Integrative Genomics Analysis Identifies Candidate Drivers at 3q26-29 Amplicon in Squamous Cell Carcinoma of the Lung

Jing Wang1, Jun Qian2, Megan D. Hoeksema2, Yong Zou2, Allan V. Espinosa2, S.M. Jamshedur Rahman2, Bing Zhang1, and Pierre P. Massion2,3

Abstract

Purpose: Chromosome 3q26-29 is a critical region of genomic amplification in lung squamous cell carcinomas (SCC). Identification of candidate drivers in this region could help uncover new mechanisms in the pathogenesis and potentially new targets in SCC of the lung.

Experimental Design: We conducted a meta-analysis of seven independent datasets containing a total of 593 human primary SCC samples to identify consensus candidate drivers in 3q26-29 amplicon. Through integrating protein–protein interaction network information, we further filtered for candidates that may function together in a network. Computationally predicted candidates were validated using RNA interference (RNAi) knockdown and cell viability assays. Clinical relevance of the experimentally supported drivers was evaluated in an independent cohort of 52 lung SCC patients using survival analysis.

Results: The meta-analysis identified 20 consensus candidates, among which four (SENP2, DCUN1D1, DVL3, and UBXN7) are involved in a small protein–protein interaction network. Knocking down any of the four proteins led to cell growth inhibition of the 3q26-29–amplified SCC. Moreover, knocking down of SENP2 resulted in the most significant cell growth inhibition and downregulation of DCUN1D1 and DVL3. Importantly, a gene expression signature composed of SENP2, DCUN1D1, and DVL3 stratified patients into subgroups with different response to adjuvant chemotherapy.

Conclusion: Together, our findings show that SENP2, DCUN1D1, and DVL3 are candidate driver genes in the 3q26-29 amplicon of SCC, providing novel insights into the molecular mechanisms of disease progression and may have significant implication in the management of SCC of the lung. Clin Cancer Res; 1–11. ©2013 AACR.

Introduction

Lung cancer is the leading cause of cancer-related death among men and women in the United States. In 2011 alone, a total of 221,130 new cases and 156,940 deaths were reported (1). Eighty-five percent of these lung cancers are non–small cell lung carcinomas (NSCLC), a heterogeneous group composed of mostly squamous cell carcinomas (SCC) and adenocarcinomas. Despite improvements in response to increasingly sophisticated combination therapy, approximately 40% of patients with stage I and 60% with stage II NSCLC will die within 5 years, mainly due to the development of distant metastases (2).

In SCC of the lung, chromosome 3 aberrations are prevalent. Recurrent deletion on the 3p arm and gain on the 3q arm are present in the majority of cases (3, 4). Chromosome 3q26-29 amplification represents a critical region of genomic alteration in lung SCC. It occurs in up to 70% of lung SCCs and is of significant magnitude (i.e., 2- to 10-fold increase in copy number; ref. 5). The amplicon develops early in lung cancer development and persists at the metastatic stage (6). Identification of genes driving the selection for the 3q26-29 amplicon has been a significant research focus during the last decade. Using array comparative genomic hybridization (CGH) and mRNA expression array technologies, a number of candidate driver genes in this amplicon have been identified and proposed to contribute to the development of lung cancer. These genes include PK3CA (7, 8), SOX2 (9), DCUN1D1 (10), TP63 (11), EIF4G1 (12), EVII (13), THPO (14), TERC (15), ECT2 (16), PRKCI (17), EPHB3 (18), MASP1 (18), and SST (18).

Despite remarkable progress on candidate driver gene identification for this amplicon, no consensus exists as which are the drivers and what are their functional implications (9, 10). Moreover, existing studies tend to focus on a single gene in the focal region of amplification. It has been
Translational Relevance

The 3q26-29 amplicon is the most common and significant chromosomal aberration in the progression of squamous cell carcinoma (SCC) of the lung from severe dysplasia to metastasis. Identification of candidate driver genes in this 3q26-29 amplicon would uncover new mechanisms in the pathogenesis and potentially new targets in SCC. Using an integrative computational analysis followed by experimental validation, we identified three consensus candidate driver genes (SENP2, DCUN1D1, and DVL3) that could stratify patients into subgroups with different response to adjuvant chemotherapy. This integrated signature could serve as a molecular assay to guide personalized treatment decisions in SCC of the lung.

suggested that different genes in an amplicon may contribute synergistically to tumor progression (19, 20). We therefore hypothesized that genomic amplification leads to increased expression of several key regulators in the 3q26-29 amplicon that cooperate with one another to promote SCC cell growth. To identify these key regulators, we took an integrative genomics approach as depicted in Fig. 1. Although existing bioinformatic methods usually require paired copy number alteration (CNA) and gene expression data for the prioritization of candidate genes in an amplicon, in this study, we have developed and validated a simple but effective method for prioritizing candidate driver genes using expression data alone. This method allowed us to integrate seven independent datasets containing a total of 593 human primary tumor samples to identify 20 consensus candidate drivers in the 3q26-29 amplicon. To further filter for genes with potential functional connectivity, we used a network-based approach and identified a small protein interaction network consisting of SENP2, DCUN1D1, DVL3, and UBXN7, which are involved in posttranslational modifications such as SUMOylation, neddylation, and ubiquitination. Experimental interrogations showed a role of the network in promoting cell proliferation and suggested potential regulatory relationships between SENP2, DCUN1D1, and DVL3. Moreover, patient stratification based on gene expression values of this three-gene signature provides potential useful information about the adjuvant chemotherapy benefit for patients with SCC of the lung.

Materials and Methods

Datasets

One hundred and fifty-two paired gene expression and copy number data for lung SCC were downloaded from The Cancer Genome Atlas (TCGA) portal, in which gene expression data were from Agilent 244 K whole genome expression array and copy number data were from Agilent CGH 415 K array. Six unpaired gene expression data were from the following studies: Bild and colleagues (GSE3141; ref. 21), Expression Project for Oncology (Expo; GSE2109), Lee and colleagues (GSE8894; ref. 22), Raponi and colleagues (GSE4573; ref. 23), Wilkerson and colleagues (GSE17710; ref. 24), and Roepman and colleagues (25). Gene expression data with survival and adjuvant chemotherapy information was from Zhu and colleagues (GSE14814 from the JBR.10 trial; ref. 26), which only contained stage I and II patients. Only lung SCC samples from these datasets were included in this study. For each of the filtered datasets, the gene expression data were subjected to quantile normalization and standardized using a gene-wise Z-score transformation.

Of note, 299 and 110 paired gene expression and copy number data for head and neck SCC and cervical SCC were also downloaded from TCGA portal, in which gene expression data were from Illumina HiSeq 2000 RNA Sequencing Version 2 analysis and copy number data were from Affymetrix Genome-Wide Human SNP Array 6.0 array.

Human protein interaction data were collected and integrated from HPRD, MINT, intact, REACTOME, BioGRID, and DIP in April 2010. Only experimentally determined interactions supported by publications were considered to assure the reliability of the network. The consolidated dataset comprised 94,146 interactions involving 11,660 genes.

Calculating the amplification scores for TCGA copy number data

On the basis of the TCGA level 2 data downloaded from the TCGA data portal, we identified the markers in the 3q26-29 regions for each sample and then calculated the average normalized log2 score for all these markers to get the amplification score for each sample.

Concordance index

Concordance index (C-index; ref. 27) was defined as the fraction of all pairs of subjects (e.g., gene expression samples or genes in the amplicon) whose predicted scores (e.g., average expression scores or correlation between gene expression and average expression score) were correctly ordered among all subjects that can actually be ordered in the observed scores (e.g., amplification scores or correlation between gene expression and amplification score).

Order statistics

We divided the ranks by the total number of ranked genes (excluding genes with no rank because of missing values) and then calculated Q statistic for each gene in the 3q26-29 regions according to the following equation:

$$V_k = \sum_{i=1}^{N} (-1)^{r_i+1} \frac{V_{k-1}}{r_i}$$

where $Q(t_1, t_2, \ldots, t_N) = N \sum_{i=1}^{N} V_0$, $t_0 = 1.0$, $r_i$ is the rank ratio for data source $i$, $N$ is the number of data sources used, and $r_N = 0$. Because the Q statistic calculated by the above equation are not uniformly distributed under the null hypothesis and thus cannot be used directly as $P$ values (28). Thus, we randomly permuted the ranks of all genes in each of seven
datasets and calculated the random Q statistics. We repeated above process 1,000,000 times and generated a null Q statistic distribution for each gene. Comparing the real Q statistic with null distribution, we got a P value for each gene. Finally, the P values were corrected by the false discovery rate (FDR) of Benjamini–Hochberg procedure (29). Under 1% FDR, we identified candidate driver genes in the 3q26-29 regions.

Random walk analysis

NetWalker (http://bioinfo.vanderbilt.edu/netwalker; refs. 30, 31) was used to run random walk analysis based on the integrated human protein–protein interaction network described earlier. NetWalker identifies genes of potential biologic importance based on the assumption that mechanistically important genes are likely to form tightly connected groups, whereas other genes tend to be randomly distributed on the network. Using a set of genes as “seeds,” NetWalker calculated a score for each gene in the network based on its overall proximity to the seed genes, where the proximity is measured by the random walk similarity (32). NetWalker evaluates the statistical significance of the scores using a global P value and a local P value. A significant global P value indicates a nonrandom association between the gene and the input seeds, whereas a significant local P value ensures that the significant association is not simply due to network topology.

Cell cultures

Six human lung squamous carcinoma cell lines H520, HCC95, H2882, HCC15, H157, and SW900 and four human lung adenocarcinoma cell lines H1819, H1648, A549, and H23, were purchased from the American Type Culture Collection. Lung cancer cell lines were maintained in RPMI with 10% FBS. All cells were grown in 1 mmol/L penicillin/streptomycin.

RNAi knockdown assays

For transient siRNA transfection, DharmaFECT 1 Transfection Reagent, nontargeting siRNA #1, siGENOME SMART pool siRNA against SENP2 (M-006033-01-005), DCUN1D1 (M-019139-01-0005), DLV3 (M-004070-01-
0005), and UBXN7 (M-023533-01-005) were purchased from Thermo Scientific. Opti-MEM I media was purchased from Life Technologies. siRNA were suspended in water at a concentration of 20 nmol/L. The transfections were carried out according to the manufacturer’s instructions or as previously described (33). Briefly, for Western blot analysis, 1 × 10^5 cells were seeded into 6-well plates with medium overnight. For each well, 5 or 10 μL of siRNA were mixed with 185 μL Opti-MEM I and then combined with another mixture prepared using 4 μL DharmaFECT 1 transfection reagent and 15 μL Opti-MEM I. The final concentration of the siRNA was 25 to 100 nmol/L. For cell viability assay in 96-well plates, 50 or 100 nmol/L final concentration of siRNA was used.

Cell viability assay
Lung cancer cells were transiently transfected in 96-well plates with target siRNA or negative control siRNA as described earlier. The CellTitier 96 AQueous One Solution Cell Proliferation Assay (Promega) was then conducted following the manufacturer’s instructions or as previously described (34) on the wells posttransfection at 3 or 6 days. A representative viability experiment is shown in fold change that was normalized to negative siRNA control group with average ± SD.

Immunoblot analysis
Rabbit anti-SENP2 antibody (HPA029248) was purchased from Sigma-Aldrich. Rabbit anti-UBXD7 (AB10037) was from Millipore. Rabbit Dishevelled 3 antibody (GTX-102509) and rabbit DCUN1D1 antibody (GTX116558) were from Genetex. Western blotting was carried out using standard procedures as described in our previous study (33), with detection using the enhanced chemiluminescence system (Thermo Scientific). Antibody dilutions for immunoblotting were 1:1,000. The blots were reprobed with an anti-β-actin antibody (Sigma-Aldrich) to correct for protein loading differences. Anti-rabbit secondary antibody was purchased from Promega.

Results
Predicting 3q26-29 gene amplification level using gene expression data
Correlation between gene expression and regional amplification level has been widely used to prioritize candidate driver genes in a predefined amplicon because mRNA overexpression can better translate the effect of elevated copy number to tumor progression and offer biologic insights that are closer to levels of protein expression and function (35). This approach requires paired CNA and gene expression data, but many large-scale lung SCC studies, including those with available clinical endpoints, comprise only gene expression data. Therefore, we first investigated the possibility of using gene expression data alone to predict 3q26-29 gene amplification level and prioritize candidate driver genes in the region.

Using the TCGA dataset with paired CNA and gene expression data (36), we examined whether gene amplification directly contribute to increased mRNA expression in 152 lung SCC samples. We first estimated a 3q26-29 amplification score for each sample based on the CNA data (Supplementary Table S1) and then covisualized CNA and gene expression data in the Integrative Genomic Viewer (IGV; http://www.broadinstitute.org/igv/home), with the samples in both datasets ordered on the basis of the 3q26-29 amplification score. As shown in Fig. 2A and B, although 3q26-29 amplification is prevalent, different samples showed different levels of amplification. Samples with higher level of amplification also showed relatively higher expression for genes in this amplicon. We also found that most genes in this amplicon showed similar expression pattern, suggesting that genomic amplification contributes directly to an increase in mRNA gene expression. Next, we calculated the Spearman correlation coefficient between amplification scores and expression values for each gene in the expression data. Genes in the 3q26-29 regions were significantly better correlated with the amplification score compared with genes outside the region (P < 2.2E−16; Fig. 2C). Therefore, we reasoned that the amplification level of a chromosomal region could be inferred on the basis of the expression of genes within the region.

Using TCGA gene expression data, we calculated an average gene expression score for each sample based on all genes in the 3q26-29 amplicon (Supplementary Table S1). The average expression score was significantly correlated with the amplification score derived from the copy number data (Spearman correlation r = 0.94 and P < 2.2E−16) and the concordance index between the two scores was 89.4% (Fig. 2D), showing the validity of using the average expression score to predict the regional amplification level. We also ordered the samples by average expression scores and visualized the copy number data in the IGV. As shown in Supplementary Fig. S1, the magnitude of 3q26-29 amplification clearly declined with the decreasing of the average expression score.

Finally, we compared the correlation between gene expression and amplification level (cor_exp_amp) with the correlation between gene expression and average expression score (cor_exp_ave) for each gene in the 3q26-29 amplicon (Supplementary Table S2). We found high consistency between the two types of correlations (Spearman correlation r = 0.95 and P < 2.2E−16; concordance index = 90.6%; Fig. 2E), providing clear support for using the correlation between gene expression and average expression score to prioritize candidate driver genes. Meanwhile, we found that previously reported candidate driver genes had a wide spread along the diagonal (Fig. 2E), suggesting that many of them are not supported by the TCGA data, and a consensus conclusion on the driver genes at the 3q26-29 amplicon requires integrative analysis of data from multiple patient cohorts.

Integrative identification of candidate driver genes
The above results opened the possibility to estimate sample-specific amplification level and prioritize candidate
driver genes using gene expression data alone. Therefore, we applied the method to another six independent lung SCC gene expression datasets available in the public domain (Supplementary Table S3) and estimated sample-specific amplification levels (Supplementary Tables S4–S9). We integrated all seven datasets to identify genes that were consistently overexpressed in the 3q26-29 amplified samples as candidate drivers of this amplicon. We first ranked genes based on the correlation between gene expression and sample amplification score for the TCGA dataset or based on the correlation between gene expression and average expression score for other six datasets, respectively. Then, using the $N$-dimensional order statistic, we combined the seven ranked lists to calculate a $Q$ statistic and

Figure 2. Genomic amplification can be inferred from gene expression data. The IGV plot of 152 paired TCGA gene expression data (A) and copy number data (B), x-axis represents chromosome 3 and y-axis represents TCGA samples that are in descending order according to the amplification scores. Different color shades represent the expression or amplification values of samples from blue (the minimum value) to red (the maximum value). C, cumulative probability distribution of spearman correlation coefficient between amplification scores of samples and expression values of genes in 3q26-29 amplicon (red) and other genes in the expression data (blue) based on 152 TCGA paired copy number and gene expression samples. D, the scatter plot of amplification scores against average expression scores for each of the 152 TCGA SCC samples. E, the scatter plot of correlation between gene expression and amplification score (corr_exp_amp) against correlation between gene expression and average expression score (corr_exp_ave) for 164 genes in the 3q26-29 amplicon. Previously published candidate driver genes are labeled in the scatter plot. The solid lines in (D) and (E) are fitted lines based on the data in the plots. Because the data points in (D) and (E) are distributed along the diagonals, the data series represented by the x- and y-axis in each figure have a high concordance index (89.4% and 96.2%).
corresponding P value (see details in Materials and Methods), which represents the random chance of getting the same or better observed rank combinations for each gene (Supplementary Table S10). Under an FDR of 1%, 20 genes were identified as candidate drivers (Table 1). Because 3q26-29 amplicon is also present in SCC from other sites such as head and neck (37) and cervical (38), we tested whether these 20 genes were also candidate drivers in other SCC. On the basis of the paired gene expression and copy number data in TCGA (see Materials and Methods), we calculated the correlation between gene expression and 3q26-29 amplification level for these 20 genes in head and neck SCC and cervical SCC. As shown in Supplementary Tables S11 and S12, gene expression for all 20 genes were significantly correlated with 3q26-29 amplification level (for head and neck SCC, Spearman correlation P < 1.27E–24; for cervical SCC, Spearman correlation P < 7.31E–07), implying the validity of these 20 genes as candidate drivers in SCC.

Colocalization of candidate driver genes in a focal region of amplification suggests that coordinate amplification and overexpression of some of these genes might be required to alter specific mechanisms and eventually provide a growth advantage to cells that overexpress them (19, 20). Therefore, using the 20 candidate driver genes as "seeds," we applied NetWalker (30) to identify genes that might function together in a coordinated manner (see details in Materials and Methods). Controlling both local and global P values at the 0.01 level, we identified four of 20 candidate genes (SENP2, DCUN1D1, DVL3, and UBXN7) that are involved in a small, connected network with 13 significant nodes (Fig. 3). Many genes in the network are involved in SUMOylation, neddylation, and ubiquitin pathways.

| Table 1. Twenty candidate driver genes identified by the order statistic method with an FDR threshold of 1% |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Symbol          | Band            | TCGA (164)      | GSE2109 (173)   | GSE17710 (160)  | GSE3141 (173)   | GSE4573 (124)   | GSE8894 (173)   | Roepman (114)   | Q statistic     |
|                 | Link            |                 |                 |                 |                 |                 |                 |                 | FDR             |
| ABCC5           | q27.1           | 6b              | 1               | 7               | 15              | 12              | 26              | 7               | 8.35E–08 <0.001 |
| ACTL6A          | q26.33          | 2               | 43              | 1               | 22              | 24              | 35              | 26              | 3.54E–06 <0.001 |
| DCUN1D1         | q26.33          | 3               | 53              | 10              | 1               | 2               | 12              | NAc             | 3.66E–07 <0.001 |
| DLG1            | q29             | 25              | 44              | 9               | 13              | 30              | 11              | 22              | 1.78E–05 <0.001 |
| MFN1            | q26.32          | 20              | 7               | 22              | 10              | 14              | 44              | 25              | 1.28E–05 <0.001 |
| NCBP2           | q29             | 28              | 23              | 15              | 4               | 15              | 23              | 16              | 1.38E–06 <0.001 |
| PRKCI           | q26.2           | 43              | 36              | 20              | 6               | 13              | 3               | 18              | 6.17E–06 <0.001 |
| PSMD2           | q27.1           | 29              | 3               | 35              | 33              | 4               | 5               | 20              | 1.34E–06 <0.001 |
| SENP2           | q27.2           | 11              | 32              | 3               | 3               | 5               | 7               | 2               | 1.27E–09 <0.001 |
| SENP5           | q29             | 42              | 19              | 12              | 24              | 23              | 1               | 5               | 2.12E–06 <0.001 |
| UBXN7           | q29             | NA              | 39              | NA              | 8               | 22              | 18              | 4               | 1.05E–04 <0.001 |
| ZNF639          | q26.32          | 1               | 34              | 11              | 38              | 95              | 21              | 6               | 2.68E–05 <0.001 |
| ATP11B          | q26.33          | 5               | 48              | 26              | 19              | 21              | 9               | 1               | 2.46E–06 0.0091 |
| DVL3            | q27.1           | 31              | 16              | 18              | 27              | 27              | 32              | 10              | 1.45E–05 0.0091 |
| EIF2B5          | q27.2           | 22              | 11              | 23              | 32              | 16              | 14              | NA              | 2.83E–05 0.0091 |
| FXR1            | q26.33          | 19              | 12              | 38              | 37              | 9               | 8               | 13              | 3.60E–06 0.0091 |
| LSG1            | q28             | 32              | 30              | 2               | 34              | 38              | 43              | NA              | 1.11E–04 0.0091 |
| NDUFB5          | q26.33          | 27              | 33              | 14              | 2               | 8               | 61              | NA              | 6.47E–05 0.0091 |
| PGIX            | q29             | 23              | 25              | 24              | 14              | NA              | 24              | 3               | 6.86E–06 0.0091 |
| PK3CA           | q26.32          | 8               | 41              | 27              | 25              | 3               | 34              | NA              | 1.61E–05 0.0091 |

aThe number of genes in both dataset and 3q26-29.

bGene rank in a dataset according to the corr_exp_amp (TCGA) or corr_exp_ave (other datasets) from largest to smallest.

cNA, the gene is not in the corresponding dataset.

Figure 3. Graphical representation of an inferred SCC related candidate driver network centered around four candidate driver genes in 3q26-29. Four candidate driver genes are shown in red.
Functional implications of four candidate drivers in SCC

Prior work has established the oncogenic role of DCUN1D1 (defective in cullin neddylation 1 domain containing 1, also known as SCCRO or SCC-related oncogene) being an important component of the neddylation E3 complex in SCC (10, 39). In contrast, the specific function of SENP2, DLV3, or UBXN7 in SCCs of the lung remains unknown. To determine the relative contribution of over-expression of four candidate driver genes in lung SCC, we first sought to examine the expression status of four proteins in a panel of NSCLC cell lines, including five cell lines that harbor 3q amplicon including H520, HCC95, H1648, H1819, and H2882 and five cell lines that do not, including H157, HCC15, SW900, A549, and H23. As shown in Fig. 4A, overall expression of the four proteins is higher in the cell lines harboring 3q amplicon. Using Cancer Cell Line Encyclopedia database (http://www.broadinstitute.org/ccle/home), we further confirmed that four genes are amplified and have higher mRNA expression in HCC95, H1648, and H520 compared with the other four cell lines that do not harbor 3q amplification (Supplementary Table S13).

We next assessed the effect on lung cancer cell growth by knocking down the four proteins expression. We measured the proliferation of H520 cells with siRNAs against four genes respectively or scramble control siRNAs over 3 and 6 days (Fig. 4B). The silencing of the four proteins led to the cell growth inhibition at varied degrees (20%–40% inhibition), among which, knocking down SENP2 resulted in most significant inhibition (40%). Interestingly, we found that cells with SENP2 knockdown induced the downregulation of DCUN1D1 and DVL3, but not UBXN7 (Fig. 4C). In contrast, knockdown of DCUN1D1, DVL3, or UBXN7 did not lead to downregulation of SENP2 protein expression (Fig. 4D). The same findings were confirmed in another 3q amplified cell line HCC95 but not in A549, which does not harbor 3q amplicon (Supplementary Fig. S2).

Three-gene signature to predict adjuvant chemotherapy sensitivity in SCC patients

To evaluate the potential clinical significance of the computationally predicted network and experimentally supported driver genes, we tested whether a gene expression signature with the three genes (SENP2, DCUN1D1, and...
DVL3) could inform patient prognosis and treatment response. On the basis of the average of standardized expression levels of the three genes, we classified 52 patients from an independent dataset (GSE14814) into two groups: a "high-expression" group with above median expression of the genes and a "low-expression" group with below median expression of the genes. For each group, we further separated patients into two subgroups based on whether they received adjuvant chemotherapy.

As a baseline comparison (i.e., for patients who did not receive adjuvant chemotherapy), the high-expression group had a relatively shorter survival outcome compared with the low-expression group [Fig. 5, red solid line vs. green solid line; HR, 1.98; 95% confidence interval (CI), 0.62–6.33; \( P = 0.24 \)]. Interestingly, adjuvant chemotherapy was significantly associated with improved survival outcome for the high-expression group (Fig. 5, red dashed line vs. red solid line; HR, 0.243; 95% CI, 0.061–0.967; \( P = 0.031 \)), with an 85.7% overall survival rate at 5 years for patients who received chemotherapy compared with 41.7% for patients who did not receive adjuvant chemotherapy. In contrast, adjuvant chemotherapy had limited effect on patients in the low-expression group (Fig. 5, green dashed line vs. green solid line; HR, 0.835; 95% CI, 0.224–3.11; \( P = 0.788 \)), with a 76.9% overall survival rate at 5 years for patients who received adjuvant chemotherapy as compared with 69.2% for patients who did not receive chemotherapy. Thus, patient stratification based on the three-gene signature may predict potential benefit from adjuvant chemotherapy in the management of patients with SCC of the lung. Similar results were obtained when the analyses using three-gene signature were conducted for stage I and II patients separately (Supplementary Fig. S3). We also conducted survival analysis by stratifying patients based on each of the three genes individually. As shown in Supplementary Fig. S4, stratification based on all three genes outperformed those based on individual genes, suggesting that cooperation of the three genes may play an important role in cancer therapy.

**Discussion**

In this report, we investigated the functional implications of the chromosome 3q amplicon in lung SCC. We show that (i) a simple but effective method can be used to estimate sample-specific amplification level and prioritize candidate driver genes using expression data alone; (ii) data fusion and network analysis techniques enable data integration across different studies and among different genes, allowing identification of driver genes that are supported by multiple studies and involved in a common biologic theme; (iii) all four computationally inferred driver genes including SENG2, DCUN1D1, DVL3, and UBXN7 were experimentally proven to regulate cellular proliferation in lung squamous cancer cells that harbor the amplicon and three of them (SENP2, DCUN1D1, and DVL3) might be functionally related to one another; and (iv) the three-gene signature may be predictive of response to conventional chemotherapy in SCC.

Although analyzing gene expression in conjunction with DNA-level changes such as CNA holds great promise in the identification of candidate driver genes of lung SCC, many findings from this type of analysis are not reproducible in a new patient cohort. For example, many previously reported candidate driver genes are not supported by the TCGA data (Fig. 2E). One potential solution is to integrate paired CNA and gene expression datasets from multiple cohorts. However, paired lung SCC molecular datasets are still limited and even fewer with clinical annotation. Because there are sufficient numbers of published lung SCC gene expression datasets, in this study, we developed a method to predict 3q26-29 amplification level for individual tumor samples from a patient cohort based on gene expression data alone. This allowed us to integrate data from one cohort with paired CNA and gene expression measurements and another six cohorts with only gene expression measurements to identify consensus candidate driver genes with consistent overexpression in amplified tumors from multiple patient cohorts. For many cancer types, genomic studies have identified amplicons without well-defined driver genes. Although paired copy number and gene expression datasets remain very limited across all cancer types, gene expression datasets are relatively abundant in public databases such as the Gene Expression Omnibus (GEO). Thus, our method could have a general application in predicting amplification level and candidate driver genes in predefined amplicons for other cancer types.
In our study, we used the average expression score of all genes in the 3q26-29 amplicon to predict 3q26-29 amplification level. Although chromosomal amplification directly contributes to an increased mRNA expression level of genes in the amplified region (see Fig. 2A), only 110 of 164 genes in this region showed significant correlation with the amplification score (Spearman $P < 0.01$). Moreover, even genes with significant correlations also showed different levels of correlation (Spearman correlation $r$ ranging from 0.87 to 0.21). Thus, it is possible that expression of some genes may better represent genomic amplification than others. To test this possibility, we extracted the first principal component from gene expression data for all genes in the amplicon and used the first principal component as an alternative approach to estimate the level of amplification. However, the results based on first principal component were very similar to that based on all genes (data not shown), suggesting the validity of using average expression score of all genes in an amplicon as an easy method for estimating amplification level.

It is generally believed that regions of recurrent genome CNA harbor genes that are essential to cancer progression, nevertheless many functional studies focus on identifying a single candidate driver gene in the region of CNA responsible for cancer progression. It has been suggested that multiple genes from an amplicon may cooperatively promote tumorigenesis (20, 40). Thus, in our study, we used network-based analysis to identify genes that might function in a coordinated manner. Through integrated computational network analysis and experimental interrogation, we identified three genes that may function together to promote tumor cell growth. These genes are involved in SUMOylation, neddylation, or ubiquitination pathway. Recently, the functional cross-talk between these similar pathways has emerged (41). Nevertheless, how three pathways contribute to lung SCC development remains to be investigated.

Our experimental data suggest that SENP2 may regulate the expression of DCUN1D1 and DVL3. SENP2 belongs to Sentrin/SUMO-specific protease family and has diverse function via SUMOylation or deSUMOylation in human cells. In cancer cell, components of the SUMOylation or deSUMOylation machinery are enhanced or reduced (42). In lung SCC cells, SENP2 has highest prevalence of amplification compared with other SENP family members (Supplementary Fig. S5), further supporting its key role in lung SCC development. Whether and how SENP gene expression levels disturb SUMO homeostasis and contribute to lung SCC cancer development and progression are yet to be determined. Many SUMOmodified proteins function in regulation of transcription, chromatin structure, maintenance of the genome, and signal transduction (41). MDM2-p53 is most well-studied SUMOmodified proteins. SENP2 promotes the deSUMOylation of Mdm2, contributing to the p53-dependent regulations (43). Interestingly, the tumour suppressor p53 has been shown to be modified at its C-terminus with ubiquitin, SUMO and NEDD8, although the consequence undergoing these posttranslational modifications may differ (44). Notably, NEDD8 is conjugated to three lysines within the C-terminus of p53 through the E3 ligase activity of MDM2 (45), adding another mechanism by which MDM2 inhibits p53 activity.

DCUN1D1 is novel identified NEDD8 E3 complex component and has been proposed be to an oncogene in lung SCC (10, 39). Mechanistically, DCUN1D1 binds to the components of the neddylation pathway consisting of Cullin-ROC1, Ubc12, and CAND1 to promote cullin neddylation via releasing the inhibitory effects of CAND1 (39). It is unclear how neddylation is regulated and how many oncogenic neddylated targets exist in cancer cell. Our data showed that knocking down SENP2 using siRNA leads to the downregulation of DCU1ND1, prompting us to propose that DCU1ND1 might be a potential SUMO-modified target regulated by SENP2 in lung SCC, a mechanism similar to SENP2-MDM2 pathway.

Disheveled (DVL) is an essential scaffold protein that transduces both the canonical (β-catenin–dependent) and the noncanonical (β-catenin–independent) Wntsignaling pathways (46). Activated DVL3 promotes Wnt/β-catenin activation and undergoes KHL12–Cul3–mediated ubiquitination for degradation (47). Wnt-β-catenin pathway is disregulated in lung cancer (48) and may lead to new treatment strategies (49). Dvl3 overexpression has also been found in NSCLC, and once silenced, inhibition of cellular proliferation is observed (50). SENP2 negatively modulates β-catenin via its deSUMOylation activity in hepatocellular carcinoma cell (51) and colon cancer cells (52). It is suggested that noncanonical Wnt/PCP pathway but not canonical Wnt/β-catenin was enhanced in lung SCCs (48). Our data showed that SENP2 knockdown leads to DVL3 downregulation. Therefore, we hypothesize that SENP2 may regulates β-catenin through DVL3 dependent or independent mechanisms in lung SCC cells and this will be tested in the context of future studies. Thus, unlike previous studies that identify individual genes from the 3q26-29 amplicon, the three genes identified in this study may represent a common biologic mechanism that provides growth advantage to tumors harboring the amplicon.

Our survival analysis further supports potential importance of these three genes in lung SCC patient prognosis, and particularly, response to adjuvant chemotherapy. Despite the small sample size of the patient cohort used for survival analysis, we observed distinct response to adjuvant chemotherapy for patients with high- and low-level expression of the three genes, respectively and independently of the disease stage. Unfortunately, no publicly available datasets in SCC with annotated clinical outcomes were identified to confirm our preliminary data. Upon further validation in independent cohorts of patients with NSCLC, the three-gene signature may provide significant value in personalizing the management of patients with SCC of the lung.

In summary, using an integrative genomics approach, we identified three functionally linked candidate driver genes at 3q26-29 that provide novel mechanistic and clinical insights into the pathobiology of lung squamous
carcinoma. Our results show the power of an integrative systems biology approach in effectively identifying candidate driver genes from a predefined tumor amplicon.

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Authors’ Contributions
Conception and design: J. Wang, J. Qian, S.M.J. Rahman, B. Zhang, P.P. Massion
Development of methodology: J. Wang, Y. Zou, B. Zhang, P.P. Massion
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J. Wang, M.D. Hoeksema, Y. Zou, P.P. Massion
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J. Wang, J. Qian, Y. Zou, A.V. Espinosa, S.M.J. Rahman, B. Zhang, P.P. Massion

References

Writing, review, and/or revision of the manuscript: J. Wang, J. Qian, A.V. Espinosa, S.M.J. Rahman, B. Zhang, P.P. Massion

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): J. Wang, J. Qian, M.D. Hoeksema, Y. Zou, P.P. Massion

Study supervision: B. Zhang, P.P. Massion

Grant Support
This work was funded in part by the NIH CA102353 (to P.P. Massion), GM088822 (to B. Zhang), and the Vanderbilt SPORE in lung CA90949.
Integrative Genomics Analysis Identifies Candidate Drivers at 3q26-29 Amplicon in Squamous Cell Carcinoma of the Lung

Jing Wang, Jun Qian, Megan D. Hoeksema, et al.

Clin Cancer Res  Published OnlineFirst August 1, 2013.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-13-0594

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2013/08/01/1078-0432.CCR-13-0594.DC1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.