Genomic variations in paired normal controls for lung adenocarcinomas

Running title: Genomic variations in normal controls of LUADs

Li-Wei Qu,[1*] Bo Zhou,[1,2*] Gui-Zhen Wang,[1] Ying Chen,[1] Guang-Biao 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;

[2]University of the Chinese Academy of Sciences, Beijing 100049;

[*]These authors contributed equally to this work.

[#]Corresponding author: Guang-Biao Zhou, Institute of Zoology, Chinese Academy of Sciences, 1 West Beichen Road, Beijing 100101, China. Tel: 86-10-64807951; Email: gbzhou@ioz.ac.cn.

Abstract

Somatic genomic mutations in lung adenocarcinomas (LUADs) have been extensively dissected, but whether the counterpart normal lung tissues that are exposed to ambient air or tobacco smoke as the tumor tissues do, harbor genomic variations, remains unclear. Here, the genome of normal lung tissues and paired tumors of 11 patients with LUAD were sequenced, the genome sequences of counterpart normal controls (CNCs) and tumor tissues of 513 patients were downloaded from TCGA database and analyzed. In the initial screening, genomic alterations were identified in the "normal" lung tissues and verified by Sanger capillary sequencing. In CNCs of TCGA datasets, a mean of 0.2721 exonic variations/Mb and 5.2885 altered genes per sample were uncovered. The C:G→T:A transitions, a signature of tobacco carcinogen N-methyl-N-nitro-N-nitrosoguanidine, were the predominant nucleotide changes in CNCs. 16 genes had a variant rate of more than 2%, and CNC variations in *MUC5B*, *ZXDB*, *PLIN4*, *CCDC144NL*, *CNTNAP3B*, and *CCDC180* were associated with poor prognosis whereas alterations in *CHD3* and *KRTAP5-5* were associated with favorable clinical outcome of the patients. This study identified the genomic alterations in CNC samples of LUADs, and further highlighted the DNA damage effect of tobacco on lung epithelial cells.

Introduction

Comprehensive sequencing efforts over the past decade have confirmed that cancer is a disease of the genome[1]. Genomic alterations in cancers include point mutations, insertions and deletions (indels), structural variations, copy number alterations, epigenetic changes, and microbial infections. Abnormalities in oncogenes and tumor suppressors act as driver mutations to initiate the onset and progression of cancers. Technically, these alterations are identified by comparing the cancer genome sequence with a reference human genome and excluding those also found in normal controls (counterpart normal tissues or peripheral blood) and single nucleotide polymorphisms (SNPs)[2]. This strategy is effective to uncover somatic genomic alterations in tumor tissues. However, whether there is any genomic alteration in counterpart normal controls (CNCs) but not cancer tissues, remains unclear.

Lung cancer is the most common cause of cancer related mortality worldwide[3] and can be divided into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC comprises of lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and large cell carcinoma[4]. It is estimated that 90% of lung cancer deaths are caused by cigarette smoke[5], whereas SCLC and LUSC are most strongly linked with smoking[4]. Since the late 1980s, LUAD has become the most common type (40%) of lung cancer that is related to cigarette smoke due to the changes in cigarette composition and filtering which favor adenocarcinoma histology[6]. There is also a proportion of LUADs that is not associated with smoking[4], while air pollution represents another cause of lung cancer which induces somatic mutations in the genomes[7]. The cellular injury produced by smoking involves the whole respiratory tract[8, 9]; in the initial phase of lung carcinogenesis, injury is repaired by stem/progenitor cells, forming a clonal group of

self-renewing daughter cells. Accumulation of genetic alterations in the premalignant field results in proliferation of the cells and expansion of the field, gradually displacing the normal epithelium and developing into a malignant neoplasm[10, 11, 12]. Genomic alterations of LUADs had been characterized and druggable targets had been unveiled[13, 14]. However, very few studies systematically dissect genomic mutations specifically occurred in "normal" lung tissues and their roles in carcinogenesis, and no study reports how the adjacent tissues escape from developing into malignant neoplasms.

In our work analyzing somatic genomic mutations in air pollution-related lung cancers[7], we noted that when compared with the reference human genome, some variations were seen in counterpart normal lung tissues but not cancer samples (Table 1). To expand these observations, we further analyzed the Cancer Genome Atlas (TCGA) lung cancer genome data of 513 CNCs and counterpart tumor samples (Figure S1). Our results showed that the CNCs did harbor some genomic variations.

Results

Identification of CNC somatic genomic variations

The genomic DNAs of paired normal lung tissues and cancer tissues were sequenced to an average of 42.89× (30.07×–78.70×) coverage and 66.04× (range, 61.02×–74.64×) coverage, respectively. Twenty nucleotide substitutions and 7 small insertions and deletions (indels) were found in the 11 LUADs genomes (Tables 1). Variations in 6 representative genes in the normal lung tissues were validated by RT-PCR assays and subsequent sequencing (see Table S1 for sequence of primers), and the results confirmed the existence of CNC alterations (Figure 1). For example, the nucleotide at chr18: 22806878 of *ZNF521* of hg19 is C; but two peaks (C and

T) of approximately equal peak height were seen in sequence of normal lung tissues, whereas a high peak of C and a low peak of T were detected in tumor samples of the patient (Figure 1A, left panel). This change might lead to P335L substitution in the encoded protein. Sequencing results using another set of primers confirmed the existence of T in normal lung rather than counterpart tumor sample of the patient (Figure 1A, right panel). Nucleotide T at this position of *ZNF521* was not found in dbSNP138 common germline variants. The nucleotide at chr17: 76121923 of *TMC6* of hg19 is C; but two peaks (C and T) of approximately equal peak height were seen in sequence of normal lung tissues, whereas a high peak of C and a low peak of T were detected in tumor samples of the patient (Figure 1B, left panel). This change might lead to T105M substitution in the encoded protein. Sequencing results using another set of primers confirmed the existence of T in normal lung rather than counterpart tumor sample of the patient (Figure 1B, right panel). Similarly, the normal lung tissues had a single nucleotide change in *CEP250* (Figure 1C), *C9orf66* (Figure 1D), and *HIST1H1D* (Figure 1E), compared to those in the tumor samples, hg19, and dbSNP138. Substitutions of GT by AC were seen in *UNC93A* in normal lung tissue of a patient with untreated LUAD (Figure 1F).

Analyses of TCGA datasets

To expand the above observations, the genome sequence of cancer-control paired samples from 513 patients were downloaded from TCGA datasets and carefully analyzed. Among these patients (Table S2), 239 (46.6%) were males and 274 (53.4%) were females, and the median age was 66 years old (range, 33 – 88 years). Smoking history was available in 499 patients, among them 425 (85.2%) were current or reformed smokers (persons who were not smoking at the time of interview but had smoked at least 100 cigarettes in their life) and 74 (14.8%) were

nonsmokers (not smoking at the time of interview and had smoked less than 100 cigarettes in their life). Adjacent normal lung tissues and peripheral blood were used as normal controls for 135 (26.3%) and 378 (73.7%) of the 513 lung adenocarcinomas, respectively.

Genomic variations were found throughout the counterpart control genomes. A mean of 0.2721 exonic variations/megabase (Mb) and 5.2885 variants/sample was recorded in the CNCs (Table 2). The normal lung tissues had approximately equal mutations to the blood cells (Table S3), while males and females had equal CNC mutations (Table S4). We compared the CNC mutations in nonsmokers, current smokers and reformed smokers, and found that smokers harbored more variations than nonsmokers, as reflected by variations/Mb (Figure S2A), altered genes/sample (Figure S2B), synonymous/nonsynonimous variations (Figure S2C), and indels/sample (Figure S2D). We also analyzed genomic mutations in tumor samples by comparing the cancer genome sequencing with reference and excluding those found in CNCs with the same calling criteria, and reported that tumor samples had much more mutations than CNCs (Table 2 and Figure S2, E-H).

Nucleotide substitutions in TCGA CNC samples

Tobacco smoke contains more than 20 lung carcinogens, e.g., nicotine-derived nitrosaminoketone (NNK)[5] and polycyclic aromatic hydrocarbons (PAHs) which are associated with the C:G→A:T transversions in the genomes[7, 15, 16]. We analyzed the nucleotide substitutions in the genome, and found that the C:G→T:A and the A:T→G:C transitions were the most and the second most prevalent nucleotide substitutions in the CNCs (Figure 2A). In tumor samples, while the C:G→A:T transversions were the most prevalent nucleotide substitutions, the C:G→T:A transitions represented the second most prevalent nucleotide

substitutions (Figure 2A). In CNC samples, the C:G→T:A transitions were the most prevalent nucleotide substitutions in nonsmokers, current smokers and reformed smokers (Figure 2B). In consistence with previous report[16], the C:G→T:A transitions were the most prevalent nucleotide substitutions in tumor samples of nonsmokers (Figure 2C), whereas the C:G→A:T transversions were the most prevalent nucleotide substitutions in tumor samples of smokers (Figure 2C).

Recurrently altered genes in CNC samples

We reported that there were 16 genes which had a variation rate of more than 2% in the 513 CNCs (Figure 2D and Table S5). *MUC4* represented the most frequently altered gene, which was mutated in 46/513 (8.97%) of the CNC samples (Figure 2D). *FAM90A1*, *ARSD*, *MADCAM1*, *ZNF814*, *CDC27*, and *HDGFRP2* were mutated in 35 (6.82%), 32 (6.24%), 26 (5.07%), 22 (4.29%), 17 (3.31%), and 16 (3.12%) of the 513 CNC samples, respectively (Figure 2D). *TP53*, *KRAS* and *EGFR* which were frequently mutated in LUADs[13], were mutated in only 2 (0.39%), 0, and 0 CNCs, respectively (Table S5).

MUC4 encodes an integral membrane glycoprotein found on the cell surface and plays a role in tumor progression[17]. We found 35 types of variations distributed throughout the entire gene of *MUC4*, including 4 recurrent ones (Figure 3). Frequent variations were found in *FAM90A1*[18], in that in all 35 patients with variant *FAM90A1*, the nucleotide G of codon (ACG) for T344 was replaced by A and a codon for valine (GUC) was inserted in 12:8374781 position at Chr 8, leading to insertion of V right after T344 (Figure 3). There were 67 variations in *Arylsulfatase D* (*ARSD*) gene[19] in 32 CNCs, with S157F, G175D, and S176K substitutions found in 9 (14.8%), 10 (16.4%), and 10 (16.4%) of the total variations. Twenty-nine alterations were found in

MADCAM1[20] in 26 CNCs (Figure 3). Variations in *ZNF814*, *CDC27*, *HDGFRP2*, *RBMX*, *WDR66*, and *ANKRD36* in CNCs were also shown in Figure 3.

Altered signaling pathways

Pathway analysis was performed using the KEGG (Kyoto Encyclopedia of Genes and Genomes) database, and the results showed that genes involved in type I diabetes mellitus, allograft rejection, graft-versus-host disease, autoimmune thyroid disease, antigen processing and present were the most frequently altered genes (Figure S3A). The GO analysis confirmed that immune response genes were altered in CNC samples, and genes associated heart, muscle, and cell development were also perturbed in CNCs (Figure S3B).

Variations associated with poor prognosis

We analyzed the potential association between the CNC variations and the prognosis of the patients using the Kaplan-Meier method, and found that variations in 8 genes were associated with the prognosis of 502 patients whose survival information was available (Figure 4). Sixteen variations were found in *MUC5B* in 12 (2.34%) of the 513 patients (Figure 4A). As compared with patients (n=490) harboring wild type *MUC5B*, those with mutant transcript had much shorter overall survival (P=0.013; Figure 4A). *ZXDB* [21] variations were seen in 10 (1.95%) of the patients, whose overall survival time was shorter than those (n=492) with wild type *ZXDB* (Figure 4A). Patients with variant *PLIN4*[22], *CCDC144NL*, *CNTNAP3B*[23], or *CCDC180*[24] in CNCs also had shorter overall survival than patients with wild type transcripts (Figure 4A).

Variations in CNC samples associated with favorable prognosis of the patients

We found 14 variations in the *Chromodomain Helicase DNA Binding Protein 3* (*CHD3*[25]) gene in CNC samples of 13 (2.53%) of the 513 patients. These variations encoded proteins with

P99Lsubstitution in 1 patient, I234L substitution in 8 patients, and S237P change in 5 patients (Figure 4B). Interestingly, none of the variant *CHD3*-bearing patients died within a follow-up of up to 2261 days, whereas the median overall survival of patients with wild type CHD3 (n=490) was 1492 days (range, 0-7248 days) (p=0.047; Figure 4B). Deletion mutations, i.e., deletion of codons for G41 – G43 and G43 – A53 of the protein product encoded by *KRTAP5-5*[26] gene, were found in CNCs of 10 patients. Of note, patients with variant *KRTAP5-5* had longer overall survival time than those with wild type transcript (p=0.048; Figure 4B).

Discussion

In our work characterizing somatic mutations in air pollution-related lung cancer[27], we unexpectedly noted that compared to reference human genome and cancerous tissues, the normal control lung tissues also had single nucleotide variations. This was the impetus for us to conduct this study. We used the "normal-tumor pairs" method to investigate the genomic variations in CNC samples, and reported that the CNCs did harbor genomic alterations that were confirmed by Sanger capillary sequencing of PCR products (Figure 1). These observations were validated in TCGA lung cancer genome data. In CNC genomes of TCGA datasets, the variation rates were 0.2721 exonic variations/Mb and 5.2885 altered genes/sample, and the recurrent variants were of intermediate frequency (~ 8.97% of the patients), whereas variations in 8 genes were associated with poor or favorable prognosis of the patients. We also dissected CNC variations in lung squamous cell carcinoma, and reported a mean of 0.5661 exonic variations/Mb and 7.7887 altered genes/sample in 478 patients. These results uncover the previously unidentified CNC variation, and suggest a role of these alterations in lung carcinogenesis.

Some human carcinogens exhibit specific mutation signatures in the genome. For instance, PAHs which are the main carcinogens of tobacco smoke and air pollution[5, 28], induce the C:G→A:T substitutions in the genome[7, 15, 16]. The alkylating agent N-methyl-N-nitro-N-nitrosoguanidine (MNNG) which is found in tobacco smoke and a representative molecule of the N-nitroso compounds formed in human stomach[29], causes the C:G→T:A nucleotide changes in the genome[15]. We found that in CNCs of LUADs of both the smokers and non-smokers, the C:G→T:A transitions were the predominant nucleotide changes (Figure 2A, B). In tumor samples of the patients, the C:G→T:A substitutions were the predominant nucleotide changes in non-smokers, while C:G→A:T transversions were the most frequently detected substitutions in smokers (Figure 2C). Though C:G→T:A substitutions may be the mutational fingerprints of ageing[30], these changes in CNCs may be the result of exposure to environmental carcinogens (including second hand smoke in nonsmoker) rather than a consequence of ageing, because in this study the SNPs including those of elder individuals had been filtered out. These data further suggest the key role of environmental factors in lung carcinogenesis.

The significance of these CNC variations in lung carcinogenesis remains unclear. One possibility was that some of the variant genes (e.g., *MUC4*) were pro-oncogenes, cells harboring these CNC variations were in a "precancerous" stage, and accumulation of other mutations would result in transformation and development of malignant neoplasms. Hence, cancer may have multi-clonal origins. Secondly, many of the CNC altered genes were associated with immune response (Figure S3), which may facilitate avoiding immune destruction and cancer initiation. These possibilities warrant further investigation, and CNC variations in other types of cancers including cancers of oral, esophagus, stomach, colorectal,

and liver, are worthy of further dissection.

Why CNC mutations were associated with prognosis represents an open question. We hypothesized that these observations may suggest the interactions between CNC tissues and tumor cells, constituting a micro-environment to foster or constrain cancer growth. Those CNC variations associated with poor outcome (*MUC5B*, *ZXDB*, *PLIN4*, *CCDC144NL*, *CNTNAP3B*, and *CCDC180*) may probably facilitate cancer initiation and progression, and may be a new clue to study cancer metastasis and evolution. *CHD3* encodes a chromatin-remodeling factor to repress transcription and is involved in breast cancer[31]. *KRTAP5-5* is an oncogene regulating cancer cell motility and vascular invasion[32]. Whether variations in these two genes modulate transcription machinery or tumor microenvironment to constrain cancer and thus contribute to favorable prognosis, needs to be determined. In addition, how to target the CNC variations to develop novel therapeutics represents another open question.

Materials and methods

Patients, whole genome sequencing, and validation by Sanger capillary sequencing

The aim of this study was to address whether there is any genomic alteration in CNCs rather than counterpart tumor tissues (Figure S1). This study was approved by the research ethics committee of our institute. Genomic DNAs were isolated from normal lung tissues (5 cm or more away from the tumors) and adjacent tumors of 11 patients with previously untreated LUAD, the sequencing libraries were constructed, and sequenced using the Illumina Hiseq2000 platform[7]. The genome sequences of counterpart normal lung tissues were analyzed by the Genome Analysis Toolkit (GATK)[33] UnifiedGenotyper and VarScan, compared with a reference human genome (hg19, downloaded from http://genome.ucsc.edu/)[34] and filtered

against dbSNP138 common germline variants (downloaded from http://genome.ucsc.edu/) and those also detected in tumor samples. Variants from sequencing artifacts were removed by VarScan fpfilter. The following criteria were used to call variations: average base quality ≥ 20, average mapping quality ≥ 20, depth ≥ 6, and variant allele frequency in counterpart sample ≥ 0.20 and ≤ 0.05 in tumor sample. The called variations were validated by polymerase chain reaction (PCR) and Sanger capillary sequencing using genomic DNAs of the patients and primers listed in Table S1. Mutations in cancer genome were also analyzed by the GATK using the above criteria.

Analyses of TCGA LUAD genome data

The TCGA genome data of 513 LUADs (Table 1) were downloaded from the Cancer Genomics Hub (CGHub) (https://cghub.ucsc.edu/) with approval by the National Institutes of Health (NIH; approval number #24437-4), and analyzed by GATK using the criteria for CNC variants.

Statistics

All of the values were evaluated using the SPSS 17.0 software for Windows (SPSS Inc., Chicago, Illinois). Statistically significant differences were determined by Student's *t*-test, and the survival curves were plotted according to the Kaplan-Meier method and compared by the log-rank test. *P* values less than 0.05 were considered statistically significant in all cases.

References

1. Lawrence MS, *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).

2. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11**, 685-696 (2010).

3. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA: A Cancer J Clin* **65**, 87-108 (2015).

4. Herbst RS, Heymach JV, Lippman SM. Lung Cancer. *N Engl J Med* **359**, 1367-1380 (2008).

5. Hecht SS. Lung carcinogenesis by tobacco smoke. *Int J Cancer* **131**, 2724-2732 (2012).

6. Thun MJ, Lally CA, Calle EE, Heath CW, Flannery JT, Flanders WD. Cigarette Smoking and Changes in the Histopathology of Lung Cancer. *J Natl Cancer Inst* **89**, 1580-1586 (1997).

7. Yu XJ, *et al.* Characterization of Somatic Mutations in Air Pollution-Related Lung Cancer. *EBioMedicine* **2**, 583-590 (2015).

8. Slaughter DP, Southwick HW, Smejkal W. "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer* **6**, 963-968 (1953).

9. Auerbach O, Hammond EC, Kirman D, Garfinkel L. Effects of cigarette smoking on dogs. II. Pulmonary neoplasms. *Arch Environ Health* **21**, 754-768 (1970).

10. Gomperts BN, *et al.* Evolving Concepts in Lung Carcinogenesis. *Semin Respir Crit Care Med* **32**, 032-043 (2011).

11. Kadara H, *et al.* 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* **6**, 8-17 (2013).

12. Ooi AT, *et al.* Molecular Profiling of Premalignant Lesions in Lung Squamous Cell Carcinomas Identifies Mechanisms Involved in Stepwise Carcinogenesis. *Cancer Prev Res* **7**, 487-495 (2014).

13. Network TCGAR. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550 (2014).

14. Imielinski M, *et al.* Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. *Cell* **150**, 1107-1120 (2012).

15. Olivier M, *et al.* Modelling mutational landscapes of human cancers in vitro. *Sci Rep* **4**, 4482 (2014).

16. Govindan R, *et al.* Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. *Cell* **150**, 1121-1134 (2012).

17. Kargı A, Dinç ZA, Başok O, Üçvet A. MUC4 expression and its relation to ErbB2 expression, apoptosis, proliferation, differentiation, and tumor stage in non-small cell lung cancer (NSCLC). *Pathol Res Pract* **202**, 577-583 (2006).

18. Bosch N, *et al.* Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. *Hum Mol Genet* **16**, 2572-2582 (2007).

19. Franco B, *et al.* A cluster of sulfatase genes on Xp22.3: Mutations in chondrodysplasia punctata (CDPX) and implications for warfarin embryopathy. *Cell* **81**, 15-25.

20. Shyjan AM, Bertagnolli M, Kenney CJ, Briskin MJ. Human mucosal addressin cell adhesion molecule-1 (MAdCAM-1) demonstrates structural and functional similarities to the alpha 4 beta 7-integrin binding domains of murine MAdCAM-1, but extreme divergence of mucin-like sequences. *J Immunol* **156**, 2851-2857 (1996).

21. Grelg GM, Sharp CB, Carrel L, Willard HF. Duplicated zinc finger protein genes on the proximal short arm of the human X chromosome: isolation, characterization and X-inactivation studies. *Hum Mol Genet* **2**, 1611-1618 (1993).

22. Wolins NE, Skinner JR, Schoenfish MJ, Tzekov A, Bensch KG, Bickel PE. Adipocyte Protein S3-12 Coats Nascent Lipid Droplets. *J Biol Chem* **278**, 37713-37721 (2003).

23. Boyadjiev SA, *et al.* A reciprocal translocation 46,XY,t(8;9)(p11.2;q13) in a bladder exstrophy patient disrupts CNTNAP3 and presents evidence of a pericentromeric duplication on chromosome 9. *Genomics* **85**, 622-629 (2005).

24. Nagase T, Kikuno R, Ishikawa K, Hirosawa M, Ohara O. Prediction of the coding sequences of unidentified human genes. XVII. The complete sequences of 100 new cDNA clones from brain which code for large proteins in vitro. *DNA Res* **7**, 143-150 (2000).

25. Ge Q, Nilasena DS, O'Brien CA, Frank MB, Targoff IN. Molecular analysis of a major antigenic region of the 240-kD protein of Mi-2 autoantigen. *J Clin Invest* **96**, 1730-1737 (1995).

26. Yahagi S, *et al.* Identification of two novel clusters of ultrahigh-sulfur keratin-associated protein genes on human chromosome 11. *Biochem Biophys Res Commun* **318**, 655-664 (2004).

27. Yu XJ, *et al.* Characterization of somatic mutations in air pollution-related lung cancer. *EBioMedicine* **2**, 583-590 (2015).

28. Huang RJ, *et al.* High secondary aerosol contribution to particulate pollution during haze events in China. *Nature* **514**, 218-222 (2014).

29. Sugimura T, Nagao M, Okada Y. Carcinogenic action of N-methyl-N'-nitro-N-nitrosoguanidine. *Nature* **210**, 962-963 (1966).

30. Dollé MET, Snyder WK, Dunson DB, Vijg J. Mutational fingerprints of aging. *Nucleic Acids Research* **30**, 545-549 (2002).

31. Yu H, *et al.* Integrative genomic and transcriptomic analysis for pinpointing recurrent alterations of plant homeodomain genes and their clinical significance in breast cancer. *Oncotarget* **8**, 13099-13115 (2016).

32. Berens EB, *et al.* Keratin-associated protein 5-5 controls cytoskeletal function and cancer cell vascular invasion. *Oncogene* **36**, 593-605 (2017).

33. McKenna A, *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

34. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).

Acknowledgements

The results shown here are in part based upon data generated by The Cancer Genome Atlas 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.

Funding

This work was supported by the National Natural Science Funds for Distinguished Young Scholar (81425025), the National Key Research and Development Program of China (2016YFC0905500), the National Natural Science Foundation of China (81672765), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA12010307), and grants from the State Key Laboratory of Membrane Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

The project was conceived and designed by GBZ. The experiments were conducted by BZ, LWQ. The genome data were analyzed by LWQ, BZ, GZW, and YC. The manuscript was written by GBZ.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

SUPPLEMENTARY INFORMATION

Supplementary information includes 3 figures and 5 tables.

Table 1. Genomic variations in CNCs of 11 patients with LUAD.

Pt #	Gene	Chr	Start	End	Ref	Alt	Function	Transcript	cDNA position	Amino acid
794	*ATG16L2*	11	72533157	72533157	A	T	NS	NM_033388	c.A461T	p.E154V
	CEP250	20	34092235	34092235	G	A	NS	NM_007186	c.G6038A	p.R2013Q
	MYO1C	17	1371330	1371330	C	A	NS	NM_033375	c.G2743T	p.V915L
	PNPLA2	11	824813	824813	G	T	NS	NM_020376	c.G1466T	p.S489I
	TMC6	17	76121923	76121923	C	T	NS	NM_007267	c.C314T	p.T105M
	ZNF521	18	22806878	22806878	C	T	NS	NM_015461	c.C1004T	p.P335L
824	*ANKS1A*	6	34857302	34857302	-	GGCGGC	nonframeshift	NM_015245	c.123_124insGGCGGC	p.G41delinsGGG
	BBS12	4	123664726	123664728	AAG	-	nonframeshift	NM_001178007	c.1679_1681del	p.560_561del
	C9orf66	9	214943	214943	C	T	NS	NM_152569	c.C454T	p.R152W
	FAM186A	12	50745703	50745703	T	G	NS	NM_001145475	c.A4912C	p.T1638P
	HIST1H1D	6	26235127	26235127	C	A	NS	NM_005320	c.C35A	p.P12H
	HLA-DRB1	6	32551957	32551957	G	T	NS	NM_002124	c.C299A	p.A100E
	UNC93A	6	167728776	167728776	GT	AC	NS	NM_018974	c.T1210C	p.F404L
760	*DMKN*	19	36002386	36002386	C	T	NS	NM_001126057	c.G845A	p.S282N
772	*ZNF259*	11	116658654	116658654	G	A	NS	NM_003904	c.C53T	p.P18L
783	*SRA1*	5	139931629	139931629	-	G	frameshift	NM_001035235	c.327_328insC	p.V110fs
	TPSAB1	16	1291454	1291454	G	A	NS	NM_003294	c.G253A	p.A85T
	TPSD1	16	1306971	1306971	A	G	NS	NM_012217	c.A428G	p.H143R
792	*GPC3*	X	133087197	133087197	A	T	NS	NM_001164617	c.T217A	p.C73S
710	*TPTE2*	13	20048214	20048214	T	G	NS	NM_130785	c.A121C	p.I41L

799	*CD33*	19	51729578	51729596	CCCAACAA CTGGTATCT TT	-	frameshift	NM_001772	c.711_729del	p.N237fs
748	*PCMTD1*	8	52733228	52733228	G	A	NS	NM_052937	c.C757T	p.R253C
	RNASE8	14	21526082	21526084	CTG	-	nonframeshift	NM_138331	c.31_33del	p.11_11del
828	*HGC6.3*	6	168376951	168376951	T	C	NS	NM_001129895	c.A382G	p.M128V
	SYNJ1	21	34003931	34003931	C	T	NS	NM_003895	c.G4213A	p.V1405I
844	*MADCAM1*	19	501762	501762	A	C	NS	NM_130760	c.A761C	p.Q254P
	NEFH	22	29885581	29885604	AGGCCAAG TCCCCAGAG AAGGAAG	-	nonframeshift	NM_021076	c.1952_1975del	p.651_659del

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

Table 2. Genomic variations in CNCs and tumor samples of TCGA LUADs.

	Exonic mutations/MB	Variant genes/sample	Nonsynonymous alterations/sample	Synonymous mutations/sample	Rearrangements/sample	
					frameshift	Inframe
CNC	0.2721	5.2885	4.3723	2.8402	0.2378	0.8129
Tumor	6.0145	141.0132	135.7290	43.6491	6.6862	1.9357
P value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Figures and Legends

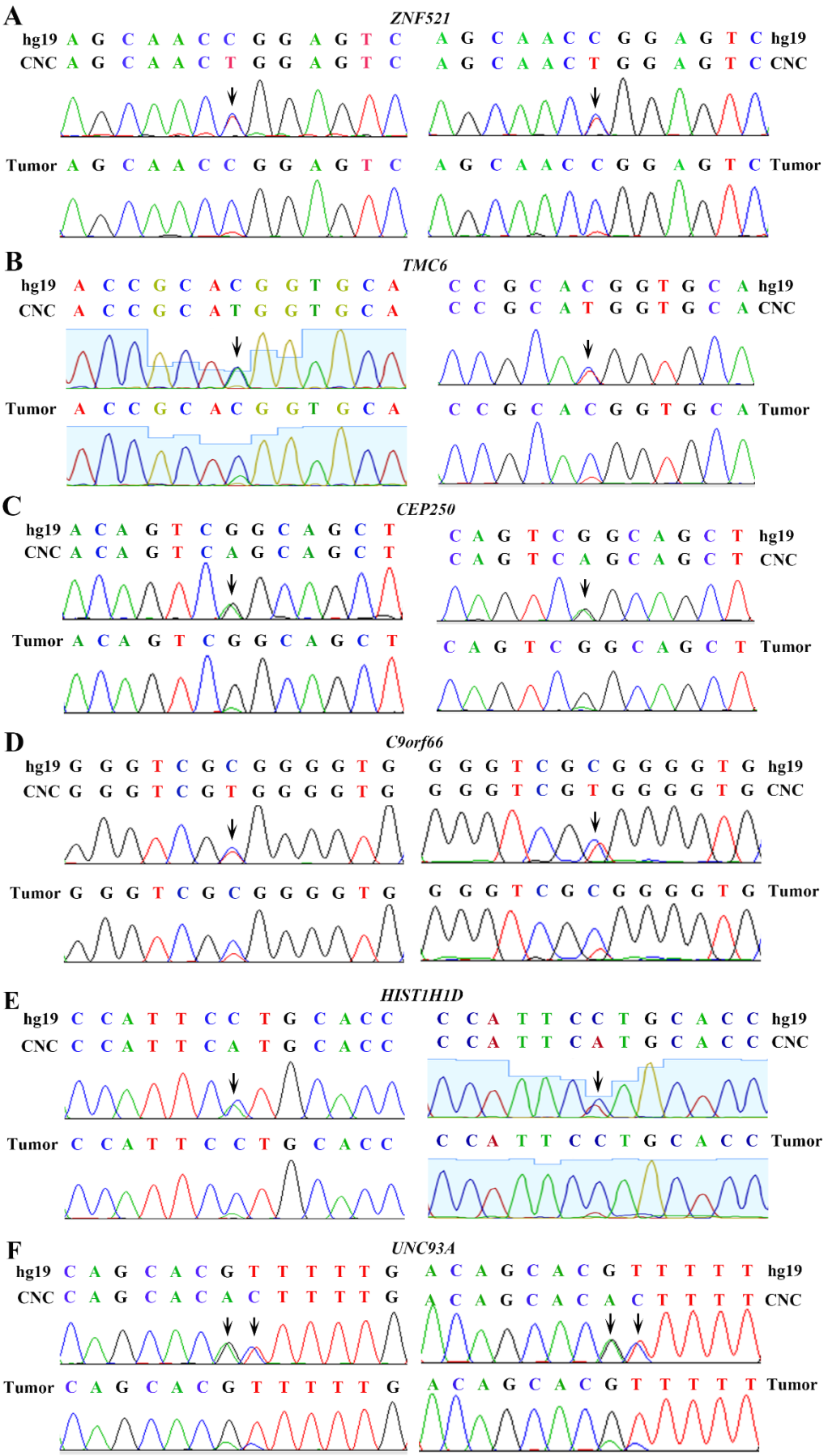

Figure 1. Validation of genomic variations in "normal" lung tissues. Polymerase chain reaction (PCR) and Sanger capillary sequencing were performed using genomic DNAs of the patients and primers listed in Table S1. (A) *ZNF521* in "normal" lung and tumor samples of the patient. Two sets of primers were used. (B) *TMC6* in "normal" lung and tumor samples of the patient. Two sets of primers were used. (C) *CEP250* in "normal" lung and tumor samples of the patient. Two sets of primers were used. (D) *C9orf66* in "normal" lung and tumor samples of the patient. (E) *HIST1H1D* in "normal" lung and tumor samples of the patient. (F) *UNC93A* in "normal" lung and tumor samples of the patient.

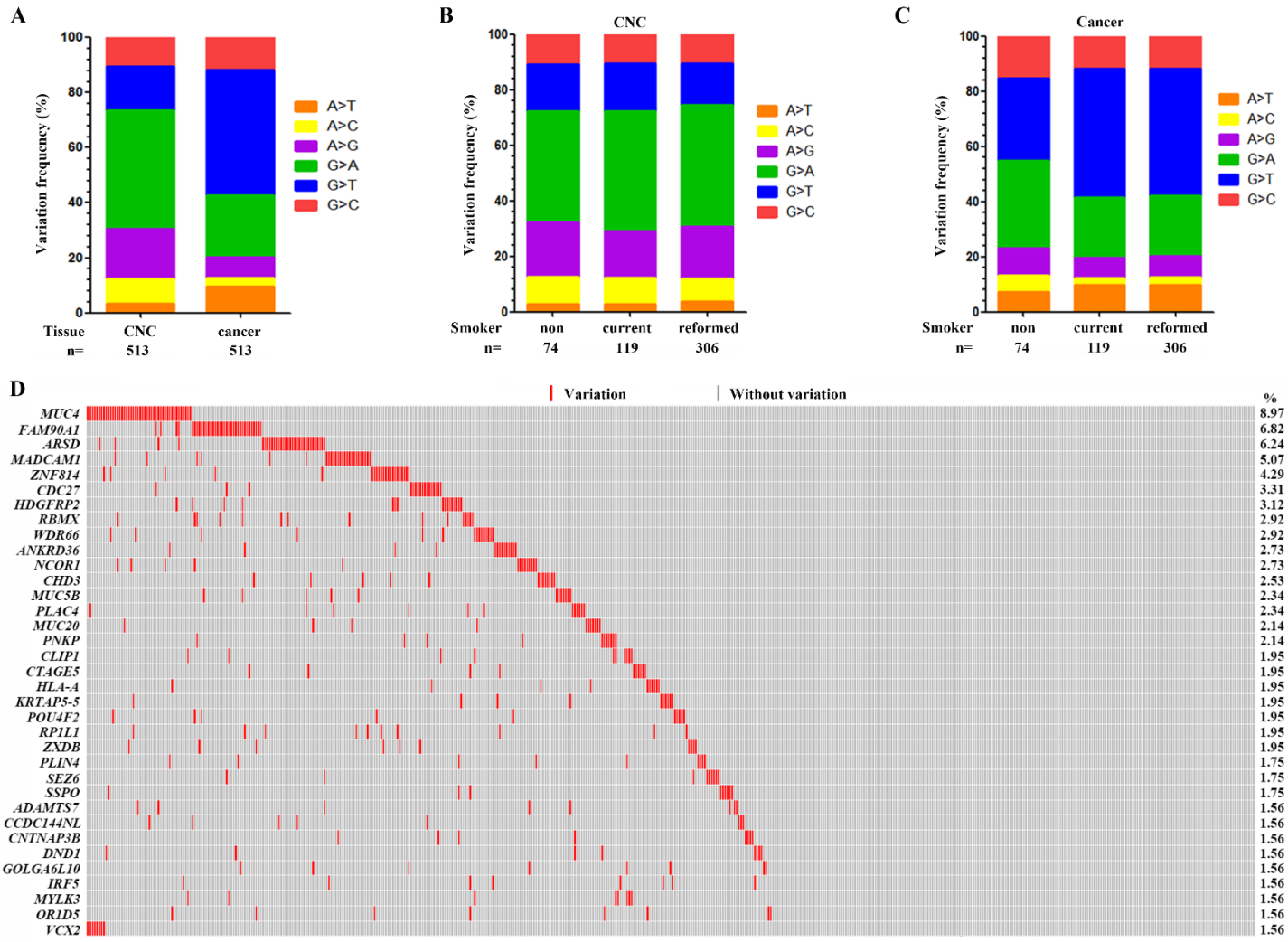

Figure 2. Genomic variations in the CNC samples. (A) The frequency of each type of single base substitution in CNCs and tumor samples of the 513 patients with LUAD. (B) The proportion of nucleotide changes in CNC samples of nonsmokers, current smokers and reformed smokers. (C) The frequency of base substitution in tumor samples of nonsmokers, current smokers and reformed smokers. (D) Significantly altered genes in CNCs of the patients. CNC samples are arranged from left to right in the top track.

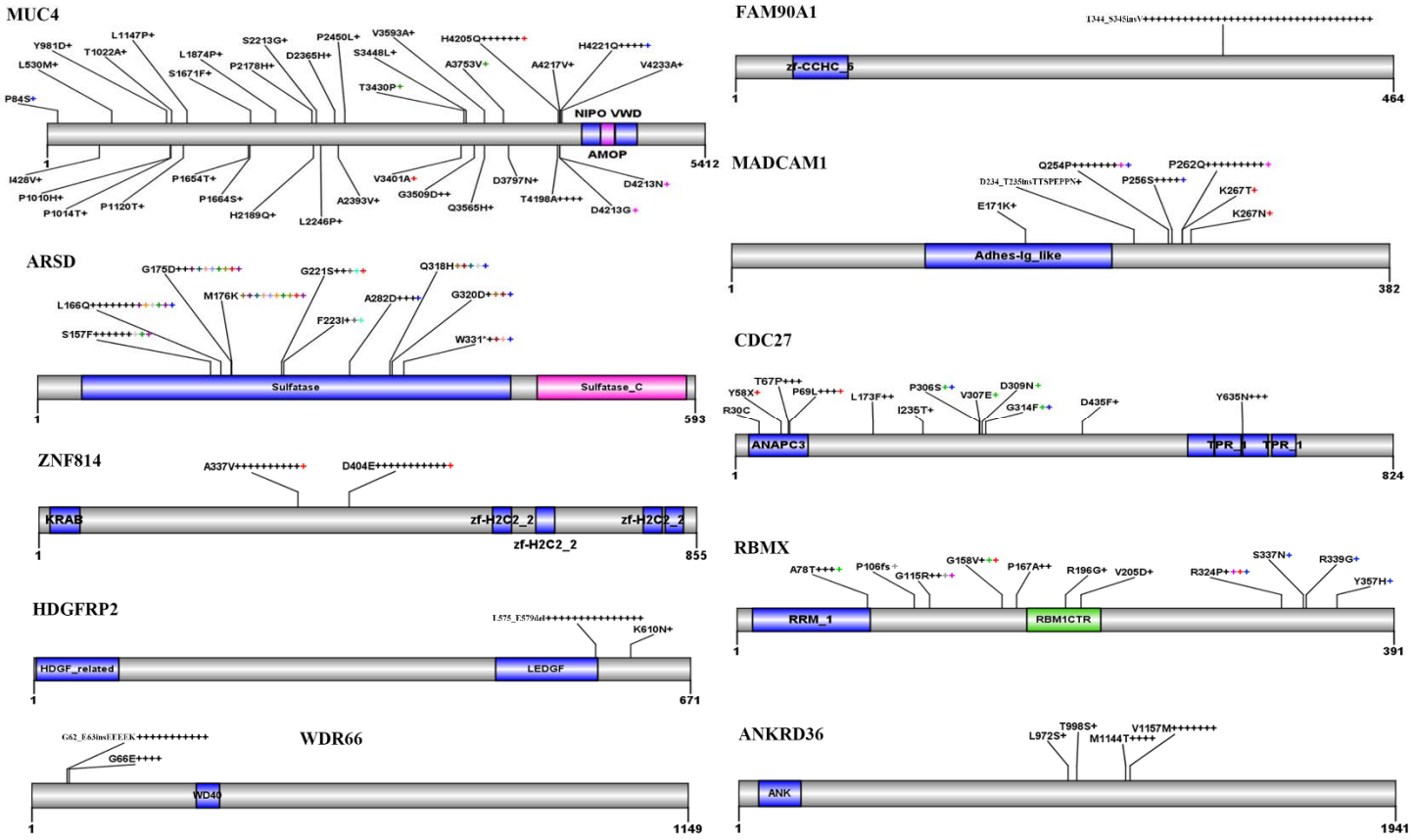

Figure 3. Recurrent variations in some representative genes in CNCs. Schematic representations of proteins encoded by the genes are shown. Numbers refer to amino acid residues. Each "+" corresponds to an independent, altered CNC sample. Variations within a gene indicated by a same-colored "+" are found in the same patient.

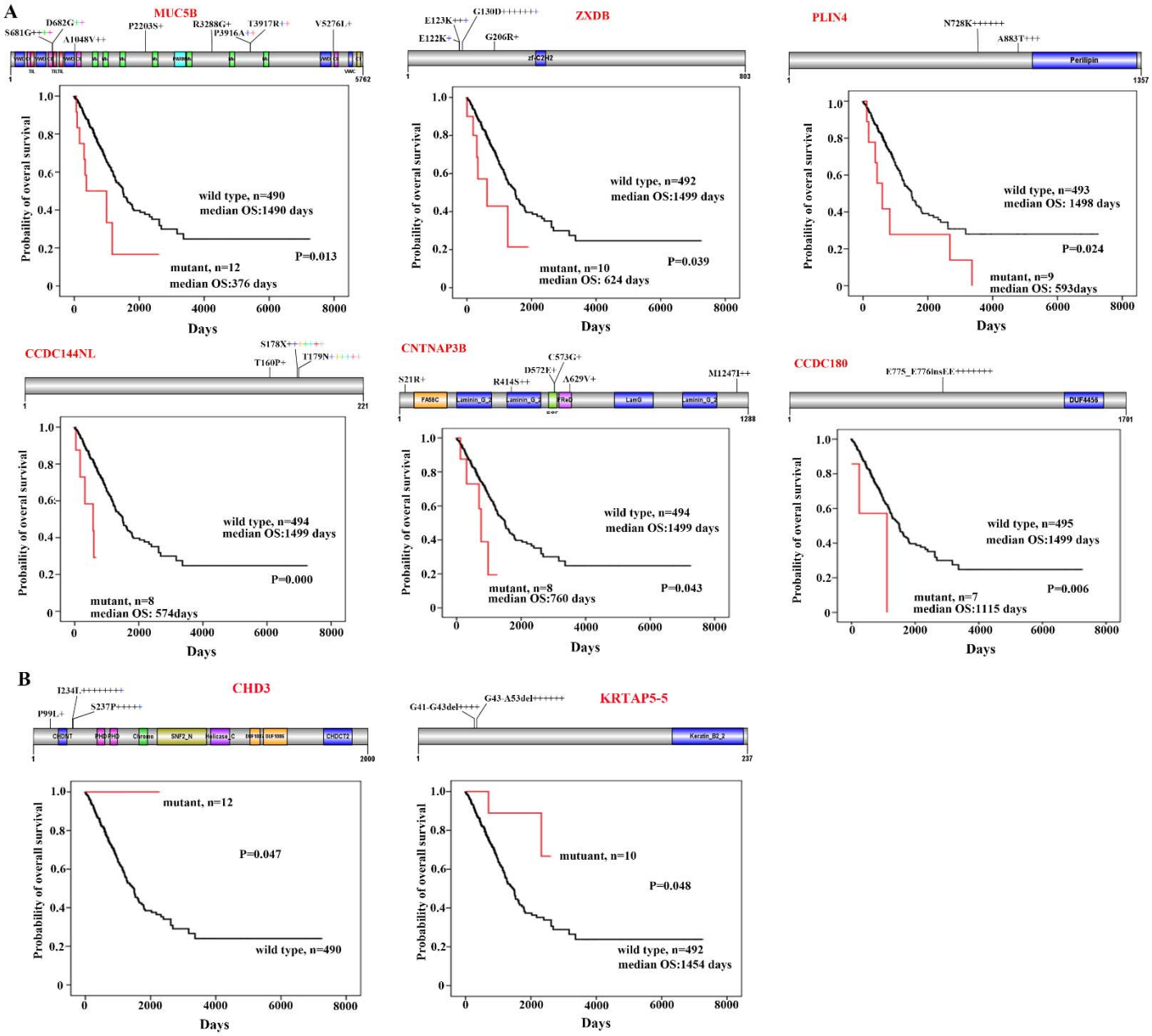

Figure 4. CNC variations associated with prognosis of the patients. (A) Variations of MUC5B, ZXDB, PLIN4, CCDC144NL, CNTNAP3B, and CCDC180 in CNCs and Kaplan–Meier curve for overall survival of the patients with wild type or variant transcripts. Schematic representations of proteins encoded by the genes are shown. (B) Overall survival of patients harboring wild type or variant CHD3 or KRTAP5-5 in their CNCs.