Data-driven patient stratification and drug target discovery by using medical information and serum proteome data of idiopathic pulmonary fibrosis patients

Yayoi Natsume-Kitatani (natsume@nibiohn.go.jp)
National Institutes of Biomedical Innovation, Health and Nutrition https://orcid.org/0000-0003-1749-3318

Mari N Itoh
Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

Yoshito Takeda
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Masataka Kuroda
Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

Haruhiko Hirata
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Kohtaro Miyake
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Takayuki Shiroyama
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Yuya Shirai
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Yoshimi Noda
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Yuichi Adachi
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Takatoshi Enomoto
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Saori Amiya
Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Jun Adachi
Laboratory of Proteomics for Drug Discovery, Center for Drug Design Research, National Institutes of Biomedical Innovation, Health and Nutrition

Ryohei Narumi
Laboratory of Proteomics for Drug Discovery, Center for Drug Design Research, National Institutes of Biomedical Innovation, Health and Nutrition

Satoshi Muraoka
Laboratory of Proteomics for Drug Discovery, Center for Drug Design Research, National Institutes of Biomedical Innovation, Health and Nutrition

Takeshi Tomonaga
Laboratory of Proteomics for Drug Discovery, Center for Drug Design Research, National Institutes of Biomedical Innovation, Health and Nutrition

Sadao Kurohashi
Graduate School of Informatics, Kyoto University

Fei Cheng
Graduate School of Informatics, Kyoto University

Ribeka Tanaka
Graduate School of Informatics, Kyoto University

Shuntaro Yada
Graduate School of Science and Technology, Nara Institute of Science and Technology (NAIST)

Eiji Aramaki
Graduate School of Science and Technology, Nara Institute of Science and Technology (NAIST)

Shoko Wakamiya
Graduate School of Science and Technology, Nara Institute of Science and Technology (NAIST)

Yi-An Chen
Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

Chihiro Higuchi
Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

Yosui Nojima
Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

Takeshi Fujiwara

Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

Chioko Nagao

Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

Toshihiro Takeda

Osaka University Graduate School of Medicine

Yasushi Matsumura

Osaka National Hospital

Kenji Mizuguchi

Laboratory of Bioinformatics, Artificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition

Atsushi Kumanogoh

Department of Respiratory Medicine and Clinical Immunology, Osaka University Graduate School of Medicine

Naonori Ueda

RIKEN Center for Advanced Intelligence Project

Research Article

Keywords: heterogeneous data integration, biomarker detection, patient stratification, idiopathic pulmonary fibrosis

Posted Date: August 10th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-405195/v4

Abstract

Medical information is valuable information obtained from humans regarding the phenotype of diseases. Omics data is informative to understand diseases at biomolecular level. We aimed to detect patient stratification patterns in a data-driven manner and identify candidate drug targets by investigating biomolecules that are linked to phenotype-level characteristics of a targeted disease. Such data integration is challenging because the data types of them are different, and these data contain many items that are not directly related to the disease. Hence, we developed an algorithm, subset binding, to find inter-related attributes in heterogeneous data. To search for potential drug targets for intractable IPF (idiopathic pulmonary fibrosis), we collected medical information and proteome data of serum extracellular vesicles from patients with interstitial pneumonia including IPF. Our approach detected 20 proteins linked with IPF characteristics, whose expression intensities were confirmed to be high in fibrotic areas of human lung tissues. Furthermore, ponatinib, which inhibits these proteins, suppressed EMT (epithelial mesenchymal transition) in vitro. This workflow paves the way for data-driven drug target discovery even for intractable diseases whose mechanisms of pathogenesis are not fully understood.

Introduction

One of the biggest problems in current drug discovery is the high failure rate of POC (Proof of concept) in Phase II clinical trials. The major cause of this failure is that no significant efficacy was observed (Arrowsmith and Miller 2013). This means that the drug effect observed in experimental animals used for drug target discovery was not observed in humans, and it is thought that the limitations of drug target discovery using experimental animals have come to the surface.

Another reason can be inadequate patient stratification. Even within a group of patients diagnosed with the same disease, the characteristics are not uniform, and they can be divided into subgroups that differ in such as prognosis, response to treatment including medication, and risks of side effects. In this case, appropriate therapeutic approaches and drug targets may differ for each subgroup.

Against this background, we have come up with an idea that it would be possible to improve the failure rate in Phase II clinical trials by searching for drug targets by using human data. Human data that is expected to be useful for this purpose includes medical information of patients with target diseases for drug target search. In addition, since molecular-level information is also necessary to search for drug targets, omics data linked to medical information is expected to be invaluable.

A canonical patient stratification approach would begin with a search for biomarkers that serve as criteria for dividing the patient population into subgroups. The biomarker biomolecules are quantified, and the patient population is divided into subgroups by their patterns, so that subgroups that differ even in phenotype level can be obtained. Therefore, if we can link disease phenotypes and biomolecules to obtain many-to-many relationships in a data-driven manner through integrated analysis of medical information and omics data, we can obtain both a list of biomolecules that can be used as biomarkers and disease phenotypes of patient subgroups stratified by these biomolecules.

If the data-driven drug target discovery using human data is possible with methods like above-mentioned, the benefit will be particularly enormous for intractable diseases for which the mechanisms of disease development are not fully understood at the molecular level. This is because it is possible to conduct drug target discovery by collecting data even if the accumulated knowledge is limited. Therefore, we selected idiopathic pulmonary fibrosis (IPF) as a target disease and conducted drug target discovery by patient stratification using medical information and omics data.

IPF is a chronic, progressive, intractable respiratory disease that is included in idiopathic interstitial pneumonias (IIPs). IIPs are interstitial pneumonias with no identifiable cause and are designated as intractable diseases in Japan. IIPs include IPF, nonspecific interstitial pneumonia (NSIP), cryptogenic organizing pneumonia (COP), idiopathic bronchiolitis obliterans organizing pneumonia (idiopathic BOOP), acute interstitial pneumonia (AIP), desquamative interstitial pneumonia (DIP), respiratory bronchiolitis associated interstitial lung disease (RB-ILD), and lymphocytic interstitial pneumonia (LIP). IPF has a very poor prognosis, with a mean survival of 3 to 5 years after diagnosis and a survival of less than 2 months after acute exacerbation. IPF patients often fail to respond to steroids, and no fundamental treatment has been established. There are only two drug treatment options: antifibrotic agents pirfenidone and nintedanib. Since the pathogenic mechanism of this disease is unknown, innovative approaches to drug target discovery that do not rely on conventional methods are required. However, the classification of IIPs is not trivial even for specialists, and multidisciplinary discussion (MDD; in which physicians of different specialties such as respiratory medicine, radiology, and pathology discuss and make a final diagnosis) is strongly recommended for reliable diagnosis. In addition, patients diagnosed with IPF do not have uniform characteristics, and there are individual differences in the response to the above-mentioned antifibrotic drugs and the severity of side effects. Therefore, stratification of patients according to the type of IPF would be effective in optimizing treatment strategies and finding clues for the development of new drugs.

Biomarker discovery is the most common method for patient stratification. For example, proteomic profiles of extracellular vesicles (EVs) are reported to be promising markers for cancer detection and cancer-type determination (Hoshino, Kim et al. 2020). We have reported several biomarkers for patient stratification or evaluation of the disease severity in refractory respiratory diseases such as chronic obstructive pulmonary disease (COPD) (Koba, Takeda et al. 2021) and sarcoidosis (Futami, Takeda et al. 2021) that were identified by proteome analysis of exosomes. Exosomes are a class of cell-derived extracellular vesicles of endosomal origin, and are typically 30-150 nm in diameter – the smallest type of extracellular vesicle. Enveloped by a lipid bilayer, exosomes are released into the extracellular environment containing a complex cargo of contents derived from the original cell, including proteins, lipids, mRNA, miRNA and DNA (L Isola and Chen 2017). Exosomes have been found to be responsible for various intercellular communication by transferring their contents in exosomes to other cells (Valadi, Ekström et al. 2007) (Zhang, Li et al. 2015). The expression state of these molecules has been shown to be deeply related to the state of cells and the progress of diseases in many diseases including cancer (Ludwig, Whiteside et al. 2019). These discoveries have attracted attention to the search for biomarkers using EVs including exosomes and their application in drug discovery. Moreover, miRNAs in serum exosomes in IPF have been suggested to reflect specific changes in miRNA expression in lung tissue of IPF patients (Njock, Guiot et al. 2019). Therefore, serum EVs including exosomes from patients with IPF, which is a multifactorial disease and shows a variety of pathological conditions due to the involvement of various cells, is likely to reflect the diversity of pathological conditions. Among a variety of biomolecules contained in EVs, proteins are the closest to phenotype and are important as direct functional molecules. Until now, the characterization of proteins in EVs from alveolar epithelium or lung tissue has been conducted in many diseases, but not in details from serum. In order to capture all the protein profiles in a certain mass range comprehensively in serum EVs from IPF patients, we utilized the cutting-edge proteomics technology called DIA (data-independent acquisition).

Herein, we constructed a database of serum EVs proteome data on interstitial pneumonia including IPF that is linked with their medical information and developed a novel machine learning algorithm, "subset binding", which can be used to detect patient stratification rules. Furthermore, when we obtain a list of biomolecules as biomarker candidates by using this method, we can identify candidate drug targets by investigating the biological responses of their quantitative changes and inferring their upstream regulators. We present a data-driven approach to drug target discovery using them and the outcomes obtained by this approach.

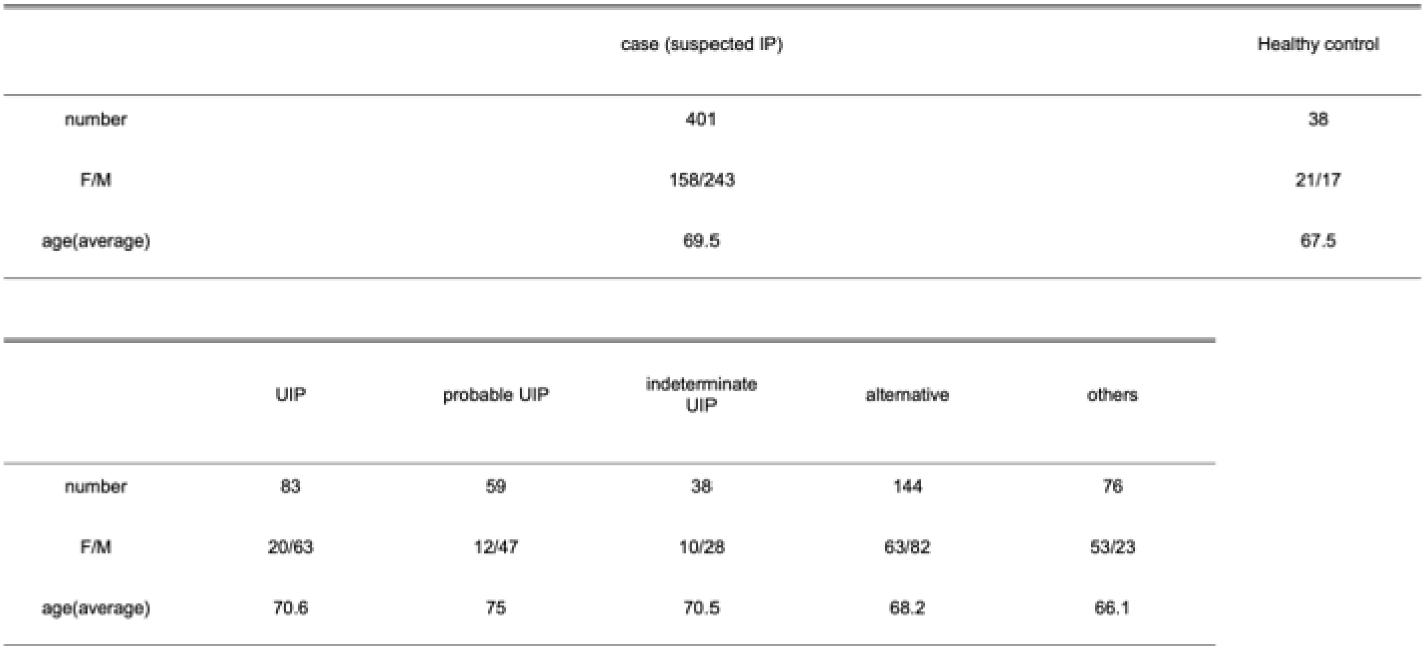
Results

Data collection from electronic medical record information and structuralization with natural language processing

The overall flow of data collection is shown in Fig. 1 a), and Fig. 1 b) outlines the procedure from collection of clinical data through electronic medical record entries (medical record and initial medical questionnaire, CT imaging interpretation report, blood test data) to input data generation. For example, the medical records are entered at the time of medical consultation using a format (referred to as a template) created with items set in advance, or the information is extracted manually from the entries and initial medical history questionnaire freely written in natural language into the template, and then the structured data were generated as input data. The CT imaging interpretation report, are paired by natural language processing with information about the entity related to the lesion and the site where it was observed, and the information about whether the lesion was observed (positive), not observed (negative), or suspected (suspected) is also added. The features were then manually modified and expressed as one-hot vectors for subsequent analysis. Blood test results were collected as structured data by manually extracting the test values for pre-selected items. The all above clinical information was collected at or near the date of blood collection for proteome data acquisition. For the proteome data, phosphatidylserine-positive extracellular vesicles were separated from serum, and the proteins contained were comprehensively measured by mass spectrometry (Fig. 1c). Each missing value was imputed with a representative value in healthy people, resulting in obtaining 6,506 (6,282 attributes from CT image interpretation reports, 171 attributes from blood test results and 53 attributes from medical records) × 602 cases (with overlap from 403 patients and 39 controls) of medical information and 2,445 protein ID × 602 cases matrices.

Basic patient characteristics and clinical items

The number of patients for whom clinical data (medical information and blood samples) was collected in this study and their basic characters are shown in Table I. The collected medical information is listed in Sup Table I, and 2,388 protein groups identified in the proteome analysis (2,445 proteins detected and 2,388 proteins were mapped to the known protein IDs) are shown in Sup Table II.



	case (suspected IP)	Healthy control
number	401	38
F/M	158/243	21/17
age(average)	69.5	67.5

	UIP	probable UIP	indeterminate UIP	alternative	others
number	83	59	38	144	76
F/M	20/63	12/47	10/28	63/82	53/23
age(average)	70.6	75	70.5	68.2	66.1

Table I patient basic characteristics

Concept of subset binding and analysis workflow

The composition of this cohort dataset used for the analysis of this study is shown in Fig. 2a), and the analysis workflow is shown in Fig. 2b). Subset binding (SB), a newly developed algorithm in this study, was used to detect patient stratification rules using structured medical information and proteome data. Subset binding outputs patient stratification rules (e.g., patients with high expression of biomolecules A, B, and C tend to show reticular shadows and traction bronchiectasis) by detecting association between phenotypic information such as medical information and biomolecular data such as omics data. SB uses fuzzy association rule mining as the underlying technology. It accepts two input matrices (e.g., proteome data and structured medical information; the number of rows must be the same, but the number of columns may be different), membership values for “Low” class and “High” class are calculated for each attribute using the membership function for each matrix, and association rules are generated so that the frequent itemsets from both data are linked (Fig. 2b; see Supplementary methods for the details of the algorithm). By using this algorithm, data with a mixture of continuous and discrete values, as is common in medical information, can be handled without any special preprocessing or prior knowledge. There are 6 possible combinations of SB analysis as shown in Fig. 2c, and we selected proteins that were included in the IPF characteristic-related association rules in the output of i) medical records (mixture of binary and numerical values) – protein (numerical values) association, ii) CT interpretation reports (binary) – protein association, and/or iii) blood test (mixture of binary and numerical values) – protein association at least once. The IPF characteristics used to select association rules include a known biomarker KL-6 (sialylated carbohydrate antigen) in blood test, respiratory difficulties during exertion in medical records.

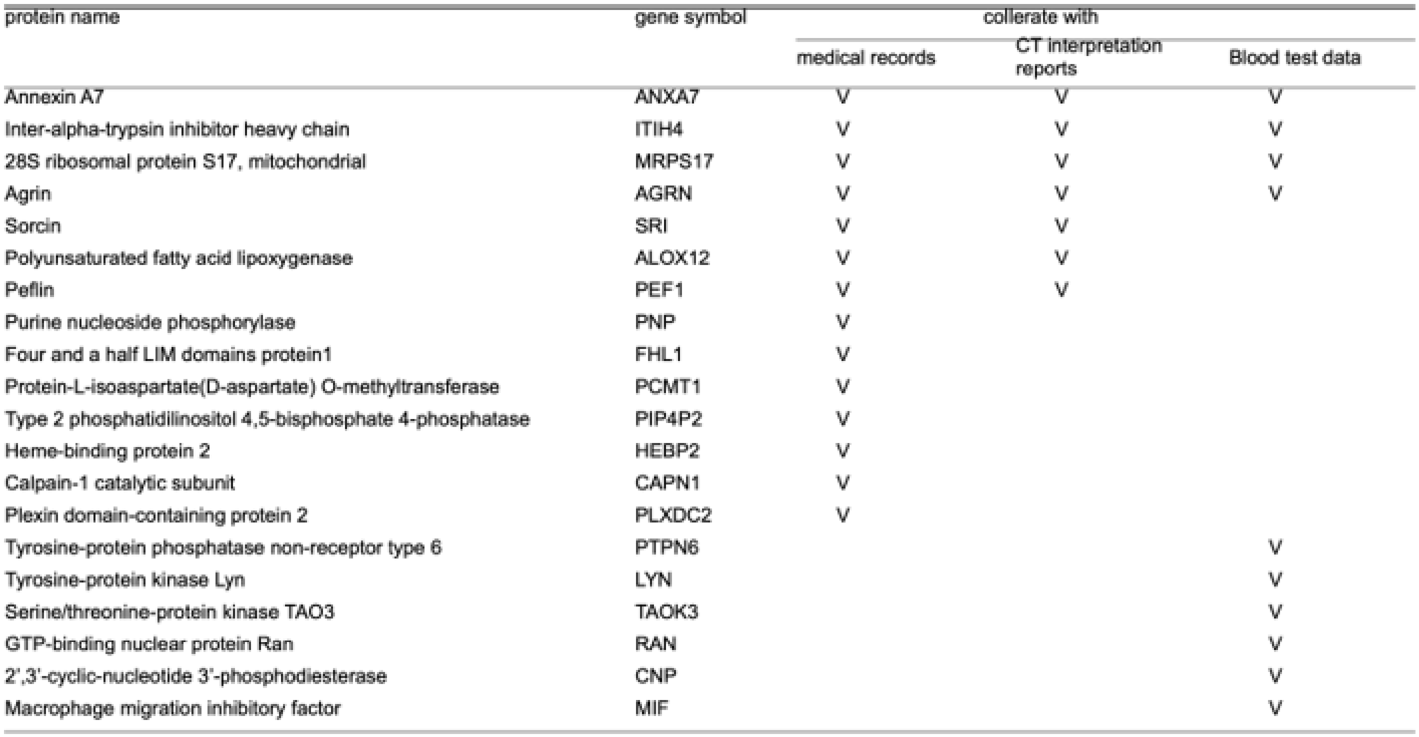
Clustering of proteome data is not suitable for patient stratification

To investigate whether the global similarities of cases in the proteomic data of serum EVs reflect the diagnosis, we visualized their quantitative patterns by heatmap with hierarchical clustering (Fig. 3a), t-SNE (Fig. 3b) and UMAP (Fig. 3c). The heatmap shows that the global similarities among cases didn't match with their diagnosis, which implied that the canonical approach such as clustering is not suitable for patient stratification. It is indicated that the proteome data contained many proteins that were not directly linked with phenotypes such as diagnosis. Fig. 3b and 3c also supported this tendency, in which several subtypes in IIPs (UIP, probable UIP, indeterminate UIP, alternative, and others) didn't show co-localization while HC (healthy control) showed weak tendency to co-localize.

Since the canonical machine learning techniques that assume the global similarities among cases are high if the diagnosis and/or phenotypic characteristics are similar, we searched proteins that linked with IPF-characteristics by SB as shown in Fig. 2b, which resulted in finding 20 proteins.

The top 20 proteins that co-occurred with characteristic findings of IPF by SB and their relationship to IPF

The 20 IPF-related proteins found by SB are shown in Table II. The protein-protein interrelationships among the 20 molecules were searched using TargetMine (Chen, Tripathi et al. 2011) (Chen, Tripathi et al. 2016) (Chen, Tripathi et al. 2019). LYN (Tyrosine-protein kinase Lyn), PTPN6 (Tyrosine-protein phosphatase non-receptor type 6), MIF (Macrophage migration inhibitory factor) and RAN (GTP-binding nuclear protein Ran) were found to be the hub molecules in the protein-protein interactions among these 20 molecules (Fig. 4a). In addition, the presence or absence of a relationship between 20 molecules was explored using TargetMine and IPA Ingenuity Pathway Analysis, QIAGEN , and the results are shown in Fig. 4b) and Sup. Table 3. As a result, molecules with no previously reported association were MRPS17 (28S ribosomal protein S17, mitochondrial) and PEF1 (Peflin), whereas molecules those were found to be associated with IPF through many other molecules were LYN (Tyrosine-protein kinase Lyn), PTPN6 (Tyrosine-protein phosphatase non-receptor type 6), MIF (Macrophage migration inhibitory factor) and RAN (GTP-binding nuclear protein Ran).



protein name	gene symbol	collerate with		
		medical records	CT interpretation reports	Blood test data
Annexin A7	ANXA7	V	V	V
Inter-alpha-trypsin inhibitor heavy chain	ITIH4	V	V	V
28S ribosomal protein S17, mitochondrial	MRPS17	V	V	V
Agrin	AGRN	V	V	V
Sorcin	SRI	V	V	
Polyunsaturated fatty acid lipoxygenase	ALOX12	V	V	
Peflin	PEF1	V	V	
Purine nucleoside phosphorylase	PNP	V		
Four and a half LIM domains protein1	FHL1	V		
Protein-L-isoaspartate(D-aspartate) O-methyltransferase	PCMT1	V		
Type 2 phosphatidilinositol 4,5-bisphosphate 4-phosphatase	PIP4P2	V		
Heme-binding protein 2	HEBP2	V		
Calpain-1 catalytic subunit	CAPN1	V		
Plexin domain-containing protein 2	PLXDC2	V		
Tyrosine-protein phosphatase non-receptor type 6	PTPN6			V
Tyrosine-protein kinase Lyn	LYN			V
Serine/threonine-protein kinase TAO3	TAOK3			V
GTP-binding nuclear protein Ran	RAN			V
2',3'-cyclic-nucleotide 3'-phosphodiesterase	CNP			V
Macrophage migration inhibitory factor	MIF			V

Table II List of proteins co-occurring with medical information with IPF characteristics

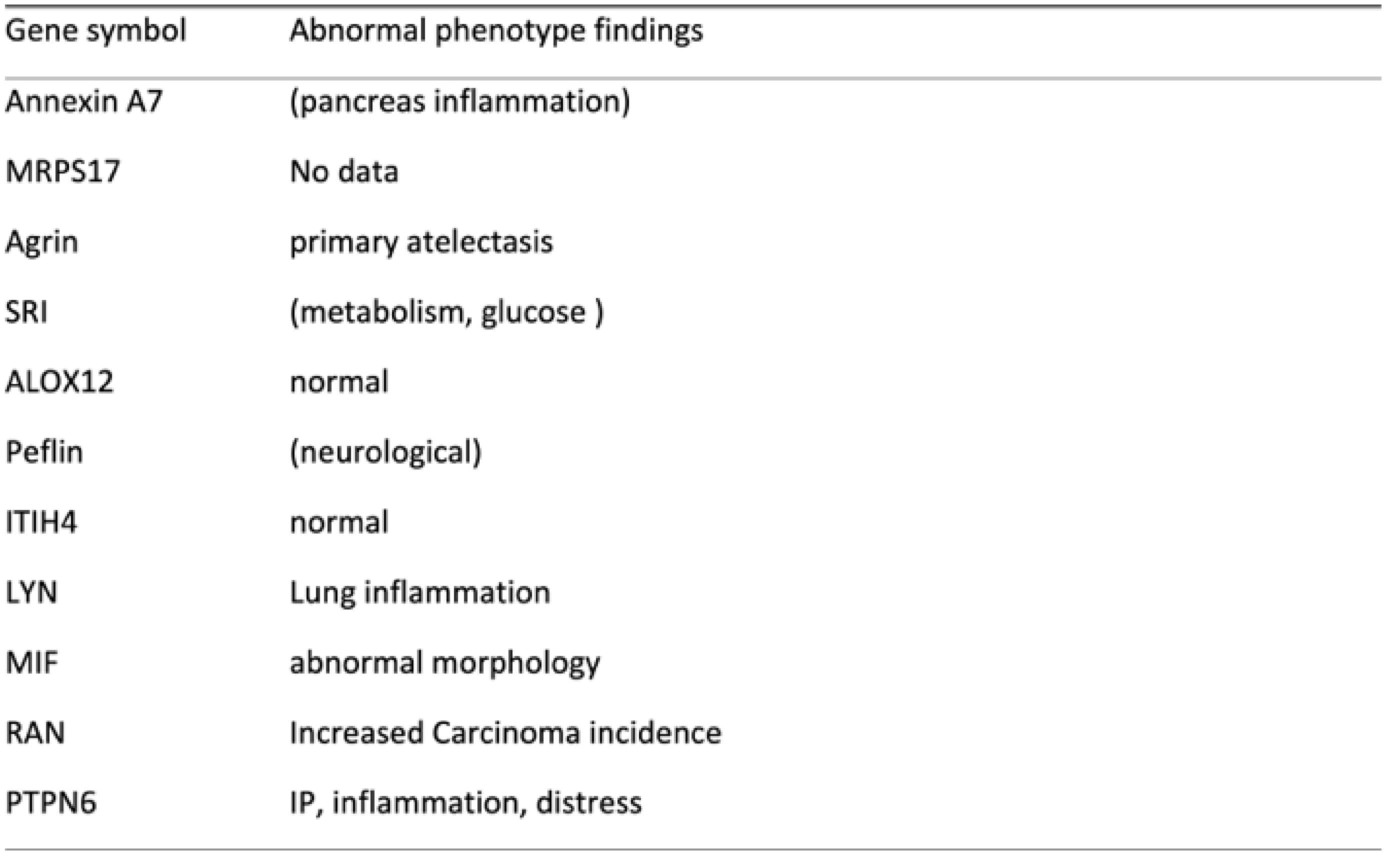
Network analysis and search for upstream control factors

In addition, we searched for molecular networks composed of seven core molecules and found pathways such as Carbohydrate Metabolism, Small Molecule Biochemistry, Cellular Assembly and Organization, where all these seven molecules are mapped. The regulatory relationships among the molecules on this network, including the seven core molecules, are depicted in Fig. 4c).

Moreover, the upstream regulatory relationships of the expression of the seven core molecules were explored using IPA causal network analysis. As shown in Fig. 4d), these molecules are regulated by molecules such as ESR1 (Estrogen receptor 1), CCND1 (Cyclin D1), CCR2 (C-C chemokine receptor type 2), NOS2 (Nitric oxide synthase 2), and MMP14 (Matrix metalloproteinase-14), which are in turn regulated by the SRC (Proto-oncogene tyrosine-protein kinase Src) family, ERK1/2 (Extracellular signal-regulated kinase 1/2) and ABL1 (ABL proto-oncogene 1) and finally ponatinib was identified as an upstream regulator.

LYN and PTPN6 knock out mice were reported to have abnormal phenotypes in the lung

The MGI database (http://www.informatics.jax.org/) and the JAXKO mouse phenotype database (https://www.jax.org/jax-mice-and-services) were used to search for KO mice and phenotypes of core and hub molecules, which are summarized in Table III. LYN and PTPN6 were found to have phenotypes such as inflammation in the lung. However, for other molecules, there are no available data or only effects on other organs have been reported.

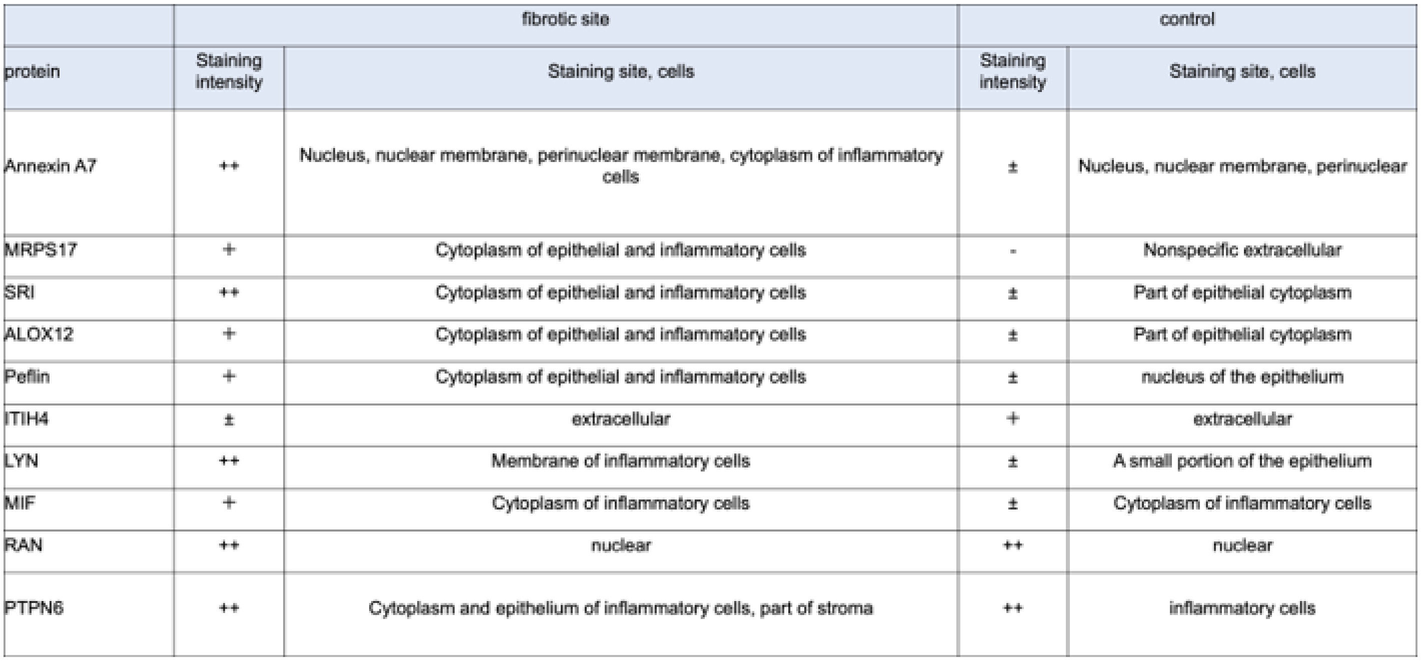


Gene symbol	Abnormal phenotype findings
Annexin A7	(pancreas inflammation)
MRPS17	No data
Agrin	primary atelectasis
SRI	(metabolism, glucose)
ALOX12	normal
Peflin	(neurological)
ITIH4	normal
LYN	Lung inflammation
MIF	abnormal morphology
RAN	Increased Carcinoma incidence
PTPN6	IP, inflammation, distress

Table III Phenotype list of KO mice for proteins found in subset binding

Immunohistochemical staining reveals many of the proteins are strongly upregulated expression in fibrotic areas, especially in epithelial cells and inflammatory cells

Of the 20 molecules presented by SB, 7 core proteins and 4 hub proteins were investigated for expression in patient lungs and for increased expression in fibrotic areas. The fibrotic and normal areas of the lungs of two IPF patients, who had concomitant cancer and were eligible for surgery, in the different cohort from that for proteome analysis were used for immunostaining using antibodies against each of the proteins. As a representative result, a clear enhancement of Lyn expression was observed in the tissues with obvious fibrosis confirmed by masson’s trichrome staining, shown in Fig. 5a). The results of immunohistochemical staining are summarized in Table IV, which shows that almost all the proteins except ITIH are upregulated in fibrotic areas, especially in epithelial cells and inflammatory cells.



	fibrotic site		control	
protein	Staining intensity	Staining site, cells	Staining intensity	Staining site, cells
Annexin A7	++	Nucleus, nuclear membrane, perinuclear membrane, cytoplasm of inflammatory cells	±	Nucleus, nuclear membrane, perinuclear
MRPS17	+	Cytoplasm of epithelial and inflammatory cells	-	Nonspecific extracellular
SRI	++	Cytoplasm of epithelial and inflammatory cells	±	Part of epithelial cytoplasm
ALOX12	+	Cytoplasm of epithelial and inflammatory cells	±	Part of epithelial cytoplasm
Peflin	+	Cytoplasm of epithelial and inflammatory cells	±	nucleus of the epithelium
ITIH4	±	extracellular	+	extracellular
LYN	++	Membrane of inflammatory cells	±	A small portion of the epithelium
MIF	+	Cytoplasm of inflammatory cells	±	Cytoplasm of inflammatory cells
RAN	++	nuclear	++	nuclear
PTPN6	++	Cytoplasm and epithelium of inflammatory cells, part of stroma	++	inflammatory cells

Table IV Summary of the results of immunohistochemical staining.

Ponatinib suppressed EMT

Epithelial mesenchymal transition (EMT) has been suggested to be important in the mechanism of pulmonary fibrosis in IPF. In this study, we succeeded in establishing a test system in which EMT is induced by TGF-b using human normal airway epithelial cells BEAS-2B, and the EMT inhibitory effect of ponatinib was confirmed (Fig. 6).

Discussion

When taking a data-driven approach using machine learning, both the quality and quantity of the data have a significant impact on the analysis results. There are many challenges to ensure both of them when utilizing disease phenotype data and biomolecular data.

In order to obtain informative knowledge from disease phenotype data by using machine learning, it is necessary to use data in a format that can be processed by machines. In order to promote the secondary use of medical information, it is important to construct a system that efficiently stores medical information as structured data while reducing the burden on clinicians, or a system that efficiently converts unstructured data into structured data. In this study, we attempted to standardize the input content of physicians and to facilitate the creation of structured data by creating a fixed format (a.k.a. template) of items that physicians consider important in advance. The physicians recorded most of the descriptions of medical interviews, examinations, and treatment results by selecting terms instead of free descriptions in order to minimize the differences in the input of electronic medical records such as fluctuations in terminology, which are often due to differences in the physicians in charge. The medical information collected before the introduction of the template was converted into structured data by manual curation while matching the format to the template. In addition, in this study, the interpretation reports of chest CT images and consultation records were used to directly incorporate the judgment of medical specialists into the analysis and to maintain high interpretability of the analysis results. We developed an automatic extraction system using natural language processing (NLP) to tag and attribute important words and phrases in the interpretation reports, which cannot be collected using the template. This attempt resulted in the creation of several resources such as annotation guideline in the field of medical language processing, including the respiratory field, where ontologies had not yet been developed (https://sociocom.naist.jp/real-mednlp/wp-content/uploads/sites/3/2021/12/Real-MedNLP_Annotation_Guidelines.pdf). On the other hand, in respiratory diseases, diagnoses are often made using images. In recent years, there has been a lot of research on AI diagnosis methods using image-processing techniques including 3D processing, and we are currently working on knowledge extraction from chest CT images instead of interpretation reports.

Data collection is a challenge not only for disease phenotype data including medical information but also for biomolecular data. Biomolecular data are not usually collected in routine clinical practice, except for blood test values, and there are many hurdles such as the cost of large-scale data collection, errors such as batch effects and noisiness that are often raised in omics analysis. In addition, selection of omics type (e.g. genome data, transcriptome data) to be linked to medical information is not trivial. In this study, we focused on proteins in serum EVs including exosomes, which are useful not only for their low invasiveness but also for their suitability as biomolecules that reflect the pathology of respiratory diseases. Exosomes are secreted by all types of cells in the body, including immune cells and tumor cells, and contain proteins, nucleic acids, and metabolites, and have been shown to function as new messengers that move between cells and organs from donor cells to recipient cells (Zhang, Li et al. 2015). Since the proteins and nucleic acids contained in exosomes are transferred to target cells, they have the potential to elucidate the pathogenesis of many diseases, including malignant diseases (Ludwig, Whiteside et al. 2019), immune diseases (Tan, Wu et al. 2016), and infectious diseases (Schorey and Harding 2016), as well as physiological conditions (L Isola and Chen 2017), and to be used for therapeutic applications (De Toro, Herschlik et al. 2015). For example, in cancer, exosomes have been shown to play an important role in various steps such as 1. metastasis, 2. immune modulation, 3. effects on peritumor fibroblasts and macrophages, 4. anticancer drug resistance, and 5. angiogenesis (Mashouri, Yousefi et al. 2019). We confirmed that the expressions of the proteins that linked to the major characters of typical IPF at phenotype level in this study were high in lung tissue, especially in fibrotic areas (Fig.5). This result strongly supports the relevance of these EV proteins to the pathogenesis, as well as that of our strategy of data-driven drug target discovery.

Although search for biomarkers by proteome analysis of serum has been attempted for many diseases, it has not been possible to identify minute amounts of proteins derived from tissues or lesion sites because many blood-derived proteins, which are present in large quantities in serum, are detected. In this study, we used EVs including exosomes to identify proteins that reflect changes in the lesion site encompassed in the EVs, and we believe that we have narrowed down the proteins that are strongly related to the pathological condition.

Even when structured medical information and omics data associated with it are collected, conventional machine learning methods do not always demonstrate their power depending on the objective of the analysis. Medical information often contains both of discrete values (e.g., items representing the presence or absence of smoking in binary form, items representing the progression of disease in terms of stages, and items representing the findings observed in CT images in terms of one-hot vectors) and continuous values (e.g., blood test values, respiratory function measurements). The existing methods are not appropriate for extracting many-to-many relationships by linking such data, a mixture of discrete and continuous values, with omics data. Therefore, a novel algorithm that can be applied to such data was required. Although clustering is commonly used for patient stratification, the characteristics commonly observed in patient subgroups defined by global similarity are not necessarily medically useful, because the majority of factors in medical information and omics data are considered to be not directly related to diseases. In addition, prior knowledge of the appropriate number of clusters (subgroups) is rarely available. Similarly, the strengths of multi-view learning, which assumes a common structure across multiple data sets, are not exploited. Furthermore, since the objective is to present biomarker candidates and drug target candidates through patient stratification, the algorithm must have high interpretability of output. The algorithm developed in this study, subset binding, is not limited to patient stratification. It extracts many-to-many items that are correlated between paired data without relying on existing knowledge. An example of applications other than patient stratification is available in the Supplementary method. Conventional approaches to identify biomarkers for patient stratification have always been based on significant differences in the rate of change relative to healthy controls, and have focused on simple statistical testing for each molecule, while our approach group several molecules to explain the combination of phenotypic characters. Our method has the advantage that it can be used even when there is limited information about the factors that are important for patient stratification, because there is no need to specify in advance groups (e.g. healthy control vs typical IPF patients) to be compared and no need to specify which metadata items (e.g. attributes in medical information) to focus on when extracting molecules of interest.

In order to validate the relevance of the IPF-related proteins we detected, we investigated the associations between these proteins and IPF using the QIAGEN database. It is noteworthy that we could detect proteins with known associations with IPF, as well as proteins with no previously reported associations with IPF were included (Sup Table III). The fact that our method detected ground truth supports the validity of our method. Furthermore, our strategy provides more opportunities to gain new insights since it uses real-world clinical information rather than a knowledge-based search. It is also expected that the IPF-related proteins in serum EVs we detected can be used to classify IPF and other IIPs without invasive techniques such as surgical lung biopsy (SLB).

It is striking that the proteins found in this study are not only linked to a common pathway, but also that ponatinib was identified as a drug that regulates this pathway. In order to verify the relevance of the proteins found in this study and ponatinib to IPF, we validated their impact on epithelial mesenchymal transition (EMT), which is involved in the mechanism of lung fibrosis in IPF. We established a TGF-b-induced EMT system using normal human bronchial epithelium-derived BEAS-2B cells in addition to A549 cells, which are commonly used for EMT. It was demonstrated that ponatinib possessed an inhibitory effect on EMT (Fig. 6). Nintedanib, one of the anti-fibrotic drugs prescribed for IPF patients, has been reported to inhibit EMT as one of its mechanisms of action. Therefore, ponatinib, which regulates IPF-related molecules, is also expected to suppress fibrosis through EMT inhibition in IPF patients, similar to nintedanib. In fact, nintedanib is a tyrosine-kinase inhibitor targeting VEGFR (vascular endothelial growth factor), PDGFR (platelet-derived growth factor) and FGFR (fibroblast growth factor), and ponatinib is also a multi-targeted tyrosine-kinase inhibitor. It has been reported that ponatinib inhibited the apoptosis of human type I alveolar epithelial cells, the proliferation of human lung fibroblasts *in vitro* and prevented fibrosis in a bleomycin-induced pulmonary fibrosis in rats by inhibiting TGF-b1/Smad3 pathway (Qu, Zhang et al. 2015). They also reported that ponatinib reversed the EMT in A549 cells, which is consistent with our result. Fibrosis occurs in many organs, such as the liver and skin, and there are many patients affected by this. In addition, it has recently been shown that fibrosis is associated with the development of cancer, and in fact, about 20% of IPF patients develop lung cancer, whose risk is equivalent to five times as high as healthy population (Tzouvelekis, Spagnolo et al. 2018). Therefore, the search for drug targets based on unraveling the true nature of fibrosis in IPF, including ponatinib, is expected to lead to the elucidation of therapeutic methods for fibrotic diseases in other organs and of the mechanisms of cancer development.

The main objective of this study is to develop a proof of concept for the feasibility of data-driven drug target discovery using medical information and omics data. Therefore, we focused on the detection of proteins associated with typical IPF features and their network analysis to present and validate drug target candidates. However, we emphasize that proteins associated with disease phenotypes that are different from the characteristics of typical IPFs are also detected in the output of SB. In other words, proteins associated with atypical IPFs and disease phenotypes of IIPs other than IPFs are also detected simultaneously, making it possible to stratify patients at the biomolecular level and search for drug targets in the resulting subgroups, rather than using conventional classification by disease type. This will lead to new drug discovery with a different mechanism of action from the two drugs already on the market. Furthermore, since the workflow of this research can be applied to diseases other than respiratory diseases, we expect that this research will accelerate the data-driven approach to drug target discovery, where the introduction of AI has not made much progress so far.

Methods

Ethics statement

Written informed consent was acquired from all patients before this study. The protocol of this study was approved by the Ethics Committee of NIBIO and Osaka University Hospital.

Protocol for collecting serum

Patients who were diagnosed or suspected with interstitial pneumonia including IPF at Osaka University Hospital were entered to this study. They were provided sufficient explanation based on the "Informed Consent Explanation Document", and then gave written consent to participate in this study. Ten mL of blood was collected and allowed to stand at room temperature for 1 hour, then centrifuged at 3000 rpm for 10 minutes, and the supernatant was separated as serum. The separated serum was immediately frozen and stored in a freezer at -80°C. Serum was also collected in the same manner for those who were diagnosed as having no organic respiratory disease as healthy control.

Protocol for proteome analysis

EV isolation and comprehensive protein measurements were performed according to the method described in (Muraoka, Hirano et al. 2022). Briefly, phosphatidylserine-positive extracellular vesicles were purified from 200 µl of serum using MagCapture isolation kit (Fujifilm Wako). Proteins in EVs were reduced with tris(2-carboxyethyl) phosphine, alkylated with iodoacetamide, trypsin digested and desalted. Pretreated samples were subjected to LC-MS/MS analysis using the Data independent acquisition (DIA) method. Data analysis was performed using DIA analysis software Spectranout, and run-wise imputation was performed for missing values. 1 commercial serum sample was added to every 15 samples as a quality control to assure quality from sample preparation to data analysis. DIA analysis of digested HeLa cells was also performed as a quality control for mass spectrometry.

Protocol for collecting medical information

Medical information securely stored in the data center of Osaka University Hospital was anonymized by patient ID and then stored in encrypted HD with the cooperation of the Medical Information Department of the Osaka University Hospital and provided to the National Institute of Biomedical Innovation (NIBIO). Medical examination records were obtained as structured data from the doctor using a template created with a list of 102 items of necessary information in advance, or by manually curating the template from free text data at the NIBIO. The CT imaging interpretation reports were tagged with key words using manual or natural language processing techniques, and were classified into site/lesion pairs and three categories: positive, negative, and suspect. Blood test values were structured by selecting and curating 173 key items. For the initial medical questionnaire and basic information, the key items were curated and added to the template items of the medical record. In structuring the data, we confirmed the meaning of missing values and used mainly the reference values for healthy subjects to impute missing values.

Protocol for reading findings NLP

Tagging Protocol

In order to assign tags to a range of expressions that appear in clinical texts such as consultation records and reading findings, which correspond to medical concepts such as names of diseases, disorders, and sites, a tag classification was performed as newly developed annotation guidelines

(URL:https://sociocom.naist.jp/real-mednlp/wp-content/uploads/sites/3/2021/12/Real-MedNLP_Annotation_Guidelines.pdf.)

Explanation of Knowledge Extraction Model

According to the tag classification described above, annotation guidelines were developed based on actual cases, and tagging and extraction of important words and phrases were performed in medical examination records and reading findings. The validity of the tagging was checked by experts with medical knowledge, and correct data was generated. Using the obtained corpus, we set up a medical expression recognition and relationship estimation system and constructed an extraction system using the Japanese BERT model.

Visualization of the proteome data

The proteome data was log-transformed (base:10) prior to visualization. The heatmap was created with *seaborn* python module (Waskom 2021) with the parameter settings as below: method='average', metric='cosine', z_score=1, standard_scale=None. For t-SNE and UMAP, the proteome data was further converted into z-score. The t-SNE was conducted with *scikit-learn* python module (Pedregosa, Varoquaux et al. 2011) with the parameter settings as below: n_components=2, perplexity=5, metric='cosine'. The UMAP was conducted with *umap* python module (McInnes, Healy et al. 2018) with the parameter settings as below: n_components=2, n_neighbors=5, metric='cosine'.

Subset binding

Advocating the idea of "subset-binding (SB)," which focuses on finding inter-related attributes in heterogeneous data according to their co-occurrence, this study developed a novel algorithm by extending fuzzy association rule mining techniques. Briefly, SB utilizes FARM (Fuzzy Association Rule Mining) approach to search for frequent itemsets (items that tend to occur) in two data (e.g. proteome data and medical information) and find association rules (patterns of co-occurrence between itemsets) so that the antecedent comes from one data and the consequent comes from the other data. The detailed algorithm is described in Supplementary methods. The proteome data which is linked with the medical information was analyzed with SB with the parameter setting as below: min support = 0.15, 0.02, 0.02, 0.02 for the proteome, the CT interpretation report, the medical records, and the blood test, respectively. min number of items = 4, 3, 3, 3 for the proteome, the CT interpretation report, the medical records, and the blood test, respectively. min lift = 2, 2, 2 for the association rules between proteome - CT interpretation report, the proteome – the medical records, and the proteome - blood test, respectively.

The analysis of protein-protein interaction using TargetMine

PPI networks for the top 20 proteins were constructed and network hubs were assigned using TargetMine (a data warehouse for drug discovery, https://targetmine.mizuguchilab.org) (Chen, Tripathi et al. 2011) (Chen, Tripathi et al. 2016) (Chen, Tripathi et al. 2019).

The network analysis and upstream regulators characterization using IPA

To identify biologically relevant molecular networks and pathways for the core and hub proteins, Ingenuity Pathways Analysis (IPA; QIAGEN, Redwood 185 City, CA), was used. We performed disease and pathway analyses and network generation using Ingenuity Knowledge Base, which relies on available publications describing the biological mechanisms, interactions and functions of proteins. The causal network analysis (CNA) (Krämer, Green et al. 2014) was performed using also IPA, for identifying novel master-regulators by creating pathways of literature-based relationships.

Immuno-histochemical staining

Fibrotic and normal areas of the lungs of 2 IPF patients who had concurrent lung cancer and were eligible for surgery at Osaka University Hospital were separated and fixed in formalin phosphate buffer solution.

Formalin-fixed lung tissue was cut along the longitudinal axis of the tissue section to prepare FFPE (formalin fixed paraffin embedded) blocks, which were thinned to 4 m in thickness using a sliding microtome to prepare unstained specimens. Unstained specimens were subjected to HE (hematoxylin-eosin) and MT (Masson's trichrome) staining to confirm inflammation and fibrosis. In addition, unstained specimens were stained by the Immunohistochemistry (IHC) method using specific antibodies for the proteins found in the analysis. The OptiView DAB Universal Kit was used as the detection reagent, and staining using antibody dilutions as a negative control for each antibody was performed simultaneously to evaluate IHC staining for each antibody.

EMT

EMT was evaluated by suppressing the expression of the epithelial marker E-cadherin and enhancing the expression of the mesenchymal markers Fibronectin and Snail. 2.5×10^4 or 3.5×10^4 cells/mL of BEAS-2B cells purchased from ATCC (The cells were seeded in 96 well plates coated with fibronectin/collagen I/BSA in BEGM™ Bu, lletKit™ ; Bronchial epithelial cell basal medium, Lonza Corporation, and after 24 hours, the test drug was added. 48 hours after TGF-β addition, cell lysis and RT were performed using the SuperPrep® II Cell Lysis & RT Kit for qPCR (Toyobo Co., Ltd.), and the expression of the above EMT marker gene group was measured. THUNDERBIRD Probe qPCR Mix (Toyobo Co., Ltd.) was used for qPCR, and TaqMan Probe (Thermo Fisher Scientific Co., Ltd.) was used for each marker probe.

Declarations

Acknowledgements

This study was supported by the Ministry of Health, Labour and Welfare [grant number JPMH20AC5001 to Y.N.K. and grant number 19AC5001 to K.M.], Cabinet Office of Japan Government for the Public/Private R&D Investment Strategic Expansion PrograM (PRISM), and the Japan Society for the Promotion of Science (JSPS) KAKENHI [grant number JP18H05282 to A.K.].

The authors thank Yoichi Kurebayashi (Kobe University, Kobe, Japan), Masanori Shindo (NIBIOHN, Osaka, Japan), Yoshinori Nakamura (NIBIOHN, Osaka, Japan), Hiromitsu Kageyama (NIBIOHN, Osaka, Japan), Michiyo Kawai (NIBIOHN, Osaka, Japan), Akiko Fukagawa (NIBIOHN, Osaka, Japan), Hideyo Kamada (NIBIOHN, Osaka, Japan), Miki Yamada (NIBIOHN, Osaka, Japan) for supporting this research project. We also thank Yuji Kosugi (Lifematics Ltd., Tokyo, Japan) for his technical support.

Author contributions

Y.N.K., M.N.I., Y.T., J.A., A.K., K.M., and N.U. conceived the study. Y.T., H.H., K.M., T.S., Y.S., Y.N., Y.A., T.E., A.A. collected samples from IP patients. T.M. collected medical information of IP patients. R.N., S.M., and T.T. performed proteome analysis. M.N.I., M.K., Y.N., T.F. created structured medical data for analysis. S.K., F.C., R.T., S.Y., E.A., S.W. created structured data from CT interpretation report by NLP. C.H. and C.N. set up and managed computing environment. Y.N.K. and N.U. developed the algorithm (subset binding). Y.N.K., M.N.I and Y.A.C performed data analysis. All authors contributed to the writing of the manuscript.

Data availability

Data deposition: The proteome data obtained in this study have been deposited to jPOST database (https://globe.jpostdb.org/) with the accession number of PXID042707.

Patents

Y.N.K. and N.U. have a patent application on the subset-binding algorithm. Y.N.K., M.N.I., M.K., K.M., J.A., T.T., Y.T., A.K., Y.M. and N.U. have a patent application on the screening method of drug target, the IPF-related proteins and their inhibitors.

Conflicts of Interest

None declared.

References

Arrowsmith, J. and P. Miller (2013). "Trial watch: phase II and phase III attrition rates 2011-2012." Nature reviews. Drug discovery **12**(8): 569.

Chen, Y.-A., et al. (2019). "The TargetMine data warehouse: Enhancement and updates." Frontiers in genetics: 934.

Chen, Y.-A., et al. (2011). "TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery." PloS one **6**(3): e17844.

Chen, Y.-A., et al. (2016). "An integrative data analysis platform for gene set analysis and knowledge discovery in a data warehouse framework." Database **2016**.

De Toro, J., et al. (2015). "Emerging roles of exosomes in normal and pathological conditions: new insights for diagnosis and therapeutic applications." Frontiers in immunology **6**: 203.

Futami, Y., et al. (2021). "CD14 and LBP as Novel Biomarkers for Sarcoidosis by Proteomics of Extracellular Vesicles."

Hoshino, A., et al. (2020). "Extracellular vesicle and particle biomarkers define multiple human cancers." Cell 182(4):1044-1061

Koba, T., et al. (2021). "Proteomics of serum extracellular vesicles identifies a novel COPD biomarker, fibulin-3 from elastic fibres." ERJ Open Research **7**(1).

Krämer, A., et al. (2014). "Causal analysis approaches in ingenuity pathway analysis." Bioinformatics **30**(4): 523-530.

L Isola, A. and S. Chen (2017). "Exosomes: the messengers of health and disease." Current neuropharmacology **15**(1): 157-165.

Ludwig, N., et al. (2019). "Challenges in exosome isolation and analysis in health and disease." International journal of molecular sciences **20**(19): 4684.

Mashouri, L., et al. (2019). "Exosomes: composition, biogenesis, and mechanisms in cancer metastasis and drug resistance." Molecular cancer **18**(1): 1-14.

McInnes, L., et al. (2018). "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426.

Muraoka, S., et al. (2022). "Comprehensive proteomic profiling of plasma and serum phosphatidylserine-positive extracellular vesicles reveals tissue-specific proteins." iScience: 104012.

Njock, M.-S., et al. (2019). "Sputum exosomes: promising biomarkers for idiopathic pulmonary fibrosis." Thorax **74**(3): 309-312.

Pedregosa, F., et al. (2011). "Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12."

Qu, Y., et al. (2015). "Ponatinib ameliorates pulmonary fibrosis by suppressing TGF-β1/Smad3 pathway." Pulmonary Pharmacology & Therapeutics **34**: 1-7.

Schorey, J. S. and C. V. Harding (2016). "Extracellular vesicles and infectious diseases: new complexity to an old story." The Journal of clinical investigation **126**(4): 1181-1189.

Tan, L., et al. (2016). "Recent advances of exosomes in immune modulation and autoimmune diseases." Autoimmunity **49**(6): 357-365.

Tzouvelekis, A., et al. (2018). "Patients with IPF and lung cancer: diagnosis and management." The Lancet Respiratory Medicine **6**(2): 86-88.

Valadi, H., et al. (2007). "Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells." Nature cell biology **9**(6): 654-659.

Waskom, M. L. (2021). "Seaborn: statistical data visualization." Journal of Open Source Software **6**(60): 3021.

Zhang, J., et al. (2015). "Exosome and exosomal microRNA: trafficking, sorting, and function." Genomics, proteomics & bioinformatics **13**(1): 17-24.

Figures

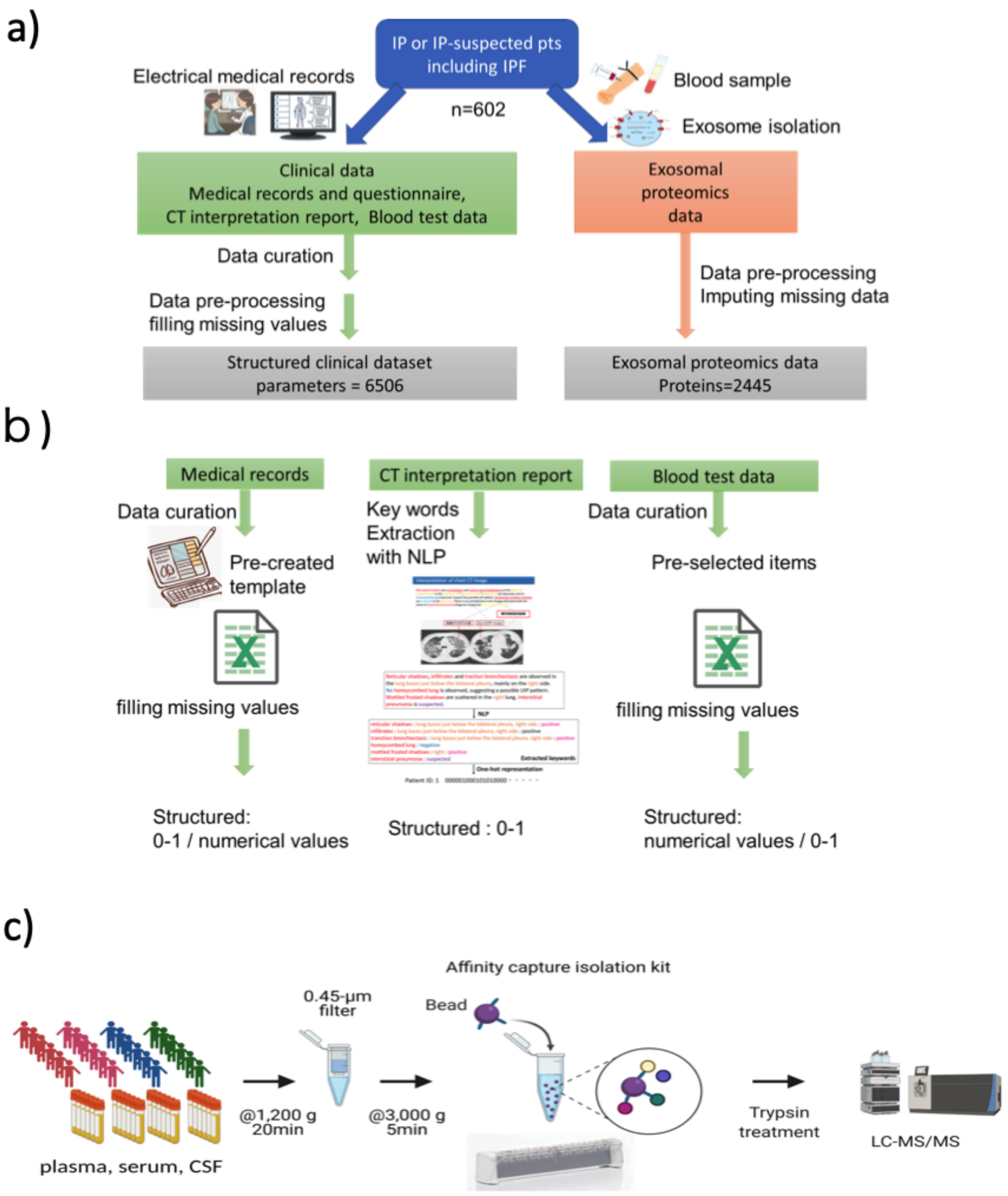
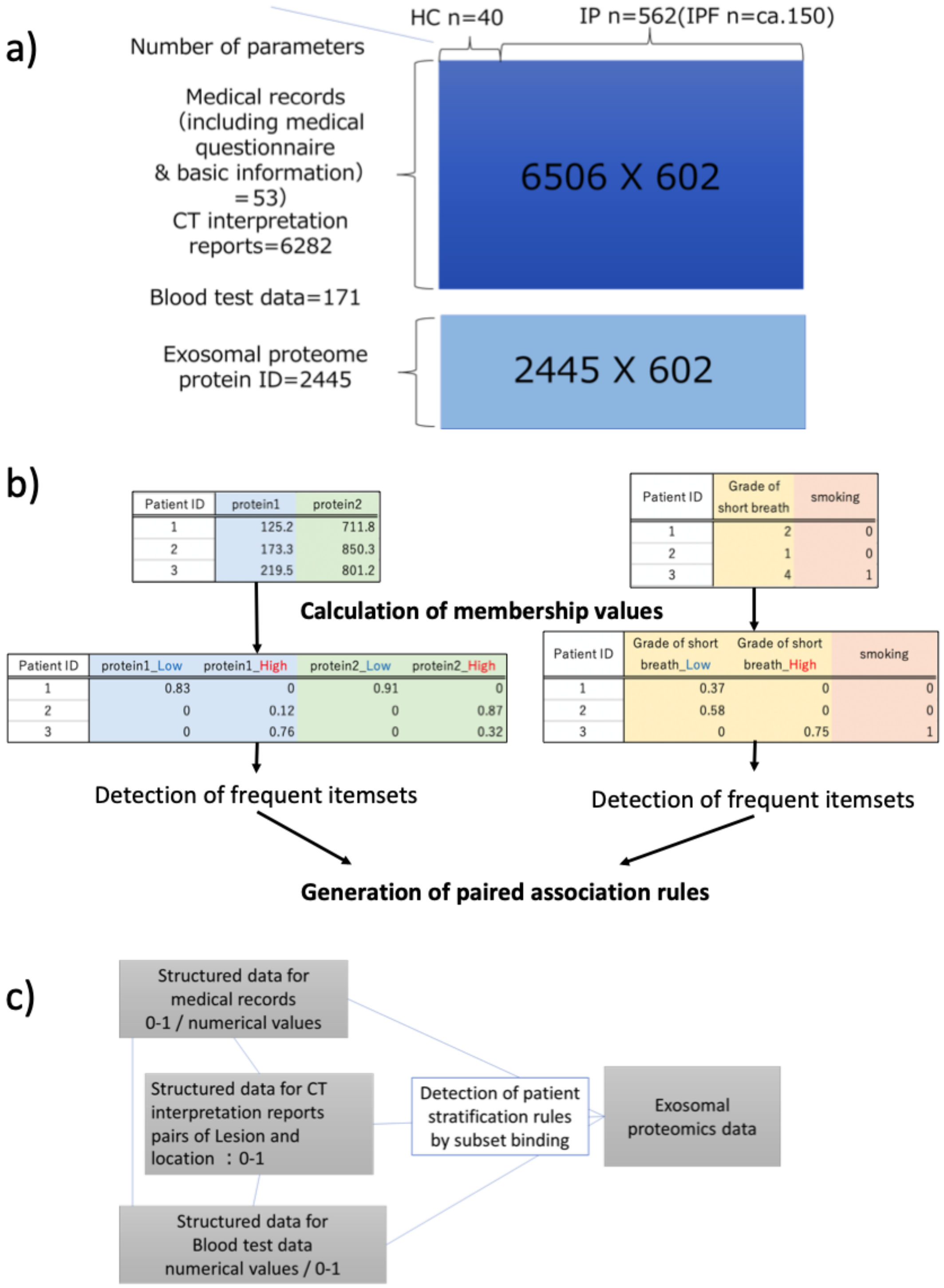

a)
IP or IP-suspected pts including IPF
n=602
Electrical medical records
Blood sample
Exosome isolation
Clinical data
Medical records and questionnaire,
CT interpretation report, Blood test data
Exosomal proteomics data
Data curation
Data pre-processing
filling missing values
Data pre-processing
Imputing missing data
Structured clinical dataset
parameters = 6506
Exosomal proteomics data
Proteins=2445
b)
Medical records
Data curation
Pre-created template
filling missing values
Structured:
0-1 / numerical values
CT interpretation report
Key words Extraction with NLP
Interpretation of chest CT image
Reticular shadows, infiltrates and traction bronchiectasis are observed in the lung bases just below the bilateral pleura, mainly on the right side.
No honeycombed lung is observed, suggesting a possible UIP pattern.
Mottled frosted shadows are scattered in the right lung. Interstitial pneumonia is suspected.
NLP
reticular shadows : lung bases just below the bilateral pleura, right side : positive
infiltrates : lung bases just below the bilateral pleura, right side : positive
tranction bronchiectasis : lung bases just below the bilateral pleura, right side : positive
honeycombed lung : negative
mottled frosted shadows : right : positive
interstitial pneumonia : suspected
Extracted keywords
One-hot representation
Patient ID: 1 000001000101010000
Structured : 0-1
Blood test data
Data curation
Pre-selected items
filling missing values
Structured:
numerical values / 0-1
c)
plasma, serum, CSF
0.45-μm filter
@1,200 g 20min
@3,000 g 5min
Affinity capture isolation kit
Bead
Trypsin treatment
LC-MS/MS

Figure 1

a) Flowchart of medical information and serum sample collection b) Conceptual diagram of medical information data generation, c) Exosome isolation and proteome analysis


a)
HC n=40
IP n=562(IPF n=ca.150)
Number of parameters
Medical records (including medical questionnaire & basic information) =53)
CT interpretation reports=6282
Blood test data=171
6506 X 602
Exosomal proteome protein ID=2445
2445 X 602
b)
Calculation of membership values
Detection of frequent itemsets
Detection of frequent itemsets
Generation of paired association rules
c)
Structured data for medical records 0-1 / numerical values
Structured data for CT interpretation reports pairs of Lesion and location ：0-1
Structured data for Blood test data numerical values / 0-1
Detection of patient stratification rules by subset binding
Exosomal proteomics data

Patient ID	protein1	protein2
1	125.2	711.8
2	173.3	850.3
3	219.5	801.2

Patient ID	Grade of short breath	smoking
1	2	0
2	1	0
3	4	1

Patient ID	protein1_Low	protein1_High	protein2_Low	protein2_High
1	0.83	0	0.91	0
2	0	0.12	0	0.87
3	0	0.76	0	0.32

Patient ID	Grade of short breath_Low	Grade of short breath_High	smoking
1	0.37	0	0
2	0.58	0	0
3	0	0.75	1

Figure 2

a) Structured data dimension. The medical information matrix is constituted with 6,282 attributes from CT interpretation reports, 171 attributes from blood test results and 53 attributes from medical records. The proteome matrix is constituted with 2,445 proteins (2,388 proteins were mapped to known protein IDs as shown in Sup Table II). Both matrices have 602 rows, which were equivalent to the number of cases (602 cases from 401 IP patients and 38 healthy controls). b) conceptual diagram of subset binding. Input data: two paired matrices (quantitative and/or categorical). First, quantitative attributes in input are converted into fuzzy categorical attributes (“Low” and “High”) with membership functions. Since membership values for “Low” and “High” categories are obtained for each attribute, this process produces matrices with double number the of columns when all attributes are quantitative. Next, these matrices are used to detect frequent itemsets independently. Thereafter, association rules are generated so that FIS derived from one matrix will be antecedent, and those from the other matrix will be consequent. User-specified threshold (e.g., lift) is used for pruning and paired (antecedent from data1 and consequent from data2) association rules are obtained as output. c) conceptual diagram of analysis workflow. Since subset binding accepts two paired matrices, the possible combinations of the data analysis were: i) proteome-medical records, ii) proteome-CT interpretation reports, iii) proteome-blood test, iv)medical records-CT interpretation reports, v) CT interpretation reports-blood test and vi)blood test-medical records. The IPF-related proteins were selected from these outputs.

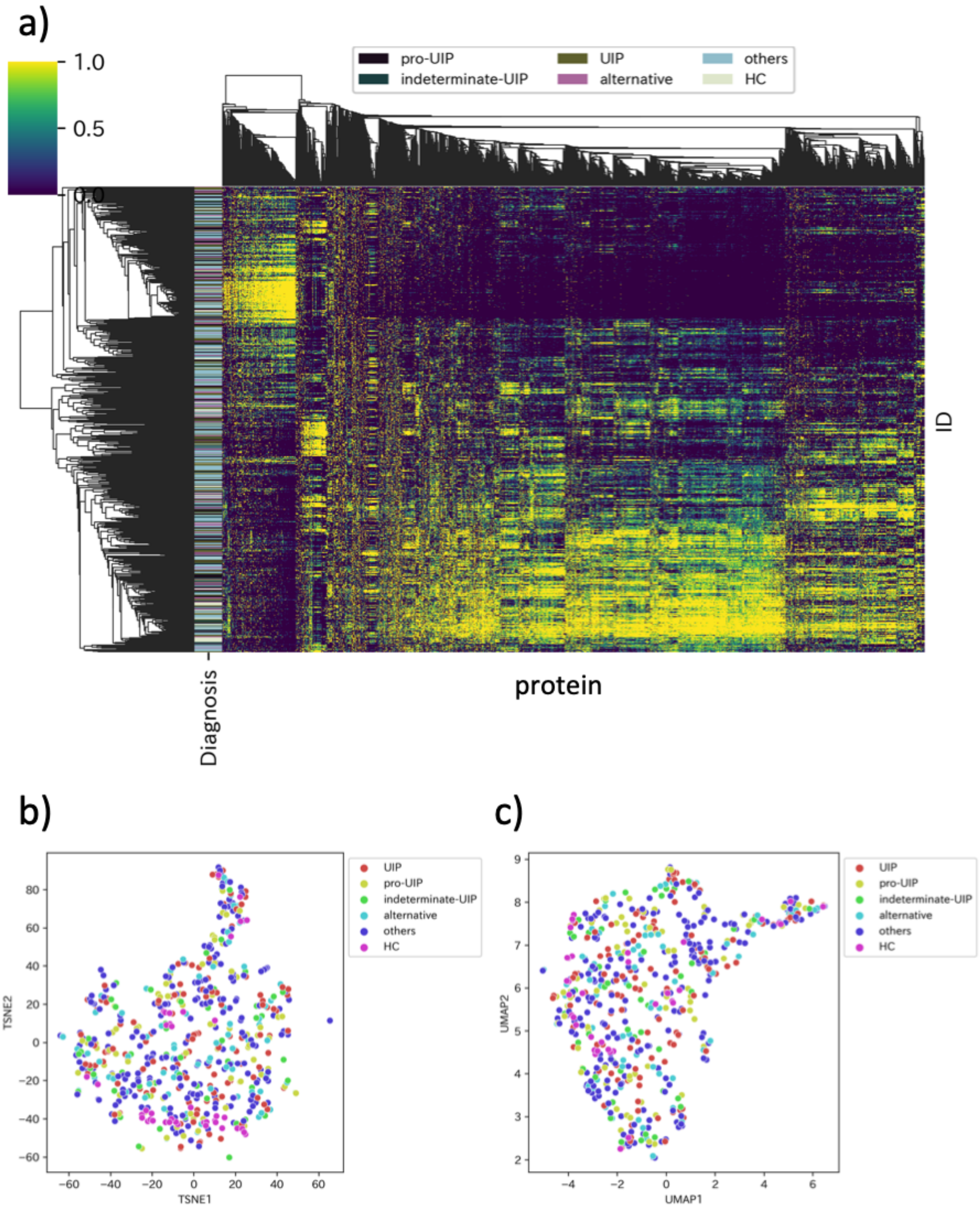


Figure 3

Visualization of the proteome data a) heat map with hierarchical clustering. The log-transformed protein amounts were scaled for each column. The x-axis represents proteins detected by DIA (n=2,445), and the y-axis represents the cases (n=602). The diagnosis of the cases were represented as six different colors as shown at the top of the figure. b) t-SNE. The log-transformed and scaled protein amounts were plotted (metric: cosine, perplexity: 5), c) UMAP. The log-transformed and scaled protein amounts were plotted (metric: cosine, perplexity: 5)

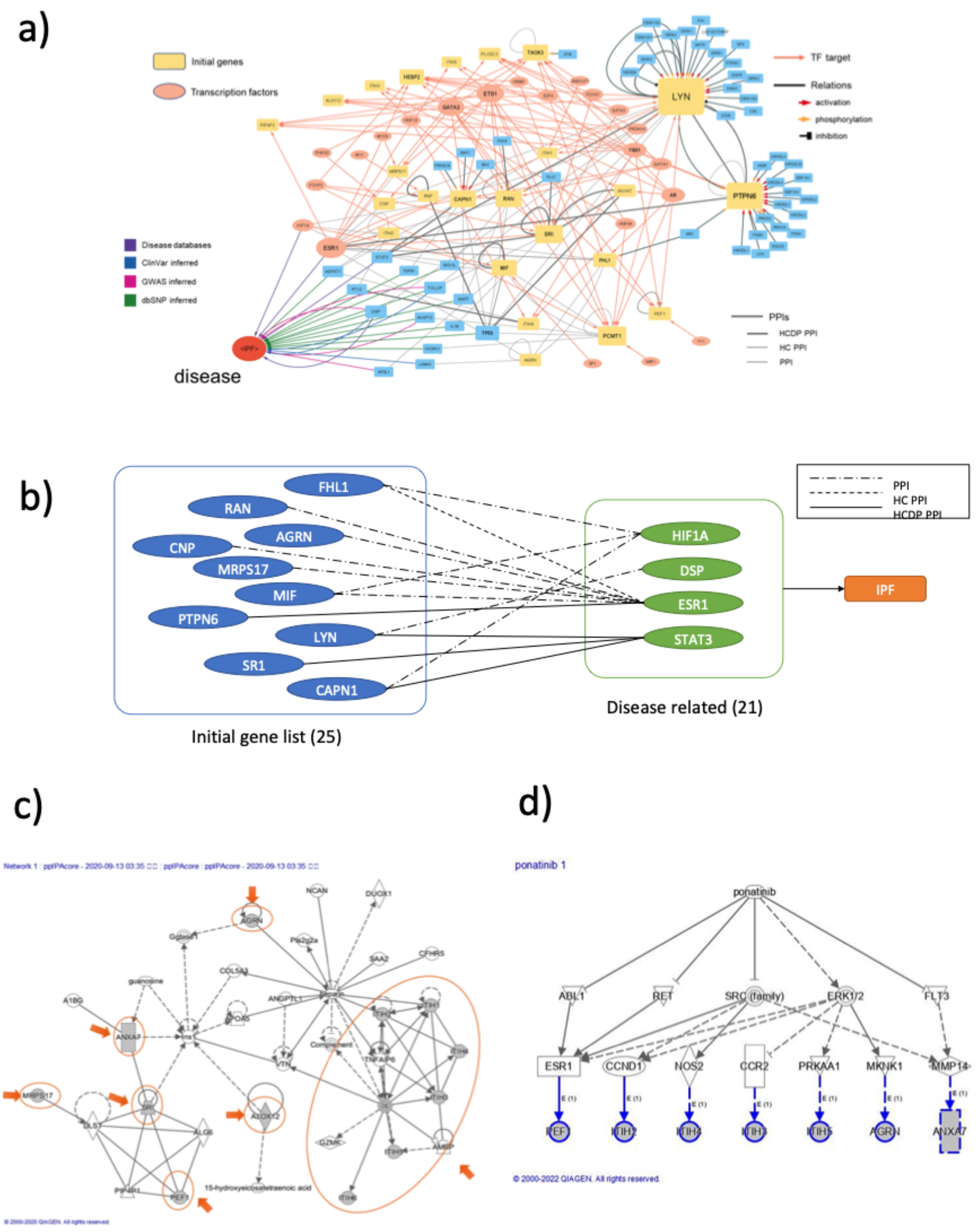


Figure 4

Proposed molecules information collected using TargetMine and IPA

a) PPI and hub molecules of proteins found TargetMine

b) Relationship with IPF TargetMine

c) Network extraction consisting of core molecules IPA

d) Upstream analysis of core molecule and relationship network of ponatinib

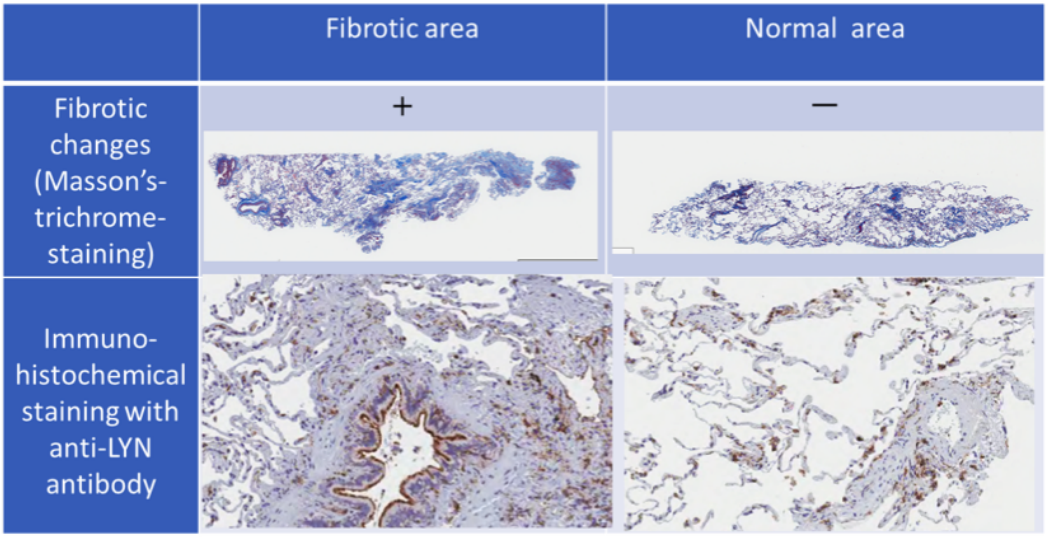

Fibrotic area
Normal area
Fibrotic changes (Masson's-trichrome-staining)
+
—
Immuno-histochemical staining with anti-LYN antibody

Figure 5

Confirmation of expression of key proteins in fibrotic lesions (immunostaining, independent cohort)

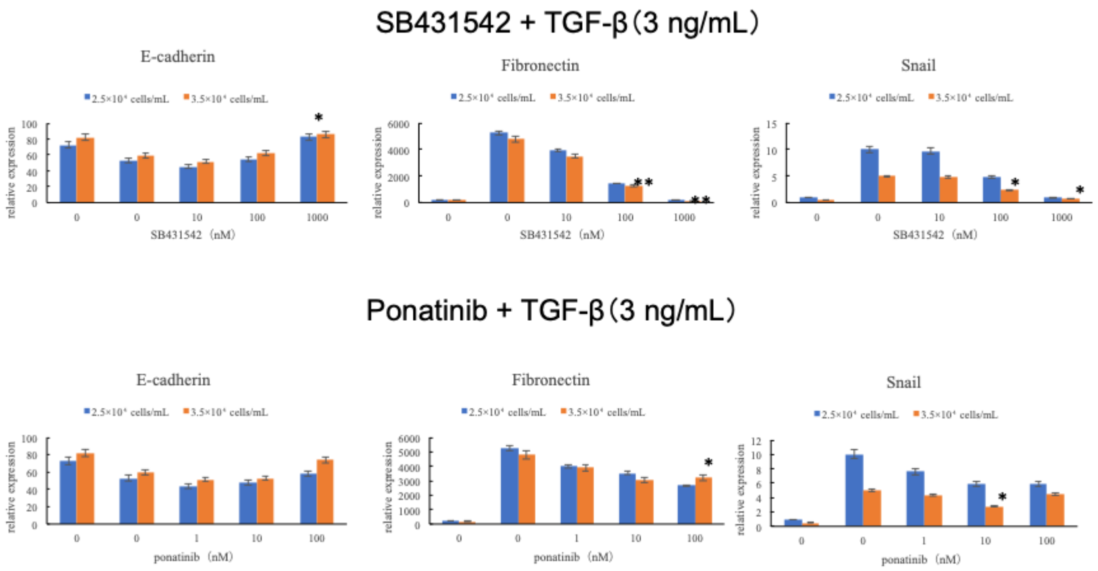


Figure 6

Ponatinib treatment attenuated TGF-β-induced expression of EMT genes/markers

Cells were treated with Ponatinib or SB431542, thereafter rTGF-β added to the cells and cultured for 48 h. Further, cells were subjected to RT-qPCR for the specified transcripts in a)E-cadherin, b)Fibronectin, c)Snail ; relative expression levels were presented as means ± SEM; significance *p<0.05, **p<0.01 vs TGF-βalone.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymethods.zip
- SupplementaryTables.pptx