

# Genetic-informed alternative RNA splicing serves an essential role in carcinogenesis and prognosis of non-small-cell lung cancer

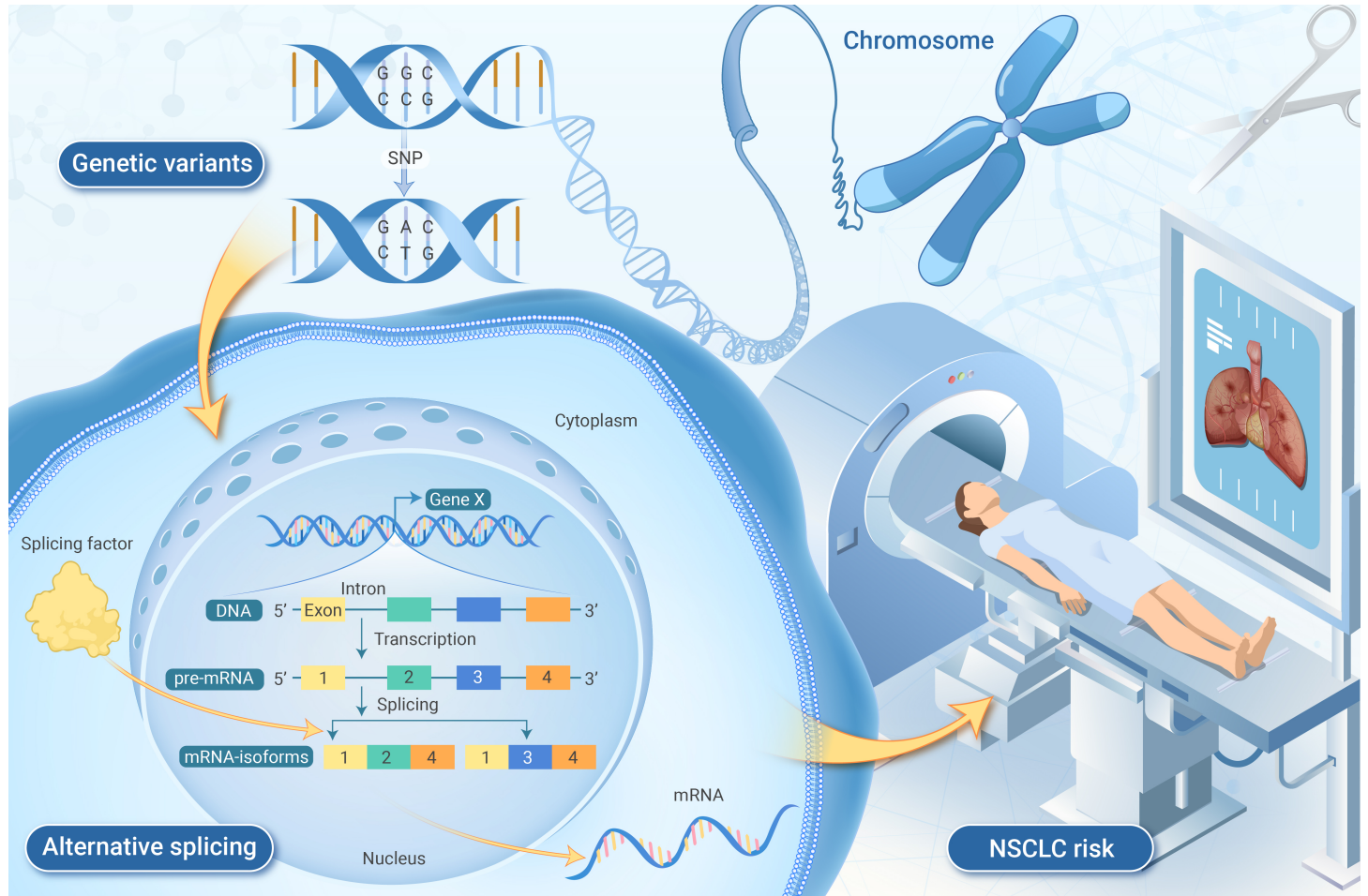
Hanyi Zhou,<sup>1,2</sup> Ming Wu,<sup>3</sup> Kui Wu,<sup>4,5</sup> Qilong Tan,<sup>1,2</sup> Xiaohang Xu,<sup>1,2</sup> Mengying Wang,<sup>1,2</sup> Liyang Sun,<sup>6</sup> Yuanxing Tu,<sup>6</sup> Weifang Zheng,<sup>7</sup> Yun Zhu,<sup>1,2</sup> Huakang Tu,<sup>1,2</sup> Dan Zhou,<sup>1,2,\*</sup> and Xifeng Wu<sup>1,2,8,\*</sup>

\*Correspondence: xifengw@zju.edu.cn (X.W.); danzhou@zju.edu.cn (D.Z.)

Received: September 20, 2024; Accepted: January 7, 2025; Published Online: January 16, 2025; <https://doi.org/10.59717/j.xinn-med.2025.100111>

© 2025 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

## GRAPHICAL ABSTRACT



## PUBLIC SUMMARY

- Lung cancer is the most commonly diagnosed malignancy both in China and globally.
- RNA splicing plays a substantial role in influencing cancer heritability.
- Genetic-based splicing dysregulation significantly contributes to the carcinogenesis and prognosis of lung cancer.

# Genetic-informed alternative RNA splicing serves an essential role in carcinogenesis and prognosis of non-small-cell lung cancer

Hanyi Zhou,<sup>1,2</sup> Ming Wu,<sup>3</sup> Kui Wu,<sup>4,5</sup> Qilong Tan,<sup>1,2</sup> Xiaohang Xu,<sup>1,2</sup> Mengying Wang,<sup>1,2</sup> Liyang Sun,<sup>6</sup> Yuanxing Tu,<sup>6</sup> Weifang Zheng,<sup>7</sup> Yun Zhu,<sup>1,2</sup> Huakang Tu,<sup>1,2</sup> Dan Zhou,<sup>1,2,\*</sup> and Xifeng Wu<sup>1,2,8,\*</sup>

<sup>1</sup>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

<sup>2</sup>Zhejiang Key Laboratory of Intelligent Preventive Medicine, Hangzhou 310058, China

<sup>3</sup>Department of Thoracic Surgery, Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China

<sup>4</sup>HIM-BGI Omics Center, Zhejiang Cancer Hospital, Hangzhou Institute of Medicine (HIM), Chinese Academy of Sciences (CAS), BGI Research, Hangzhou 310018, China

<sup>5</sup>Guangdong Provincial Key Laboratory of Human Disease Genomics, Shenzhen Key Laboratory of Genomics, BGI Research, Shenzhen 518083, China

<sup>6</sup>The Second Affiliated Hospital of Zhejiang University School of Medicine, Lanxi Branch (Lanxi People's Hospital), Lanxi 321100, China

<sup>7</sup>Lanxi Hospital of Traditional Chinese Medicine, Lanxi 321102, China

<sup>8</sup>National Institute for Data Science in Health and Medicine, Zhejiang University, Hangzhou 310058, China

\*Correspondence: xifengw@zju.edu.cn (X.W.); danzhou@zju.edu.cn (D.Z.)

Received: September 20, 2024; Accepted: January 7, 2025; Published Online: January 16, 2025; <https://doi.org/10.59717/j.xinn-med.2025.100111>

© 2025 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Zhou H., Wu M., Wu K., et al. (2025). Genetic-informed alternative RNA splicing serves an essential role in carcinogenesis and prognosis of non-small-cell lung cancer. *The Innovation Medicine* 3:100111.

Dysregulated RNA splicing is a post-transcriptional molecular feature that significantly influences tumor progression and prognosis. However, the role of alternative splicing in the development of non-small cell lung cancer (NSCLC) within the Chinese population remains poorly understood. In this study, we investigated the genetic regulation of splicing in 245 tumor and 297 normal lung tissue samples from Chinese NSCLC patients. By integrating splicing data with a meta-analyzed genome-wide association study (GWAS) for NSCLC in East Asians (7,035 cases and 185,413 controls), we identified 14 novel NSCLC-associated splicing events ( $FDR < 0.05$ ) through a splicing transcriptome-wide association study (spTWAS). Additionally, we validated the involvement of the splicing gene *FARP1* and the *EIF3* family, both of which have been associated with NSCLC risk. By combining the results of differential splicing analysis and spTWAS, followed by colocalization analysis and putative splicing factor predictions, we highlighted the critical roles of splicing events in *TP63* (1st exon skipping) and *TPM1* (6th exons mutually exclusive) in NSCLC, bridging the missing biology between SNP-NSCLC association. Furthermore, we underscored several splicing events in genes including *ILK*, which were also associated with NSCLC prognosis. In conclusion, this study elucidated the genetic architecture of splicing in lung tissues and revealed the significant contribution of splicing dysregulation to the carcinogenesis and prognosis of NSCLC.

## INTRODUCTION

Lung cancer is the most frequently diagnosed malignancy in China and worldwide, representing a leading cause of cancer-related mortality and posing significant public health challenges.<sup>1-3</sup> Non-small cell lung cancer (NSCLC) accounts for approximately 85% of all lung cancer cases.<sup>4</sup> Genetic factors contribute significantly to the etiology of lung cancer,<sup>5-7</sup> and genome-wide association studies (GWAS) have identified numerous loci associated with NSCLC. However, most of these loci lack clear molecular mechanisms.<sup>8,9</sup> Given that over 90% of these loci are located in non-coding regions, the potential mediating role of gene transcriptional regulation in SNP-NSCLC associations has been increasingly explored.<sup>10,11</sup>

Expression quantitative trait loci (eQTLs) have been extensively studied to elucidate the associations between genetic variants and diseases.<sup>12,13</sup> However, empirical data suggest that eQTLs account for only a limited proportion of cancer susceptibility loci,<sup>14</sup> leaving much of the underlying biology unexplained (missing biology), particularly within the context of non-coding regulation.<sup>15</sup> Recent studies have indicated that a substantial portion of cancer heritability may be attributable to biological processes beyond the genetic regulation of gene expression. For example, alternative splicing, which can result in the absence of key domains encoded by skippable exons,<sup>16</sup> highlights the complementary role of splicing in gene regulation.<sup>17,18</sup>

Alternative splicing is a critical component of post-transcriptional regulation, enabling pre-mRNA molecules to produce various mRNA isoforms. This mechanism affects more than 95% of human genes.<sup>19</sup> Abnormal splicing can

disrupt essential biological processes, forming the basis of various pathological conditions, including carcinogenesis and prognosis.<sup>20,21</sup> Aberrant splicing in tumors can arise from multiple factors, including (1) alterations in the expression or function (due to somatic mutations) of splicing factors and (2) genetic variants that affect splicing sequences.<sup>22</sup> These mechanisms underline the importance of splicing regulation in cancer biology.

The role of alternative splicing in SNP-disease associations has been explored through the mapping of splicing quantitative trait loci (sQTLs).<sup>23-25</sup> Additionally, disease-associated splicing events have been identified using a splicing-based transcriptome-wide association study (spTWAS).<sup>23,26,27</sup> Transcriptome-wide association studies (TWAS) are widely used frameworks for projecting SNP-level signals from GWAS onto molecular-trait associations. Recently, Wang et al. successfully mapped lung cancer-associated splicing events using spTWAS in Han Chinese.<sup>27</sup> However, the reference panel for the splicing prediction model in that study was relatively small ( $N = 116$ ), which limited the number of predictable splicing events.

Leveraging data from the GTEx project, we hypothesized that using a larger reference panel could significantly enhance the predictability of splicing events and increase the likelihood of identifying disease-associated molecular mechanisms.<sup>28</sup> TWAS approaches can be underpowered due to the limited proportion of variance in molecular traits explained by genetic variants.<sup>29,30</sup> Differential splicing analysis between normal and tumor tissues, on the other hand, is highly effective in identifying up- or down regulated events at the PSI (percent-spliced-in) level.<sup>31</sup> However, these events may represent consequences rather than causes of tumorigenesis. Therefore, integrating spTWAS with differential splicing analysis can complement their respective strengths and weaknesses, offering deeper mechanistic insights into NSCLC-associated alternative splicing.

In this study, we conducted a comprehensive investigation to identify NSCLC-associated splicing events. First, we analyzed the genetic architecture of splicing and identified sQTLs in 245 tumor and 297 normal lung tissue samples from NSCLC patients of Chinese descent. Second, we performed differential splicing analysis to compare splicing events between tumor and normal tissues. Third, we developed a genetic variant-based splicing prediction model, which was applied to bridge SNP-NSCLC associations and identify NSCLC-associated splicing events through spTWAS, followed by colocalization analysis. Fourth, we assessed the impact of splicing events on prognosis and investigated potential upstream regulatory mechanisms. Overall, this study aimed to provide a comprehensive evidence chain for the role of alternative splicing events in NSCLC pathogenesis and prognosis.

## MATERIALS AND METHODS

### Cohorts and clinical specimens

A total of 417 patients diagnosed with NSCLC were recruited from the Second Affiliated Hospital of Zhejiang University School of Medicine (SAH), Zhejiang University (Hangzhou, China), between 2015 and 2020. Matched tumor and adjacent normal tissue samples were collected from these

patients, who had not received prior systemic therapy and underwent surgical resection for NSCLC. The demographic characteristics of the participants, abstracted from medical chart records, are summarized in Table S1. Informed consent was obtained from all patients, and the study protocol was approved by the Institutional Review Board of SAH (2020-829).

All specimens were obtained from tissue samples resected during surgery, immediately frozen in liquid nitrogen, and subsequently stored at  $-80^{\circ}\text{C}$ . Genomic DNA and total RNA were extracted from the frozen tissues using the DNA/RNA AllPrep Kit (Qiagen). DNA concentrations were measured using a Qubit Fluorometer (Invitrogen), and DNA integrity was evaluated using the TapeStation system (Agilent Technologies). Total RNA quantity and integrity were assessed with the RNA Nano 6000 Assay Kit on the Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

A study design diagram illustrating the use of patient cohorts and the main procedures for data processing in this study is provided in Figure S1.

### DNA genotyping

The library was prepared using the MGIEasy FS DNA Library Prep Set (cat. 100006987; MGI, China), and whole-genome sequencing (WGS) was conducted on the DIPSEQ platform (BGI, Shenzhen). Prior to imputation, the following quality control criteria were evaluated: (1) contamination rate exceeding 5%; (2) duplication or relatedness greater than 10%; (3) less than 80% of regions covered at  $5\times$  depth; and (4) discrepancies in sex concordance between genetic and reported sex.

The GATK Best Practice pipeline was employed for germline short variant discovery (SNPs and Indels) during sequencing alignment.<sup>32</sup> Joint variant calling was performed using GATK GenotypeGVCFs. Variants exhibiting excessive heterozygosity were excluded. Low-quality variants were filtered based on quality metrics, including QD, MQRankSum, ReadPosRankSum, FS, SOR, and DP, followed by Variant Quality Score Recalibration (VQSRR).

To enhance genotyping accuracy, refinement was conducted using Beagle 5.4.<sup>33,34</sup> Variants with a Hardy-Weinberg equilibrium  $P$ -value ( $P_{\text{HWE}}$ )  $< 1 \times 10^{-6}$  were filtered out after refinement. Additionally, potentially monozygotic twins or duplicate samples were identified and removed.

### RNA sequencing and processing

Ribosomal RNA was removed using the MGIEasy 32 RXN Kit (MGIEasy, China). Subsequently, sequencing libraries suitable for paired-end sequencing were prepared using the MGIEasy Total RNA Library Prep Kit (MGIEasy, China). RNA samples were pooled and sequenced on the MGI HiSeq platform (BGI, Shenzhen), generating 100 bp paired-end reads.

The sequencing quality of each sample was assessed using FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Sample-relatedness and mismatched sequencing datasets from different donors were evaluated using Picard Tools CheckFingerprint (Broad Institute, Cambridge, MA).<sup>35</sup> Ribosomal RNA reads were identified and removed using RiboDetector.<sup>36</sup> Following the hmftools pipeline (<https://github.com/hartwigmedical/hmftools>), RNA-seq alignment was performed with STAR,<sup>37</sup> duplicates were marked using sambamba,<sup>38</sup> and gene counts were generated with Isofox (<https://github.com/hartwigmedical/hmftools/tree/master/isofox>). After quality control, 747 samples were retained for analysis, meeting the criteria of low ribosomal RNA content ( $\leq 30\%$ ) and high sequencing depth ( $\geq 20$  million reads).

### Splicing & expression quantification

Leafcutter v0.2 was utilized to identify splicing events and quantify the intron usage ratio for 346 tumor and 401 normal tissue samples from NSCLC patients.<sup>39</sup> Initially, the qualified BAM files were converted into splicing junction files. Introns were then clustered using the default parameters:  $-m\ 50\ -l\ 500000$ , which retains clusters with at least 50 supported split reads and allows for introns up to 500 kb in length. Introns with minimal variation, those utilized in fewer than 40% of individuals, or those located on non-autosomal chromosomes were filtered out for subsequent analyses.

Isofox v1.6.1 (<https://github.com/hartwigmedical/hmftools/tree/master/isofox>) was employed to generate gene count data and Transcript Per Million (TPM) values for the 747 samples. To adjust TPM values, Isofox excluded certain genes from GC bias calculations and applied normalization factors.

For a given sample, the adjusted TPM was calculated as:  $\text{AdjustedTPM} = \text{rawTPM} \times \text{constant}$ , where the constant was determined by the normalization factor. An expression filter was then applied, retaining genes with an expression value of at least 1 TPM in a minimum of 20% of the samples within the dataset.

### DSG and DEG identification

Differential splicing gene (DSG) analysis was conducted using a Dirichlet multinomial model implemented in leafcutter\_ds.R. Subsequently, leafviz v0.1.0 (<https://leafcutter.shinyapps.io/leafviz/>) was employed to classify detected introns as annotated or cryptic and to generate preliminary visualizations. Potential batch effects were adjusted during the analysis. Significant DSGs were identified based on a threshold of  $FDR < 0.05$ .

Differential expression gene (DEG) analysis was performed using a negative binomial generalized linear model implemented in edgeR.<sup>40</sup> Potential batch effects were similarly adjusted, and significant DEGs were identified at  $FDR < 0.05$ .

All 346 tumor tissue samples and 401 normal tissue samples were included in the DSG and DEG analyses.

### QTL mapping and characterization

Linear regression was conducted to identify genetic variants associated with splicing (sQTLs) and expression (eQTLs) using the R package MatrixE-QTL v2.3.<sup>41</sup> Considering the local regulation of molecular traits, cis-regulation regions were defined with window sizes of 100 kb and 1 Mb.

To account for latent confounders, including potential batch effects, the probabilistic estimation of expression residuals (PEER) method was implemented to estimate PEER factors.<sup>42</sup> Age, Age<sup>2</sup>, sex, the top 20 principal components (PCs), and PEER factors (30 for tumor tissues and 45 for normal tissues) were regressed out to generate splicing/expression residuals for subsequent analyses.

Variants identified with  $FDR < 0.05$  in the association analysis were classified as sQTLs or eQTLs. Introns containing at least one sQTL were defined as slntrons.

Annotation of QTLs was performed using SnpEff v5.2,<sup>43</sup> and these were subsequently mapped to genes. Genes containing sQTLs were categorized as sGenes, while those containing eQTLs were categorized as eGenes.

Functional annotation of splicing-related genes, including sGenes and DSGs, was performed using Gene Ontology (GO) for biological processes and the Kyoto Encyclopedia of Genes and Genomes (KEGG).

Venn diagrams were generated using the R package Venn Diagram (v1.6.0).

### Signal sharing between sQTL and GWAS

To quantify signal sharing between sQTLs and GWAS, the R package qvalue (<https://bioconductor.org/packages/qvalue>) was employed to estimate the proportion of shared signals. GWAS summary statistics were derived by meta-analyzing data from two sources: a GWAS of NSCLC in 13,722 Chinese individuals (2,591 cases and 11,131 controls) and a GWAS of lung cancer from the Biobank Japan (BBJ),<sup>44</sup> totally 7,035 cases and 185,413 controls.

First, tumor-specific and normal-specific sQTLs were categorized into five groups: normal, normal-specific, tumor-normal-shared, tumor-specific, and tumor. Next,  $\pi_0$  was estimated for each group based on the GWAS  $P$ -value distribution of SNP-NSCLC risk associations. The  $\pi_0$  value reflects the proportion of true null  $P$ -values within the distribution. Finally,  $\pi_1 = 1 - \pi_0$  was calculated to estimate the lower bound of true positive associations between sQTLs and NSCLC risk.

### Splice-transcriptome-wide association study (spTWAS)

To identify NSCLC-associated splicing events, we adapted the TWAS framework into a splicing-specific approach, termed spTWAS, using normal lung tissue data. A total of 297 individuals with both DNA genotyping and RNA-seq data were included in the analysis.

First, we reduced the computational burden for model construction by narrowing down splicing events in two steps: (1) Introns with limited evidence of genetic regulation ( $FDR \geq 0.1$  from sQTL mapping in normal lung tissues)



were excluded. (2) Using GCTA,<sup>45</sup> we estimated the cis-heritability of intron usage ratios, adjusting for age, age<sup>2</sup>, sex, the top 20 principal components (PCs), PEER factors, and a cis-window defined as  $\pm 100$  kb around the splice site. Introns with significant cis-heritability ( $P < 0.05$ ) were retained for subsequent analysis.

Second, prediction models for the qualified introns were trained using cis-SNPs. To capture the genetic architecture of splicing comprehensively, we applied four statistical models: single best SNP, best linear unbiased predictor (BLUP), elastic net regression, and least absolute shrinkage and selection operator (LASSO) regression. Prediction performance for each model was evaluated using 5-fold cross-validation, and the model with the highest cross-validated  $R^2$  was selected for the association test.

Third, the splicing-NSCLC association test was conducted using GWAS summary statistics and SNP-SNP covariance matrices estimated from the 1000 Genomes Project Phase 3 reference panel, which included 504 East Asian (EAS) subjects.<sup>46</sup>

Models with  $R^2 > 0.01$  and  $P$ -value  $< 0.05$  were retained. Introns with adjusted  $FDR < 0.05$  in the association tests were considered transcriptome-wide significant. Additionally, introns with promising significance ( $FDR < 0.2$ ) were included for further downstream analyses. Genetic colocalization analysis for these introns was conducted using the R package COLOC.<sup>47</sup> Posterior probabilities for hypothesis 4 (PPH4) were generated through Fully Bayesian colocalization analysis using Bayes Factors, indicating whether the intron usage ratio and NSCLC risk were influenced by the same causal variant(s). A PPH4  $> 0.8$  was considered strong evidence of colocalization.

Differential intron usage ratios between tumor and normal tissues were tested using the t.test function in R, with the Welch method applied to correct degrees of freedom. For comparisons across different genotypes, ANOVA followed by LSD tests was performed. Correlations between intron usage ratios were evaluated using Spearman's correlation test.

The splice-transcriptome-wide association study was conducted using FUSION.<sup>48</sup> The human reference genome version 38 (GRCh38) was used as the reference for exon annotations and variant coordinates.

### Identification for splicing factors modifying the SNP-splicing associations

Utilizing the genome and transcriptome data from lung tissue samples, we investigated potential splicing regulators that modulate SNP-splicing associations. A curated list of 404 splicing factor genes with established biological relevance was considered in this analysis.<sup>49</sup> Using an interaction model, we estimated the modifying effects of splicing factor gene expression levels on SNP-splicing associations to identify splicing factors that regulate splicing events and sQTLs of interest.

Circos plots were generated using the R package Circlize (v.0.4.15).<sup>50</sup>

### PPI network analysis

A protein-protein interaction network was constructed using STRING v12.0 with default settings.<sup>51</sup> The network included genes meeting the following criteria: (1) genes harboring introns with adjusted  $FDR < 0.2$  in the spTWAS association test and (2) a curated list of NSCLC-related genes reported in the COSMIC Cancer Gene Census (CSC). Genes located within the major histocompatibility complex (MHC) region were excluded from the analysis.

### Survival analysis

Survival analysis was conducted for NSCLC patients with available transcriptome measurements and non-missing survival data. Demographic characteristics were obtained from medical chart records. The intron usage ratio was dichotomized into high and low groups based on the median value.

A Cox regression model was used to estimate hazard ratios (HRs) with the R packages survival (<https://CRAN.R-project.org/package=survival>) and survminer (<http://rpkgs.datanovia.com/survminer/>). Kaplan-Meier survival curves were generated to compare survival outcomes between the groups. Both overall survival (OS) and progression-free survival (PFS) were employed to define survival status.

For independent validation, data from The Cancer Genome Atlas (TCGA) database were utilized. Clinical characteristics and gene expression quantification for lung adenocarcinoma and lung squamous cell carcinoma patients in TCGA were downloaded using the R package TCGAbiolinks.<sup>52</sup> Splicing

events in TCGA were quantified using SpliceSeq, and processed data were obtained from the publicly available repository (<https://bioinformatics.mdanderson.org/TCGASpliceSeq/PSIdownload.jsp>). Additionally, the online tool OncoSplicing was used for pan-cancer splicing survival analysis within TCGA.<sup>53</sup>

### Tumor MicroEnvironment (TME) and drug sensitivity analysis

To assess immune cell infiltration in the NSCLC immune microenvironment, the in silico immunophenotyping algorithm CIBERSORT was employed to estimate the proportions of different immune cell types within cancer tissues.<sup>54</sup>

The Oncopredict package was utilized for drug response prediction and drug-gene association analysis between PSI groups.<sup>55</sup> Gene expression data (GDSC2-Expr [RMA Normalized and Log Transformed].rds) and drug sensitivity data (GDSC2-Res.rds) from the GDSC2 dataset, provided by the official package, were used for model construction. The required files were downloaded from <https://osf.io/c6tfx/files/osfstorage/624e056b9fcdf400554fe1ae>.

The PSI values of introns in tumor tissues ( $n = 346$ ) were divided into high and low groups based on the median value.

## RESULTS

### Study overview

Figure 1 provides an overview of the study design. Initially, RNA sequencing data from all lung tissue samples ( $n = 747$ ) were processed to quantify splicing events. Differential splicing analysis was conducted to compare splicing differences between NSCLC tumor tissues ( $n = 346$ ) and adjacent normal lung tissues ( $n = 401$ ). Subsequently, QTL mapping was performed for individuals with both high-quality DNA and RNA sequencing data, encompassing tumor ( $n = 245$ ) and normal tissue samples ( $n = 297$ ). Genes involved in sQTLs, eQTLs, differential splicing, or differential expression were included in the enrichment analysis.

The genetic architecture of splicing was studied through the following approaches: (1) estimating the regional heritability of splicing, (2) annotating and visualizing the relative locations of sQTLs, (3) comparing identified sQTLs with eQTLs, (4) estimating the shared proportion of sQTLs and GWAS signals for NSCLC, and (5) determining the number of identifiable sQTLs under varying sample sizes through down-sampling. Regionally heritable introns were subsequently incorporated into the splicing-specific transcriptome-wide association study (spTWAS).

The spTWAS was employed to investigate NSCLC-associated splicing events within the context of genetic regulation. Splicing events identified by spTWAS were prioritized through colocalization analysis. Several splicing events showing concordant signals (from both spTWAS and differential splicing analysis) were further analyzed to explore potential upstream regulators. Protein-protein interaction (PPI) network analysis integrated the novel findings into a well-established framework by pooling spTWAS results with well-documented NSCLC genes. Prognostic analyses offered an additional perspective on the roles of splicing events in NSCLC. Overall, the study comprehensively investigated the role of splicing in NSCLC from multiple dimensions.

### Identification and characterization of sQTL in lung tissue

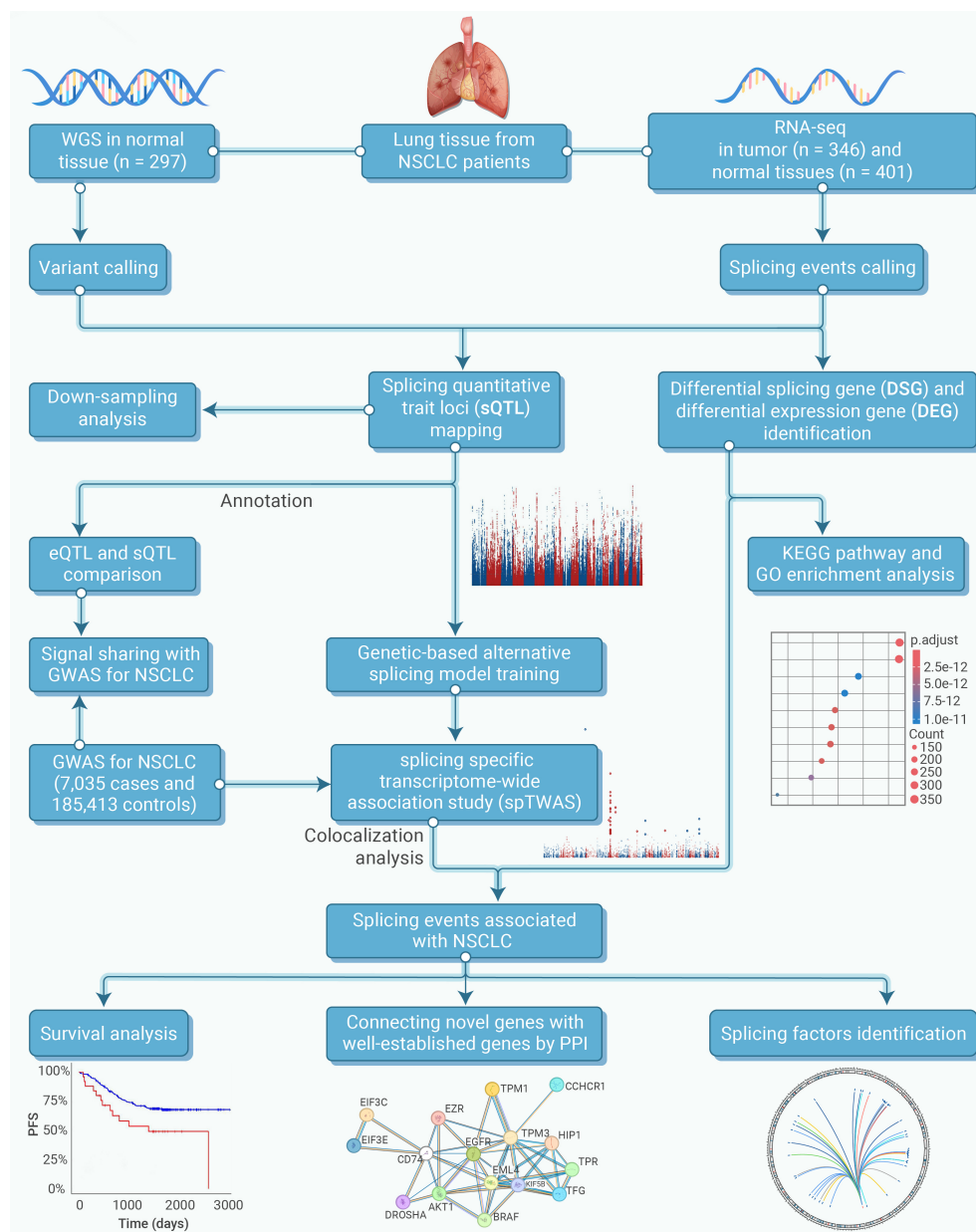
In this study, splicing events were quantified, and individuals with both DNA genotyping and RNA-seq data were included for sQTL mapping in tumor ( $n = 245$ ) and normal ( $n = 297$ ) tissues. To examine the distribution of sQTLs, the distance between each sQTL and its corresponding intron splice site was calculated.

For each splicing event, an  $FDR$  threshold of  $< 0.05$  was applied to determine significance. In normal tissues, a total of 363,853 sQTLs and 46,183 sIntrons were identified (Figure S2A), while tumor tissues yielded 209,463 sQTLs and 22,661 sIntrons (Figure S2B).

Figure S3A illustrates the enrichment of sQTLs near splice sites, highlighting the strong local effect of genetic regulation. This finding aligns with previous research indicating that sQTLs are predominantly located in close proximity to their associated splice sites.<sup>56</sup> Based on these results, a cis-window size of 100 kb was used for sQTL mapping.

The identified sQTLs were classified into five groups: normal, normal-





**Figure 1. Study overview** A genome-transcriptome reference map was constructed for both tumor and normal tissues from Chinese NSCLC patients. Alternative splicing profiles were characterized using RNA-seq data, leading to the identification of differentially spliced genes (DSGs) and differentially expressed genes (DEGs). Splicing quantitative trait loci (sQTLs) and expression quantitative trait loci (eQTLs) were mapped and systematically compared across the transcriptome. The degree of signal sharing between GWAS for NSCLC and sQTLs was estimated, and the relationship between sample size and the number of identified sQTLs was evaluated through down-sampling. A splice-transcriptome-wide association study (spTWAS) was conducted to identify potential causal splicing events associated with NSCLC. Further analyses explored the roles of splicing factors, constructed a protein-protein interaction (PPI) network, and assessed the prognostic significance of NSCLC-associated splicing events.

S4). This discrepancy may be partially attributed to noise in tumor tissues introduced by somatic mutations, structural variations, and other factors, which competitively reduce the proportion of phenotypic variation that can be attributed to genetic components.

### sQTLs and eQTLs serve different roles in NSCLC

We systematically compared the genetic architecture of sQTLs and eQTLs in lung tissue samples (Figure S5A). As depicted in the QQ plot, similar strong inflation was observed for sQTLs and eQTLs in normal tissues, suggesting comparable genetic regulatory strength for splicing and expression. In tumor tissues, sQTLs exhibited a slightly higher level of signals than eQTLs, potentially reflecting the influence of regulators such as abnormal copy number variations in tumors, which may explain a larger proportion of variance in gene expression, thereby reducing the relative contribution of genetic variants.

Figure S5B illustrates the overlap between sQTLs and eQTLs, as well as between sGenes and eGenes. In normal tissues, 25.6% of sQTL

and eQTL SNPs overlapped, while this proportion decreased to 15.0% in tumor tissues, as shown in the Venn plot. It is worth noting that sQTLs and eQTLs were defined using thresholds adjusted for multiple comparisons, which may underestimate the shared proportion. When mapping these QTLs to their corresponding genes, 57.7% of sGenes overlapped with eGenes in normal tissues, compared to 29.7% in tumor tissues. These results suggest that while a significant number of genes are genetically regulated by both expression and splicing, there are relatively independent regulatory mechanisms for sQTLs and eQTLs.

We further examined the overlap between differentially spliced/expressed genes and genetically regulated sGenes/eGenes (Figure S5C). Approximately 35.2% of DEGs overlapped with eGenes, whereas 32.0% of DSGs overlapped with sGenes. When focusing on differential analysis and comparing DSGs with DEGs, a 53.7% overlap was observed, indicating that a substantial number of genes undergo significant changes in both splicing and expression during carcinogenesis.

In summary, these findings highlight the distinct yet interconnected regulatory mechanisms underlying splicing and expression, as well as their significant roles in the genetic and functional landscape of NSCLC.

Compared to sQTLs identified in normal tissues, those identified in tumor tissues exhibited a lower level of signal sharing with GWAS for NSCLC (Figure

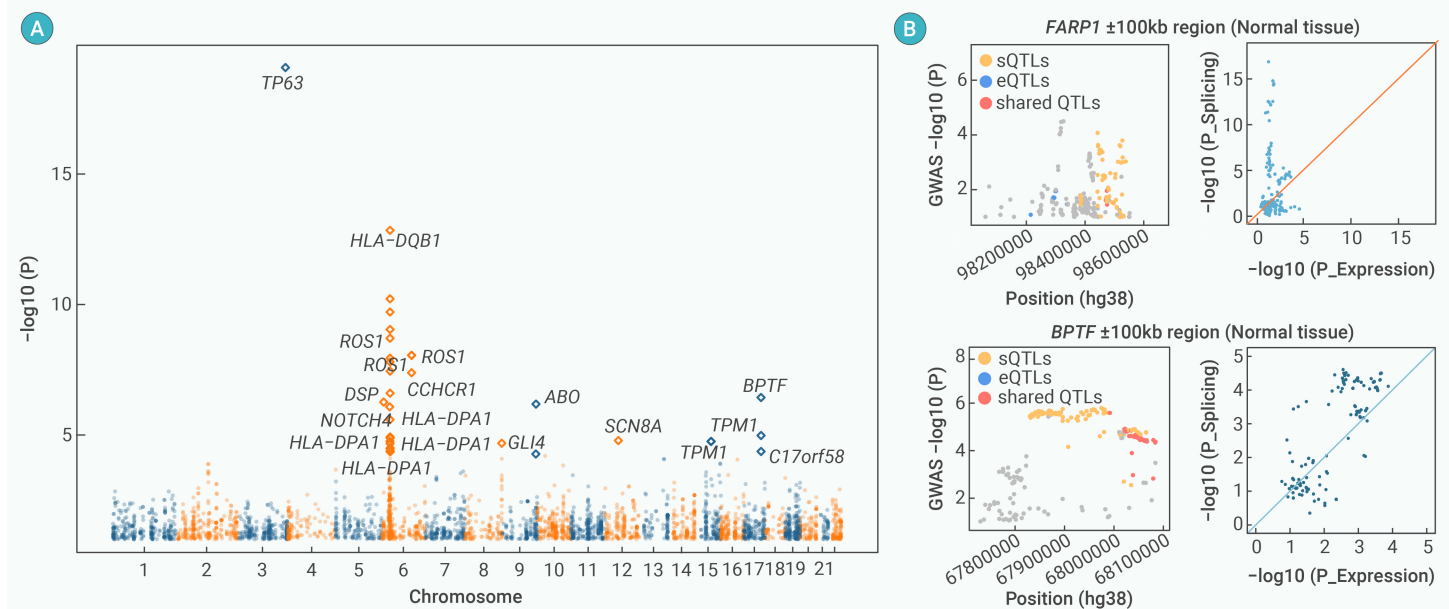
specific, tumor-normal-shared, tumor-specific, and tumor. The genomic localization of sQTLs was consistent across these groups, with the majority situated within intronic regions (Figure S3B). Among the sQTLs, 39.6% ( $n = 162,625$ ) were shared between tumor and normal tissues. A larger proportion of sQTLs were specific to normal tissues (49.0%,  $n=201,228$ ) compared to tumor-specific sQTLs (11.4%,  $n=46,838$ ) (Figure S3C).

To evaluate the impact of sample size on sQTL discovery, we performed down-sampling on both normal and tumor tissues. Six sampling points ( $n=100, 125, 150, 175, 200$  and  $297$ ) were considered to estimate the relationship between sample size and the number of significant sQTLs identified (Figures S3D–E). The results demonstrated an approximately linear correlation between sample size and the number of identified sQTLs.

To assess the degree of signal sharing between sQTLs and GWAS, we utilized the metric  $\pi_1$ . Among the five defined groups of sQTLs, the  $P$ -value distributions for GWAS correlation tests were generally similar, with the normal-related group displaying more significant signals. The highest  $\pi_1$  (8.87%) was observed in the normal-specific sQTL group, indicating a higher true positive rate of these sQTLs in GWAS SNP-NSCLC association tests. Conversely, the  $\pi_1$  for tumor-specific sQTLs was much lower (2.52%), underscoring the predominant role of normal tissue sQTLs in NSCLC (Figure S4).

### Enrichment analysis for DSGs and sGenes

We conducted enrichment analyses to uncover biological pathways and



**Figure 2. The splicing-based transcriptome-wide association study (spTWAS)** (A) The Manhattan plot illustrates the results of the spTWAS analysis, and genes successfully mapped by top-ranked significance introns ( $FDR < 0.05$ ) are labeled. (B) Compared to eQTL regulation, sQTLs exhibit stronger signals for spTWAS-identified NSCLC-associated genes, including *FARP1* (top) and *BPTF* (bottom). Local GWAS signals are displayed and color-coded to reflect eQTL and sQTL signals. SNPs associated with expression or splicing with  $FDR < 0.05$  were classified as QTLs.

processes associated with differentially spliced genes (DSGs) and genes harboring sQTLs (sGenes). Gene Ontology (GO) enrichment analysis revealed that DSGs were significantly enriched in terms such as histone modification, small GTPase-mediated signal transduction, cilium organization, positive regulation of cell projection organization, and regulation of small GTPase-mediated signal transduction. Additionally, DSGs were significantly enriched in KEGG pathways, including the MAPK signaling pathway, focal adhesion, axon guidance, cell cycle, and efferocytosis. Notably, significant enrichment was observed in pathways related to small cell lung cancer and pancreatic cancer, suggesting potential shared mechanisms (Figure S6A).

Similar GO terms and pathways were enriched in sGenes identified in tumor and normal tissues. Interestingly, sGenes showed enrichment in immune response activation and regulation signaling pathways, highlighting the potential role of genetic-based splicing regulation in immune system processes. Furthermore, both DSGs and sGenes were enriched in GO terms related to small GTPase-mediated signal transduction and cilium organization. Among KEGG pathways, efferocytosis and phagocytosis emerged as top-ranked pathways for sGenes (Figures S6B-C).

To ensure robustness, we performed a sensitivity analysis by limiting the enrichment analysis to the top one-fifth of genes (ranked by statistical significance). The results confirmed the reproducibility of top terms, including novel insights such as transcriptional misregulation in cancer for DSGs (Figure S7A) and antigen processing pathways for sGenes in both normal and tumor tissues (Figures S7B-C).

These findings underscore the importance of small GTPase-mediated signal transduction and the MAPK signaling pathway in the regulation of DSGs in NSCLC. Additionally, the results highlight immune-related signaling pathways, efferocytosis, and phagocytic pathways as central to splicing-related regulation in lung tissues.

### spTWAS identified the splicing susceptibility loci of NSCLC

To identify susceptibility genes for NSCLC associated with splicing regulation in lung tissues, we adapted the expression-based TWAS framework into a splicing-specific TWAS (spTWAS). Prediction models were trained for each of the 32,948 cis-heritable splicing events quantified by intron usage ratios. Four regression models were applied to each splicing event, with the best-performing model selected based on cross-validation. GWAS summary statistics were then used to probe NSCLC-associated splicing events. Figure S8 demonstrates the performance of the best models, with LASSO showing superior prediction performance for the majority of splicing events.

In the Manhattan plot (Figure 2A), we identified 19 alternative splicing events significantly associated with NSCLC ( $FDR < 0.05$ ), spanning 13 genes, including *TP63*, *ROS1*, *NOTCH4*, and *TPM1*. Notably, 14 of these 19 splicing events have not been previously linked to lung cancer (Table S2). Among the novel splicing events, some were located in genes already known to be associated with NSCLC, such as *TP63* and *ROS1*, suggesting that alternative splicing represents an additional mechanism by which these genes may influence NSCLC susceptibility.

By relaxing the  $FDR$  threshold to 0.2 to account for the relatively limited sample size in spTWAS, we identified 41 alternative splicing events in 30 genes (Table S3). Among these, *FARP1* ( $FDR = 0.068$ ) and the *EIF3* family (*EIF3C*,  $FDR = 0.18$ ), previously reported in NSCLC studies, were replicated under this threshold. Including these suggestive results ( $FDR < 0.2$ ) broadens the scope for identifying promising NSCLC-associated splicing events.

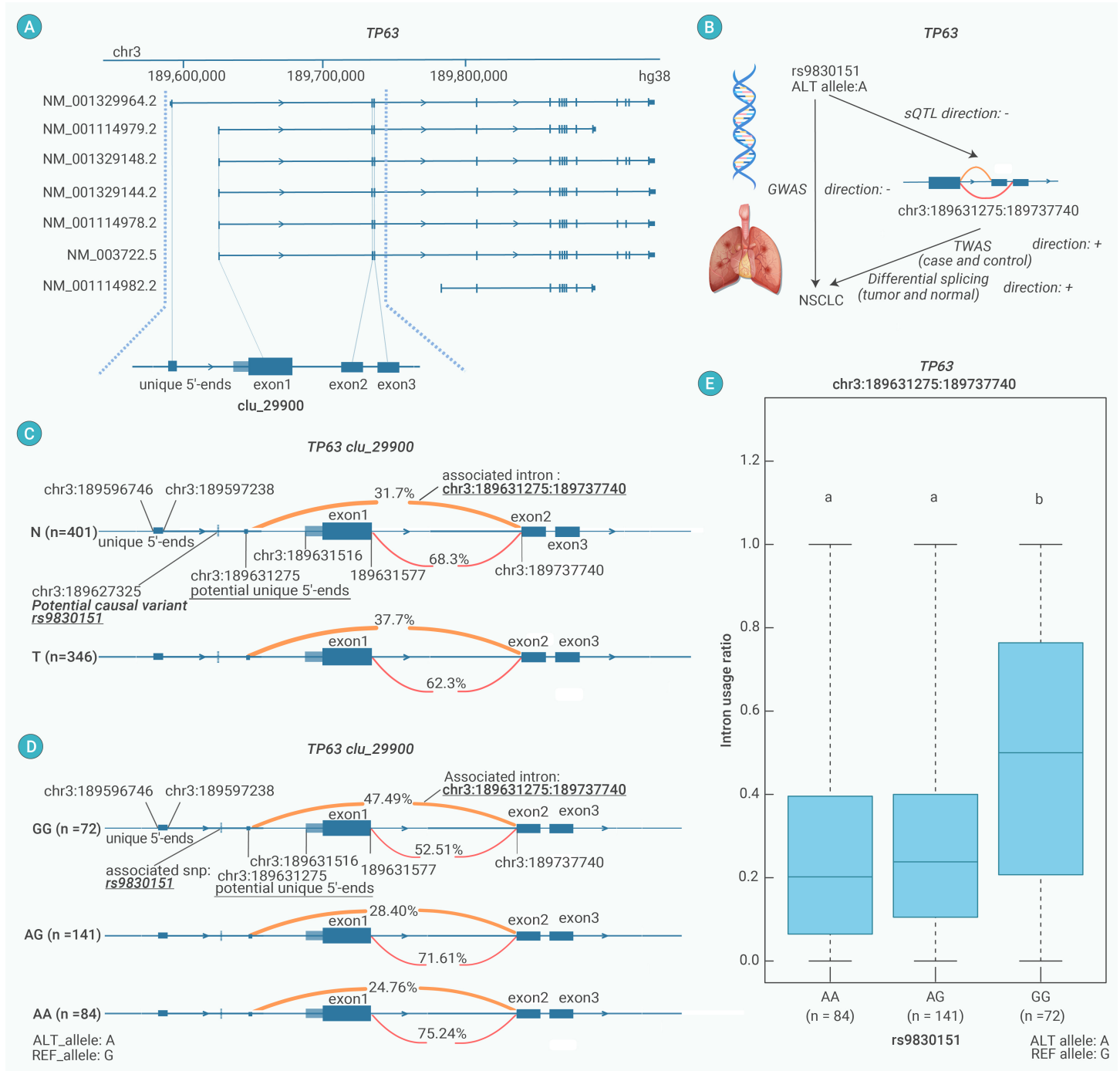
Remarkably, spTWAS identified unique signals in *BPTF* ( $P = 3.7E-07$ ) and *FARP1* ( $P = 8.4E-05$ ), which were not detected using expression-based TWAS analysis. Figure 2B highlights stronger local signals for sQTLs compared to eQTLs, emphasizing the independent regulation and potentially more significant role of splicing in NSCLC carcinogenesis.

To further prioritize spTWAS results, we conducted colocalization analysis as a complementary approach. Among the 40 introns with promising spTWAS significance ( $FDR < 0.2$ ), 7 intron regions showed evidence of colocalization (regional  $PPH4 > 0.8$ , Table S3). These findings underscore the utility of integrating splicing-specific approaches to identify and characterize NSCLC-associated splicing events.

### *TP63* and *TPM1* in the context of splicing regulation

Here, we highlighted *TP63* and *TPM1*, two genes demonstrating significance in spTWAS, colocalization analysis, and tumor-normal differential splicing analysis.

Figure 3A illustrates the complexity of post-transcriptional regulation in *TP63*, with the blue dashed line marking the splicing event identified in this study. Increased usage of an intron in *TP63* at chr3:189631275:189737740, identified in spTWAS, was significantly associated with a higher risk of NSCLC ( $P = 8.3E-20$ , Table S3). This intron represents a skipping event of the first exon of *TP63*, resulting in a novel transcript that connects at the 5' end of chr3:189631275. Colocalization analysis further supported the co-regulation of splicing and NSCLC (regional  $PPH4 = 0.98$ ) with the potential causal variant rs9830151. Figure 3B summarizes our integrative analysis, demonstrating that the first exon-skipping event mediates the effect of SNP rs9830151



**Figure 3. Alternative splicing of *TP63* is associated with the risk of NSCLC** (A) A regional map illustrates multiple isoforms of *TP63*, with the blue dashed line highlighting the splicing event identified in this study. (B) A summary of evidence for the splicing-NSCLC association demonstrates that the splicing event in *TP63* mediates the effect of rs9830151 on NSCLC risk. This conclusion is supported by spTWAS, colocalization analysis, and DSG analysis. (C) Sashimi plots depict differences in intron usage between tumor and normal tissues. (D) and (E) Visualizations of intron usage variations among samples with different genotypes of rs9830151 in normal tissues.

on NSCLC risk.

We observed that the A allele of rs9830151 was associated with reduced usage of the intron at chr3:189631275:189737740 ( $P = 3.4\text{E-}07$ , Figures 3D–E) and a decreased risk of NSCLC ( $P = 5.1\text{E-}21$ ). Differential splicing analysis further revealed that tumor tissues exhibited increased usage of the intron at chr3:189631275:189737740 ( $P = 5.4\text{E-}03$ , Figure 3C and Table S3).

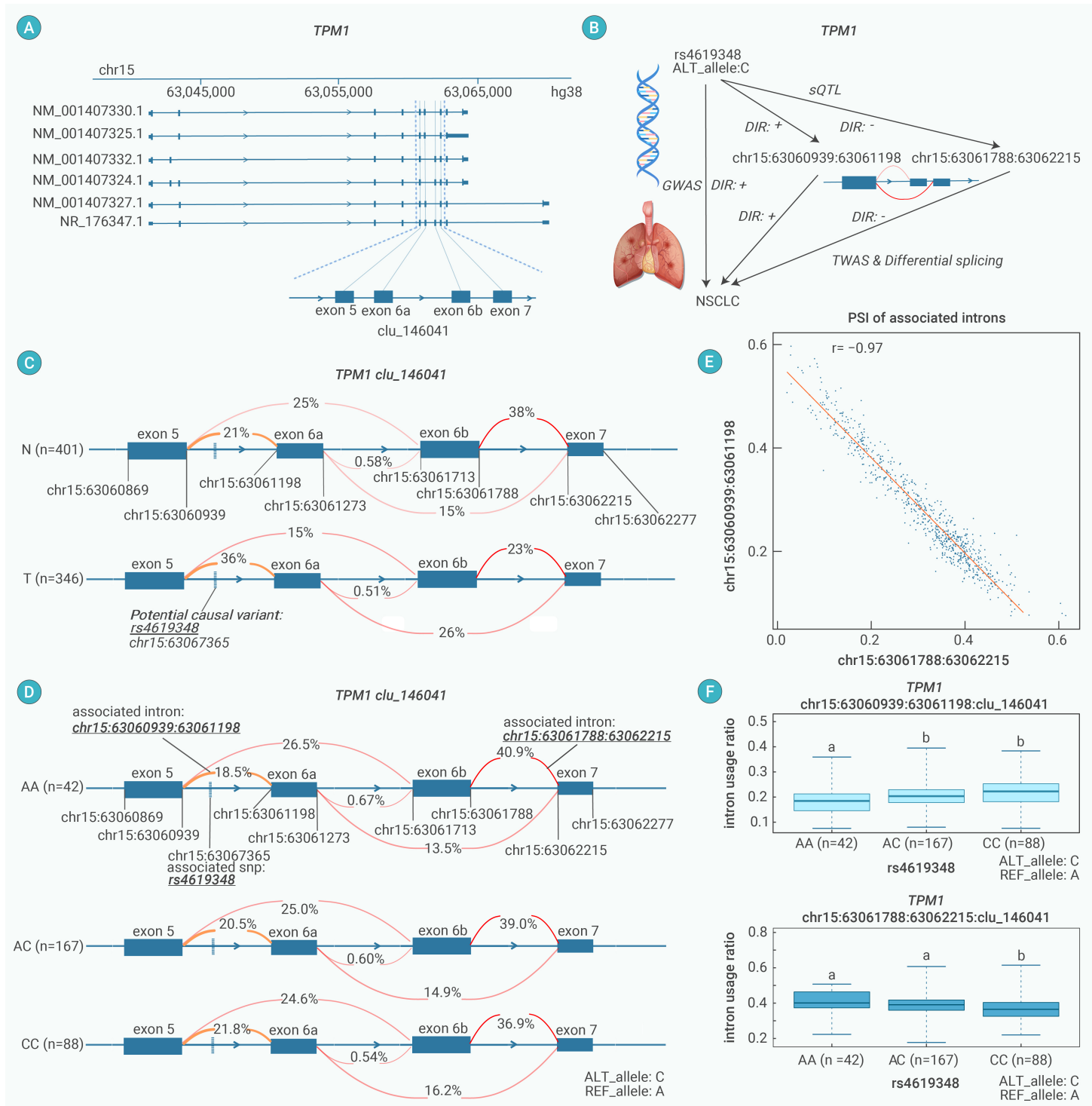
In conclusion, although skipping of the first three exons has previously been reported in *TP63*, our findings provide novel evidence for the intricate post-transcriptional regulation of this gene.

Our spTWAS association test identified a mutually exclusive splicing event in *TPM1* that was significantly associated with NSCLC risk. Figure 4A illus-

trates different splicing patterns within the *TPM1* gene region, with blue dashed lines marking the genomic regions of interest identified in this study. Increased usage of the intron chr15:63060939:63061198 was associated with a higher risk of NSCLC ( $P = 1.8\text{E-}05$ , Table S3), whereas increased usage of the intron chr15:63061788:63062215 was associated with a lower risk of NSCLC ( $P = 1.8\text{E-}05$ , Table S3). Figure 4B highlights our integrative analysis, demonstrating that the mutually exclusive splicing event mediates the effect of SNP rs4619348 on NSCLC risk.

We observed a strong negative correlation between the usage ratios of these two introns ( $r = -0.97$ , Figure 4E), indicating that the corresponding exons are mutually exclusive. Colocalization analysis further supported a potentially causal role of rs4619348, with shared genetic signals for both





**Figure 4. Alternative splicing of *TPM1* was associated with the risk of NSCLC** (A) A regional map displays multiple isoforms of *TPM1*, with the blue dashed line indicating the splicing event identified in this study. (B) A summary of evidence for the splicing-NSCLC association demonstrates that the splicing events in *TPM1* mediate the effect of rs4619348 on NSCLC risk, supported by sPTWAS, colocization analysis, and DSG analysis. (C) Sashimi plots illustrate differences in intron usage between tumor and normal tissues. (D) Sashimi plots depict differences in intron usage among samples with different genotypes of rs4619348 in normal tissues. (E) scatter plot reveals the mutually exclusive usage of exon6a and exon6b. (F) Box plots show differences in intron usage among samples with different genotypes of rs4619348 in normal tissues.

splicing and NSCLC (regional PPH4 = 0.92). For rs4619348, the C allele was associated with increased usage of the intron chr15:63060939:63061198 ( $P = 2.5 \times 10^{-3}$ , Figures 4D & F) and an elevated risk of NSCLC ( $P = 1.8 \times 10^{-5}$ ).

Concordant results were observed in differential splicing analysis for the intron pair. Tumor tissues exhibited increased usage of chr15:63060939:63061198 compared to adjacent normal tissues ( $P = 2.1 \times 10^{-100}$ , Figure 4C and Table S3). Similarly, increased usage of chr15:63061788:63062215 was also detected in tumor tissues ( $P = 1.7 \times 10^{-100}$ , Table S3).

These findings underscore the role of mutually exclusive splicing events in

*TPM1* as a potential mechanism influencing NSCLC susceptibility.

To evaluate the roles of expression and splicing regulation in NSCLC, we visualized genetic regulation signals for splicing (sQTL), expression (eQTL), and GWAS within the *TP63* and *TPM1* regions (Figure S9). In the *TP63* region, we identified a cluster of signals shared between sQTL and eQTL regulation. For rs9830151, the splicing signal exhibited slightly stronger association strength compared to the expression signal. In tumor tissue, only SNP-splicing associations were observed, while expression regulation signals were minimal (Figure S9A). For *TPM1*, comparable eQTL and sQTL signals were

detected within the cis region in both normal and tumor tissues (Figure S9B). These findings highlight the indispensable role of splicing in gene expression regulation and its contribution to the etiology of NSCLC.

To identify potential splicing factors that modulate the genetic regulation of NSCLC-associated splicing events in *TP63* and *TPM1*, we examined the effects of 404 known splicing factors on SNP-splicing associations in normal tissues. Using an interaction model, we identified 39 potential splicing factors for the *TP63* first exon skipping event and 33 for the *TPM1* exon 6 mutually exclusive event ( $P_{\text{interaction}} < 0.05$ , Figures 5A–B). A significant interaction was observed between the expression of *CLASRP* and the genotype of rs9830151 on intron usage at *TP63*:chr3:189631275:189737740 ( $P_{\text{interaction}} = 0.002$ ). Stratified analysis revealed a significant association between intron usage and its potential regulator *CLASRP* in samples with the GG genotype of rs9830151, whereas no significant association was observed in samples carrying the A allele. This suggests that the A allele may reduce the affinity of *CLASRP* binding to the target splicing site (Figure 5C). Similarly, *HSPB1* was found to modify the genetic regulation of rs4619348 on intron usage at *TPM1*:chr15:63061788:63062215 ( $P_{\text{interaction}} = 0.002$ ). A stronger effect of *HSPB1* on the splicing event in *TPM1* was observed in rs4619348 A allele carriers, with evidence of a dosage response trend (Figure 5D). These results underline the critical role of splicing factors, such as *CLASRP* and *HSPB1*, in modulating the genetic regulation of splicing events associated with NSCLC.

We further extended the list of splicing factors to identify additional potential modification effects in normal and tumor tissues. For the *TPM1* splicing event, more genes were identified, including the pseudogene *RBM22P2* (Figure S10), which belongs to the RBM family, a group known to include many critical splicing factors. Additionally, we identified the putative modification effect of *RNF34* on the sQTL for *TP63* and *DDX3X* on the sQTL for *TPM1* (Figure S11). The proposed mechanisms underlying splicing regulation are illustrated in Figures 5E–F. These findings, supported by in silico mapping of splicing factors, provide deeper insights into the regulatory processes driving the identified NSCLC-associated splicing events.

### spTWAS result genes were linked to known NSCLC genes in the PPI network

To investigate potential links between the newly identified NSCLC splicing-related genes and a list of widely recognized NSCLC-associated genes, we constructed a protein-protein interaction (PPI) network. Genes identified by spTWAS ( $FDR < 0.2$ ), excluding those located in the MHC region, were evaluated for connections with known NSCLC-related genes summarized by the COSMIC Cancer Gene Census (CGC, Figure 6A). Among the 30 genes identified by spTWAS, 14 were directly connected to known NSCLC genes within the PPI network (Figure 6B). These genes also exhibited widespread splicing dysregulation (Table S3) and expression dysregulation (Table S4) between tumor and normal tissues. Additionally, inner connections were observed among the spTWAS-identified genes, including the pairs *BPTF-C17orf58*, *LIMS2-ILK*, and *DSP-CDH11* (Figure 6B). Notably, *ROS1* and *TP63* are well-established NSCLC-related genes, while *DDR1* and *TPM1* are family members of known NSCLC-related genes *DDR2* and *TPM3*, respectively. Similarly, *EIF3E*, a family member of the spTWAS-identified gene *EIF3C*, has been reported to be associated with lung cancer. Remarkably, the pair *BPTF-C17orf58* and *ZIM22* were found to connect with known NSCLC-associated genes such as *SMARCA4* and *KEAP1* ( $P = 1.0E-16$ , Figure 6C). The *LIMS2-ILK* pair and *DDR1* were linked to core NSCLC genes *EGFR* and *DDR2* ( $P = 3.8E-05$ , Figure 6C). Connections between the spTWAS-identified pair *DSP-CDH11* and NSCLC-related genes *EGFR*, *SOX2*, and *TP63* were also established ( $P = 3.3E-03$ , Figure 6C). Furthermore, a network involving spTWAS-identified genes (*EIF3C*, *TPM1*, and *CCHCR1*) and well-known genes (*EIF3E*, *TPM3*, and *CD74*) was revealed ( $P = 8.8E-12$ , Figure 6C). Overall, these PPI analyses connect the novel splicing-related genes to the NSCLC network, providing new insights and expanding the knowledge base of NSCLC etiology.

### Survival analysis provided evidence for *ILK* exon 3 skipping

We further investigated the potential prognostic role of the identified splicing events in NSCLC. Using transcriptome data from tumor tissues, we applied a Cox regression model to estimate the impact of alternative splicing

on overall survival (OS) and progression-free survival (PFS). Thirteen introns with promising spTWAS significance ( $FDR < 0.2$ ) and concordant DSG results were included in the survival analysis. Among these, three introns—associated with *ILK*, *DSP*, and *GLI4*—were significantly linked to NSCLC survival outcomes (Table S5).

We focused on the splicing event in the gene *ILK*. Increased intron usage at chr11:6608211:6608394 was associated with a higher risk of NSCLC ( $P = 3.5E-04$ , Table S3). Figure 7A illustrates the splicing structure of the region near intron chr11:6608211:6608394 in *ILK*, which involves the skipping of exon 3. This intron connects exon 3 and exon 4 of *ILK*, and its elevated PSI level indicates increased usage of exon 3.

Survival analysis revealed that higher intron usage at chr11:6608211:6608394 was significantly associated with worse OS (HR = 2.45,  $P = 2.0E-04$ , Figure 7B). Furthermore, differential splicing analysis supported an oncogenic role for increased usage of *ILK*:chr11:6608211:6608394, although the difference between tumor and normal tissues was nominally significant ( $P = 0.028$ , Figure 7C and Table S3). To validate these findings, we used independent data from the TCGA database. In TCGA, we confirmed that asid:14173 corresponds to the same *ILK* exon 3 skipping event highlighted in our analysis. Increased PSI values for this event also indicated elevated usage of *ILK* exon 3. The oncogenic potential and poor prognostic association of increased *ILK* exon 3 usage were replicated in TCGA, particularly in LUAD, aligning with our conclusions (Figures S12A & S13A).

To assess the contribution of *ILK* expression and splicing events, we conducted both differential expression analysis and survival analysis for *ILK* gene expression. Differential expression analysis suggested a potential tumor suppressor role for *ILK*, consistent with findings in the TCGA dataset (Figure S12B). However, no significant association was observed between *ILK* expression and NSCLC prognosis, a result also mirrored in the TCGA dataset (Figures 7D & S13B). Similarly, analysis of PFS also showed that *ILK* expression was not predictive of NSCLC prognosis (Figure S14).

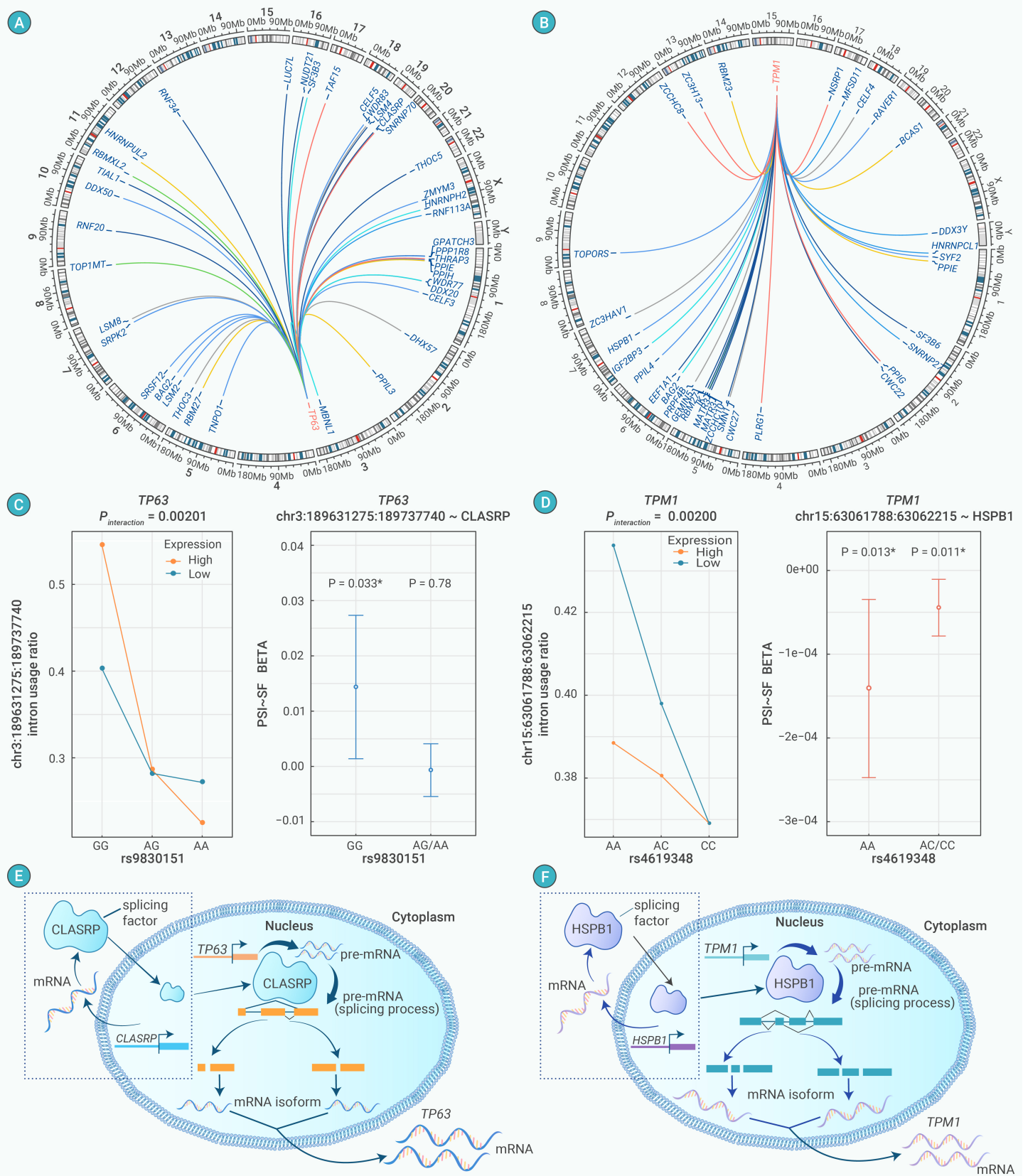
Taken together, these findings highlight that the splicing event involving *ILK* exon 3 is consistently associated with NSCLC risk, supported by genetic-based TWAS and DGE analyses (from both our dataset and TCGA), as well as prognostic outcomes (from both our dataset and TCGA). In contrast, overall *ILK* gene expression does not provide prognostic value for NSCLC in either dataset.

### Potential biological role of *ILK* exon 3 skipping and drug candidate prediction

The tumor microenvironment (TME) is of particular interest<sup>57</sup> in the context of splicing regulation, aligning with our observation that genes harboring sQTLs are enriched in immune-related pathways. To explore this, we estimated the immune cell composition of NSCLC patients and identified differences in T cells, M0 and M2 macrophages, dendritic cells, mast cells, and neutrophils between high and low PSI groups for *ILK*:chr11:6608211:6608394. A higher skipping rate of *ILK* exon 3 was observed in samples with a higher proportion of T cells and M0 macrophages, whereas a lower skipping rate was noted in samples with increased proportions of M2 macrophages, dendritic cells, mast cells, and neutrophils (Figure S15A).

To assess the broader impact of *ILK* exon 3 skipping, we performed a pan-cancer analysis across multiple cancer types. Differential splicing between tumor and normal tissues was detected in LUAD, LUSC, PRAD, KICH, THCA, COAD, BLCA, and HNSC (Figure S15B). Poor prognosis associated with increased usage of *ILK* exon 3 was observed in LGG and LUAD (Figure S15C). Notably, lung cancer exhibited one of the most prominent effects, highlighting the unique prognostic value of *ILK* exon 3 skipping in this context.

In addition, we explored potential drug targets based on cellular drug response data from the GDSC dataset, with a particular focus on the *ILK* exon 3 skipping event. Drug sensitivity analysis revealed the top ten drugs (ranked by statistical significance) most sensitive to cell growth in relation to the PSI levels of *ILK* exon 3 skipping (Figure S15D). For drugs such as Doramapimod\_1042, JQ1\_2172, NU7441\_1038, RO-3306\_1052, Ruxolitinib\_1507, SB216763\_1025, and SB505124\_1194, lower IC50 values were observed in the high PSI group, indicating increased exon 3 usage. Conversely, for drugs like ERK\_6604\_1714, Pictilisib\_1058, and Uprosertib\_2106, lower IC50 values were found in the low PSI group, reflecting reduced exon 3 usage.



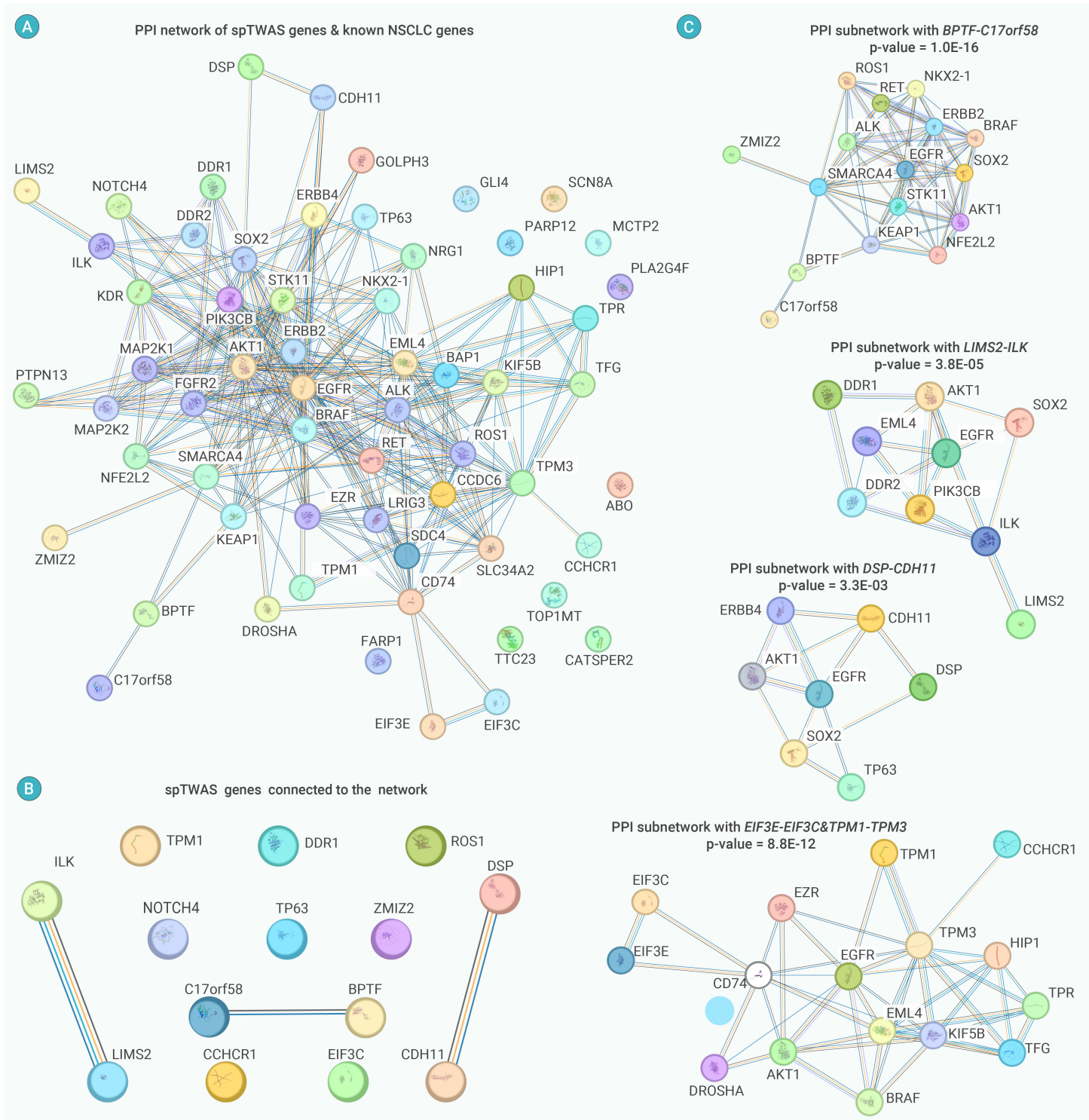
**Figure 5. Potential upstream regulators and mechanisms for the alternative splicing events** Potential splicing factors that modify the association between genetic variants and NSCLC-associated splicing events are visualized for *TP63* (A) and *TPM1* (B). Examples illustrating the modification effects of CLASRP and HSPB1 are presented in panels C and D, respectively. The proposed mechanisms underlying genetic-based splicing regulation are depicted in panels E and F.

These findings provide valuable insights into the prognostic significance of *ILK* exon 3 skipping in lung cancer and suggest potential avenues for drug development targeting this specific splicing event.

## DISCUSSION

In this study, we constructed a comprehensive reference panel for alternative splicing in lung tissues and identified NSCLC-associated splicing events



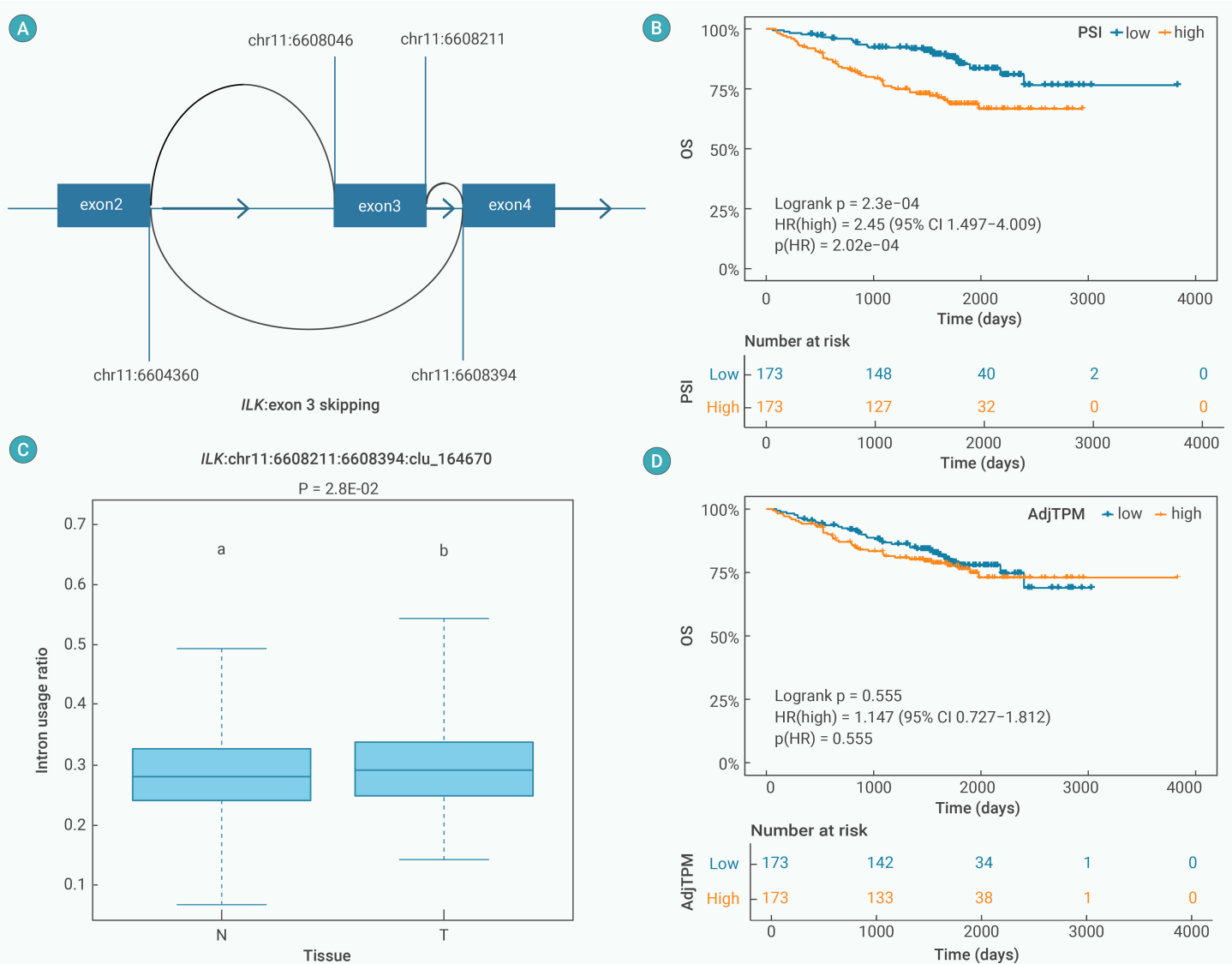


**Figure 6. The protein-protein interaction (PPI) network analysis connects novel NSCLC-associated genes identified by spTWAS with documented cancer genes** (A) The PPI network of all NSCLC-associated genes identified by spTWAS, excluding those located in the MHC region, alongside NSCLC-related genes from the COSMIC CGC gene list. (B) spTWAS-identified genes that are connected to the PPI network. (C) Subnetworks formed by the identified spTWAS genes.

through a GWAS-based scan and differential splicing analysis between tumor and normal tissues. Fourteen novel NSCLC-associated splicing events were highlighted, involving genes such as *TP63* and *TPM1*. Furthermore, we identified potential upstream regulators, connections with well-established cancer genes, and prognostic roles for these splicing events, underscoring the critical role of alternative splicing in NSCLC.

We characterized the genetic architecture of sQTLs and demonstrated the importance of larger sample sizes for sQTL identification. This increased identification of sQTLs significantly expanded the catalog of mappable regulatory elements, offering new insights into potential risk factors for NSCLC.

Across the transcriptome, we observed a diverse array of sGenes, highlighting the widespread impact of genetic-based regulation as a valuable resource for uncovering NSCLC-associated molecular mechanisms. Our comparative analysis of sQTLs and eQTLs, which are more commonly studied, revealed that most sQTLs are independent of eQTLs, consistent with previous findings.<sup>17,24,27,58</sup> This independence suggests that alternative splicing provides a unique contribution to explaining SNP-NSCLC associations. By integrating GWAS results with tumor- and normal-specific sQTLs, we demonstrated that while sQTLs in normal tissues explained a larger proportion of lung cancer-related associations, tumor-specific sQTLs also contributed significantly to



**Figure 7. Alternative splicing of *ILK* was a predictor for the prognosis of NSCLC** (A) A regional map highlights the splicing event of *ILK* identified in this study. (B) Using the splicing events of *ILK* in tumor tissue samples from 346 NSCLC patients, we performed a survival analysis with Cox regression. A significant difference in OS was observed between high and low PSI levels for intron chr11:6608211:6608394, with higher PSI levels indicating increased usage of *ILK* exon 3. (C) Differential analysis between tumor and normal tissues revealed higher PSI levels in tumor tissues. (D) The overall expression of *ILK* did not provide informative prognostic predictions.

the interpretation of GWAS results. Notably, some GWAS signals could not be colocalized with sQTLs in normal tissues, indicating spatiotemporal tumor-specific regulation of splicing.<sup>24,59,60</sup> These findings enhance our understanding of the genetic and regulatory landscape of alternative splicing in NSCLC and its potential as a distinct mechanism underlying disease etiology.

Differential splicing analysis identified a substantial number of differentially spliced genes (DSGs). The enrichment analysis of these DSGs highlighted the critical roles of small GTPase-mediated signal transduction and the mitogen-activated protein kinase (MAPK) pathway, particularly the Ras-Raf-MEK-ERK signaling cascade. In this study, aberrant splicing was observed in 18 members of the MAPK family within tumor tissues, including *MAPK1* and *MAPK3* (also known as ERK2 and ERK1, respectively). The ERK signaling pathway influences cancer cell survival, invasion, and migration through various mechanisms.<sup>61</sup> ERK itself represents a promising therapeutic target alongside Raf and MEK, and the development of ERK1/2 inhibitors offers potential for treating a variety of malignancies driven by activated pathway mechanisms.<sup>62</sup> Enrichment of sGenes underscored the importance of immune response activation and regulatory signaling pathways. Many genes identified in this study as undergoing genetically regulated splicing were associated with immune responses, including *CELF2*, *TRAF3*, *BRD8*, and *TRIM*.<sup>63</sup> Notably, abnormal splicing of splicing factors can drive transcriptome-wide changes. A notable example is *SF3A1*, an sGene encoding a splicing

factor that regulates the splicing of genes involved in Toll-like receptor (TLR) signaling pathways in macrophages.<sup>64</sup> Immune cells undergo repeated division and differentiation throughout their lifespan, with adaptive immune system cells representing highly evolved entities in vertebrates.<sup>65</sup> As such, genetically regulated alternative splicing likely plays a pivotal role in immune cell differentiation, potentially explaining the enrichment of sGenes in immune-related pathways.<sup>66</sup> Remarkably, a considerable proportion of DSGs overlapped with sGenes, reinforcing the connection between tumorigenesis-related alternative splicing and genetic-based splicing regulation in lung tissues. A representative example is *CD44*, which was identified as both a DSG and an sGene. Consistent with our findings, *CD44* splicing is involved in immune-related pathways, with its splicing regulation influenced by the Ras-Raf-MEK-ERK signaling pathway.<sup>65,67,68</sup> *CD44* isoforms are associated with various cancers,<sup>69–71</sup> and studies have demonstrated that the *CD44v* isoform in turn, promotes ERK activation in Kras-mediated signaling in human H358 and mouse LKR13 lung cancer cells.<sup>72</sup> In summary, we observed significant enrichment of DSGs and sGenes in cancer- and immunity-related pathways, providing a comprehensive view of the regulatory roles of alternative splicing in NSCLC. These findings highlight the intricate interplay between splicing regulation, tumorigenesis, and immune modulation.

While differential splicing analysis is a powerful tool for identifying changes in PSI (percent-spliced-in) levels in tumor tissues, it is susceptible to reversed

causality and confounding factors. To address these limitations, we incorporated genetic-informed spTWAS to further prioritize NSCLC-associated splicing genes. In spTWAS, four statistical models were simultaneously applied for model construction (see Methods section). However, given the modest sample size, the final model selected may not always represent the optimal choice. The models in Table S3 may raise concerns due to their low  $R^2$  values, but this is consistent with the expectation for genetic-based prediction models, which typically yield small  $R^2$  values. Empirical data indicate that common variants explain approximately 10% of the variance in gene expression levels, with an average  $R^2$  of ~10%.<sup>29</sup> Given the limited sample size of xQTL datasets, observing  $R^2$  values in the range of 1%-10% is reasonable.<sup>23,27</sup>

Notably, some genes identified by spTWAS in this study, such as *FARP1*, have been previously reported to be associated with NSCLC risk through splicing regulation, but the majority are novel discoveries. This novelty may stem from several factors: (1) The use of Leafcutter's non-predefined clustering framework, which allows greater flexibility in defining splicing events and identifying more novel events; (2) The larger sample size in our sQTL mapping, enabling the identification of additional genetic regulations of splicing events and providing a more comprehensive understanding of SNP-splicing-NSCLC associations. By integrating results from TWAS, GWAS, differential splicing analysis, and colocalization, we highlighted the genetic-based alternative splicing regulation of *TP63* and *TPM1*. The *TP63* gene encodes the transcription factor p63, which is crucial for epidermal differentiation.<sup>73</sup> Structurally, *TP63* is highly complex, producing multiple protein isoforms via alternative promoters, splicing, or polyadenylation. Its splicing events have been closely linked to squamous cell carcinoma.<sup>74,75</sup> Studies have also revealed the existence of a *TP63* transcript with a unique 5' end originating upstream of the currently mapped gene.<sup>76</sup> However, the role of this alternative first exon splicing event in NSCLC development remains unclear. Our study demonstrated that the well-characterized association between the lead SNP rs9830151 and NSCLC may be mediated by alternative splicing at intron *TP63:chr3:189631275:189737740*. Additionally, we uncovered the modifying role of rs9830151 in regulating the interaction between the splicing factor CLASRP and splicing events at *TP63*. These findings expand the current understanding of *TP63* splicing regulation and its implications in NSCLC.

Three of the four mammalian tropomyosin (TM) genes—*TPM1*, *TPM2*, and *TPM3*—are known to undergo alternative splicing. Notably, many of these events involve mutually exclusive exons. In *TPM1*, exons 2 and 6 exhibit mutually exclusive splicing patterns.<sup>77</sup> The mutually exclusive exon 6 of *TPM1* has been implicated in several cancers, including colon, bladder, kidney, lung, prostate, and uterine cancers, as demonstrated through differential splicing analyses.<sup>78-81</sup> For NSCLC, a prior exon array-based study reported that skipping of exon 6b in *TPM1* was more frequently observed in 11 tumor tissues compared to paired normal tissues.<sup>82</sup> Our study confirmed this observation for differentially spliced genes (DSGs) with a significantly larger sample size and further provided evidence through genetic-based approaches. Guided by the potentially causal variant rs4619348 identified in colocalization analysis, our spTWAS results validated the role of the exon 6 splicing event in the risk of developing NSCLC. Additionally, HSPB1 and other splicing factors were identified as regulators whose effects on exon 6 skipping could be influenced by the associated sQTL. These findings highlight the genetic and regulatory underpinnings of exon 6 splicing in *TPM1* and its critical role in NSCLC pathogenesis.

We also validated the NSCLC-associated splicing genes *FARP1* ( $P = 8.4 \times 10^{-5}$ ) and the *EIF3* family (*EIF3C*,  $P = 3.4 \times 10^{-4}$ ), which have previously been identified as associated with NSCLC risk in Han Chinese through splicing events.<sup>27</sup> *FARP1* is a Rac guanine nucleotide exchange factor (Rac GEF) known to be upregulated in EpCAM+ cells of human lung adenocarcinoma. It serves as a key driver of motor signal transduction in response to tyrosine kinase receptor (e.g., EGFR, c-MET, AXL) stimulation in lung adenocarcinoma cells.<sup>83</sup> While the splicing event we identified differs from previously reported findings due to the complex splicing patterns within *FARP1*, the concordance in results further supports the role of splicing regulation in *FARP1* for NSCLC. Moreover, our analysis suggests that splicing, rather than gene expression, is a stronger mediator of the SNP-NSCLC association in *FARP1*. The eukaryotic translation initiation factor 3 (eIF3) is the largest and most complex mammalian initiation factor, weighing 800 kDa and composed of 13 distinct

subunits (eIF3a-m). Genes encoding translation initiation factors are frequently altered in human tumors. eIF3 plays a pivotal role in cancer due to its involvement in protein synthesis and multifaceted functions.<sup>84-86</sup> *EIF3E*, a family member of *EIF3C*, has been reported to associate with NSCLC through alternative splicing.<sup>27</sup> Additionally, *EIF3C* itself has been shown to exhibit overexpression in seminoma and high mRNA levels in meningioma and colorectal cancer.<sup>87-89</sup> The validation of *FARP1* and the *EIF3* family not only underscores the reliability of our findings but also highlights the potential specificity of these novel splicing events to Chinese patients, as they may not have been detected in other populations. Comparing our results with spTWAS findings in European populations would be highly informative; however, this effort is currently limited by the lack of individual-level genotype and RNA-seq data required to construct models consistent with the event-calling processes used in this study.

We further demonstrated the importance of the splicing event in the NSCLC-associated gene *ILK* by integrating survival analysis. Increased usage of the intron *ILK:chr11:6608211:6608394* was associated with an elevated risk of NSCLC. Survival analysis revealed that patients with higher intron usage exhibited worse survival outcomes. Integrin-linked kinase (ILK), an ankyrin repeat-containing serine/threonine protein kinase, plays a critical role in the initiation and progression of various cancers, including lung cancer.<sup>90,91</sup> *ILK* has been shown to induce cancer cell migration and invasion in multiple cancer types via the NF- $\kappa$ B pathway,<sup>90,92,93</sup> and to form a regulatory feedback loop with KRAS, promoting oncogenic KRAS signaling and interacting with the tumor microenvironment to drive invasive phenotypes.<sup>94,95</sup> Despite these findings, the biological functions underlying *ILK* splicing events remain unclear. Notably, differential splicing events in *ILK* have been associated with melanoma progression in mouse models.<sup>96</sup> Our study provides a significant observation of the role of *ILK* splicing in the prognosis of NSCLC. Importantly, this prognostic effect was only observed through alternative splicing, not gene expression regulation, underscoring the distinct role of splicing in prognostic prediction.

Our study has several key strengths. First, as shown in the down-sampling analysis, we mapped a larger number of sQTLs, resulting in a more comprehensive identification of NSCLC-associated splicing events. Second, by leveraging both spTWAS and DSG analyses, we systematically probed NSCLC-associated splicing events. Third, we connected identified splicing events and their relevant genes to potential splicing factors and well-established cancer networks, offering insights into plausible biological mechanisms. Despite these strengths, our study has certain limitations. While we increased the number of sQTLs detected in the Han Chinese population, the sample size of 297 remains insufficient to reach a detection plateau, limiting the statistical power of spTWAS. Future studies with larger sample sizes are necessary to address this limitation. Although we provided robust statistical evidence for the association between splicing-mediated SNPs and NSCLC, functional validations are required to confirm the underlying mechanisms.

## CONCLUSION

In summary, we highlighted the critical role of alternative splicing and identified novel splicing events significantly associated with NSCLC in genes such as *TP63* and *TPM1*. Furthermore, we uncovered potential upstream splicing regulators and integrated these genes into a network of well-established cancer-related genes. Notably, the prognostic predictive value was observed for splicing events, rather than gene expression, in genes such as *ILK*. This study underscores the importance of RNA splicing dysregulation in NSCLC and provides additional insights into the etiology of the disease.

## REFERENCES

- Sung H, Ferlay J, Siegel R. L., 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**:209–249. DOI:10.3322/caac.21660
- Han B, Zheng R, Zeng H, et al. (2024). Cancer incidence and mortality in China, 2022. *J. Natl. Cancer Cent.* **4**:47–53. DOI:10.1016/j.jncc.2024.01.006
- Huang W, Zhai G, Dong H, et al. (2023). Geographical and sexual disparities of lung cancer mortality trends in China: A population-based study. *Innov. Med.* **1**:100032. DOI:10.59717/j.xinn-med.2023.100032
- Duma N, Santana-Davila R, and Molina J. R. (2019). Non-small cell lung cancer: Epidemiology, screening, diagnosis, and treatment. *Mayo Clin. Proc.* **94**:1623–1640.



- DOI:10.1016/j.mayocp.2019.01.013
5. Benusiglio P. R., Fallet V., Sanchis-Borja M., et al. (2021). Lung cancer is also a hereditary disease. *Eur. Respir. Rev.* **30**:210045. DOI:10.1183/16000617.0045-2021
  6. Du X., Yang B., An Q., et al. (2021). Acquired resistance to third-generation EGFR-TKIs and emerging next-generation EGFR inhibitors. *The Innovation* **2**:100103. DOI:10.1016/j.xinn.2021.100103
  7. Shi X., Young S., Cai K., et al. (2022). Cancer susceptibility genes: update and systematic perspectives. *The Innovation* **3**:100277. DOI:10.1016/j.xinn.2022.100277
  8. Bosse Y. and Amos C. I. (2018). A decade of GWAS results in lung cancer. *Cancer Epidemiol. Biomarkers Prev.* **27**:363–379. DOI:10.1158/1055-9965.EPI-16-0794
  9. Long E., Patel H., Byun J., et al. (2022). Functional studies of lung cancer GWAS beyond association. *Hum. Mol. Genet.* **31**:R22–R36. DOI:10.1093/hmg/ddac140
  10. Martinez-Ruiz C., Black J. R. M., Puttick C., et al. (2023). Genomic-transcriptomic evolution in lung cancer and metastasis. *Nature* **616**:543–552. DOI:10.1038/s41586-023-05706-4
  11. Group P. T. C., Calabrese C., Davidson N. R., et al. (2020). Genomic basis for RNA alterations in cancer. *Nature* **578**:129–136. DOI:10.1038/s41586-020-1970-0
  12. Hao K., Bosse Y., Nickle D. C., et al. (2012). Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8**:e1003029. DOI:10.1371/journal.pgen.1003029
  13. McKay J. D., Hung R. J., 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**:1126–1132. DOI:10.1038/ng.3892
  14. Yao D. W., O'Connor L. J., Price A. L., et al. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**:626–633. DOI:10.1038/s41588-020-0625-2
  15. Li Y. I., van de Geijn B., Raj A., et al. (2016). RNA splicing is a primary link between genetic variation and disease. *Science* **352**:600–604. DOI:10.1126/science.aad9417
  16. Wan L. D., Yu W. Y., Shen E. H., et al. (2019). SRSF6-regulated alternative splicing that promotes tumour progression offers a therapy target for colorectal cancer. *Gut* **68**:118–129. DOI:10.1136/gutjnl-2017-314983
  17. Tian J. B., Chen C., Rao M. L., et al. (2022). Aberrant RNA splicing is a primary link between genetic variation and pancreatic cancer risk. *Cancer Res.* **82**:2084–2096. DOI:10.1158/0008-5472.Can-21-4367
  18. Li B., Cai Y. M., Chen C., et al. (2023). Genetic variants that impact alternative polyadenylation in cancer represent candidate causal risk loci. *Cancer Res.* **83**:3650–3666. DOI:10.1158/0008-5472.Can-23-0251
  19. Wang E. T., Sandberg R., Luo S., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**:470–476. DOI:10.1038/nature07509
  20. de Figueiredo-Pontes L. L., Wong D. W., Tin V. P., et al. (2014). Identification and characterization of ALK kinase splicing isoforms in non-small-cell lung cancer. *J. Thorac. Oncol.* **9**:248–253. DOI:10.1097/JTO.000000000000050
  21. Zhang S., Mao M., Lv Y., et al. (2022). A widespread length-dependent splicing dysregulation in cancer. *Sci. Adv.* **8**:eabn9232. DOI:10.1126/sciadv.abn9232
  22. Bradley R. K. and Anczukow O. (2023). RNA splicing dysregulation and the hallmarks of cancer. *Nat. Rev. Cancer* **23**:135–155. DOI:10.1038/s41568-022-00541-7
  23. Raj T., Li Y. I., Wong G., et al. (2018). Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat. Genet.* **50**:1584–1592. DOI:10.1038/s41588-018-0238-1
  24. Zhang M., Chen C., Lu Z. Q., et al. (2023). Genetic control of alternative splicing and its distinct role in colorectal cancer mechanisms. *Gastroenterology* **165**:1151–1167. DOI:10.1053/j.gastro.2023.07.019
  25. Jin M., Liu B., Chen C., et al. (2023). Genome-wide splicing quantitative expression locus analysis identifies causal risk variants for non-small cell lung cancer. *Cancer Res.* **83**:1742–1756. DOI:10.1158/0008-5472.CAN-22-3184
  26. Gusev A., Lawrenson K., Lin X., et al. (2019). A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nat. Genet.* **51**:815–823. DOI:10.1038/s41588-019-0395-x
  27. Wang Y., Ding Y., Liu S., et al. (2023). Integrative splicing-quantitative-trait-locus analysis reveals risk loci for non-small-cell lung cancer. *Am. J. Hum. Genet.* **110**:1574–1589. DOI:10.1016/j.ajhg.2023.07.008
  28. Lonsdale J., Thomas J., Salvatore M., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**:580–585. DOI:10.1038/ng.2653
  29. Gamazon E. R., Wheeler H. E., Shah K. P., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**:1091–1098. DOI:10.1038/ng.3367
  30. Wainberg M., Sinnott-Armstrong N., Mancuso N., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**:592–599. DOI:10.1038/s41588-019-0385-z
  31. Wang F. R., Yang X. J., Ren Z. Y., et al. (2023). Alternative splicing in mouse brains affected by psychological stress is enriched in the signaling, neural transmission and blood-brain barrier pathways. *Mol. Psychiatry* **28**:4707–4718. DOI:10.1038/s41380-023-02103-1
  32. Van der Auwera G. A., Carneiro M. O., Hartl C., et al. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**:11.10.1–11.10.33. DOI:10.1002/0471250953.bi1110s43
  33. Browning B. L., Tian X. W., Zhou Y., et al. (2021). Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**:1880–1890. DOI:10.1016/j.ajhg.2021.08.005
  34. Browning B. L., Zhou Y. and Browning S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**:338–348. DOI:10.1016/j.ajhg.2018.07.015
  35. Javed N., Farjoun Y., Fennell T. J., et al. (2020). Detecting sample swaps in diverse NGS data types using linkage disequilibrium. *Nat. Commun.* **11**:3697. DOI:10.1038/s41467-020-17453-5.
  36. Deng Z. L., Münch P. C., Mreches R., et al. (2022). Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Res.* **50**:e60. DOI:10.1093/nar/gkac112.
  37. Dobin A., Davis C. A., Schlesinger F., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**:15–21. DOI:10.1093/bioinformatics/bts635
  38. Tarasov A., Vilella A. J., Cuppen E., et al. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31**:2032–2034. DOI:10.1093/bioinformatics/btv098
  39. Li Y. I., Knowles D. A., Humphrey J., et al. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**:151–158. DOI:10.1038/s41588-017-0004-9
  40. Robinson M. D., McCarthy D. J. and Smyth G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139–140. DOI:10.1093/bioinformatics/btp616
  41. Shabalin A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**:1353–1358. DOI:10.1093/bioinformatics/bts163
  42. Stegle O., Parts L., Piipari M., et al. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**:500–507. DOI:10.1038/nprot.2011.457
  43. Cingolani P., Platts A., Wang le L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**:80–92. DOI:10.4161/fly.19695
  44. Sakaue S., Kanai M., Tanigawa Y., et al. (2021). A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**:1415–1424. DOI:10.1038/s41588-021-00931-x
  45. Yang J., Lee S. H., Goddard M. E., et al. (2011). GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**:76–82. DOI:10.1016/j.ajhg.2010.11.011
  46. Genomes Project C., Auton A., Brooks L. D., et al. (2015). A global reference for human genetic variation. *Nature* **526**:68–74. DOI:10.1038/nature15393
  47. Giambartolomei C., Vukcevic D., Schadt E. E., et al. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**:e1004383. DOI:10.1371/journal.pgen.1004383
  48. Gusev A., Ko A., Shi H., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**:245–252. DOI:10.1038/ng.3506
  49. Seiler M., Peng S. Y., Agrawal A. A., et al. (2018). Somatic mutational landscape of splicing factor genes and their functional consequences across 33 cancer types. *Cell Rep.* **23**:282–296.e4. DOI:10.1016/j.celrep.2018.01.088
  50. Gu Z., Gu L., Eils R., et al. (2014). Circize Implements and enhances circular visualization in R. *Bioinformatics* **30**:2811–2812. DOI:10.1093/bioinformatics/btu393
  51. Szklarczyk D., Gable A. L., Lyon D., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**:D607–D613. DOI:10.1093/nar/gky1131
  52. Colaprico A., Silva T. C., Olsen C., et al. (2016). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**:e71. DOI:10.1093/nar/gkv1507
  53. Zhang Y., Yao X., Zhou H., et al. (2022). OncoSplicing: An updated database for clinically relevant alternative splicing in 33 human cancers. *Nucleic Acids Res.* **50**:D1340–D1347. DOI:10.1093/nar/gkac851
  54. Chen B., Khodadoust M. S., Liu C. L., et al. (2018). Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol. Biol.* **1711**:243–259. DOI:10.1007/978-1-4939-7493-1\_12
  55. Maeser D., Gruener R. F. and Huang R. S. (2021). oncoPredict: An R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. *Brief. Bioinform.* **22**:bbab260. DOI:10.1093/bib/bbab260
  56. Walker R. L., Ramaswami G., Hartl C., et al. (2020). Genetic control of expression and splicing in developing human brain informs disease mechanisms (vol 179, 750.e1, 2019). *Cell* **181**:745–745. DOI:10.1016/j.cell.2020.04.016
  57. Zhang Z., Gao X., Ren X., et al. (2023). Characterization of intratumor microbiome in cancer immunotherapy. *The Innovation* **4**:100482. DOI:10.1016/j.xinn.2023.100482
  58. Qi T., Wu Y., Fang H. L., et al. (2022). Genetic control of RNA splicing and its distinct role in complex trait variation. *Nat. Genet.* **54**:1355–1363. DOI:10.1038/s41588-022-01154-4
  59. Kahles A., Lehmann K. V., Toussaint N. C., et al. (2018). Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* **34**:211–224.e6. DOI:10.1016/j.ccell.2018.07.001
  60. Strober B. J., Elorbany R., Rhodes K., et al. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**:1287–1290. DOI:10.1126/

- science.aaw0040
61. Kim E. K. and Choi E. J. (2010). Pathological roles of MAPK signaling pathways in human diseases. *Biochim. Biophys. Acta* **1802**:396–405. DOI:10.1016/j.bbdis.2009.12.009
  62. Roskoski R., Jr. (2019). Targeting ERK1/2 protein-serine/threonine kinases in human cancers. *Pharmacol. Res.* **142**:151–168. DOI:10.1016/j.phrs.2019.01.039
  63. Martinez N. M., Pan Q., Cole B. S., et al. (2012). Alternative splicing networks regulated by signaling in human T cells. *RNA* **18**:1029–1040. DOI:10.1261/rna.032243.112
  64. O'Connor B. P., Danhorn T., De Arras L., et al. (2015). Regulation of toll-like receptor signaling by the SF3a mRNA splicing complex. *PLoS Genet.* **11**:e1004932. DOI:10.1371/journal.pgen.1004932
  65. Schaub A. and Glasmacher E. (2017). Splicing in immune cells-mechanistic insights and emerging topics. *Int. Immunol.* **29**:173–181. DOI:10.1093/intimm/dxx026
  66. Kong Y., Yu J., Ge S., et al. (2023). Novel insight into RNA modifications in tumor immunity: Promising targets to prevent tumor immune escape. *The Innovation* **4**:100452. DOI:10.1016/j.xinn.2023.100452
  67. Matter N., Herrlich P. and Konig H. (2002). Signal-dependent regulation of splicing via phosphorylation of Sam68. *Nature* **420**:691–695. DOI:10.1038/nature01153
  68. Weg-Remers S., Ponta H., Herrlich P., et al. (2001). Regulation of alternative pre-mRNA splicing by the ERK MAP-kinase pathway. *EMBO J.* **20**:4194–4203. DOI:10.1093/emboj/20.15.4194
  69. Hassn Mesrati M., Syafruddin S. E., Mohtar M. A., et al. (2021). CD44: A multifunctional mediator of cancer progression. *Biomolecules* **11**:1850. DOI:10.3390/biom11121850
  70. Ishimoto T., Nagano O., Yae T., et al. (2011). CD44 variant regulates redox status in cancer cells by stabilizing the xCT subunit of system xc(-) and thereby promotes tumor growth. *Cancer Cell* **19**:387–400. DOI:10.1016/j.ccr.2011.01.038
  71. Ni J., Cheung B. B., Beretov J., et al. (2020). CD44 variant 6 is associated with prostate cancer growth and chemo-/radiotherapy response. *Exp. Cell Res.* **388**:111850. DOI:10.1016/j.yexcr.2020.111850
  72. Zhao P., Damerow M. S., Stern P., et al. (2013). CD44 promotes Kras-dependent lung adenocarcinoma. *Oncogene* **32**:5186–5190. DOI:10.1038/onc.2012.542
  73. Mills A. A., Zheng B. H., Wang X. J., et al. (1999). p63 is a p53 homologue required for limb and epidermal morphogenesis. *Nature* **398**:708–713. DOI:10.1038/19531
  74. Yi M., Tan Y. X., Wang L., et al. (2021). TP63 links chromatin remodeling and enhancer reprogramming to epidermal differentiation and squamous cell carcinoma development (vol 77, pg 4325, 2020). *Cell. Mol. Life Sci.* **78**:7915. DOI:10.1007/s00018-021-03982-9
  75. Taylor W., Deschamps S., Reboutier D., et al. (2022). The splicing factor PTBP1 represses  $\gamma$  isoform production in squamous cell carcinoma. *Cancer Res. Commun.* **2**:1669–1683. DOI:10.1158/2767-9764.Crc-22-0350
  76. Beyer U., Moll-Rocek J., Moll U. M., et al. (2011). Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. *Proc. Natl. Acad. Sci. U. S. A.* **108**:3624–3629. DOI:10.1073/pnas.1016201108
  77. Gooding C. and Smith C. W. J. (2008). Tropomyosin exons as models for alternative splicing. *Adv. Exp. Med. Biol.* **644**:27–42. DOI:10.1007/978-0-387-85766-4\_3
  78. Okumura M., Kondo S., Ogata M., et al. (2005). Candidates for tumor-specific alternative splicing. *Biochem. Biophys. Res. Commun.* **334**:23–29. DOI:10.1016/j.bbrc.2005.06.057
  79. Gardina P. J., Clark T. A., Shimada B., et al. (2006). Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* **7**:325. DOI:10.1186/1471-2164-7-325
  80. Thorsen K., Sorensen K. D., Brems-Eskildsen A. S., et al. (2008). Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Mol. Cell Proteomics* **7**:1214–1224. DOI:10.1074/mcp.M700590-MCP200
  81. Ding J. Y., Li C. Q., Cheng Y. W., et al. (2021). Alterations of RNA splicing patterns in esophagus squamous cell carcinoma. *Cell Biosci.* **11**:36. DOI:10.1186/s13578-021-00546-z
  82. Langer W., Sohler F., Leder G., et al. (2010). Exon array analysis using re-defined probe sets results in reliable identification of alternatively spliced genes in non-small cell lung cancer. *BMC Genomics* **11**:676. DOI:10.1186/1471-2164-11-676
  83. Cooke M., Kreider-Letterman G., Baker M. J., et al. (2021). FARP1, ARHGEF39, and TIAM2 are essential receptor tyrosine kinase effectors for Rac1-dependent cell motility in human lung adenocarcinoma. *Cell Rep.* **37**:109905. DOI:10.1016/j.celrep.2021.109905
  84. Yin Y. T., Long J. L., Sun Y. Q., et al. (2018). The function and clinical significance of eIF3 in cancer. *Gene* **673**:130–133. DOI:10.1016/j.gene.2018.06.034
  85. Gomes-Duarte A., Lacerda R., Menezes J., et al. (2018). eIF3: A factor for human health and disease. *RNA Biol.* **15**:26–34. DOI:10.1080/15476286.2017.1391437
  86. Bhat M., Robichaud N., Hulea L., et al. (2015). Targeting the translation machinery in cancer. *Nat. Rev. Drug Discovery* **14**:261–278. DOI:10.1038/nrd4505
  87. Rothe M., Ko Y., Albers P., et al. (2000). Eukaryotic initiation factor 3 p110 mRNA is overexpressed in testicular seminomas. *Am. J. Pathol.* **157**:1597–1604. DOI:10.1016/S0002-9440(10)64797-9
  88. Song N., Wang Y., Gu X. D., et al. (2013). Effect of siRNA-mediated knockdown of gene on survival of colon cancer cells. *Journal of Zhejiang University-Science B* **14**:451–459. DOI:10.1631/jzus.B1200230
  89. Lee A. S. Y., Kranzusch P. J., Doudna J. A., et al. (2016). eIF3d is an mRNA cap-binding protein that is required for specialized translation initiation. *Nature* **536**:96–99. DOI:10.1038/nature18954
  90. Zhao M., Gao Y., Wang L., et al. (2013). Overexpression of integrin-linked kinase promotes lung cancer cell migration and invasion via NF-kappaB-mediated upregulation of matrix metalloproteinase-9. *Int. J. Med. Sci.* **10**:995–1002. DOI:10.7150/ijms.5963
  91. Jia Z. Y. (2015). Role of integrin-linked kinase in drug resistance of lung cancer. *Oncotargets and Therapy* **8**:1561–1565. DOI:10.2147/Ott.S81447
  92. Tseng P. C., Chen C. L., Shan Y. S., et al. (2014). An increase in integrin-linked kinase non-canonically confers NF-kappaB-mediated growth advantages to gastric cancer cells by activating ERK1/2. *Cell Commun. Signal.* **12**:69. DOI:10.1186/s12964-014-0069-3
  93. Liang C. H., Chiu S. Y., Hsu I. L., et al. (2013). alpha-Catulin drives metastasis by activating ILK and driving an alphavbeta3 integrin signaling axis. *Cancer Res.* **73**:428–438. DOI:10.1158/0008-5472.CAN-12-2095
  94. Chu P. C., Yang M. C., Kulp S. K., et al. (2016). Regulation of oncogenic KRAS signaling via a novel KRAS-integrin-linked kinase-hnRNPA1 regulatory loop in human pancreatic cancer cells. *Oncogene* **35**:3897–3908. DOI:10.1038/onc.2015.458
  95. Nikou S., Arbi M., Dimitrakopoulos F. D., et al. (2020). Integrin-linked kinase (ILK) regulates KRAS, IPP complex and Ras suppressor-1 (RSU1) promoting lung adenocarcinoma progression and poor survival. *J. Mol. Histol.* **51**:385–400. DOI:10.1007/s10735-020-09888-3
  96. Watahiki A., Waki K., Hayatsu N., et al. (2004). Libraries enriched for alternatively spliced exons reveal splicing patterns in melanocytes and melanomas. *Nat. Methods* **1**:233–239. DOI:10.1038/nmeth719

## FUNDING AND ACKNOWLEDGMENTS

This work was supported by Healthy Zhejiang One Million People Cohort (K-20230085), the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (2019R01007), Zhejiang Key Laboratory of Intelligent Preventive Medicine (2020E10004), and the Key Research and Development Program of Zhejiang Province(2020C03002). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We appreciate the support and cooperation of all patients and their families who participated in the project.

## AUTHOR CONTRIBUTIONS

X.W. and D.Z. contributed to the study conceptualization and supervised the project. H.Z. and D.Z. made the analysis plan and conducted the statistical analyses. M.W. collected the lung cancer samples, and K.W. contributed to the sequencing. Q.T., X.X. verified the underlying data and provided suggestions for the research. M.W., L.S., Y.T., W.Z. were responsible for data collection. Y.Z. contributed to sample processing and management work. H.Z., H.T., D.Z. and X.W. drafted the manuscript. All authors revised the manuscript. All authors contributed to and approved the final manuscript.

## DECLARATION OF INTERESTS

Xifeng Wu is an Editorial Board member of The Innovation Medicine. She was blinded from reviewing or making final decisions on the manuscript. The article was subject to the journal's standard procedures, with peer review handled independently of this Editorial Board member and their research groups. The other authors declare no competing interests.

## ETHICAL STATEMENT AND PATIENT CONSENT

Totally 417 patients diagnosed with NSCLC were recruited from the second affiliated hospital Zhejiang University School of Medicine (SAH), Zhejiang University (Hangzhou, China) during 2015–2020. Informed consent was obtained from all patients, and the study protocol was approved by the Institutional Review Board of SAH (2020-829).

## DATA AND CODE AVAILABILITY

The results of GWAS for NSCLC have been deposited in the National Genomics Data Center (NGDC, <https://www.cnbc.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 the 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. The code for this paper is available on the GitHub ([https://github.com/zdangm/NSCLC\\_splicing](https://github.com/zdangm/NSCLC_splicing)).

## SUPPLEMENTAL INFORMATION

It can be found online at <https://doi.org/10.59717/j.xinn-med.2025.100111>