21 Mori K, Suzuki T, Uozaki H *et al.* Detection of minimal gastric cancer cells in peritoneal washings by focused microarray analysis with multiple markers: clinical implications. *Ann Surg Oncol* 2007; **14**: 1694–702.
22 Markman M, Brady MF, Spirtos NM, Hanjani P, Rubin SC. Phase II trial of intraperitoneal paclitaxel in carcinoma of the ovary, tube, and peritoneum: a gynecologic oncology group study. *J Clin Oncol* 1998; **16**: 2620–4.
23 Nishiyama M, Wada S. Docetaxel: its role in current and future treatments for advanced gastric cancer. *Gastric Cancer* 2009; **12**: 132–41.
24 Haren L, Remy MH, Bazin I, Callebaut I, Wright M, Merdes A. NEDD1-dependent recruitment of the gamma-tubulin ring complex to the centrosome is necessary for centriole duplication and spindle assembly. *J Cell Biol* 2006; **172**: 505–15.
25 Inoue M, Matsumoto S, Saito H, Tsujitani S, Ikeguchi M. Intraperitoneal administration of a small interfering RNA targeting nuclear factor-kappa B with paclitaxel successfully prolongs the survival of xenograft model mice with peritoneal metastasis of gastric cancer. *Int J Cancer* 2008; **123**: 2696–701.
26 Takeshita F, Patrawala L, Osaki M *et al.* Systemic delivery of synthetic microRNA-16 inhibits the growth of metastatic prostate tumors via downregulation of multiple cell-cycle genes. *Mol Ther* 2010; **18**: 181–7.
27 Ueda T, Volinia S, Okumura H *et al.* Relation between microRNA expression and progression and prognosis of gastric cancer: a microRNA expression analysis. *Lancet Oncol* 2010; **11**: 136–46.
28 Yanagihara K, Tsumuraya M. Transforming growth factor β1 induces apoptotic cell death in cultured human gastric carcinoma cells. *Cancer Res* 1992; **52**: 4042–5.
29 Saeki N, Kim DH, Usui T *et al.* GASDERMIN, suppressed frequently in gastric cancer, is a target of LMO1 in TGF-β-dependent apoptotic signalling. *Oncogene* 2007; **26**: 6488–98.

doi: 10.1111/cas.12054

nature genetics

Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia

Qing Lan[1,68], Chao A Hsiung[2,68], Keitaro Matsuo[3,68], Yun-Chul Hong[4,68], Adeline Seow[5,68], Zhaoming Wang[6,68], H Dean Hosgood III[1,7,68], Kexin Chen[8,68], Jiu-Cun Wang[9,10,68], Nilanjan Chatterjee[1], Wei Hu[1], Maria Pik Wong[11], Wei Zheng[12], Neil Caporaso[1], Jae Yong Park[13], Chien-Jen Chen[14], Yeul Hong Kim[15], Young Tae Kim[16], Maria Teresa Landi[1], Hongbing Shen[17,18], Charles Lawrence[19], Laurie Burdett[6], Meredith Yeager[6], Jeffrey Yuenger[6], Kevin B Jacobs[6], I-Shou Chang[20], Tetsuya Mitsudomi[21], Hee Nam Kim[22], Gee-Chen Chang[23,24], Bryan A Bassig[1,25], Margaret Tucker[1], Fusheng Wei[26], Zhihua Yin[27], Chen Wu[28,29], She-Juan An[30], Biyun Qian[8], Victor Ho Fun Lee[31], Daru Lu[9,10], Jianjun Liu[32,33], Hyo-Sung Jeon[34], Chin-Fu Hsiao[2], Jae Sook Sung[15], Jin Hee Kim[35], Yu-Tang Gao[36], Ying-Huang Tsai[37], Yoo Jin Jung[16], Huan Guo[38], Zhibin Hu[17,18], Amy Hutchinson[6], Wen-Chang Wang[2], Robert Klein[39], Charles C Chung[1], In-Jae Oh[40,41], Kuan-Yu Chen[42], Sonja I Berndt[1], Xingzhou He[43], Wei Wu[27], Jiang Chang[28,29], Xu-Chao Zhang[30], Ming-Shyan Huang[44], Hong Zheng[8], Junwen Wang[45,46], Xueying Zhao[9,10], Yuqing Li[32], Jin Eun Choi[34], Wu-Chou Su[47], Kyong Hwa Park[15], Sook Whan Sung[48], Xiao-Ou Shu[12], Yuh-Min Chen[23,49], Li Liu[50], Chang Hyun Kang[16], Lingmin Hu[17,18], Chung-Hsing Chen[20], William Pao[51], Young-Chul Kim[40,41], Tsung-Ying Yang[24], Jun Xu[52], Peng Guan[27], Wen Tan[28,29], Jian Su[30], Chih-Liang Wang[53], Haixin Li[8], Alan Dart Loon Sihoe[54], Zhenhong Zhao[9,10], Ying Chen[5], Yi Young Choi[34], Jen-Yu Hung[44], Jun Suk Kim[55], Ho-Il Yoon[56], Qiuyin Cai[12], Chien-Chung Lin[47], In Kyu Park[16], Ping Xu[57], Jing Dong[17,18], Christopher Kim[1], Qincheng He[27], Reury-Perng Perng[49], Takashi Kohno[58], Sun-Seog Kweon[59,60], Chih-Yi Chen[61], Roel Vermeulen[62], Junjie Wu[9,10], Wei-Yen Lim[5], Kun-Chieh Chen[24], Wong-Ho Chow[1], Bu-Tian Ji[1], John K C Chan[63], Minjie Chu[17,18], Yao-Jen Li[14], Jun Yokota[64], Jihua Li[65], Hongyan Chen[9,10], Yong-Bing Xiang[36], Chong-Jen Yu[42], Hideo Kunitoh[66], Guoping Wu[26], Li Jin[9,10], Yen-Li Lo[2], Kouya Shiraishi[58], Ying-Hsiang Chen[2], Hsien-Chih Lin[2], Tangchun Wu[38,69], Yi-Long Wu[30,69], Pan-Chyr Yang[67,69], Baosen Zhou[27,69], Min-Ho Shin[60,69], Joseph F Fraumeni Jr[1,69], Dongxin Lin[28,29,69], Stephen J Chanock[1,69] & Nathaniel Rothman[1,69]

To identify common genetic variants that contribute to lung cancer susceptibility, we conducted a multistage genome-wide association study of lung cancer in Asian women who never smoked. We scanned 5,510 never-smoking female lung cancer cases and 4,544 controls drawn from 14 studies from mainland China, South Korea, Japan, Singapore, Taiwan and Hong Kong. We genotyped the most promising variants (associated at $P < 5 \times 10^{-6}$) in an additional 1,099 cases and 2,913 controls. We identified three new susceptibility loci at 10q25.2 (rs7086803, $P = 3.54 \times 10^{-18}$), 6q22.2 (rs9387478, $P = 4.14 \times 10^{-10}$) and 6p21.32 (rs2395185, $P = 9.51 \times 10^{-9}$). We also confirmed associations reported for loci at 5p15.33 and 3q28 and a recently reported finding at 17q24.3. We observed no evidence of association for lung cancer at 15q25 in never-smoking women in Asia, providing strong evidence that this locus is not associated with lung cancer independent of smoking.

It is estimated that 25% of lung cancer cases arise in individuals who never smoked. Lung cancer in never smokers ranks as the seventh most common cause of cancer death worldwide[1]. A number of observations suggest that the molecular pathogenesis of lung cancer differs by smoking status. Differences have been reported by smoking status in cellular and molecular carcinogenic pathways, distinct profiles of oncogenic mutations (for example, in *EGFR*) and response to targeted therapy[2,3]. Compared to lung cancer in smokers, cases in never smokers are more likely to arise in women at a younger age, and there is a greater proportion of cases with the adenocarcinoma histology subtype[3]. Epidemiological studies of lung cancer in never smokers have shown that the incidence of lung cancer in women is particularly high in Asia[4], which is partially attributed to exposure to environmental tobacco smoke, combustion products from indoor heating and cooking fuel, and cooking oil fumes[4–10].

A full list of affiliations appears at the end of the paper.

Received 31 May; accepted 5 October; published online 11 November 2012; doi:10.1038/ng.2456

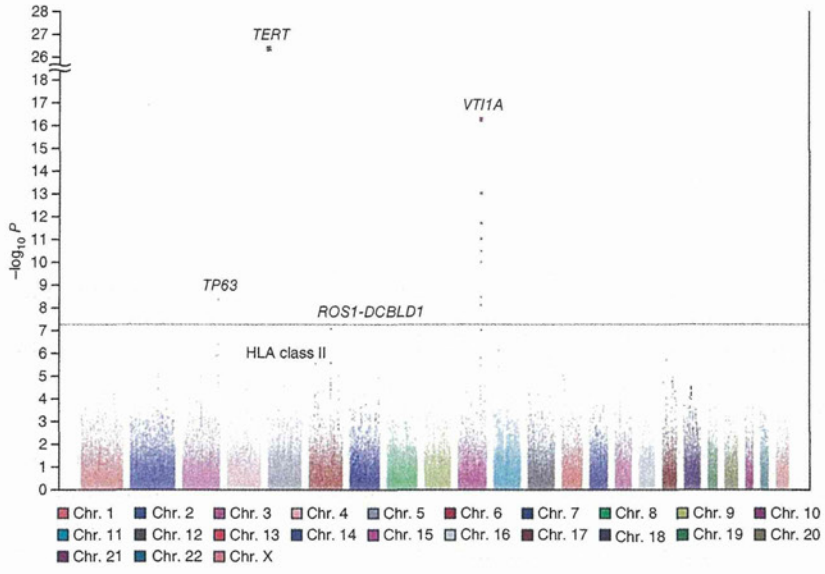

Figure 1 Association results from a GWAS of never-smoking women in Asia. Manhattan plot based on *P* values derived from 1-degree-of-freedom tests of genotype trend effect in unconditional logistic regression analysis adjusted for study, age and three eigenvectors in a GWAS of lung cancer in never-smoking Asian females, including 5,510 lung cancer cases and 4,544 controls. The *x* axis represents chromosomal location, and the *y* axis shows *P* values on a negative logarithmic scale. The red horizontal line represents the genome-wide significance threshold of $P = 5 \times 10^{-8}$. Labeled are two previously associated loci (*TERT* at 5p15.33 and *TP63* at 3q28) together with three newly identified loci (*VTI1A* on chromosome 10 and *ROS1-DCBLD1* and the HLA class II region on chromosome 6).

To gain insight into the etiology of lung cancer in never-smoking women, we formed the Female Lung Cancer Consortium in Asia (FLCCA), which includes studies drawn from mainland China, South Korea, Japan, Singapore, Taiwan and Hong Kong. Previously, we published the first genome-wide association study (GWAS) of lung cancer in never-smoking Asian women, including 584 cases and 585 controls with large-scale replication, reporting an association at 5p15.33 near the *TERT* gene[11]; in this study, it was also notable that the estimated effect of the associated locus was greater in nonsmoking Asian women than the reported effect size observed in primarily smokers of European ancestry[12]. We also confirmed an association signal in *TP63* at 3q28 (ref. 13), replicating the report from a GWAS conducted in Japan[14].

To identify new susceptibility loci in Asian never-smoking women, we conducted a lung cancer GWAS in 14 studies (13 case-control studies and 1 cohort study; **Supplementary Table 1** and **Supplementary Note**). Samples were scanned at six centers (Online Methods): the US National Cancer Institute (NCI) Cancer Genomic Research (CGR) Laboratory, the Genome Institute of Singapore, the Memorial Sloan-Kettering Cancer Center (MSKCC), GeneTech Biotech in Taiwan, Gene-Square Biotech in Beijing and deCODE Genetics in Iceland. After stringent quality control analysis of genotypes (Online Methods), we combined data sets for 5,510 lung cancer cases and 4,544 controls using a previously described clustering algorithm[15]. The primary analysis was performed using logistic regression for genotype trend effect (with 1 degree of freedom) adjusted for study center, age and three eigenvectors (on the basis of principal-components analysis). A comparison of the observed and expected *P* values in the quantile-quantile plot showed an enrichment of observed signals with small *P* values compared to the null distribution of no association, with little evidence for genomic inflation (unscaled $\lambda = 1.014$, $\lambda_{1000} = 1.003$; **Supplementary Fig. 1**)[16].

The overall association results are shown in a Manhattan plot, in which we observed both new and known loci that exceeded the threshold for genome-wide significance ($P < 5 \times 10^{-8}$; **Fig. 1**). We observed association at two previously established loci, rs2736100 at 5p15.33 (refs. 11,12,14,17–19) and rs4488809 at 3q28 (refs. 13,14). We also observed support for association of a recently reported locus marked by rs7216064 at 17q24.3 (ref. 20) (**Supplementary Table 2**). Notably, there was no evidence for association across the 15q25 region, which has been associated with smoking-related lung cancer[12,19,21–24]. We did not observe strong association signals for other loci reported in either European[25] or Asian[17,26] populations (**Supplementary Table 2**).

In our primary scan, we observed one new locus at 10q25.2, marked by rs7086803, that substantially exceeded the threshold for genome-wide significance (odds ratio (OR) = 1.32, 95% confidence interval (CI) = 1.24–1.41; $P = 5.04 \times 10^{-17}$) (**Fig. 1** and **Table 1**). We developed assays to genotype 13 SNPs associated at $P < 5 \times 10^{-6}$ in the initial scan, using analysis of all cases or the most common subtype in nonsmokers, adenocarcinoma. We genotyped 1,099 new cases and 2,913 controls drawn from the same studies as in the initial scan. In a combined analysis of 6,609 cases and 7,457 controls, 3 new loci achieved associations at genome-wide significance (**Table 1**): 10q25.2 (rs7086803: OR = 1.28, 95% CI = 1.21–1.35; $P = 3.54 \times 10^{-18}$), 6q22.2 (rs9387478: OR = 0.85, 95% CI = 0.81–0.90; $P = 4.14 \times 10^{-10}$)

Table 1 New loci associated with lung cancer in a GWAS of never-smoking Asian females

SNP	Plausible candidate gene(s)	Chromosome position	Subset	Allele[a]	MAF[b]		Subjects		OR (95% CI)	P_{trend}
					Control	Case	Control	Case		
rs7086803	*VTI1A*	10q25.2	GWAS	G/A	0.26	0.32	4,492	5,457	1.32 (1.24–1.41)	5.04×10^{-17}
			Replication	G/A	0.27	0.31	2,887	1,085	1.23 (1.10–1.37)	3.36×10^{-4}
			Combined	G/A	0.27	0.31	7,379	6,542	1.28 (1.21–1.35)	3.54×10^{-18}
rs9387478	*ROS1, DCBLD1*	6q22.2	GWAS	C/A	0.50	0.46	4,542	5,510	0.85 (0.81–0.90)	7.79×10^{-8}
			Replication	C/A	0.49	0.47	2,891	1,091	0.92 (0.83–1.01)	0.088
			Combined	C/A	0.50	0.46	7,433	6,601	0.85 (0.81–0.90)	4.14×10^{-10}
rs2395185[c] (rs28366298)	HLA class II region	6p21.32	GWAS	G/T	0.35	0.38	4,541	5,504	1.16 (1.09–1.23)	2.60×10^{-6}
			Replication	A/C	0.37	0.42	2,880	1,008	1.20 (1.08–1.33)	7.93×10^{-4}
			Combined	Meta			7,421	6,512	1.17 (1.11–1.23)	9.51×10^{-9}

[a]Minor allele listed second. [b]Minor allele frequency. [c]For the HLA class II region, because a TaqMan assay could not be designed for rs2395185, we instead genotyped rs28366298, its perfect surrogate ($r^2 = 1.0$), by TaqMan. The reported *P* value is based on meta-analysis of the rs2395185 results in the GWAS and the rs28366298 results in the TaqMan set.

and 6p21.32 (rs2395185: OR = 1.17, 95% CI = 1.11–1.23; $P = 9.51 \times 10^{-9}$) (**Fig. 2**, **Table 1**, **Supplementary Fig. 2** and **Supplementary Tables 3** and **4**).

Analysis by histological subtype of lung cancer showed that both the 6q22.2 (rs9387478) and 6p21.32 (rs2395185) loci were associated with adenocarcinoma only, which comprised 71% of cases (**Table 2**). The estimated effects were consistent across studies (**Supplementary Fig. 2**). We note that rs7086803 showed a somewhat larger effect for squamous carcinoma compared to adenocarcinoma (**Table 2**), but, as the number of squamous carcinoma cases analyzed was small, we consider this a preliminary observation requiring independent replication.

To explore the relationship between these three regions and lung cancer in populations of European ancestry, we analyzed data from a previously reported GWAS of 5,718 lung cancer cases and 5,739 controls, including men and women who were primarily ever smokers[12]. We found no evidence for association at the three newly associated loci. In a subanalysis of 350 never-smoker cases and 1,379 never-smoker controls drawn from this study, we observed some evidence of association for rs2395185 (M.T.L., unpublished data), but larger studies are warranted.

We imputed SNPs catalogued in the 1000 Genomes Project March 2012 release and the Division of Cancer Epidemiology and Genetics

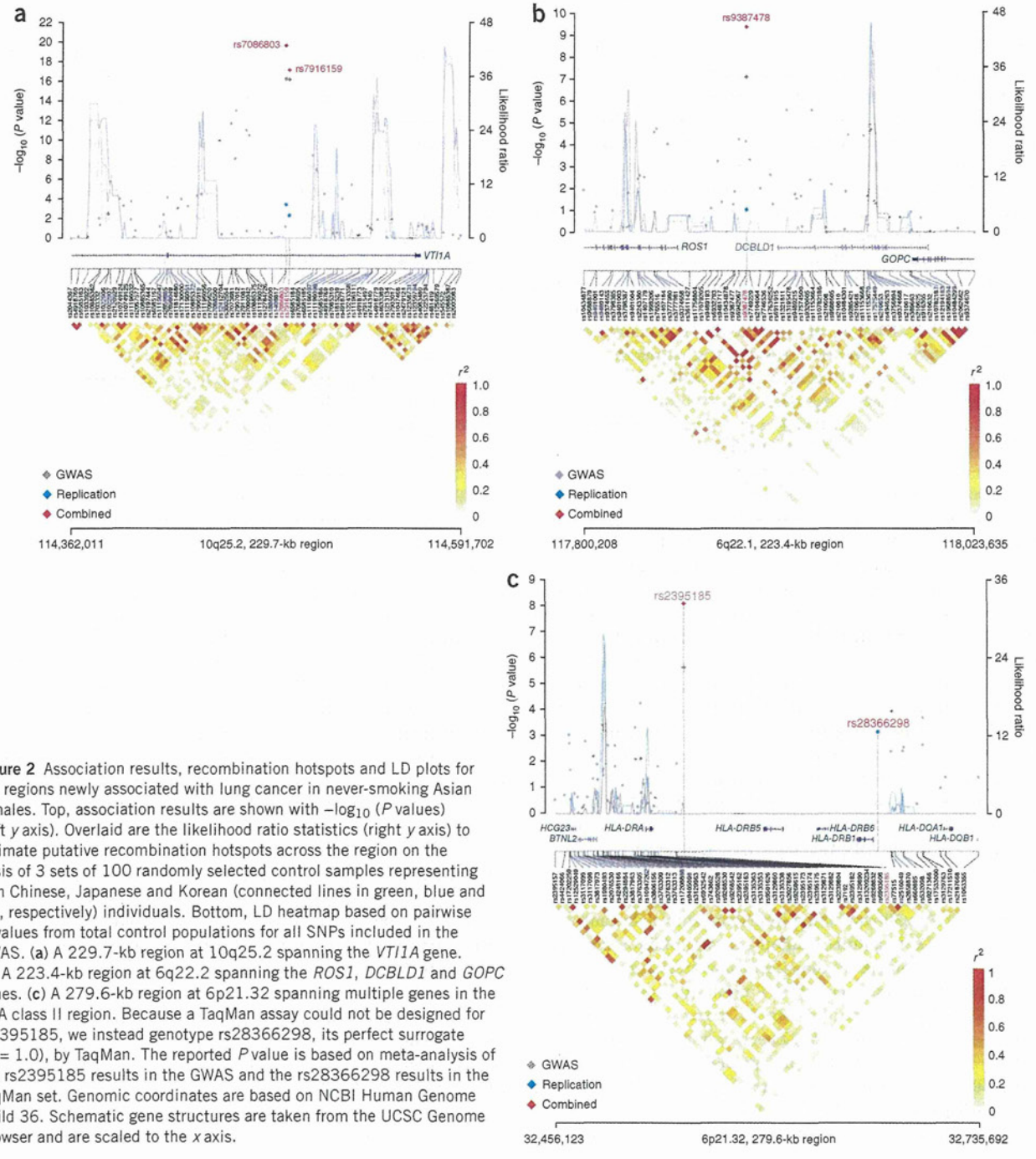

Figure 2 Association results, recombination hotspots and LD plots for the regions newly associated with lung cancer in never-smoking Asian females. Top, association results are shown with $-\log_{10}$ (*P* values) (left *y* axis). Overlaid are the likelihood ratio statistics (right *y* axis) to estimate putative recombination hotspots across the region on the basis of 3 sets of 100 randomly selected control samples representing Han Chinese, Japanese and Korean (connected lines in green, blue and red, respectively) individuals. Bottom, LD heatmap based on pairwise r^2 values from total control populations for all SNPs included in the GWAS. (**a**) A 229.7-kb region at 10q25.2 spanning the *VTI1A* gene. (**b**) A 223.4-kb region at 6q22.2 spanning the *ROS1*, *DCBLD1* and *GOPC* genes. (**c**) A 279.6-kb region at 6p21.32 spanning multiple genes in the HLA class II region. Because a TaqMan assay could not be designed for rs2395185, we instead genotype rs28366298, its perfect surrogate ($r^2 = 1.0$), by TaqMan. The reported *P* value is based on meta-analysis of the rs2395185 results in the GWAS and the rs28366298 results in the TaqMan set. Genomic coordinates are based on NCBI Human Genome Build 36. Schematic gene structures are taken from the UCSC Genome Browser and are scaled to the *x* axis.

Table 2 New loci associated with adenocarcinoma and squamous carcinoma of the lung in a GWAS of never-smoking Asian females

SNP	Putative gene	Chromosome position	Allele[a]	MAF[b] 1	2	3	Adenocarcinoma Subjects Control	Case	OR (95% CI)	P_{trend}	Squamous carcinoma Subjects Control	Case	OR (95% CI)	P_{trend}	$P_{heterogeneity}$[c]
rs7086803	*VTI1A*	10q25.2	G/A	0.27	0.31	0.34	7,035	4,666	1.24 (1.17–1.32)	1.19×10^{-11}	6,714	756	1.36 (1.21–1.54)	7.11×10^{-7}	0.014
rs9387478	*ROS1*, *DCBLD1*	6q22.2	C/A	0.50	0.46	0.48	7,089	4,726	0.84 (0.80–0.89)	1.55×10^{-9}	6,768	755	0.90 (0.81–1.01)	0.078	0.060
rs2395185[d] (rs28366298)	HLA class II region	6p21.32	Meta				7,390	4,696	1.20 (1.13–1.28)	9.47×10^{-10}	7,211	742	1.05 (0.93–1.18)	0.42	0.56

[a]Minor allele listed second. [b]Minor allele frequency. 1, MAF in controls; 2, MAF in adenocarcinoma; 3, MAF in squamous carcinoma. [c]Tested by case-case analysis. [d]For the HLA class II region, because a TaqMan assay could not be designed for rs2395185, we instead genotyped rs28366298, its perfect surrogate ($r^2 = 1.0$), by TaqMan. The reported *P* value is based on meta-analysis of the rs2395185 results in the GWAS and the rs28366298 results in the TaqMan set.

Imputation Reference Set version 1 (ref. 27) using the IMPUTE2 program[28] across a 1-Mb region centered on the index SNP (Online Methods). For the two regions outside of the human leukocyte antigen (HLA) region, the association analysis did not identify new signals that were substantially stronger than those found for the genotyped SNPs (**Supplementary Fig. 3a,b**). Although there seem to be stronger signals in the imputed data for the HLA class II region (**Supplementary Fig. 3c**), HLA typing will be necessary to unravel the specific haplotypes involved.

At the 6q22 locus, six SNPs were highly correlated with rs9387478 (r^2 = 0.99–1.00). Two SNPs, rs9387478 and rs6937083 (pairwise $r^2 = 1$), were observed within a region defined by the Encyclopedia of DNA Elements (ENCODE) as containing both chromatin state segmentation and enhancer- and promoter-associated histone marks. Although the evidence for evolutionary conservation is weak (that is, a cross-species sequence alignment comparison indicated conservation at the site of ~29.2 million years since divergence from a common ancestor), rs6937083 falls within an ENCODE-predicted transcription factor–binding site and an exon of the AceView-predicted gene, *DCBLD1*. The architecture of the region on chromosome 10q25 is more complicated because there are 23 perfectly correlated SNPs ($r^2 = 1$) and 1 highly correlated SNP ($r^2 = 0.99$). All localize to intron 7 or the UTR of one transcript of the *VTI1A* gene (encoding vesicle transport through interaction with t-SNAREs homolog 1A (yeast)). Sixteen fall within putatively functional regions, defined as ENCODE DNase I hypersensitivity clusters, chromatin state segmentation, the UTR of *VTI1A*, ENCODE enhancer- and promoter-associated histone marks and/or highly conserved (that is, a cross-species sequence alignment comparison indicated conservation at the site of 300 million years since divergence from a common ancestor) regions (**Supplementary Table 5**). rs11196080 is noteworthy because many of the functionally predicted areas converge on this SNP, making this a high-priority variant for functional follow-up studies.

The strongest new association signal, rs7086803 at 10q25.2, maps to intron 7 of the *VTI1A* gene, which has been implicated in lung carcinogenesis. Loss of VTI1A activity has been reported to reduce high-frequency spontaneous neurotransmitter release[29] and rapid progressive neurodegeneration in the peripheral ganglia[30]. VTI1A is also involved in Acrp30-containing vesicles in adipocytes, and lower amounts of VTI1A in cultured adipocytes can inhibit adiponectin secretion[31]. Lower amounts of adiponectin have previously been associated with advanced lung cancer[31,32]. A recent study reported recurrent *VTI1A-TCF7L2* fusions in colorectal cancers, and a colorectal carcinoma cell line with the fusion gene was shown to be dependent on *VTI1A-TCF7L2* for anchorage-independent growth[33].

The rs9387478 SNP at 6q22.2 is located in an interval that contains two candidate genes: *DCBLD1* (encoding discoidin, CUB and LCCL domain containing 1) and *ROS1* (encoding the ROS proto-oncogene receptor tyrosine kinase). ROS1 functions as both an integral membrane protein and a receptor tyrosine kinase[34]. Expression of *Ros1* is specifically increased in lung cancer tissue in mouse models, and *ROS1* expression levels are higher in non-small cell lung cancer (NSCLC)[35]. *ROS1* fusions in lung adenocarcinoma and NSCLC, particularly in Asian never smokers, have been identified as drivers of oncogenesis[36–38]. *ROS1* rearrangements were found to be more common in lung adenocarcinomas from never smokers and younger affected individuals[39]. There is limited evidence concerning the functional role of the protein encoded by *DCBLD1*; a related gene at 3q12.2, *DCBLD2* (encoding discoidin, CUB and LCCL domain containing 2; also known as *CLCP1*) regulates cellular proliferation and invasion and may have an important role in cancer metastasis[40–42].

The third locus, marked by rs2395185 at 6p21.3, is located within 20 kb of *HLA-DRA* (encoding major histocompatibility complex, class II, DRα) and 52 kb downstream of *HLA-DRB5* (encoding major histocompatibility complex, class II, DRβ5). There was no evidence for strong linkage disequilibrium (LD) between this SNP and other SNPs at 6p21.32 reported to be associated with lung cancer[17,23]. There was little LD with a recently reported SNP at 6p21.3, rs3817963, which was associated with lung cancer in a Japanese population[20]; the r^2 in Han Chinese and Japanese HapMap samples was 0.18 and 0.10, respectively, and *D* was 0.57 and 0.43, respectively. These data suggest that our locus probably represents a new HLA class II–related finding for nonsmoking lung cancer susceptibility. Further mapping across the complex HLA region is required to characterize the specific susceptibility alleles or haplotypes involved in nonsmoking lung cancer risk. We also note that rs2395185 has been previously associated with ulcerative colitis[43], Hodgkin lymphoma[44] and type 1 diabetes[45].

In previous GWAS of lung cancer, in which a majority of cases were smokers, SNPs across a region at 15q25 have been associated with lung cancer risk[12,19,21–24]. However, studies of smoking-related behavior have also identified variants at 15q25, raising the possibility that the variants previously identified by GWAS for lung cancer could mediate risk through effects on tobacco use[46]. We previously genotyped additional SNPs across 15q25 in Asian studies and observed no evidence of association with lung cancer in never-smoking Asian females[11]. Notably, in our current, larger study, there was no evidence for association with lung cancer at 15q25 in the never-smoking population overall or in the major subtypes. These data provide strong evidence that this locus is not associated with lung cancer independent of smoking in never-smoking females in Asia, which contrasts with the results from a smaller Asian study[24] but is consistent with previous reports from smaller studies conducted in populations of European ancestry[12,47,48].

We investigated the relationship between our new loci and known environmental exposures. The association between exposure to

environmental tobacco smoke in the home and adenocarcinoma in the five studies with data available yielded an OR of 1.36 ($P = 1.2 \times 10^{-4}$) in an analysis of 1,770 cases and 2,675 controls, consistent with previous reports[8]. The effect of environmental tobacco smoke was stronger for subjects with the GG genotype at rs2395185, with OR = 1.78 ($P = 1.15 \times 10^{-5}$), compared to subjects with the GT or TT genotypes, OR = 1.16 ($P = 0.15$), with $P_{interaction} = 0.002$. The association between the T allele at rs2395185 and risk of adenocarcinoma in subjects with and without exposure to environmental tobacco smoke yielded OR = 1.13 ($P = 0.031$) and OR = 1.43 ($P = 5.6 \times 10^{-4}$), respectively, with $P_{interaction} = 0.037$. There was no evidence of interaction with the other two new loci reported here.

In summary, we conducted a GWAS of lung cancer in never-smoking females in Asia and identified three new susceptibility loci at 10q25.2, 6q22.2 and 6p21.32. We also confirmed associations with two previously reported regions at 5p15.3 and 3q28 and a recently reported locus at 17q24.3. It is notable that our strongest finding at 10q25.2 has not been reported previously in lung cancer GWAS. This observation suggests that the etiology of lung cancer in never smokers in Asia may have unique genetic characteristics. This is consistent with the distinct pattern of environmental risk factors that have been causally linked to lung cancer in never-smoking females in Asia[4–8,10] and the distinct molecular phenotypes of lung cancer in never smokers[2,3]. Further work is warranted to map the new regions. Functional work is required to identify the variants that directly account for the underlying association, as well as to study how the genetic variants interact with established environmental risk factors, including environmental tobacco smoke, cooking fumes and fuel use, in never-smoking females in Asia.

URLs. CGF, http://cgf.nci.nih.gov/; GLU, http://code.google.com/p/glu-genetics/; EIGENSTRAT, http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm; Structure, http://pritch.bsd.uchicago.edu/structure.html; IMPUTE2, http://mathgen.stats.ox.ac.uk/impute/impute_v2.html; SNPTEST, https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html; liftOver, http://hgdownload.cse.ucsc.edu/downloads.html; SAS v9.2 (used to generate forest plots), http://support.sas.com/kb/43/855.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The CGEMS data portal provides access to individual-level data for investigators from certified scientific institutions after approval of their submitted Data Access Request.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J.-J. Yang, X.-N. Yang, Q. Zhou, W.-B. Guo, S.-L. Chen, Y. Huang, Z. Xie, J.-G. Chen, H.-H. Yan, K. Tajima, Y. Yatabe, T. Hida, K.-L. Chuah, A. Ng, P. Eng, S.-S. Leong, M.-K. Ang, E. Lim, T.-K. Lim, M. Teh, W.-T. Poh and A. Tee. The overall GWAS project was supported by the intramural program of the US National Institutes of Health/National Cancer Institute. A list of support provided to individual studies is provided in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

Q.L., N.R., S.J.C., D. Lin, C.A.H., Y.-C.H., K.M., A.S., H.D.H., J.Y.P., C.-J.C., Y.H.K., Y.T.K., C.L., Y.-L.W., P.-C.Y., B.Z., M.-H.S., J.F.F., K.C., W.Z., T.W., H.S., I.-S.C., D. Lu, N. Caporaso, W.P., R.K., J. Liu, M.T.L., N. Chatterjee, M.T. and M.Y. organized and designed the study. S.J.C., D. Lin, R.K., J. Liu, C.A.H., K.M., T.W., L.B., M.Y., J. Yuenger, Z.Y., C.W., H.G., A.H., W.W., Y.L., W.P., H.-C.L. and B.Z. conducted and supervised the genotyping of samples. Z.W., K.B.J., N.R., Q.L., S.J.C., N. Chatterjee, C.A.H., H.D.H., W.H., M.Y., I.-S.C., C.-F.H., W.-C.W., C.C.C., S.I.B., C.-H.C., R.V. and Y.-H.C. contributed to the design and execution of statistical analysis. Q.L., N.R., S.J.C., Z.W., W.H., C.C.C., C.A.H., K.M., Y.-C.H., A.S., H.D.H., N. Chatterjee, N. Caporaso, C.L., M.Y., B.A.B., M.T., S.-J.A., S.I.B., M.T.L., C.K., R.V., Y.-L.W., J.F.F. and I.-S.C. wrote the first draft of the manuscript. C.A.H., Q.L., B.Z., Y.-C.H., K.M., A.S., K.C., J.-C.W., M.P.W., W.Z., J.Y.P., W.H., C.-J.C., Y.H.K., Y.T.K., T.W., H.S., I.-S.C., T.M., H.N.K., F.W., Z.Y., C.W., S.-J.A., G.-C.C., B.Q., V.H.F.L., D. Lu, H.-S.J., J.S.S., J.H.K., Y.-T.G., Y.-H.T., Y.J.J., H.G., Z.H., I.-J.O., K.-Y.C., X.H., W.W., J.C., X.-C.Z., M.-S.H., H.Z., J. Wang, X.Z., J.E.C., W.-C.S., K.H.P., S.W.S., X.-O.S., Y.-M.C., L.L., C.H.K., L.H., Y.-C.K., T.-Y.Y., J.X., P.G., W.T., J.S., C.-L.W., H.L., A.D.L.S., Z.Z., Y.C., Y.Y.C., J.-Y.H., J.S.K., H.-I.Y., Q.C., C.-C.L., I.K.P., P.X., J.D., Q.H., R.-P.P., T.K., S.-S.K., C.-Y.C., R.V., J. Wu, W.-Y.L., K.-C.C., W.-H.C., B.-T.J., J.K.C.C., M.C., Y.-J.L., J. Yokota, J. Li, H.C., Y.-B.X., C.-J.Y., H.K., G.W., L.J., Y.-L.L., K.S., Y.-L.W., P.-C.Y., M.-H.S., J.F.F., D. Lin, S.J.C. and N.R. conducted the epidemiological studies and contributed samples to the GWAS and/or follow-up genotyping. All authors contributed to the writing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2456.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Sun, S., Schiller, J.H. & Gazdar, A.F. Lung cancer in never smokers—a different disease. *Nat. Rev. Cancer* **7**, 778–790 (2007).
2. Sun, Y. *et al.* Lung adenocarcinoma from East Asian never-smokers is a disease largely defined by targetable oncogenic mutant kinases. *J. Clin. Oncol.* **28**, 4616–4620 (2010).
3. Rudin, C.M. *et al.* Lung cancer in never smokers: molecular profiles and therapeutic implications. *Clin. Cancer Res.* **15**, 5646–5661 (2009).
4. Thun, M.J. *et al.* Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. *PLoS Med.* **5**, e185 (2008).
5. Gao, Y.T. *et al.* Lung cancer among Chinese women. *Int. J. Cancer* **40**, 604–609 (1987).
6. Gu, D. *et al.* Cigarette smoking and exposure to environmental tobacco smoke in China: the international collaborative study of cardiovascular disease in Asia. *Am. J. Public Health* **94**, 1972–1976 (2004).
7. Lan, Q., Chapman, R.S., Schreinemachers, D.M., Tian, L. & He, X. Household stove improvement and risk of lung cancer in Xuanwei, China. *J. Natl. Cancer Inst.* **94**, 826–835 (2002).
8. Couraud, S., Zalcman, G., Milleron, B., Morin, F. & Souquet, P.J. Lung cancer in never smokers—a review. *Eur. J. Cancer* **48**, 1299–1311 (2012).
9. Samet, J.M. *et al.* Lung cancer in never smokers: clinical epidemiology and environmental risk factors. *Clin. Cancer Res.* **15**, 5626–5645 (2009).
10. Lo, Y.L. *et al.* Risk factors for primary lung cancer among never smokers by gender in a matched case-control study. *Cancer Causes Control* published online, doi:10.1007/s10552-012-9994-x (22 May 2012).
11. Hsiung, C.A. *et al.* The 5p15.33 locus is associated with risk of lung adenocarcinoma in never-smoking females in Asia. *PLoS Genet.* **6**, e1001051 (2010).
12. Landi, M.T. *et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.* **85**, 679–691 (2009).
13. Hosgood, H.D. III *et al.* Genetic variant in *TP63* on locus 3q28 is associated with risk of lung adenocarcinoma among never-smoking females in Asia. *Hum. Genet.* **131**, 1197–1203 (2012).
14. Miki, D. *et al.* Variation in *TP63* is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. *Nat. Genet.* **42**, 893–896 (2010).
15. Amundadottir, L. *et al.* Genome-wide association study identifies variants in the *ABO* locus associated with susceptibility to pancreatic cancer. *Nat. Genet.* **41**, 986–990 (2009).
16. de Bakker, P.I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
17. Hu, Z. *et al.* A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nat. Genet.* **43**, 792–796 (2011).
18. McKay, J.D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40**, 1404–1406 (2008).
19. Truong, T. *et al.* Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J. Natl. Cancer Inst.* **102**, 959–971 (2010).
20. Shiraishi, K. *et al.* A genome-wide association study identifies two new susceptibility loci for lung adenocarcinoma in the Japanese population. *Nat. Genet.* **44**, 900–903 (2012).
21. Amos, C.I. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* **40**, 616–622 (2008).
22. Hung, R.J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633–637 (2008).

23. Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.* **40**, 1407–1409 (2008).
24. Wu, C. *et al.* Genetic variants on chromosome 15q25 associated with lung cancer risk in Chinese populations. *Cancer Res.* **69**, 5065–5072 (2009).
25. Shi, J. *et al.* Inherited variation at chromosome 12p13.33, including *RAD52*, influences the risk of squamous cell lung carcinoma. *Cancer Discov.* **2**, 131–139 (2012).
26. Dong, J. *et al.* Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat. Genet.* **44**, 895–899 (2012).
27. Wang, Z. *et al.* Improved imputation of common and uncommon SNPs with a new reference set. *Nat. Genet.* **44**, 6–7 (2012).
28. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
29. Ramirez, D.M., Khvotchev, M., Trauterman, B. & Kavalali, E.T. Vti1a identifies a vesicle pool that preferentially recycles at rest and maintains spontaneous neurotransmission. *Neuron* **73**, 121–134 (2012).
30. Kunwar, A.J. *et al.* Lack of the endosomal SNAREs vti1a and vti1b led to significant impairments in neuronal development. *Proc. Natl. Acad. Sci. USA* **108**, 2575–2580 (2011).
31. Bose, A. *et al.* The v-SNARE Vti1a regulates insulin-stimulated glucose transport and Acrp30 secretion in 3T3-L1 adipocytes. *J. Biol. Chem.* **280**, 36946–36951 (2005).
32. Petridou, E.T. *et al.* Circulating adiponectin levels and expression of adiponectin receptors in relation to lung cancer: two case-control studies. *Oncology* **73**, 261–269 (2007).
33. Bass, A.J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion. *Nat. Genet.* **43**, 964–968 (2011).
34. Lemmon, M.A. & Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117–1134 (2010).
35. Acquaviva, J., Wong, R. & Charest, A. The multifaceted roles of the receptor tyrosine kinase ROS in development and cancer. *Biochim. Biophys. Acta* **1795**, 37–52 (2009).
36. Li, C. *et al.* Spectrum of oncogenic driver mutations in lung adenocarcinomas from East Asian never smokers. *PLoS ONE* **6**, e28204 (2011).
37. Rikova, K. *et al.* Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. *Cell* **131**, 1190–1203 (2007).
38. Takeuchi, K. *et al. RET, ROS1* and *ALK* fusions in lung cancer. *Nat. Med.* **18**, 378–381 (2012).
39. Bergethon, K. *et al. ROS1* rearrangements define a unique molecular class of lung cancers. *J. Clin. Oncol.* **30**, 863–870 (2012).
40. Kim, M. *et al.* Epigenetic down-regulation and suppressive role of *DCBLD2* in gastric cancer cell proliferation and invasion. *Mol. Cancer Res.* **6**, 222–230 (2008).
41. Koshikawa, K. *et al.* Significant up-regulation of a novel gene, *CLCP1*, in a highly metastatic lung cancer subline as well as in lung cancers *in vivo*. *Oncogene* **21**, 2822–2828 (2002).
42. Nagai, H. *et al.* CLCP1 interacts with semaphorin 4B and regulates motility of lung cancer cells. *Oncogene* **26**, 4025–4031 (2007).
43. Silverberg, M.S. *et al.* Ulcerative colitis–risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.* **41**, 216–220 (2009).
44. Urayama, K.Y. *et al.* Genome-wide association study of classical Hodgkin lymphoma and Epstein-Barr virus status–defined subgroups. *J. Natl. Cancer Inst.* **104**, 240–253 (2012).
45. Nakanishi, K. & Shima, Y. Capture of type 1 diabetes–susceptible HLA DR-DQ haplotypes in Japanese subjects using a tag single nucleotide polymorphism. *Diabetes Care* **33**, 162–164 (2010).
46. Chanock, S.J. & Hunter, D.J. Genomics: when the smoke clears. *Nature* **452**, 537–538 (2008).
47. Spitz, M.R., Amos, C.I., Dong, Q., Lin, J. & Wu, X. The *CHRNA5-A3* region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer. *J. Natl. Cancer Inst.* **100**, 1552–1556 (2008).
48. Wang, Y., Broderick, P., Matakidou, A., Eisen, T. & Houlston, R.S. Chromosome 15q25 (*CHRNA3-CHRNA5*) variation impacts indirectly on lung cancer risk. *PLoS ONE* **6**, e19085 (2011).

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA. [2]Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan. [3]Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan. [4]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea. [5]Saw Swee Hock School of Public Health, National University of Singapore, Singapore. [6]Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA. [7]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York, USA. [8]Department of Epidemiology and Biostatistics, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China. [9]Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China. [10]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China. [11]Department of Pathology, Li Ka Shing (LKS) Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [12]Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Institute for Medicine and Public Health, Vanderbilt University, Nashville, Tennessee, USA. [13]Lung Cancer Center, Kyungpook National University Medical Center, Daegu, Republic of Korea. [14]Genomic Research Center, Taipei, Taiwan. [15]Department of Internal Medicine, Division of Oncology/Hematology, College of Medicine, Korea University Anam Hospital, Seoul, Republic of Korea. [16]Department of Thoracic and Cardiovascular Surgery, Cancer Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea. [17]Ministry of Education Key Laboratory of Modern Toxicology, Nanjing Medical University, Nanjing, China. [18]Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment, Nanjing Medical University, Nanjing, China. [19]Westat, Rockville, Maryland, USA. [20]National Institute of Cancer Research, National Health Research Institutes, Zhunan, Taiwan. [21]Division of Thoracic Surgery, Kinki University School of Medicine, Sayama, Japan. [22]Genome Research Center for Hematopoietic Diseases, Chonnam National University Hwasun Hospital, Hwasun-eup, Republic of Korea. [23]Department of Medicine, School of Medicine, National Yang-Ming University, Taipei, Taiwan. [24]Division of Chest Medicine, Department of Internal Medicine, Taichung Veterans General Hospital, Taichung, Taiwan. [25]Division of Environmental Health Sciences, Yale School of Public Health, New Haven, Connecticut, USA. [26]China National Environmental Monitoring Center, Beijing, China. [27]Department of Epidemiology, School of Public Health, China Medical University, Shenyang, China. [28]Department of Etiology & Carcinogenesis, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [29]State Key Laboratory of Molecular Oncology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [30]Guangdong Lung Cancer Institute, Medical Research Center and Cancer Center of Guangdong General Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China. [31]Department of Clinical Oncology, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [32]Department of Human Genetics, Genome Institute of Singapore, Singapore. [33]School of Life Sciences, Anhui Medical University, Hefei, China. [34]Cancer Research Center, Kyungpook National University Medical Center, Daegu, Republic of Korea. [35]Institute of Environmental Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea. [36]Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China. [37]Department of Respiratory Therapy, Chang Gung Memorial Hospital, Chiayi, Taiwan. [38]Institute of Occupational Medicine and Ministry of Education Key Laboratory for Environment and Health, School of Public Health, Huazhong University of Science and Technology, Wuhan, China. [39]Program in Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [40]Lung and Esophageal Cancer Clinic, Chonnam National University Hwasun Hospital, Hwasun-eup, Republic of Korea. [41]Department of Internal Medicine, Chonnam National University Medical School, Gwangju, Republic of Korea. [42]Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan. [43]Chinese Center for Disease Control and Prevention, Beijing, China. [44]Department of Internal Medicine, Kaohsiung Medical University Hospital, School of Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan. [45]Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [46]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [47]Department of Internal Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan. [48]Department of Thoracic and Cardiovascular Surgery, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. [49]Chest Department, Taipei Veterans General Hospital, Taipei, Taiwan. [50]Cancer Center, Union Hospital, Huazhong University of Science and Technology, Wuhan, China. [51]Division of Hematology and Oncology, Vanderbilt University Medical Center, Nashville, Tennessee, USA. [52]School of Public Health, The University of Hong Kong, Hong Kong, China. [53]Department of Pulmonary and Critical Care, Chang Gung Memorial Hospital, Taoyuan, Taiwan. [54]Department of Surgery, Division of Cardiothoracic Surgery, Queen Mary Hospital, Hong Kong, China. [55]Department of Internal Medicine, Division of Medical Oncology, College of Medicine, Korea University Guro Hospital, Seoul, Republic of Korea. [56]Department of Internal Medicine, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. [57]Department of Oncology, Wuhan Iron and Steel Corporation Staff Worker Hospital, Wuhan, China. [58]Division of Genome Biology, National Cancer Center Research Institute, Tokyo, Japan. [59]Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Hwasun-eup, Republic of Korea. [60]Department of Preventive Medicine, Chonnam National University Medical School, Gwangju, Republic of Korea. [61]Cancer Center, China Medical University and Hospital, Taichung, Taiwan. [62]Division of Environmental Epidemiology, Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht, The Netherlands. [63]Department of Pathology, Queen Elizabeth Hospital, Hong Kong, China. [64]Division of Multistep Carcinogenesis, National Cancer Center Research Institute, Tokyo, Japan. [65]Qujing Center for Diseases Control and Prevention, Sanjiangdadao, Qujing, China. [66]Department of Respiratory Medicine, Mitsui Memorial Hospital, Tokyo, Japan. [67]Department of Internal Medicine, National Taiwan University College of Medicine, Taipei, Taiwan. [68]These authors contributed equally to this work. [69]These authors jointly directed this work. Correspondence should be addressed to Q.L. (qingl@mail.nih.gov).

ONLINE METHODS

Study participants. Participants were drawn from 14 studies (**Supplementary Table 1**). Cases had histologically confirmed lung cancer. Each participating study obtained informed consent from study participants and approval from its respective institutional review board for this study. Studies obtained institutional certification permitting data sharing in accordance with the US National Institutes of Health (NIH) Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS), with the exception of the component of the GELAC study that was not scanned at the NCI.

Genotyping and quality control. Genome-wide scanning data came from two sources. Internal sets (HKS, SNU, CNULCS, SWHS, YLCS and components of samples from Japan and GELAC) were genotyped at the NCI CGR Laboratory using the Illumina 660W SNP microarray. External sets were genotyped as follows: (i) samples from CAMSCH, FLCS, GDS, SLCS, TLCS and WLCS were genotyped on contract at Gene-Square Biotech in Beijing using the Illumina 660W SNP microarray; (ii) samples from GELAC were genotyped on contract at GeneTech Biotech in Taiwan on the Illumina 370K SNP microarray in a pilot project, and remaining samples were genotyped on contract at deCODE Genetics in Iceland using the Illumina 610Q SNP microarray and initially reported elsewhere[11]; (iii) a subset of samples from Japan were genotyped at MSKCC using the Illumina 610Q SNP microarray; and (iv) samples from Singapore were genotyped at the Genome Institute of Singapore on the Illumina 660W SNP microarray. The scanned intensity data from external sources were collected, and genotypes were clustered and called at the CGR using Illumina Genome Studio v2011.1 on the basis of the GenTrain2 calling algorithm. Genotype clusters were estimated from samples with preliminary completion rates of greater than 98% per cluster group.

Genotyping was attempted for a total of 5,568 samples on the Illumina 660W SNP microarray at the CGR. Six samples could not be loaded into Illumina Genome Studio because of their low intensities, and 16 samples failed to scan because of broken chips. In addition, a total of 5,946 samples were genotyped at Gene-Square Biotech (3,828), deCODE Genetics and GeneTech Biotech (1,232), MSKCC (374) and the Genome Institute of Singapore (512); the distribution of samples genotyped per SNP microarray chip was as follows: Illumina 660W (4,340), Illumina 610Q (1,494) and Illumina 370K (112). Seven samples (all from Gene-Square Biotech) could not be loaded into Illumina Genome Studio because of their low intensities. In addition, 111 samples from 4 studies (FLCS, GDS, SLCS and TLCS) were excluded due to laboratory processing errors. The combined 11,374 samples with genotypes mapped to 11,025 unique individuals drawn from 14 studies.

We subsequently performed quality control filtering at the sample level in 19 quality control groups (**Supplementary Table 6**). Samples were exclude that had low completion rates ($n = 725$ samples) and extreme mean heterozygosity rates ($n = 116$). Thresholds were chosen on the basis of the sample completion rate or sample mean heterozygosity distribution for each quality control group (**Supplementary Table 6**) and on the basis of discordant expected duplicate samples ($n = 6$). There were samples that were excluded for multiple reasons, and the total number of unique samples excluded was 761 (**Supplementary Table 6b**). Genotype data for the remaining 10,613 samples were merged, resulting in data from 10,312 unique individuals. The genotype concordance rate for expected duplicates ($n = 311$) was greater than 99.9%. Further quality control analysis at the individual level led to the exclusion of samples with (i) gender discordance ($n = 94$); (ii) less than 86% Asian ancestry ($n = 3$); (iii) first-degree relatives who were also genotyped in the study ($n = 136$ subjects); and (iv) incomplete phenotype or unknown histology, as well as those who had ever smoked or were deemed ineligible ($n = 15$). Thus, the total number of scanned subjects after both quality control and analytic exclusions was 10,054 (5,510 cases and 4,544 controls). A summary of the number of excluded loci by study is shown in **Supplementary Table 6c**.

TaqMan custom genotyping assays (Applied Biosystems) were designed and optimized for 13 SNPs, including 9 in the NCI scan data and 4 surrogates not in this scan. In an analysis of 385 samples from 7 studies, comparison of the Illumina calls with the results from TaqMan assays conducted at the NCI CGR showed an average concordance rate of 99.97% (with a range of 99.7–100%) for the overlapping 9 SNPs. The Cancer Institute and Hospital at the Chinese Academy of Medical Sciences also conducted TaqMan genotyping for 7 SNPs on 201 previously scanned samples from 5 studies. Comparison of the Illumina calls with the results of TaqMan assays showed an average concordance rate of 99.93% (with a range of 99.5–100%). In examining the concordance between rs2395185 (scan) and its perfect surrogate rs28366298 (TaqMan), we applied genotype mapping GG→AA, GT→AC and TT→CC to confirm reproducibility of genotyping between platforms.

For the replication phase, we analyzed an additional 3,933 individuals (1,023 cases and 2,910 controls) with TaqMan data, and an additional 79 individuals (76 cases and 3 controls) genotyped using the Illumina 660W array at Gene-Square Biotech were available for analysis. Thus, the final number of subjects included in the analyses was 14,066 (6,609 cases and 7,457 controls; **Supplementary Table 1**). SNP assays with locus call rates lower than 90% or Hardy-Weinberg equilibrium P values less than 1.0×10^{-7} in each quality control group were excluded. In total, 596,032 SNPs remained in the analytic data set. After setting the minimum minor allele frequency (MAF) to 0.01, we excluded 83,806 loci from the association analysis. Thus, 512,226 SNPs were analyzed in the association studies reported here.

Statistical analyses. Data analysis and management were performed with GLU (Genotyping Library and Utilities version 1.0), a suite of tools available as an open-source application for the management, storage and analysis of GWAS data. Assessment of the population structure of study participants was performed with the GLU struct.admix module using the Japanese in Tokyo, Japan (JPT) and Han Chinese in Beijing, China (CHB), Utah residents of Northern and Western European ancestry (CEU) and Yoruba from Ibadan, Nigeria (YRI) samples as the reference populations (HapMap Build 28). A set of 33,165 SNPs with low pairwise correlation ($r^2 < 0.01$) was selected for this analysis. Three individuals were estimated to have less than 86% Asian ancestry (**Supplementary Fig. 4**).

The genotypes for all subject pairs were computed for cryptic relatedness using the GLU qc.ibds module with the same set of selected SNPs. In addition to 68 pairs of unexpected duplicates, we detected 33 parent-offspring and 41 full-sibling pairs. For the 142 unexpected duplicates and first-degree relative pairs, 1 subject from each simple pair was excluded. For each family with multiple relative pairs detected, only one randomly chosen subject was included in the principal-components analysis (PCA). To address the underlying population substructure, PCA was conducted using the GLU struct.pca module, a program similar to EIGENSTRAT[49,50], with the same set of SNPs (**Supplementary Fig. 5a,b**). Three samples with less than 86% Asian ancestry were excluded on the basis of PCA.

Association analysis. Association analyses were conducted using logistic regression, adjusted for age (in 10-year categories), study group and eigenvectors, if they were significant when analyzed in the base models. For analysis of all cases versus controls, we adjusted for EV1, EV2 and EV4. For analysis of adenocarcinoma cases versus controls, we adjusted for EV2 and EV4. For analysis of squamous cell cases versus controls, we adjusted for EV8. Each SNP genotype was coded as a count of minor alleles (trend effect). A score test with 1 degree of freedom was performed on all genetic parameters in each model to determine statistical significance. The unscaled λ value for all cases versus controls in the main effect model was 1.014, and λ_{1000} was 1.003, with corrected λ calculated as $\lambda_{corrected} = 1 + (\lambda - 1) \times (n_{case}^{-1} + n_{control}^{-1})/(2 \times 10^{0-3})$.

We assessed heterogeneity in genetic effects across studies using the Cochran's Q statistic, which conforms to a χ-squared distribution with $k - 1$ degree of freedom, where k is the number of studies.

For the inclusion of TaqMan data for the SNPs that failed assay design (rs2395185 and rs10197940), we conducted a fixed-effects meta-analysis by combining the aggregate results from their perfect surrogates (rs28366298 and rs2290368, respectively) scanned in the GWAS with their own results based only on the additional TaqMan samples not used in the GWAS association analyses.

Genotype-environment interactions with environmental tobacco smoke were assessed using logistic regression for studies with such information available and adjusted by age, study group, the main effect of the SNP and environmental tobacco smoke, and the interaction term.

doi:10.1038/ng.2456

Estimate of recombination hotspots. To identify recombination hotspots in the region, we used SequenceLDhot[51], a program that uses the approximate marginal likelihood method[52] and calculates likelihood ratio statistics at a set of possible hotspots. Drawn from scanned controls, 100 individuals were randomly sampled from Han Chinese, Japanese and Korean samples. Three independent recombination hotspot inferences were analyzed and are represented as three different colored lines in **Figure 1.** Specifically, for the *VTI1A* regional plot, genotypes of 70 SNPs spanning chromosome 10 114,362,000–114,593,000 (UCSC Genome Build hg18) were phased using PHASE v2.1 (ref. 53) to calculate background recombination rates. The PHASE outcome was used as direct input for the SequenceLDhot program, and LD was estimated as r^2 for 70 SNPs within a ~230-kb region, and a heatmap was drawn using the snp.plotter program[54]. Similarly, we started with the genotypes of 63 SNPs for the *ROS1-DCBLD1* regional plot and the genotypes of 59 SNPs for the HLA class II locus.

Imputation analysis. To begin to fine map newly identified regions, we imputed all the SNPs catalogued in 1000 Genomes Project data, March 2012 release, and the DCEG Imputation Reference Set version 1 (ref. 27). The IMPUTE2 program[28] was used to impute a 1-Mb region centered on the index SNP for each of the three regions, using recommended default settings. Imputed SNPs with INFO of <0.3 were excluded from association analysis using the SNPTEST program v2.3 (see URLs), which considered probabilistic genotypes out of imputation. Because 1000 Genomes Project data was based on the NCBI Build 37 reference genome, we conducted liftOver (see URLs) on our scan data from Build 36 to 37 before imputation.

49. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
50. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
51. Fearnhead, P. SequenceLDhot: detecting recombination hotspots. *Bioinformatics* **22**, 3061–3066 (2006).
52. Fearnhead, P., Harding, R.M., Schneider, J.A., Myers, S. & Donnelly, P. Application of coalescent methods to reveal fine-scale rate variation and recombination hotspots. *Genetics* **167**, 2067–2081 (2004).
53. Abnet, C.C. *et al.* Genotypic variants at 2q33 and risk of esophageal squamous cell carcinoma in China: a meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **21**, 2132–2141 (2012).
54. Luna, A. & Nicodemus, K.K. snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. *Bioinformatics* **23**, 774–776 (2007).

Genetic Epidemiology 36 : 340–351 (2012)

Smoking and Genetic Risk Variation Across Populations of European, Asian, and African American Ancestry—A Meta-Analysis of Chromosome 15q25

Li-Shiun Chen,[1*] Nancy L. Saccone,[2] Robert C. Culverhouse,[3] Paige M. Bracci,[4] Chien-Hsiun Chen,[5,6] Nicole Dueker,[7] Younghun Han,[8] Hongyan Huang,[9] Guangfu Jin,[10] Takashi Kohno,[11] Jennie Z. Ma,[12] Thomas R. Przybeck,[1] Alan R. Sanders,[13] Jennifer A. Smith,[14] Yun Ju Sung,[9] Angie S. Wenzlaff,[15] Chen Wu,[16] Dankyu Yoon,[17,18] Ying-Ting Chen,[5] Yu-Ching Cheng,[19] Yoon Shin Cho,[20,21] Sean P. David,[22,23,24] Jubao Duan,[13] Charles B. Eaton,[24] Helena Furberg,[25] Alison M. Goate,[1] Dongfeng Gu,[26,27] Helen M. Hansen,[28] Sarah Hartz,[1] Zhibin Hu,[10] Young Jin Kim,[17,22] Steven J. Kittner,[29] Douglas F. Levinson,[30] Thomas H. Mosley,[14] Thomas J. Payne,[31] D. C. Rao,[9] John P. Rice,[1] Treva K. Rice,[9] Tae-Hwi Schwantes-An,[2] Sanjay S. Shete,[8] Jianxin Shi,[32] Margaret R. Spitz,[8] Yan V. Sun,[14] Fuu-Jen Tsai,[33] Jen C. Wang,[1] Margaret R. Wrensch,[28] Hong Xian,[3] Pablo V. Gejman,[13] Jiang He,[34] Steven C. Hunt,[35] Sharon L. Kardia,[14] Ming D. Li,[36] Dongxin Lin,[16] Braxton D. Mitchell,[20] Taesung Park,[37] Ann G. Schwartz,[15] Hongbing Shen,[10] John K. Wiencke,[28] Jer-Yuarn Wu,[5,6] Jun Yokota,[38] Christopher I. Amos,[8] and Laura J. Bierut[1]

[1] *Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri*
[2] *Department of Genetics, Washington University School of Medicine, St. Louis, Missouri*
[3] *Department of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri*
[4] *Department of Epidemiology and Biostatistics, UCSF, San Francisco, California*
[5] *National Genotyping Center Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan*
[6] *Graduate Institute of Chinese Medical Science, China Medical University, Taichung, Taiwan*
[7] *Department of Epidemiology and Public Health, University of Maryland, Baltimore, Maryland*
[8] *Department of Epidemiology Anderson Cancer Center, University of Texas M.D., Houston, Texas*
[9] *Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri*
[10] *Department of Epidemiology and Biostatistics, Nanjing Medical University, Nanjing, China*
[11] *Division of Genome Biology, National Cancer Center Research Institute,, Tokyo, Japan*
[12] *Department of Public Health Sciences, University of Virginia, Charlottesville, Virginia*
[13] *Department of Psychiatry and Behavioral Sciences North Shore University Health System Research Institute, University of Chicago, Chicago, Illinois*
[14] *Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan*
[15] *Karmanos Cancer Institute, Wayne State University, Detroit, Michigan*
[16] *Department of Etiology & Carcinogenesis Cancer Institute, Chinese Academy of Medical Sciences, Beijing, China*
[17] *Interdisciplinary Program in Bioinformatics College of Natural Science, Seoul National University, Seoul, Korea*
[18] *Center for Immunology and Pathology, National Institute of Health, Seoul, Korea*
[19] *Department of Medicine, University of Maryland Medical Center, Baltimore, Maryland*
[20] *Center for Genome Science, National Institute of Health, Seoul, Korea*
[21] *Department of Biomedical Science, Hallym University, Chuncheon, Korea*
[22] *Center for Health Sciences, SRI International, Menlo Park, California*
[23] *Department of Medicine, Stanford University School of Medicine, California*
[24] *Department of Family Medicine, Brown University, Providence, Rhode Island*
[25] *Department of Epidemiology, Memorial Sloan Kettering Cancer Center, New York, New York*
[26] *Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China*
[27] *Chinese National Center for Cardiovascular Disease Control and Research, Beijing, China*
[28] *Neurological Surgery Division of Epidemiology, Helen Diller Family Cancer Center, San Francisco, California*
[29] *Department of Neurology University of Maryland Baltimore, Maryland*
[30] *Department of Psychiatry and Behavioral Sciences, Stanford University, Palo Alto, California*
[31] *University of Mississippi Medical Center, Jackson, Mississippi*
[32] *Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland*
[33] *School of Post-Baccalaureate Chinese Medicine, China Medical University, Taiwan*
[34] *Department of Epidemiology, Tulane School of Public Health and Tropical Medicine, New Orleans, Louisiana*
[35] *Department of Internal Medicine, University of Utah, Salt Lake City, Utah*
[36] *Department of Psychiatry and Neurobehavioral Sciences, University of Virginia, Charlottesville, Virginia*
[37] *Department of Statistics College of Natural Science, Seoul National University, Seoul, Korea*
[38] *Division of Multistep Carcinogenesis, National Cancer Center Research Institute, Tokyo, Japan*

Recent meta-analyses of European ancestry subjects show strong evidence for association between smoking quantity and multiple genetic variants on chromosome 15q25. This meta-analysis extends the examination of association between distinct genes in the *CHRNA5-CHRNA3-CHRNB4* region and smoking quantity to Asian and African American populations to confirm and refine specific reported associations. Association results for a dichotomized cigarettes smoked per day phenotype in 27 datasets (European ancestry (N = 14,786), Asian (N = 6,889), and African American (N = 10,912) for a total of 32,587 smokers) were meta-analyzed by population and results were compared across all three populations. We demonstrate association between smoking quantity and markers in the chromosome 15q25 region across all three populations, and narrow the region of association. Of the variants tested, only rs16969968 is associated with smoking ($P < 0.01$) in each of these three populations (odds ratio [OR] = 1.33, 95% CI = 1.25–1.42, $P = 1.1 \times 10^{-17}$ in meta-analysis across all population samples). Additional variants displayed a consistent signal in both European ancestry and Asian datasets, but not in African Americans. The observed consistent association of rs16969968 with heavy smoking across multiple populations, combined with its known biological significance, suggests rs16969968 is most likely a functional variant that alters risk for heavy smoking. We interpret additional association results that differ across populations as providing evidence for additional functional variants, but we are unable to further localize the source of this association. Using the cross-population study paradigm provides valuable insights to narrow regions of interest and inform future biological experiments. *Genet. Epidemiol.* 36:340–351, 2012.

Key words: smoking; genetics; meta-analysis; cross-population

Supporting Information is available in the online issue at wileyonlinelibrary.com.
*Correspondence to: Li-Shiun Chen, Department of Psychiatry (Box 8134), Washington University School of Medicine, 660 S. Euclid Ave., St. Louis, MO 63110. E-mail: chenli@psychiatry.wustl.edu
Received 24 August 2011; Revised 24 January 2012; Accepted 30 January 2012
Published online 16 April 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21627

INTRODUCTION

Recent genetic meta-analyses, including tens of thousands of subjects of European ancestry, show strong evidence of association between smoking quantity (cigarettes smoked per day; CPD) and multiple genetic markers on chromosome 15q25 [Liu et al., 2010; Saccone et al., 2010; TAG, 2010; Thorgeirsson et al., 2010]. Those studies synthesized evidence across many independent datasets to highlight specific variants in the region of the *CHRNA5-CHRNA3-CHRNB4* gene cluster associated with smoking behavior in European ancestry subjects. It is important to determine the biological mechanisms underlying these associations; however, the high linkage disequilibrium (LD) in this region among individuals of European ancestry makes it difficult to differentiate potentially causal variants from the many correlated variants. Because the genetic architecture of chromosome 15q25 varies across populations, comparing associations across diverse populations with differing genetic architecture can help refine the region of association and point to variants more likely to have functional relevance [Rotimi and Jorde 2010; Saccone et al., 2008; Zaitlen et al., 2010].

The most robust genetic finding on chromosome 15q25 in subjects of European ancestry is the region tagged by rs16969968, rs1051730, and other correlated variants. This finding has been replicated for smoking-related traits in multiple distinct datasets [Baker et al., 2009; Berrettini et al., 2008; Keskitalo et al., 2009; Saccone et al., 2007, 2009; Sherva et al., 2008; Stevens et al., 2008; Thorgeirsson et al., 2008; Weiss et al., 2008] and has now been reported as the most significant genome-wide association in recent meta-analyses of European ancestry subjects (e.g. rs16969968, $P = 5.57 \times 10^{-72}$, or rs1051730, $P = 2.75 \times 10^{-73}$) [Liu et al., 2010; Saccone et al., 2010; TAG, 2010; Thorgeirsson et al., 2010]. We will use the term "bin" to denote a group of correlated single nucleotide polymorphisms (SNPs) ($r^2 \geq 0.7$) that may constitute the same association signal in European ancestry samples [Carlson et al., 2004]. Under this definition and using the 1000 Genomes Pilot 1 CEU as the European ancestry reference sample [Durbin et al., 2010], the single bin tagged by rs16969968 and rs1051730 includes 52 known variants. This bin, which we will call bin A, groups together and unifies the most significant meta-analysis findings as well as individual dataset reports of SNPs associated with nicotine dependence, heavy smoking, lung cancer, and other smoking-related diseases in European ancestry datasets.

There are additional markers of interest in this region that are not strongly correlated with bin A. Because of the clear association between smoking behavior and bin A, each of the large-scale meta-analyses of European ancestry samples carried out association tests conditional on bin A variants for other SNPs to determine whether additional genetic markers in 15q25 are associated after adjusting for effects of bin A [Liu et al., 2010; Saccone et al., 2010; TAG, 2010; Thorgeirsson et al., 2010]. After conditioning on bin A, the meta-analyses identified additional SNPs in this region associated with smoking behavior. These SNPs can be grouped into three distinct bins (B, C, D) (Table I). Bin B, tagged by rs588765 and rs880395, is associated with genome-wide significance among heavy vs. light smokers but only in analyses conditioning on bin A ($P = 1.2 \times 10^{-9}$) [Saccone et al., 2010]. Notably, bin B is also associated with mRNA levels of *CHRNA5* in brain and lung [Falvella et al., 2010; Smith et al., 2010; Wang et al., 2009]. Bin C, tagged by rs6495308 [Liu et al., 2010], rs2036534 [Thorgeirsson et al., 2010], rs7163730, rs9788682, rs684513 [TAG, 2010], and rs578776 [Saccone et al., 2010], is associated with heavy smoking after conditioning on bin A (*P*-values from 9.1×10^{-5} to 6.3×10^{-9}). In contrast to bin B, bin C is less significant in conditional analysis compared to single

TABLE I. Genetic variants associated with smoking quantity reported in meta-analyses in subjects of European ancestry

Bin	SNP	References	Coded allele	Coded allele frequency		
				CEU	JPT/CHB	ASW/YRI
A	rs16969968	Saccone et al. (2010)	A	0.42	0.03	0.07
A	rs1051730	TAG (2010), Liu et al. (2010), Thorgeirsson et al. (2010)	T	0.42	0.03	0.12
B	rs588765/rs880395	Saccone et al. (2010)	T	0.39	0.05	0.23
C	rs6495308	Liu et al. (2010)	C	0.20	0.80	0.31
C	rs2036534	Thorgeirsson et al. (2010)	C	0.19	0.50	0.24
C	rs7163730	TAG (2010)	G	0.19	0.49	0.25
C	rs9788682	TAG (2010)	A	0.19	0.28	0.20
C	rs684513	TAG (2010)	G	0.20	0.25	0.21
C	rs578776	Saccone et al. (2010)	T	0.24	0.87	0.55
D	rs2869046	Thorgeirsson et al. (2010)	C	0.45	0.39	0.17

Allele frequencies based on 1000 Genomes Pilot 1 and HapMap 3 Release 2.

SNP analysis. Bin D is represented by rs2869046, which also displayed residual association after conditioning on bin A ($P = 4.8 \times 10^{-5}$) [Thorgeirsson et al., 2010]. Markers from these different bins (A, B, C, and D) are only modestly correlated with one another, with $r^2 \leq 0.52$ in the 1000 Genomes Pilot 1 CEU (N = 180; Table II).

Differences in the correlational structure of markers spanning the region 15q25 between populations result in distinct sub-bins of correlated markers among Asian and African American populations that provide an opportunity to refine the source of the previously reported signals. For example, bin A, consisting of 52 variants including rs16969968, separates into 20 sub-bins in Asians (based on 1000 Genomes Pilot 1 JPT/CHB) and 38 sub-bins in African Americans (based on combined information from the 1000 Genomes Pilot 1 YRI and HapMap 3 Release 2 ASW) [Altshuler et al., 2010]. In particular, rs16969968 and rs1051730 are highly correlated in European ancestry ($r^2 = 1$) and Asian populations ($r^2 = 1$), but display only moderate correlation ($r^2 = 0.40$) in the African American population. These differences in genetic architecture can be used to dissect the association signals.

The purpose of this meta-analysis is to determine if bins A, B, C, and D shows consistent association with smoking behavior across populations and, if so, to leverage these differences in genetic correlation across populations to refine the genetic associations in this region previously reported in subjects of European ancestry. We expect a sub-bin showing consistent evidence across all three populations to be more likely to contain a variant altering a biological mechanism. We performed meta-analyses of results from a total of 27 datasets: nine European ancestry samples (used to evaluate consistency with previous results), seven Asian samples, and 11 African American samples. We tested for association between smoking phenotypes and the four distinct bins (A through D) across all three populations. This cross-population study therefore improves our understanding of genetic risk for smoking by highlighting potentially functional variants.

METHODS

SAMPLES

Results from 27 datasets, containing a total of 32,587 smokers with measures of CPD, contributed to the meta-analyses. Of these datasets, nine consisted of European ancestry subjects (N = 14,786), seven consisted of Asians (N = 6,889), and 11 consisted of African Americans (N = 10,912). Twenty datasets were samples of unrelated individuals. The remaining seven datasets were family-based studies, for which the primary analyses involved an extraction of unrelated individuals. To be included in the analyses, each subject was required to have reported smoking cigarettes in his/her lifetime. Genotyping varied among studies from extensive coverage based on genome-wide association genotyping to only a limited number of candidate SNPs genotyped in this 15q25 region. Text S1 provides additional details for each dataset, including recruitment, primary phenotypes, definitions for smokers and CPD, DNA source, genotyping platforms, and genotyping quality control. Table S1 shows the sample size and demographics for each participating dataset. Four of nine datasets of European ancestry were included in the previous report [Saccone et al., 2010] (see Table S1 for the overlap, which involves only European-ancestry samples). The informed consent from participants and approval from the appropriate institutional review boards were obtained.

PHENOTYPES

Smoking quantity was assessed with cigarettes smoked per day (CPD). The primary phenotype for analysis was a dichotomous trait contrasting light smoking controls (CPD $\leq$ 10) to heavy smoking cases (CPD $>$ 20). In addition, a four-level ordered trait (CPD $\leq$ 10; 11 $\leq$ CPD $\leq$ 20; 21 $\leq$ CPD $\leq$ 30; CPD $\geq$ 31 coded as 0, 1, 2, 3, respectively) was developed for confirmatory analysis. The only exception was one study (Women's Health Initiative) that measured smoking amount with different threshold levels (CPD $\leq$ 14, 15 $\leq$ CPD $\leq$ 24, 25 $\leq$ CPD $\leq$ 34, CPD $\geq$ 35), and CPD $\leq$ 14 defined the light smoking controls that was contrasted with CPD $\geq$ 25 as heavy smoking cases.

VARIANTS FOR ANALYSES

Multiple SNPs in 15q25 have been identified as associated with smoking behavior in studies of European ancestry subjects. We focused on the results highlighted in the most powerful studies, namely the large meta-analyses [Liu et al., 2010; Saccone et al., 2010; TAG, 2010; Thorgeirsson

TABLE II. Correlation between the examined variants (r^2) in European, Asian, and African American ancestry populations: Bins A, B, C, and D denote four groups of correlated SNPs ($r^2 \geq 0.7$) in the European ancestry reference sample (1000 Genomes Pilot 1 CEU)

	rs16969968	rs1051730	rs588765	rs880395	rs6495308	rs2036534	rs7163730	rs9788682	rs684513	rs578776	rs2869046
European ancestry											
rs16969968	1.0										
rs1051730	1.0	1.0									
rs588765	0.44	0.44	1.0								
rs880395	0.46	0.46	0.76	1.0							
rs6495308	0.18	0.18	0.11	0.05	1.0						
rs2036534	0.17	0.17	0.1	0.15	0.75	1.0					
rs7163730	0.17	0.17	0.1	0.15	0.75	1.0	1.0				
rs9788682	0.17	0.17	0.1	0.15	0.75	1.0	1.0	1.0			
rs684513	0.1	0.1	0.11	0.16	0.7	0.85	0.85	0.85	1.0		
rs578776	0.23	0.23	0.04	0.01	0.78	0.66	0.66	0.66	0.61	1.0	
rs2869046	0.18	0.18	0.54	0.52	0.07	0.14	0.14	0.14	0.15	0.04	1.0
Asian ancestry											
rs16969968	1.0										
rs1051730	1.0	1.0									
rs588765	0	0	1.0								
rs880395	0	0	0.1	1.0							
rs6495308	0.1	0.1	0.21	0.08	1.0						
rs2036534	0.03	0.03	0.05	0.04	0.07	1.0					
rs7163730	0.03	0.03	0.05	0.04	0.06	0.97	1.0				
rs9788682	0.01	0.01	0.02	0	0.1	0.38	0.39	1.0			
rs684513	0.01	0.01	0.02	0.02	0.08	0.28	0.29	0.64	1.0		
rs578776	0.17	0.17	0.11	0.06	0.62	0.02	0.02	0.02	0.02	1.0	
rs2869046	0.04	0.04	0	0.01	0.01	0.11	0.13	0.13	0.09	0.02	1.0
African American ancestry											
rs16969968	1.0										
rs1051730	0.4	1.0									
rs588765	–	0.04	1.0								
rs880395	–	0.02	0.68	1.0							
rs6495308	0.04	0.06	0.14	0.11	1.0						
rs2036534	0.03	0.05	0.06	0.04	0.09	1.0					
rs7163730	0.03	0.05	0.06	0.04	0.09	1.0	1.0				
rs9788682	0.02	0.03	0.06	0.05	0.03	0.6	0.6	1.0			
rs684513	–	0.04	0.08	0.06	0.16	0.62	0.62	0.48	1.0		
rs578776	0.1	0.01	0.55	0.34	0.22	0.08	0.08	0.05	0.13	1.0	
rs2869046	–	0.03	0.03	0.01	0.13	0	0	0.04	0.02	0.02	1.0

We used SNAP to obtain LD values from HapMap 3 ASW and 1000 Genomes Pilot 1 reference populations. –, data unavailable.

et al., 2010] (Table I). Table II lists the 11 targeted SNPs and illustrates how LD (measured by r^2) structure varies across different populations. We used SNAP [Johnson et al., 2008] with 1000 Genomes Pilot 1 reference samples [Altshuler et al., 2010; Durbin et al., 2010] and HapMap3 ASW to obtain LD estimates for our three populations: CEU for European ancestry, JPT/CHB for Asians, and ASW/YRI for African Americans.

It is important to examine not just the 11 previously identified SNPs listed in Table I, but all SNPs correlated with these 11 SNPs in Europeans. We used a two-step process to define distinct groups of correlated SNPs, which we call bins. First, we grouped previously identified SNPs by their correlation in the 1000 Genomes Pilot 1 CEU (i.e. European ancestry) reference sample, using $r^2 \geq 0.7$ as our threshold [Durbin et al., 2010]. Under this strategy, the 11 previously identified SNPs listed in Table I are partitioned into four groups: Group A (rs16969968, rs1051730), Group B (rs588765, rs880395), Group C (rs6495308, rs2036534, rs7163730, rs9788682, rs684513, rs578776), and Group D (rs2869046) (Table II). From these four groups, we established the bins by including all SNPs correlated ($r^2 \geq 0.7$) in the European ancestry reference sample with at least one of the SNPs defining the bin. The threshold of 0.7 was chosen to provide an inclusive collection of tested SNPs. Using SNAP to obtain correlated variants in a bin based on 1000 Genomes Pilot 1 CEU, we identified 52 SNPs in bin A, 111 SNPs in bin B, 82 SNPs in bin C, and 15 SNPs in bin D.

Next, we partitioned these SNPs within a bin into "sub-bins" based on $r^2 \geq 0.8$ in the Asian and African American populations. The higher threshold of 0.8 was used for sub-bins in the other populations to refine the focus of the

analyses. In Asians, we identified 20 sub-bins for bin A, 39 sub-bins for bin B, 24 sub-bins for bin C, and seven sub-bins for bin D. In African Americans, we identified 38 sub-bins for bin A, 37 sub-bins for bin B, 26 sub-bins for bin C, and seven sub-bins for bin D.

STATISTICAL ANALYSES AND META-ANALYSES

We evaluated the genetic associations between heavy smoking and each genotyped SNP in three populations. Standardized scripts were developed centrally by the coordinating site (Washington University) for analyses of all participating datasets at each individual research center. Results were returned to the coordinating site for quality checks and meta-analyses. Individual SNP analyses were performed using SAS (SAS Institute, Cary, NC).

In each dataset, association between heavy vs. light smoking based on CPD and all SNPs was evaluated with logistic regression models as the primary analysis. Genotypes were coded additively as the number of nonreference alleles, where the reference allele was defined as the major allele in the European ancestry population in dbSNP [Sherry et al., 2001]; consistency of allelic coding was confirmed by comparing allele labels and allele frequencies across all datasets within each population. Age as a continuous variable and gender were included as covariates. Secondary analyses of the four-level CPD trait used linear regression models with the same covariates, assuming that the trait has a simple linear relationship with the predictors.

Analyses were stratified by ancestry: European, Asian, and African American. We evaluated the effect of each bin A SNP using single SNP association analyses. For bins B, C, and D, both single SNP association and conditional analyses controlling for bin A were performed. Analyses conditional on bin A (rs16969968) served as our primary analysis model for bins B, C, and D because they were targeted due to previously reported results of analyses conditional on bin A in European ancestry meta-analyses.

For each ancestry group, every dataset with at least one genotyped SNP in a given sub-bin contributed to the meta-analysis of that sub-bin. For each sub-bin, a SNP was selected as the target. In samples where the target SNP was missing, we used the results from the SNP with highest correlation (r^2) with the target SNP in the sub-bin defined by the 1000 Genome Pilot 1 JPT/CHB for Asians, and the 1000 Genome Pilot 1 YRI or HapMap3 ASW project for African Americans.

We used PLINK to perform meta-analyses and generate overall summary odds ratios (ORs), standard errors, and *P*-values [Purcell et al., 2007]. The R package, rmeta, was used to confirm results and generate meta-analysis plots [Lumley, 2009]. Meta-analyses results were based on fixed effect models to determine the evidence for association within our collected samples, so we are not making a general inference about what might be observed in other samples.

MULTIPLE TEST CORRECTION

Our primary analysis was to determine if any intersecting sub-bins across Asians and African Americans would display evidence of consistent association when comparing heavy vs. light smokers, where we defined a consistent association as having the same direction and *P*-value < 0.01 in both populations. Our binning strategy resulted in 100 single sub-bin tests and 67 conditional association tests across the four bins: a total of 167 tests. Because the probability of any particular test resulting in a *P*-value < 0.01 in both non-European populations by chance would be 0.0001 ($=0.01 \times 0.01$), results consistently associated in both populations would remain significant after Bonferroni correction ($167 \times 0.0001 < 0.05$).

RESULTS

GENETIC ASSOCIATIONS IN BIN A

Bin A (tagged by rs16969968 and rs1051730 in Europeans) includes 52 SNPs correlated ($r^2 \geq 0.7$) in the European ancestry reference sample. This bin separates into 20 sub-bins in Asian populations and 38 sub-bins in African American populations. We had adequate coverage to test nine of these 20 sub-bins in Asian data and 27 of these 38 sub-bins in African American data.

We detected a strong association between the dichotomous phenotype of heavy smoking vs. light smoking and bin A in European ancestry data (OR = 1.31, 95% CI = 1.22–1.40, $P = 1.3 \times 10^{-14}$). The only sub-bin showing consistent association with heavy smoking across the other two populations is tagged by rs16969968 (Asian population: OR = 1.64, 95% CI = 1.15–2.32, $P = 5.8 \times 10^{-3}$; African American population: OR = 1.62, 95% CI = 1.21–2.17, $P = 1.1 \times 10^{-3}$). As noted in the "Methods," because the probability of any particular test resulting in a $P < 0.01$ in both populations by chance alone would be 0.0001, this result of consistent association in both populations remained significant after Bonferroni correction.

Figure 1 shows all SNPs in bin A, and the only consistently associated sub-bins ($P < 0.01$ in both Asians and African Americans). Bin A variants span six genes in the European ancestry population, the sub-bin tagged by rs16969968 in the Asian population spans three genes, and the sub-bin tagged by rs16969968 in the African American population spans only one gene (CHRNA5). Figure 2 provides a forest plot summary of the stratified meta-analyses for the bin/sub-bin tagged by rs16969968, the only consistent association for bin A, in all three populations. Each plot lists ORs for each contributing sample. The overall cross-population meta-analysis across all datasets gave an OR of 1.33 (95% CI = 1.25–1.42, $P = 1.1 \times 10^{-17}$).

In European and Asian populations, rs16969968 and rs1051730 are highly correlated. However, due to the different LD structure in African Americans, rs16969968 and rs1051730 represent two different sub-bins ($r^2 = 0.40$ in HapMap 3 Release 2 ASW). In our analysis of African Americans, there is stronger evidence of association between the dichotomous phenotype heavy smoking vs. light smoking and the sub-bin tagged by rs16969968 (OR = 1.62, 95% CI = 1.21–2.17, $P = 1.1 \times 10^{-3}$), compared to the sub-bin tagged by rs1051730 (OR = 1.15, 95% CI = 1.03–1.28, $P = 1.1 \times 10^{-2}$). This stronger finding is seen despite the lower minor allele frequency (MAF) and much smaller available sample and for rs16969968 (MAF = 0.06, 667 cases/1,140 controls) compared to that for rs1051730 (MAF = 0.12, 1,712 cases/5,640 controls).

For bin A, no tested sub-bin other than the one tagged by rs16969968 shows consistent association across populations. The meta-analyzed genetic associations between all

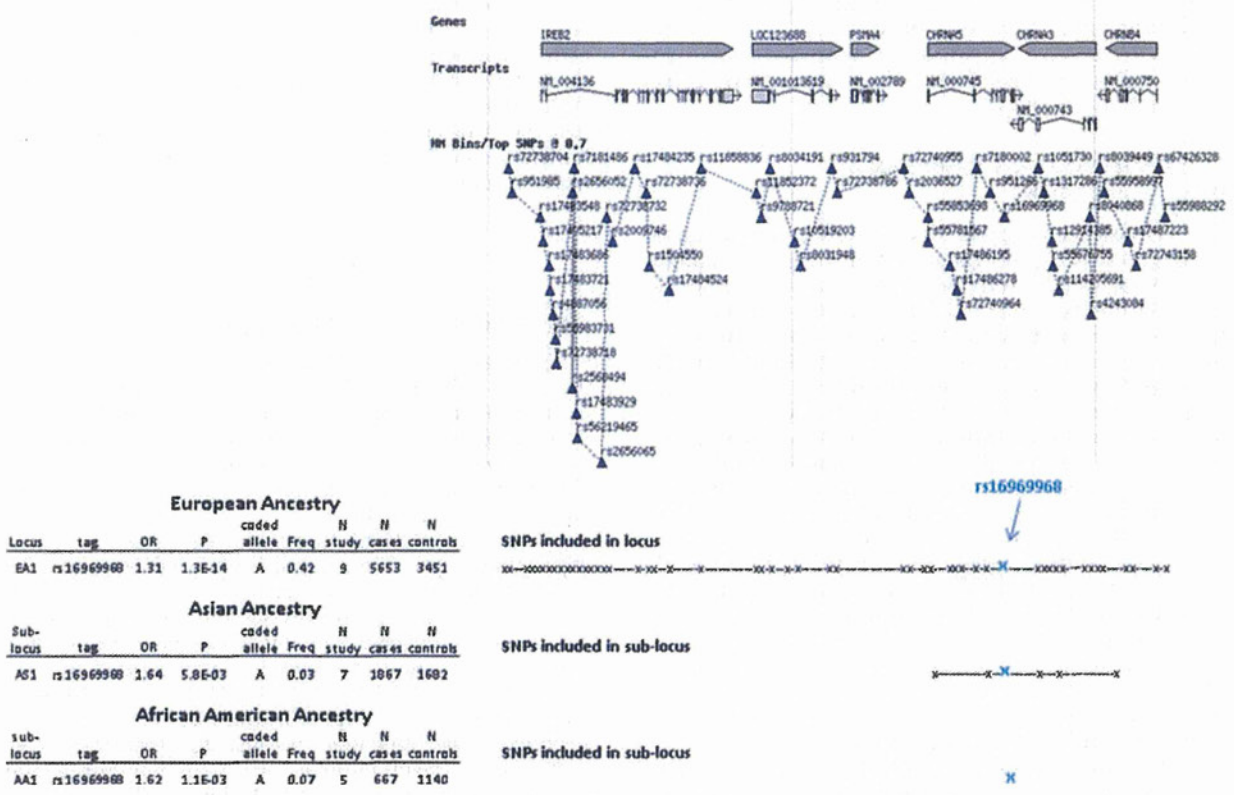

Fig. 1. Top associations with heavy vs. light smoking: Bin A across three populations. This figure shows all 52 SNPs in bin A, and also the only consistently associated sub-bins ($P < 0.01$ in both Asians and African Americans). This figure also lists corresponding odds ratios for the association between the sub-bins and heavy smoking. Bin A variants span across six genes in the European ancestry population, the sub-bin tagged by rs16969968 in the Asian population spans across two genes, and the sub-bin tagged by rs16969968 in the African American population spans across only one gene (CHRNA5).

available constituent sub-bins and heavy smoking are shown in Table S2.

GENETIC ASSOCIATIONS IN BIN B

Bin B (tagged by rs588765 and rs880395 in Europeans) includes 111 SNPs correlated ($r^2 \geq 0.7$) in the European ancestry reference sample, which was partitioned into 39 sub-bins in Asian and 37 sub-bins in African American ancestry reference samples. We had adequate coverage to test 10 of these 39 sub-bins in Asian samples and 22 of these 37 sub-bins in African American samples. Consistent with the previous report [Saccone et al., 2010] that used some of these same data (see Table S1 for the overlap, which involves only European-ancestry samples), we find that in European ancestry samples, bin B is associated (OR = 1.27, 95% CI = 1.16–1.38, $P = 8.7 \times 10^{-8}$) with heavy smoking in conditional analyses with rs16969968; bin B is not associated in single SNP analyses (OR = 1.0, 95% CI = 0.94–1.07, $P = 0.99$). In Asian samples, testing for SNP association conditioning on rs16969968 show an association between heavy smoking and bin B, with the strongest result for the sub-bin tagged by rs514743 (OR = 1.30, 95% CI = 1.07–1.58, $P = 9.7 \times 10^{-3}$), which is similar to the single SNP test (OR = 1.28, 95% CI = 1.05–1.56, $P = 0.014$). In African American subjects, there is a trend of association for the same sub-bin in conditional association (OR = 1.16, 95% CI = 0.99–1.36, $P = 0.064$; Table S3), compared to the single SNP association (OR = 1.05, 95% CI = 0.96–1.15, $P = 0.24$). Thus, we found evidence of association in the Asian samples consistent with the association observed in the samples of European ancestry, but only a trend toward association in the African American subjects. The meta-analyzed conditional and single SNP associations between these constituent sub-bins and heavy smoking are shown in Tables S3 and S6.

GENETIC ASSOCIATIONS IN BIN C

Bin C (tagged by rs6495308, rs2036534, rs7163730, rs9788682, rs684513, and rs578776 in Europeans) includes 82 SNPs correlated ($r^2 \geq 0.7$) in the European ancestry reference sample, which was partitioned into 24 sub-bins in Asian and 26 sub-bins in African American reference samples. We had adequate coverage to test 12 of these 24 sub-bins in Asian samples and 19 of these 26 sub-bins in African American samples. Consistent with the previous studies [Liu et al., 2010; Saccone et al., 2010; TAG, 2010; Thorgeirsson et al., 2010], in European ancestry samples, there is an association between heavy smoking and bin C (OR = 0.79, 95% CI = 0.72–0.86, $P = 2.5 \times 10^{-7}$) in association tests conditioning on rs16969968 as well as an association in a

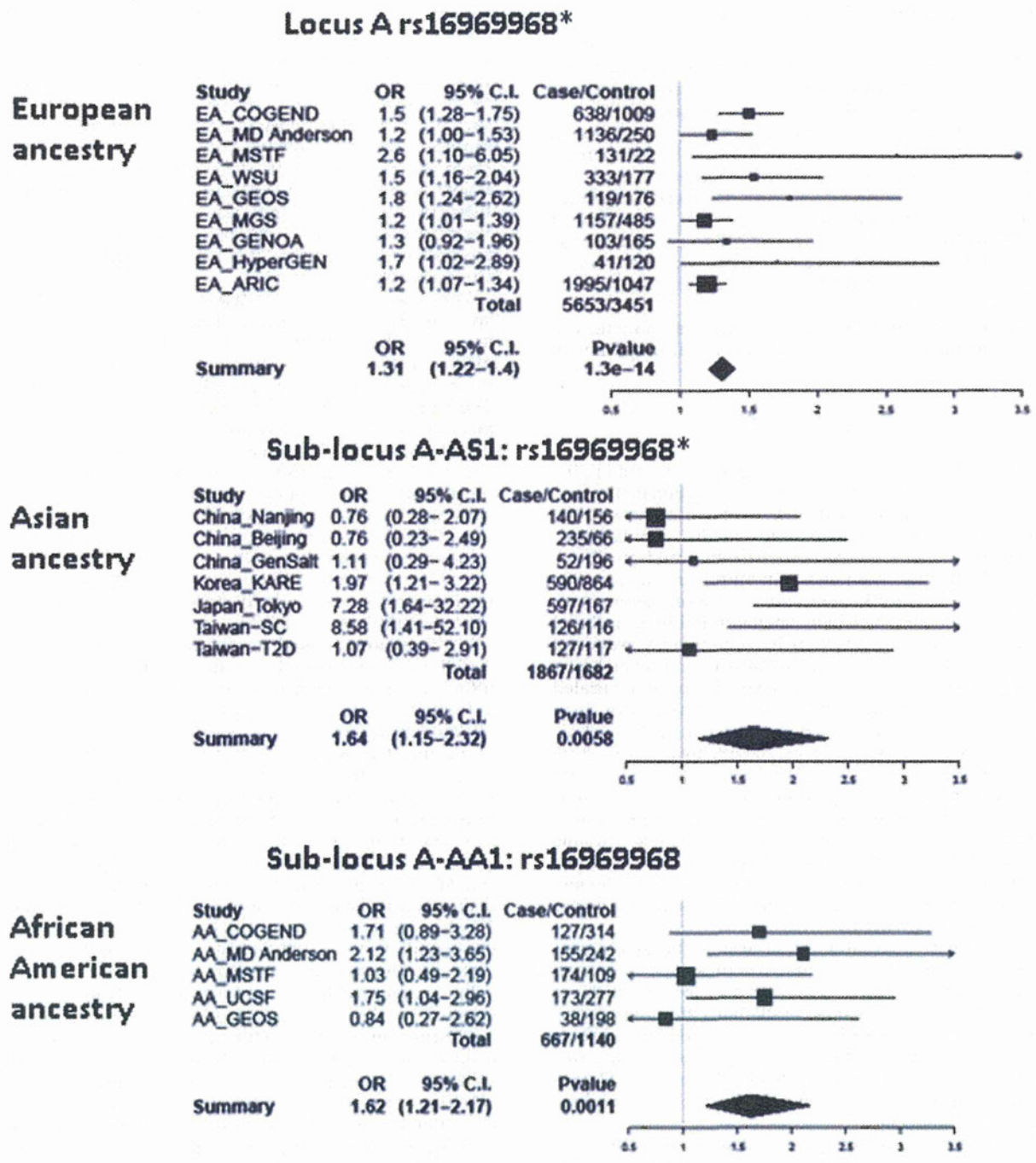

Fig. 2. Rs16969968 and heavy smoking in samples of European, Asian, and African American ancestry. The ORs and 95% CIs are for the effect per allele using additive coding in the logistic regression with age and sex as covariates. *When rs16969968 is not available, rs1051730 and rs951266 are used as proxy SNPs in European and Asian ancestry samples.

single SNP analysis (OR = 0.77, 95% CI = 0.71–0.83, $P = 4.0 \times 10^{-11}$).

Neither the Asian nor African American populations provide strong evidence of association with heavy smoking in any tested sub-bin in bin C under conditional association tests (all $P > 0.01$). In the Asian data, the strongest single SNP signal was the sub-bin tagged by rs6495308 (OR = 0.83, 95% CI = 0.72–0.96, $P = 9.8 \times 10^{-3}$). In the African American data, there was no evidence of consistent association in either single SNP or conditional analyses ($P > 0.01$) for the sub-bin tagged by rs6495308 or any other sub-bin. The meta-analyzed conditional and single SNP associations between tested sub-bins and heavy smoking are shown in Tables S4 and S7.

GENETIC ASSOCIATIONS IN BIN D

Bin D (tagged by rs2869046 in Europeans) includes 15 SNPs correlated ($r^2 \geq 0.7$) in the European ancestry reference sample, which was partitioned into seven sub-bins in Asians and seven sub-bins in African Americans. We had adequate coverage to test two of these seven sub-bins in Asian samples and three of the seven sub-bins in African American samples. We found no evidence of association between bin D and heavy smoking in European ancestry data, or across populations in single SNP or conditional association analyses ($P > 0.1$). The meta-analyzed genetic associations between available sub-bins and heavy smoking conditional and single SNP associations are shown in Tables S5 and S8.

All bins were tested in secondary analyses using the four level phenotype measured by CPD and results were similar.

DISCUSSION

This collaborative genetic meta-analysis of smoking behavior is the first to show consistent association in the chromosome 15q25 region with heavy smoking, across samples representing three genetically distinct populations—European ancestry, Asian, and African American. Previous meta-analyses examined only European ancestry data to definitively identify associations between chromosome 15q25 and smoking behavior. Smaller individual studies of Asians and African Americans have previously examined this region for association with smoking and related phenotypes. Smoking quantity has been reported as associated with variants correlated with rs16969968 in subjects of Asian and African American descent [Amos et al., 2010; Li et al., 2005; Li et al., 2010; Saccone et al., 2009; Schwartz et al., 2010; Shiraishi et al., 2009; Wu et al., 2009]. Our meta-analysis synthesizes reported findings of individual SNP associations and compares genetic associations across multipopulation samples to take the correlations between genetic variants within each population into account. Our meta-analysis strengthens the evidence of association between the specific SNP rs16969968 in bin A and heavy smoking across these diverse populations.

The strongest association signal seen in this gene cluster in European ancestry populations is represented by a group of 52 correlated variants, including rs16969968, which we call bin A. Due to these high correlations, the ability to statistically refine the association between smoking and these SNPs is very limited when using only European ancestry subjects. However, the LD structure between these 52 variants breaks down into 20 sub-bins in Asians and 38 sub-bins in African Americans.

By requiring consistent genetic effects across the three populations, we can refine a genetic association to variants that are more likely to reflect potential functional variants. Two SNPs in bin A are the most frequently reported from previous meta-analyses of smoking behavior in European ancestry subjects: rs16969968 and rs1051730. They are highly correlated ($r^2 = 1$) in European ancestry and Asian populations, but display only modest correlation in African Americans ($r^2 = 0.40$; HapMap 3 Release 2 ASW). We can leverage this difference in LD architecture to differentiate the association of heavy smoking with these two variants.

In our meta-analysis of African Americans, rs16969968 is more strongly associated with heavy smoking (OR = 1.62, 95% CI = 1.21–2.17, $P = 0.0011$, N = 1,807) than rs1051730 (OR = 1.15, 95% CI = 1.03–1.28, $P = 0.011$, N = 7,352). SNP rs16969968 is the most strongly associated polymorphism across all three populations and the only variant meeting the consistent association threshold in our study. In addition, SNP rs16969968 causes an amino acid change in the nicotinic receptor α5 subunit and alters function of its receptor [Bierut et al., 2008]. The observed consistent associations across diverse populations, combined with the results of biological experiments on rs169669968, provide converging evidence that rs16969968, rather than rs1051730, is most likely one causative variant in this region driving the strongest association signal.

Prior meta-analyses in European ancestry populations have reported additional association signals distinct from bin A, and they cluster into three groups. Bin B, a group of 111 variants highly correlated in Europeans, includes the previously reported associated SNPs rs588765 and rs880395. The association with bin B previously reported in Europeans was seen only in association analyses conditioning on rs16969968. Bin B consists of 39 sub-bins in Asian subjects and 37 sub-bins in African American subjects. In conditional analyses, we found evidence of association between bin B and heavy smoking in the Asian data (OR = 1.30, 95% CI = 1.07–1.58, $P = 9.7 \times 10^{-3}$) as well as reproducing the European ancestry finding (OR = 1.27, 95% CI = 1.16–1.38, $P = 8.7 \times 10^{-8}$). In the African American data, there was a trend toward association in the same direction (OR = 1.16, 95% CI = 0.99–1.36, $P = 0.064$).

Bin B variants, located upstream of the coding region of *CHRNA5*, are associated with variability in *CHRNA5* mRNA levels in European ancestry samples [Falvella et al., 2010; Smith et al., 2010; Wang et al., 2009]. Low levels of *CHRNA5* mRNA expression are associated with lower risk for nicotine dependence. No data exist on *CHRNA5* mRNA expression in other populations, and further work to examine expression data and smoking behavior in other populations is needed. Because the risk allele of rs16969968 occurs primarily on the low mRNA expression alleles represented by bin B, conditional SNP analysis controlling for bin A (rs16969968) is important to distinguish between these two distinct mechanisms [Saccone et al., 2010; Wang et al., 2009]. This is an important example to demonstrate how a genetic effect could be better detected and characterized when additional related variants are taken into account.

Bin C, a group of 82 variants correlated in Europeans, consisted of 24 sub-bins in subjects of Asians and 26 sub-bins in African Americans. Variants reported in previous meta-analyses of European ancestry (rs6495308 [Liu et al., 2010], rs2036534 [Thorgeirsson et al., 2010], rs7163730, rs9788682, rs684513 [TAG, 2010], rs578776 [Saccone et al., 2010]) were all examined. However, no tested SNP in bin C was consistently associated with heavy smoking with $P < 0.01$ in both Asians and African Americans. Similarly, we have no consistent associations with bin D, which contains 15 SNPs correlated in Europeans and consists of seven sub-bins in Asians and seven sub-bins in African Americans.

In undertaking this project, we faced numerous challenges. First, smoking behavior differs substantially across populations. Smoking quantity distributions differ across populations; smokers of European ancestry smoke more heavily than do Asians or African Americans. As a result, we decided to compare heavy (>20 CPD) vs. light smoking (≤10 CPD) in our primary association analysis to more closely capture the contrast between nicotine

dependent smokers and nondependent smokers. We then confirmed the consistency of results using the full distribution of smoking quantity in subsequent analyses.

Second, genotyping coverage varied between studies, and several studies in our meta-analysis had only a few variants genotyped. As a result, not all sub-bins were tested and the sample size varied across the tested sub-bins. For example, SNP rs55853698 that was imputed and reported as highly associated with smoking quantity in a previous meta-analysis of European ancestry subjects [Liu et al., 2010], lies in bin A, but no genotyping data were available for testing this SNP or the sub-bin it represents in Asian and African American populations. Use of imputed data has the potential to mitigate these problems. However, imputation was not possible for our low-coverage studies. Therefore, the concerns about having untested SNPs and unequal subjects in the region would remain even with imputed data. We believe it is important to report our findings based on directly genotyped variants, and the interpretation of the consistent associations is not expected to change with imputation.

In addition, we were not able to perform thorough admixture tests in all datasets due to variable genotyping. Population stratification unaccounted for by our stratified analyses of self-identified ancestry—European, Asian, and African American—could be a confounder in our results. Although we are leveraging the admixture to separate the effects of different genetic variants, there may be differential admixture in the cases and controls among African Americans. The impact of varied genetic architecture within given, broadly defined populations as well as within and across populations represented by individual sites (e.g., Japanese, Chinese, and Korean) needs to be elucidated in future larger scale studies with sufficient representation of individuals from different population backgrounds and more comprehensive genotyping.

Lastly, an association seen in one population that is not consistent across all three may nonetheless represent a true biological signal. Lack of consistency for the association may simply reflect differences in power for our population samples. Issues that can affect power across our three diverse populations include sample size, MAF, and even population-specific effect size. The last factor could arise in a variety of ways, including differences in LD structure, background variation, and marker information content. Thus, we suggest caution when interpreting the negative or non-consistent association results from this study. Though these results strengthen the evidence for rs16969968 as a likely causal variant, this region remains in need of further interrogation with additional genotyping and standardized imputation across all populations.

Despite the limitations of this study, this meta-analysis refines the association signals with heavy smoking across samples representing European ancestry, Asian, and African American populations. In particular, for bin A, we present evidence showing rs16969968 is a likely causal variant for heavy smoking among the common SNPs in the bin. Our evidence also suggests there are additional distinct genetic variants in the chromosome 15q25 region associated with smoking, but we are unable to clearly identify these other associations across all three populations. For example, we extend the finding of association with bin B in European ancestry samples to an association in Asians, and a trend toward association in African Americans.

This consistent pattern of cross-population association despite many unmeasured genetic and environmental differences has provided important evidence to support true causal variants. It also provides critical information by narrowing a region of interest so laboratory experiments that must follow association studies can focus on a smaller number of variants. Thus, this study represents an important step on the pathway from association to function.

ACKNOWLEDGMENTS

Collaborative Genetic Study of Nicotine Dependence (COGEND): We thank the subjects who participated in this study. We wish to thank Hilary Davidson, Sherri Fisher, Tracey Richmond, and Heidi Kromrei for administrative support; and Louis Fox for data analysis. The Collaborative Genetic Study of Nicotine Dependence (COGEND) study investigators are Laura Bierut (PI), Michael Brent, Naomi Breslau, Robert Culverhouse, Alison Goate, Richard Grucza, Dorothy Hatsukami, Anthony Hinrichs, Eric Johnson, Sharon Murphy, John Rice, Nancy Saccone, Scott Saccone, Joe Henry Steinbach, Jerry Stitzel, and Jen-Chyong Wang. *MD Anderson*: We are grateful for the invaluable contributions of clinical information and tissue samples by the participants in this study, as well as for the dedicated work of the research staff at different clinical sites. Also, we wish to thank Dr. Qing Xu for performing SNP-typing on our samples. *Mid-south Tobacco Family Study (MSTF)*: We are grateful for the invaluable contributions of clinical information and tissue samples by the participants in this study, as well as for the dedicated work of the research staff at different clinical sites. *Molecular Genetics of Schizophrenia*: We thank the study participants and the research staff at the study sites. *Genetic Epidemiology Network of America (GENOA)*: Mayo Clinic (Rochester Field Center and Genotyping Center): Stephen T. Turner, Mariza de Andrade, Julie Cunningham. University of Texas Health Sciences Center (DNA lab): Eric Boerwinkle, Megan L. Grove-Gaona. University of Michigan (Analysis Center): Patricia Peyser, Lawrence Bielak, Wei Zhao. *Hypertension Genetic Epidemiology Network (HyperGEN)*: University of Utah (Network Coordinating Center, Field Center, and Molecular Genetics Lab): Steven C. Hunt, Ph.D. (Network Director and Field Center P.I.); Mark F. Leppert, Ph.D. (Molecular Genetics P.I.); Jean-Marc Lalouel, M.D., D.Sc; Robert B. Weiss, Ph.D.; Roger R. Williams, M.D. (late); Janet Hood. University of Alabama at Birmingham (Field Center): Cora E. Lewis, M.D., M.S.P.H. (P.I.); Albert Oberman, M.D., M.P.H.; Donna Arnett, Ph.D.; Phillip Johnson; Christie Oden. Boston University (Field Center): Richard H. Myers, Ph.D. (P.I.); R. Curtis Ellison, M.D.; Yuqing Zhang, M.D.; Jemma B. Wilk, D.Sc.; Luc Djouss, M.D., D.Sc.; Jason M. Laramie; Greta Lee Splansky, M.S. University of Minnesota (Field Center and Biochemistry Lab): James S. Pankow, Ph.D. (Field Center P.I.); Michael B. Miller, Ph.D.; Michael Li, Ph.D.; John H. Eckfeldt, M.D., Ph.D.; Anthony a. Killeen, M.D., Ph.D.; Catherine Leiendecker-Foster, M.S.; Jean Bucksa; Greg Rynders. University of North Carolina (Field Center): Kari E. North, Ph.D. (P.I); Barry I. Freedman, M.D.; Gerardo Heiss, M.D. Washington University (Data Coordinating Center): D.C. Rao, Ph.D. (P.I.); Charles Gu, Ph.D.; Treva Rice, Ph.D; Aldi T. Kraja, D.Sc., Ph.D.; Gang Shi, Ph.D.; Yun Ju Sung, Ph.D.; Karen L. Schwander, M.S.;

Matthew Brown; Michael A. Province, Ph.D.; Ingrid Borecki, Ph.D. Weil Cornell Medical College (Echo Reading Center): R.B. Devereux, M.D.; Giovanni de Simone, M.D., Jonathan N. Bella, M.D. National Heart, Lung, & Blood Institute: Cashell Jaquish, Ph.D.; Dina Paltoo, Ph.D. *ARIC*: The authors thank the staff and participants of the ARIC study for their important contributions. *Japan*: We thank Dr. Kouya Shiraishi for his help on genotype data processing. *GenSalt*: The GenSalt Study Steering Committee: Dongfeng Gu, Jiang He (Chair), James E. Hixson, Cashell E. Jaquish, Depei Liu, DC Rao, Paul K. Whelton, and Zhijian Yao. *GenSalt Collaborative Research Group*: Tulane University Health Sciences Center, New Orleans, USA: Jiang He (PI), Lydia A. Bazzano, Chung-Shiuan Chen, Jing Chen, Lee Hamm, Paul Muntner, Kristi Reynolds, Jaqueline R. Reuben, Paul K. Whelton, and Wenjie Yang. Washington University School of Medicine, St. Louis, USA: DC Rao (PI), Matthew Brown, Charles Gu, Hongyan Huang, Treva Rice, Karen Schwander, Gang Shi, and Yun Ju Sung. Chinese Academy of Medical Sciences, Beijing, China: Dongfeng Gu (PI), Jie Cao, Jichun Chen, Xiufang Duan, Jianfeng Huang, Jinghan Huang, Jianxin Li, Depei Liu, Donghua Liu, Enchun Pan, Yang Wei, and Xiqui Wu. Shandong Academy of Medical Sciences, Shandong, China: Fanghong Lu (PI), Shikuan Jin, Qingjie Meng, Fan Wu, and Yingxin Zhao; Shandong Center for Diseases Control and Prevention, Shandong, China: Jixiang Ma (PI), Weika Li, and Jiyu Zhang; Zhengzhou University: Dongsheng Hu (PI), Yaxin Ding, Hongwei Wen, Meixi Zhang, and Weidong Zhang; Xinle Traditional Chinese Medicine Hospital, Hebei, China: Xu Ji (PI), Rongyan Li, Haijun Zu; Nanjing University of Medical Sciences, Jiangsu, China: Cailiang Yao (PI), Yongchao Li, Chong Shen, and Jiayi Zhou; Xi'an Jiaotong University, Shanxi, China: Jianjun Mu (PI), Enrang Chen, Qinzhou Huang, and Man Wang. Chinese National Human Genome Center at Beijing: Zhi-Jian Yao (PI), Shufeng Chen, Dongfeng Gu, Hongfan Li, Laiyuan Wang, Penghua Zhang, Qi Zhao. University of Texas Health Sciences Center at Houston: James E. Hixson (PI) and Lawrence C. Shimmin. National Heart, Lung, and Blood Institute: Cashell E. Jaquish. *Women's Health Initiative*: This manuscript was prepared in collaboration with investigators of the WHI and has been reviewed and approved (MS1453) by the Women's Health Initiative (WHI) Publications & Presentations Committee. The authors thank Charles Kooperberg and the WHI investigators and staff and study participants for making the program possible. WHI investigators are listed at http://www.whiscience.org/publications/WHI_investigators_shortlist.pdf. *Collaborative Genetic Study of Nicotine Dependence (COGEND)*: The COGEND contribution was supported by the National Cancer Institute (NCI; P01 CA089392), The National Human Genome Research Institute (NHGRI; U01 HG04422-01), and the National Institute on Drug Abuse (NIDA;K02 DA021237). COGEND genotyping was in part performed under NIDA Contract HHSN271200477471C; phenotypic and genotypic data are stored in the NIDA Center for Genetic Studies (NCGS) at http://zork.wustl.edu/ under NIDA Contract HHSN271200477451C (PIs J Tischfield and J Rice); genotyping services were also provided by the Center for Inherited Disease Research (CIDR), which is fully funded through a federal contract from the National Institutes of Health (NIH) to The Johns Hopkins University, contract number HHSN268200782096. Dr. LiShiun Chen was supported by KL2RR024994 and K08DA030398. Support was also provided by R01DA026911, R03DA023166, and R21DA033827. *MD Anderson*: Research was supported by NIH grants U19CA148127, R01CA121197S2, R01CA141716, R01CA127219, CA121197, R01CA133996, P30CA16672, P50CA70907, R01CA55769 and Cancer Prevention Research Institute of Texas grant RP10043. *Midsouth Tobacco Family Study (MSTF)*: This project is supported in part by NIH Grant R01 DA012844 (PI: Ming Li). *Wayne State University and the Karmanos Cancer Institute: Family Health Study; Women's Epidemiology of Lung Disease; EXHALE (Exploring Health Ancestry and Lung Epidemiology)*: This research was supported through NIH grants R01CA060691 and R01CA87895, and NIH contract PC35145. *University of Maryland: GEOS*: The GEOS Study was supported by the National Institutes of Health Genes, Environment and Health Initiative (GEI) Grant U01 HG004436, as part of the GENEVA consortium under GEI, with additional support provided by the Mid-Atlantic Nutrition and Obesity Research Center (P30 DK072488); and the Office of Research and Development, Medical Research Service, and the Baltimore Geriatrics Research, Education, and Clinical Center of the Department of Veterans Affairs. Genotyping services were provided by the Johns Hopkins University Center for Inherited Disease Research (CIDR), which is fully funded through a federal contract from the National Institutes of Health to the Johns Hopkins University (contract number HHSN268200782096C). Assistance with data cleaning was provided by the GENEVA Coordinating Center (U01 HG 004446; PI Bruce S Weir). Study recruitment and collection of datasets were supported by a Cooperative Agreement with the Division of Adult and Community Health, Centers for Disease Control and by grants from the National Institute of Neurological Disorders and Stroke (NINDS) and the NIH Office of Research on Women's Health (R01 NS45012, U01 NS069208-01). *Molecular Genetics of Schizophrenia*: This study was supported by NIH R01 grants (MH67257 to Nancy G. Buccola, MH59588 to Bryan J. Mowry, MH59571 to Pablo V. Gejman, MH59565 to Robert Freedman, MH59587 to Farooq Amin, MH60870 to William F. Byerley, MH59566 to Donald W. Black, MH59586 to Jeremy M. Silverman, MH61675 to Douglas F. Levinson, MH60879 to C. Robert Cloninger, and MH81800 to Pablo V. Gejman), NIH U01 grants (MH46276 to C. Robert Cloninger, MH46289 to Charles Kaufmann, MH46318 to Ming T. Tsuang, MH79469 to Pablo V. Gejman, and MH79470 to Douglas F. Levinson), the Genetic Association Information Network (GAIN), and by The Paul Michael Donovan Charitable Foundation. Genotyping was carried out by the Center for Genotyping and Analysis at the Broad Institute of Harvard and MIT (Stacy Gabriel and Daniel B. Mirel), which is supported by grant U54 RR020278 from the National Center for Research Resources. Genotyping of half of the EA sample and almost all the AA sample was carried out with support from GAIN. The GAIN quality control team (Gonçalo R. Abecasis and Justin Paschall) made important contributions to the project. We thank Shaun Purcell for assistance with PLINK. *Genetic Epidemiology Network of America (GENOA)*: The Genetic Epidemiology Network of Arteriopathy phenotyping and genome-wide genotyping is supported by the National Heart Lung and Blood Institute (NHLBI) of the National Institutes of Health (HL54457, HL68737, and HL087660). *Hypertension Genetic Epidemiology Network (HyperGEN)*: The Hypertension Genetic Epidemiology Network is funded by cooperative agreements (U10) with NHLBI: HL54471, HL54472, HL54473, HL54495, HL54496, HL54497, HL54509, HL54515. *ARIC*: The Atherosclerosis

Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute (NHLBI) contracts N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022, R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. *Nanjing and Beijing*: This work was supported by the Chinese National Natural Science Foundation grant 30230080 (Hongbing Shen) and the State Key Basic Research Program grants 2002CB512902 (Hongbing Shen). Dr. Dongxin Lin was supported by State Key Basic Research Program grant 2004CB518701. *KARE (Korea Association Resource)*: The KARE data analyzed in this study were obtained from the Korean Genome Analysis Project (4845-301) which was funded by a grant from the Korea National Institute of Health (Korea Center for Disease Control, Ministry for Health, Welfare and Family Affairs), Republic of Korea. The work of TP was supported by the Consortium for Large Scale Genome Wide Association Study, the National Research Foundation (KRF-2008-313-C00086) and the Brain Korea 21 Project of the Ministry of Education. *Japan*: Grants-in-Aid from the Ministry of Health, Labor and Welfare for the 3rd-term Comprehensive 10-year Strategy for Cancer Control and for Cancer Research (19-9 and 19S-1). *Taiwan*: This study was supported by Academia Sinica Genomic Medicine Multicenter Study and National Research Program for Genomic Medicine, National Science Council, Taiwan (National Clinical Core, NSC97-3112-B-001-014 and National Genotyping Center, NSC97-3112-B-001-015). *GenSalt*: The Genetic Epidemiology Network of Salt Sensitivity is supported by a cooperative agreement project Grant (U01HL072507) from the National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland. *University of California San Francisco*: This work was supported by the National Institute of Environmental Health Sciences [R01 ES06717]; and the National Cancer Institute [R01 CA 52689 to MW]. *Women's Health Initiative*: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119, 32122, 42107-26, 42129-32, and 44221. SPD is supported by DA 017441 and 02733. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. *CONFLICT OF INTEREST DISCLOSURE*: LB Bierut, AM Goate, JP Rice, and JC Wang are listed as inventors on Issued U.S. Patent. 8,080,371, "Markers for Addiction" covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction. Ming Li serves as a scientific advisor to ADial Pharmaceuticals.

REFERENCES

Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Bonnen PE, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Gonzaga-Jauregui C, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VA, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58.

Amos CI, Gorlov IP, Dong Q, Wu X, Zhang H, Lu EY, Scheet P, Greisinger AJ, Mills GB, Spitz MR. 2010. Nicotinic acetylcholine receptor region on chromosome 15q25 and lung cancer risk among African Americans: a case-control study. J Natl Cancer Inst 102(15):1199–1205.

Baker TB, Weiss RB, Bolt D, von Niederhausern A, Fiore MC, Dunn DM, Piper ME, Matsunami N, Smith SS, Coon H, McMahon WM, Scholand MB, Singh N, Hoidal JR, Kim SY, Leppert MF, Cannon DS. 2009. Human neuronal acetylcholine receptor A5-A3-B4 haplotypes are associated with multiple nicotine dependence phenotypes. Nicotine Tob Res 11(7):785–796.

Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H, Waterworth D, Muglia P, Mooser V. 2008. Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. Mol Psychiatry 13(4):368–373.

Bierut LJ, Stitzel JA, Wang JC, Hinrichs AL, Grucza RA, Xuei X, Saccone NL, Saccone SF, Bertelsen S, Fox L, Horton WJ, Morgan SD, Breslau N, Budde J, Cloninger CR, Dick DM, Foroud T, Hatsukami D, Hesselbrock V, Johnson EO, Kramer J, Kuperman S, Madden PAF, Mayo K, Nurnberger JJ, Pomerleau O , Porjesz B, Reyes O, Schuckit M, Swan GT, J. A, Edenberg HJ, Rice JP, Goate AM. 2008. Nicotine dependence and the a5-a3-b4 nicotinic receptor gene cluster: variants in the nicotinic receptors alter the risk for nicotine dependence. Am J Psychiatry 9(165):1163–1171.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74(1):106–120.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073.

Falvella FS, Galvan A, Colombo F, Frullanti E, Pastorino U, Dragani TA. 2010. Promoter polymorphisms and transcript levels of nicotinic receptor CHRNA5. J Natl Cancer Inst 102(17):1366–1370.

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24(24):2938–2939.

Keskitalo K, Broms U, Heliovaara M, Ripatti S, Surakka I, Perola M, Pitkaniemi J, Peltonen L, Aromaa A, Kaprio J. 2009. Association of serum cotinine level with a cluster of three nicotinic acetylcholine receptor genes (CHRNA3/CHRNA5/CHRNB4) on chromosome 15. Hum Mol Genet 18(20):4007–4012.

Li MD, Beuten J, Ma JZ, Payne TJ, Lou XY, Garcia V, Duenes AS, Crews KM, Elston RC. 2005. Ethnic- and gender-specific association of the nicotinic acetylcholine receptor alpha4 subunit gene (CHRNA4) with nicotine dependence. Hum Mol Genet 14(9):1211–1219.

Li MD, Yoon D, Lee JY, Han BG, Niu T, Payne TJ, Ma JZ, Park T. 2010. Associations of variants in CHRNA5/A3/B4 gene cluster with smoking behaviors in a Korean population. PLoS One 5(8): e12183.

Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Berrettini W, Knouff CW, Yuan X, Waeber G, Vollenweider P, Preisig M, Wareham NJ, Zhao JH, Laos RJ, Barroso I, Khaw KT, Grundy S, Barter P, Mahley R, Kesaniemi A, McPherson R, Vincent JB, Strauss