Investigation of the mechanism for intratumor heterogeneity in non-small cell lung cancer

Myung Jin Song

Department of Medicine

The Graduate School, Yonsei University

Investigation of the mechanism for intratumor heterogeneity in non-small cell lung cancer

Directed by Professor Yoon Soo Chang

The Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science

Myung Jin Song

December 2022

This certifies that the Doctoral Dissertation of Myung Jin Song is approved.

Thesis Supervisor : Yoon Soo Chang

Thesis Committee Member#1 : Eun Young Kim

Thesis Committee Member#2 : Seung Joon Kim

Thesis Committee Member#3: Yoon Jin Cha

Thesis Committee Member#4: Sungsoon Fang

The Graduate School
Yonsei University

December 2022

ACKNOWLEDGEMENTS

This study was supported by the National Research Foundation of the Republic of Korea (Grant No. NRF-2020R1A2B5B01001883).

<TABLE OF CONTENTS>

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Investigation of the mechanism for intratumor heterogeneity in non-small cell lung cancer

Myung Jin Song

Department of Medicine

The Graduate School, Yonsei University

(Directed by Professor Yoon Soo Chang)

"Intratumor heterogeneity (ITH)" is defined as the uneven distribution of genetically diverse tumor subpopulations within a tumor. We investigated the clinical implications of ITH, inferred from the number of subclones, and determined the mechanism of subclonal expansion. Single nucleotide variation, clinical data, copy number variation, and RNA sequencing data from The Cancer Genome Atlas-Lung Squamous Cell Carcinoma (TCGA-LUSC) and Lung Adenocarcinoma (LUAD) cases were obtained from the Genomic Data Commons data portal. The clonal status was estimated from the variant allele frequency of the mutated genes using the SciClone package. Candidate biomarkers for clonal expansion were inferred by analyzing the differentially expressed genes between the high and low clone groups, and their impact was evaluated in NSCLC cell lines. Finally, single-cell RNA sequencing analysis was performed to evaluate the impact of ITH on myeloid cells in the tumor microenvironment. Data from 481 LUSCs and 493 LUADs in stages I–IV that had not received any treatment for lung cancer were collected from the TCGA database. The number of subclones was positively correlated with the number of somatic variants and the cancer stage. The number of subclones was significantly higher in males (vs. females) and

smokers (vs. never-smokers). MTA1, FDG1, and MSI1 were selected in the DEG analysis as candidate genes for inferring clonal expansion. In subsequent experiments, NSCLC cells transfected with each candidate gene were not viable. Single-cell RNA sequencing data showed a decrease in the proportion during tumorigenesis of smokers' M2-like Macrophage 1. In contrast, classical monocytes, nonclassical monocytes, and proinflammatory macrophages increased in smokers during tumorigenesis. The differences between smokers and never smokers in proportion shift of M2-like Macrophage 1, classical monocytes, nonclassical monocytes, and proinflammatory macrophage were statistically significant. In conclusion, the findings from this study indicated that ITH is positively associated with the tumor mutational burden and cancer stage. Male sex and smoking are associated with high clonality. ITH may induce inflammation in the tumor microenvironment, promoting the proliferation of proinflammatory myeloid cells.

Keywords: non-small cell lung cancer, intratumor heterogeneity, myeloid cell, inflammation

Investigation of the mechanism for intratumor heterogeneity in non-small cell lung cancer

Myung Jin Song

Department of Medicine

The Graduate School, Yonsei University

(Directed by Professor Yoon Soo Chang)

I. INTRODUCTION

Lung cancer is the second most commonly diagnosed cancer and the leading cause of cancer-related death. In 2020, there were an estimated 2.2 million cases of lung cancer and 1.8 million lung cancer-associated deaths worldwide.[1] Non-small cell lung cancer (NSCLC) is histologically divided into adenocarcinoma, squamous cell lung cancer, and large cell carcinoma. Lung adenocarcinoma (LUAD) is located more peripherally and is more common among never smokers. In contrast, lung squamous cell carcinoma (LUSC) is more commonly located in the central lung and frequently invades the proximal bronchus because its pathogenesis is strongly associated with airway lesions that arise with smoking.[2] Targetable activating mutations, such as the epidermal growth factor receptor and anaplastic lymphoma kinase fusion, which lead to remarkable changes in LUAD treatment, are typically absent in LUSC, and targeted agents used with adenocarcinoma are largely ineffective with LUSC.[3-6]

Through tumor genome profiling, detailed information on carcinogenesis, including tumor development, progression, therapeutic response, and drug resistance, has been

obtained via the development of next-generation sequencing. Multiple studies based on tumor sequencing suggest that "intratumoral heterogeneity (ITH)," which describes the uneven distribution of genetically diverse tumor subpopulations within a tumor, plays a crucial role in treatment failure and drug resistance.[7-10] However, genetic biomarkers that can be used to infer ITH remain largely unknown.

Using whole-genome sequencing data, the number of subclones can be inferred from variant allele frequency (VAF), the percentage of sequence reads that match a specific DNA variant, divided by the overall coverage at that locus.[11] VAF represents the percentage of tumor cells harboring a specific mutation, assuming a relatively pure tumor sample [12,13]. VAF clustering can help infer the number of subclones in the tumor and also estimate the heterogeneity.

In this study, we investigated the relationship between intratumor heterogeneity inferred by the number of subclones and clinical characteristics and evaluated its clinical implications. We also investigated highly expressed genes in high clonality cancers and evaluated whether their expression in NSCLC cell lines induced clonal expansion. Finally, we used single-cell RNA sequencing (scRNA-seq) to depict the impact of clonal expansion in the tumor microenvironment.

II. MATERIALS AND METHODS

1. Analysis of The Cancer Genome Atlas (TCGA) data

A. Data Acquisition

The following data were downloaded from 504 LUSC and 585 LUAD cases shared in the TCGA project (https://www.cancer.gov/tcga): 1) Mutation Annotation Format (MAF) files for single nucleotide variants (SNV) analyzed with VarScan 2 variant Aggregation and

Masking workflow; 2) Masked Copy Number Segment analyzed using Affymetrix SNP 6.0; 3) RNA sequencing analyzed using HTSeq; and 4) clinical information. The Tumor Sample Barcodes of these cases were confirmed and analyzed based on the data obtained from the primary solid tumor. Of the total LUSC and LUAD cases, the following cases were excluded:1) cases whose records did not include all four data mentioned above (16 cases from LUSC, 85 cases from LUAD); 2) cases with exceptionally high numbers of mutations (3 cases from LUSC, 1 case from LUAD); and 3) cases in which the SNVs were 0 (4 cases from LUSC, 6 cases from LUAD). Finally, 481 LUSC and 493 LUAD cases were included in the analysis. To improve the positive predictive value of low allele frequency, SNVs with a total read depth of less than 40 and SNV data that did not meet the detection limit, as suggested by Shin et al.[13]

B. Calculations of subclone numbers and differentially expressed gene (DEG) analysis

To estimate the number of subclones, we used the SciClone package (http://github.com/genome/sciclone), which helps estimate the number of subclones by clustering variants with similar allele frequencies.[14] Computational efficiency was achieved by clustering VAFs using a variational Bayesian mixture model.[15] To identify genes related to intratumor heterogeneity, which were represented by the number of subclones, DEG analysis was performed between the high and low-clone groups. Patients with five or more subclones were paired with those with one subclone. Potential confounding variables, such as age, sex, stage, and smoking status, were adjusted using the propensity matching method while pairing the two groups using the "MatchIt" package from R. Three independent bootstrapped propensity score matching was performed in each LUSC and LUAD cohort. The DEG between the high- and low-clone groups was analyzed using the "DEseq2" package in each propensity score-matched cohort. When the gene expression ratio in the

experimental groups to the control group was more than 2 or less than 1/2, and the P-adjusted value was less than 0.05, the gene was considered to have significant differential expression, and further analysis was performed. The intersection of the three DEG results was defined as the final DEG result for LUSC and LUAD. Genes commonly upregulated in the DEG results of LUSC and LUAD were identified, and genes thought to be involved in the clonal expansion were selected as candidates for genetic biomarkers. The ontology of differentially expressed genes was confirmed using ToppGene (https://toppgene.cchmc.org/).

C. Statistical analysis

The distribution of variables was examined using the Shapiro–Wilk test. Continuous variables of three or more groups were analyzed using the Kruskal–Wallis test. Categorical variables were analyzed using the chi-square distribution and Fisher's exact test. In all cases, p-values < 0.05 were considered statistically significant. Statistical analyses were performed using the R statistical software, version 4.1.0 (R Foundation for Statistical Computing, Vienna, Austria).

2. In vitro cell culture experiment

A. Cells, plasmids, transfection

NSCLC lines, A549 and H460, were obtained from the American Type Culture Collection (ATCC) (Manassas, VA, USA). The cells were maintained in RPMI-1640 medium containing 10% fetal bovine serum and cultured at 37 °C in a humidified atmosphere of 5% CO2.

Plasmids containing the candidate biomarkers and a paired control vector (pCMV6 empty plasmid) were purchased from OriGene Tech (Rockville, MD, USA). The three candidate

genes were metastasis-associated 1 (MTA1), FYVE, RhoGEF, PH domain containing 1 (FGD1), and Musashi RNA-binding protein 1 (MSI1). The details of how these genes were selected as candidate genes are described in the Results section. MTA1, FGD1, MSI1, and the control vector were transfected into A549 and H460 cell lines using Lipofectamine® 2000 (Invitrogen, Carlsbad, USA) according to the manufacturer's instructions. Western blotting was performed 48 hours after transfection. Plasmid-transfected NSCLC cells were screened for 3–4 weeks with 500 μg/ml geneticin after a 48 h transfection.

B. Immunohistochemistry (IHC)

The expression of MTA1, FGD1, and MSI1 in NSCLC and tissue samples was analyzed by IHC staining. Briefly, sections were deparaffinized, rehydrated, immersed in H_2O_2 methanol solution, and incubated overnight with primary antibodies against MTA1, FGD1, and MSI1. Incubation was performed in an antibody diluent at dilutions of 1:500, 1:2000, 1:400, and 1:100. The sections were incubated for 10 min with a biotinylated linker and processed using avidin/biotin IHC techniques. 3,3′-Diaminobenzidine (DAB) was used as a chromogen in conjunction with the Liquid DAB Substrate kit (Novacastra, UK).

C. Western blotting

Cells were harvested using 2×LSB lysis buffer containing protease and phosphatase inhibitors (Sigma-Aldrich, St. Louis, MO, USA) on ice. Proteins (20 mg) were subjected to sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), electrophoresed, and transferred onto a nitrocellulose membrane. After blocking with 5% non-fat milk in Tris-buffered saline, the membrane was washed and incubated with the indicated primary antibody (MTA-1, Invitrogen, PA5-79699, rabbit; FGD1 Invitrogen,

PA5-40416, rabbit; MSI-1, thermos, 14-9896-82, rat) and subsequently incubated with anti-mouse, anti-rabbit, or anti-rat IgG coupled with horseradish peroxidase. All experiments were repeated at least three times.

3. scRNA-seq analysis

A. scRNA-seq data collection from public databases.

The scRNA-seq data were downloaded from the National Center for Biotechnology Information/Gene Expression Omnibus (GEO) with accession number GSE134174,[16] the Genome Sequence Archive of the Beijing Institute of Genomics Data Center under accession number HRA000154,[17] and Sequence Read Archive (SRA) under the accession number PRJNA773987.[18] Tumor-adjacent normal paired lung tissues from three current smokers and three never smokers were obtained from GSE134174. Tumor tissues of one current smoker and one never-smoker were obtained from HRA000154. Tumor-adjacent normal paired lung tissue of three never smokers and tumor tissue of one never-smoker were retrieved from our previously published scRNA-seq dataset (PRJNA773987). A total of 21 samples (4 tumor lung tissues of current smokers, 3 normal lung tissues of current smokers, 8 tumor lung tissues of never smokers, and 6 normal lung tissues of never smokers) were collected; the details of the collected samples are shown in Table 1. The histological type of all the tumor samples was adenocarcinoma.

Table 1. Characteristics of the study samples.

Patient ID	Samples	Source	Tissue origins	Age	Sex	Smoking	Stage
Case_08	Case_08_NL	GSE134174	NL	NA	NA	Nev	IB

Case_19	Case_19_NL	GSE134174	NL	NA	NA	Cur	IA
Case_20	Case_20_NL	GSE134174	NL	NA	NA	Cur	IA
Case_28	Case_28_NL	GSE134174	NL	NA	NA	Cur	IIIA
Case_30	Case_30_NL	GSE134174	NL	NA	NA	Nev	IA
Case_34	Case_34_NL	GSE134174	NL	NA	NA	Nev	IA3
Case_08	Case_08_Tu	GSE134174	Tu	NA	NA	Nev	IB
Case_19	Case_19_Tu	GSE134174	Tu	NA	NA	Cur	IA
Case_20	Case_20_Tu	GSE134174	Tu	NA	NA	Cur	IA
Case_28	Case_28_Tu	GSE134174	Tu	NA	NA	Cur	IIIA
Case_30	Case_30_Tu	GSE134174	Tu	NA	NA	Nev	IA
Case_34	Case_34_Tu	GSE134174	Tu	NA	NA	Nev	IA3
Case_24	Case_24_Tu	HRA000154	Tu	67	M	Cur	IB
Case_27	Case_27_Tu	HRA000154	Tu	66	M	Nev	IA2
Case_03	Case_03_NL	PRJNA773987	NL	71	F	Nev	IA2
Case_04	Case_04_NL	PRJNA773987	NL	73	M	Nev	IA1
Case_05	Case_05_NL	PRJNA773987	NL	68	F	Nev	IA2
Case_03	Case_03_Tu	PRJNA773987	Tu	71	F	Nev	IA2
Case_04	Case_04_Tu	PRJNA773987	Tu	73	M	Nev	IA1
Case_05	Case_05_Tu	PRJNA773987	Tu	68	F	Nev	IA2
Case_06	Case_06_Tu	PRJNA773987	Tu	67	F	Nev	IA1

Cur, current; NA, not accessible; Nev, never; NL, normal; Tu, tumor

B. Data quality control

Raw FASTQ files were collected for all 21 samples. Gene expression matrices were generated per sample using CellRanger (v6.1.2), and the output-filtered gene expression matrices were converted to a Seurat object using the R package Seurat version 4.0.6.[19] Low-quality cells were removed if they fell within the following criteria: (i) > 10% unique

molecular identifiers derived from the mitochondrial gene, (ii) ribosomal percentage (< 5%), and (iii) gene count (< 200). The presence of a doublet was identified using the DoubletFinder R package.[20]

C. Integration

After normalizing the individual datasets, we used integration methods described in the tutorial on the Seurat website to assemble distinct scRNA-seq datasets into an unbatched dataset (https://satijalab.org/seurat/articles/integration_introduction.html). Briefly, FindVariableFeatures, FindIntegrationAnchor, and IntegrateData functions were adopted serially to create a "batch-corrected" expression matrix for all cells.

D. Clustering and annotation

Scaling and principal component analyses were performed using an integration slot. The ElbowPlot and JackStrawPlot functions were used to identify the true dimensionality of the dataset, as recommended by Seurat developers. The FindNeighbors and FindClusters functions were employed to identify clusters and perform nonlinear dimensional reduction with the RunUMAP function. The FindAllMarkers function in Seurat was used to identify markers for each identified cluster. Clusters were then classified and annotated based on the expression of canonical markers in particular cell types (Table 2).

Clusters expressing high levels of CD68 and LYZ were annotated as myeloid clusters. Only cells with expression levels of CD68 and LYZ exceeding 2 (log_2 Fold Change) were selected as the final myeloid lineage to minimize the possibility of including nonspecific cells.

Table 2. Canonical marker genes of major cell clusters

Major cell type	Canonical marker gene
Epithelial cell (EP)	EPCAM, KRT18, SLPI
Fibroblast (FB)	LUM DCN, COL1A2
Endothelial cell (EC)	CLDN5, EPAS1, VWF
T cell (TC)	CCL5, NKG7, GNLY
B cell (BC)	MZB1, IGHG1, JCHAIN
Myeloid cell (MY)	LYZ, CD68, AIF1
Mast cell (MA)	MS4A2, TPSAB1, TPSB2

E. Subclustering of myeloid lineage

Subclustering of myeloid cells was performed as described in our previous publication.[18] Briefly, by adjusting the dims and resolution parameters of Seurat's FindNeighbors and FindCluster functions, approximately 1.2 multiples of the number of clinically explainable subclusters was obtained. The subclusters were annotated by checking the enrichment of the identified genes described in Table 3 and referring to panglaoDB (https://panglaodb.se/search.html). The similarity between each cluster was evaluated using Seurat's BuildClusterTree function and Jaccard Index.

Table 3. Signature genes used for the nomenclature of each subcluster.

Annotated cell type	Cluster number	Number of cells	Marker gene
Clacciscal monocyte	5	1517	FCN1, S100A8, S100A12, IL1B
Nonclassical monocyte	6	1287	CDKN1C, ZNF703, FCGR3A

Mono-mac 1	2	3337	RNASE1, LYZ, CCL3, VCAN
Mono-mac 2	4	1849	LGMN, SELENOP, CCL13, CCL2, RNASE1
M2-like Macrophage 1	0	5109	RBP4, GCHFR, CES1, C1QB
M2-like Macrophage 2	1	4674	INHBA, FABP4, LPL
Proinflammatory mac	11	432	CXCL10, CXCL9
Proliferating mac	8	722	MKI67, TOP2A, NUSAP1
cDC1	10	438	CPVL, CLEC9A
cDC2	3	2408	CD1C, FCER1A, CD1A
Activated DC	12	163	FSCN1, LAMP3, CCR7
NK cell	9	521	NKG7, GNLY, CCL5
Nonspecific	7	975	

DC, dendritic cell; NK, natural killer; Mac, macrophage; Mono-mac, monocyte-derived macrophage

F. Trajectory analysis

Trajectories with pseudotime were computed using the slingshot R package (version 2.4.0).[21] The inputs were the Uniform Manifold Approximation and Projection (UMAP) coordinates and cluster annotations.

III. RESULTS

1. Analysis of The Cancer Genome Atlas (TCGA) data

A. Study cohort

Data from 481 LUSCs and 493 LUADs in stages I–IV that had not received any treatment for lung cancer were collected from the TCGA database. Demographic characteristics are summarized in Table 4.

Table 4. Demographic characteristics of the TCGA study cases.

		LUSC (n=481)	LUAD (n=493)
Age (years)		68.6 (62.3–73.9) (n=471)	66.0(59.0–72.5) (n=475)
Sex			
Male		356 (74.0%)	229 (46.5%)
Female		125 (26.0%)	264 (53.5%)
Smoking status			
Ever-smoker		453 (94.2%)	413 (83.8%)
Never-smoker		18 (3.7%)	67 (13.6%)
Unknown		10 (2.1%)	13 (2.6%)
Pack-years		54.0 (21.0–70.0) (n=407)	41.8 (21.0–50.0) (n=341)
Tumor stage			
I		232 (48.2%)	266 (54.0%)
II		157 (32.6%)	113 (22.9%)
III		81 (16.8%)	82 (16.6%)
IV		7 (1.4%)	25 (5.1%)
Unknown		4 (0.8%)	7 (1.4%)
Number of subclones			
	1	143 (29.7%)	200 (40.6%)
	2	182 (37.8%)	162 (32.9%)
	3	107 (22.2%)	97 (19.7%)
	4	37 (7.7%)	29 (5.9%)
	5	12 (2.5%)	4 (0.8%)
	6	–	1 (0.2%)
EGFR		3 (0.6%)	40 (8.1%)
KRAS		7 (1.5%)	73 (14.8%)

EGFR, epidermal growth factor receptor; KRAS, Kirsten rat sarcoma viral oncogene

Values are expressed as median (interquartile range) or number (%).

B. The number of subclones and their association with genomic mutations

Whole exome sequencing of 481 LUSCs identified a total of 117,869 variants with a median of 209.0 (128.0–319.0) variants per tumor, and those of 493 LUADs identified a total of 84,796 variants with a median of 1172.0 (59.0–354.0) (Fig. 1A, Fig. 1B). The total variation in a tumor showed a significant positive correlation with the number of subclones in a tumor in both the LUSC and LUAD cohorts ($\sigma = 0.388$, P-value < 0.001, Fig. 1C; $\sigma = 0.352$, P-value < 0.001, Fig. 1D). When these variations were classified into SNVs and indels, they showed a significant positive correlation with the number of subclones (Table 5).

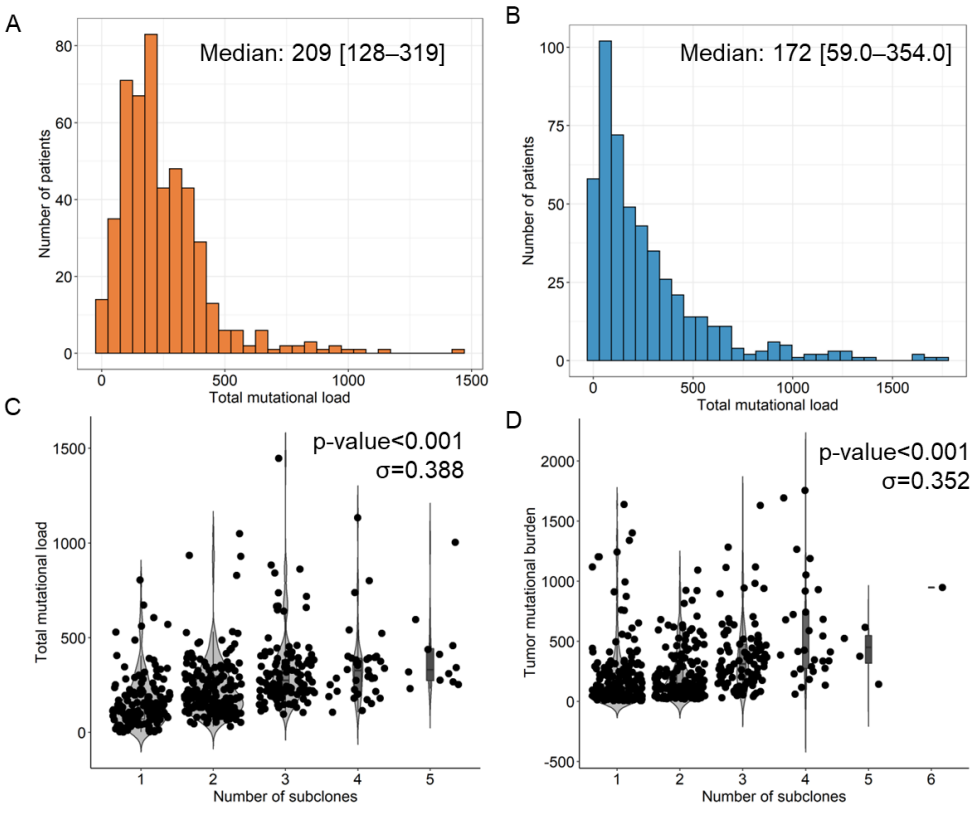


Figure 1. Relationship between mutations and subclones that constituted a tumor. (A) Distribution of total mutations in TCGA-LUSC. (B) Distribution of total mutations in TCGA-LUAD. (C) A violin plot showing the total number of mutations according to the number of subclones in TCGA-LUSC. The total number of mutations and the number of subclones were positively correlated (Pearson's correlation efficiency, $\sigma = 0.376$, P-value < 0.001). (D) A violin plot showing the total number of mutations according to the number of subclones in TCGA-LUAD. The total number of mutations and the number of subclones were positively correlated (Pearson's correlation efficiency, $\sigma = 0.352$, P-value < 0.001).

Table 5. Relationship between the number of subclones and variants.

LUSC	Number of subclones					Correlation coefficiency (σ)	P-value*
	1 (n=143)	2 (n=182)	3 (n=107)	4 (n= 37)	5 (n=12)		
Subtypes of mutations							
SNV	126 (68.5–205.5)	183 (123.5–287.3)	266 (200.5–373.0)	314 (210.0–378.0)	318.5 (261.8–436.0)	0.389	<0.001
Indel	1.0 (0.0–1.5)	1.0 (0.0–2.0)	1.0 (0.0–3.0)	2.0 (1.0–3.0)	3.0 (1.0–4.0)	0.241	<0.001
LUAD	1 (n=200)	2 (n=162)	3 (n=97)	4 (n=29)	5 and 6 (n=5)		
SNV	85.5 (31.0–225.5)	142 (67.3–280.8)	303 (176–460)	402 (261–713)	496 (359–594)	0.351	<0.001
Indel	4 (2–8)	5 (3–10)	10 (6–16)	14 (9–29)	21 (17–28)	0.266	<0.001

*P-value obtained using the Kruskal–Wallis test.

SNV, single nucleotide variants

C. The number of subclones and their association with clinical characteristics

(1) Tumor stage

The staging system is still the most useful parameter for predicting the clinical outcome in patients with lung cancer. We compared the number of subclones with the lung cancer stage and discovered that there was a significant positive correlation between the number of subclones and stage in both LUSC and LUAD (σ = 0.099, P-value = 0.030, Fig. 2A; σ =

0.091, P-value = 0.044, Fig. 2B). The fact that the number of subclones comprising a primary tumor is positively related to the number of variants and the stage of the corresponding tumor suggests that the number of subclones reflects the biological aspect of a tumor.

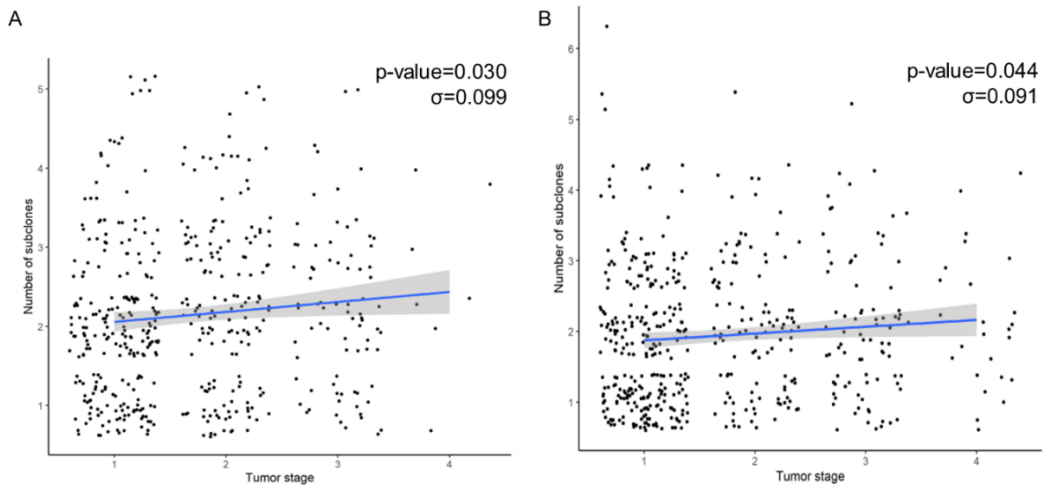


Figure 2. Scatter plot showing the relationship between the tumor stage and the number of subclones. (A) The tumor stage and the number of subclones were positively correlated in LUSC (Pearson's correlation efficiency, σ = 0.099, P-value = 0.030). (B) The tumor stage and the number of subclones were positively correlated in LUAD (Pearson's correlation efficiency, σ = 0.091, P-value = 0.044).

(2) Smoking

Since smoking is one of the leading causes of lung cancer induced by C>A transversions of DNA, we divided the patients into never smokers and ever-smokers and compared the number of subclones between the two groups. Tumors of ever-smokers comprised a

significantly larger number of subclones than never smokers in both LUSC and LUAD (Fig. 3A, Fig. 3B, and Table 6).

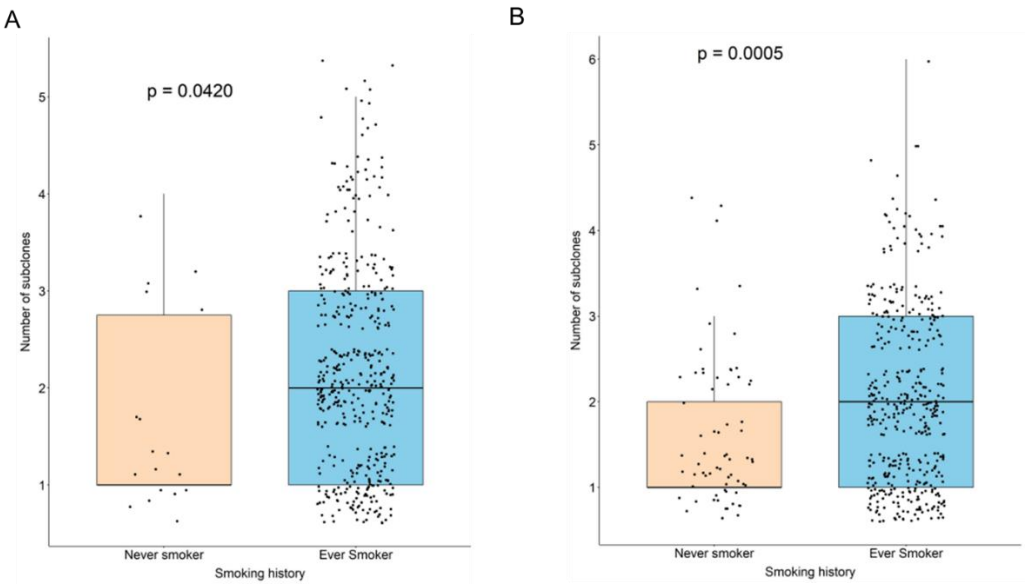


Figure 3. The relationship between smoking and the number of subclones.
(A) A box plot showing the difference in the number of subclones according to smoking history in LUSC. (B) A box plot showing the difference in the number of subclones according to smoking history in LUAD.

Table 6. The relationship between smoking and the number of subclones.

Number of subclones	LUSC			LUAD		
	Never-smoker (n=18)	Ever-smoker (n=453)	P-value*	Never-smoker (n=67)	Ever-smoker (n=413)	P-value*

1	11 (61.1%)	128 (28.3%)	0.037	39 (58.2%)	156 (37.8%)	<0.001
2	2 (11.1%)	177 (39.1%)		20 (29.9%)	136 (32.9%)	
3	4 (22.2%)	100 (22.1%)		5 (7.5%)	90 (21.8%)	
4	1 (5.6%)	36 (7.9%)		3 (4.5%)	26 (6.3%)	
5	0 (0.0%)	12 (2.6%)		0 (0.0%)	4 (1.0%)	
6	–	–		0 (0.0%)	1 (0.2%)	

Values are expressed as numbers (%)

*P-value was obtained using the Kruskal–Wallis test.

(3) Age

As age-related mutations were observed in most malignancies, including lung cancer, and there was a significant positive correlation between lung cancer incidence and age [22], the relationship between age at diagnosis of lung cancer and the number of subclones was investigated. Age at diagnosis and the number of subclones did not show a significant correlation with either LUSC or LUAD (Fig. 4).

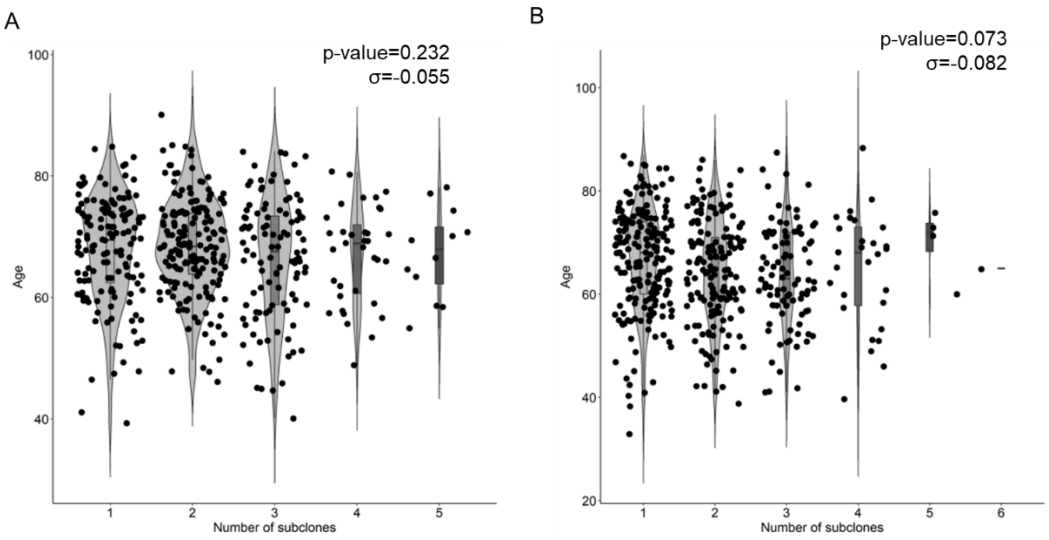


Figure 4. The relationship between age and the number of subclones.

(A) A violin plot showing age according to the number of subclones in LUSC. (B) A violin plot showing age according to the number of subclones in LUAD.

(4) Sex

Finally, the influence of sex on the increase in the number of subclones was investigated. The number of subclones constituting a tumor was significantly higher in male patients than in female patients with both LUSC and LUAD (P = 0.001, Fig. 5A; P = 0.005, Fig. 5B).

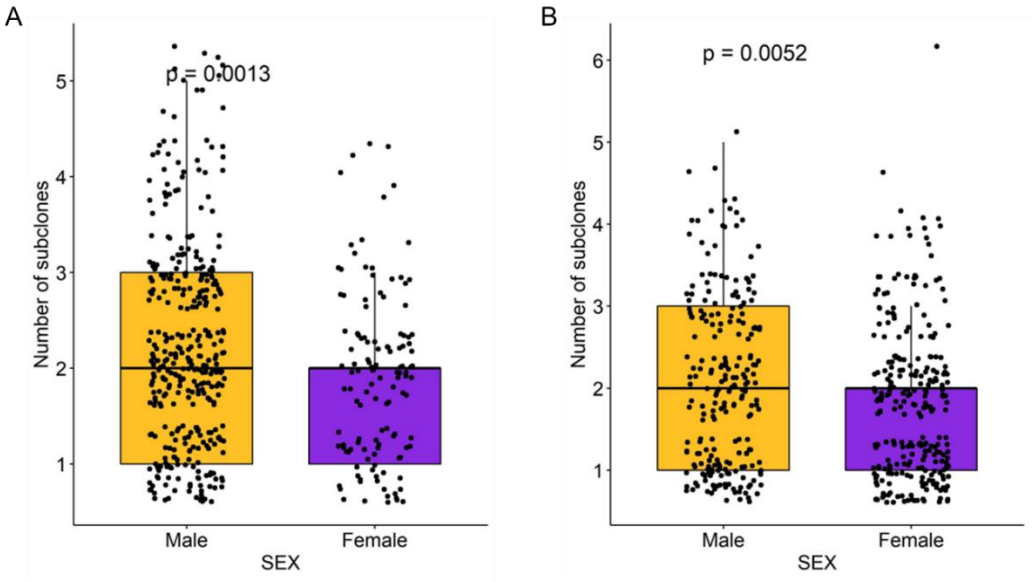


Figure 5. The relationship between sex and the number of subclones.

(A) A box plot illustrating the difference in the number of subclones according to sex in LUSC. (B) A box plot illustrating the difference in the number of subclones according to Sex in LUSC

D. Inferring clonal expansion via DEG

An additional analysis was performed using RNA sequencing data to uncover the possible etiology by increasing the number of subclones that constituted a tumor. Using propensity score matching of the “MatchIt” R package, the cases in which tumors were composed of five or more subclones were matched with those in which the tumor was composed of a single subclone by age, sex, pack-years, and stage. After performing three independent matching and DEG analyses, we obtained a set of intersections from the results of each analysis and performed gene ontology analysis using ToppGene. DEG analysis revealed 707 genes upregulated in tumors composed of five subclones compared to single subclone

tumors in LUSC and 851 genes upregulated in tumors composed of five and six subclones compared to single subclone tumors in LUAD. Among these genes, 49 were upregulated in LUSC and LUAD. Three genes commonly having a low p-value and a high log 2-fold change value in the DEG analysis of LUSC and LUAD were selected as candidate genes involved in clonal expansion: MTA1, which is involved in the cellular component organization or biogenesis; FGD1, which is involved in cell death; and MSI1, which is involved in developmental processes.

Table 7. Gene ontology of high-ranked commonly upregulated genes in LUSC and LUAD

Gene ontology of biological process	Gene
Metabolic process	FOLH1, NAT8L
Developmental process	FBN2, HOXB9, MSI1, HOXD8
Regulation of secretion by cell	RAB3B, STXBP5L, NOS1
Regulation of cell signal transduction	FRRS1L, GNG4, OTX2
Cell component organization or biogenesis	MTA1
Cell death	FGD1

E. Clinical implications of three candidate genes

Survival correlations with these genes were evaluated using oncoLnc (http://www.oncolnc.org). The upper 25% and lower 25% of patients were extracted from TCGA-LUSC and LUAD cohorts, respectively.

In LUSC, the upper and lower 25% of patients were matched by age, sex, pack-years, and stage. Patients with high expression of MTA1 showed significantly better survival probability (log-rank p-value = 0.042). In contrast, FGD1 and MSI1 did not show

differences between the high and low gene expression groups (Fig. 6). However, although not statistically significant, the high gene expression group tended to have a better survival probability in both FGD1 and MSI1.

In LUAD, the upper and lower 25% of patients were matched by age, sex, pack-years, stage, and major driver mutations (EGFR and KRAS). There was no significant difference between the high and low expression groups for all three candidate genes, i.e., MTA1, FGD1, and MSI1 (Fig. 7).

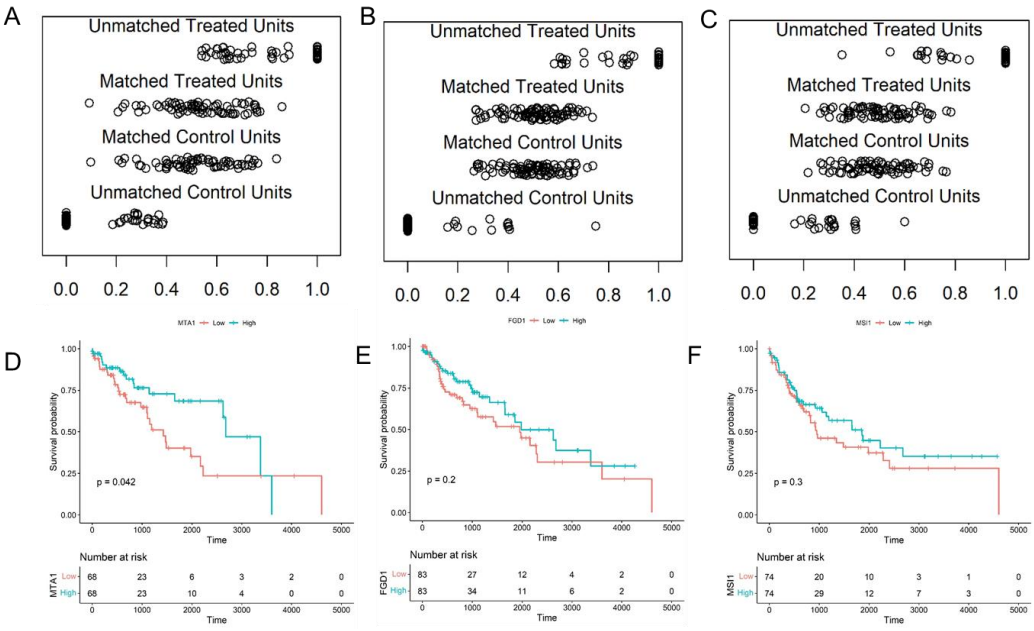


Figure 6. Survival probability of TCGA-LUSC patients, stratified by MTA1, FGD10. and MSI1 expression level.

(A–C) Distribution of propensity scores of the high (> 25%) and low (< 25%) gene expression groups (MTA1, FGD1, and MSI1).

(D–F) Kaplan–Meier survival plot stratified by gene expression levels.

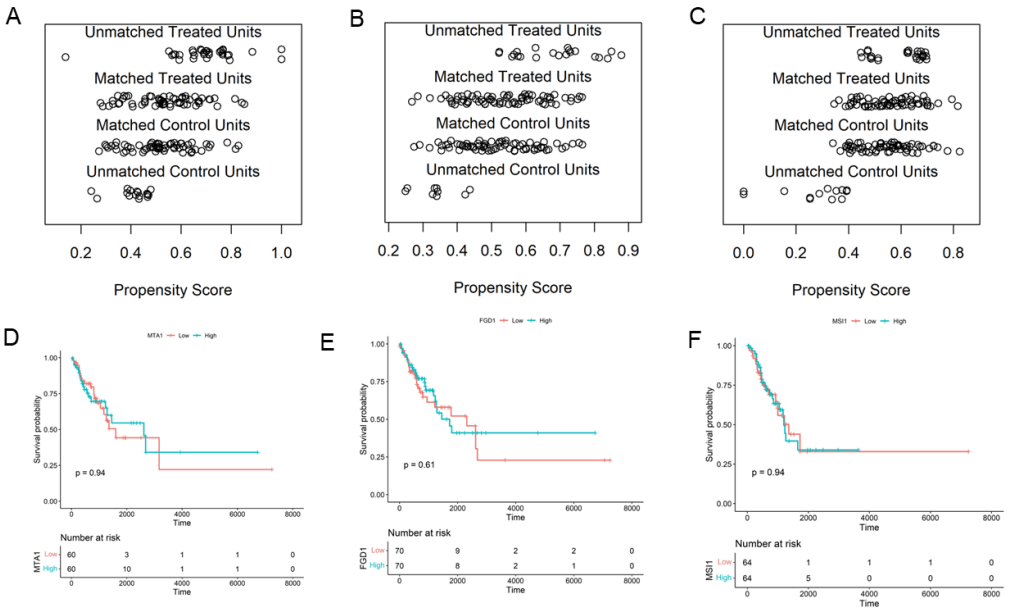


Figure 7. Survival probability of TCGA-LUAD patients, stratified by MTA1, FGD1, and MSI1 expression level.

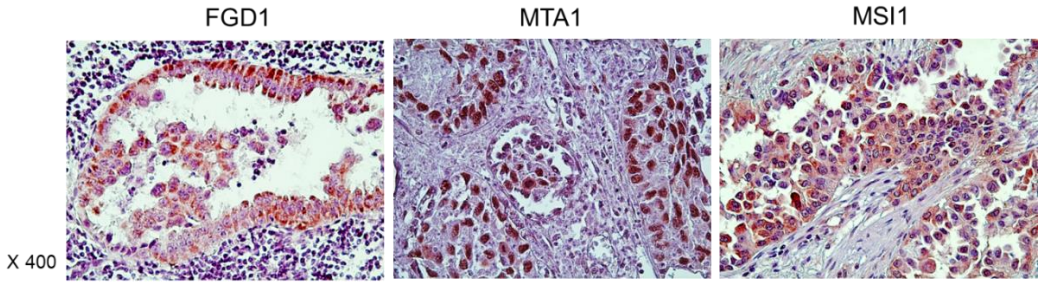
(A–C) Distribution of propensity scores of the high (>25%) and low (<25%) gene expression groups (MTA1, FGD1, and MSI1).

(D–F) Kaplan–Meier survival plot stratified by gene expression levels.

2. In vitro cell culture experiment

A. Immunohistochemistry (IHC)

To confirm whether the three genes selected as candidate biomarkers involved in the clonal expansion were expressed in NSCLC tissues, IHC analysis was performed on formalin-fixed paraffin-embedded NSCLC tissues. FGD1 and MSI1 were well expressed in the cytoplasm of tumor cells, and MTA1 was well expressed in the nucleus (Fig. 8).



Age	64	76	56
Sex	Female	Male	Female
Smoking	Never	Former	Current
Pathology	Adenocarcinoma	Squamous	Adenocarcinoma
Stage	III	III	IV

Figure 8. Representative photographs of IHC staining in NSCLC tissues.

B. Culture of gene-transfected NSCLC cell

MTA1, FGD1, and MSI1 genes were transfected into NSCLC cell lines (H460 and A549), and the expression of specific proteins was confirmed by western blotting (Fig. 9). However, the cells were not viable in three repetitive culture experiments.

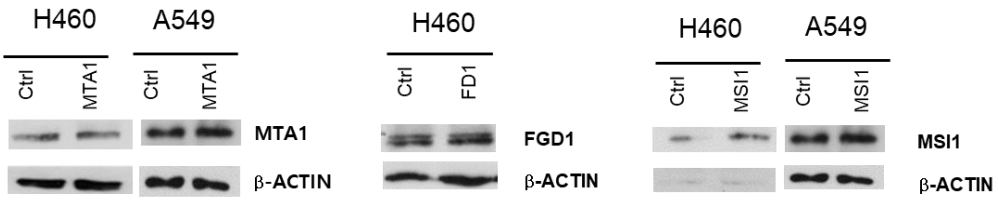


Figure 9. Western blot analysis showing the expression of three candidate genes in H460 and A549 cells.

The results of cell culture experiments showed that cells do not survive after the transfection of candidate genes. Survival analysis of TCGA-LUSC data showed better survival trends in patients with higher expression levels of candidate genes. We speculated that cancers with high clonality might induce more inflammation to promote cancer immune surveillance. In this process, we hypothesized that innate immunity, which serves as the first-line host defense against neoantigens, plays an important role. Therefore, we analyzed the scRNA-seq data to depict the myeloid lineage and determine the influence of smoking exposure, which was found to be associated with clonal expansion through TCGA data analysis.

3. scRNA-seq analysis

A. Single-cell transcriptomic profiling of LUAD in relation to smoking

A total of 21 samples were obtained from the 12 patients (Fig. 10A). After quality control, 109,962 cells were initially divided into seven major cell groups: epithelial, endothelial, myeloid, B cell, T cell, fibroblast, and mast cells (Fig. 10B, Fig. 10C).

To refine cancer cells and determine the proportion of myeloid cells in the tumor microenvironment, the proportion of myeloid cells was evaluated in all six major cell types, except for clusters with epithelial features (high expression of EPCAM and KRT18). T cells account for the largest portion of the TME, followed by myeloid cells. Myeloid cells were more frequently detected in current smokers than in never smokers in both normal and tumor lungs (Fig. 10D–G). In normal lungs, the proportion of myeloid cells was 29% higher in current smokers than in never smokers. In contrast, in the tumor lung, the proportion of myeloid cells was 14% higher in current smokers than in never smokers. These results suggest that the increase in the proportion of myeloid cells caused by smoking was more

pronounced in the normal lung than in the tumor-affected lung.

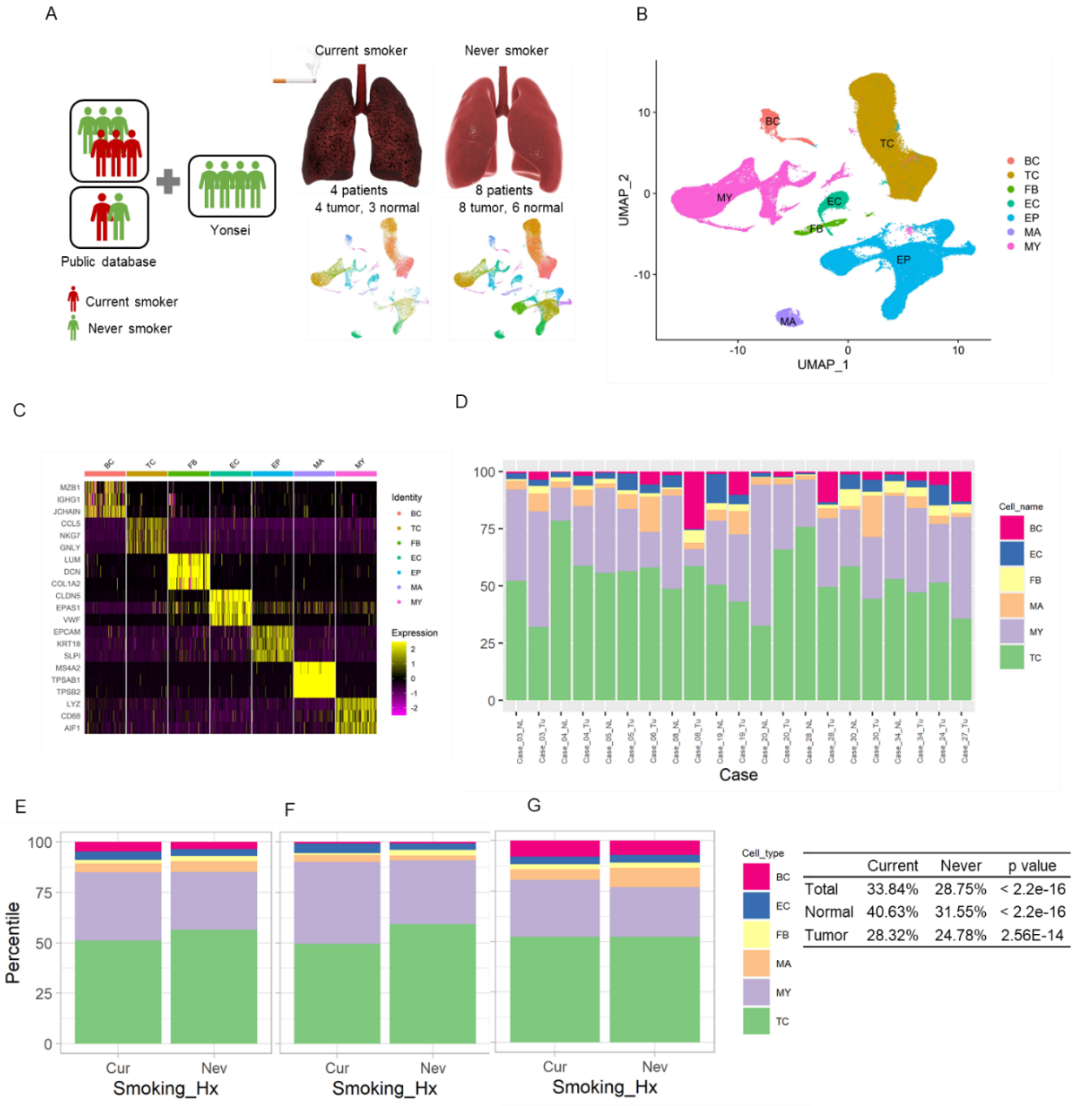


Figure 10. Overview of the dataset, single-cell RNA sequencing analysis, and clustering of major cell types. (A) Overview of dataset construction. (B) UMAP (Uniform Manifold Approximation and Projection) embedding of 109,962 cells clustered by major lung cell lineages (C) Heatmap showing representative marker genes of the major cell lineages (D)

Horizontal bar plots showing the number of cells belonging to each cluster by tissue origin (E-G) Stacked column chart and table showing the proportion of cells belonging to the major cell cluster by smoking status (from left to right total, normal, tumor)

B. Cellular diversity in the myeloid lineage

Overall, 23,432 myeloid cells were clustered into 12 subclusters (Fig. 11). Classic monocytes express high levels of alarmins (S100A8 and S100A12) and proinflammatory cytokines (IL1B), which play a significant role in inflammation. Nonclassical monocytes express high levels of cyclin-dependent kinase inhibitor 1C (CDKN1C), a potent inhibitor of G1cyclin/cdk complexes, which negatively regulate cell proliferation. CDKN1C acts as a tumor suppressor gene.[23] The adjacent ‘‘monocyte-derived macrophage (mono-mac)’’ population showed a less distinct phenotype, which is indicative of a transitory differentiation state.

We identified four different types of macrophages. M2-like Macrophage-1 expresses high levels of the scavenger receptor macrophage receptor with collagenous structure (MARCO), which regulates macrophage polarization toward the immunosuppressive M2 phenotype.[24,25] M2-like Macrophage-2 expressed high levels of ‘inhibin beta A’’ (INHBA), a member of the TGF-β family that also contributes to the M2 phenotype. Pro-inflammatory macrophages express pro-inflammatory cytokines (CXCL10). Finally, “proliferating macrophages” express cell-cycle-related genes (MKI67, TOP2A, and NUSAP1).

We identified three types of dendritic cells (DCs). The conventional DC (cDC) 1 highly expressed CLEC9A. In contrast, cDC2 highly expressed CD1 and CLEC10A. CLEC9A and CLEC10A belong to the family of C-type lectin-like receptors (CTLR), responsible for recognizing sugar structures in bacteria and cancer cells. After capturing antigens, DCs are

activated, express higher levels of chemokine receptors, such as CCR7, and secrete cytokines, which are essential for T cell activation.[26] DCs expressing high levels of CCR7 are referred to as activated DCs.[27]

We also found NK cells that highly expressed cytotoxic genes (NKG7, GNLY, GZMA, GZMB, and GZMH) and chemokine gene CCL5, which recruit T cells and other immune cells by binding to CCR5.[28]

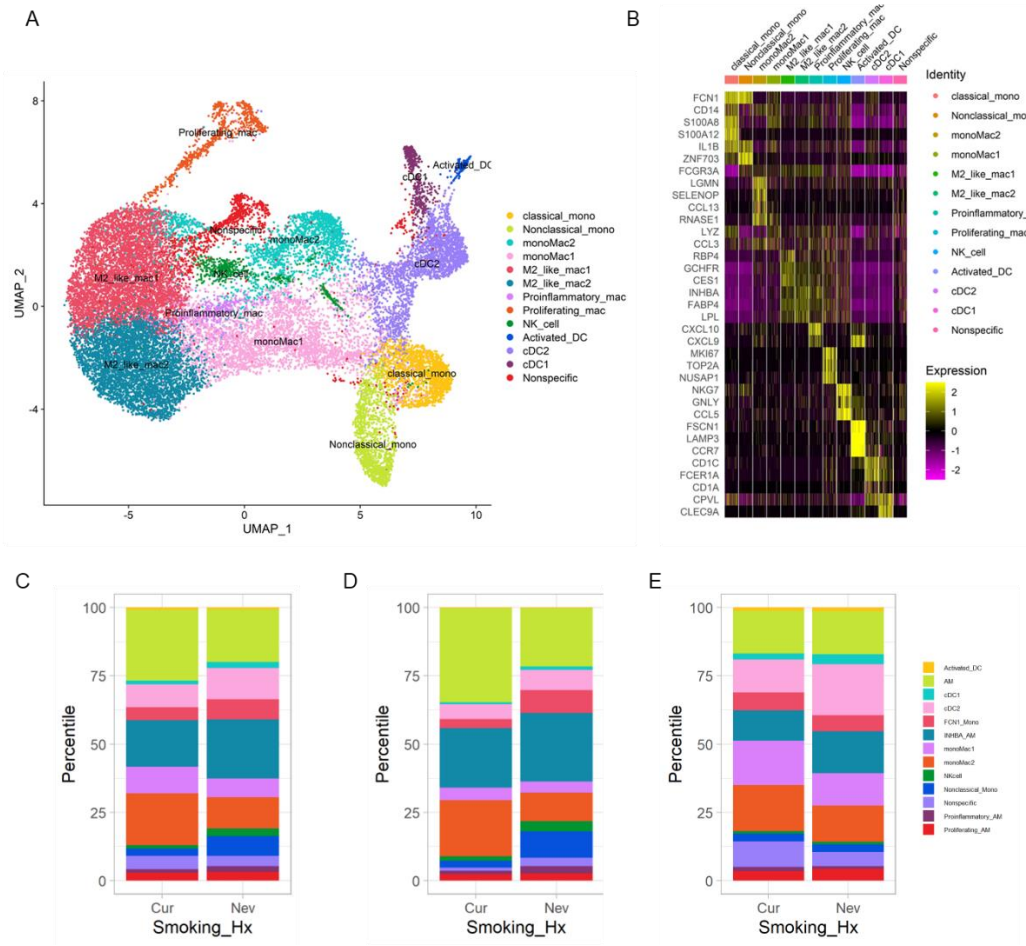


Figure 11. Subclustering of myeloid cell lineages. (A) UMAP visualization of myeloid cell lineages colored by cell type. (B) Heatmap showing representative marker genes of the

myeloid subclusters. (C–E) Stacked column chart showing the proportion of cells belonging to the major cell cluster by smoking status (from left to the right, total, normal, tumor)

C. Effect of smoking on the myeloid cells in tumorigenesis

To understand the effect of smoking on the myeloid cell population during tumorigenesis, we explored the myeloid cell proportion shift in tumors compared to normal tissues by smoking status. Among the 12 subclusters in myeloid cells, classical monocytes, non-classical monocytes, M2-like macrophages 1, and pro-inflammatory macrophages showed significant differences between current smokers and never smokers in the myeloid cell proportion shift during tumorigenesis (Fig. 12).

The level of M2-like Macrophage 1 decreased during tumorigenesis in smokers. In contrast, no significant change was observed in never smokers. Classical monocytes increased during tumorigenesis in smokers, while those in never smokers were unaffected by tumorigenesis. The proportion of non-classical monocytes and pro-inflammatory macrophages in smokers was unaffected by tumorigenesis, whereas that of never smokers decreased during tumorigenesis.

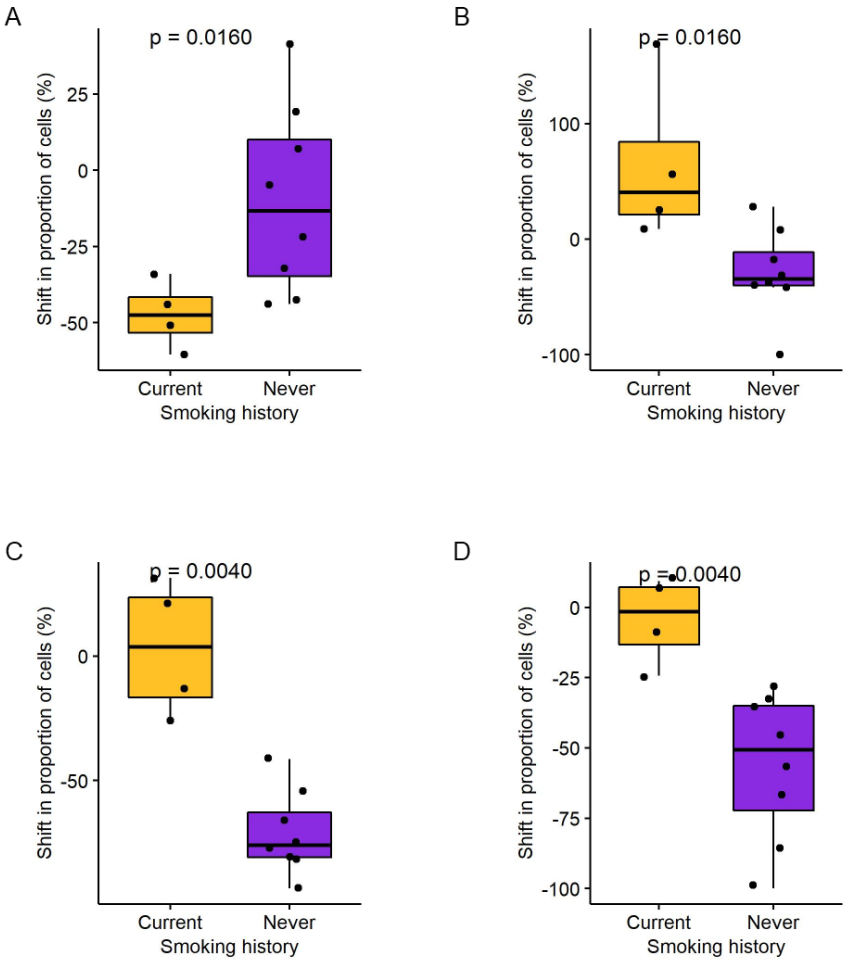


Figure 12. The relative proportion of myeloid cell population shift during tumorigenesis (A) M2-like Macrophage 1, (B) Classical monocyte, (C) Nonclassical monocyte, (D) Proinflammatory macrophage.

IV. DISCUSSION

In this study, we discovered that intratumor heterogeneity inferred from the number of subclones constituted a tumor positively correlated with the number of somatic variants

and cancer stage. The number of subclones was significantly higher in males (vs. females) and smokers (vs. never smokers), whereas age did not significantly correlate with the number of subclones. Through DEG analysis, we selected MTA1, FDG1, and MSI1 as candidate genes for inferring clonal expansion. We could not determine whether these genes induced ITH in NSCLC cell lines because of the inviability of the gene-transfected cell lines. With the speculation that high clonality of cancer could act as neoantigens that increase the inflammatory response, we further analyzed myeloid lineage in scRNA-seq data. We found that the proportion shift during tumorigenesis of monocytes and pro-inflammatory macrophages increased in smokers, a clinical characteristic associated with high clonality.

Inflammation is involved in cancer development and progression, as well as in anti-cancer treatment.[29,30] Inflammation has two opposing roles in cancer: promotion and inhibition. Chronic inflammation induces immunosuppression, creating a favorable microenvironment for tumorigenesis. Epidemiologic studies have reported that up to 25% of cancers are related to chronic inflammatory disease.[31] In contrast, acute inflammation contributes to cancer cell death.[30] The immune system can recognize and destroy tumor cells in cancer surveillance, and acute inflammation can promote this process.

The tumor cell heterogeneity could be associated with treatment failure and drug resistance. However, from the perspective of the TME, the heterogeneity of tumor cells acts as neoantigens, causing inflammation in the tumor microenvironment and promoting cancer immune surveillance, leading to better survival outcomes.[30,32]

Our results of TCGA-LUSC survival analysis suggest that a better survival outcome in high clonality cancer might be explained by the inflammatory response caused by innate immunity that serves as an initial defense against nascent tumor neoantigens. Further

scRNA-seq analysis revealed that pro-inflammatory myeloid subclusters significantly increased and anti-inflammatory myeloid subclusters significantly decreased during tumorigenesis in current smokers. Such crosstalk between inflammatory processes in the tumor microenvironment and cancer cells may explain the better survival outcome in high clonality cancers.

Moreover, upon analyzing the survival of the TCGA-LUAD cohort according to FCN1 expression level, which is highly expressed in classical monocytes, mortality was found to be significantly lower in patients with high expression of FCN1. Classical monocytes have a high expression of pro-inflammatory cytokine genes, such as IL-B, and are immature cells with high plasticity that can differentiate into DCs and macrophages.[23] Further research is needed to determine whether the enhancement of the pro-inflammatory response can be used as a treatment with antitumor effects.

V. CONCLUSION

Our study revealed that intratumor heterogeneity inferred from the number of subclones was positively correlated with the number of somatic variants and cancer stage. Male sex and smoking were found to be associated with high clonality. ITH may induce inflammation in the microenvironment. However, the association between ITH and better survival outcomes should be interpreted cautiously, and further studies are warranted to clarify the theoretical basis of ITH and inflammation.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-249.

2. Rosado-de-Christenson ML, Templeton PA, Moran CA. Bronchogenic carcinoma: radiologic-pathologic correlation. RadioGraphics. 1994;14(2):429-446.

3. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519-525.

4. Zhao W, Choi Y-L, Song J-Y, et al. ALK, ROS1 and RET rearrangements in lung squamous cell carcinoma are very rare. Lung cancer (Amsterdam, Netherlands). 2016;94:22-27.

5. Miyamae Y, Shimizu K, Hirato J, et al. Significance of epidermal growth factor receptor gene mutations in squamous cell lung carcinoma. Oncology Reports. 2011;25(4):921-928.

6. Rekhtman N, Paik PK, Arcila ME, et al. Clarifying the spectrum of driver oncogene mutations in biomarker-verified squamous carcinoma of lung: lack of EGFR/KRAS and presence of PIK3CA/AKT1 mutations. Clinical cancer research : an official journal of the American Association for Cancer Research. 2012;18(4):1167-1176.

7. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168(4):613-628.

8. Greaves M. Evolutionary determinants of cancer. Cancer discovery. 2015;5(8):806-820.

9. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nature reviews Cancer. 2012;12(5):323-334.

10. Garraway LA, Lander ES. Lessons from the cancer genome. Cell. 2013;153(1):17-37.

11. Strom SP. Current practices and guidelines for clinical next-generation sequencing oncology testing. Cancer biology & medicine. 2016;13(1):3-11.

12. Sallman DA, Padron E. Integrating mutation variant allele frequency into clinical practice in myeloid malignancies. Hematology/oncology and stem cell therapy. 2016;9(3):89-95.

13. Shin H-T, Choi Y-L, Yun JW, et al. Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nature communications. 2017;8(1):1377-1377.

14. Miller CA, White BS, Dees ND, et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. PLoS computational biology. 2014;10(8):e1003665.

15. CM B. Pattern recognition and machine learning. Information science and statistics. . 2006:738.

16. Kim N, Kim HK, Lee K, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11(1):2285.

17. Xing X, Yang F, Huang Q, et al. Decoding the multicellular ecosystem of lung adenocarcinoma manifested as pulmonary subsolid nodules by single-cell RNA sequencing. Sci Adv. 2021;7(5).

18. Kim EY, Cha YJ, Lee SH, et al. Early lung carcinogenesis and tumor microenvironment observed by single-cell transcriptome analysis. Transl Oncol. 2022;15(1):101277.

19. Belinsky SA. Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer. 2004;4(9):707-717.

20. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019;8(4):329-337.e324.

21. Street K, Risso D, Fletcher RB, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477.

22. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-421.

23. Anbazhagan K, Duroux-Richard I, Jorgensen C, Apparailly F. Transcriptomic network support distinct roles of classical and non-classical monocytes in human. Int Rev Immunol. 2014;33(6):470-489.

24. Georgoudaki AM, Prokopec KE, Boura VF, et al. Reprogramming Tumor-Associated Macrophages by Antibody Targeting Inhibits Cancer Progression and Metastasis. Cell Rep. 2016;15(9):2000-2011.

25. La Fleur L, Boura VF, Alexeyenko A, et al. Expression of scavenger receptor MARCO defines a targetable tumor-associated macrophage subset in non-small cell lung cancer. Int J Cancer. 2018;143(7):1741-1752.

26. Patente TA, Pinho MP, Oliveira AA, Evangelista GCM, Bergami-Santos PC, Barbuto JAM. Human Dendritic Cells: Their Heterogeneity and Clinical Application Potential in Cancer Immunotherapy. Front Immunol. 2018;9:3176.

27. Macri C, Pang ES, Patton T, O'Keeffe M. Dendritic cell subsets. Semin Cell Dev Biol. 2018;84:11-21.

28. de Andrade LF, Lu Y, Luoma A, et al. Discovery of specialized NK cell populations infiltrating human melanoma metastases. JCI Insight. 2019;4(23).

29. Crusz SM, Balkwill FR. Inflammation and cancer: advances and new agents. Nat Rev Clin Oncol. 2015;12(10):584-596.

30. Zhao H, Wu L, Yan G, et al. Inflammation and tumor progression: signaling pathways and targeted intervention. Signal Transduct Target Ther. 2021;6(1):263.

31. Murata M. Inflammation and cancer. Environ Health Prev Med. 2018;23(1):50.

32. Zhao W, Zhu B, Hutchinson A, et al. Clinical Implications of Inter- and Intratumor Heterogeneity of Immune Cell Markers in Lung Cancer. J Natl Cancer Inst. 2022;114(2):280-289.

ABSTRACT (IN KOREAN)

비소세포폐암의 종양내 이질성 기전 탐구

<지도교수 장윤수>
연세대학교 대학원 의학과
송명진

내용

종양내 이질성은 항암제 약제 내성과 치료 실패의 원인으로 알려져 있다. 이 연구를 통해 종양내 이질성과 관련된 임상적 특징과 클론 확장을 일으키는 바이오 마커를 찾고자 하였다. TCGA 편평세포폐암과 선암 코호트에서 단일염기변이, 복제수 변이, RNA 시퀀싱, 임상데이터를 얻었다. 서브클론의 수는 Sciclone 패키지를 통해 분석하였다. 차등 발현 유전자 분석을 통해 클론 확장을 유도하는 후보 유전자를 선정하였으며 이 후보유전자의 플라스미드를 비소세포암 세포주에 형질감염 시켜 세포배양을 하였다. 마지막으로 단일세포 RNA 분석을 통하여 종양내 이질성이 골수세포에 미치는 영향을 확인하였다. TCGA 데이터에서 병기 I-IV, 치료를 받지 않은 481 명의 편평세포폐암, 493 명의 선암 환자 정보를 얻었다. 서브 클론의 수는 체세포 변이 수, 종양 병기와 양의 상관 관계를 보였다. 남자, 흡연력이 있는 환자에서 서브클론의 수가 더 많았다. 차등 발현 유전자 분석을 통해 MTA1, FDG1 그리고 MSI1 을 클론확장을 유도하는 후보 유전자로 선정하였으나 후속 세포배양 실험에서 해당 유전자를 형질감염 시킨 비소세포폐암 세포는 생존하지 못하였다. 단일세포 RNA 시퀀싱 분석에서 흡연자의 M2-유사 대식세포 1 은 종양형성과정에서 감소하였으나, 고전적 단핵구, 비고전적 단핵구, 그리고 염증성 대식세포는 종양형성과정에서 증가하였다. 또한 M2-유사 대식세포 1, 고전적 단핵구, 비고전적 단핵구, 그리고 염증성

대식세포의 종양 형성과정에서의 비율 변화가 흡연 여부에 따라 통계적으로 유의한 차이를 보였다. 이 연구를 통해 종양내 이질성이 종양 돌연변이 부하, 병기, 성별, 흡연과 관련이 있음을 확인하였다. 종양내 이질성은 종양내미세환경에 염증을 유발하여 염증성 골수 세포를 증가시킬 수 있다.

핵심 되는 말: 비소세포 폐암, 종양내 이질성, 골수세포, 염증

PUBLICATION LIST

Song MJ, Lee SH, Kim EY, Chang YS. Increased number of subclones in lung squamous cell carcinoma elicits overexpression of immune related genes. Transl Lung Cancer Res. 2020;9(3):659-69.