Investigational New Drugs (2019) 37:384–400
https://doi.org/10.1007/s10637-018-0664-z

SHORT REPORT

Identification of differentially expressed genes and signaling pathways using bioinformatics in interstitial lung disease due to tyrosine kinase inhibitors targeting the epidermal growth factor receptor

Yuan Lu[1] · Ang Li[2] · Xiaofeng Lai[3] · Jun Jiang[4] · Lihong Zhang[1] · Zhicheng Zhong[1] · Wen Zhao[1] · Ping Tang[1] · Hu Zhao[5] · Xinling Ren[1]

Received: 21 June 2018 / Accepted: 28 August 2018 / Published online: 10 September 2018

Summary

Interstitial lung disease (ILD) is a rare but lethal adverse effect of epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors (TKIs) treatment. The specific mechanism of this disease is not fully understood. To systematically analyze genes associated with EGFR-TKI induced ILD, gene data of EGFR-TKI induced ILD were extracted initially using text mining, and then the intersection between genes from text mining and Gene Expression Omnibus (GEO) dataset was taken for further protein-protein interaction (PPI) analysis using String-bd database. Go ontology (GO) and pathway enrichment analysis was also conducted based on Database of Annotation, Visualization and Integrated Discovery (DAVID) platform. The PPI network generated by STRING was visualized by Cytoscape, and the topology scores, functional regions and gene annotations were analyzed using plugins of CytoNCA, molecular complex detection (MCODE) and ClueGo. 37 genes were identified as EGFR-TKI induced ILD related. Gene enrichment analysis yield 18 enriched GO terms and 12 associated pathways. A PPI network that included 199 interactions for a total of 35 genes was constructed. Ten genes were selected as hub genes using CytoNCA plugin, and four highly connected clusters were identified using MCODE plugin. GO and pathway annotation analysis for the cluster one revealed that five genes were associated with either response to dexamethasone or with lung fibrosis, including CTGF, CCL2, IGF1, EGFR and ICAM1. Our data might be useful to reveal the pathological mechanisms of EGFR-TKI induced ILD and provide evidence for the diagnosis and treatment in the future.

Keywords EGFR-TKI · Interstitial lung disease · Bioinformatics · Signaling pathway

Yuan Lu and Ang Li contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10637-018-0664-z) contains supplementary material, which is available to authorized users.

✉ Hu Zhao
zhaohubear@163.com

✉ Xinling Ren
majrenxl@szu.edu.cn

[1] Department of Respiratory, Shenzhen University General Hospital, Shenzhen University Clinical Medical Academy, Xueyuan AVE 1098, Xili University Town, Shenzhen 518055, Guangdong, People's Republic of China

[2] The State Key Laboratory of Cancer Biology, Department of Immunology, Air Force Military Medical University (Fourth Military Medical University), 169 Changle West Road, Xi'an 710032, People's Republic of China

[3] Department of Clinical Genetics and Experimental Medicine, Fuzhou General Hospital, Xiamen University School of Medicine, Fuzhou, Fujian 350025, People's Republic of China

[4] Department of Respiratory, Xijing Hospital, Air Force Military Medical University (Fourth Military Medical University), ChangleWest Road 127, Xi'an, 710032, People's Republic of China

[5] Department of Urology, Fuzhou Dongfang Hospital, Xiamen University, Xierhuan Northern Road 156, Fuzhou 350025, People's Republic of China

Introduction

Lung cancer is the leading cause of cancer related deaths in China, and the mortality and morbidity are still rising due to tobacco use and air pollution during the past decades [1]. Non-small-cell lung cancer (NSCLC) is the most common form of lung cancer, which account for 85% of total cases. Tyrosine kinase inhibitors (TKIs), such as gefitinib and erlotinib, benefits patients whose tumor harboring activating mutations of epidermal growth factor receptor (EGFR), and have been recommended as first line treatment for them in National Comprehensive Cancer Network (NCCN) guidelines [2].However, long-term exposure to EGFR TKIs often causes the occurrence of adverse effects (AEs), which distress two third of patients and seriously degrade patient quality of life. Furthermore, severe AEs might cause treatment withdrawal thereby reducing EGFR-TKI clinical efficacy [3].

Interstitial lung disease (ILD), known as diffuse parenchymal lung disease, encompass a wide and heterogeneous group of respiratory disorders, which are classified by specific clinical, radiologic and pathologic features. Although the pathogenesis of ILD remain poorly understood, increasing evidence suggests a specific environmental exposure or underlying disease may play a potent role in ILD development [4]. Prior studies have identified many therapeutic drugs responsible for the development of ILD, because the large contact surface of lung provide metabolic platform for certain substances [5–8]. EGFR TKIs that reversibly or irreversibly block EGFR downstream signaling by inhibiting tyrosine kinase phosphorylation of EGFR, which is usually activated in the response to lung injury, impede pulmonary repair by alleviating alveolar and bronchial epithelial migration, proliferation, differentiation, and extracellular matrix formation [9]. ILD was one of the most common withdrawal AEs with an incidence rate of 2.5 and 0.9% in Asian and non-Asian patients, respectively [10]. Importantly, ILD was estimated as the leading cause of treatment-related death (TRD) with an overall death rate of 38.6% [11]. It is clinically important to explore the possible mechanisms for the development of ILD in patients underwent EGFR-TKI treatment.

Nowadays, tremendous data are being generated, transmitted and stored, and text is the most common vehicle for the formal exchange of data. For the sake of processing immense quantities of data to extract useful information and construct disease-related interaction networks, text mining technology has been widely used in biomedical research [12].In this study, we use text mining to identify genes associated with the pathogenesis of EGFR-TKI-induce ILD. In addition, we performed GO and pathway analysis to explore the functions and pathways involved.

Materials and methods

Data collection

Text mining was performed by using the web-based service pubmed2ensembl (http://pubmed2ensembl.ls.manchester.ac.uk/), which is a publicly available resource links and integrates genomic data from 50 species and biological literature from MEDLINE life science journals, and online books [13]. We performed two queries: one with the concept "EGFR tyrosine kinase inhibitor", and another with the concept "interstitial lung disease" from 100,000 relevant document IDs. We choose "*Homo sapiens* (GRCh37)" as species dataset and "filter on MEDLINE" to constrain the query result. After that, the intersection of unique gene hits from the two gene sets was retrieved.

Differential expression analysis

Gene expression dataset (GSE40839) containing 11 ILD patients and 10 normal controls, which was based on [HG-U133A] Affymetrix Human Genome U133A Array, was obtained from the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and normalized for further analysis [14]. The differential expressed genes (DEGs) between the ILD patients and normal controls were extract based on empirical Bayes t-tests provided by the R/bioconductor Limma package, DEGs with fold change value>1, $P<0.05$) were selected. We then obtained the intersection of genes from text-mining and DEGs for the next steps.

Gene ontology and pathway enrichment analysis

Protein-protein interactive (PPI) network is the basic skeleton for proteins to determine their functions in the system biology. By setting the minimum interaction score to 0.4, the web-based STRING database (version 10.5) (http://string-db.org/) was used to produce PPI predictions after uploading the intersection gene list to the search bar. The PPI network was derived from experimentally proved statistics like automated text mining of the scientific literature, experimental data, available pathway and PPI database, systematic co-expression, phylogenetic co-occurrence, observed genome neighborhood, and gene fusion events [15]. We also perform Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis using the on-line Database of Annotation, Visualization and Integrated Discovery (DAVID) platform (version 6.8, https://david.ncifcrf.gov/) to reveal bio-information of the identified DEGs, both up- and down- regulated. A false discovery rate (FDR) less than 0.05, which represents the adjusted p value, was set as significance threshold to determine enrichment terms.

The PPI network generated by STRING was imported into Cytoscape software (version 3.60, http://www.cytoscape.org), which is a JAVA-based network visualization and analysis tool [16]. The topology scores of the nodes, including degree, subgraph centrality, betweenness centrality and closeness centrality, were measured using CytoNCA plugin (version 2.1.6) [17]. "Without weight" was set as the parameter. Then, we used molecular complex detection (MCODE) plugin (version 1.5.1) of Cytoscape to detect functional regions within the whole PPI network based on topology. The default parameters of MCODE was used: degree cutoff≥2, node score cutoff≥0.2, k-score ≥ 2, maximum depth = 100 [18]. The MCODE-generated gene cluster having highest score was integrated and quantitatively assessed (term *P* value) by ClueGo, another cytoscape plugin for GO and pathway annotations originated from WikiPathways, Reactome and KEGG database [19, 20].

Results

Identification of EGFR TKI and ILD related genes

As shown in Fig. 1, 675 unique genes were related to EGFR tyrosine kinase inhibitor, 1030 were related to interstitial lung disease, and 247 were commonly related to both queries (More information in Table 1). To verify the selected 247 genes were truly related to the two concepts, we download ILD-expression microarray dataset GSE40839. After being screened by the R/limma package (adjusted *p* value<0.05, |logFC|>1). 738 DEGs with statistical significance between normal controls and ILD patients were identified including 341 upregulated genes and 397 downregulated ones though volcano plot filtering (Fig. 2). A comparison between the gene lists derived from text mining and the ILD gene set further revealed an overlap of genes, with a total of 37 genes involved. Of them, 17 genes upregulated and 20 genes downregulated as shown in Fig. 3. A detailed list of the 37 genes and the corresponding statistics is shown in S.Table. 1.

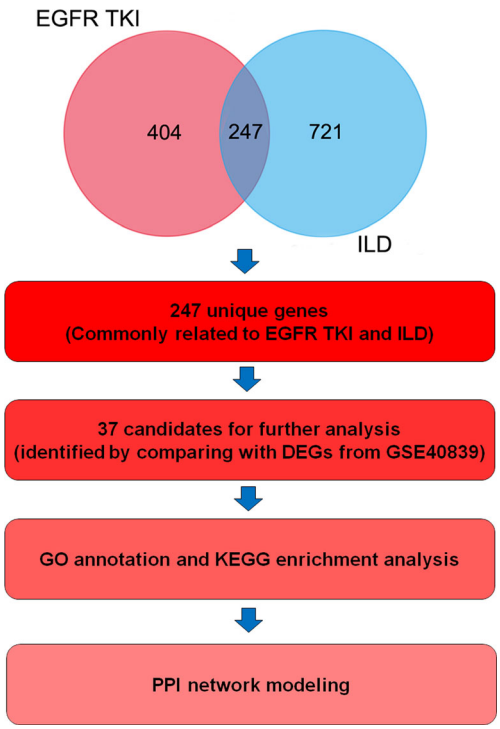

Fig. 1 Overall bioinformatics analysis strategy. Gene lists correlated with EGFR-TKI and ILD were identified by pubmed2ensemble and then intersected with DEGs from GSE40839. Further GO annotations and KEGG enrichment was obtained by DAVID. Final PPI network was modeled by STRING and Cytoscape

GO annotation and KEGG enrichment analysis

To determine the biological features of the common genes, we performed GO annotation and KEGG enrichment analysis by querying DAVID database [21]. DAVID' algorithm can do bioinformatic analysis by clustering gene groups across multiple databases, for example, GO and KEGG database. As a result, the top 18 significantly enriched GO terms, containing biological process, cell component and molecular function, were demonstrated in Fig. 4 and S. Table. 2. 11 genes fall into the category of "positive regulation of cell proliferation" (FDR = 9.53E-06), 10 falls into the category of "positive regulation of cell migration" (FDR = 1.19E-07), and 9 falls into the category of "response to hypoxia" (FDR = 1.06E-06), which were the top 3 terms of GO biological process (BP) having the largest number of genes enriched. GO cellular component (CC) annotation showed that the gene products mainly enriched in the extracellular space (59.5%), extracellular region (51.4%) and cell surface (29.7%). As to molecular function (MF) annotation, "growth factor activity" and "protein homodimerization activity" were the most enriched term with 9 genes hits, respectively.

The KEGG enrichment analysis revealed 12 terms were significantly associated with EGFR-TKI-induced ILD as shown Fig. 5 and S. Table. 3. The top enriched term was "cytokine-cytokine receptor interaction" with a FDR value of 3.08E-08(EGFR, CCL11, VEGFC, TNFRSF11B, TNFSF10, CCL2, CSF1, VEGFA, MET, PDGFRA, FAS, CXCL12, TGFB1). Besides "cytokine-cytokine receptor interaction", which involve several important signaling pathways (including Rap1 signaling pathway) having a critical influence on tumorigenesis and fibrogenesis, "Rap1 signaling pathway" was the most significantly enriched term containing 8 genes with an FDR value of 0.001368 (EGFR, VEGFC, CSF1, VEGFA, MET, PDGFRA, IGF1, NGF). Importantly, the 8 hit genes were also genes that highly enriched in the categories of "phosphatidylinositol 3-kinase/Akt (PI3K/Akt) signaling pathway" and "Focal adhesion", with relatively lower FDR values of 0.001456 and 0.011626, respectively.

Table 1 Detailed gene lists related with EGFR-TKI and ILD from text-mining

EGFR-TKI related genes	ILD related genes	Common genes
NME1	GSTM1	CD44
WNT3A	TAP2	MMP2
LBR	NLRP3	MLC1
TGFB2	LGALS8	XDH
IRF6	NID1	EDN1
IL24	ACTA1	CXCR4
IL10	PARP1	AIFM1
AVPR1B	STMN1	IGF1
PIK3C2B	EPHX1	TGFB2
REN	SRP9	PLAU
PTGS2	LBR	CYCS
NCF2	TGFB2	MMP7
LAMC2	KCNH1	NOS2
SOAT1	C1orf107	BAK1
PSMD1	IRF6	MET
FASLG	HSD11B1	MAX
IRS1	PLXNA2	NGF
PAX3	CD34	MSN
CD247	CD46	LEP
DES	CR1	GJA1
ABCB6	CR2	NOS1
USF1	FAIM3	PRL
NHLH1	IL10	ACACA
PNKD	PDCD1	ELANE
GPBAR1	CTSE	SMAD3
IL8RA	REN	PLCG1
IL8RB	SNRPE	NPPB
NTRK1	CHI3L1	SMUG1
FN1	CHIT1	LGALS1
NES	RNPEP	MUC16
ERBB4	TNNT2	FAS
PIKFYVE	PTPRC	GRB2
CREB1	GLRX2	F2RL1
MUC1	PTGS2	HBEGF
CD28	NCF2	BDNF
SHC1	LAMC2	CSF1
IL6R	NPHS2	TNFSF10
CASP8	SOAT1	MUC1
CFLAR	PSMD1	BCL2
S100A4	TNR	TNFRSF10A
S100A7	TNFSF4	PTHLH
FLG	FASLG	NOVA2
HSPD1	CCL20	TNFRSF11B
STAT1	COL4A3	COL18A1
MCL1	COL4A4	NCF2
ECM1	PRRX1	TF
ITGAV	SELE	NEFM
FCGR1A	SELL	KCNH2
FRZB	F5	CASP8
CD160	SELP	TAP1
PIAS3	SLC19A2	TSC2
TNK2	XCL1	NCAM1
ATF2	DPT	PDP1
SP3	SLC4A3	AGT
NGF	CD247	IL6R
TP63	UCK2	PPARG
PDK1	DES	IRF6
KNG1	FCGR3B	CALCA
KCNA3	FCGR3A	MYBL2
CSF1	ABCB6	GSTP1
PIK3CA	CYP27A1	CCR3
TNFSF10	F11R	EGR1
VCAM1	CD244	FN1
F3	ITLN1	ABCB1

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
GFI1	CD48	AHR
HNMT	VANGL2	WT1
CXCR4	SLC11A1	MAPK14
CLCA1	PNKD	PDGFRB
MGAT5	CRP	SMPD1
PLG	PYHIN1	TNF
LPA	IGFBP2	LCK
DIRAS3	CD1C	C5
LEPR	CD1D	EGF
ESR1	INSRR	IL1A
A1BG	FN1	FYN
IL1B	NES	VEGFC
IL1A	LMNA	PDGFB
RHO	RIT1	ELN
JUN	GBA	VIM
MAP3K7IP2	MTX1	NUDT6
BCL2L11	MUC1	DES
SEC61A1	CTLA4	IL8
PLAGL1	CD28	PCNA
SHH	IL6R	KIT
MAP3K5	CASP8	SHH
KALRN	CFLAR	JAK2
FASN	S100A1	TNFRSF1A
SLC9A2	S100A6	STAT3
MYLK	S100A12	SPP1
CDKN2C	S100A9	PTGS2
CTGF	IVL	OSM
NOS3	TCHH	SLPI
KCNH2	CGN	CPM
ENPP1	STAT4	CTNNB1
CASR	CTSK	NT5E
GSK3B	MCL1	VTN
PTPRK	TFPI	CCL4
NR1I2	FCGR1A	ATM
ARHGEF12	DNAJC10	ESR1
CD80	ITGA4	TAC1
CBL	HFE2	CDK4
FABP7	TFRC	TFF2
GJA1	NFE2L2	TYMS
TIMP2	PHGDH	CAMP
UPK2	HES1	HRAS
TK1	IGSF2	VEGFA
ROS1	CD2	IL3
TAS2R38	WIPF1	TGFB1
FPR1	CD58	MYC
MKI67	SP3	TIMP1
BRAF	CLDN1	PI3
CD33	NGF	SOAT1
ADAM12	TP63	IL13
CD3E	SST	MCL1
FYN	KNG1	CCL2
TMPRSS4	DNAJB11	MMP1
SPHK1	PTPN22	IL6
DUSP2	SSB	CTGF
KLK15	PPIG	ANXA5
KLK1	PSMD2	SYP
FUT7	ABCC5	NDE1
TRAF2	CSF1	HMGCR
YBX1	GSTM3	ITGAM
FOXO3	GALNT3	MST1R
NOTCH1	IFIH1	FASLG
FGFR2	DPP4	GGT1
AKR1B1	TNFRSF13B	CCL11
GRB2	TNFSF10	KDR
NCAM1	RNPC3	FLT1

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
CPA1	VCAM1	SMAD2
SLC9A3R1	DPYD	MMP3
ATM	F3	KRIT1
MAP3K7	ABCA4	F13A1
ADRB1	GCLM	BCL2L11
CALU	CXCR4	MT1E
MAP2K6	PFN2	TGFB3
CASP1	CP	NTRK2
LEP	CPB1	IL10
BAX	AGTR1	MAPK1
MMP3	CCR6	IGF2
TCF7L1	PLSCR1	CD80
MMP1	CFC1	TSC1
MMP10	PLG	ALOX5AP
MMP7	IGF2R	MMP14
BIRC2	MAS1	PLAT
BIRC3	SOD2	MDM2
SPHK2	VIL2	FGF7
NT5E	PROC	EGFR
TSC1	TMEM37	SP1
PGR	CCRL1	NODAL
CACNG4	MARCO	SOD1
PRKCA	ESR1	CXCR3
PDHB	IL1RN	BCR
TACR1	TRH	IL2
GNA13	RAET1E	ERBB2
MET	IL1B	IGFBP1
CAV1	IL1A	DDIT3
MTNR1B	BCL2L11	CDH1
SH3PXD2A	B3GAT1	IFNG
GH1	STX11	FOS
ABL1	NLRP5	PLAG1
SLC17A5	SHH	IL4
PXN	OPCML	CD28
NRCAM	IL11	CLU
NOVA2	HECA	CXCL12
DLD	IL22RA2	MAPK8
PPP2R4	CD7	IGF1R
NOS1	AHI1	ICAM1
FOSB	TAP2	CAV1
PIK3CG	IL18R1	BCL2L1
ERCC1	TCF21	ABL1
LCK	EPS15	MAPK3
PAK1	IL1R1	CCL28
DTX1	CTGF	SLC9A1
PVR	NOS3	TERT
PTPN11	KCNH2	CD247
EMD	ENPP1	HIF1A
ERP29	ESAM	NOS3
RPS6KB1	ARG1	PTH
OPN1LW	CASR	RB1
TGFA	CD86	SRC
PLAUR	STIL	KNG1
XRCC1	PDZK1IP1	CASR
CISH	CFHR1	NR0B2
SERPINE1	PRKCG	DSG3
LPO	CD80	GFAP
PTAFR	GJA1	ACTB
MUT	ADAM8	TLR4
TGFB1	TIMP2	SP3
DNTT	UROD	PLG
SLC9A1	ROS1	SERPINE1
NR0B2	TMC8	VCAN
MST1R	TMC6	MKI67
SFN	MGMT	VCAM1

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
CTTN	MKI67	PSMD1
XPO1	CD200	TP53
RPS6KA1	WISP3	SFTPC
CCND1	FYN	LOX
MIA	APOA1	MSLN
VEGFA	UNC13D	IFNA1
MAPKAP1	SLC2A1	DLD
DAG1	AKT1S1	CDKN1A
MBTPS1	SEC23IP	CAT
EME1	GRB2	LAMC2
RHOA	NCAM1	ADAM17
IGF1	CD8A	LYZ
GSTP1	IL18	ENPP1
CYP3A4	C2	CFLAR
PSMC4	EIF3A	AKT1
BCAR1	PAEP	GRP
FAS	FCGRT	PDGFRA
TAC1	MCHR2	ABCB6
PTEN	FLT3LG	FCGR1A
AIFM1	TSPAN33	NES
C5	STX2	GH1
METAP2	ATM	NFKB1
EPHB2	SFTPB	KRT5
FRS3	PROK2	TIMP2
SOCS2	SLN	IL1B
CDC42	CASP8AP2	REN
SH2D1A	SNRNP70	KITLG
PLXNB1	CARD18	IGFBP3
MSH2	LEP	ALB
TLR4	BAX	F3
CDK6	POU3F1	MLH1
VAC14	ABO	CD69
HSPG2	MMP13	CDKN2A
CAMP	FUT1	SERPINB3
CDC25A	MMP12	CCL3
FOSL1	MMP3	NF2
MAP3K14	MMP1	EPHB2
XIAP	MMP8	PLAUR
PIGF	FUT2	CISH
KRIT1	MMP7	HSPG2
EXOSC6	NDUFA5	TWIST1
KITLG	CEL	BAX
GFAP	NT5E	MIB1
NTS	TSC1	SLC17A5
DDOST	PSMB2	HGF
CDH1	CFTR	BTG3
PIM1	FUT4	TP63
PTK6	DNASE1L3	APP
PLA2G2A	MET	ROS1
ABCB1	FLNB	PNKD
CDKN1A	CAV1	CSF2
CCR3	C5AR1	MMP9
PYY	GH1	LBR
FLT4	ABL1	
MAPK14	INA	
MAPK9	ASS1	
SOS1	NOX4	
PLAU	DYNLL1	
HGF	CTSC	
BRCA1	DNAJB9	
NOL3	SLC17A5	
BECN1	TOR1A	
TRADD	CHDH	
FRS2	NOVA2	
LYZ	DLD	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
CPM	DMPK	
EIF2AK2	SLC26A3	
MDM2	DKC1	
IFNG	NOS1	
CLDN2	ZBTB8OS	
ABCA1	SYPL1	
PTRF	G6PD	
BAD	POLL	
BAK1	ACE	
STAT3	LCK	
GOT2	BCAM	
STAT5A	DNASE1L1	
STAT5B	SCD	
FGFR4	SERPINH1	
IGF1R	CD207	
EPHA2	FLNA	
XDH	OPN1MW2	
CTNNB1	OPN1MW	
TLR3	CPN1	
NR4A3	PLAUR	
MMP15	HPS1	
NODAL	GCLC	
AIFM2	CISH	
CDK4	LCN2	
DUSP1	CEACAM8	
NPM1	SERPINE1	
HK1	MPO	
TAP1	ABCD1	
KRT10	ZMYND10	
DDIT3	CIC	
VEGFC	RASSF1	
TIMM8A	IL17A	
TNS4	CRISP3	
MT1E	BGN	
TOP2A	CD79A	
RARA	EPO	
ELN	C2	
GPT	FGR	
HPGD	GALNS	
PTPN1	IL18BP	
CEBPB	TAP2	
PTTG1	C2	
CCAR1	MEP1A	
SLC25A16	TGFB1	
FANCC	MVD	
MYD88	CYBA	
MAPK11	GABRQ	
MMP2	SLC9A1	
SCRT1	AC005921.3	
AKTIP	NR0B2	
ITK	CA5A	
RBL2	MST1R	
UCN	RUNX2	
GRB7	FADD	
ERBB2	FOXF1	
NPPB	CD52	
MLC1	SHKBP1	
SPRY2	VEGFA	
DCT	HSPA5	
CCNH	COL1A1	
CYP19A1	ASCL1	
EIF2S1	PAH	
RAC1	PLCG2	
IFNA1	IGF1	
CDKN2A	GSTP1	
EDN1	CD40LG	
CDKN2B	GNPTAB	
OTC	AIP	
COPS5	AC005393.1	
MUC16	MAF	
PCSK1	SAGE1	
MAP2K4	C2	
NCOA2	SRP9L1	
NDE1	KARS	
NEDD4	NR1H4	
ABCC1	ADRBK1	
NME2	HPRT1	
ID1	GRHL3	
BCL2L1	CBX1	
BHLHE40	ELF3	
PRL	FAS	
AQP3	TAC1	
ICAM1	LIPF	
VIM	GSN	
HEY1	PRPH2	
AHR	HP	
AGT	TRERF1	
HCK	AIFM1	
ADAM10	UCN2	
OXTR	ACTN4	
SNX16	C5	
TWIST1	EPHB2	
RAPGEF3	TAT	
BDNF	COL1A2	
ARAF	CALB2	
DSPP	CALCR	
TIMP1	MAPT	
IRS2	TLR4	
SPP1	TREM1	
RAPGEF5	EEA1	
IL6	HSPG2	
ABCG2	CAMP	
LOX	DCN	
WT1	ERVWE1	
CCL27	TNC	
BMI1	KRIT1	
E2F1	AARS	
WNT1	KITLG	
SNX1	SFTPD	
PDP1	GFAP	
SLC12A2	GLO1	
HDAC6	COG8	
FOS	CDA	
EIF4E	CDH1	
CYCS	SLC4A1	
CAT	SLC7A6	
TGFB3	ABCB1	
CD44	CDKN1A	
VHL	LTF	
CA9	CCR5	
IL3	CCR3	
CSF2	AGTR2	
PROCR	CXCR6	
MMP24	TYROBP	
NFKB1	MAPK14	
IRF1	VCL	
LGALS1	RNF130	
IL13	PLAU	
IL4	AGRP	
MAP2K1	PRODH2	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
SYP	HGF	
MAP2K3	CANX	
FKBP14	CD36	
PPARG	MUSK	
PRDX2	LAMA5	
NOS2	TXN	
SMAD3	AOC3	
PAPSS1	COL4A5	
GRHL2	SS18L1	
RAF1	NHP2	
MAP3K8	LYZ	
DVL1	CPM	
VTN	MDM2	
ZEB1	RRAD	
EGF	DDX41	
ENPEP	TSC22D3	
MDC1	USF2	
PDGFB	MLN	
PKM2	VIT	
ADAM17	IL22	
MAP3K7IP1	FXYD5	
KRT5	IFNG	
RBL1	EDN3	
EGR1	SLC44A1	
HSPA9	TH1L	
ODC1	BAK1	
CTNNA1	STAT3	
PKN1	SMC2	
PTGER1	HCRT	
SRC	DAXX	
RALA	IGF1R	
ANXA5	GPI	
TNFRSF11B	CYP4V2	
HES5	XDH	
NRG2	CTNNB1	
HBEGF	PRF1	
EP300	NODAL	
IL2	PPA1	
CHGA	RING1	
RET	CCR8	
NUDT6	RXRB	
IGFBP6	CYP27B1	
INTU	CX3CR1	
TP73	XIRP1	
NF1	CDK4	
CXCL12	ACR	
PLCG1	HLA-DPB1	
MYCN	KRT19	
MAPK3	FGF18	
CSK	CCL17	
SP1	BLM	
MSN	CX3CL1	
TNF	CCL22	
PPYR1	AURKA	
RPL22	GTF2I	
MYC	IQGAP1	
RHOB	TAP1	
BDKRB2	GLA	
BDKRB1	SCN5A	
MYBL2	LAT2	
TOP2B	DDIT3	
SMUG1	AGA	
CCL2	C2	
CCL11	ANPEP	
GAB1	VEGFC	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
HNF4A	CCR7	
FGF1	HLA-DQB1	
MAPK8	PLIN	
F2RL3	HLA-DQA1	
IGFBP1	ARHGAP9	
IGFBP3	MT1E	
TNS3	HLA-DRB1	
PI3	BTNL2	
UTS2	ELN	
SLPI	ROM1	
PPP1R1A	WIPF2	
ERRFI1	PTCH1	
TGFBR2	IFT172	
UPP1	MTA2	
CCL3	SNAI1	
PTK2	IL12B	
CCL4	LRP1	
SPINK1	AGER	
NTRK2	TNFRSF1B	
TF	CSF3	
NCOA4	MMP2	
HSP90AA2	TNFRSF8	
ITGAM	NTRK3	
CDK2	SCGB1A1	
HSPA1A	TAC3	
HSPA1B	CHM	
JAK3	FTHL16	
CXCR3	ERBB2	
SEC61G	PLOD1	
ERBB3	NPPB	
ACACA	CPE	
EGFR	FEN1	
TLR2	CYLD	
CCRK	NOD2	
PDGFRB	MLC1	
DFFA	APPL1	
MMP9	DNM1L	
LRAT	CCL25	
MTOR	HAPLN1	
SYK	CXCL9	
TNIP1	CXCL10	
MLH1	S100B	
CDC2	CXCL11	
AKT1	PES1	
NCOA3	PIK3C2A	
VCAN	NUCB2	
MAX	CRH	
CALCA	ABCC8	
MAP2K7	IFNB1	
DSP	THBD	
FGF7	CXCL13	
DHFRP1	ARG2	
BTC	CYBB	
OSM	IFNA2	
AREGB	IFNA1	
AREG	CDKN2A	
ACTB	EDN1	
EREG	MYOD1	
ESR2	ANTXR2	
PTH	VSX1	
IL8	MUC16	
F13A1	SAA1	
PLAG1	CD83	
NF2	GLRX	
PSIP1	NDE1	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
PDXP	SLC10A1	
ALB	TNFRSF6B	
F2RL1	RTEL1	
F2R	REM1	
PTHLH	HM13	
COL18A1	TPMT	
GNA12	BCL2L1	
EWSR1	LY96	
RB1CC1	CDS1	
HIF1A	PRL	
INSR	ICAM1	
HMGCR	VIM	
SNAI2	IL7	
PTTG1IP	FER	
RIPK1	SMG1	
DUOX1	AHR	
BHLHE41	ICAM3	
SH3KBP1	TIMP3	
PRKDC	AGT	
TP53	HDAC9	
PDHA1	FABP4	
TUBB4	MRC1	
JAG1	APC	
ETNK1	TWIST1	
GNRHR	HMOX1	
C21orf33	FLCN	
JAK2	BDNF	
IKBKB	TNFSF13B	
GGT1	ANXA2	
PLAT	MB	
GRPR	DNAI1	
OCLN	VDR	
KDR	TIMP1	
KIT	SPP1	
PDGFRA	COL2A1	
SFRP1	UMOD	
ILK	HFE	
PCNA	OGG1	
PIK3R1	IL6	
PTGER2	RIPK2	
RB1	GALT	
SMPD1	LOX	
PTPRS	ACSM3	
TFF1	DECR1	
TFF2	WT1	
TFF3	CCL19	
BCR	CCL21	
CDC25B	MMRN1	
FGFR1	DLSTP	
TMPRSS2	PGF	
MAPK1	LMNB1	
CDKN1B	MCF2L	
TNFSF11	WAS	
IL6ST	PDP1	
TSC2	NCF4	
MAP2K2	FBN2	
PLD2	TST	
GFER	FOS	
TLR6	F10	
ARRB2	CYCS	
MATK	CAT	
FOXO1	HINT1	
NRG1	IL2RB	
TBXA2R	TGFB3	
CCL28	CD44	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
TAP1	SCNN1G	
GHR	IL3	
NFKBIA	CSF2	
PSMA6	GSTT1	
DUSP4	ALDH3A2	
CD69	AP003117.1	
PPARGC1A	GSTZ1	
CARD6	NFKB1	
PTGER4	TAP2	
RAD51	IL5	
SERPINB3	LGALS1	
TAP1	AQP2	
FYB	ALKBH1	
SERPINB5	IL13	
TAP1	IL4	
BCL2	HLA-A	
FKBP1A	SYP	
C4A	TNFRSF18	
IGF2	TNFRSF4	
CLU	PPARG	
GRP	FOXP3	
CSNK2A1	LGALS9	
PTK2B	GARS	
TAP1	TPO	
ADCY10	NOS2	
SKP2	HSPA4	
MSLN	SMAD3	
GNRH1	IL4R	
NEDD4L	GCDH	
NEFM	HADH	
MUC2	C2	
SEC14L2	HLA-E	
MUC6	VTN	
ALOX5AP	RPL32	
TNFRSF10A	ADI1	
SMAD4	CFI	
HSPA1B	CACNA1A	
BID	EGF	
IFNGR2	NEUROG1	
FLT1	ITGB1	
TNFRSF10B	ACVRL1	
PDZD2	GRASP	
HSPA1B	KRT7	
HSPA1A	PDGFB	
AXIN1	ADAM17	
TAP1	IER3	
STK11	RABEP2	
DYM	DDR1	
HSPA1B	KRT5	
HSPA1A	CD19	
CHRNA7	EGR1	
SMAD2	GRAP2	
SFTPC	GHRH	
SOD1	KRT8P9	
HRAS	SRC	
NME1-NME2	SULT1A4	
CHRFAM7A	ANXA5	
TJP1	SPI1	
BCL2L2	TNFRSF11B	
ELANE	CTNNBL1	
SRD5A1	CCNA2	
TNF	SPN	
TNFRSF1A	LGMN	
CD9	FXN	
APP	CD276	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
TERT	PML	
TNF	BPI	
GJB2	COL14A1	
MMP14	SLC6A4	
TNF	HBEGF	
BTG3	IL2	
SLC9A3	FGF2	
TNF	CDIPT	
DSG2	NUDT6	
DSG3	HLA-C	
TNF	HLA-B	
TNF	ADAP2	
CDH2	CXCL12	
MDC1	PLCG1	
MDC1	CD14	
MIB1	ALOX5	
MDC1	MAPK3	
ANGPT2	SERPINA6	
MDC1	SERPINA1	
TAP1	ANXA1	
HSPA1B	SP1	
HSPA1A	HARS	
TYMS	HARS2	
MDC1	SERPINA5	
MT-CO2	SERPINA3	
	MSN	
	SULT1A3	
	LTA	
	TNF	
	MYC	
	SDC1	
	NDUFA6	
	AR	
	TCL1A	
	ITGAL	
	IL15	
	MYBL2	
	SMUG1	
	CBX5	
	CCL2	
	CCL7	
	CCL11	
	TNFRSF25	
	LRRC6	
	CCL1	
	GYPB	
	A4GALT	
	MIF	
	ADA	
	WISP1	
	ITGA5	
	MAPK8	
	NR3C1	
	IGFBP1	
	EVL	
	IGFBP3	
	IL2RG	
	LARS	
	PI3	
	TNFRSF9	
	SLPI	
	ENO1	
	CCL5	
	DDAH2	
	CCL16	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
	CCL15	
	CCL18	
	CCL3	
	CCL4	
	POMC	
	TLR9	
	NTRK2	
	TF	
	LRBA	
	CD63	
	ITGAM	
	SLURP1	
	ITGAX	
	GIF	
	CCR4	
	GLB1	
	CXCR3	
	MS4A2	
	ACACA	
	EGFR	
	DAPK1	
	RPS4X	
	IL12RB1	
	CSF1R	
	PDGFRB	
	IL16	
	MS4A1	
	MMP9	
	MBL2	
	CD40	
	GPR44	
	SYNPO	
	EMILIN1	
	GUSB	
	MLH1	
	RBM12	
	AKT1	
	MTHFR	
	VCAN	
	MAX	
	DMD	
	CALCA	
	FSCN1	
	SOCS1	
	FGF7	
	TAP2	
	CIITA	
	OSM	
	LIF	
	ACTB	
	PTH	
	CXCL5	
	PPBP	
	PF4	
	FCER2	
	CXCL1	
	IL8	
	F13A1	
	PLAG1	
	NF2	
	PER1	
	AFP	
	FBN1	
	ALB	
	DDX53	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
	PHEX	
	F2RL1	
	PTHLH	
	COL18A1	
	PPL	
	OTOR	
	HIF1A	
	GC	
	XBP1	
	HMGCR	
	ITGB2	
	C16orf5	
	ADM	
	ODAM	
	MN1	
	HTN1	
	IL2RA	
	SERPINB1	
	B2M	
	CSN1S1	
	TP53	
	TNFSF9	
	CD68	
	IRF4	
	PDCD1LG2	
	CTSD	
	SERPINB2	
	CD274	
	KIR2DL3	
	JAK2	
	TUB	
	HOPX	
	PLCB1	
	IL32	
	GGT1	
	PLAT	
	FUT3	
	LGALS3	
	OR10A4	
	ACE2	
	TAF9	
	KDR	
	AC145132.3	
	KIT	
	TPP1	
	PDGFRA	
	BMP4	
	IDO1	
	CABIN1	
	FANCB	
	PITRM1	
	PCNA	
	CD180	
	RB1	
	ADRA1D	
	APBB1	
	SMPD1	
	TFF2	
	GTPBP4	
	IL13RA1	
	ABCA3	
	BCR	
	SIGLEC1	
	EMP1	
	DAXX	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
	UCHL1	
	MAPK1	
	ITPA	
	TRIM21	
	GP1BA	
	CD99	
	TSC2	
	STK19P	
	AVP	
	DDX4	
	RING1	
	ETV6	
	MBP	
	RXRB	
	DAXX	
	CXCL16	
	GZMA	
	IL3RA	
	DAXX	
	TLR1	
	CSF2RA	
	SHOX	
	TBC1D1	
	ITGA1	
	AC022075.1	
	KLRD1	
	NKX2–1	
	FGF10	
	RING1	
	ERG	
	CCL28	
	DAXX	
	TAP1	
	RXRB	
	RING1	
	GSR	
	CD226	
	SLC34A2	
	TAP2	
	RXRB	
	CD69	
	KCNQ1	
	SOD3	
	KLRB1	
	HLA-DPB1	
	BTNL2	
	EXOSC8	
	SERPINB3	
	RING1	
	A2M	
	ITGAE	
	TAP1	
	RXRB	
	TAP1	
	AGER	
	BCL2	
	TNFRSF11A	
	IGF2	
	CLU	
	TPSAB1	
	PMAIP1	
	CD38	
	HLA-DRB3	
	HLA-DPB1	
	BTNL2	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
	BTNL2	
	GRP	
	AMH	
	CLEC4C	
	CLDN5	
	TAP1	
	KCNE1	
	CD163	
	GZMB	
	CTSG	
	CMA1	
	AGER	
	MSLN	
	CLNK	
	THBS1	
	PNMA2	
	NEFL	
	NEFM	
	C1S	
	U47924.1	
	NARS	
	STUB1	
	SERPINF1	
	SERPINF2	
	TINF2	
	ENO2	
	ALOX5AP	
	RING1	
	C4B	
	AGER	
	SLC7A1	
	TARS	
	RXRB	
	TNFRSF10A	
	TNFRSF10D	
	NOP10	
	FLT1	
	CD4	
	YWHAE	
	FCGR2B	
	ABR	
	DDAH2	
	TAP1	
	GREM1	
	TAP2	
	ING4	
	SMAD7	
	SMAD2	
	MRAP	
	SFTPC	
	HUNK	
	LRPAP1	
	CENPJ	
	HBA1	
	SOD1	
	HBA2	
	HRAS	
	DDAH2	
	C16orf35	
	AGER	
	MPG	
	DDAH2	
	GAPDH	
	ELANE	
	PRTN3	

Table 1 (continued)

EGFR-TKI related genes	ILD related genes	Common genes
	AZU1	
	SETBP1	
	CD27	
	LPL	
	LTBR	
	TNF	
	TNFRSF1A	
	DDAH2	
	PALM	
	C2	
	NAT2	
	DDAH2	
	IL17D	
	VWF	
	APP	
	TERT	
	NTF3	
	BSG	
	DDAH2	
	MICA	
	BGLAP	
	TNF	
	MMP14	
	SLC7A7	
	SLC7A2	
	TNF	
	BTG3	
	IDUA	
	TNF	
	TTR	
	DSG3	
	DSG1	
	TNF	
	LTA	
	DDR1	
	TNF	
	LTA	
	PDE6B	
	CTSB	
	DDR1	
	RNASE2	
	RNASE3	
	DDR1	
	NPC1	
	APEX1	
	MIB1	
	DEFB1	
	DDR1	
	HLA-E	
	TAP1	
	BTNL2	
	HLA-E	
	NDC80	
	TYMS	
	USP9Y	
	AC217779.1	

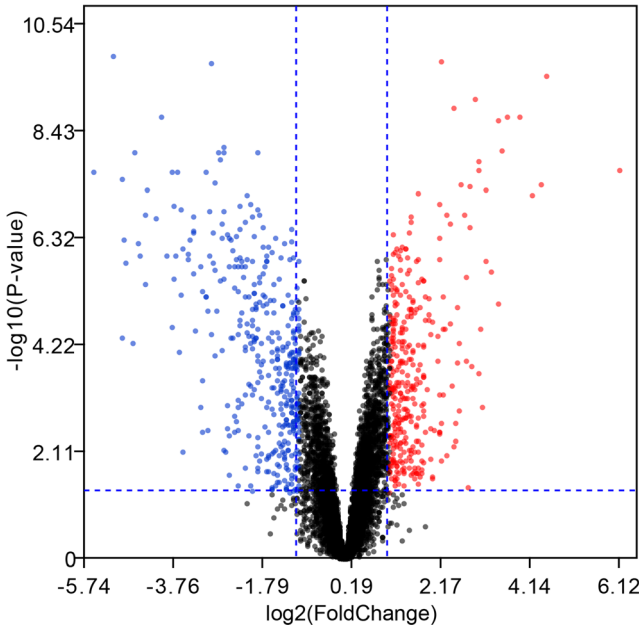

Fig. 2 Volcano plot analysis of gene expression changes between fibroblasts cultured from normal and fibrotic human lung tissues in GSE40839. Red circles represent upregulated genes that passed the statistical and fold-change cutoffs in fibrotic human lung tissues, while blue circles represent downregulated genes that meeting the above-mentioned requirements

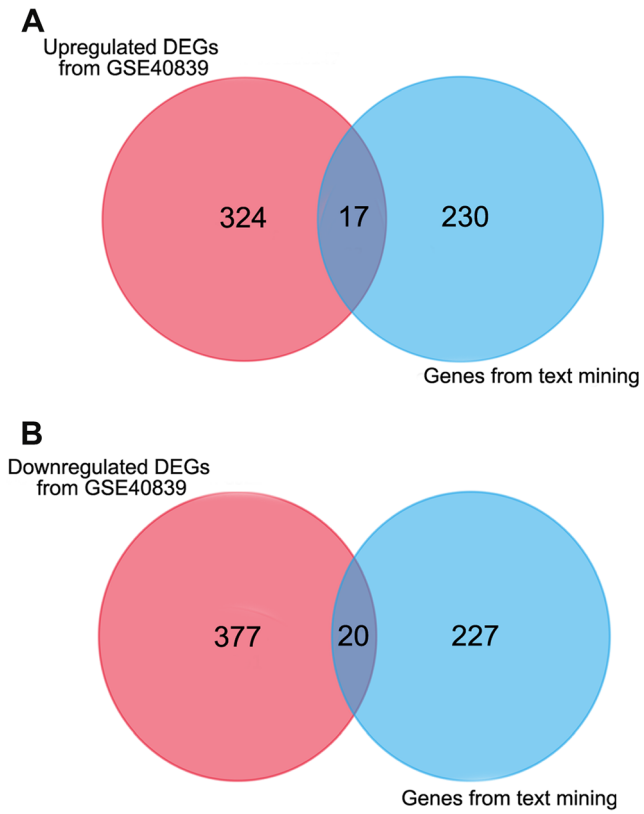

Fig. 3 Venn diagram of DEGs from GSE40839 and genes from text mining. **a** Venn diagram showing common gene expression between upregulated DEGs and genes from text mining (17 genes). **b** Venn diagram of 20 common genes between downregulated DEGs and genes from text mining

PPI network modeling

The STRING database is web-based tool that can identify the interactions between known proteins and predicted proteins with assigned confidence scores by uploading specific DEGs list [15]. The interaction map revealed deeper insights of protein interactions among corresponding proteins of DEGs. After filtering out disconnected genes, a PPI network containing 199 interactions for a total of 35 DEGs was generated by STRING, including 16 upregulated and 19 downregulated genes. Next, we imported the data into Cytoscape and the PPI network was visualized as shown in Fig. 6. The essential genes were selected based on the analysis of degree centrality, subgraph centrality, betweenness centrality and closeness centrality using CytoNCA. VEGFA scored highest in all four analysis methods, while four genes EGFR, TGFB1, IGF1 and CCL2 were the top-ranked genes in three methods (degree, betweenness centrality and closeness centrality). In addition to these five genes, another five with less scores were also selected as hub nodes in the PPI network (ICAM1, CXCL12, SERPINE1, NGF, VCAM1) as shown in S. Table.4.

MCODE algorithm was used to detect subnets of the PPI network that are likely to represent molecular complexes. There were four highly connected clusters being identified as shown in Fig. 7 and S. Table.5. After validating clusters using ClueGo plugin of Cytoscape, the most significant GO-BP category associated with cluster one was: "cellular response to dexamethasone stimulus" including three genes (CCL2, EGFR and ICAM1), with a corrected term *P* value of 1.48E-06 as shown in S. Table. 6. Pathway enrichment analysis conducted by screening WikiPathways, REACTOME and KEGG database. Notably, three genes were enriched in the category of lung fibrosis: CCL2, CTGF and IGF1 with a corrected term *P* value of 5.94E-05, which was one of the main pathological features of ILD [22] (S.Table. 7).

Discussion

ILD can be induced by a wide variety of drugs and demonstrate various types of pulmonary complications [23]. Although not too frequent in comparison with the most common adverse events of EGFR-TKI treatment, such as skin rash and diarrhea, patients with ILD have a worse prognosis and an increased risk of death. The mortality of ILD events due to anti-EGFR-therapy was significantly higher than that of traditional chemotherapies or best supportive care (22.8 and 7.1%, respectively) [24]. Previous study demonstrated that the imbalance of destruction and repair is fundamental in ILD events as acute phase proteins in the plasma of patients were significantly altered upon EGFR-TKI treatment [25]. In addition, ILD is regarded as a comorbid disease of lung cancer, being observed in patients at the time of diagnosis, and conversely, lung cancer is more common in

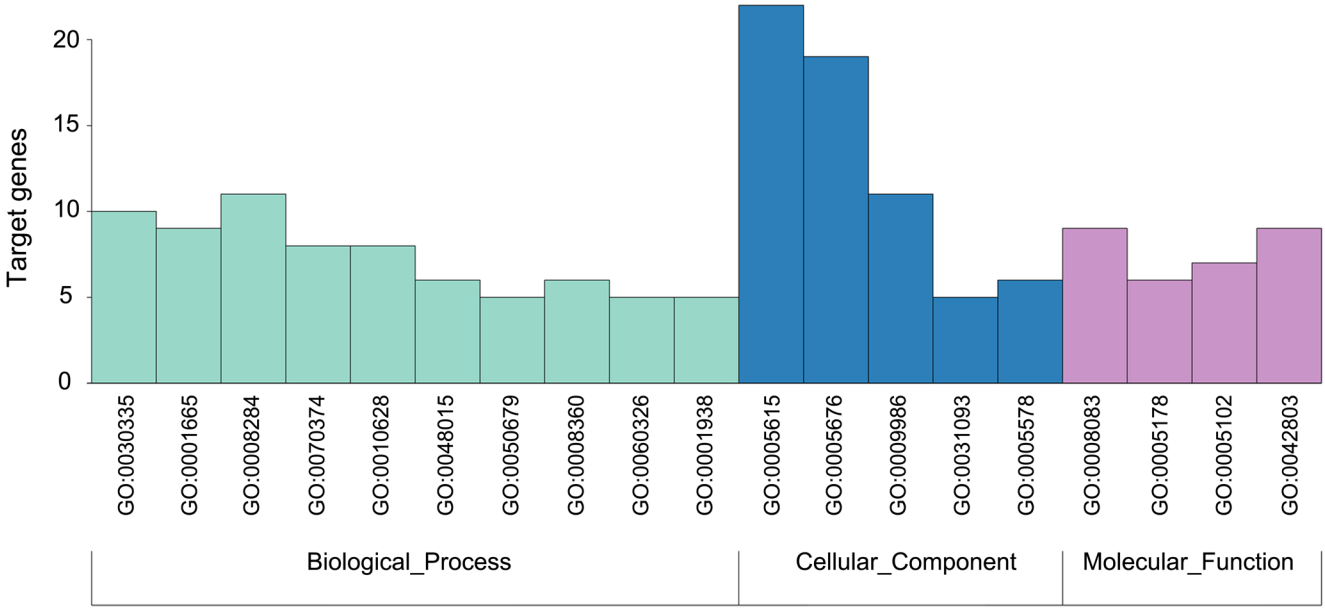

Fig. 4 Go analysis of common genes associated with ILD and EGFR-TKI

patients with various types of ILD [26]. ILD and lung cancer share several common pathological features: firstly, the continuously proliferated myofibroblasts and fibroblasts in the chronic injury regions are resistant to apoptosis, which is similar with characters of cancer cells [27]; secondly, epithelial mesenchymal transition (EMT) as an etiologic factor in fibrosis can lead to the increasing produce of reactive oxygen species (ROS), which promotes tumorigenesis and chemoresistance [28]; the similar patterns of DNA hypomethylation and telomerase deficiency have been implicated in the common pathogenesis of both diseases [29]. However, the specific mechanisms underlying the pathogenesis of EGFR-TKI induced ILD are still poorly understood. Besides EGFR-TKI treatment, studies revealed several other risk factors contribute less to this adverse event, including male sex, ethnicity, smoking history, older than 40 years, poor performance status, radiation, drug interaction and the presence of pulmonary fibrosis [30, 31].

In the present study, we connected the two concepts of "EGFR tyrosine kinase inhibitor" and "ILD" using online text mining tool and identified 247 genes might involve in this disease. To ensure that our analysis found the most relevant genes, we further took a sub-intersection between genes from text mining and mRNA expression dataset using gene microarray analysis. Finally, we constructed a PPI network of these DEGs and classified the nodes by GO annotation and KEGG enrichment analysis. Nineteen enriched GO terms revealed fundamental pathophysiological alterations of ILD, including fibroblasts proliferation, migration and inflammation [32]. Furthermore, eight genes identified by KEGG analysis were common members of top three significantly enriched pathway terms, including Rap1 pathway, Ras pathway and PI3K/Akt pathway. Rap1 is a small G protein belong to the Ras-like family of GTPases, which can be activated by cAMP-GEF-I (Epac1), a cAMP effector. Previous study demonstrated that the activation of Rap1 inhibits fibroblast

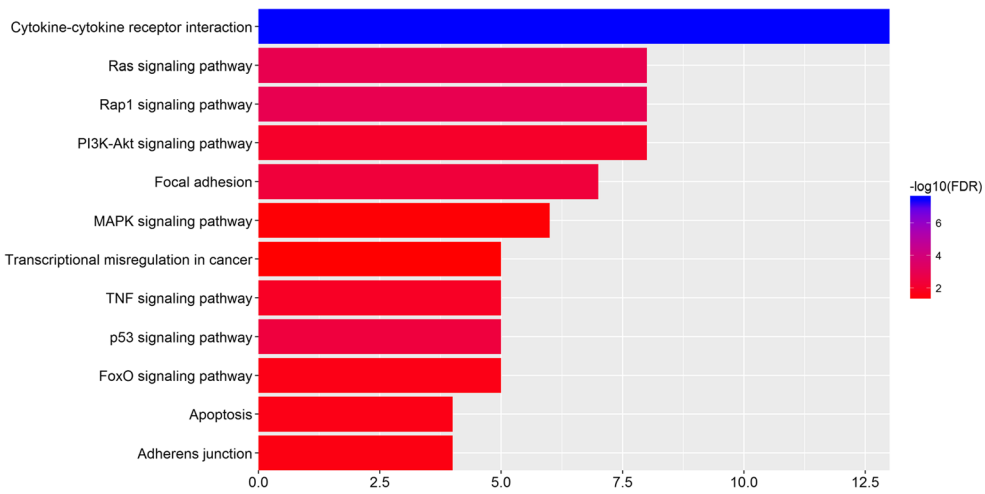

Fig. 5 KEGG analysis of common genes from ILD and EGFR-TKI

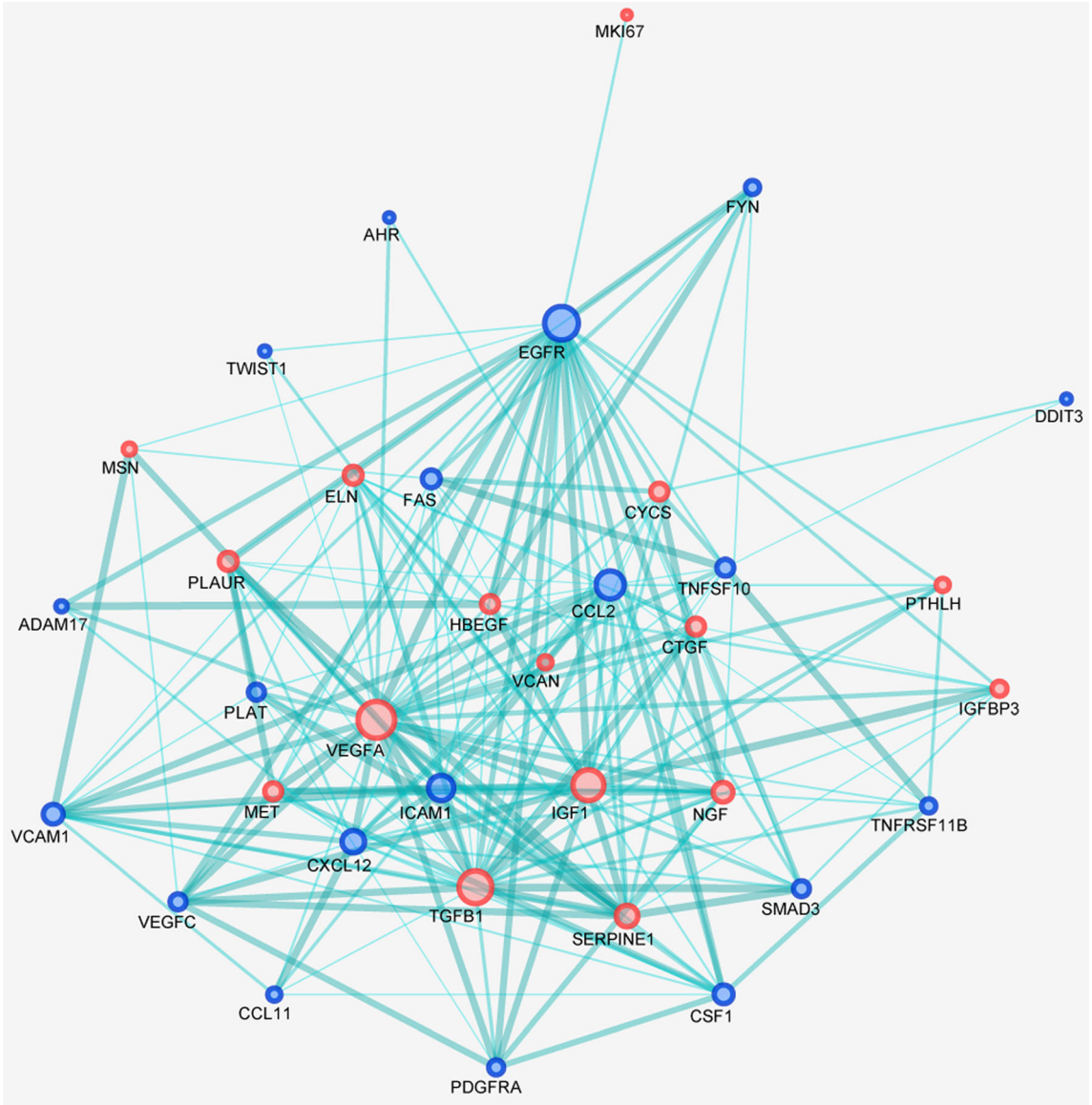

Fig. 6 PPI network of DEGs visualized by STRING

proliferation, which was mediated by the direct interaction of microtubule with Epac1 [33]. However, Rap1 has been suggested might play a dual function of controlling Ras/mitogen-activated protein kinase (MAPK) signaling by binding to different downstream effectors. Study showed that B-raf, a Rap1-activating protein kinase, was highly expressed in lung homogenates of ILD patients compared with normal controls [34]. Moreover, in vitro data supported the pro-fibrosis property of Rap1 in ILD that tryptase produced by mast cell can stimulate the proliferation of human lung fibrosis in a Raf1/external signal-regulated kinase (ERK) signaling depended manner [33]. Ras belongs to the same superfamily of GTPase with Rap1, and shared downstream signaling molecules of them, such as MAPK/ERK kinase (MEK) and ERK, had been identified [35]. The activation of Ras results in significant loss of endothelial-specific markers in endothelial cells, which empower lung capillary endothelial cells to be an extra source of fibroblasts [36]. Importantly, Ras proteins are potent activators of PI3K subunits, while TGFB1-induced transformation of fibroblasts into myofibroblasts requires the activation of PI3K/AKT pathway [37, 38].

A set of 10 high ranking genes were identified using CytoNCA based on the importance connectivity of them in the whole network we built. Most of them have established well-known association with the etiology and pathology of ILD. TGFB1 has significant profibrotic activity both in vitro and in vivo [37]. Upon binding to the cell-surface type II TGFB receptor, TGFB1 could trigger and coordinate the recruitment and activities of type I receptor, and subsequently phosphorylate Smad proteins to promote expression of fibrogenic intermediate effectors [39]. IGF1 acts as a vital survival factor for lung fibroblasts under adverse conditions by inducing the transcription of a bunch of pro-proliferative and antiapoptotic genes, and thereby

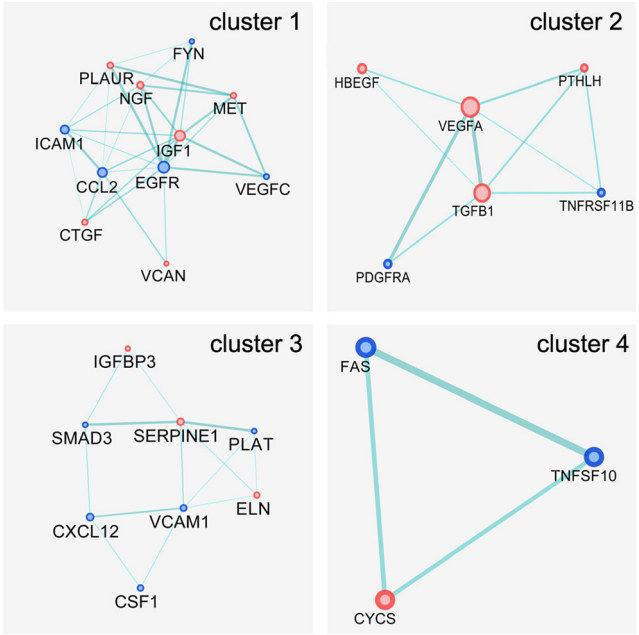

Fig. 7 Four highly connected clusters identified by MCODE algorithm and visualized by ClueGo plugin of Cytoscape. Interactions are color coded according to combined scores with darker edges corresponding to higher scores

stimulate the formation of myofibroblasts and collagen deposition in a PI3K-dependent way [40]. CCL2 (also known as monocyte chemoattractant protein-1/MCP1) is one of the key chemokines that regulate accumulation, migration and infiltration of inflammatory cells in responses to tissue injury [41]. More importantly, highly increased CCL2 levels in bronchoalveolar lavage fluid (BALF) and serum from patients with ILD was significantly correlate with disease severity and prognosis [42]. CXCL12 is the main chemokines produced by aberrant alveolar epithelial cells in ILD, which likely expanse the population of fibroblast/myofibroblast by attracting circulatory fibrocytes through the CXCR4/CXCL12 axis [43]. SERPINE1, also known as plasminogen activator inhibitor 1 (PAI-1), is a target gene of TGFB1 and highly expressed in type II alveolar cells. Previous in vivo study has shown that SERPINE1 regulates the fibrin accumulation after lung damage and cell senescence induced by bleomycin [44].

Although the relationship with ILD remains unclear, besides its functions in airway neurons of enhancing cough reflex, NGF exerts profibrotic properties and pro-remodeling activities in lung fibroblasts from mouse chronic asthma model by inducing collagen production though TGFB1/Smad pathway [45, 46]. ICAM-1, VCAM-1 and VEGFA were also speculated to be associated with ILD, and were considered to be potent predictors of survival in ILD and thus helpful in disease monitoring [47, 48]. ICAM-1 and VCAM-1 were two inflammatory and adhesion molecules that play critical roles in leukocyte kinetics during collagen deposition and fibrosis, while VEGFA (also known as VEGF) is a key regulator for angiogenesis that is assumed to be responsible for aberrant angiogenesis in ILD. However, discrepant results were also reported in studies that there was no statistically significant difference in serum levels of ICAM-1 and VCAM-1 and BALF levels of VEGFA between patients and healthy volunteers [49, 50]. Further in vitro and in vivo investigations are needed to define their pathophysiological roles in ILD.

MCODE and ClueGo were used as discovery tools to further analyze the network and extract hub genes of EGFR-TKI induced ILD, which might be helpful to identify new potential drug targets. The former analyzes network based on topological structure, while the latter provides an estimate of functions based on gene annotations. MCODE detected four densely connected clusters, which were then uploaded in ClueGo, and only cluster one genes retrieved potential functional annotations. Cluster one includes 11 nodes and 30 edges, and CTGF was introduced as seed protein. CTGF, as a downstream effector of TGFB1 and a secreted matricellular protein that modulate cellular functions and signaling depend on the cell type and context without unique receptor to bind, is regarded as playing central role in tissue remodeling and fibrosis [51]. Clinical study shown that plasma CTGF levels were significantly higher in ILD patients and negatively correlated with pulmonary function [52]. Furthermore, five genes from cluster one including CTGF, CCL2, IGF1, EGFR and ICAM1 were associated with either response to dexamethasone or with lung fibrosis as represented in S. Table. 5 and S. Table. 6, and notably, the latter four also ranked top 10 highest-score genes in the topological analysis. EGFR plays key roles in integrating the signaling events leading to cell growth, differentiation. Infant with inherited loss-of-function mutation of EGFR showed lifelong respiratory distress due to airway epithelial inflammation [53]. Although, in vitro study demonstrated that the EGFR ligand amphiregulin and an intact EGFR/MAPK/ERK signaling pathway are required for TGF-β1 dependent pulmonary fibrosis to enhance the proliferation of ILD fibroblasts, blocking the signaling of EGFR and its family members with targeted drugs exacerbate lung injury [54, 55]. Interestingly, EGFR mutation-selective inhibitor osimertinib, designed to irreversibly inhibit the phosphorylation of EGFR with T790 M drug-resistant mutation and less potently inhibits wild type EGFR, thereby minimizes adverse effects, was also associated with the development of ILD in patients who had no pulmonary toxicities during a prior treatment with first- or second-generation EGFR-TKI, which highlights the complexity underlying EGFR-induced ILD [56].

In our study, the expression of EGFR, TGFB1, IGF1, SERPINE1, VEGFA and CTGF were upregulated, which were concordant with their roles in ILD reasoned by prior studies. Otherwise, the expression of CCL2, ICAM1 and VCAM1 were downregulated in this dataset, which need

further basic and clinical research to provide solid evidence, especially in the context of EGFR-TKI treatment.

Conclusions

ILD is the most severe form of EGFR-TKI treatment-related adverse effects for patients with NSCLC. To date, very few studies have been published in this field and the pathogenetic pathways and mediators of this disease are poorly understood. In the present study, we started with systematically analysis of associated genes using text mining, and then we identified essential genes and pathways by data mining and functional annotation. We underline a central role of some genes such as EGFR and TGFB1, and specific pathways such as Rap1 pathway in the pathogenesis of EGFR-TKI induced ILD, which may have diagnostic and therapeutic value in the future.

Author contributions Yuan Lu, Ang Li, Hu Zhao and Xinling Ren conceived and designed the study. Yuan Lu and Ang Li were responsible for the collection, analysis and interpretation of the data. Yuan Lu, Wen Zhao and Ping Tang were responsible for data visualization. Yuan Lu, Ang Li, XiaoFeng Lai drafted the manuscript, and all authors critically revised it for important intellectual content. Xinling Ren obtained funding, and Hu Zhao was responsible for logistic support. All authors had final approval of the submitted manuscript.

Compliance with ethical standards

Funding This study was funded by National Natural Science Foundation of China (No. 81871880), and Director Fund of Shenzhen University General Hospital (No. 0000040546).

Conflict of interest All authors have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Chen W, Zheng R, Zeng H, Zhang S (2015) Epidemiology of lung cancer in China. Thorac cancer 6(2):209–215
2. Ettinger DS, Wood DE, Aisner DL, Akerley W, Bauman J, Chirieac LR, D'Amico TA, Decamp MM, Dilling TJ, Dobelbower M (2017) Non-small cell lung Cancer, version 5.2017, NCCN clinical practice guidelines in oncology. J Natl Compr Cancer Netw 15(4):504–535
3. Kimura K, Takayanagi R, Fukushima T, Yamada Y (2017) Theoretical method for evaluation of therapeutic effects and adverse effects of epidermal growth factor receptor tyrosine kinase inhibitors in clinical treatment. Med Oncol 34(10):178
4. Bagnato G, Harari S (2015) Cellular interactions in the pathogenesis of interstitial lung diseases. European respiratory review : an official journal of the European Respiratory Society 24(135):102–114
5. Ramos-Casals M, Perez-Alvarez R, Perez-de-Lis M, Xaubet A, Bosch X (2011) Pulmonary disorders induced by monoclonal antibodies in patients with rheumatologic autoimmune diseases. Am J Med 124(5):386–394
6. Peerzada MM, Spiro TP, Daw HA (2010) Pulmonary toxicities of biologics: a review. Anti-Cancer Drugs 21(2):131–139
7. Davies EC, Green CF, Taylor S, Williamson PR, Mottram DR, Pirmohamed M (2009) Adverse drug reactions in hospital in-patients: a prospective analysis of 3695 patient-episodes. PLoS One 4(2):e4439
8. Bast A, Weseler AR, Haenen GR, den Hartog GJ (2010) Oxidative stress and antioxidants in interstitial lung disease. Curr Opin Pulm Med 16(5):516–520
9. de Boer WI, Hau CM, van Schadewijk A, Stolk J, van Krieken JHJM, Hiemstra PS (2006) Expression of epidermal growth factors and their receptors in the bronchial epithelium of subjects with chronic obstructive pulmonary disease. Am J Clin Pathol 125(2): 184–192
10. Takeda M, Okamoto I, Nakagawa K (2015) Pooled safety analysis of EGFR-TKI treatment for EGFR mutation-positive non-small cell lung cancer. Lung Cancer 88(1):74–79
11. Sakuma K, Nakamura H, Nakamura T, Hoshino Y, Ueda S, Ichikawa M, Tabata C, Fujita S, Masago K, Yodoi J et al (2007) Elevation of serum Thioredoxin in patients with Gefitinib-induced interstitial lung disease. Intern Med 46(23):1905–1909
12. Inzalkar S, Sharma J (2015) A survey on text mining- techniques and application. Int J Eng Sci 24:1–14
13. Baran J, Gerner M, Haeussler M, Nenadic G, Bergman CM (2011) pubmed2ensembl: a resource for mining the biological literature on genes. PLoS One 6(9):e24716
14. Lindahl GE, Stock CJ, Xu SW, Leoni P, Sestini P, Howat SL, Bou-Gharios G, Nicholson AG, Denton CP, Grutters JC (2013) Microarray profiling reveals suppressed interferon stimulated gene program in fibroblasts from scleroderma-associated interstitial lung disease. Respir Res 14(1):1–14
15. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P (2017) The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res 45(Database issue):D362–D368
16. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
17. Tang Y, Li M, Wang J, Pan Y (2015) Wu F-X: CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. Biosystems 127:67–72
18. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, Pico AR, Bader GD, Ideker T (2012) A travel guide to Cytoscape plugins. Nat Methods 9(11):1069–1076
19. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J (2009) ClueGO a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25(8):1091–1093
20. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42(Database issue):472–477
21. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44
22. Casoni GL, Tomassetti S, Cavazza A, Colby TV, Dubini A, Ryu JH, Carretta E, Tantalocco P, Piciucchi S, Ravaglia C (2014) Transbronchial lung cryobiopsy in the diagnosis of fibrotic interstitial lung diseases. PLoS One 9(2):e86716
23. Schwaiblmair M, Behr W, Haeckel T, Markl B, Foerg W, Berghaus T (2012) Drug induced interstitial lung disease. Open Respir Med J 6:63–74

24. Ashiq U, Jamal RA, Mesaik MA, Mahroof-Tahir M, Shahid S, Khan KM (2014) Synthesis, immunomodulation and cytotoxic effects of vanadium (IV) complexes. Med Chem 10(3):287–299
25. F N, A O, G HC (2011) K N: proteomic biomarkers for acute interstitial lung disease in gefitinib-treated Japanese lung cancer patients. PLoS One 6(7):e22062
26. Tsuboi M, Le Chevalier T (2006) Interstitial lung disease in patients with non-small-cell lung cancer treated with epidermal growth factor receptor inhibitors. Med Oncol 23(2):161–170
27. Drakopanagiotakis F, Xifteri A, Polychronopoulos V, Bouros D (2008) Apoptosis in lung injury and fibrosis. Eur Respir J 32(6):1631
28. Costa A, Scholer-Dahirel A, Mechta-Grigoriou F (2014) The role of reactive oxygen species and metabolism on cancer cells and their microenvironment. Semin Cancer Biol 25:23–32
29. Archontogeorgis K, Steiropoulos P, Tzouvelekis A, Nena E, Bouros D (2012) Lung cancer and interstitial lung diseases: a systematic review. Pulm Med 2012(315918):1–11
30. Ando M, Okamoto I, Yamamoto N, Takeda K, Tamura K, Seto T, Ariyoshi Y, Fukuoka M (2006) Predictive factors for interstitial lung disease, antitumor response, and survival in non-small-cell lung cancer patients treated with gefitinib. J Clin Oncol: Off J Am Soc Clin Oncol 24(16):2549–2556
31. Akamatsu H, Inoue A, Mitsudomi T, Kobayashi K, Nakagawa K, Mori K, Nukiwa T, Nakanishi Y, Yamamoto N (2013) Interstitial lung disease associated with gefitinib in Japanese patients with EGFR-mutated non-small-cell lung cancer: combined analysis of two phase III trials (NEJ 002 and WJTOG 3405). Jpn J Clin Oncol 43(6):664–668
32. Fischer A, West SG, Swigris JJ, Brown KK, Bois RMD (2013) Connective-tissue disease-associated InterstitialLung disease. J Intensive Care Med 84(4):498
33. Huang SK, Wettlaufer SH, Chung J, Peters-Golden M (2008) Prostaglandin E2 inhibits specific lung fibroblast functions via selective actions of PKA and Epac-1. Am J Respir Cell Mol Biol 39(4):482–489
34. Yoshida K, Kuwano K, Hagimoto N, Watanabe K, Matsuba T, Fujita M, Inoshima I, Hara N (2002) MAP kinase activation and apoptosis in lung tissues from patients with idiopathic pulmonary fibrosis. J Pathol 198(3):388–396
35. Goitre L, Trapani E, Trabalzini L, Retta SF (2014) The Ras Superfamily of Small GTPases: The Unlocked Secrets. Ras Signaling 1120:1–18
36. Hashimoto N, Phan SH, Imaizumi K, Matsuo M, Nakashima H, Kawabe T, Shimokata K, Hasegawa Y (2010) Endothelial-mesenchymal transition in bleomycin-induced pulmonary fibrosis. Am J Respir Cell Mol Biol 43(2):161–172
37. Fernandez IE, Eickelberg O (2012) The impact of TGF-β on lung fibrosis: from targeting to biomarkers. Proc Am Thorac Soc 9(3): 111–116
38. Kulkarni AA, Thatcher TH, Olsen KC, Maggirwar SB, Phipps RP, Sime PJ (2011) PPAR-γ ligands repress TGFβ-induced myofibroblast differentiation by targeting the PI3K/Akt pathway: implications for therapy of fibrosis. PloS One 6(1):e15909
39. Andrianifahanana M, Wilkes MC, Gupta SK, Rahimi RA, Repellin CE, Edens M, Wittenberger J, Yin X, Maidl E, Becker J, Leof EB (2013) Profibrotic TGFbeta responses require the cooperative action of PDGF and ErbB receptor tyrosine kinases. FASEB journal 27(11):4444–4454
40. Hung CF, Rohani MG, Lee SS, Chen P, Schnapp LM (2013) Role of IGF-1 pathway in lung fibroblast activation. Resp Res 14:102
41. Deshmane SL, Kremlev S, Amini S, Sawaya BE (2009) Monocyte chemoattractant protein-1 (MCP-1): an overview. J Interf Cytok Res 29(6):313–326
42. Assassi S, Wu M, Tan FK, Chang J, Graham TA, Furst DE, Khanna D, Charles J, Ferguson EC, Feghali-Bostwick C et al (2013) Skin gene expression correlates of severity of interstitial lung disease in systemic sclerosis. Arthritis Rheum 65(11):2917–2927
43. Anderssonsjöland A, de Alba CG, Nihlberg K, Becerril C, Ramírez R, Pardo A, Westergrenthorsson G, Selman M (2008) Fibrocytes are a potential source of lung fibroblasts in idiopathic pulmonary fibrosis. Int J Biochem Cell Biol 40(10):2129
44. Jiang C, Liu G, Luckhardt T, Antony V, Zhou Y, Carter AB, Thannickal VJ, Liu RM (2017) Serpine 1 induces alveolar type II cell senescence through activating p53-p21-Rb pathway in fibrotic lung disease. Aging Cell 16(5):1114–1124
45. Harrison NK (2013) Cough, sarcoidosis and idiopathic pulmonary fibrosis: raw nerves and bad vibrations. Cough 9(1):9
46. Kilic A, Sonar SS, Yildirim AO, Fehrenbach H, Nockher WA, Renz H (2011) Nerve growth factor induces type III collagen production in chronic allergic airway inflammation. J Allergy Clin Immunol 128(5):1058–1066 e1051–1054
47. Richards TJ, Kaminski N, Baribaud F, Flavin S, Brodmerkel C, Horowitz D, Li K, Choi J, Vuga LJ, Lindell KO (2012) Peripheral blood proteins predict mortality in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 185(1):67–76
48. Ando M, Miyazaki E, Ito T, Hiroshige S, Nureki SI, Ueno T, Takenaka R, Fukami T, Kumamoto T (2010) Significance of serum vascular endothelial growth factor level in patients with idiopathic pulmonary fibrosis. Lung 188(3):247–252
49. Kennedy B, Branagan P, Moloney F, Haroon M, O'Connell OJ, O'Connor TM, O'Regan K, Harney S, Henry MT (2015) Biomarkers to identify ILD and predict lung function decline in scleroderma lung disease or idiopathic pulmonary fibrosis. Sarcoidosis Vasculitis & Diffuse Lung Diseases Official Journal of Wasog 32(3):228
50. Yamashita M, Mouri T, Niisato M, Nitanai H, Kobayashi H, Ogasawara M, Endo R, Konishi K, Sugai T, Sawai T (2015) Lymphangiogenic factors are associated with the severity of hypersensitivity pneumonitis. Bmj Open Respiratory Research 2(1): e000085
51. Lipson KE, Wong C, Teng Y, Spong S (2012) CTGF is a central mediator of tissue remodeling and fibrosis and its inhibition can reverse the process of fibrosis. Fibrogenesis Tissue Repair 20125(Suppl 1):S24
52. Kono M, Nakamura Y, Suda T, Kato M, Kaida Y, Dai H, Inui N, Hamada E, Miyazaki O, Kurashita S (2011) Plasma CCN2 (connective tissue growth factor; CTGF) is a potential biomarker in idiopathic pulmonary fibrosis (IPF). Clin Chim Acta 412(23–24): 2211–2215
53. Campbell P, Morton PE, Takeichi T, Salam A, Roberts N, Proudfoot LE, Mellerio JE, Aminu K, Wellington C, Patil SN et al (2014) Epithelial inflammation resulting from an inherited loss-of-function mutation in EGFR. J Investig Dermatol 134(10):2570–2578
54. Andrianifahanana M, Wilkes MC, Gupta SK, Rahimi RA, Repellin CE, Edens M, Wittenberger J, Yin X, Maidl E, Becker J (2013) Profibrotic TGFβ responses require the cooperative action of PDGF and ErbB receptor tyrosine kinases. Faseb Journal Official Publication of the Federation of American Societies for Experimental Biology 27(11):4444–4454
55. Harada C, Kawaguchi T, Ogatasuetsugu S, Yamada M, Hamada N, Maeyama T, Souzaki R, Tajiri T, Taguchi T, Kuwano K (2011) EGFR tyrosine kinase inhibition worsens acute lung injury in mice with repairing airway epithelium. Am J Respir Crit Care Med 183(6):743–751
56. Matsumoto Y, Kawaguchi T, Yamamoto N, Sawa K, Yoshimoto N, Suzumura T, Watanabe T, Mitsuoka S, Asai K, Kimura T (2017) Interstitial lung disease induced by Osimertinib for epidermal growth factor receptor (EGFR) T790M-positive non-small cell lung Cancer. Intern Med 56(17):2325–2328