based upon all combinations of the nine omics datasets, ranging from dual to septuple networks. Group belonging was then predicted using leave-one-out






	mRNA	miRNA	Proteins		Metabolites	
	Micro-array	Micro-array	DIGE	iTRAQ/TMT	Metabolome	Eicosanoids
 Serum					*	*
 BALF						*
 BAL cells	*	*	*	*		
 BEC				*		
 Exosome		*				

FIGURE 1 Nine omics data blocks collected from multiple molecular levels (mRNA, microRNA, proteomes and metabolomes) and multiple anatomical locations (airway epithelium, lung resident immune cells, airway exudates, exosomes and serum) from subjects from the Karolinska COSMIC cohort were used to explore how integration of multiple omics data blocks can improve the statistical power of group classification. The detailed methods used for sample collection as well as the analytical platforms used for data collection are described in the supplementary methods. The overlap of omics datasets for the 52 included subjects is shown in figure E1. DIGE: 2-D difference gel electrophoresis proteomics; iTRAQ: isobaric tags for relative and absolute quantitation proteomics; TMT: tandem mass tag proteomics; BALF: bronchoalveolar lavage fluid; BAL: bronchoalveolar lavage; BEC: bronchial epithelial cell; exosome: exosomes from BALF.

cross-validation (LOOCV) [31] with random sampling using label propagation (figure E2), or with spectral clustering [14].

The SNF parameters hyperparameter ( $\alpha$ ), number of iterations ( $t$ ) and the number of neighbours ( $K$ ) were set to  $\alpha=0.5$ ,  $t=30$  and  $K=5$ , and sampling times  $N$  in LOOCV was set to  $N=10000$  based on optimisation for robustness (figures E3–E4). In addition, the effect of subgroup sample size on multi-omics fusion was evaluated. Subject networks were visualised both as a fixed-position network, with clustering according to group belonging defined by clinical parameters (Healthy, Smoker, COPD), as well as with subjects clustered according to network similarity [32]. All networks were generated by Cytoscape 3.1.1 [33].

### **Strategies for handling missing data in network integration**

Three strategies for handling missing data blocks in SNF prediction were evaluated: 1) the conservative strategy included the 24 subjects with the most comprehensive coverage of omics data blocks across all subjects (figure E1); 2) the equal sample size strategy included all 52 subjects (figure E1), but with equal subgroup sizes ( $n=4$ ) in each iteration of training sets in LOOCV; and 3) the unequal sample size strategy included all 52 subjects, allowing for different group sizes ( $n=5-12$ ) in the training sets, thereby utilising the maximum information of each omics integration (figure E1). The three evaluated approaches for handling missing omics data blocks showed similar mean performance, with marginally higher mean performance for the unequal sample size strategy (figure E5). Given that this strategy is also the most liberal in terms of allowing inclusion of subjects with missing omics data blocks, results from the unequal sample size strategy are presented below. Results from the other two strategies are presented in the supplementary material.

### **Accuracy and power calculations**

The accuracy of group prediction/classification was calculated as the ratio of subjects correctly classified into the three study groups (Healthy, Smoker, COPD) by the respective SNF multi-omics workflow described above. Correct study groups were defined by COPD diagnosis (according to the GOLD initiative;  $FEV_1/FVC < 0.70$ ), as well as by smoking history and current smoking status (as confirmed by exhaled carbon monoxide measured at all four clinical visits [26]). The resulting normalised mutual information (NMI) represents the ratio of correctly classified subjects, where 0 equals all subjects misclassified, and 1 equals all subjects correctly classified. To assess the improvement in accuracy resulting from an increased number of predictors (here, an increased number of omics data blocks), the analyses were repeated following permutation of the subject labels. Power curves for the mean accuracy of each omics  $n$ -tuple (number of integrated omics datasets) were calculated based on equal allocation sample sizes (representing the utilised study design for the Karolinska COSMIC cohort). Required group sizes ( $n$ ) were calculated at the 80% and 95% statistical power level. For investigations of the accuracy of sub-classification of COPD patients, chronic bronchitis diagnosis was used as a ground-truth for calculating the NMI between clinical diagnosis and SNF-based prediction using spectral clustering. Chronic bronchitis diagnosis was determined as self-reported cough and sputum production for  $\geq 3$  months in each of at least two consecutive years [34].

## **Results**

### **Improvement in accuracy of group prediction by multi-omics integration**

We investigated whether multi-omics data fusion improves the statistical power and accuracy of unsupervised molecular classification of COPD in the presence of a strong confounder such as smoking. The mean accuracy of group prediction (Healthy, Smoker, COPD) increased in a linear fashion with the omics  $n$ -tuple, from a mean accuracy of 0.28 for the nine single-omics platforms to 0.90 for the septuple omics networks when using the label propagation approach (figure 2a, solid line). For the small cohort utilised here, group prediction using LOOCV appeared marginally more robust than group prediction using spectralClustering (figure E6A). A permutation test showed that the improvement in accuracy occurring by chance as a result of an increased number of predictors (*i.e.* number of omics data blocks) was negligible, increasing from 0.09 to 0.13 from single to septuple omics (figure E6B).

Power curves corresponding to the mean accuracy of each omics  $n$ -tuple indicate that septuple omics integration decreased the required subgroup size from  $n=30$  for single omics to  $n=6$  for septuplet omics at the 95% accuracy level (figure 2b, table 1). At the 0.80 accuracy level, the required subgroup size decreased from  $n=18$  to  $n=4$ . The mean accuracy achieved for each  $n$ -tuple omics integration using even smaller subgroup sizes ( $n=1-5$ ), relevant for personalised medicine or very rare subgroups of patients, are displayed in figure 2c.

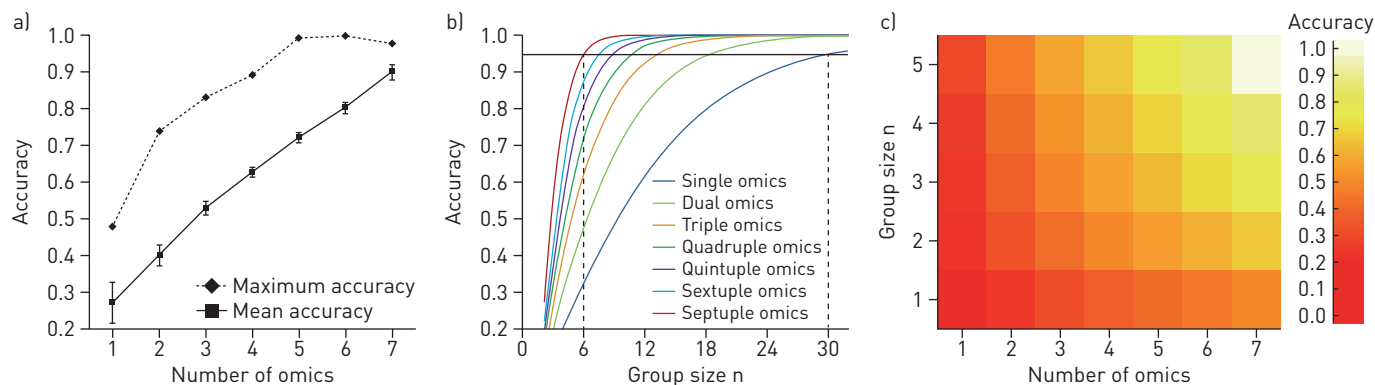


FIGURE 2 a) Accuracy of group prediction as a function of the number of omics datasets included in the SNF-mediated omics integration using nine omics datasets from the Karolinska COSMIC cohort (see figure 1). Values are displayed as mean accuracy±SE (solid line) as well as maximum accuracy (dashed line) for all possible omics combinations for each respective number of omics platforms, ranging from single to septuple omics integration. The presented data are based on the unequal sample size strategy (for other sampling strategies, see figure E5). b) Individual power curves corresponding to mean accuracy levels for each omics n-tuple. The graphs are showing group size (n) versus accuracy of group prediction for each respective omics n-tuple, ranging from single to septuple omics integration. Solid black horizontal line indicates 95% accuracy level, dashed vertical lines indicate n required at 95% accuracy level for single versus septuple omics integration. c) Heatmap displaying the mean accuracy levels achieved for subgroup sizes of n=1–5 for each n-tuple omics integration. Accuracy was calculated as the similarity network fusion-based accuracy compared to classification by chronic obstructive pulmonary disease diagnosis (according to the Global Initiative for Chronic Obstructive Lung Disease criteria) and current smoking status (defined by exhaled carbon monoxide monitoring).

**Peak performance networks**

Although the mean performance described above was highly correlated with the included number of omics platforms, reaching 90% accuracy at best, a number of specific network combinations reached better accuracy. Most notably, a sextuple omics combination consisting of BAL cell microRNA, BAL cell DIGE proteomics, BAL cell iTRAQ proteomics, serum metabolomics, BALF eicosanoids and serum eicosanoids resulted in 100% correct prediction of all subjects, both in terms of COPD diagnosis and smoking status (figure 3, figure E7). As a comparison, the best unsupervised single-omics prediction was achieved by the BEC TMT proteomics data, resulting in an accuracy of NMI=0.46. These results were achieved with the smallest samples group size of n=6. Comparison of the six individual single-omics similarity networks with the fused sextuple network demonstrated the improved power and reduced noise achieved by aggregating across multiple types of molecular data; sextuple integration distinguished molecular alterations due to smoking-related COPD in the presence of the confounding effects of current smoking status. None of the single-omics data blocks had the power to separate both current smoking status and COPD diagnosis. The trajectory of increased predictive information flow with all possible n-tuple omics fusions for the sextuple omics network providing 100% accuracy displayed in figure 3 are shown in figure 4.

The highest performing prediction network using the conservative sampling strategy was achieved through a septuple network, with 91% accuracy for group prediction (figure E8).

TABLE 1 Subgroup size required to reach 80% and 95% accuracy of group classification

	n-tuple <sup>#</sup>	n at 80% <sup>¶</sup>	n at 95% <sup>¶</sup>
Single omics	1	18	30
Dual omics	2	11	18
Triple omics	3	8	13
Quadruple omics	4	7	11
Quintuple omics	5	6	9
Sextuple omics	6	5	8
Septuple omics	7	4	6

<sup>#</sup>: the number of omics datasets included in the respective integration; e.g. quadruple analysis integrated four different omics datasets; <sup>¶</sup>: accuracy of group classification for all included subjects, across all three groups.

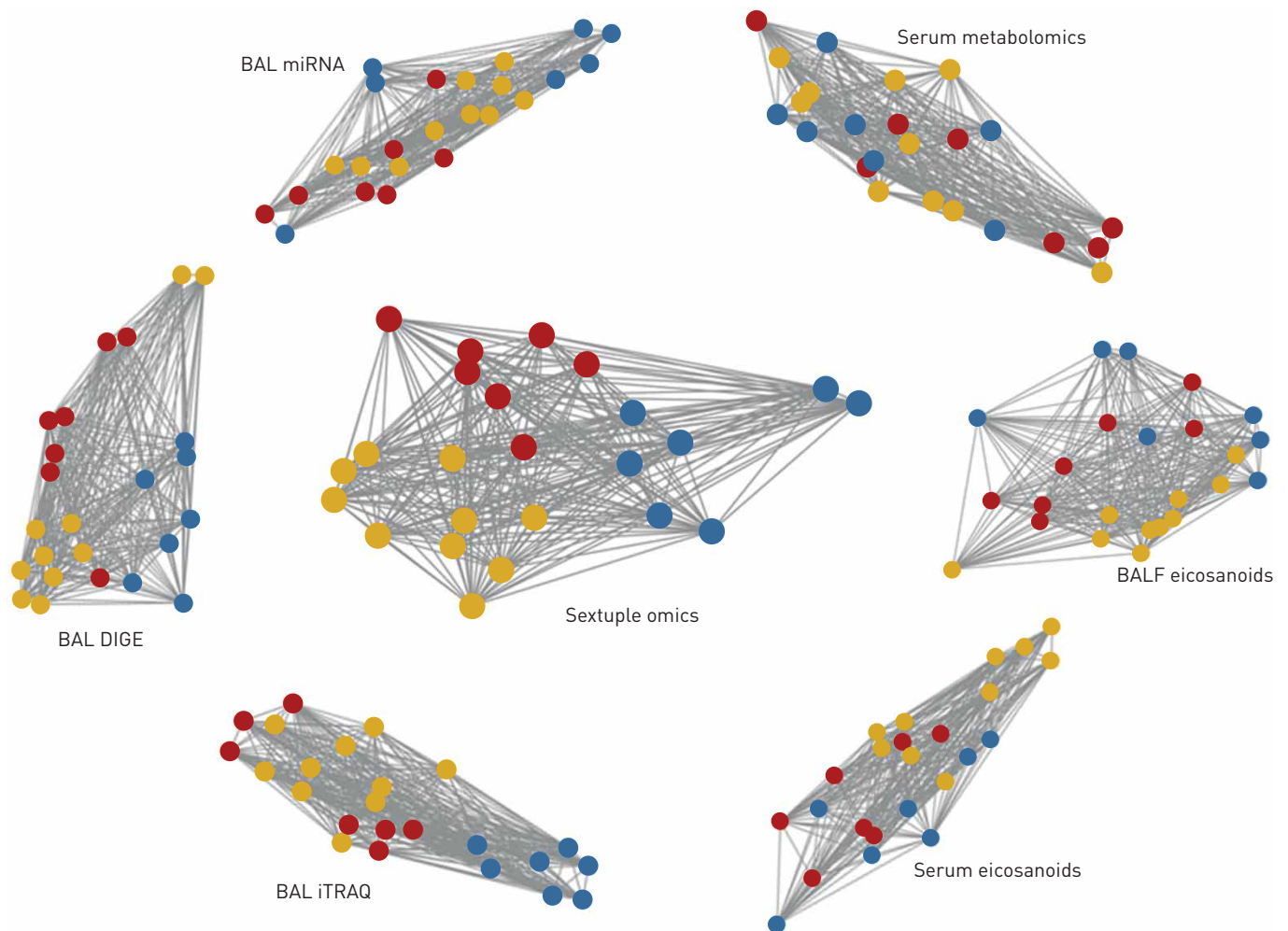


FIGURE 3 The best performing subject similarity network, consisting of a sextuple omics integration similarity network fusion similarity network (centre), provided 100% correct classification of the three subject groups of Healthy, Smokers and COPD. Similarity networks for each of the included single-omics data (periphery) are shown for reference. Nodes represent subjects. The networks are displayed with subjects clustered according to network similarity. The accuracy of 100% is based on 10000-times leave-one-out cross-validation permutation test using training data iteratively selecting six samples from each group. The same network displayed as a fixed-position network, with clustering according to the sextuple fused network preserved for all seven networks to facilitate visual comparison, is available in figure E7. Blue: healthy never-smoker; yellow: smoker with normal spirometry; red: smoker with COPD. BAL: bronchoalveolar lavage; DIGE: 2-D difference gel electrophoresis proteomics; iTRAQ: isobaric tags for relative and absolute quantitation proteomics; BALF: bronchoalveolar lavage fluid.

Out of the 303 possible single to septuple omics combinations, 25 different quadruple to septuple omics combinations reached a prediction accuracy  $>85\%$ , with seven network combinations providing an accuracy  $>95\%$  (figure 5), indicating some plasticity in the selection of omics platforms for optimal classification.

#### *Sub-phenotyping of COPD using multi-omics integration*

In an effort to investigate the statistical power of multi-omics integration for further sub-phenotyping of COPD patients, the accuracy of unsupervised molecular classification of chronic bronchitis diagnosis in the COPD group was investigated using eight of the omics datasets displayed in figure 1 and figure E1. One omics dataset (mRNA from BAL cells) was excluded because it did not fulfil the criterion of a minimum coverage of  $n=4$  subjects in each of the subgroups with/without chronic bronchitis. The mean accuracy of group prediction (COPD with *versus* without chronic bronchitis) increased in a linear fashion with the omics  $n$ -tuple, from a mean accuracy of  $<0.1$  for the eight single-omics platforms, to 0.75 for the septuple omics networks using spectral clustering (figure E9, solid line). However, 57 of the 254 possible dual to septuple omics combinations reached an accuracy of 100% (figure E9, dashed line).

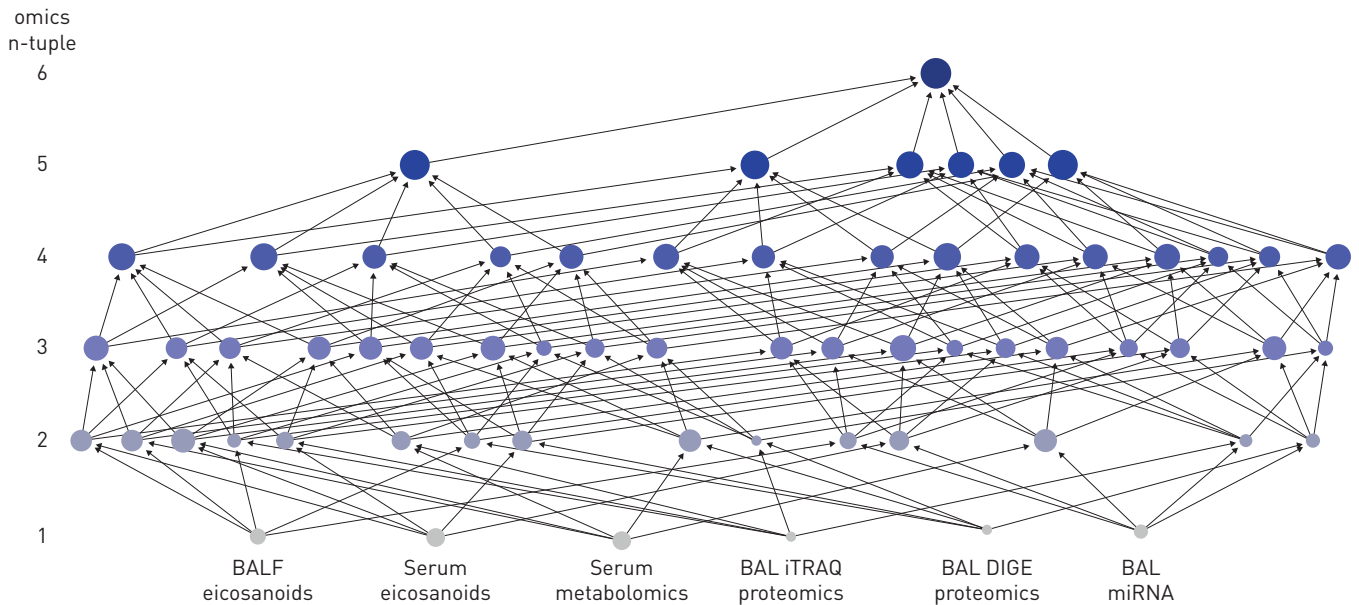


FIGURE 4 Example of accuracy increasing with omics n-tuple in similarity network fusion integration. The accuracy of prediction is indicated by the node size, ranging from the smallest (5% accuracy; bronchoalveolar lavage fluid (BAL) isobaric tags for relative and absolute quantitation proteomics (iTRAQ) single omics) to the largest representing 100% accuracy (sextuple omics). The n-tuple of omics datasets fused is shown from single (bottom) to sextuple omics (top). The n-tuple is also indicated by colour coding in grey to blue. BALF: bronchoalveolar lavage fluid; DIGE: 2-D difference gel electrophoresis proteomics.

### Discussion

The primary objective of the current study was to investigate whether integration of large-scale omics data from multiple molecular levels and anatomical locations increases the power and accuracy of molecular classification in complex disease, here exemplified by COPD. Our proof-of-concept investigations applying SNF network integration [14] to nine omics data blocks from 52 female subjects from the Karolinska

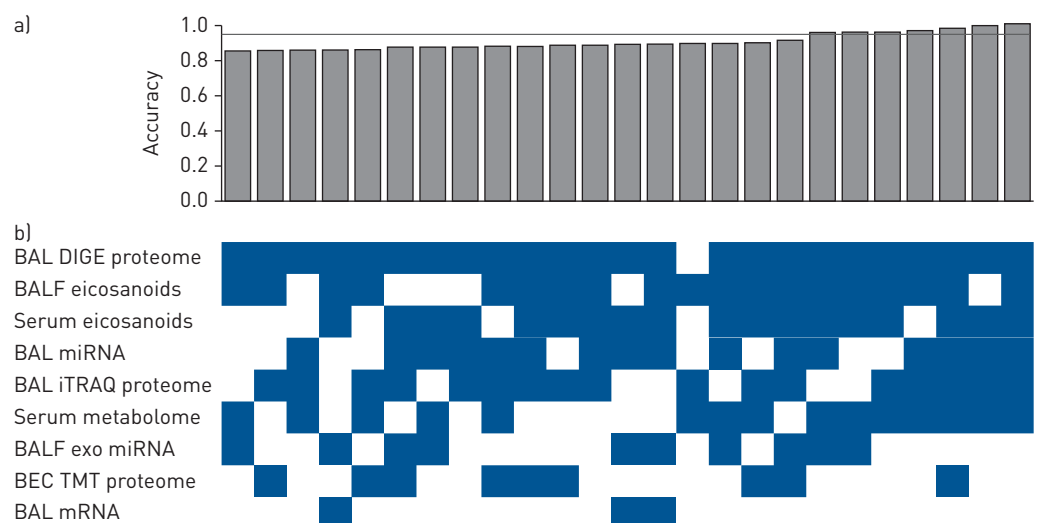


FIGURE 5 Thirty different similarity network fusion multi-omics network combinations reached an accuracy of group prediction >85%, calculated as the normalised mutual information compared to chronic obstructive pulmonary disease (COPD) diagnosis and smoking status (three groups: Healthy, Smoker, COPD). Seven of the network combinations reached an accuracy >95% [grey line]. a) Accuracy of prediction achieved for the respective combination; b) omics datasets included in the specific fused network corresponding to the respective bar graph shown above. Please note that each bar represents the accuracy of a single, specific network, hence the lack of error bars. BAL: bronchoalveolar lavage cells; DIGE: 2-D difference gel electrophoresis proteomics; BALF: bronchoalveolar lavage fluid; iTRAQ: isobaric tags for relative and absolute quantitation proteomics; exo: exosomes isolated from BALF; BEC: bronchial epithelial cells; TMT: tandem mass tag proteomics.

COSMIC cohort clearly show that integrating multi-omics data greatly improves the accuracy of unsupervised classification. The mean accuracy of prediction of the three groups of COPD, smokers with normal lung function and never-smokers increased in a near-linear fashion for the 303 evaluated network combinations, from a mean accuracy of 28% for the single-omics platforms to 90% for septuple omics integration. However, a large degree of variation was observed depending on the specific omics platforms included. Larger n-tuple omics did not automatically equate with better prediction for the specific omics combinations, and 100% accurate classification was achieved with sextuple omics integration (figure 3, figure E7). An accuracy >85% was reached by 25 different quadruple to septuple omics combinations, indicating that there is a large degree of plasticity in the combination of omics data required to optimise the accuracy of prediction (figure 5). Notably, the seven networks with >95% accuracy of prediction all contained omics datasets from several different molecular levels (miRNA, proteomes, metabolomes and eicosanoids) and anatomical locations (BAL cells, BAL fluid and serum), implying that combining data from different anatomical compartments and molecular levels is advantageous.

Smoking alone induces significant alterations of up to 50% of all biomolecules in the lung, as demonstrated in the BAL cell and BEC proteomes of the same cohort [28, 29]. As such, the true challenge in this cohort is to distinguish the subtle molecular effects associated with COPD pathology from the confounding effects of acute smoking in the group of current-smoker COPD patients. None of the single-omics data blocks had the power to separate out the molecular alterations due to mild-to-moderate COPD from the confounding effects of current smoking in an unsupervised fashion (figure 3, figure E7). In our previous investigations of the single-omics data blocks [17, 21, 30] as in most COPD investigations, stratification by current smoking status in combination with supervised analysis has therefore been mandatory to identify COPD-related alterations. In contrast, quintuple to septuple omics integration provided the power to classify COPD diagnosis from both never-smoker and current-smoking controls with 100% accuracy (figure 2a). It should be emphasised that this was achieved in an unsupervised, data-driven manner, with subgroups as small as n=6. The ability to sub-phenotype the COPD group based on chronic bronchitis diagnosis with 100% accuracy following integration of two or more omics platforms using subgroups as small as n=4 subjects (figure E9) further demonstrates the clinical utility of these methods. The integration methods employed here may thus enable systems medicine-based approaches to be performed in small, focused cohorts, which is desirable given the prohibitive costs of performing nonuple omics analyses on larger cohorts. Integration of the full arsenal of omics characterisations from multiple compartments could provide the power to detect rare molecular sub-phenotypes of disease from statistically relatively small cohorts, which is the general case for translational multi-omics studies.

The homogeneous COPD population selected for these proof-of-principle investigations, consisting of female patients with mild-to-moderate COPD who were free from co-morbidities or treatment, reflects a somewhat artificial scenario that poses limitations on the ability to extrapolate the findings to a more general disease population. COPD is a heterogeneous disease that can consist of 10–15 distinct molecular sub-phenotypes [35–37], few of which have been defined to date [16]. It is clear that the current clinical characterisation scheme does not provide the necessary resolution to classify or even identify these disease subgroups [16]. While the end goal of the unsupervised workflow presented here is to perform molecular sub-classification of all or at least several of the existing COPD sub-phenotypes, any study evaluating the performance of this method absolutely must focus on a COPD subgroup that can be expected to have molecular similarities to ours. The existence of a female-dominated molecular COPD phenotype in the Karolinska COSMIC cohort has been well established in our findings from supervised analyses of the single-omics data blocks [17, 21, 23, 30]. The female part of the Karolinska COSMIC cohort thus provides a set of molecularly distinct ground-truth study groups to evaluate the unsupervised classification.

Missing data is a common issue in studies of human subjects, particularly in multi-omics, multi-compartment studies such as the Karolinska COSMIC study. It is common for omics data blocks to be missing from a given subject for a number of reasons, *e.g.* omitted sampling of selected biospecimens because of safety considerations for the patient, or for individual omics experiments to be excluded owing to quality control criteria. The result is a data matrix with gaps, where every subject has some missing omics data blocks (figure E1). The original SNF data integration approach does not accommodate missing data blocks [14]. As such, developing approaches for dealing with missing data in the network construction was a secondary aim of the study. Out of three evaluated approaches, the most liberal of the three in terms of allowing inclusion of subjects with missing data (unequal sample size strategy; figure E5) performed the best. Accordingly, this strategy appears to provide a robust way to allow inclusion of all subjects and data collected in spite of the missing data block issue, which is inevitable in this type of translational study with invasive sampling at the site of inflammation.

Although the utilised cross-validation design assures that the accuracy is estimated independent of the training set, the relative homogeneity and limited sample size of this cohort poses some constraints. A



larger cohort may facilitate the true goal of multi-omics data integration: to identify previously unknown, molecularly distinct sub-phenotypes of health and disease. The results of these proof-of-principle investigations may provide some guidance in the design of future systems medicine studies, in which five to seven omics data blocks collected from complementary molecular levels and anatomical locations, with cohort sizes of six to ten subjects times the expected number of distinct molecular subgroups, may represent a good starting point for molecular sub-phenotyping of complex diseases. For example, the current diagnostic criteria for COPD represent a range of sub-phenotypes, driven by sex, smoking history, premature birth, environmental exposures *etc.* If we were to postulate that these aetiologies give rise to 15 molecularly distinct phenotypes, each represented by a different subset of biomarkers and pharmaceutical targets, then based on the subgroup sizes indicated from this cohort (table 1), an investigation using one omics dataset of choice would require 30 study subjects per COPD subgroup, *i.e.* 450 patients plus relevant control groups. The increased molecular resolution afforded by septuple omics integration reduces the required number of study subjects to  $n=6$  per subgroup to be identified. Thus, for the postulated study design aiming to identify 15 molecularly distinct subgroups of COPD, it would be sufficient to include 90 COPD patients. This dramatic reduction in the number of patients needed to achieve the study aims will greatly increase the clinical feasibility of molecular phenotyping studies. A significant bottleneck in clinical studies is often the ability to recruit a sufficient patient population in a timely fashion. The use of septuple omics integration can vastly reduce the time necessary for the cohort collection, facilitating the identification of unknown molecular sub-phenotypes.

In conclusion, these examples from the integration of nine omics data blocks from the Karolinska COSMIC cohort demonstrate an extraordinary increase in statistical power and accuracy of group classification, achieved by integrating data from multiple molecular levels and anatomical locations. For the first time, we have quantified the improvements in statistical power afforded by multi-omics integration, with the classification powers increasing on average from 28% to 90% with septuple omics integration. From the perspective of computational systems medicine, the mechanism of disease is not caused by independent subsets of genes/proteins/metabolites identifiable by traditional univariate statistics, but rather by their interactions, which makes it vital to develop a system-level understanding of disease. As demonstrated here, bridging and integrating data from multiple molecular levels and anatomical compartments relevant for disease pathology from the same individual could at last provide the statistical power of unsupervised classification. The unsupervised aspect is mandatory to facilitate identification of unknown molecular sub-phenotypes of complex diseases such as COPD and asthma. The unsupervised identification of molecularly distinct subgroups of disease represents a first, crucial step in elucidating treatable traits and biomarker subsets. The molecularly distinct subgroups identified by SNF can then be interrogated for handprints of diagnostic or prognostic biomarkers by supervised multivariate modelling approaches, *i.e.* orthogonal projections to latent structures, that provide a filter of variables of interest. In addition, we are currently developing multi-omics integration approaches at the pathway level. These leverage on the increased power achieved by integrating across several molecular levels in the downstream steps of identifying mechanistic features and treatable traits associated with identified patient subgroups. Combining the identified subsets of biomarkers, or “handprints of disease” [38], from multiple molecular levels may bring forth a long-awaited paradigm shift in precision- and personalised medicine.

Author contributions: Conception and design: C-X. Li, Å.M. Wheelock; analysis and interpretation: C-X. Li, Å.M. Wheelock; drafting of manuscript: C-X. Li, Å.M. Wheelock; principal investigator of systems medicine, proteomics and transcriptomics segments: Å.M. Wheelock; principal investigator of clinical segment, including clinical characterisation, bronchoscopies and sample collection: C.M. Sköld; principal investigator of metabolomics and eicosanoid segments: C.E. Wheelock.

Support statement: The Karolinska COSMIC study was funded by the Swedish Heart-Lung Foundation, Swedish Foundation for Strategic Research (SSF), VINNOVA (VINN-MER), EU FP6 Marie Curie, Karolinska Institutet, AFA Insurance, the King Oscar II Jubilee Foundation, the King Gustaf V and Queen Victoria’s Freemasons Foundation, the Swedish Research Council, the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Centre for Allergy Research, and the Karolinska Institutet and AstraZeneca Joint Research Program in Translational Science. C-X. Li is supported by an ERS/EU Marie Curie RESPIRE2 postdoctoral fellowship. C.E. Wheelock is supported by the Swedish Heart-Lung Foundation. Å.M. Wheelock is supported by a Swedish Heart-Lung Foundation senior researcher position. Funding information for this article has been deposited with the Crossref Funder Registry.

Conflict of interest: None declared.

## References

- 1 Soler Artigas M, Loth DW, Wain LV, *et al.* Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet* 2011; 43: 1082–1090.

- 2 Obeidat M, Hao K, Bosse Y, *et al.* Molecular mechanisms underlying variations in lung function: a systems genetics analysis. *Lancet Respir Med* 2015; 3: 782–795.
- 3 Martinez FD. Early-life origins of chronic obstructive pulmonary disease. *N Engl J Med* 2016; 375: 871–878.
- 4 Ghosh N, Dutta M, Singh B, *et al.* Transcriptomics, proteomics and metabolomics driven biomarker discovery in COPD: an update. *Expert Rev Mol Diagn* 2016; 16: 897–913.
- 5 Gomez-Cabrero D, Menche J, Cano I, *et al.* Systems medicine: from molecular features and models to the clinic in COPD. *J Transl Med*. 2014; 12: Suppl. 2, S4.
- 6 Davidsen PK, Turan N, Egginton S, *et al.* Multilevel functional genomics data integration as a tool for understanding physiology: a network biology perspective. *J Appl Physiol (1985)* 2016; 120: 297–309.
- 7 Chen H, Wang X. Significance of bioinformatics in research of chronic obstructive pulmonary disease. *J Clin Bioinforma* 2011; 1: 35.
- 8 Hobbs BD, Hersh CP. Integrative genomics of chronic obstructive pulmonary disease. *Biochem Biophys Res Commun* 2014; 452: 276–286.
- 9 Liu Z, Li W, Lv J, *et al.* Identification of potential COPD genes based on multi-omics data at the functional level. *Mol Biosyst* 2016; 12: 191–204.
- 10 Bowler RP, Bahr TM, Hughes G, *et al.* Integrative omics approach identifies interleukin-16 as a biomarker of emphysema. *OMICS* 2013; 17: 619–626.
- 11 Kim S, Herazo-Maya JD, Kang DD, *et al.* Integrative phenotyping framework (iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genomics* 2015; 16: 924.
- 12 Chang Y, Glass K, Liu YY, *et al.* COPD subtypes identified by network-based clustering of blood gene expression. *Genomics* 2016; 107: 51–58.
- 13 Azimzadeh Jamalkandi S, Mirzaie M, Jafari M, *et al.* Signaling network of lipids as a comprehensive scaffold for omics data integration in sputum of COPD patients. *Biochim Biophys Acta* 2015; 1851: 1383–1393.
- 14 Wang B, Mezlini AM, Demir F, *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014; 11: 333–337.
- 15 Bersanelli M, Mosca E, Remondini D, *et al.* Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016; 17: Suppl. 2, 15.
- 16 Castaldi PJ, Benet M, Petersen H, *et al.* Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax* 2017; 72: 998–1006.
- 17 Kohler M, Sandberg A, Kjellqvist S, *et al.* Gender differences in the bronchoalveolar lavage cell proteome of patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol* 2013; 131: 743–751.
- 18 Mikko M, Forsslund H, Cui L, *et al.* Increased intraepithelial (CD103+) CD8+ T cells in the airways of smokers with and without chronic obstructive pulmonary disease. *Immunobiology* 2013; 218: 225–231.
- 19 Forsslund H, Mikko M, Karimi R, *et al.* Distribution of T-cell subsets in BAL fluid of patients with mild to moderate COPD depends on current smoking status and not airway obstruction. *Chest* 2014; 145: 711–722.
- 20 Karimi R, Tornling G, Forsslund H, *et al.* Lung density on high resolution computer tomography (HRCT) reflects degree of inflammation in smokers. *Respir Res* 2014; 15: 23.
- 21 Balgoma D, Yang M, Sjodin M, *et al.* Linoleic acid-derived lipid mediators increase in a female-dominated subphenotype of COPD. *Eur Respir J* 2016; 47: 1645–1656.
- 22 Levänen B. Mechanisms of inflammatory signalling in chronic lung diseases: transcriptomics & metabolomics approaches. Doctoral Thesis. Solna, Karolinska Institutet, 2012.
- 23 Forsslund H, Yang M, Mikko M, *et al.* Gender differences in the T-cell profiles of the airways in COPD patients associated with clinical phenotypes. *Int J Chron Obstruct Pulmon Dis* 2017; 12: 35–48.
- 24 Karimi R, Tornling G, Forsslund H. Differences in regional air trapping in current smokers with normal spirometry. *Eur Respir J* 2017; 49: 1600345.
- 25 Yang M, Kohler M, Heyder T. Proteomic profiling of lung immune cells reveals dysregulation of phagocytotic pathways in female-dominated molecular COPD phenotype. *Respir Res* 2018; 19: 39.
- 26 Sandberg A, Skold CM, Grunewald J, *et al.* Assessing recent smoking status by measuring exhaled carbon monoxide levels. *PLoS One* 2011; 6: e28864.
- 27 Levanen B, Bhakta NR, Torregrosa Paredes P, *et al.* Altered microRNA profiles in bronchoalveolar lavage fluid exosomes in asthmatic patients. *J Allergy Clin Immunol* 2013; 131: 894–903.
- 28 Yang M, Kohler M, Heyder T, *et al.* Long-term smoking alters abundance of over half of the proteome in bronchoalveolar lavage cell in smokers with normal spirometry, with effects on molecular pathways associated with COPD. *Respir Res* 2018; 19: 40.
- 29 Heyder T. Between two lungs: proteomic and metabolomic approaches in inflammatory lung diseases. Doctoral thesis. Solna, Karolinska Institutet, 2017.
- 30 Naz S, Kolmert J, Yang M, *et al.* Metabolomics analysis identifies sex-associated metabolotypes of oxidative stress and the autotaxin-lysoPA axis in COPD. *Eur Respir J* 2017; 49: 1602322.
- 31 Zhu X, Ghahramani Z. 2002 Technical Report CMU-CALD-02-107. Carnegie Mellon University, 951: 1–8.
- 32 de Leeuw J. Convergence of the majorization method for multidimensional scaling. *J Classification* 1988; 5: 163–180.
- 33 Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; 13: 2498–2504.
- 34 Definition and classification of chronic bronchitis for clinical and epidemiological purposes. A report to the Medical Research Council by their Committee on the Aetiology of Chronic Bronchitis. *Lancet* 1965; 1: 775–779.
- 35 Agusti A, Bel E, Thomas M, *et al.* Treatable traits: toward precision medicine of chronic airway diseases. *Eur Respir J* 2016; 47: 410–419.
- 36 Vestbo J, Agusti A, Wouters EF, *et al.* Should we view chronic obstructive pulmonary disease differently after ECLIPSE? A clinical perspective from the study team. *Am J Respir Crit Care Med* 2014; 189: 1022–1030.
- 37 Lee JH, Cho MH, McDonald ML, *et al.* Phenotypic and genetic heterogeneity among subjects with mild airflow obstruction in COPD. *Respir Med* 2014; 108: 1469–1480.
- 38 Wheelock CE, Goss VM, Balgoma D, *et al.* Application of 'omics technologies to biomarker discovery in inflammatory lung diseases. *Eur Respir J* 2013; 42: 802–825.