Genome-Wide Association Study of Genetic Predictors of Overall Survival for Non–Small Cell Lung Cancer in Never Smokers

Xifeng Wu, Liang Wang, Yuanqing Ye, Jeremiah A. Aakre, Xia Pu, Gee-Chen Chang, Pan-Chyr Yang, Jack A. Roth, Randolph S. Marks, Scott M. Lippman, Joe Y. Chang, Charles Lu, Claude Deschamps, Wu-Chou Su, Wen-Chang Wang, Ming-Shyan Huang, David W. Chang, Yan Li, V. Shane Pankratz, John D. Minna, Waun Ki Hong, Michelle A.T. Hildebrandt, Chao Agnes Hsiung, and Ping Yang

Abstract

To identify the genetic factors that influence overall survival in never smokers who have non–small cell lung carcinoma (NSCLC), we conducted a consistency meta-analysis study using genome-wide association approaches for overall survival in 327 never smoker patients with NSCLC from The University of Texas MD Anderson Cancer Center (Houston, TX) and 293 cases from the Mayo Clinic (Rochester, MN). We then conducted a two-pronged validation of the top 25 variants that included additional validation in 1,256 patients with NSCLC from Taiwan and assessment of expression quantitative trait loci (eQTL) and differential expression of genes surrounding the top loci in 70 tumors and matched normal tissues. A total of 94 loci were significant for overall survival in both MD Anderson and Mayo studies in the consistency meta-analysis phase, with the top 25 variants reaching a P value of 10^-4. Two variants of these 25 were also significant in the Taiwanese population: rs6901416 [HR, 1.44; 95% confidence interval (CI), 1.01–2.06] and rs10766739 [HR, 1.23; 95% CI, 1.00–1.51]. These loci resulted in a reduction of median survival time of at least eight and five months in three populations, respectively. An additional six variants (rs4237904, rs7976914, rs4970833, rs954785, rs485411, and rs10906104) were validated through eQTL analysis that identified significant correlations with expression levels of six genes (LEAD3, TMBIM, ATXN7L2, SHE, ITIH2, and NUDT5, respectively) in normal lung tissue. These genes were also significantly differentially expressed between the tumor and normal lung tissue. These findings identify several novel, candidate prognostic markers for NSCLC in never smokers, with eQTL analysis suggesting a potential biologic mechanism for a subset of these observed associations. Cancer Res; 73(13): 4028–38. ©2013 AACR.

Introduction

Lung cancer is the number one cause of cancer-related mortality in the United States, with an estimated 157,300 deaths in 2010 (1). Non–small cell lung carcinoma (NSCLC) comprises approximately 80% of all lung cancers, with 5-year survival rate for all stages ranging from 11% to 17% (2). It is well known that the majority of lung cancer deaths are due to tobacco smoking; however, 10% to 20% of lung cancer-related deaths involve people who have never smoked (3). It is increasingly clear that lung cancer in never smokers represents a unique disease entity separate from smoking-related lung cancer, highlighting the need for investigation and discovery of novel genetic factors influencing survival in this population.

Tumor-based studies have shown that lung cancer in never smokers has a distinct profile from that in smokers. Never smokers have very low rates of mutations in K-ras and p53 genes...
as compared with smokers (4–7). Inactivation of p16 by promoter methylation also has been reported to be more frequent in tobacco-associated lung cancer (8–10), whereas EGFR mutations were more common in never smokers (11). Indeed, better efficacy and survival has been reported for EGFR tyrosine kinase inhibitors in the treatment of NSCLC among never smokers (11–13). Furthermore, differences in allelic imbalances have been observed with higher chromosomal aberrations on 3p, 6q, 9p, 17p, and 19p from smokers than never smokers (14). However, these loci were commonly altered in both smoking subgroups in a Chinese population (15). Patients from East Asia, particularly women never smokers, also have a high frequency of EGFR mutations (13). These results indicate that there are also differences in tumor characteristics across ethnic/racial groups that may further influence outcomes.

With these differences between lung cancer in smokers and never smokers in mind, we previously conducted a genome-wide association study (GWAS) to identify genetic variants of single-nucleotide polymorphisms (SNP) associated with the risk of developing lung cancer in never smokers (16). In contrast with the candidate genes involving nicotinic receptors and other chromosomal regions (i.e., 5p15, 15q25, and 6p21) previously reported as susceptibility loci for overall lung cancer risk in smoking populations (17–21), we found that GC5 genetic variants were associated with lung cancer in never smokers only and this association was supported by expression quantitative trait loci (eQTL) analysis. Because the GWAS approach itself is only able to map disease risk to specific loci, not necessarily the causal gene, additional clues about the biologic mechanism responsible for the association can be provided by the incorporation of eQTL analysis into GWAS (22–24).

Individual differences in survival among patients treated with identical treatment regimens for same-stage tumors strongly implicate a genetic basis also for survival in NSCLC. Identifying the genetic basis that contributes to this variation would be potentially valuable to the clinician for several reasons, including stratification of the population into different prognostic groups to assist in selection of optimal treatment modalities. This may also facilitate risk-stratified clinical trials for novel treatments in those patients who are predicted to not respond to standard treatment. In addition, there is the potential to identify novel targets for therapeutic development to improve treatment response/survival. Few studies have comprehensively examined the influence of germline genetic variants on lung cancer survival. A previous genome-wide study of variation within NSCLC tumors identified polymorphisms within STK39, PCDH7, A2BP1, and EYA2, as well as copy number changes on 3p, 5p, and 8q that were prognostic of overall survival in early-stage patients (25). In addition, our group identified a genetic predictor of survival in patients with NSCLC treated with platinum-based therapy (26). However, these studies were primarily of smokers and did not explore genetic variation that would predict overall survival in never-smoker patients with lung cancer. In this study, we conducted the first GWAS among never-smoker patients with lung cancer to identify germline genetic variants that are associated with overall survival in this distinct population of NSCLC. To increase the sample size and to minimize false discovery, a multistage study design using a dual genome-wide scan was conducted in two independent populations for consistency meta-analysis. The top 25 variants were then validated using a two-pronged approach that included genotyping in a large population of never-smoker patients with lung cancer from Taiwan to better understand the effects of these candidate prognostic loci in a population of non-European descent and eQTL mapping to identify prognostic loci based on potential functional effects within normal lung tissues.

Materials and Methods

Study populations

This study involved three independent patient populations from The University of Texas MD Anderson Cancer Center (MDACC; Houston, TX) and Mayo Clinic (Mayo; Rochester, MN) in the United States and multiple institutions in Taiwan (Fig. 1). All participants included in this study were patients with NSCLC who were never smokers. Written informed consent was obtained from all participants and the study was approved by the Institutional Review Boards of each participating institution.

MD Anderson study. All cases were newly diagnosed, histologically confirmed patients with NSCLC enrolled from 1995 to 2008 in an ongoing epidemiologic lung cancer study at MDACC. Comprehensive epidemiologic data were collected, including detailed smoking history. At the end of each interview, a blood sample was collected for DNA extraction. Clinical and follow-up data were collected by trained chart reviewers, including date of diagnosis, pathology, grade, clinical stage, surgery, chemotherapy, radiotherapy, chemoradiotherapy, and mortality.

Mayo Clinic study. Never-smoker patients with lung cancer were identified and enrolled between 1997 and 2007. A detailed subject enrollment process has been reported previously (27, 28). Briefly, new cases diagnosed with lung cancer are identified by a daily electronic pathology reporting system. Once identified, study consents are obtained from the patients for enrollment, their medical records abstracted (from Mayo Clinic and outside records if treated elsewhere), and interviews conducted. Detailed information on demographics and smoking history was collected, vital status and cause of death were determined by reviewing the Mayo Clinic registration database and medical records, correspondence from patients’ next of kin, death certificates, obituary documents, the Mayo Clinic Tumor Registry, and the Social Security Death Index.

Taiwan study. All cases were never smokers drawn from the ongoing cooperative study, the Genetic Epidemiological Study of Lung Adenocarcinoma (GELAC) in Taiwan, which includes patients recruited between January 2002 and May 2010 from 6 hospitals (29). Cases were over the age of 18 years with incident primary lung cancer and confirmed by hospital pathologists. Blood sample and epidemiologic data on demographic characteristics and smoking history were collected by a trained research nurse at recruitment.

Genotyping

Genotyping and quality control for the MDACC and Mayo datasets have been previously described (16). Briefly, genotyping data for MDACC patients were obtained from Illumina HumanHap 660k BeadChips and genotyping data for Mayo...
patients were obtained from Illumina HumanHap 370k and 610k BeadChips. All the individuals had a call rate of 95% or more and SNPs passed the quality control filtering with a call rate of 95% or more and a minor allele frequency of 0.01 or more. Genotyping of the top 25 SNPs was conducted using TaqMan Genotyping assays (Applied Biosystems) at the National Health Research Institutes (Zhunan, Taiwan).

**Microarray eQTL analysis**

Total RNA was extracted from a total of 70 tumor tissues and 70 matched normal lung tissues from Mayo Clinic using miRNeasy Mini Kit (QIAGEN) under the manufacturer’s guidelines. All 70 patients were analyzed in the GWAS. Total RNA integrity was assessed using an Agilent 2100 Bioanalyzer (19). We used the Whole-Genome DASL Assay (Illumina) for gene expression profiling according to the manufacturer’s recommendation. Two hundred nanogram of RNA from each subject was labeled and hybridized to each array using standard Illumina protocols and scanned on a BeadArray reader. Arrays were normalized using quantile normalization implemented in Bioconductor (www.bioconductor.org) and samples with replicates were averaged. We filtered out probes with median detection $P > 0.01$ based on reported values from BeadStudio. For the eQTL analysis, we identified genes within 600 kb surrounding each selected SNP for genotype-phenotype association studies. Gene expression levels in normal lung tissues were regressed to the number of minor alleles for each SNP representing that region. Genotype-phenotype associations with $P < 0.05$ were considered significant. Differences between tumor and normal tissues were assessed by t test with false discovery-corrected $P$ values.

**Statistical analysis**

Overall survival time was defined as the time from date of diagnosis to date of death or date of last follow-up, whichever came first. For the U.S. populations, a “never smoker” was defined as an individual who had smoked less than 100 cigarettes during his/her lifetime. ‘Never smokers’ in the Taiwan study were individuals who had never smoked or not smoked one cigarette a day or one cigarette a week for 6 months at any period during his/her lifetime. MDACC and Mayo independently assessed three genetic models of inheritance (dominant, recessive, and additive) for each SNP using their own GWAS datasets. HR and 95% confidence intervals (95% CI) were estimated using multivariate Cox proportional hazard regression analysis for each SNP. The covariates included were age, gender, stage, and treatment information at MDACC and were gender, stage, grade, and treatment information at Mayo. The model with the smallest $P$ value was used to measure the statistical significance of each SNP. Only the dominant model was considered when the rare homozygous genotype was less than 5 in patients with or without event of interest. Each site identified the top 1,000 SNPs with overall survival (excluding SNPs with pairwise $R^2 > 0.9$) that were overlapping in Illumina HumanHap 660k and HumanHap 370k BeadChips. Of note, 991 of the top 1,000 SNPs from the MDACC scan passed quality control measures in the Mayo dataset and 996 of the top 1,000 SNPs from the Mayo scan passed quality control measures in the MDACC dataset. The combined SNP list contained 6 overlapping SNPs, resulting in a final of 1,981 SNPs in the combined list. A meta-analysis was conducted using a fixed-effects model and estimated HR and 95% CI derived from the Cox regression analysis adjusted for the appropriate covariates in the two datasets. Heterogeneity was
tested with the Cochran Q test. The Cochran Q test was calculated by summing the squared deviations of each study’s effect estimate from the overall effect estimate weighted by inverse variance from each study. Kaplan–Meier curves and log-rank tests were used to assess the 3-year survival difference by individual polymorphisms. The quantile–quantile (Q–Q) plot displays the observed test statistic against the expected test statistic. For the MDACC population, we also selected the first principal component to include in the adjustment based on the scree plot of the top 60 eigenvalues and the Tracy–Widom P values (data not shown). Only the top eigenvalue was more than 1.50 and showed clear separation from the other eigenvalues, which were all less than 1.20. The λ(GC) for adjusting the first principle component (PC) was 1.159 compared with 1.175 without adjusting for PC. The effect of population stratification is predicted to be minimal. Therefore, no PC was adjusted in the subsequent association analyses. Analyses at MDACC were conducted using STATA software (version 10.1; STATA Corporation) and data manipulations at both MDACC and Mayo were completed using the PLINK and R software packages.

For the top 25 SNPs from the meta-analysis and genotyped in the Taiwanese population, one variant (rs4263860) was not polymorphic in the population and was removed from further analysis. A multivariate Cox proportional hazard regression analysis was conducted to estimate the HR and 95% CI while adjusting for age, gender, stage, and treatment information. For the Taiwan study, analyses were done using R software.

Results

Among all the participants, 327 were from MDACC, 293 from the Mayo Clinic, and 1,256 from Taiwan (Table 1). For the MDACC population, 13% of the patients were alive but with follow-up less than 3 years, 10.4% of the living patients had less than 3 years of follow-up at the Mayo Clinic, and 12.6% of the patients were alive but with follow-up less than 3 years in Table 1.

Table 1. Host characteristics

<table>
<thead>
<tr>
<th></th>
<th>MD Anderson</th>
<th>Mayo Clinic</th>
<th>Taiwan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%)a</td>
<td>N (%)a</td>
<td>N (%)a</td>
</tr>
<tr>
<td>Total</td>
<td>327</td>
<td>293</td>
<td>1,256</td>
</tr>
<tr>
<td>Age, mean (SD)</td>
<td>61.40 (13.15)</td>
<td>62.6 (13.2)</td>
<td>59.6 (11.9)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>104 (31.8)</td>
<td>98 (33.4)</td>
<td>204 (16.2)</td>
</tr>
<tr>
<td>Female</td>
<td>223 (68.2)</td>
<td>195 (66.6)</td>
<td>1,051 (83.7)</td>
</tr>
<tr>
<td>Clinical stage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage I</td>
<td>92 (28.1)</td>
<td>105 (35.8)</td>
<td>97 (7.7)</td>
</tr>
<tr>
<td>Stage II</td>
<td>15 (4.6)</td>
<td>15 (5.1)</td>
<td>28 (2.2)</td>
</tr>
<tr>
<td>Stage III</td>
<td>76 (23.2)</td>
<td>78 (26.6)</td>
<td>271 (21.6)</td>
</tr>
<tr>
<td>Stage IV</td>
<td>144 (44.0)</td>
<td>95 (32.4)</td>
<td>753 (60.0)</td>
</tr>
<tr>
<td>Pathology</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>244 (74.6)</td>
<td>230 (78.5)</td>
<td>1,011 (80.5)</td>
</tr>
<tr>
<td>Squamous cell</td>
<td>26 (8.0)</td>
<td>15 (5.1)</td>
<td>87 (6.9)</td>
</tr>
<tr>
<td>NSCLC, not otherwise specified</td>
<td>31 (9.5)</td>
<td>18 (11.3)</td>
<td>99 (7.9)</td>
</tr>
<tr>
<td>Bronchioloalveolar</td>
<td>20 (6.1)</td>
<td>15 (5.1)</td>
<td>5 (0.4)</td>
</tr>
<tr>
<td>Other</td>
<td>6 (1.8)</td>
<td>15 (5.1)</td>
<td>34 (2.7)</td>
</tr>
<tr>
<td>Surgery</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>135 (41.3)</td>
<td>158 (53.9)</td>
<td>291 (23.2)</td>
</tr>
<tr>
<td>No</td>
<td>192 (58.7)</td>
<td>135 (46.1)</td>
<td>964 (76.7)</td>
</tr>
<tr>
<td>Radiation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>88 (26.9)</td>
<td>76 (25.9)</td>
<td>519 (41.3)</td>
</tr>
<tr>
<td>No</td>
<td>239 (73.1)</td>
<td>217 (74.1)</td>
<td>731 (58.2)</td>
</tr>
<tr>
<td>Chemotherapy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>197 (60.2)</td>
<td>177 (60.4)</td>
<td>1,105 (88.0)</td>
</tr>
<tr>
<td>No</td>
<td>130 (39.8)</td>
<td>116 (39.6)</td>
<td>107 (12.0)</td>
</tr>
<tr>
<td>Chemoradiation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>34 (10.4)</td>
<td>69 (23.5)</td>
<td>n/a</td>
</tr>
<tr>
<td>No</td>
<td>293 (89.6)</td>
<td>224 (76.5)</td>
<td>n/a</td>
</tr>
<tr>
<td>Vital status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead</td>
<td>217 (66.4)</td>
<td>187 (63.8)</td>
<td>875 (69.7)</td>
</tr>
<tr>
<td>Alive</td>
<td>110 (33.6)</td>
<td>106 (36.2)</td>
<td>377 (30.0)</td>
</tr>
</tbody>
</table>

*Percentages may not be equal to 100% due to missing values. Abbreviation: n/a, not applicable.
NOTE: The model is a model of inheritance. The top significant target gene is the gene displaying the most significant eQTL within a 600-kb region flanking each SNP.

Abbreviations: add, additive; chr, chromosome; dom, dominant; FDR, false-discovery rate; MAF, minor allele frequency.

Taiwan. The median age at diagnosis was approximately 60 years, and a majority of the patients were female, up to 83.7% in the Asian population. The MDACC and Mayo studies were similar in the distributions of stage, whereas the Taiwan study had an excess of patients with stage IV disease. There were also differences between the U.S.-based studies and the Taiwan study with regard to treatment regimens. These differences were taken into account in the data analysis through multivariable adjustment for sex, stage, and treatment regimens.

The initial genome-wide scans were done independently by MDACC and Mayo and identified several promising loci (Supplementary Fig. S1). From the 1,000 SNPs identified from each site, the number of overlapping SNPs was 6, which is more than the 3.17 that would be expected by chance ($P = 0.10$). A meta-analysis was conducted for the 1,981 SNPs identified in the combined list (Fig. 1). The number of SNPs that showed association with overall survival ($P < 0.05$) in both the MDACC and Mayo studies with a combined $P < 1.67 \times 10^{-3}$ (Supplementary Table S1) was 94, which is more than the 48.4 variants that would be expected by chance alone ($P = 9.86 \times 10^{-11}$). No heterogeneity was detected between MDACC and Mayo study samples ($P_{\text{heterogeneity}} > 0.05$). Of these 94 loci, there were a total of 25 SNPs at a combined $P$ value of $10^{-6}$, including 7 reaching $10^{-7}$.

These top 25 candidates were genotyped in the Taiwanese population for further validation and to explore the effect of these variants in a population with a different genetic background. A SNP on chromosome 6q16 located approximately 38 kb downstream of EPHA7 (EPH receptor A7), rs6901416, displayed a similar effect on overall survival in the Taiwanese population as the U.S. populations (HR, 2.42; 95% CI, 1.66–3.52; combined $P = 4.10 \times 10^{-6}$). This variant was associated with a 1.44-fold increased risk (95% CI, 1.01–2.06; $P = 0.043$) under the recessive model (Table 2) in the Taiwanese population and resulted in at least an 8-month reduction in median survival time in the three populations (Table 3; Fig. 2A). An additional variant, rs10766739 [chromosome 11p15.1, intron of NELL1 (Nel-like 1)], also one of the top loci in the U.S. populations (combined $P = 3.66 \times 10^{-7}$), reached borderline significance with $P = 0.051$ (HR, 1.23; 95% CI, 1.00–1.51) in the Taiwanese population and had a significant effect on reduction of median survival time by at least 6 months in MDACC and Taiwanese populations (log-rank $P, \text{MDACC} = 0.029, \text{Taiwan} = 0.024$; Fig. 2B). However, it would be expected to have one or two variants reach statistical significance ($P < 0.05$) by chance when testing 25 individual SNPs. Therefore, these candidate variants will need to be further analyzed in other Asian populations to solidify their role in mediating overall survival for never smokers with lung cancer.

To identify potential prognostic loci based on phenotypic validation, we conducted cis-eQTL analysis in 600 kb regions flanking each of the 25 candidate SNPs. We observed six SNPs that showed a genotype-expression association with nearby genes (Table 2). Three of these variants reached genome-wide significance at $10^{-7}$ in the initial consistency meta-analysis (rs7976914, rs4237904, and rs4970833). rs7976914 and rs4237904 are approximately 48 kb apart and in strong linkage disequilibrium ([LD]) $r^2 = 0.88$ representing a single unique locus. Subjects with the rare homozygous genotype for either locus
had a more than 2.6-fold significantly increased risk of death in the consistency meta-analysis (HR, 2.67; 95% CI, 1.86–3.85 and HR, 2.77; 95% CI, 1.89–4.06, respectively) as compared with subjects with the common homozygous/heterozygous genotype. Furthermore, patients with the variant genotype for either SNP had significant reductions in survival times of 14.21 months ($P = 1.52 \times 10^{-4}$) and 16.15 months ($P = 0.0017$), respectively, in the MDACC population with similar dramatic reductions in the Mayo population (Fig. 2C and D; Table 3). Both SNPs are located in the middle of an approximately 300-kb gene desert region on chromosome 12, yet genotypes for these variants for the Mayo population were associated with $TMBIM4$ ($P = 0.013$ and 0.014, respectively) and $LEMD3$ expression ($P = 0.027$ and 0.022, respectively). The SNPs are over 500 kb from the 3′-UTR of $TMBIM4$, a gene encoding for transmembrane BAX inhibitor motif containing 4, and over 325 kb to the 3′-UTR of $LEMD3$ (LEM domain containing 3). The third significant genotype-expression association for a variant for which the $P$ value reached $10^{-7}$ in the consistency meta-analysis (HR, 1.81; 95% CI, 1.43–2.29) was for rs4970833 located within intron 5 of $CELSR2$, a gene encoding a receptor with EGF-domains. However, rs4970833 genotypes were associated with $ATXN7L2$ (ataxin 7-like 2) expression ($P = 0.037$) located 222 kb upstream.

Similarly, an additional 3 SNPs located within known genes had significant eQTL relationships with distant genes (Table 2). rs954785 was associated with a 1.62-fold increase in risk in the consistency meta-analysis (95% CI, 1.32–1.98) and corresponding significant decreases in median survival time for both the U.S. populations ($P = 0.038$ and 0.046, respectively; Table 3; Fig. 2E). Interestingly, this effect was in the opposite direction for the Taiwanese population (HR, 0.46; 95% CI, 0.23–0.92; $P = 0.029$; Table 2). This variant is located in an intron of $KCNN3$, a potassium channel, yet a significant genotype–phenotype relationship was identified with $SHE$ (Src homology 2 domain containing E). $P = 3.19 \times 10^{-5}$, located over 195 kb 3′ of $KCNN3$. $FLJ45983$ encodes for a gene with unknown function and harbors a nonsynonymous variant, rs485411 that results in a His–Arg substitution. This variant resulted in a 1.44-fold increase in risk (95% CI, 1.23–1.69). This SNP was not associated with $FLJ45983$ expression, but expression of $ITIH2$ (inter-$\alpha$-trypsin inhibitor heavy chain 2; $P = 7.10 \times 10^{-5}$) nearly 302 kb away on chromosome 10. The final eQTL was identified between rs10906104 (consistency meta-analysis HR, 2.37; 95% CI, 1.62–3.47) and $NUDT5$ [nudix (nucleoside diphosphate–linked moiety X)-type motif 5; $P = 0.034$]. rs10906104 is an intronic variant in $CDC123$, a cell division cycle protein, and located over 42 kb 5′ from $NUDT5$. Interestingly, all six genes were significantly differentially expressed in 70 tumor tissues as compared with adjacent normal tissues (Table 2).

Discussion

Although tobacco use has been established as the main etiologic factor for lung cancer, a substantial fraction of lung cancer cases involves never smokers. A wide range of survival rates seen in patients of this population suggests heterogeneity of prognoses that may be independent of cancer type and disease stage. To identify genetic loci that are associated with overall survival in never smokers with NSCLC, we conducted the first dual genome-wide scan, consistency meta-analysis in 327 never smokers recruited from MDACC and 293 never smokers at Mayo Clinic. An additional validation analysis was conducted in a group of 1,256 NSCLC cases in never smokers from Taiwan, and through exploration of eQTL and differential expression in tumor and adjacent normal lung tissues. This multiphase

**Table 3.** Survival durations for significant SNPs by study site

<table>
<thead>
<tr>
<th>SNP</th>
<th>Genotype</th>
<th>MDACC MST</th>
<th>$P$</th>
<th>Mayo Clinic MST</th>
<th>$P$</th>
<th>Taiwan MST</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs6901416</td>
<td>TT+TG</td>
<td>25.3</td>
<td>0.15</td>
<td>&gt;36.0</td>
<td>0.026</td>
<td>28.1</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>GG</td>
<td>17.3</td>
<td></td>
<td>22.0</td>
<td></td>
<td>18.7</td>
<td></td>
</tr>
<tr>
<td>rs10766739</td>
<td>GG+GA</td>
<td>25.3</td>
<td>0.029</td>
<td>&gt;36.0</td>
<td>0.42</td>
<td>28.3</td>
<td>0.024</td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>9.8</td>
<td></td>
<td>31.3</td>
<td></td>
<td>22.3</td>
<td></td>
</tr>
<tr>
<td>rs4237904</td>
<td>GG+GT</td>
<td>26.6</td>
<td>1.52x10^-4</td>
<td>&gt;36.0</td>
<td>0.19</td>
<td>27.5</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>TT</td>
<td>12.4</td>
<td></td>
<td>20.4</td>
<td></td>
<td>28.3</td>
<td></td>
</tr>
<tr>
<td>rs7976914</td>
<td>AA+AC</td>
<td>26.0</td>
<td>0.0017</td>
<td>&gt;36.0</td>
<td>0.057</td>
<td>27.8</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>9.9</td>
<td></td>
<td>16.8</td>
<td></td>
<td>27.6</td>
<td></td>
</tr>
<tr>
<td>rs4970833</td>
<td>GG</td>
<td>28.5</td>
<td>0.30</td>
<td>&gt;36.0</td>
<td>0.014</td>
<td>29.1</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>GG+CC</td>
<td>22.7</td>
<td></td>
<td>34.1</td>
<td></td>
<td>27.1</td>
<td></td>
</tr>
<tr>
<td>rs954785</td>
<td>TT</td>
<td>30.4</td>
<td>0.038</td>
<td>&gt;36.0</td>
<td>0.046</td>
<td>27.7</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>TG+GG</td>
<td>18.9</td>
<td></td>
<td>32.4</td>
<td></td>
<td>33.8</td>
<td></td>
</tr>
<tr>
<td>rs485411</td>
<td>CC</td>
<td>28.5</td>
<td>0.084</td>
<td>&gt;36.0</td>
<td>0.57</td>
<td>28.1</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>CT</td>
<td>22.3</td>
<td></td>
<td>35.9</td>
<td></td>
<td>25.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TT</td>
<td>12.7</td>
<td></td>
<td>34.1</td>
<td></td>
<td>29.4</td>
<td></td>
</tr>
<tr>
<td>rs10906104</td>
<td>GG+GA</td>
<td>26.6</td>
<td>0.0033</td>
<td>&gt;36.0</td>
<td>0.088</td>
<td>28.1</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>16.2</td>
<td></td>
<td>22.0</td>
<td></td>
<td>27.7</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: MST, median survival time in months.
approach incorporating a total of 1,876 NSCLC cases and phenotypic analysis identified several novel loci that are associated with overall survival in this distinct patient population. One of the challenges of GWAS has been the investigation of the genetic effects in non-European populations. It is well established that population substructure can mask true associations that are population specific, necessitating restriction to a single ethnic/racial population to avoid this effect. However, it is also known that NSCLC manifests differently in different populations, particularly in individuals from East

Figure 2. Kaplan-Meier 3-year survival curves for selected genetic loci in the three study populations. rs6901416 (A), rs10766739 (B), rs4237904 (C), rs7976914 (D), and rs954785 (E). MST, median survival time in months.
Asia (30). Therefore, we assessed the effect of the top 25 candidate variants identified from the European-descent GWAS studies in a large set of never-smoker NSCLC cases from Taiwan. Two variants, rs6901416 and rs10766739, were both associated with a 1.2- to 1.5-fold increase in risk and corresponding reduction in survival durations in the Asian population. However, as stated in the results, only these two of 25 variants genotyped replicated, which is no more than would be expected by chance and these results should be viewed with this in mind. Nevertheless, these two variants are candidates for further consideration. rs6901416 is located at 6q16.1, approximately 58 kb downstream of EPHA7. Ephrin receptors are receptor tyrosine kinases that are known to play a role in many facets of cancer and many cancer sites (31–34). EPHA7 is highly expressed in lung cancer, and a secreted form of EPHA7, generated from a splice variant, was also detected and seemed to be unique to lung cancer (35). rs10766739 was one of the top 7 loci in the U.S. populations (combined P = 3.66 × 10^{-7}) and is located in an intron of NELL1 on chromosome 11p15.1. NELL1 encodes a secreted growth factor containing EGF-like repeats. Significantly, Jin and colleagues have reported that hypermethylation of NELL1 was a common and early event and was associated with poor prognosis in early-stage esophageal adenocarcinoma (36). Compared with normal tissue, methylation of NELL1 promoter increased with pathologic progression from Barrett metaplasia to esophageal adenocarcinoma and correlated with decrease survival in stage I and II patients. A genome-wide scan also identified epigenetic silencing of NELL1 in colon cancer cell lines (37). Taken together, these studies suggest NELL1 as a candidate tumor suppressor gene. No significant eQTL relationships were observed for either of these SNP-gene pairs.

Interestingly, the most significant SNP in the Taiwanese replication (rs954785) had an opposite effect from that in the U.S. populations, further emphasizing the differences variants can have in different populations. Together, the findings from the Taiwanese population suggest that there may be commonality among risk factors in a European-descent population and an Asian population, with each also having unique genetic profiles influencing clinical outcomes in never smokers with lung cancer. The results from the Taiwanese population will require further validation in an independent Asian population and additional follow-up for confirmation of the findings presented here.

eQTL analysis validated six variants as being associated with overall survival through the identification of genotype–phenotype relationships. rs4237904 and rs7976914 were highly significant in the consistency meta-analysis and associated with 14 to 20 month reduction in 3-year survival durations. Although they are located in a large gene desert, the expression of two genes in the vicinity of 12q14, LEMD3 and TMBIM4, were found to be significantly correlated with rs4237904 and rs7976914 genotypes (Table 2). Both SNPs are located 3’ distal of these two genes (300–500 kb downstream). Because the two SNPs show strong LD and are relatively close to each other, it is possible that they might influence nearby cis-regulatory elements that can affect the transcription of the two surrounding genes from a distance. LEMD3 (also known as MAN1) encodes a LEM domain-containing protein that can antagonize TGF-β signaling (38). Despite its role in altering TGF-β signaling, association of this gene with cancer or its alteration in tumorigenesis has not been previously reported. TMBIM4 (or GAAP) is a novel antiapoptotic regulator that belongs to a family of Bax-inhibitor-1 proteins that can bind to Bcl-2 family members and regulate cell death (39). TMBIM4 is ubiquitously expressed in all tissues, including the lungs. Because the endogenous functions of both aforementioned genes are relatively unstudied, their influence on cancer development and progression remains to be characterized. Moreover, both TMBIM4 and LEMD3 were found to be significantly differentially expressed between lung tumor and adjacent normal tissue, suggesting the potential involvement of this locus in lung tumorigenesis.

eQTL analysis also identified a correlation between genotypes for rs4970833 and ATXN7L2 expression. The variant is an intronic SNP within CELSR2, a receptor with EGF domains and member of the flamingo subfamily of cadherins, on chromosome 1p21. CELSR2 was previously identified in multiple GWAS studies as being a candidate gene for various cardiovascular phenotypes and blood lipid levels (40–42). No studies have investigated CELSR2 in the setting of cancer. However, no eQTL association was evident between CELSR2 and the variant. ATXN7L2 encodes for the ataxin 7-like 2 protein and is a member of the ataxin family. The exact function of ataxin 7 and ataxin 7–like 2 is not clear. The current study provides the first hint of ATXN7L2’s potential role in cancer, which was provided by the significantly differential expression of the gene in lung cancer tumors and adjacent normal tissue samples (Table 2).

KCNN3 is a small conductance calcium-activated potassium channel that has been shown to enhance the migration ability of melanoma cells and thus metastatic ability. An intronic variant in this gene, rs954785, was identified in the consistency meta-analysis as being highly significant for poor overall survival and significantly reduced median survival time. rs954785 was not associated with KCNN3 expression but was associated with expression of SHE. SHE has not been well characterized and little is known about its function, except for containing a Src homology 2 domain (SH2). This current study is the first to link SHE with cancer.

Another gene with unknown function, FLJ495983, was implicated in mediating overall survival in never smokers due to the significant association of a nonsynonymous variant, rs485411, with a poor outcome. This gene is close to GATA3, which encodes for a transcription factor that has been implicated in the development of many cancers, particularly luminal breast cancer (reviewed in ref. 43). eQTL analysis did not identify significant relationships between rs485411 and either of these two genes. The significant association was with ITIH2 located nearly 302 kb telomeric. ITIH2, also known as SHAP, complexes with hyaluronan (HA) and plays a role in the inflammatory response. Serum levels of SHAP-HA have been associated with a poor outcome in ovarian and endometrial cancers (44, 45). High Glasgow scores, a measure of systemic inflammation, have been shown to be associated with poor outcomes for NSCLC in both smokers and never smokers, although more pronounced in smokers. It is possible that rs485411 is
mediating ITH2 and thus altering the levels of inflammation that would result in a poor outcome.

rs10906104, located in an intron of CDC123, was highly significant in the consistency meta-analysis and was an eQTL for NUDT5 located over 42 kb away and not with CDC123 expression. NUDT5 is a hydrolase that functions to remove toxic nucleotide derivatives from the cell to protect from mutagenesis (46). This includes 8-hydroxy-dGTP, a known mutagen (47), which is produced by reactive oxygen species in the cell. Tumor levels of 8-hydroxy-dGTP incorporated into DNA were found to be predictive of survival in NSCLC with lower 8-hydroxy-dGTP being associated with a better prognosis (48). Studies correlating rs10906104 and 8-hydroxy-dGTP levels would be of interest to determine the link between rs10906104, NUDT5, and 8-hydroxy-dGTP.

The significant eQTL relationships identified in this study were not between the identified SNP and the host gene; instead, they were often for genes some distance from the genetic variant. However, the lack of eQTL associations between these host genes and the variants does not preclude them as playing a role in overall survival. eQTL analysis only investigates one mechanism by which a variant can have a phenotypic effect, not taking into consideration potential effects on splicing or allelic imbalance or more subtle effects through gene regulatory elements that may not reach a significance threshold. Future studies are warranted to explore these relationships in greater detail. Furthermore, although associations were identified between candidate SNPs and both overall survival and cis-gene regulation, we cannot prove causality or functionality of the genotyped SNP in mediating variation in survival. The true causal variant(s) may be in linkage disequilibrium with the observed candidate SNPs. Future experiments are needed to explore this possibility and determine the underlying mechanisms responsible for the association. However, the observed associations with cis-gene regulation provides additional support that these genomic regions identified initially through GWAS play a role in variation in overall survival in never smokers with NSCLC.

Previous genome-wide studies of germline genetic variants have generally focused on cancer predisposition rather than the identification of prognostic markers due to the requirement for large sample sizes and the likely presence of subject heterogeneity in clinical characteristics and treatment regimens. In addition, very little overlap between novel loci for susceptibility and clinical outcomes (survival) has been identified through both candidate gene and genome-wide approaches. This suggests that there are distinct biologic pathways that mediate the development of disease from response to therapy and clinical outcomes. Recently, three genome-wide studies have been reported with the goal of identifying germline genetic variants associated with survival for NSCLC (25, 26, 49). Candidate loci were found in several genes: STK39, PCDH7, A2BP1, and EYA2 for early-stage lung cancer and CMKLRI, EIF4E2, ETS2, and DSCAM for advanced disease. We did not observe any significant associations with survival for SNPs within the genes implicated in these previous studies, suggesting that the genetic influence on mortality may be different for smokers and never smokers. This would be in line with the hypothesis that lung cancer in never smokers is a distinct disease from that in smokers. Alternatively, variances in population structure and/or clinical characteristics could contribute to differences in risk association.

In this study, we conducted a consistency meta-analysis using genome-wide scan data from two independent populations consisting of individuals of European-descent followed by a two-pronged validation that included replication in a non-European population and analysis of eQTL relationships to minimize potential for false discovery. All populations had detailed clinical information and were similar in clinicopathologic characteristics, which minimized study site heterogeneity and allowed us to adjust for potential confounding variables. Moreover, Q–Q plots of the observed and expected associations in the studies indicate little deviation from the identity line except for the tail end, with inflation factors for the studies being 1.175 and 1.094 for MDACC and Mayo, respectively. This provides confidence in the quality and robustness of the consistency meta-analysis. Together, these results identify potential prognostic loci for NSCLC in never smokers and implicate several novel genes as having a role in influencing lung cancer survival among never smokers.

Disclosure of Potential Conflicts of Interest

J.D. Minna has a commercial research grant from Genentech Pharmaceuticals. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions

Conception and design: X. Wu, G.-C. Chang, P.-C. Yang, R.S. Marks, V.S. Pankratz, M.A.T. Hildebrandt, C.A. Hsiung, P. Yang

Development of methodology: X. Wu, G.-C. Chang, P.-C. Yang, P. Yang

 Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): X. Wu, L. Wang, G.-C. Chang, P.-C. Yang, J.A. Roth, R.S. Marks, J.Y. Chang, C. Lu, C. Deschamps, W.-C. Su, M.-S. Huang, Y. Li, C.A. Hsiung, P. Yang

Analysis and interpretation of data (e.g., statistical analysis, bios-statistics, computational analysis): L. Wang, Y. Ye, J.A. Aakre, X. Pu, P.-C. Yang, J.A. Roth, S.M. Lippman, W.-C. Chang, Y. Li, V.S. Pankratz, M.A.T. Hildebrandt, C.A. Hsiung, P. Yang

Writing, review, and/or revision of the manuscript: X. Wu, X. Pu, G.-C. Chang, J.A. Roth, S.M. Lippman, J.Y. Chang, C. Lu, C. Deschamps, D.W. Chang, V.S. Pankratz, J.D. Minna, M.A.T. Hildebrandt, C.A. Hsiung, P. Yang

 Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): X. Wu, Y. Ye, J.A. Aakre, G.-C. Chang, P.-C. Yang, M.-S. Huang, Y. Li, C.A. Hsiung, P. Yang

 Study supervision: X. Wu, W.K. Hong, P. Yang

Grant Support

This study was supported by U.S. NIH grants R01-CA111646 (X. Wu), P50CA070907 (J.D. Minna, J.A. Roth, and X. Wu), R01-CA127615 (X. Wu), R01-CA80127 (P. Yang), R01-CA115857 (P. Yang), and R03-CA77118 (P. Yang). The GELAC study was supported by D0H98-TD-0-111-015 from the National Research Program on Genomic Medicine (NRPGM) in Taiwan and D01H00-TD-PB-111-TM013 (C.A. Hsiung) from the National Research Program for Biopharmaceuticals (NRPB) in Taiwan. Additional support was provided by the Center for Translational and Public Health Genomics of the Duncan Family Institute for Cancer Prevention and Risk Assessment (X. Wu), MD Anderson Cancer Center Research Trust (X. Wu), and Mayo Foundation Funds (P. Yang). Mayo Genomic Analyses Shared Resources directed by Drs. Julie Cunningham and Jin Jen carried out the genotyping and gene expression experiments for the Mayo Clinic study.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received October 31, 2012; revised March 19, 2013; accepted April 5, 2013; published OnlineFirst May 23, 2013.
References

43. Chou J, Provot S, Werb Z. GATA3 in development and cancer differ-
entiation: cells GATA have it! J Cell Physiol 2010;222:42–9.
Clinicopathological role of serum-derived hyaluronan-associated pro-
tein (SHAP)-hyaluronan complex in endometrial cancer. Obstet Gyne-
hyaluronan complex in ovarian cancer. Oncol Rep 2008;19: 
1245–51.
46. Hori M, Satou K, Harashima H, Kamiya H. Suppression of mutagenesis 
by 8-hydroxy-2′-deoxyguanosine 5′-triphosphate (7,8-dihydro-8-
oxo-2′-deoxyguanosine 5′-triphosphate) by human MTH1, MTH2, and 
47. Satou K, Kawai K, Harashima H, Kamiya H. Mutagenic effects of 
8-hydroxy-dGTP in live mammalian cells. Free Radic Biol Med 
48. Shen J, Deininger P, Hunt JD, Zhao H. 8-Hydroxy-2′-deoxyguanosine 
(8-OH-dG) as a potential survival biomarker in patients with nonsmall-
Genome-wide association study on overall survival of advanced non-
small cell lung cancer patients treated with carboplatin and paclitaxel. 