Cigarette Smoking Increases Copy Number Alterations in Non-Small Cell Lung Cancer

Yen-Tsung Huang[1,2], Xihong Lin[2], Yan Liu[3], Lucian R. Chirieac[4], Ray McGovern[2], John C. Wain[5,6], Rebecca S. Heist[5,7], Vidar Skaug[8], Shanbeh Zienolddiny[8], Aage Haugen[8], Li Su[7], [9]Edward A. Fox, Kwok-Kin Wong[3] and David C. Christiani[1,7,10]

Affiliations: Departments of [1]Epidemiology, [2]Biostatistics and [7]Environmental Health, Harvard School of Public Health, 667 Huntington Avenue, Boston, MA 02115; [3]Department of Medical Oncology and [9]Molecular Diagnostics Laboratory, Dana-Farber Cancer Institute, 44 Binney Street Boston, MA 02115; [4]Department of Pathology, Brigham and Women's Hospital, 75 Francis Street, Boston, MA 02115; [5]Cancer Center, [6]Thoracic Surgery Unit and [10]Pulmonary and Critical Care Unit, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; and [8]Department of Biological and Chemical Working Environment, National Institute of Occupational Health, P.O. Box 8149 Dep, N-0033 Oslo, Norway.

Corresponding author: David C. Christiani, MD, MPH, Department of Environmental Health, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115; e-mail: dchris@hsph.harvard.edu

Classification:

Major: Biological Sciences; Minor: Genetics

ABSTRACT

Cigarette smoking has been a well-established risk factor of lung cancer for decades. How smoking contributes to tumorigenesis in lung remains not fully understood. Here we report the results of a genome-wide study of DNA copy number and smoking pack-years in a large collection of non-small cell lung cancer (NSCLC) tumors. Genome-wide analyses of DNA copy number and pack-years of cigarette smoking were performed on 264 NSCLC tumors, which were divided into discovery and validation sets. The copy numbers-smoking associations were investigated in three scales: whole-genome, chromosome/arm and focal regions. We found that heavy cigarette smokers (>60 pack-years) have significantly more copy number gains than non-/light smokers (≤60 pack-years) ($p=2.46\times10^{-4}$), especially in 8q and 12q. Copy number losses tend to occur away from genes in non-/light smokers ($p=5.15\times10^{-5}$) but not in heavy smokers (p=0.52). Focal copy number analyses show that there are strong associations of copy number and cigarette smoking pack-years in 12q23 ($p=9.69\times10^{-10}$) where *IGF1* (insulin-like growth factor 1) is located. All of the above analyses were tested in the discovery set and confirmed in the validation set. DNA double-strand break assays using human bronchial epithelial cell lines treated with cigarette smoke condensate (CSC) were also performed, and indicated that CSC leads to genome instability in human bronchial epithelial cells. We conclude that cigarette smoking leads to more copy number alterations, which may be mediated by the genome instability.

\body INTRODUCTION

Lung cancer, of which 85% is non-small cell lung carcinoma (NSCLC), is the second most common cancer and the leading cause of cancer-related death in the United States.(1) The epidemiologic evidence supporting that cigarette smoking is an important factor in causing lung cancer was reported almost six decades ago.(2-4) Moreover, lung cancer mortality mirrors trends in tobacco use(5). Carcinogens derived from cigarette smoking damage lung epithelium by oxidative stress and direct DNA damage.(6) Although there has been progress in our understanding of lung carcinogenesis over the past few decades, the knowledge of mechanisms by which cigarette smoking causes lung cancer remains incomplete.

Profiles of copy number alterations (CNAs) in NSCLC have been studied.(7, 8) However, what causes copy number (CN) changes remain unknown. Several mechanisms of copy number changes have been proposed including homologous recombinations and non-homologous mechanisms.(9, 10) Bacteria, yeast and human seem to share similar mechanisms.(10) In bacteria, copy number alterations can be induced by environmental stress to enable swifter evolution in response to such stress. In the cell population within a tumor or precancerous lesion, similar stress such as hypoxia may induce copy number change. Thus, it is plausible to hypothesize that cigarette smoking serves as an environmental stress on the cells that leads to tumorigenesis by means of copy number alterations.

Using the tumor cells separated from malignant pleural effusions, it was found that gains of 11p were more frequent in smoking men than non-smoking men.(11) Furthermore, another study identified a copy number-based genomic signature in resected lung tumors for current smokers as compared to never smokers.(12) However, these studies had significant limitations. First, discrete smoking status (smokers vs. non-smokers) may not be an optimal indicator to capture the dose-response relationship between cigarette smoking and copy number changes. Second, smoking may have different implications on copy number depending on whether it induces gains or losses. Third, the conclusions in the previous studies were drawn based on modest sample sizes. Lastly, none of previous studies provide a biological explanation on how cigarette smoking causes CNAs. To better investigate the relationship between cigarette smoking and copy number alterations, we conducted a genome-wide study of copy numbers and smoking pack-years in a large collection of resected NSCLC tumors. Our analyses cover the association of cigarette smoking with copy numbers on three different scales: whole-genome, chromosome/arm and focal copy numbers. The causal mechanism behind such smoking-CNAs association was further explored in a human non-tumorigenic bronchial cell line.

RESULTS

A total of 264 subjects were randomly divided into two data sets: discovery and validation sets. The characteristics of the populations are similar (Table 1), indicating the balance of the two sets. Two alternative data splittings were pursued to prevent from the possibility that the results presented here are simply due to chance or to multiple comparisons. (SI Appendix (Tables S1-S3)) To account for batch effects, we also performed batch-adjusted analyses by normalization and explicitly adjusting for the batch identity as a covariate in the regression. The batch effect-adjusted analyses showed similar patterns to those without adjustment. (SI Appendix (Tables S4 and S5; Figure S1)) The analyses of smoking vs. copy number associations are outlined as three parts: on the genome-wide scale, on the chromosome/arm specific scale and on the focal region scale.

Cigarette smoking and whole-genome copy number pattern

There is a significant increase in total events of copy number gains among heavy smokers (>60 pack-years) (p=0.0080, 0.0095, and 2.5×10^{-4} for discovery, validation and both sets, respectively), but no difference in copy number losses (Figure 1A and 1B). No significant difference was observed in age, clinical stage, histology and gender between heavy and light/non-smokers.

For copy number losses, G/T ratios in light/non-smokers (≤60 pack-years) are significantly lower than the null ratio (i.e., the ratio when CNAs occur at random with respect to the gene location) (p=0.011, 9.80×10^{-4}, and 5.15×10^{-5} for discovery, validation and both sets, respectively) but heavy smokers (>60 pack-years) show no difference (p=0.78, 0.31 and 0.52, respectively) (Figure 1C and 1D). These results suggest that copy number losses tend to occur away from genes but such tendency disappears in heavy smokers. In contrast, there is no consistent pattern for copy number gains. Heavy smokers seem to have more genes with copy number changes, especially in gains. (SI Appendix (Figure S2))

Cigarette smoking and copy number pattern by chromosome/arm

The chromosome/arm-specific analyses suggest the majority responsible for the genome-wide difference comes from chromosomes 8q (p $=1.19\times10^{-5}$ for total events of CN gains between light and heavy smokers) and 12q (p $=2.1\times10^{-4}$) (Figure 2) as well as many others (chromosomes 1, 3, 7, 10, 11, 16 and 17) (SI Appendix (Figure S3)). Similar results were observed when genomic location was taken into account, especially in 8q and 12q. (SI Appendix (Figure S4A)) The dose-response relationships between continuous copy numbers and smoking pack-years are also significant in 8q ($p=0.015$) and 12q ($p=0.0025$). (SI Appendix (Figure S4B)) These two regions are also found the most signals in focal copy number analyses, as will be shown in the following.

Cigarette smoking and focal copy numbers

As stated in Materials and Methods, we performed single- and multiple-marker analyses to investigate the association of cigarette smoking and focal copy numbers. In the moving window 10-marker analyses, we selected the top 50 sets with smallest p values in the discovery set and tested the 50 sets using the validation set ($p<0.05$). Using such criteria, we identified one 10-marker set in 12q23 with p values of 9.69×10^{-10}, which reached the genome-wide significance. (Figure 3A) The region harbors a gene, *IGF1* (insulin-like growth factor 1) that plays an important role in tumorigenesis. (Figure 3B) In the single-marker analyses, the most significant signals are also in the same region of 12q23: two loci are in the intron between the last two exons of *IGF1* and two loci are located downstream of *IGF1*. (SI Appendix (Figure S5A and S5B, and Table S6)) The p value of the 10-marker set identified in the 10-marker analyses and the corresponding p values and R^2 from the single-marker analyses were shown in Table 2. Compared with the single-marker analyses, statistical power was gained from the 10-marker analyses by borrowing information in the neighboring markers, accounting for correlation among the CNVs in the marker set, reducing degrees of freedom of the test, and reducing the total number of tests across the genome.

The dose-response relationship of copy number and smoking pack-years for the four loci in 12q23.2 is shown in SI Appendix (Figure S5C-F), indicating a J-shape curve. That is, beyond a certain threshold, the higher smoking pack-years, the more departure from the neutral copy number. Notably, the threshold, about 60 pack-years, is consistent with the cut-off used in the above analyses of whole-genome CNAs pattern.

In addition to 12q23, 3q24 and 8q24 are two additional regions that are potentially associated with the pack-years of cigarette smoking from single-marker analyses. (SI Appendix (Figures S6 and S7)) We also performed the analyses in the dichotomous version, detail of which can be found in SI Appendix. (Table S7 and Figure S8)

DNA double-strand break assay

To investigate further the results of our statistical analyses, we determined whether cigarette smoke could induce DNA double-stand breaks in cultured cells. To mimic longer and heavier cigarette smoking conditions, we treated human nontumorigenic bronchial epithelial cell HBEC 3KT with 0.04 and 0.4 μg/ml CSC for 24 hours. Under these conditions, the survival rates are 96.9% and 95.7%, respectively, indicating the dose of CSC and the length of treatment used in this study are not toxic to the cells (Figure 4A). To minimize background DNA double-strand breakage, we treated cells with CSC right after the growth had reached confluence. Under these conditions, ~5% of non-CSC treated cells still display double stand breaks (Figure 4C). When treated with 0.04 μg/ml CSC for 24 hours, the percentage of cells with double strand breaks increased to 15%. This percentage doubled with the application of more concentrated 0.4 μg/ml CSC (Figure 4C). We also treated the cells with 0.4 μg/ml CSC for 2 hours, and observed a similar DNA double-strand break ratio as that of the non-CSC treated control cells, suggesting DNA double-strand break occurring after a longer time of CSC treatment.

To determine the effects of CSC on induction of cellular apoptosis which indirectly contributes to DNA double-strand breaks, the same set of cells (as used in Figure 4B and 4C) were lysed for apoptotic-specific Caspase-3/7 activity. As shown in Figure 4D, there is a basal level of Caspase-3/7 activity in non-CSC treatment cells. Upon CSC treatment, the value of relative fluorescence unit (RFU) increased in a dose-dependent matter. However, the extent to which the RFU value increased in response to CSC treatment was much less than the corresponding increase in DNA double strand breaks in Figure 4C. Collectively, these results indicate that higher CSC leads to genome instability in bronchial epithelial cells. As such, theses data provide biological evidence to bridge the associations between CNAs and smoking observed in the above human data.

DISCUSSION

We show that heavy smokers (>60 pack-years) have more copy number gains than light/non-smokers but not copy number losses and that light/non-smokers (≤60 pack-years) have copy number losses away from the gene location, in contrast to heavy smokers. The discrepancy between gains and losses suggests that different mechanisms may exist for the genome impact of cigarette smoking. For gains, smoking executes its oncogenic effect by increasing the event of copy number changes. For losses, in contrast, smoking does not increase CNAs events but increase the proportion of genes being affected. Because losing a fragment of DNA is less favorable than gaining one(13), two separate mechanisms may be developed to hit the genes responsible for tumorigenesis.

The phenomenon may be a consequence of selection during cancer development and cell proliferation. Because different cells possess different CNAs, selection by a nutrient-limited environment makes those clones that can grow without regulatory control become dominant.

For heavy smokers, there were more copy number gains compared to non- or light smokers and no tendency for copy number losses to occur away from the gene location. We have also found that genes with gains are more likely to be oncogenes or to be involved in pathways that are associated with tumor growth, which suggests that lung cancer cells in heavy smokers tend to acquire the growth advantage via copy number gains.(14) As a result, copy number losses within genes have less unfavorable impact on such cells since it is compensated by the fact that they can grow without regulation. This explains our observation that the proportion of losses within gene among heavy smokers is not different from that at random.

Previous studies have shown that copy number alterations are more frequent in smokers than in non-smokers,(11, 12) consistent with our findings based on pack-years. Copy number-based genomic signature has also been identified to discriminate current smokers and never smokers(12), which, however, does not include *IGF1*. Smoking status may not necessarily reflect the same oncogenic feature as pack-years of smoking, a measure of cumulative exposure. Furthermore, the large sample size and discovery-validation process in this study increase robustness of the findings.

Smoking causes lung cancer through numerous carcinogens derived from cigarette combustion. There are two parts of the carcinogenic effect: early damage of oxidative stress by reactive oxygen species and late damage by DNA adduct and DNA mutation.(6) Both kinds of damage can serve as initiators of copy number changes, especially oxidative stress. It has recently been proposed that cellular stress coming from environmental agents can induce copy number changes, which seems to be a common mechanism in bacteria, yeast and humans.(10)

The most significant region on 12q23 is at the junction of the last two exons and the downstream of *IGF1*. The two loci within *IGF1* are located in the intron between the last two exons of *IGF1*. The protein product of the aberrant genomic DNA can exert its undue influence on the cellular physiology. On the other hand, if the key player is the downstream rather than the coding region of *IGF1*, it is still possible that *IGF1* function is affected because the downstream fragment can serve as a regulatory element of *IGF1* transcription. That is, the alterations of the regulatory element can lead to the abnormal gene expression of *IGF1*.

IGF1/*IGF1R* signaling pathway can induce many effects, including cell proliferation, differentiation, transformation and inhibition of apoptosis.(15) Because of the overlap with downstream signaling pathways of epidermal growth factor receptor (*EGFR*) signaling, *IGF1*/*IGF1R* signaling may modulate the *EGFR* pathway, a critical pathway in lung tumorigenesis(16), and it may explain, in part, clinical resistances to *EGFR* inhibitors(17).

Several studies have provided the links among smoking, *IGF1* and cancer. For example, it has been reported that smoking may affect *IGF1* serum level and its signaling.(18, 19) On the other hand, *IGF1* and the risk of developing cancer have also been extensively studied in lung cancer(20-23), breast cancer(24), prostate cancer(25) and colorectal cancer(26). Our analysis supports the hypothesis that smoking can act through increasing the copy number of *IGF1* to induce its over-expression and subsequent oncogenesis.

MATERIALS AND METHODS

Study population, specimens and data collection

A series of 264 snap-frozen tumor samples from NSCLC patients with complete information on cigarette smoking was collected during surgery or biopsy from the Massachusetts General Hospital (MGH), Boston, MA, and the National Institute of Occupational Health, Oslo, Norway. An additional 50 paired specimens of non-neoplastic lung parenchyma and 63 paired blood samples were included as the reference group for copy number estimation. Demographic and smoking information was collected by a trained research assistant using a modified standardized American Thoracic Society respiratory questionnaire.(27) A similar approach was used for the Norwegian cohort.(28) Written informed consents were obtained from all patients. The study was approved by the institutional review boards of MGH, the Harvard School of Public Health, and the Norwegian Data Inspectorate, and Local Regional Committee for Medical Research Ethics.

DNA quality, histopathology and genechip

DNA was extracted from tumor and non-neoplastic lung parenchyma after manual microdissection from 5-μm thick histopathologic sections. Each specimen was evaluated for amount and quality of tumor cells. Tumors were reviewed and classified using the WHO criteria. Specimens with lower than 70% tumor cellularity, inadequate DNA concentration, or not intact genomic DNA were not included for chip hybridization. The platform of genechip is Affymetrix 250K Nsp GeneChip®.

Data preprocessing

Copy numbers were obtained with dChip software by invariant set normalization and median smoothing with the window of 11 loci.(29) Only 256,554 probes on somatic chromosomes were analyzed. We further classified the continuous inferred copy number into a discrete variable of CNAs: copy number gains defined as copy numbers ≥2.7 and copy number losses defined as copy numbers ≤1.3, to detect copy number ≥3 and ≤1 by tolerating 30% normal tissue contamination. The probes were mapped to the RefSeq genes with 2 kb extension both upstream and downstream using the UCSC Genome Browser. Among the 256,554 probes on somatic chromosomes, 104,256 (40.64%) were mapped to 11,700 genes.

Statistical analysis

Only early stage tumors were analyzed here because we have found that late stage tumors have more CNAs. The number of pack-years is defined as the packs of cigarette smoked per day multiplied by the years of smoking. 60 pack-years of cigarette smoking was chosen as the cut-off for heavy and light/non-smokers according to the observation of total CNAs events by the interval of 10 pack-years in both discovery and validation sets (SI Appendix (Figure S9)). Using the cut-off, we had 203 light/non-smokers and 61 heavy smokers. We developed three methods to test the genome-wide or chromosome/arm-specific copy number patterns between heavy and light/non-smokers and one method to test the association of the chromosome/arm-specific or focal CNs and smoking pack-years.

First, we calculated the total events of copy number gains and losses and compared them between the two smoking groups by the two-sided t test, which provides a convenient summary index but collapses CNAs information over the genomic locations.

The second method is to apply two-sample tests for continuous copy number by calculating the standardized difference of two average copy numbers for each locus as: $c_i = (m_{1i} - m_{2i}) \big/ \sqrt{v_{1i}/n_1 + v_{2i}/n_2}$ where m_{ji} and v_{ji} is the estimated mean and variance, respectively, of copy number for group j at locus i, and n_j is the sample size in group j. We summed up $c_i^{\ 2}$ over i across the loci in the arm of a chromosome to calculate the observed total standardized squared difference ($C_{observed}$). By permuting the two groups and carrying out the above procedure for 10,000 times, we obtained a non-parametric null distribution ($\boldsymbol{C}_{null}$). Then p values were obtained by comparing $C_{observed}$ and $\boldsymbol{C}_{null}$. This permutation procedure provides a valid global test for the overall difference by accounting for multiple comparisons and correlation of CNAs between different loci.

The third is a similar method extended to the discrete variable of CNAs (CN≥2.7 or not; CN≤1.3 or not) as mentioned above, but with advantage of testing gains and losses separately. We applied two-sample tests for binomial data by calculating the standardized difference of two proportions for each locus as: $d_i = (p_{1i} - p_{2i}) \big/ \sqrt{p_{1i}(1 - p_{1i})/n_1 + p_{2i}(1 - p_{2i})/n_2}$ where p_{ji} is the estimated proportion (stabilized by adding 0.5 in the numerator) of CN gains (or losses) for group j at locus i and n_j is the sample size in group j. We summed up $d_i^{\ 2}$ over i across the arm of a chromosome to calculate the observed total standardized squared difference ($D_{observed}$). Non-parametric null distribution ($\boldsymbol{D}_{null}$) was approximated by 10,000 permutations, and p values were obtained by comparing $D_{observed}$ and $\boldsymbol{D}_{null}$.

The above three methods require the smoking exposure to be dichotomized. To fully capture the continuous dose-response relationship between copy numbers and pack-years of cigarette smoking, we also developed a test to summarize such association in a chromosome/arm-specific fashion. We obtained the test statistics, $F_{observed} = \sum_i f_i$ where f_i is the F statistics of regressing continuous copy numbers on the smoking pack-years (square-root transformed) up to the quadratic term at locus *i*. Again, the non-parametric null distribution ($\boldsymbol{F}_{null}$) was generated by 10,000 permutations and p values were obtained as the tail probability of $F_{observed}$ in $\boldsymbol{F}_{null}$.

The proposed test is equivalent to the powerful score test for testing the variance of coefficients in a multivariate regression by assuming regression coefficients have an arbitrary distribution with mean 0 and variance τ (30), in which copy numbers of a region or chromosome (as a vector) are regressed on smoking pack-years. The null hypothesis of our proposed test is that all the coefficients relating pack-years to copy numbers are zero, or equivalently, copy numbers at all loci have no association with smoking pack-years, which is equivalent to H_0: τ=0. The alternative hypothesis would be that copy numbers at some loci have association with smoking pack-years. This variance component test is a powerful test by borrowing information in multiple markers and effectively accounting for correlation among the CNVs in a marker set, and reducing the degrees of freedom of the test.

We used another method to investigate gene selection of CNAs between heavy and light/non-smokers. Both the total probes (T) in which CNAs were detected and the probes locating within genes (G) in which CNAs were detected were calculated for each individual. We proposed a ratio of G vs. T (termed as G/T ratio) to estimate the selection of CNAs with respect to the gene location. Under the null hypothesis that CNAs occur randomly relative to where genes locate, we would expect the null ratio of 40.64% (104,256/256,554), where 104,256 is the number of probes located within genes on the chip. By comparing the G/T ratios to the null ratio, 40.64%, with two-sided t test, we were able to test whether CNAs occur preferentially away from genes.

To investigate the association of focal copy numbers and smoking pack-years, we analyzed copy number >2 and ≤2 separately. Two outliers (>2.5 standard deviations) of smoking pack-years were excluded to eliminate the potential spurious result driven by them. Square root was taken for smoking pack-years to transform a right skewed distribution into an approximately normal distribution.(31) Both multiple-marker and single-marker analyses were performed. In the multiple-marker analyses, we grouped the consecutive ten SNPs (markers) as a set and calculated $F_{observed}$ and $\boldsymbol{F}_{null}$ by the methods mentioned above to obtain the p values for each set of markers. For those with $F_{observed}$ much greater than $\boldsymbol{F}_{null}$ (from 10,000 permutations), the null distributions were obtained by the Satterthwaite approximation(32), in which the first two moments of scaled χ^2 distribution were matched with those of $\boldsymbol{F}_{null}$. In total, we performed 25,655 hypothesis tests for copy number >2 and ≤2 separately. Such multiple-marker analyses had better statistical power than the single-marker analyses when the markers were correlated, which is the case in the copy number data.

For single-marker analyses, 256,554 regressions for both copy number >2 and ≤2 were performed in the discovery set with continuous copy number at each locus as a dependent variable and square root of smoking pack-years and its quadratic term as independent covariates. For the validated candidates (p<0.05 in the validation set), pooled results were generated with linear regressions (with up to quadratic term of square root of smoking pack-years), spline regressions (with spline of square root of smoking pack-years) and locally weighted scatter plot smoothing (LOWESS). Adjusted linear and spline regressions were performed with adjustment of age at diagnosis, gender, two cohorts, clinical stage and histology.

DNA double-strand break assay

For the cytotoxicity analysis of cigarette smoke condensate, a human nontumorigenic bronchial epithelial cell line HBEC 3KT was cultured in 12-well plates to confluence and then treated with indicated concentration of cigarette smoke condensate (CSC) for 24 hours. Viable cells were monitored by MTT assay using the CellTiter 96 AQueous One Solution Cell Proliferation Assay kit (Promega, Madison, WI). All assays were performed in triplicate. For the neutral comet assay, HBEC 3KT was cultured in 100-mm plates to confluence and then treated with indicated concentration of CSC for 24 hours. Cells having DNA double-stand break were analyzed by Neutral comet assay using CometAssay kit (TREVIGEN, Gaithersburg, MD). About 600~800 cells were viewed per treatment. For the apoptosis analysis, HBEC 3KT was cultured in 100-mm plates to confluence and then treated with indicated concentration of CSC for 24 hours. The status of cellular apoptosis was determined using SendoLyteTM Homogeneous Rh110 Caspase-3/7 Assay kit (ANASPEC, Fremont, CA). All apoptosis assays were performed in triplicate.

ACKNOWLEDGEMENTS

We are indebted to the participants of Molecular and Genetic Analysis of Lung Cancer Study; to Dr. Cheng Li of HSPH for his advice on data preprocessing; to the HSPH Lung Cancer Study Group: Dr. Kofi Asomaning, Dr. Eugene Mark, Dr. Matthew Kulke, Dr. Wei Zhou, Dr. Geoffrey Liu, Marcia Chertok, Andrea Shafer, Lauren Cassidy, Maureen Convery, Salvatore Mucci; to Drs. Panos Fidias and Bruce A. Chabner and the physicians and surgeons of the Massachusetts General Hospital Cancer Center; and to Dr. Lodve Stangeland of Haukeland University Hospital, Norway for recruiting patients. This study is supported by US National Institutes of Health Grants No. CA076404 and CA134294 (Y.-T. H. and X.L.), CA092824 (D.C.C.), CA074386 (D.C.C.) and CA090578 (D.C.C.), and Norwegian Cancer Society (A.H.).

REFERENCES

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J*, et al.* (2009) Cancer statistics, 2009 *CA Cancer J Clin* **59,** 225-249.
2. Bach PB (2009) Smoking as a factor in causing lung cancer *JAMA* **301,** 539-541.
3. Wynder EL & Graham EA (1950) Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases *J Am Med Assoc* **143,** 329-336.
4. Doll R & Hill AB (1954) The mortality of doctors in relation to their smoking habits; a preliminary report *Br Med J* **1,** 1451-1455.
5. Dubey S & Powell CA (2009) Update in lung cancer 2008 *Am J Respir Crit Care Med* **179,** 860-868.
6. Alavanja MC (2002) Biologic damage resulting from exposure to tobacco smoke and from radon: implication for preventive interventions *Oncogene* **21,** 7365-7375.
7. Kim TM, Yim SH, Lee JS, Kwon MS, Ryu JW*, et al.* (2005) Genome-wide screening of genomic alterations and their clinicopathologic implications in non-small cell lung cancers *Clin Cancer Res* **11,** 8235-8242.
8. Weir BA, Woo MS, Getz G, Perner S, Ding L*, et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma *Nature* **450,** 893-898.
9. van Gent DC & van der Burg M (2007) Non-homologous end-joining, a sticky affair *Oncogene* **26,** 7731-7740.
10. Hastings PJ, Lupski JR, Rosenberg SM, & Ira G (2009) Mechanisms of change in gene copy number *Nat Rev Genet* **10,** 551-564.
11. Yen CC, Liang SC, Jong YJ, Chen YJ, Lin CH*, et al.* (2007) Chromosomal aberrations of malignant pleural effusions of lung adenocarcinoma: different cytogenetic changes are correlated with genders and smoking habits *Lung Cancer* **57,** 292-301.
12. Massion PP, Zou Y, Chen H, Jiang A, Coulson P*, et al.* (2008) Smoking-related genomic signatures in non-small cell lung cancer *Am J Respir Crit Care Med* **178,** 1164-1172.
13. Emerson JJ, Cardoso-Moreira M, Borevitz JO, & Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster *Science* **320,** 1629-1631.
14. Huang Y-T, Lin X, Chirieac LR, McGovern R, Wain JC*, et al.* (2011) Impact of Disease Development, Genomic Location and Biological Function on Copy Number Alterations of Non-Small Cell Lung Cancer *PLoS One* (in press).
15. Furstenberger G & Senn HJ (2002) Insulin-like growth factors and cancer *Lancet Oncol* **3,** 298-302.
16. Herbst RS, Heymach JV, & Lippman SM (2008) Lung cancer *N Engl J Med* **359,** 1367-1380.
17. Morgillo F, Kim WY, Kim ES, Ciardiello F, Hong WK*, et al.* (2007) Implication of the insulin-like growth factor-IR pathway in the resistance of non-small cell lung cancer cells to treatment with gefitinib *Clin Cancer Res* **13,** 2795-2803.
18. Kaklamani VG, Linos A, Kaklamani E, Markaki I, & Mantzoros C (1999) Age, sex, and smoking are predictors of circulating insulin-like growth factor 1 and insulin-like growth factor-binding protein 3 *J Clin Oncol* **17,** 813-817.

19. Tannheimer SL, Ethier SP, Caldwell KK, & Burchiel SW (1998) Benzo[a]pyrene- and TCDD-induced alterations in tyrosine phosphorylation and insulin-like growth factor signaling pathways in the MCF-10A human mammary epithelial cell line *Carcinogenesis* **19,** 1291-1297.
20. Yu H, Spitz MR, Mistry J, Gu J, Hong WK*, et al.* (1999) Plasma levels of insulin-like growth factor-I and lung cancer risk: a case-control analysis *J Natl Cancer Inst* **91,** 151-156.
21. Wu X, Yu H, Amos CI, Hong WK, & Spitz MR (2000) Joint effect of insulin-like growth factors and mutagen sensitivity in lung cancer risk *J Natl Cancer Inst* **92,** 737-743.
22. Lukanova A, Toniolo P, Akhmedkhanov A, Biessy C, Haley NJ*, et al.* (2001) A prospective study of insulin-like growth factor-I, IGF-binding proteins-1, -2 and -3 and lung cancer risk in women *Int J Cancer* **92,** 888-892.
23. Spitz MR, Barnett MJ, Goodman GE, Thornquist MD, Wu X*, et al.* (2002) Serum insulin-like growth factor (IGF) and IGF-binding protein levels and risk of lung cancer: a case-control study nested in the beta-Carotene and Retinol Efficacy Trial Cohort *Cancer Epidemiol Biomarkers Prev* **11,** 1413-1418.
24. Hankinson SE, Willett WC, Colditz GA, Hunter DJ, Michaud DS*, et al.* (1998) Circulating concentrations of insulin-like growth factor-I and risk of breast cancer *Lancet* **351,** 1393-1396.
25. Chan JM, Stampfer MJ, Giovannucci E, Gann PH, Ma J*, et al.* (1998) Plasma insulin-like growth factor-I and prostate cancer risk: a prospective study *Science* **279,** 563-566.
26. Ma J, Pollak MN, Giovannucci E, Chan JM, Tao Y*, et al.* (1999) Prospective study of colorectal cancer risk in men and plasma levels of insulin-like growth factor (IGF)-I and IGF-binding protein-3 *J Natl Cancer Inst* **91,** 620-625.
27. Zhou W, Heist RS, Liu G, Asomaning K, Miller DP*, et al.* (2006) Second hand smoke exposure and survival in early-stage non-small-cell lung cancer patients *Clin Cancer Res* **12,** 7187-7193.
28. Zienolddiny S, Campa D, Lind H, Ryberg D, Skaug V*, et al.* (2008) A comprehensive analysis of phase I and phase II metabolism gene polymorphisms and risk of non-small cell lung cancer in smokers *Carcinogenesis* **29,** 1164-1169.
29. Zhao X, Li C, Paez JG, Chin K, Janne PA*, et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays *Cancer Res* **64,** 3060-3071.
30. Liu D, Lin X, & Ghosh D (2007) Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models *Biometrics* **63,** 1079-1088.
31. Zhou W, Liu G, Park S, Wang Z, Wain JC*, et al.* (2005) Gene-smoking interaction associations for the ERCC1 polymorphisms in the risk of lung cancer *Cancer Epidemiol Biomarkers Prev* **14,** 491-496.
32. Satterthwaite FE (1946) An Approximate Distribution of Estimates of Variance Components *Biometrics Bulletin* **2,** 110-114.
33. Kwee L, Liu D, Lin X, Ghosh D, and Epstein M. (2008) A powerful and flexible multilocus association test for quantitative traits *Am J of Human Genetics,* **82**, 386-397.

FIGURE LEGENDS

Figure 1. Association of cigarette smoking and whole-genome copy numbers.

A and **B**, Among the 256,554 total probes, the proportion (%) with CNAs (**A**, gains; **B**, losses) events by pack-years of cigarette smoking (NS/LS: non-/light smokers, HS: heavy smokers). **C** and **D**, Mean and its 95% confidence interval of G/T ratios in the heavy smokers (HS) and non-/light smokers (NS/LS) for copy number gains (**C**) and losses (**D**); and the dashed lines represent the null G/T ratio on the chip (104,256/256,554=40.64%). P values were used to test the indicated indices between HS and NS/LS with methods described in Methods.

Figure 2. Association of cigarette smoking and chromosome/arm-specific copy numbers. A, the p values are from analyses testing the association of copy number-gain events with the smoking group (heavy vs. light-/non-smokers). B, the p values are from analyses testing the association of copy numbers (>2) with pack-years of cigarette smoking. The dashed line indicates p=0.05.

Figure 3. Association of cigarette smoking and 25,655 moving window10-marker focal copy numbers.

A. A plot of -$\log_{10}P$ of the association between smoking pack-years and 10-marker set focal copy number, which is analyzed for copy number >2 (upper half) and ≤2 (lower half), separately. B, P values of focal copy number analyses in 12q23. The black dots and line indicate p values from 10-marker analyses, and the superimposed gray dots and line indicate the corresponding ones from single-marker analyses.

Figure 4. Effects of cigarette smoke condensate (CSC) treatment on HBEC 3KT DNA double-stand breaks and apoptosis. **A**, Cytotoxic effect of CSC on HBEC3KT survival. HBEC3KT cells were cultured in 12 well plates to confluence and then incubated with 0, 0.04, 0.4, 0.8, 4, and 40 μg/mL CSC for 24 hours. Viable cells were monitored with MTT assay. Live cells treated with 0 μg/mL CSC were defined as 100%. Percentage of live cells verses CSC concentration was plotted. **B** and **C**, CSC treatment induces DNA single/double-stand breaks in HBEC 3KT cells. HBEC 3KT cells were culture to confluence in 100-mm plates and then treated with 0, 0.04, and 0.4 μg/mL CSC for 24 hours. Cells were harvested by trypsinization and DNA single/double-stand breaks were analyzed by neutral comet assay. A representative photo with undamaged-DNA (bright dot) and DNA with single/double-strand breaks (bright dot with an elongated tail) was shown (**B**). About 600~800 cells per treatment were viewed, and percentage of DNA single/double-strand breaks verses CSC dose was plotted (**C**). **D**, CSC treatment induces apoptosis in HBEC 3KT cells. The same set of cells used in **B** and **C** was also analyzed for caspase-3/7 activity. Columns are mean value of relative fluorescence unit (RFU). A larger RFU value represents a higher caspase-3/7 activity and thus a stronger apoptotic response. Standard deviations are provided in **C** and **D**.

TABLE LEGENDS

Table 1. Characteristics of study populations

Table 2. Summary of the ten candidate loci at 12q23 from both 10-marker and single-marker analyses.

Table 1. Characteristics of study populations

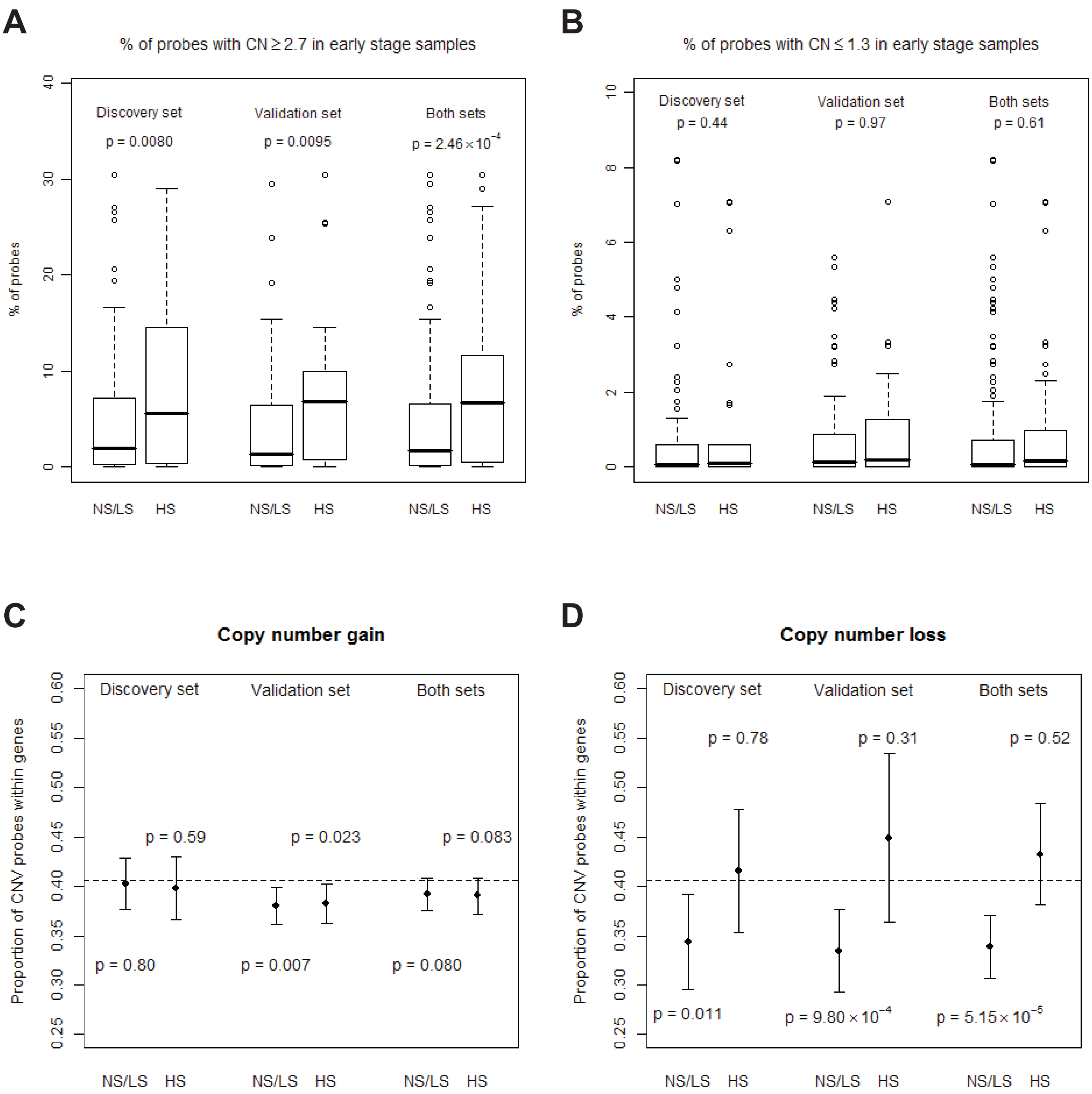
	Discovery set	Validation set	P value*
Sample size	134	130	
Male (%)	65.67	56.92	0.18
Age,			
Mean ± standard deviation	67.27 ± 8.17	67.59 ± 8.39	0.75
Cigarette smoking pack-years,			
Median ± interquartile range	34.25 ± 39.64	38 ± 35.93	0.28
Clinical stage			0.43
Stage 1 (%)	77.27	70.00	
Stage 2 (%)	15.15	19.23	
Stage 3 or 4 (%)	7.58	10.77	
Cigarette smoking status			0.23
Never smokers (%)	7.46	6.15	
Ex-smokers (%)	43.28	53.85	
Current smokers (%)	49.25	40.00	
Adenocarcinoma (%)	67.91	64.62	0.66

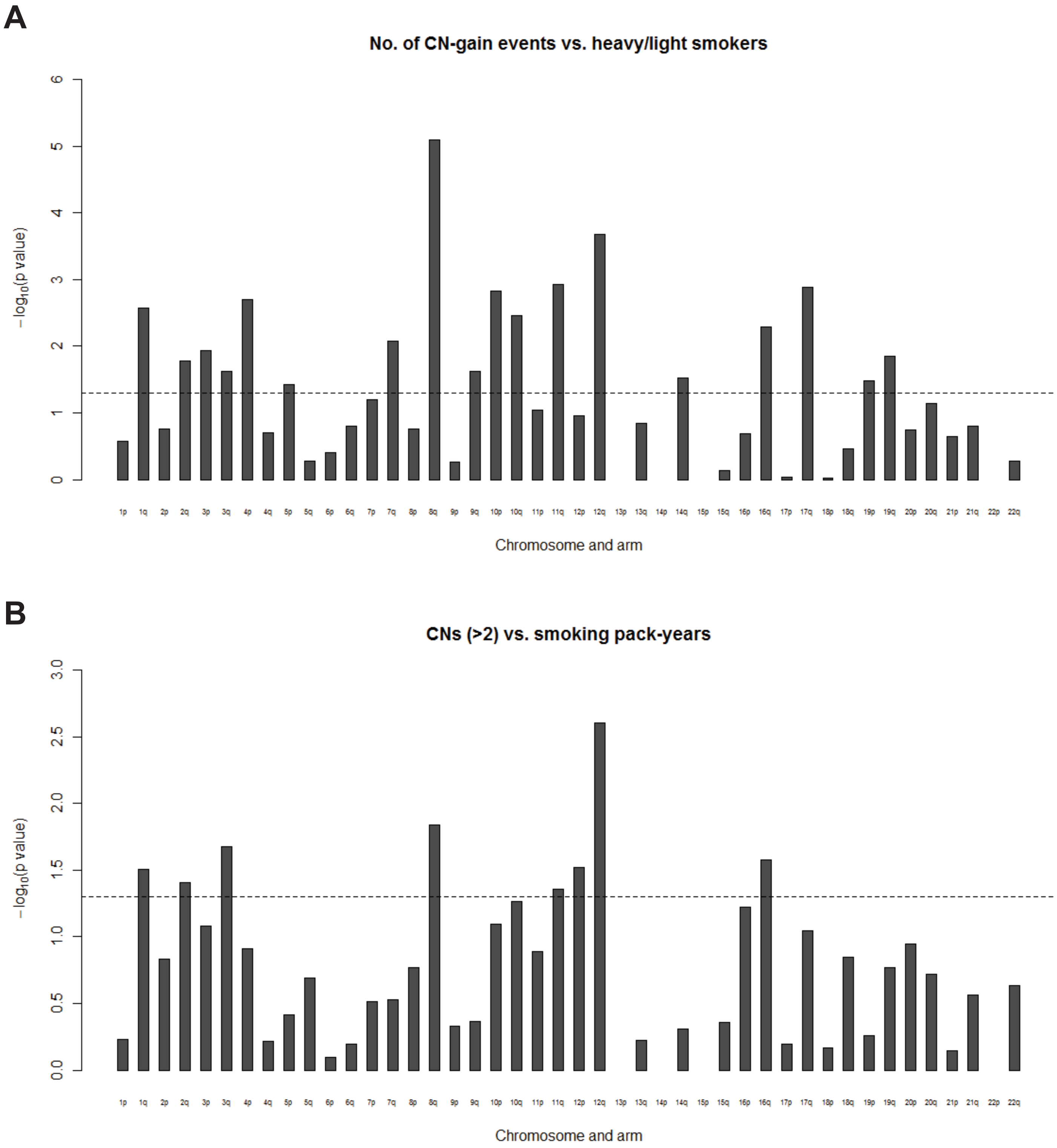
*P values were calculated with X^2 test for percentage of male (1 degree of freedom, d.f.), adenocarcinoma, patentis from MGH (1 d.f.), clinical stage (2 d.f.) and cigarette smoking status (2 d.f.); with t test for age; and with Wilcoxon test for cigarette smoking pack-years.

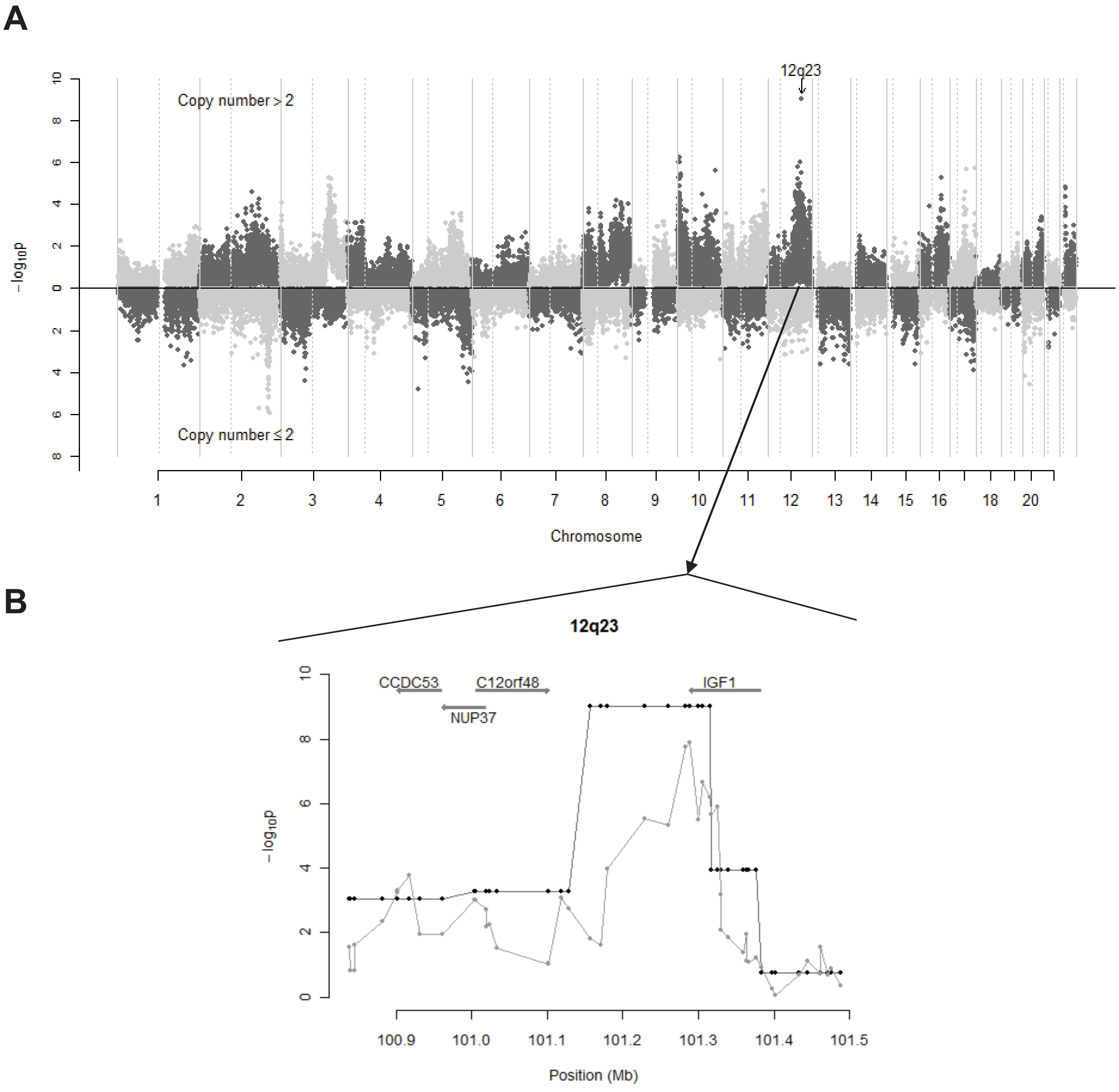
Table 2. Summary of the ten candidate loci at 12q23 from both 10-marker and single-marker analyses.

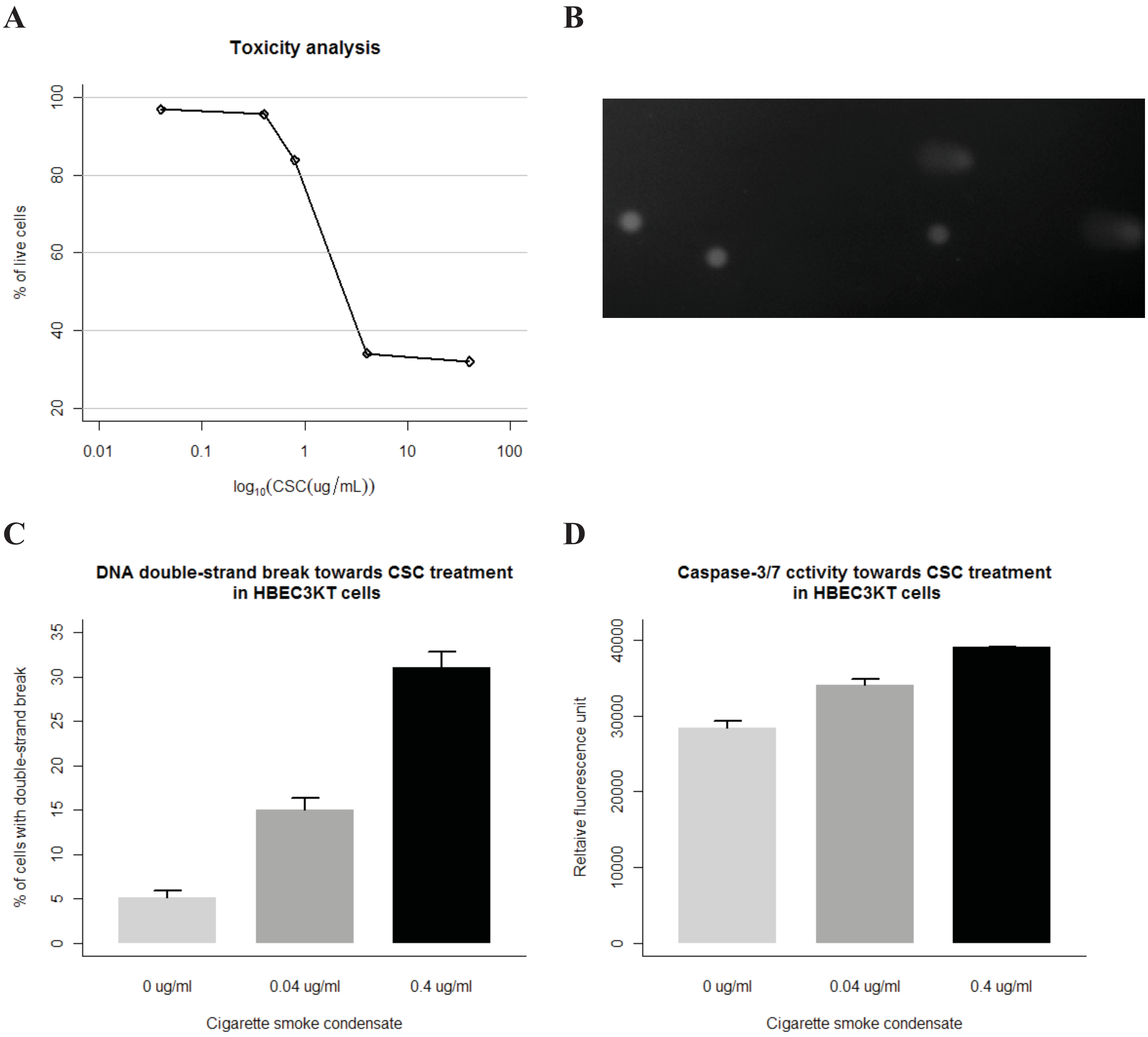
					Focal copy number-smoking association					
					10-marker analyses			Single-marker analyses		
Affy ID	dbSNP	Cyto-band	Position (Mb)	Gene	p value, discovery set	p value, validation set	p value, pooled	p value, pooled	R^2	p value, adjusted*
SNP_A-2002985	rs5011687	12q23	101.157	-	3.17×10^{-8}	0.0291	9.69×10^{-10}	0.0152	0.065	0.0168
SNP_A-2125858	rs17439974	12q23	101.171	-				0.0239	0.060	0.0285
SNP_A-4222341	rs17032384	12q23	101.179	-				0.000110	0.126	0.000263
SNP_A-1899321	rs1520223	12q23	101.229	-				2.92×10^{-6}	0.175	8.07×10^{-6}
SNP_A-4222344	rs4764695	12q23	101.260	-				4.55×10^{-6}	0.167	1.33×10^{-5}
SNP_A-4228436	rs10860860	12q23	101.283	-				1.79×10^{-8}	0.223	9.78×10^{-8}
SNP_A-2106083	rs2946831	12q23	101.289	-				1.29×10^{-8}	0.235	2.63×10^{-8}
SNP_A-2255731	rs10745940	12q23	101.300	*IGF1*				3.26×10^{-6}	0.163	8.20×10^{-6}
SNP_A-2092658	rs9308315	12q23	101.306	*IGF1*				2.10×10^{-7}	0.202	7.74×10^{-7}
SNP_A-2271065	rs2072592	12q23	101.316	*IGF1*				6.17×10^{-7}	0.200	4.42×10^{-6}

*P values of smoking pack-years were calculated from linear models with up to quadratic term of square root-transformed smoking pack-years, adjusting for age, gender, clinical stage, and cell type.


A
% of probes with CN ≥ 2.7 in early stage samples
% of probes
0
10
20
30
40
Discovery set
p = 0.0080
Validation set
p = 0.0095
Both sets
p = 2.46 × 10^-4
NS/LS
HS
NS/LS
HS
NS/LS
HS
B
% of probes with CN ≤ 1.3 in early stage samples
% of probes
0
2
4
6
8
10
Discovery set
p = 0.44
Validation set
p = 0.97
Both sets
p = 0.61
NS/LS
HS
NS/LS
HS
NS/LS
HS
C
Copy number gain
Proportion of CNV probes within genes
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
Discovery set
Validation set
Both sets
p = 0.59
p = 0.023
p = 0.083
p = 0.80
p = 0.007
p = 0.080
NS/LS
HS
NS/LS
HS
NS/LS
HS
D
Copy number loss
Proportion of CNV probes within genes
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
Discovery set
Validation set
Both sets
p = 0.78
p = 0.31
p = 0.52
p = 0.011
p = 9.80 × 10^-4
p = 5.15 × 10^-5
NS/LS
HS
NS/LS
HS
NS/LS
HS


A
No. of CN-gain events vs. heavy/light smokers
$-\log_{10}$(p value)
0
1
2
3
4
5
6
1p 1q 2p 2q 3p 3q 4p 4q 5p 5q 6p 6q 7p 7q 8p 8q 9p 9q 10p 10q 11p 11q 12p 12q 13p 13q 14p 14q 15p 15q 16p 16q 17p 17q 18p 18q 19p 19q 20p 20q 21p 21q 22p 22q
Chromosome and arm
B
CNs (>2) vs. smoking pack-years
$-\log_{10}$(p value)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1p 1q 2p 2q 3p 3q 4p 4q 5p 5q 6p 6q 7p 7q 8p 8q 9p 9q 10p 10q 11p 11q 12p 12q 13p 13q 14p 14q 15p 15q 16p 16q 17p 17q 18p 18q 19p 19q 20p 20q 21p 21q 22p 22q
Chromosome and arm


A
12q23
Copy number > 2
Copy number ≤ 2
−log10p
10
8
6
4
2
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
Chromosome
B
12q23
CCDC53
C12orf48
IGF1
NUP37
−log10p
10
8
6
4
2
0
100.9
101.0
101.1
101.2
101.3
101.4
101.5
Position (Mb)


A
Toxicity analysis
% of live cells
20
40
60
80
100
0.01
0.1
1
10
100
$\log_{10}$(CSC(ug/mL))
B
C
DNA double-strand break towards CSC treatment in HBEC3KT cells
% of cells with double-strand break
0
5
10
15
20
25
30
35
0 ug/ml
0.04 ug/ml
0.4 ug/ml
Cigarette smoke condensate
D
Caspase-3/7 cctivity towards CSC treatment in HBEC3KT cells
Reltaive fluorescence unit
0
10000
20000
30000
40000
0 ug/ml
0.04 ug/ml
0.4 ug/ml
Cigarette smoke condensate