scientific reports

OPEN

Identification of potential susceptibility loci for non-small cell lung cancer through whole genome sequencing in circadian rhythm genes

Xiaohang Xu[1,2,5], Luopiao Xu[1,2,5], Zeyong Lang[1], Gege Sun[1], Junlong Pan[1], Xue Li[1,2], Zilong Bian[1] & Xifeng Wu[1,2,3,4]✉

Lung cancer is a malignant tumor with a high morbidity and mortality rate worldwide, causing an increasing disease burden. Of these, the most common type is non-small cell lung cancer (NSCLC), which accounts for 80–85% of all lung cancer cases. Genetic research is crucial for continuously discovering susceptibility genes related to lung cancer for in-depth study. The role of genetic predisposition in the development of NSCLC, particularly within circadian rhythm pathways known to govern various physiological processes, is increasingly acknowledged. Yet, the association between genetic variants of circadian rhythm-related genes and NSCLC susceptibility among Chinese populations is not fully understood. This study carried out a two-phase (discovery and validation stages) research design to identify genetic variants associated with NSCLC risk within the circadian rhythm pathway. We employed extensive whole-genome sequencing (WGS) for 1,104 NSCLC cases and 9,635 controls. FastGWA-GLMM was used for single-locus risk association analysis of NSCLC, and we screened candidate SNPs in the validation set that comprised 4,444 cases and 174,282 controls from the Biobank Japan Project (BBJ). Furthermore, GCTA-COJO conditional analysis was utilized to confirm SNPs related to NSCLC risk. Finally, potential genetic variations that may regulate gene expression were explored in GTEx and QTLbase. RNA sequencing data were utilized for transcriptomic verification. Our study identified eight candidate SNPs associated with NSCLC susceptibility within the circadian rhythm pathway that met the requirement with $P < 0.05$ in both the discovery and validation populations. After conditional analysis, five of these SNPs remained. The A allele of *CUL1* rs78524436 ($OR_{meta} = 1.18$, 95%CI: 1.09–1.29, $P_{meta} = 7.99e\text{-}5$) and the A allele of *TEF* rs9611588 ($OR_{meta} = 1.06$, 95%CI: 1.02–1.10, $P_{meta} = 1.28e\text{-}3$) were associated with an increased risk of NSCLC. The A allele of *FBXL21* rs2069868 ($OR_{meta} = 0.86$, 95%CI: 0.80–0.96, $P_{meta} = 4.78e\text{-}4$), the T allele of *CSNK1D* rs147316973 ($OR_{meta} = 0.76$, 95%CI: 0.65–0.88, $P_{meta} = 5.93e\text{-}4$), and the A allele of *RORA* rs1589701 ($OR_{meta} = 0.94$, 95%CI: 0.91–0.98, $P_{meta} = 3.40e\text{-}3$) were associated with a lower risk of NSCLC, separately. The eQTL results revealed an association between *RORA* rs1589701 and *TEF* rs9611588 with the expression levels of *RORA* and *TEF*, respectively. Transcriptome data indicated that *RORA* and *TEF* showed lower expression levels in tumor tissues compared to normal tissues ($P < 0.001$). Moreover, poorer survival was observed in patients with lower *RORA* and *TEF* expressions (log-rank $P < 0.05$). Our findings spotlight potential susceptibility loci within circadian rhythm pathway genes that modulate NSCLC carcinogenesis, which enriches the understanding of the genetic susceptibility of NSCLC in the Chinese population and provides a more solid basis for exploring the biological mechanism of circadian rhythm genes in NSCLC.

Keywords Non-small cell lung cancer, Circadian rhythm genes, Genome-wide association study, Gene expression

[1]Center of Clinical Big Data and Analytics of the Second Affiliated Hospital and School of Public Health, Zhejiang University School of Medicine, Hangzhou 310058, China. [2]Zhejiang Key Laboratory of Intelligent Preventive Medicine, Hangzhou 310058, China. [3]National Institute for Data Science in Health and Medicine, Zhejiang University,

Hangzhou 310058, Zhejiang, China. [4]School of Medicine and Health Science, George Washington University, Washington, DC, USA. [5]Xiaohang Xu and Luopiao Xu contributed equally to this work ✉email: xifengw@zju.edu.cn

The global incidence and mortality rates of cancer are escalating, with lung cancer emerging as a leading malignancy and a primary cause of cancer-related deaths worldwide[1,2]. In the United States, projections indicate that by 2023, there will be approximately 238,340 new lung cancer cases and about 127,070 deaths[2], presenting significant socioeconomic challenges and highlighting its urgency as a public health issue. Of these, the most common type is non-small cell lung cancer (NSCLC), which accounts for 80–85% of all lung cancer cases[2]. While smoking is recognized as the primary risk factor for lung cancer[3–5], the occurrence of the disease in individuals without a smoking history points to the role of genetic susceptibility[6,7].

In the last decade, genome-wide association studies (GWAS) focused on lung cancer have unearthed susceptibility loci across diverse populations[8]. These studies have significantly advanced our understanding of the genetic foundations of lung cancer, opening avenues for personalized prevention strategies. Moreover, there is an expanding evidence base linking genetic variations in cancer-related pathways to the onset of cancer[9–12] offering deeper insights into the disease's pathogenesis. The link between genetic variations in circadian rhythm genes and elevated cancer risk, including lung cancer[13–17], is increasingly supported by evidence.

The circadian rhythm pathway plays a crucial role in regulating metabolism[18], influencing the expression and functionality of enzymes involved in key metabolic processes like glucose metabolism, fatty acid metabolism, and cholesterol synthesis[19–22]. At the heart of the circadian rhythm pathway are the clock genes, such as CLOCK, BMAL1, PER, and CRY, which establish a transcriptional-translational feedback loop[23,24]. This regulatory mechanism ensures the rhythmic expression of clock-controlled genes (CCGs) that oversee a plethora of metabolic pathways and physiological activities[25]. Perturbations in circadian rhythms, due to factors such as irregular sleep patterns or shift work, have been linked to metabolic disorders, including obesity, diabetes, and cardiovascular diseases[26–29].Beyond its regulatory role in metabolism, the circadian rhythm pathway is instrumental in managing critical cellular processes, including cell division and proliferation, to ensure proper cell cycle progression and DNA replication[30]. Disturbances in circadian rhythms are associated with increased genomic instability and susceptibility to DNA damage-induced mutations, contributing to cancer development[31,32]. Nevertheless, the genetic factors contributing to NSCLC risk, particularly in the Chinese population, remain to be fully elucidated.

Given the stringent significance thresholds established for published studies, additional susceptibility loci with moderate association significance may be overlooked. Traditionally, published associations have relied on genotyping arrays to assess fixed panels containing hundreds of thousands to millions of common genetic markers, supplemented by imputation using population reference panels to enhance the coverage of genetic variation studies. This method, however, does not offer a comprehensive and unbiased approach to exploring a significant portion of genetic variation[33,34]. Moreover, there has been a scarcity of replication of these associations in previous studies and a thorough examination of false-positive results within their designs.

In our study, we systematically assessed the association of genetic variation in circadian pathway genes with susceptibility to NSCLC and validated it in an independent population. Therefore, we conducted large-scale whole-genome sequencing (WGS) on 1,104 individuals diagnosed with NSCLC and 9,635 cancer-free control individuals from the Chinese population to pinpoint potential susceptibility loci for lung cancer within circadian rhythm genes. Further, based on the expression level of susceptibility sites, the biological mechanism of NSCLC was preliminarily explored.

Methods

Sample collection for discovery population

The study's protocol received approval from the Ethics Committee of the Second Affiliated Hospital of Zhejiang University School of Medicine. Informed consent was obtained from all subjects and/or their legal guardian. A workflow chart illustrating the study's design is presented in Supplementary Fig. 1.

For the discovery phase, the study population included 1,104 NSCLC patients from the Second Affiliated Hospital of Zhejiang University School of Medicine (SAHZU) and 9,635 cancer-free controls from the Healthy Zhejiang One Million People Cohort in China. All lung cancer cases were histologically verified as NSCLC by pathologists, with no history of other cancers. Data on pathological classification and TNM staging were also gathered. Cancer-free controls were matched with cases on age, gender, and geographical region, excluding those with self-reported cancer diagnoses.

Essential data, including age, gender, BMI, and smoking status were collected from all participants through structured questionnaire-based in-person interviews by trained personnel. Smoking is defined as cumulative smoking of 100 cigarettes or more, otherwise, no smoking. Peripheral blood samples were collected and preserved at -80 °C in the laboratory for subsequent DNA extraction and whole-genome sequencing (WGS).

Data collection for validation population

The validation phase utilized data from the Biobank Japan (BBJ)[35]. This phase included 4,444 lung cancer patients and 174,282 controls. Comprehensive GWAS summary statistics were accessible on the PheWeb.jp website (https://pheweb.jp/).

DNA sample preparation

Genomic DNA was isolated from peripheral blood samples of participants using the QIAamp 96 DNA QIAcube HT Kit (Qiagen, Germany). Subsequently, DNA libraries were prepared with the MGIEasy FS DNA Library Prep Set (MGI, China).

Whole genome sequencing and quality control

Whole-genome sequencing was executed on the DIPSEQ platform (BGI, China). We adhered to the GATK Best Practices for the identification of germline short variants, including SNPs and insertions/deletions (Indels) up to 50 bp, for sequencing alignment. Joint calling was facilitated by GATK GenotypeGVCFs. Variants showing excessive heterozygosity were removed. Quality scores, including QualByDepth (QD), MappingQualityRankSumTest (MQRankSum), ReadPosRankSumTest (ReadPosRankSum), FisherStrand (FS), StrandOddsRatio (SOR), and depth (DP), were utilized for Variant Quality Score Recalibration (VQSR) to filter low-quality variants. Genotyping accuracy was enhanced through refinement using Beagle 5.4. Based on the common variants ($MAF \geq 0.01$), we estimated the sample relatedness and removed samples that failed in sex check, potentially monozygotic twins or duplicated samples using KING.

Association analysis

The software based on the Generalized Linear Mixed Model (GLMM) can specifically analyze genome-wide association studies of binary traits, and its operation efficiency is much higher than that of similar binary traits association analysis methods[36]. In this study, a common variant-based ($MAF \geq 0.01$) genome-wide association study was performed using the generalized linear mixed effect model (GLMM-fastGWA), adjusting for Age, Age^2, Sex, and PC1-PC10. Meta-analysis was performed with METAL (https://csg.sph.umich.edu/abecasis/metal/)[37], employing a fixed-effect model to integrate findings from Japanese populations (BBJ).

The candidate SNPs inclusion criteria were as follows: (1) The direction of the effect was consistent in both the discovery and validation populations; (2) *P* value was less than 0.05 in both populations; (3) In the meta-analysis, FDR_{meta} was less than 0.10 after correction (based on the SNPs less than 0.05 in both discovery and validation populations).

Conditional analysis

Conditional and Joint Multiple-SNP Analysis (COJO) was performed by GCTA (Genome-wide Complex Trait Analysis) software. COJO screens for SNPs independently associated with NSCLC susceptibility in the circadian pathway, which eliminates the recurrence of two or more mutually explainable sites in subsequent analyses due to a high Linkage Disequilibrium (LD). That is, any two loci with a distance of less than 1 Mb on the same chromosome were analyzed conditionally to observe whether other loci were also associated with the phenotype. If a point is not found to be statistically significant after conditional analysis, it means that its effect can be explained by other SNP sites and is filtered.

Circadian pathway genes and SNP selection

Our study reviewed the research strategies for screening candidate genes by referring to published literature[12,13,38,39]. We used the Kyoto Encyclopedia of Genes and Genomes, KEGG, https://www.kegg.jp/) and Reactome channel database (https://reactome.org/), search target channel names “Circadian Rhythm” and “Circadian Clock” by keyword. The genes associated with the circadian rhythm pathway of both KEGG and Reactome were collected. At the same time, the existing literature was supplemented[18,24,40–44]. Finally, the list of circadian pathway genes was sorted out and summarized, which encompasses 14 core circadian rhythm genes (the clock genes) and 32 key genes related to circadian rhythms (Supplementary Table 1).

We then used the Genome Reference Consortium Human Genome37 (GRCh37) data as the standard to identify the initiation sites of each gene in the list of circadian pathway genes. In this study, SNPs in the upstream and downstream 50 kb range of candidate genes were extracted for subsequent analysis, resulting in 15,462 independent SNPs. Tagging SNPs were pinpointed within 50 kb regions flanking each selected gene using the Clumping program (https://www.cog-genomics.org/plink/1.9/postproc#clump).

Gene-based analysis (MAGMA)

To conduct gene-based analysis of GWAS data, we used MAGMA to analyze summary SNP *P*-values from meta GWAS statistics. We first integrated individual SNP data (15,462 independent SNPs) at the gene level (annotation), and then used gene data (integrating information from individual SNPs) and lung cancer for association analysis to quantify the degree of association between each gene and phenotype.

eQTL analysis

To elucidate the effects of target SNPs, we the following publicly available eQTL database to explore the association between SNPs and the expression levels of corresponding genes in lung tissue, blood, and other tissues: Genotype-Tissue Expression Project[45] (GTEx, v8) (https://www.gtexportal.org/home/); QTLbase[46] (http://mulinlab.tmu.edu.cn/qtlbase/index.html). These resources enabled the investigation of associations between genetic variants and tissue-specific gene expression patterns.

Transcriptome data collection

A total of 417 NSCLC patients were recruited from the SAHZU from February 2013 to June 2022. The inclusion criteria and data collection are the same as sample collection for the discovery population. All specimens (tumor and matched normal tissue samples from the same patient) were cut fresh from surgically resected tissue samples. We followed the survival status of these NSCLC patients every six months.

Total RNA was extracted using AllPrep DNA/RNA Universal Kit (Qiagen, Germany) and sequencing libraries were generated using the MGIEasy Total RNA Library Prep Kit (MGI, China). The total RNA of the tumor and paired normal tissue samples were sequenced on the MGI HiSeq platform (BGI, Shenzhen) and 100 bp paired-end reads were generated. After quality control, we only included 747 tissue samples with low ribosome RNA ($\leq 30\%$) and high sample sequencing reads (≥ 20 million reads).

We also download gene expression data, age, gender, follow-up time, survival status, survival time, and other clinical information of NSCLC patients from The Cancer Genome Atlas database (TCGA, https://portal.gdc.cancer.gov/).

Gene expression analysis

Wilcoxon rank-sum test was used to compare the gene expression level between tumor and normal tissues and between different stages. A Cox regression model was implemented to estimate the hazard ratio (HR) using the R package survival and surviminer. We also conducted multi-variable cox regression analysis adjusting for gender, age, smoking, and clinical stage. We used both progression-free survival (PFS) and overall survival (OS) as survival status. Kaplan-Meier method and Log-rank test were used to compare the differences in survival status between the high and low-expression groups divided by the median values.

Statistical analysis

Statistical analyses were conducted using R version 4.3.0. A nominal significance level was set at a $P < 0.05$. *P*-values for each SNP and gene were adjusted using the false discovery rate (FDR) method to account for multiple testing, considering FDR < 0.10 as statistically significant. All methods were performed in accordance with the relevant guidelines and regulations.

Results

Host characteristics of the discovery populations

Characteristics of cases and controls within the discovery cohort are detailed in Table 1, encompassing 1,104 lung cancer cases and 9,635 controls. The average age was 56.80 years for cases and 58.80 years for controls, with cases having a higher proportion of current smokers (22.92% vs. 20.53%, $P < 0.001$). The predominant histological type was adenocarcinoma (90.67%), followed by squamous cell carcinoma (6.07%), with most patients diagnosed at stage I.

Association analysis of circadian pathway-related SNPs and NSCLC susceptibility

A total of 6,109,887 SNPs were analyzed for their association with NSCLC risk adjusting for Age, Age2, Sex, and PC1-PC10. We then tagged 15,462 SNPs within 46 genes of the circadian rhythm pathway. In the discovery phase, we defined $P < 0.05$ as the statistical significance threshold of potential association to screen for candidate SNPs for the next stage. A total of 888 SNPs were nominally associated with NSCLC risk ($P < 0.05$, Supplementary Table 2) and were then included in the validation populations of BBJ. After removing SNPs with discordant effect directions in the two datasets, eight SNPs met the requirement of $P < 0.05$ in both the discovery and validation populations and FDR < 0.10 in the meta-analysis (Fig. 1; Table 2).

In the meta-analysis, the A allele of *CUL1* rs78524436 was associated with increased susceptibility to NSCLC (OR_{meta} = 1.18, 95%CI: 1.09–1.29, P_{meta} = 7.99e-5), while the A allele of *FBXL21* rs2069868 was associated with reduced susceptibility to NSCLC (OR_{meta} = 0.86, 95%CI: 0.80–0.96, P_{meta} = 4.78e-4). The T allele of *CSNK1D* rs147316973 was associated with reduced susceptibility to NSCLC (OR_{meta} = 0.76, 95%CI: 0.65–0.88, P_{meta} = 5.93e-4). The A allele of *RORA* rs1589701 was associated with reduced susceptibility to NSCLC (OR_{meta} =

	Case (*N*; %)	Control (*N*; %)
Num	1104 (100.00)	9635 (100.00)
Gender		
Female	693 (62.77)	5974 (62.00)
Male	411 (37.23)	3661 (38.00)
Age (years, mean±sd)	56.80±12.70	58.83±9.75
BMI (kg/m2, mean±sd)	23.30±4.91	23.60±2.94
Smoking history		
No	717 (64.95)	7616 (79.05)
Yes	253 (22.92)	1978 (20.53)
Missing	134 (12.10)	41 (0.43)
Histology		
LUAD	1001 (90.67)	-
LUSC	67 (6.07)	-
Others	36 (3.26)	-
Pathological stage		
I	872 (79.00)	-
II	42 (3.80)	-
III	35 (3.17)	-
IV	12 (1.09)	-
Others	143 (12.95)	-

Table 1. Characteristics of discovery populations in the SAHZU population.

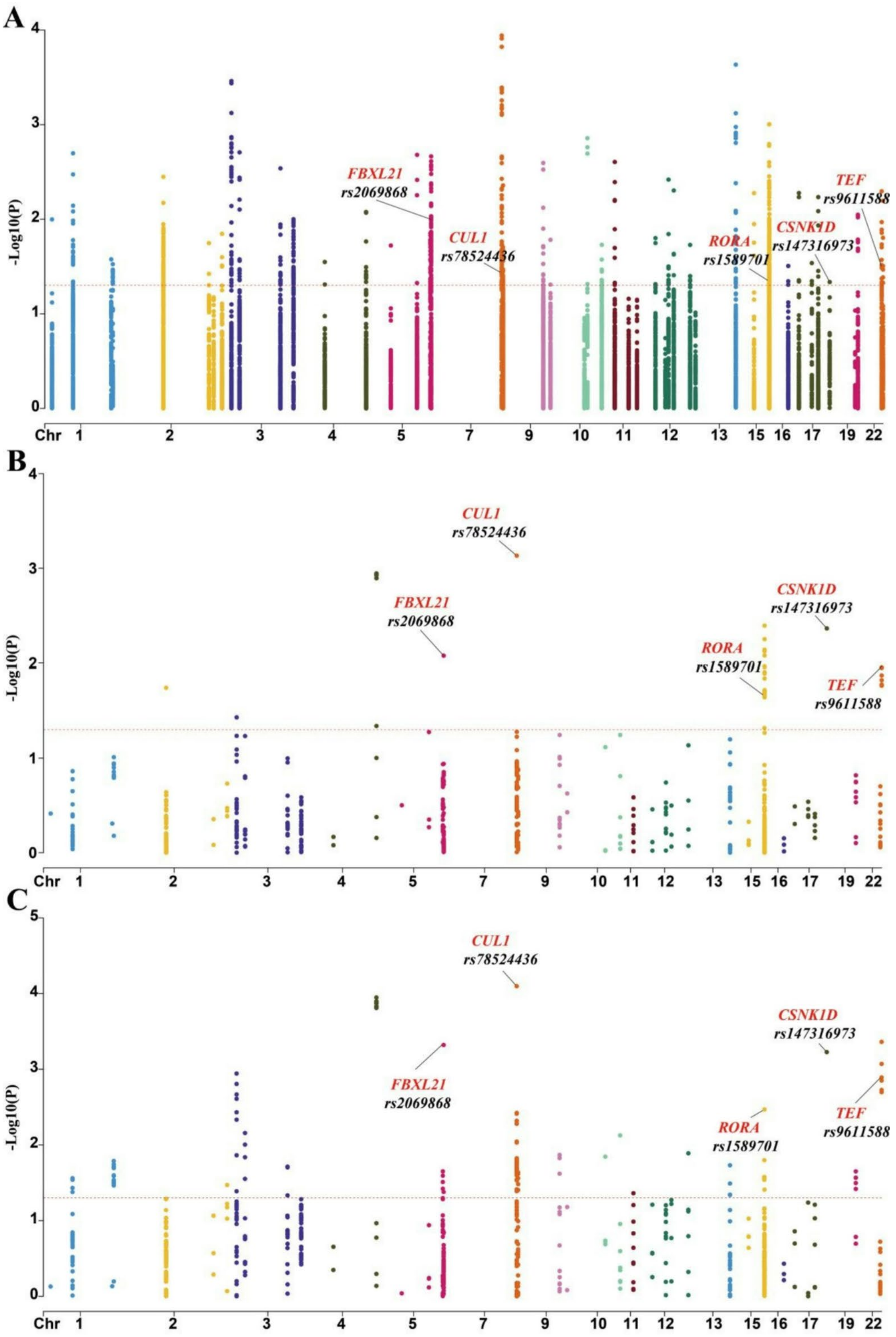

Fig. 1. GWAS results in the discovery, validation, the meta-analysis. SNPs were nominally associated with NSCLC risk (**A**, $P < 0.05$) and were then validated in the validation populations of BBJ (**B**). After removing SNPs with discordant effect directions in the two datasets, eight SNPs met the requirement with $P < 0.05$ in both the discovery and validation populations and FDR < 0.10 in the meta-analysis (**C**).

SNP	Chr	Pos_hg19	Gene	Alle	MAF	phase	OR (95%CI)	*P*	Heterogeneity
rs78524436	7	148372374	CUL1	A/G	0.07	Discovery	1.17 (1.00, 1.35)	3.71E-02	0.222
						Validation	1.19 (1.07, 1.31)	7.38E-04	
						Mata	1.18 (1.09, 1.29)	7.99E-05	
						FDR meta		7.76E-03	
rs2069868	5	135232966	FBXL21	A/G	0.08	Discovery	0.84 (0.74, 0.95)	9.77E-03	0.06
						Validation	0.88 (0.80, 0.96)	8.41E-03	
						Mata	0.86 (0.80, 0.93)	4.78E-04	
						FDR meta		1.96E-02	
rs147316973	17	80253346	CSNK1D	T/C	0.02	Discovery	0.70 (0.50, 0.99)	4.61E-02	0.209
						Validation	0.77 (0.65, 0.92)	4.33E-03	
						Mata	0.76 (0.65, 0.88)	5.93E-04	
						FDR meta		2.22E-02	
rs9611588	22	41823803	TEF	A/G	0.4	Discovery	1.12 (1.01, 1.24)	3.11E-02	0.136
						Validation	1.05 (1.01, 1.10)	1.12E-02	
						Mata	1.06 (1.02, 1.10)	1.28E-03	
						FDR meta		3.85E-02	
rs1883828	22	41739370	TEF	T/C	0.41	Discovery	1.13 (1.02, 1.25)	1.57E-02	0.075
						Validation	1.05 (1.00, 1.10)	1.72E-02	
						Mata	1.06 (1.02, 1.10)	1.41E-03	
						FDR meta		3.98E-02	
rs2073167	22	41791536	TEF	C/T	0.41	Discovery	1.11 (1.00, 1.23)	3.41E-02	0.139
						Validation	1.05 (1.01, 1.10)	1.50E-02	
						Mata	1.06 (1.02, 1.10)	1.88E-03	
						FDR meta		4.71E-02	
rs9611579	22	41800882	TEF	A/G	0.41	Discovery	1.10 (1.00, 1.22)	4.70E-02	0.18
						Validation	1.05 (1.01, 1.10)	1.34E-02	
						Mata	1.05 (1.02, 1.10)	2.01E-03	
						FDR meta		4.77E-02	
rs1589701	15	61272284	RORA	A/G	0.44	Discovery	0.90 (0.82, 0.99)	4.53E-02	0.162
						Validation	0.95 (0.91, 0.99)	2.21E-02	
						Mata	0.94 (0.91, 0.98)	3.40E-03	
						FDR meta		6.97E-02	

Table 2. Significant circadian pathway-related SNPs in the discovery, validation, and meta-analysis. The P-values in the discovery, validation, and meta are original values. The FDR meta P-values are adjusted using the false discovery rate (FDR) method to account for multiple testing.

0.94, 95%CI: 0.91–0.98, P_{meta} = 3.40e-3). The SNP in the *TEF* gene with the strongest association with NSCLC susceptibility was rs9611588 (OR_{meta} = 1.06, 95%CI: 1.02–1.10, P_{meta} = 1.28e-3). There was no heterogeneity in the SNPs between the two groups ($P_{heterogeneity}$ > 0.05).

We further replicated the eight identified SNPs in European populations to see whether these SNPs are specific to East Asian populations. To this end, we used summary statistics from the FinnGen cohort. We found that only *CUL1* rs78524436 was still significantly associated with susceptibility to lung cancer in the same direction as our results ($P < 0.05$, Supplementary Table 3).

Conditional analysis

To exclude the recurrence of two or more mutually explainable SNPs due to highly linked disequilibrium, we performed a GCTA-COJO conditional analysis of eight candidate SNPs associated with NSCLC susceptibility. The final screening results showed that *FBXL21* rs2069868, *CUL1* rs78524436, *RORA* rs1589701, *CSNK1D* rs147316973, and *TEF* rs9611588 were independently correlated with NSCLC susceptibility ($P < 0.05$, Table 3).

Gene-based analysis

To aggregate the associations of multiple SNPs in the vicinity of candidate genes revealing consistent results in both the discovery and validation phases, we performed gene-based analysis using MAGMA. Although none surpassed the significance threshold after adjusting for multiple tests, *CUL1* were nominally associated with lung cancer case-control status ($P < 0.05$, Supplementary Table 5).

eQTL mapping

To further explore the relationship between target SNPs and their corresponding gene expression levels, we reviewed the GTEx and QTLbase database results. The GTEx eQTL result indicated a significant downregulation

SNP	Gene	Region	Alle	OR(95%CI)	*P*
rs2069868	FBXL21	5q31.1	A/G	0.94 (0.90, 0.97)	4.78E-04
rs78524436	CUL1	7q36.1	A/G	1.07 (1.03, 1.12)	8.05E-05
rs1589701	RORA	15q22.2	A/G	0.97 (0.95, 0.99)	3.40E-03
rs147316973	CSNK1D	17q25.3	T/C	0.89 (0.83, 0.95)	5.94E-04
rs9611588	TEF	22q13.2	A/G	1.29 (1.24, 1.35)	<1.00E-8

Table 3. Conditional analysis of candidate SNPs and NSCLC susceptibility.

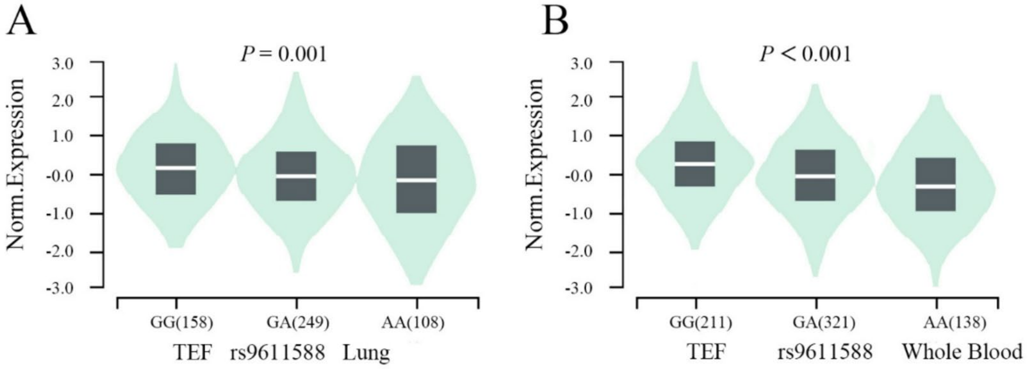

Fig. 2. The GTEx eQTL results of TEF rs9611588. The GTEx eQTL results indicated a significant downregulation of TEF expression with increased rs9611588 effect allele A in the lung (**A**) and blood samples (**B**).

of TEF expression in lung and blood tissue by the A allele of *TEF* rs9611588 ($P = 0.001$ for lung tissue; $P < 0.001$ for blood tissue). With the increase of rs9611588 effect allele A, the expression level of *TEF* in lung tissue (Fig. 2a) and blood samples (Fig. 2b) showed a decreasing trend. *CUL1* rs78524436 was closely associated with *CUL1* expression level in blood samples but was not significant in lung tissue (Figure S2). There was no significant correlation between *RORA* rs1589701 and *RORA* expression in lung tissue and blood samples (Figure S3). *FBXL21* rs2069868 and *CSNK1D* rs147316973 have no corresponding information in GTEx. In the QTLbase database, *RORA* rs1589701 and *TEF* rs9611588 were associated with expression levels of *RORA* and *TEF* in adipose, brain, or other tissues, respectively ($P < 0.01$, Fig. 3). *FBXL21* rs2069868, *CUL1* rs78524436, and *CSNK1D* rs147316973 do not have corresponding information. These results suggest that *RORA* rs1589701 and *TEF* rs9611588 regulate the expression of *RORA* and *TEF*, respectively.

Gene expression analysis

To explore the potential relationship between the expression level of candidate genes with eQTL effect and NSCLC risk, we performed differential expression analysis and survival analysis using the transcriptome data from SAHZU. These results were validated in an additional 1008 NSCLC patients from TCGA. For lung cancer samples from SAHZU, clinical and demographic information of the RNA sequencing population is provided in Supplementary Table 4.

Consistent differential gene expression results were observed from both SAHZU-NSCLC and TCGA-NSCLC. *RORA* and *TEF* showed lower expression levels in tumor tissues compared to normal tissues ($P < 0.001$). In comparison, the expression levels of *RORA* and *TEF* in early-stage I /II were lower than those in late-stage III /IV tumors ($P < 0.001$, Fig. 4).

In the survival analysis, we also observed that *RORA* and *TEF* expression were positively correlated with better survival in SAHZU-NSCLC patients. In contrast, only high RORA expression was significantly associated with overall survival rate ($P < 0.05$, Figs. 5 and 6). After adjusting for gender, age, smoking, and clinical stage, patients with high *RORA* expression or *TEF* expression still had higher clinical survival rates than those with low expression (HR_{RORA} = 0.61, $95\%CI_{RORA}$: 0.38–0.96; HR_{TEF} = 0.37, $95\%CI_{TEF}$: 0.22–0.60). Patients with high *RORA* expression (HR = 0.69, 95%CI: 0.49–0.98) or high *TEF* expression (HR = 0.54, 95%CI: 0.38–0.77) had a lower risk of disease progression than those with low expression levels.

Discussion

This study systematically assessed the association between genetic variants within circadian rhythm pathway genes and NSCLC risk in a large Chinese population and obtained validation in the BBJ data. We employed whole-genome sequencing and identified 888 common genetic variants with significant associations, with eight of these SNPs across five genes (*CUL1*, *FBXL21*, *CSNK1D*, *TEF*, and *RORA*) confirmed in the validation populations and meta-analysis. Additionally, we reviewed the results in the GTEx and QTLbase databases, which suggested that *RORA* rs1589701 and *TEF* rs9611588 regulate the expression of *RORA* and *TEF*, respectively.

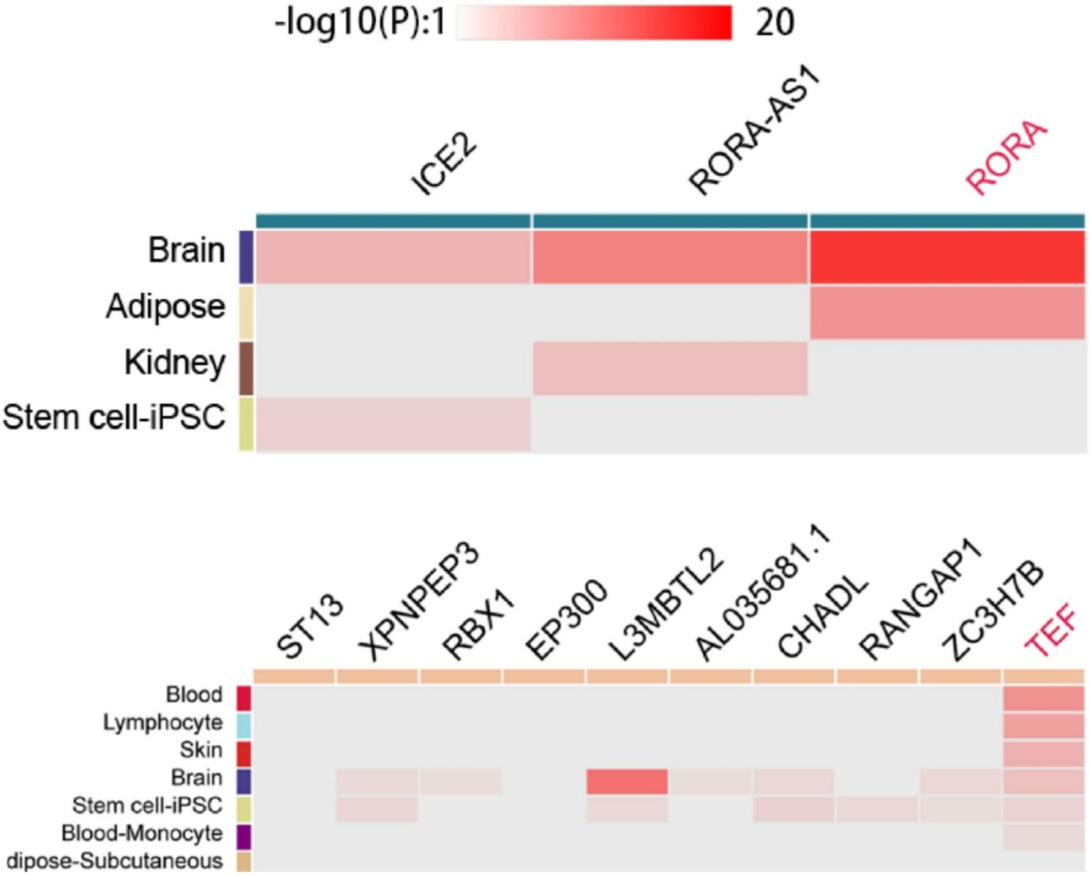

Fig. 3. The QTLbase eQTL results of TEF rs9611588 and RORA rs1589701. The QTLbase eQTL results suggest that RORA rs1589701 and TEF rs9611588 regulate the expression of RORA and TEF, respectively.

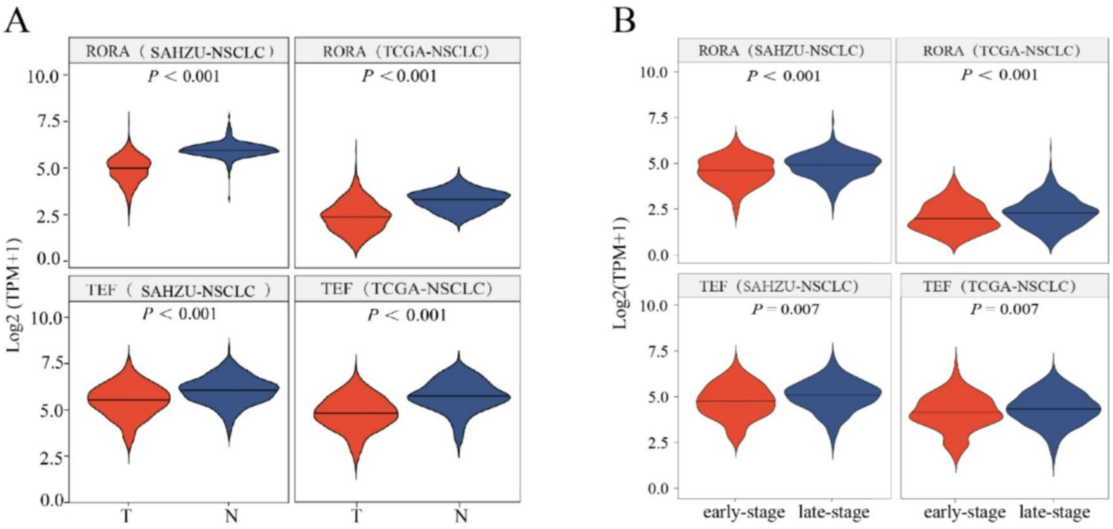

Fig. 4. Differential gene expression analysis of RORA and TEF in SAHZU and TCGA cohorts. Wilcoxon rank-sum test was used to compare the gene expression level between tumor and normal tissues (**A**) and between different stages (**B**).

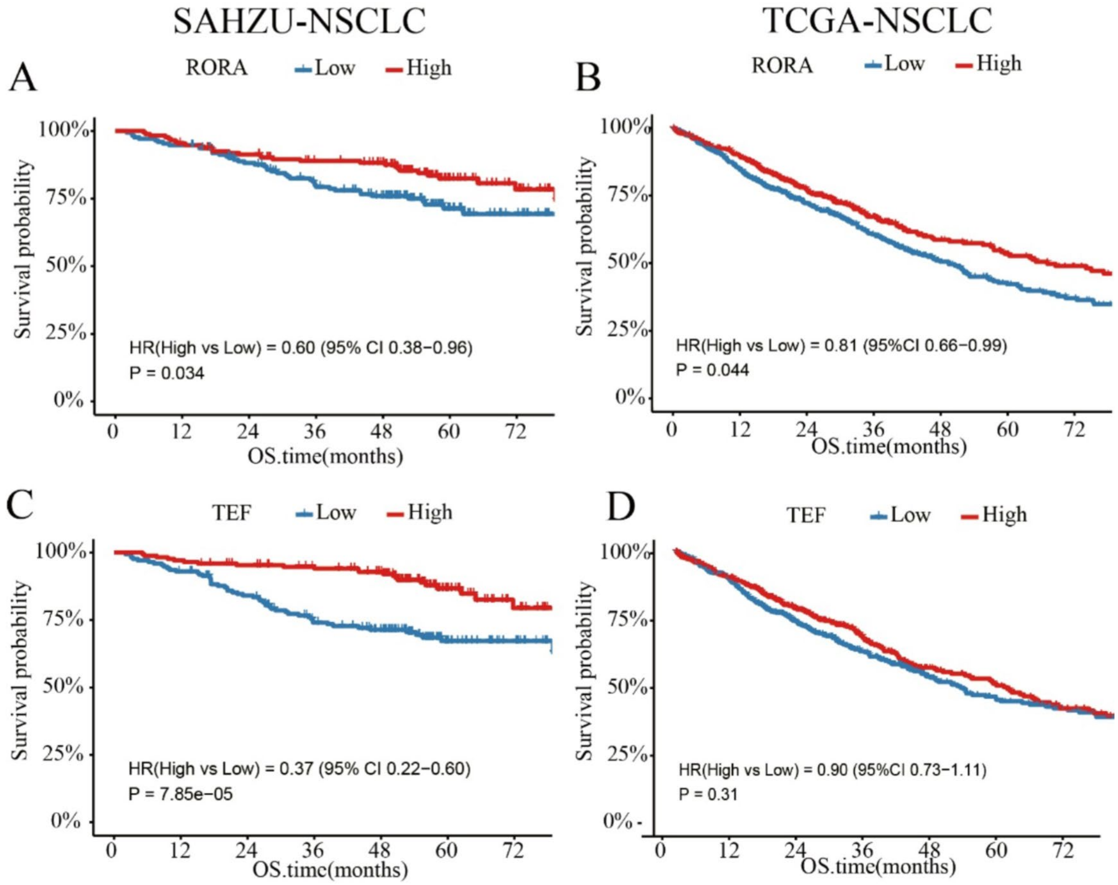

Fig. 5. Overall survival analysis of RORA and TEF in SAHZU and TCGA cohorts. Kaplan-Meier survival plots for two patient subgroups divided by the median gene expression values of RORA (**A**, in SAHZU; **B**, in TCGA) and TEF (**C**, in SAHZU; **D**, in TCGA). OS: overall survival.

Results of our transcriptome data from SAHZU revealed that *RORA* and *TEF* showed lower expression levels in tumor tissues ($P < 0.001$) and poorer survival was observed in patients with lower *RORA* and *TEF* expressions. Overall, our study assessed the association between circadian rhythm genetic polymorphisms and NSCLC risk in the East Asian population, which underscores the potential role of circadian rhythm in NSCLC carcinogenesis and prognosis.

Epidemiological evidence suggests that disruption of circadian rhythms may increase susceptibility to various cancers, including those of the breast, lung, prostate, and colorectum[47–50]. An investigation into occupational lung cancer among U.S. women highlighted the elevated risk associated with night-shift work in nursing, even after adjusting for smoking habits[51]. Animal studies[52] have further validated the impact of genetic and environmental disruption of circadian rhythms on lung cancer development[53], employing models to study the effects of circadian misalignment on tumor growth and proliferation[53]. Furthermore, the circadian rhythm pathway interacts with several signal transduction pathways, including MAPK, PI3K-Akt, and Wnt, modulating their activity and affecting downstream cellular outcomes[54–57]. Circadian genes are implicated in lung carcinogenesis through diverse mechanisms, including regulation of the c-myc oncogene, metastatic factors, immune responses, and cell cycle regulation[58–60], suggesting their utility as biomarkers for lung cancer. Notably, targeting circadian genes has shown promise in tumor suppression. For instance, *RORC* inhibitors have demonstrated efficacy in pancreatic cancer[61], while the role of *RORA* in inducing apoptosis in lung epithelial cells warrants further investigation[64].

Specifically, our investigations revealed that rs1589701 within *RORA* played a prominent role in NSCLC. In the association analysis, the A allele of *RORA* rs1589701 was associated with reduced susceptibility to NSCLC (OR_{meta} = 0.94, 95%CI: 0.91–0.98, P_{meta} = 3.40e-3), which was consistent in the discovery, validation, the meta-analysis, and conditional analysis. Our results supported that *RORA* rs1589701 regulated the expression of *RORA*, whereas *RORA* under-expression was related to NSCLC carcinogenesis and worse survival. *RORA*, a nuclear hormone receptor gene, spanning a 730 kb region with 15 exons on chromosome 15q22.2, has been identified as a potential tumor suppressor[62]. It regulates *SEMA3F*, a tumor-suppressive factor, and plays a crucial role in cellular stress responses. Its inactivation in multiple tumor types underscores its significance in cancer[63]. Therefore, our findings align with previous research indicating *RORA*'s association with lung cancer risk[13], reinforcing the potential of circadian rhythm genetic polymorphisms as critical determinants of NSCLC susceptibility.

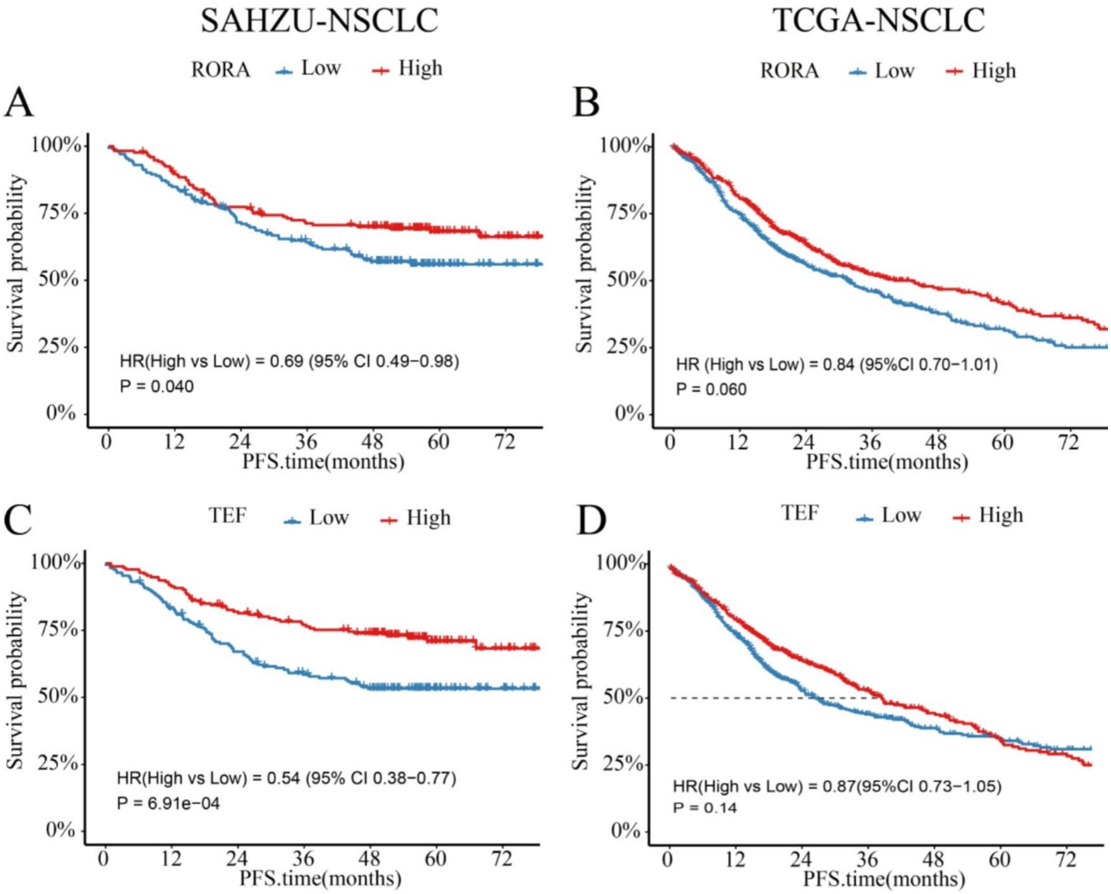

Fig. 6. Progress-free survival analysis of RORA and TEF in SAHZU and TCGA cohorts. Kaplan-Meier survival plots for two patient subgroups divided by the median gene expression values of RORA (**A**, in SAHZU; **B**, in TCGA) and TEF (**C**, in SAHZU; **D**, in TCGA). PFS: progress-free survival.

Furthermore, this study also identified additional SNPs in *TEF* associated with NSCLC risk, marking the new findings of these associations in the context of lung cancer. Following this, we reviewed the results of transcriptome-wide association studies (TWAS) that were conducted to identify genes implicated in the pathogenesis of lung cancer by correlating genotype and gene expression with phenotypic traits. A Japanese study[64] involved integrating lung cancer GWAS summary results with gene expression profiles from 28 immune cells highlighted that *RORA* and *TEF* were significantly correlated with lung cancer risk across immune cells. Notably, lower expression of *TEF* (Z = -3.20, P=1.41e-3) in mDC and Tfh cells was predicted to increase lung cancer risk. Conversely, higher expression of *RORA* (Z=2.72, P=6.62e-3) in Naive B cells was predicted to elevate lung cancer risk. These findings necessitate experimental validation through further studies.

This study's strengths include a large patient cohort, a targeted pathway-based approach encompassing major circadian rhythm genes, and the novelty of exploring these associations in East Asian populations. Nonetheless, there are several limitations in our study. First, while the statistical evidence for circadian rhythmin in NSCLC has been demonstrated, further extensive functional studies are required to elucidate the precise mechanisms. Second, the focus on common variants may overlook the potential impact of rare genetic variations. Future research on rare genetic variations of circadian clock genes may unveil new biomarkers and therapeutic targets for lung cancer.

Conclusion

In this study, we undertook a thorough examination of potential susceptibility loci associated with lung cancer within the circadian rhythm pathway by employing whole-genome sequencing in East Asian cohorts. Our research identified significant correlations between lung cancer risk and specific single nucleotide polymorphisms (SNPs) in circadian rhythm-related genes, including *CUL1*, *FBXL21*, *CSNK1D*, *TEF*, and *RORA*. These correlations were further corroborated in an independent case-control study, offering strong evidence of their connection to lung cancer susceptibility. Additionally, our comprehensive analysis encompassing quantitative trait loci (QTL), transcriptome-wide association studies (TWAS), gene-based, and transcriptomic evaluations, underscored the role of these genes in lung cancer's etiology, suggesting their utility as potential biomarkers and their influence on the tumor's immune microenvironment. Despite the progress made, our findings necessitate further investigation to solidify and expand upon these initial observations. To confirm the relationships we observed, it is imperative to conduct Large-scale genome-wide association studies (GWAS) that include diverse and geographically varied populations. Furthermore, detailed functional studies are essential to decode the exact molecular pathways through which the circadian rhythm genetic variants affect lung cancer risk. Such endeavors will significantly advance our comprehension of how genetic variations interact with cancer pathways, leading to the development of tailored prevention and treatment modalities for lung cancer in East Asian populations, thereby contributing to the broader field of personalized medicine.

Data availability

The result of GWAS for the NSCLC have been deposited in the National Genomics Data Center (NGDC, https://www.cncb.ac.cn/?lang=en) with OMIX ID: OMIX005344. The raw sequencing data from the study have been deposited in the database of the Data Centre of School of Public Health, Zhejiang University, following the regulations of the Human Genetic Resources Administration of China (HGRAC). The data are available under controlled access after approval by the research team for educational and research purposes only.

Received: 10 October 2024; Accepted: 25 February 2025
Published online: 06 March 2025

References

1. Sung, H. et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71** (3), 209–249 (2021).
2. Siegel, R. L. et al. Cancer statistics, 2023. *CA Cancer J. Clin.* **73** (1), 17–48 (2023).
3. Malhotra, J. et al. Risk factors for lung cancer worldwide. *Eur. Respir J.* **48** (3), 889–902 (2016).
4. Wynder, E. L. Tobacco as a cause of lung cancer: some reflections. *Am. J. Epidemiol.* **146** (9), 687–694 (1997).
5. Stayner, L. et al. Lung cancer risk and workplace exposure to environmental tobacco smoke. *Am. J. Public. Health.* **97** (3), 545–551 (2007).
6. Lorenzo Bermejo, J. & Hemminki, K. Familial lung cancer and aggregation of smoking habits: a simulation of the effect of shared environmental factors on the Familial risk of cancer. *Cancer Epidemiol. Biomarkers Prev.* **14** (7), 1738–1740 (2005).
7. Smolle, E. & Pichler, M. Non-Smoking-Associated lung cancer: A distinct entity in terms of tumor biology, patient characteristics and impact of hereditary Cancer predisposition. *Cancers (Basel)* **11**(2), 204. https://doi.org/10.3390/cancers11020204 (2019).
8. Bossé, Y. & Amos, C. I. A decade of GWAS results in lung Cancer. *Cancer Epidemiol. Biomarkers Prev.* **27** (4), 363–379 (2018).
9. Shu, X. et al. Potential susceptibility loci identified for renal cell carcinoma by targeting Obesity-Related genes. *Cancer Epidemiol. Biomarkers Prev.* **26** (9), 1436–1442 (2017).
10. Li, R. et al. Hypoxia pathway genetic variants predict survival of non-small-cell lung cancer patients receiving platinum-based chemotherapy. *Carcinogenesis* **38** (4), 419–424 (2017).
11. Coscio, A. et al. Genetic variants of the Wnt signaling pathway as predictors of recurrence and survival in early-stage non-small cell lung cancer patients. *Carcinogenesis* **35** (6), 1284–1291 (2014).
12. Lin, J. et al. Systematic evaluation of apoptotic pathway gene polymorphisms and lung cancer risk. *Carcinogenesis* **33** (9), 1699–1706 (2012).
13. Mocellin, S. et al. Circadian pathway genetic variation and cancer risk: evidence from genome-wide association studies. *BMC Med.* **16** (1), 20 (2018).
14. Wendeu-Foyet, M. G. et al. Circadian genes and risk of prostate cancer: findings from the EPICAP study. *Int. J. Cancer*. **145** (7), 1745–1753 (2019).
15. Hoffman, A. E. et al. CLOCK in breast tumorigenesis: genetic, epigenetic, and transcriptional profiling analyses. *Cancer Res.* **70** (4), 1459–1468 (2010).
16. Zhu, Y. et al. Non-synonymous polymorphisms in the circadian gene NPAS2 and breast cancer risk. *Breast Cancer Res. Treat.* **107** (3), 421–425 (2008).
17. Gery, S. et al. The circadian gene per1 plays an important role in cell growth and DNA damage control in human cancer cells. *Mol. Cell.* **22** (3), 375–382 (2006).
18. Takahashi, J. S. Transcriptional architecture of the mammalian circadian clock. *Nat. Rev. Genet.* **18** (3), 164–179 (2017).
19. Eckel-Mahan, K. & Sassone-Corsi, P. Metabolism and the circadian clock converge. *Physiol. Rev.* **93** (1), 107–135 (2013).
20. Masri, S. & Sassone-Corsi, P. The emerging link between cancer, metabolism, and circadian rhythms. *Nat. Med.* **24** (12), 1795–1803 (2018).
21. Asher, G. & Sassone-Corsi, P. Time for food: the intimate interplay between nutrition, metabolism, and the circadian clock. *Cell* **161** (1), 84–92 (2015).
22. Marcheva, B. et al. Circadian clocks and metabolism. *Handb. Exp. Pharmacol.* **217**, 127–155 (2013).
23. Koike, N. et al. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science* **338** (6105), 349–354 (2012).
24. Patke, A., Young, M. W. & Axelrod, S. Molecular mechanisms and physiological importance of circadian rhythms. *Nat. Rev. Mol. Cell. Biol.* **21** (2), 67–84 (2020).
25. Bass, J. & Takahashi, J. S. Circadian integration of metabolism and energetics. *Science* **330** (6009), 1349–1354 (2010).
26. Scheer, F. A. et al. Adverse metabolic and cardiovascular consequences of circadian misalignment. *Proc. Natl. Acad. Sci. U S A*. **106** (11), 4453–4458 (2009).
27. Lim, Y. C. et al. Association between night-shift work, sleep quality and metabolic syndrome. *Occup. Environ. Med.* **75** (10), 716–723 (2018).
28. Vetter, C. et al. Night shift work, genetic risk, and type 2 diabetes in the UK biobank. *Diabetes Care*. **41** (4), 762–769 (2018).
29. Sooriyaarachchi, P. et al. Shift work and the risk for metabolic syndrome among healthcare workers: A systematic review and meta-analysis. *Obes. Rev.* **23** (10), e13489 (2022).
30. Sancar, A. et al. Circadian clock control of the cellular response to DNA damage. *FEBS Lett.* **584** (12), 2618–2625 (2010).
31. Kondratov, R. V. & Antoch, M. P. Circadian proteins in the regulation of cell cycle and genotoxic stress responses. *Trends Cell. Biol.* **17** (7), 311–317 (2007).
32. Koritala, B. S. C. et al. Night shift schedule causes circadian dysregulation of DNA repair genes and elevated DNA damage in humans. *J. Pineal Res.* **70** (3), e12726 (2021).
33. Liu, F. et al. Association between three polymorphisms in BMAL1 genes and risk of lung Cancer in a Northeast Chinese population. *DNA Cell. Biol.* **38** (12), 1437–1443 (2019).
34. Couto, P. et al. Association between CLOCK, PER3 and CCRN4L with non–small cell lung cancer in Brazilian patients. *Mol. Med. Rep.* **10** (1), 435–440 (2014).
35. Sakaue, S. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53** (10), 1415–1424 (2021).
36. Jiang, L. et al. A generalized linear mixed model association tool for biobank-scale data. *Nat. Genet.* **53** (11), 1616–1621 (2021).

37. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26** (17), 2190–2191 (2010).
38. Wang, Q. et al. Genetic associations of T cell cancer immune response with tumor aggressiveness in localized prostate cancer patients and disease reclassification in an active surveillance cohort. *Oncoimmunology* **8** (1), e1483303 (2019).
39. Wang, Q. et al. Genetic associations of T cell cancer immune response-related genes with T cell phenotypes and clinical outcomes of early-stage lung cancer. *J. Immunother Cancer* **8**(2), e000336. https://doi.org/10.1136/jitc-2019-000336 (2020).
40. Allada, R. & Bass, J. Circadian mechanisms in medicine. *N Engl. J. Med.* **384** (6), 550–561 (2021).
41. Scheiermann, C. et al. Clocking in to immunity. *Nat. Rev. Immunol.* **18** (7), 423–437 (2018).
42. Panda, S. Circadian physiology of metabolism. *Science* **354** (6315), 1008–1015 (2016).
43. Musiek, E. S. & Holtzman, D. M. Mechanisms linking circadian clocks, sleep, and neurodegeneration. *Science* **354** (6315), 1004–1008 (2016).
44. Honma, S. et al. Dec1 and Dec2 are regulators of the mammalian molecular clock. *Nature* **419** (6909), 841–844 (2002).
45. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501** (7468), 506–511 (2013).
46. Zheng, Z. et al. QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic Acids Res.* **48** (D1), D983–d991 (2020).
47. Travis, R. C. et al. Night shift work and breast Cancer incidence: three prospective studies and Meta-analysis of published studies. *J. Natl. Cancer Inst.* **108**(12), djw169. https://doi.org/10.1093/jnci/djw169 (2016).
48. Schernhammer, E. S. et al. Rotating night-shift work and lung cancer risk among female nurses in the united States. *Am. J. Epidemiol.* **178** (9), 1434–1441 (2013).
49. Behrens, T. et al. Shift work and the incidence of prostate cancer: a 10-year follow-up of a German population-based cohort study. *Scand. J. Work Environ. Health.* **43** (6), 560–568 (2017).
50. Papantoniou, K. et al. Rotating night shift work and colorectal cancer risk in the nurses' health studies. *Int. J. Cancer.* **143** (11), 2709–2717 (2018).
51. Robinson, C. F. et al. Occupational lung cancer in US women, 1984–1998. *Am. J. Ind. Med.* **54** (2), 102–117 (2011).
52. Shi, Y. et al. Retinoic acid-related orphan receptor-α is induced in the setting of DNA damage and promotes pulmonary emphysema. *Am. J. Respir Crit. Care Med.* **186** (5), 412–419 (2012).
53. Papagiannakopoulos, T. et al. Circadian rhythm disruption promotes lung tumorigenesis. *Cell. Metab.* **24** (2), 324–331 (2016).
54. Goldsmith, C. S. & Bell-Pedersen, D. Diverse roles for MAPK signaling in circadian clocks. *Adv. Genet.* **84**, 1–39 (2013).
55. Beker, M. C. et al. Interaction of melatonin and Bmal1 in the regulation of PI3K/AKT pathway components and cellular survival. *Sci. Rep.* **9** (1), 19082 (2019).
56. Cui, L. et al. Circadian rhythm-associated Rev-erbα modulates polarization of decidual macrophage via the PI3K/Akt signaling pathway. *Am. J. Reprod. Immunol.* **86** (3), e13436 (2021).
57. Boucsein, A. et al. Photoperiodic and diurnal regulation of WNT signaling in the arcuate nucleus of the female Djungarian hamster, Phodopus sungorus. *Endocrinology* **157** (2), 799–809 (2016).
58. Partch, C. L., Green, C. B. & Takahashi, J. S. Molecular architecture of the mammalian circadian clock. *Trends Cell. Biol.* **24** (2), 90–99 (2014).
59. Cederroth, C. R. et al. Medicine in the fourth dimension. *Cell. Metab.* **30** (2), 238–250 (2019).
60. Gaucher, J., Montellier, E. & Sassone-Corsi, P. Molecular Cogs: interplay between circadian clock and cell cycle. *Trends Cell. Biol.* **28** (5), 368–379 (2018).
61. Lytle, N. K. et al. A multiscale map of the stem cell state in pancreatic adenocarcinoma. *Cell* **177** (3), 572–586e22 (2019).
62. Xiong, G. et al. RORα suppresses breast tumor invasion by inducing SEMA3F expression. *Cancer Res.* **72** (7), 1728–1739 (2012).
63. Zhu, Y. et al. RORA, a large common fragile site gene, is involved in cellular stress response. *Oncogene* **25** (20), 2901–2908 (2006).
64. Ota, M. et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184** (11), 3006–3021e17 (2021).

Author contributions

Concept and design: Xifeng Wu; Acquisition of data: Xifeng Wu; Statistical analysis: Xiaohang Xu and Luopiao Xu; Interpretation of data: Xiaohang Xu, Luopiao Xu, Zeyong Lang, Xue Li, and Xifeng Wu; Drafting of the manuscript: Xiaohang Xu, Luopiao Xu, Xifeng Wu; Critical revision of the manuscript for important intellectual content: All authors; Obtained funding: Xifeng Wu; Administrative, technical, or material support: Xifeng Wu; Supervision: Xifeng Wu.

Funding

This study was supported by Healthy Zhejiang One Million People Cohort (K20230085), Key Research and Development Program of Zhejiang Province (2020C03002) and the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (2019R01007). The funders had no role in the conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-92083-9.

Correspondence and requests for materials should be addressed to X.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.