Proteogenomic analysis of air-pollution-associated lung cancer reveals prevention and therapeutic opportunities

Honglei Zhang[1,*,#], *Chao Liu*[2,3,*], *Shuting Wang*[2,*], *Qing Wang*[4,*], *Xu Feng*[1], *Huawei Jiang*[5], *Yong Zhang*[6,#], *Xiaosan Su*[1,#] *and Gaofeng Li*[2,#]

[1]Center for scientific research, Yunnan University of Chinese Medicine, Kunming, Yunnan, 650500, China;

[2]Department of Thoracic Surgery II, Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming 650106, China;

[3]Department of Nuclear Medicine, Third Affiliated Hospital of Kunming Medical University, Yunnan Cancer Hospital, Kunming 650106, China;

[4]Department of Oncology, Qujing First People's Hospital, Qujing, 202150, China;

[5]Department of Plastic Surgery, First Affiliated Hospital of Kunming Medical University, Kunming 650032, China;

[6]Department of Nephrology, Institutes for Systems Genetics, Frontiers Science Center for Disease-Related Molecular Network, West China Hospital, Sichuan University, Chengdu 610041, China.

[*]These authors contributed equally to this work.

[#]Corresponding authors. Yong Zhang (nankai1989@foxmail.com), Xiaosan Su (suxs163@163.com); Gaofeng Li (ligaofeng@kmmu.edu.cn); Honglei Zhang (hlzhang2014@163.com).

Abstract

Air pollution significantly impact lung cancer progression, but there is a lack of a comprehensive molecular characterization of clinical samples associated with air pollution. Here, we performed a proteogenomic analysis of lung adenocarcinoma (LUAD) in 169 female never-smokers from the Xuanwei area (XWLC cohort), where coal smoke is the primary contributor to the high lung cancer incidence. Genomic mutation analysis revealed XWLC as a distinct subtype of LUAD separate from cases associated with smoking or endogenous factors. Mutational signature analysis suggested that Benzo[a]pyrene (BaP) is the major risk factor in XWLC. The BaP-induced mutation hotspot, EGFR-G719X, was present in 20% of XWLC which endowed XWLC with elevated MAPK pathway activations and worse outcomes compared to common *EGFR* mutations. Multi-omics clustering of XWLC identified four clinically relevant subtypes. These subgroups exhibited distinct features in biological processes, genetic alterations, metabolism demands, immune landscape, tumor microbiota composition and radiomic features. Finally, *MAD1 and TPRN* were identified as novel potential therapeutic targets in XWLC. Our study provides a valuable resource for researchers and clinicians to explore prevention and treatment strategies for air-pollution-associated lung cancers.

Introduction

Lung cancer is the leading cause of cancer deaths globally[1]. Though the most common cause of lung cancer is tobacco smoking, studies estimate that approximately 25% of lung cancers worldwide occur in individuals who have never smoked[2]. Recently, lung cancer in never smokers (LCINS) were molecular profiled and new genomic features were revealed[3-9]. For now, further stratification of LCINS based on different risk factors would be helpful to reveal the oncogenic mechanisms and develop more targeted therapies. Air pollutants, which can directly affect the pulmonary airway, play crucial roles in promoting lung adenocarcinoma[10-12]. More than 20 environmental and occupational agents are lung carcinogens[13, 14] and amount of studies have been made to investigate molecular mechanisms in tumor progression of air pollution chemicals or components using cell lines or mouse/rat models[15-17]. However, a comprehensive molecular characterization of clinical lung cancer samples associated with air pollution is still lacking.

The Xuanwei area has the highest rate of lung cancer in China, and extensive research has established a strong link between lung cancer and exposure to domestic coal smoke[18-23]. Specifically, etiologic link between smoky coal burning and cancer was epidemiologically established[18, 19] and association between household stove improvement and lower risk of lung cancer was observed[23]. Moreover,

genomic evidence of lung carcinogenesis associated with coal smoke in Xuanwei area, China was provided in our previous study[22]. Thus, lung cancer in Xuanwei areas exemplifies the ideal disease to study characteristics of lung cancers associated with air pollution. In recent years, the molecular features of Xuanwei lung cancer have been gradually revealed[24-26] [22]. A large sample size with multi-comic molecular profiling is urgent needed to explicit the air pollution chemicals and furthermore propose more targeted therapies.

To better understand the molecular mechanisms and heterogeneity of XWLC and to advance precision medicine, we expanded the sample size of our next-generation sequencing dataset to 169 sample size and performed proteomic and phosphoproteomic profiling to 112 samples. Furthermore, we integrated 107 radiomic features derived from X-ray computed tomography (CT) scans to 115 samples to non-invasively distinguish molecular subtypes. This allowed us to identify potential major risk factor, distinguish the genomic features and establish clinically relevant molecular subtypes. Our study provides an exceptional resource for future biological, diagnostic, and drug discovery efforts in the study of lung cancer related to air pollution.

Results

To investigate unique biological features of LUAD associated with air pollution, three previous LUAD datasets related to different carcinogens were used for comparison (Fig. 1a). CNLC is the subset of lung adenocarcinoma from non-smoking patients in Chinese Human Proteome Project (CNHPP project) [8] (n=77). TSLC is the subset of lung adenocarcinoma from smoking female in TCGA-LUAD project [27] (n=168). TNLC is the subset of lung adenocarcinoma from non-smoking female in TCGA-LUAD project[27] (n=102). The clinicopathological characteristics of patients from CNLC, TSLC and TNLC cohorts were supplied in Supplementary Table 1a and 1b.

Proteogenomic landscape in Xuanwei lung cancer (XWLC)

The present study prospectively collected primary samples of lung adenocarcinoma (LUAD) from 169 never-smoking women from the Xuanwei area in China (Supplementary Table 1c). The XWLC cohort had a median age of 56 years (Fig. 1b), and the majority of tissue samples were in the early stages of the disease (145 were stage I/II, and 24 were stage III/IV, Fig. 1c). A total of 135, 136, 102 and 102 tumor samples were profiled with whole-exome sequencing (WES), RNA-seq, label-free protein quantification, and label-free phosphorylation quantification, respectively (Fig.1d-e and Extended Fig.1a and 1b). Analysis of the WES data from the paired tumor and normal tissue samples revealed 37,149

somatic mutations, including 1,797 InDels, 32,972 missense mutations, 2,345 nonsense mutations, and 35 nonstop mutations (Supplementary Table 2). Copy number analysis showed 140,396 gene-level amplifications and 67,605 deletions across 40 cytobands (Supplementary Table 3). The mRNA-seq data characterized the transcription profiles of 19,182 genes (Supplementary Table 4). The label-free global proteomics identified 9,152 proteins (encoded by 6,864 genes) with an average of 6,457 proteins per sample (Supplementary Table 5). The label-free phosphoproteomics identified 24,990 highly reliable phosphosites from 5,832 genes with an average of 10,478 phosphosites per sample (Supplementary Table 6). The quality and reproducibility of the mass spectrometry data were maintained throughout the study (Extended Data Fig. 1c-e).

The air pollutant Benzo[a]pyrene (BaP) primarily contributes to the mutation landscape of XWLC

To infer the primary risk factor responsible for the progression of XWLC, we used SomaticSignatures[28] to identify mutational signatures from single nucleotide variants. Mutational signatures were identified in each cohort and a cosine similarity analysis was performed against mutational signatures in COSMIC mutational signatures[29, 30] and environmental agents mutational sigantures[31] allowing for inference of the underlying causes (Fig. 1f-i and Extended Fig. 2). Generally, exposure tobacco

smoking carcinogens (COSMIC signature) and chemicals such as BaP (environmental agent signature) were identified as the most significant contributing factors in both the XWLC and TSLC cohorts (Fig. 1f and 1g). In contrast, defective DNA mismatch repair and endogenous mutational processes initiated by spontaneous deamination of 5-methycytosine were inferred as the major causes in the TNLC and CNLC cohorts (Fig. 1h and 1i). Therefore, the XWLC and TSLC cohorts are more explicitly influenced by environmental carcinogens, while the TNLC and CNLC cohorts may be more affected by age or endogenous risk factors. BaP, a representative compound of polycyclic aromatic hydrocarbons (PAHs), is found in both cigarette smoke and coal smoke and is recognized as a major environmental risk factor for lung cancer[32-34]. Upon metabolism, BaP forms the carcinogenic metabolite 7β,8α-dihydroxy-9α,10α-epoxy-7,8,9,10-tetrahydrobenzo[a]pyrene (BPDE), which creates DNA adducts leading to mutations and malignant transformations. This process involves two key regulators: *CYP1A1* and *AhR*. CYP1A1 plays a crucial role in BaP epoxidation at the 7,8 positions, which is the most critical step in BPDE formation[35]. AhR is a ligand-activated transcription factor that responds to various chemicals, including chemical carcinogens, and is activated by BaP[36]. Accordingly, our results demonstrated significantly higher expression of CYP1A1 and AhR in tumor samples compared to normal samples (Fig. 1j

and 1k). To validate the association between BaP and the elevated incidence of lung cancer in the Xuanwei area, we assessed the serum BPDE levels in individuals residing in this region (Methods). Our findings revealed that the serum BPDE content in individuals from the Xuanwei area averaged around 0.85 nm/ml, and there was no difference between younger and older cases (Fig. 1l and Supplementary Table 7). Furthermore, we treated hiPSC (human induced pluripotent stem cell (hiPSC) with BaP in vitro and found that sites such as Y1173 and Y1068 of EGFR were more phosphorylated in BaP treated cells whereas the total abundance of EGFR showed no significant differences (Fig. 1m). All these results provided further evidence for the involvement of BaP and its metabolite in the development of lung cancer.

Though coal-smoke related lung cancer (XWLC cohort) and cigarette-smoke related lung cancer (TSLC cohort) showed similar environmental carcinogens, we found that downstream pathway activation and therapeutic targeted potential showed distinctive features. Firstly, the correlation of genomic mutations between XWLC and TSLC was found to be low (Fig. 2a). Secondly, there was a remarkable difference in the fraction of samples affected by pathway mutations between two cohorts (Fig. 2b). Notably, the TSLC cohort exhibited a higher fraction of samples affected by oncogenic pathways comparing to XWLC cohort. Thirdly, mutation frequencies, such as *EGFR*, *TP53*,

RBM10, and *KRAS* (Fig. 2c), as well as the distribution of amino acid changes in *EGFR* and *TP53*, showed significant differences between the XWLC and other cohorts (Fig. 2d). Specifically, the XWLC cohort exhibited a higher mutation rate in G719C/A/D/S and E746_A750del within the *EGFR* gene compared to the other three cohorts. For the *TP53* gene, frame shift mutations including D9Gfs*3 (n=1), S15Ifs*28 (n=1), D22Efs*61 (n=1), and V34Wfs*49 (n=2) were exclusively detected in the XWLC cohort, whereas tetramer domain mutations were only found in the other three cohorts (Fig. 2d). Finally, there was a noticeable disparity in the percentage of samples with actionable targets among the cohorts. (Fig. 2e). Actionable targets in XWLC cohort were mainly focus on *EGFR* mutations including pG719S, pG719C, pG719A, pL858R, p.E746_A750del whereas TSLC cohort had more actionable targets in CHEK2 p.K373E, KRAS p.G12V/D/C/A (Fig. 2e).

Taken together, we found that the XWLC and TSLC cohorts, which are smoke-related lung adenoma groups, demonstrated distinct etiology compared to the TNLC and CNLC cohorts which may be influenced by endogenous risk factors to a greater extent. The blood serum assay provided support for BaP as a promising risk factor specifically in the XWLC cohort. Additionally, significant disparities were observed between XWLC and TSLC in terms of downstream pathway activations and specific oncogene loci. Consequently, we conclude that air pollution-

202 associated lung cancer represents a distinct subtype within LUAD.

203 **The EGFR-G719X mutation, which is a hotspot associated with BaP**
204 **exposure, possesses distinctive biological features**

205 Notably, the XWLC cohort displayed a distinguishable mutation pattern
206 in specific *EGFR* mutation sites compared to the other cohorts (Fig.3a
207 and Fig.2d). In particular, the G719C/A/D/S (G719X) mutation was the
208 most prevalent *EGFR* mutation in the XWLC cohort (20%), while it was
209 rarely found in the other three cohorts (CNLC: 1.9%; TSLC: 1.9%;
210 TNLC: 0) (Fig. 3b). Notably, we found it was a hot spot associated with
211 BaP exposure (Fig. 3c and 3d). Specifically, **GG**C is the 719 codon, the
212 first G can be converted to T (pG719C, n=13) or A (pG719S, n=5), the
213 second G can be converted to A (pG719D, n=1) or C (pG719C, n=8).
214 Thus, pG719C was the most detected mutation type (Fig. 3c). G>T/C>A
215 transversion can be induced by several compounds such as BaP or
216 dibenz(a,h)anthracene (DBA), unlike other compounds, the tallest peak
217 induced by BaP occurs at GpGpG, reflecting how their DNA adducts are
218 formed principally at N^2-guanine[31]. Our result showed that the most
219 frequently detected pG719C AAchange was correspond to
220 G**G**GC>G**T**GG transversion (Fig. 3d). Thus, pG719C is a hot spot
221 associated with BaP exposure.

222 We conducted further investigations into the biological characteristics
223 of samples carrying the G719X mutations. Notably, we observed a

moderate to high expression of MAPK signaling components, MAP2K2 (MEK), and MAPK3 (ERK1), in tumors harboring the EGFR-G719X mutation compared to other EGFR statuses (Fig. 3e-3h). Utilizing hallmark capability analysis and RNA-seq-based estimation of immune cell infiltration, we found that tumors with G719X mutations exhibited similarities to those with L858R mutations (Extended Data Fig. 3a-b). However, patients with G719X mutations were notably younger than those with L858R mutations, indicating a higher occurrence rate of G719X in younger female patients (Fig. 3i). Analysis of overall survival and progression-free interval (PFI) revealed that patients with the G719X mutation had worse outcomes compared to other EGFR mutation subtypes (Fig. 3j and 3k). Furthermore, there were no significant differences in mutation burden or the number of neoantigens between tumors with G719X mutations and tumors with other *EGFR* mutation statuses (Extended Data Fig. 3c).

To explore the heterogeneity of signaling pathways activated by different *EGFR* mutation statuses, we conducted Kinase-Substrate Enrichment Analysis (KSEA) [37, 38] based on the XWLC phosphoproteomics dataset. Our analysis of the phosphoproteome across various *EGFR* mutation types revealed distinct activation patterns of kinases. Specifically, the G719X mutation was associated with the activation of PRKCZ, CDK2, AURKB, CSNK1A1, CDK4, and HIPK2.

The L858R mutation showed activation of PRKCZ, MAPK7, MAPK12, HIPK2, and CSNK2A1. The Exon19del mutation exhibited activation of CHUK, TTK, PRKCZ, PLK1, NEK2, MAP2K2, CDK2, PRKDC, and MAP2K6. Other EGFR mutations were associated with the activation of AURKB, NEK2, TTK, PLK1, PRKACB, and PRKACG. EGFR-WT mutations showed activation of CSNK1E, PRKCZ, AURKB, CDK2, AURKC, CDK1, CSNK1A1, PRKDC, and CSNK2A1 (Fig. 3l). In Extended Data Fig. 3d, we provide a list of FDA-approved drugs that target the activated kinases in tumors harboring the G719X mutation. Currently, afatinib is widely regarded as a first-line therapy for patients with the G719X mutation[39-42]. However, reports indicate that 80% of patients with this mutation may develop resistance to afatinib, even in the absence of T790M[43], underscoring the need for a deeper understanding of the downstream pathways associated with the G719X mutation. Therefore, a promising approach to overcome resistance in tumors with this mutation could involve combining afatinib, which targets activated EGFR, with FDA-approved drugs that specifically target the activated kinases associated with G719X.

Clinically relevant Subtyping in XWLC

To uncover the inherent subgroups within air-pollution-associated tumors, we employed unsupervised Consensus Clustering[44] on integrated RNA, protein, and phosphoprotein profiles of XWLC tumor samples. This

analysis led to the identification of four distinct intrinsic clusters, denoted as MC-I, II, III, and IV (Fig. 4a, Extended Data Fig. 4a and Methods). Further survival analysis demonstrated that patients belonging to the MC-IV group exhibited the poorest overall survival compared to the other three subgroups, thus indicating the prognostic potential of multi-omic clustering (Fig. 4b). Notably, there were no significant differences in clinical features such as age and stage observed among the four subgroups (Fig. 4a). Since BaP has the potential to induce the expression of CYP1A1 and AhR, we utilized the expression levels of CYP1A1 and AhR as indicators of BaP activity. Consequently, we compared the expression levels of CYP1A1 and AhR among the four subtypes to identify the tumor type most strongly associated with air pollution. Our findings revealed that the MC-II subtype exhibited higher expression of CYP1A1 (Fig. 4c) and moderately elevated expression level of AhR compared to the other three subtypes (Fig. 4d) Moreover, the MC-II possessed more G719X mutations (MC-1:0.39, MC-II:0.42, MC-III: 0.20, MC-IV: 0.08). These results collectively suggest a stronger association between the MC-II subtype and BaP exposure. Notably, there was a significant correlation between CYP1A1 and EGFR expression (Fig. 4e), with EGFR being more highly expressed in the MC-II subtype (Fig. 4f). These results indicated that MC-II was more associated with air-pollution.

Through Kinase Substrate Enrichment Analysis (KSEA) of the phosphoproteome in tumor samples compared to normal adjacent tissues (NATs), we identified specific kinase activations within the four subgroups. In MC-I samples, kinase activations included PRKDC, PRKCZ, CSNK1A1, NEK2, GSK3A, and ROCK1. MC-II samples showed activations of CDK2, CDK1, AURKA, TTK, CDK6, and CHUK. MC-III samples exhibited activations of AKT1, AKT3, RPS6KB1, CSNK2A1, and PAK2. Finally, CDK2 and ROCK1 were activated in MC-IV samples (Fig. 4g). Particularly noteworthy is the enrichment of CDK1/2/6 kinases, which regulate cell cycle checkpoints, in the MC-II subtype, indicating its high proliferation capabilities. These findings imply that distinct kinase pathways are activated within each subgroup, suggesting the presence of specific therapeutic targets for each subgroup. Consequently, we proceeded to explore therapeutic strategies for each subgroup as outlined below:

The **MC-IV** subtype exhibited the poorest overall survival compared to the other three subtypes (Fig. 4b). Given the crucial role of epithelial-mesenchymal transition (EMT) in malignant progression, our first evaluation focused on the EMT process across the four subtypes. We observed higher expression levels of mesenchymal markers such as *VIM*, *FN1*, *TWIST2*, *SNAI2*, *ZEB1*, *ZEB2*, and others in the MC-IV subtype (Fig. 5a). To comprehensively assess the EMT capability of the MC-IV

subtype, we calculated EMT scores using the ssGSEA enrichment method based on protein levels and GSEA hallmark gene set (M5930)[45]. The results confirmed the elevated EMT capability of the MC-IV subtype at the protein level (Fig. 5b). Furthermore, Fibronectin (FN1), an EMT marker that promotes the dissociation, migration, and invasion of epithelial cells, was found to be highly expressed in the MC-IV subtype at the protein level (Fig. 5c). Additionally, β-Catenin, a key regulator in initiating EMT, was highly expressed in the MC-IV subtype at the protein level (Fig. 5d). Collectively, our findings demonstrate that the MC-IV subtype possesses an enhanced EMT capability, which may contribute to the high malignancy observed in this subtype.

The **MC-II** subtype demonstrated the second-worst outcome and was found to be more strongly associated with air pollution (Fig. 4). This subtype exhibited dysregulation of cell cycle processes, including cell division, glycolysis, and cell cycle biological processes (Fig. 4a). The KSEA analysis revealed that the CDK1 and CDK2 pathways, which are closely linked to cell cycle regulation, were predominantly activated in the MC-II subtype (Fig. 5e). Consistently, we observed higher expression levels of CDK1 and CDK2 at both the protein and phosphoprotein levels in the MC-II subtype, indicating specific elevation of the G2M phase in the cell cycle (Fig. 5e). The cell cycle and glycolysis processes are tightly coordinated, allowing cells to synchronize their metabolic state and

energy requirements with cell cycle progression to ensure proper cell growth and division[46-48]. In line with this, we found that key enzymes involved in glycolysis regulation, such as Hexokinase 1 (HK1), Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and Glyceraldehyde-3-phosphate dehydrogenase-like protein (GPL), were highly expressed in the MC-II subtype (Fig. 5f). Additionally, the MC-II subtype was enriched with EGFR mutations (MC-II vs. others: 18/24 vs. 51/110; Fisher's exact p = 0.013) and TP53 mutations (MC-II vs. others: 14/24 vs. 35/110; Fisher's exact p = 0.019), consistent with the characteristic loss of control over cell proliferation. In summary, the MC-II subtype exhibited dysregulated cell cycle processes accompanied by an elevated glycolysis capability, indicating a distinct metabolic and proliferative phenotype.

The **MC-I** subtype exhibited enrichment in various biological processes including angiogenesis, the cAMP signaling pathway, complement and coagulation cascades, PDL1 expression, the PD-1 checkpoint pathway, leukocyte transendothelial migration, and actin cytoskeleton processes (Fig. 4a). In-depth exploration of key components involved in angiogenesis revealed that vascular endothelial growth factor A (VEGFA), a growth factor crucial for both physiological and pathological angiogenesis, was highly expressed in the MC-I subtype (Fig. 5g). Additionally, phosphorylation of vascular endothelial growth factor

receptor 1 (VEGFR1), a receptor tyrosine kinase essential for angiogenesis and vasculogenesis, was also highly expressed in the MC-I subtype (Fig. 5h). The angiogenesis scores, calculated using the ssGSEA method based on protein levels and the hallmark gene set (M5944), were relatively high in the MC-I and MC-IV subtypes (Fig. 5i). Furthermore, the relationship between the Notch signaling pathway and angiogenesis is well-established[49]. Notch signaling plays a role in multiple aspects of angiogenesis, including endothelial cell sprouting, vessel branching, and vessel maturation [50, 51]. In the MC-I subtype, the expression of Notch receptors (Notch1-4) and ligands (DLL1, DLL4, JAG1, and JAG2) was highly elevated, indicating increased activation of Notch signaling (Fig. 5j). KEGG pathview analysis demonstrated that key regulators of the VEGF signaling pathway were highly expressed in the MC-I subtype (Extended Fig.5a). Therefore, manipulating Notch signaling could potentially serve as a strategy to regulate angiogenesis and control pathological angiogenesis in the MC-I subtype.

The **MC-III** subtype is characterized by the upregulation of various metabolic processes, including oxidative phosphorylation, peroxisome function, adipogenesis, fatty acid metabolism, and xenobiotic metabolism-related processes (Fig. 5k). Additionally, we conducted further investigations into the immune features across the subtypes. Interestingly, we observed higher expression of genes associated with

PD-1 signaling (GSEA, SYSTEMATIC_NAME M18810) in the MC-III subtype (Fig. 5l). Since PD-1 is primarily expressed on the surface of certain immune cells, particularly activated T cells, we inferred the immune cell infiltration using the ssGSEA method based on immune cell-specific gene sets. We found that activated CD8+ T cells exhibited higher infiltration levels in the MC-III subtype compared to the other three subtypes (Fig. 5m and Supplementary Table 8), which may explain the elevated PD-1 signaling in the MC-III subtype. Furthermore, we examined the expression of receptor-ligand pairs involved in both anti-tumor and pro-tumor lymphocyte recruitment. Remarkably, the MC-III subtype exhibited specific high expression of anti-tumor lymphocyte receptors and ligands, while the expression of pro-tumor lymphocyte receptors and ligands was relatively lower (Fig. 5n). In general, the MC-I subtype showed the reverse expression trend in terms of anti-tumor and pro-tumor receptor-ligand pairs (Fig. 5n).

In conclusion, our classification of lung adenocarcinoma associated with air pollution resulted in the identification of four subtypes, each exhibiting distinct biological pathway activation and immune features. The MC-I subtype demonstrated elevated angiogenesis processes, while the MC-II subtype showed a high capacity for cell division and glycolysis. The MC-III subtype exhibited a notable infiltration of CD8+ cells, and the MC-IV subtype was characterized by high EMT capability, which

may contribute to its poor outcome. These findings have significant implications for the development of precision treatments for XWLC (Extended Fig.5b).

Tumor microbiota composition and radiomic features across subtypes

Host-bacterial immune interactions profoundly influence tumorigenesis, cancer progression, and response to therapy[52-54]. Thus, we further generated species-level resolution data allowing the identification of specific bacteria across subgroups. Firstly, we extracted and mapped high-quality reads of bacterial origin from RNA-Seq datasets across normal and four tumor subgroups of the XWLC cohort using PathSeq[55] while carefully mitigating the influences of microbial contamination (Supplementary Table 9 and Methods). We found that the subgroups showed distinct bacterial abundances (Fig. 6a). Principal coordinate analysis (PCoA)[56] of UniFrac distances of species abundances across subgroups revealed that the first PCoA axis generally captured differences between tumor and adjacent normal tissues, while the second axis mainly captured the differences between the MC-IV subgroup and the other three subgroups (Fig. 6b). Thus, the MC-IV subgroup showed distinguishable bacterial composition compared to the other three subgroups. We further identified 16 bacterial species that were enriched in the MC-IV subgroup, mainly belonging to the *Pseudomonas*,

Stenotrophomonas, *Acidovorax*, *Rheinheimera*, *Brevundimonas*, *Sphingomonas*, and *Methyloversatilis genera* (Fig. 6c). Among them, *Acidovorax* and *Brevundimonas* have been identified within lung tumor sections[57]. Further investigation is warranted to determine whether and how the unique microbiota composition of MC-IV contributes to its malignancy.

Furthermore, we built a noninvasive method to distinguish MC subtypes with radiomics which entails the extensive quantification of tumor phenotypes by utilizing numerous quantitative image features. In the initial step, we defined 107 quantitative image features that describe various characteristics of tumor phenotypes, including tumor image intensity, size, shape, and texture. These features were derived from X-ray computed tomography (CT) scans of 155 patients with XWLC (Methods). The baseline characteristics of this cohort can be found in Supplementary Table 10. Firstly, all features were compared among the four subtypes, and notably, eight features showed significant differences between the MC-II subtype and the other three subtypes (Fig. 6d), suggesting a denser image in the MC-II subtype. We further established a signature using a multivariate linear regression model with five image features to distinguish MC-II from the other three subgroups (Extended Fig.6). The performance of the five-feature radiomic signature was validated using the AUC value, which is a generation of the area under

the ROC curve. The radiomic signature had an AUC value of 0.94 in the training set and 0.83 in the validation set (Fig. 6e). The confusion matrix revealed an overall accuracy of 0.875 for sample classification using the signature, indicating proficient performance. However, it exhibited suboptimal performance in terms of false-negative classification (Fig. 6f). Taken together, we found that MC-II showed a dense image phenotype, which can be noninvasively distinguished using radiomic features.

Identification of novel targets based on mutation-informed protein-protein interface (PPI) analysis

The integration of genomics and interactomics has enabled the discovery of functional and biological consequences of disease mutations[58, 59]. To explore novel targets with the concept, we created PPI networks with structural resolution using missense mutations from the XWLC, CNLC, TSLC, and TNLC cohorts (Fig.7a and Methods). OncoPPIs, defined as a significant enrichment of interface mutations in either of the two protein-binding partners across individuals, were identified in each cohort and were provided in Supplementary Tables 11. The OncoPPIs from the four cohorts are named XWLC_oncoPPIs, CNLC_oncoPPIs, TSLC_oncoPPIs, and TNLC_oncoPPIs, respectively (Fig. 7b and Extended Fig. 7a-c). Initially, the nodes from these four OncoPPIs were subjected to biological process enrichment analysis (Extended Fig. 7d). The analysis revealed that biological processes such as regulation of

466 mitotic cell cycle, TGF-beta signaling pathway, and immune system were
467 predominantly enriched in the genes related to OncoPPIs. Moreover, the
468 processes disrupted by interface mutations showed a relatively higher
469 similarity between the XWLC and TSLC cohorts (Extended Fig. 7d)
470 suggesting convergent targets or pathways affected by smoke-induced
471 mutations.

472 To refine the novel targets from XWLC_oncoPPs, we performed
473 molecular dynamics simulations to predict the binding affinity change by
474 the interface-located mutations (Methods). Mitotic Arrest Deficient 1
475 Like 1 (MAD1), a crucial component of the mitotic spindle-assembly
476 checkpoint[60, 61], forms a tight core complex with MAD2, facilitating
477 the binding of MAD2 to CDC20[62] , which plays a critical role in sister
478 chromatid separation during the metaphase-anaphase transition[63].
479 Specifically, MAD1 Arg558His has been identified as a susceptibility
480 factor for lung cancer[64] and colorectal cancer[65]. Here, we found that
481 MAD1 allele carrying a p.Arg558His substitution may disrupt the
482 interaction between MAD1 and MAD2 (Fig.7b-c). To assess this, we
483 performed molecular dynamics simulations and found that the binding
484 affinity between Arg558His MAD1 and MAD2 was -195.091 kJ/mol and
485 that of wild type was -442.712 kJ/mol (Fig.7c). Furthermore, on a per-
486 residue basis, the predicted binding affinity (ΔΔG) of Arg558His
487 (Extended Fig.7e) was projected to increase by 118.319 kJ/mol relative to

the wild type (Extended Fig.7f), indicating that the substitution of Arg558His in MAD1 perturbs the binding affinity. Thus, our findings suggest that the MAD1 Arg558His attributed to lung cancer progression by disrupting of the interaction between MAD1 and MAD2, which showed potential to be explored as a target.

Notably, we identified TPRN as a novel significantly mutated gene in the XWLC cohort (Extended Fig.7g and 7h) whose status was also associated with patients' outcomes (Fig.7e). Previous studies have reported that TPRN interacts with PPP1CA[66]. Thus, we assessed the binding affinity of the TPRN-PPP1CA complex affected by the mutant variant His550Gln. Our result showed that the binding affinities of the complex were -694.372 kJ/mol and -877.570 kJ/mol in mutant and WT cases, respectively (Fig.7d). On a per-residue basis, the predicted ΔΔG of His550Gln compared to the wild type exhibited an increase of 96.774 kJ/mol (Extended Fig. 7i-j). All these results indicated that TPRN His550Gln increase the binding affinity of the TPRN-PPP1CA complex. To investigate the effect of TPRN His550Gln mutation on tumor progression, we examined proliferation and migration capabilities in both A549 and H1299 lung adenocarcinoma cell lines. CCK-8 assay showed significantly enhanced cell growth after transfection of TPRN mutant allele in both A549 (Fig.7f) and H1299 cells (Extended Fig. 8a). Moreover, wound-healing assay showed that TPRN mutant cell had

achieved enhanced migration capacity (Fig.7g-h and Extended Fig. 8b-c). Finally, more cell clones in TPRN His550Gln mutation cells were observed in both TPRN-mutant A549 and H1299 cells (Fig.7i). All these results supported that TPRN His550Gln could be explored as a target in XWLC.

Taken together, our integrated analysis of oncoPPIs and molecular dynamics simulations showed potential to explore novel therapeutic vulnerabilities.

Discussion

In this study, we conducted proteogenomic and characterized air-pollution-related lung cancers. We found that Benzo[a]pyrene (BaP) influenced the mutation landscape, particularly the EGFR-G719X hotspot found in 20% of cases. This mutation correlated with elevated MAPK pathway activation, worse clinical outcomes, and younger patients. Multi-omics clustering identified four subtypes with unique biological pathways and immune cell patterns. Moreover, our analysis of protein-protein interfaces unveiled novel therapeutic targets. These findings have significant implications for preventing and developing precise treatments for air-pollution-associated lung cancers.

Previously considered uncommon, the EGFR-G719X mutation was detected in only 1-2% of CNLC or TCGA-LUAD cohort samples. Limited knowledge exists from G719X, mostly based on isolated case

reports or small series studies[67-71]. In vitro experiments using G719X mutant cell lines and patient-derived xenografts (PDX) demonstrated that osimertinib effectively inhibits signaling pathways and cellular growth, leading to sustained tumor growth inhibition[72]. However, in silico protein structure analysis suggests that G719 alterations may confer osimertinib resistance due to reduced EGFR binding[73]. Presently, afatinib is proposed as the first-line therapy for G719X mutation patients[39-42]. Unfortunately, 80% of G719X patients develop acquired resistance to afatinib without detecting the T790M mutation40. Hence, further mechanistic studies are warranted for G719X. Our study reveals that the G719X mutation is prevalent in the XWLC cohort, significantly impacting treatment selection. Additionally, the large number of G719X samples allowed us to uncover variations in biology and pathway activation, which may facilitate the development of more precise targeted therapies for these patients.

There is substantial evidence linking lung cancer in the Xuanwei area to coal smoke[18, 19, 21, 22, 74, 75]. In addition, we conducted a rat model study that demonstrated the induction of lung cancer by local coal smoke exposure[22]. However, the specific chemical compound in coal smoke responsible for causing lung cancer remains largely unknown. Previous research has mainly focused on studying indoor concentrations of airborne particles and BaP[18, 19, 21, 23, 76]. For instance, studies

554 have shown an association between the concentration of BaP and lung
555 cancer rates across counties[18]. Moreover, improvements in household
556 stoves have led to reduced exposure to benzopyrene and particulate
557 matter, benefiting people's health[23, 76]. However, these studies
558 primarily relied on epidemiological data, which may be influenced by
559 confounding factors. In our study, we used clinical samples and linked the
560 mutational signatures of XWLC to the chemical compound BaP, which
561 advanced the etiology and mechanism of air-pollution-induced lung
562 cancer.

563 In summary, our proteogenomic analysis of clinical tumor samples
564 provides insights into air-pollution-associated lung cancers, especially
565 those induced by coal smoke and offers an opportunity to expedite the
566 translation of basic research to more precise diagnosis and treatment in
567 the clinic.

568 **Declarations**

569 **Ethical Approval**

570 The study protocol was reviewed and approved by the ethical committees
571 of Yunnan Cancer Hospital & The Third Affiliated Hospital of Kunming
572 Medical University (KYCS2022067) and conformed to the ethical
573 standards for medical research involving human subjects, as laid out in
574 the 1964 Declaration of Helsinki and its later amendments.

575 **Consent to participate**

Participants provided written informed consent prior to taking part in the study.

Consent for publication

Written informed consent was obtained from the patient and the ethical committees of Yunnan Cancer Hospital & The Third Affiliated Hospital of Kunming Medical University for publication of this study.

Competing interests

The authors declare no competing interests.

Authors' contributions

Honglei Zhang, Gaofeng Li, Xiaosan Su and Yong Zhang conceived the study. Honglei Zhang, Chao Liu, Qing Wang and Shuting Wang performed data analysis. Honglei Zhang wrote the manuscript. Xu Feng, Heng Li, Zhiyong Deng, Huawei Jiang, Yunyan He, Chao Luo, Rou Qian, Minjun Zhou, Zhi Li, Li Xiao, Yong Zhang interpreted the data analysis and drafted the manuscript and critically revised the manuscript. All authors critically revised and gave final approval of the manuscript.

Funding

This work was supported by National Natural Science Foundation (Nos. 82273501, 81960322, 82160343, 82060519, 81860171); Yunnan young and middle-aged academic and technical leaders Reserve Talents Project (No. 202005AC160048); Yunnan basic research program (Nos.202101AZ070001-002, 202001AY070001-277, 2019FE001(-236));

598 Medical Reserve Personnel Training Program of Yunnan Provincial
599 Health Commission(No. H-2018097); “Famous Doctor” Special Project
600 of Ten Thousand People Plan of Yunnan Province(Nos.CZ0096, YNWR-
601 MY-2020-095); Medical Leading Talents Training Program of Yunnan
602 Provincial Health Commission (No.L-2019028).

603 **Acknowledgements**

604 We thank Novogene Co., Ltd. and Beijing Qinglian Biotech Co., Ltd. for
605 the analysis of mass spectrometric data.

606 **Availability of data and materials**

607 Raw sequencing data have been deposited in the Genome Sequence
608 Archive for Human (GSA-Human, https://ngdc.cncb.ac.cn/gsa-human/)
609 under accession codes HRA000124, HRA001481 and HRA001482.
610 Proteomics and phosphoproteomics data have been deposited in the Open
611 Archive for Miscellaneous Data (OMIX, https://ngdc.cncb.ac.cn/omix/)
612 under accession codes OMIX001292. The raw lung CT images used in
613 this paper are available from OMIX under accession codes OMIX002491.

614

615

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: **Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA Cancer J Clin* 2018, **68:**394-424.
2. Parkin DM, Bray F, Ferlay J, Pisani P: **Global cancer statistics, 2002.** *CA Cancer J Clin* 2005, **55:**74-108.
3. Chen J, Yang H, Teo ASM, Amer LB, Sherbaf FG, Tan CQ, Alvarez JJS, Lu B, Lim JQ, Takano A, et al: **Genomic landscape of lung adenocarcinoma in East Asians.** *Nat Genet* 2020, **52:**177-186.
4. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, Maher CA, Fulton R, Fulton L, Wallis J, et al: **Genomic landscape of non-small cell lung cancer in smokers and never-smokers.** *Cell* 2012, **150:**1121-1134.
5. Chen YJ, Roumeliotis TI, Chang YH, Chen CT, Han CL, Lin MH, Chen HW, Chang GC, Chang YL, Wu CT, et al: **Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression.** *Cell* 2020, **182:**226-244.e217.
6. Zhang T, Joubert P, Ansari-Pour N, Zhao W, Hoang PH, Lokanga R, Moye AL, Rosenbaum J, Gonzalez-Perez A, Martinez-Jimenez F, et al: **Genomic and evolutionary classification of lung cancer in never smokers.** *Nat Genet* 2021, **53:**1348-1359.
7. Zhang XC, Wang J, Shao GG, Wang Q, Qu X, Wang B, Moy C, Fan Y, Albertyn Z, Huang X, et al: **Comprehensive genomic and immunological characterization of Chinese non-small cell lung cancer patients.** *Nat Commun* 2019, **10:**1772.
8. Xu JY, Zhang C, Wang X, Zhai L, Ma Y, Mao Y, Qian K, Sun C, Liu Z, Jiang S, et al: **Integrative Proteomic Characterization of Human Lung Adenocarcinoma.** *Cell* 2020, **182:**245-261.e217.
9. Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, Petralia F, Li Y, Liang WW, Reva B, et al: **Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma.** *Cell* 2020, **182:**200-225 e235.
10. Hill W, Lim EL, Weeden CE, Lee C, Augustine M, Chen K, Kuan FC, Marongiu F, Evans EJ, Jr., Moore DA, et al: **Lung adenocarcinoma promotion by air pollutants.** *Nature* 2023, **616:**159-167.
11. Turner MC, Andersen ZJ, Baccarelli A, Diver WR, Gapstur SM, Pope CA, 3rd, Prada D, Samet J, Thurston G, Cohen A: **Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations.** *CA Cancer J Clin* 2020.
12. Fajersztajn L, Veras M, Barrozo LV, Saldiva P: **Air pollution: a potentially modifiable risk factor for lung cancer.** *Nat Rev Cancer* 2013, **13:**674-678.
13. Cogliano VJ, Baan R, Straif K, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Freeman C, et al: **Preventable exposures associated with human cancers.** *J Natl Cancer Inst* 2011, **103:**1827-1839.
14. Humans IWGotEoCRt: **Some non-heterocyclic polycyclic aromatic hydrocarbons and some related exposures.** *IARC Monogr Eval Carcinog Risks Hum* 2010, **92:**1-853.
15. Shi Q, Godschalk RWL, van Schooten FJ: **Inflammation and the chemical carcinogen benzo[a]pyrene: Partners in crime.** *Mutat Res Rev Mutat Res* 2017, **774:**12-24.

16. Saravanakumar K, Sivasantosh S, Sathiyaseelan A, Sankaranarayanan A, Naveen KV, Zhang X, Jamla M, Vijayasarathy S, Vishnu Priya V, MubarakAli D, Wang MH: **Impact of benzo[a]pyrene with other pollutants induce the molecular alternation in the biological system: Existence, detection, and remediation methods.** *Environ Pollut* 2022, **304:**119207.
17. Abd El-Fattah EE, Abdelhamid AM: **Benzo[a]pyrene immunogenetics and immune archetype reprogramming of lung.** *Toxicology* 2021, **463:**152994.
18. Mumford JL, He XZ, Chapman RS, Cao SR, Harris DB, Li XM, Xian YL, Jiang WZ, Xu CW, Chuang JC, et al.: **Lung cancer and indoor air pollution in Xuan Wei, China.** *Science* 1987, **235:**217-220.
19. Barone-Adesi F, Chapman RS, Silverman DT, He X, Hu W, Vermeulen R, Ning B, Fraumeni JF, Jr., Rothman N, Lan Q: **Risk of lung cancer associated with domestic use of coal in Xuanwei, China: retrospective cohort study.** *BMJ* 2012, **345:**e5414.
20. Chapman RS, He X, Blair AE, Lan Q: **Improvement in household stoves and risk of chronic obstructive pulmonary disease in Xuanwei, China: retrospective cohort study.** *BMJ* 2005, **331:**1050.
21. Lin H, Ning B, Li J, Ho SC, Huss A, Vermeulen R, Tian L: **Lung cancer mortality among women in Xuan Wei, China: a comparison of spatial clustering detection methods.** *Asia Pac J Public Health* 2015, **27:**NP392-401.
22. Zhang H, Liu C, Li L, Feng X, Wang Q, Li J, Xu S, Wang S, Yang Q, Shen Z, et al: **Genomic evidence of lung carcinogenesis associated with coal smoke in Xuanwei area, China.** *Natl Sci Rev* 2021, **8:**nwab152.
23. Lan Q, Chapman RS, Schreinemachers DM, Tian L, He X: **Household stove improvement and risk of lung cancer in Xuanwei, China.** *J Natl Cancer Inst* 2002, **94:**826-835.
24. Hosgood HD, Pao W, Rothman N, Hu W, Pan YH, Kuchinsky K, Jones KD, Xu J, Vermeulen R, Simko J, Lan Q: **Driver mutations among never smoking female lung cancer tissues in China identify unique EGFR and KRAS mutation pattern associated with household coal burning.** *Respiratory Medicine* 2013, **107:**1755-1762.
25. Wang J, Duan Y, Meng QH, Gong R, Guo C, Zhao Y, Zhang Y: **Integrated analysis of DNA methylation profiling and gene expression profiling identifies novel markers in lung cancer in Xuanwei, China.** *PLoS One* 2018, **13:**e0203155.
26. Wang X, Li J, Duan Y, Wu H, Xu Q, Zhang Y: **Whole genome sequencing analysis of lung adenocarcinoma in Xuanwei, China.** *Thorac Cancer* 2017, **8:**88-96.
27. Cancer Genome Atlas Research N: **Comprehensive molecular profiling of lung adenocarcinoma.** *Nature* 2014, **511:**543-550.
28. Gehring JS, Fischer B, Lawrence M, Huber W: **SomaticSignatures: inferring mutational signatures from single-nucleotide variants.** *Bioinformatics* 2015, **31:**3673-3675.
29. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR: **Deciphering signatures of mutational processes operative in human cancer.** *Cell Rep* 2013, **3:**246-259.
30. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, et al: **Signatures of mutational processes in human cancer.** *Nature* 2013, **500:**415-421.
31. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, Gomez C, Degasperi A, Harris R, Jackson SP, et al: **A Compendium of Mutational Signatures of Environmental Agents.** *Cell* 2019, **177:**821-836 e816.

32. Petit P, Maitre A, Persoons R, Bicout DJ: **Lung cancer risk assessment for workers exposed to polycyclic aromatic hydrocarbons in various industries.** *Environment International* 2019, **124:**109-120.
33. Widziewicz K, Rogula-Kozlowska W, Majewski G: **Lung Cancer Risk Associated with Exposure to Benzo(A)Pyrene in Polish Agglomerations, Cities, and Other Areas.** *International Journal of Environmental Research* 2017, **11:**685-693.
34. Mangal D, Vudathala D, Park JH, Lee SH, Penning TM, Blair IA: **Analysis of 7,8-dihydro-8-oxo-2'-deoxyguanosine in cellular DNA during oxidative stress.** *Chem Res Toxicol* 2009, **22:**788-797.
35. Chung JY, Kim JY, Kim YJ, Jung SJ, Park JE, Lee SG, Kim JT, Oh S, Lee CJ, Yoon YD, et al: **Cellular defense mechanisms against benzo[a]pyrene in testicular Leydig cells: implications of p53, aryl-hydrocarbon receptor, and cytochrome P450 1A1 status.** *Endocrinology* 2007, **148:**6134-6144.
36. Hidaka T, Fujimura T, Aiba S: **Aryl Hydrocarbon Receptor Modulates Carcinogenesis and Maintenance of Skin Cancers.** *Front Med (Lausanne)* 2019, **6:**194.
37. Wiredja DD, Koyuturk M, Chance MR: **The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics.** *Bioinformatics* 2017, **33:**3489-3491.
38. Casado P, Rodriguez-Prados JC, Cosulich SC, Guichard S, Vanhaesebroeck B, Joel S, Cutillas PR: **Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells.** *Sci Signal* 2013, **6:**rs6.
39. Ettinger DS, Wood DE, Aisner DL, Akerley W, Bauman JR, Bharat A, Bruno DS, Chang JY, Chirieac LR, D'Amico TA, et al: **Non-Small Cell Lung Cancer, Version 3.2022, NCCN Clinical Practice Guidelines in Oncology.** *J Natl Compr Canc Netw* 2022, **20:**497-530.
40. Janning M, Suptitz J, Albers-Leischner C, Delpy P, Tufman A, Velthaus-Rusik JL, Reck M, Jung A, Kauffmann-Guerrero D, Bonzheim I, et al: **Treatment outcome of atypical EGFR mutations in the German National Network Genomic Medicine Lung Cancer (nNGM).** *Ann Oncol* 2022, **33:**602-615.
41. Yang JC, Schuler M, Popat S, Miura S, Heeke S, Park K, Marten A, Kim ES: **Afatinib for the Treatment of NSCLC Harboring Uncommon EGFR Mutations: A Database of 693 Cases.** *J Thorac Oncol* 2020, **15:**803-815.
42. Cho JH, Lim SH, An HJ, Kim KH, Park KU, Kang EJ, Choi YH, Ahn MS, Lee MH, Sun JM, et al: **Osimertinib for Patients With Non-Small-Cell Lung Cancer Harboring Uncommon EGFR Mutations: A Multicenter, Open-Label, Phase II Trial (KCSG-LU15-09).** *J Clin Oncol* 2020, **38:**488-495.
43. Harada T, Futamura S, Inoue Y, Sawada R, Okuda T, Kagawa K: **P2.03-13 Acquired Resistance to Afatinib in Non-Small Cell Lung Cancer with EGFR G719X Mutation.** *Journal of Thoracic Oncology* 2019, **14:**S687.
44. Wilkerson MD, Hayes DN: **ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.** *Bioinformatics* 2010, **26:**1572-1573.
45. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P: **The Molecular Signatures Database (MSigDB) hallmark gene set collection.** *Cell Syst* 2015, **1:**417-425.
46. Vander Heiden MG, Cantley LC, Thompson CB: **Understanding the Warburg effect: the metabolic requirements of cell proliferation.** *Science* 2009, **324:**1029-1033.

47. DeBerardinis RJ, Thompson CB: **Cellular metabolism and disease: what do metabolic outliers teach us?** *Cell* 2012, **148:**1132-1144.
48. Cairns RA, Harris IS, Mak TW: **Regulation of cancer cell metabolism.** *Nat Rev Cancer* 2011, **11:**85-95.
49. Carmeliet P: **Angiogenesis in life, disease and medicine.** *Nature* 2005, **438:**932-936.
50. Pitulescu ME, Schmidt I, Giaimo BD, Antoine T, Berkenfeld F, Ferrante F, Park H, Ehling M, Biljes D, Rocha SF, et al: **Dll4 and Notch signalling couples sprouting angiogenesis and artery formation.** *Nat Cell Biol* 2017, **19:**915-927.
51. Gridley T: **Notch signaling in vascular development and physiology.** *Development* 2007, **134:**2709-2718.
52. Davar D, Dzutsev AK, McCulloch JA, Rodrigues RR, Chauvin JM, Morrison RM, Deblasio RN, Menna C, Ding Q, Pagliano O, et al: **Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients.** *Science* 2021, **371:**595-602.
53. Dzutsev A, Badger JH, Perez-Chanona E, Roy S, Salcedo R, Smith CK, Trinchieri G: **Microbes and Cancer.** *Annu Rev Immunol* 2017, **35:**199-228.
54. Finlay BB, Goldszmid R, Honda K, Trinchieri G, Wargo J, Zitvogel L: **Can we harness the microbiota to enhance the efficacy of cancer immunotherapy?** *Nat Rev Immunol* 2020, **20:**522-528.
55. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M: **PathSeq: software to identify or discover microbes by deep sequencing of human tissue.** *Nat Biotechnol* 2011, **29:**393-396.
56. Lozupone C, Knight R: **UniFrac: a new phylogenetic method for comparing microbial communities.** *Appl Environ Microbiol* 2005, **71:**8228-8235.
57. Greathouse KL, White JR, Vargas AJ, Bliskovsky VV, Beck JA, von Muhlinen N, Polley EC, Bowman ED, Khan MA, Robles AI, et al: **Interaction between the microbiome and TP53 in human lung cancer.** *Genome Biol* 2018, **19:**123.
58. Sahni N, Yi S, Taipale M, Fuxman Bass JI, Coulombe-Huntington J, Yang F, Peng J, Weile J, Karras GI, Wang Y, et al: **Widespread macromolecular interaction perturbations in human genetic disorders.** *Cell* 2015, **161:**647-660.
59. Cheng F, Zhao J, Wang Y, Lu W, Liu Z, Zhou Y, Martin WR, Wang R, Huang J, Hao T, et al: **Comprehensive characterization of protein-protein interactions perturbed by disease mutations.** *Nat Genet* 2021, **53:**342-353.
60. Wang Y, Skibbe JR, Hu C, Dong L, Ferchen K, Su R, Li C, Huang H, Weng H, Huang H, et al: **ALOX5 exhibits anti-tumor and drug-sensitizing effects in MLL-rearranged leukemia.** *Sci Rep* 2017, **7:**1853.
61. Tsukasaki K, Miller CW, Greenspun E, Eshaghian S, Kawabata H, Fujimoto T, Tomonaga M, Sawyers C, Said JW, Koeffler HP: **Mutations in the mitotic check point gene, MAD1L1, in human cancers.** *Oncogene* 2001, **20:**3301-3305.
62. Yang M, Li B, Liu CJ, Tomchick DR, Machius M, Rizo J, Yu H, Luo X: **Insights into mad2 regulation in the spindle checkpoint revealed by the crystal structure of the symmetric mad2 dimer.** *PLoS Biol* 2008, **6:**e50.
63. Sironi L, Mapelli M, Knapp S, De Antoni A, Jeang KT, Musacchio A: **Crystal structure of the tetrameric Mad1-Mad2 core complex: implications of a 'safety belt' binding mechanism for the spindle checkpoint.** *EMBO J* 2002, **21:**2496-2506.

64. Guo Y, Zhang X, Yang M, Miao X, Shi Y, Yao J, Tan W, Sun T, Zhao D, Yu D, et al: **Functional evaluation of missense variations in the human MAD1L1 and MAD2L1 genes and their impact on susceptibility to lung cancer.** *J Med Genet* 2010, **47:**616-622.

65. Zhong R, Chen X, Chen X, Zhu B, Lou J, Li J, Shen N, Yang Y, Gong Y, Zhu Y, et al: **MAD1L1 Arg558His and MAD2L1 Leu84Met interaction with smoking increase the risk of colorectal cancer.** *Sci Rep* 2015, **5:**12202.

66. Ferrar T, Chamousset D, De Wever V, Nimick M, Andersen J, Trinkle-Mulcahy L, Moorhead GB: **Taperin (c9orf75), a mutated gene in nonsyndromic deafness, encodes a vertebrate specific, nuclear localized protein phosphatase one alpha (PP1alpha) docking protein.** *Biol Open* 2012, **1:**128-139.

67. Chiu CH, Yang CT, Shih JY, Huang MS, Su WC, Lai RS, Wang CC, Hsiao SH, Lin YC, Ho CL, et al: **Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitor Treatment Response in Advanced Lung Adenocarcinomas with G719X/L861Q/S768I Mutations.** *J Thorac Oncol* 2015, **10:**793-799.

68. Kunishige M, Ichihara S, Kadota N, Okano Y, Machida H, Hatakeyama N, Naruse K, Shinohara T, Takeuchi E: **Non-small cell lung cancer with EGFR (L858R and E709X) and CNNB1 mutations responded to afatinib.** *Thorac Cancer* 2022.

69. Yang JC, Sequist LV, Geater SL, Tsai CM, Mok TS, Schuler M, Yamamoto N, Yu CJ, Ou SH, Zhou C, et al: **Clinical activity of afatinib in patients with advanced non-small-cell lung cancer harbouring uncommon EGFR mutations: a combined post-hoc analysis of LUX-Lung 2, LUX-Lung 3, and LUX-Lung 6.** *Lancet Oncol* 2015, **16:**830-838.

70. Massarelli E, Johnson FM, Erickson HS, Wistuba, II, Papadimitrakopoulou V: **Uncommon epidermal growth factor receptor mutations in non-small cell lung cancer and their mechanisms of EGFR tyrosine kinase inhibitors sensitivity and resistance.** *Lung Cancer* 2013, **80:**235-241.

71. Han SW, Kim TY, Hwang PG, Jeong S, Kim J, Choi IS, Oh DY, Kim JH, Kim DW, Chung DH, et al: **Predictive and prognostic impact of epidermal growth factor receptor mutation in non-small-cell lung cancer patients treated with gefitinib.** *J Clin Oncol* 2005, **23:**2493-2501.

72. Floc'h N, Lim S, Bickerton S, Ahmed A, Orme J, Urosevic J, Martin MJ, Cross DAE, Cho BC, Smith PD: **Osimertinib, an Irreversible Next-Generation EGFR Tyrosine Kinase Inhibitor, Exerts Antitumor Activity in Various Preclinical NSCLC Models Harboring the Uncommon EGFR Mutations G719X or L861Q or S768I.** *Mol Cancer Ther* 2020, **19:**2298-2307.

73. Yang Z, Yang N, Ou Q, Xiang Y, Jiang T, Wu X, Bao H, Tong X, Wang X, Shao YW, et al: **Investigating Novel Resistance Mechanisms to Third-Generation EGFR Tyrosine Kinase Inhibitor Osimertinib in Non-Small Cell Lung Cancer Patients.** *Clin Cancer Res* 2018, **24:**3097-3107.

74. Wong JYY, Downward GS, Hu W, Portengen L, Seow WJ, Silverman DT, Bassig BA, Zhang J, Xu J, Ji BT, et al: **Lung cancer risk by geologic coal deposits: A case-control study of female never-smokers from Xuanwei and Fuyuan, China.** *Int J Cancer* 2019, **144:**2918-2927.

75. Vermeulen R, Downward GS, Zhang J, Hu W, Portengen L, Bassig BA, Hammond SK, Wong JYY, Li J, Reiss B, et al: **Constituents of Household Air Pollution and Risk of Lung**

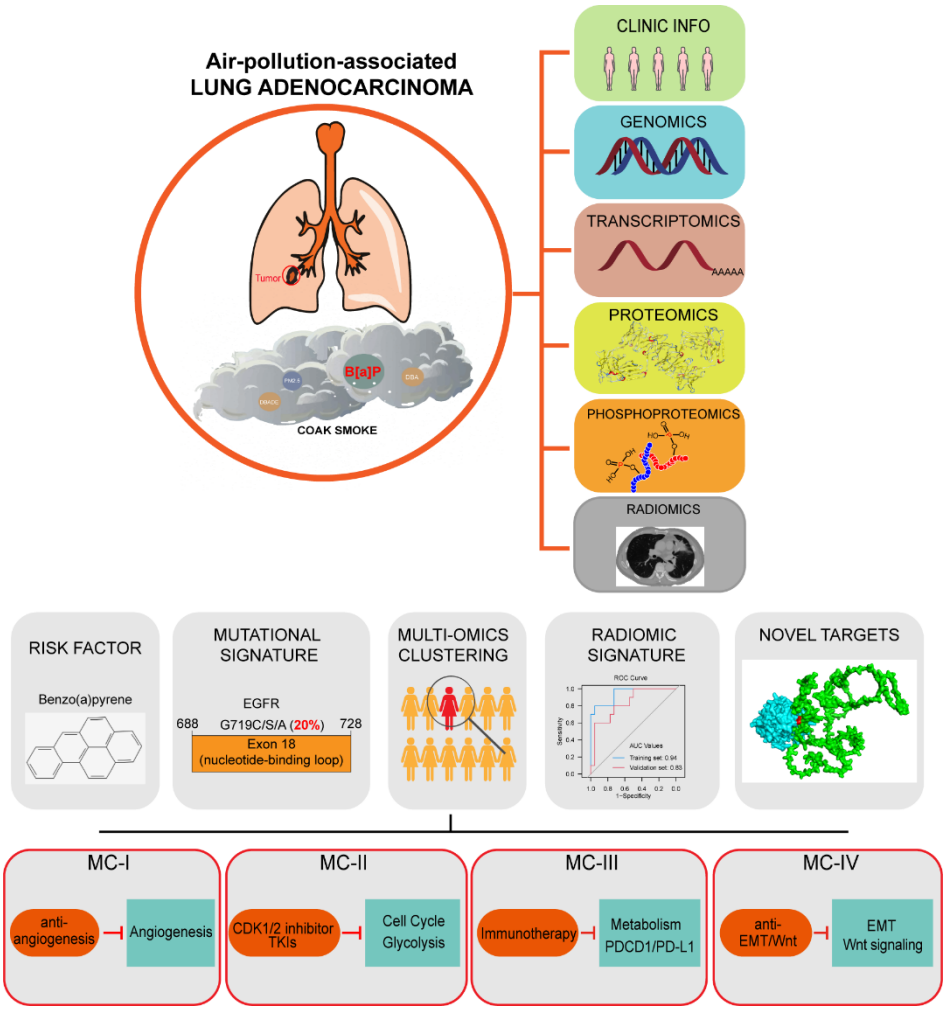
Cancer among Never-Smoking Women in Xuanwei and Fuyuan, China. *Environ Health Perspect* 2019, **127:**97001.

76. Kim C, Chapman RS, Hu W, He X, Hosgood HD, Liu LZ, Lai H, Chen W, Silverman DT, Vermeulen R, et al: **Smoky coal, tobacco smoking, and lung cancer risk in Xuanwei, China.** *Lung Cancer* 2014, **84:**31-35.

77. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z: **Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade.** *Cell Rep* 2017, **18:**248-262.

846

Graphical Abstract



847

848

849 **Highlights:**

850 - We conducted comprehensive multi-omic profiling of air-pollution-
851 associated LUAD.
852 - Our study revealed the significant roles of the air pollutant BaP and
853 its induced hot mutation G719X in lung cancer progression.
854 - Multi-omic clustering enabled the identification of personalized
855 therapeutic strategies.
856 - Through mutation-informed interface analysis, we identified novel
857 targets for therapeutic intervention.

858

859

Figure legends

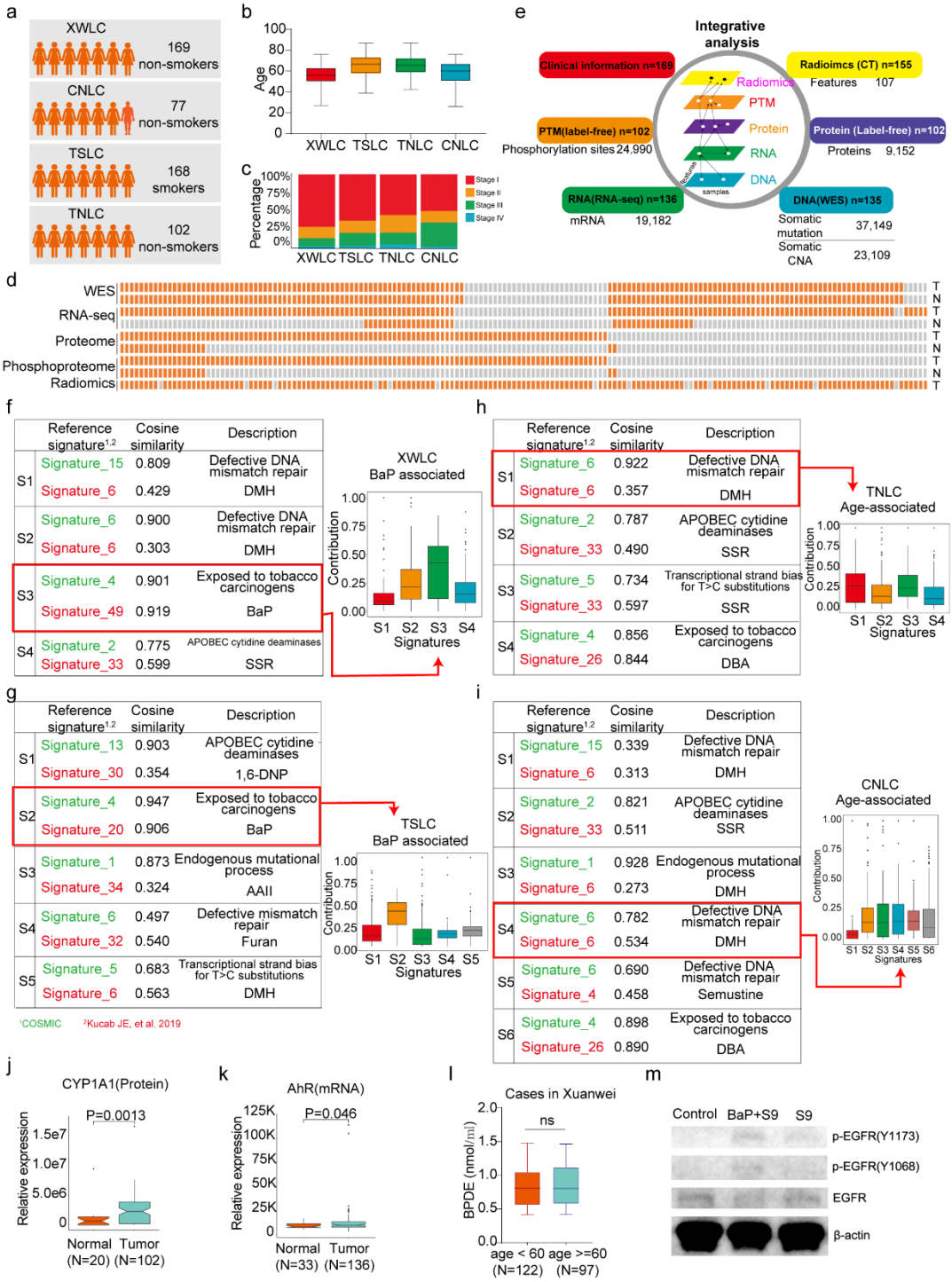


Fig 1.| Proteogenomic profiling and mutational signatures in XWLC

a. Four cohort datasets used in this study: XWLC (Lung adenocarcinoma from non-smoking females in Xuanwei area), CNLC (subset of lung adenocarcinoma from non-smoking patients in Chinese Human Proteome Project), TSLC (subset of lung adenocarcinoma from smoking females in

867 TCGA-LUAD project), TNLC (subset of lung adenocarcinoma from non-
868 smoking females in TCGA-LUAD project); b. Age distribution of
869 patients at the time of operation in the four cohorts; c. Distribution of
870 tumor stages across the cohorts; d. Data availability for the XWLC
871 datasets. Each bar represents a sample, with orange bars indicating data
872 availability and gray bars indicating data unavailability; e. Summary of
873 data generated from the XWLC cohort; f-i. Mutational signatures
874 identified in XWLC (f), TSLC (g), TNLC (h), and CNLC (i) cohorts.
875 Cosine similarity analysis of the signatures compared to well-established
876 COMIC signatures (in green) and Kucab et al. signatures (in red).
877 Contribution of signatures in each cohort provided on the right; j. Protein
878 abundance of CYP1A1 in tumor and normal samples within the XWLC
879 cohort; k. Expression levels of the AhR gene in tumor and normal
880 samples within the XWLC cohort; l. Comparison of serum BPDE content
881 in individuals from the Xuanwei area, categorized by young cases and
882 older cases. m. western blotting of EGFR-Y1173, EGFR-Y1068 and
883 EGFR abundance in hiPSC (Control), BaP+S9 treated hiPSC (BaP+S9),
884 and S9 treated hiPSC (S9). Liver S9 is a variety of biological sources
885 represent the post-mitochondrial supernatant fraction from homogenized
886 liver and is known to be a rich source of drug metabolizing enzymes
887 including P-450. Two-tailed Wilcoxon rank sum test used to calculate p-
888 values in j-l.

889

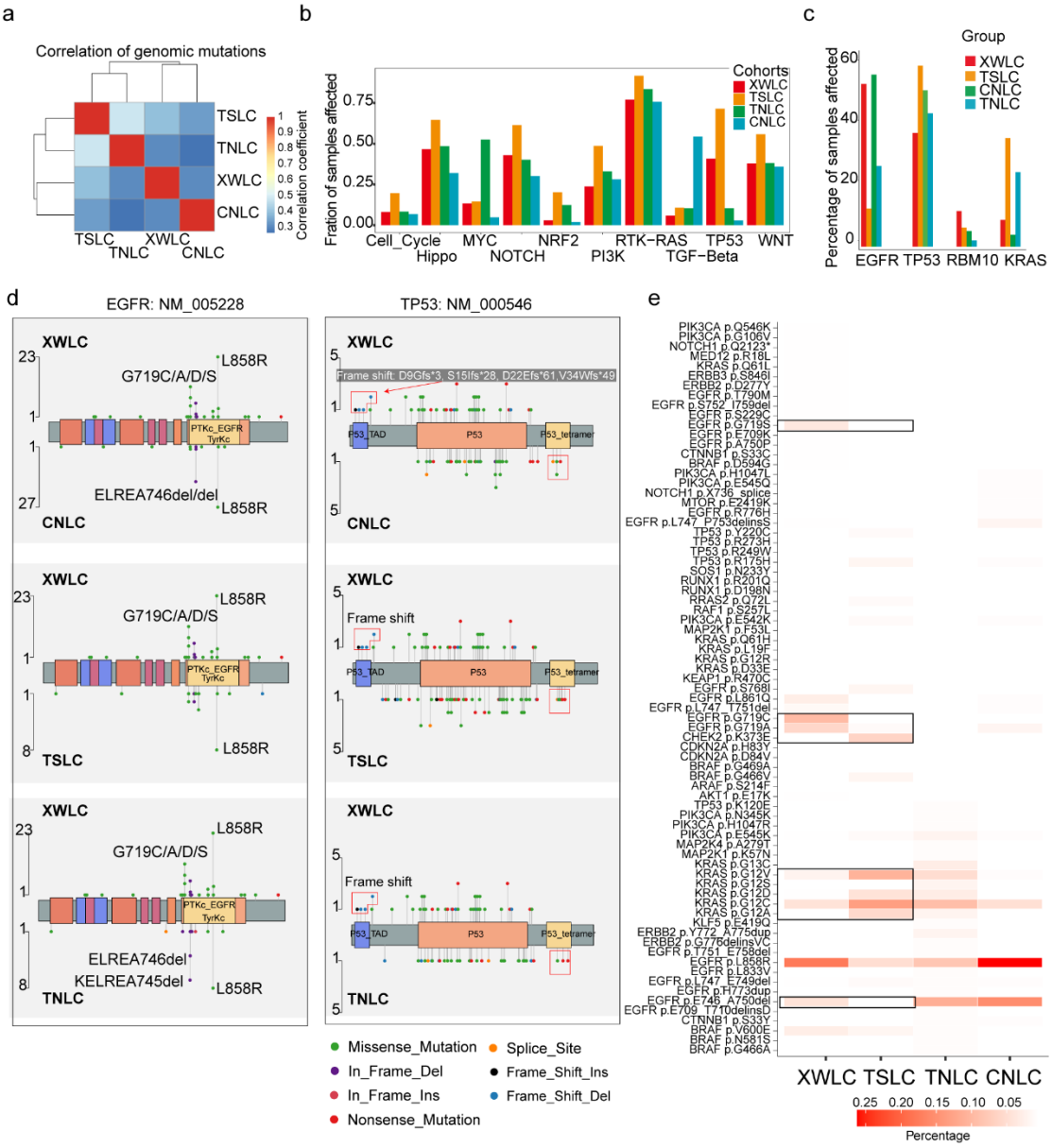


Fig 2.| Genomic and genetic features in XWLC cohort

a. Correlation of genomic mutations among cohorts, determined using the Pearson correlation coefficient; b. Comparison of oncogenic pathways affected by mutations in each cohort; c. Comparison of mutation frequency of four key genes across cohorts; d. Lollipop plot illustrating differences in mutational sites within EGFR (left) and TP53 (right) across XWLC/CNLC, XWLC/TSLC, and XWLC/TNLC pairs; e. Analysis of the percentage of samples with actionable alterations, with a focus on significant variations between XWLC and TSLC cohorts, highlighted by black boxes.

901

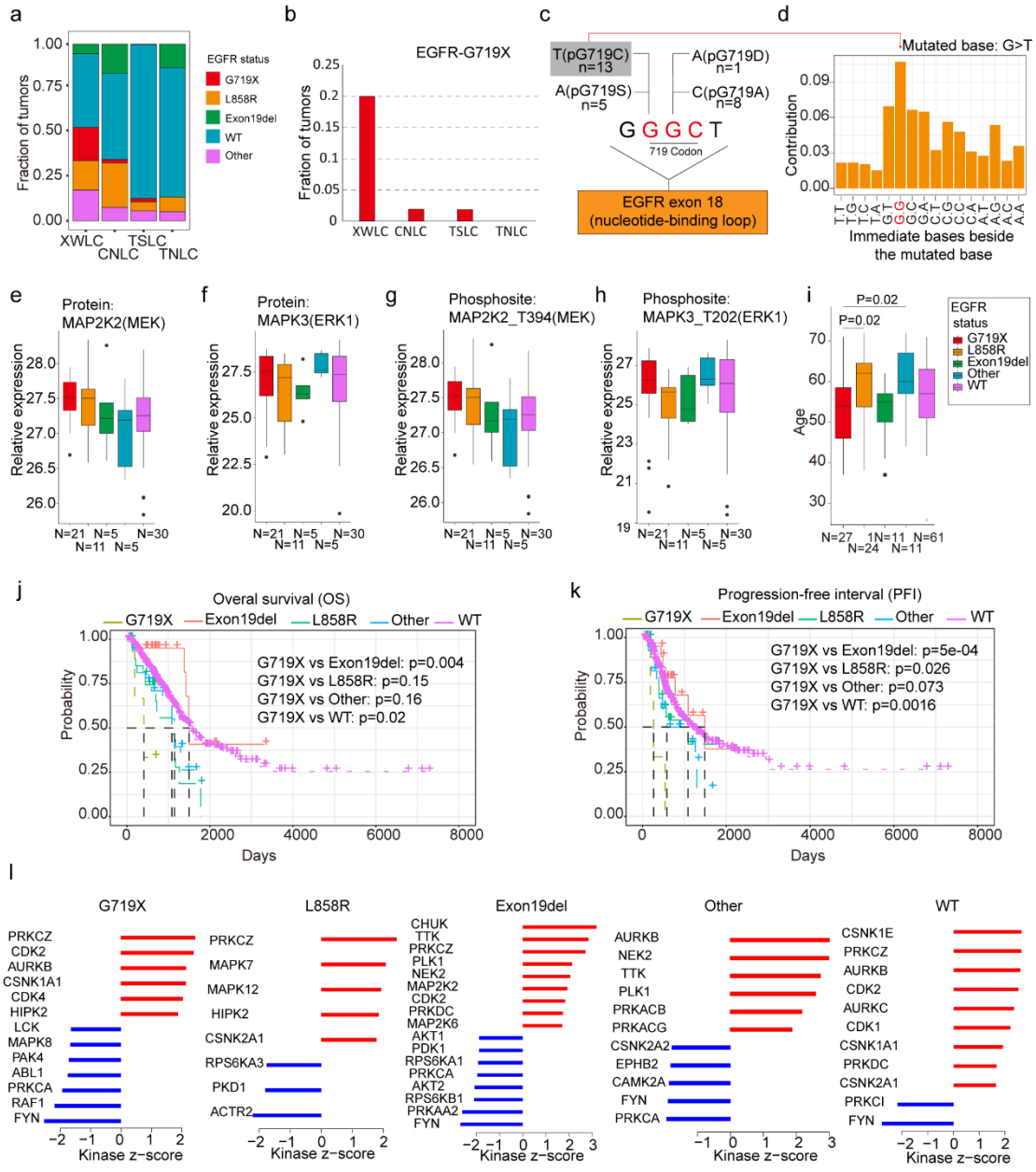


Fig 3.| EGFR-G719X in the XWLC cohort

a. Distribution of different EGFR mutation statuses across the four cohorts; b. Comparison of the fraction of G719X mutations across the four cohorts; c. Detailed information on pG719X (pG719/A/D/C/S) mutations in the XWLC cohort. The number of each mutation type is labeled; d. Distribution of nucleotide pairs surrounding the most common G>T transversion site in the XWLC cohort. The x-axis represents the immediate bases surrounding the mutated base. For BaP, the tallest G>T peak occurs at GpGpG; e-h. Comparison of activation levels of key

components in the MAPK pathway across different EGFR mutation statuses in the XWLC cohort; i. Comparison of patient ages across different EGFR mutation statuses in the XWLC cohort; j-k. Presentation of overall survival (OS, j) and progression-free interval (PFI, k) analysis across different EGFR mutation statuses in the TCGA-LUAD cohort; l. Evaluation of kinase activities by KSEA in tumors across different EGFR mutation statuses in the XWLC cohort.

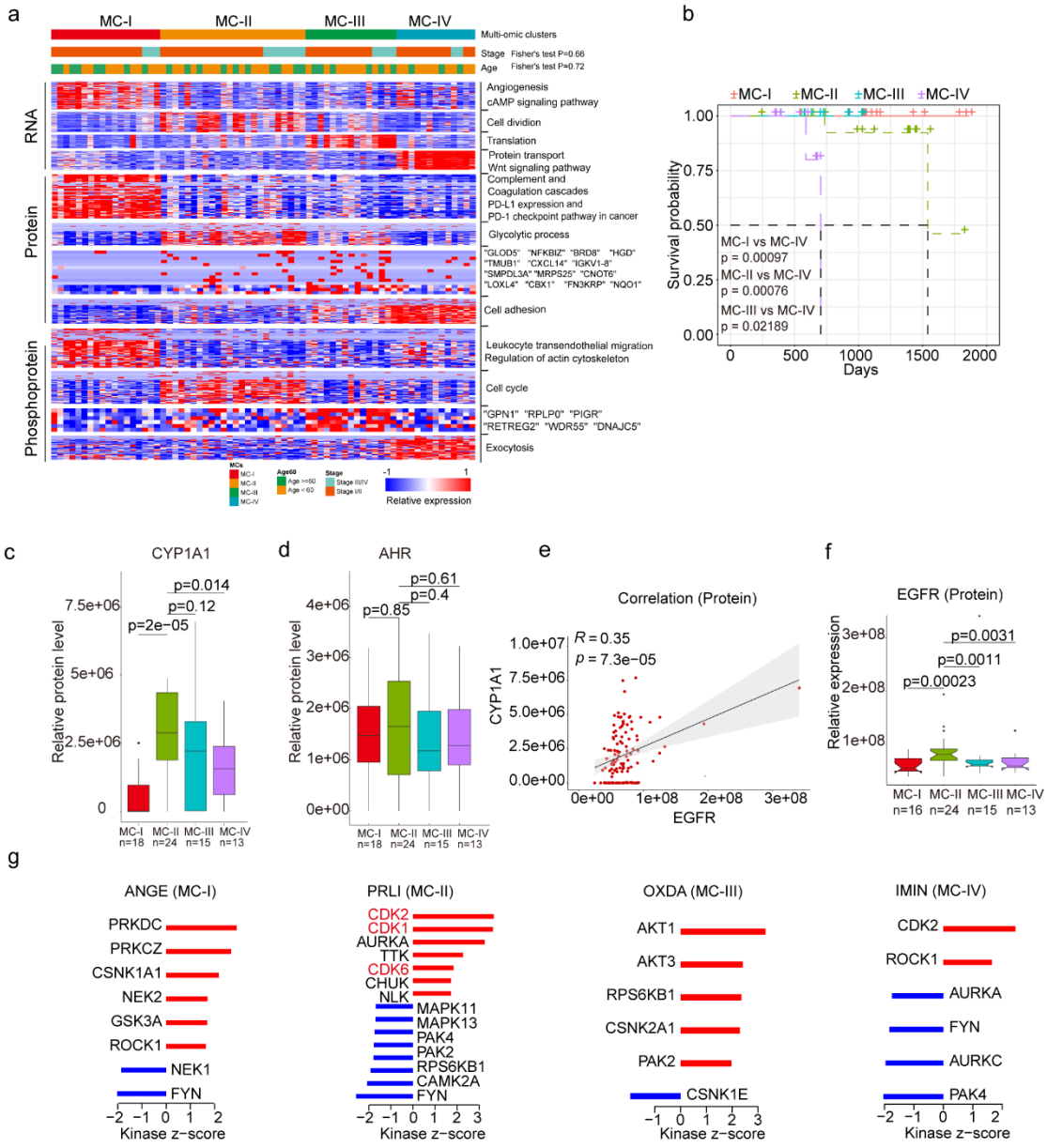


Fig 4.| Subtyping of XWLC

a. Integrative classification of tumor samples into four ConsensusClusterPlus-derived clusters (MC-I to MC-IV). The heatmap displays the top 50 features, including mRNA transcripts, proteins, and phosphoproteins, for each multi-omic cluster. The features are annotated with representative pathways or genes. If a cluster has fewer than 50 features, all features are shown. If no significant GO biological processes are associated with cluster features, all features are displayed; b. Comparison of overall survival between MC-IV and the other three subtypes; c-d. Protein abundance comparison of CYCP1A1 (c) and AHR

(d) across subtypes; e. Protein-level correlation between CYCP1A1 and EGFR; f. Protein-level comparison of EGFR across subtypes; g. Evaluation of kinase activities by KSEA in tumors across subtypes in the XWLC cohort.

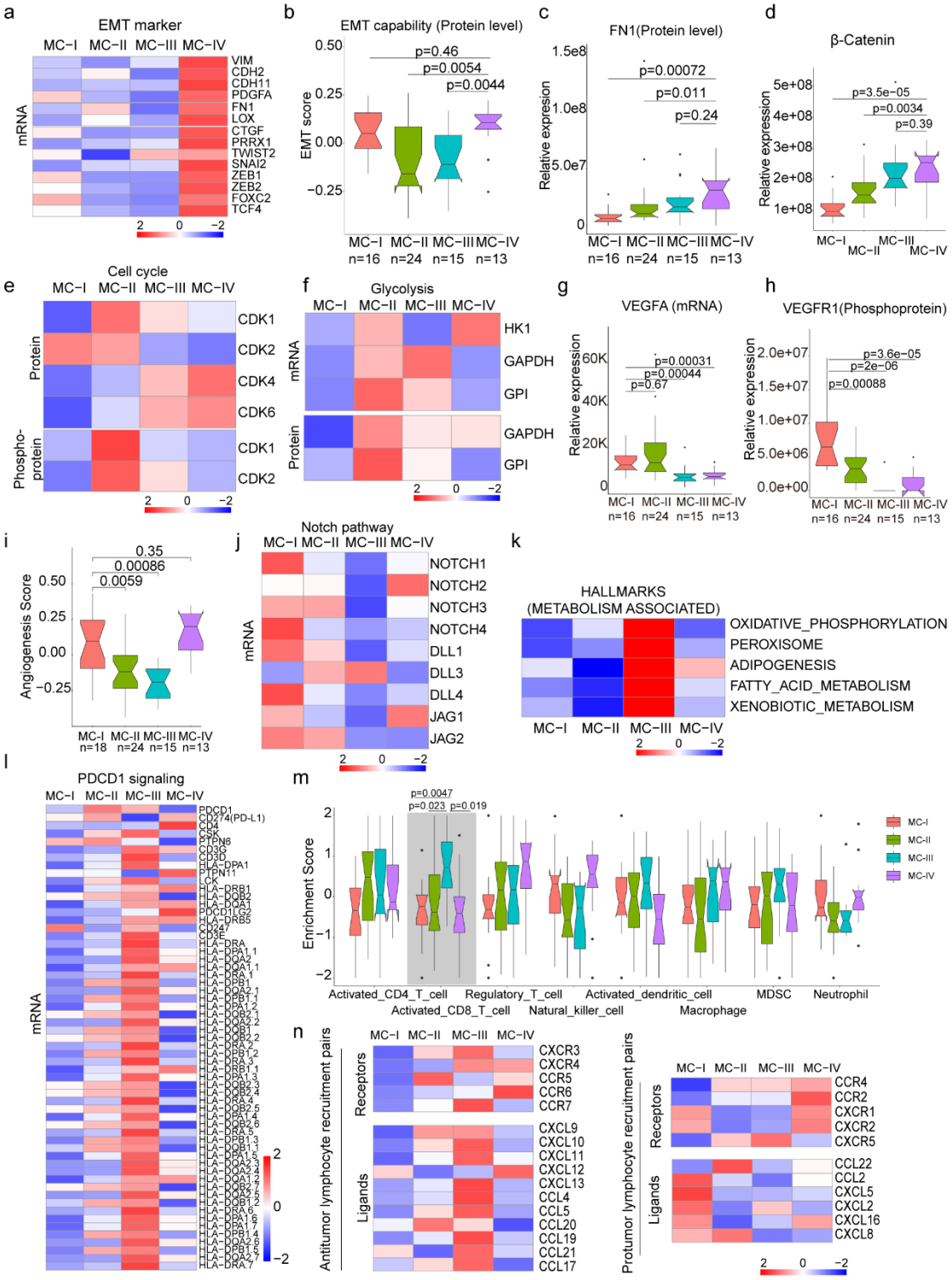


Fig5.| Biological and immune features across MC subtypes

a. Relative expression of epithelial-mesenchymal transition (EMT) markers across subtypes; b. EMT scores across subtypes using a gene set derived from MsigDB (M5930); c-d. Protein abundance comparison of FN1 (c) and β-Catenin (d) across subtypes; e. Protein and phosphoprotein

levels of key cell cycle kinases across subtypes; f. Expression of mRNA and protein levels of glycolysis-associated enzymes; g. mRNA expression of VEGFA across subtypes; h. Phosphoprotein abundance of VEGFR1 across subtypes; i. Angiogenesis score across subtypes using a gene set derived from MsigDB (Systematic name M5944); j. Expression comparison of key regulators of the Notch pathway across subtypes; k. Metabolism-associated hallmarks across subtypes. Gene sets for oxidative phosphorylation, peroxisome, adipogenesis, fatty acid metabolism, and xenobiotic metabolism were derived from MsigDB hallmark gene sets; l. Expression of PD-1 signaling-associated genes across subtypes. PD-1 signaling-associated genes were derived from MsigDB (Systematic name M18810) ; m. Immune cell infiltration across subtypes. Gene sets for each immune cell type were derived from a previous study[77]; n. Expression of anti-tumor/pro-tumor lymphocyte receptors and ligands across subtypes. The two-tailed Wilcoxon rank sum test was used to calculate p-values in panels b, c, d, g, h, i, and m.

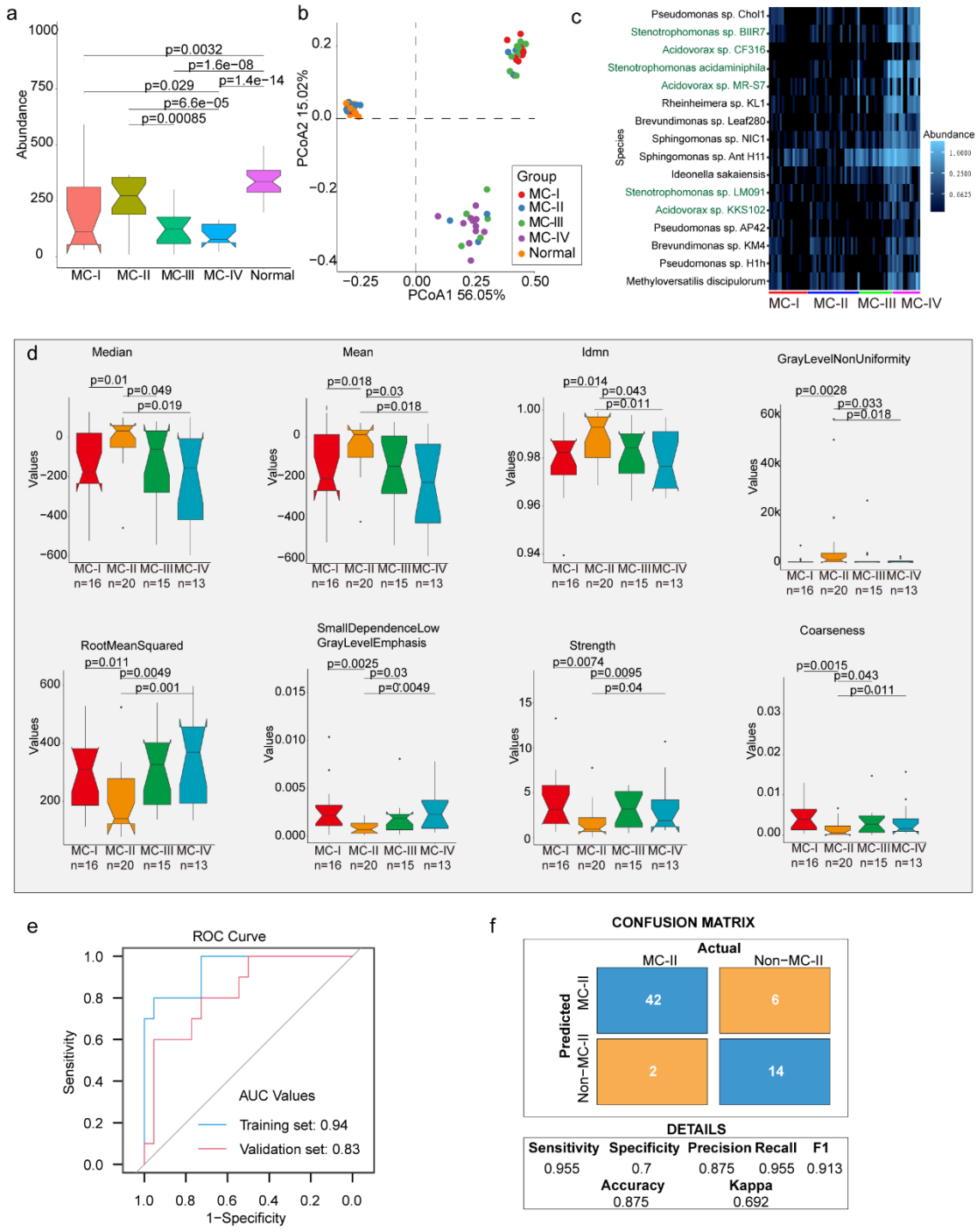


Fig 6.| Microbiota composition and radiomic features across subtypes.

a. Abundance of bacterial sequencing reads per million (RPM) across tumor molecular subtypes and adjacent normal tissues. Significance was determined using a paired two-sided Wilcoxon rank sum test; b. Principal coordinates analysis (PCoA) plot of mRNA data for XWLC samples shows significant variation between tumor and adjacent normal samples

along the first axis of variation, and variation between MC-IV subtypes and the other three tumor subtypes along the second axis; c. Heatmap of bacterial species' abundance highly expressed in the MC-IV subtype compared to the other three subtypes in the XWLC cohort. Bacterial species labeled green prefer an acidic living environment; d. Eight features showing significant differences between MC-II and the other three subtypes. The Wilcoxon rank sum test was used to calculate the p-values; e. A receiver operating characteristic (ROC) curve was used to evaluate the performance of the radiomic signature in distinguishing MC-II from the other three subtypes; f. Confusion matrix allows visualization of the performance of the algorithm in separating MC-II from other subtypes.

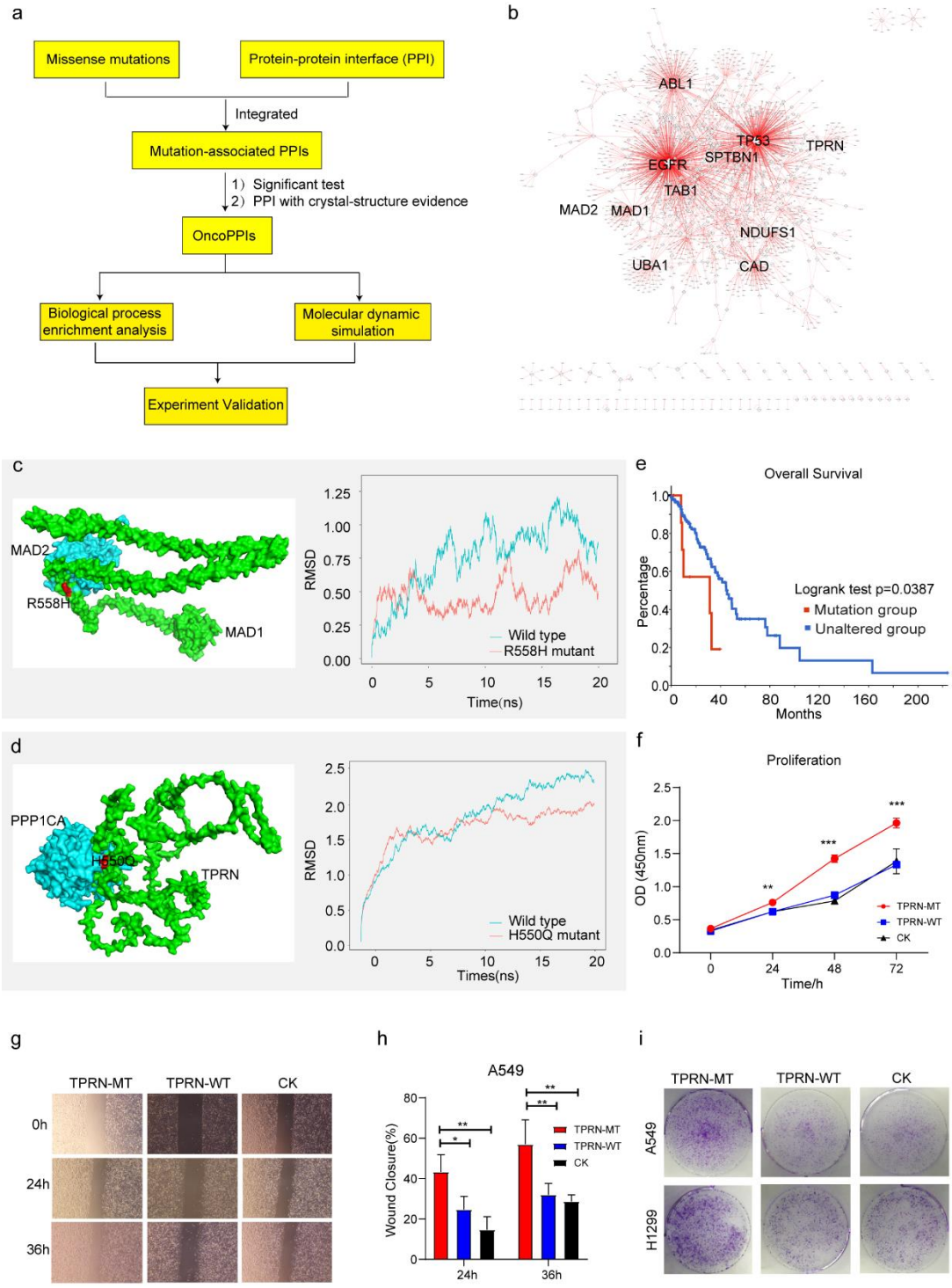


Fig.7|Identification of novel targets in XWLC

a. Flow chart showing the integration of mutation-informed PPI analysis, molecular dynamic simulation and experiment validation to identify novel targets; b. Network visualization of XWLC_oncoPPIs. Edge thickness represents the number of missense mutations at the protein-protein interaction (PPI) interface, while node size indicates connectivity;

c. MAD1-MAD2 interaction model and the p.Arg558His mutation at the interface (left). The complex model was generated using Zdock protein docking simulation. The right distribution showing root-mean-squared deviation (RMSD) during a 20 ns molecular dynamics simulation of MAD1 wild type vs. MAD1 p.Arg558His in the complex; d. Model showing the p.His550Gln alteration within the TPRN-PPP1CA complex (left). The right distribution showing root-mean-squared deviation (RMSD) during a 20 ns molecular dynamics simulation for TPRN wild type vs. TPRN p.His550Gln (H550Q) in the complex; e. Survival analysis of TPRN mutation group and unaltered group derived from cbioportal using TCGA-LUAD cohort (https://www.cbioportal.org/); f. CCK8 assay for TPRN-MT, TPRN-WT, and CK cell lines in A549 cells which was transfected by mutant TPRN, wild-type and empty vector, respectively. g. Transwell assay for TPRN-MT, TPRN-WT, and CK after 24h and 36h in A549 cells. Magnification was set to 40x; h.Bar chart showing the statistical results of transwell assay; i. Cell colon assay for TPRN-MT, TPRN-WT and CK in A549 and H1299 cell line. The two-tailed Wilcoxon rank sum test was used to calculate p-values in f and h. *.$p<0.05$; **,$p<0.01$; ***, $p<0.001$;