RESEARCH ARTICLE

Editorial Process: Submission:06/17/2022 Acceptance:10/24/2022

SNP rs9387478 at ROS1-DCBLD1 Locus is Significantly Associated with Lung Cancer Risk and Poor Survival in Indian Population

Jonita Chongtham[1], Namita Pandey[1,2], Lokesh Kumar Sharma[3], Anant Mohan[4], Tapasya Srivastava[1]*

Abstract

Objective: Receptor tyrosine kinases (RTK) are relevant therapeutic targets in the treatment of lung cancer. Germline susceptibility variants that influence these RTKs may provide new insights into their regulation. rs9387478 is located in the genomic interval between two RTK-genes ROS1/DCBLD1, of which ROS1 alterations are implicated in lung carcinogenesis and treatment response while the latter remains poorly understood. **Materials and methods:** Venous blood was drawn from 100 control and 231 case subjects. Genotype was scored by restriction fragment length polymorphism (RFLP), PCR amplification followed by HindIII digestion. Logistic regression was applied to compare the association between variables. Survival curve was plotted to draw a correlation between the genotype and overall survival. Also, eQTL and chromatin state changes were analyzed and correlated with the survival of patients using available datasets. **Results:** In our population smoking correlated significantly with lung cancer [OR= 2.607] with the presence of the minor allele 'A' enhancing the nicotine dependence [CA (OR=3.23)]. Individuals with homozygous risk allele 'A' had a higher chance of developing lung cancer [OR=2.65] than individuals with CA/CC implying a recessive model of association. Patients with CC/CA genotype had better overall survival than patients with AA genotype [161 days/142 days vs 54 days, p=0.005]. The homozygous risk allele was significantly associated with increased DCBLD1 and ROS1 expression in lung cancer, with enriched active histone marks due to the polymorphism. Interestingly, increased DCBLD1 expression was associated with poor outcomes in lung cancer. **Conclusion:** Overall, our study provides strong evidence that rs9387478 is significantly associated with both nicotine dependence and lung cancer in our North Indian cohort. The association of the SNP with prognostic genes, DCBLD1 and ROS1 make rs9387478 a promising prognostic marker in the North Indian population. The results obtained are significant, however, the study needs to be performed in a larger sample size.

Keywords: Lung cancer- smoking- overall survival- rs9387478- ROS1- DCBLD1

Asian Pac J Cancer Prev, **23 (10)**, 3553-3561

Introduction

Lung cancer which accounts for 11.4% (one in 10 cancers diagnosed) of the total cancer incidence and 18% (one in 5 deaths) of the total cancer mortality is one of the most common malignancies in adults (Sung et al., 2021). According to GLOBOCAN 2021, India reported 1,324,413 new lung cancer cases and 851,678 mortalities (Sung et al., 2021). While a decline in lung cancer is seen globally (Islam et al., 2015), the incidence in India is increasing with most cases exhibiting an advanced disease state at the time of diagnosis explaining the high lung cancer morbidity in India (Mohan et al., 2020). The decrease in global lung cancer incidence can be attributed to the decline in overall tobacco use (WHO Global report) the highest risk factor for lung cancers. However, 14.8% of adults in Indian population use tobacco in different forms (WHO 2020), which along with other environmental factors such as pollution and lifestyle contribute to the growing burden of lung cancer in India. Various high-throughput genome-wide association studies (GWAS) of SNPs and candidate genes have powered our understanding of the genetic component of lung cancer (Stadler et al., 2010; Wang et al, 2017). Additionally, expression quantitative trait loci (eQTL) studies have shown the influence of SNP on gene expression and

*[1]Department of Genetics, University of Delhi South Campus, New Delhi, India. [2]Current affiliation: Clinical Genomic Knowledgebase, PerianDx, Pune, Maharashtra, India. [3]Ram Manohar Lohia Hospital, New Delhi, India. [4]Department of Pulmonary, Critical Care and Sleep Medicine, All India Institute of Medical Sciences (AIIMS), New Delhi, India. *For Correspondence: tapasya@south.du.ac.in*

disease risk. Chromosome 15q25.1 locus (nAChR encoding gene cluster) (Thorgeirsson et al., 2008; Pandey et al. 2017) has emerged as a locus strongly associated with nicotine dependence and lung cancer in these association studies. These genetic variations have a substantial contribution to smoking initiation, nicotine dependence, and cessation as well as influence the expression of genes in proximity (Quach et al., 2020). Other SNPs such as intronic DNMT3B rs910083 (Hancock et al., 2017), rs10865246 NRXN1 (Pérez-Rubio et al., 2016) and rs1137115, rs1801272 and rs28399433 rs4105144 in CYP2A6 (Lopez Flores et al., 2017) have also emerged as prominent risk markers in various studies. Predictive genetic markers for lung cancer susceptibility and nicotine dependence can benefit individuals to ascertain potential risks and make significant changes to improve addiction, withdrawal severity associated response to treatment, and other health-related consequences (Gu et al., 2020).

The objective of this study was to explore the effect of a genetic variant rs93874478 at chromosome 6q22.2 which was identified as a lung cancer susceptibility locus in the Asian women population (Lan et al., 2012). rs9387478 is positioned at the genetic interval between two receptor tyrosine kinases- DCBLD1 (Discoidin, CUB and LCCL containing Protein 1) and ROS1 (ROS proto-oncogene receptor kinase 1), two very important plausible candidate genes in primary lung adenocarcinogenesis (Lan et al., 2012). ROS1, a tyrosine kinase insulin receptor gene is often mutated and transformed in many cancers including ovarian cancer, brain tumors, and lung cancer (Jones et al., 2013; Cilloni et al., 2013; Acquaviva et al., 2019).. Genetic rearrangement of ROS1 (ROS1 fusion) presumably leading to constitutive ROS1 expression was identified as a distinct molecular signature of NSCLC found in 2.54% of the patients with lung adenocarcinoma (Uguen and Braekeleer, 2016) and a relatively higher frequency of 2.8% in north Indian patients (Shukla et al., 2019; Mehta et al., 2020). ROS1 dephosphorylation occurs through direct binding with the SH2 domain of PTPN6 (Protein Tyrosine Phosphatase Non-Receptor Type6) and interacts with PTPN11 and activates the PI3K/mTOR signaling cascade (Charest et al., 2006). Therefore, ROS1 is a validated druggable target in NSCLC (Lin and Shaw, 2017). Effective targeted therapies for ROS1-positive lung cancer include Crizotinib, Lorlatinib, Entrectinib, and Ceritinib of which Crizotinib and Entrectinib, have received Food and Drug Administration (FDA) approval (Sehgal et al., 2020). Patients with ROS1 mutation respond better to chemotherapy compared to other driver mutations such as EGFR, KRAS, and ALK with an estimated 60% objective response rate (ORR), 89.5% disease control rate (DCR), and a 7-month progression-free survival (PFS) (D'Angelo et al., 2020). However, the majority of patients treated with Crizotinib develop resistance within a few years of treatment and advance to secondary mutations in the ROS1 tyrosine kinase domain (Gainor et al., 2017). The other gene, DCBLD1 (Discoidin, CUB, and LCCL Containing protein 1) in the locus is linked with higher risks of never-smoking HPV negative head and neck cancer, lung cancer, and lung adenocarcinoma in females (Yoo et al., 2017). Cardin et al also performed a retrospective study of the TCGA cohort of several cancer types and reported that DCBLD1 is associated with an augmented integrin signaling pathway affecting focal adhesion and consequently cell migration (Cardin et al., 2021). It was concluded that increased expression of DCBLD1 leads to poor overall survival in NSCLC and invasive breast carcinoma (Cardin et al., 2021). Although the DCBLD1 remains largely understudied, a related highly conserved protein DCBLD2 or CLCP1 is widely studied and known to enhance cell proliferation and invasion (Mirang et al., 2008) and significantly increase lung cancer progression and metastasis (Koshikawa et al., 2002). Furthermore, xenograft studies using A549 cells showed that knockdown of DCBLD1 showed a reduction in tumor growth (Wang et al., 2020). These preliminary studies signifying its prognostic role indicate that DCBLD1 is a potential therapeutic target.

rs9387478 has not been previously investigated as a risk factor for lung cancer in the Indian population although there are studies in other Asian populations. Given reported differences in ethnicity within the Asian population, it would be to the benefit of the large Indian population that independent association studies are undertaken. Our study, therefore, aims to find the previously unexplored association of rs9387478 with lung cancer, nicotine dependence, and overall survival in North Indian lung cancer patients.

Materials and Methods

Patient data and sample collection, follow up and relevant clinical characteristics

We performed a case-control study with 331 individuals. 100 patients diagnosed with Lung cancer at the Outpatient Department clinic and ward at the Department of Pulmonary Medicine and Sleep Disorders, All India Institute of Medical Sciences (AIIMS), New Delhi were recruited. Disease diagnosis was carried out in AIIMS using imaging and histopathological analysis of appropriate specimens. Age-matched 231 subjects (control) were also recruited at AIIMS and Ram Manohar Lohia (RML) Hospital, New Delhi admitted for other ailments, following strict exclusion criteria of absence of chronic lung-associated, smoking-associated diseases and diagnosis of any cancer at other sites.

The study was approved by the institutional ethics committee of University of Delhi, South Campus (4/IEC/TS/Gen/UDSC/18.2.2014 and 4/IEC/TS/Gen/UDSC/18.2.2020, AIIMS (IEC/NP-73/2013) and RML Hospital {File 118(19/2015)/IEC/PGIMER/RMLH}. Written consent for participation was obtained from all participants. In case the patient is unable to sign the consent, a family member of the patient signed the consent. Follow-up was done and records were maintained at AIIMS for date of confirmed diagnosis/registration/treatment, definitive date of death/last follow-up and smoking status.

DNA genotyping

About 2ml blood was drawn from the subjects and collected in vacutainers containing EDTA to avoid

coagulation. The gDNA from blood were isolated using a kit (DNA Blood Minikit, Qiagen) according to manufacturer's protocol. Quality and quantity were determined by spectrophotometry (Nanodrop, ThermoScientific). The 260/280 ratio of the samples were greater than or equal to 1.8.

A region spanning rs9387478 was amplified using a primer set- FP: CTCCAACAAAAGCTTCAATTCC and RP: GATTCAGAGCATAAGCTTAGG resulting in an amplicon of size 500bp. The PCR product was digested using HindIII and the products were separated on a 3% agarose gel. The SNP rs9387478 was genotyped using the DNA band pattern obtained after digestion.

Statistical Analysis

The statistical power and minor allele frequency (MAF) in cases and controls were calculated using Bioinformatics Institute's Online Size Estimator (OSSE). Logistic Regression models were built for estimating the contribution of smoking and genetic variant to lung cancer. The effects were estimated using odds ratios (OR), chi ($\chi 2$) square test, and corresponding 95% confidence intervals (CI).

Survival analysis: Survival was estimated using Kaplan Meier curves and results were compared using the log-rank test. Survival analysis was performed in lung cancer patients based on the three genetic models. The graph was plotted using Graphpad Prism and a two-tailed p-value of <0.05 was considered statistically significant.

eQTL and expression analysis

To examine the functional consequences of possible genotypes, gene expression data of DCBLD1 and ROS1 flanking the SNP was extracted from GTex database (Carithers et al., 2015). The tissue-specific expression quantitative trait loci (eQTL) were investigated using the GTex (GTEx, release v7 and human genome build 37) portal (https://www.gtexportal.org/home). The p-value and the Normalized Effect Size (NES) were considered while evaluating the results. NES is the slope of the linear regression and is computed as the effect of the alternative allele relative to the reference allele in the human genome reference GRCh38/hg38. The expression of DCBLD1 and ROS1 was compared in never-smokers and lung cancer patients using the TCGA dataset. The changes in the chromatin state as a result of the polymorphism was analyzed using Haploreg v4 (Lucas et al., 2016). The Kaplan Meier curve for survival vs gene expression was constructed using KM Plotter (http://www.kmplot.com) which is an online public database that integrate gene expression and clinical data in lung, breast, ovarian or gastric cancers. The survival curve for adenocarcinoma patients (Gyorffy et al., 2013) were plotted separately for DCBLD1 and ROS1 expression.

Table 1. Characteristics of Study Participants

Variables	Cases	Controls
Total No.	100	231
Sex		
Male	90 (90%)	172 (74.46%)
Female	10 (10%)	59 (25.54%)
Smoking Status		
Non-smoker	30 (30%)	105 (45.55%)
Smoker	70 (70%)	126 (54.55%)

Results

The study was performed in 331 samples comprising 100 lung cancer patients and 231 age-matched control samples. Participants consisted of both gender (males and females) and similar age groups of 40-75 years. The samples consisted of patients diagnosed with Lung adenocarcinoma, Non-Small Cell Lung Carcinoma (NSCLC) and Small cell Lung Carcinoma (SCLC) cases. The main clinical characteristics of the subjects are presented in Table 1.

The amplified PCR product was digested with HindIII and the product was resolved on an agarose gel. The three possible genotypes can be visualized to represent i) 3 bands at 500bp, 350bp and 55bp corresponding to heterozygote CA, ii) a double band of 350bp and 55bp corresponding to the homozygous genotype AA and iii) a single band of 500bp corresponding to the genotype CC (Figure 1).

The region spanning rs9387478 at chromosome 6q22.2 consists of 129 SNPs with MAF >0.2 and r2≥0.8. The SNP addressed in this study, rs9387478 is in high Linkage Disequilibrium (LD) with 9 other SNPs. The

Table 2. Allele Frequency of rs9387478 in Different Population

Genome Project	Population	Allele Frequency	
1000 GENOME	Global	C=35.6%	A=64.4%
	African	C=9.8%	A=90.2%
	American	C=56.1%	A=34.9%
	East Asian	C=50.2%	A=49.8%
	European	C=50.0%	A=50.0%
	South Asian	C=26.6%	A =73.4%
	Gih	C=26.2%	A=73.8%
ALFA	Global	C=50.6%	A=49.4%
	African	C=52.6%	A=47.4%
	American	C=53.0%	A=47.4%
	East Asian	C=54.9%	A=45.1%
	South Asian	C=35.1%	A=64.9%
HAPMAP	Global	C=34.9%	A=65.0%
	African	C=16.9%	A=83.1%
	American	C=59.4%	A=40.6%
	Asian	C=53.5%	A=46.5%
	European	C=54.5%	A=45.5%
GNOMAD	Global	C=39.8%	A=60.2%
	African	C=25.0%	A=85.0%
	American	C=55.1%	A=44.9%
	East Asian	C=50.2%	A=49.8%
	South Asian	C=31.4%	A=68.6%

A, adenosine (minor allele); C, cytosine (major allele)

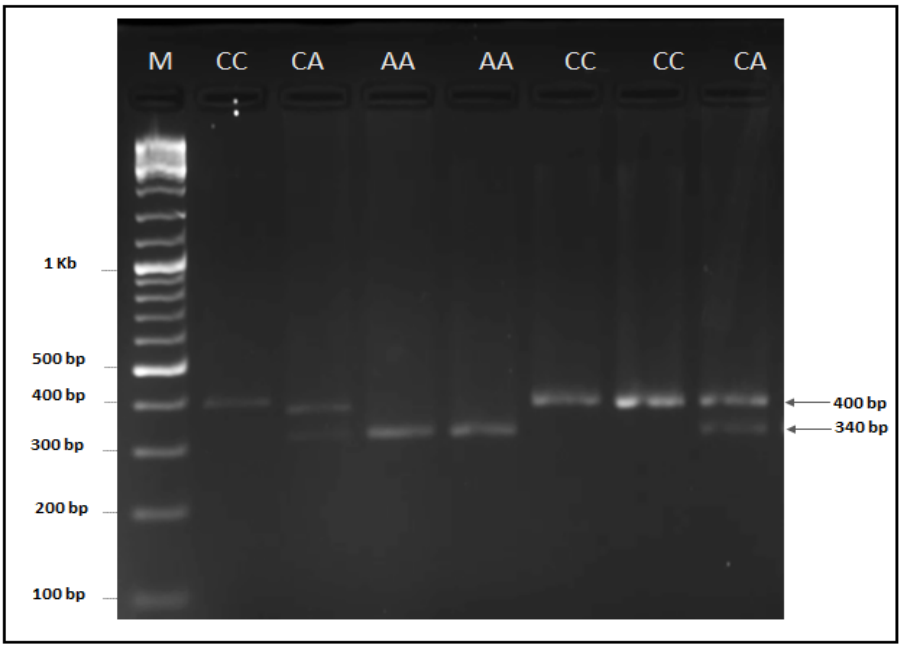

Figure 1. Representation of Genotypes Obtained from Patient Samples. The SNP (C>A) generates a restriction site for HindIII. When the gDNA samples were PCR amplified and digested with HindIII, three fragments separate on agarose gel - a single band at 400bp (in CC), triple bands of 400bp, 340bp and 60bp (in CA) and single band at 340 bp (in AA).

Table 3. Association of rs9387478 with Smoking (Nicotine Dependence)

SNP ID	Variable	Reference Genotype	OR	95% CI
	Genotype Model			
rs9387478	Het (CA) Vs Common Hz(CC)	CC	3.2344	1.8163-5.7597
	Rare Hz (AA) Vs Common Hz (CC)	CC	0.5921	0.3441-1.0189
	Dominant Model			
rs9387478	Het (CA) + Rare Hz (AA) Vs Common Hz (CC)	CC	1.375	0.8717-2.1688
	Recessive Model			
rs9387478	Rare Hz (AA) Vs Het (CA) + Common Hz (CC)	CC	0.3509	0.2156- 0.5711

A, adenosine (minor allele); C, cytosine (major allele); OR, Odds Ratio; CI, Confidence Interval

allele frequencies of rs9387478 (chr6-117465017-C-A) in different populations according to 1000Genome project, ALFA (Allele Frequency Aggregator), HapMap and gnomAD (Genome Aggregation database) are represented in Table 2. We have also calculated the allele frequency of rs9387478 in our present study. The minor allele frequency (MAF) of the SNP in the North Indian population is A=0.4018. Unexpectedly, the minor allele frequency of our population deviates from the MAF of the GIH (Gujarati in Houston) population.

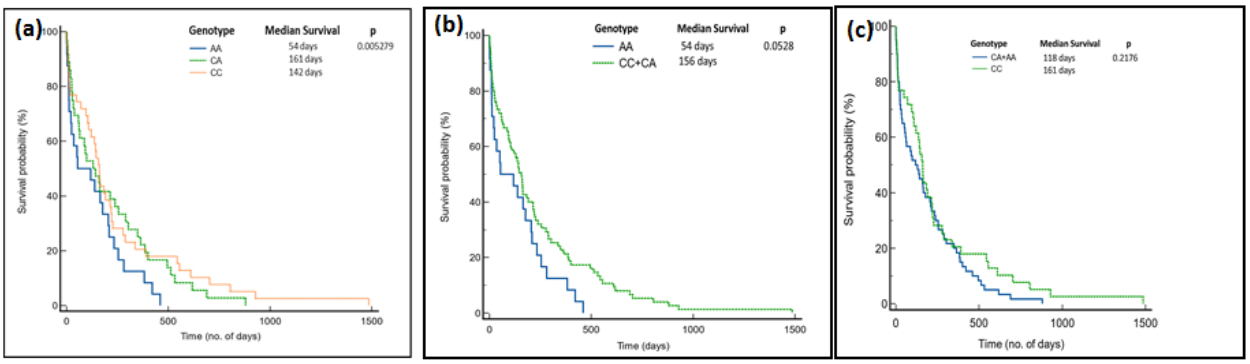

Figure 2. Survival Analysis of Patients with Different rs9387478 Genotypes. The relationship between rs9387478 polymorphism genotypes and survival probability was determined. Comparison of the overall survival of patients by Kaplan Meier analysis and Log Rank test shows that patients with AA genotype have the lowest overall survival in AA vs CA vs CC (a) and in AA vs CC+CA (b) but no significant change in CC vs CA+AA (c).

Table 4. Association of rs9387478 with Lung Cancer (Crude Odds Ratio)

SNP ID	Variable	Reference Genotype	OR	95% CI
	Genotype Model			
rs9387478	Het (CA) Vs Common Hz(CC)	CC	0.5672	0.3317 - 0.9697
	Rare Hz (AA) Vs Common Hz (CC)	CC	1.9506	1.0070 - 3.7785
	Dominant Model			
rs9387478	Het (CA) + Rare Hz (AA) Vs Common Hz (CC)	CC	0.8129	0.5004 - 1.3205
	Recessive Model			
rs9387478	Rare Hz (AA) Vs Het (CA) + Common Hz (CC)	CC	2.6547	1.4560 - 4.8402

A, adenosine (minor allele); C, cytosine (major allele); OR, Odds Ratio; CI, Confidence Interval

Genetic models for association studies by logistic regression

The minor allele 'A' was taken as the risk allele in our study and its association with smoking (Table 3) as well as lung cancer (Table 4) was analyzed using different genetic models: the genotype model, dominant model, and the recessive model. In the genotype model, the presence of the risk (minor) allele was compared to the wildtype common homozygous (CA Vs CC, AA Vs CC). In the dominant model, the risk homozygotes (AA) and the heterozygotes (CA) was combined and compared to the wildtype (CC). And in the recessive model presuming that the presence of both alleles will render susceptibility, the risk genotype (AA) was compared to the wildtype (CC) and the heterozygote (CA) taken together.

Association of the single minor allele A with smoking

In our population (cases and controls together), association of the SNP rs9387478 was analyzed with smoking and we observed that rs9387478 is associated with smoking in concordance with the genotype model (OR= 3.2344, 95%CI= 1.8163-5.7597, z score= 3.987 and p-value = 0.0001) and the dominant model (OR=1.357, 95%CI= 0.8717-2.1688, z score= 1.37 and p-value= 0.1708) (Table 3). While a single copy of the risk allele 'A' is sufficient for significantly increasing nicotine dependence, the presence of both copies of the allele 'AA' does not show an association with nicotine dependence in our dataset.

We have analyzed the effect of smoking on susceptibility to lung cancer development independent of other factor(s). Expectedly, we found that exposure to nicotine in the form of bidi/cigarette (a major form

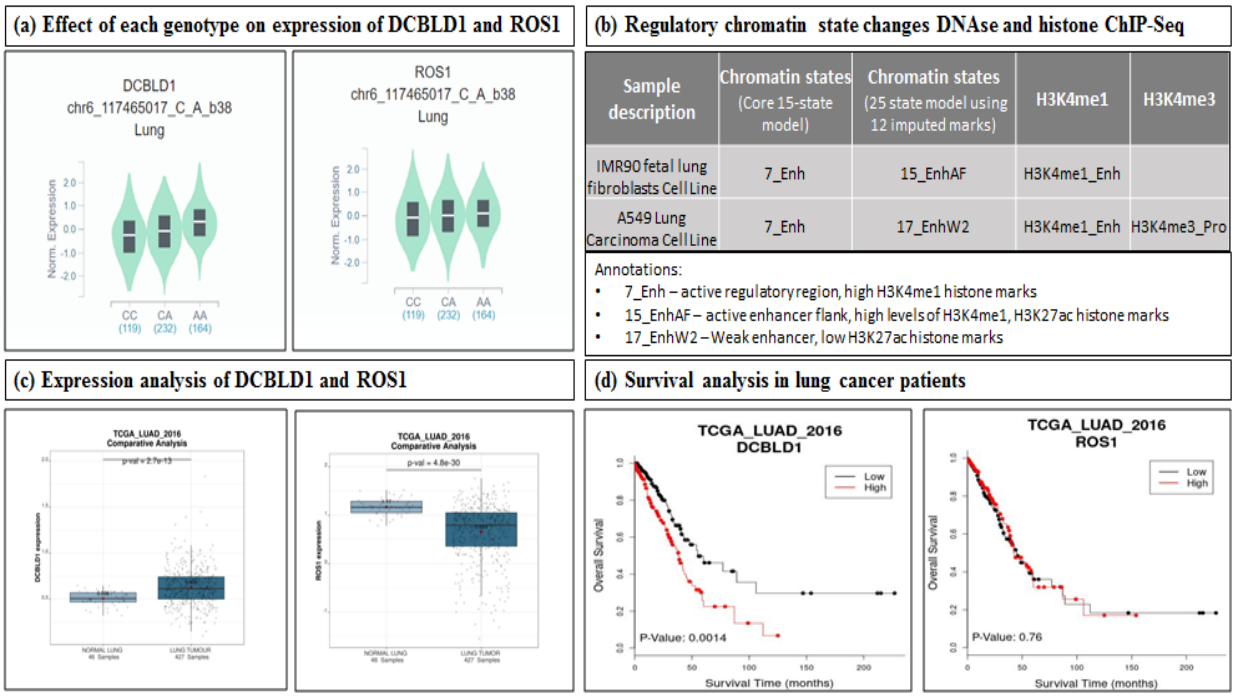

Figure 3. Association of rs9387478 Genotypes on Regulation and Expression of ROS1 and DCBLD1 (a) Effect of each genotype on expression of DCBLD1 and ROS1. The median normalised expression of DCBLD1 transcript associated with each genotype are CC= -0.2554, CA= -0.06806 and AA= 0.2932 while the same for ROS1 are CC = -0.1022, CA= -0.004858 and AA= 0.08268) (b) Regulatory chromatin state changes DNAse and histone ChIP-Seq using Roadmap Epigenetics Consortium, 2015 and HaploReg v4 (https://pubs.broadinstitute.org/mammals/haploreg) reveals increase in active transcription marks of H3K4 with cell-lines of AA genotype (c) Expression analysis of DCBLD1 and ROS1 in ever smoking (former and current smoker) normal and lung cancer patients (d) Survival plot of lung adenocarcinoma patients of TCGA datasets with low and high expression of DCBLD1 and ROS1.

Table 5. Association of rs9387478 with Lung Cancer (OR adjusted to Smoking)

SNP ID	Variable	Reference Genotype	OR	95% CI
	Genotype Model			
rs9387478	Het (CA) Vs Common Hz (CC)	CC	0.3571	0.1774 - 0.7189
	Rare Hz (AA) Vs Common Hz (CC)	CC	1.5714	0.7097 - 3.4797
	Dominant Model			
rs9387478	Het (CA) + Rare Hz (AA) Vs Common Hz (CC)	CC	0.6	0.3244 - 1.1096
	Recessive Model			
rs9387478	Rare Hz (AA) Vs Het (CA) + Common Hz (CC)	CC	2.75	1.3525 - 5.5917

A, adenosine (minor allele); C, Cytosine (major allele); OR, Odds Ratio; CI, Confidence Interval

of nicotine intake in our population) increased the risk of lung cancer (OR=1.9444, 95%CI=1.1794 to 3.2059, z score=2.607 and p-value=0.0091) in the North Indian population being studied.

Significant association of risk allele AA with lung cancer

By using the genetic models, we explored the effect of the minor allele A of SNP rs9387478 on lung cancer in the North Indian population. In our study, rs9387478 was found to be highly associated with lung cancer as shown in Table 4. The Crude Odds Ratio (OR) in the genotype model shows that the homozygous risk genotype AA is associated with lung cancer (OR=1.95 95%CI=1.0070-3.7787, z Score= 1.9881, p-value= 0.0476). While the Odds Ratio (OR) in the dominant model is less than 1 which implies that the presence of a single risk allele has lower odds of association with lung cancer, the recessive model demonstrates that presence of both copies of the risk allele of rs9387478 can significantly confer susceptibility to lung cancer as shown by the Crude OR (OR= 2.6547, 95%CI= 1.4560-4.8402, z score= 3.186, p-value= 0.0014) inferring a recessive effect of the risk allele A.

There have been extensive studies on the synergistic effect of smoking and genetic predisposition to lung cancer in many populations including the population considered in this study (Pandey et al. 2017) and therefore, we extended our study by considering smoking as a contributing factor. Similar to the crude OR, rs9387478 shows an association to lung cancer when exposure to tobacco is taken as an additional factor. The genotype model (OR=1.5714, 95%CI= 0.7097-3.4797, z score= 1.114, p-value= 0.2651) and the recessive model (OR= 2.75, 95%CI= 1.3525-5.5917, z score= 2.794, p-value= 0.0052) explains the susceptibility of individuals with AA genotype in rs9387478 to lung cancer in smokers. In our study, among the non-smokers 54 out of 135 (40%) are homozygous (CC), 24 out of 135(17.78%) are heterozygous (CA) and 57 out of 135 (42%) are homozygous for the risk allele (AA).

Survival analysis of patients with different rs9387478 genotypes

The lung cancer patients in our study were classified into the three subtypes (CC, CA and AA) to check if the different genotypes of rs9387478 have any effect on the survival time of the lung cancer patients. The median survival time (MST) of the patients with CC, CA and AA genotypes were 161 days, 142 days, and 54 days respectively {$\chi2$=10.48793, p=0.005279, Kaplan Meier analysis, Log Rank test; Figure 2 (a)}. On comparing the overall survival (OS) of the patients with genotype AA vs CC/CA, it was found that the median survival time were 54 days and 156 days respectively {$\chi2$ =0.394943, p= 0.0528, Kaplan Meier Analysis, Log Rank test; Figure 2 (b)}. We also compared the overall survival of genotype CC with genotype AA/CA. There was no significant difference in the median survival between the two subtypes {MST= 161 days vs 118 days, $\chi2$=1.5204, p= 0.2176 Figure 2 (c)}.

Effect of rs9387478 in ROS1 and DCBD1 expression, changes in chromatin signatures and analysis of LUAD samples and overall survival analysis based on gene expression

We extended our study to analyze the effect of the SNP on the expression of DCBLD1 and ROS1. eQTL analysis of rs9387478 was performed using GTex and QTL effect was determined in DCBLD1 and ROS1{Figure 3(a)} which predicted differential expression of the two genes due to polymorphism at the locus. The NES, which estimates the effect of the alternative allele compared to the reference allele was the highest for DCBLD1 in AA genotype (CC=-0.2554, CA=-0.06806, AA=0.2932) exhibiting a significant effect of the SNP on DCBLD1 expression in lung tissues (NES=0.368, p-value=2.8e-31). However, the effect of SNP on ROS1 expression is not significant (NES=0.071, p-value=0.46) evident from observed minor increase of ROS1 expression in AA genotype (CC= -0.1022, CA= -0.004858 and AA= 0.08268) (Figure 3(a)).

Encyclopaedia of DNA Elements (ENCODE) shows that the locus contains chromatin state segmentation and enhancer histone mark binding sites. We analyzed changes in the chromatin state in the region due to the SNP using Haploreg v4 and observed active histone mark enrichment because of rs9387478 C>A polymorphism. H3K4me1, H3K4me1, and H3K27ac were predicted to enrich if the genotype at the locus is AA as per the Roadmap Epigenomics Consortium, 2015 in lung cancer cell lines A549 and IMR90 {Figure 3 (b)}.

We analyzed the expression of DCBLD1 and ROS1 in normal lung tissue samples vs lung cancer tissue samples using the TCGA dataset. While lung cancer samples exhibit an increased expression of DCBLD1 (Normal= 0.5, Lung tumor= 0.632, p-value=2.7e-13),

ROS1 expression was lower in lung cancer sample (Normal=1.17, Lung Tumor= 0.654, p-value= 4.8e-30) {Figure 3(c)}. These differences in the expression level of DCBLD1 and ROS1 in normal vs tumor in the TGCA dataset guided us to investigate the association of expression of these two genes with the overall survival of lung cancer patients. The results of the Kaplan Meier plot indicate that higher expression of DCBLD1 is associated with poor survival (p-value=0.0014) while ROS1 had no significant effect on the overall survival of the patients (p-value=0.76) {Figure 3(d)}.

Discussion

Genetic variants are known to influence the development of diseases and are a promising approach for the study of genetic changes in lung cancer (Cai et al., 2013). The identification of risk genetic markers can aid in early disease detection and treatment of lung cancer but among the large number of SNPs identified by GWAS for lung cancer susceptibility, very few are validated as predictive biomarkers. rs9387478 was previously identified in a GWAS case-control study as a lung cancer susceptibility locus in Asian population (Wang et al., 2016), Korean male population (Yoo et al., 2020) and European population (McKay et al., 2017), however, no study has been conducted in the Indian population.

In our cohort, the SNP rs9387478 located in chromosome 6q22.2 is associated with nicotine dependence, susceptibility to lung cancer, and disease prognosis. rs9387478 C>A increases nicotine dependence in a dominant model. Also, rs9387478 is associated with lung cancer (recessive model); individuals with the genotype AA are 2.7 times more susceptible to developing lung cancer than individuals with CC/CA genotypes. Interestingly, an extrapolation of our study to test if smoking further augments its association with lung cancer revealed that smokers with genotype AA are 2.75 times more susceptible than the other genotypes. In our cohort, patients with AA genotype have a significantly lesser overall survival than CC/CA genotype. Our result is similar to a study (Han et al., in 2016) where it was observed that patients with CC genotype at rs9387478 locus had better overall survival compared to other genotypes (CA/AA) (Han et al., 2016). This observation may reflect altered expressions of ROS1 and DCBLD1 which are known to aid in the progression of many types of tumors including lung cancer progression.

Expression levels of ROS1 and DCBLD1 were correlated with different possible genotypes (CC, CA, and AA) using GTex and it was observed that the mRNA level of DCBLD1 was significantly higher in AA genotype while a marginal increase was seen in ROS1 expression. C>A polymorphism in the locus was associated with chromatin state segmentation along with enrichment of promoter and enhancer histone marks, such as H3K27ac, H3K4me1 and H3K4me3 (ENCODE analysis) which is in concordance with the increased mRNA levels of DCBLD1 and ROS1. However, this interesting observation needs further validation.

DCBLD1 and ROS1 upregulation has been shown to lead to an increased proliferative and invasive capacity of cancer cells. We have therefore explored the TCGA dataset to analyze the overall survival of lung adenocarcinoma patients based on low/high expression of DCBLD1 and ROS1. A lower expression of DCBLD1 is associated with significantly improved overall survival while no such change in overall survival was observed with ROS1 expression in lung cancer patients. These observations are previously unreported. Functional attributes of altered rs9387478 polymorphism on the expression of the two genes, disease progression, and prognosis will help in validation and confirmation.

Overall, we have addressed lung cancer risk, nicotine dependence, and poor survival outcomes as an effect of rs9387478 SNP in the North Indian population. We recognize a need to expand the sample size for a decisive conclusion to the study as well as evaluate if the observations are affected (or not) by sex. However, with the current study, we conclude that the SNP is a promising prognostic marker/genetic risk marker in our population. Additionally, rs9387478 was associated with EGFR status of lung cancer patients wherein the possibility of ROS1 or EGFR as therapeutic targets in lung cancer was discussed (Seow et al., 2017). Therefore, further investigation to determine a correlation of the SNP with EGFR status in our population shall be attempted. Our study shows the potential in using rs9387478 as a predisposition marker facilitating the early genetic diagnosis of lung cancer in the North Indian population.

Author Contribution Statement

JC: Concept and design, methodology and experiments, analysis, writing manuscript; NP: sample collection, methodology; LKS and AM: Resource, Data acquisition, storage, and analysis; TS: Concept and design, resource and funding, writing manuscript and editing, correspondence.

Acknowledgements

JC is a recipient of ICMR-Senior Research fellowship and DBT-JRF fellowship. NP was a recipient of UGC-BSR fellowship. We thank Mr. Shakeel for help in lab maintenance and Ms. Shabana Ansari and Ms. Kasthuri for help with maintaining lab accounts.

Funding

This work was supported with grants from Science and Engineering Research Board (SERB-CRG and SERB-POWER) and University funded Faculty Research Project (FRP) and Major Research Project (MRP) grant, IOE scheme to TS. Funding support from DBT, ICMR, UGC for SAP (BRS III), DST for FIST (level II) program, DU-DST PURSE (phase II) is gratefully acknowledged.

Ethical Declaration

The authors state that they obtained appropriate institutional review board approval. The study was approved by the institutional ethics committee of University of Delhi, South Campus (4/IEC/TS/Gen/UDSC/18.2.2014 and 4/IEC/TS/Gen/UDSC/18.2.2020),

AIIMS (IEC/NP-73/2013) and RML Hospital {File 118(19/2015)/IEC/PGIMER/RMLH}.

Conflict of interest

The authors declare no conflict of interests.

References

Acquaviva J, Wong R, Charest A (2009). The multifaceted roles of the receptor tyrosine kinase ROS in development and cancer. *Biochim Biophys Acta*, **1795**, 37-52.

Cai M, Dai H, Qiu Y, et al (2013). SNP set association analysis for genome-wide association studies. *PLoS One*, **8**, e62495.

Cardin GB, Bernard M, Rodier F, Christopoulos A (2021). DCBLD1 is associated with the integrin signaling pathway and has prognostic value in non-small cell lung and invasive breast carcinoma. *Sci Rep*, **11**, 12753.

Charest A, Wilker EW, McLaughlin ME, et al (2006). ROS fusion tyrosine kinase activates a SH2 domain-containing phosphatase-2/phosphatidylinositol 3-kinase/mammalian target of rapamycin signaling axis to form glioblastoma in mice. *Cancer Res*, **66**, 7473-81.

Cilloni D, Carturan S, Bracco E, et al (2013). Aberrant activation of ROS1 represents a new molecular defect in chronic myelomonocytic leukemia. *Leuk Res*, **37**, 520-30.

Gainor JF, Tseng D, Yoda S, et al (2017). Patterns of metastatic spread and mechanisms of resistance to crizotinib in ROS1-positive non-small-cell lung cancer. *JCO Precis Oncol*, **1**, PO.17.00063.

Gu F, Zhao C, Jiang T, et al (2020). Association between nicotine-dependent gene polymorphism and smoking cessation in patients with lung cancer. *Clin Lung Cancer*, **21**, 171-6.

Győrffy B, Surowiak P, Budczies J, Lánczky A (2013). Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One*, **8**, e82241.

Han JF, An SJ, Su WM, et al (2016). Polymorphism of rs9387478 correlates with overall survival in female nonsmoking patients with lung cancer. *Int J Biol Markers*, **31**, e144-52.

Hancock DB, Guo Y, Reginsson GW, et al (2018). Genome-wide association study across European and African American ancestries identifies a SNP in DNMT3B contributing to nicotine dependence. *Mol Psychiatry*, **23**, 1911-9.

Islami F, Torre LA, Jemal A (2015). Global trends of lung cancer mortality and smoking prevalence. *Transl Lung Cancer Res*, **4**, 327-38.

Jones MR, Rose AM, Baillie DL (2013). The ortholog of the human proto-oncogene ROS1 is required for epithelial development in C. elegans. *Genesis*, **51**, 545-61.

Koshikawa K, Osada H, Kozaki K, et al (2002). Significant up-regulation of a novel gene, CLCP1, in a highly metastatic lung cancer subline as well as in lung cancers in vivo. *Oncogene*, **21**, 2822-8.

Lan Q, Hsiung CA, Matsuo K, et al (2012). Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet*, **44**, 1330-5.

Carithers LJ, Ardlie K, Barcus M, et al (2015). A novel approach to high-quality postmortem tissue procurement: The GTEx Project. *Biopreserv Biobank*, **13**, 311-9.

Lin JJ, Shaw AT (2017). Recent advances in targeting ROS1 in lung cancer. *J Thorac Oncol*, **12**, 1611-25.

López-Flores LA, Pérez-Rubio G, Falfán-Valencia R (2017). Distribution of polymorphic variants of CYP2A6 and their involvement in nicotine addiction. *EXCLI J*, **16**, 174-96.

McKay JD, Hung RJ, Han Y, et al (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet*, **49**, 1126-32.

Mehta A, Saifi M, Batra U, Suryavanshi M, Gupta K (2020). Incidence of ROS1-rearranged non-small-cell lung carcinoma in India and efficacy of Crizotinib in lung adenocarcinoma patients. *Lung Cancer (Auckl)*, **11**, 19-25.

Kim M, Lee KT, Jang HR, et al (2008). Epigenetic down-regulation and suppressive role of DCBLD2 in gastric cancer cell proliferation and invasion. *Mol Cancer Res*, **6**, 222-30.

Mohan A, Garg A, Gupta A, et al (2020). Clinical profile of lung cancer in North India: A 10-year analysis of 1862 patients from a tertiary care center. *Lung India*, **37**, 190-7.

Pandey N, Pal S, Sharma LK et al (2017). SNP rs16969968 as a strong predictor of nicotine dependence and lung cancer risk in a north Indian populatio. *Asian Pac J Cancer Prev*, **18**, 3073-9.

Pérez-Rubio G, Pérez-Rodríguez ME, Fernández-López JC, et al (2016). SNPs in NRXN1 and CHRNA5 are associated to smoking and regulation of GABAergic and glutamatergic pathways. *Pharmacogenomics*, **17**, 1145-58.

Quach BC, Bray MJ, Gaddis NC, et al (2020). Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. *Nat Commun*, **11**, 5562.

Sehgal K, Piper-Vallillo AJ, Viray H, et al (2020). Cases of ROS1-rearranged lung cancer: when to use crizotinib, entrectinib, lorlatinib, and beyond. *Precis Cancer Med*, **3**, 17.

Seow WJ, Matsuo K, Hsiung CA, et al (2017). Association between GWAS-identified lung adenocarcinoma susceptibility loci and EGFR mutations in never-smoking Asian women, and comparison with findings from Western populations. *Hum Mol Genet*, **26**, 454-65.

Shukla S, Pandey RK, Kant S, Garg R, Husain N (2019). ROS-1 re-arrangements and c-MET amplifications in adenocarcinoma lung: A tertiary care center study from North India. *Indian J Pathol Microbiol*, **62**, 433-6.

Stadler ZK, Thom P, Robson ME, et al (2010). Genome-wide association studies of cancer. *J Clin Oncol*, **28**, 4255-67.

Sung H, Ferlay J, Siegel RL, et al (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, **71**, 209-49.

Thorgeirsson TE, Geller F, Sulem P, et al (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, **452**, 638-42.

Uguen A, De Braekeleer M (2016). ROS1 fusions in cancer: a review. *Future Oncol*, **12**, 1911-28.

Wang J, Liu Q, Yuan S, et al (2017). Genetic predisposition to lung cancer: comprehensive literature integration, meta-analysis, and multiple evidence assessment of candidate-gene association studies. *Sci Rep*, **7**, 8371.

Wang Y, Ma R, Liu B, et al (2020). SNP rs17079281 decreases lung cancer risk through creating an YY1-binding site to suppress DCBLD1 expression. *Oncogene*, **39**, 4092-4102.

Wang Z, Seow WJ, Shiraishi K, et al (2016). Meta-analysis of genome-wide association studies identifies multiple lung cancer susceptibility loci in never-smoking Asian women. *Hum Mol Genet*, **25**, 620-9.

Ward LD, Kellis M (2016). HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res*, **44**, 877-81.

World Health Organization Tobacco (2020) https://www.who.int/news-room/fact-sheets/detail/tobacco. Published May 27, 2020. Accessed December 20, 2020.

Yoo SS, Kang HG, Choi JE, et al (2017). Effects of polymorphisms identified in genome-wide association studies of never-smoking females on the prognosis of non-small cell lung

cancer. *Cancer Genet*, **2017**, 212-213, 8-12.
Yoo SS, Kang HG, Choi, et al (2020). The effect of susceptibility variants, identified in never-smoking female lung cancer cases, on male smokers. *Korean J Intern Med*, **35**, 929-35.