Front. Med. 2018, 12(3): 280–288
https://doi.org/10.1007/s11684-017-0580-1

RESEARCH ARTICLE

Genomic variations in the counterpart normal controls of lung squamous cell carcinomas

Dalin Zhang[1,*], Liwei Qu[1,*], Bo Zhou[1,2,*], Guizhen Wang[1], Guangbiao Zhou (✉)[1]

[1]Division of Molecular Carcinogenesis and Targeted Therapy for Cancer, State Key Laboratory of Membrane Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; [2]University of the Chinese Academy of Sciences, Beijing 100049, China

Abstract **Lung squamous cell carcinoma (LUSC) causes approximately 400 000 deaths each year worldwide. The occurrence of LUSC is attributed to exposure to cigarette smoke, which induces the development of numerous genomic abnormalities. However, few studies have investigated the genomic variations that occur only in normal tissues that have been similarly exposed to tobacco smoke as tumor tissues. In this study, we sequenced the whole genomes of three normal lung tissue samples and their paired adjacent squamous cell carcinomas. We then called genomic variations specific to the normal lung tissues through filtering the genomic sequence of the normal lung tissues against that of the paired tumors, the reference human genome, the dbSNP138 common germline variants, and the variations derived from sequencing artifacts. To expand these observations, the whole exome sequences of 478 counterpart normal controls (CNCs) and paired LUSCs of The Cancer Genome Atlas (TCGA) dataset were analyzed. Sixteen genomic variations were called in the three normal lung tissues. These variations were confirmed by Sanger capillary sequencing. A mean of 0.5661 exonic variations/Mb and 7.7887 altered genes per sample were identified in the CNC genome sequences of TCGA. In these CNCs, C:G→T:A transitions, which are the genomic signatures of tobacco carcinogen N-methyl-N-nitro-N-nitrosoguanidine, were the predominant nucleotide changes. Twenty five genes in CNCs had a variation rate that exceeded 2%, including *ARSD* (18.62%), *MUC4* (8.79%), and *RBMX* (7.11%). CNC variations in *CTAGE5* and *USP17L7* were associated with the poor prognosis of patients with LUSC. Our results uncovered previously unreported genomic variations in CNCs, rather than LUSCs, that may be involved in the development of LUSC.**

Keywords lung cancer; counterpart normal control; genomic variations

Introduction

Tens of thousands of cancer genomes have been sequenced, and numerous point mutations, insertions and deletions (indels), structural variations, copy number alterations, epigenetic changes, and microbial infections have been uncovered [1] through tumor–normal pairs method. In this method, the cancer genome sequence is compared with a reference human genome and subtracted with those found in counterpart normal controls (CNCs, counterpart normal tissues or peripheral blood) and single-nucleotide polymorphisms (SNPs) [2]. However, whether genomic alterations occur only in CNCs and not in tumor tissues remain unclear.

Approximately 90% of lung cancer deaths are caused by cigarette smoke, which contains more than 20 lung carcinogens including nicotine-derived nitrosaminoketone (NNK) and polycyclic aromatic hydrocarbons (PAHs) [3]. Tobacco smoke induces cellular injury throughout the entire respiratory tract [4] and causes all types of lung cancer but is most strongly linked with small-cell lung cancer and lung squamous-cell carcinoma (LUSC), which accounts for approximately 25%–30% of all lung cancer cases [5,6]. Characterizing the genomic landscapes of LUSCs through the tumor–normal pairs method has shown the presence of a large number of exonic mutations, genomic rearrangements, and segments of copy number alterations [7,8]. Long-term exposure to air pollution, another cause of lung cancer, also induces the development

Received May 16, 2017; accepted August 1, 2017

Correspondence: gbzhou@ioz.ac.cn

[*] These authors contributed equally to this work.

of numerous genomic mutations in patients [9]. However, whether genomic alterations exist in the counterpart normal lung tissues that have been similarly exposed to tobacco smoke or polluted air as tumor tissues remains unclear.

In this study, we report previously unidentified CNC variations found in 481 patients with LUSC. We sequenced the whole genomes of three normal lung tissue samples and their paired adjacent squamous cell carcinomas to characterize CNC-specific genomic variations. We also analyzed the genome sequences of 478 CNCs and paired LUSCs of The Cancer Genome Atlas (TCGA) datasets. We compared the genomic sequence of CNCs with that of the reference human genome and filtered out alterations found in counterpart tumors or germline variants (normal–tumor pairs). Variations derived from sequencing artifacts were removed by VarScan fpfilter.

Materials and methods

Patients, genomic data, and analytical method

This study was approved by the research ethics committee of the Institute of Zoology, Chinese Academy of Sciences. The diagnosis of LUSC was confirmed by three pathologists. The tumor samples contained a tumor cellularity greater than 80%, and the paired normal lung tissues had no tumor content. Genomic DNA samples were isolated from normal lung tissues obtained 5 cm or more away from tumors. Sequencing libraries were constructed and sequenced with the Illumina Hiseq2000 platform [9]. The raw sequencing data were processed with the FASTX-Toolkit to retrieve high-quality paired reads, which were then aligned to the reference human genome (hg19) [11] by Burrows–Wheeler Alignment (BWA) with default parameters. After marking the duplicates with Picard, Binary Alignment (BAM) files were subjected to base recalibration and indel realignment with Genome Analysis Toolkit (GATK) [10] (https://www.broadinstitute.org/gatk/). Variants in CNCs were called by the UnifiedGenotyper of GATK and filtered against dbSNP138 common germline variants (http://genome.ucsc.edu/) and those detected in the tumor samples. The false-positive filter incorporated in VarScan v2.3.9 was used to remove sequencing artifacts and filter false-positive variants. The readcount files of variants in CNC BAMs were constructed by bam-readcount and entered into the filter with default parameters other than the following options: –min-var-basequal: 20; –min-var-mapqual: 20; –min-var-freq: 0.2; –min-var-count: 2. Variants that passed the false-positive filter with the allele frequency $\geqslant 0.20$ in the normal counterpart and $\leqslant 0.05$ in tumor samples were reserved as variants in CNCs. The called variants were validated by polymerase chain reaction (PCR) and Sanger capillary sequencing using the primers listed in Table S1 and the genomic DNA samples of the patients. Mutations in the cancer genome were also analyzed through GATK with the above criteria.

The use of the TCGA genome data was approved by the National Institutes of Health of the United States of America with the approval number of #24437-4. The dbGaP accession number is phs000178.v9.p8. The TCGA genome data of 478 patients with LUSC (Table S2) were downloaded from the Cancer Genomics Hub (https://cghub.ucsc.edu/) and analyzed as described above.

Table 1 Genomic variations in the normal lung tissues of three patients with LUSCs

Patient ID	Gene	Chr	Start	End	Ref	Alt	Function	Transcriptor	cDNA position	Amino acid
712	*C10orf95*	10	104210754	104210754	G	C	Stopgain	NM_024886	c.C234G	p.Y78X
	CNTNAP3	9	39287976	39287976	C	T	Splicing	NM_033655	c.85 + 1G>A	
	DPPA4	3	109046864	109046864	C	A	NS	NM_018189	c.G886T	p.V296F
	GLB1L	2	220102416	220102416	T	C	NS	NM_024506	c.A1507G	p.I503V
	MACF1	1	39913789	39913789	C	A	NS	NM_012090	c.C13876A	p.P4626T
	NCL	2	232325382	232325384	TCC	–	NF	NM_005381	c.807_809del	p.E271del
	NCOR2	12	124810072	124810072	G	A	NS	NM_006312	c.C7421T	p.A2474V
805	*IGFN1*	1	201178904	201178904	A	G	NS	NM_001164586	c.A4883G	p.E1628G
	MADCAM1	19	501801	501802	AG	CC	NS	NM_130760	c.[A800C; G801C]	p.K267T
	ZP3	7	76069902	76069902	G	C	NS	NM_001110354	c.G1034C	p.R345T
831	*ACAP3*	1	1233970	1233970	G	T	NS	NM_030649	c.C840A	p.S280R
	CEL	9	135947032	135947032	C	A	NS	NM_001807	c.C2152A	p.P718T
	SLAMF9	1	159923185	159923185	C	A	NS	NM_001146172	c.G305T	p.W102L
	MUC4	3	195513461	195513461	G	A	NS	NM_018406	c.C4990T	p.P1664S
	UBE2Q1	1	154530881	154530881	G	T	NS	NM_017582	c.C149A	p.S50Y
	KDM4B	19	5032981	5032981	A	T	NS	NM_015015	c.A80T	p.D27V

Alt, alterations; Chr, chromosome; NF, nonframeshift; NS, nonsynonymous; Ref, reference human genome.

Statistics

Differences between data groups were evaluated for significance using the software SPSS 17.0 for Windows (Chicago, IL, USA) and Student's *t*-test. The survival curves were plotted in accordance with the Kaplan–Meier method and compared through the log-rank test. *P* values < 0.05 were considered statistically significant.

Results

Identification and validation of CNC genomic variations in patients with LUSC

In the initial screening, the genomic DNA samples of the paired normal lung tissues and cancer tissues of three patients with LUSC were sequenced to an average of 44.15× (35.83× – 50.68×) coverage and 64.63× (range, 62.03× – 65.97×) coverage, respectively. Nucleotide substitutions and small indels were found in the three LUSC genomes, including 14 single nucleotide substitutions, one dinucleotide substitution (AG→CC in *MADCAM1*), and one indel (TCC deletion in *NCL*) (Table 1). PCR assays and subsequent sequencing were performed to verify the identified alterations in six genes in the normal lung tissues of the patients, and the results confirmed the existence of the alterations (Fig. 1). For example, the nucleotide at chr12:124810072 of *NCOR2* of hg19 is C. However, two peaks (C and T) of equal peak height were seen in the sequence of normal lung tissues, whereas a high peak of C and a very low peak of T were detected in the tumor samples of a patient (Fig. 1A, left panel). This change might lead to A2474V substitution in the encoded protein. Sequencing results using another set of primers confirmed the existence of T in the normal lung rather than in the counterpart tumor sample of the patient (Fig. 1A, right panel). Nucleotide T at this position of *NCOR2* was absent in dbSNP138 common germline variants. Similarly, the normal lung tissues had variations in *GLB1L* (Fig. 1B), *MACF1* (Fig. 1C), *C10orf95* (Fig. 1D), and *DPPA4* (Fig. 1E) compared with those in the tumor samples, hg19, and dbSNP138. TCC deletion in *NCL* in normal lung tissues was also confirmed by Sanger capillary sequencing using genomic DNA and two sets of primers (Fig. 1F).

Analyses of TCGA datasets

We expanded the observations in TCGA datasets by analyzing the genome sequences of the normal–tumor pairs from 478 patients with LUSC. Of these patients (Table S2), 353 (73.85%) were males and 125 (26.15%) were females, and the median age was 68 years old (range, 39–90 years). The smoking histories of 468 patients were available. Among these patients, 450 (96.15%) were current smokers or reformed smokers (not smoking at the time of interview but had smoked at least 100 cigarettes in their life), and 18 (3.85%) were nonsmokers (not smoking at the time of the interview and had smoked less than 100 cigarettes in their life). Adjacent normal lung tissues and peripheral blood were used as normal controls for 224 (46.86%) and 254 (53.14%) of the 478 patients with LUSC, respectively.

Variations were found throughout the genomes of CNCs and tumor tissues. A mean of 0.566 exonic alterations per megabase (Mb) was recorded in the CNC samples. This value is considerably less than that in tumor tissues (7.067 mutations/Mb, $P < 0.0001$; Table 2). A total of 0.588 exonic mutations/Mb was found in normal lung tissues. This value is approximately equal to that in peripheral blood (0.547 mutations/Mb; Table S3). CNCs from male patients had 0.598 exonic mutations/Mb, which is slightly more than that of CNCs from females (0.476 mutations/Mb; Table S4). Black people had more CNC mutations than white people (Table S5). Only 18 patients were nonsmokers in this cohort (Table S2). This finding might provide an explanation for the observation that mutations in CNCs, as well as tumors in smokers, were not significantly higher than that in nonsmokers, as reflected by mutations/Mb, mutated genes/sample, synonymous/nonsynonymous mutations, and indels/sample (Fig. S2A–S2H).

Nucleotide substitutions in TCGA datasets

The nucleotide variations of the CNC genomes were analyzed. The results showed that the C:G→T:A transitions were the most predominant nucleotide substitutions, followed by A:T→G:C transitions (Fig. 2A). C:G→A:T transversions were the most predominant nucleotide substitutions in the tumor samples of the patients, and C:G→T:A transitions were the second most prevalent nucleotide changes in the genomes of the patients (Fig. 2A). We further showed that the C:G→T:A transitions were the most prevalent nucleotide changes in CNCs of nonsmokers and smokers (Fig. 2B) and males and females (Fig. 2C).

Altered genes in CNCs of TCGA datasets

We found a mean of 7.7887 altered genes per CNC sample. This value is considerably less than that in tumor samples (164.8159 mutated genes/sample, $P < 0.0001$; Table 2). In the 478 CNC samples, 25 genes had a variation rate of more than 2% (Fig. 2D and Table S6). *ARSD* [12] represented the most frequently altered gene and was altered in 89/478 (18.62%) of the CNC samples (Fig. 2D). In the 89 CNCs, 192 variations were found in *ARSD*, and

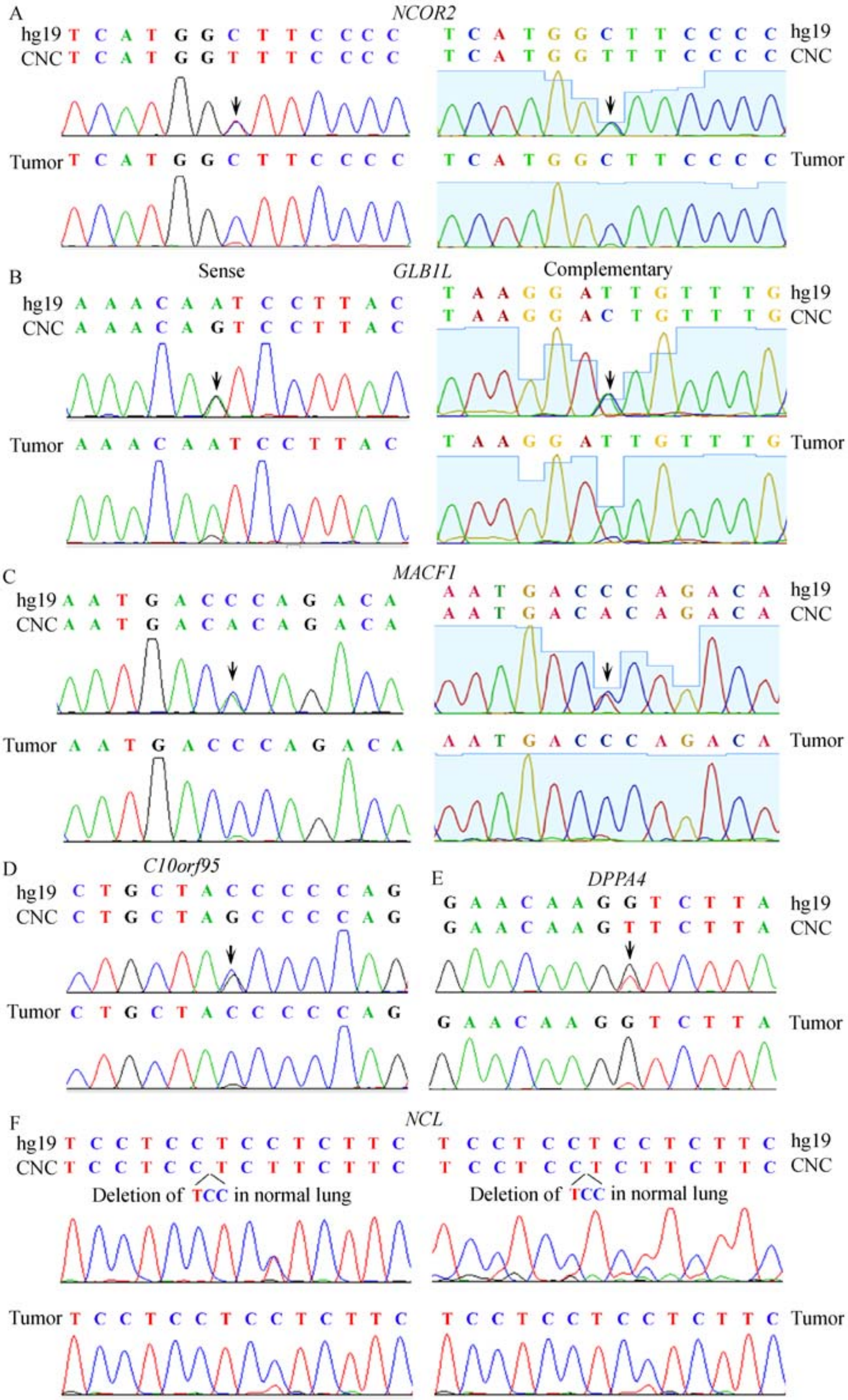

Fig. 1 Validation of genomic variations in normal lung tissues. Genomic variations were identified through the analyses of whole-genome sequencing data. Polymerase chain reaction and Sanger capillary sequencing were performed using the primers listed in Table S1and genomic DNA samples from three patients with LUSC. (A) *NCOR2* in the normal lung and tumor samples of a patient with LUSC. Two sets of primers were used. (B) *GLB1L* in the normal lung and tumor samples of a patient with LUSC. Two sets of primers were used. (C) *MACF1* in the normal lung and tumor samples of a patient with LUSC. Two sets of primers were used. (D) *C10orf95* in the normal lung and tumor samples of a patient with LUSC. (E) *DPPA4* in the normal lung and tumor samples of a patient with LUSC. (F) *NCL* in the normal lung and tumor samples of a patient with LUSC.

Table 2 Somatic mutations in CNCs and the tumor samples of TCGA LUSCs

	Exonic mutations/MB	Mutated genes/sample	Nonsynonymous mutations/sample	Synonymous mutations/sample	Rearrangements/sample	
					Frameshift	Inframe
CNC	0.5661	7.7887	6.9184	4.1339	0.3661	1.0774
Tumor	7.0671	164.8159	152.5272	53.3745	12.2950	3.1088
P value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

28 (14.583%), 28 (14.583%), and 27 (14.06%) of these alterations led to G175D, L166Q, and M176K amino acid substitutions (Fig. 3A), respectively. *MUC4*, *RBMX*, *MUC5B*, *RP1L1*, and *CDC27* were mutated in 42 (8.79%), 34 (7.11%), 18 (3.77%), 18 (3.77%), and 17 (3.56%) of the 478 CNC samples, respectively. Variations and small indels, which resulted in single amino acid substitutions or the truncation of the encoded proteins, were frequently seen in CNC variations. Some genes (e.g., *ARSD*) also had variation hotspots (Fig. 3). *TP53*, *MLL2*, *PIK3CA*, *CDKN2A*, and *NFE2L2* were frequently mutated in LUSC [7]. However, no alteration in these genes was detected in these CNC samples (Table S6).

Altered signaling pathways

The affected signaling pathways were analyzed through Gene Ontology analysis [13]. The results showed that genes involved in interferon-γ (IFN-γ)-mediated signaling pathway and O-glycan processing, antigen processing and presentation were altered in CNC samples (Fig. S3A). Assays using the Kyoto Encyclopedia of Genes and Genome database showed that allograft rejection, cell adhesion molecules, and asthma pathways were affected (Fig. S3B).

CNC variations associated with poor prognosis of the patients

We analyzed the potential association between variations in CNCs and the prognosis of the patients using Kaplan–Meier method. Variations in two genes were associated with poor clinical outcome (Fig. 4). Alternative splicing variations in *CTAGE5* (for CTAGE Family Member 5)

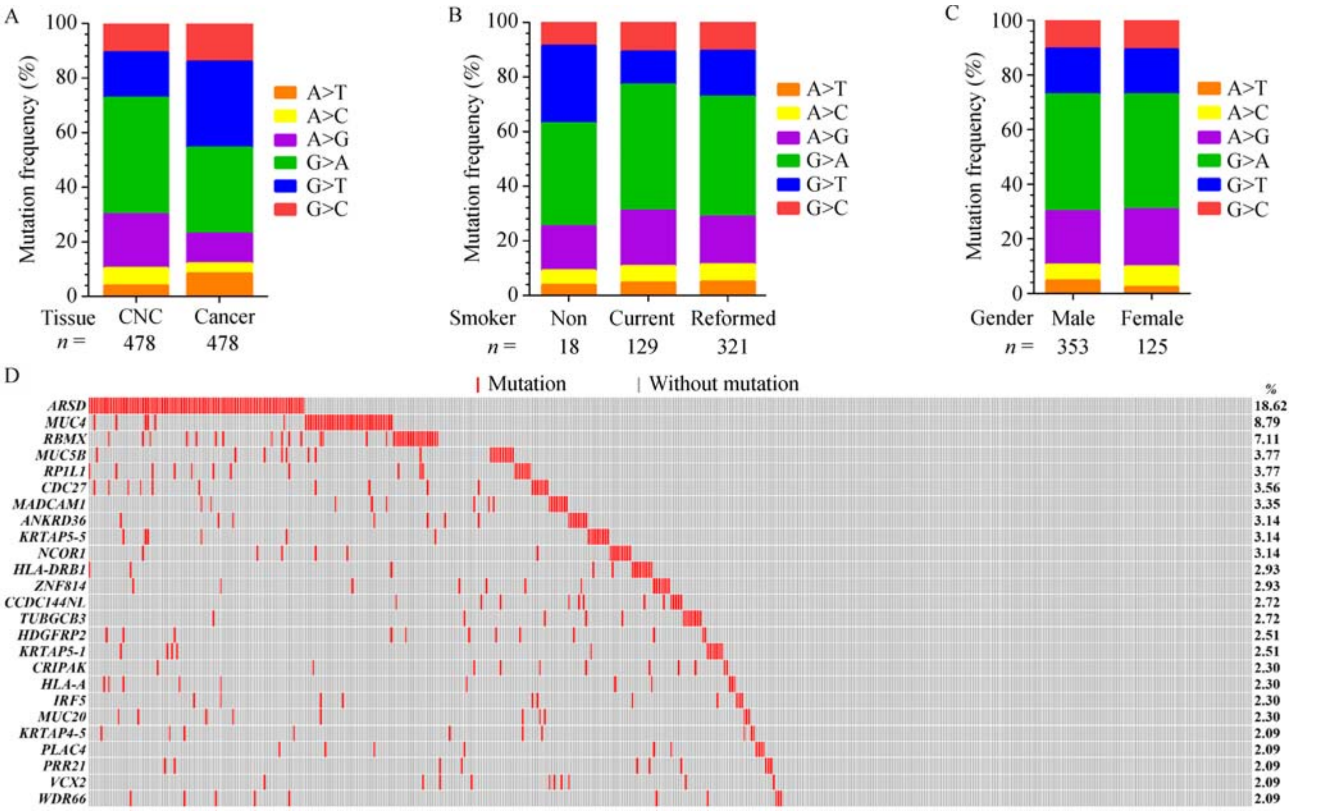

Fig. 2 Genomic variations in the CNCs of 478 patients with LUSC. (A) Frequency of each type of single base substitution in CNC and tumor samples of 478 patients with LUSC. (B) Proportion of nucleotide changes in the CNCs of nonsmokers, current smokers, and reformed smokers. (C) Frequency of base substitution in the CNCs of male and female patients with LUSC. (D) Altered genes in the CNCs of patients with LUSC. CNC samples are arranged from left to right in the top track.

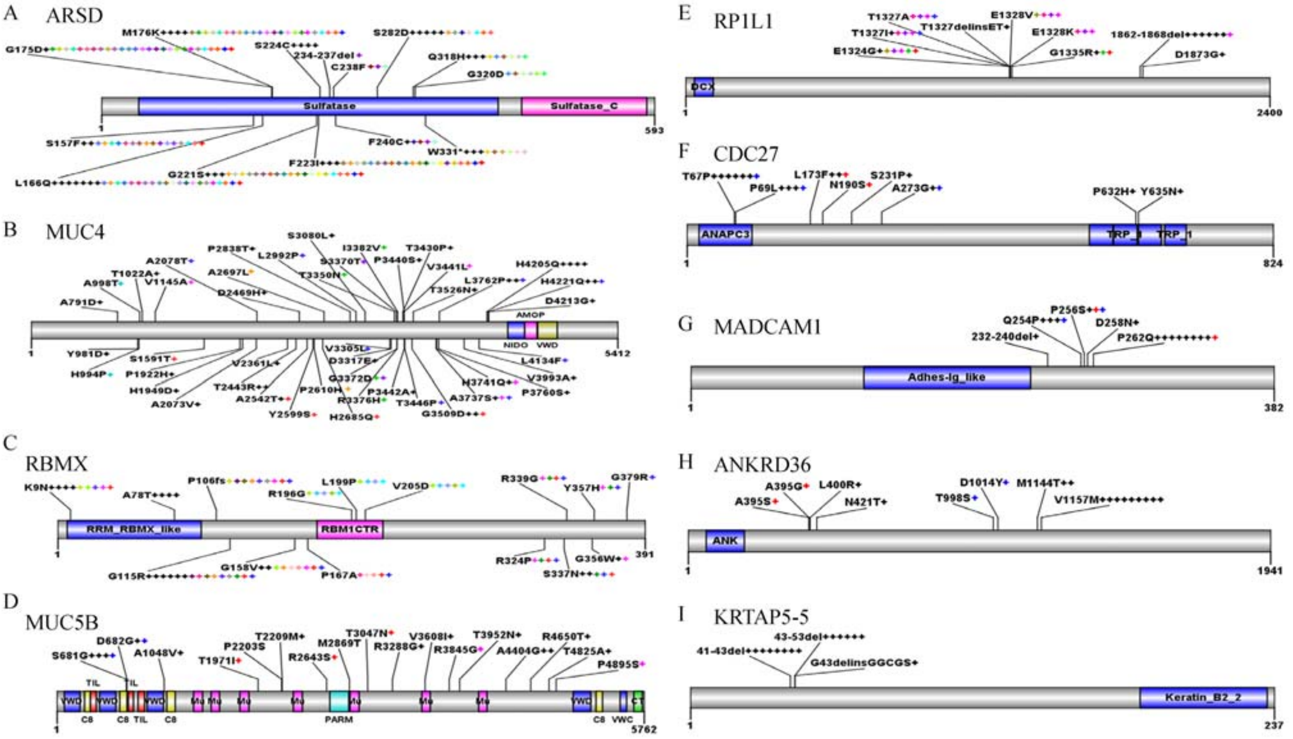

Fig. 3 CNC variations in nine representative genes. Schematic representations of proteins encoded by the genes are shown. Numbers refer to amino acid residues. Each "+" corresponds to an independent, mutated CNC sample, and mutations in a nonred "+" with the same color are found in the same patient. (A) Variations in ARSD. (B) Variations in MUC4. (C) Variations in RBMX. (D) Variations in MUC5B. (E) Variations in RP1L1. (F) Variations in CDC27. (G) Variations in MADCAM1. (H) Variations in ANKRD36. (I) Variations in KRTAP5-5.

[14], c.1356 + 2_1356 + 3delTA, were observed in the CNCs of seven (1.46%) of the 478 patients (Fig. 4A, upper panel). In the 473 patients with available survival information, the overall survival of the seven patients with the splicing variant of *CTAGE5* in CNCs was considerably shorter than those with wild type *CTAGE5* ($P < 0.0001$; Fig. 4A). Nucleotide changes that result in R301L substitution in Ubiquitin Specific Peptidase 17-Like Family Member 7 (*USP17L7*) [15] gene were seen in six CNCs (Fig. 4B). Patients with these CNC variations

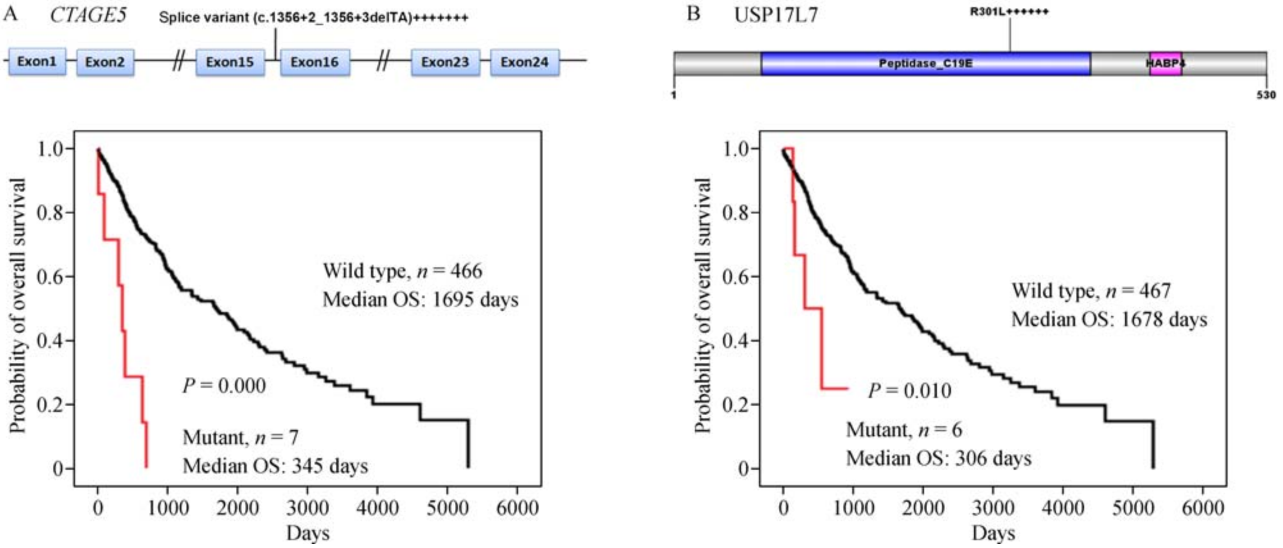

Fig. 4 CNC variations associated with poor patient prognosis. (A) Variations of *CTAGE5* in CNC samples and Kaplan–Meier curve for the overall survival of the patients. (B) Variations of USP17L7 in the CNCs and overall survival of the patients.

had considerably shorter survival time than those with wild type *USP17L7* (Fig. 4B).

Discussion

Chronic exposure to tobacco smoke causes the development of LUSC in the central airway. The development of LUSC follows a stepwise progression, e.g., from hyperplasia, metaplasia, dysplasia, and to carcinoma *in situ* [16]. Molecular lesions (e.g., genetic mutations and somatic copy number variations) are present in premalignant patches [16–19], and somatic genomic mutations have been found in lung tumors [7,8]. In this study, we dissected the whole genome sequence of normal lung tissues from three patients with LUSC using the normal–tumor pairs method to characterize the genomic alterations present in normal lungs that have been exposed to tobacco smoke. We found that the normal lung tissues of three patients with LUSC harbored genomic variations that have not been observed in their counterpart tumor samples, hg19, and dbSNP138 (Table 1). The six identified genomic variations were obvious upon validation through Sanger capillary sequencing (Fig. 1, A through E). However, the tumor samples also exhibited very low peaks of respective nucleotides (Fig. 1), suggesting the presence of normal lung epithelial cells in tumor samples or allelic loss in the tumor cells. Genomic variations were also detected in CNCs of lung adenocarcinomas (LUADs) in our own genome sequencing data and TCGA dataset [20]. In the TCGA datasets, the CNC genomic alterations displayed a frequency of up to 18.62%, and variations in two genes were associated with poor prognosis. Although we were unable to verify these variations because of the unavailability of the TCGA samples, our results provide new opportunities for the investigation of cigarette smoke-induced genomic mutations in normal lungs and the elusive lung carcinogenesis.

We found that the C:G→T:A transitions are the most prevalent nucleotide changes in CNCs and the second most prevalent substitutions in LUSCs. Meanwhile, C:G→A:T transversions are the predominant nucleotide substitutions in LUSCs (Fig. 2A). Previous studies have shown that C:G→T:A transitions are the genomic signature of the tobacco carcinogen N-methyl-N-nitro-N-nitrosoguanidine (MNNG) [21] and the mutational fingerprints of aging [22]. C:G→A:T transversions represent a genomic signature of PAHs, which are found in tobacco smoke and act as air pollutants [9,21,23]. Given that SNPs, including those of elderly individuals, had been filtered in this study, C:G→T:A transitions in CNCs may reflect the exposure of the patients to environmental carcinogens such as MNNG (from tobacco smoke and second-hand smoke for nonsmokers). Our results further indicate that the genotoxicity of tobacco smoke triggers lung carcinogenesis.

Genomic variations are frequently seen in CNCs, suggesting that these alterations may perturb the biological function of relevant proteins and are involved in lung tumorigenesis. We hypothesized that some of these variants are pro-oncogenes (e.g., *CDC27* [24] and *MADCAM1* [25]) or tumor suppressors (e.g., *NCOR1* [26]). Cells that harbor the gain-of-function or loss-of-function mutations of these genes are in a precancerous stage, and the accumulation of other mutations will result in the transformation and development of malignant neoplasms. Thus, LUSCs may have multiclonal origins with genetic variants. In addition, many of the CNC-altered genes are associated with immune response (Fig. S3), which may help avoid immune destruction and cancer initiation. CNC cells may interact with tumor cells to provide an environment that either fosters or constrains carcinogenesis. In addition, variations in the components of the DNA-damage response system, such as RBMX [27], are also frequently seen in CNCs (Fig. 2D), suggesting their role in maintaining genome stability. Some CNC variations, i.e., *CTAGE5* and *USP17L7*, are associated with poor patient prognosis (Fig. 4). This association further suggests their significance in lung carcinogenesis. Notably, some CNC variations that are similar to passenger mutations in tumor samples may have a minimal role in lung tumorigenesis. Further works required to investigate the roles of CNC variations in the initiation and progression of LUSC.

Acknowledgements

This work was supported by the National Natural Science Funds of China for Distinguished Young Scholar (No. 81425025), the National Key Research and Development Program of China (No. 2016YFC0905500), the National Natural Science Foundation of China (No. 81672765), the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA12010307), and grants from the State Key Laboratory of Membrane Biology. The funders had no role in the design, data collection, and analysis of the study or in the preparation of and the decision to publish the manuscript. The results shown here are partly based on the data generated by The Cancer Genome Atlas, which is managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. The dbGaP accession number is phs000178.v9.p8.

Compliance with ethics guidelines

Dalin Zhang, Liwei Qu, Bo Zhou, Guizhen Wang, and Guangbiao Zhou declare that they have no conflict of interest. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the *Helsinki Declaration* of 1975, as revised in 2000. Additional informed consent was obtained from all patients for which identifying information is included in this article.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://doi.org/10.1007/s11684-017-0580-1 and is accessible for authorized users.

References

1. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014; 505(7484): 495–501
2. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010; 11(10): 685–696
3. Hecht SS. Lung carcinogenesis by tobacco smoke. Int J Cancer 2012; 131(12): 2724–2732
4. Auerbach O, Hammond EC, Kirman D, Garfinkel L. Effects of cigarette smoking on dogs. II. Pulmonary neoplasms. Arch Environ Health 1970; 21(6): 754–768
5. Herbst RS, Heymach JV, Lippman SM. Lung cancer. N Engl J Med 2008; 359(13): 1367–1380
6. Lemjabbar-Alaoui H, Hassan OUI, Yang YW, Buchanan P. Lung cancer: biology and treatment options. Biochim Biophys Acta 2015; 1856(2): 189–210
7. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012; 489 (7417): 519–525
8. Li C, Gao Z, Li F, Li X, Sun Y, Wang M, Li D, Wang R, Li F, Fang R, Pan Y, Luo X, He J, Zheng L, Xia J, Qiu L, He J, Ye T, Zhang R, He M, Zhu M, Hu H, Shi T, Zhou X, Sun M, Tian S, Zhou Y, Wang Q, Chen L, Yin G, Lu J, Wu R, Guo G, Li Y, Hu X, Li L, Asan, Wang Q, Yin Y, Feng Q, Wang B, Wang H, Wang M, Yang X, Zhang X, Yang H, Jin L, Wang CY, Ji H, Chen H, Wang J, Wei Q. Whole exome sequencing identifies frequent somatic mutations in cell-cell adhesion genes in Chinese patients with lung squamous cell carcinoma. Sci Rep 2015; 5: 14237
9. Yu XJ, Yang MJ, Zhou B, Wang GZ, Huang YC, Wu LC, Cheng X, Wen ZS, Huang JY, Zhang YD, Gao XH, Li GF, He SW, Gu ZH, Ma L, Pan CM, Wang P, Chen HB, Hong ZP, Wang XL, Mao WJ, Jin XL, Kang H, Chen ST, Zhu YQ, Gu WY, Liu Z, Dong H, Tian LW, Chen SJ, Cao Y, Wang SY, Zhou GB. Characterization of somatic mutations in air pollution-related lung cancer. EBioMedicine 2015; 2(6): 583–590
10. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20(9): 1297–1303
11. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010; 26(5): 589–595
12. Franco B, Meroni G, Parenti G, Levilliers J, Bernard L, Gebbia M, Cox L, Maroteaux P, Sheffield L, Rappold GA, Andria G, Petit C, Ballabio A. A cluster of sulfatase genes on Xp22.3: mutations in chondrodysplasia punctata (CDPX) and implications for warfarin embryopathy. Cell 1995; 81(1): 15–25
13. Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009; 4(1): 44–57
14. Saito K, Yamashiro K, Ichikawa Y, Erlmann P, Kontani K, Malhotra V, Katada T. cTAGE5 mediates collagen secretion through interaction with TANGO1 at endoplasmic reticulum exit sites. Mol Biol Cell 2011; 22(13): 2301–2308
15. Burrows JF, McGrattan MJ, Johnston JA. The DUB/USP17 deubiquitinating enzymes, a multigene family within a tandemly repeated sequence. Genomics 2005; 85(4): 524–529
16. Ooi AT, Gower AC, Zhang KX, Vick JL, Hong L, Nagao B, Wallace WD, Elashoff DA, Walser TC, Dubinett SM, Pellegrini M, Lenburg ME, Spira A, Gomperts BN. Molecular profiling of premalignant lesions in lung squamous cell carcinomas identifies mechanisms involved in stepwise carcinogenesis. Cancer Prev Res (Phila) 2014; 7(5): 487–495
17. Gomperts BN, Spira A, Massion PP, Walser TC, Wistuba II, Minna JD, Dubinett SM. Evolving concepts in lung carcinogenesis. Semin Respir Crit Care Med 2011; 32(1): 32–43
18. Kadara H, Shen L, Fujimoto J, Saintigny P, Chow CW, Lang W, Chu Z, Garcia M, Kabbout M, Fan YH, Behrens C, Liu DA, Mao L, Lee JJ, Gold KA, Wang J, Coombes KR, Kim ES, Hong WK, Wistuba II. Characterizing the molecular spatial and temporal field of injury in early-stage smoker non-small cell lung cancer patients after definitive surgery by expression profiling. Cancer Prev Res (Phila) 2013; 6(1): 8–17
19. Wistuba II, Behrens C, Milchgrub S, Bryant D, Hung J, Minna JD, Gazdar AF. Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma. Oncogene 1999; 18(3): 643–650
20. Qu LW, Zhou B, Wang GZ, Chen Y, Zhou GB. Genomic variations in paired normal controls for lung adenocarcinomas. Oncotarget 2017 (in press)
21. Olivier M, Weninger A, Ardin M, Huskova H, Castells X, Vallée MP, McKay J, Nedelko T, Muehlbauer KR, Marusawa H, Alexander J, Hazelwood L, Byrnes G, Hollstein M, Zavadil J. Modelling mutational landscapes of human cancers *in vitro*. Sci Rep 2014; 4: 4482
22. Dollé MET, Snyder WK, Dunson DB, Vijg J. Mutational fingerprints of aging. Nucleic Acids Res 2002; 30(2): 545–549
23. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, Maher CA, Fulton R, Fulton L, Wallis J, Chen K, Walker J, McDonald S, Bose R, Ornitz D, Xiong D, You M, Dooling DJ, Watson M, Mardis ER, Wilson RK. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell 2012; 150(6): 1121–1134
24. Qiu L, Wu J, Pan C, Tan X, Lin J, Liu R, Chen S, Geng R, Huang W. Downregulation of CDC27 inhibits the proliferation of colorectal cancer cells via the accumulation of p21Cip1/Waf1. Cell Death Dis 2016; 7(1): e2074
25. Wang J, Ma L, Tang X, Zhang X, Qiao Y, Shi Y, Xu Y, Wang Z, Yu Y, Sun F. Doxorubicin induces apoptosis by targeting Madcam1 and AKT and inhibiting protein translation initiation in hepatocellular carcinoma cells. Oncotarget 2015; 6(27): 24075–24091
26. Wang W, Song XW, Bu XM, Zhang N, Zhao CH. PDCD2 and NCoR1 as putative tumor suppressors in gastric gastrointestinal stromal tumors. Cell Oncol (Dordr) 2016; 39(2): 129–137

27. Adamson B, Smogorzewska A, Sigoillot FD, King RW, Elledge SJ. A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. Nat Cell Biol 2012; 14(3): 318–328