Global profiling of alternative splicing in non-small cell lung cancer reveals novel histological and population differences

Saman Zeeshan

Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey

<https://orcid.org/0000-0001-9278-6204>

Rony A. Mendeley

Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute. NIH

Adriana Zingone

Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute. NIH

Hossein Khiabani

Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey

<https://orcid.org/0000-0003-1446-4394>

Sharon R. Pine (✉ piners@cinj.rutgers.edu)

Rutgers Cancer Institute of New Jersey, Rutgers, The State University of New Jersey

<https://orcid.org/0000-0001-5318-0277>

Brid M. Ryan (✉ ryanb@mail.nih.gov)

Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute. NIH

<https://orcid.org/0000-0003-0038-131X>

Research Article

Keywords:

Posted Date: January 18th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1272617/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Lung cancer is one of the most frequently diagnosed cancers with higher incidence and mortality rates in African-American (AA) men compared to European-American (EA) men. Here, we report high-confidence alternative splicing (AS) events from high-throughput, high-depth total RNA-sequencing of lung tumors and non-tumor adjacent tissues (NATs) from two independent cohorts of patients with adenocarcinoma and squamous cell carcinoma. We uncover tumor subtype- and population-specific AS events different between AAs and EAs, associated with oncogenic signaling pathways, cell surface proteins, and cancer driver genes. We highlight significant AS events in SYNE2 specific to LUAD in both populations as well as those in CD44 from EAs and TMBIM6 from AAs specific to LUAD. Our large survey of lung tumors presents a rich data resource that can help uncover new therapeutic vulnerabilities that potentially advance health equity.

Introduction

Lung cancer remains the leading cause of cancer-related deaths in the US [1]. Non-small cell lung cancer (NSCLC) forms the larger group, with lung adenocarcinoma (LUAD) as the most common subtype, followed by lung squamous cell carcinoma (LUSC). African American (AA) men are about 15% more likely to develop lung cancer than European American (EA) men [2]. Disparities also exist in lung cancer survival, where AAs consistently have worse outcomes, at least partly attributed to access to care and treatment decisions [3][4]. In recent years, vast progress has been made in understanding the global molecular architecture of lung tumors, but AAs have been largely underrepresented in these studies. The inclusion of minority populations has the potential to reveal novel signatures and tumor vulnerabilities, and advance cancer health equity. We previously reported population differences in lung tumor biology, including circulating inflammation proteins, gene expression signatures, somatic mutations, and ancestry-associated single nucleotide polymorphisms [5–14]. However, work remains to complete the global comparison of tumor biology across racial and ethnic groups.

Recent genome-wide studies suggested that alternative splicing (AS), or aberrant splicing of precursor (pre)-mRNA, is associated with critical oncogenic processes [15, 16], indicating that AS plays an important, but largely underappreciated, role in cancer. However, previous work has been largely saturated in the analysis of splicing from low coverage mRNA-seq [17] rather than total RNA-seq [18] which permits the investigation of RNA-processing and transcriptional regulation. Studies indicate that isoform-level expression can more precisely define tumor molecular subtypes [19, 20]. In addition to revealing the role of AS events as potential cancer drivers, AS profiling can also uncover novel molecular biomarkers and neoantigens for cancer diagnosis, treatment, and prognosis [21–23]. Indeed, recent work suggests that AS is a key driver of novel neoantigens that influences immune checkpoint inhibitor sensitivity [24] and escape from immunotherapy [25].

Aberrant AS events have been linked to tumorigenesis in cancers of the lung, gastrointestinal tract, brain, ovary, breast, and other tissues [26–33], though the majority of these studies did not include populations

of African descent. Recently, Deveaux *et al.* reported differences in AS between AAs and EAs in a small subset of LUSC tumors, determined by transcriptome array [34]. However, to date, differences in lung tumor-specific AS events between racial and histological groups, the impact of AS events in cancer driver genes on health disparities, or the association between tumor-specific AS events and certain external exposures, such as smoking, have remained unexamined. Our goal was to study the genome-wide differences in AS between LUAD and LUSC tumors from AAs and EAs and, by generating high-throughput, high-depth, total RNA-seq on tumor and non-tumor adjacent tissue (NAT) samples from two independent datasets, to identify such differences with high confidence. These new data indicate a complex system of AS that furthers our understanding of population differences in the biology underlying NSCLC.

Results

Histology-specific differences in splicing events in lung cancer

We first examined the AS events in tumors compared with their respective NATs from the discovery dataset and identified 990 AS events in 773 genes in LUAD, and 668 events in 529 genes in LUSC (Fig. 1a, 1b; Tables S3, S4). The majority of these events were in protein coding genes (LUAD 90%, LUSC 91%). The AS events represented all five major classes, and the proportions of AS events across the classes were similar between LUAD and LUSC, with SE as the most common and MXE as the least common (Fig. 1c). Multiple AS classes were observed for the same gene exons. Approximately 60% of the AS events (30% of genes) were overlapping between LUSC and LUAD (Fig. 1d).

To validate our observations, we generated a second dataset from independent tumor and NAT samples. Overall, 40% (398 of 990) and 50% (347 of 688) of detected AS events from the discovery dataset were also detected in the validation dataset for LUAD and LUSC, respectively, and represented all five major classes of AS (Fig. 1e; Tables S3, S4). Of these, 209 validated AS events in 160 genes were observed in both LUAD and LUSC (Table S5), suggesting that both histological subtype-specific and shared splicing events in NSCLC are common.

Most validated differentially spliced genes (DSGs) were involved in single splicing events in LUAD (82%) and LUSC (81%), while approximately 18% had two or more splice events. Twenty-one genes with three or more events, were common to LUAD and LUSC, including *TPM1* (with 9 and 8 AS events, respectively), *VEGFA* (with 6 and 6 AS events, respectively) and *CTNND1* (with 6 and 4 AS events, respectively) (Tables S3-S5).

To identify the major signaling pathways affected by aberrant AS in NSCLC, we performed IPA analysis on the validated DSGs (Fig. 1f). The DSGs common to both LUAD and LUSC (23 genes) were enriched in the *IL-15* production, *ILK*, *STAT3*, and *VEGF* signaling pathways. Validated DSGs specific to LUAD were enriched in protein kinase A and HIPPO signaling pathways, whereas validated DSGs specific to LUSC were enriched in clathrin-mediated endocytosis signaling and remodeling epithelial adherens junctions

pathways (Table S6). These data suggest that while AS contributes to common pathway perturbation in NSCLC, distinct pathways are also perturbed by AS in each clinical subtype.

Because gene expression patterns in NATs may be influenced by the type of tumor tissue or cell of origin [35], we directly compared the tumor tissues in LUAD vs. LUSC, without consideration of NATs. We identified 94 AS events in the discovery dataset, 15 of which were validated in the second dataset (Table S7). Among these, four AS events in *CD44* had a significantly lower PSI in LUAD compared with LUSC. This indicates the presence of subtype specific tumor isoforms. In contrast, when comparing NATs, 88 detected AS events were observed in the tissues adjacent to LUAD or LUSC tumors (Table S8); only two of these, in *GLS2* and *LRP8*, were observed in the validation dataset. Neither *GLS2* nor *LRP8* were found differentially spliced in LUAD versus LUSC tumor tissue. This suggests that most of the AS events observed in NSCLC, particularly those that differ by histological subtype, are somatic and tumor-specific in nature.

RNA splicing differences in NSCLC between African Americans and European Americans

As we previously reported on key differences in genomic and gene expression changes in NSCLC from AAs vs. EAs [5–14], we explored potential population-related differences in AS. When populations were analyzed separately, over 1,000 AS events from more than 800 genes were identified in LUAD compared with NATs, 33% of which were also present in the validation datasets (Fig. 2a; Tables S9, S10). These validated AS events had similar distributions of splicing classes between the AA and EA populations (Fig. 2b). More than half (54%) of AS events were shared between the populations in LUAD (Fig. 2c, 2d). Half of the genes with more than three AS events (13 of 25) were unique to AAs or EAs (e.g., *TPM1* and *VEGFA* were shared, *ADAM15* and *FGFR2* were unique to AAs, and *IL32* and *STYL2* were unique to EAs), further highlighting population differences (Table S11).

In LUSC, we identified more than 300 validated AS events in both AAs and EAs compared with NATs, representing approximately 40% of the AS events from the discovery datasets (Fig. 3a, 3b; Tables S16, S17). There were both shared (~50%) and distinct (~50%) splicing events in AAs and EAs (Fig. 3d) suggesting that both common and population-specific AS events occur in all modes of splicing in LUSC as well (Fig. 3c). Again, we saw multiple splicing events in both AAs and EAs within *TPM1* and *VEGFA*. Genes with more than three AS events were also unique to AAs or EAs in LUSC, such as *DLG1* and *METTL26* in AAs and *CTNND1* and *MACF1* in EAs, further highlighting key shared and unique population-related splicing events in LUSC (Table S18).

To understand population differences in the cellular pathways affected by AS in AAs and EAs with LUAD, we analyzed on the genes with validated AS events (Fig. 2f). In AAs, 37 significant pathways were affected by AS, including HGF related Met receptor signaling, and CSDE1 pathway regulation (Table S12). In EAs, we observed 40 pathways, including those associated with *IL7*, *STAT3*, and *ERK/MAPK* signaling (Table S12). Twenty-two pathways commonly perturbed in AAs and EAs included *IL-15*, *ILK*, and *VEGF*

signaling. In LUSC, there were 25 pathways affected by AS only in AAs with LUSC including *VEGF*, iron hemostasis, and *GP6* signaling (Table S20). In EAs, we observed 32 pathways regulated related to epithelial remodeling and HIPPO signaling (Table S20). Sixteen pathways were shared between the populations (Fig. 3f).

When comparing AS events between AA and EA LUAD tumors, not accounting for NATs, we identified 246 AS events, of which 45 were validated (Fig. 2e; Table S13). We also compared AS events in the NATs between the two populations and identified 325 events, of which 44 were observed in the validation dataset (Table S14). Thirteen validated events overlapped between AA and EA LUAD tumors and NATs (Table S15), suggesting that these events are not tumor specific. In LUSC tumors, we identified 226 AS events between AAs and EAs, of which 37 were validated (Fig. 3e; Table S20). Nine of these population specific AS events were observed in both LUAD and LUSC, suggesting that they are independent of histology. We also compared AS events in the NATs from LUSC between the populations and identified 238 events, and 42 of these events were validated (Table S21). We found 11 validated events overlapped between AA and EA LUSC tumors and NATs (Table S22), suggesting that these events are non-tumor specific. Ultimately, the results emphasize the importance of quantitatively profiling AS in cancer and the potential of this approach to uncover novel, and perhaps population-specific, tumor biology.

Alternative Splicing of Driver Genes and Surface Proteins in NSCLC

Growing evidence has demonstrated that AS is involved in oncogenic processes including proliferation, apoptosis, hypoxia, angiogenesis, immune escape and metastasis [36]. Specific isoforms can act as cancer drivers [37]. To determine which cancer-specific splicing events may drive lung cancer or reflect the tissue of origin, we selected events from COSMIC cancer driver genes. We identified AS events with $\geq 20\%$ PSI differences shared between the subtypes, including those effecting *AFDN*, *FGFR2*, *GCOM1*, and *PPFIBP1* (Fig. 4a), while there were also validated AS events in cancer driver genes with a $\geq 10\%$ PSI difference unique to LUAD and LUSC (Fig. 4a; Table S23). Similarly, we identified 17 validated events in cancer driver genes that were shared between AAs and EAs (Fig. 4b; Table S24). Of note, an SE event in *AFDN* was validated in both histological subtypes. (Fig. 4c, 4d). Further, this event in exon 16 in *AFDN* had a $>20\%$ PSI difference between tumor and NATs (PSI = 0.59 and 0.84, respectively) and was significantly more spliced out from tumor among both AAs and EAs for both histologies (Fig. 4d; Table S24). This event occurs in the protein's dilute DIL domain, responsible for actin stress fiber formation involved in tumor induced cell migration, invasiveness, and growth [38].

Splicing variants in genes encoding for surface proteins have been recognized as a class of therapeutic targets called alternative tumor-specific antigens, which provide potential surface targets for antitumor therapy. These neoantigens have a distinct advantage over other classes of tumor antigens because they are twice as common as single-nucleotide variants across the TCGA pan-cancer dataset and are highly correlated with *PD-L1* expression [39]. To determine if surface proteins associate with a certain population profile, we first identified validated events from surface molecules in LUAD and LUSC (Fig. 5a;

Table S25). Of these, 26 involved the extracellular region of the protein and among these, eight involved extracellular along with transmembrane region, while seven events spanned across the extracellular, transmembrane and the cytoplasmic region. Fourteen events occurred exclusively in the cytoplasmic region, four of which involved an alteration in a signal peptide. Seven events occurred exclusively in the transmembrane region which may have a role in structural modification and localization. A notable cell surface event specific to LUAD with a PSI difference >20% between tumor vs. NAT was in *SYNE2* (Fig. 5b). The SE event in *SYNE2*, also known as Nesprin-2, occurred in exon 107 with respect to reference transcript and was validated in LUAD only. In addition, this event involved splicing out 159 bp (53 amino acids) of the KASH domain in tumor cells, with a nearly 30% PSI difference between tumor and NATs (PSI = 0.32 and 0.59, respectively). This event was validated in LUAD from both AAs and EAs (Fig. 5c).

An additional event worth of note occurred in *CD44* (Fig. 5a, 6a). In comparison to the full-length reference transcript, the loss of exon 11 leads to splicing out of 132bp (44 amino acids) in the tumor isoform (Fig. 6b). Skipping of exon 11 had a 12% PSI difference between tumors and NATs (PSI = 0.47 and 0.61 respectively) which was validated in LUAD (Fig. 6c). In contrast, skipping of exon 11 was the reverse for LUSC, where there was significantly less skipping in tumor as compared to NAT, and the magnitude of the PSI difference in LUSC was small (<10%). These observations were extended similarly to only EAs in LUAD tumors (Fig. 6c). In LUSC, noteworthy events included *CD47* and *EPHA1*. These data suggest that an increased frequency of *CD44* exon loss may drive splicing activity in LUAD that is enriched in AAs. Another SE event in *TMBIM6* was evidently validated only in LUAD (Supplementary Fig. 1c).

Alternative splicing has also emerged as a central regulatory process during the epithelial–mesenchymal transition (EMT) controlled by splicing factors. Differentially expressed splicing factors may be responsible for regulating tumor-associated splicing events in cancer tissues [40–42]. So, we also explored EMT markers and RNA binding proteins (RBP) for their association with tumor subtypes (Supplementary Fig. 1a, 1b; Tables S27 and S28). An SE event was validated only in LUSC in the EMT gene, *FIP1L1* (FIP1-like 1 gene), involved in oncogenic activity through genomic fusion events [43] (Supplementary Fig. 1d). We also noted involvement of EMT–associated splicing factors *MBNL1* and *MBNL2* in NSCLC tumor biology while *hnRNPC* and *hnRNPL* were observed in only regulating LUAD (Supplementary Fig. 1a, 1b; Tables S27, S28).

Effect of environmental and host factors on alternative splicing in lung tumor tissues

Although lung carcinogenesis is a multifactorial process driven by exogenous exposures (e.g., cigarette smoking), inherited genetic variations, and an accumulation of somatic genetic events, our results suggest that it is characterized by population differences in biology. To explore whether host factors impact these differences, we performed linear regression to model PSI as a continuous outcome variable based on multiple predictor variables (i.e., environmental and host factors, including age, gender, smoking and aspirin use).

First, we assessed how many events were associated with race in both tumor and NAT samples. Following adjustment for age and gender, at $P<0.01$, we detected 174 AS events associated with race (Table S29). Smoking intensity is known to modulate the transcriptome specifically regulating immunity, inflammation, and cell death gene networks, and is a key demographic variable that differs between populations with a longer cumulative exposure to tobacco in AAs leading to a greater cancer risk [44]. After adjusting for age and gender, 623 AS events were significantly associated ($P<0.01$) with the number of cigarettes smoked per day (smoking intensity) and for the smoking duration (smoking years) (Table S30). Eight hundred and thirteen events were significantly associated with smoking intensity following adjustment for race, age, and gender at $P<0.01$ (Table S31), and 882 events were associated with smoking duration in tumor tissues after adjustment for race, age, and gender (Table S32). Notably, 57% of these events were shared between smoke intensity and duration indicating that majority of the AS is modulated similarly, while the remaining 43% unique events are modulated independently by intensity and duration. Collectively, this suggests that some population-associated AS events in lung cancer are attributable to population differences in smoking patterns.

We previously reported a relationship between aspirin and lung cancer survival among AAs, but not EAs [45], which suggested that aspirin differentially influenced host tumor biology. We found 1130 AS events were associated with aspirin use ($P<0.01$) in tumor tissues after adjustment for age and gender (Table S33). These findings suggest that exposure-associated alterations observed in lung tumor tissue may reflect, in part, the altered transcriptome observed in lung tumors and NATs to these tumors, supporting the concept that external exposures such as aspirin use have strong relationships to gene-splicing changes in lung tumor tissues.

Discussion

Complexity of the splicing network is influenced by splicing machinery, splicing regulators, and RNA structure. The dysregulation of alternative splicing is closely associated with tumor progression and leads to the formation of multiple isoforms with diverse functions. Indeed, recent work has highlighted both the functional impact of AS and its implications for immune and targeted therapies. Here, we report the first comprehensive genome-wide data analysis of isoform-level AS and define both population and histological differences in lung cancer. A key strength of our study is the generation of two independent discovery and validation datasets that included samples from AAs and EAs patients with LUAD and LUSC histological sub-types and the inclusion of matched NATs. Our findings on splicing modulations by environmental and host factors are consistent with previous studies linking exogenous exposures with AS modulation [46–48]. Tobacco exposure and aspirin use are two exposures that demonstrate striking differences in AA and EA [44, 45], therefore it is notable that these exposures also leave a footprint in terms of tumor biology and highlights the importance of considering an individual's lifestyle and exposures when assessing tumor biology and inter-tumor heterogeneity.

We tested the hypothesis that AS events define histology-specific diver-like properties and observed many alternatively spliced cancer driver genes including Afadin/*AF-6* (*AFDN*) with the canonical isoform

ENST00000351017.8, which may have a role in tumor progression and invasion [38]. Another similar AS driver-like event was observed in *SYNE2*, a multi-isomeric cancer gene. Here we observed a differential exon inclusion rate (PSI) favoring exon skipping in exon 107. The latter is involved in a domain required for scaffolding for protein-protein interactions, and thus may influence the shape and migration properties of cells [49].

In addition, our analysis also uncovers splicing changes in cell surface proteins that may completely transform their structure or function are common in lung cancer. For example, the human *CD44* gene encodes a polymorphic group of transmembrane glycoproteins generated by AS, which have been reported to be associated with poor prognosis in NSCLC [50, 51]. We detected skipped exon 11 in reference to the full-length *CD44* transcript (ENST00000428726.8) in LUAD but not in LUSC. This event may result in an altered extracellular region in the variant domain; thus impacting the structure of the protein. The alternatively spliced *CD44* variants that have exon 11 spliced-in (*CD44v6*) are required for activation of the receptors. Cooperation between *CD44v6* and *MET* is essential for tumor growth and metastasis of cancer cells [52], and *CD44v6* may be involved in promoting the properties of lung tumor stem-like cells [53]. Further exploration of the function and impact of this isoform in LUAD is warranted. Similarly, dysregulated AS within tumor microenvironment may be orchestrated by a limited number of cancer specific EMT-associated splicing factors [41]. For example, we identified an EMT-associated splicing event in *FIP1L1*, which is known to be associated with two oncogenic fusion genes: *FIP1L1*-retinoic acid receptor alpha (*RARA*) and *FIP1L1*-platelet-derived growth factor receptor alpha (*PDGFRA*) [43].

Our approach also allowed us to elaborate on the variations in the lung cancer transcriptome as we identified consistent and recurrent AS events that can have an impact on protein functions through gain or loss of exon(s). A relatively small change in transcript abundance can be sufficient to trigger an oncogenic effect in cells [54] and our data clearly show that cancer cells frequently exploit AS for their own advantage.

In addition to adding granularity to the layers of aberrant molecular biology in cancer, analysis of alternative splicing has led to its development as a therapeutic target. For example, steric block oligonucleotides have been developed that hide specific sequences within a target transcript from the spliceosome and thereby interfere with transcript RNA-RNA and / or RNA-protein interactions, leading to the inclusion, or exclusion, of exons [55]. An example of this drug class, Nusinersen, is FDA-approved for the treatment of muscular dystrophy and alters splicing of the *SMN2* gene. Another class of ASO, steric blockers, promote isoform switching and can diminish the expression of harmful protein isoforms and / or promote the expression of beneficial ones [19, 30]. Nearly 200 oligonucleotide therapeutics for cancer using ASOs are in phase 1 / 2 clinical trials [56]. Our data highlighting cancer-associated exons in *CD44* suggest that this splicing event could also be targeted with this technology.

In addition to driving proteomic diversity, the inclusion, and exclusion, of exon splicing 'errors' in cancer cells can also expand the neoantigen repertoire and enhance immunotherapy sensitivity [24]. Spliced

exons with a high fold difference between tumor and NATs, in genes that encode for proteins exposed on the cell surface have the potential to be specific targets for immunotherapy. These neoepitopes can also help us understand mechanism of antigen escape to improve drug sensitivity [57, 58]. Our findings provide the specificity and therapeutic window to explore the potential of AS to mediate response to immunotherapy. The PSI levels of the validated events show a maximum of 30% difference between the tumors and NATs in the subtypes. Interestingly, LUAD underwent splicing at a much higher frequency than LUSC, and future studies are needed to determine if these subtype differences translate to clinical implications for immunotherapy. Indeed, mounting evidence indicates that AS-derived neoepitopes can be promising targets for inhibiting tumor growth and enhancing checkpoint blockade [59, 60].

In summary, we performed a systematic characterization of the complexity of isoform-level alternative splicing variation in lung cancer. We identified high-confidence, novel alternative splicing (AS) biomarkers with differential splicing patterns between tumor and NATs in discovery and validation cohorts. Our approach afforded sensitive detection of relatively pronounced tumor-associated AS events for cancer driver genes and surface molecules that occur in the majority of patients in lung cancer from AAs and EAs in histological subtypes. Our inclusion of tumor samples from EAs and AAs and the biological differences we subsequently identified, underscores the importance of increasing the representation of minority and under-represented populations in cancer genomics, as such studies not only expand our knowledge of cancer biology but advance health equity and can uncover new cancer vulnerabilities. Our study has some potential limitations. Accurate determination of differential transcript usage in genes with many isoforms likely requires high sequencing coverage and sufficient samples per population [19]. This was, in part, mitigated by our use of total RNAseq in independent datasets at high coverage and depth in tumors from AAs and EAs with matching non-tumor adjacent tissues, and across histological sub-types. Our study signifies the potential to further define lung cancer subtypes through splice variations and provides a new path toward a more personalized approach to therapy.

Methods

Patient samples

Patients with NSCLC and living in the Baltimore Metropolitan area were prospectively recruited from seven Baltimore City hospitals to the ongoing NCI-MD Case Control Study, as described previously [11]. This study was conducted in accordance with the Declaration of Helsinki. Institutional review boards at all participating institutions approved the study and written informed consent was obtained from all patients. Eligibility was based on patients that had a histologically confirmed NSCLC diagnosis, had been born in the United States, was a current resident of the state of Maryland, spoke English well enough to be interviewed, was physically and mentally capable of performing the interview (i.e., must be able to hear the interviewer, mentally comprehend the interviewer's questions, and verbally respond), and did not currently reside in an institution such as a prison, nursing home, or shelter. Eligible participants that self-described as non-Hispanic AA or non-Hispanic EA received a structured in-person interview. Clinical and questionnaire data were obtained from medical records and pathology reports. Fresh human lung tumors

and non-tumor adjacent lung tissues (NAT) were obtained from patients directly after surgery. After resection, each tumor was macroscopically dissected by a pathologist to remove normal tissue. The tissues were transferred to sample collection tubes, flash frozen, and stored at -80°C . Specimens were transported to the NCI on dry ice within 24 hours and stored at -80°C until molecular analyses were performed.

Discovery samples

Initially, 75 evaluable primary tumors from 35 AAs and 40 EAs and 77 NATs were used in this study. Samples included 38 primary LUAD tumors and 37 primary LUSC tumors. Of these, 54 tumor/NATs were match paired. (Table S1).

Validation samples

This dataset consisted of 191 tissues including 94 evaluable tumor tissues from 23 AAs and 71 EAs and 95 NATs. Samples included 60 primary LUAD tumors, 25 primary LUSC tumors, and 10 other histological subtypes. Of these 93 tumors had patient-matched NAT samples (Table S2).

RNA Extraction

Total RNA was extracted using Trizol according to the manufacturer's instructions. RNA quality was assessed on a bioanalyser before including for sequencing.

RNA Sequencing

Libraries were prepared according to the TruSeq Stranded Total RNA kit. cDNA libraries that passed size and purity criteria were retained for sequencing. Sequencing for the discovery dataset was conducted using the Illumina TruSeq v4.0 chemistry (Illumina, Inc., San Diego, CA), paired-end for 126 cycles and at a read length of 125 bp. Sequencing for the validation dataset was conducted using Illumina HiSeq3000/4000 chemistry, paired-end for 150 cycles and at a read length of 150 bp. The quality of raw RNA-seq data was evaluated using FastQC v0.11.7

(<https://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>). Paired end raw reads were aligned to the human reference genome (hg38) using HISAT2 software (version 2.1.0) [61].

Differential alternative splicing analysis

To assess the differential splicing landscape, replicate multivariate analysis of transcript splicing (rMATS, version 3.2.5) [62] was used to identify differential AS events. The percent spliced in (PSI) value, defined as the ratio of inclusive reads over inclusive and exclusive reads normalized by read length and scored from zero to one, as well as associated statistics, were computed for each identified splicing event. Five AS classes were assessed: skipped exon (SE), retained intron (RI), alternative 5'splice site (A5SS), alternative 3' splice site (A3SS), and mutually exclusive exons (MXE). For each tumor/NAT comparison, high-confidence AS events were identified as those with a PSI difference >0.1 (10%) and false discovery

rate (FDR) <0.1. AS events were “validated” if they were significant and they pass these thresholds in both the discovery and validation datasets.

Pathway enrichment analysis of differentially spliced genes

To associate cellular functions with differentially spliced genes (DSGs), QIAGEN Ingenuity Pathway Analysis (IPA) [63] was used utilizing the curated gene sets of the QIAGEN’s Ingenuity Knowledge Base. The Core Analysis that includes the enrichment pathway analysis was performed. The canonical pathway visualization is a list of enriched pathways ranked by *p*-value and percentage of overlapping genes mapped against the total number of those in that pathway. Significantly enriched terms with similar descriptions and functions were further grouped into distinct biological categories by IPA to better reflect the biology in context, and top pathways were schematically projected on the network of enriched terms. Validated events were tested for statistical overrepresentation in IPA’s canonical signaling pathways using Fisher’s exact test.

Protein function and structure annotation:

Genes with AS events were annotated for cancer drivers using COSMIC Cancer Gene Census [64] and for surface proteins subcellular localization database using COMPARTMENTS [65]. The protein topology of altered transcripts was assessed using PROTTTER [66], which provided proteoform visualization of the AS protein sequence.

Regression analysis:

Tumor splicing alterations were characterized for association with multiple exposure variables and lung cancer in patients for whom demographic data was available (Table S2). The AS events were tested for differences in splicing between tumor and NATs using linear models implemented by Limma (v3.42.0) [67]. Linear models were trained across PSI data matrix for AS events using standard linear regression against a continuous indicator (PSI) of AS events absent or present in a gene. The independent variables included continuous variables (age, smoking status, smoke packs, smoking duration, smoking intensity, menthol cigarette usage, aspirin use, aspirin per day, aspirin duration, and BMI) and categorical variables (gender and race). Statistical analyses were performed using the R software package.

Data availability statement

The NCI-MD data that supports the findings of this study was derived from patients enrolled in the NCI-MD Case-Control Study [11].

Code availability statement

We used open-source R version 3.6 to generate the figures. Wherever required, commercially available Adobe Illustrator 23.0.3 (2019) was used to create the figure grids. All of the scripts for analysis and figure production were built in-house and are provided on GitHub (<https://github.com/iamsamanzeeshan/LungSplicingAAvsEA>)

Declarations

Author Contribution

B.M.R. conceived the project, supervised the project, wrote and revised the manuscript, and acquired funding for the present study. S.Z. designed the study, performed bioinformatic analyses including analyzing the data and carried out writing and revising the manuscript as well as preparing figures. S.R.P. conceived the project, supervised the project, revised, and edited the manuscript, and acquired funding for the present study. H.K. supervised the project, revised and edited the manuscript, and acquired funding for the present study. R.A.M. and A.Z. designed experiments and provided technical and sequencing assistance.

Competing Interests

The authors declare no competing interests.

Acknowledgements:

This work was supported by a grant from the National Cancer Institute (R01 CA239093) to SRP, HK and BMR, and Rutgers Cancer Institute of New Jersey Cancer Health Equity Pilot Award to SRP and HK, and Biomedical Informatics Shared Resource at Rutgers Cancer Institute (P30CA072720-5917).

References

1. Siegel, R.L., Miller, K.D., & Jemal, A. Cancer statistics, 2020. *CA: a cancer journal for clinicians*, **70**(1), 7–30 (2020).
2. National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention, (2019).
3. Ryan, B.M. Lung cancer health disparities. *Carcinogenesis*, **39**(6), 741–751 (2018).
4. Howlader, N., et al. (eds). SEER Cancer Statistics Review, 1975-2016, National Cancer Institute, Bethesda, MD, https://seer.cancer.gov/csr/1975_2016/, based on November 2018 SEER data submission, posted to the SEER website, April 2019.
5. Schwartz, A.G., et al. Admixture mapping of lung cancer in 1812 African-Americans. *Carcinogenesis*, **32**(3), 312–317 (2011).
6. Meaney, C.L. et al. Circulating Inflammation Proteins Associated With Lung Cancer in African Americans. *Journal of thoracic oncology: official publication of the International Association for the Study of Lung Cancer*, **14**(7), 1192–1203 (2019).
7. Mitchell, K.A., et al. Recurrent PTPRT/JAK2 mutations in lung adenocarcinoma among African Americans. *Nature communications*, **10**(1), 5735 (2019).
8. Mitchell, K.A., et al. Relationship between West African ancestry with lung cancer risk and survival in African Americans. *Cancer causes & control: CCC*, **30**(11), 1259–1268 (2019).

9. Meaney, C.L., et al. Identification of serum inflammatory markers as classifiers of lung cancer mortality for stage I adenocarcinoma. *Oncotarget*, **8(25)**, 40946–40957 (2017).
10. Arauz, R.F., et al. Whole-Exome Profiling of NSCLC Among African Americans. *J Thorac Oncol*, **15(12)**, 1880–1892 (2020).
11. Mitchell, K.A., Zingone, A., Toulabi, L., Boeckelman, J., & Ryan, B.M. Comparative Transcriptome Profiling Reveals Coding and Noncoding RNA Differences in NSCLC from African Americans and European Americans. *Clin Cancer Res*, **23(23)**, 7412–7425 (2017).
12. Pine, S.R., et al. Differential Serum Cytokine Levels and Risk of Lung Cancer Between African and European Americans. *Cancer Epidemiol. Biomarkers Prev*, **25**, 488–497 (2016).
13. Campbell, J.D., et al. Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations. *JAMA oncology*, **3(6)**, 801–809 (2017).
14. Zingone, A., et al. A comprehensive map of alternative polyadenylation in African American and European American lung cancer patients. *Nature communications*, **12(1)**, 5605 (2021).
15. Climente-González, H., Porta-Pardo, E., Godzik, A., & Eyras, E. The Functional Impact of Alternative Splicing in Cancer. *Cell reports*, **20(9)**, 2215–2226 (2017).
16. Oltean, S., & Bates, D.O. Hallmarks of alternative splicing in cancer. *Oncogene*, **33(46)**, 5311–5318 (2014).
17. Li, Y., et al. Prognostic alternative mRNA splicing signature in non-small cell lung cancer. *Cancer letters*, **393**, 40–51 (2017).
18. Liu, Y., Nie, H., Liu, H., & Lu, F. Poly(A) inclusive RNA isoform sequencing (PAIso-seq) reveals widespread non-adenosine residues within RNA poly(A) tails. *Nature communications*, **10(1)**, 5292 (2019).
19. Kahraman, A., et al. Pathogenic impact of transcript isoform switching in 1,209 cancer samples covering 27 cancer types using an isoform-specific interaction network. *Scientific reports*, **10(1)**, 14453 (2020).
20. Danan-Gotthold, M. et al. Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic acids research*, **43(10)**, 5130–5144 (2015).
21. Zhang, D., Duan, Y., Cun, J., & Yang, Q. Identification of Prognostic Alternative Splicing Signature in Breast Carcinoma. *Frontiers in genetics*, **10**, 278 (2019).
22. Kahles, A., et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer cell*, **34(2)**, 211–224.e6 (2018).
23. Bonnal, S.C., López-Oreja, I., & Valcárcel, J. Roles and mechanisms of alternative splicing in cancer - implications for care. *Nature reviews. Clinical oncology*, **17(8)**, 457–474 (2020).
24. Ding, L., & Odunsi, K. RNA Splicing and Immune-Checkpoint Inhibition. *The New England journal of medicine*, **385(19)**, 1807–1809 (2021).
25. Bourcier, J. & Abdel-Wahab, O. Splicing-Mediated Antigen Escape from Immunotherapy for B-cell Malignancies. *Blood Cancer Discovery*, **3**, 1–3 (2021).

26. Chang, Y. S., et al. Genome-Wide Analysis of Prognostic Alternative Splicing Signature and Splicing Factors in Lung Adenocarcinoma. *Genes*, **11**(11), 1300 (2020).
27. Wei, C., et al. Profiles of alternative splicing events in the diagnosis and prognosis of Gastric Cancer. *Journal of Cancer*, **12**(10), 2982–2992 (2021).
28. Pal, S., et al. Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes. *Nucleic acids research*, **42**(8), e64 (2014).
29. Zhu, J., Chen, Z., & Yong, L. Systematic profiling of alternative splicing signature reveals prognostic predictor for ovarian cancer. *Gynecologic oncology*, **148**(2), 368–374 (2018).
30. Zhao, W., Hoadley, K.A., Parker, J.S., & Perou, C.M. Identification of mRNA isoform switching in breast cancer. *BMC genomics*, **17**, 181 (2016).
31. Al Abo, M., et al. Differential alternative RNA splicing and transcription events between tumors from African American and White patients in The Cancer Genome Atlas. *Genomics*, **113**(3), 1234–1246 (2021).
32. Li, S., Hu, Z., Zhao, Y., Huang, S., & He, X. Transcriptome-Wide Analysis Reveals the Landscape of Aberrant Alternative Splicing Events in Liver Cancer. *Hepatology (Baltimore, Md.)*, **69**(1), 359–375 (2019).
33. Li, H., et al. Identification of prognostic alternative splicing events in sarcoma. *Scientific reports*, **11**(1), 14949 (2021).
34. Deveaux, A.E., et al. RNA splicing and aggregate gene expression differences in lung squamous cell carcinoma between patients of West African and European ancestry. *Lung cancer (Amsterdam, Netherlands)*, **153**, 90–98 (2021).
35. Aran, D., et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature communications*, **8**(1), 1077 (2017).
36. Sciarrillo, R., et al. The role of alternative splicing in cancer: From oncogenesis to drug resistance. *Drug resistance updates: reviews and commentaries in antimicrobial and anticancer chemotherapy*, **53**, 100728 (2020).
37. Robinson, T. J., et al. Alternative RNA Splicing as a Potential Major Source of Untapped Molecular Targets in Precision Oncology and Cancer Disparities. *Clinical cancer research: an official journal of the American Association for Cancer Research*, **25**(10), 2963–2968 (2019).
38. Fournier, G., et al. Loss of AF6/afadin, a marker of poor outcome in breast cancer, induces cell migration, invasiveness and tumor growth. *Oncogene*, **30**(36), 3862–3874 (2011).
39. Smith, C.C., et al. Alternative tumour-specific antigens. *Nature reviews. Cancer*, **19**(8), 465–478 (2019).
40. Brosseau, J.P., et al. Tumor microenvironment-associated modifications of alternative splicing. *RNA (New York, N.Y.)*, **20**(2), 189–201 (2014).
41. Shapiro, I.M., et al. An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS genetics*, **7**(8), e1002218 (2011).

42. Yang, Y., et al. Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition. *Molecular and cellular biology*, **36**(11), 1704–1719 (2016).
43. Iwasaki, J., et al. FIP1L1 presence in FIP1L1-RARA or FIP1L1-PDGFR α differentially contributes to the pathogenesis of distinct types of leukemia. *Annals of hematology*, **93**(9), 1473–1481 (2014).
44. Pintarelli, G., et al. Cigarette smoke alters the transcriptome of non-involved lung tissue in lung adenocarcinoma patients. *Scientific reports*, **9**(1), 13039 (2019).
45. Stahre, M., Okuyemi, K.S., Joseph, A.M., & Fu, S.S. Racial and Ethnic Differences in the Relationship between Aspirin Use and Non-Small Cell Lung Cancer Risk and Survival. *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, **27**(12), 1518–1526 (2018).
46. Cates, H.M., et al. Transcription Factor E2F3a in Nucleus Accumbens Affects Cocaine Action via Transcription and Alternative Splicing. *Biological psychiatry*, **84**(3), 167–179 (2018).
47. Donadoni, M., Cicalese, S., Sarkar, D.K., Chang, S.L., & Sariyer, I.K. Alcohol exposure alters pre-mRNA splicing of antiapoptotic Mcl-1L isoform and induces apoptosis in neural progenitors and immature neurons. *Cell death & disease*, **10**(6), 447. (2019).
48. Lamba, V., et al. PXR (NR1I2): splice variants in human tissues, including brain, and identification of neurosteroids and nicotine as PXR activators. *Toxicology and applied pharmacology*, **199**(3), 251–265 (2004).
49. Rajgor, D., Mellad, J.A., Autore, F., Zhang, Q., & Shanahan, C.M. Multiple novel nesprin-1 and nesprin-2 variants act as versatile tissue-specific intracellular scaffolds. *PloS one*, **7**(7), e40098 (2012).
50. Hu, B., et al. CD44 promotes cell proliferation in non-small cell lung cancer. *Oncology letters*, **15**(4), 5627–5633 (2018).
51. Wang, C. Y., et al. The subpopulation of CD44-positive cells promoted tumorigenicity and metastatic ability in lung adenocarcinoma. *Journal of the Chinese Medical Association: JCMA*, **82**(3), 196–201 (2019).
52. Afify, A.M., Tate, S., Durbin-Johnson, B., Rocke, D.M., & Konia, T. Expression of CD44s and CD44v6 in lung cancer and their correlation with prognostic factors. *The International journal of biological markers*, **26**(1), 50–57 (2011).
53. Thapa, R., & Wilson, G.D. The Importance of CD44 as a Stem Cell Biomarker and Therapeutic Target in Cancer. *Stem cells international*, 2016, 2087204 (2016).
54. Anczuków, O., et al. BRCA2 deep intronic mutation causing activation of a cryptic exon: opening toward a new preventive therapeutic strategy. *Clinical cancer research: an official journal of the American Association for Cancer Research*, **18**(18), 4903–4909. (2012).
55. Roberts, T. C., Langer, R., & Wood, M. Advances in oligonucleotide drug delivery. *Nature reviews. Drug discovery*, **19**(10), 673–694 (2020).
56. Xiong, H., Veedu, R. N., & Diermeier, S. D. Recent Advances in Oligonucleotide Therapeutics in Oncology. *International journal of molecular sciences*, **22**(7), 3295 (2021).

57. Bourcier, J. & Abdel-Wahab, O. Splicing-Mediated Antigen Escape from Immunotherapy for B-cell Malignancies. *Blood Cancer Discovery*, **3**, 1–3 (2021)
58. Peng, Q., et al. Impacts and mechanisms of alternative mRNA splicing in cancer metabolism, immune response, and therapeutics. *Molecular therapy: the journal of the American Society of Gene Therapy*, S1525-0016(21)00581–5 (2021).
59. Frankiw, L., Baltimore, D., & Li, G. Alternative mRNA splicing in cancer immunotherapy. *Nature reviews. Immunology*, **19**(11), 675–687 (2019).
60. Lu, S. X., et al. Pharmacologic modulation of RNA splicing enhances anti-tumor immunity. *Cell*, **184**(15), 4032–4047.e31 (2021).
61. Kim, D., Langmead, B., & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, **12**(4), 357–360 (2015).
62. Shen, S., et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(51), E5593–E5601 (2014).
63. QIAGEN Ingenuity Pathway Analysis (IPA) <https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/>
64. Sondka, Z., et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature reviews. Cancer*, **18**(11), 696–705 (2018).
65. Binder, J. X., et al. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database: the journal of biological databases and curation*, *2014*, bau012. (2014).
66. Omasits, U., Ahrens, C.H., Müller, S., & Wollscheid, B. Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics (Oxford, England)*, **30**(6), 884–886 (2014).
67. Ritchie, M. E., et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, **43**(7), e47 (2015).

Figures

Fig.1: Profiling of AS events in lung cancer.

Figure 1

Fig 1: Profiling of AS events in lung cancer. **(a)** Schematic representation of five modes of AS events: skipped exon (SE), retained intron (RI), mutually exclusive exons (MXE), alternate 5' splice site (A5SS), and alternate 3' splice site (A3SS). **(b)** Volcano plot of validated AS events with FDR<0.1 and delta PSI of $\geq 10\%$ in LUAD and LUSC tumor tissues. Color represents the filtered AS events. Grey represents the non-significant AS events. **(c)** Pie chart showing proportion of AS classes in LUAD and LUSC. **(d)** Venn diagram showing overlap of DSGs between LUAD and LUSC. **(e)** Bar plot of shared and unique validated candidate events in alternative splicing modes in LUAD and LUSC. **(f)** Top canonical pathways from validated events enriched in LUAD and LUSC.

a

AA-LUAD PSI expression

EA-LUAD PSI expression

-Log₁₀(FDR)

PSI Difference (Tumor-Control)

Alteration
● Not significant
● Significant

b

AA-LUAD (n=1160) EA-LUAD (n=1119)

SE RI MXE ASSS A3SS

c

AA LUAD Validated events (377)

EA LUAD Validated events (355)

171 206 149

d

Count

A3SS ASSS MXE RI SE

■ AA-LUAD
■ EA-LUAD
■ Shared

e

PSI difference

A3SS ASSS MXE RI SE

■ Validation
▲ Skipping

Fig 2: AA and EA population differences of alternatively spliced events in LUAD. **(a)** Volcano plot of candidate AS events with FDR<0.1 and delta PSI of $\geq 10\%$ in LUAD tumors from AAs and EAs. Color represents the filtered AS events. Grey scale represents the AS events not filtered. **(b)** Venn diagram showing overlap of validated AS events in LUAD between AAs and EAs. **(c)** Pie chart showing proportion of candidate events in LUAD from AAs and EAs. **(d)** Bar plot of shared and unique candidate events in alternative splicing modes identified in corresponding DSGs in LUAD from AAs and EAs. **(e)** Top canonical pathways from validated events enriched in LUAD from AAs and EAs.

Fig.3: Population differences in alternatively spliced events in LUSC.

Figure 3

Fig 3: Population differences of alternatively spliced events in LUSC. **(a)** Volcano plot of candidate AS events with $\text{FDR} < 0.1$ and ΔPSI of $\geq 10\%$ in AA and EA populations in LUSC tumors. Color represents the filtered AS events. Grey scale represents the AS events not filtered using stringent filters. **(b)** Venn diagram showing overlap of validated AS events between AA-LUSC and EA-LUSC. **(c)** Pie chart showing proportion of candidate events in AA-LUSC and EA-LUSC. **(d)** Bar plot of shared and unique candidate events in alternative splicing modes identified in corresponding DSGs in AA-LUSC and EA-LUSC. **(e)** Violin plot of validated events showing the number of inclusion and skipping events for all splicing modes between the AA and EA tumor tissues in LUSC. **(f)** Top canonical pathways from validated events enriched in AA-LUSC and EA-LUSC.

Fig.4: Alternative spliced isoforms in cancer driver genes in tumor subtypes and populations.

Figure 4

Fig 4: Alternative spliced isoforms detected in cancer driver genes in tumor subtypes and populations. **(a)** Validated events of alternatively spliced cancer driver genes in tumor subtypes. **(b)** Validated events of alternatively spliced cancer driver genes in populations. **(c)** *AFDN* alteration in exon 16 involving skipped exon (SE) in the DIL domain. **(d)** PSI difference in tumor vs. NAT for *AFDN* event in discovery and validation datasets in tumor subtypes and populations.

Fig.5: Alternative spliced isoforms detected in surface proteins in subtypes.

Fig 5: Alternative spliced isoforms detected in surface proteins in subtypes. **(a)** Validated events of alternatively spliced surface proteins in tumor subtypes. **(b)** *SYNE2* alteration in exon 107 involving skipped exon (SE) in the KASH domain. **(c)** PSI difference in tumor vs. NAT for *SYNE2* AS event in discovery and validation datasets in tumor subtypes and populations.

a

Heatmap showing MeanPSI for CD44 across various cancer types and tissues. The y-axis lists cancer types: BLCA, BRCA, CESC, COADREAD, HNSC, KIPAN, LIHC, LUAD, LUSC, OV, PAAD, PRAD, SARC, SKCM, STAD, THCA, UCEC. The x-axis lists tissues: Adipose, Blood, Brain, Colon, Esophagus, Kidney, Liver, Muscle, Pancreas, Skin, Stomach, Testis, Uterus, Vagina, Whole Blood, Whole Tissue. The color scale ranges from 0 to 1.0.

b

CD44 GeneCode Reference ENS00000428726.8 [Exon 11]. The protein structure shows domains: Link domain, Stalk domain, Variant domain, Transmembrane domain, ERM binding domain, Ankyrin binding domain. A splice event in exon 11 is highlighted, showing 44 amino acids. The event is labeled "SE Splicing Event exon 11 132 bp" and "Tumor" and "NAT".

c

Violin plots showing PSI for CD44 event across Discovery and Validation datasets for various cancer types. The y-axis is PSI for CD44 event (0.0 to 1.0). The x-axis shows Discovery Validation and Discovery Validation. The legend indicates NAT (blue) and Tumor (orange). The plots show that the PSI for CD44 event is generally higher in Tumor samples compared to NAT samples across all cancer types and datasets.

Fig 6: Alternative spliced isoforms detected in surface proteins in populations. **(a)** Validated events of alternatively spliced surface proteins in populations. **(b)** *CD44* alteration in exon 11 involving skipped exon (SE) in the variant domain. **(c)** PSI Difference in tumor vs. NAT for *CD44* event in discovery and validation datasets in tumor subtypes and populations.

This is a list of supplementary files associated with this preprint. Click to download.

- Page 21/21