Online Data Supplement

Integration of multi-omics datasets enables molecular classification of COPD

Chuan-xing Li[1], Craig E. Wheelock[2], C. Magnus Sköld[3] and Åsa M. Wheelock[1,*]

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

[*] Corresponding author:
Åsa M. Wheelock
Lung Research Lab L4:01
Respiratory Medicine Unit
Department of Medicine
Karolinska Institute
SE-171 76 Stockholm
Email: asa.wheelock@ki.se

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 ($\geq$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×*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 5µm, 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 µl, 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 5µm, 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 µg 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 TMT_{126} 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 vs1.4 (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 (*alpha*) 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) ± 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 (*alpha*), 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 *alpha* 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 *alpha* parameter was evaluated from 0.3 to 0.8, by an increment of 0.1 (Figure E3 shows *alpha* 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 x 10^{-4} and standard deviation as 7.6 x 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 (1001, 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 [×10^9/L]	5.6 (4.5, 6.5)	6.8 (6.4, 8.0)	8.2 (6.1, 9.4)
Blood platelets [×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, FEV$_1$ [%]= 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

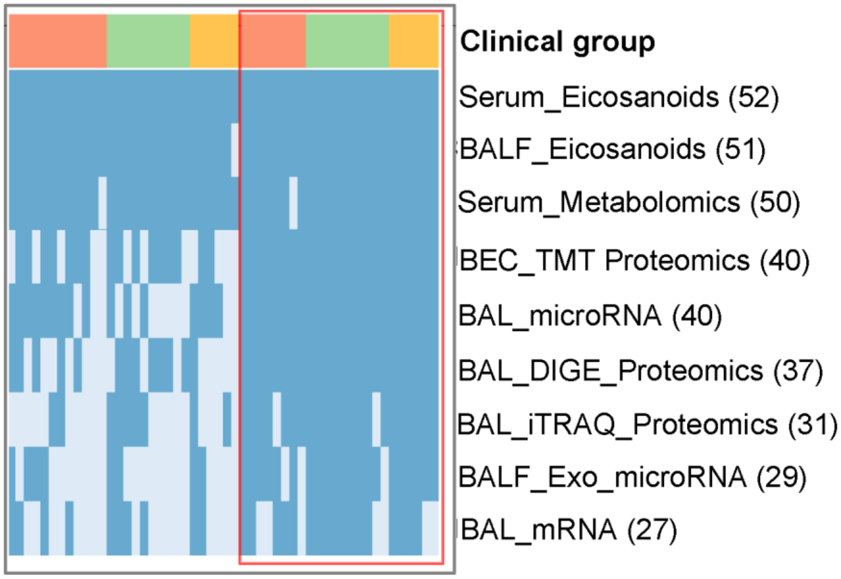


Figure E1: Overview of omics data blocks-subject matrix.
Overview of overlap of omics data collected from the 52 female subjects in the Karolinska COSMIC cohort that were included in the SNF performance evaluations. Each row represents a platform, and each column is a subject. Row one indicated subject 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*.

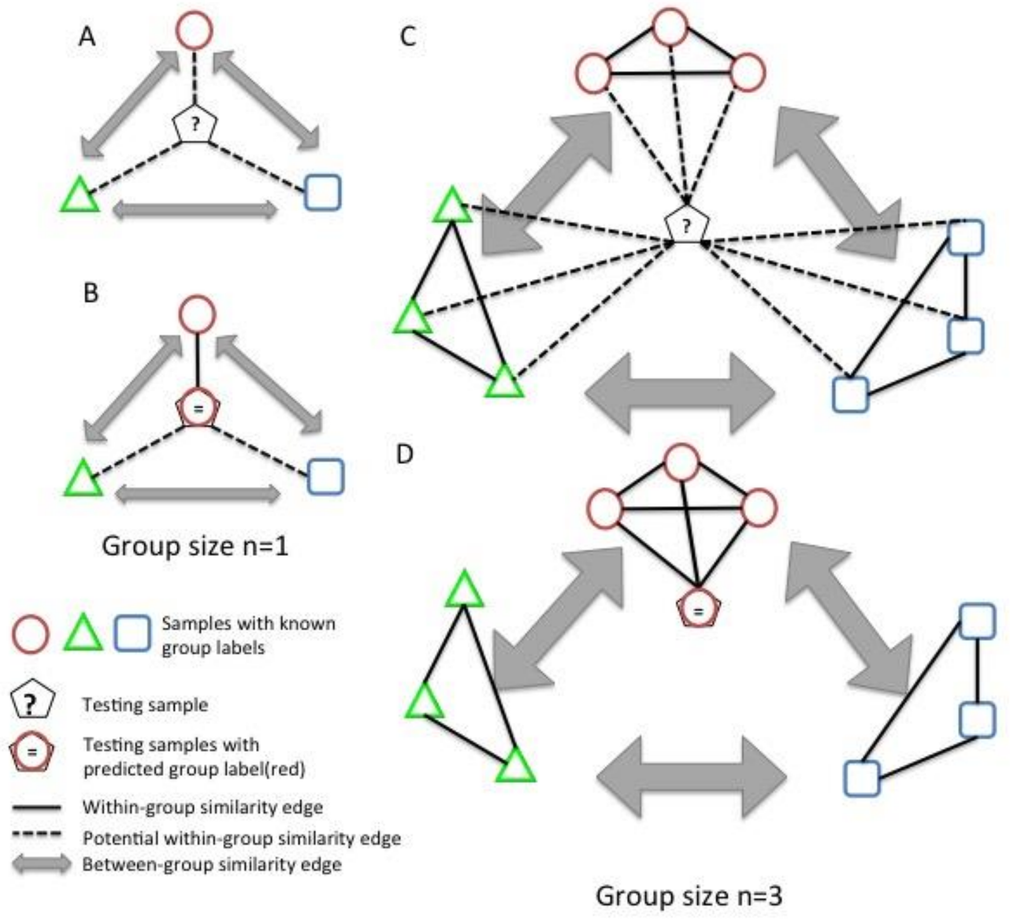
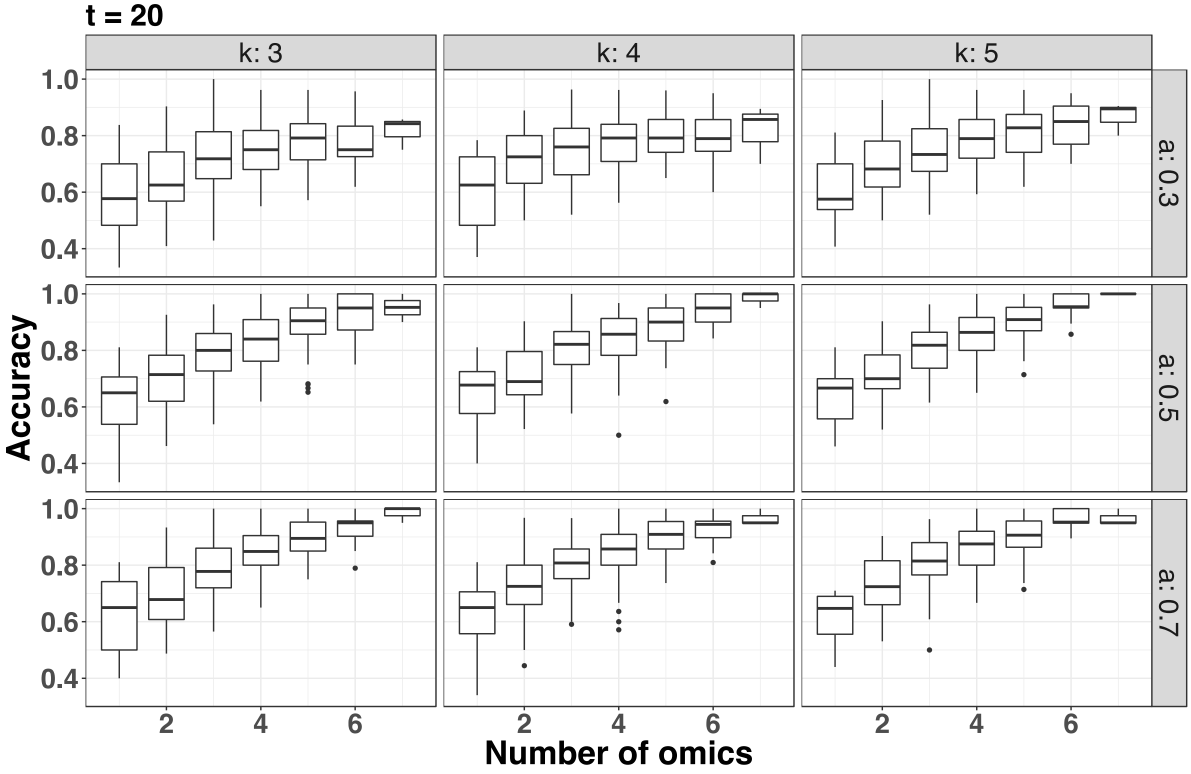

A
?
B
=
Group size n=1
C
?
D
=
Group size n=3
Samples with known group labels
Testing sample
Testing samples with predicted group label(red)
Within-group similarity edge
Potential within-group similarity edge
Between-group similarity edge

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


t = 20
k: 3
k: 4
k: 5
a: 0.3
a: 0.5
a: 0.7
Accuracy
1.0
0.8
0.6
0.4
2
4
6
Number of omics

B

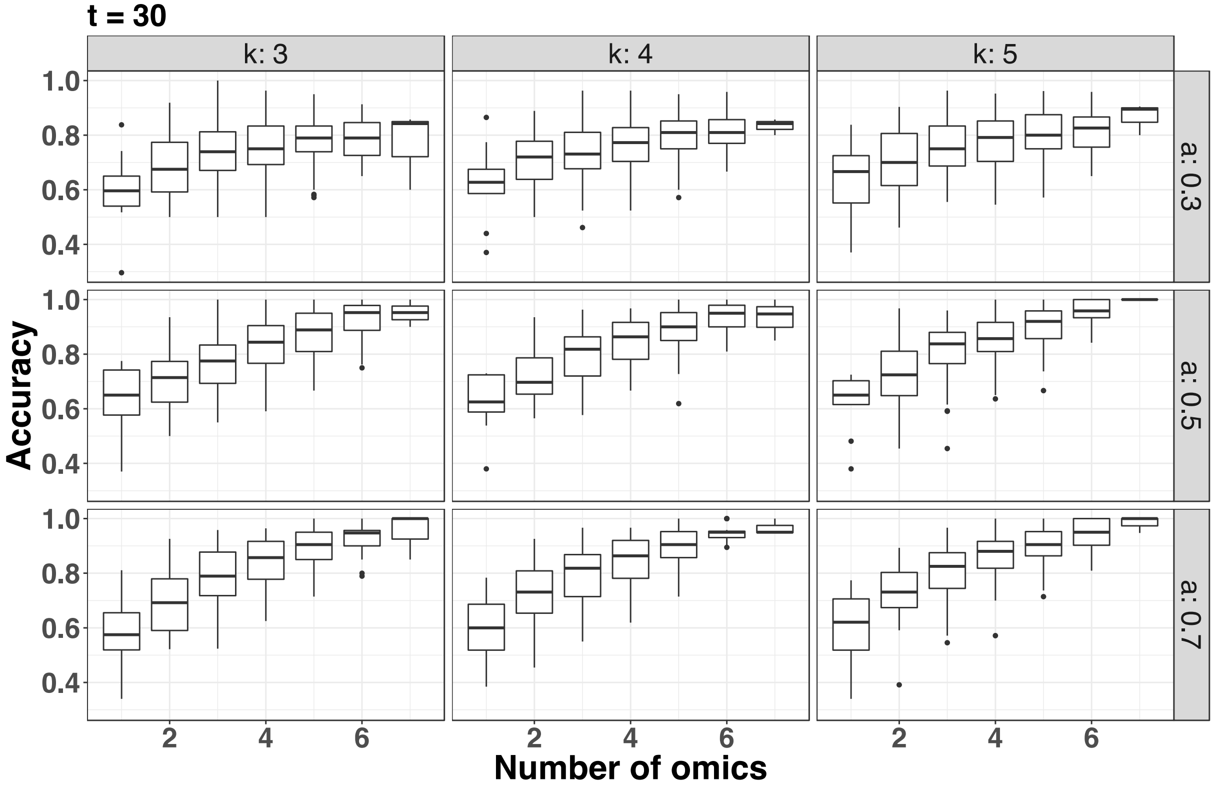

t = 30
k: 3
k: 4
k: 5
a: 0.3
a: 0.5
a: 0.7
Accuracy
1.0
0.8
0.6
0.4
2
4
6
Number of omics

Figure E3: SNF parameter optimization.

Evaluation of the three critical parameters used in SNF: the number of neighbors (*K*, in columns), hyperparameter (*alpha* 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.

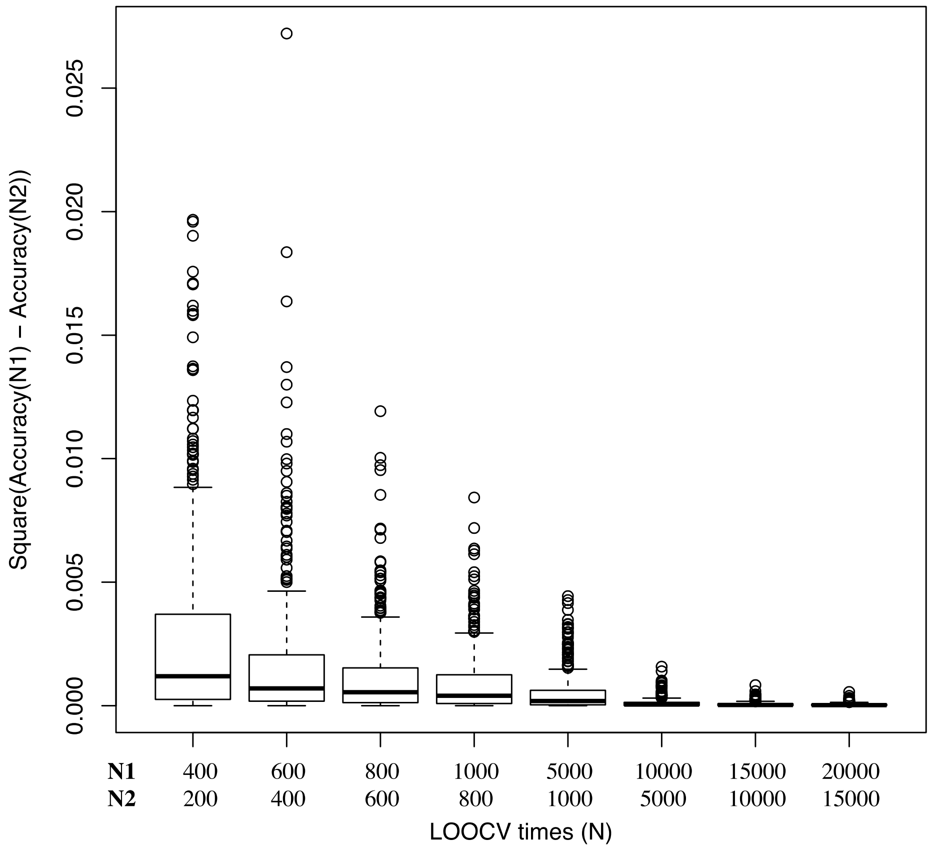


Figure E4: Optimization of number of sampling in LOOCV

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.

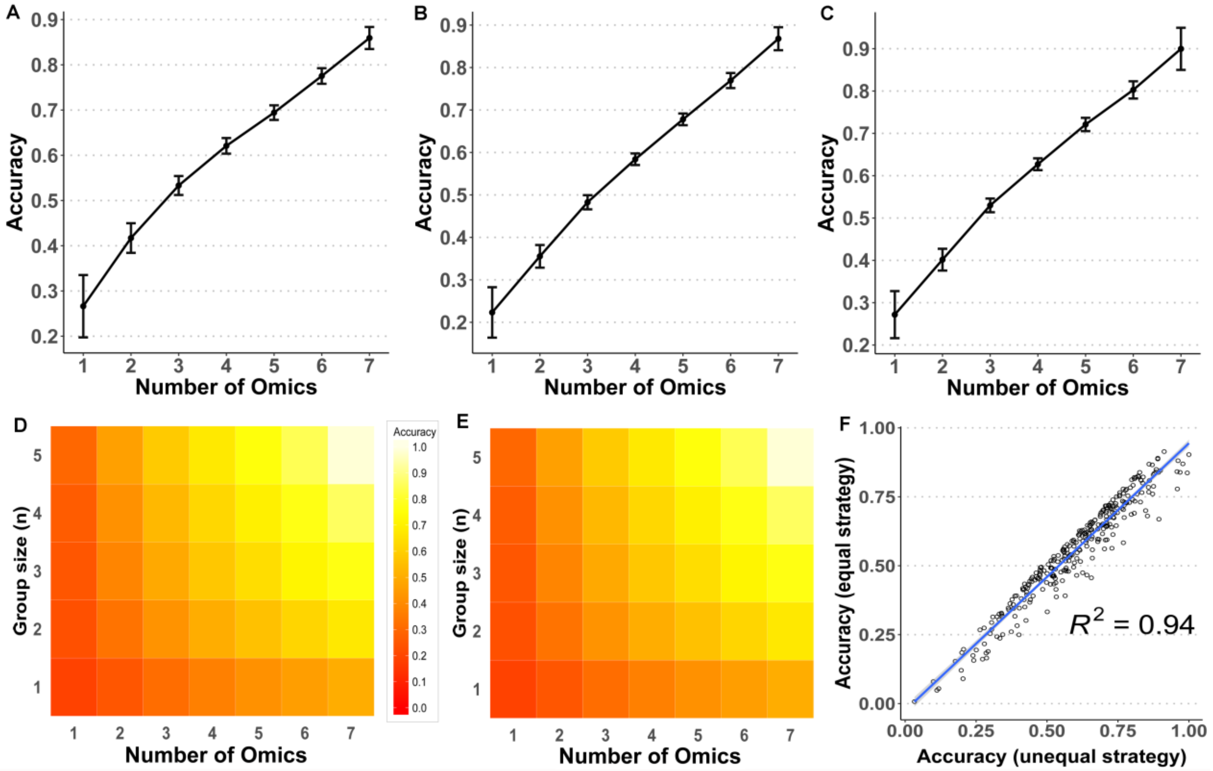


Figure E5: Comparison of strategies for handling missing omics data blocks
Panels A-C display the mean accuracy of group prediction for SNF-mediated omics integration using 9 omics data sets from the Karolinska COSMIC cohort, as displayed in Figure E1. Values are displayed as mean accuracy ± SE of all 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).

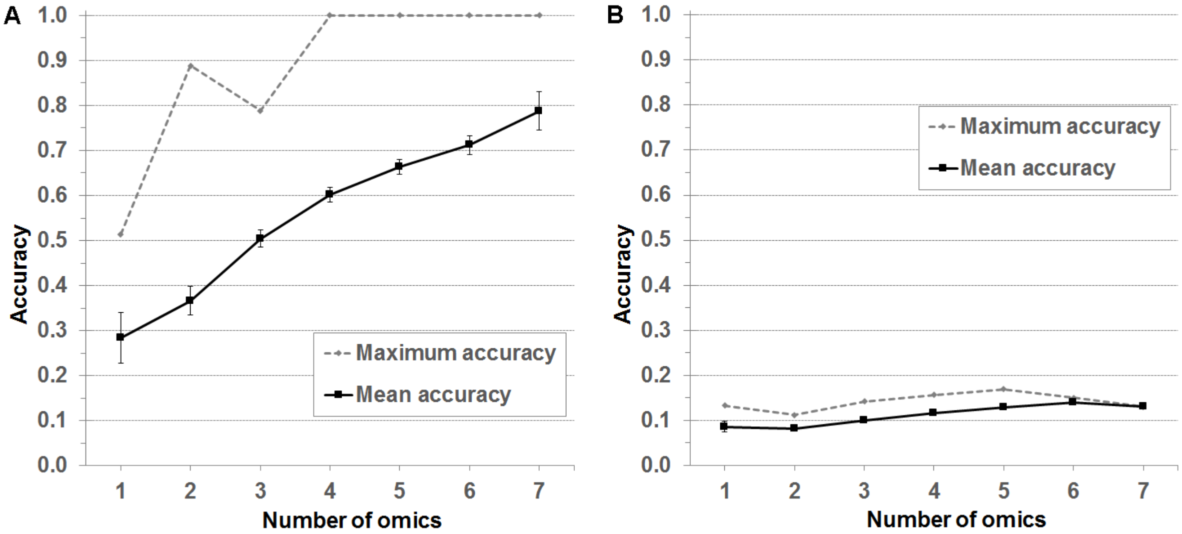


Figure E6: Accuracy of group prediction using spectralClustering

The accuracy of group prediction using the unsupervised spectralClustering alogorithm 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 ± 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.

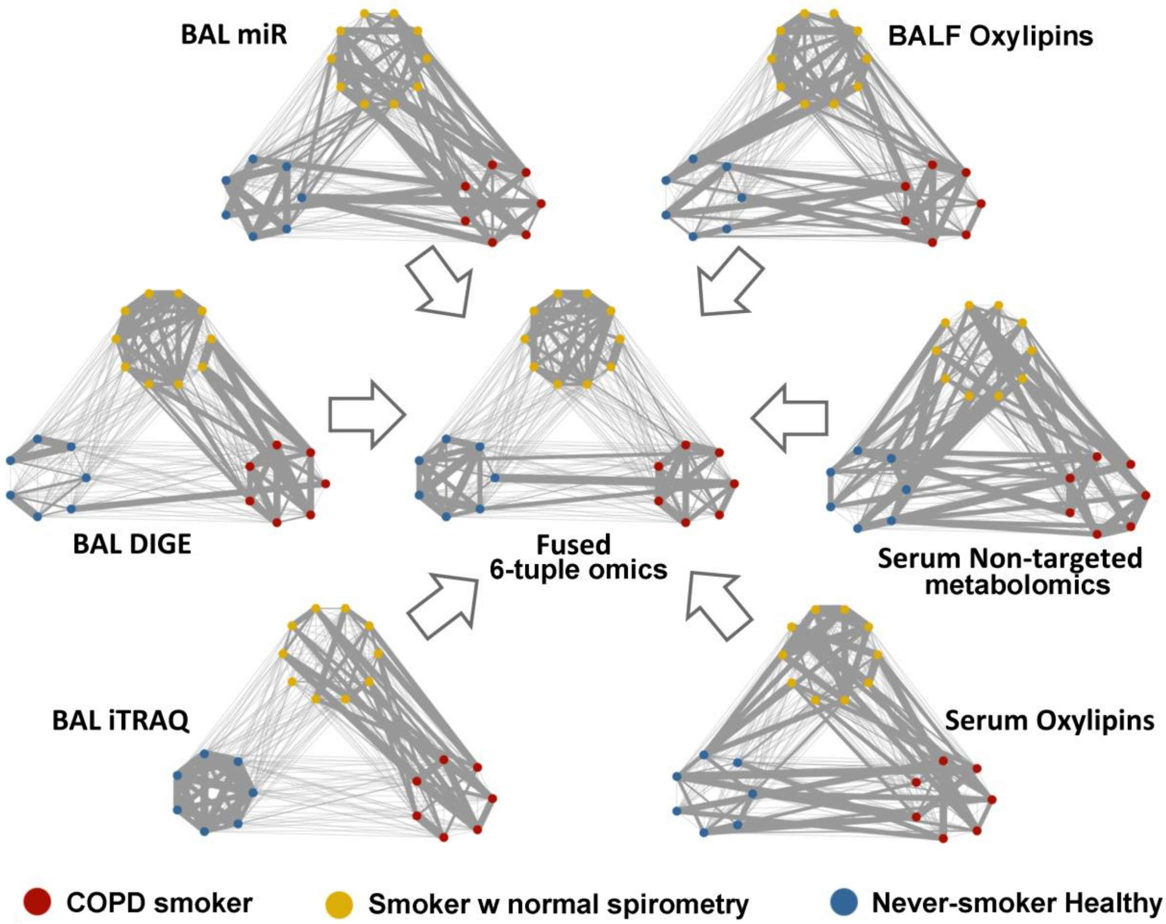


Figure E7: Sextuple SNF network with 100% accuracy of prediction

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.

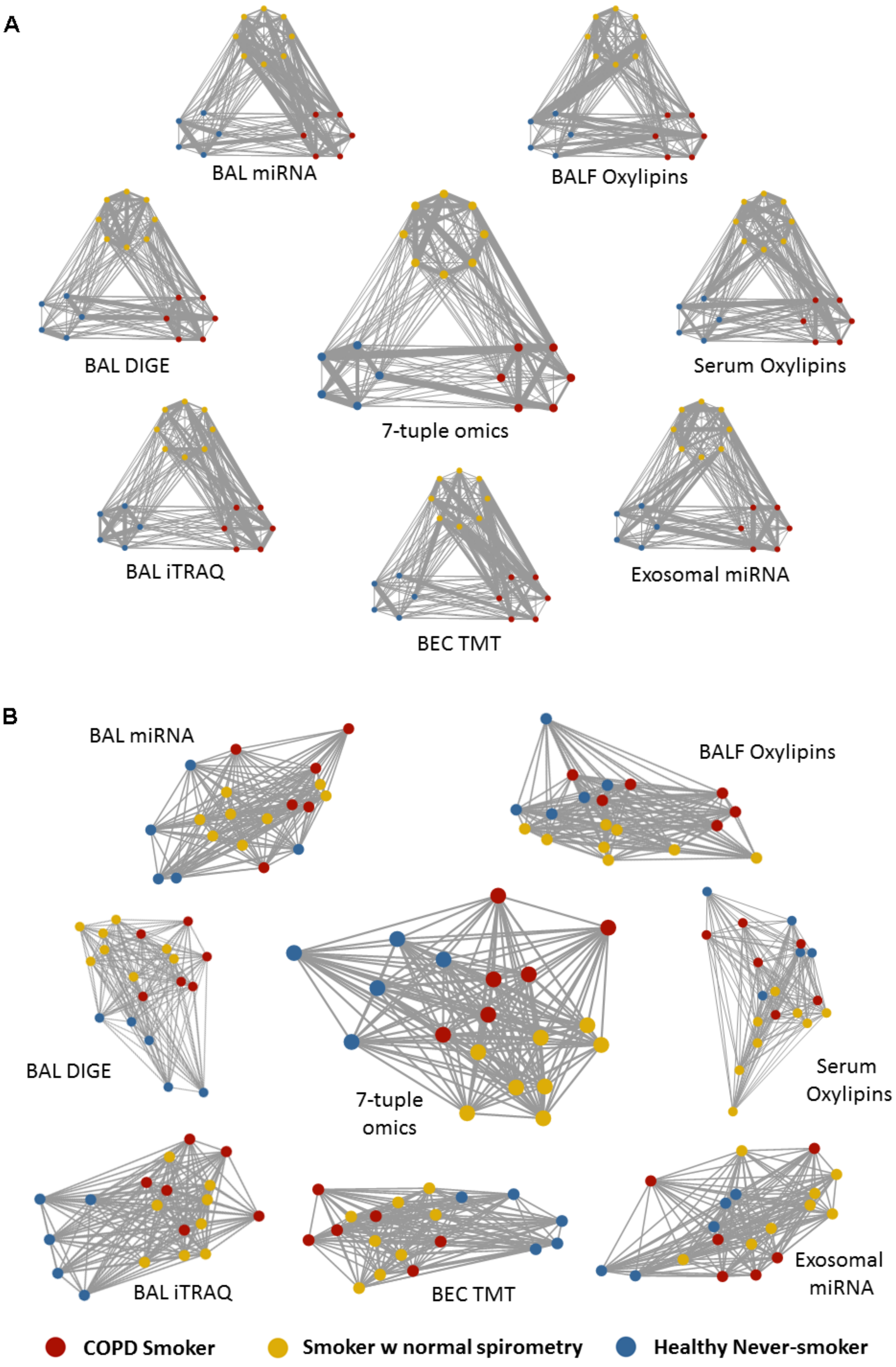

A
BAL miRNA
BALF Oxylipins
BAL DIGE
7-tuple omics
Serum Oxylipins
BAL iTRAQ
BEC TMT
Exosomal miRNA
B
BAL miRNA
BALF Oxylipins
BAL DIGE
7-tuple omics
Serum Oxylipins
BAL iTRAQ
BEC TMT
Exosomal miRNA
● COPD Smoker
● Smoker w normal spirometry
● Healthy Never-smoker

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.

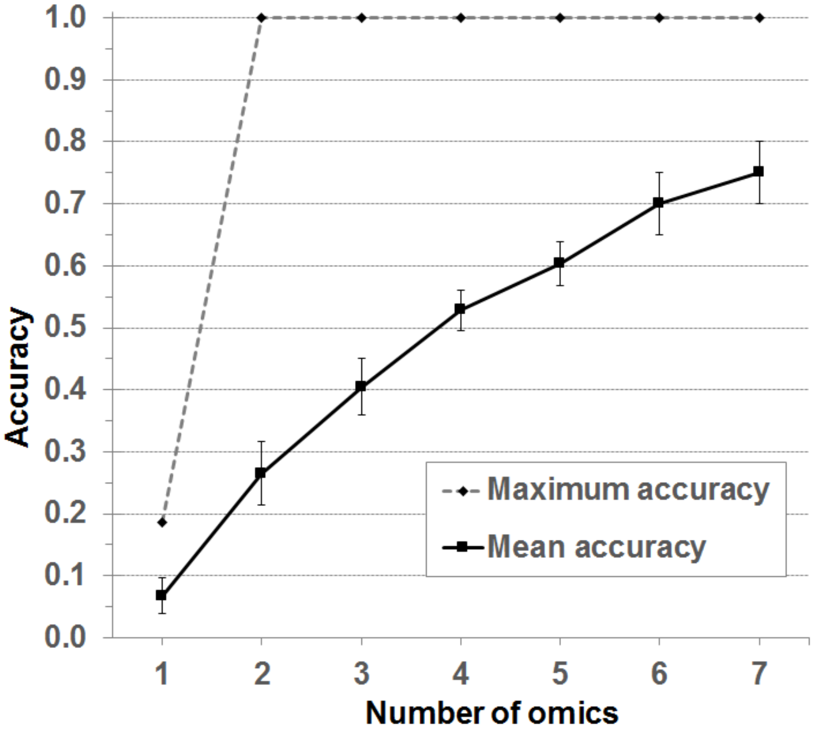


Figure E9: Accuracy of prediction of chronic bronchitis in COPD patients
The accuracy of group prediction of chronic bronchitis diagnosis among the female COPD group using the unsupervised, data driven prediction based on SNF multi-omics integration. The graphs display the mean (solid line) and maximum (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 ≥3months 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.