



Early View

Original article

Integration of multi-omics datasets enables molecular classification of COPD

Chuan-Xing Li, Craig E. Wheelock, C. Magnus Sköld, Åsa M. Wheelock

Please 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; in press (<https://doi.org/10.1183/13993003.01930-2017>).

This manuscript has recently been accepted for publication in the *European Respiratory Journal*. It is published here in its accepted form prior to copyediting and typesetting by our production team. After these production processes are complete and the authors have approved the resulting proofs, the article will move to the latest issue of the ERJ online.

Copyright ©ERS 2018

Li and Wheelock

Integration of multi-omics datasets enables molecular classification of COPD

Chuan-Xing Li¹, Craig E. Wheelock², C. Magnus Sköld³ and Åsa M. Wheelock^{1,*}

¹Pulmonomics group, Respiratory Medicine Unit, Department of Medicine & Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden.

²Integrative Molecular Phenotyping laboratory, Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institutet

³ Lung Allergy Clinic, Karolinska University Hospital, Stockholm, Sweden

Author contribution:

Conception and design: CXL, ÅMW; Analysis and interpretation: CXL, ÅMW; Drafting of manuscript: CXL, ÅMW; PI of systems medicine, proteomics, and transcriptomics segments: ÅMW; PI of clinical segment, including clinical characterization, bronchoscopies, and sample collection: CMS; PI of metabolomics and eicosanoid segments: CEW.

Corresponding author:

Åsa M. Wheelock

Pulmonomics group, Lung Research Lab L4:01

Respiratory Medicine Unit, Department of Medicine

Karolinska Institutet

SE-171 76 Stockholm

Email: asa.wheelock@ki.se

Keywords: chronic obstructive pulmonary disease, systems medicine, multi-omics network integration, molecular sub-phenotyping

Abstract

Rationale: COPD represents an umbrella diagnosis caused by a multitude of underlying mechanisms, and molecular sub-phenotyping is needed to facilitate the development of molecular diagnostic/prognostic tools and efficacious treatments.

Objectives: To investigate whether multi-omics integration improves the accuracy of molecular classification of COPD in small cohorts.

Measurements and Results: Nine omics data-blocks (mRNA, miRNA, proteomes, 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 smokers with normal spirometry, reducing required group sizes from $n=30$ to $n=6$ at 95% power. Seven different combinations of 4-7 omics platforms achieved >95% accuracy.

Conclusions: For the first time, a quantitative relationship between multi-omics data integration and accuracy of data-driven classification power was demonstrated across 9 omics data-blocks. Integration of 5-7 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 integration of 5-6 data-blocks from several molecular levels and anatomical locations suffices to facilitate unsupervised molecular classification in small cohorts.

Take-home message (117 char, for social media)

Multi-omics integration drastically improves unsupervised molecular prediction of COPD;

100% accuracy, subgroups n=6

Introduction

Chronic obstructive pulmonary disease (COPD) is an umbrella diagnosis currently defined by spirometry and symptoms alone. COPD, however, is caused by a multitude of etiologies including environmental exposures, genetic predispositions and developmental factors. Genome-wide association studies indicate that polygenic variance can explain a portion of the variance of both FEV₁ and FEV₁/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 a limited effect of genetic variants 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.

Due to the large number of COPD sub-phenotypes giving rise to similar clinical characteristics, molecular sub-phenotyping of COPD represents an essential first step for the identification and classification of these sub-groups, before diagnostic/prognostic tools and treatment options can be established for the respective patient sub-group. Large scale profiling such as genomics, proteomics, lipidomics, metabolomics and breathomics provide the means to elucidate global alterations in complex inflammatory diseases such as COPD. However, since dysregulation at different molecular levels and anatomical locations may dominate in different disease sub-groups, multi-omics integration may be necessary to facilitate the diagnosis and understanding of disease mechanisms involved in the underlying COPD disease sub-groups (4-8). Several recent studies integrating two or three omics data blocks have indicated that integration from multiple molecular levels improves the identification of biomarkers (9, 10), sub-phenotype prediction (11, 12), or mechanistic

understanding (13) of COPD. While these specific examples provide convincing evidence for 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 integration of 9 multi-molecular level omics data blocks (genotyping, mRNA, microRNA, 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 3-7 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 utilized Similarity Network Fusion (14), representing 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. In 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 first is created for every single omics data block, then on 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 of disease status or other group belonging is included. This is an essential aspect in the context of COPD classification, as the underlying assumption is that clinical characterization as it stands does not provide sufficient resolution to facilitate COPD sub-phenotyping (16).

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

Materials and Methods

For detailed descriptions of Methods, see Online Supplement.

Clinical cohort

Omics data-blocks from the Karolinska COSMIC cohort (ClinicalTrials.gov ID: NCT02627872), a three group cross-sectional study (17-23) with age- (45-65 years) and sex matched groups of healthy never-smokers (“Healthy”), smokers with normal lung function (“Smokers”), and COPD patients (“COPD”; GOLD stage I-II/A-B; $FEV_1=51-97\%$; $FEV_1/FVC<70$) was utilized (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 no exacerbations for at least 3 months prior to study inclusion. Current-smokers were matched in terms of smoking history (>10 pack years) and current smoking habits (>10 cigarettes/day the past 6 months). Current smoking status and abstinence for >8 hrs. prior to BAL was verified through exhaled carbon monoxide (24). 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 utilized: mRNA from BAL cells collected by microarray (22); miRNA from BAL cells and from exosomes from BALF collected by microarray (22, 25); Difference Gel Electrophoresis (DIGE) proteomics from BAL cells (26); Shotgun proteomics data from BAL cells collected by isobaric tags for relative and absolute quantitation (iTRAQ) mass spectrometry (MS) (27); Shotgun proteomics data from BEC collected by means of tandem mass tag (TMT)-MS (28); Eicosanoid profiling data from serum and BALF (21); and metabolomics data from serum (29). For details regarding data collection platforms and data preprocessing, see previous publications and Online Supplement. 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 (SNF)

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) (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, where the number of predictors depends on the number of study subjects in the analysis, not the number of variables (i.e., mRNA, proteins etc.). The vastly different numbers of variables between omics platforms (40,000 for mRNA vs. 100 in eicosanoids) does therefore 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 therefore 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 vastly different numbers of variables (e.g. mRNA data with 40,000 variables vs. eicosanoid data with 100 variables). Fused subject similarity graphs were constructed based upon all combinations of the 9 omics data sets, ranging from dual to 7-tuple networks. Group belonging was then predicted using leave-one-out cross-validation (LOOCV) (30) with random sampling using label propagation (Figure E2), or with spectralClustering (14).

The SNF parameters hyperparameter (*alpha*), number of iterations (*t*) and the number of neighbors (*K*) were set to $K=5$, $\alpha=0.5$, and $t=30$, and sampling times *N* in LOOCV was set to $N=10,000$ based on optimization for robustness (Figures E3-E4). In addition, the effect of sub-group sample size on multi-omics fusion was evaluated. Subject networks were visualized 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 (31). All networks were generated by Cytoscape 3.1.1 (32).

Strategies for handling missing data in network integration

Three strategies for handling missing data blocks in SNF prediction were evaluated: 1)

Conservative strategy: Including the 24 subjects with the most comprehensive coverage of omics data blocks across all subjects (Figure E1); 2) *Equal sample size strategy*: Including all 52 subjects (Figure E1), but with equal sub-group sizes ($n=4$) in each iteration of training sets

in LOOCV; 3) *Unequal sample size strategy*: Including all 52 subjects, allowing for different group-sizes ($n=5-12$) in the training sets, thereby utilizing 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 also is 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 Online Supplement.

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 4 clinical visits (24)). The resulting Normalized Mutual Information (NMI) represents the ratio of correctly classified subjects, where 0 equals all subjects misclassified, and 1 equals all subjects correctly classified. In order to assess the improvement in accuracy occurring as a result of 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 utilized 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 calculation of 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 ≥ 3 months in each of at least two consecutive years (33).

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, increasing from a mean accuracy of 0.28 for the 9 single-omics platforms, to 0.90 for the 7-tuple omics networks when using the label propagation approach (Figure 2A, solid line). For the small cohort utilized here, group prediction using LOOCV appeared marginally more robust than group prediction using SpectralClustering (Figure E6 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 7-tuple omics (Figure E6 B).

Power curves corresponding to the mean accuracy of each omics n-tuple indicate that 7-tuple omics integration decreased the required sub-group size from $n=30$ for single omics to $n=6$ for 7-tuplet omics at the 95% accuracy level (Figure 2B, Table 1). At the 0.80 accuracy level, the required sub-group size decreased from $n=18$ to $n=4$. The mean accuracy achieved for each n -tuple omics integration using even smaller sub-group sizes ($n=1-5$), relevant for personalized medicine or very rare sub-groups of patients, are displayed in Figure 2C.

Peak performance networks

While the *mean* performance described above was highly correlated to the number of omics platforms included, reaching 90% accuracy at best, a number of specific network combinations reached better accuracy. Most notably, a 6-tuple 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 (Figures 3, 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 6-tuple network demonstrates the improved power and reduced noise achieved by aggregating across multiple types of molecular data; 6-tuple integration facilitated distinction of 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 (Figure 4).

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

Out of the 303 possible single- to 7-tuple omics combinations, 25 different 4-to-7-tuple omics combinations reached an accuracy of prediction >85%, with 7 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 8 of the omics data sets displayed in Figure 1 and Figure E1. One omics data set (mRNA from BAL cells) was excluded due to not fulfilling the criteria of a minimum coverage of $n=4$ subjects in each of the sub-group with/without chronic bronchitis. The *mean* accuracy of group prediction (COPD with vs. without chronic bronchitis) increased in a linear fashion with the omics n-tuple, increasing from a mean accuracy of <0.1 for the 8 single-omics platforms, to 0.75 for the 7-tuple omics networks using the Spectral Clustering (Figure E9, solid line). However, 57 of the 254 possible 2-to-7-tuple omics combinations reached an accuracy of 100% (Figure E9, dashed line).

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 will increase 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 9 omics data-blocks from 52 female subjects from the Karolinska COSMIC cohort clearly show that integration of multi-omics data greatly improves the accuracy of unsupervised classification. The *mean* accuracy of prediction of the groups of COPD, smokers with normal lung function, and never-smokers increased in a near-linear fashion for the 303 evaluated

network combinations, from 28% mean accuracy for the single omics platforms, to 90% mean accuracy for 7-tuple omics integration. However, a large degree of variation was observed depending on the specific omics platforms included. Larger n -tuple did not automatically equate better prediction for the specific omics-combinations, and 100% accurate classification was achieved with 6-tuple omics integration (Figure 3, Figure E7). Twenty-five different 4-to-7-tuple omics combinations reached an accuracy >85%, indicating that there is a large degree of plasticity in the combination of omics data required to optimize accuracy of prediction (Figure 5). Notably, the seven networks with >95% accuracy of prediction all contained omics data sets 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 (27, 28). 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 (Figures 3, E7). In our previous investigations of the single-omics data-blocks (21, 26, 34), as in most COPD investigations, stratification by current smoking status in combination with supervised analysis has therefore been mandatory in order to identify COPD-related alterations. In contrast, 5-to-7-tuple 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 emphasized 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 2 or more omics platforms using sub-groups as small as $n=4$ subjects (Figure E9) further demonstrate 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 9-tuple omics analyses on larger cohorts. Integration of the full arsenal of omics characterizations from multiple compartments may thus provide the power to detect rare molecular sub-phenotypes of disease also 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 mild-to-moderate COPD patients 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 may consist of up to 10-15 distinct molecular sub-phenotypes (35-37), few of which have been defined to date (16). It is clear that the current clinical characterization scheme does not provide the necessary resolution to classify or even identify these disease sub-groups (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, it is an absolute necessity to focus the study to a COPD sub-group that can be expected to have molecular similarities when evaluating the performance of the method. 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 (21, 23, 26, 34). 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. Missing omics data-blocks from a given subject due to reasons such as omitted sampling of selected biospecimens due to safety considerations for the patient, or individual omics experiments being excluded due to quality control criteria are a common scenario. 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 is not accommodating of 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 for inclusion of subjects with missing data (*unequal sample size strategy*; Figure E5), performed the best. Accordingly, this strategy appears to provide a robust way that allows for 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 utilized 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, where 5-7 omics data blocks collected from complementary molecular levels and anatomical locations, with cohort sizes of 6-10 subjects times the expected number of distinct molecular sub-groups may represent a

good starting point for molecular sub-phenotyping of complex diseases. The current diagnostic criteria for COPD represent a range of sub-phenotypes, driven by gender, smoking history, premature birth, environmental exposures etc. Postulate that these etiologies give rise to 15 molecularly distinct phenotypes, each represented by a different subset of biomarkers and pharmaceutical targets. Based on the sub-group sizes indicated from this cohort (Table 1), an investigation using one omics data set of choice would require 30 study subjects per COPD sub-group, i.e. 450 patients plus relevant control groups. The increased molecular resolution afforded by 7-tuple omics integration reduces the required number of study subjects to $n=6$ per sub-group to be identified. For the postulated study design aiming to identify 15 molecularly distinct sub-groups of COPD, it would be sufficient to include 90 COPD patients. This dramatic reduction in the number of necessary patients 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 7-tuple 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 integration of 9 omics data blocks from the Karolinska COSMIC cohort demonstrate an extraordinary increase in statistical power and accuracy of group classification achieved from integration of 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 7-tuple 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 may 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 sub-groups of disease represents a first, crucial step in elucidation of treatable traits and biomarker subsets. The molecularly distinct sub-groups identified by SNF can then be interrogated for handprints of diagnostic or prognostic biomarkers, proposedly by supervised multivariate modeling approaches such as orthogonal projections to latent structures (OPLS), that provide a filter of variables of interest. In addition, we are currently developing multi-omics integration approaches at the pathway level, to leverage on the increased power achieved through integration across several molecular levels also in the downstream steps of identifying mechanistic features treatable traits associated with identified patient sub-groups. Combining the identified subsets of biomarkers from multiple molecular levels, also referred to as handprints of disease (38), may bring forth a long-awaited paradigm shift in precision- and personalized medicine.

Acknowledgement

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

Institiutet and AstraZeneca Joint Research Program in Translational Science. Dr. Chuan-xing Li is supported by ERS/EU Marie Curie RESPIRE2 postdoctoral fellowship. Dr. Craig Wheelock is supported by the Swedish Heart-Lung Foundation. Dr. Åsa Wheelock is supported by a Swedish Heart-Lung Foundation senior researcher position.

References

1. Soler Artigas M, Loth DW, Wain LV, Gharib SA, Obeidat M, Tang W, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet.* 2011;43(11):1082-90.
2. Obeidat M, Hao K, Bosse Y, Nickle DC, Nie Y, Postma DS, et al. Molecular mechanisms underlying variations in lung function: a systems genetics analysis. *Lancet Respir Med.* 2015;3(10):782-95.
3. Martinez FD. Early-Life Origins of Chronic Obstructive Pulmonary Disease. *N Engl J Med.* 2016;375(9):871-8.
4. Ghosh N, Dutta M, Singh B, Banerjee R, Bhattacharyya P, Chaudhury K. Transcriptomics, proteomics and metabolomics driven biomarker discovery in COPD: an update. *Expert Rev Mol Diagn.* 2016.
5. Gomez-Cabrero D, Menche J, Cano I, Abugessaisa I, Huertas-Miguelanez M, Tenyi A, 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, Falciani F. Multilevel functional genomics data integration as a tool for understanding physiology: a network biology perspective. *J Appl Physiol (1985).* 2016;120(3):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, Hersch CP. Integrative genomics of chronic obstructive pulmonary disease. *Biochem Biophys Res Commun.* 2014;452(2):276-86.
9. Liu Z, Li W, Lv J, Xie R, Huang H, Li Y, et al. Identification of potential COPD genes based on multi-omics data at the functional level. *Mol Biosyst.* 2016;12(1):191-204.
10. Bowler RP, Bahr TM, Hughes G, Lutz S, Kim YI, Coldren CD, et al. Integrative omics approach identifies interleukin-16 as a biomarker of emphysema. *OMICS.* 2013;17(12):619-26.
11. Kim S, Herazo-Maya JD, Kang DD, Juan-Guardela BM, Tedrow J, Martinez FJ, 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, Silverman EK, Crapo JD, Tal-Singer R, et al. COPD subtypes identified by network-based clustering of blood gene expression. *Genomics.* 2016;107(2-3):51-8.
13. Azimzadeh Jamalkandi S, Mirzaie M, Jafari M, Mehrani H, Shariati P, Khodabandeh M. Signaling network of lipids as a comprehensive scaffold for omics data integration in sputum of COPD patients. *Biochim Biophys Acta.* 2015;1851(10):1383-93.
14. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods.* 2014;11(3):333-7.
15. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, 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, Rafaels N, Finigan J, Paoletti M, et al. Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax.* 2017;72(11):998-1006.
17. Kohler M, Sandberg A, Kjellqvist S, Thomas A, Karimi R, Nyrén S, et al. Gender differences in the bronchoalveolar lavage cell proteome of patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol.* 2013;131(3):743-51.

18. Mikko M, Forsslund H, Cui L, Grunewald J, Wheelock AM, Wahlstrom J, et al. Increased intraepithelial (CD103+) CD8+ T cells in the airways of smokers with and without chronic obstructive pulmonary disease. *Immunobiology*. 2013;218(2):225-31.
19. Forsslund H, Mikko M, Karimi R, Grunewald J, Wheelock AM, Wahlstrom J, 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(4):711-22.
20. Karimi R, Tornling G, Forsslund H, Mikko M, Wheelock A, Nyren S, et al. Lung density on high resolution computer tomography (HRCT) reflects degree of inflammation in smokers. *Respiratory research*. 2014;15:23.
21. Balgoma D, Yang M, Sjodin M, Snowden S, Karimi R, Levanen B, et al. Linoleic acid-derived lipid mediators increase in a female-dominated subphenotype of COPD. *Eur Respir J*. 2016;47(6):1645-56.
22. Levanen B. Mechanisms of inflammatory signalling in chronic lung diseases : transcriptomics & metabolomics approaches [Doctoral Thesis]. Karolinska Institutet: Karolinska Institutet; 2012.
23. Forsslund H, Yang M, Mikko M, Karimi R, Nyren S, Engvall B, 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. Sandberg A, Skold CM, Grunewald J, Eklund A, Wheelock AM. Assessing recent smoking status by measuring exhaled carbon monoxide levels. *PLoS One*. 2011;6(12):e28864.
25. Levanen B, Bhakta NR, Torregrosa Paredes P, Barbeau R, Hiltbrunner S, Pollack JL, et al. Altered microRNA profiles in bronchoalveolar lavage fluid exosomes in asthmatic patients. *The Journal of allergy and clinical immunology*. 2013;131(3):894-903.
26. Kohler M, Sandberg A, Kjellqvist S, Thomas A, Karimi R, Nyren S, et al. Gender differences in the bronchoalveolar lavage cell proteome of patients with chronic obstructive pulmonary disease. *The Journal of allergy and clinical immunology*. 2013;131(3):743-51 e9.
27. Yang M, Kohler M, Heyder T, Forsslund H, Garberg HK, Karimi R, 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. *Respiratory research*. 2018;In press.
28. Heyder T. Between two lungs: proteomic and metabolomic approaches in inflammatory lung diseases [Doctoral thesis]: Karolinska Institutet; 2017.
29. Naz S, Kolmert J, Yang M, Reinke SN, Kamleh MA, Snowden S, et al. Metabolomics analysis identifies gender-associated metabolotypes of oxidative stress and the autotaxin-lysoPA axis in COPD. *Eur Respir J*. 2017;In press.
30. Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation. 2002.
31. de Leeuw J. Convergence of the majorization method for multidimensional scaling. *Journal of Classification*. 1988;5(2):163-80.
32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-504.
33. 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(7389):775-9.

34. Naz S, Kolmert J, Yang M, Reinke SN, Kamleh MA, Snowden S, et al. Metabolomics analysis identifies sex-associated metabotypes of oxidative stress and the autotaxin-lysoPA axis in COPD. *Eur Respir J*. 2017;49(6).
35. Agusti A, Bel E, Thomas M, Vogelmeier C, Brusselle G, Holgate S, et al. Treatable traits: toward precision medicine of chronic airway diseases. *Eur Respir J*. 2016;47(2):410-9.
36. Vestbo J, Agusti A, Wouters EF, Bakke P, Calverley PM, Celli B, 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(9):1022-30.
37. Lee JH, Cho MH, McDonald ML, Hersh CP, Castaldi PJ, Crapo JD, et al. Phenotypic and genetic heterogeneity among subjects with mild airflow obstruction in COPD Gene. *Respir Med*. 2014;108(10):1469-80.
38. Wheelock CE, Goss VM, Balgoma D, Nicholas B, Brandsma J, Skipp PJ, et al. Application of 'omics technologies to biomarker discovery in inflammatory lung diseases. *Eur Respir J*. 2013;42(3):802-25.

Figure Legends

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 Supplemental Methods. BAL: bronchoalveolar lavage, BALF: BAL fluid, BEC: bronchial epithelial cell, exosomes: exosomes from BALF, DIGE: 2-D Difference Gel Electrophoresis proteomics; iTRAQ: Isobaric tags for relative and absolute quantitation proteomics; TMT: Tandem mass tag proteomics. The overlap of omics datasets for the 52 included subjects is shown in Figure E1.






Figure 2: Panel A displays the accuracy of group prediction for as a function of the number of omics datasets included in the SNF-mediated omics integration using 9 omics data sets from the Karolinska COSMIC cohort (see Figure 1). Values are displayed as mean accuracy \pm 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 7-tuple omics integration. The presented data is based on the unequal sample size strategy (for other sampling strategies, see Figure E5). Panel B displays the individual power curves corresponding to mean accuracy levels for each omics n-tuple. The graphs are showing group size (n) vs. accuracy of group prediction for each respective omics n-tuple, ranging from single to 7-tuple omics integration. Solid black horizontal line indicates 95% accuracy level, dashed vertical lines indicate n required at 95% accuracy level for single vs. 7-tuple omics integration. C) Heatmap displaying the mean accuracy levels achieved for sub-group sizes of n=1-5 for each n-tuple omics integration. Accuracy was calculated as the SNF-based accuracy compared to classification by COPD diagnosis (according to the GOLD criteria) and current smoking status (defined by exhaled carbon monoxide monitoring).

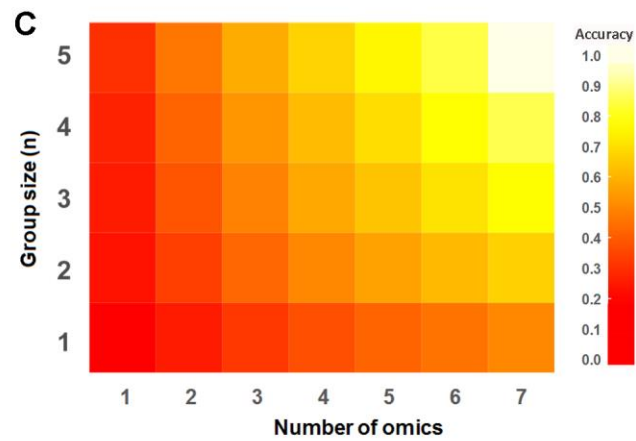
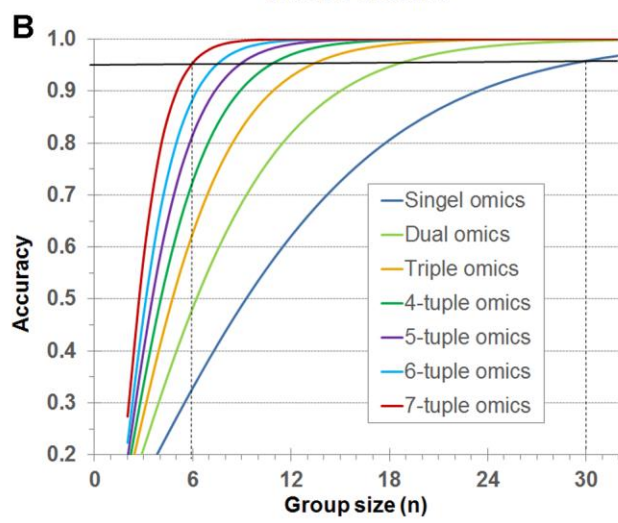
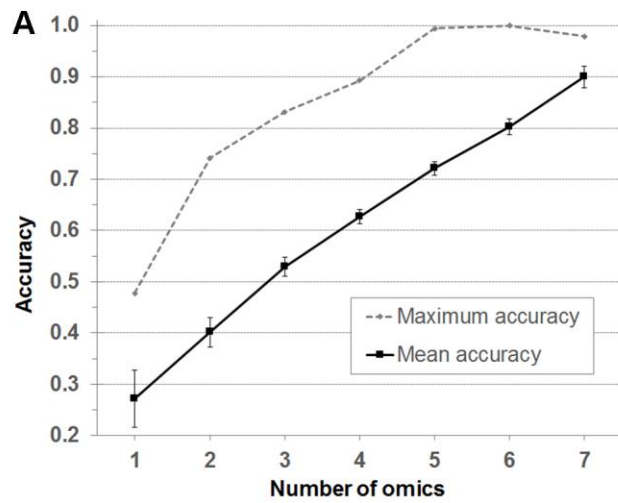
Figure 3: The best performing subject similarity network, consisting of a 6-tuple omics integration SNF similarity network (center), provided 100% correct classification of the three subject groups of Healthy, Smokers and COPD. Similarity networks for each of the included single-omics data (periphery) are shown for reference. Nodes represent subjects (red: COPD current smokers, yellow: Current smokers with normal lung function, blue: healthy never-smokers). The networks are displayed with subjects clustered according to network similarity. The accuracy of 100% is based on 10,000-times LOOCV permutation test using training data iteratively selecting 6 samples from each group. The same network displayed as fixed-position network, with clustering according to the 6-tuple fused network preserved for all seven networks to facilitate visual comparison, is available in Figure E7.

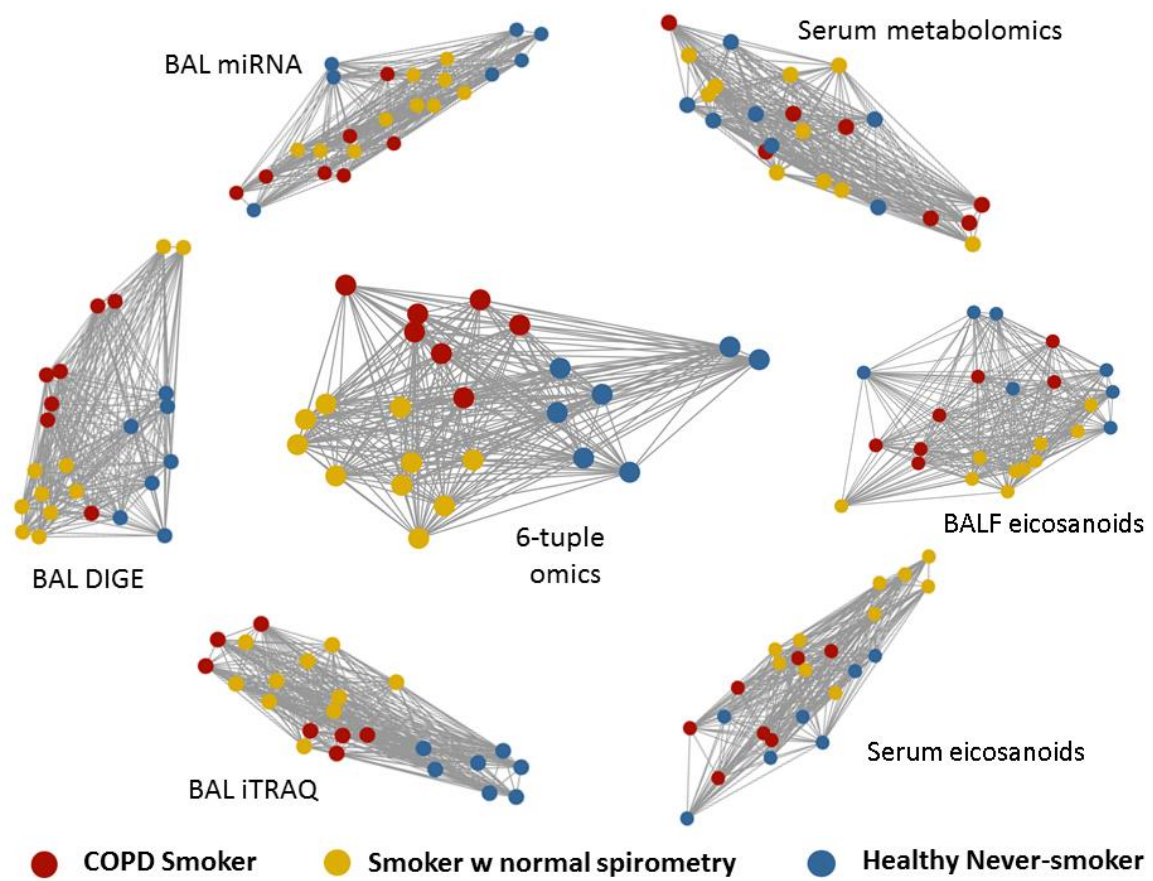
Figure 4. Example of accuracy increasing with omics n-tuple in SNF integration. The accuracy of prediction is indicated by the node size, ranging from the smallest (5% accuracy; BAL iTRAQ single omics) to the largest representing 100% accuracy (6-tuple omics). The n-tuple of omics datasets fused is shown from single (bottom) to 6-tuple omics (top). The n-tuple is also indicated by color-coding in grey-to-blue. BAL: bronchoalveolar lavage cells;

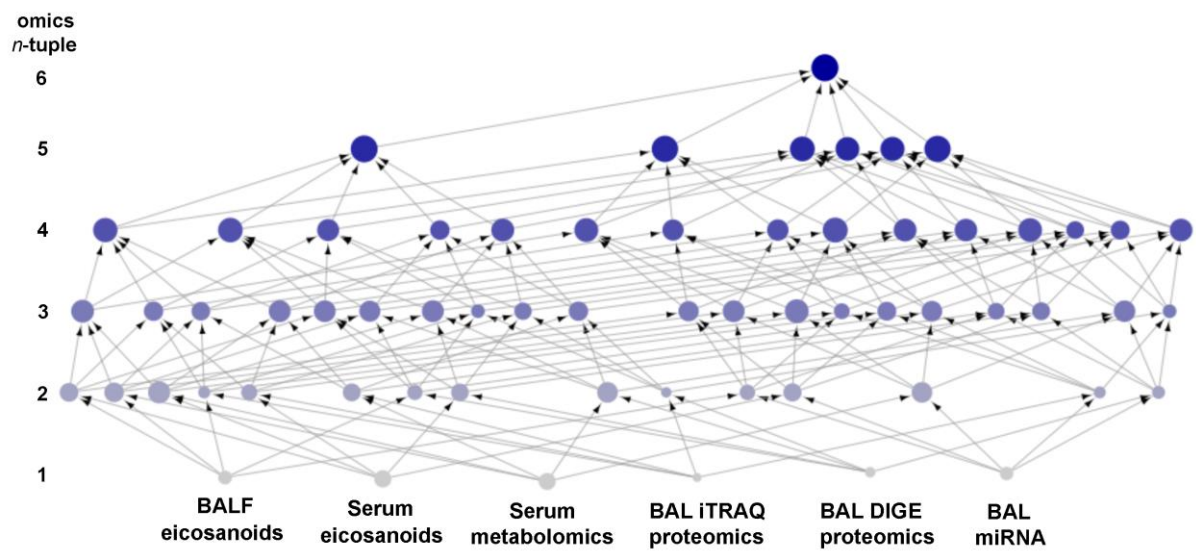
BALF: BAL fluid; DIGE: 2-D Difference Gel Electrophoresis proteomics; iTRAQ: Isobaric tags for relative and absolute quantitation proteomics.

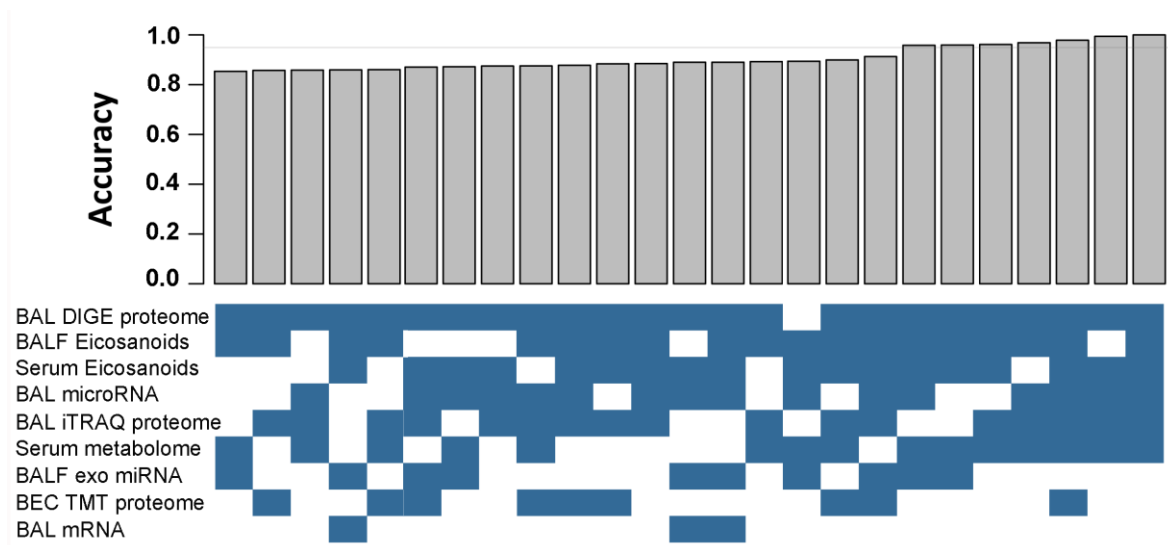
Figure 5: Thirty different SNF multi-omics network combinations reached an accuracy of group prediction $>85\%$, calculated as the normalized mutual information (NMI) compared to COPD diagnosis and smoking status (3 groups: Healthy, Smoker, COPD). Seven of the network combinations reached an accuracy of $>95\%$ (grey line). The upper panel shows the accuracy of prediction achieved for the respective combination, and the lower panel shows which omics data set were included in the specific fused network. Please note that each bar represents the accuracy of a single, specific network, hence the lack of error bars. BAL: bronchoalveolar lavage cells; BALF: BAL fluid; BEC: bronchial epithelial cells; Exo: exosomes isolated from BALF; DIGE: 2-D Difference Gel Electrophoresis proteomics; iTRAQ: Isobaric tags for relative and absolute quantitation proteomics; TMT: Tandem mass tag proteomics.

	mRNA	miRNA	Proteins		Metabolites	
	Micro-array	Micro-array	DIGE	iTRAQ /TMT	Meta- bolome	Eicosa- noids
 serum					*	*
 BALF						*
 BAL cells	*	*	*	*		
 BEC				*		
 Exosome		*				









Tables

Table 1: Sub-group size required to reach 80% and 95% accuracy of group classification

	<i>n</i> -tuple	<i>n</i> at 80%	<i>n</i> at 95%
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

%: Accuracy of group classification for all included subjects, across all three groups

n-tuple: the number of omics datasets included in the respective integration. E.g., 4-tuple analysis integrated 4 different omics datasets.

Online Data Supplement

Supplemental Methods

Clinical cohort

Nine omics data blocks collected from 52 subjects from the Karolinska COSMIC (Clinical & Systems Medicine Investigations of Smoking-related COPD) cohort (ClinicalTrials.gov ID: NCT02627872) were utilized. The COSMIC study is a three group cross-sectional study in which each group was stratified by gender with the aim of investigating the differentiation between the genders in early stage COPD and integrating several aspects of COPD and smoking through the use of imaging, transcriptomics, proteomics, metabolomics, and lymphocyte profiling in the context of clinical phenotypes (1-7). The COSMIC cohort consists of healthy never-smokers ("Healthy"), smokers with normal lung function ("Smokers"), and patients with COPD (GOLD stage I-II/A-B; $FEV_1=51-97\%$; $FEV_1/FVC<70$). For the purpose of this study, the female groups of Healthy (n=20), Smokers (n=20) and current-smoker COPD patients (n=12) were included. The three female study groups were selected based on minimal missing data blocks across the maximal number of omics platforms. Previous single-omics analyses have also shown a more homogeneous intra-group molecular profiles in the female population with regards to COPD diagnosis and current smoking status, which is a necessity to provide a ground-truth reference for evaluation of the SNF unsupervised classification performance. Groups were matched in terms of age (45-65 years) and gender, as well as smoking history and the number of cigarettes per day where relevant. Bronchoscopy was performed as previously described for the collection of bronchoalveolar lavage (BAL), and bronchial epithelial cell (BEC) through brushings (1, 3). Peripheral blood was also collected through venipuncture.

Study participants were recruited from individuals performing spirometry during "The World Spirometry Day," through advertisements in the daily press and via primary care centers. The majority of the individuals with COPD were smokers who were found to have an obstructive spirometry upon screening. Participants had no history of allergy or asthma, did not use inhaled or oral corticosteroids and had no exacerbations for at least 3 months prior to study inclusion. In vitro screenings for the presence of specific IgE antibodies (Phadiatop; Pharmacia Corp) were negative. Reversibility was tested after inhalation of two doses of 0.25 mg terbutaline (Bricanyl; Turbuhaler®; AstraZeneca). Medications (including oral contraceptives, estrogen replacement, and NSAIDs or other potential lipid mediator-modifying drugs) were recorded by means of a questionnaire. Lung function parameters were calculated as post-bronchodilator percent of predicted using the European Community of Coal and Steel (ECCS) normal values. COPD patients and smokers were matched in terms of smoking history (>10 pack years) and current smoking habits (>10 cigarettes/day the past 6 months). Self-reported current smoking status as well as abstinence for at least 8 hrs. prior to BAL was verified through exhaled carbon monoxide (8). The COPD group consisted of both current smokers and ex-smokers (≥ 2 years since smoking cessation). COPD ex-smokers were excluded for the purpose of SNF evaluation. Blood was drawn between 7-9 AM from fasting individuals by venipuncture and allowed to stand at

room temperature for 30 min before centrifugation at $1695\times g$ for 10 min at room temperature, and stored at -80°C until use. 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

Based on maximal overlap of omics data blocks across all subjects, 9 omics data blocks from 52 female subjects were utilized for the purpose of the performance evaluation of the SNF n-tuple omics integration. The 9 omics data blocks (Figure 1, Figure E1) consisted of mRNA from BAL cells collected by microarrays containing 41,000 probes corresponding to 19,596 genes as previously described (9); miRNA from BAL cells as well as from exosomes isolated from BALF collected by Agilent custom arrays as previously described (9, 10); Difference Gel Electrophoresis (DIGE) proteomics from BAL cells collected as previously described (11); Shotgun proteomics data from BAL cells collected by isobaric tags for relative and absolute quantitation (iTRAQ) mass spectrometry (MS) based proteomics (12, 13); Shotgun proteomics data from BEC collected by means of tandem mass tag (TMT)-MS as previously described (14); Oxylipin (eicosanoid) data from serum and BALF collected by LC-MS/MS as previously described (5); and non-targeted metabolomics data from serum collected as previously described (15). Each data collection platform is briefly described below:

RNA isolation

RNA from BAL cells, BEC cells, and the exosomal pellet from ultra-centrifugation of 100 ml of BAL fluid was extracted into two fractions containing small RNAs (including miRNAs) and large RNAs (containing mRNA) using the NucleoSpin® miRNA kit according to the manufacturer's instructions (Macherey-Nagel, Düren, Germany). RNA quality and quantity was assessed for concentration and purity by determining UV 260/280 and 230/260 absorbance ratios obtained by the Nanodrop ND-1000 spectrophotometer (Nanodrop, Wilmington, DE). RNA integrity and size distribution was examined by gel electrophoresis on RNA Pico LabChips (Agilent Technologies, Palo Alto, CA) processed on the Agilent 2100 Bioanalyzer. The content of miRs and mRNA in the exosomes was measured by bioanalyser.

mRNA Microarrays (1, 2)

RNA was amplified using the Low Input Quick Amplification Kit (Agilent Technologies) according to the manufacturer's protocol, and subsequent Cy3-CTP labeling was performed by using one-color labeling kits (Agilent Technologies). Clean-up of the labeled and amplified probe was performed (Zymo Research Corporation, Irvine, CA). The size distribution and quantity of the amplified product was assessed by Nanodrop. Equal amounts of Cy3-labeled target were hybridized to Agilent human whole-genome 4x44K Ink-jet arrays containing a total of 41,000 probes corresponding to 19,596 entrez genes. Hybridizations were performed at 65°C for 17 hours at a rotation of 10 rpm. Arrays were scanned by using the Agilent microarray G2565BA scanner (Agilent Technologies) with Scan region: Agilent HD (61x21.6) and a resolution of $5\mu\text{m}$, TIFF: 16 bit, XDR: 0.10. Raw signal intensities were extracted with Feature Extraction v10.1 software (Agilent Technologies). Flagged outliers were not included in any subsequent analyses. Microarray datasets were normalized using the *quantile* normalization method according to Bolstad et al (3). No background subtraction was performed, and the median feature pixel intensity was used as the raw signal before normalization. All procedures were carried out using functions in the R package *limma* in *Bioconductor* (4, 5).

miR Microarrays (1, 6)

The exosomal small RNA extracts were concentrated using a Speed-Vac, and the entire amount, except 1 μ l, was used for the amplification. The BAL and BEC samples were diluted to a working concentration prior to labeling. Small RNA was labeled with Cy3-CTP using the miRCURY LNA microRNA power labeling kit (Exiqon, Inc, Woburn, MA), according to manufacturer's protocol. Briefly, dephosphorylation of 5' end was performed in 37°C for 30 min followed by 95°C for 5 min to stop the enzyme reaction and denature the RNA. Dye labeling of 3' end with fluorochrome Cy3 was performed in a thermal cycler for 3 hrs in 16°C, 15 min 65°C and kept at 4°C until the next step. The reaction was stopped by blocking agent at 100°C, thereafter samples were snap-frozen before hybridization overnight (16 hrs) at 55°C with a rotation of 20 rpm. Labeled RNA was hybridized to one-color Agilent custom UCSF miRNA, v3.5 containing 894 miRs (BAL samples) or v4.0 containing 1223 miRs (exosomal samples), multi-species 8x15K Ink-jet arrays (Agilent Technologies). Arrays were washed in Agilent gene expression wash buffer 1 & 2 before scanning on the Agilent G2565BA laser scanner (Agilent Technologies) with Scan region: Agilent HD (61x21.6) and a resolution of 5 μ m, TIFF: 16 bit and an extended dynamic range (XDR) of 0.10. Raw signal intensities were extracted with Feature Extraction v10.1 software (Agilent Technologies). Flagged outliers were not included in any subsequent analyses. Microarray datasets were normalized using the *quantile* normalization method according to Bolstad et al (3). No background subtraction was performed, and the median feature pixel intensity was used as the raw signal before normalization. All procedures were carried out using functions in the R package *limma* in *Bioconductor* (4, 5).

DIGE proteome analysis of BAL cells (7) and BEC cells (8)

After lysis of cells in 8M urea, 2M thiourea, 4% Chaps, 33 mM Tris, aliquots of 50 μ g sample were labeled with minimal DIGE according to the supplier's recommendations (GE Healthcare). A triplex of 2 samples and 1 internal standard were co-separated by isoelectric focusing using 18 cm strips, pH 4-7, for 86 kVhrs and sodium dodecyl sulfate (SDS)-PAGE was performed on lab-casted 10% tris-glycine gels prior to image acquisition using a FLA Typhoon 9000 laser scanner. Image analysis and univariate statistics were performed using SameSpots version 4.0 (Nonlinear Dynamics, Newcastle, U.K.)

iTRAQ proteome analyses of BAL cells (9, 10)

Trypsinized protein extracts from 1.5×10^6 BAL cells were labeled with 4-plex iTRAQ reagents, with the 114 isobaric tag dedicated to a pooled reference sample used for ratiometric normalization to reduce the variance between batches(11), while the subject samples were randomized and labeled with the 115, 116 or 117 isobaric tags. Labeled peptides were fractionated into 5 mix-mode fractions, and analyzed on an LTQ-Orbitrap Velos Pro (Thermo Scientific, Sunnyvale, California, USA) connected to a Dionex Ultimate NCR-3000RS (LC system, Sunnyvale, California, USA). Full scan MS spectra were acquired with resolution $R=120,000$ at m/z 400. Peak integration of iTRAQ MS/MS spectra was performed by Proteome discoverer 2.1 (Thermo Fisher Scientific) searched against the UniProt human database (2015_12). Ratio data of samples to reference was log2 transformed.

TMT proteome analysis of BEC cells (8)

BECs were lysed and subjected to 10-plexed TMT® (Thermo Fisher Scientific), with TMT₁₂₆ dedicated to a reference pool. LC-Easy-nLCII interfaced to an Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific) were used for LC-MS/MS analysis. Database matching was performed using Mascot (Matrix Science) in Proteome Discoverer v1.4 (Thermo Fisher Scientific) using the Homo Sapiens Swissprot Database (04/2015). The ratios of TMT-reporter ion intensities for unique peptides and the reference pool were used for relative quantification.

Eicosanoid analysis of BAL fluid and serum

A liquid chromatography-mass spectrometry (LC-MS/MS) method was developed to quantify the reported lipid mediators. The complete method is described in the online supplement, with lipid mediator nomenclature provided in Table E1. Briefly, 3.3 mL of bronchoalveolar lavage fluid (BALF) were mixed with 10 µL of internal standards (Table E2, ref (2)) and loaded onto Waters Oasis HLB solid phase extraction (SPE) cartridges. SPE cartridges were air-dried, and lipid mediators eluted with organic solvent, evaporated under vacuum and reconstituted in 100 µL of methanol. Following spin filtering, 7.5 µL were injected onto an Acquity UPLC with a BEH C18 column (2.1x150 mm, 1.7 µm, Waters) and analyzed on a Waters Xevo TQ-MS in negative mode. The calibration levels and method parameters of all analyzed compounds are provided in Table E2 and Table E3, ref (2). Isoprostanes were screened via LC-MS/MS as previously reported (12).

Metabolomics analyses of serum (13)

Briefly, for non-targeted metabolomics, 50 µL of serum was used for both hydrophilic interaction liquid chromatography (HILIC) and reversed-phase chromatography. Samples were analyzed on an Ultimate 3000 UHPLC coupled to a Q-Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific, Bremen). Mass spectrometry data were acquired (full scan mode) in both positive and negative ionization. Molecular features were extracted using the software XCMS (<https://metlin.scripps.edu/xcms/index.php>). Putative metabolite annotation was performed using the Human Metabolome Database (HMDB) (14), and output matched to an in-house accurate mass/retention time library of reference standards (15). The chromatographic signal drift (if any) was normalized with a QC normalization algorithm in MATLAB vR2015a (Mathworks, Natick, MA, USA) (16). Only metabolites that were present in ≥70% of the samples in any group and had a coefficient of variance <30% in the QC samples were included in the SNF analyses.

Data processing

Proteomic data from BAL and BEC were log₂ transformed and normalized to a pooled internal reference sample. Features detected in <75% of the subjects in *each* sub-group were excluded. MicroRNA and mRNA profiles from BAL, BEC and exosomes were log₂ transformed and quantile normalized. MicroRNA and mRNA below the lowest limit of quantification (LLOQ; defined as 5 x SD of the noise above the background fluorescence) (16) (RFU<2^{5.5}) were excluded. Missing values in the non-targeted metabolomics platform, deemed to be associated with technical limitations rather than the detection limit, were imputed by KNN (K-nearest neighborhood) method with K = 10 by Euclidean distance. Oxylipin analytes present at levels below the limit of detection (LOD; defined as 3 x SD of the noise above) were set to 25% of

LOD (5). Data blocks were mean-centered and scaled to unit variance across features prior to SNF.

Similarity Network Fusion (SNF) construction and group prediction

Network-based multi-omics data fusion analysis was performed by Similarity Network Fusion analysis, followed by clustering of subjects (17). The analysis includes four major steps: 1) The subjects' distance matrices based on each single-omics data was calculated using Euclidean Distance; 2) Subject similarity graphs were constructed for each single-omics based on their distance matrices; 3) Subjects' similarity graphs from different omics platforms were iteratively fused to one similarity network representing all the omics data blocks included in the specific evaluation at hand; 4) Based on the resulting fused similarity graph, prediction of each subject's group label was performed using the label propagation method proposed in the SNFtool. Leave-one-out cross-validation (LOOCV, $N=10,000$), using random sampling with replacement was performed, with the added constraint that a minimum of $n=5$ subjects with full overlap of all omics platforms included in the specific network had to be available (see Illustrations in Figure E2). The accuracy of each fused similarity network was evaluated by comparison between the predicted label and the known group label (by clinical diagnosis) by Normalized mutual information (NMI), with value range from 0 to 1, in which 1 means 100% correct prediction of test subjects' group belonging, as defined by COPD diagnosis and current smoking status. The three parameters used in the SNF algorithm, the number of neighbors (K) and hyperparameter (α) when construct similarity graph from distance matrix, number of iterations (t) and also the number of neighbors (K) when fused similarity graphs, were optimized (see Results). In addition, the sampling times N in LOOCV was optimized for robustness. To decrease the impact of unbalance sample sizes among different groups, we use the equal number of sample size for all groups in training set to construct the fused similarity network. We select parameters as $K = 5$, $\alpha = 0.5$, $t = 30$, and $N = 10,000$ based on the robustness analysis (Figure E3-E4). The rational for choosing the label propagation and LOOCV methods as the primary prediction approach was based on the risk of overfitting associated with the small n . However, comparison evaluations using the spectralClustering approach indicated similar results (Figure E6).

Evaluation of strategies for handling missing omics data blocks in SNF analysis

Out of the 52 subjects included in the SNF evaluations, some of the 9 omics data blocks were missing from certain subjects, resulting in sample size variation from 27 to 52 across the 9 omics platforms. In order to evaluate the influence on varying overall sample sizes as well as sub-group sample sizes in the SNF construction, we compare three strategies for handling missing data blocks in SNF prediction: 1) A *conservative strategy* including only the 24 subjects with the most complete set of omics data blocks; 2) An *equal sample size strategy*, where all 52 subjects were included, but to avoid influence on sub-group sizes on the performance of different n -tuple omics combinations, equal sub-group sizes ($n=4$) were used as training data in the LOOCV training; 3) An *unequal sample size strategy* allowing for sampling of the different sub-group sizes (i.e. maximal number of subjects with full coverage of the particular omics combination minus one: with the overall n ranging from 18-52, and the smallest sub-group size ranging from $n=5$ -12) as training data in the LOOCV prediction approach to utilize the maximum information of each omics integration. For all these three strategies, the same SNF analysis procedure is applied with equal group sample size for all three groups within each omics integration and the same parameters (Parameter values $K = 5$, $\alpha = 0.5$, and $t = 30$, as well as $N = 10,000$ with the

optimization results are described in the Online Supplement Result section as well as in Figures E3-E4).

Estimation of the performance of multi-omics fusion in small sample size data

A subset of the multi-omics data with 24 subjects with the majority of the data blocks available for every subject was used to estimate the performance of multi-omics fusion in different sample size data. We set different sample size in training data to construct different similarity networks, and then predict the test sample based on SNF with LOOCV random sampling without replacement. To decrease the impact of unbalance sample sizes among different groups, we used the same sample size across all groups for this evaluation. Results were plotted as mean accuracy (NMI) \pm SE for each omics n-tuple. Theoretical power curves were generated based on equal allocation sample sizes on the calculated mean accuracy for each n-tuple in order to allow estimation of the n required to reach relevant accuracies for the various omics n-tuples.

Subject network visualization

All subject-based network visualizations were made with nodes representing subjects and node color reflecting known diagnostic groups, as defined by GOLD COPD diagnosis and current smoking status. The positioning of the subjects in the network visualizations are made in two different manners: 1) Similarity networks, with subjects clustered according to *network similarity*, thereby facilitating visual inspection of the clustering performance of the network, with edge-weighted spring embedded layout (18). All edges are displayed with the same width, and proximity of subjects (length of edge) represent similarity. This applies to Figure 3 and Figure E7 (panels B). 2) Fixed-position network, with clustering according to subjects' known group belonging (Healthy, Smoker, COPD) facilitate visual comparison. Edge thickness reflects the strength of the similarity between each pair of subjects, with ranked similarities <75% displayed as a thin line, and ranked similarities 75-100% proportional to edge thickness. This applies to Online Supplement Figure E7 and E8, panel A. All subjects networks are generated by Cytoscape 3.1.1 (19).

Supplemental Results

Robustness of SNF parameters

Robustness analysis for the three main parameters used in the SNF algorithm (the number of neighbors (K), hyperparameter (α), and number of iterations (t)) was performed for the ranges recommended by Wang et al. (17) for our multi-omics data set, to assure that α and t are within their optimal ranges. Different levels and combinations of the three parameters were compared by the accuracy distributions based on 303 single- to 7-tuple omics similarity networks. Twenty-four subjects from the three groups of female current-smoker COPD patients (COPD, $n=6$), smokers with normal lung function (Smokers, $n=10$) and healthy never-smoker controls (Healthy, $n=8$) were selected for the parameter evaluation, based on the availability of the 9 omics data sets across most subjects, as well as the relative homogeneity of intra-group molecular profiles of the individual omics data sets, as evident from results from the individual omics data sets (5, 11). The 9 omics data blocks (Figure E1) were used to construct fused similarity networks and predict test groups by LOOCV with random sampling without replacement. The K parameter was evaluated from 2 to 5, as the least group size is 6 (Figure E3

shows K from 3 to 5). The α parameter was evaluated from 0.3 to 0.8, by an increment of 0.1 (Figure E3 shows α of 0.3, 0.5 and 0.7). The t parameter was evaluated for $t=20$ and $t=30$ (Figure E3). Our results agreed with those of Wang et al. (17), with a high level of robustness for all three parameters. We selected $K = 5$, $\alpha = 0.5$, and $t = 30$ in all further analyses. In addition, the robustness of the number of random sampling with replacement in the LOOCV test to construct and evaluate different SNF predictor was evaluated. We tested the $N = [200, 400, 800, 1000, 5000, 10000, 15000, 20000]$ (Figure E4), and compared the accuracy difference between adjacent N pairs in the same 24 sample dataset (Figure E4). At $N = 10000$, the accuracy is very robust with the mean of squared differential accuracy as 5.0×10^{-4} and standard deviation as 7.6×10^{-4} . We select $N = 10,000$ in all further analyses.

Evaluation of strategies for handling missing data blocks in SNF analysis

The SNF data integration approach requires that data is available across *all* omics platforms from *all* included subjects in a particular network. As such, developing approaches for dealing with missing data in the network construction is essential. For the purpose to deal with missing data, we evaluated three approaches: 1) A *conservative strategy* including only the 24 subjects with the most complete omics data across all 9 platforms (Figure E1, red box), using a fixed sub-group size of $n=4$ as training sets for the iterative LOOCV prediction; 2) An *equal sample size strategy*, where all 52 subjects were included (Figure E1, grey box), but to avoid influence on sub-group sizes on the performance of different n -tuple omics combinations, equal sub-group sizes ($n=4$) were used for training purposes; 3) An *unequal sample size strategy* where all 52 subjects were included (Figure E1, grey box), but allowing for sampling of different sub-group sizes in training sets (range: $n=5-12$) in order to maximize the information utilized in the training. The mean performance of the three methods was very robust, indicating that the *unequal sample size strategy* is the optimal strategy for addressing the missing omics data block issue, while at the same time making use of all collected data.

Supplementary Tables

Table E1. Clinical parameters of female individuals from the Karolinska COSMIC cohort included in the current study

Parameters	Healthy never-smokers	Smokers	COPD
Group size (n)	20	20	12
Age	55.5 (49.5, 62.0)	54.0 (48.0, 58.0)	59.0 (57.0, 63.0)
BMI	25.4 (23.3, 28.2)	24.2 (22.7, 25.8)	23.5 (21.7, 25.9)
Smoking [pack-years]	0.0 (0.0, 0.0)	33.0 (27.8, 40.0)	40.5 (37.3, 45.8)
FEV ₁ [% predicted]	120 (113, 127)	110 (100.1, 116)	78.5 (74.8, 90.5)
FEV ₁ /FVC	0.83 (0.77, 0.84)	0.79 (0.75, 0.82)	0.62 (0.57, 0.63)
GOLD Stage (1/2)	N.A.	N.A.	6/6
GOLD-2011 (A/B)	N.A.	N.A.	9/3
Blood leucocytes [$\times 10^9$ /L]	5.6 (4.5, 6.5)	6.8 (6.4, 8.0)	8.2 (6.1, 9.4)
Blood platelets [$\times 10^9$ /L]	267 (245, 304)	288 (245, 343)	281 (238, 327)
Serum albumin [g/L]	40.0 (38.0, 41.0)	39.0 (37.8, 39.3)	39.5 (38.0, 41.0)
Antitrypsin [g/L]	1.4 (1.3, 1.5)	1.6 (1.4, 1.7)	1.6 (1.5, 1.7)
Menopause (yes/no)	14/6	12/8	12/0
Emphysema (yes/no)	0/20	12/8	11/1
Chronic bronchitis (yes/no)	0/20	3/17	5/7

Definition of abbreviations: BMI = body mass index, COPD =current-smokers with chronic obstructive pulmonary disease, FEV₁ [%]= post-bronchodilator forced expiratory volume in one second as % of predicted based on ECCS reference values, FVC = post-bronchodilator forced vital capacity, GOLD = Global Initiative for Obstructive Lung Disease, N.A. = not applicable, Smoker = current-smokers with normal spirometry. Values are presented as median and IQR

Supplementary Figures

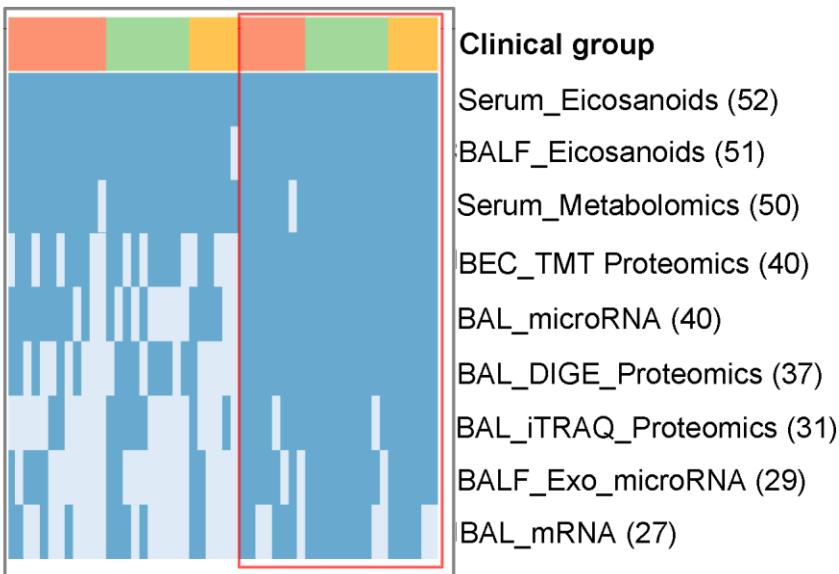


Figure E1: Overview of omics data blocks-subject matrix.

Overview of overlap of omics data collected from the 52 female subjects in the Karolinska COSMIC cohort that were included in the SNF performance evaluations. Each row represents a platform, and each column is a subject. Row one indicated subject groups (red=Never-smoker healthy, green=smoker with normal lung function, yellow=current-smoker COPD). Dark blue cells indicate available data blocks, and light blue indicates missing data blocks. The number in brackets following the data block name indicates total number of subjects that the respective data is available for. Anatomical locations: Serum; BAL: bronchoalveolar lavage cells; BALF: BAL fluid; BEC: bronchial epithelial cell; Exo: exosomes isolated from BAL fluid. Data types: DIGE: 2-D Difference Gel Electrophoresis proteomics; iTRAQ: Isobaric tags for relative and absolute quantitation proteomics; TMT: Tandem mass tag proteomics; mRNA: mRNA microarray; miRNA; miRNA microarray. Red box: Subjects with maximal omics block overlap, included in the *conservative sampling strategy*. Grey box: Subjects included in *equal-* and *unequal sampling strategies*.

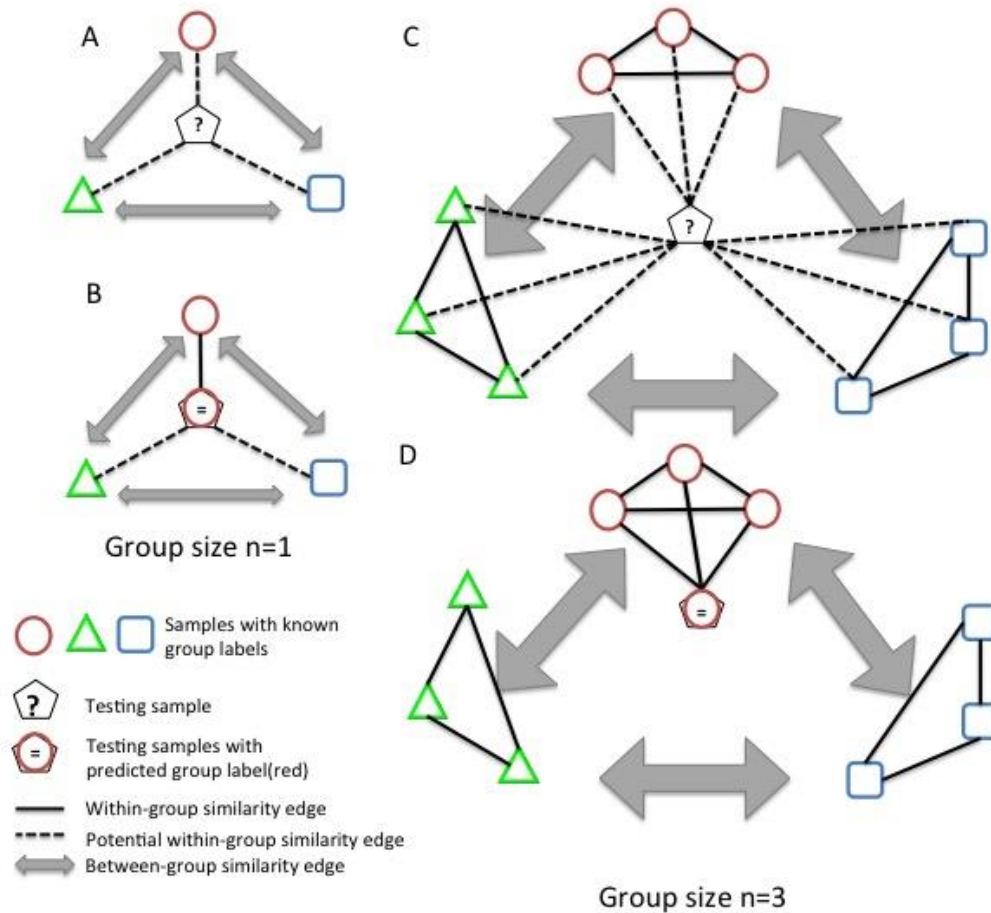


Figure E2: Prediction of group label using leave-one-out cross validation

Illustration of two different scenarios for prediction of group belonging for a subject, using label propagation and LOOCV. First, a test set consisting of one subject (marked as “?”) is randomly selected from all subjects. A subject-based similarity network is constructed based on pairwise subject similarities from the fused multi-omics similarity matrix of all remaining subjects, excluding the test set. The between-group similarity edges are calculated based on the connections within- (C, black lines) and between (C, grey arrows) groups. Second, the group label of the test set consisting of one subject (C; marked as “?”) is predicted based on the similarities to all samples in the network using the label propagation method (D). The unique example of personalized medicine, with subgroups sizes of $n=1$ is illustrated in panels A-B.

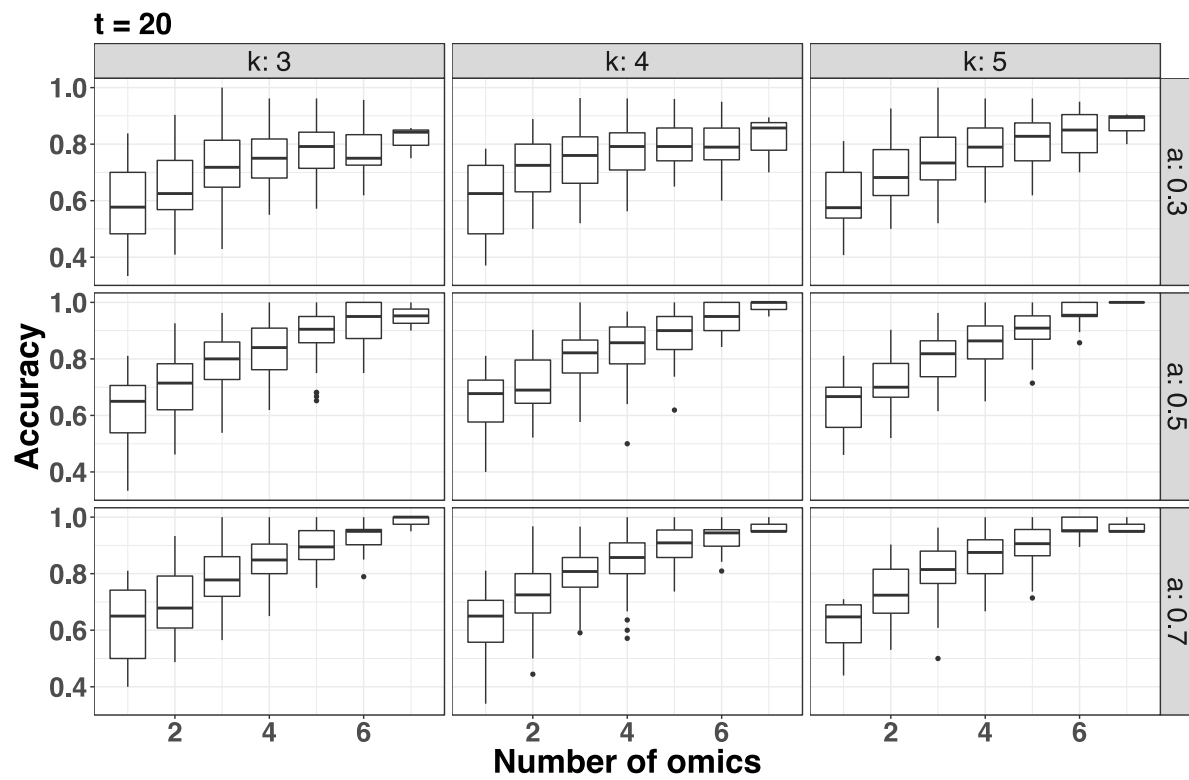
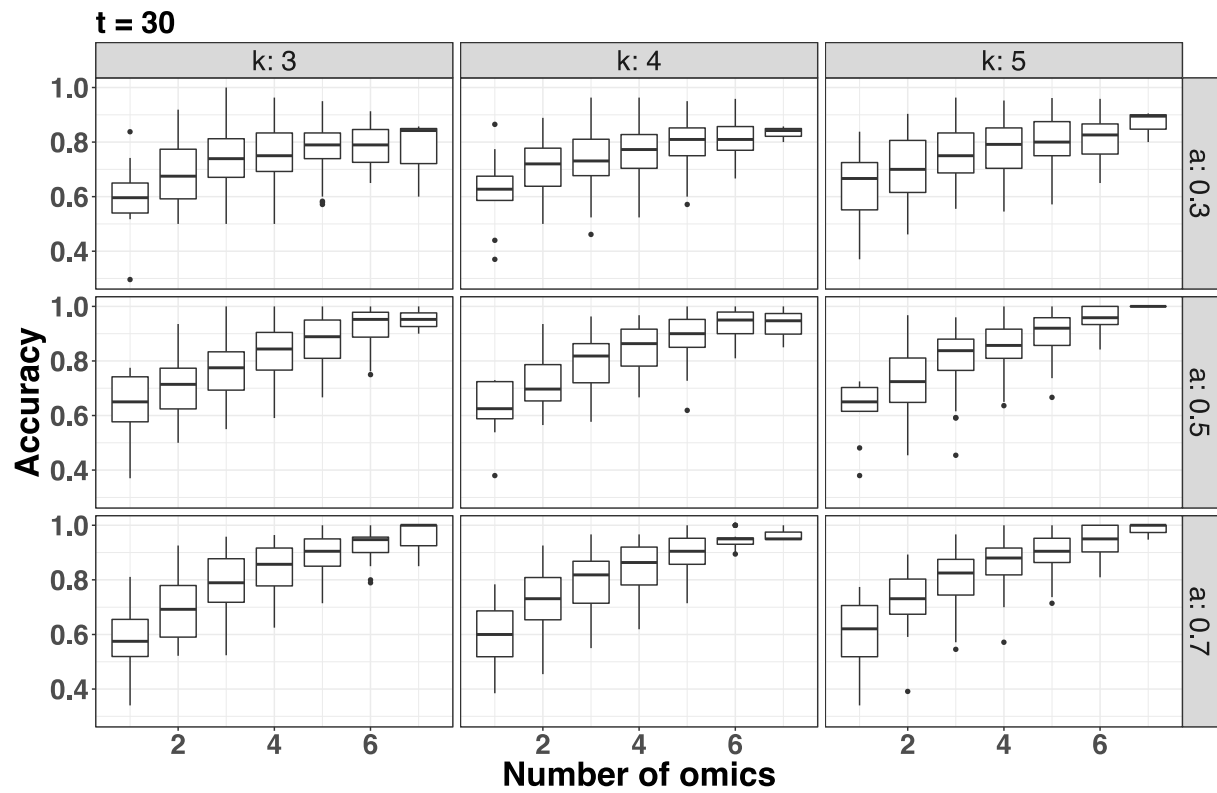
A**B**

Figure E3: SNF parameter optimization.

Evaluation of the three critical parameters used in SNF: the number of neighbors (K , in columns), hyperparameter (α as a in rows), and the number of iterations ($t = 20$ in panel A and $t = 30$ in panel B). The x-axis represents the n-tuple of multi-omics fusion, and y-axis is the accuracy of prediction (NMI). Box plots showing median (horizontal solid line), interquartile range (IQR; boxes), and range (whiskers). The accuracy is based on 303 single- to 7-tuple omics similarity networks using 24 samples from the three groups of female current-smoker COPD patients (6), smokers with normal lung function (10) and healthy never-smoker controls (8). Both use LOOCV with random sampling without replacement. As discussed by Wang et al. (17), these parameters are quite robust. We selected $K = 5$, $\alpha = 0.5$, and $t = 30$ in all further analyses.

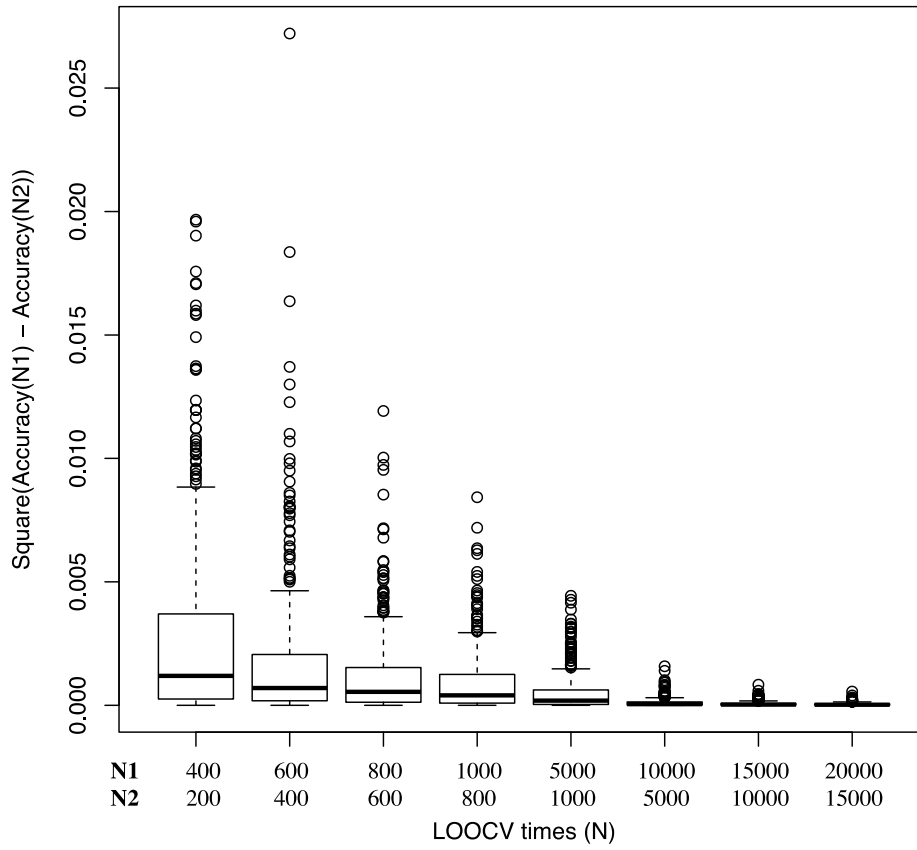


Figure E4: Optimization of number of sampling in LOOCV

Estimation of the variation in the robustness of accuracy of prediction, calculated as NMI, with the N -times LOOCV random sampling test. Boxplot displaying median (solid line), IQR (boxes), and range (whiskers) of the squared differences in accuracy (NMI) between each pair of permutation tests with $N1$ and $N2$ times sampling. The accuracy is based on 303 single- to 7-tuple omics similarity networks using 24 samples from the three groups of female current-smoker COPD patients (6), smokers with normal lung function (10) and healthy never-smoker controls (8). We use LOOCV random sampling with replacement with $K = 5$, $\alpha = 0.5$, and $t = 30$. Based on these results, 10,000-times LOOCV was utilized in all further analyses.

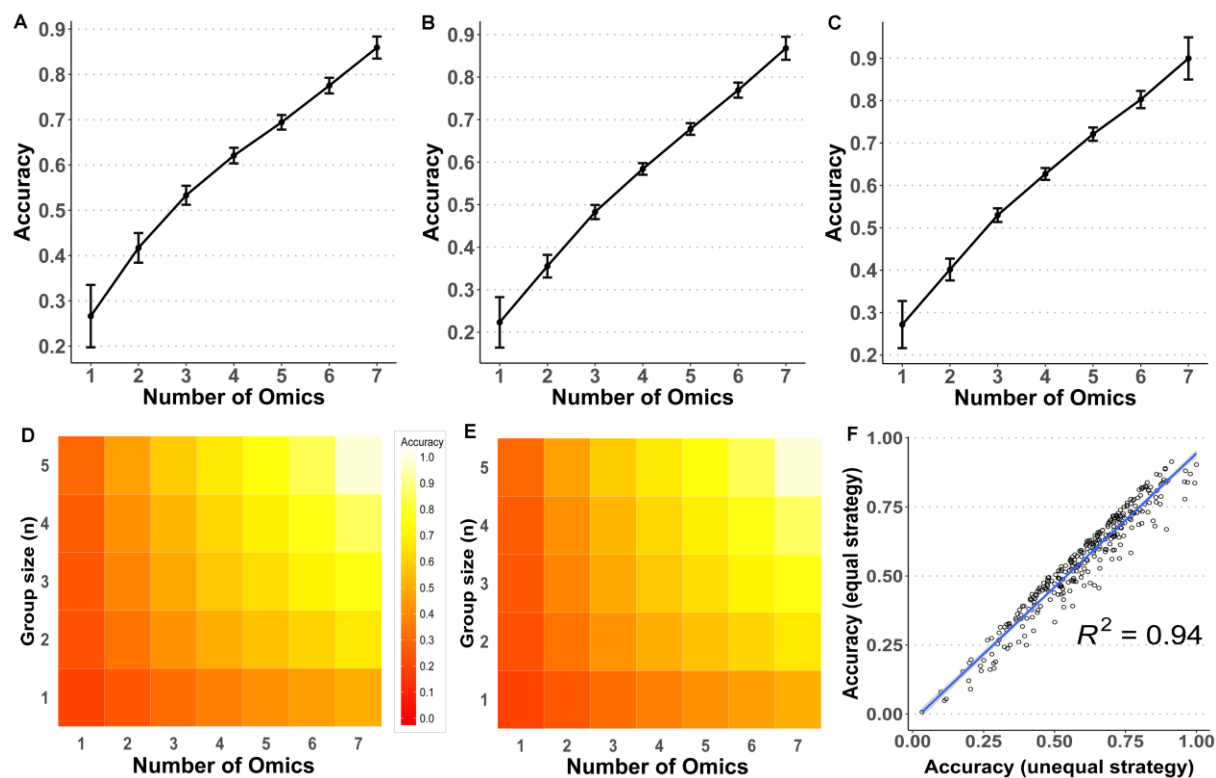


Figure E5: Comparison of strategies for handling missing omics data blocks

Panels A-C display the mean accuracy of group prediction for SNF-mediated omics integration using 9 omics data sets from the Karolinska COSMIC cohort, as displayed in Figure E1. Values are displayed as mean accuracy \pm SE of all possible omics combinations for each respective number of omics (n-tuple) combination based on the *Conservative* (A), *equal* (B) and *unequal sampling strategies* (C; identical to Figure 2A). The heat maps in panels D-E are displaying the accuracy of group prediction achieved when using sub-group sizes of $n=1-5$ (y-axis) for each number of omics platforms integrated (x-axis) for are displayed for the *conservative* (D) as compared to *equal* or *unequal sampling strategy* (E, identical to Figure 2C). Accuracy of prediction was calculated by comparing prediction using the SNF with COPD diagnosis according to the GOLD criteria as well as current smoking status to define correct reference groups. Panel F displays the correlation of accuracy of between *equal* vs. *unequal sampling strategy* ($R^2 = 0.94$).

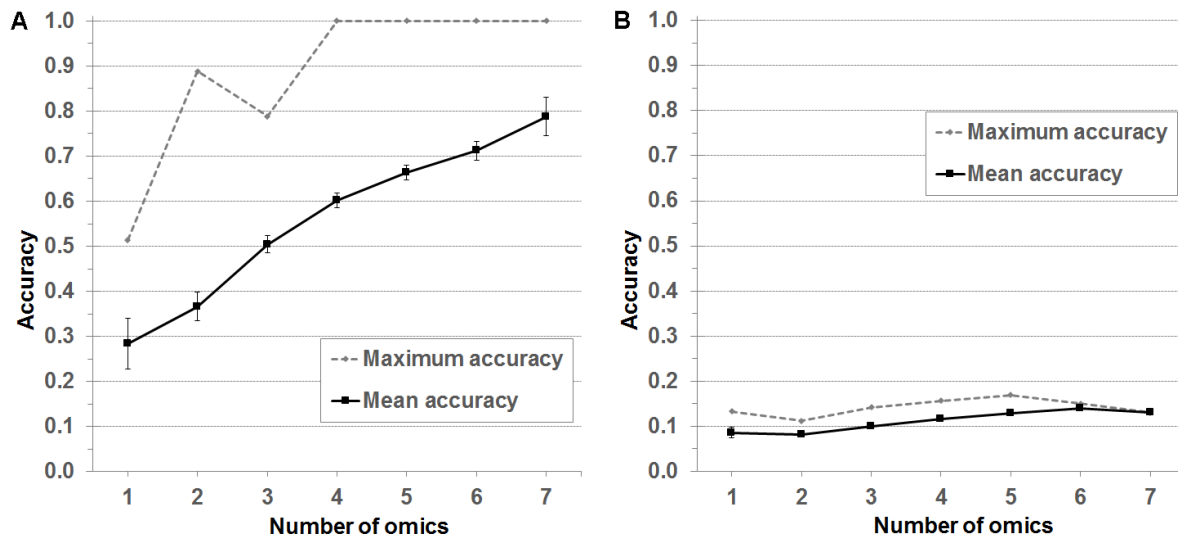


Figure E6: Accuracy of group prediction using spectralClustering

The accuracy of group prediction using the unsupervised spectralClustering algorithm provided in the SNFtool, to be contrasted to Figure 2A displaying the accuracy of group prediction using the label propagation method. A) Accuracy of prediction of the three study groups (Healthy never-smokers, Smokers with normal spirometry, and smokers with COPD). The graphs display the mean (solid line) and maximum (dashed line) accuracy of group prediction for n-tuple SNF-mediated omics integration using 9 omics data sets from the Karolinska COSMIC cohort, as displayed in Figure 1 and Figure E1. Values are displayed as mean accuracy \pm SE of all possible omics combinations for each respective n-tuple combination. Group belonging was predicted using spectralClustering, and accuracy of group prediction was calculated as NMI compared with COPD diagnosis according to the GOLD criteria as well as current smoking status to define correct reference groups. The mean performance was lower compared to the LOOCV (Figure 2A), with a higher variation between networks for the higher n-tuples. However, peak performing networks were achieved already at 4-tuple omics integration, as compared to 5-tuple integration required for 100% accurate prediction for the LOOCV (Figure 2A). Panel B shows the corresponding results following permutation of original omics data across all features for each subject separately (which means the feature-subject relationships are randomized), thereby corresponding to the accuracy of prediction that can occur by random in data sets of the same size. The improvement in accuracy observed 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 7-tuple omics.

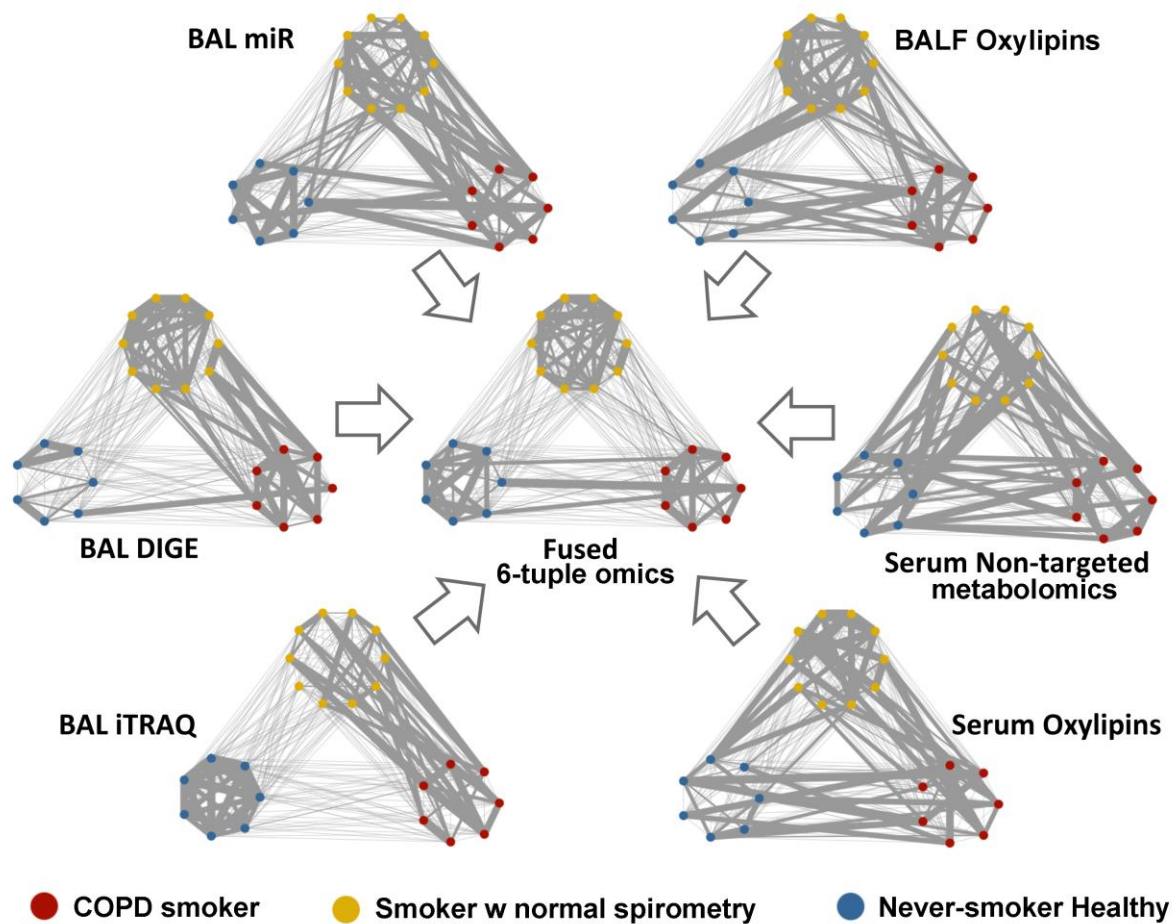
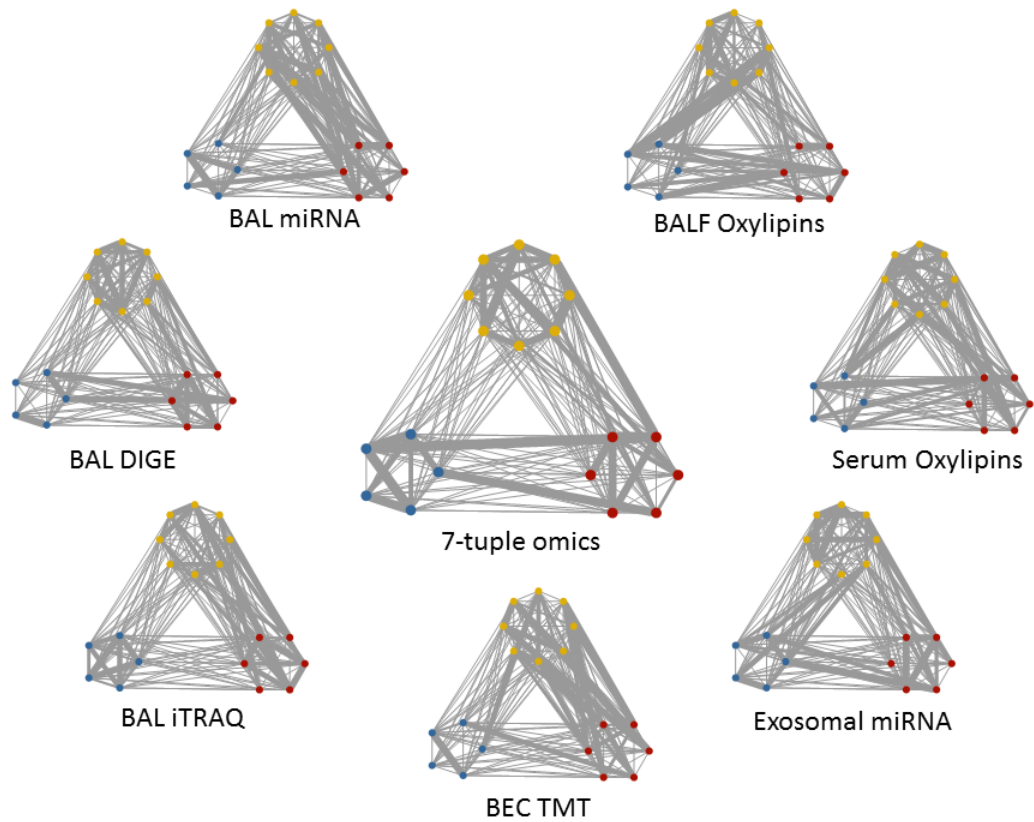


Figure E7: Sextuple SNF network with 100% accuracy of prediction

Subject similarity networks for each of the individual single-omics data blocks, compared to the optimal 6-tuple fused SNF similarity network (center), which resulted in 100% correct classification of the three groups. Nodes represent subjects (red: COPD current smokers, yellow: Current smokers with normal lung function, blue: Healthy never-smokers). Edge thickness reflects the strength of the similarity between each pair of subjects, with similarity ranks <75% displayed as a thin line, and similarity ranks 75-100% proportional to edge thickness. The accuracy of 100% is based on 10,000-times LOOCV permutation test using training data iteratively selecting 6 samples from each group.

A



B

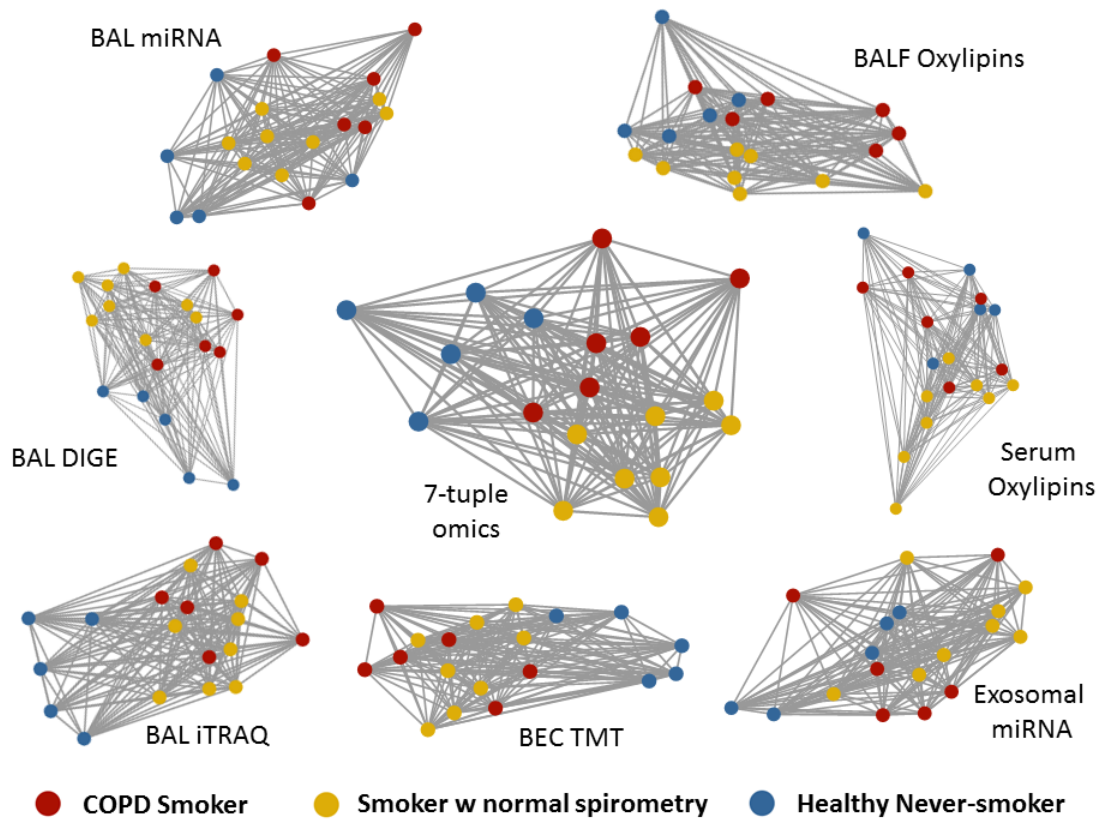


Figure E8: Optimal SNF network based on the conservative sampling strategy

Subject similarity networks for each of the individual single-omics data blocks, compared to the optimal septuple fused SNF similarity network (center) achieved from the conservative sampling strategy, which resulted in 91% correct classification of the three groups. Nodes represent subjects (red: COPD current smokers, yellow: current smokers with normal lung function, blue: healthy never-smokers; all female subjects). The upper panel (A) displays as fixed-position network, with clustering according to known groups preserved for all six networks to facilitate visual comparison. Edge thickness reflects the strength of the similarity between each pair of patients, with similarities rank in each network $<75\%$ displayed as a thin line, and similarities rank 75-100% proportional to edge thickness. The lower panel (B) displays the corresponding networks with subjects clustered according to network similarity. The accuracy of 91% is based on 10,000-times LOOCV permutation test using training data with 4 samples in each group.

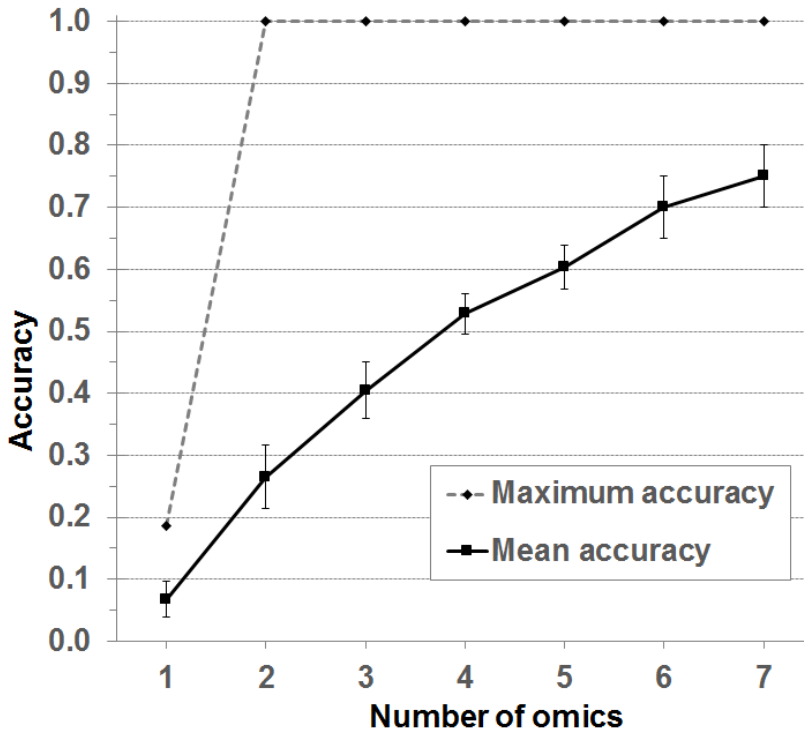


Figure E9: Accuracy of prediction of chronic bronchitis in COPD patients

The accuracy of group prediction of chronic bronchitis diagnosis among the female COPD group using the unsupervised, data driven prediction based on SNF multi-omics integration. The graphs display the mean (solid line) and maximum (dashed line) accuracy of prediction for each respective n-tuple combination using 8 omics data sets from the Karolinska COSMIC cohort as displayed in Figure 1 and Figure E1. One omics data set (mRNA from BAL cells) was excluded due to not fulfilling the criteria of a minimum coverage of $n=4$ subjects in each of the sub-group with/without *chronic bronchitis*. The mean accuracy increased in a near-linear fashion from <0.10 for the single omics data blocks to 0.75 for 7-tuple omics integration. Out of 254 possible single to 7-tuples omics networks, 57 networks of 2-7 omics combinations achieved an accuracy of 100% (dashed line) with group sizes as small as $n=4$. Group belonging was predicted using spectralClustering, and accuracy of group prediction was calculated as NMI compared with chronic bronchitis diagnosis as determined by self-reported cough and sputum production for ≥ 3 months in each of at least two consecutive years.

References

1. Kohler M, Sandberg A, Kjellqvist S, Thomas A, Karimi R, Nyrén S, et al. Gender differences in the bronchoalveolar lavage cell proteome of patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol*. 2013;131(3):743-51.
2. Mikko M, Forsslund H, Cui L, Grunewald J, Wheelock AM, Wahlstrom J, et al. Increased intraepithelial (CD103+) CD8+ T cells in the airways of smokers with and without chronic obstructive pulmonary disease. *Immunobiology*. 2013;218(2):225-31.
3. Forsslund H, Mikko M, Karimi R, Grunewald J, Wheelock AM, Wahlstrom J, 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(4):711-22.
4. Karimi R, Tornling G, Forsslund H, Mikko M, Wheelock A, Nyren S, et al. Lung density on high resolution computer tomography (HRCT) reflects degree of inflammation in smokers. *Respiratory research*. 2014;15:23.
5. Balgoma D, Yang M, Sjodin M, Snowden S, Karimi R, Levanen B, et al. Linoleic acid-derived lipid mediators increase in a female-dominated subphenotype of COPD. *Eur Respir J*. 2016;47(6):1645-56.
6. Forsslund H, Yang M, Mikko M, Karimi R, Nyren S, Engvall B, 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.
7. Karimi R, Tornling G, Forsslund H, Mikko M, Wheelock AM, Nyren S, et al. Differences in regional air trapping in current smokers with normal spirometry. *Eur Respir J*. 2017;49(1).
8. Sandberg A, Skold CM, Grunewald J, Eklund A, Wheelock AM. Assessing recent smoking status by measuring exhaled carbon monoxide levels. *PLoS One*. 2011;6(12):e28864.
9. Levanen B. Mechanisms of inflammatory signalling in chronic lung diseases : transcriptomics & metabolomics approaches [Doctoral Thesis]. Karolinska Institutet: Karolinska Institutet; 2012.
10. Levanen B, Bhakta NR, Torregrosa Paredes P, Barbeau R, Hiltbrunner S, Pollack JL, et al. Altered microRNA profiles in bronchoalveolar lavage fluid exosomes in asthmatic patients. *The Journal of allergy and clinical immunology*. 2013;131(3):894-903.
11. Kohler M, Sandberg A, Kjellqvist S, Thomas A, Karimi R, Nyren S, et al. Gender differences in the bronchoalveolar lavage cell proteome of patients with chronic obstructive pulmonary disease. *The Journal of allergy and clinical immunology*. 2013;131(3):743-51 e9.
12. Yang M, Kohler M, Heyder T, Forsslund H, Garberg HK, Karimi R, et al. Proteomic profiling of lung immune cells reveals dysregulation of phagocytotic pathways in female-dominated molecular COPD phenotype. *Respiratory research*. 2017;In press.
13. Yang M, Kohler M, Heyder T, Forsslund H, Garberg HK, Karimi R, 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. *Respiratory research*. 2017;In press.
14. Heyder T. Between two lungs: proteomic and metabolomic approaches in inflammatory lung diseases [Doctoral thesis]: Karolinska Institutet; 2017.

15. Naz S, Kolmert J, Yang M, Reinke SN, Kamleh MA, Snowden S, et al. Metabolomics analysis identifies gender-associated metabotypes of oxidative stress and the autotaxin-lysoPA axis in COPD. *Eur Respir J*. 2017;In press.
16. U.S.FDA. Guidance for Industry; Bioanalytical Method Validation. 2001.
17. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*. 2014;11(3):333-7.
18. de Leeuw J. Convergence of the majorization method for multidimensional scaling. *Journal of Classification*. 1988;5(2):163-80.
19. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-504.