Exome sequencing identifies MXRA5 as a novel cancer gene frequently mutated in non–small cell lung carcinoma from Chinese patients

Donghai Xiong†, Guangming Li†‡, Kezhen Li†‡, Qinzi Xu†‡, Zhongjie Pan†, Peng Ding†, Peter Vedell†, Pengyuan Liu†, Peng Cui†, Xing Hua†, Hui Jiang†, Yuxin Yin†, Ze Zhu†, Xiaomian Li†, Bin Zhang†, Ding Ma‡§, Yian Wang†‡ and Ming You†‡*

1Department of Pharmacology and Toxicology and the Cancer Center, Medical College of Wisconsin, Milwaukee, WI 53226, USA, 2Basic Medical School, Tianjin Medical University, Tianjin, China, 3Department of Surgery, Washington University in St. Louis, St. Louis, MO 63110, USA, 4Cancer Biology Research Center, 5Department of Thoracic Surgery, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China, 6Vascular Department, Tianjin Union Medical Centre, Tianjin, China; and 7Department of Physiology and the Cancer Center, Medical College of Wisconsin, Milwaukee, WI 53226, USA.

Abbreviations: (ADCs), adenocarcinomas; (BAC), bronchioloalveolar carcinoma; (COSMIC), Catalogue of Somatic Mutations in Cancer; (ECM), extracellular matrix; (indels), insertions and deletions; (MXRA5s), matrix-remodeling associated 5; (ncRNA), noncoding RNA; (NS/S), nonsilent-to-silent; (NSCLCs), non–small cell lung carcinomas; (SCLC), small cell lung carcinoma; (SCCs), squamous cell carcinomas; (UTR), untranslated regions.

Introduction

Lung cancer has become the top killer among malignant tumors in China and around the world. In China, approximately 300,000 patients with lung cancer and more than 250,000 deaths from the disease are predicted each year (1). Lung cancer consists of two major types, non–small cell lung carcinoma (NSCLC) and small-cell lung carcinoma (SCLC). NSCLC is the most common type and accounts for approximately 85% of lung cancer. NSCLCs usually grow and spread more slowly than SCLCs but are relatively insensitive for approximately 85% of lung cancer. NSCLCs usually grow and spread more slowly than SCLCs but are relatively insensitive.

DNA library preparation

Paired-end libraries were prepared following the manufacturer’s protocols (Illumina and Agilent). Briefly, 3 µg of genomic DNA was fragmented to 150–200 bp using the Covaris E220 sonicator. The ends were repaired, and an “A” base was added to the 3′ ends. Paired-end DNA adaptors (Illumina) with a single “T” base overhang at the 3′ end were ligated, and the resulting constructs were purified using AMPure beads from Agencourt. The adapter-modified DNA fragments were enriched by four cycles of PCR using PE 1.0 forward and PE 2.0 reverse (Illumina) primers. The concentration and the size distribution of the libraries were determined on an Agilent Bioanalyzer DNA 1000 chip.

Massively parallel sequencing

Sequencing was carried out for the captured libraries with the HiSeq 2000 using 100 bp paired-end reads. Libraries were loaded onto paired-end flow cells at concentrations of 4–5 pM (HiSeq 2000) to generate cluster densities of 300 000–500 000/mm² (HiSeq 2000) following Illumina’s standard protocol using the Illumina cluster station and the Illumina cBot and HiSeq
Table I. Distribution of somatic SNVs learned from the exome sequencing of 14 Chinese lung cancer patients

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Tumor type</th>
<th>Position</th>
<th>Ref</th>
<th>Var</th>
<th>Region</th>
<th>Function</th>
<th>Seq_change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1250660</td>
<td>ADC</td>
<td>3242284</td>
<td>G</td>
<td>T</td>
<td>Exonic</td>
<td>Non-synonymous</td>
<td>NM_015419:p.P1491A</td>
</tr>
<tr>
<td>1270737</td>
<td>ADC</td>
<td>3229435</td>
<td>C</td>
<td>T</td>
<td>Exonic</td>
<td>Non-synonymous</td>
<td>NM_015419:p.P1491A</td>
</tr>
<tr>
<td>1273036</td>
<td>ADC</td>
<td>3241141</td>
<td>C</td>
<td>T</td>
<td>Exonic</td>
<td>Non-synonymous</td>
<td>NM_015419:p.P1491A</td>
</tr>
<tr>
<td>1273036</td>
<td>ADC</td>
<td>3227666</td>
<td>AT</td>
<td>TA</td>
<td>3'-UTR</td>
<td>Messenger RNA stability</td>
<td></td>
</tr>
</tbody>
</table>

*Sample 1269558 is from a male nonsmoker, sample 1273036 is from a female smoker, and all the others are from male smokers.

Table II. MXRA5 (on chromosome X) mutations detected in exome sequencing and screening samples

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Tumor type</th>
<th>Position</th>
<th>Ref</th>
<th>Var</th>
<th>Region</th>
<th>Function</th>
<th>Seq_change</th>
</tr>
</thead>
<tbody>
<tr>
<td>259091</td>
<td>ADC</td>
<td>3240644</td>
<td>C</td>
<td>A</td>
<td>Exonic</td>
<td>Non-synonymous</td>
<td>NM_015419:p.V1028F</td>
</tr>
<tr>
<td>259610</td>
<td>SCC</td>
<td>324710</td>
<td>G</td>
<td>T</td>
<td>Exonic</td>
<td>Non-synonymous</td>
<td>NM_015419:p.A981D</td>
</tr>
<tr>
<td>260211</td>
<td>SCC</td>
<td>3329255</td>
<td>G</td>
<td>C</td>
<td>Exonic</td>
<td>Non-synonymous</td>
<td>NM_015419:p.P1491A</td>
</tr>
<tr>
<td>345480</td>
<td>BAC</td>
<td>3242284</td>
<td>G</td>
<td>T</td>
<td>Exonic</td>
<td>Non-synonymous</td>
<td>NM_015419:p.P4181H</td>
</tr>
</tbody>
</table>

Downloaded from https://academic.oup.com/carcin/article-abstract/33/9/1797/2463747 by guest on 05 March 2019
Results

Overview of somatic mutation profiles

In total, we identified 3321 high-confidence somatic mutations in the 14 Chinese tumor-normal pairs sequenced (Table I, Supplementary Tables 3 and 4). These 3321 mutations are distributed in various genomic regions, including exonic (48.7%), splicing (1.2%), non-coding RNA (ncRNA; 10.2%), 5′/3′-untranslated regions (UTR) (2.6%), intronic (22.8%), upstream/downstream (0.8%), and intergenic regions (13.7%) (Table I). There are 1659 somatic mutations of potential biological significance as they reside in protein coding regions or splicing sites (Supplementary Table 7). Among them, 447 are silent (synonymous) substitutions, 1038 are missense substitutions causing amino acid changes, 88 are non-sense mutations leading to truncated proteins, 2 are stop-loss mutations causing abnormal protein extension, and 41 SNVs occurred at splicing sites. We also found 39 indels causing frameshift mutations and 4 indels that did not alter reading frame, ranging from 1 to 6 bp in length. Non-sense, frameshift indels and splicing mutations generally lead to the inactivation of the protein products. To evaluate missense mutations, we used four algorithms to make a consensus prediction to identify putative driver mutations. These functional prediction algorithms are based on phylogenetics, structural biology, bioinformatics, or population genetics. Of 1038 missense mutations, 91 (8.8%) were predicted to be deleterious by all four algorithms and 216 (20.8%) were predicted by three or four algorithms (Supplementary Table 8).

The somatic mutations occurring in exonic regions or splicing sites are distributed in 1454 genes. A comparison of the 1454 mutated genes with the Catalogue of Somatic Mutations in Cancer (COSMIC) database revealed that 1015 of these genes have been previously archived in COSMIC database to be somatically mutated in human tumors.

We observed, on average, approximately 115 somatic SNVs in exonic regions per Chinese exome with the most frequently altered exome having 493 SNVs. No somatic mutations in exonic regions were identified in two BAC samples. On average, SCC had 243 somatic exonic SNVs, ADC had 111 somatic exonic SNVs, and BAC had only 11 somatic SNVs in exonic regions (Supplementary Figures 1 and 2). The observed ratio of non-synonymous to synonymous change (dNd/s) did not show significant difference across different types of lung tumors (ADC versus BAC versus SCC: 3.3 versus 2 versus 1.9; P > 0.12) (Supplementary Figure 3). Among 3321 SNVs identified by exome sequencing, somatic variants occurred predominantly at G:C base pairs (67.5% in ADC, 70.1% in SCC, and 59.8% in BAC). G:C > T:A and G:C > A:T transversions are the two most prevalent categories in ADC and SCC (32.5% and 70.1% in SCC, and 59.8% in BAC). G:C > T:A and G:C > A:T transversions occurred predominantly at G:C base pairs (67.5% in ADC, Figure 3). Among 3321 SNVs identified by exome sequencing, somatic mutations occurring in exonic or splicing sites are distributed in 1454 genes.

The KRAS mutation we identified was a missense mutation in a lung ADC sample. We also confirmed two novel somatic mutations in a lung ADC sample (Supplementary Table 5). Three tumors are SCC, two are ADC, and two are BAC. There are two non-sense mutations causing truncated TP53 protein in one ADC and one BAC tumor (Supplementary Table 5). All of these seven TP53 somatic mutations have been reported in COSMIC, with six (genomic position at chr17: 7576885, 7577120, 7577538, 7577548, 7578455, and 7578534—hg19/GRCh37 coordinates) found in 143 lung tumors (mainly of NSCLC type) according to COSMIC (Supplementary Table 6). EGFR had a five amino acid non-frameshift deletion in one BAC sample at chr7: 55242464. In COSMIC, there is a one amino acid deletion (p.E746del) at this position in a lung ADC sample. We also confirmed two novel somatic mutations in ncRNA coding regions within EGFR, which had not been reported before. We confirmed these two novel somatic mutations in a patient with BAC at chr12:25398284, reported by COSMIC to be existent in 781 lung tumors (most are NSCLC; Supplementary Table 6). Two

| Table III. Several most significant pathways for Chinese NSCLC and their associated genes |
|-----------------------------------------------|------------------|------------------------------------------|--------------------------------|-----------------------|
| Gene | ECM remodeling | ECM remodeling | Vogelstein Core Cancer Pathways | Other Cancer-Related Pathways | Downstream Pathways |
|-----------------------------------------------|------------------|------------------------------------------|--------------------------------|-----------------------|
| MXRA5 | FN1 | Yes | Yes | Integrins, TGF-β | Integrins, MAPK, Notch, TGF-β, | α5/β1 integrin |
| NID1 | LAMA3 | Yes | Yes | Cell adhesion, integrins | Cell adhesion, integrins |
| MMP16 | COL4A6 | Yes | Yes | Invasion | Invasion |
| CDH13 | NRXN3 | Yes | Yes | G1/S phase transition, cell adhesion, small GTPases | Cell adhesion (Kyoto Encyclopedia of Genes and Genomes) |
| COL4A2 | VCAN | Yes | Integrins | Integrins, antiapoptosis, cytoskeleton remodeling |
| LAMA1 | | | | | EMT, cell adhesion |

Somatic variants validation

To evaluate false-positive rates of the identified somatic mutations, we selected a subset of 106 mutations out of our high-confidence somatic mutation list for Sanger sequencing validation. The selected mutations are distributed in two categories of genes: (i) five previously known lung cancer driver genes (TP53, EGFR, KRAS, PIK3CA, and ROS1) that have non-silent somatic mutations or mutations located in ncRNA coding regions; and (b) some novel genes that are predicted to have non-silent somatic mutations in at least 2 tumors in our Chinese cohort. Novel means that these genes remain seldom studied. We validated 98 somatic mutations out of the initial 106 variants, indicating that the false-positive rate of our high-confidence somatic mutations is only approximately 7.5%.

TP53 was the most frequently mutated one, with seven tumor samples (50%) harboring non-silent TP53 mutations (two non-sense and five missense mutations; Supplementary Table 5). Three tumors are SCC, two are ADC, and two are BAC. There are two non-sense mutations causing truncated TP53 protein in one ADC and one BAC tumor (Supplementary Table 5). All of these seven TP53 somatic mutations have been reported in COSMIC, with six (genomic position at chr17: 7576885, 7577120, 7577538, 7577548, 7578455, and 7578534—hg19/GRCh37 coordinates) found in 143 lung tumors (mainly of NSCLC type) according to COSMIC (Supplementary Table 6). EGFR had a five amino acid non-frameshift deletion in one BAC sample at chr7: 55242464. In COSMIC, there is a one amino acid deletion (p.E746del) at this position in a lung ADC sample. We also confirmed two novel somatic mutations in ncRNA coding regions within EGFR, which had not been reported before. The KRAS mutation we identified was a missense mutation in a patient with BAC at chr12:25398284, reported by COSMIC to be existent in 781 lung tumors (most are NSCLC; Supplementary Table 6). Two
missense PIK3CA mutations were confirmed in an ADC sample of ours, both of which have been reported by COSMIC as existent in multiple types of tumors. The seven novel somatic mutations in ncRNA coding regions within ROS1 in an SCC sample of ours were validated as well.

As for novel genes, Sanger sequencing validated that 19 genes were mutated in at least 2 of 14 samples, with a mutation rate of >14%, which are as follows: MXRA5, STAG2, ZFHX4, HMCN1, ABCA12, CSMD3, EIF4G3, EVI1/MECOM, FCRL4, FN1, PCSK5, SLC8A1, TSHZ3, KIAA1109, ZBTB41, MUC16, PKHD1L1, USH2A, and XRN1 (Supplementary Table 5). Except for TP53, MXRA5 was the most frequently mutated gene in our sample set, which had four missense mutations in four tumors including one ADC, two SCC and one BAC, with mutation frequency being 28.6% (4 of 14). The genes mutated in three tumors were ZFHX4 and PKHD1L1. FCRL4 was the only gene with somatic missense mutations in two BAC samples (Supplementary Table 5).

**Sanger screening for MXRA5 somatic mutations**

Because that MXRA5 was seldom studied before and the second most frequently mutated gene in our initial exome sequencing sample, we want to further examine this gene and determine its somatic mutation frequency using larger sample set. Therefore, we performed Sanger sequencing of all the exons and UTRs of this gene using a validation cohort composed of tumor/normal paired DNA samples extracted from another 52 Chinese patients with NSCLC (Supplementary Table 1). Incorporating the original 14 patients with NSCLC analyzed by exome sequencing, we found that 15% of Chinese NSCLCs (10 of 66) contained 12 somatic mutations in MXRA5. Specifically, MXRA5 had mutations in 7 of 29 ADCs (24%), in 2 of 26 SCCs (8%), and in 1 of 6 BAC (17%) (Table II and Figures 2 and 3). More MXRA5 mutations were seen in smokers, 9 of 48 (18.8%), versus non-smokers, 1 of 16 (6.3%). Furthermore, 9 of 48 (18.8%) men versus 1 of 18 (5.6%) women had mutations in MXRA5. These suggest that smoking and gender may predict somatic MXRA5 mutations in NSCLC, similar to observations on somatic BRAF mutations in colorectal cancer (12).

More MXRA5 mutations were seen in smokers, 9 of 48 (18.8%), versus non-smokers, 1 of 16 (6.3%). Furthermore, 9 of 48 (18.8%) men versus 1 of 18 (5.6%) women had mutations in MXRA5. These suggest that smoking and gender may predict somatic MXRA5 mutations in NSCLC, similar to observations on somatic BRAF mutations in colorectal cancer (12). The 12 somatic MXRA5 mutations that we identified in 10 NSCLCs were not present in the dbSNP135 database (http://www.ncbi.nlm.nih.gov/projects/SNP/), which contains approximately 40 million SNPs (single nucleotide polymorphisms) submitted by the 1000 Genomes Project. Similarly,

---

**Fig. 1.** Sanger sequencing chromatograms of somatic MXRA5 mutations in Chinese patients with NSCLC. Mutations are shown in the indicated tumor compared with matched normal DNA and are marked by arrows.

---

| Tumor CTGGGANACCTGG | Tumor CTAAANCTCCAT | Tumor ATGCTGNCTCCTT |

N: the presence of both C and A allele for the p.V1028F alteration

| Tumor CCACTANGCTCAA | Tumor TGTTCCAGATGGT | Tumor GCATGCNTCTGGA |

N: the presence of both G and T allele for the p.P481H alteration

A: the change from G to A allele for the p.R2349W alteration

N: the presence of both A and G allele (on reverse strand) for the p. H2678R alteration
MXRA5 (p.R2300H)
Patient 1270737
Normal
CCAAGGCTATGT

Tumor
CCAAGCNCTATGT

N: the presence of both G and A allele (on reverse strand) for the p. R2300H alteration

MXRA5 (p.A2763G)
Patient 1270737
Normal
CCAAAGCTGACAT

Tumor
CCAAAGNTGACAT

N: the presence of both C and G allele (on reverse strand) for the p.A2763G alteration

MXRA5 (p.E2716A)
Patient 1270737
Normal
AGACGGAAGTACGG

Tumor
AGACGGNGTACGG

N: the presence of both A and C allele (on reverse strand) for the p.E2716A alteration

MXRA5 (p.S862N)
Patient 1273036
Normal
CAGCCAGCATGGG

Tumor
CAGCCANCATGGG

N: the presence of both G and A allele (on reverse strand) for the p.S862N alteration

MXRA5 (p.W611C)
Patient 1275369
Normal
TAGCTGGATTCTT

Tumor
TAGCTGNATTCTT

N: the presence of both G and T allele (on reverse strand) for the p.W611C alteration

MXRA5 (in 3'-UTR)
Patient 1250660
Normal
AAATAATGTGTCAG

Tumor
AAATANNTGTCAG

N: the presence of both AT (wild-type) and TA (mutant) alleles for the alteration in 3'-UTR

Fig. 1. Continued

Fig. 2. Schematic drawing of MXRA5 gene structure. MXRA5 encoded protein consists of 2824 amino acids. The conserved domains and 11 missense somatic mutations validated in our 9 Chinese NSCLC tumor samples were drawn with MXRA5 structure. Seven missense mutations labeled in red are located in the conserved domains of MXRA5.
these mutations were not any of the SNVs of MXRA5 identified by the exome sequencing project of the National Heart, Lung and Blood Institute, which sequenced 5400 exomes (http://evs.gs.washington.edu/EVS/). Therefore, our reported somatic mutation loci in the MXRA5 gene were not due to germ line polymorphisms or germ line mutations.

**Pathway and gene set analysis**

Table III and Supplementary Table 9 shows the most significantly mutated pathways we have identified. The ratios of non-silent to silent somatic (NS/S) mutations across the 10 individuals with at least one non-silent somatic mutations are significantly higher than expected ($P \leq 0.05$) in several pathways, including a DNA damage pathway (“GeneGo role of SUMO in p53 regulation,” 8:0), apoptosis-related pathways (“GeneGo antiapoptosis mediated by external signals” and “GeneGo apoptosis stimulation by external signals,” 20:3; “GeneGo role of CDK5 in neuronal death and survival,” 7:0), and AKT signaling (7:0). The “HGNC major cadherins” (9:0) and “HGNC olfactory receptors” (46:9) gene families also showed high NS/S ratios. In addition, several cancer-specific pathways and gene sets had significantly high NS/S ratios (“Stransky 2011 Table S8,” “Liu/You ADC/SCC Mutated,” “Tissue factor signaling in Lung Cancer,” “Cancer Census”).

A second approach for gene set enrichment was used, which measures the presence or absence of at least one alteration for each gene set by sample. Among gene sets with at least two different genes

---

**Fig. 3.** The GeneGo “Cell Adhesion_ECM remodeling” pathway map. The legend for the symbols of this figure is standard for GeneGo pathway maps and is available at http://www.GeneGo.com. Note that additionally, the golden hexagons with the letter G indicate network objects (e.g. proteins) associated with the genes that are mutated in our samples.
mutated in our samples, the two with the greatest statistical significance ($P < 5\times10^{-10}$) in this measure were associated with extracellular matrix (ECM) remodeling (11 of 81 genes mutated in 9 of 10 samples; $P = 1\times10^{-10}$). The mutated genes in the most significant ECM remodeling gene group include two genes mutated in multiple samples, MXRA5 (4 samples) and FN1 (2 samples), as well as CDH13, COL4A2, COL4A6, LAMA1, LAMA3, MMP16, NID1, NRXN3, and VCAN (Table III).

Discussion

Similar to the findings in Caucasians (2–4), Chinese smokers have much higher mutation rates than never smokers, and G:C $\rightarrow$ T:A transitions (28.4%) were the most frequent somatic mutation pattern observed in Chinese smokers with lung cancer. The numbers of somatic mutations were also significantly different across different NSCLC subtypes in Chinese (SCC > ADC > BAC). There exists wide variation in the number of somatic mutations even within the group of patients having the same NSCLC subtype and smoking status (Supplementary Figure 1), suggesting the high heterogeneity in the genetic causes of NSCLC in Chinese. COSMIC database searches for our identified 1454 genes with functionally important somatic mutations revealed that 30% (439) of them were not present in COSMIC, further highlighting the complexity of lung cancer genomes and the strong carcinogenic effects of tobacco exposure on mutagenesis.

We identified recurrently mutated genes in our Chinese patients with NSCLC. In addition to TP53 that has been well established as a tumor suppressor, MXRA5 was the most striking NSCLC candidate gene with frequent somatic mutations in our sample. MXRA5, also known as adican, encodes an adhesion protoglycan and belongs to a group of genes involved in ECM remodeling and cell–cell adhesion (13). Although the function of MXRA5 in cancer is unknown, somatic mutations in MXRA5 have been identified in tumors obtained from a variety of tissues such as skin, brain, lung, ovary (reported in the COSMIC database; http://www.sanger.ac.uk/genetics/CGP/cosmic/) and parietal pleura (14). By combining initial discovery cohort and subsequent screening cohort, we identified 11 missense mutations that may affect MXRA5 protein structure and 1 mutation in the 3′-UTR that may conceivably affect messenger RNA stability or microRNA interaction. The somatic mutations are distributed throughout the length of the gene. The associations of MXRA5 somatic mutations with lung cancer seem stronger in men because more than 90% (11 of 12) of the mutations were found in tumors from Chinese male patients with NSCLC. This may be partly because men contain a single X chromosome and thus will lose their sole “good” copy of the MXRA5 gene in lung tumors with MXRA5 somatic mutations. Given this and the distribution of mutations throughout the gene, it is reasonable to hypothesize that MXRA5 may act as a tumor suppressor. However, the up-regulation of MXRA5 messenger RNA expression at a modest extent was observed in ovarian and esophageal cancer (15,16). Yet, the increased gene expression of MXRA5 in certain cancer may be merely a consequence rather than a cause or a contributing factor of tumorigenesis. Answers remain elusive for the exact role that the ECM structural protein, MXRA5, could play in tumorigenesis.

Among MXRA5 mutations, 64% (7 of 11) are inactivating mutations that may well contribute to tumor migration and metastasis via ECM remodeling pathway in a similar way to loss-of-function mutations within PTPp and LKB1.

Pathway analysis highlighted the importance of ECM remodeling genes in the genetic etiology of Chinese NSCLC. The ECM remodeling pathway was mutated in 9 of the 10 Chinese patients with lung cancer who have at least one non-silent somatic mutation. This pathway is also enriched with genes associated with the wound healing process. Given that smokers constitute the majority of our Chinese sample, exposure to smoking may play a significant role in the development of NSCLC in Chinese smokers by promoting somatic mutations in genes involved in lung injury healing and cell–cell adhesion processes. Many of the genes of this pathway were previously found to have significantly higher expression in tumor fibroblasts than in normal fibroblasts (16) (MXRA5, FN1, CDH13, COL4A6, and NRXN3). The mutated genes of this pathway also belong to multiple specifically defined core cancer pathways (11), including integrin signaling (COL4A2, COL4A6, FN1, LAMA1, and LAMA3), cell adhesion (MXRA5, CDH13, LAMA1, and LAMA3), invasion (MMP16), small GTPases (CDH13), G1/S phase transition (CDH13), transforming growth factor $\beta$ (TGF-$\beta$) signaling (FN1), and other cancer-associated processes such as tumor metastasis, antiapoptosis, and cytoskeleton remodeling (Table III). FN1, which was mutated in two samples of the present study, belongs to integrins, MAPK, Notch, TGF-$\beta$, and antiapoptosis signaling pathways and is involved in tumor metastasis. COL4A2, COL4A6, and MMP16 are all upstream of $\alpha 1/\beta 1$ integrin signaling pathway, whereas FN1 is upstream of the $\alpha 5/\beta 1$ integrin signaling pathway (Figure 3). Interestingly, VCAN is upstream of the EGF, ERBB2 family, and cytoskeleton remodeling signaling pathways and associated with the epithelial to mesenchymal transition process (Figure 3).

We investigated the potential drugs that may target the most frequently mutated genes and pathways in our Chinese NSCLC sample (Supplementary Table 10). In our gene list, TP53 and several mutated genes in ECM Remodeling category (CDH13, FN1, LAMA1, LAMA3, NID1, and MMP16) may interact with known drugs (20). For example, the gene expression of FN1, which has been reported to be an oncogene (21), is increased in response to the anticancer drug tamoxifen (22) and the anticancer combination of altretinoin and ascorbic acid (23). Similarly, FN1 activity is associated with chemoresistance to drugs such as doxorubicin (24,25). It has been reported that such chemoresistance affects FN1 in NSCLC work, in part, through the activation of the Akt/mTOR/p70S6K pathway (26). However, the significant drug responses of certain genes do not adequately justify them as feasible drug targets for treating diseases. In this study, an open translational research question then is whether targeting the promising genes identified earlier is effective in preventing recurrence of NSCLC.

We sequenced six BAC samples in this study. BAC is one of four histological distinct subtypes of lung ADC. It possesses unique clinical and pathological features and prognosis and responds to different treatments. FCRL4 is the only gene with missense mutations in two BAC samples. This gene is one of Fc receptor-like glycoproteins and encodes a member of the Ig receptor super family. FCRL4 aberrations were associated with lymphoma and myeloma previously (27,28).

The four truncating mutations we identified in BAC samples occurred in four genes, APC, KIDINS220, KIAA1211, and STAG2. APC is a tumor suppressor gene acting as an antagonist of the Wnt signaling pathway. It is also involved in other processes including cell migration and adhesion, transcriptional activation, and apoptosis. Defects in this gene have been linked to lung cancer before (29,30), consistent with our observation that APC protein is truncated in lung cancer. STAG2 is a gene encoding a subunit of the cohesion complex that regulate the separation of sister chromatids during cell division. STAG2 is located on the X chromosome so a single mutation event like the truncating mutation we identified here in a male patient with BAC (patient 345954) will totally abolish its function. Several recent papers (31–34) showed that inactivating mutations of STAG2 led to chromtid cohesion defect, which is the cause of aneuploidy.
that drives mutator phenotypes in human cancer. Our observation of STAG2 truncating mutation in male lung tumor supports the importance of STAG2 in human cancer.

In summary, this study systematically investigated the somatic genetic variants associated with Chinese NSCLC based on whole exome sequencing. The current exome sequencing technology and analytical methods are inefficient to identify certain structural gene changes such as gene fusions. For example, we missed the detection of recently established NSCLC relevant ALK-PTPN3 gene fusion (35) in our samples. However, exome sequencing is especially powerful in detect meaningful point mutations at the genome level, and thus we still made certain novel discoveries. Particularly, we found that the ECM remodeling/cell-cell adhesion gene set is the most significantly altered gene set in Chinese NSCLC. Functionally, ECM modification and remodeling is one of the most frequent cellular events in cancer progression. Before tumor cells leave their sites of origin and become metastatic, they must first detach from neighboring cells and remodel ECM structure and activity to gain the invasion and metastasis phenotypes (36). ECM, on the other hand, is not only a passives structure acting as a physical support/barrier but also a dynamic structure that contains ECM proteins that interact with cell surface receptors to initiate or modulate signal transduction of cancer cells residing on or within it (37). Our findings of frequent MXRA5 somatic mutations in Chinese NSCLCs contribute to the body of evidence implicating altered ECM remodeling in the genetic etiology of NSCLC. Other novel genes identified in this study also provide new clues regarding molecular targets for treating non-small cell lung cancer, especially in Chinese.

Supplementary Data

Supplementary Figure 1. Number of somatic mutations in protein coding regions detected in 14 lung cancer genomes. Samples were organized according to their histology and smoking status. FS, frameshift; indel, insertion and deletion; SMK, smokers; Non-SMK, non-smokers.

Supplementary Figure 2. Number of SNVs across the Chinese lung cancer samples. (A) ADC, BAC, and SCC; (B) smokers and non-smokers.

Supplementary Figure 3. dN/dS across the samples. (A) ADC, BAC, and SCC; (B) smokers and non-smokers.

Supplementary Figure 4. Somatic single-nucleotide mutation trends and patterns in Chinese lung cancer samples. Distribution of specific nucleotide changes among germ line and somatic variations in the lung cancer exome was presented for different histology (ADC, BAC, and SCC) and smoking status (SMK and Non-SMK).

Supplementary Tables

Supplementary Table 1. Clinical information of Chinese patients with NSCLC.

Supplementary Table 2. Sequence reads and coverage among 14 pairs of matched normal and tumor samples.

Supplementary Table 3. Summary of somatic SNVs in 14 Chinese patients with lung cancer.

Supplementary Table 4. All somatic SNVs in 14 Chinese lung cancer genomes.

Supplementary Table 5. Somatic SNVs validated by the Sanger sequencing.

Supplementary Table 6. Our validated somatic SNVs that are reported in the COSMIC database.

Supplementary Table 7. All somatic mutations in protein coding regions or splicing sites in 14 Chinese patients with lung cancer.

Supplementary Table 8. Missense mutations predicted to affect protein function.

Supplementary Table 9. The gene set enrichment analysis results for all of the 14 Chinese NSCLC tumors in the initial exome sequencing study. The columns of this table contain gene set name (column name: category), patient-oriented permutation P value (p-values, perm.null), non-silent/silent alteration ratio (ratio), number of mutated genes (nG), number of mutated subjects (nS), total number of genes in each gene set (nSet), number of somatic mutations in each gene set (nSOS), the mutated genes (geneSet), the mutated samples (sampleSet), the mutated genes and corresponding samples (geneSet/sampleSet). The significance cutoff is 1E-05.

Supplementary Table 10. Drugs (rows) and their druggable target genes identified in this study (columns). We limit the druggable targets to genes that were mutated in more than two individuals or were part of the most enriched pathway, the ECM remodeling pathway. The green-shaded entries indicate the drug by target interactions (labeled “1”). Entries in black indicate that the corresponding genes are not drug targets. The pink- and orange-shaded drug names indicate drugs that have been applied, at least in clinical trials for treatment of lung cancer (pink) or some other type of cancer (orange).

Acknowledgements

This work was supported in part by the National Institutes of Health (grant nos. 7R01AT003203, 7R01AT005522, 5R01CA134433, 5R01CA134682, 5R01CA113793, and 5R01CA129533) and Advancing a Healthier Wisconsin Fund. The authors thank Gary D. Stoner, Michael James, Haris Vikis, and Jay Tichelaar for their comments on the manuscript.

References