SNP rs4142441 and MYC co-modulated long non-coding RNA OSER1-AS1 suppresses non-small cell lung cancer by sequestering RNA-Binding Protein ELAVL1

Supplementary Methods

Weijia Xie[1*], Youhao Wang[1*], Yao Zhang[1*], Ying Xiang[1], Na Wu[1], Long Wu[1], Chengying Li[1], Tongjian Cai[1], Xiangyu Ma[1], Zubin Yu[2#], Li Bai[3#], Yafei Li[1#]

Selection of SNPs and genotyping (detailed descriptions)

As described in **Supplementary Figure 1**, we selected 16 SNPs located near potential lung cancer-related lncRNAs by following this procedure: first, we downloaded 85,273,413 SNPs from 1000 Genome Project (phase 3 v5b). We only included bi-allelic SNPs with minor allele frequency (MAF) greater than 1% in East Asian population; then we screened this list of SNPs against lncRNAs that were either known to be associated with lung cancer risks, or showing differentiated expression or associated with overall survival (OS) in The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) or The Cancer Genome Atlas Lung Squamous Cell Carcinoma (TCGA-LUSC) patients in public database, or showing tissue-specific expression in lung (UCSC). After this step, we narrowed down to a list of 6,214 SNPs located ±1000bp near 620 lung cancer-related lncRNAs. Among these SNPs, we further selected 16 SNPs with potential function, which included at least two of the following functional annotations: (1) known GWAS loci or eQTL loci; (2) promoter/enhancer/DNaseI hypersensitivity region in A549 cell line; (3) highly conserved in mammals; (4) transcription factor binding sites (**Supplementary Table S1)**. SNP were genotyped by SNPscan™ technology (Genesky Inc. Shanghai, China). The criteria of the quality control included: (1) Genotyping rate >0.95; (2) Call rate >98%; (3) Hardy Weinberg equilibrium test (P>0.01). The qualified SNPs were used for case-control analysis.

Case-control analysis (detailed descriptions)

We collected 778 cases with NSCLC from Xinqiao Hospital of the Third Military Medical University in Chongqing, China. We excluded patients with non-specified NSCLC subtypes (103), or missing information on age (2), sex (1), ethnicity (17) or repeated enrollment (10). Then, we matched up 645 patients on age and sex with 748 healthy controls from the annual physical examination group in the same hospital. We excluded controls with no DNA sample, missing information on sex and age, or with pre-existing lung diseases. In the end, 645 cases and 748 controls were genotyped by SNPscanTM technology (Genesky Inc. Shanghai, China). Nineteen cases and 12 controls failed genotyping.

Bioinformatics analysis for TCGA-LUAD data (detailed descriptions)

TCGA data (http://cancergenome.nih.gov/) was used to assess the clinical significance of OSER1-AS1 in NSCLC. TCGA-LUAD clinical data was downloaded from the GDC portal (https://gdc-portal.nci.nih.gov/). The OSER1-AS1 expression data for TCGA-LUAD patients were downloaded from the TANRIC website (http://ibl.mdanderson.org/tanric/_design/basic/index.html). In total, 456 out of the 546 TCGA-LUAD patients have expression data available on OSER1-AS1. TCGA-LUAD patients were removed if they have missing values in age (19), sex (0), tumor stage (9), smoking status (0), follow-up days (22) or if they have started treatment (chemotherapy/radiotherapy/molecular

targeted therapy) before the procurement date of biopsy samples (4). The expression levels of OSER1-AS1 were normalized by natural log transformation. Individuals were divided into two groups by high or low expression levels of OSER1-AS1, using median OSER1-AS1 expression value as the cutoff. The survival curves were calculated using the Kaplan–Meier method and the differences were evaluated using the log-rank test. Univariate and multivariate Cox proportional hazards regression models were used to examine the independent factors of overall survival for LUAD patients.