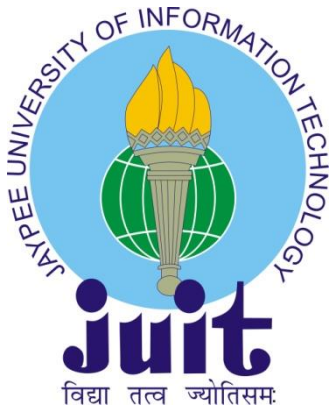
COMPREHENSIVE KNOWLEDGE GENERATION FOR RADIOLOGY AND GENETICS OF INTERSTITIAL LUNG DISEASES (ILDs)

Thesis submitted in fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

By

SMRITI MISHRA

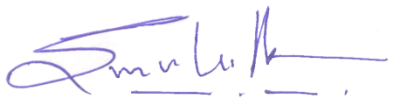


DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT, SOLAN, HP, INDIA
NOVEMBER, 2019

TABLE OF CONTENTS

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled "**Comprehensive Knowledge Generation for Radiology and Genetics of Interstitial Lung Diseases (ILDs)**" submitted at **Jaypee University of Information Technology, Waknaghat, Solan (HP), India** is an authentic record of my work carried out under the supervision of **Dr. Jayashree Ramana**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. thesis.



Smriti Mishra

Enrollment No. 136505

Department of Biotechnology & Bioinformatics

Jaypee University of Information Technology, Waknaghat, Solan (HP), India

Date:

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled **"Comprehensive Knowledge Generation for Radiology and Genetics of Interstitial Lung Diseases (ILDs)"**, submitted by **Smriti Mishra** at **Jaypee University of Information Technology, Waknaghat, Solan (HP), India** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

Dr. Jayashree Ramana
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology, Waknaghat, Solan (HP), India
Date:

ACKNOWLEDGEMENT

In the endeavor of the journey towards my Ph.D., this thesis has been kept on track and seen through to completion with the support and encouragement of numerous people including my mentors, well-wishers, friends, colleagues and various laboratories. It is a pleasant opportunity to express my heartfelt thanks to all those people who made this thesis possible and contributed in different ways.

My first debt of gratitude must go to Late **Dr. Chittaranjan Rout**. He was assigned as my first Ph.D. supervisor. His support and guidance throughout the research work has been incomprehensible. Dr. Rout always has been the first and foremost constructive critic for every idea we came up to consider as a research objective. He always had been there for us through thick and thin. His dedication towards work has been remarkable in his entire career.

I would like to express my heart full thanks to my supervisor, **Dr. Jayashree Ramana**. The insuperable contribution and peer support rendered by her is incomprehensible. She patiently provided me with the encouragement and advice necessary for me to proceed through the doctoral program and complete my dissertation. I want to thank **Dr. Ramana** for her unflagging encouragement and caring supports she has provided me whenever required.

I wish to thank my parents, **Mr. Ramchran Mishra** and **Mrs. Sarita Mishra,** and **Mrs. Ananti Sharma (Amma)**and my sister **Ms. Vijaya Mishra**. Their love, support and motivation provided me inspiration and encouragement, and they were the driving force to successfully complete my work. I am fortunate to have friends who have always stood beside me. I owe them everything and wish I could show them just how much I love and appreciate them.

I emphatically express my sincere thanks to JUIT administration, **Prof (Dr) Vinod Kumar (Vice Chancellor), Prof (Dr) S.D. Gupta (Director & Dean, Academics & Research), Maj. Gen (Retd.) Rakesh Bassi (Registrar), Prof (Dr) Y. Medury (Former COO, JES), Prof (Dr) S.K. Kak (Former vice-chancellor), Brig (Retd.) K.K. Marwah (Former Registrar), Brig (Retd.) Balbir Singh (Former Registrar), Prof (Dr) T.S. Lamba (Former Dean, Academic & Research), Prof. (Dr.) R.S. Chauhan (Former Head, Department of Biotechnology and Bioinformatics)** and **Prof. (Dr.) Sudhir Kumar (Head, Department of Biotechnology and Bioinformatics)** for providing an opportunity to pursue a Doctorate

Degree, teaching assistantship and advanced lab infrastructure to accomplish this scientific venture of my life.

I am also highly beholden to **Prof (Dr) Malay Sarkar, Prof Sanjeev Sharma, Indira Gandhi Medical College, Shimla** for their guidance, valuable suggestions and tremendous cooperation. I am also thankful to **Mr. Vinod Chauhan**, PG students and Technical Staff of Pulmonary Medicines and Radio-diagnosis Departments, Indira Gandhi Medical College, Shimla for their support in collection and interpretation of real-time patients' data.

I am also thankful to **Prof. (Dr.) Satya Prakash Ghrera, Prof. (Dr.) Ghanshyam Singh, Dr. Tiratha Raj Singh and Dr. Rahul Shrivastava** for their guidance and valuable suggestions throughout my research work. I wish to convey my sincere thanks to all the faculty members of Department of Biotechnology and Bioinformatics, for their help and guidance at the various stages of this study.

I am also thankful to all the members of technical and non-technical staff of the department, especially **Mrs. Somlata Sharma, Mr. Baleshwar, Mrs. Sonika, Mrs. Mamta and Mr. Ismail** for their assistance and valuable contributions.

It is my pleasure to express my gratitude to all research scholars of the Biotechnology & Bioinformatics Department for keeping me blessed with best wishes. I extend my heartfelt thanks to all my friends for their sustained support and ever needed cooperation. I would like to express my heartfelt gratitude to all those who have contributed directly or indirectly towards obtaining my doctorate degree and apologize if have missed out anyone.

Last, but not the least, I thank the one above all of us, omnipresent God, for answering my prayers, for giving me the strength to plod on during each and every phase of my life.

All may not be mentioned, but no one is forgotten.

Smriti Mishra

LIST OF ABBREVIATIONS

Abbreviations	Full Form
AIP	Acute Interstitial Pneumonia
BAL	Bronchoalveolar lavage
CAD	computer–aided diagnosis
CDSS	clinical decision support system
ceRNA	Competing endogenous RNA
ChILD	Children's Interstitial Lung Disease
COP/BOOP	Cryptogenic organizing pneumonia
COPD	Chronic Obstructive Pulmonary Disease
CPEF	Combined Pulmonary Fibrosis and Emphysema
CTD-ILD	Connective Tissue Disorder Related-ILDs
CVD-ILDs	Collagen vascular disease associated interstitial lung diseases
CXR	Chest X-Ray
DCGs	Disease candidate genes
DICOM	Digital Imaging and Communications in Medicine
DIP	Desquamative interstitial pneumonia
DPLD	Diffuse parenchymal lung disease
EHR	Electronic health record
FVC	Forced vital capacity
GERD	gastroesophageal reflux disease
GGO	ground glass opacity
GIP	Giant-cell interstitial pneumonia
GO	Gene ontology
HC	Honey-combing
HP	Hypersensitivity pneumonitis
HRCT	High-resolution computed tomography
IIP	Idiopathic Interstitial Pneumonia
ILD	Interstitial lung disease
ILDsi	ILDs specificity index

ILST	intra and interlobular septal thickening
IPF	Idiopathic pulmonary fibrosis
LAM	Lymphangioleiomyomatosis
LCH	langerhans cell histiocytosis
LIP	Lymphocytic interstitial pneumonia
lncRNA	Long non-coding RNAs RNA
MCTD-ILDs	Mixed connective tissue disease - ILDs
MDD	Multidisciplinary discussion
miRNA	microRNAs
ncRNA	noncoding RNAs
NSIP	Nonspecific interstitial pneumonia
PACS	Picture archiving and communication system
PAM	Pulmonary alveolar microlithiasis
PAP	Pulmonary alveolar proteinosis
PFT	Pulmonary function testing
pLCH	Pulmonary Langerhans Cell Histiocytosis
PS	Pulmonary Sarcoidosis
pSD	Pulmonary Surfactant dysfunction
RA-ILD	Rheumatoid arthritis-ILD
RB-ILD	Respiratory bronchiolitis interstitial lung disease
ROC	receiver operating characteristics
ROI	Region of interest
SLB	Surgical lung biopsy
SLE	Systemic lupus erythematosus
SNP	Single-nucleotide polymorphism
SSc	Systemic sclerosis
SSc-ILDs	Scleroderma-associated ILDs
TB	Tuberculosis
UIP	Usual interstitial pneumonia

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Interstitial lung diseases (ILDs) are the diverse group of about 200 pulmonary disorders (acute and chronic), which can cause inflammation and/or fibrosis to variable extent. ILDs cases are linked with considerable morbidity and mortality due to the delayed diagnosis or misdiagnosis. Clinical symptoms and radiological tests such as 2D chest radiograph (CXR) and 3D high resolution computed tomography (HRCT) hold more promises in developing countries due to their availability. Previous studies revealed the potential roles of genetic alterations of disease candidate genes (DCGs) in association with miRNAs and/or SNPs in ILDs pathogenesis. The non-coding RNAs, especially long-noncoding RNAs (lncRNAs) and micro RNAs (miRNAs) work as regulatory molecules, which are important for disease diagnostics and/or monitoring. The cross-regulation between these molecules for competing-endogenous RNAs (ceRNAs) activity is also important in ILDs pathogenesis. In the first objective of this research work, the focus has been given to the radiological assessment of ILDs to facilitate effective diagnosis through comparative analysis of clinical and radiological patterns. A comprehensive web-based resource of ILDs (ILD-DB) has been established for the radiological assessment of ILDs. Radiological (HRCT and CXR) and clinical profiles, and follow-up data from eleven ILD-subtypes including, a rare disease, pulmonary alveolar microlithiasis (PAM) were anonymized and incorporated. Dominant radiological patterns and the most reported diseases mainly from North-Western Himalayan region are presented in this study. All the data is freely accessible as a public repository in the healthcare domain (http://14.139.240.55/ilddb/). This study would be beneficial for characterizing new-onset ILDs, benchmarking of referral dataset, image processing algorithm development and radiological training. In the second objective, the first genetic knowledgebase of ILDs, ILDgenDB, is developed to provide ILDs associated genes, their genomic information, and integrated analyses. This resource would help to identify the role of genetics in disease pathogenesis and identification of diagnostics-based biomarkers. The catalog of DCGs along with associated regulatory molecules and processes are produced using an integrated pipeline of literature curation, databases searches and *In-silico* bioinformatics analyses. The DCGs' association with SNPs, miRNAs and pathways involved in ILDs and/or its subtypes' pathogenesis are also provided. These integrated analyses recommended that defense mechanisms and immune system are potentially involved in ILDs pathogenesis. This knowledge can assist in genomic level disease diagnosis and monitoring. ILDgenDB is publicly available at https://14.139.240.55/ildgendb. In the third objective, the role of ncRNAs as therapeutic/biomarker targets is investigated. The ncRNAs (miRNAs and lncRNAs) with higher interactions to target DCGs, biological pathways and ceRNAs are proposed as potential candidate biomarkers. This study suggested the potential role of ncRNAs and their target interactions in disease pathogenesis and may serve as less/non-invasive molecular biomarkers for ILDs diagnosis and monitoring. All the studies have been performed to address the current limitations in the diagnostics and therapeutics of the ILDs. This work may assist researchers, pulmonologists and radiologists working for the improvement in the diagnostics and therapeutics of ILDs.

CHAPTER 1

INTRODUCTION

1.1 Background

Interstitial lung diseases (ILDs), also known as diffuse parenchymal lung disease, are a diverse group of chronic pulmonary disorders that mainly involves interstitial tissue of the lung [1]. ILDs may cause considerable morbidity and mortality and recognized as a crucial health issue [2-4]. ILDs major characteristics features include inflammation and thickening of the interstitium and lung parenchyma. The interstitium represents the area from the epithelium of the alveolar sac to the capillaries' endothelium [5]. Mostly, the inflammation and thickening in interstitium lead to the irregular lung function [1]. Lung parenchyma comprises of alveolar walls, which consist of collagen, capillary endothelium, elastin and interstitium [5]. Thickening could be because of irritation, scarring or additional liquid (edema) [6]. ILDs believe to be an autoimmune disorder and may occur as an abnormal healing response triggered by lung injury. When the lung injury occurs, the repair damage mechanism goes askew, which leads to the scarred and thickened interstitial tissue [5]. The thickened walls make the blood-oxygen transfusion difficult by disabling full expansion of lung with air. Because of the restriction of lungs from fully expanding, these diseases also referredas "Restrictive lung diseases". Other similar class of disease also exists which characterized by shortness of breath with exertion [5].

Nearly 200 etiologic factors have been identified for ILDs [7]. The cause of the ILDs may not always be apparent, and nearly 70% of ILDs are idiopathic ("of unknown origin") in nature [6]. However, few of the known causes can be an environmental factor, exposure to a certain type of dust or allergens (silica, asbestos, coal etc) and/or autoimmune (or rheumatological) diseases. Furthermore, a small percentage of ILDs patients were reported due to few drug side-effects (eg. amiodarone, nitrofurantoinand other drugs) over the long period of time. Several unique forms of ILD not only exist in adults but also reported specifically in children [4]. Additionally, ILDs can develop due to the prolonged occupational or environmental history. Occupational ILDs like silicosis and pneumoconiosis can occur to sandblasters, granite workers or miners. automobile mechanics, electricians, pipe fitters, shipyard workers, and welders can develop occupational ILD called asbestosis. Farmers, poultry workers, bird fanciers and bird breeders can get affected by hypersensitivity

pneumonitis (HP). Drug-induced ILDs may also evoke airway (bronchiolar or lung) injury and are an important cause of ILDs [8]. For example, the reaction of drug Amiodarone can be involved in mononuclear cell infiltrates, fibrosis, septal thickening, and textural changes in lung that occur in nonspecific interstitial pneumonia (NSIP), cryptogenic organizing pneumonia (COP) or usual interstitial pneumonia (UIP) [8]. Similarly, it has been commonly observed that patients undergoing cancer treatment (e.g. bleomycin) can develop respiratory symptoms and lung infiltrates [8, 9]. Early diagnosis of drug-induced ILD, timely discontinuation of medications and treatment with corticosteroid is required to reduce the disease progression. Furthermore, many other rarer ILDs types' exist and affecta smaller number of patients which may or may not show typical disease patterns [6].

No specific findings have been reported yet for specific causes of interstitial lung disease. The disease or most of its subtypes are having the common feature of either scarring in the lung tissue or maybe inflammation, and the bottom line of the disease is lack of oxygen. Treatments depend on an accurate diagnosis, but sufficient numbers of distinguishing features are important to understand the disease exactly. The patterns, course, and prognosis of ILDs vary in different area. Some of the conventional biomedical tests include imaging tests (CXR, HRCT), pulmonary function test (spirometry, exercise stress test, oximetry, etc.), lung tissue analysis (bronchoalveolar lavage (BAL), bronchoscopy, surgical lung biopsy (SLB). Sometimes, only HRCT scan doesn't give enough information in such condition; therefore, appropriate and managed diagnostics that include radiological features, pathological reports with integrated patient history and physical examinations are required.

It has been studied and shown that a computer-aided diagnosis (CAD)-based system for ILDs can perform potentially and has many advantages [10]. The symptoms and critical information for diagnosis do not differ at greater extent among ILDs patients (all ILD patients features look similar), but a disease can exhibit different radiological and clinical patterns whereas in some cases multiple diseases can encompass common patterns. The study performed by Radiological Society of North America (RSNA) on CAD (based on receiver operating characteristics (ROC) curves) promoted the use of CAD system to boost radiologist's effectiveness [11]. Doctor's diagnostics in combination with CAD, an instrument for second opinions, can serve better and efficient diagnosis of disease. It could also help to achieve a good survival rate for the ILD patients. [12, 13].

The patterns, mode of action, and prognosis of ILDs differ in different area and make it difficult to achieve a confident diagnosis. Several efforts have been made and research is in progress to understand the necessity of higher prognosis and better care to the ILDs patient. British thoracic society, American thoracic society, French network for ILDs, Australian and New Zealand's thoracic society, Irish thoracic society and several other healthcare societies are operating for the disease management [14, 15]. However, several ILDs patients may not respond to the provided treatment and frequently require oxygen support. Developed countries carry out biopsy to confirm ILD, while in low income and developing countries like India radiological features are the most important method for diagnosis [16]. Radiology provides a very good diagnostic rate, but it has higher radiation exposure, uncertain rates of diagnostics in early disease stages and less prognostic values [17, 18]. Thus,the requirement of a multidisciplinary diagnostic process is an important tool mainly for diseases with nonspecific etiologies.

However, the establishment of patient-centered and personalized ILDs management plan is really a challenging task for clinicians mainly in developing country [19]. The main challenges the clinician's face is the availability of supporting research, genetic information and clinical data from patients [20]. Although there are several online resources and information available, there's a large gap between world knowledge and medical registries from developing countries. We have tried to fill this gap and provided a comprehensive knowledge of ILDs. A detailed description of ILDs such as history, ILD's clinical and radiological presentations, genetics and mechanism of disease, current diagnostics and management, limitation, and importance of present work to overcome these limitations has been discussed in succeeding sections.

1.2 History of ILD

Starting from the 19th century, lung fibrosis was referred as 'cirrhosis of the lung', while Louis Hamman and Arnold Rich referred it as pulmonary fibrosis in 1944 [21]. Gradual deaths after subacute respiratory failure were reported, and interstitial fibrosis was reported in the autopsy. This was the first reported pathological description for interstitial fibrosis. Furthermore, ILD was named after the doctors and called as 'Hamman-Rich syndrome' for the next few years. However, clinicians distinguished a chronic form of fibrosis different from the description of 'Hamman-Rich syndrome'. To better describe the disease, a new term "Cryptogenic fibrosing alveolitis" was coined which is synonymous to IPF. After so many

disguising discussions over the disease and symptoms in 1965, Stack et al. suggested that "Diffused Interstitial lung disease" is the most acceptable and accurate term to describe disease phenotype [22].

1.3 Classification of ILD

The first classification came in 1969 when ILDs were differentiated into five major types naming as bronchiolitis obliterans interstitial pneumonia (BIP), giant-cell interstitial pneumonia (GIP), desquamative interstitial pneumonia (DIP),usual interstitial pneumonia (UIP) and lymphoid interstitial pneumonia (LIP) by Liebow and Carrington [23]. The American Thoracic Society and European Respiratory Society (ATS/ERS) issued guidelines in 2002 to standardize the approach towards ILD as various classifications were available without standardization. They broadly referred ILDs as diffuse parenchymal lung disease (DPLD). The DPLDs was further categorized into DPLD of a known cause, DPLDs of unknown cause (idiopathic interstitial pneumonia (IIP)), granulomatous DPLDs and other DPLD such as histiocytosis, LAM, etc. The IIP was further sub-categorized as idiopathic pulmonary fibrosis (IPF), cryptogenic organizing pneumonia (COP), non-specific interstitial pneumonia (NSIP), acute interstitial pneumonia (AIP), respiratory desquamative interstitial pneumonia (DIP), bronchiolitis-associated interstitial lung disease (RB-ILD) and lymphocytic interstitial pneumonia (LIP). The International Multidisciplinary Consensus Classification of ILDs by ATS/ERS was also proposed (Figure 1.1).

Radiological changes in HRCT due to the ILDs can be categorized as increased and decreased attenuation of lung tissue (Table 1.1). Some characteristics of these ILDs based on their classifications are described in the following subsections:

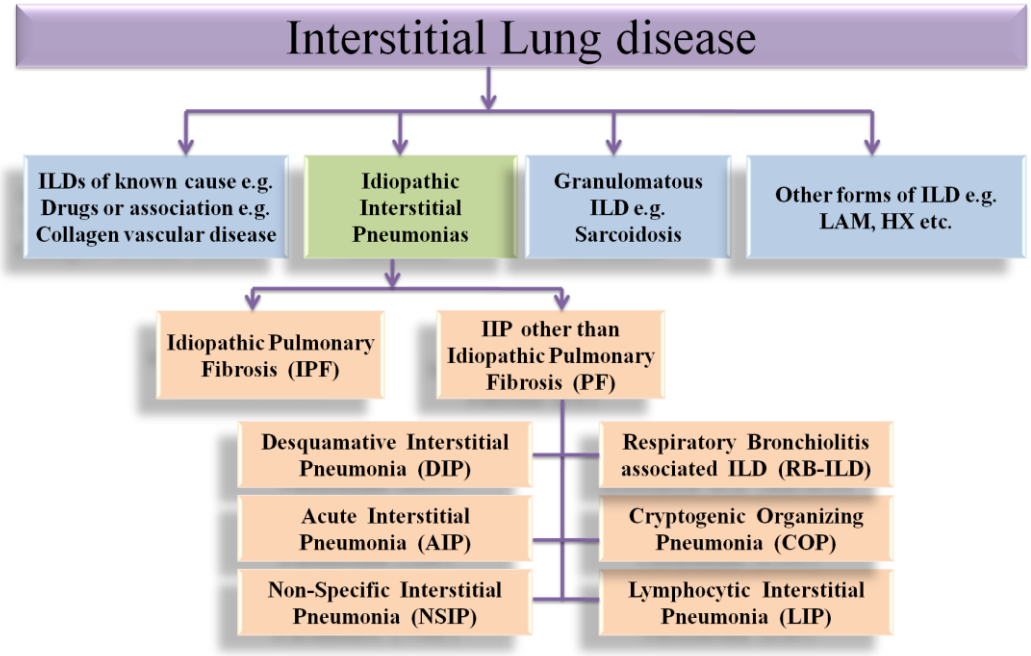


Figure 1.1 Multidisciplinary classification of ILDs by American Thoracic Society and European Respiratory Society [24]

Table 1.1 Increased and decreased attenuation found in HRCT and associated diseases*.

Increased attenuation					Decreased attenuation	
Reticular Opacities	**Ground-glass Pattern**	**Consolidation**	**Nodules**	**Linear Opacities**	**Honeycombing**	**Cystic Diseases**
UIP NSIP Collagen vascular disease *Asbestosis* Drug-related pulmonary fibrosis	NSIP *AIP (acute, subacute)* DIP RB-ILD LIP COP *Subacute HP* Acute exacerbation of ILD	COP Polymyositis/ Dermatomyositis Acute exacerbation of ILD *AIP* *Acute HP* Drugs Induced *Sarcoidosis*	**CENTRILOBULAR NODULARITY** *Subacute HP* RB-ILD Langerhans cell histiocytosis **PERILYMPHATIC NODULES** *Sarcoidosis* Silicosis Coal worker pneumoconiosis Berylliosis LIP **RANDOM NODULES** Hematogenous metastases Miliary fungal infection Miliary tuberculosis Silicosis (mimic) Coal worker pneumoconiosis (mimic) Sarcoidosis (mimic)	**SMOOTH** Hydrostatic edema Lymphatic congestion **IRREGULAR** Fibrosis Lymphoma Secondary solid tumor **NODULAR** Sarcoid Lymphoma Secondary tumor	UIP/idiopathic pulmonary fibrosis (most common) Fibrotic NSIP Drugs-induced *Late asbestosis* *Sarcoidosis (stage IV)* *Chronic HP* Collagen vascular disease	Langerhans cell histiocytosis Lymphangioleiomyomatosis LIP DIP Birt-Hogg-Dube´ Light-chain deposition disease

*Diseases exhibit more than one pattern is presented with underline and the diseases exhibit different patterns in distinctive stages are presented with italic letters. AIP: acute interstitial pneumonia, COP: cryptogenic organizing pneumonia, DIP: desquamative interstitial pneumonia, HP: hypersensitivity pneumonitis, LIP: lymphoid interstitial pneumonia, NSIP: nonspecific interstitial pneumonia, RB-ILD: respiratory bronchiolitis-ILD, UIP: usual interstitial pneumonia.

Note: Table reproduced with the help of guidelines provided by Hamza et al. and other internet sources [25].

1.3.1. ILD of known causes

Known causes of ILDs are mainly some etiologic factors like hypersensitivity pneumonitis, collagen vascular disease, drugs etc.

1.3.1.1 Collagen vascular disease associated ILD (CVD-ILD)

The CVD-ILDs can be a manifestation of collagen vascular disease. Comparatively less chronic symptoms can be seen in disease early stages. In imaging modalities, UIP or NSIP pattern can be commonly seen.

1.3.1.2 Hypersensitivity pneumonitis(HP)

HP can be an infection from of the exposures such as fungus, aspergillus, moisture and birds feathers. Majorly, physical examinations include flu-like symptoms such as fever, chronic cough, dyspnea etc.

1.3.2. Idiopathic interstitial pneumonia (IIP)

1.3.2.1 Idiopathic pulmonary fibrosis (IPF) and Familial IPF

IPFs mainly affects patients from older age groups (>50yrs). Symptoms mainly start with cough and shortness of breath. Respiratory crackles (Bibasilar) and standard restrictive lung symptoms can be present in spirometry. Predominant HRCT findings are Basal, subpleural, and focal ground glass. In the case of familial IPF, the disease can be hereditary and can be carried out by one or more family members. The HRCT finding of such cases has less honeycombing and less lower-lobe predominance. ATS/ERS standard for diagnosis of IPF is given in Table 1.2:

Table 1.2 UIP pattern of IPF in HRCT: OFFICIAL DOCUMENT –ATS-ERS-JRS-ALAT [26]

UIP pattern (All four features)	**Possible UIP pattern (All three features)**	**Inconsistent with UIP (Any of the seven features)**
Subpleural, basal predominance Reticulrar abnormality Honeycombing with or without traction bronchiectasis Absence of features listed as inconsistent with UIP pattern (See third column)	Subpleural, basal predominance Reticular abnormality Absence of features listed as inconsistent with UIP pattern (See third column)	Upper or mid-lung predominance Peribronchovascular predominance Extensive ground-glass abnormality (extent > reticular abnormality) Profuse micronodules (bilateral, predominantly upper lobes) Discrete cysts (multiple, bilateral, away from areas of honeycombing) Diffuse mosaic attenuation/ air-trapping (bilateral, in three or more lobes) Consolidation in bronchopulmonary segment(s)/ lobe (s)

1.3.2.2 IIP other than IPF

i. Nonspecific interstitial pneumonia (NSIP):

Progression is higher in NSIP. The HRCT patterns show basal, consolidation and ground glass opacity (subpleural, peripheral). Further associated conditions include HP, CVD-ILDs, HIV and drug-induced pneumonitis.

ii. Acute interstitial pneumonia (AIP): The AIP is a fast progressive disorder and its mean age is approximately 50 years. HRCT finding suggests progressive ground glass, diffuse consolidation, lobular sparing. Many AIP patients can have a prodromal viral illness.

iii. Respiratory bronchiolitis induced interstitial lung disease (RB-ILD): tobacco smoking is strongly associated with RB-ILDs. HRCT representations of the diseases are mainly bronchial wall thickening, centrilobular nodules, and diffuse ground glass.

iv. Cryptogenic organizing pneumonia (COP): The COP is comparatively common in non-smokers. The COP patients have shown responses to corticosteroids; however, relapse is a common side effect after stopping steroids. The HRCT findings suggest bilateral consolidation (patchy) and nodules.

v. Desquamative interstitial pneumonia (DIP): The DIP mainly affects smokers and the mean age of affected patients is 45 years. HRCT findings mainly suggest ground glass and reticular lines.

vi. Lymphocytic interstitial pneumonia (LIP): The LIP is another rare form of ILDs characterize by the accumulation of lymphocytes and plasma cell. The disease symptoms may exhibit cough, fever etc. HRCT findings can be nonspecific, sometimes interstitial infiltrates can be seen.

1.3.2.3 Granulomatous diseases

i. Sarcoidosis

Other system involvement can be found in sarcoidosis. Staging of sarcoidosis ranges from Normal chest radiograph to extreme pulmonary fibrosis. HRCT findings suggest hilar lymphadenopathy, parenchymal fibrosis, peribronchovascular nodules.

1.3.2.4 Miscellaneous

i. Lymphangioleiomyomatosis

Commonly seen in premenopausal women, and can be associated with multiple bilateral cysts.

ii. Histiocytosis

Histiocytosis is associated with farming and cigarette smoking. The HRCT findings are interstitial thickening, cysts and nodules patterns.

iii. Pulmonary alveolar proteinosis (PAM)

PAM is a very rare ILD form. Mainly affects patients of the mean age group of around 20-40 years. The HRCT findings include crazy paving pattern, mid and lower lobe involvement.

1.4Mechanisms of Disease

The exact phenomenon for the beginning of pathogenic mechanism is not known. Nearly half of the ILDs are of unknown cause, and the exact mechanism for the initiation of lung injury and fibrosis is also unidentified [27]. Due to the unidentifiable underlying causes, most of the ILDs are defined as "idiopathic" [5, 28]. However, occupational, environmental and prolonged drug courses were found to be associated with ILDs. Varying extent of inflammation, fibrosis, and increased amount and abnormal structure of the connective tissue is frequently observed in ILDs [8]. Whenever causatives exposed to the respiratory tract, lungs inflammation occurred and subsequently followed to lung scarring and fibrosis. Fibrosis can develop in lung because of many reasons ranging from short- and long-term injuries, or other pre-existing diseases. Some forms of fibrosis are partially reversible but the majority of them are irreversible and fatal. Initiation of fibrosis can be caused due to the accruement of fibroblasts and extracellular matrix (ECM). The origin of fibroblasts activation is yet not clear and poorly explained. One possible scenario is when a drug, disease or such cause exposed to the lung; it causes epithelial damage and inflammations. These inflammations then produce the signals to activate the resident fibroblasts which cause the interstitial thickening and fibrosis to develop mild to chronic lung disease symptoms [5]. A pictorial representation of the mechanism is given in Figure 1.2.

1.5 Genetics of Interstitial Lung Disease (ILD)

Genetic basis of ILDs has been already demonstrated in many studies [4, 27, 29, 30]. Many studies demonstrated the aberrations of different pathways, mutations, and links to noncoding RNAs [31-33]. The molecular basis and importance of genetics in ILDs pathogenesis are also studied, and many classes of ILDs have also shown the role of inheritances in the development of ILDs [34]. A detailed discussion about the background, the role of mutations and ncRNAs, and clinical management by genetic of ILDs is provided in subsequent sections.

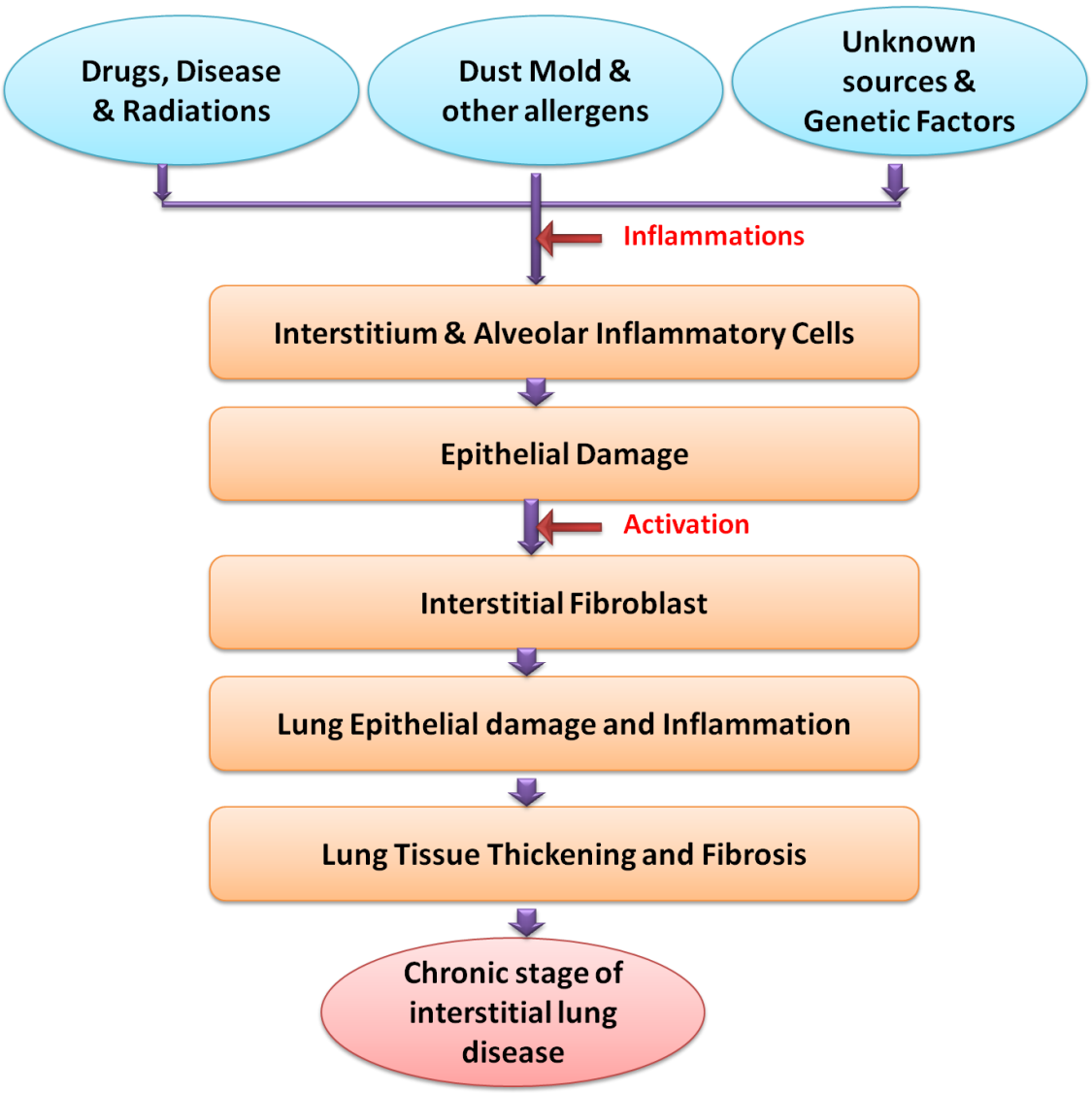


Figure 1.2 Underlying mechanism of interstitial lung disease

1.5.1 Background of Genetic of ILDs

Role of genetics in ILDs are not new insight in the research field, and there was a huge history of related research that gives significant support to the fact that genetics always have a suspected role in etiologic and predisposition of ILDs [4]. Reported literature made a strong reference that genetics may have a very crucial role in prognosis and etiology of the disease groups. Statistical significances in the hereditary nature of emphysema and bronchitis have already been studied [35]. An investigation research paper reported that inheritance through father's genes was predominating the predisposition of familial silicosis in 250 out of 391 families [36]. Another research published at the same time stated that three families having a history of the chronic pulmonary disease had five infants suffering from "interstitial pneumonitis". In order to make the better understanding about the disease, researchers

suggested the term familial cystic fibrosis to distinguish the familial cases from the rest [2]. Studies are not limited to these references, and approximately 150+ published studies in early 19 century suggested the evidence of familial trait in ILDs.

Idiopathic pulmonary fibrosis also has such evidence of genetic inheritances. A study reported a case of IPF transmitted through autosomal dominant trait. The disease traits had been carried by 3 generation of the family and about 8 persons were suffering from the disease [37]. Furthermore, many studies also suggested the population-based disease traits of the disease. Linking of excesses human leucocyte antigens (HLA) is predominantly found in a community-based population suffering from silicosis. This study suggested the possible genetic susceptibility to silicosis [38]. Another study in the Russian population found that 15 systems of hereditary polymorphism showed resistance and susceptibility of silicosis in addition to HLA system. The study suggested a prior genetic test in occupational selection could be helpful in avoidance of the disease [29].

There is a substantial increment in the number of studies in the past few decades. Many factors can have roles in disease cause and progression. Genetic factor, environmental factors, and geographical cofactors can be important in the clinical representation of the disease. In recent studies, scientists made strong substantiation that genetics has a vital part in the pathogenesis of IPF. Genetic studies performed on model animals like rat can give a clear insight to the importance of few genes on disease pathogenesis [39].

1.5.2 Role of mutations in ILDs pathogenesis

Mutation in genes associated with surface protein, telomeric length, etc. have been reported and proven a strong connection with the ILDs pathogenesis. Mutational studies on the set of genes have been current thrust in the past decade. Genes like SFTPA2, SFTPC, TERT, and TERC carry inheritance in familial IPF [40-43]. Mutations in these genes provide significant association in disease-causing factors. There is also evidence that pathways related to aging and pathogenesis of IPF are interrelated to some extent [44]. Also, mutations in ABCA3 gene augment the chance of CHILD by altering the function of the gene [43]. These mutations have provided an idea that surface protein mutation testing could be done in CHILD diagnostics.

1.5.3 Role of ncRNAs in ILDs pathogenesis

Not only mutation but micro RNA (miRNAs), small noncoding RNAs conjointly proved to have potential roles in many biological processes and pathways involving in ILDs [45].

Although the research scenario is comparatively new for miRNA-mRNA regulation in LDs, many studies since the last 5 years are published for the role of miRNAs in ILDs. Prior researchers have identified the role of miRNAs in PF for mice. The involvement of miRNA in susceptibility of human lung fibroblast severity was also reported [46]. Scientists use miRNAs microarray to the interrogated role of miRNAs in lung fibrosis. The study reported the confident roles of miRNA up-regulation over genes and that could provide a beneficial therapeutic and testing target for IPF and other ILDs [47]. Many studies emphasized on the fact that miRNA profile based biomarkers can be promising and serve as non-invasive diagnostic tools in many lung diseases including lung cancer [48].

1.5.4. Clinical Management of ILDs by Genetics

The genetic of interstitial diseases demonstrated the dysregulation of different pathways, differential expressions, ncRNAs regulations or mutations can lead to one or many forms of ILDs [27]. Pulmonary phenotype and associated symptoms can provide the clue for the underlying disorder. Aberrant gene expression by mutation or by interaction with other non-coding regions may consequence in a range of clinical manifestations. Identification of pathogenic mutations can assist to understand different clinical phenotypes. Furthermore, these analyses may lead to discovery of a rare variant with unidentified clinical importance. This mutation can either be novel or know for irregular functional effects on the coded protein. These investigations can help the researchers to discover the underlying pathogenesis and novel therapeutic targets for enhanced management of ILDs [49].

1.6 Diagnostics of ILDs

A multidisciplinary clinical evaluation can accomplish a certain diagnosis of a specific ILD subtype [3].The earlier diagnosis can lead to the early initiation of therapy which could result in reduced health risk for ILDs patients. The ATS/ERS/JRS/ALAT recommendations are useful in simplification of the diagnostic procedure. However, these guidelines do not recommend diagnosis coming from only imaging or singular form of a diagnostic method. Additionally, information regarding the treatment recommendations is required to be added significantly. Therefore, multidisciplinary diagnostics system is in urgent need to include pulmonology, radiology, and pathology to serve early and accurately confident diagnostics and provide effective therapies [14, 26, 50]. A traditional way of ILD diagnosis protocol is given in Figure 1.3.

1.6.2. Chest imaging

Chest imaging has an essential role in confirming the diagnosis, and in the case of consistent and typical cases, often useful to avoid other invasive tests such as SLB [51]. CXR is the most primary examinations and most popular mainly in developing countries due to

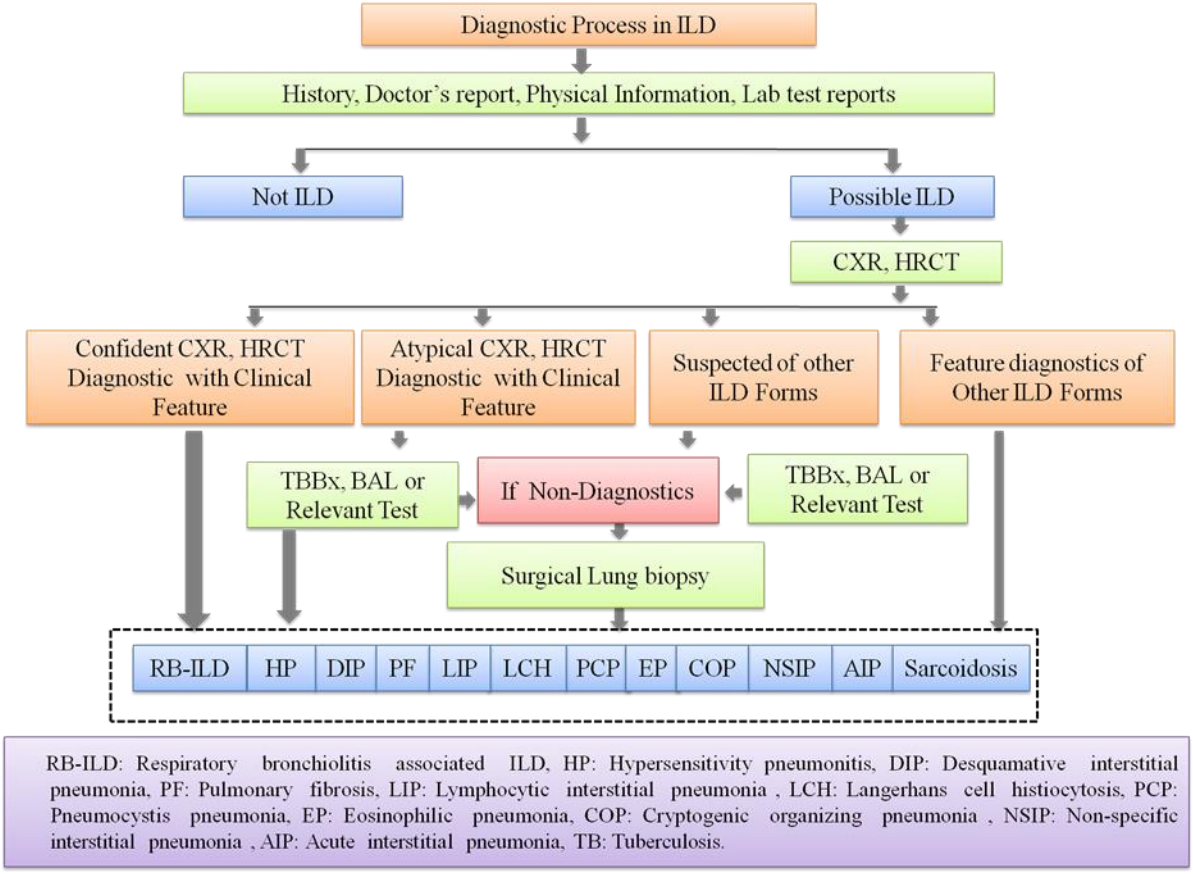


Figure 1.3 Multidisciplinary standard approaches for ILD diagnosis and management

affordable cost and comparatively less radiation [1, 51]. However, CXRs do not provide much information in some ILDs subtypes and reflect normal radiological findings. Thus, this method yields comparatively less level of accuracy (nearly 25%) [51]. In order to acquire better and accurate assessment of the lung tissue, a 3D form of data from chest HRCT were considered as former gold standard protocol. HRCTs provide efficient way and quantified the distribution of the histological patterns of the lung [24, 51]. However, many ILDs represent very nonspecific and atypical radiological features. A single form of ILDs can encompass different radiological patterns, and a single pattern can be present in many ILDs; careful investigation of radiological patterns is required. Considering the mimic and overlapping of the radiological patterns, integration of HRCT with clinical findings are expectant for accurate and confirmed diagnosis [52].

1.6.3. Pulmonary function test

As discussed earlier, difficulty in breathing is a common symptom in ILDs. Pulmonary function test (PFT) is very useful to identify any alterations in normal lung capacity and functions. Generally, pulmonary abnormalities are a sign of restrictive and decreased lung volumes in ILDs. Results of PFT mainly consider the comparison to the average amount of air inhaled ("Total lung capacity (TLC)") and exhaled ("Vital capacity (VC)") during the test. Many studies tried to establish the use of PFT results for ILDs differential patterns, but considerable overlap among the diseases limits its practical implication [53]. Although PFT doesn't offer any precise information, it is a valuable examination for ILDs diagnosis and management [54]. As TLC usually decreased, hypoxia becomes a consistent and sensitive sign of early disease stages [54].

1.6.4. Bronchoalveolar lavage(BAL)

The BAL has been a conventional and reliable diagnosis process where a bronchoscope is introduced into the lung via nose or mouth to collects lung specimen for cytological examination, molecular analysis and microbial cultures for ILD testing.. BAL results assist to investigate patients when clinical and radiological investigations did not clinch a diagnosis [55].

1.6.5. Tissue biopsies

In addition to clinical patterns, ILDs examination using histological patterns has become increasingly important due to varying and differential patterns of disease. ILDs diagnosis can often be made after reviewing of HRCTs. However, surgical lung biopsy (SLB) can be required to verify the histological diagnosis in order to achieve multidisciplinary diagnostics. Thus, SLB has been a "gold standard" for most type of ILDs [24]. It is most commonly performed methods in developed countries and requires general anesthesia [56]. Moreover, a multidisciplinary approach can decrease the need for surgical lung biopsy and can serve as new "gold standards" for ILDs diagnosis and prognostics [57].

1.6.6. Genetic testing

As discussed earlier, the role of inheritances in bronchitis, cystic fibrosis, silicosis, and IPF has already been studied [2, 24, 33, 36, 37]. Genetic testing conducted on the model animal has given a lucid insight into the involvement of genetic factors in etiology of the ILDs [27]. Use of these factors in improved disease diagnosis processes is also popular these days [26]. Hence, an integrated analysis with the help ofdisease-associated genes and other genetic data

can help to identify candidate biomarkers and new therapeutic targets. Recently, non or less invasive techniques are being popular to trim down the side effects of biopsies, other co-morbidities and heavy radiation exposure [58]. Critical involvement of genetics in the etiology and prognosis of ILDs were already identified [30]. Improvements in the current ILDs diagnosis process are possible with the help of genetic biomarkers [59].

The regulatory role of different type of noncoding-RNAs (ncRNAs), associated pathways and genes were already studied in many diseases pathogenesis including ILDs [32, 60-63]. Recent studies have revealed that the immune system is an intervening factor in ILDs pathogenesis, and it could be dysregulated by ncRNAs [64-67]. These ncRNAs were already established as promising therapeutic targets for the prognosis of IPF and other ILDs [10, 68, 69]. Furthermore, the role of cross-regulation between the ncRNAs referred as CeRNA (competing endogenous RNA) was also found in many diseases including cancers [32, 70, 71]. It is also involved in post-transcriptional regulation and biological processes. Interaction analysis of ncRNAs with disease candidate gene, and their role in disease pathogenesis can be explored. Furthermore, ncRNA as biomarker or therapeutic targets could be further explored for ILDs and other respiratory diseases [72, 73].

1.7 Limitation of current diagnostics

1.7.1. Misdiagnosis of ILD

The developing countries mainly relies on comprehensive and complete clinical history (including the history of drug intake, familial history, occupational and environmental exposures, and multi-systemic examination), CXR and HRCT for the diagnosis of ILDs. In developing countries, there is always a looming possibility of under-diagnosis and under-reporting of ILDs [74, 75]. It may be due to the lack of acquaintance among doctors, or non-availability and/or non-affordability of various diagnostic modalities like high-resolution computed tomography (HRCT) scanning, bronchoscopy and video-assisted thoracoscopic surgery (VATS) [75, 76]. Another significant problem is the over-diagnosis of tuberculosis in ILD patients [77, 78].

1.7.2. Lack of infrastructure for ILD

In India, most of the district level hospitals (private or government) have chest X-ray (CXR) machines but do not have CT machines. The patient can go for CT but lack of trained and efficient radiologists hinders the diagnosis of ILDs. Most of the diagnosis is currently made

on the basis of CXR and HRCTs, and doctors mostly do not (radiologist or doctor) suggests further treatment regimen [75] mainly due to improper infrastructure and heavy patients burden [79]. Though SLB is the "gold-standard in ILDs diagnosis", lack of specialist equipment and its affordability dissuade patients to explore this test [75].

1.7.3. Difficulty in diagnosis

The symptoms of an extensive range of medical conditions can mimic ILDs; therefore, confirmation of ILD can be extremely challenging even for highly experienced radiologists [52]. An unusually large number of lung disorders fall into this broad category. Furthermore, doctors must rule out other cases before making a state-of-the-art diagnosis [50].

1.7.4. Lack of sufficient representative cases

As recommended by many official thoracic societies, accurate and confirmatory diagnosis can be achieved by collaborative efforts from pathologist, pulmonologist, and radiologist, which is not in common practice in India. A proper investigation and interpretation of medical data are always required for proper diagnosis and treatment of any disease [80]. Accurate guidelines, disease identification or management always depended on an accurate compilation of medical data. These datasets were compiled, drafted and kept as medical literature from ancient time. Gathering, analyzing and compiling diversified medical data in order to achieve "decision making" is currently acknowledged as a key requirement for the entire healthcare [80]. Many diseases including ILDs encounters rare cases casually even in big university's hospitals. Incidentally, most of these do not have easy availability of referral data to overcome the lack of experience [51].

1.7.5. Unknown etiology and radiation exposures

As discussed earlier, the ILDs are very less known disease and most of the causes are still idiopathic. Also, ILDs encompasses diverse pathological processes so the treatment and regimen are different for each ILD subtype [81]. An accurate and specific diagnosis is always required for successful management of ILDs. Many diseases have a similar presentation and many have considerably varying symptoms. The CXR and lung HRCT can offer valuable information that strongly supports a precise diagnosis. Many scientific committees including the American College of Radiology (ACR), European Commission's Radiation Protection Actions Committee (EUR16262), International Council of Radiation Protection (ICRP) and United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) are continuously raising serious concerns on radiation exposure from HRCT and CXRs. Studies

are ongoing for radiation exposure and their provable biological effects and potential risks [82].

After 2008, ATS recommended the use of SLB to confirm the histological diagnosis. Again these tests are critical with associated risk factors and very well trained practitioners are required to perform the task [26]. Sampling results of the body fluid indicated that the important role of proteins secretion and their alteration in ILDs. These proteins are also present in bloodstream and serum even in small amount. These results suggest that genetic analysis of this protein, blood serum, and associated genetic factors might represent a novel approach in the assessment of ILDs [3]. Genetic testing is in common practice in many countries; however, proper guidelines are not available [3].

Aforementioned limitations in current diagnostics approach could be addressed if patient's complete medical data made available. Multi-disciplinary diagnostics (MDD) approach could be used to achieve more accurate diagnostics of ILDs as MDD required patients' data from a different domain (doctor's reports, radiologist's report, imaging data, patients' history, pathological tests, and symptoms, etc). Furthermore, genetics of ILDs could be used for better and less invasive approaches. Importance, type and current status of medical data are described in subsequent sections.

1.8. Importance of medical data in healthcare

Data generated by healthcare are usually recorded for a variety of reasons ranging from evidence-based treatment to knowledge generation for future training. Patient's clinical data can be essential for the appropriate treatment. Further, this data can be provided for research scientists to contribute to clinical research. Results of these researches can provide insightful knowledge to benefit the practinores and researchers to learned from the data itself [80, 83]. In many cases, there isthe occurrence of very atypical and rare cases. Avilibity of this kind of data from worldwide can be difficult due to different demographic reasons. Many disease types only occur in a specific type of environmental or geographical regions. The transferability of these datasets are required to make the knowledge availability worldwide. The major usages of medical data are as follows:

i. Provide the base for the medical knowledge

Medical records can be created and projected as a comprehensive compilation of clinical data about individual patients to provide evidence for many aspects such as:

- Detailed history about the patient's condition (if any prior disease histories; present illness; familial history for genetic illness; environmental and occupational exposures; social, and demographic information).
- To understand the similar existence of patient's symptoms.
- Clues towards possible causes.
- History of first physical signs, the duration of symptoms and the changes in symptoms and signs over the time if any.
- Types of examinations and patterns reported for the patients or disease.
- Other relevant tests history and utility of the test requirement as many diseases may have undetectable symptoms for specific tests.
- Any specific radiological or pathological test performed and if any specific pattern was detected.
- History of medication and prevention are being prescribed; and are there any allergies reported?

ii. Support Communication among Providers

Over the years(since technology takes place in the healthcare sector), the emergence of specialization in medical practitioners has emphasized the crucial role of the medical data [80]. Now the medical data is not only a list of observations by a physician for reference, but also a knowledgebase for diverse data. These datasets are successfully helping in transferability of knowledge among physicians and medical personals [80]. Easy availability of such data can benefit the patients receiving treatment over time from a variety of physicians to avoid repetitive tests and correct assessment.

iii. Future approaches to common health problems and identification of atypical behavior of patient's

The patterns, courses and symptoms of many diseases such as ILDs, cancers etc. vary patients to patients with changing the environment, geographical factors or lifestyle changes [84]. Formal information to the patients about their risk factors, lifestyle changes to reduce the risk of future incidents or progression of the lethal disease can be useful. Similar information can also assist the medical practitioner which can save the time and efforts toward successful diagnosis [85]. Similarly, intervention, case studies and medical history of patients can assist in new methodology development, or for treatment of existing cases where evidence-based medication could help [86]. The easy availability of diverse clinical data is

significant to identify the risk factors, and guiding the patient for preventive interventions. For example, an overview of patient's medications, allergies or clinical history would be required if a patient needs an immediate blood transfusion. Availability of medical data can easily reduce unnecessary costs or risky treatments [80].

iv. Provide basis for clinical research

Conduction of a successful clinical research depends upon so many challenging factors including recruitment challenges, troublesome implementations for data collections and raising funding issues mainly in developing countries [87]. These factors add to the increasing the difficulties in clinical research often do not provide much significant if time, manpower and money has not been invested wisely and sufficiently. Till date, availability of such clinical data is considerably less mainly for diseases like ILDs. If the clinical data is available and drafted carefully it can be used for reproduction of the clinical research and significantly reduce the research cost [87]. Clinical data can be used to establish standard clinical patterns through the statistics and further analysis of annotations collected from patients [73, 88].

1.8.1. Type of medical data

The medical data can mainly be of two types: textual and imaging. The data obtained from clinical history, histopathological tests, or physical examination can be in the form of numerical values or descriptive. Imaging data mainly comes from recorded signals, radiographs, HRCTs and pathological images (Figure 1.4). These two data sets are the major categories required for successful disease diagnosis. The textual data is traditionally collected either by asking several questions like symptoms of present illness, occupational or family history, and ongoing/previous medication of patients by doctors. This could be collected verbally or by preparing specific questionnaire to create patient's profile. The later is nowadays a common practice mainly in developed countries. Similarly, medical images acquired from different sources (radiological or sketches) are the most significant tools for disease diagnostics or prognostics measures. The imaging data are usually preferred as a primary test because of its availability and non-invasive nature. Traditionally, textual data were handwritten and stored in the patient's medical records (mainly as files or transcripts) by physicians and clinicians. These records were then transcribed by a typist or medical coders in electronic form. With the advancement of technology, patients data is now recorded directly as electronic medical health records in modern patient's care [80].

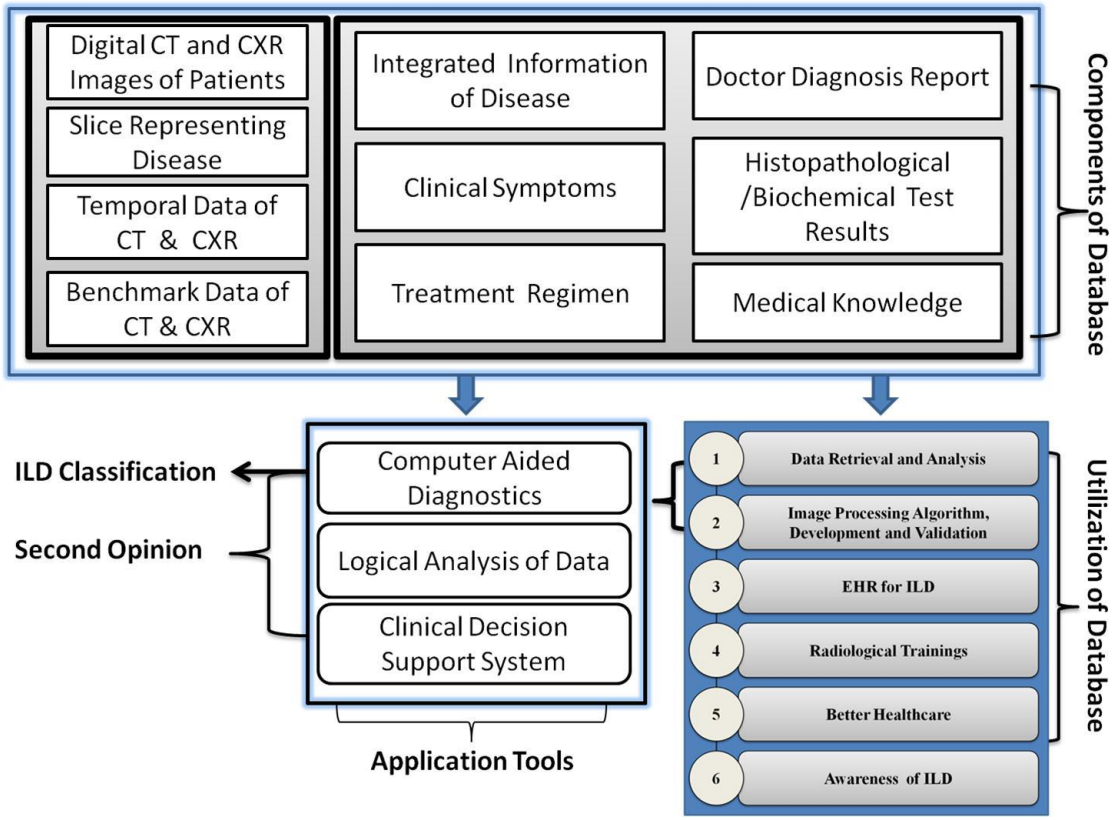


Figure 1.4 Types of medical data, utilization of database and applications tools

1.9. Review of current status of research and development

In the early eighties, "the Kurt Rossmann Laboratories for Radiologic Image research in the Department of Radiology at the University of Chicago" established large scale and systematic research and implementation of varied computer-aided diagnosis (CAD) schemes [89]. These CAD schemes include web-based knowledge resources and tools to assist in early disease diagnostics. After this, use of CAD has been accelerated and recognized as a valuable means for improved performance and decision-making for detection and analysis of complicated imaging features in the chest. The sole aim of developing these systems was not only to offer a unified centralized location but conjointly address the problem from a totally different perspective and serve to physicians and radiologist to improve the disease treatment in clinical routine by using these systems as the second opinion. There are many successful and unsuccessful implementations of these systems around the national and international research communities. Here we are representing a number of the successful and proven systems that serve successfully.

1.9.1. International Status

i. Healthy People 2020 [89]

The United States Department of Health and Human Services set targets to guide national health promotion and disease prevention efforts for improving the healtcare. The “Topic Areas of Healthy People 2020” recognize and group objectives of specific issues and populations. Some of these objectives also include

1. Public health infrastructure
2. Respiratory diseases
3. Health communication and health information technology

ii. The lung imaging database consortium (LIDC) [90]

In total, five different institutions at the United States were collaborated to establish LIDC. The database contains abnormal and healthy HRCTs with annotated nodules and major clinical data of the patient. The database is publicly available from the national biomedical image archive (NBIA). However, LIDC does not specifically mention ILDs cases and only targeted nodules in HRCs.

iii. Lung Tissue Research Consortium (LTRC) [91]

The LTRC offers human lung tissues for research purpose. The program enrolls volunteers that are about to go under lung surgery, and collects blood specimens, tissues samples and phenotypic data for research. Most of the cases are diagnosed with interstitial fibrotic lung disease or COPD. High quality and uniformity of the biospecimens are achieved by complying standard protocols. The tissue samples are processed to allocate biomarker measurements, genetic and gene expression analyses, immune histochemistry and pathological study. Phenotypic data include clinical and pathological diagnoses, HRCT, PFT, exposure and symptom questionnaires, and exercise tests.

iv. Japanese Society of Radiological Technology (JSRT) Database [92]

JSRT was developed in association with “The Japanese Radiological Society (JRS)" in 1998. This standard image database contains digital images with and without chest lung nodules. Since foundation, the resource has been used for image processing, CAD, and training. This database has also been used for developing algorithms in imaging.

v. University of California San Francisco (UCSF) Interstitial Lung Disease Program [93]

This program was established to improve ILDs' management by focusing on patient care, education, and research. The program provides up to date diagnosis and treatment recommendations.

vi. Idiopathic Pulmonary Fibrosis Clinical Research Network (CRN) [94]

This program was sponsored by "The National Institutes of Health (NIH)". A major goal of the program is to conduct and extend clinical research for the advancement of potential IPF therapies.

vii. Reference Centre for Rare Lung Diseases [95]

Reference Centre for Rare Lung Diseases was established in 2006. The center includes extensive data on major forms of rare lung disorders. These lung disorders grouped mainly in four main categories: ILD, ciliary dyskinesia, respiratory tract malformations, and other rare chronic respiratory insufficiency diseases.

vii. Veterans Health Information Systems and Technology Architecture (VistA) [96]

*The VistA*is an electronic health record (EHR) enabled information system. *VistA* has been used by "The United States Department of Veterans Affairs (VA) Medical System", known as the "Veterans Health Administration (VHA)".VistA is the most popular EHR in the US, which contains various tools for clinical care [21].

ix. MEDgle [97]

MEDgle was started in late 2006 by data scientists and artificial intelligent experts from MIT with a focus on the organization of health data. This resource has been considerable in data curation, data-models building, and the algorithms to enable real-time predictive clinical analytics. This resource was developed to resolve the supply-demand imbalance of clinical data worldwide.

1.9.2. National Status

i. Interstitial Lung Disease (ILD), India, Registry [16]

The "Indian Chest Society" initiated the establishment of a registry to congregate epidemiological data from different ILDs patients. The registry provides the necessary knowledge to identify and understand the associated risks of ILDs for the improvement of ILD management. The registry has been published in 2017 and collected data from different Indian states.

ii. Online Graphics Communicator (OGC) [98]

This is a tool of Telemedicine Server, iMediK4, developed at Indian Institute of Technology (IIT), Kharagpur, India. This system provides an online session for doctors to connect remotely. During such sessions, doctors can access patients' images and perform various operations. Currently, OGC is being used in a test bench environment with iMediK server at IIT Kharagpur as well as in another installation at Calcutta Medical College and Hospital, Kolkata, India. A version of OGC has been integrated with the iMediK server and hosted on http://tmportal.iitkgp.ernet.in/. This system currently deals with still images and image profiles only.

iii. IIIT-Delhi Contact Lens Iris Database (CLI) [99]

IIIT-D CLI database contains total 6570 Iris images (both left and right) from 101 subjects. There are 202 Iris classes and CLI that contain at least three images for every Iris class.

iv. National Cancer Registry Programme (NCRP) [100]

The NCRP was commenced in December 1981 by the "Indian Council of Medical Research (ICMR)" with cancer registries network of across the country. The main objectives of this program were to generate reliable data on the magnitude and patterns of cancer.

1.10 Databases and their impact on diagnosis

ILDs are a less understood group of diseases with a heterogeneous set of symptoms and conditions. The amount and quality of the data are very crucial in the performance of CAD implementations and evaluations. To augment the quality of care of patients, an enriched and population representative case library of the previous data should be available [51]. As the advancement in the technologies has been made, different modalities are introduced in the systems in digital form. This data needs to be collected and stored in an appropriate manner. This gives rise to the need of a database. Many commercial and non-commercial agencies are implementing a database system for the management, storage, and retrieval of the data.

An improved understanding of the disease and its condition comes from the variety of case studies for the disease. In order to understand disease process, primary and necessary requirement will be a collection of a sufficient amount of data that can cover the maximum spectrum of the disease. Sometimes doctors encounter atypical and rare cases which could aid in the study reference process of the disease. A multimedia database enriched with multiple

case studies and different dimensions of the disease understanding process could provide a better understanding of the disease.

In common practice, instrumentations and treatment process followed by healthcare providers is almost the same. However, the management and learning from the previous case studies differ in large scale in comparison of established developed countries versus developing one. This factor plays a key role in the qualitative and efficient treatment process. This process could only get better when we have sufficient amount of data and knowledge in a proper protocol of access and storage. Access to diverse representation as well as well-characterized imaging data remains a common limitation [51]. The storedsets of information not only play an important role in the disease understanding process but also aid in the development and evaluations process of new technologies and software. A well-characterized set of data enable early identification of abnormalities and progression of disease [101].

In developing countries and commercial research organizations, the research and clinical data have already been captured in databases. These datasets are increasingly being used in the development and improvement of different projects, drug or methodologies convertible to market shares. However, the majority of the real world patient data is still unreachable to a large number of researchers or medical practitioners due to privatization and/or high subscription costs. While the availability of these datasets in the public domain can be used to advance healthcare processes, adherence to standard clinical practice and guidelines, and cost of healthcare is important [102]. These databases play an imperative role in medical research by providing training data and platform for small and randomized clinical trials, prioritization of new drug targets and improvement of current methodologies. A comprehension of genomic, biomarker, clinical and imaging data together can improve the research opportunities. Keeping these points into considerations, the goal of presents thesis works were designed. Greater details of the importance of the thesis work are discussed in following sections.

1.11 Goal of present research

In recent years, thoracic societies recommendation of a multi-disciplinary discussion based diagnostics (MDD) has been clearly identified [103]. These approaches rely on a cumulative evaluation of radiological patterns, clinical history, and pathological test results. After 2008, ATS recommended the use of SLB to confirm the histological diagnosis [14]. Again these tests are critical with associated risk factors, and very well trained practitioners are required

to perform the task. Furthermore, sampling results of the body fluid indicated the important role of proteins secretion and their alteration in ILDs. These proteins are also present in bloodstream and serum even in small amount. These results suggest that genetic analysis of the protein, blood serum, and associated genetic factors might represent a novel approach in the assessment of ILDs [3]. The genetic testing is in common practice in many countries; however, the proper protocol to implement is not yet available. “The American Thoracic Society, European Respiratory Society, Japanese Respiratory Society, and the Latin American Thoracic Association statement” for interstitial pulmonary fibrosis advocates genetic testing [26].

The successful establishment of these practices relies on the availability of biological and clinical data from public resources. There are very less specialized databases/datasets available for ILDs in public domain. Thus, the needfor centralized databases and analysis for both genetic and clinical aspects of ILDs can be clearly identified. As stated before there is limited availability of ILDs medical data and analysis in public domain, a wide-range meta-analysis of differential/representative radiological and clinical patterns from the northwestern Himalayan region is performed. A comparative analysis among the northwestern Himalayan region, other Indian published study and ILDs referral dataset from western countries is also performed. Additionally, keeping in considerations the role of genetics in disease diagnosis, ILDs’ genetics was also explored. To aid more insight into the topic, enrichment analysis using pathways and gene ontology (GO) is performed, and the role of mutations and ncRNAs is demonstrated. The overarching goal of the genetic analysis is also the identification and integration of ncRNAs biomarkers to assist ILDs diagnosis, prognosis, and management.

In this thesis work, we mainly aimed to establish the comprehensive knowledge resources generated by combining radiological, clinical and genetic data of ILDs. We have developed two different web resources ILD-DB (radiological data, clinical data, symptoms, history, etc.) and ILDgenDB (genetic data of ILDs). Furthermore, based on ILD-DB, common clinical and radiological patterns are demonstrated. Identification of molecular biomarkers is performed using genetic analysis of ILDgenDB data. This study can be utilized to advance the diagnostic of ILDs. A pictorial representation of the research objectives and workflow has been given in Figure 1.5.

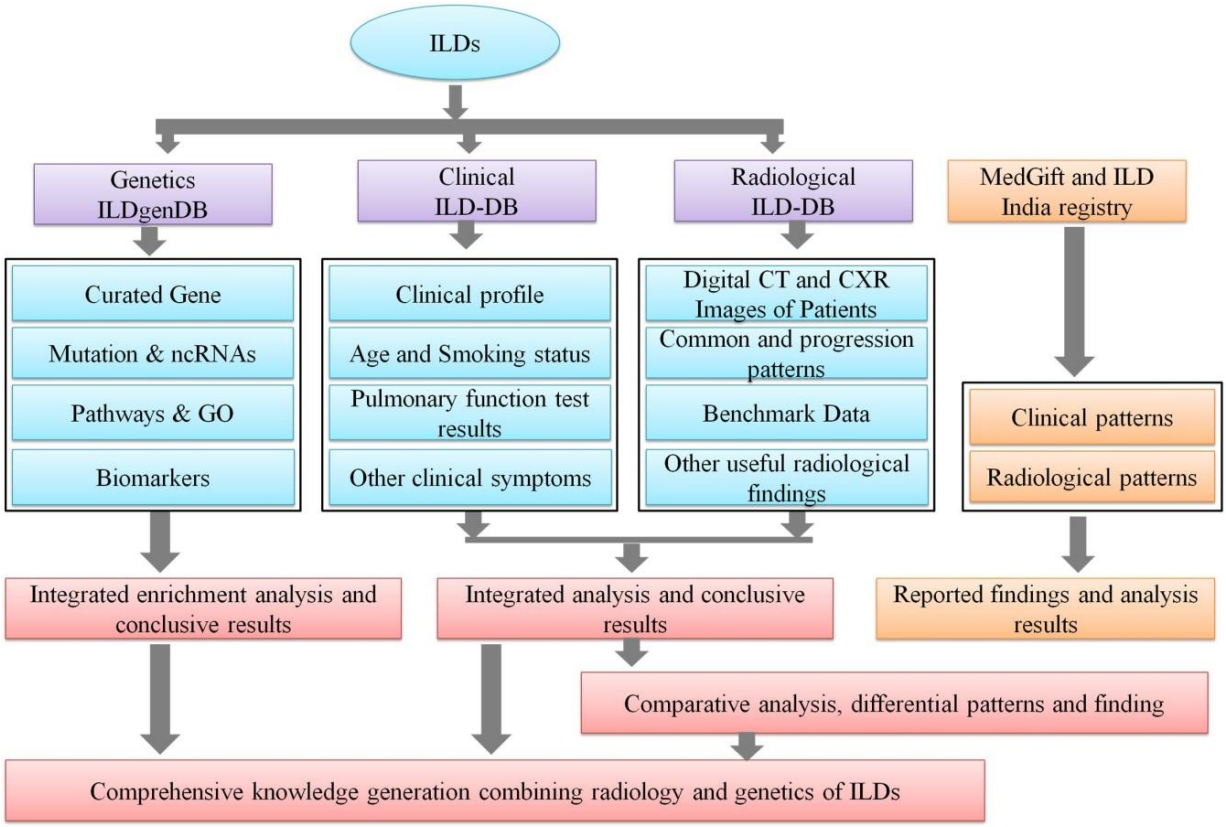


Figure 1.5 Workflow to establish comprehensive knowledge generation combining radiology and genetics of ILDs

1.12 Objectives of the thesis

Following three objectives have been accomplished to address the current limitations.

1.12.1 Development of radiological and clinical information based integrated resource for Interstitial Lung Diseases (ILDs): ILD-DB

The main objective was to create an integrated resource for ILDs referential data set that will be a collection of lung HRCT, 2D X-ray and other supporting information that can be used to develop decision support mechanism for ILDs' detection and determination of image lesions. Main utilizations of this resource are as follow:

i. This reference database could work as a uniform established system in order to provide anintegrated resource for point to point access that could aid in diagnostics process in the nationalized health services.
ii. This resource can be used for the evaluation of different CAD systems. Data can be directly uploaded to match and correlate the significant characters for a particular medical condition to check for presence or signs of disease. Quantitative evaluation and images' retrieval can be done for images similar to those of unknown lesions.

iii. It could be used by a radiologist for disease and other patient-related information from a radiological training point of view that could follow uniform public health recommendations.
iv. Query systems could be used to provide ILDs related data at an efficient customization level according to user's requirement.

Different levels of information have been added with radiological patterns correlated with other diagnostics reports and test data. This proposed database will be a platform and workbench for the development of new algorithms and tools for understanding and analyzing image modalities. Researchers can utilize this data and analyzed results to validate newly developed hypothesis and techniques. Developers can use this data as test material to build new analysis tools and techniques for developing and validating algorithms.

1.12.2 Development of genetics based integrated resource for Interstitial Lung Diseases (ILDs): ILDgenDB

ILDgenDB is a knowledgebase that integrates different types of functional and structural genetic annotation and their interactions with sixteen different classes of ILDs. The ILDs associated candidate genes (DCGs) are curated from published articles. The DCGs are annotated for different genetic information, disease relationships, SNPs, gene function, and gene ontology. Several molecular biomarkers are also curated from literature and provided in this resource. Additionally, novel biomarkers are also proposed based on the analysis performed on genetic data of ILDgenDB. Keeping in the vision of the substantial utility of the genetic data, this study discusses the major applications of ILDgenDB, a highly curated comprehensive knowledgebase for ILDs diagnostics and therapeutics.

1.12.3 Identification of noncoding-RNAs (ncRNAs) and differential pathways as potential diagnostic biomarkers for ILDs pathogenesis

The third objective of this thesis work focuses on identification of ncRNAs and associated pathways as a potential biomarker for ILDs diagnostics. The potential miRNAs and lncRNAs in ILDs pathogenesis and associated pathways were first identified with the help of the previous objective. Integrated and association analyses of these potential ncRNAs with genes and pathways involved in ILDs are performed. Additionally, enrichment analysis of the ncRNAs using ontology analysis, protein class mapping and pathways association are also performed. The miRNAs altered expression level (Disease versus control) is also determined for ILDs pathogenesis. The results of this objective provided the significant association

among genes, ncRNAs and pathways involved in ILDs. This study may assist to establish ncRNAs based biomarkers and therapeutic targets in clinical practice for better ILDs management.

1.13 Organization of the thesis

The thesis is structured into five different chapters as follow.

In this chapter (Chapter 1), the introductory part and the literature's review of the work is presented. In Chapter 2, a newly developed integrated resource of radiological and clinical information for interstitial lung diseases (ILD-DB) has been presented which can provide the radiological assistance and diagnostics of ILDs' patients. In Chapter 3, first integrated resource of genetic data for interstitial lung diseases (ILDgenDB) has been presented, and systematic analysis on "literature curated disease candidate genes" and their association with different regulatory molecules is also discussed. In chapter 4, the novel biomarkers, differential pathways and the role of noncoding RNAs in ILD pathogenesis have been identified and presented. In chapter 5, problems addressed by each objective are mentioned. Also, future work is proposed to extend this thesis work.

CHAPTER 2

INTEGRATED MEDICAL DATA RESOURCE AND META-ANALYSIS FOR DIFFERENTIAL PATTERNS IN INTERSTITIAL LUNG DISEASES (ILDs) FROM NORTH-WESTERN HIMALAYAN REGION: INTERSTITIAL LUNG DISEASE-DATABASE (ILD-DB)

Summary

Interstitial lung diseases (ILDs) are chronic disorders characterized by a variable amount of inflammation and/or fibrosis and associated with substantial morbidity and mortality due to misdiagnosis or delayed diagnosis. Clinical symptoms and radiological tests such as high resolution computed tomography (HRCT) and chest X-ray (CXR) hold more promises in developing countries due to their availability. To facilitate effective diagnosis through comparative analysis of these patterns, a comprehensive web-based resource of ILDs (ILD-DB) has been established from the North-Western Himalayan region. Dominant radiological patterns and their distribution pointing to disease presentation were manually annotated and validated with the help of radiologists. A total of 165 radiological profiles, 128 clinical profiles, and follow-up data from eleven ILD-subtypes including, a rare disease, pulmonary alveolar microlithiasis (PAM) were anonymized and incorporated in ILD-DB. Idiopathic pulmonary fibrosis (IPF) and connective tissue disease-associated ILDs (CTD-ILDs) were the most reported diseases in this study. Honey-combing and intra/interlobular septal thickening patterns were predominant in the reported cases. This study would be beneficial for characterizing new-onset ILDs, benchmarking of referral dataset, image processing algorithm development and radiological training. The ILD-DB is accessible as a public repository (http://14.139.240.55/ilddb/).

2.1 Background

Interstitial lung diseases (ILDs) primarily affect the interstitium of the lungs but may also involve the airway, pulmonary vasculature, and other structures. These diseases may be acute, sub-acute or chronic, and may be reversible or irreversible. The diseases majorly cause

variable degrees of inflammation and fibrosis, and can be broadly classified into following four categories [57]:

i. Idiopathic interstitial pneumonias (IIPs) (*e.g.* idiopathic pulmonary fibrosis (IPF), non-specific interstitial pneumonia (NSIP), etc.) [57].
ii. ILDs with a known cause such as drugs, environmental exposure (*e.g.* connective tissue disorder related-ILDs (CTD-ILD), silicosis, etc.);
iii. Granulomatous ILDs (*e.g.* sarcoidosis);
iv. Rare ILDs with well-defined clinical-pathological features (*e.g.* pulmonary langerhans cell histiocytosis and lymphangioleiomyomatosis)

IPF is an example of a chronic fibrosing type of IIP, characterized by a pathological and/or radiological usual interstitial pneumonia (UIP) pattern. The IPF occurs mainly in elderly people and adults, particularly in male smokers. Another variety of chronic fibrosing IIP is NSIP. Apart from being idiopathic, the NSIP pattern can also be seen in several other conditions like collagen vascular diseases (CVD), hypersensitivity pneumonitis (HP), and drug toxicity. Thus, multi-disciplinary discussions (MDD) involving clinicians, radiologists and pathologists are considered as the gold-standardfor ILDs' diagnosis.

In developing countries, there is always a looming possibility of under-diagnosis and under-reporting of ILDs [75]. It may be due to the lack of acquaintance among doctors, or non-availability and/or non-affordability of various diagnostic modalities like high-resolution computed tomography (HRCT) scanning, bronchoscopy and video-assisted thoracoscopic surgery (VATS) [57, 75, 76]. Another significant problem is the over-diagnosis of tuberculosis in ILD patients [104]. Though surgical lung biopsy is the gold-standard in ILDs diagnosis, lack of specialist equipment and its affordability dissuade patients to explore this test [75]. The ATS/ERS/JRS/ALAT guidelines have emphasized on evidence-based ILDs diagnosis and management and recommended that biopsy can be omitted in the case of typical clinical and radiological patterns [26]. Therefore, the clinical and radiological tests (HRCT and chest X-ray (CXR)) hold more promises in developing countries [75]. As overlapping and mimicking of the radiological patterns is frequent in ILD-subtypes (e.g. IPF, NSIP, etc.), diagnosis of any atypical or new onset presentation of patterns solely relies on assumptions or individual clinician's judgments which lead to empirical treatment and diagnostics [75]. Therefore, efficient ILD diagnosis needs to be based on integrated clinical

history and physical examinations, radiological features and pathological reports of the patient [52, 57].

Additionally, managing domestic and environmental factors, co-morbidities and mislabeled HRCT disease presentations are often crucial for accurate ILDs diagnosis. Estimation of precise incidence and prevalence, the basis of etiology, natural history and outcomes of ILDs can be acquired from integrations of the patient's data [105]. Web availability of medical registries/data could offer diverse disease presentation, and it can provide better understanding and treatment outcomes which improve clinical efficiency and reduces medication errors [16, 106].

Unlike earlier studies, the most common ILD reported in a recent study was hypersensitivity pneumonitis (HP) with 47.3% cases, followed by connective tissue disease-associated ILD (CTD-ILD) and IPF in 13.9% and 13.7% cases, respectively [105]. However, web-availabilities of these datasets are rare [74], and the role of varying altitudes in ILDs patterns have not been studied well. Studies have indicated that patterns, course, and prognosis of ILDs vary in different areas. The prevalence rate of ILDs in India is very broad, but more comparative analyses of epidemiological studies from different geographical locations are needed to record for its exact incidence and prevalence [75]. Additionally, overlapping of radiological features and lack of proper knowledge and diagnostic facilities are major reasons for misdiagnosis or diagnosis delay of ILD-subtypes at primary health care level [104].

Keeping these factors into consideration, this study integrated clinical and radiological data of ILDs patients from North-Western Himalayan regions as a repository. This resource represents a validated multidisciplinary data, and eleven ILDs-subtypes information was collected with the help of a team of expert pulmonologist and radiologists. A very rare ILD-subtype called pulmonary alveolar microlithiasis (PAM-ILD) is also incorporated into this resource [107]. Meta-analysis was performed to discover differential/representative radiological and clinical patterns observed in each ILDs-subtype. All the validated clinical and radiological data were incorporated into ILD-DB, and the public repository is freely accessible at http://14.139.240.55/ilddb/.

2.2 Materials and methods

2.2.1 Ethics Statement and study design

Approval from the Ethical Committee of Indira Gandhi Medical College (IGMC), Shimla, India was obtained with all requisite study design protocols. In the project duration, a team of three expert radiologists, one pulmonologist, and other researchers were involved in data collection and annotation, disease diagnosis and treatment regime, clinical parameters selection, questionnaire designing, data conversions, anonymization, and designing and development of database architecture.

2.2.2 Diseases selection & patient examination form

Under the guidance of specialist doctors, a comprehensive and informative questionnaire was designed to collect the radiological (HRCT, CXRs,) and clinical data from patients. In total, eleven ILDs-subtypes *i.e.* idiopathic interstitial pneumonia (IIP), namely idiopathic pulmonary fibrosis (IPF) and nonspecific interstitial pneumonia (NSIP); Granulomatous ILDs, namely sarcoidosis and hypersensitivity pneumonitis (HP); ILDs of known cause, namely silicosis, connective tissue diseases (CTDs) related ILDs (CTD-ILD) consisting of mixed connective tissue disease (MCTD-ILD), progressive systemic sclerosis (PSS-ILD), systemic lupus erythematosus (SLE) and rheumatoid arthritis-associated ILDs); Other ILDs, namely pulmonary alveolar microlithiasis-ILD (PAM-ILD) (Rare ILD), and combined pulmonary fibrosis and emphysema (CPFE) were considered. Other lung diseases such as obstructive lung diseases (*i.e.* chronic obstructive pulmonary disease (COPD) and bronchiectasis); infectious disease (*i.e.* tuberculosis (TB) and pneumonia) and other non-ILD cases were also considered. Non-ILD patients' data were labeled in the category of non-ILDs. All the patient data were acquired, validated and incorporated into the data resource.

2.2.3 Data collection

The following data were acquired from patients reported at Indira Gandhi Medical College (IGMC), Shimla, India. The architecture of ILD-DB is depicted in Figure 2.1. No exclusive HRCT, CXR and other pathological tests were performed. The patient's recruitment, data collection, and analysis for this study were performed over thirty-month. Verified datasets were included in ILD-DB resource for future references. Data collection steps were as follows:

i. Clinical data: General clinical symptoms of every patient were included. The Pulmonary function test, pathological test and diagnostic reports of patients were also collected.
ii. Digital imaging data: The HRCT and CXR of patients going through standard treatment were collected. Cases with multiple HRCT were also traced and incorporated.

iii. Data entry protocol: All the imaging data and diagnostics reports were first anonymized for patients' privacy protection. After filtration and verification, data entry was made using basic guidelines provided by Indian EHR recommendations so that the data can be retrieved and submitted through an efficient and specified query.

2.2.4 Data annotation and validation

Digital Imaging and Communications in Medicine (DICOM) images were extracted from the picture archiving and communication system (PACS), and inbuilt graphical software was used to visualize and demarcate the region of interest (ROI) in entire HRCT slices. Dominant patterns (annotated ROIs) and their distribution pointing to disease presentation were annotated with the help of radiologists. These ROIs are available in 2D image format (JPEG) with a resolution of 1500 X 700. For better representation of the ROIs, online image viewer is implemented in ILD-DB (Figure 2.2). Cases with multiple HRCT were traced and incorporated as temporal data. Data validation was conducted manually by expert doctors and radiologists in a time-to-time manner as data were captured for the study.

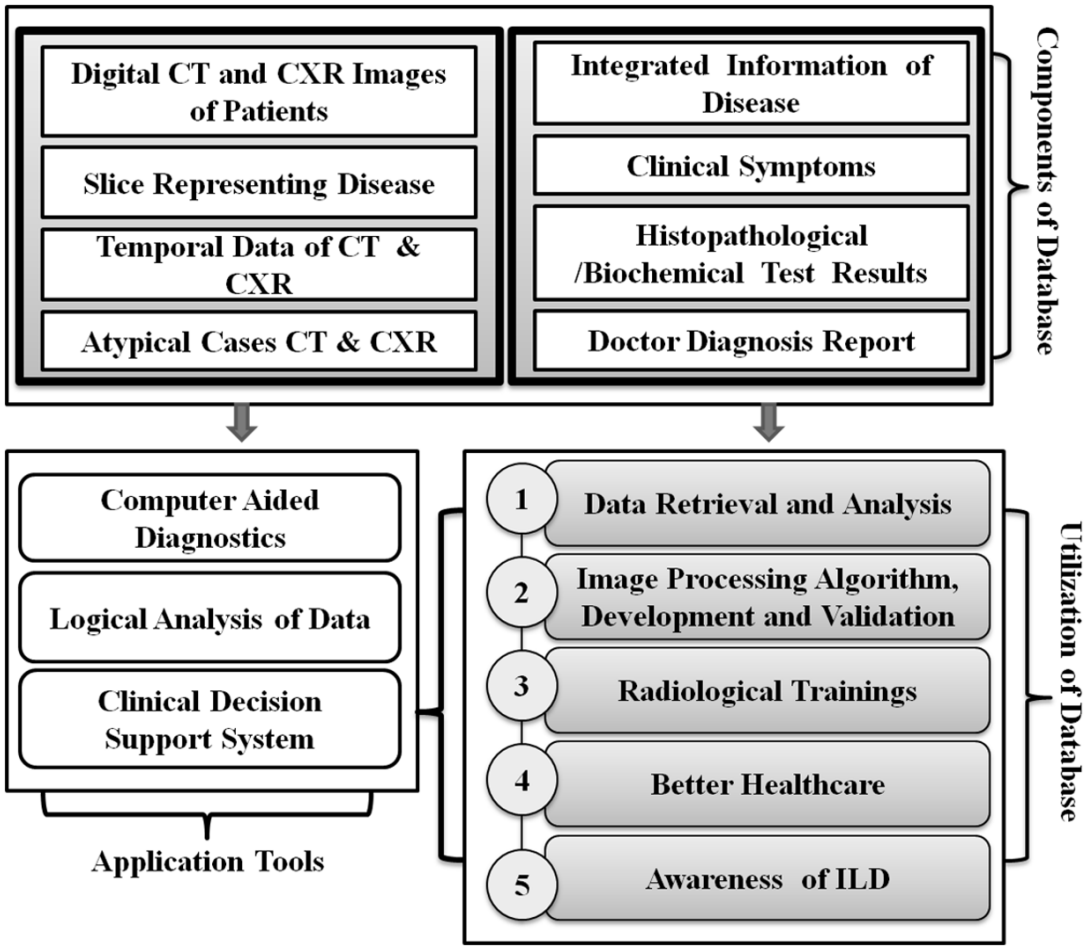


Figure 2.1 Architecture and potential applications of the ILD-DB resource

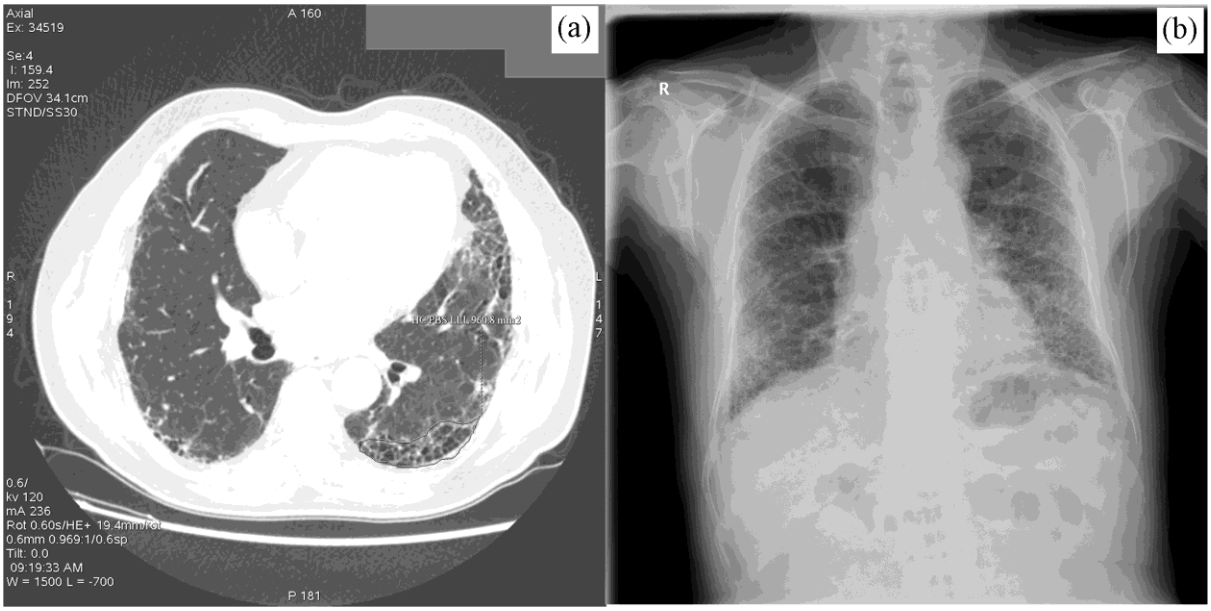


Figure 2.2 Representation of HRCT and CXR of IPF patient. (a) Marked HRCT image showing subpleural reticular shadow and honeycombing, and (b) CXR image showing interstitial thickening and ground-glass opacity in bilateral lung fields predominantly in lower zones.

2.2.5 Integrated ILD resource development and data access

Architecture and application of the proposed ILD-DB unified resource are provided in Figure 2.1. The HRCT-DICOM images, CXRs, ROIs and clinical data were stored in the database using MySQL server. HTML, PHP, and JavaScripts were used for graphical user interface (GUI) development. This resource facilitates advance options for efficient retrieval of the data i.e. “Browse” and “Query” (Figure 2.3, A and B). More options to extract relevant information are discussed in Section 3.2.Detailed information and data accessibility options can be explored from ILD-DB tutorials accessible at http://14.139.240.55/ilddb/tutorial1.php.

2.3 Results and discussion

This study presents a knowledge-resource and meta-analysis of different ILD-subtypes based on clinical and radiological patterns of patients. The meta-analysis presents a comparative analysis of symptoms and radiological patterns of this resource with earlier published studies from India and MedGift group [51]. All the data were systematically collected, analyzed and provided in the form of EHR-enabled web resource named ILD-DB. Currently, this resource describes 238 cases with different lung diseases from the different parts of North-Western Himalayan region (Section 2.1). Out of the total 238 cases, following ILD sub-types consisting of 128 patients were reported: CPFE, HP, PAM, IIP including IPF and NSIP, sarcoidosis, silicosis, CTD-ILD including MCTD-ILD, PSS-ILD, SLE and RA-ILD, and Unclassifiable-ILD (Figure 2.4, A) (Table 2.1). Rest of the cases (110) belongs to obstructive

lung diseases (*i.e.* COPD and bronchiectasis), infective diseases (*i.e.* TB and pneumonia) and other non-ILD (Figure 2.4, B).

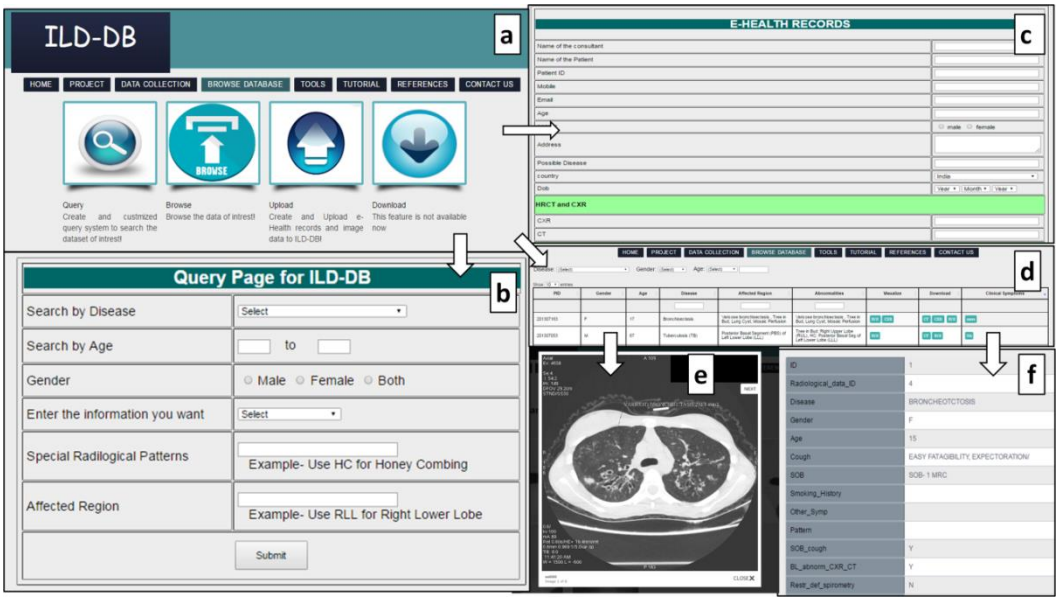


Figure 2.3 Overview of ILD-DB resource. (a) Access of information from ILD-DB using 'Search', 'Query', 'Upload' and 'Download' options; (b) Accessing different categories of data using 'Query' options, data associated with disease by choosing 'Disease', 'Radiological Patterns' and/or "Affected Region'; (c) ILD individual patient data submission through "Upload" option for new cases using EHR form; (d) Patients data in the tabular format with hyperlinks to access more specific data for a single patient.; (e) Visualization window for marked radiological patterns of HRCT; and (f) Clinical information of ILDs patient.

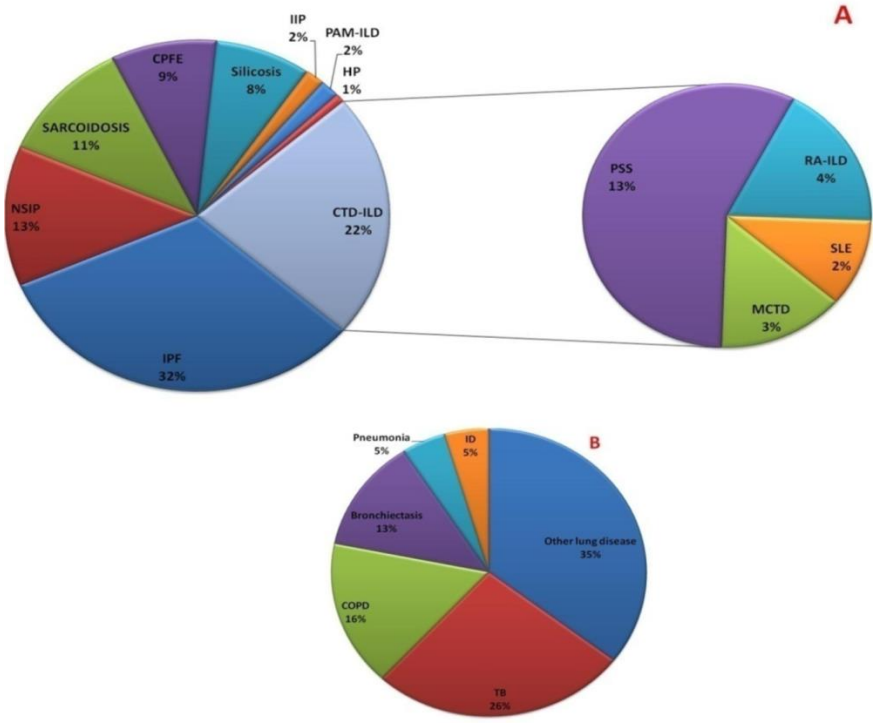


Figure 2.4 Distribution of the total cases included in ILD-DB. (a) 128 cases distribution among different ILD-subtypes and the enlarged portion indicates sub-grouping within the CTD-ILD and (b) 110 cases distribution among obstructive lung and infective diseases.

Table 2.1 Clinical pattern analysis of different ILD-subtypes present in ILD-DB.

S. No.	ILDs	Subgroups*	Number of Cases	Mean Age#	% Female#	% GERD*	% Clubbing *	Mean (FEV1/FVC) *	Mean FVC*	% of Smoking history
1	IIP	Unclassifiable IIP	2	68	50	50	50	NA	NA	50
2		IPF	40	61.7 (71.4)	27.5 (50)	35.5	46	93.7	64.32	38.4
3		NSIP	15	52.2 (61.5)	81.2 (50)	40	50	89.5	65.2	NO
4	ILDs of known cause	Silicosis	10	51.7	0	40	40	77.3	73	40
5		PSS-ILD	17	51.2	80	69.23	25	0.8	60.68	NO
6		RA-ILD	5	54.6	60	75	0	84.9	55.26	33.3
7		MCTD	3	43.6	100	100	66.6	80.1	67	NO
8		SLE	3	48	33.3	50	0	62	62	NO
9	Granulomatous ILDs	Sarcoidosis	14	45.1 (48.5)	78.5 (30)	47	66.6	83.9	70.1	50
10		HP	1	32 (65.3)	0 (80.3)	0	0	NA	NA	NO
11	Other ILDs	PAM	3	49	100	NA	NA	NA	NA	NA
12		CPFE	11	69.9	10	44.4	50	90.3	90.33	74.9
13	Unclassifiable-ILDs		5	64.8	80	60	33.3	85.5	63.3	NO

Value in parentheses represents number in previous study [51]; * NA/NO: Clinical history was normal or test was not performed for the disease. GERD: Gastroesophageal reflux disease.

2.3.1 ILD-DB resource

2.3.1.1 Data access

ILD-DB provides an extended interface for efficient and customized retrieval of data through "Query", "Browse" and "Download" options (Figure 2.3, A). The image (HRCT, CXR and ROIs) data is also available for download (Figure 2.1). The resource can be queried by "ILD subtype" and/or radiological pattern information. For example, the user can use disease name "Idiopathic Pulmonary Fibrosis" or a pattern name "Honeycombing" or the abbreviated form "HC" from "Search by Disease" or "Specific radiological feature" option, respectively (Figure 2.3, B). Users from the different locations can deposit the ILD data by using 'Upload' option (Figure 2.3, C). Output page (Figure 2.3, D) gives detailed information about the patient's data i.e. clinical symptoms and radiological features of patients which may be retrieved by clicking on "CT", "CXR" or "ROI" buttons present in "Download" column (Figure 2.3, D). Marked patterns of each patient can be visualized by clicking on "ROI" (Figure 2.3, E). Detailed information related to clinical symptoms can be viewed in a

different window (Figure 2.3, F) by clicking on “more” tab provided in the “Clinical Symptom” column.

2.3.1.2 Statistics

In total, 238 cases are captured in ILD-DB resource. Out of these cases, 165 contained an annotated HRCT corresponding to the period of this study. In total, 128 ILDs and 110 other lung disease cases are provided in this database. A total of 31 cases also contain CXRs and/or raysum (2D presentation of CT) of same patients. In total, thirteen types of textural HRCT patterns containing 242 ROIs were delineated in 149 HRCT image series (Table 2.2). These distinct ROIs were marked in each HRCT image (Figure 2.3). Currently, in ILD-DB, IPF is the most common ILD with 40 (31.0%) patients, followed by CTD-ILDs, idiopathic NSIP and sarcoidosis with 28 (21.7%), 15 (11.6%) and 14 (10.85%) patients, respectively. A rare ILD, pulmonary alveolar microlithiasis (PAM) was also reported in 3 (2.3%) patients. All the cases with marked radiological patterns and diagnostics information are available in the public domain to streamline training and understanding on ILDs. This resource may also be useful for the development of novel algorithms related to computer-aided diagnosis (CAD). Comparative analysis of the radiological patterns present in HRCT/CXRs can help in the differential diagnosis of the diseases. The regions of interests (ROIs) of ILDs and other obstructive lung diseases were also included in this resource as the reference for predominant HRCT patterns.

In the following sections, a meta-analysis of radiological and clinical patterns and comparative analysis of ILDs cases are presented.

2.3.2 Meta-analysis of ILD cases

The distribution of clinical and radiological characteristics of observed patients are described in Table 2.1, Table 2.2 and Figure 2.4. In total patients of ILD-DB, honey-combing (HC) and intra/interlobular septal thickening (ILST) were observed predominantly. In total, 46% of patients have dry eye/mouth and heartburn symptoms, and 35.5 % of patients were observed to have gastroesophageal reflux disease (GERD). Studies have reported a close association and higher prevalence between GERD-ILDs patients in comparison to normal GERD patients. However, more evidence-based studies are needed to establish a causal relationship between GERD and ILDs [108]. Textural patterns exhibited in the HRCT of the lung are very useful for the differential diagnosis of ILDs. The locations of fibrosis in the apical and anterior segment of the right lower lung, right middle & left lower lung, and bilateral upper lobe indicated sarcoidosis, IPF, and CPFE, respectively. Thus, the characteristic location of a radiological pattern is also very helpful in differential diagnosis. However, some patterns are ubiquitous and not uniquely related to any particular ILD-subtypes; therefore,

Table 2.2 Comparative analysis of HRCT patterns in ILD-DB with other studies

S No.	ILDs	Subgroups*	Depeursinge et al., (2012)[#] [51]	Singh et al., (2017) [#] [105]	ILD-DB[#]
1	IIP	Idiopathic pulmonary fibrosis (IPF)	Fibrosis, Bronchiectasis, GGO Peripheral, Subpleural, Basal, Posterior	UIP (Subpleural and basilar (P), Reticular abnormality, Honeycombing,	Honeycombing (F), Bilateral lower lobe (P), Intra and interlobular upper lobar septal thickening (F), Fibrosis, Cyst, minimal GGO (Bilateral Lower Lobe (P)), Traction bronchiectasis.
2		Non-specific interstitial pneumonia (NSIP)	GGO, Consolidation, Reticulation, Fibrosis (uncommon) Peripheral, Basal	Basilar GGO (P) with or without sub-pleural sparing, Reticular abnormality (infrequent)	Intra and interlobular upper lobar septal thickening, GGO (Bilateral lower lobe, Right middle lobe), minimal Honeycombing (Right lower lobe), Nodules (Right middle lobe, Right lower lobe)
3	ILDs of known cause	Silicosis	NA	NA	Fibrosis or progressive massive fibrosis (F), Nodules (BUL), Alveolar opacities, Plural thickening, Intra and inter lobular upper lobar septal thickening, Bulla, Fibrocalcification
4		Progressive systemic sclerosis associated ILDs (PSS-ILDs)	NA	Interstitial pneumonia patternscompatible with CTD: UIP, NSIP, LIP, OP	Intra and interlobular upper lobar septal thickening, GGO, Honeycombing and Emphysema
5		ILD secondary to collagen vascular disease, MCTD	NA		Intra and interlobular upper lobar septal thickening, GGO
6		Systemic lupus erythematosus (SLE)	NA		Intra and interlobular upper lobar septal thickening, Alveolar infiltrates (AI)
7	Granulomatous ILDs	Hypersensitivity pneumonitis (HP)	GGO, emphysema, Fibrosis, Diffuse	Upper lobe GGO (P), poorly defined centrilobular nodules; mosaic attenuation	Intra and interlobular upper lobar septal thickening & GGO (Diffuse)
8		Sarcoidosis	Micronodules, Consolidation, Macronodules, Ground glass, Fibrosis (end-stage), Peribronchovascular, Subpleural, Peripheral	Upper lung perilymphatic distribution of nodules (P) with or without mediastinal or hilar lymphadenopathy	GGO (F), Mediastinal and hilar Lymph adenopathy (F), Tree in bud, Cyst, Intra and interlobular upper lobar septal thickening, Fibrosis, Traction bronchiectasis
9	Other ILDs	Combined Pulmonary Fibrosis and Emphysema (CPFE)	NA	NA	Fibrosis (F), Emphysema (F), Traction bronchiectasis, Honeycombing

[#]GGO: Ground glass opacity F: Frequent pattern, P: Predominant location. [*]Desquamative interstitial pneumonia (DIP), acute interstitial pneumonia (AIP), cryptogenic organizing pneumonia (COP /BOOP), lymphocytic interstitial pneumonia (LIP) and langerhans cell histiocytosis (LCH) were excluded from this study as no cases were reported.

critical evaluation of clinical symptoms and radiological patterns is required for confirmatory diagnosis. For example, progressive massive fibrosis alongside the occupational history of stone crasher confirms the diagnosis as silicosis. Comparative analysis of radiological patterns in different lung locations and clinical manifestations for all four classes of ILDs (i.e. IIP, ILDs of known cause, granulomatous, and others ILDs) is performed. Comparative analysis of radiological patterns and their distribution in different lung locations is presented in Table 2.2.

2.3.3 Clinical and radiological patterns analysis of ILDs cases

Clinical and radiological features of different ILDs cases are investigated in this study and presented in the following subsections. Furthermore, radiological features associated with different ILDs and other lung diseases were reported in the various literature. In this study, the classification of ILDs based on reported radiological patterns along with their location and predominance are presented in Figure 2.5.

2.3.3.1 Idiopathic interstitial pneumonia (IIP)

*i. Idiopathic pulmonary fibrosis (IPF)***:**Prevalence of IPF in this study is 32% (Table 2.1). Majority of the IPF patients are male with a mean age of 62 and found to have a history of gradually progressive shortness of breath (SOB). Clubbing is present in 46% of patients. Bi-basilar crackles are present in 83.3% of patients. GERD symptoms are observed in 35.5% of patients in this study, while Singh et al. reported that 43.3% of their patients had GERD symptoms [105]. Radiological presentation of IPF is recognized by peripheral and subpleural lower lobe reticular opacities in association with subpleural honeycombing. In this study, honeycombing (64%), and intra and interlobular septal thickening (ILST)(28%) are the most frequent patterns. Cyst, ground glass opacity (GGO) and emphysema patterns have medium predominance. Cavity, mosaic perfusion, and traction bronchiectasis patterns are also observed but rare in IPF patients.

*ii. Nonspecific interstitial pneumonia (NSIP):*Higher prevalence (81.25%) of female patients is observed for NSIP in this study. GERD symptoms (40%) are relatively higher than the previously published study (36.6%) [105]. Joint pain is observed in 60% of patients, and 53.3% of patients are exposed to dust, mold and/or coal smoke. No correlation between tobacco smoking and NSIP is observed; however, 53.84% of patients were exposed to fire-wood smoke. Clubbing is present in 50% of patients. Bibasilar-crackles are present in 30% of patients. In the radiological presentation, ILST, GGO, and nodules are commonly seen patterns.

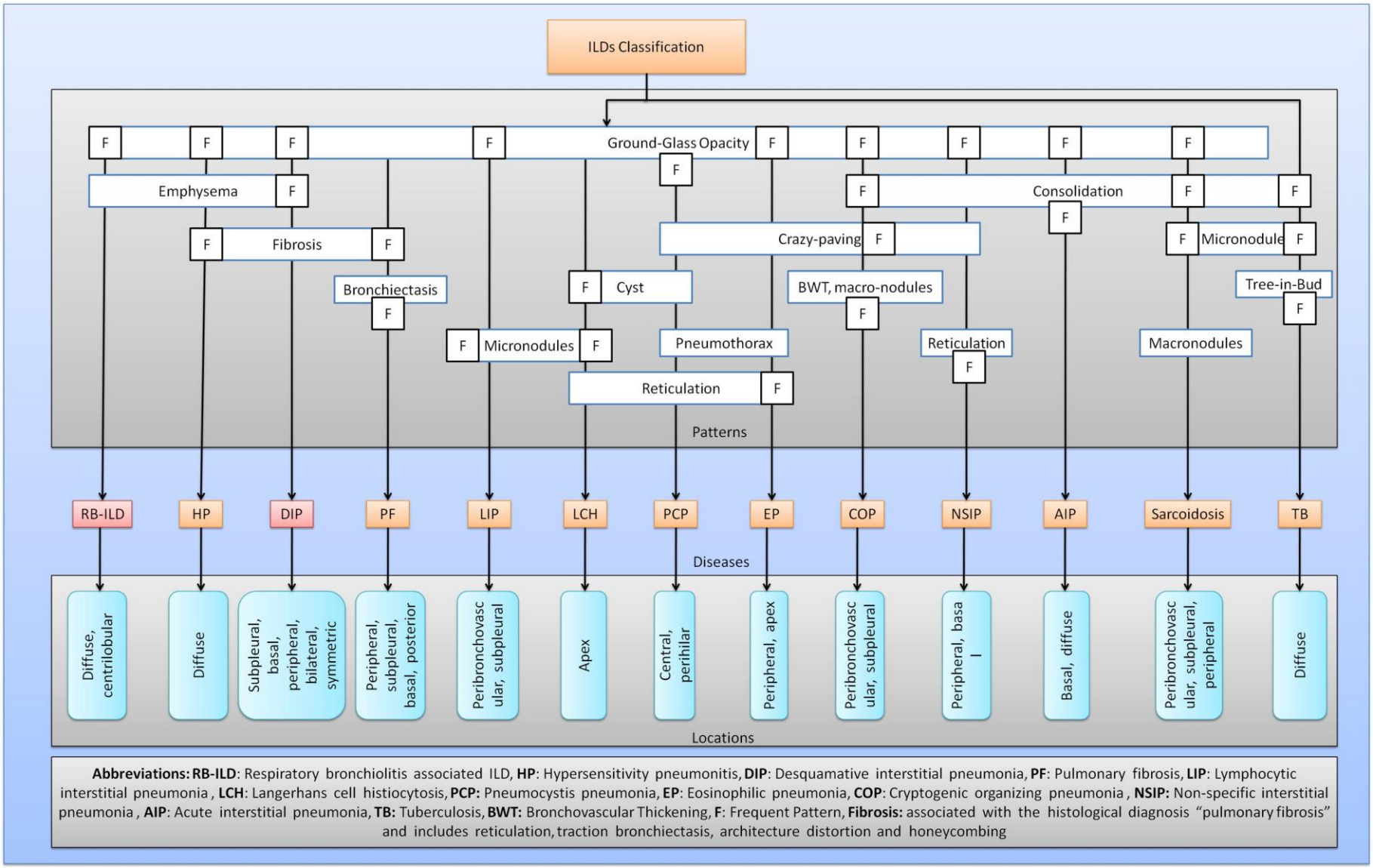


Figure 2.5 ILDs classification based on radiological (HRCT) patterns, their locations and predominance.

2.3.3.2 ILDs of known cause

i. Silicosis: In this disease, all the patients are male in ILD-DB database. Cough with expectoration is observed in 50% of patients while others are having a dry cough. Smoke exposure (fire-wood cooking) are found in 60% of patients, and 35% are exposed to visible dust or molds. All patients have worked in stone mines or related occupations, and exposed to the dusty environment. In HRCT, fibrosis patterns are present in all the patients. Progressive massive fibrosis, nodules, alveolar opacities, plural thickening, ILST, bulla, and fibrocalcification are also observed in a few patients.

ii. Connective tissue disease related ILDs (CTD-ILDS): In total, 28 cases of CTD-ILDs are reported to the hospital during the study. PSS-ILDs (17) are the most common followed by RA-ILDs (5), MCTD (3) and SLE (3), respectively (Figure 2.4, A).

a. Progressive systemic sclerosis-associated ILDs (PSS-ILDs): Female predominance is observed in PSS-ILDs. Raynaud's phenomenon is also observed in patients (68%). Exposure to fire-wood smoke (73.3%), and dust and mold (53.3%) are also recorded. Honeycombing (100%), intra/interlobular septal thickening (50%) and emphysema (33.3%) are the most common pattern detected in PSS-ILD. UIP patterns are observed in 33% of patients.

*b. Mixed connective tissue disease (MCTD-ILD):*In MCTD-ILD, 100% of patients are female. UIP patterns are observed in 50% of patients. Joint pain and Reynaud's phenomenon are observed in 100% and 50% of patients, respectively. All the patients are exposed to heavy smoke. In HRCT patterns, intra/interlobular septal thickening and alveolar infiltrates are observed.

*c. Rheumatoid arthritis-associated ILD (RA-ILD):*RA-ILD was the most common and specific CTD-ILD reported in the earlier study [105]. In contrary to the previously published study (29% female) [109], female predominance (60%) are reported in the present study. UIP pattern is observed in 60% of the patients.

*iii. Systemic lupus erythematosus (SLE):*Comparatively late onset (48 yrs) of disease is observed in this study compare to the previous one [109]. In HRCT, GGO and ILST pattern are observed, while in the previously published study [109], fibrosis was present in 63% patients.

2.3.3.3 Granulomatous ILDs

i. Sarcoidosis: Female predominance (78.57%) is observed in this study, and it is comparatively higher than previous studies (60% & 50% in [105] and [109], respectively). GERD symptoms are also observed in a higher number of patients (47 %) than earlier study

(33%) [105]. In total, 80% of patients are observed with significant weight loss. Joint pain is observed in 60% of patients, whereas 40% of patients have dry mouth. A total of 50% of patients are exposed to smoke from firewood, and 60% of patients are having pet animals or birds. In radiological patterns, hilar and mediastinal lymphadenopathy, GGO and parenchymal nodules are predominantly seen, while reticular/reticulo-nodular pattern was predominantly marked in previous studies [109].

ii. Hypersensitivity pneumonitis (HP): In HP, shortness of breath, night fever, and exposure to smoke are observed, which is similar to the earlier study [109]. The HP was diagnosed in maximum patients in ILD-India registry, and associations with air-cooler were delineated [105]. Only one case in the duration of our study is reported to the hospital. The patient is found to have a heavy smoking history. In HRCT, ILST and GGO (diffuse) patterns are marked.

2.3.3.4 Other ILDs

i. Combined pulmonary fibrosis and emphysema (CPEF): In CPEF, 100% of patients are observed with dry mouth. In total, 57.14% of patients are observed with leg edema and heartburn. Exposure to fire-wood smoke is observed in 83.3% of patients, and 66.6% of patients are tobacco smokers. Results indicate an association of smoke with CPEF etiology, and 44.4% of patients have previous medical history of TB, COPD and/or GERD. In HRCT, fibrosis and emphysema patterns are present in all the cases.

ii. Pulmonary alveolar microlithiasis (PAM): Three cases of PAM are observed in this study duration. In chest radiograph, numerous small dense nodules scattered in bilateral lung fields, predominantly in bilateral lower zones, have been noticed (Figure 2.6). In HRCT, several small nodules are observed in bilateral lung fields (predominantly in sublpeural regions and basal segments of bilateral lower lobes) (Figure 2.7). GGO and ILST are presented in bilateral lung fields. Emphysematous changes were also observed along anterior subcostal pleura.

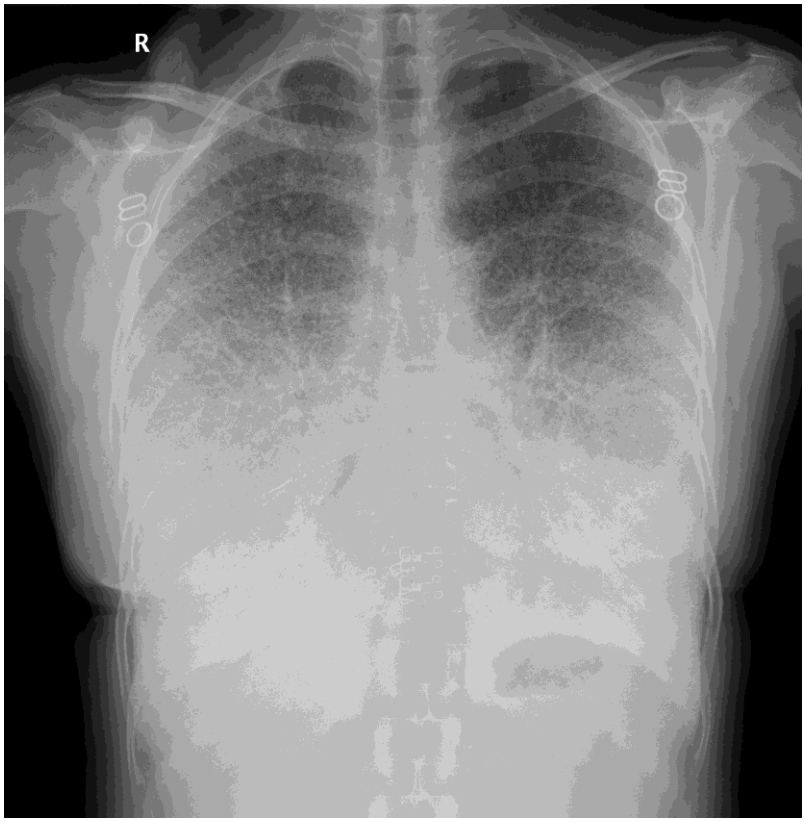


Figure 2.6 The CXR of pulmonary alveolar microlithiasis (PAM). It is showing numerous small dense nodules in bilateral lung fields (predominant in lower zones). Nodules are obscuring mediastinal borders and domes of diaphragm.

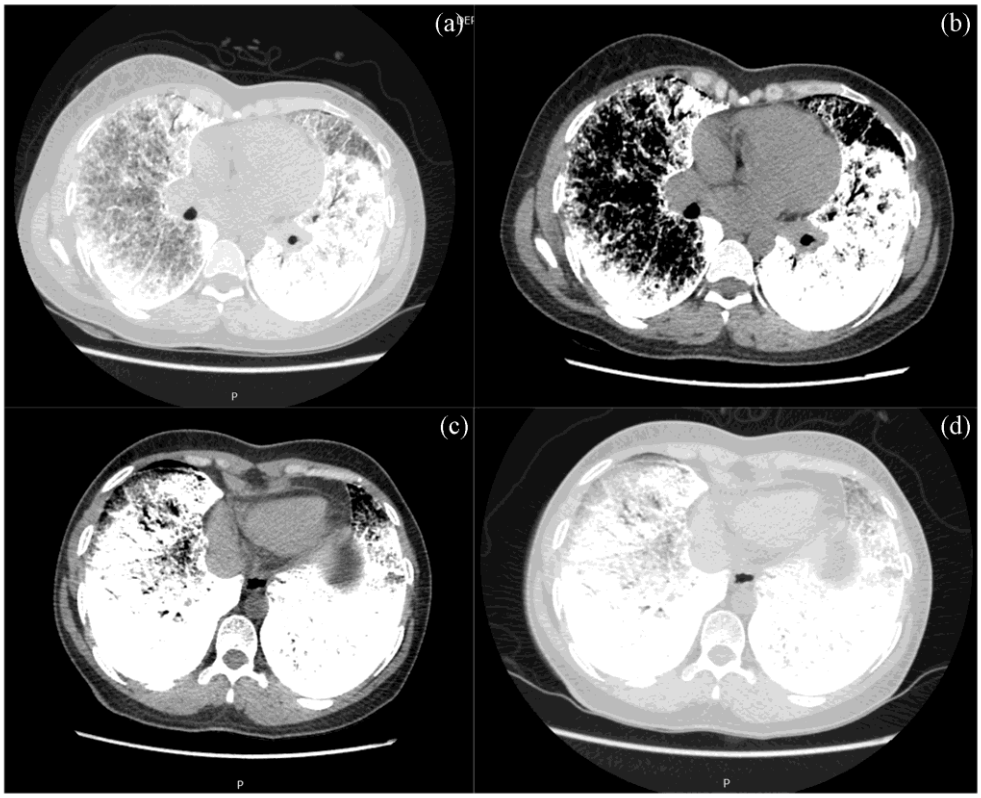


Figure 2.7 HRCT of pulmonary alveolar microlithiasis (PAM). Subpleural calcification inmediastinal window (a) and lung window (b). Confluent calcification in mediastinal window (c) and Lung window (d).

2.3.3.5 Unclassifiable-ILDs

Unspecified patterns are observed in these patients group. In HRCT, consolidation pattern is marked in the patients. The difficulty to mark these cases was acknowledged in many studies but not formally individualized. These cases are observed with a heterogeneous clinical course and corroborate the essential role of multidisciplinary evaluation in the characterization of unclassifiable disease.

2.3.3.6 Other lung diseases

Patterns of obstructive lung diseases such as COPD and bronchiectasis are also incorporated in this study. Additionally, infectious diseases *i.e.* tuberculosis (TB) and pneumonia are also included. In bronchiectasis, multiple years of shortness of breath and coughing symptoms are observed. Chest pain is observed in 80% of the patients. In HRCT of cystic bronchiectasis, tree-in-bud, cyst, and GGO are observed. In case of indistinct CXR presentation of disease, the patients are recommended to go through HRCT. In TB patients, tree-in-bud, cavity, and nodules are predominantly observed in HRCT, while emphysema and bronchiectasis are also observed.

2.3.4 Comparative analysis of lung tissue patterns and clinical features with previous studies

In this study, the eleven most common diagnosed ILDs associated HRCT patterns with their locations on lungs are presented (Figure 2.4 and Table 2.2). These findings were also compared with the previously published studies [51][105, 109][81]. Silicosis, CPFE, and subtypes of CTD-ILDs such as MCTD, PSS-ILDs and PAM are not mentioned discretely in ILD-India registry [105] and MedGIFT database [51], while the CTD-ILDs are reported with second maximum number of cases in this study. In HP ILD, pleural based soft tissues nodules are reported in this study. In NSIP, the presence of ILST, GGO and nodular patterns is predominantly observed in ILD-DB, while the absence of these patterns was marked in ILD-India registry. The ILSTs are present in 19.3% of ILD patients, while previously it was reported in 13.3% [81] and 15.1% [105] patients. In IPF, the presence of ILST pattern was often seen as early stages of honeycombing and can occasionally mimic pulmonary fibrosis patterns [52]. The GGO patterns are present in 13.8% of patient, and the predominance of this pattern over honeycombing favors the diagnosis of NSIP. Different nodular patterns (13.8%) are also observed in various ILDs. The peri-bronchovascular nodular patterns are classically seen in sarcoidosis. Different size and form of the nodular patterns can assist in ILDs'

subtype diagnosis. Tractional bronchiectasis is observed in many patients (12.8 %) of IPF. Bulla and cyst are also observed in a comparatively lesser number of IPF patients (8.2% and 6.4% respectively).

In comparison with earlier studies, no significant gender predominance is reported (54% female, 46% males) which is similar to ILD-India registry and slightly higher (36.7% female) than MedGIFT (Table 2.1). In comparison with lower altitudes ILDs cases, clubbing, and bibasilar crackles are found significantly high in this study. Furthermore, joint pain reported in this study is four times higher than previous studies. The comparatively early onset of sarcoidosis disease is observed (43.9 years). Infrequent weight-loss are reported and suggested that the feature may be uniquely co-related with altitude differences. Symptoms of GERD are reported slightly high (47.3%) in earlier studies in comparison to this study (maximum 40%). The FVC observed among patients is significantly high in the present study in comparison with earlier results [105].

As discussed earlier, three PAM cases were reported in this study. The first time reported in 1962, only a few cases (approximately, 30-50) from India and less than 800 cases from worldwide are reported till date [52, 110, 111]. Most of these cases are reported from Europe (42.7%) and Asia (40.6%). Respiratory complications are usually not common in these PAM cases and often misdiagnosis with pulmonary tuberculosis or sarcoidosis. Similar to the previous study, no significant clinical patterns were observed [112]. In this study, small nodules, GGO and ILST in bilateral lung fields are present, while in the previous study the predominant distribution of bilateral diffuse micronodular opacities over basal and posterior regions with thickening of interlobar septa were reported. Differential radiological presentations are observed in our study.

CHAPTER 3

INTEGRATED GENETIC KNOWLEDGEBASE OF INTERSTITIAL LUNG DISEASES (ILDs): ILDGENDB

Summary

Interstitial lung disease (ILD) is a broad and diverse term that represents approximately ~200 chronic and acute pulmonary disorders. These disorders can cause inflammation, fibrosis, and architectural distortion at variable extents with substantial morbidity and mortality. Delayed and inaccurate diagnoses may augment the risk, particularly in developing or low/ middle-income countries. The potential roles of genetics in ILDs pathogenesis have reported in many studies. Thus, the first knowledgebase for ILDs genetic data has been developed which also provide the results of integrated analyses for the improvement in the diagnosis process. This knowledgebase is referred as ILDgenDB and may provide a better understanding of disease pathogenesis based on proposed diagnostics-based biomarkers. The ILDgenDB provides a collection of disease candidate genes (DCGs) along with diverse regulatory elements produced by an integrated pipeline of databases searches, literature curation and association analyses of DCGs with miRNAs and SNPs using *in silico* bioinformatics methods. These DCGs are literature-curated. Hypergeomatric test along with ILD-specificity index scores are also incorporated for the statistically significant disease-gene association. Manually curated and verified DCGs (299) along with significantly associated SNPs, miRNAs and miR-polymorphisms (1932, 2966 and 9170, respectively) are incorporated. Association among DCGs, SNPs, miRNAs and pathways involved in the pathogenesis of ILDs' subtypes are also identified and incorporated. Moreover, 216 literature-curated and few proposed biomarkers are incorporated. The ILDgenDB knowledgebase offers interactive query and browsing options for easy accessibility of the data. Additionally, this knowledgebase also provides an interactive force directed graph of DCGs interactions with SNPs/miRNAs, and DCGs-ILDs association based on reported literature for each ILD's subtype. This study indicated the potential role of immune and defense system in ILDs pathogenesis. This knowledgebase may assist in better disease diagnosis and monitoring on a genetic level. ILDgenDB is available at http://14.139.240.55/ildgendb.

3.1 Background

ILDs are a diverse group of chronic and acute diffuse lung diseases, which primarily affects the interstitium of the lungs [28]. Interstitial lung diseases (ILDs) could be of known or unknown causes. Sometimes, ILDs also involve peripheral airways, alveoli, pleura and the vessels [34]. ILDs are included in the top ten reasons of increased global deaths from 1990–2013 in "Global Burden of Disease" report [113]. The ILDs' mortality rate has been significantly increased by 50% in past decades. About 15% of total cases reported by a pulmonologist are account for ILDs, and reported cases are even higher for developing and low-income countries due to excessive tobacco use, exposure to environmental hazards (i.e. asbestos), and cooking by coal. Studies have revealed that the incidence and mortality rate of ILDs increase with age, exposures and other associated risk factors. Several different studies from India in between 1979-2016 reported that majority of ILDs patients suffered from IPF and hypersensitivity pneumonitis [40, 105, 114-116].

In developing and low-income countries, the ILDs are mostly diagnosed using evaluation of patient's symptoms, multi-systemic examination, familial disease, occupational and environmental exposures and history, drug history and medical imaging (2D chest radiograph (CXR) and high resolution computed tomography (HRCT)). Multidisciplinary discussion (MDD) of radiologists, pulmonologists, and pathologists dealing with ILDs are essential for the accurate diagnosis of disease (ATS/ETS- 2004). However, MDD is not a common practice in developing countries. Misdiagnosis or wrong-diagnosis is reported in many cases that lead to delayed treatment procedure. Furthermore, over-dependency on chest radiographs is considered as an important reason for wrong-diagnosis as it cannot predict more than ten percent of ILDs cases [51]. Also, the ILDs patients are often wrongly diagnosed as tuberculosis (TB) patients [75].

The scarcity of specialized diagnostics techniques such as HRCT, bronchoscopy, diffusion capacity, spirometry, surgical or video-assisted lung biopsy, etc. are another bottleneck in accurate and early diagnosis. The lack of expert pathologists to read biopsy specimens correctly is another limitation [75]. Additionally, lung biopsies are not a good option due to co-morbidities in elderly persons. All these factors contribute to delayed diagnosis as well as delayed initiation of proper therapy. The inconsistency in the clinical course is also an important factor in ILDs diagnosis. For example, progression is often slow in some ILDs cases, while it can be rapid in others [117]. Use of molecular biomarkers in the management of ILDs may address many of these issues by providing early and accurate diagnosis by

identification of rapidly progressive phenotypes. It would also help in the better monitoring of disease prognosis. This may help in initiating early and proper treatment of ILDs in an efficient way [118].

The role of genetic inheritance is already known in emphysema, silicosis, bronchitis, cystic fibrosis [2, 36] and IPF [37]. Involvement of genes on ILDs pathogenesis, prognosis, and etiology have already verified by performing genetics studies on model animals [27]. Therefore, essential and significant molecular biomarkers can lead to the better treatment outcomes. Thus, a complete and comparative genomic study can offer molecular based candidate biomarkers as diagnostic tools. Considering the importance of genetics in ILDs pathogenesis, a literature-curated comprehensive knowledgebase, ILDgenDB, is developed. The key utilities of this knowledgebase are also discussed. ILDgenDB contains genes identified through literature curation (299), along with their functional/structural annotations, association analysis with other regulatory elements, and their involvement in ILDs pathogenesis. Gene ontology (GO) and pathways associations are performed for functional annotation and enrichment of the DCGs. Furthermore, DCGs' association with microRNAs (miRNAs) and single nucleotide polymorphism (SNP) are also predicted. Furthermore, ILDs molecular biomarkers for disease diagnosis are provided. All these analyzed data are integrated into the ILDgenDB knowledgebase. ILDgenDB facilitates cross-linking to well-established knowledgebase like NCBI, Ensembl, UniProt, PDB, dbSNP, KEGG, OMIM, etc. for advanced searching in ILDgenDB. Protocols used to identify potential molecular biomarkers for ILDs diagnosis and monitoring are also proposed. This literature curated knowledgebase would help in understanding important traits for ILDs pathogenesis. The whole database including analyzed results and proposed biomarkers can be accessed easily via browsing and extensive query options.

3.2 Material and Methods

3.2.1 Data collection and integration

3.2.1.1 ILDs Genes identification

Graphical representations of genetic data incorporated in the ILDgenDB are depicted in Figure 3.1. A systemic literature's review using advanced query is performed in PubMed to retrieve the ILDs' genes. The specific role of these genes in ILDs is validated using experimental findings of literature. Furthermore, these validated genes are enriched using resources such as KEGG, CTD, GHR, GAD, DISEASE, OMIM, and GeneCards. The final

list of literature validated genes having roles in ILDs were referredas disease candidate genes (DCGs) and integrated into the knowledgebase. Genes that do not have direct or significant association along with ambiguous or redundant information were discarded. These DCGs are further categorized into five different groups based on their specific roles in disease pathogenesis. These categories are differential expression, mutation, therapeutic targets, biomarker & genetic testing, and others. Standard public resources like NCBI, PDB, UniProt, Ensemble, etc. are used to retrieve genomics and proteomics information of collected DCGs. A sample advanced query for the retrieval of ILDs associated genes is provided below:

"Interstitial lung disease" [All Fields] AND "genes" [MeSH Terms] AND "genetics" [MeSH Terms]

3.2.1.2 ILDs specificity index (ILDsi)

Sometimes, one or more DCGs are involved in more than one subtype of ILDs (*e.g.* SFTPC has a role in non-specific interstitial pneumonia (NSIP), children's ILD (ChILD), IPF, etc.). The extent of DCGs and ILD association is determined by calculating ILDs specificity index (ILDsi) for each DCG (eq. 3.1). This ILDsi is adopted from DisGeNET [119].

$$\boldsymbol{ILDsi} = ((\boldsymbol{log_2}\left(\frac{\boldsymbol{Nd}}{\boldsymbol{Nt}}\right))/(\boldsymbol{log_2}\left(\frac{\boldsymbol{1}}{\boldsymbol{Nt}}\right)) \quad (3.1)$$

Where,

- Nd represent the number of ILDs' subtypes associated with a given gene
- Nt represents the total number of ILDs available in the knowledgebase.

This score varies from 0 to 1. Score near to zero represents a higher number of diseases associations with the given DCG, while score near to 1 represents a smaller number of diseases associations with the given DCG.

3.2.1.3 DCG Enrichment using Hypergeometric Test

The hypergeometric test has been performed to determine the significant association between any DCG and ILDs's subtype based on reported literature [120]. This statistical test provides a score for each disease-gene pair and ranges from 0 to 1. It provides statistical probability for identifying significant involvement of DCG in ILD subtype's pathogenesis. The lower score indicates a significant role of DCGs.

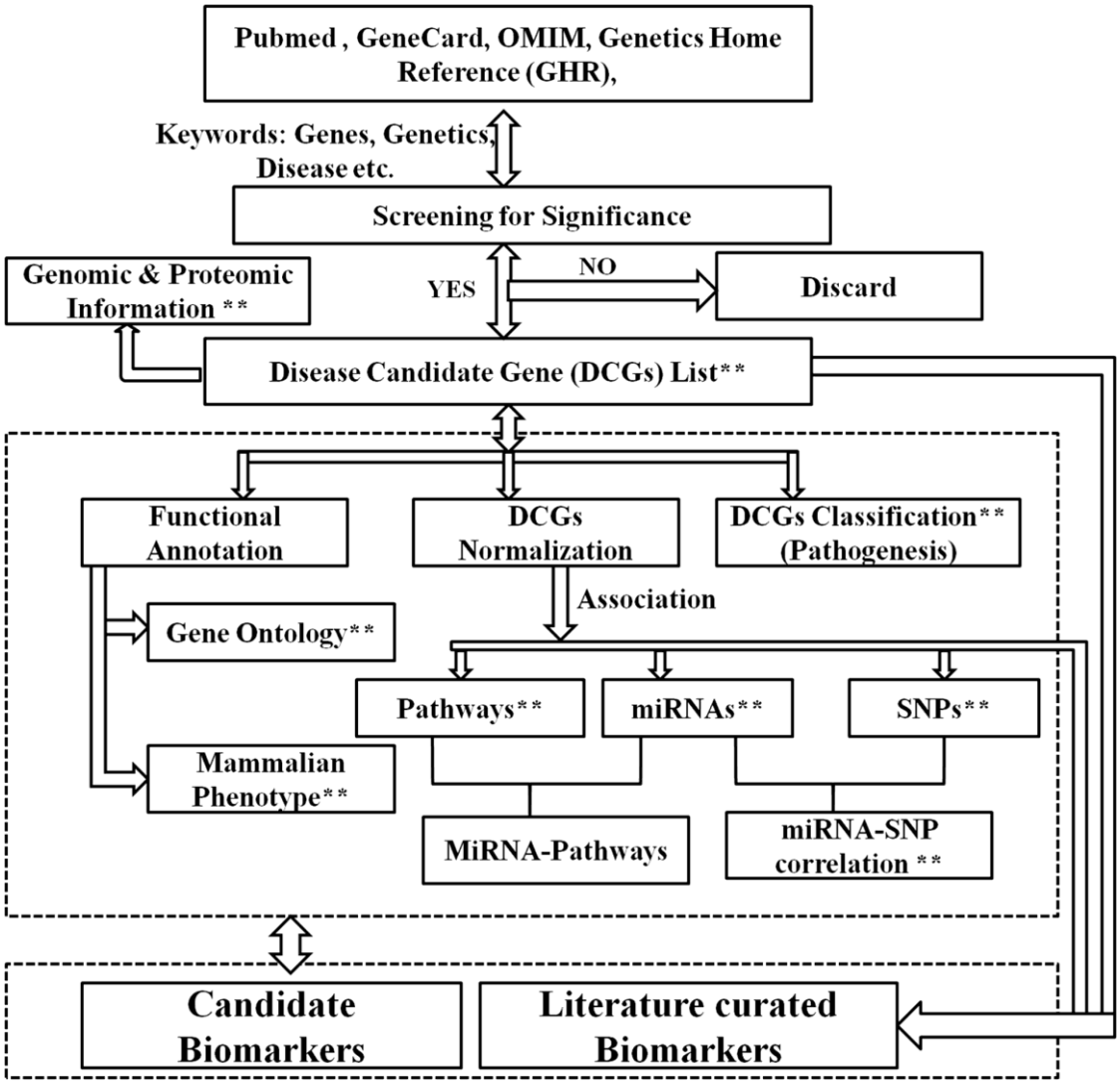


Figure 3.1 Architecture of ILDgenDB knowledgebase. **These can be used as query terms to explore the knowledgebase. Gene, disease, biomarker, category, SNP, miRNA, phenotype, etc. can be used as query.

3.2.1.4 Gene normalization and annotation using GRCh38.p7

Currently approved symbol, aliases, previous symbols and the accession of different resources like Entrez, HGNC, Ensembl, InterPro, UniProt, VEGA, OMIM, MGI, etc., are incorporated (Figure 3.1). Important information of DCGs such as nucleotide sequences, amino acid sequences, their functions, cryptogenic location, pathways, domains, mRNAs, mammalian phenotypes, protein family, pathogenesis, etc. are also incorporated (Figure 3.2).

3.2.1.5 Gene Ontology analysis, and DCGs' association with miRNA and SNPs

The GO analysis is performed to identify the functional importance of DCGs. Different GO enrichment tools such as AmiGO, DAVID, Ensembl BioMart, and PANTHER have been used for GO analysis. A cut-off P-value <=0.05 (provided optimal results) is used to filter the significant GO terms, and the redundant information is removed. Molecular functions,

biological processes and cellular component terms mapped to DCGs are verified for ILDs' association and incorporated into the knowledgebase.

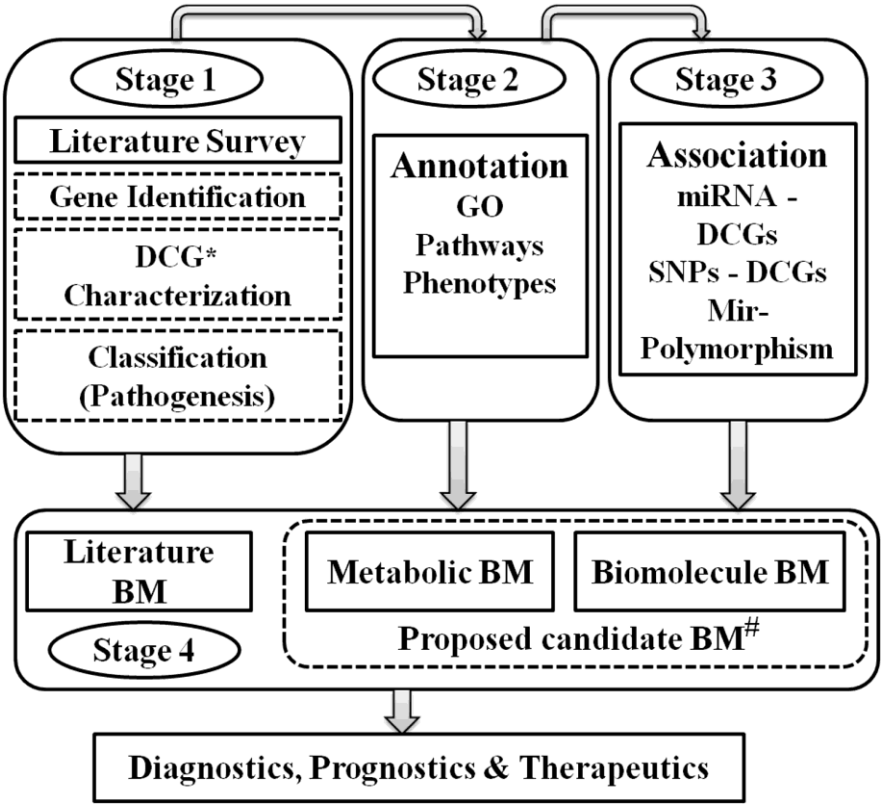


Figure 3.2 Workflow of ILDgenDB knowledgebase in different stages. Disease candidate gene (DCG),biomarker (BM)

miRNA prediction servers and databases such as miRwalk, PITA, starBase, mirBase and targetScan are used to predict the interaction with mRNA (DCGs). Experimental verification status and clinical significance of miRNA-mRNA (DCGs) association were also identified. The identified potentially significant miRNA list is further validated using mirBase (Figure 3.1). Ambiguous and redundant values are excluded. Chromosomal positions and miRNAs' mature sequences are also incorporated in knowledgebase.

SNPs associated to DCGs are identified using dbSNP resource. Potential SNPs are filtered out using clinical significance and annotation available in dbSNP. Mir-Polymorphism (miRNA-SNPs-DCGs associations) is also retrieved from miRdSNP resource. Analysis of clinical significance of SNPs is performed, and insignificant SNPs were discarded to include unique miRNA-SNPs-DCGs associations with clinical significance.

3.2.1.6 Biomarkers

The PubMed is subsequently queried using specific MeSH terms for identifying established molecular biomarkers of different ILDs. Different keywords such as biomarkers, disease

diagnostics, genetic testing and molecular diagnostics are used separately to retrieve data on reported biomarkers. Following is the sample of the standard query used in PubMed:

"Interstitial lung disease"[All Fields] AND "biomarker"[All Fields]

These molecular biomarkers are categorized into different groups according to their biological sample and cellular sources (e.g. serum, plasma, etc.). Biomarkers identified with specific roles in ILDs' pathogenesis were subcategorized as gene, disease, protein and/or disease-specific using literature (Figure 3.2). The literature references and brief description for each biomarker are also provided in ILDgenDB.

3.2.2 Data access

Users can use "Browse" and "Search" options to efficiently explore and retrieve the data from knowledgebase (Figure 3.3). Detailed steps to efficiently use these options are provided in ILDgenDB tutorial. Integrated and cross-linked information of ILDs data available in the knowledgebase can be accessed by ILDs' subtypes or diverse query with multiple fields can be used (Figure 3.1 and Figure 3.3). Datasets from the different categories are interlinked to provide easy navigation system.

3.2.3 Web implementation

The ILDgenDB knowledgebase is built on different web-languages such as HTML, JavaScript, PHP, CSS, etc. (Figure 3.1). The MySQL server is used to store data and provide faster results from users' queries. The client side programming languages such as HTML, JavaScript, CSS are used for the development of graphical user interface (GUI), while server-side PHP script is used to communicate and retrieve the data from MySQL server. The ILDgenDB is comprised of data and application layer. The data layer contains basic genomic and proteomic information of the DCGs, whereas application layer provides the outcomes of analyses performed on DCGs. Data from both the layers are easy to access and download.

3.2.3.1 Disease-DCGs network

The disease-DCG network viewer termed as NetViz is a web-based visualization tool which provides an interactive network graph of gene and disease association. The force-directed layout is used to build the network. The NetViz provides options to visualize two types of the data; first, the association of ILD with DCGs and SNPs/miRNAs; and second, Association of ILD with DCGs determined by hypergeometric test.

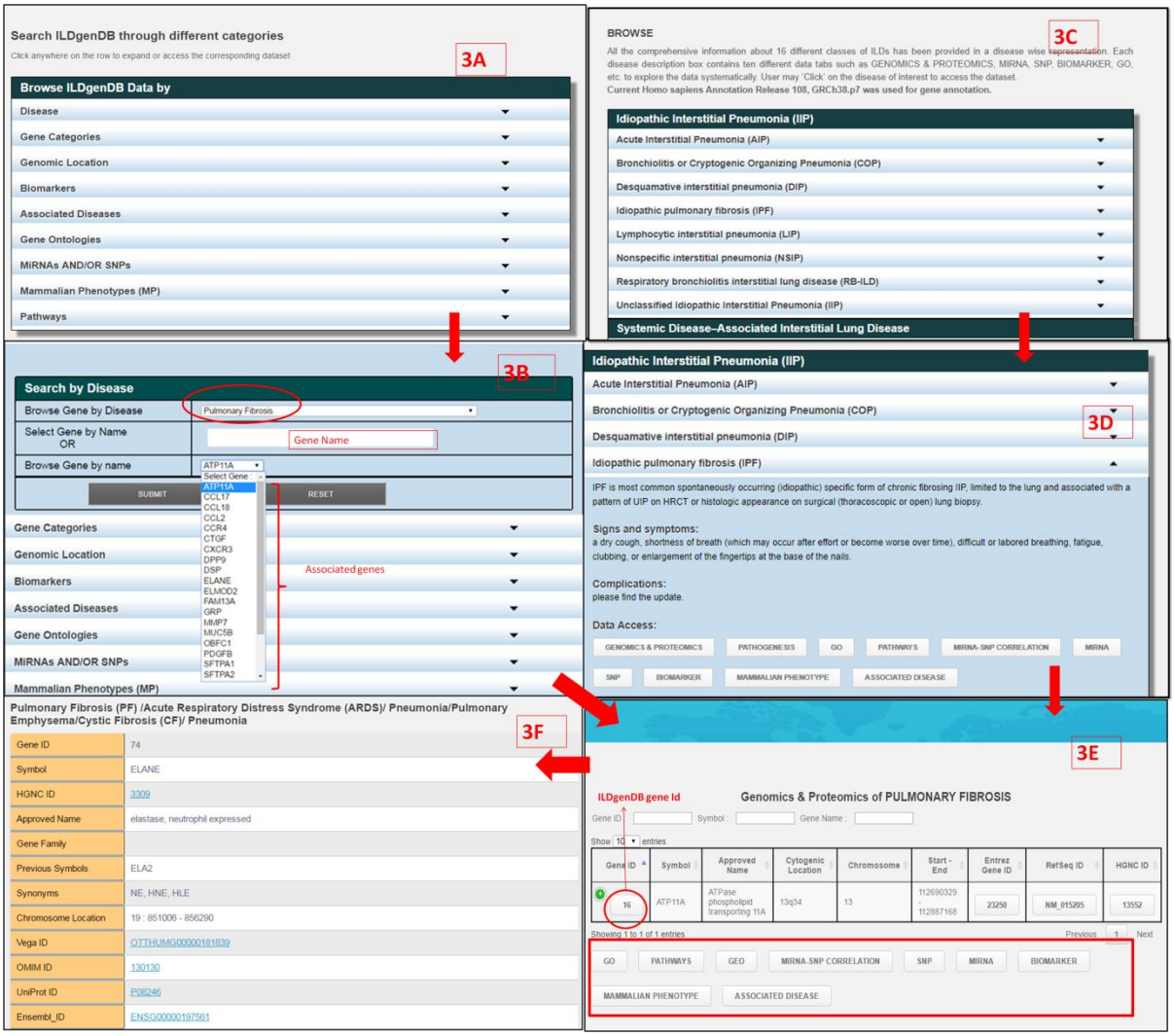


Figure 3.3 Retrieval of genetic data from ILDgenDB knowledgebase; (3A) "Search" option to explore data using different user specific queries. (3B) Query the knowledgebase using disease IPF and associated 'ATP11A' gene; (3C) "Browse" option for disease-wise knowledgebase exploration; (3D) Every disease in "Browse" option contain 10 different ways to explore the knowledgebase; (3E) Output from 3B & "genomic and proteomic" tab under IPF in 3D; (3F) Complete information page of DCG accessed by "Gene ID" hyperlink on 3E.

3.3 Results and Discussion

Disease-oriented genomic knowledgebase comprising of genetic elements such as genes, SNPs, pathways, miRNAs, etc., can assist in characterizing molecular biomarkers/network, and therapeutic targets [121]. Several databases/tools have already been in use for better and efficient disease diagnosis and management [101]. Nevertheless, the majority of the publically available database are either developed for cancers or infective diseases [122]. ILDgenDB is the primary knowledgebase comprising of DCGs, miRNAs, SNPs, pathways, biomarkers, and other genomic data for different subtypes of ILDs (Figure 3.2, Stage 1). An extensive literature survey is carried out in PubMed to retrieve the ILDs' associated genes.

Key molecular biomarkers involved in disease pathogenesis are also identified. Hyperlinks to the PubMed articles presenting the association between DCG and disease are also incorporated.

Annotation, enrichment analysis and characterization of DCGs are carried out by analyses of GO terms, phenotypes, and pathways (Figure 3.2). The DCGs associated SNPs and miRNAs are also determined and validated to incorporate in ILDgenDB (Figure 3.2). Molecular biomarkers are also mined from reported literature to incorporate into the ILDgenDB. Novel biomarkers are also determined by performing downstream analysis of DCGs (Figure 3.2). Following sections provide the main applications of the developedknowledgebase.

3.3.1 ILDgenDB web application

3.3.1.1 "Search" utility

ILDgenDB offers an interactive query system to access the data of interest using ten different query options such as gene/disease name, miRNA/SNP IDs, etc. (Figure 3.1 and Figure 3.3:3A). For instance, a user can usethe drop-down box to select disease name "Idiopathic Pulmonary Fibrosis" from "Browse Gene by Disease", then genes involved in IPF can be selected from drop-down menu or gene symbol(ATP11A, etc.) can be typed in text-box (Figure 3.3: 3B). DCGs associated GO terms, pathways, phenotypes, SNPs and miRNAs are also accessible through cross-linking (Figure 3.3: 3E). Hyperlink on "Gene ID" of a result page (Figure 3.3: 3F) can be used to access detailed genomic/proteomic information (Figure 3.3: 3E). DCGs detailed information like mRNA/protein sequence, CDS, coordinates, and hyperlinks to external databases/tools are also incorporated (Figure 3.3: 3F). ILDgenDB's tutorial page can be used to understand all the available options to query knowledgebase. An alternative to "Search" is provided in the form of "Browse" utility where the user can directly select the disease of interest to explore the related data in knowledgebase (Figure 3.3: 3C).

3.3.1.2. Disease-gene network visualization (NetViz)

The NetViz is an interactive force-directed network for a subtype of ILDs. The "ILD-Gene-miRNA/SNP Association" option can be used to view the interaction of disease with associated genes and SNPs/miRNAs (Figure 3.4:a). This interactive force directed graph can assist in identifying SNPs/miRNAs associations with DCGs. This graph is created using edges and nodes. The smaller edges between nodes represents the greater number of miRNA or SNPs association with DCGs of ILDs' subtype (Figure 3.4:b). These associations could be

used to prioritize the regulatory elements for their potential use as biomarkers. The "ILD-Disease-Category Literature Association" option can be used to check the association of a DCGs with different ILDs based on reported literature (Figure 3.4:c). The size of disease node in force directed graph is directly proportional to the extent of association (minimal hypergeometric p-value) between gene and disease determined using a number of reported literature (Figure 3.4:d). The larger is the size, better is the association between DCG and disease. These results are also provided in the tabular format.

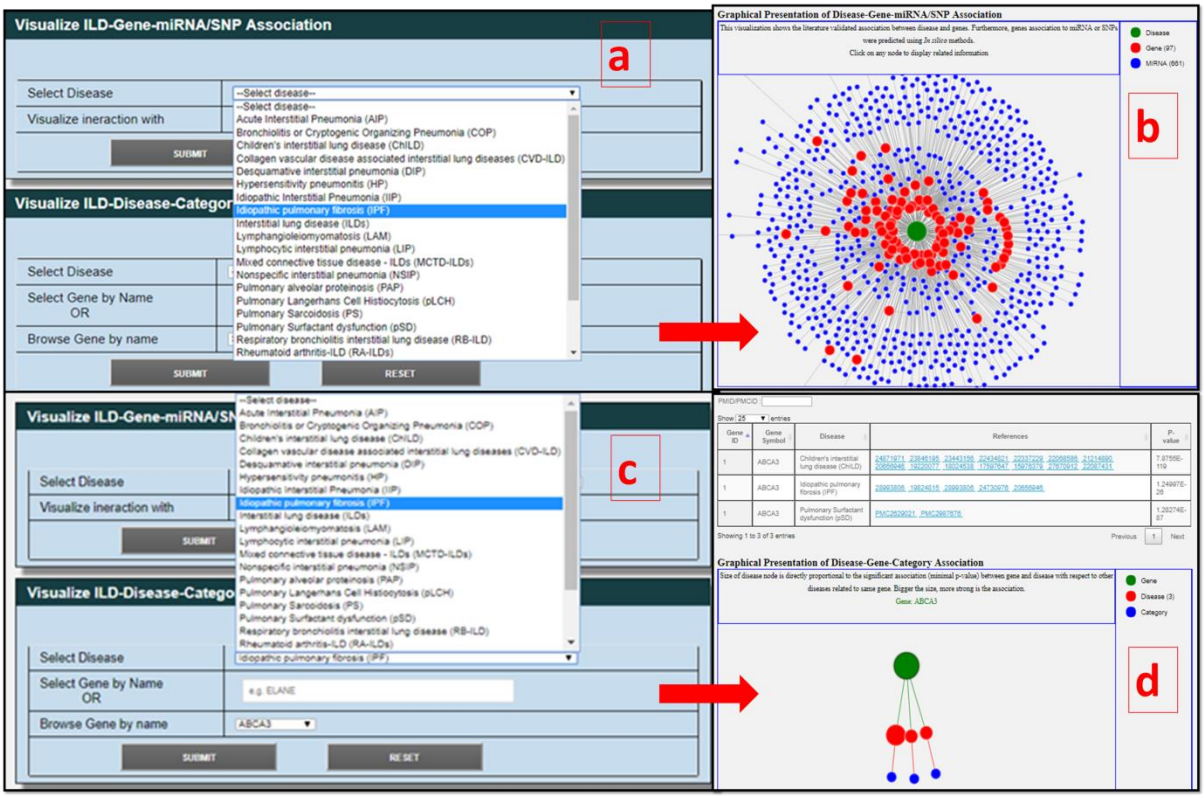


Figure 3.4 "Interactive network visualization (NetViz). (A) Select disease name and "miRNA" or "SNP" option from query page; (B) Force-directed graph for selected disease and regulatory element; (C) Select disease and an associated DCG name for DCG's category wise association identified by literature survey; (D) Force-directed graph showing association of a DCG with different ILDs based on reported literature.

3.3.2 DCGs functional annotation, categorization, and their associations with other genetic elements

The DCGs were categorized based on their role in pathogenesis. They are categorized as therapeutic targets, biomarker-and-genetic testing, mutations, altered expression and other group. DCGs such as IL-10, TGF-β, etc., are categorized as therapeutic targets based on their association with pathways and regulatory elements. The up/down regulated DCGs in diseased conditions are categorized as "differential expression". BDNF along with several other genes were experimentally validated for their altered expression in pulmonary diseases [123]. Few

established DCGs like SFTPB, TERT, ABCA3, etc. were validated for their structural and functional significance in lung diseases [42, 43, 124]. Smoking-related ILDs association with mutations in few DCGs were also reported [40]. These findings have suggested that experimental validation of altered DCGs role may help in exposure-induced ILDs pathogenesis. Validated DCGs can be used as potential molecular biomarkers for the pathogenesis and diagnosis of ILDs' subtype. Few established and promising biomarkers are CCL18, KL-6, SP-D and SP-A [125]. Bimolecular biomarkers (DCGs, SNPs, and miRNA) may assist in the detection of patients' subtypes for more specific treatment and monitoring [125]. All the DCGs' categories are not mutually exclusive, and single DCG can be associated with more than one category. The "other" category contains DCGs with indirect role or DCGs associated with other primary genes.

3.3.2.1 DCGs' functional annotation using GO analysis

"Functional annotation is performed using GO enrichment analysis to determine the corresponding DCG-products' biology. The GO terms for different categories such as biological processes (BP), cellular components (CC) and molecular functions (MF) are identified for all the DCGs. The majority of DCGs (IL-4, IL-8, CCL-ligand, etc.) were mapped to the BPs like defense response (90), response to external stimulus (84) and immune system process (74) (Appendix 1: Figure A1, C).Enrichment with CC for the majority of DCGs has provided that theses DCGs belong to the extracellular region and extracellular space (Appendix 1: Figure A1, B). The MFs of the majority of DCGs are cytokine activity, chemokine receptor binding and small molecule binding (Appendix 1: Figure A1, A). These findings indicate the significant connection of immune system and molecular binding to ILDs pathogenesis." [126]

3.3.2.2 Pathways association to DCGs

"ILDs are progressive diseases which mainly cause inflammation and fibrosis in advanced stages. Targeting fibrotic and inflammatory pathways may help in disease pathogenesis and therapeutics [44, 127]. Pathways associated with DCGs are incorporated in this knowledgebase. The DCGs like SMAD, MMP-2, TGF-β,and MMP-9 are found to be associated with signaling cascades, stimulation and pro-fibrotic protein expression pathways. The potential involvement of these biological pathways was already verified in ILDs pathogenesis [44, 128]. Analysis of DCGs associated pathways has also indicated their potential role in the immune/innate-immune system, signal transduction and cytokine

signaling (134, 91, 67 and 59 DCGs, respectively). The cytokine and Chemokine signaling pathways are also potentially associated with key DCGs like IL10, IL9, CCL11,and TNF. The cytokine-cytokine receptors interaction is involved in development & repair processes, and adaptive inflammatory host defense that is crucial for the ILDs' progression. Therefore, these interactions can be studied for disease diagnosis and pathogenesis. Selective immune processes can be targeted to gain new insight in ILDs' therapeutics (Appendix 1: Table A1)." [126]

3.3.3 DCGs' Association with miRNAs and SNPs

3.3.3.1 miRNAs-DCGs association

"The miRNAs-DCGs' associations may assist in disease pathogenesis, diagnosis, and prognosis [31, 129]. The current analyses revealed that the top-ranked miRNAs such as hsa-miR-335, hsa-miR-26b and hsa-let-7have targeted about 65, 34 and 33 DCGs, respectively(Figure 3.4 and Appendix 1: Table A4). Furthermore, hsa-miR-30 has targeted 26 DCGs. These targeted DCGs was verified for their dysregulation in many ILDs' subtypes like systemic sclerosis (SSC), IPF and acute lung injuries [45, 130]. Up-regulation in has-miR-145a, has-miR-126, has-miR-21, has-miR-106a, has-miR-221/222, has-miR-155, and down-regulation in has-let-7, has-miR-133a and has-miR-20b were already reported for their roles in lung's inflammatory responses. Similarly, has-miR-29, has-miR-155, has-miR-200, has-miR-326,and has-miR-21 were reported for their important role in IPF [131]. Down-regulation in has-Let-7f, has-miR-22, has-miR-30c, and up-regulation of has-miR-322 and has-miR-451 were significantly associated with pulmonary hypertension (PH). The PH has a similar representation to ILDs. miRNAs of has-let-7 family are mapped to pathways that revealed the influence of hepatic fibrogenesis during the activation of TGF- β signaling in hepatic satellite cells [132]. Six miRNAs (has-miR-16, has-miR-21, has-miR-146a, has-miR-126, has-miR-155 and has-miR-223) were previously identified as potential molecular biomarkers for various diseases [133]." [126]

"Additionally, the analyses of DCGs that have shown the interaction with multiple miRNAs are also performed. The top five DCGs namely, DNAL1, MICA, FGF2, C3 and ITGA3 are targeted by the highest number of miRNAs (175, 143, 99, 94 and 79, respectively) (Figure 3.5). The miRNA-mediated-immune-effect in lung tissues is mainly carried out due to the MICA gene [134]. Few DCGs such as STAT3, HMOX1, etc., are also reported for their potential involvement in IPF pathogenesis, [135] immune deficiency and autoimmune ILDs [4]. Consecutively, altered expressions of interleukin families' DCGs (especially IL6)

and TGF-β1 were experimentally validated [136, 137]. Findings of this study can be used as starting point to validate the role of genetics in ILDs. Identified genetic targets may work as potential molecular biomarker for better disease diagnostics and management. The role of miRNAs (miR-26b, miR-335, let-7and miR-30 etc.) and biological pathways associated to the top-ranked DCGs may work as potential biomarkers for ILDs pathogenesis & diagnosis (Figure 3.5)." [126]

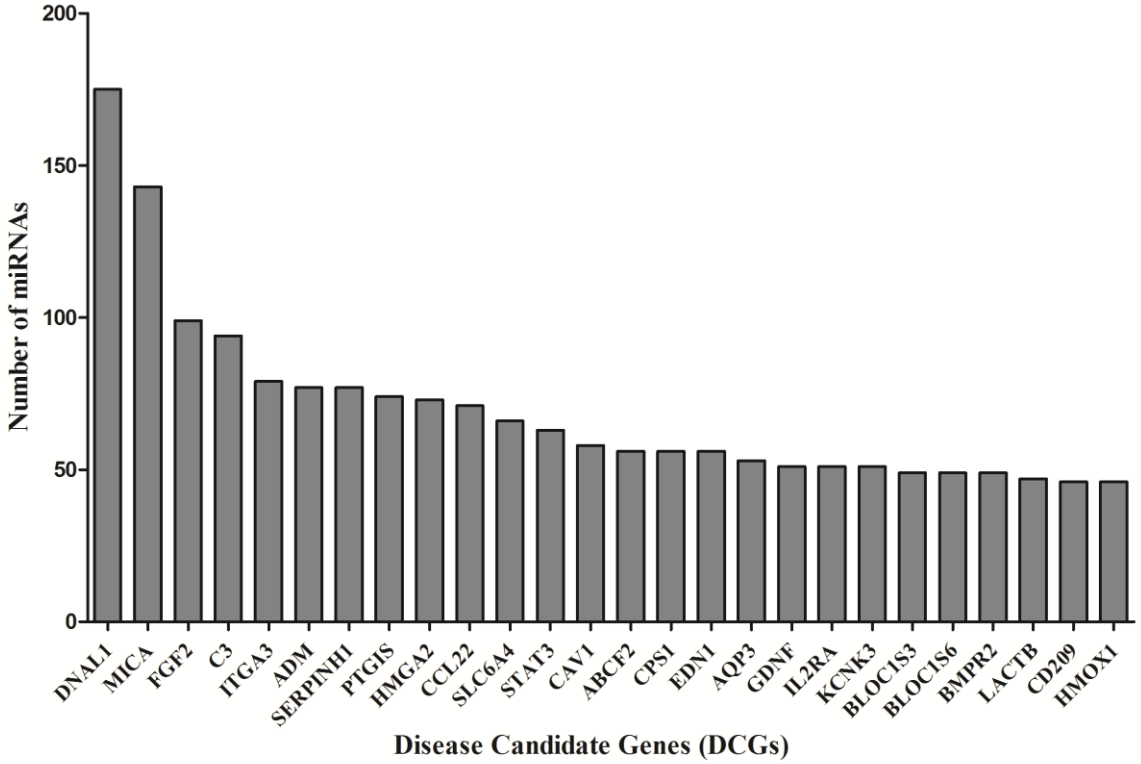


Figure 3.5 Significant DCGs identified by interaction with higher number of miRNAs.

3.3.3.2 SNP-DCGs association

SNPs and DCGs association are determined with the help of dbSNP and Ensemble knowledge resources. Clinical significance of these associations is also incorporated into the knowledgebase. Studies have revealed that the association between SNP and gene can play a potential role in ILDs pathogenesis and progression (Tochimoto et al., 2015). In total, fifty clinically significant SNPs were identified in nine genes that have potentialrole in IPF pathogenesis (Appendix 1: Table A2). The highly associated DCGs are namely, VWF, NOD2, TERT and ABCA3, which have shown the association with 29, 6, 5, and 3 SNPs, respectively (Appendix 1: Table A2). Mutations in DCGs related to surfactant catabolism, surfactant-production-and-function and transcription factors (involved in lung development)

are the major factors in several ILDs including children's ILDs [138]. Mutations in the telomerase genes were also contributed significantly to ILDs pathogenesis [41].

3.3.3.3 miR-polymorphisms and DCGs association

SNPs in miRNAs are comparatively novel and effective way to understand disease pathogenesis, diagnosis, and prognosis [139]. These polymorphisms are referred as miR-polymorphisms. Ongoing studies have verified that miR-polymorphisms can be associated with diverse pathological processes including cell growth, apoptosis, differentiation and tumorigenesis [140]. The miR-polymorphisms associated with DCGs are identified and provided in ILDgenDB. In total, 91 DCGs are found to be associated with 9170 miR-polymorphisms. These results can be explored using "miRNA-SNP correlation" tab under "Browse" or "miRNAs AND/OR SNPs Search" option of the knowledgebase. Polymorphism in miRNAs may produce illicit binding sites by alterations in existing binding sites. Furthermore, this distorted binding of miRNA with mRNA may leads to differential expression level of gene, which may involve in disease pathogenesis [68]. Apart from ILDs, the SNP rs17281995 in has-miR-582, has-miR-337, has-miR-184, has-miR-200a and has-miR-212 [141], and SNP rs11614913 in hsa-mir-196a2 were found to have a potential role in cancers pathogenesis [142]. Few top-ranked genes with the highest number of associations are HMGA2, PLCG1, FKBP1A, DLG1, SPRED1, FAM13A, etc. Top 20 DCGs associated with 5706 miR-polymorphisms (229 miRNAs, 340 SNPs) are provided in Appendix 1: Table A3. These findings can be verified experimentally for their significant roles in ILDs or their subtypes pathogenesis and diagnosis.

3.3.4 Biomarkers characterization

Molecular biomarkers have shown potential in the disease diagnosis/prognosis and can be used to develop molecular-based therapeutics [143]. Integrated genomic knowledgebase for a particular group of diseases may provide a better prospect to identify novel molecular biomarkers for developing novel diagnostic strategies [144]. ILDs biomarkers were collected from published literature to provide integrated information in ILDgenDB. Functional annotations of these biomarkers were performed through literature and provided in "Biomarker" section of knowledgebase (Figure 3.3: 3D). Total of 216 biomarkers are grouped into categories such as disease-specific (135), disease-specific proteins (5), proteins of pathogenic pathways (57), protein (10), disease-specific serum (7), serum (40), gene (12)and disease-specific miRNA (3). Most of the reported biomarkers are used for

differentiating ILDs patients from others. However, very fewer biomarkers such as SP-A, KL6, SP-D, etc. are specific to the ILDs' subtypes. Furthermore, "Source" (literature reference) and a description of each biomarker are also incorporated for the better understanding.

Mutations or altered expressions of DCGs can perturb different biological pathways and cellular processes and can be used as molecular biomarkers for disease diagnostics and prognostics [145, 146]. Furthermore, miRNAs' altered expression level, SNPs, miR-polymorphism, pathways and molecular functions of DCGs are identified to propose them as potential novel biomarkers. These biomarkers are referred as metabolic and bio-molecular biomarkers. DCGs that are mapped with the significant number of GO terms and pathways are categorized as metabolic biomarkers. Furthermore, miRNAs, SNPs and miR-polymorphisms with significant associations with DCGs are referred as bio-molecular biomarkers (Appendix 1: Table A4). These potentially significant biomarkers may need experimental verification (Figure 3.2).

3.4 Discussion and Future Developments

Several knowledge resources have been developed and reported based on associations of different genetic factors with the disease. For instance, ClinVar is the database of genetic variations, phenotypes, and their clinical significance. Similarly, SNPs-3D is the database of SNPs-genes association [147]. T-HOD [148] and PubMath [149] are the text-mining servers associated with disease-specific data. "DISEASES" database provides manually curated associations between diseases and genes, cancerous mutations and genome-wide association studies [122]. Similarly, DisGeNET database contains the disease and associated genes and SNPs; However, associated miRNAs and molecular biomarkers are not available [147]. Nevertheless, comprehensive information on different ILDs and associated genes and other genomic data is very inadequate. The ILDgenDB knowledgebase is the only integrated resource that provides literature-mined DCGs and their association's analysis with genetic elements such as miRNAs, SNPs, pathways, GO, and molecular biomarkers. Furthermore, this knowledgebase provides cross-associations among miRNAs, SNPs, and DCGS. Potentially significant biomarkers based on strong associations with DCGs are identified to validate and improve the utility of knowledgebase. This knowledgebase would be greatly useful to get an overview of DCGs and associated genetic factors involved in ILDs. Most of the DCGs reported in the literature are incorporated in the current version; however, the primary prospect is to incorporate novel DCGs in the knowledgebase. Additionally, literature

references will be updated in the next version for the greater significance of the knowledgebase**.**

APPENDIX 1

Gene Ontology (GO) mapping of all DCGs in all three domains: Molecular function (A), Cellular component (B) and Biological process (C)

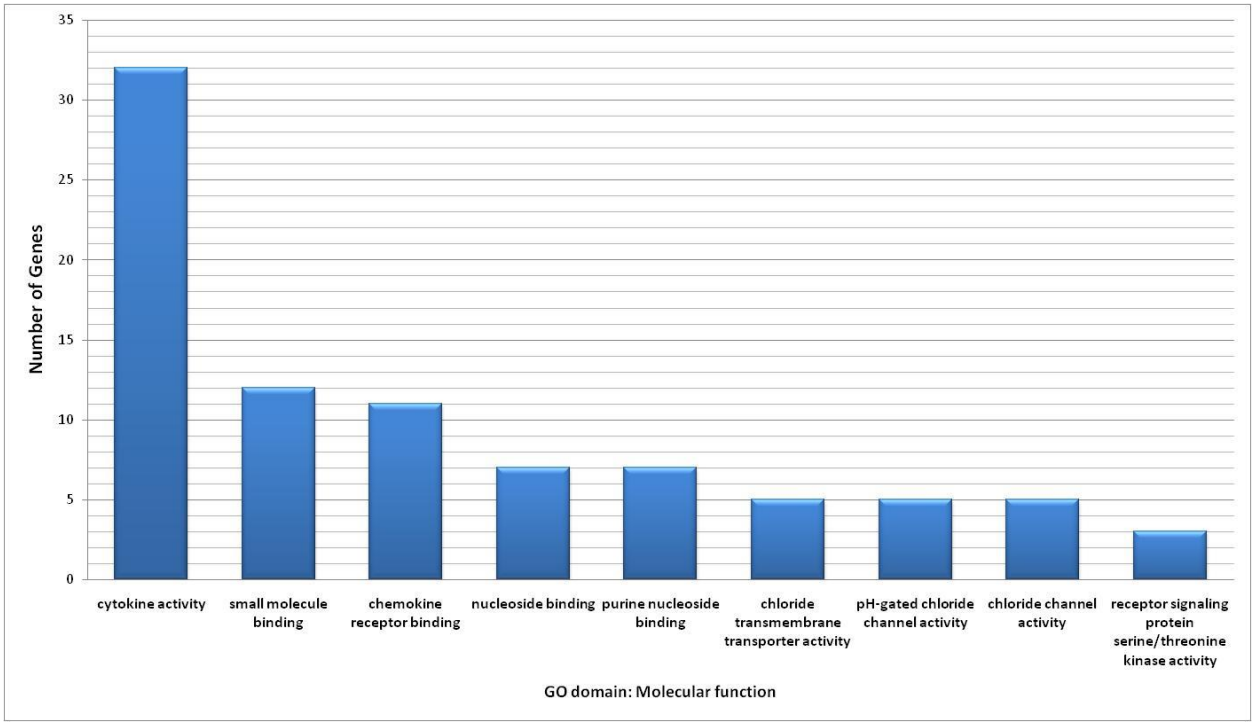


Figure A1, A: Distribution of DCGs with different molecular function

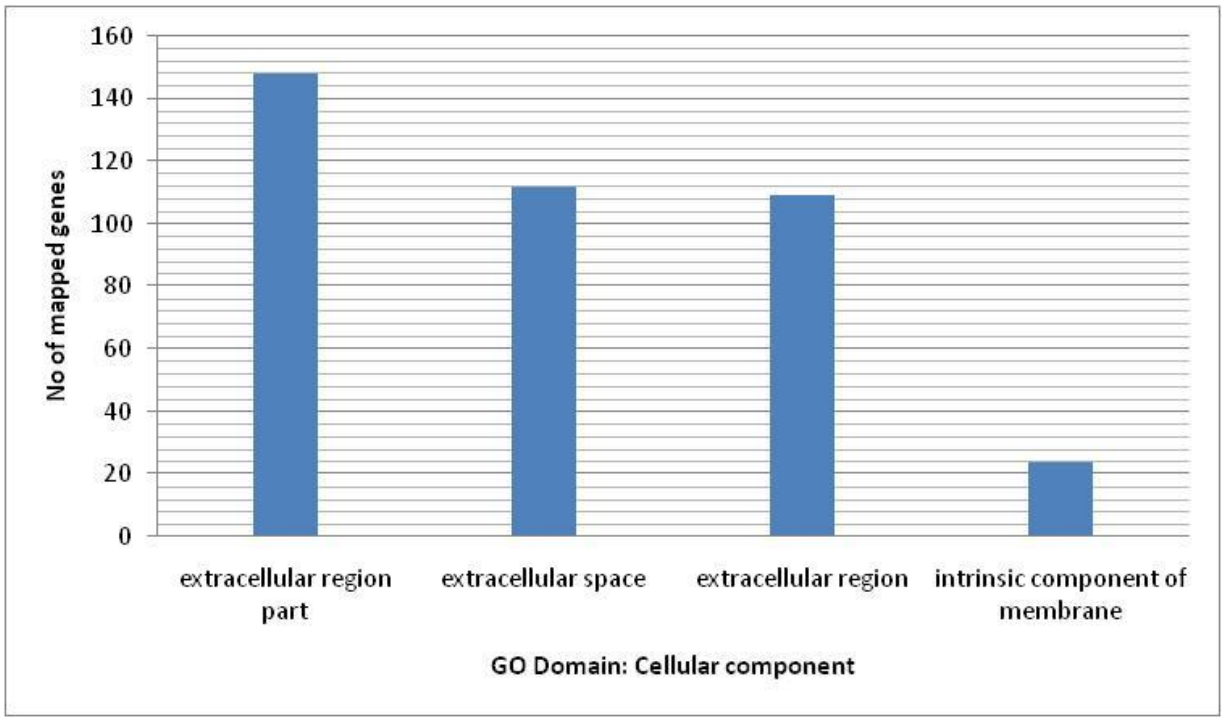


Figure A1, B: Distribution of DCGs with different cellular component

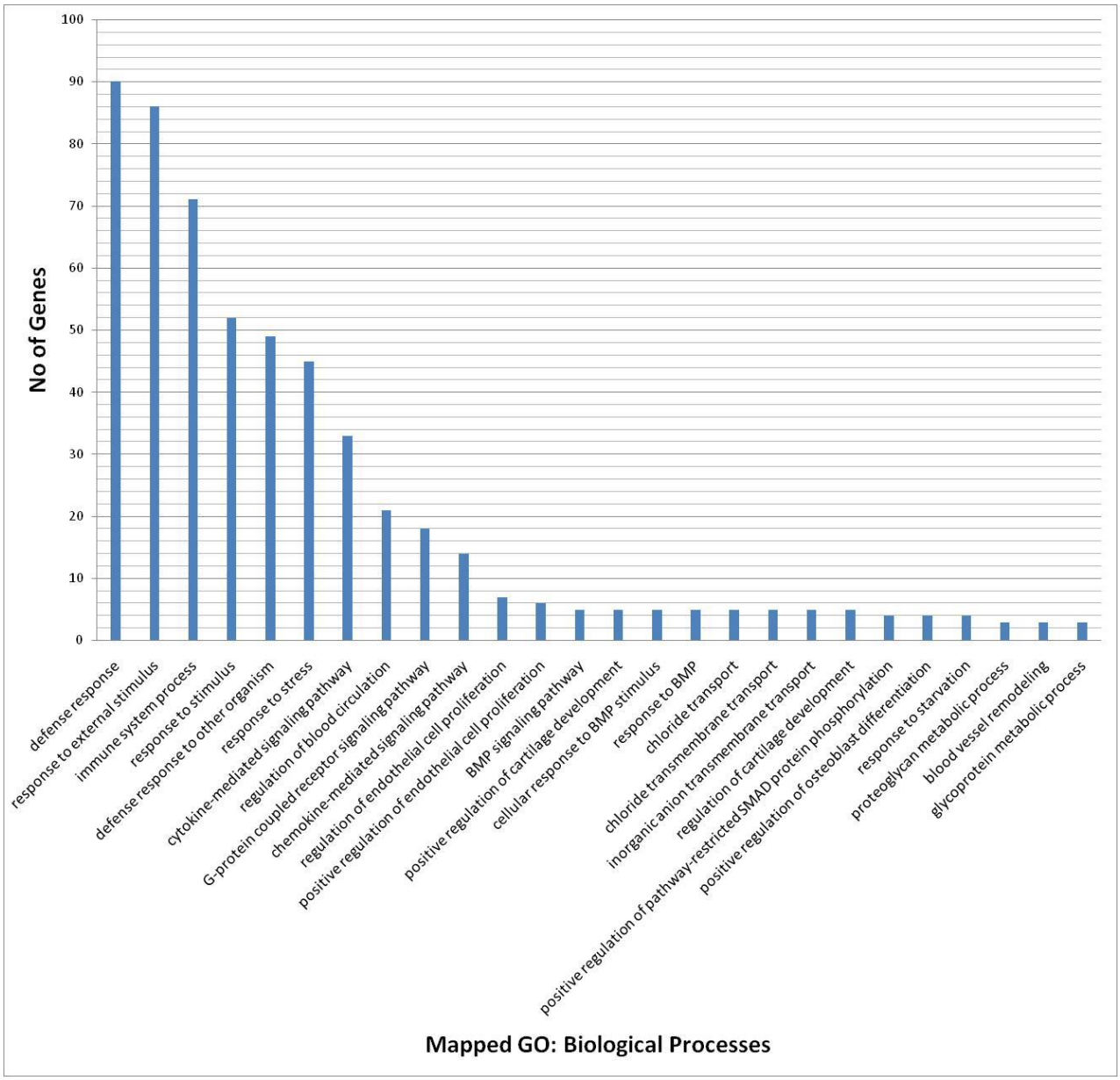


Figure A1, C: Distribution of DCGs with different biological processes

Table A1 Top 10 potential Pathways mapped by disease candidate genes (DCGs)

S. No	Pathway ID	Pathway Name	Number of DCGs mapped
1	P00031	Inflammation mediated by chemokine and cytokine signaling pathway	26
2	P00036	Interleukin signaling pathway	18
3	P06664	Gonadotropin-releasing hormone receptor pathway	13
4	P00052	TGF-beta signaling pathway	9
5	P00054	Toll receptor signaling pathway	8
6	P00011	Blood coagulation	8
7	P00006	Apoptosis signaling pathway	4
8	P00005	Angiogenesis	4
9	P00034	Integrin signalling pathway	4
10	P00057	Wnt signaling pathway	4

Table A2 Manually curated SNPs and associated DCGs potentially involved in IPF pathogenesis

S.No.	DCGs*	Count of SNPs
1	VWF	29
2	NOD2	6
3	TERT	5
4	ABCA3	3
5	PRF1	2
6	TNC	2
7	IL2RA	1
8	TGFB1	1
9	VDR	1
Grand Total		50

*Top targeted DCGs and SNP frequency in those DCGs involved in IPF. The analysis suggested that the maximum number of SNPs have been predicted in VWF (29).

Table A3 Co-occurrence of miRNAs-SNPs in DCGs

S. No.	DCGs	No of SNP	No of miRNA
1	HMGA2	34	11
2	FGF2	19	1
3	VDR	18	1
4	CTGF	17	5
5	CAV1	13	1
6	STAT3	12	2
7	CXCL12	11	1
8	BDNF	9	3
9	HOXA5	9	1
10	FBN1	8	1
11	IL6	5	1
12	ELMOD2	4	1

Table A4 Proposed candidate biomarkers on the basis of functional and structural mapping with DCGs

Candidate biomarkers	Number of interactions	Remarks
DNAL1	175	DCGs targeted by multiple number of miRNAs (Interaction= DCGs-miRNA interactions)
MICA	143	
FGF2	99	
C3	94	
ITGA3	79	
ADM	77	
SERPINH1	77	
PTGIS	74	
HMGA2	73	
CCL22	71	
SLC6A4	66	
STAT3	63	
CAV1	58	
ABCF2	56	
CPS1	56	
EDN1	56	
AQP3	53	
GDNF	51	
IL2RA	51	
KCNK3	51	
BLOC1S3	49	
BLOC1S6	49	
BMPR2	49	
LACTB	47	
CD209	46	
HMOX1	46	
hsa-miR-335	65	miRNAs having multiple numbers of DCGS target (Interaction= DCGs-miRNA interactions)
hsa-miR-26b	34	
hsa-let-7	33	
hsa-miR-30	26	
hsa-miR-20	24	
hsa-miR-124	24	
hsa-miR-17	17	
hsa-miR-93	17	
hsa-miR-106b	16	
hsa-miR-4722	15	
hsa-miR-1	14	
hsa-miR-155	14	
hsa-miR-16	14	
hsa-miR-6778	14	
hsa-mir-8485	14	
hsa-miR-98	14	

hsa-miR-128	13	
hsa-miR-19a	13	
hsa-miR-21	13	
hsa-miR-3653	13	
hsa-miR-519d	13	
hsa-miR-4768	12	
hsa-miR-92a	12	
hsa-miR-106a	11	
hsa-miR-125b	11	
hsa-miR-149	11	
hsa-miR-186	11	
hsa-miR-192	11	
hsa-miR-193b	11	
hsa-miR-548c	11	
Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	26	Potential Pathways mapped maximum number of disease candidate genes (DCGs) ((Interaction= DCGs mapped with pathways)
Interleukin signaling pathway (P00036)	18	
Gonadotropin-releasing hormone receptor pathway (P06664)	13	
TGF-beta signaling pathway (P00052)	9	
Toll receptor signaling pathway (P00054)	8	
Blood coagulation (P00011)	8	
Apoptosis signaling pathway (P00006)	4	
Angiogenesis (P00005)	4	
Integrin signalling pathway (P00034)	4	
Wnt signaling pathway (P00057)	4	
HMGA2	1183	Top targeted DCGs mapped with maximum number of clinically confirmed mir-polymorphism ((Interaction= DCGs-miRNA-SNP interactions)
SPRED1	496	
PLCG1	450	
FKBP1A	357	
FAM13A	304	
DLG1	300	
TSC1	257	
NF1	251	
FBN1	220	
PIK3C2A	213	
BDNF	206	
FOXF1	203	
MDGA2	196	
CTGF	194	
CXCL12	194	
IL10	191	
CAV1	174	
NOG	161	
FASLG	156	

ITGA3	153	
Defense Response (Go:0006952)	90	Top Gene Ontology (Biological process) terms mapped with maximum number of DCGs (Interaction= Number of DCGs with mapped GO term)
Response To External Stimulus(Go:0009605)	86	
Immune System Process (Go:0002376)	71	
Response To Stimulus (Go:0050896)	52	
Defense Response To Other Organism (Go:0098542)	49	
Response To Stress(Go:0006950)	45	
Cytokine-Mediated Signaling Pathway(Go:0019221)	33	
Regulation Of Blood Circulation(Go:1903522)	21	
G-Protein Coupled Receptor Signaling Pathway(Go:0007186)	18	
Chemokine-Mediated Signaling Pathway(Go:0070098)	14	
Cytokine Activity(Go:0005125)	32	Top Gene Ontology (Moleculer function) terms mapped with maximum number of DCGs (Interaction= Number of DCGs with mapped GO term)
Small Molecule Binding(Go:0036094)	12	
Chemokine Receptor Binding(Go:0042379)	11	
Nucleoside Binding(Go:0001882)	7	
Purine Nucleoside Binding(Go:0001883)	7	
Chloride Transmembrane Transporter Activity(Go:0015108)	5	
Ph-Gated Chloride Channel Activity(Go:0061797)	5	
Chloride Channel Activity(Go:0005254)	5	
Receptor Signaling Protein Serine/ Threonine Kinase Activity (Go:0004702)	3	
Extracellular Region Part(Go:0044421)	148	Top Gene Ontology (Cellular component) terms mapped with maximum number of DCGs (Interaction= Number of DCGs with mapped GO term)
Extracellular Space (Go:0005615)	112	
Extracellular Region (Go:0005576)	109	
Intrinsic Component Of Membrane (Go:0031224)	24	

CHAPTER 4

IDENTIFICATION OF POTENTIAL ILDS BIOMARKERS AND THERAPEUTIC TARGETS THROUGH COMPREHENSIVE ANALYSIS OF NON-CODING RNAs

Summary

ILDs encompass nearly ~200 chronic lung disorders with impenetrable pathological mechanisms. The potential role of deregulation in many molecular mechanisms related to immunity and defense response has been studied in the progression of ILDs. Non-coding RNAs especially microRNA (miRNAs) and long-noncoding RNAs (lncRNAs) are reported for their significance in different diseases pathogeneses including ILDs. Furthermore, these miRNAs and lncRNAs cross-regulate each other towards ceRNAs (competing endogenous RNAs) activity". These factors can be potentially used for disease diagnostics and observation. Efficient analysis or examination of these regulatory molecules could facilitate promising molecular biomarkers for ILDs. In this study, the roles of significant ncRNAs, differential expression and potential regulatory role in different biological functions and pathways have been studied. Outcomes of this study are validated using reported literature and different customary web resources. This research work revealed the noteworthy involvement of ncRNAs interaction with their target in ILDs pathogenesis. These interactions could lead to novel directions for ILDs management through less-invasive procedures.

4.1 Background

ILDs are a group of numerous chronic respiratory disorders with intricated underlying mechanism of pathogenesis [24]. A multidisciplinary discussion (MDD) among the "pulmonologist, radiologist, and pathologist" is required for the efficient diagnosis of ILDs [24]. Primary ILDs diagnostics approaches rely on the clinical symptoms, histopathological tests results and radiological patterns [103]. Availability of alternative testing such as surgical lung biopsies or bronchoalveolar lavage is difficult. Thus, HRCTs and CXRs are usually considered as a key diagnostic feature for ILDs diagnosis. However, The specificity of CXR and HRCT is low as ILDs can encompass many nonspecific and mimicking patterns associated with other lung diseases as well [103]. Additionally, many thorasic societies recommend low dose CT-protocols to avoid radiation hazards. In the past few decades, use of

less-invasive techniques in ILDs diagnostics and other disease has been popular. These techniques are proven to avoid the side effects of current diagnostic processes such as radiations, other co-morbidities and biopsies [150]. Past researches emphasized the vital role of genetic factors in ILDs etiology and prognosis [30]. Studies have shown the positive impact of molecular biomarkers for ILDs diagnosis, prognosis and treatment response to improve current diagnostic approaches [59].

Many noncoding-RNAs (ncRNAs) were reported for their regulatory roles in pathways and target genes associated with ILDs pathogenesis and diagnosis [32, 60-63]. The ncRNAs can be grouped into many subclasses according to their transcript lengths such as microRNAs (miRNAs), long non-coding RNAs (lncRNAs), siRNAs etc. [151]. Generally, miRNAs are made up of length of 20 to 22 nucleotides, and lncRNAs are made up of approximately 200 nucleotides. These ncRNAs also cross-regulated and the phenomenon is known as ceRNA activity. Many ongoing studies evidently supported the involvement of ncRNAs in the dysregulation of the immune system and defense response. These factors are reported to have potential intervention in many disease progression and pathogenesis including ILDs [61, 64-67].

Role of the differentially regulated miRNAs was already reported in signaling pathways and several biological processes and disease pathogenesis [46]. These miRNAs were already identified to increase lung fibroblast susceptibility and severity [47]. Additionally, few miRNAs are recognized as promising less-invasive therapeutic/biomarkers targets for ILDs (*e.g.* IPF) and cancers [10, 69]. The lncRNAs are also important for alteration of regular biological processes, genes dysregulation and disease progression to different stages [152, 153]. Role of lncRNAs in adaptive and innate immunity was already reported [61]. Integrated analysis of lncRNAs and disease association were found to promote lncRNAs-based management for many respiratory diseases [72, 73]. Additionally, an association of the ceRNA activities, their post-transcriptional regulation with cancers and fibrotic lungs progression were also implicated [32, 70, 71].

"The reported outcomes give clear insight on the comprehensive identification of miRNAs, lncRNAs, pathways and ceRNAs that affect molecular pathways, which can be characteristic factors in the etiology of ILDs [32]. The immune mediation and defense response have shown crucial role in ILDs pathogenesis and progression in the previous objective of this research work. Immunological processes were found to have associations with ILDs candidate genes, associated pathways and regulatory networks [65, 126]. The role of ncRNAs

in immunity was also established through previous objective and other supported literature [126]. Thus, the role of biomarkers related to autoimmunity has been well identified for ILDs pathogenesis and diagnosis [61]. Hence, miRNAs, lncRNAs, and ceRNAs are anticipated to be involved in ILDs progression. Nevertheless, very limited miRNAs are in clinical practice as biomarkers or used as therapeutic targets for ILDs subtypes [47].” [154]

“Therefore, functional characterization and verification of ncRNAs and their vital role as ILDs genetic biomarkers and potential therapeutic targets can together formulate significant progress in ILDs management [61, 151]. Considering these factors, this chapter emphasize on comprehensive study of ncRNAs and their associations with ILDs. Enrichment analysis and functional annotation of these ncRNAs, pathways and target genes are also performed. Integrated expression analysis of miRNA is also performed to identify the crucial role of miRNAs in different ILDs. The outcome of this objective would assist in the identification of promising biomarkers and therapeutic targets for ILDs.” [154]

4.2 Materials and Method

The stepwise systematic analysis was performed to identify potential miRNA and lncRNAs as therapeutic targets/biomarkers (Figure 4.1). Subsequent sections provide a better idea about the process, data collection, and their analysis.

4.2.1 Identification of ncRNAs and their association with ILDs

The miRNAs and their most recent information are collected from miRBase [155]. The lncRNAs are collected from lncRNA databases (such as IRNdb [67] and lncRNAdb [156]), for “Homo sapiens”. Identification of ILDs association with the ncRNAs is performed with the help of ILDgenDB [126].

4.2.2 Association analysis of ncRNAs and ILDs-pathways

4.2.2.1 Analysis of miRNAs associations with GO and pathways

Pathways datasets are downloaded from WikiPathways, KEGG, and Reactome. A consistence annotation is produced after removal of redundancy. The miRWalk2.0 comprehensive atlas and literature data are used to perform miRNAs association with pathways [157]. Top 10 ILDs associated pathways and their interactions with miRNAs are studied and identified (FIG 2). Analysis of biological processes, molecular mechanism, and cellular component is done using Gene ontology (GO) enrichment analysis of these miRNAs (Figure 4.1). A cut-off P-value of <0.05 is used to obtain significant enrichment results.

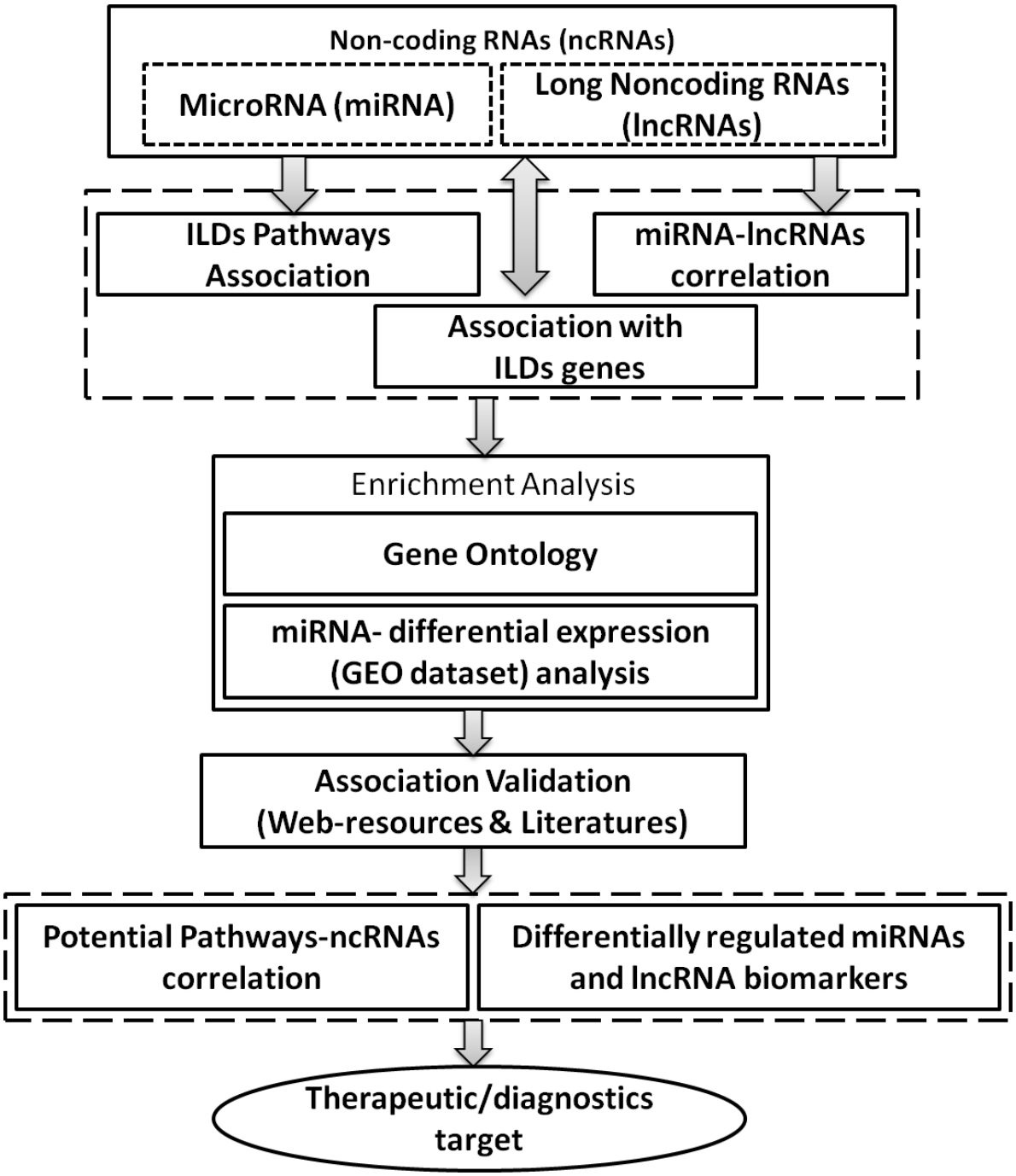


Figure 4.1 "Methodology followed for the identification of potential ILDs ncRNAs therapeutic target/biomarkers" [154]

4.2.2.2 Differential expression analysis of miRNAs

To perform miRNAs expression analysis, Gene expression omnibus (GEO) datasets are used to extract the expression profiles of miRNAs.

"Advance query options were used with keywords like *miRNAs, interstitial lung diseases, idiopathic pulmonary fibrosis*, etc. After checking the availability and type of data available, finally, a total of six datasets (GSE13316, GSE8555, GSE75647, GSE21394, GSE81293, and

GSE27430) are selected based on study design. The GEO2R is used to compare groups of samples(such as control v. disease) for differential expression. A combination ranking approach (proposed by McCarthy et al. [158]) by combining log fold change (LogFC) value (>1.5 for up-regulation and < -1.5 for down-regulation) and P-Value (<0.05) is applied to determine the biologically significant differentially regulated miRNAs." [154]

4.2.2.3 miRNAs, lncRNAs and pathways association analysis

To analyze the impact of ncRNAs crosstalk, the starBase v2.0 is used [159]. The ILDs association with pathways and lncRNAs are identified using ILDgenDB [126], IRNdb [67] and lncRNAdb [156]. The significantly associated lncRNAs are considered (Figure 4.1) for further analysis.

4.2.2.4 Validation

All the significant associations such as miRNAs/lncRNAs-pathways and miRNAs-lncRNAs are validated with the help of Pubmed. This literature-based approach has established confidence by validating the interactions in ILDs pathogenesis. A systematic literature survey is done for the evaluation of ncRNAs–target gene interaction responsible for ILDs. Advanced search option of PubMed is used to retrieve the potential associations of ILDs:

"microRNAs"[MeSH Terms] OR "miRNAs"[All Fields]) AND "interstitial lung disease "[All Fields]

Only the most significant terms are presented here. Several other query terms are also used such as "pathway name", "gene name", "ILD name", etc. The final conclusion is established after careful screening of the literature. Top-ranked associations have produced few potential ncRNAs that could be used as ILDs specific biomarkers. Few of these biomarkers were already experimentally verified in the literature.

4.3 Results and Discussion

Non-coding RNAs were found to have an important role in disease mechanism and pathogenesis, including ILDs [69, 159-161]. However, the majority of the studies were conducted mainly for cancers, and only a small number of ncRNAs-based clinical trials are carried out for ILDs and other disease diagnosis and prognosis [61]. This association analysis of ncRNAs along with a role in mRNAs regulation can result in less-invasive ncRNAs based diagnostic biomarker for pulmonary diseases like ILDs [10, 59, 151, 162]. To develop improved indulgent on these regulatory elements associations and their role in ILDs,

integrated analyses of ncRNAs, genes and pathways are performed and provided in a comprehensive manner (Figure 4.1).

In-house PERL scripting is used for identifying the ncRNAs and ceRNAs associations with ILDs specific genes and pathways. Many relevant bioinformatics tools, web resources are used for the enrichment analysis and literature from Pubmed are used for the results validation.

"In total, 2701, 669 and 4568 associations are identified in miRNA-gene-pathways, lncRNAs-pathways and miRNA-lncRNA (ceRNAs), respectively. The miRNAs interacting with many genes can alter their expressions and are believed to have a potential role as biomarkers and therapeutic targets in different studies [153, 163]. Potential diagnostics/therapeutic targets were proposed based on their implication in ncRNAs-pathways associations, lncRNAs-miRNAs associations, and dysregulated miRNAs (Figure 4.1)." [154]

4.3.1 Association analysis among ncRNAs, ILD genes, and pathways

4.3.1.1 miRNAs and pathways associations

Role of miRNAs in the alteration of several regulatory pathways and genes expression in ILDs cases are already studied [32]. These associations are identified by validating the curated miRNAs, genes and pathways associated with ILDs (Figure 4.2). In total, 127 ILD-specific genes reported in ILDgenDB are found to be associated with 228 different pathways. These pathways were retrieved from "KEGG", "Reactome" and "WikiPathways" knowledgebase [126].

"Total of 10different ILDs pathways and associated ncRNAs are identified and demonstrated to prove their significant role in ILDs pathogenesis (Table 4.2 and Figure 4.2). Outcomes of this study indicated that majority of the miRNAs are involved in inflammatory pathways, which suggest the potential role of inflammation in disease pathogenesis. Immunological processes are also identified for their potential role in ILDs pathogenesis. ncRNAs, ceRNAs and associated pathways are provided in Figure 4.2. Pathways such as TGF-β Signaling, Chemokine signaling, Cytokine-cytokine receptor interaction are identified as highly associated pathways to ncRNAs. Many pathways such as- JAK/STAT, MAPK and TGF-β signaling pathways were already studied for their role in fibroblast activation which leads to ILDs [67, 162]. These results substantiate the impending role of miRNAs in immunological processes. The miRNAs like miR-335-5p, miR-1, etc. have shown the

significant association to ten different ILDs pathways. These miRNAs can be explored for their significant roles in future studies." [154]

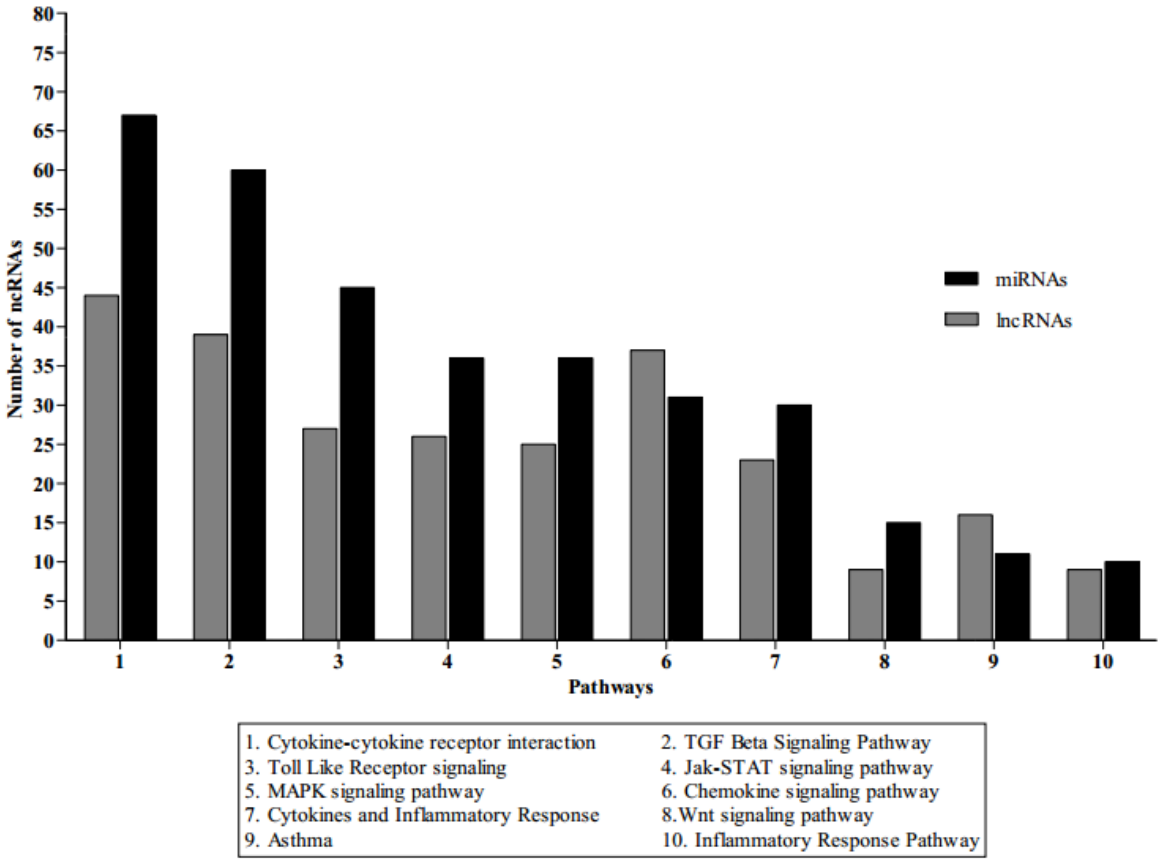


Figure 4.2 "Number of non-coding RNAs associated with pathways predicted to be involved in ILDs pathogenesis" [154]

4.3.1.2 The miRNAs' differential expressions

Role of differently expressed miRNAs in ILD pathogenesis is already studied. Rapid and precise measurability of miRNAs along with their highly stable nature makes them a promising molecular biomarker. miRNAs are proved to have significant roles not only for diagnosis but also in different phases of disease mechanism [48, 164]. To further explore the potential role of differentially expressed miRNAs in ILDs, seven different gene expression omnibus (GEO) datasets are analyzed (Table 4.1). The miRNAs which passed through the combined ranking were then analyzed for their potential roles in ILDs pathways (Section 4.3.1.1)**.**

"Several miRNAs (44) that are identified as significant in this study were also reported in previous studies for their promising roles in lung disease. For example, hsa-miR-1821 and has-miR-199a are studied for their potential role in ILDs and IPF, respectively [163][165]. Similarly, hsa-miR-363 and miR-588 were studied for their role in lung cancer [166, 167]. In

all the seven datasets of this study, mir-575, mir-4417, mir-665, etc. have shown the significant up-regulation in diseased condition. This altered expression of miRNAs makes them promising candidates for ILD-specific biomarkers. Similar studies could be performed on other ILDs to identify subtype-specific biomarkers. Therefore, differential expression analysis of miRNAs might be a great tool for ILDs diagnostics." [154]

4.3.1.3 lncRNA and ILD pathways association

Reported case studies have shown the negative impact of lncRNAs in ample of dysregulated pathways. These results suggested the use of lncRNAs as a novel tool for diagnostics, prognostics, and treatment of different cancers [69]. Identification of regulatory lncRNAs and their pathway association analysis provide links to promising lncRNAs that are involved in various ILDs pathways (Figure 4.2.).

"Total of seventy-three lncRNAs is found to interact with 91 ILDs genes and 196 pathways, which leads to 669 interactions among lncRNAs, pathways, and genes. Some potential lncRNAs such as XIST (5), MALAT1 (4), CTA-204B4.6 (5), etc. are mapped with many immunological pathways and ILDs genes. The detailed summary of lncRNAs and miRNAs association is given in Table 4.2. The genes targeted by these lncRNAs were found to be associated with ILDs pathogenesis (Figure 4.3), and they have exhibited the lncRNA-mediated differential expression in host defense and inflammation [161, 168]. Promising roles of these lncRNAs in a number of pathophysiological processes is already suggested in ongoing studies [61]. However, the involvement of lncRNAs in the pathogenesis of ILDs is not well known. Identified lncRNAs mediated regulation of genes and pathways related to ILDs could provide a greater insight into the disease pathogenesis; therefore, these associations could be a good therapeutic or diagnostic target for ILDs." [154]

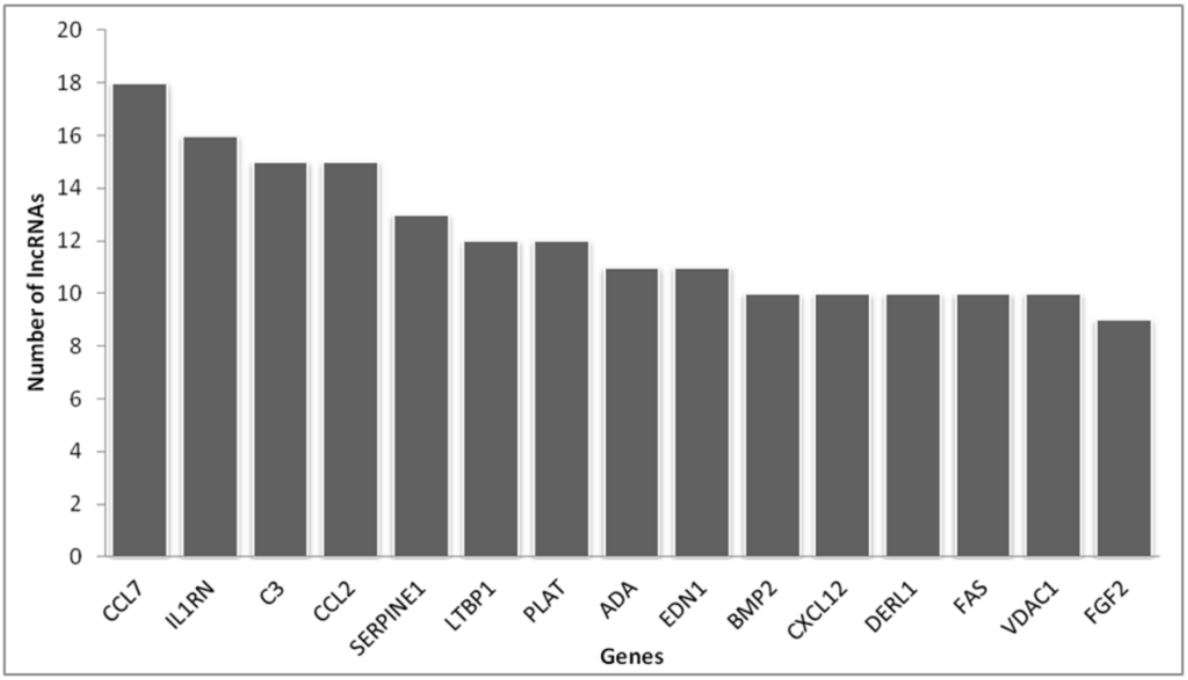


Figure 4.3 "Top 15 potential lncRNAs-ILDs genes target" [154]

4.3.1.4 Interaction between miRNA- lncRNA (ceRNA)

"The crosstalk between miRNA and lncRNA knew as "ceRNAs" interrupt the miRNAs bindings to the target [161]. Regulatory role of this ceRNAs activity was already studied in ILD like IPF [32, 168]. Total of 108 miRNAs associated to ILDs and 909 lncRNAs have produced the 4568 ceRNAs interactions in this study. Targeted ILDs genes used in this study are literature curated and have a potential role in disease pathogenesis and progression (Table 4.1). Table 4.1 also provides a number of lncRNAs important for ceRNAs interactions, and the number of literature verified pathways involved in ILDs. Identified top ten miRNAs associated pathways related to ILDs are namely, Asthma, Cytokine-cytokine receptor interaction, Chemokine signaling pathway, Inflammatory response pathway, Cytokines and inflammatory response, MAPK signaling pathway, JAK/STAT signaling pathway, Toll-like receptor signaling, TGF-β signaling pathway and Wnt signaling pathway. The potentially significant lncRNAs with the higher number of miRNAs interactions (ceRNAs interaction) are identified (Table 4.2). These top-ranked lncRNAs identified as biomarkers are associated with five different pathways involved in ILDs. These five pathways are namely, MAPK signaling, Cytokine-cytokine receptor, TGF-β signaling, Toll-like receptor signaling, and Chemokine signaling. The significant involvement of ceRNAs activities in cellular immune responses, reduction in anti-thrombogenic agents, inflammatory/pro-inflammatory pathways, growth factors, etc. is already reported [63, 160]." [154]

Table 4.1 The miRNAs predicted as potential biomarkers or therapeutic targets

S. No.	miRNA	Targeting ILD genes	No. of associated lncRNAs	Number of associated Pathways
1	hsa-miR-1	AP3B1, BDNF, CCL2, EDN1, F2, HMOX1, IL6, IL8, ITGA3, TLR4	33	11
2	hsa-miR-124-3p	BMP6, CAMP, CAV1, CCL2, DNAH5, EDN1, HMOX1, IL6, IL8, ITGA3, NME4, PGM1, PIK3C2A, PLA2G7, SERPINE1, SERPINH1	30	10
3	hsa-miR-125b-5p	BMPR1B, IL1RN, ADM, BMPR1B, SLC9A3R2, STAT3, VDAC1, VDR	26	8
4	hsa-miR-155-5p	CAT, CCL2, EDN1, F5, FGF2, IFNGR1, IL6, IL8, NEU1, SLC9A3R2, SMAD3, STAT3, TTF1	46	10
5	hsa-miR-21-5p	BMPR2, DERL1, FAS, FASLG, PIK3C2A, PLAT, SOD3, STAT3, TGF-β1, TLR4	19	10
6	hsa-miR-26b-5p	ADM, BMP2, BMPR2, C3, CAV1, CCL2, CCL7, CXCL9, CXCR1, DERL1, FASLG, FGF23, GUSB, HMOX1, IFNG, IFNGR2, INHA, ITGA3, MMP8, PDGFB, SERPINH1, STX1A, TLR1	48	10
7	hsa-miR-335-5p	ABCA3, ACE, ATF6, BMP2, CCR4, CCR7, CD14, CD27, CFTR, CLCA4, CXCL9, CXCR1, CXCR2, CYP2E1, FGF23, GC, GUCA2B, HMOX1, HSPG2, IL17A, IL1A, IL4, IL5, IL6, IL8, LTBP1, P2RY2, PDE5A, PIK3C2A, PLA2G7, PLAT, SERPINA1, SHH, SLC6A4, SMAD3, STX1A, TERT, THBD, TLR1, TLR2, TLR4, TNC, TREH, TSC1, VIP	32	12
8	hsa-miR-93-5p	BMPR2, DCTN4, STAT3, TOLLIP	74	8
9	hsa-miR-19a	BMPR2, TLR2, BMPR2, DERL1, TLR2	-	8
10	hsa-miR-15a-5p	IFNG, PLA2G2D	145	7

Table 4.2 Top ranked lncRNAs predicted as potential biomarkers or therapeutic target

S. No.	lncRNAs	Association with No. of miRNAs	Association with No. of Pathways	Association with No. of ILD-Genes
1	XIST	87	5	45
2	CTA-204B4.6	58	5	36
3	MALAT1	42	4	27
4	KCNQ1OT1	41	3	28
5	ZNF518A	40	4	25
6	NEAT1	34	2	24
7	SNHG16	33	4	24
8	HCG18	32	4	22
9	OIP5-AS1	29	2	17
10	RP11-361F15.2	29	5	21

Imperative role of these processes in ILDs was also established [126]. Further analyses of the proposed ceRNAs interactions may assist in functional similarity networks and ILDs-lncRNAs associations, which may assist to comprehend the mechanism of ILDs [69]. Top 10

ceRNAs interactions obtained from 10 different disease candidate genes and pathways are anticipated as potential biomarkers (Table 4.1, Table 4.2).

4.3.2 Association of ILDs genes and ncRNAs involved in immune processes

Association analysis of miRNAs, GO and ILDs genes are performed to identify the miRNAs' regulatory roles in various biological processes. Analysis of miRNAs with GO terms yielded a total of 2364 interactions. The miRNAs such as miR-335-5p and miR-26b-5p are found to be associated with the highest number of GO terms, which are 20 and 9, respectively. Hypoxia and cytokine activities related GO terms are predominant in this study, where former is the characteristic symptom of lung disorder like ILDs [169]. Inflammation, wounding and defense response is the predominant biological processes identified through miRNAs-GO term analysis. These outcomes have suggested a lucid indication that ncRNAs are associated with immunological processes in ILDs pathogenesis.

CHAPTER 5

CONCLUSION & FUTURE WORK

5.1 Conclusion

Current work advances the knowledge of interstitial lung disease, radiological patterns, the predominance of the disease and patterns, disease candidate gene identification, comprehensive enrichment analysis and candidate biomarker discovery. The significant findings of this thesis would expect to improve current diagnostics, prognostics and monitoring procedures in ILDs patients. It could be done by building genetic-based molecular models that would assist in novel therapeutic interventions. Additionally, the effort has been made to develop EHRs system to share and access this information considering security and human rights issues to serve for public health. This research is expected to offer veritable potentials, particularly for the developing countries, including web-based monitoring to provide personalized effective treatments and a better understanding of disease condition.

The first objective of the presented research work was to create an integrated resource for radiological and clinical ILDs presentation. In many developing countries, healthcare services are a system of segregated blocks. Integration of symptoms, pathological and imaging data to provide an efficient diagnosis and treatment is beyond the reach of healthcare systems. Although high-quality health services are available, the majority of the populations cannot afford it. Therefore, efforts should be made to provide better diagnosis through state healthcare services. In this study, EHRs based web-repository was developed to share and access ILDs data. This resource gives due attention to security and human rights issues, and provides anonymized data. Comparative analysis of patterns from different geographical regions offers the opportunity for better understanding of ILDs risks and management. This study also presented a rare lung disease, PAM, with its radiological and clinical pattern analysis, which may update us about the current knowledge of this disease. This web accessible resource for ILDs data (radiological patterns reported in CXRs and HRCTs, relevant patient's history and lifestyle data) can facilitate computer-aided diagnostics through the development/evaluation of medical image processing algorithms and clinical decision support system (CDSS) which may serve as the second opinion to assist better ILDs monitoring. This study would be valuable to clinicians, especially for radiological training. This web resource, ILD-DB, is expected to work as a uniform resource for public awareness, and it can assist governments and international organizations for framing better policies.

The second objective was to create an integrated genetic knowledgebase of ILDs. ILDgenDB is a unique and centralized knowledgebase that provides diverse genetic data and their analyses related to ILDs. The knowledgebase aims to endow the researchers with a complete and unique platform with unrestricted access to contemporary genetic data and its annotations. This knowledgebase contains 299 literature curetted disease candidate genes and their association with biological processes, miRNAs, pathways, and SNPs, etc. The potential role of DCGs in disease pathogenesis is validated using standard resources (GAD, GHR, OMIM, DISEASE, GeneCards, etc.). To incorporate a complete knowledge about ILD pathogenesis, the DCGs involvements in biological processes were identified through gene ontology, pathways and phenotype analyses. The outcomes of downstream analyses of DCGs have revealed the potentially significant role of "molecular binding and inflammatory host defenses", "immune systems", and "selective immune processes pathways" in ILDs' pathogenesis. Several miRNAs, SNPs, and miR-polymorphisms are identified in association studies that account for altered expression of DCGs. The miR-polymorphisms have already verified for their role in cancers, and their role in ILDs can be verified by adopting the similar protocol. Cytokines, interleukins, surfactant proteins, etc., are some of the potential identified DCGs that may have potential involvement in ILDs pathogenesis. These DCGs, pathways, SNPs, etc. can be verified experimentally to serve as diagnostics biomarkers. The main objective of this work is to provide contemporary genetic data in one place to significantly improve the efficacy of ILDs' management and monitoring through patient-specific therapy and also by providing tools for therapeutic response. Datasets and analysis provided here will be helpful for researchers and scientists working for the betterment of ILDs management.

The third objective was to identify and establish the role of ncRNAs as less-invasive clinical biomarkers of ILDs. The results of this objective have provided the integrated overview of non-coding RNAs, biological pathways and associated genes which are involved in ILDs pathogenesis. This objective has also discussed top regulatory associations. In total, 20 non-coding RNAs and 10 pathways are identified as potential molecular biomarkers and therapeutic targets (Table 1, Table 2, Figure 2). Signaling pathways, inflammatory responses, and cytokine-cytokine receptor interactions are identified as key processes involved in ILDs pathogenesis which could be used as a potential tool for disease diagnostics. Furthermore, this research objective established the association among ncRNAs (lncRNAs & miRNAs), ILDs pathways and immunological processes. These ncRNAs along with pathways could lead us to a novel therapeutic intervention for ILDs. The association between lncRNAs and miRNAs (ceRNAs) is

also established as a potential cause for disease pathogenesis and progression. Non-coding RNAs have already established as a potential tool for disease diagnosis and management. Several ongoing studies have reported that approximately 30-50% of lung biopsies could be prevented by ncRNAs based disease monitoring and management. Sensitivity and specificity of ncRNAs based disease diagnosis is significantly higher than other biomarkers. ncRNAs from body fluid could be used as less-invasive biomarkers for diseases like cancers and ILDs [170]. The outcomes of this research objective are projected to deliver better diagnostics, prognostics, and monitoring of ILDs cases. It would also allow building molecular model using ncRNAs, ILDs-genes, and pathways, which could be used to in new therapeutic interventions.

5.2 Future Work

This research could lead us to explore many possibilities for advancement of radiology and genetics of ILDs. In the future, the two data resources namely ILD-DB and ILDgenDB can be used as a training set and the data can be used for future application development and biomarker discovery. Following enhancement can be incorporated in the existing approach:

- Incorporation of machine learning approaches for supervised classification and identification of ILD's specific patterns.
- Comparative computer-aided analysis of radiological patterns to differentiate between ILDs and non-ILDs cases, and identify ILDs specific patterns among many ILDs classes.
- Disease candidate genes from ILDgenDB and ncRNAs biomarker can be used as training cases and similar gene expression patterns can be mined for ILDs specific genetic evolution for disease diagnosis and prognosis.

REFERENCES

[1]. Raghu G, Brown KK. Interstitial lung disease: clinical evaluation and keys to an accurate diagnosis. Clinics in Chest Medicine. 2004;25(3):409-19.

[2]. Donohue W, Laski B, Uchida I, et al. Familial fibrocystic pulmonary dysplasia and its relation to the Hamman-Rich syndrome. Pediatrics. 1959;24(5):786-813.

[3]. Borie R, Kannengiesser C, Debray MP, et al. The genetic diagnosis of interstitial lung disease: a need for an international consensus. American Journal of Respiratory and Critical Care Medicine. 2017;195(11):1538-9.

[4]. Devine MS, Garcia CK. Genetic interstitial lung disease. Clinics in Chest Medicine. 2012;33(1):95-110.

[5]. Bagnato G, Harari S. Cellular interactions in the pathogenesis of interstitial lung diseases. European Respiratory Review. 2015;24(135):102-14.

[6]. Leslie K. My approach to interstitial lung disease using clinical, radiological and histopathological patterns. Journal of clinical pathology. 2009;62(5):387-401.

[7]. Cottin V. Interstitial lung disease: new challenges and evolving phenotypes. Eur Respiratory Soc; 2010.

[8]. Camus P, Kudoh S, Ebina M. Interstitial lung disease associated with drug therapy. British Journal of Cancer. 2004;91(S2):S18.

[9]. White DA, Stover DE. Severe bleomycin-induced pneumonitis: clinical features and response to corticosteroids. Chest. 1984;86(5):723-8.

[10]. Kim Y-K. Extracellular microRNAs as biomarkers in human disease. Chonnam medical journal. 2015;51(2):51-7.

[11]. Abe H, MacMahon H, Engelmann R, et al. Computer-aided diagnosis in chest radiography: results of large-scale observer tests at the 1996–2001 RSNA scientific assemblies. Radiographics. 2003;23(1):255-65.

[12]. Arzhaeva Y, Prokop M, Tax DM, et al. Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography. Medical Physics. 2007;34(12):4798-809.

[13]. Suzuki K. A review of computer-aided diagnosis in thoracic and colonic imaging. Quantitative imaging in medicine and surgery. 2012;2(3):163.

[14]. Bradley B, Branley H, Egan J, et al. Interstitial lung disease guideline: the British Thoracic Society in collaboration with the Thoracic Society of Australia and New Zealand and the Irish Thoracic Society (vol 63, Suppl V, pg v1, 2008). Thorax. 2008;63(11):1029-.

[15]. Wells A, Hirani N. Interstitial lung disease guideline. Thorax. 2008;63(Suppl 5):v1-v58.

[16]. Raghu G, Mehta S. Interstitial lung disease (ILD) in India: Insights and lessons from the prospective, landmark ILD-India registry. Lung India: official organ of Indian Chest Society. 2016;33(6):589.

[17]. Tanizawa K, Handa T, Nakashima R, et al. The prognostic value of HRCT in myositis-associated interstitial lung disease. Respiratory Medicine. 2013;107(5):745-52.

[18]. Jiang F, Todd NW, Qiu Q, et al. Combined genetic analysis of sputum and computed tomography for noninvasive diagnosis of non-small-cell lung cancer. Lung Cancer. 2009;66(1):58-63.

[19]. Toledo RA, Sekiya T, Longuini VC, et al. Narrowing the gap of personalized medicine in emerging countries: the case of multiple endocrine neoplasias in Brazil. Clinics. 2012;67:3-6.

[20]. Borry P. Coming of age of personalized medicine: challenges ahead. BioMed Central; 2009.

[21]. Hamman L. Acute diffuse interstitial fibrosis of the lung. Bull Johns Hopkins Hosp. 1944;74:177-212.
[22]. Stack BH, Grant IW, Irvine WJ, et al. Idiopathic diffuse interstitial lung disease: a review of 42 cases. American Review of Respiratory Disease. 1965;92(6P1):939-48.
[23]. Liebow A. The interstitial pneumonias. Frontiers of pulmonary radiology. 1969:102-41.
[24]. Travis WD, King TE, Bateman ED, et al. American Thoracic Society/European Respiratory Society international multidisciplinary consensus classification of the idiopathic interstitial pneumonias. American Journal of Respiratory and Critical Care Medicine. 2002;165(2):277-304.
[25]. Jawad H, Chung JH, Lynch DA, et al. Radiological approach to interstitial lung disease: a guide for the nonradiologist. Clinics in Chest Medicine. 2012;33(1):11-26.
[26]. Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. American Journal of Respiratory and Critical Care Medicine. 2011;183(6):788-824.
[27]. Furukawa H, Oka S, Shimada K, et al. Genetics of interstitial lung disease: vol de nuit (night flight). Clinical Medicine Insights: Circulatory, Respiratory and Pulmonary Medicine. 2015;9:CCRPM. S23283.
[28]. Crystal RG, Gadek JE, Ferrans VJ, et al. Interstitial lung disease: current concepts of pathogenesis, staging and therapy. The American journal of medicine. 1981;70(3):542-68.
[29]. Gorbatovskiĭ I, Morozova O, Lotosh E, et al. Role of genetic factors in the etiology of silicosis. Meditsina truda i promyshlennaia ekologiia. 1996(7):13-6.
[30]. Kitazawa H, Kure S. Interstitial Lung Disease in Childhood: Clinical and Genetic Aspects: Supplementary Issue: Current Developments in Interstitial Lung Disease. Clinical Medicine Insights: Circulatory, Respiratory and Pulmonary Medicine. 2015;9:CCRPM. S23282.
[31]. Cho J-H, Gelinas R, Wang K, et al. Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. BMC medical genomics. 2011;4(1):8.
[32]. O'Reilly S. Epigenetics in fibrosis. Molecular Aspects of Medicine. 2017;54:89-102.
[33]. Tochimoto A, Kawaguchi Y, Yamanaka H. Genetic susceptibility to interstitial lung disease associated with systemic sclerosis. Clinical Medicine Insights: Circulatory, Respiratory and Pulmonary Medicine. 2015;9:CCRPM. S23312.
[34]. Garcia CK, Raghu G. Inherited interstitial lung disease. Clinics in Chest Medicine. 2004;25(3):421-34.
[35]. CALLAHAN WP, Sutherland JC, Fulton JK, et al. Acute diffuse interstitial fibrosis of the lungs. AMA archives of internal medicine. 1952;90(4):468-82.
[36]. Feldmann A. Familial Predisposition to Silicosis. Zentralblatt fur Arbeitsmedizin und Arbeitsschutz. 1960;10(10):229-33.
[37]. Bonanni PP, Frymoyer JW, Jacox RF. A family study of idiopathic pulmonary fibrosis: a possible dysproteinemic and genetically determined disease. The American journal of medicine. 1965;39(3):411-21.
[38]. Kreiss K, Danilovs J, Newman L. Histocompatibility antigens in a population based silicosis series. Occupational and Environmental Medicine. 1989;46(6):364-9.
[39]. Verleden G, Du Bois R, Bouros D, et al. Genetic predisposition and pathogenetic mechanisms of interstitial lung diseases of unknown origin. European Respiratory Journal. 2001;18(32 suppl):17S-29s.
[40]. De Leon AD, Cronkhite JT, Katzenstein A-LA, et al. Telomere lengths, pulmonary fibrosis and telomerase (TERT) mutations. PloS one. 2010;5(5):e10680.

[41]. Marrone A, Sokhal P, Walne A, et al. Functional characterization of novel telomerase RNA (TERC) mutations in patients with diverse clinical and pathological presentations. Haematologica. 2007;92(8):1013-20.
[42]. Tsakiri KD, Cronkhite JT, Kuan PJ, et al. Adult-onset pulmonary fibrosis caused by mutations in telomerase. Proceedings of the National Academy of Sciences. 2007;104(18):7552-7.
[43]. Shulenin S, Nogee LM, Annilo T, et al. ABCA3 gene mutations in newborns with fatal surfactant deficiency. New England Journal of Medicine. 2004;350(13):1296-303.
[44]. Lear T, Chen BB. Therapeutic targets in fibrotic pathways. Cytokine. 2016;88:193-5.
[45]. Sessa R, Hata A. Role of microRNAs in lung development and pulmonary diseases. Pulmonary circulation. 2013;3(2):315-28.
[46]. Pottier N, Maurin T, Chevalier B, et al. Identification of keratinocyte growth factor as a target of microRNA-155 in lung fibroblasts: implication in epithelial-mesenchymal interactions. PloS one. 2009;4(8):e6718.
[47]. Lawson WE, Blackwell TS, Gauldie J. Let it be: microRNAs impact interstitial lung disease. American Thoracic Society; 2011.
[48]. Kupczyk M, Kuna P. MicroRNAs—new biomarkers of respiratory tract diseases. Advances in Respiratory Medicine. 2014;82(2):183-90.
[49]. Nathan N, Borensztajn K, Clement A. Genetic causes and clinical management of pediatric interstitial lung diseases. Current opinion in pulmonary medicine. 2018;24(3):253-9.
[50]. Raghu G, Rochwerg B, Zhang Y, et al. An official ATS/ERS/JRS/ALAT clinical practice guideline: treatment of idiopathic pulmonary fibrosis. An update of the 2011 clinical practice guideline. American Journal of Respiratory and Critical Care Medicine. 2015;192(2):e3-e19.
[51]. Depeursinge A, Vargas A, Platon A, et al. Building a reference multimedia database for interstitial lung diseases. Computerized Medical Imaging and Graphics. 2012;36(3):227-38.
[52]. Oikonomou A, Prassopoulos P. Mimics in chest disease: interstitial opacities. Insights into imaging. 2013;4(1):9-27.
[53]. Martinez FJ, Flaherty K. Pulmonary function testing in idiopathic interstitial pneumonias. Proceedings of the American Thoracic Society. 2006;3(4):315-21.
[54]. Clement A, Nathan N, Epaud R, et al. Interstitial lung diseases in children. Orphanet journal of rare diseases. 2010;5(1):22.
[55]. Meyer KC, Raghu G. Bronchoalveolar lavage for the evaluation of interstitial lung disease: is it clinically useful? European Respiratory Journal. 2011:erj00695-2009.
[56]. Hutchinson JP, McKeever TM, Fogarty AW, et al. Surgical lung biopsy for the diagnosis of interstitial lung disease in England: 1997–2008. European Respiratory Journal. 2016:ERJ-00378-2016.
[57]. Travis WD, Costabel U, Hansell DM, et al. An official American Thoracic Society/European Respiratory Society statement: update of the international multidisciplinary classification of the idiopathic interstitial pneumonias. American Journal of Respiratory and Critical Care Medicine. 2013;188(6):733-48.
[58]. Tzouvelekis A, Kouliatsis G, Anevlavis S, et al. Serum biomarkers in interstitial lung diseases. Respiratory research. 2005;6(1):78.
[59]. De Lauretis A, Renzoni EA. Molecular biomarkers in interstitial lung diseases. Mol Diagn Ther. 2014;18(5):505-22.

[60]. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annual Review of Biochemistry. 2012;81:145-66.
[61]. Cui H, Xie N, Thannickal VJ, et al. The code of non-coding RNAs in lung fibrosis. Cellular and Molecular Life Sciences. 2015;72(18):3507-19.
[62]. Yarmishyn AA, Kurochkin IV. Long noncoding RNAs: a potential novel class of cancer biomarkers. Frontiers in genetics. 2015;6:145.
[63]. Chen T, He P, Tan Y, et al. Biomarker identification and pathway analysis of preeclampsia based on serum metabolomics. Biochemical and Biophysical Research Communications. 2017;485(1):119-25.
[64]. Marsland B, Königshoff M, Saglani S, et al. Immune system dysregulation in chronic lung disease. Eur Respiratory Soc; 2011.
[65]. Heward JA, Lindsay MA. Long non-coding RNAs in the regulation of the immune response. Trends in immunology. 2014;35(9):408-19.
[66]. Ben-Hamo R, Efroni S. MicroRNA regulation of molecular pathways as a generic mechanism and as a core disease phenotype. Oncotarget. 2015;6(3):1594.
[67]. Denisenko E, Ho D, Tamgue O, et al. IRNdb: the database of immunologically relevant non-coding RNAs. Database. 2016;2016.
[68]. Chen K, Song F, Calin GA, et al. Polymorphisms in microRNA targets: a gold mine for molecular epidemiology. Carcinogenesis. 2008;29(7):1306-11.
[69]. Chen X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Scientific Reports. 2015;5:13186.
[70]. Dai Q, Li J, Zhou K, et al. Competing endogenous RNA: A novel posttranscriptional regulatory dimension associated with the progression of cancer. Oncology letters. 2015;10(5):2683-90.
[71]. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. Nature. 2014;505(7483):344.
[72]. Martens-Uzunova ES, Böttcher R, Croce CM, et al. Long noncoding RNA in prostate, bladder, and kidney cancer. European Urology. 2014;65(6):1140-51.
[73]. Schaukowitch K, Kim T-K. Emerging epigenetic mechanisms of long non-coding RNAs. Neuroscience. 2014;264:25-38.
[74]. Singh V, Sharma BB. Laying the ground for research of interstitial lung disease in our country: ILD India registry. Lung India: official organ of Indian Chest Society. 2014;31(4):320.
[75]. Richeldi L, Rubin AS, Avdeev S, et al. Idiopathic pulmonary fibrosis in BRIC countries: the cases of Brazil, Russia, India, and China. BMC medicine. 2015;13(1):237.
[76]. Rosas IO, Dellaripa PF, Lederer DJ, et al. Interstitial lung disease: NHLBI workshop on the primary prevention of chronic lung diseases. Annals of the American Thoracic Society. 2014;11(Supplement 3):S169-S77.
[77]. Cosgrove GP, Bianchi P, Danese S, et al. Barriers to timely diagnosis of interstitial lung disease in the real world: the INTENSITY survey. BMC pulmonary medicine. 2018;18(1):9.
[78]. Akhter N, Rizvi NA. Interstitial Lung Diseases misdiagnosed as tuberculosis. Pakistan journal of medical sciences. 2018;34(2):338.

[79]. Deo MG. Doctor population ratio for India-The reality. The Indian journal of medical research. 2013;137(4):632.

[80]. Shortliffe EH, Cimino JJ. Biomedical informatics: computer applications in health care and biomedicine: Springer Science & Business Media; 2013.

[81]. Gagiya AK, Suthar H, Bhagat G. Clinical profile of interstitial lung diseases cases. Natl J Med Res. 2012;2:2-4.

[82]. Kalra MK, Maher MM, Rizzo S, et al. Radiation exposure from chest CT: issues and strategies. Journal of Korean medical science. 2004;19(2):159-66.

[83]. Karki R. Medical Record and its Importance. http://www.healthnet.org.np/reports/bpklcos/mrecord.html2017 [cited 2018].

[84]. Pinheiro L, Blake K, Januskiene J, et al. Geographical variation in reporting Interstitial Lung Disease as an adverse drug reaction: findings from an European Medicines Agency analysis of reports in EudraVigilance. Pharmacoepidemiology and drug safety. 2016;25(6):705-12.

[85]. Dresselhaus TR, Peabody JW, Lee M, et al. Measuring compliance with preventive care guidelines. Journal of general internal medicine. 2000;15(11):782-8.

[86]. Sharif R. Overview of idiopathic pulmonary fibrosis (IPF) and evidence-based guidelines. The American journal of managed care. 2017;23(11 Suppl):S176-S82.

[87]. Cowie MR, Blomster JI, Curtis LH, et al. Electronic health records to facilitate clinical research. Clinical Research in Cardiology. 2017;106(1):1-9.

[88]. Collett D. Modelling survival data in medical research: Chapman and Hall/CRC; 2015.

[89]. Healthy People 2020 US: U.S. Department of Health and Human Services; 2018 [updated 01/10/2018; cited 2018 01/10/2018]; Available from: https://www.healthypeople.gov/.

[90]. Armato III SG, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical Physics. 2011;38(2):915-31.

[91]. Dowling J, Malaterre M, Greer P, et al. The lung tissue research consortium: an extensive open database containing histological, clinical, and radiological data to study chronic lung disease. 2006.

[92]. Ogura A, Funabashi M, Nishiki S, et al. Japanese Society of Radiological Technology. Japan: Japanese Society of Radiological Technology; [cited 2018 01/10/2018]; Available from: https://www.jsrt.or.jp/data/english/.

[93]. The Pulmonary, Critical Care, Allergy and Sleep Medicine Program Department of Medicine, The University of California San Francisco [cited 2018 01/10/2018]; Available from: https://pulmonary.ucsf.edu.

[94]. De Andrade J, Schwarz M, Collard HR, et al. The idiopathic pulmonary fibrosis clinical research network (IPFnet): Diagnostic and adjudication processes. Chest. 2015;148(4):1034-42.

[95]. Reference centre for rare respiratory diseases. CHRU, BREST; [cited 2018 01/10]; Available from: https://www.chu-brest.fr/fr/notre-offre-soins/nos-specialites/femme-mere-enfant/pediatrie.

[96]. Veterans Health Information Systems and Technology Architecture (VistA). US: US Department of Veterans Affairs (VA); [updated 2018; cited 2018 01/10/2018]; Available from: http://worldvista.org.

[97]. MEDgle: Health, Wellness and Fitness. [cited 2018 01/10]; Available from: www.medgle.com.

[98]. Prasad DD, Ray S, Majumdar AK, et al. Real time medical image consultation system through Internet. Journal of Healthcare Engineering. 2010;1(1):141-54.
[99]. Kumar A, Passi A. Comparison and combination of iris matchers for reliable personal authentication. Pattern recognition. 2010;43(3):1016-26.
[100]. Rath GK, Gandhi AK. National cancer control and registration program in India. Indian journal of medical and paediatric oncology: official journal of Indian Society of Medical & Paediatric Oncology. 2014;35(4):288.
[101]. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. Journal of Clinical Epidemiology. 2005;58(4):323-37.
[102]. Hoque DME, Kumari V, Hoque M, et al. Impact of clinical registries on quality of patient care and clinical outcomes: A systematic review. PloS one. 2017;12(9):e0183667.
[103]. Meyer KC. Multidisciplinary discussions and interstitial lung disease diagnosis: how useful is a meeting of the minds? The Lancet Respiratory Medicine. 2016;4(7):529-31.
[104]. Sen T, Udwadia ZF. Retrospective study of interstitial lung disease in a tertiary care centre in India. The Indian journal of chest diseases & allied sciences. 2010;52(4):207.
[105]. Singh S, Collins BF, Sharma BB, et al. Interstitial lung disease in India. Results of a prospective registry. American Journal of Respiratory and Critical Care Medicine. 2017;195(6):801-13.
[106]. Staggers N, Weir C, Phansalkar S. Patient safety and health information technology: Role of the electronic health record. 2008.
[107]. Jönsson ÅLM, Simonsen U, Hilberg O, et al. Pulmonary alveolar microlithiasis: two case reports and review of the literature. European Respiratory Review. 2012;21(125):249-56.
[108]. Soares RV, Forsythe A, Hogarth K, et al. Interstitial lung disease and gastroesophageal reflux disease: key role of esophageal function tests in the diagnosis and treatment. Arquivos de Gastroenterologia. 2011;48(2):91-7.
[109]. Kumar R, Gupta N, Goel N. Spectrum of interstitial lung disease at a tertiary care centre in India. Advances in Respiratory Medicine. 2014;82(3):218-26.
[110]. Kashyap S, Mohapatra PR. Pulmonary alveolar microlithiasis. Lung India: official organ of Indian Chest Society. 2013;30(2):143.
[111]. Khaladkar SM, kumar Kondapavuluri S, Kamal A, et al. Pulmonary alveolar microlithiasis-clinico-radiological dissociation-a case report with radiological review. Journal of radiology case reports. 2016;10(1):14.
[112]. Devi HG, Rao KM, Prathima K, et al. Pulmonary alveolar microlithiasis. Lung India: official organ of Indian Chest Society. 2011;28(2):139.
[113]. Vos T, Barber RM, Bell B, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet. 2015;386(9995):743-800.
[114]. Jindal S, Malik S, Deodhar S, et al. Fibrosing alveolitis: a report of 61 cases seen over the past five years. The Indian journal of chest diseases & allied sciences. 1979;21(4):174-9.
[115]. Sharma S, Pande J, Verma K, et al. Bronchoalveolar lavage fluid (BALF) analysis in interstitial lung diseases--a 7-year experience. The Indian journal of chest diseases & allied sciences. 1989;31(3):187-96.

[116]. Subhash H, Ashwin I, Solomon S, et al. A comparative study on idiopathic pulmonary fibrosis and secondary diffuse parenchymal lung disease. 2004.
[117]. Disayabutr S, Calfee CS, Collard HR, et al. Interstitial lung diseases in the hospitalized patient. BMC medicine. 2015;13(1):245.
[118]. Bauer PR, Kalra S, Osborn TG, et al. Influence of autoimmune biomarkers on interstitial lung diseases: A tertiary referral center based case-control study. Respiratory Medicine. 2015;109(3):397-405.
[119]. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research. 2016:gkw943.
[120]. Falcon S, Gentleman R. Hypergeometric testing used for gene set enrichment analysis. Bioconductor case studies: Springer; 2008. p. 207-20.
[121]. McDermott JE, Wang J, Mitchell H, et al. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert opinion on medical diagnostics. 2013;7(1):37-51.
[122]. Pletscher-Frankild S, Pallejà A, Tsafou K, et al. DISEASES: Text mining and data integration of disease–gene associations. Methods. 2015;74:83-9.
[123]. Prakash Y, Martin RJ. Brain-derived neurotrophic factor in the airways. Pharmacology and Therapeutics. 2014;143(1):74-86.
[124]. Weaver TE. Synthesis, processing and secretion of surfactant proteins B and C. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease. 1998;1408(2-3):173-9.
[125]. Campo I, Zorzetto M, Bonella F. Facts and promises on lung biomarkers in interstitial lung diseases. Expert review of respiratory medicine. 2015;9(4):437-57.
[126]. Mishra S, Shah MI, Sarkar M, et al. ILDgenDB: integrated genetic knowledge resource for interstitial lung diseases (ILDs). Database. 2018;2018.
[127]. Richeldi L, Du Bois RM, Raghu G, et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis. New England Journal of Medicine. 2014;370(22):2071-82.
[128]. Shi Y, Massagué J. Mechanisms of TGF-β signaling from cell membrane to the nucleus. Cell. 2003;113(6):685-700.
[129]. Mørk S, Pletscher-Frankild S, Palleja Caro A, et al. Protein-driven inference of miRNA–disease associations. Bioinformatics. 2013;30(3):392-7.
[130]. Christmann RB, Wooten A, Sampaio-Barros P, et al. miR-155 in the progression of lung fibrosis in systemic sclerosis. Arthritis research & therapy. 2016;18(1):155.
[131]. Brown D, Rahman M, Nana-Sinkam SP. MicroRNAs in respiratory disease. A clinician's overview. Annals of the American Thoracic Society. 2014;11(8):1277-85.
[132]. Matsuura K, De Giorgi V, Schechterly C, et al. Circulating let-7 levels in plasma and extracellular vesicles correlate with hepatic fibrosis progression in chronic hepatitis C. Hepatology. 2016;64(3):732-45.
[133]. Haider BA, Baras AS, McCall MN, et al. A critical evaluation of microRNA biomarkers in non-neoplastic disease. PloS one. 2014;9(2):e89565.
[134]. Xie J, Liu M, Li Y, et al. Ovarian tumor-associated microRNA-20a decreases natural killer cell cytotoxicity by downregulating MICA/B expression. Cellular & molecular immunology. 2014;11(5):495.
[135]. Selman M, Vicens V, Mendoza C, et al. Subsets of fibroblasts show resistance to apoptosis independently of their interstitial lung disease origin. Federation of American Societies for Experimental Biology; 2013.

[136]. Aschner Y, Khalifah AP, Briones N, et al. Protein tyrosine phosphatase α mediates profibrotic signaling in lung fibroblasts through TGF-β responsiveness. The American journal of pathology. 2014;184(5):1489-502.
[137]. Rincon M, Irvin CG. Role of IL-6 in asthma and other inflammatory pulmonary diseases. International journal of biological sciences. 2012;8(9):1281.
[138]. Nogee LM. Genetic basis of children's interstitial lung disease. Pediatric allergy, immunology, and pulmonology. 2010;23(1):15-24.
[139]. Tan Z, Randall G, Fan J, et al. Allele-specific targeting of microRNAs to HLA-G and risk of asthma. The American Journal of Human Genetics. 2007;81(4):829-34.
[140]. George G, Mittal RD. MicroRNAs: Potential biomarkers in cancer. Indian Journal of Clinical Biochemistry. 2010;25(1):4-14.
[141]. Landi D, Gemignani F, Barale R, et al. A catalog of polymorphisms falling in microRNA-binding regions of cancer genes. DNA and cell biology. 2008;27(1):35-43.
[142]. Hu Z, Liang J, Wang Z, et al. Common genetic variants in pre-microRNAs were associated with increased risk of breast cancer in Chinese women. Human mutation. 2009;30(1):79-84.
[143]. Kelloff GJ, Sigman CC. Cancer biomarkers: selecting the right drug for the right patient. Nature reviews Drug discovery. 2012;11(3):201.
[144]. Bravo A, Cases M, Queralt-Rosinach N, et al. A knowledge-driven approach to extract disease-related biomarkers from the literature. BioMed research international. 2014;2014.
[145]. Goossens N, Nakagawa S, Sun X, et al. Cancer biomarker discovery and validation. Translational cancer research. 2015;4(3):256.
[146]. Bhatt AN, Mathur R, Farooque A, et al. Cancer biomarkers-current perspectives. Indian Journal of Medical Research. 2010;132(2):129-49.
[147]. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC bioinformatics. 2006;7(1):166.
[148]. Dai H-J, Wu JC-Y, Tsai RT-H, et al. T-HOD: a literature-based candidate gene database for hypertension, obesity and diabetes. Database. 2013;2013.
[149]. Ongenaert M, Van Neste L, De Meyer T, et al. PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Research. 2007;36(suppl_1):D842-D6.
[150]. R Lammi M, P Baughman R, S Birring S, et al. Outcome measures for clinical trials in interstitial lung diseases. Current respiratory medicine reviews. 2015;11(2):163-74.
[151]. Vencken SF, Greene CM, McKiernan PJ. Non-coding RNA as lung disease biomarkers. Thorax. 2015;70(5):501-3.
[152]. Bhan A, Mandal SS. Long noncoding RNAs: emerging stars in gene regulation, epigenetics and human disease. ChemMedChem. 2014;9(9):1932-56.
[153]. Okada Y, Muramatsu T, Suita N, et al. Significant impact of miRNA–target gene networks on genetics of human complex traits. Scientific Reports. 2016;6:22223.
[154]. Mishra S, Shah MI, Sarkar M, et al. Integrated analysis of non-coding RNAs for the identification of promising biomarkers in interstitial lung diseases. Gene Reports. 2018;11:87-93.
[155]. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Research. 2013;42(D1):D68-D73.

[156]. Quek XC, Thomson DW, Maag JL, et al. lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Research. 2014;43(D1):D168-D73.
[157]. Dweep H, Gretz N. miRWalk2. 0: a comprehensive atlas of microRNA-target interactions. Nature methods. 2015;12(8):697.
[158]. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. Bioinformatics. 2009;25(6):765-71.
[159]. Li J-H, Liu S, Zhou H, et al. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Research. 2013;42(D1):D92-D7.
[160]. Ballantyne M, McDonald R, Baker A. lncRNA/MicroRNA interactions in the vasculature. Clinical Pharmacology and Therapeutics. 2016;99(5):494-501.
[161]. Paraskevopoulou MD, Hatzigeorgiou AG. Analyzing miRNA–lncRNA interactions. Long Non-Coding RNAs: Springer; 2016. p. 271-86.
[162]. Dang X, Qu X, Wang W, et al. Bioinformatic analysis of microRNA and mRNA Regulation in peripheral blood mononuclear cells of patients with chronic obstructive pulmonary disease. Respiratory research. 2017;18(1):4.
[163]. Lu J, Clark AG. Impact of microRNA regulation on variation in human gene expression. Genome Research. 2012:gr. 132514.111.
[164]. Huang Z, Luo Q, Yao F, et al. Identification of differentially expressed long non-coding RNAs in polarized macrophages. Scientific Reports. 2016;6:19705.
[165]. Cardenas CLL, Henaoui IS, Courcot E, et al. miR-199a-5p Is upregulated during fibrogenic response to tissue injury and mediates TGFbeta-induced lung fibroblast activation by targeting caveolin-1. PLoS genetics. 2013;9(2):e1003291.
[166]. Qian L, Lin L, Du Y, et al. MicroRNA-588 suppresses tumor cell migration and invasion by targeting GRN in lung squamous cell carcinoma. Molecular medicine reports. 2016;14(4):3021-8.
[167]. Siquan Z, Zhu L, Liu H, et al. Over-Expressions of Serum miR-182-5p, miR-363-3p, and miR-378a-3p serve as Biomarkers in Hepatocellular Carcinoma. Molecular Biology. 2016;5:154.
[168]. Song X, Cao G, Jing L, et al. Analysing the relationship between lnc RNA and protein-coding gene and the role of lnc RNA as ce RNA in pulmonary fibrosis. Journal of cellular and molecular medicine. 2014;18(6):991-1003.
[169]. Sarkar M, Niranjan N, Banyal P. Mechanisms of hypoxemia. Lung India: official organ of Indian Chest Society. 2017;34(1):47.
[170]. Li Y, Huang J, Guo M, et al. MicroRNAs regulating signaling pathways: potential biomarkers in systemic sclerosis. Genomics, proteomics & bioinformatics. 2015;13(4):234-41.

PUBLICATIONS

Journal Publications

1. **Mishra S**, Shah MI, Sarkar M, Asati N, Rout C. ILDgenDB: integrated genetic knowledge resource for interstitial lung diseases (ILDs). Database. 2018. **DOI: 10.1093/database/bay053**
2. **Mishra S**, Shah MI, Sarkar M, Rout C. Integrated analysis of non-coding RNAs for the identification of promising biomarkers in interstitial lung diseases. Gene Reports. 2018; 11:87-93. **DOI: 10.1016/j.genrep.2018.03.002**
3. **Mishra S**, Shah MI, Sarkar M, Sharma S, Kumar N, Shaikh M, Rout C. Integrated medical data resource and meta-analysis for differential patterns in interstitial lung diseases (ILDs) from North-Western Himalayan region: Interstitial Lung Disease-Database (ILD-DB). (Under Process)

Conference Publications

4. **Mishra S**, Shah MI, Sarkar M. Characterisation of radiological imaging biomarkers for the identifications of interstitial lung diseases (ILDs) for developing countries. In Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on 2016 Dec 22 (pp. 348-350). IEEE. **DOI: 10.1109/PDGC.2016.7913174**
5. **Mishra S**, Shah MI, Sarkar M, Chaudhary N, Sharma S, Rout C. Clinical and radiological decision support system prototype for characterisation of interstitial lung disease (ILDS). **DOI: 10.1049/cp.2016.1461**
6. **Mishra S**, Shah MI, Sarkar M, Chauhan A, Sharma S, Rout C. Computer Aided health care system for interstitial lung disease, S Mishra et al., R&D expo by IEEE-JUIT- 2016
7. **Mishra S**, Shah MI, Sarkar M, Chauhan A, Sharma S, Rout C. Integration of medical Data for affordable healthcare, S Mishra et al. SciDataCon 2014.

Databases

1. ILDgenDB (Interstitial Lung Disease Genetic Database): A comprehensive knowledgebase of the genes contributing to the etiology and pathogenesis of interstitial lung diseases (ILDs). Accessible at- http://14.139.240.55/ildgendb/index.php.
2. ILD-DB (Interstitial lung disease database): An integrated image and clinical profile resource of interstitial lung disease, Shimla, India. Accessible at- http://14.139.240.55/ilddb/