Citation: Transl Psychiatry (2012) 2, e119, doi:10.1038/tp.2012.41

www.nature.com/tp

Genome-wide meta-analyses of smoking behaviors in African Americans

SP David[1,2,3], A Hamidovic[4,51], GK Chen[5,51], AW Bergen[1], J Wessel[1,6,7], JL Kasberger[8], WM Brown[9], S Petruzella[10], EL Thacker[11], Y Kim[12], MA Nalls[13], GJ Tranah[14], YJ Sung[15], CB Ambrosone[16], D Arnett[17], EV Bandera[18], DM Becker[19], L Becker[19], SI Berndt[20], L Bernstein[21], WJ Blot[22,23], U Broeckel[24], SG Buxbaum[25], N Caporaso[20], G Casey[5], SJ Chanock[20], SL Deming[23], WR Diver[26], CB Eaton[3], DS Evans[14], MK Evans[27], M Fornage[28], N Franceschini[29], TB Harris[30], BE Henderson[5], DG Hernandez[13], B Hitsman[4], JJ Hu[31], SC Hunt[32], SA Ingles[5], EM John[33,34], R Kittles[35], S Kolb[36], LN Kolonel[37], L Le Marchand[37], Y Liu[38], KK Lohman[9], B McKnight[39], RC Millikan[40], A Murphy[41], C Neslund-Dudas[42], S Nyante[40], M Press[5], BM Psaty[43,44], DC Rao[15], S Redline[45], JL Rodriguez-Gil[31], BA Rybicki[42], LB Signorello[22,23], AB Singleton[13], J Smoller[46], B Snively[9], B Spring[4], JL Stanford[36], SS Strom[47], GE Swan[1], KD Taylor[48], MJ Thun[26], AF Wilson[12], JS Witte[49], Y Yamamura[47], LR Yanek[19], K Yu[20], W Zheng[23], RG Ziegler[20], AB Zonderman[50], E Jorgenson[8,52], CA Haiman[5,52] and H Furberg[10,52]

The identification and exploration of genetic loci that influence smoking behaviors have been conducted primarily in populations of the European ancestry. Here we report results of the first genome-wide association study meta-analysis of smoking behavior in African Americans in the Study of Tobacco in Minority Populations Genetics Consortium ($n = 32\,389$). We identified one non-coding single-nucleotide polymorphism (SNP; rs2036527[A]) on chromosome 15q25.1 associated with smoking quantity

[1]Center for Health Sciences, Policy Division, SRI International, Menlo Park, CA, USA; [2]Center for Education and Research in Family and Community Medicine, Division of General Medical Disciplines, Stanford University School of Medicine, Stanford, CA, USA; [3]Department of Family Medicine, Center for Primary Care and Prevention, Brown Alpert Medical School, Pawtucket, RI, USA; [4]Department of Preventative Medicine, Northwestern University, Chicago, IL, USA; [5]Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, USA; [6]Department of Public Health, Division of Epidemiology and Environmental Health, Indiana University School of Medicine, Indianapolis, IN, USA; [7]Department of Medicine, Division of Cardiology, Indiana University School of Medicine, Indianapolis, IN, USA; [8]Department of Neurology, Ernest Gallo Clinic and Research Center, University of California, San Francisco, CA, USA; [9]Department of Biostatistical Sciences, Wake Forest School of Medicine, Winston-Salem, NC, USA; [10]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA; [11]Department of Epidemiology, University of Washington, Seattle, WA, USA; [12]Genometrics Section, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD, USA; [13]Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA; [14]California Pacific Medical Center Research Institute, San Francisco, CA, USA; [15]Division of Biostatistics, Washington University School of Medicine, St Louis, MO, USA; [16]Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA; [17]Department of Epidemiology, University of Alabama, Birmingham, AL, USA; [18]The Cancer Institute of New Jersey, New Brunswick, NJ, USA; [19]Department of Medicine, The Johns Hopkins GeneSTAR Research Program, The Johns Hopkins University School of Medicine, Baltimore, MD, USA; [20]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA; [21]Department of Population Science, Division of Cancer Etiology, Beckman Research Institute, City of Hope, Duarte, CA, USA; [22]International Epidemiology Institute, Rockville, MD, USA; [23]Department of Medicine, Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt University and the Vanderbilt-Ingram Cancer Center, Nashville, TN, USA; [24]Department of Medicine, Medical College of Wisconsin, Milwaukee, WI, USA; [25]Jackson Heart Study, Jackson State University, Jackson, MS, USA; [26]Epidemiology Research Program, American Cancer Society, Atlanta, GA, USA; [27]Health Disparities Research Section, Clinical Research Branch, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA; [28]Division of Epidemiology, Brown Foundation Institute of Molecular Medicine, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA; [29]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [30]Laboratory of Epidemiology, Demography and Biometry, National Institute on Aging, Bethesda, MD, USA; [31]Department of Epidemiology and Public Health, Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, USA; [32]Department of Internal Medicine, University of Utah, Salt Lake City, UT, USA; [33]Cancer Prevention Institute of California, Fremont, CA, USA; [34]Stanford University School of Medicine, Stanford Cancer Institute, Stanford, CA, USA; [35]Department of Medicine, Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, IL, USA; [36]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA; [37]Epidemiology Program, Cancer Research Center, University of Hawaii, Honolulu, HI, USA; [38]Sticht Center on Aging, Wake Forest University School of Medicine, Winston-Salem, NC, USA; [39]Department of Biostatistics, University of Washington, Seattle, WA, USA; [40]Department of Epidemiology, Gillings School of Global Public Health, and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA; [41]Department of Urology, Northwestern University, Chicago, IL, USA; [42]Department of Public Health Sciences, Henry Ford Hospital, Detroit, MI, USA; [43]Departments of Epidemiology, Medicine and Health Services, University of Washington, Seattle, WA, USA; [44]Group Health Research Institute, Group Health Cooperative, Seattle, WA, USA; [45]Department of Medicine and Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA; [46]Department of Psychiatry, Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; [47]Department of Epidemiology, The University of Texas MD, Anderson Cancer Center, Houston, TX, USA; [48]Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA; [49] Departments of Epidemiology and Biostatistics, and Urology, Institute for Human Genetics, University of California, San Francisco, CA, USA and [50]Laboratory of Personality and Cognition, National Institute on Aging, National Institutes of Health, Baltimore, MD

Correspondence: Dr SP David, Center for Education and Research in Family and Community Medicine, Division of General Medical Disciplines, Stanford University School of Medicine, 1215 Welch Road, Modular G, Stanford, CA 93405-5408, USA or CA Haiman, Department of Preventive Medicine, University of Southern California Keck School of Medicine, Harlyne Norris Research Tower, 1450 Biggy Street, Room 1504A, Los Angeles, CA 90033, USA or E Jorgenson, Enest Gallo Clinic and Research Center, University of California, San Francisco, 5858 Horton Street, Suite 200, Emeryville, San Francisco, CA 94608, USA.

E-mail: spdavid@stanford.edu or haiman@usc.edu or ejorgenson@gallo.ucsf.edu

[51]Joint first authors.

[52]Joint senior authors.

Keywords: African American; genome-wide association; health disparities; nicotine; smoking; tobacco

Received 13 March 2012; accepted 10 April 2012

(cigarettes per day), which exceeded genome-wide significance ($\beta = 0.040$, s.e. $= 0.007$, $P = 1.84 \times 10^{-8}$). This variant is present in the 5′-distal enhancer region of the *CHRNA5* gene and defines the primary index signal reported in studies of the European ancestry. No other SNP reached genome-wide significance for smoking initiation (SI, ever vs never smoking), age of SI, or smoking cessation (SC, former vs current smoking). Informative associations that approached genome-wide significance included three modestly correlated variants, at 15q25.1 within *PSMA4*, *CHRNA5* and *CHRNA3* for smoking quantity, which are associated with a second signal previously reported in studies in European ancestry populations, and a signal represented by three SNPs in the *SPOCK2* gene on chr10q22.1. The association at 15q25.1 confirms this region as an important susceptibility locus for smoking quantity in men and women of African ancestry. Larger studies will be needed to validate the suggestive loci that did not reach genome-wide significance and further elucidate the contribution of genetic variation to disparities in cigarette consumption, SC and smoking-attributable disease between African Americans and European Americans.
Translational Psychiatry (2012) **2,** e119; doi:10.1038/tp.2012.41; published online 22 May 2012

Introduction

Smoking is influenced by genetic and environmental factors.[1,2] Genome-wide association studies (GWAS) in populations of European ancestry have identified genetic variation associated with smoking behaviors, including smoking initiation (SI), smoking quantity and smoking cessation (SC). An initial, large ($n = 10\,995$) GWAS of smoking quantity identified associations with genetic variants in the nicotinic acetylcholine receptor α5, α3 and β4 subunit cluster on chromosome 15q25.1.[3] Genome-wide meta-analyses in three large consortia ($n = 74\,053$, 31 226 and 41 150) of smoking behaviors confirmed the finding at 15q25.1 and refined the association signal within the locus.[4–6] Additional studies in diverse populations also have revealed independent signals in this region, suggesting multiple biologically functional variants.[7,8] This locus has also been reported as a susceptibility locus for lung cancer; however, whether this effect is independent of smoking behavior is unclear.[9,10] Additional regions have been identified for smoking quantity (*CHRNB3*/*CHRNA6*) on 8p11,[4] *CYP2A6* on 19q13[4,6] and *LOC100188947* on 10q25[6]), SI (*BDNF* on 11p13)[6] and SC (*DBH* on 9q34).[6]

To date, all published GWAS for smoking behaviors have been conducted in populations of European descent.[11] Conducting GWAS in non-European populations, such as African ancestry populations is important because of their greater genetic diversity and population differences in disease allele frequency, linkage disequilibrium patterns and phenotype prevalence.[12] For smoking behaviors, the need for GWAS in African American populations is particularly clear; African Americans, on average, initiate smoking later, smoke fewer cigarettes per day, yet are less likely to successfully quit smoking. Further, they have a higher risk of smoking-related lung cancer than many other populations.[13] Ethnic differences in the clearance of nicotine, cotinine and other metabolites have been shown to contribute to the observed differences in cigarette consumption across populations, mediated in part by genetic variants in the cytochrome *p450 2A6* gene.[14–16]

The genetic architecture of smoking-related traits is not well described in non-European ancestral groups, but there is evidence that genetic determinants have important implications for multiple addictive behaviors in populations globally.[17] We established the Study of Tobacco in Minority Populations (STOMP) Genetics Consortium, which represents 13 GWAS studies of men and women of African ancestry, to search for risk loci for smoking behaviors in this population.

Materials and methods

Study description. The STOMP Genetics Consortium is comprised of the following studies: the Women's Health Initiative SNP Health Association Resource ($n = 8208$), the African American GWAS consortia of Breast Cancer ($n = 5061$) and Prostate Cancer ($n = 5556$), the Candidate Gene Association Resource Consortium (including the Atherosclerosis Risk in Communities ($n = 2916$) study, the Cleveland Family Study ($n = 632$), the Coronary Artery Risk Development in Young Adults ($n = 953$) study, the Jackson Heart Study ($n = 2145$) and the Multi-Ethnic Study of Atherosclerosis ($n = 1646$)), the Cardiovascular Health Study ($n = 801$), the Healthy Aging in Neighborhoods across the Life Span Study ($n = 918$), the Health ABC Study ($n = 1137$), the Genetic Study of Atherosclerosis Risk ($n = 1175$) and the Hypertension Genetic Epidemiology Network ($n = 1241$). A description of each participating study as well as details regarding the measurement and collection of smoking data for each study are provided in Supplementary Materials. All studies had local Institutional Review Board approval for the present study and all participants provided written informed consent.

Smoking phenotypes. We examined four smoking phenotypes previously shown to be heritable in the African and European ancestry samples[18–21] and used in prior GWAS of smoking behavior.[4–6] SI contrasted individuals who reported having smoked 100 cigarettes during their lifetime (ever smokers) with those who reported having smoked between 0 and 99 cigarettes during their lifetime (never smokers), consistent with the Centers for Disease Control classification.[22] Among smokers, the age of SI (AOI) represented the age individuals began smoking. Some studies captured the age they first tried smoking, whereas others collected the age they began smoking regularly. As prior research suggests similar heritabilities and high genetic correlation between these phenotypes, we justified using either value in a general assessment of AOI. Similarly, for cigarettes smoked per day (CPD), some studies collected maximum CPD, whereas others collected average CPD. Longitudinal twin data suggests a high correlation between these variables over time, which supported using either value in our analyses. For studies that collected CPD as ranges, the mid-point of the interval was used as the data point; for example, individuals who reported the CPD category 0–4

were assigned a CPD value of 2. SC contrasted individuals who had quit smoking at interview (former smokers) with those who were current smokers. As relapse to smoking is highest within the first year after quitting,[23] we tried to reduce misclassification by excluding smokers who quit within 1 year of interview within studies with available data. Table 1 presents distributions of smoking phenotypes across participating studies.

Genotyping and quality control. Each study performed its own genotyping using Illumina (San Diego, CA, USA) or Affymetrix GWAS arrays (Santa Clara, CA, USA). Supplementary Tables 1 and 2 present the details of the arrays, genotyping quality control procedures and sample exclusions (i.e., sex mismatch, call rate failure, relatedness, missing smoking and ancestry outliers) for each study. The quality control filters applied by each study were comparable; single-nucleotide polymorphisms (SNPs) with call rates $<95\%$ (except the Genetic Study of Atherosclerosis Risk, $<90\%$), $<1\%$ minor allele frequency or significant ($P<10^{-6}$) departure from Hardy–Weinberg equilibrium were excluded, as were individuals with excess autosomal heterozygosity, mismatch between reported and genetically determined sex, or first- or second-degree relatedness. Genome-wide imputation[24] was carried out in each study using the software MACH, IMPUTE, BEAGLE or BIMBAM v0.99,[25–32] to infer genotypes for SNPs that were not genotyped directly on the platforms, but were genotyped on the HapMap phase 2 CEU and YRI samples.[33] SNPs with imputation quality scores <0.5 were excluded.

Data analyses. *Study-specific GWAS analysis*. Each study conducted uniform cross-sectional analyses for each smoking phenotype using an additive genetic model. Logistic regression was used for discrete traits (SI and SC) and linear regression was used for quantitative traits (CPD and AOI). Continuous, quantitative traits were normalized by transformation to Z scores, owing to heavy tails and non-normality. Outliers were removed within each study, where abs $(Z)>2$. Link $(Y)=Z$ scores were fit using ordinary least squares regression. To investigate potential sources of heterogeneity across studies, we examined the distribution of African ancestry in each cohort (Supplementary Figure 1). To account for population stratification and admixture, all studies adjusted for an appropriate number of eigenvectors[3–10] from a study-specific principal components analysis.[34] In addition, study-specific analyses included adjustment for age and case status or study site, when appropriate. Genomic control inflation factors were computed using standard methods.[35,36]

Meta-analyses of GWAS results. We performed fixed-effect meta-analysis for each smoking phenotype by computing pooled inverse-variance-weighted β-coefficients, s.e. and Z scores for each SNP.[37] All GWAS results were corrected via genomic control before the meta-analysis. The study-specific lambda values utilized in this step ranged from 1.01 to 1.08 for SI (Supplementary Table 1). Heterogeneity across studies was investigated using the I^2 statistic.[38] The results presented herein are corrected by a second GC correction based on λ of the meta-analyses ($\lambda<1.02$). A significance threshold of $P<5\times10^{-8}$ was considered to indicate genome-wide significance. Linkage disequilibrium statistics for the largest of the STOMP cohorts (Women's Health Initiative, $n=8208$) were calculated using DPRIME (http://www.phs.wfubmc.edu/public/bios/gene/downloads.cfm). Linkage disequilibrium statistics for CEU and YRI were obtained from HapMap phase 2 33. Statistical power analysis was performed using QUANTO.[39]

Results

The meta-analysis included 32 389 genotyped men and women of African ancestry from 13 studies with sample sizes ranging from $n=632$ to $n=8208$ (Table 1). Our meta-analysis

Table 1 Descriptive characteristics of the 13 studies participating in the STOMP Consortium

Study	*N (% female)*	*Age, mean (s.d.)*[a]	*Ever smokers (%)*	*CPD, mean (s.d.)*[b]	*AOI*[a]*, mean (s.d.)*[b]	*Former smokers (%)*[b]
AABC	5061 (100)	56.6 (12.6)	47.2	11.9 (8.4)	23.3 (9.0)	58.8
AAPC	5556 (0)	63.7 (9.6)	68.7	14.6 (9.9)	23.2 (9.0)	64.9
CHS	801 (63.2)	72.9 (5.6)	51.2	13.9 (11.2)	19.0 (5.2)	66.8
CARe						
ARIC	2916 (61.2)	54.1 (5.7)	52.2	14.4 (9.8)	19.5 (6.4)	28.1
CARDIA	953 (61.4)	24.4 (3.8)	39.2	11.8 (8.7)	17.3 (5.1)	4.6
CFS	632 (59.0)	35.5 (19.8)	45.1	13.1 (10.3)	19.0 (5.5)	13.3
JHS	2145 (60.7)	55.2 (12.8)	33.2	14.9 (10.8)	19.3 (5.7)	17.0
MESA	1646 (54.7)	62.2 (10.1)	53.5	14.6 (18.2)	18.3 (5.4)	35.0
GeneSTAR	1175 (61.7)	47.4 (12.3)	57.2	11.5 (10.3)	18.3 (5.4)	44.0
HANDLS	918 (54.5)	48.6 (9.0)	65.4	15.7 (32.8)	17.4 (6.2)	29.0
Health ABC	1137 (57.2)	73.4 (2.9)	56.4	15.7 (12.6)	19.5 (7.0)	69.5
HyperGEN	1241 (67.3)	45.2 (13.3)	48.7	12.1 (9.8)	19.5 (5.5)	58.0
WHI (SHARe)	8208 (100)	61.6 (7.0)	50.6	11.5 (9.5)	20.5 (5.9)	39.1

Abbreviations: STOMP, Study of Tobacco in Minority Populations; CPD, cigarettes smoked per day; AOI, age of smoking initiation; AABC, African American GWAS consortia of Breast cancer; AAPC, African American GWAS consortia of Prostate Cancer; CHS, Cardiovascular Health Study; CARe, Candidate Gene Association Resource; ARIC, Atherosclerosis Risk in Communities; CARDIA, Coronary Artery Risk Development in Young Adults; CFS, Cleveland Family Study; JHS, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis; GeneSTAR, Genetic Study of Atherosclerosis Risk; HANDLS, Healthy Aging in Neighborhoods across the Life Span Study; HyperGEN, Hypertension Genetic Epidemiology Network; WHI, Women's Health Initiative; SHARe, SNP Health Association Resource. Descriptive statistics for smoking behaviors included ever smokers only.
[a]Age in years. [b]Calculated among ever smokers.

sample was 66.1% female, the mean age when smoking information was collected ranged from 35.5 to 73.4 years, and 52.7% were ever smokers. Among smokers, mean CPD ranged from 11.5 to 15.7, the mean AOI ranged from 17.3 to 23.3 years, and 44.8% were former smokers.

Sample sizes for the four smoking phenotype analyses (i.e., with complete genotype and phenotype data) were $n=32\,389$ for SI, $n=16\,877$ for AOI, $n=15\,547$ for CPD and $n=16\,215$ for SC. Manhattan plots for the four smoking phenotypes after double-GC scaling are shown in Figure 1. In the entire analysis, only one SNP, rs2036527, achieved genome-wide significance for one trait, CPD ($\beta=0.04$, s.e. $=0.007$, $P=1.84\times10^{-8}$, $I^2=41.6\%$, Table 2; study-specific results are show in Supplementary Table 3). This variant is located 6246 bp 5′ of the *CHRNA5* gene on chromosome 15q25.1. We observed multiple SNPs with *P*-values of 10^{-7} associated with CPD: rs3101457, located in intron 2 (IVS2) of *C1orf100* on 1q44, and rs547843, located 63 kb 5′ of a non-coding RNA sequence (*LOC503519*) on 15q12. Three highly correlated SNPs ($r^2>0.95$, YRI) in the *SPOCK2* gene on 10q22.1 exhibited a *P*-value of 10^{-7} with AOI (Table 2). The most significant associations for SI and SC were observed at rs566973 ($\sim$20 kb 3′ of *CRCT1* on 1q21.3) and rs3813637 (in the 3′-untranslated region of *C1orf49* on 1q25.2), respectively (data not shown).

Four top SNPs associated with CPD span approximately 100 kb (76.6–76.7 Mb) at 15q25.1; from rs3813570, located in the 5′-untranslated region (c.-72T>C) of *PSMA4*, to rs938682, located in IVS4 (c.378-1941C>T) of *CHRNA3* (Table 2 and Figure 2). The most significant SNP, rs2036527, is located between *PSMA4* and *CHRNA5*, and is correlated with the index signals (rs1051730, rs16969968) for CPD reported in previous European ancestry studies. In CEU, the r^2 is 0.84 between rs2036527 and rs1051730, and 0.93 between rs2036527 and rs16969968. The r^2 between rs2036527 and 1051730 is 0.44 in YRI, and 0.502 in STOMP, whereas rs16969968 is non-polymorphic. Rs2036527 is also correlated with SNPs in the European Americans that tag a haplotype associated with increased expression of *CHRNA5* in prefrontal cortex brain samples from European Americans and African Americans,[40] but is not correlated with this haplotype in African ancestry samples (r^2 between rs2036527 and rs1979905 $=0.443$ in CEU, 0.045 in YRI and 0.064 in STOMP). The additional signals at 15q25.1 with near genome-wide significance in our study are represented by rs667282, rs938682 and rs3813570, which are weakly correlated with rs2036527 ($r^2$0.2 in CEU, 0.12 in YRI and 0.084 in STOMP). These three SNPs are correlated with each other ($r^2$0.60 in CEU and 0.32 in YRI) as well as with rs578776 and other SNPs at 15q25.1 that define a signal for smoking intensity in the European ancestry populations that is independent of rs2036527.[8] However, when conditioning on rs2036527 in the four largest study populations in our sample (the African American GWAS consortia of Prostate Cancer, African American GWAS consortia of Breast Cancer, Candidate Gene Association Resource and Women's Health Initiative; $n=13\,113$), the association between these three SNPs and CPD diminished (*P*-values of 10^{-3} after conditioning on rs2036527; Supplementary Figure 2). Assuming the GWAS arrays utilized in this study provide adequate coverage of common alleles at 15q25.1, this suggests there are not

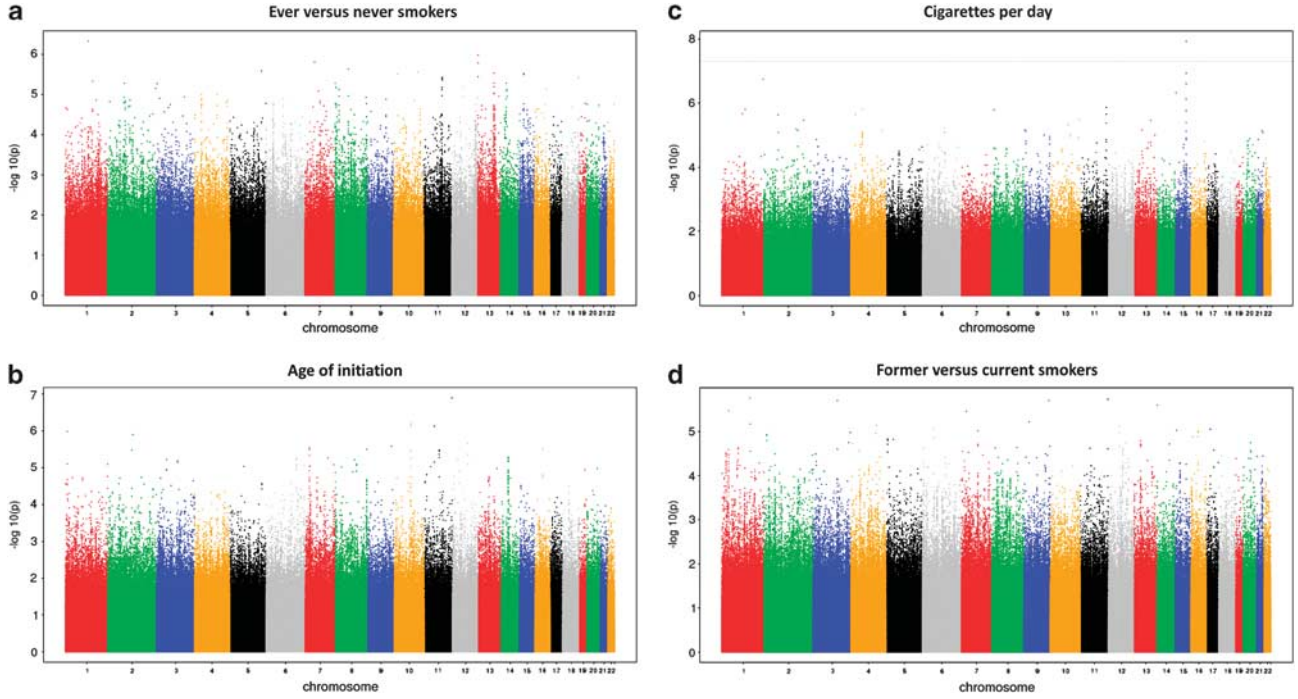


Figure 1 Double genomic control (GC)-corrected Manhattan plots showing significance of association of all single-nucleotide polymorphisms (SNPs) for four smoking phenotypes. (**a**–**d**). SNPs plotted on the *x* axis according to their position on each chromosome against, on the *y* axis (shown as −log10 *P*-value), the association with (**a**) smoking initiation (SI, ever vs never smokers), (**b**) age of SI, (**c**) cigarettes smoked per day, and (**d**) smoking cessation (former vs current smokers). Dotted red line indicates genome-wide significance threshold of $P<5\times10^{-8}$.

Table 2 SNPs with meta-analytic *P*-values of $<1\times10^{-6}$ for CPD and AOI

Phenotype	*SNP*	*Chromosome (bp position)*	*Nearby genes*	*Alleles**	*Coded AF*	*Sample size (N)*	*β*	*s.e.*	P-*value*	I^2 *(%)*
CPD	**rs2036527**	**15 (76638670)**	***CHRNA5***	**A/G**	**0.22**	**15 554**	**0.040**	**0.007**	$\mathbf{1.84\times10^{-8}}$	**41.6**
CPD	rs667282	15 (76650527)	*CHRNA5*	C/T	0.29	15 536	0.033	0.006	1.81×10^{-7}	21.7
CPD	rs3101457	1 (242599837)	*C1orf100*	A/G	0.75	15 513	0.041	0.008	2.63×10^{-7}	1.1
CPD	rs938682	15 (76683602)	*CHRNA3*	A/G	0.71	15 475	0.033	0.006	3.75×10^{-7}	17.4
CPD	rs547843	15 (23975140)	*LOC503519*	C/G	0.65	12 701	−0.035	0.007	6.16×10^{-7}	24.2
CPD	rs3813570	15 (76619887)	*PSMA4*	C/T	0.26	15 543	0.033	0.007	9.85×10^{-7}	0.0
AOI	rs1678618	10 (73476294)	*SPOCK2*	A/G	0.74	16 874	−0.060	0.012	8.25×10^{-7}	0.0
AOI	rs1245577	10 (73480920)	*SPOCK2*	C/G	0.26	16 877	0.060	0.012	8.30×10^{-7}	2.6
AOI	rs1612028	10 (73475296)	*SPOCK2*	C/G	0.75	16 798	−0.060	0.012	9.28×10^{-7}	6.3

Abbreviations: AF, allele frequency; AOI, age of smoking initiation; CPD, cigarettes smoked per day; SNP, single-nucleotide polymorphism.
First named allele is coded allele. Coded AF refers to the allele analyzed as the predictor allele; it is not necessarily the minor allele. All SNPs coded to NCBI Build 36/UCSC hg18 forward strand. One SNP (rs2036527) highlighted in bold text achieved genome-wide significance.

multiple independent signals for CPD in this region in African Americans or the frequencies of the functional alleles and/or their effect sizes are much smaller than the signal defined by rs2036527.

Supplementary Table 4 presents how the variants associated with smoking behaviors in European ancestry populations performed in STOMP (rs1051730 in *CHRNA3*; rs16969968 in *CHRNA5*; rs1329650 and rs1028936 in *LOC100188947*; rs3733829 in *EGLN2*, near *CYP2A6*; rs6265, rs1013443, rs4923457, rs4923460, rs4074134, rs1304100, rs6484320 and rs879048 in *BDNF*; and rs3025343, near *DBH*). We observed modest nominally statistically significant associations for CPD with rs1051730 ($P=0.0079$) and rs16969968 ($P=0.027$), and for SC with rs3025343 ($P=0.03$).

Discussion

Investigating whether there are genetic variants associated with smoking behavior among African Americans is important, given that smoking prevalence and smoking-attributable mortality differ by race/ethnicity. Smoking prevalence and smoking intensity are lower for African Americans than European Americans, yet African Americans are less likely to successfully quit smoking.[41]

To our knowledge, this is the first meta-analysis of GWAS data for smoking behaviors in African Americans. The single genome-wide significant association we observed between rs2036527 and CPD is the same signal that was reported previously at 15q25.1 for nicotine dependence, smoking intensity and lung cancer in European ancestry samples.[4–6,42,43] The strong association that we found for this SNP supports studies suggesting that it is highly correlated with the functional allele(s) in populations of African ancestry. The fact that we did not observe a strong second association signal in this region after conditioning on rs2036527 suggests that rs2036527 and correlated SNPs in the African ancestry populations may define a single common haplotype at chr15q25.1 with sufficient effect size to be detected in our sample. After back transformation of the beta estimate, mean CPD values for each rs2036527 genotype were 14.6 for AA, 13.5 for AG and 12.8 for GG, suggesting that there is an increase of less than one cigarette smoked per day for each copy of the A allele. This SNP accounted for approximately 0.20% of the phenotypic variance of CPD in our sample. This effect is similar to that reported for rs1051730, which is correlated with rs2036527, where each copy of the rs1051730 A allele corresponds to a approximately one CPD increase and accounts for 0.5% of the phenotypic variance in smoking quantity in populations of European ancestry.

A study of *CHRNA5* knock-out mice showed that re-expressing this gene in the medial habenula, which extends projections to a brain region shown to mediate nicotine withdrawal,[44] abolished the inhibitory effects of nicotine while maintaining the reinforcing effects of nicotine.[45] In a functional magnetic resonance study of smokers, genetic variation in *CHRNA5* appeared to also affect reactivity to smoking cues in the insula, hippocampus and dorsal striatum, regions implicated in addictive behavior and memory.[46] Thus, it is biologically plausible that rs2036527, as a correlate of increased expression of the *CHRNA5* gene, could be associated with smoking quantity as a consequence of neuro-adaptations resulting from complex interactions between genes and environment that alter positive and negative reinforcement.[47]

To our knowledge, no SNPs in the *SPOCK2* gene, which encodes a protein that forms part of the extracellular matrix, have been reported previously in association with smoking behaviors or smoking-related cancer phenotypes. Variants at the *SPOCK2* locus have been linked to bronchopulmonary dysplasia, a respiratory condition observed in premature infants[48] that has been linked to intrauterine smoke exposure.[49] These variants are weakly correlated with the SNPs identified at this locus for AOI in Europeans ($r^2<0.25$ in CEU), but are not correlated in the African ancestry populations ($r^2=0$). The top SNP associated with SC (rs3813637) is located at 1q25 in the *C1orf49* gene. This locus has been linked to late-onset Alzheimer's disease, but genetic variation at this locus has not been reported in association with smoking behavior.[50] We are not aware of any smoking-related, other behavioral or pathological phenotypes associated with the variants we detected at 1q44 (*C1orf100)* and 15q12 (*LOC503519*) or *CTCT1* for CPD.

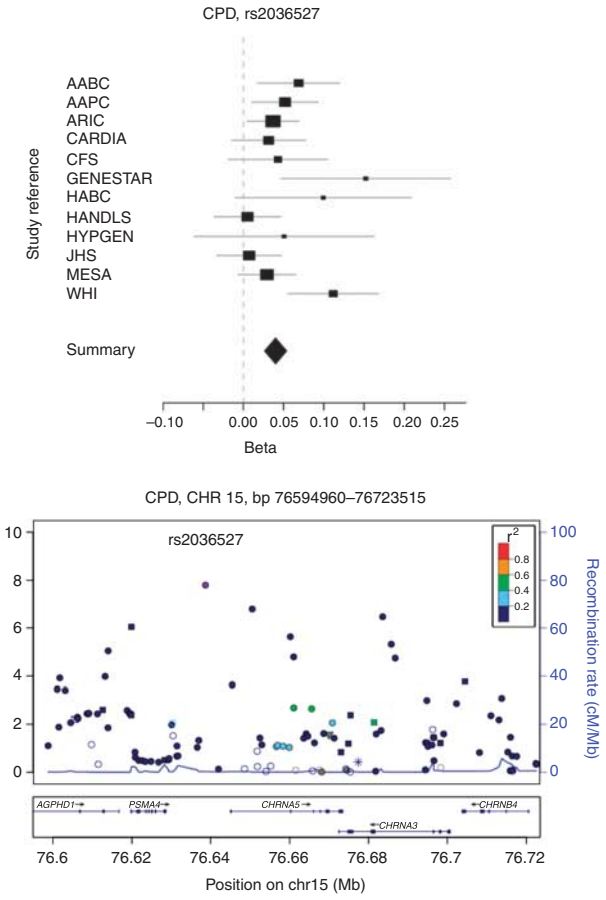


Figure 2 Forest and regional plot of rs2036527 with cigarettes smoked per day (CPD) from meta-analyses of the Study of Tobacco in Minority Populations (STOMP) consortia. Forest plot showing effect sizes across studies; $I^2 = 41.6\%$. Regional association plot show single-nucleotide polymorphisms (SNPs) plotted by position on chromosome against −log10 *P*-value. Estimated recombination rates (from HapMap-CEU) are plotted in light blue to reflect the local linkage disequilibrium (LD) structure on a secondary *y* axis. The SNPs surrounding the most significant SNP (purple) are color-coded to reflect their LD with this SNP (using pairwise r^2 values from HapMap-CEU): orange, $r^2$0.8, red; 0.6–0.8, orange; 0.6–0.8; green, 0.4–0.6, light blue, 0.2–0.4; dark blue, <0.2. The blue bars at the bottom of the plot represent the relative size and location of genes in the region. AABC, African American GWAS consortia of Breast cancer; AAPC, African American GWAS consortia of Prostate Cancer; ARIC, Atherosclerosis Risk in Communities; CARDIA, Coronary Artery Risk Development in Young Adults; CFS, Cleveland Family Study; JHS, Jackson Heart Study; MESA, Multi-Ethnic Study of Atherosclerosis; HANDLS, Healthy Aging in Neighborhoods across the Life Span Study; HYPGEN, Hypertension Genetic Epidemiology Network; WHI, Women's Health Initiative.

Although this is the largest GWAS meta-analysis of smoking phenotypes conducted to date in men and women of African ancestry, statistical power was a significant limitation. We had 80% power (for a mean allele frequency of 0.15 and α of 5×10^{-8}) to detect effect sizes of 1.25 for SI, AOI and SC, and a β of 0.15 for CPD. Notably, effect sizes for variants reported with many of these smoking phenotypes reported in the larger GWAS of the European ancestry were much smaller. For example, TAG, ENGAGE and Ox-GSK consortia reported β for SI of 0.015 for SNPs in *BDNF* and 0.026 for rs3025343 in *DBH*. Thus, we cannot rule out the possibility of additional loci that influence smoking behavior among African Americans that may be detected with larger sample sizes.

This analysis was limited by the fact that we were not able to adjust for local admixture, and the chip coverage of common variants (>5%) is less complete compared with the European populations,[51] which applies to most GWAS of African American populations. However, the use of a global adjustment for population genetic variation in the regression analysis using the principal components approach provided some measure of control for potential confounding because of population admixture.[34,52] Additionally, we acknowledge the limited precision of the smoking phenotypes. Smoking quantity is a highly heritable trait: estimates for CPD, heavy versus light smoking and/or pack-years range from 40 to 70% heritability in the European, African and Asian ancestry twin and family studies. Other studies have estimated that shared environmental factors account for 50% or more of the observed variation in SI, AOI and SC.[1,18,20,53–57]

We were unable to directly assess more refined phenotypes and highly heritable traits such as nicotine metabolism,[58] given our reliance on existing data originally collected for other purposes. Moreover, we were unable to examine gene × environment interactions using meta-GWAS analytic approach. Our analyses did not incorporate environmental covariate analyses, such as type of cigarettes smoked, mentholated or non-mentholated, dietary factors, socioeconomic status and other factors that might influence one or more of the phenotypes analyzed—data were not uniformly available and beyond the scope of the planned analyses we undertook in this discovery investigation. Future prospective studies with more detailed characterizations of smoking phenotypes and relevant environmental covariates are needed to identify additional variants that may be associated with smoking behaviors.

In summary, collective findings from GWAS among the African and European ancestry populations implicate chromosome 15q25 region as the most significant for smoking quantity. However, for both populations, SNPs in this region are associated with very small changes in smoking quantity and explain a small proportion of the variance, which suggests that conventional GWAS approaches may not be adequate to discover the likely hundreds of variants contributing small increments in risks of the additive genetic effects for heritable traits or so-called 'missing heritability' of complex diseases.[59] The use of more refined, specific and harmonized phenotypes capturing the complex behavior of SI, trajectories of progression and cessation, and environmental effect-modifiers are also needed to detect the genetic architecture of smoking behavior in different ancestral populations. Larger studies utilizing next-generation SNP arrays, whole-exome or whole-genome sequencing will be required to investigate lower-frequency variation, which may contribute to unexplained heritability for common traits.[60]

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements. We wish to acknowledge the many contributors from multiple institutions and funders who contributed to this project. Detailed acknowledgements are described in the supplementary information available at *Translational Psychiatry*'s website.

1. Lessov CN, Martin NG, Statham DJ, Todorov AA, Slutske WS, Bucholz KK *et al.* Defining nicotine dependence for genetic research: evidence from Australian twins. *Psychol Med* 2004; **34**: 865–879.
2. Broms U, Silventoinen K, Madden PA, Heath AC, Kaprio J. Genetic architecture of smoking behavior: a study of Finnish adult twins. *Twin Res Hum Genet* 2006; **9**: 64–72.
3. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008; **452**: 638–642.
4. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F *et al.* Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* 2010; **42**: 448–453.
5. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L *et al.* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010; **42**: 436–440.
6. Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, Ardissino D *et al.* Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**: 441–447.
7. Saccone NL, Schwantes-An TH, Wang JC, Grucza RA, Breslau N, Hatsukami D *et al.* Multiple cholinergic nicotinic receptor genes affect nicotine dependence risk in African and European Americans. *Genes Brain Behav* 2010; **9**: 741–750.
8. Saccone NL, Wang JC, Breslau N, Johnson EO, Hatsukami D, Saccone SF *et al.* The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer Res* 2009; **69**: 6848–6856.
9. Bierut LJ. Convergence of genetic findings for nicotine dependence and smoking related diseases with chromosome 15q24-25. *Trends Pharmacol Sci* 2010; **31**: 46–51.
10. Thorgeirsson TE, Stefansson K. Commentary: gene-environment interactions and smoking-related cancers. *Int J Epidemiol* 2010; **39**: 577–579.
11. Hindorff LA, Junkins HA, Hall PN, Mehta JP, Manolio TA. *A Catalog of Published Genome-Wide Association Studies*, Available at:##www.genome.gov/gwastudies##Accessed 25 July 2011. 2011.
12. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. *Nat Rev Genet* 2010; **11**: 356–366.
13. Haiman CA, Stram DO, Wilkens LR, Pike MC, Kolonel LN, Henderson BE *et al.* Ethnic and racial differences in the smoking-related risk of lung cancer. *N Engl J Med* 2006; **354**: 333–342.
14. Benowitz NL, Dains KM, Dempsey D, Wilson M, Jacob P. Racial differences in the relationship between number of cigarettes smoked and nicotine and carcinogen exposure. *Nicotine Tob Res* 2011; **13**: 772–783.
15. Mwenifumbo JC, Sellers EM, Tyndale RF. Nicotine metabolism and CYP2A6 activity in a population of black African descent: impact of gender and light smoking. *Drug Alcohol Depend* 2007; **89**: 24–33.
16. Moolchan ET, Berlin I, Robinson ML, Cadet JL. Characteristics of African American teenage smokers who request cessation treatment: implications for addressing health disparities. *Arch Pediatr Adolesc Med* 2003; **157**: 533–538.
17. Bierut LJ. Genetic vulnerability and susceptibility to substance dependence. *Neuron* 2011; **69**: 618–627.
18. Whitfield KE, King G, Moller S, Edwards CL, Nelson T, Vandenbergh D. Concordance rates for smoking among African-American twins. *J Natl Med Assoc* 2007; **99**: 213–217.
19. Li MD, Payne TJ, Ma JZ, Lou XY, Zhang D, Dupont RT *et al.* A genomewide search finds major susceptibility loci for nicotine dependence on chromosome 10 in African Americans. *Am J Hum Genet* 2006; **79**: 745–751.
20. Li MD, Cheng R, Ma JZ, Swan GE. A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins. *Addiction (Abingdon, England)* 2003; **98**: 23–31.
21. True WR, Heath AC, Scherrer JF, Waterman B, Goldberg J, Lin N *et al.* Genetic and environmental contributions to smoking. *Addiction (Abingdon, England)* 1997; **92**: 1277–1287.
22. CDC. Cigarette smoking among adults–United States, 2007. *Morb Mortal Wkly Rep* 2008; **57**: 1221–1226.
23. Hughes JR, Keely J, Naud S. Shape of the relapse curve and long-term abstinence among untreated smokers. *Addiction (Abingdon, England)* 2004; **99**: 29–38.
24. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet* 2009; **10**: 387–406.
25. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 2003; **165**: 2213–2233.
26. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.
27. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 2006; **78**: 629–644.
28. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 2009; **84**: 210–223.
29. Browning SR. Multilocus association mapping using variable-length Markov chains. *Am J Hum Genet* 2006; **78**: 903–913.
30. Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genet* 2008; **124**: 439–450.
31. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**: 1084–1097.
32. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 2010; **11**: 499–511.
33. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
34. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.
35. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 2005; **37**: 1243–1246.
36. Devlin B, Bennett P, Dawson G, Figlewicz DA, Grigorenko EL, McMahon W *et al.* Alleles of a reelin CGG repeat do not convey liability to autism in a sample from the CPEA network. *Am J Med Genet B Neuropsychiatr Genet* 2004; **126B**: 46–50.
37. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Gen* 2008; **17**(R2): R122–R128.
38. Ioannidis JP, Patsopoulos NA, Evangelou E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* 2007; **2**: e841.
39. Gauderman WJ, Morrison JM. *QUANTO 1.1: A Computer Program for Statistical Power and Sample Size Calculations for Genetic-Epidemiology Studies* 2006.
40. Smith RM, Alachkar H, Papp AC, Wang D, Mash DC, Wang JC *et al.* Nicotinic alpha5 receptor subunit mRNA expression is associated with distant 5′ upstream polymorphisms. *Eur J Hum Genet* 2011; **19**: 76–83.
41. Trinidad DR, Perez-Stable EJ, White MM, Emery SL, Messer K. A nationwide analysis of US racial/ethnic disparities in smoking behaviors, smoking cessation, and cessation-related factors. *Am J Public Health* 2011; **101**: 699–706.
42. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008; **40**: 616–622.
43. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008; **452**: 633–637.
44. Salas R, Sturm R, Boulter J, De Biasi M. Nicotinic receptors in the habenulo-interpeduncular system are necessary for nicotine withdrawal in mice. *J Neurosci* 2009; **29**: 3014–3018.
45. Fowler CD, Lu Q, Johnson PM, Marks MJ, Kenny PJ. Habenular alpha5 nicotinic receptor subunit signalling controls nicotine intake. *Nature* 2011; **471**: 597–601.
46. Janes AC, Smoller JW, David SP, Frederick BD, Haddad S, Basu A *et al.* Association between CHRNA5 genetic variation at rs16969968 and brain reactivity to smoking images in nicotine dependent women. *Drug Alcohol Depend* 2012; **120**: 7–13.
47. Robinson TE, Berridge KC. The psychology and neurobiology of addiction: an incentive-sensitization view. *Addiction (Abingdon, England)* 2000; **95**(Suppl 2): S91–S117.
48. Hadchouel A, Durrmeyer X, Bouzigon E, Incitti R, Huusko J, Jarreau PH *et al.* Identification of SPOCK2 as a Susceptibility Gene for Bronchopulmonary Dysplasia. *Am J Respir Crit Care Med* 2011; **184**: 1164–1170.
49. Antonucci R, Contu P, Porcella A, Atzeni C, Chiappe S. Intrauterine smoke exposure: a new risk factor for bronchopulmonary dysplasia? *J Perinat Med* 2004; **32**: 272–277.
50. Liu F, Arias-Vasquez A, Sleegers K, Aulchenko YS, Kayser M, Sanchez-Juan P *et al.* A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. *Am J Hum Genet* 2007; **81**: 17–31.
51. Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. *Nat Rev Genet* 2006; **7**: 885–891.
52. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
53. Kaprio J, Koskenvuo M, Langinvainio H. [Finnish twins reared apart. IV: smoking and drinking habits. A preliminary analysis of the effect of heredity and environment]. *Acta Genet Med Gemellol* 1984; **33**: 425–433.
54. Kaprio J, Koskenvuo M, Sarna S. Cigarette smoking, use of alcohol, and leisure-time physical activity among same-sexed adult male twins. *Prog Clin Biol Res* 1981; **69**(Pt C): 37–46.
55. Lessov-Schlaggar CN, Pang Z, Swan GE, Guo Q, Wang S, Cao W *et al.* Heritability of cigarette smoking and alcohol use in Chinese male twins: the Qingdao twin registry. *Int J Epidemiol* 2006; **35**: 1278–1285.

56. Carmelli D, Swan GE, Robinette D, Fabsitz RR. [Heritability of substance use in the NAS-NRC Twin Registry]. *Acta Genet Med Gemellol (Roma)* 1990; **39**: 91–98.
57. Hettema JM, Corey LA, Kendler KS. A multivariate genetic analysis of the use of tobacco, alcohol, and caffeine in a population based sample of male and female twins. *Drug Alcohol Depend* 1999; **57**: 69–78.
58. Swan GE, Lessov-Schlaggar CN, Bergen AW, He Y, Tyndale RF, Benowitz NL. Genetic and environmental influences on the ratio of 3′hydroxycotinine to cotinine in plasma and urine. *Pharmacogenet Genomics* 2009; **19**: 388–398.
59. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ *et al*. Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
60. Goldstein DB. Growth of genome screening needs debate. *Nature* 2011; **476**: 27–28.

Supplementary Information accompanies the paper on the Translational Psychiatry website (http://www.nature.com/tp)

Supplementary material for Study of Tobacco in Minority Populations (STOMP) Genetics Consortium

Table of Contents:

I. Supplementary Table 1. Genotyping, imputation & statistical analysis for STOMP studies……..2

II. Supplementary Table 2. Sample quality control for STOMP studies……..3

III. Supplementary Table 3. Meta-analytic results for SNPs $<1x10^{-6}$ for any smoking phenotype...4

IV. Supplementary Figure 1. Estimated African ancestry across STOMP studies.....10

V. Supplementary Table 4. Associations of known variants for smoking behaviors ……..11

VI. Supplementary Figure 2. Conditional analysis results.....12

VII. Summary of the studies participating in STOMP……..14

VIII. Acknowledgements……..26

IX. Literature cited……..29

Supplementary Table 1. Genotyping, imputation & statistical analysis for STOMP studies.

Study	Genotyping					Imputation			Association Analyses		
	Platform	Inclusion Criteria			SNPs met QC criteria	Imputation software	Inclusion criteria		SNPs in meta-analysis	λ GC*	Analysis software
		MAF	Call rate	P HWE			MAF	Imputation quality			
AABC	1M duo	≥ .01	≥ .95	--	1 043 036	MACH	≥ .01	Rsq> 0.5	2 886 710	1.08	in-house C++ program
AAPC	1M duo	≥ .01	≥ .95	--	1 047 986	MACH	≥ .01	Rsq> 0.5	2 969 774	1.08	in-house C++ program
CHS	Illumina Human Omni-Quad_v1 BeadChip	--	≥ .97	$\geq 10^{-5}$	963 248	BEAGLE 3.2.1	n/a	n/a	2 770 583	1.03	R
ARIC	Affy 6.0	≥ .01	≥ .95	--	796 384	MACH	≥ .01	Rsq> 0.5	2 600 605	1.02	PLINK 1.06
CARDIA	Affy 6.0	≥ .01	≥ .95	--	839 912	MACH	≥ .01	Rsq> 0.5	2 599 949	1.03	PLINK 1.06
CFS	Affy 6.0	≥ .01	≥ .95	--	867 495	MACH	≥ .01	Rsq> 0.5	2 602 914	1.14	PLINK 1.06
JHS	Affy 6.0	≥ .01	≥ .95	--	796 384	MACH	≥ .01	Rsq> 0.5	2 610 328	1.04	PLINK 1.06
MESA	Affy 6.0	≥ .01	≥ .95	--	881 666	MACH	≥ .01	Rsq> 0.5	2 621 586	1.02	PLINK 1.06
GENESTAR	Illumina1 Mv1_C	≥ .01	≥ .90	$>10^{-8}$	938 240	MACH	≥ .01	Rsq> 0.3	2 296 036	1.03	R- GEE/GEEGLM for family structures
HANDLS	Illumina 1M	> .01	≥ .95	$>10^{-7}$	907 763	MACH	≥ .01	Rsq≥ 0.3	2 862 300	1.02	MACH2DAT, MACH2QTL, R, PLINK
HABC	Illumina Human1 M-Duo	≥ .01	≥ .97	$\geq 10^{-6}$	1 007 948	MACH	≥ .01	none	1 958 375	1.01	R
HYPERGEN	Affy 6.0	≥ .01	≥ .95	$\geq 10^{-6}$	846 813	MACH	≥ .01	Rsq> 0.3	2 846 152	1.04	GWAF (R)
WHI	Affy 6.0	≥ .01	≥ .95	--	855 294	MACH	≥ .01	Rsq> 0.5	2 424 494	1.02	SNPGWA, PLINK, PROBABLE (R)

* Lambda is shown for smoking initiation (ever versus never smokers). MAF=minor allele frequency, HWE= Hardy Weinberg Equilibrium

Supplementary Table 2. Sample quality control for STOMP studies.

Study	Sample QC		Samp inc
	Call rate	Other exclusions	
AABC	≥ 95%	1) ancestry outliers 2) suspected relatives 3) suspected males	
AAPC	≥ 95%	1) ancestry outliers 2) suspected relatives 3) suspected females	
CHS	> 95%	1) genotype discordant with known sex or prior genotyping	
ARIC	> 95%	1) Mendelian errors 2) missingness 3) cluster outliers 4) missing gender 5) replicates	
CARDIA	> 95%	1) Mendelian errors 2) missingness 3) cluster outliers 4) discordance with alternate platform	
CFS	> 95%	1) Mendelian errors 2) missingness 3) cluster outliers 4) discordance with alternate platform	
JHS	> 95%	1) Mendelian errors 2) missingness 3) cluster outliers 4) replicates	
MESA	> 95%	1) Mendelian errors 2) missingness 3) cluster outliers	
GENESTAR	≥ 95%	1) ancestry outliers 2) gender inconsistencies 3) Mendelian inconsistency rate >5%	
HANDLS	≥ 95%	1) ancestry outliers 2) suspected relatives 3) suspected gender discrepancies	
HABC	≥ 97%	1) sample failure 2) genotypic sex mismatch 3) first-degree relative of an included individual based on genotype data	
HYPERGEN	≥ 95%	1) blood sample mixed up 2) unknown smoking status 3) quit smoking for less than a year	
WHI	≥ 95%	1) ancestry outliers 2) suspected relatives 3) suspected males	

Supplementary Table 3. Meta-analytic results for SNPs <1x10-6 with AOI and CPD.

Phenotype	SNP	Chromosome (bp position)	Nearby genes	Alleles	Study	Coded AF	Sample Size (N)	β	s.e.	P-value
AOI	rs1245577	10 (73480920)	*SPOCK2*	C/G	AABC	0.26	2221	0.0934	0.0318	0.0033
					AAPC	0.26	3501	0.0470	0.0273	0.0850
					ARIC	0.26	1803	0.0673	0.0439	0.1255
					CARDIA	0.26	643	0.1118	0.1035	0.2809
					CFS	0.27	219	0.2685	0.1228	0.0299
					GENESTAR	0.29	551	0.0338	0.0482	0.4825
					HABC	0.27	586	0.1377	0.0433	0.0015
					HANDLS	0.24	613	-0.0040	0.0440	0.9184
					HYPERGEN	0.24	569	0.0112	0.0531	0.8336
					JHS	0.24	1043	0.0560	0.0805	0.4868
					MESA	0.24	919	0.0672	0.0569	0.2379
					WHI	0.26	4209	0.0483	0.0257	0.0598
					Overall	**0.26**	**16,877**	**0.0605**	**0.0122**	**8.30×10^{-7}**
AOI	rs1678618	10 (73476294)	*SPOCK2*	A/G	AABC	0.74	2219	-0.0960	0.0316	0.0023
					AAPC	0.75	3500	-0.0473	0.0269	0.0788
					ARIC	0.74	1803	-0.0682	0.0439	0.1205

					GENESTAR	0.71	551	-0.0336	0.0486	0.4888
					HABC	0.74	586	-0.1334	0.0427	0.0019
					HANDLS	0.76	613	0.0040	0.0450	0.9218
					HYPERGEN	0.75	569	0.0056	0.0546	0.9178
					JHS	0.75	1043	-0.0622	0.0809	0.4419
					MESA	0.75	919	-0.0666	0.0566	0.2400
					WHI	0.74	4209	-0.0471	0.0251	0.0598
					Overall	**0.74**	**16,874**	**-0.060**	**0.0121**	**8.25×10^{-7}**
AOI	rs1612028	10 (73475296)	*SPOCK2*	C/G	AABC	0.74	2221	-0.0955	0.0316	0.0025
					AAPC	0.75	3501	-0.0475	0.0270	0.07802
					ARIC	0.74	1803	-0.0697	0.0439	0.1126
					CARDIA	0.74	643	-0.1099	0.1029	0.2862
					CFS	0.73	219	-0.2713	0.1211	0.0262
					GENESTAR	0.71	551	-0.0173	0.0496	0.7271
					HABC	0.74	586	-0.1332	0.0429	0.0020
					HANDLS	0.76	613	0.0050	0.0450	0.9184
					HYPERGEN	0.76	491	-0.0054	0.0599	0.9286
					JHS	0.76	1043	-0.0552	0.0804	0.4929
					MESA	0.76	919	-0.0617	0.0566	0.2760
					WHI	0.74	4208	-0.0484	0.0250	0.0529
					Overall	**0.75**	**16,798**	**-0.0601**	**0.0122**	**9.28x10-7**

CPD	rs3101457	1 (242599837)	*C1orf100*	A/G	AABC	0.73	2208	0.0285	0.0251	0.2557
					AAPC	0.73	3541	0.0144	0.0199	0.4681
					ARIC	0.75	1028	0.0763	0.0218	0.0005
					CARDIA	0.77	409	0.0285	0.0303	0.3475
					CFS	0.76	276	0.0553	0.0374	0.1399
					GENESTAR	0.79	607	0.1052	0.0450	0.0194
					HABC	0.71	619	-0.0062	0.0482	0.8980
					HANDLS	0.75	636	0.0380	0.0210	0.0755
					HYPERGEN	0.75	577	0.0016	0.0643	0.9801
					JHS	0.75	653	0.0422	0.0241	0.0812
					MESA	0.76	882	0.0691	0.0223	0.0020
					WHI	0.76	4077	0.0017	0.0333	0.9582
					Overall	**0.75**	**15,513**	**0.0410**	**0.0079**	**2.63×10^{-7}**
CPD	rs2036527	15 (76638670)	*CHRNA5*	A/G	AABC	0.23	2246	0.0686	0.0259	0.0082
					AAPC	0.22	3542	0.0518	0.0211	0.0138
					ARIC	0.25	1028	0.0367	0.0163	0.0249
					CARDIA	0.22	409	0.0314	0.0233	0.1787
					CFS	0.23	276	0.0430	0.0315	0.1734
					GENESTAR	0.21	607	0.1523	0.0537	0.0046
					HABC	0.22	621	0.0994	0.0559	0.0759
					HANDLS	0.23	636	0.0050	0.0210	0.8063

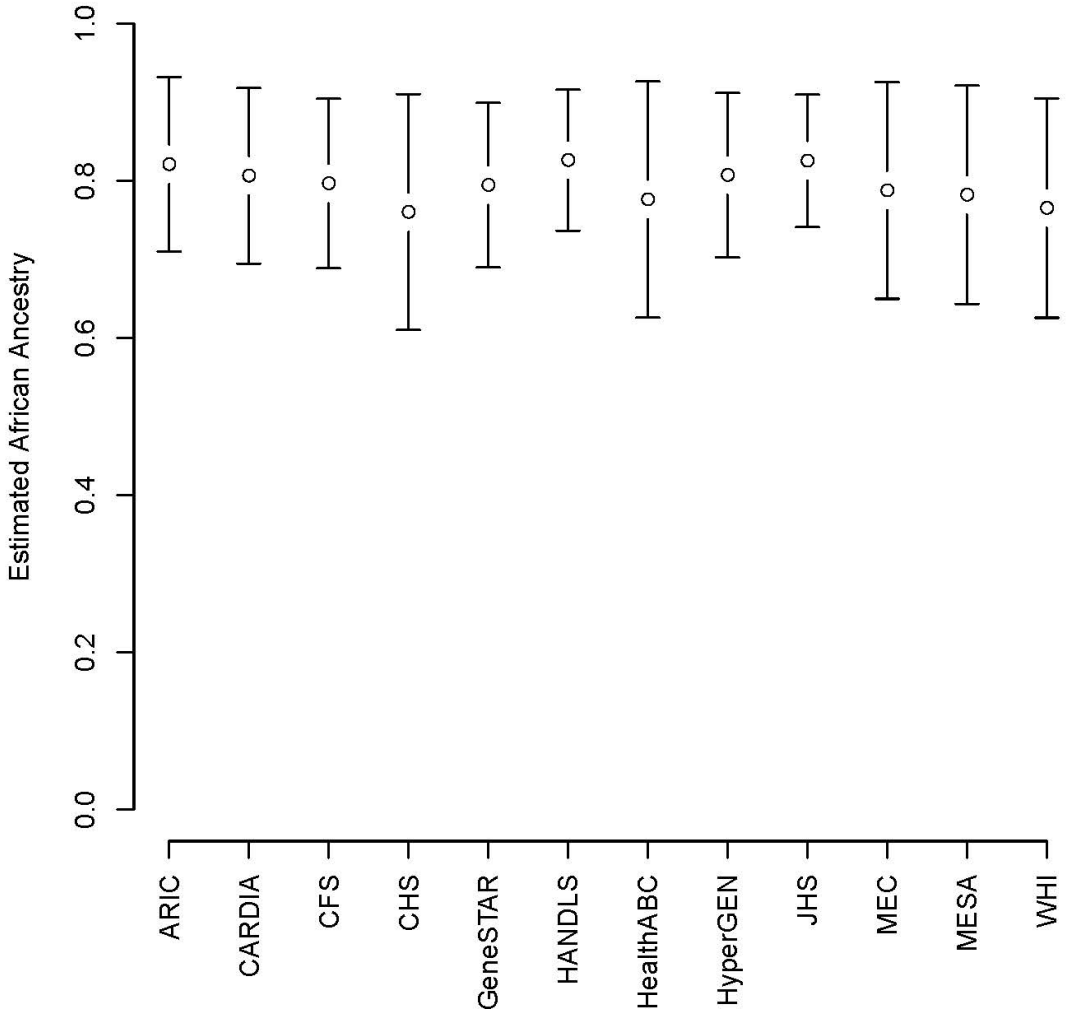
					HYPERGEN	0.22	577	0.0505	0.0571	0.3765
					JHS	0.23	653	0.0071	0.0200	0.7276
					MESA	0.24	882	0.0293	0.0180	0.1089
					WHI	0.21	4077	0.1117	0.0286	9.7×10^{-5}
					Overall	**0.22**	**15,554**	**0.0400**	**0.0071**	**1.84×10^{-8}**
CPD	rs667282	15 (76650527)	*CHRNA5*	C/T	AABC	0.29	2244	-0.0476	0.0242	0.0491
					AAPC	0.30	3533	-0.0653	0.0200	0.0009
					ARIC	0.28	1028	-0.0098	0.0149	0.5121
					CARDIA	0.30	409	-0.0029	0.0202	0.8860
					CFS	0.31	276	-0.0702	0.0272	0.0104
					GENESTAR	0.23	600	-0.0892	0.0436	0.0411
					HABC	0.29	621	-0.0632	0.0496	0.2035
					HANDLS	0.29	636	-0.0230	0.0190	0.2383
					HYPERGEN	0.30	577	-0.0077	0.0484	0.8731
					JHS	0.29	653	-0.0165	0.0181	0.3619
					MESA	0.26	882	-0.0431	0.0166	0.0096
					WHI	0.29	4077	-0.0576	0.0245	0.0186
					Overall	**0.29**	**15,536**	**0.0333**	**0.0064**	**1.81×10^{-7}**

CPD	rs938682	15 (76683602)	*CHRNA3*	A/G	AABC	0.71	2249	0.0408	0.0242	0.0923
					AAPC	0.71	3539	0.0636	0.0198	0.0012
					ARIC	0.72	1028	0.0080	0.0148	0.5869
					CARDIA	0.70	409	0.0025	0.0201	0.8993
					CFS	0.69	276	0.0722	0.0272	0.0084
					GENESTAR	0.76	606	0.0732	0.0445	0.0995
					HABC	0.71	621	0.0710	0.0500	0.1557
					HANDLS	0.72	636	0.0270	0.0200	0.1621
					HYPERGEN	0.68	500	0.0312	0.0495	0.5279
					JHS	0.71	653	0.0144	0.0180	0.4248
					MESA	0.73	882	0.0419	0.0165	0.0115
					WHI	0.71	4076	0.0575	0.0242	0.0176
					Overall	**0.71**	**15,475**	**0.0325**	**0.0064**	**3.4×10^{-7}**
CPD	rs547843	15 (23975140)	*LOC503519*	C/G	AABC	NA				
					AAPC	0.67	3542	-0.0373	0.0253	0.1403
					ARIC	0.65	1028	-0.0444	0.0148	0.0028
					CARDIA	0.63	409	-0.0099	0.0210	0.6367
					CFS	0.64	276	-0.0299	0.0260	0.2512
					GENESTAR	NA				
					HABC	0.68	621	-0.1862	0.0638	0.0036
					HANDLS	0.66	636	-0.0020	0.0250	0.9361

					HYPERGEN	0.64	577	-0.0430	0.0450	0.3390
					JHS	0.64	653	-0.0211	0.0184	0.2525
					MESA	0.67	882	-0.0384	0.0172	0.0262
					WHI	0.64	4077	-0.0674	0.0227	0.0029
					Overall	**0.65**	**12,701**	**-0.0352**	**0.0070**	**6.16×10^{-7}**
CPD	rs3813570	15 (76619887)	*PSMA4*	C/T	AABC	0.27	2247	-0.0397	0.024	0.1073
					AAPC	0.27	3535	-0.0392	0.020	0.0498
					ARIC	0.25	1028	-0.0178	0.017	0.2936
					CARDIA	0.27	409	0.0034	0.021	0.8728
					CFS	0.25	276	-0.0409	0.030	0.1804
					GENESTAR	0.14	603	-0.0631	0.044	0.1559
					HABC	0.27	620	-0.0498	0.050	0.3219
					HANDLS	0.27	636	-0.0290	0.021	0.1672
					HYPERGEN	0.27	577	0.0153	0.049	0.7577
					JHS	0.25	653	-0.0240	0.018	0.2052
					MESA	0.26	882	-0.0515	0.017	0.0030
					WHI	0.26	4077	-0.0816	0.025	0.0013
					Overall	**0.26**	**15,543**	**0.0333**	**0.0068**	**9.85×10^{-7}**

CPD (cigarettes per day): AOI (age of initiation). After double-GC correction, λ =1.008 for CPD and λ =1.018 for AOI.

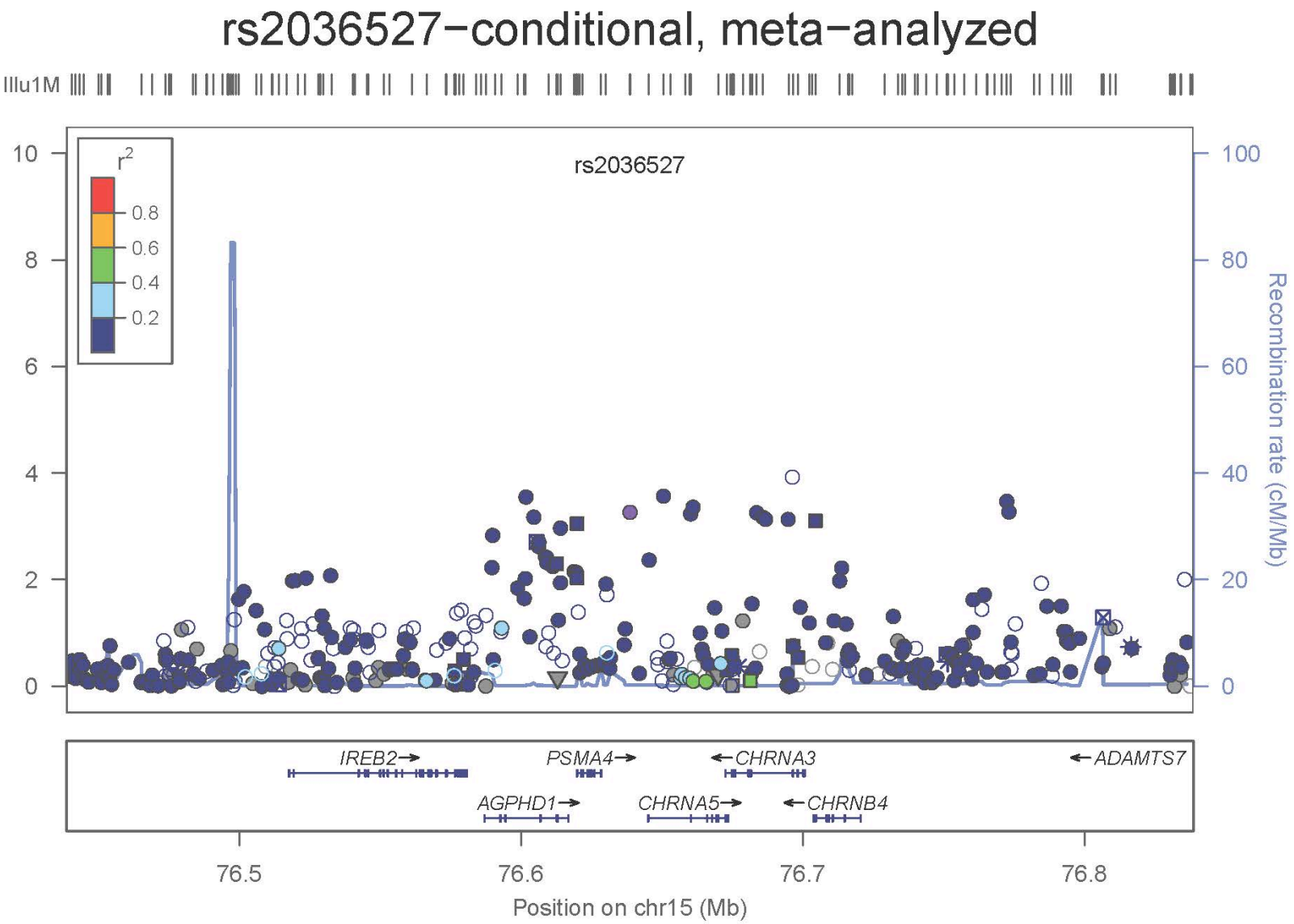
Supplementary Figure 1. Estimated African ancestry for participants across STOMP studies.



Supplementary Table 4. Variants associated with smoking phenotypes reported in largest meta-GWAS of European-ancestry (Tobacco and Genetics Consortium/ENGAGE/Ox-GSK)[1] and African-ancestry (STOMP Consortium).

				TAG				STOMP			
Phenotype	**SNP**	**Nearest Genes**	**Alleles***	**n**	**Frequency**	**β**	***P*-value**	**n**	**Frequency**	**β**	***P*-value**
CPD	rs1051730	*CHRNA3*	G/A	38 181	0.65	-1.021	$2.8x10^{-75}$	15 552	0.88	0.025	0.0079
CPD	rs16969968	*CHRNA5*	G/A	38 181	0.65	-1.003	$5.6x10^{-72}$	11 480	0.93	0.028	0.027
CPD	rs1329650	*LOC100188947*	T/G	38 181	0.28	-0.367	$5.7x10^{-10}$	15 551	0.11	0.006	0.54
CPD	rs1028936	*LOC100188947*	C/A	37 284	0.18	-0.446	$1.3x10^{-9}$	15 551	0.09	0.011	0.30
CPD	rs3733829	*EGLN2, CYP2A6*	G/A	38 181	0.36	0.333	$1.0x10^{-8}$	15 471	0.08	-0.018	0.091
Ever vs. Never	rs6265	*BDNF*	T/C	74 035	0.21	-0.061	$1.8x10^{-8}$	30 620	0.04	-0.036	0.42
Ever vs. Never	rs1013443	*BDNF*	T/A	74 035	0.26	-0.055	$3.3x10^{-8}$	NA	NA	NA	NA
Ever vs. Never	rs4923457	*BDNF*	T/A	74 035	0.23	-0.059	$3.3x10^{-8}$	31 798	0.17	-0.029	0.21
Ever vs. Never	rs4923460	*BDNF*	T/G	74 035	0.23	-0.058	$4.1x10^{-8}$	31 721	0.17	-0.029	0.21
Ever vs. Never	rs4074134	*BDNF*	T/C	74 035	0.23	-0.058	$4.1x10^{-8}$	31 796	0.17	-0.037	0.11
Ever vs. Never	rs1304100	*BDNF*	G/A	74 035	0.26	-0.055	$4.4x10^{-8}$	31 798	0.39	0.280	0.12
Ever vs. Never	rs6484320	*BDNF*	T/A	74 035	0.24	-0.057	$4.9x10^{-8}$	31 798	0.09	0.010	0.75
Ever vs. Never	rs879048	*BDNF*	C/A	74 035	0.23	-0.058	$4.9x10^{-8}$	31 606	0.37	0.010	0.59
Former vs. Current	rs3025343	*DBH*	G/A	41 278	0.84	0.121	$3.6x10^{-8}$	11 644	0.97	-0.207	0.03

'CPD' = cigarettes per day; 'Ever vs. Never' smoking (smoking initiation)
NA: Not available in STOMP Consortium


rs2036527−conditional, meta−analyzed
Illu1M
rs2036527
r2
0.8
0.6
0.4
0.2
10
8
6
4
2
0
100
80
60
40
20
0
Recombination rate (cM/Mb)
IREB2
PSMA4
CHRNA3
ADAMTS7
AGPHD1
CHRNA5
CHRNB4
76.5
76.6
76.7
76.8
Position on chr15 (Mb)

Supplemental Figure 2. Regional plot of chromosome 15q25 CPD meta-analysis association signals, conditioning on rs2036527. The top association signal remaining after the conditional analysis is for rs3743076 (p-value = 1.1×10^{-4}) in *CHRNA3*. SNPs are plotted by position on chromosome against −log10 *P* value. Estimated recombination rates (from HapMap-CEU) are plotted in light blue to reflect the local LD structure on a secondary *y* axis. The SNPs surrounding the most significant SNP (pink row & column) are color coded to reflect their LD with this SNP: orange, $r^2 \geq 0.8$, red; 0.6-0.8, orange; 0.6–0.8; green, 0.4–0.6, light blue, 0.2–0.4; dark blue, <0.2. The blue bars at the bottom of the plot represent the relative size and location of genes in the region.

Summary of the studies participating in STOMP

This study includes samples from two ongoing GWAS of breast (AABC) and prostate cancer (AAPC) in African Americans. A description of the studies participating in each of these scans and providing data to the meta-analysis of smoking traits is provided below.

1. Genome-wide Association Study of Breast Cancer in African Americans (AABC)

The Multiethnic Cohort Study (MEC): The MEC consists of over 215,000 men and women in Hawaii and Los Angeles (with additional African-Americans from elsewhere in California)[2]. The cohort is comprised predominantly of African Americans, Native Hawaiians, Japanese Americans, Latinos and European Americans who entered the study between 1993 and 1996 by completing a 26-page self-administered questionnaire that requested detailed information about dietary habits, demographic factors, personal behaviors, history of prior medical conditions, family history of common cancers, and for women, reproductive history and exogenous hormone use. The participants were between the ages 45 and 75 at enrollment. Incident cancers in the MEC are identified by cohort linkage to population-based cancer Surveillance, Epidemiology and End Results (SEER) registries covering Hawaii and Los Angeles County, and to the California State cancer registry covering all of California. Beginning in 1994, blood samples were collected from incident breast cancer cases and a random sample of MEC participants to serve as a control pool for genetic analyses in the cohort. In 2000, we established a large biorepository of blood and urine samples from incident cases of breast, prostate and colorectal cancer and from ~67,000 members of the cohort. Eligible cases in the African American breast cancer case-control study consisted of women with incident breast cancer diagnosed after enrollment in the MEC through December 31, 2007. Controls were participants without breast cancer prior to entry into the cohort and without a diagnosis up to December 31, 2007. Controls from the MEC included in the stage 1 sample were frequency matched to cases based age (in 5-year intervals). All participants provided written informed consent. This case-control study in the MEC includes 556 African American cases (2 in situ) and 1,003 African American controls. Genomic DNA was extracted for all samples using the QIAamp Blood Mini Kit (Qiagen, Valencia, CA).

Self-reported information on smoking was collected at cohort baseline via questionnaire. An ever smoker was defined as ≥20 pack-years in a lifetime. Information on the number of cigarettes smoked per day was collected as categorical data (≤5, 6-10, 11-20, 21-30, 31+), and was converted to a continuous variable for analysis. Detailed information about age at smoking initiation was not collected.

The Carolina Breast Cancer Study (CBCS): The CBCS is a population-based, case-control study conducted between 1993 and 2001 in 24 counties of central and eastern North Carolina [3]. Incident cases of breast cancer were identified using a rapid case ascertainment system in cooperation with the North Carolina Central Cancer Registry. Controls were selected from the North Carolina Division of Motor Vehicles for women younger than 65 years old and United States Health Care Financing Administration beneficiary lists for women aged 65 years or older. Controls were frequency matched to cases based upon age and race (± 5 years). Randomized recruitment was employed to over-sample African-American women and women under the age of fifty. In-person interviews were conducted in participants' homes and included informed consent, structured questionnaire, body measurements and blood draws. For cases, the average time interval between data of diagnosis and in-person interview was 4 months. A total of 1808 invasive breast cancer cases (788 African Americans, 1020 whites) and 1564 controls (718 African-Americans, 846 whites) were enrolled. Age of participants ranged from 20 to 74. [4]. The compliance rates for blood draws among African Americans were 86% for cases and 87% for controls. DNA was extracted from peripheral blood lymphocytes using the automated Applied Biosystems Nucleic Acid Purification

System. For the GWAS, DNA samples were provided from 656 African American cases with invasive breast cancer and 608 controls. African American participants with DNA available did not differ from the remaining participants based upon age, body mass index, waist hip ratio, family history or other breast cancer risk factors, or among cases for stage at diagnosis.

Self-reported information on smoking was collected from cases and controls via questionnaire at the time of interview, which for cases was after their date of diagnosis. The number of cigarettes per day was collected as categorical data (number of packs per day: <0.5, 0.5-1, >1), and was converted to a continuous variable for analysis.

The Los Angeles component of The Women's Contraceptive and Reproductive Experiences Study (CARE): The NICHD Women's Contraceptive and Reproductive Experiences (Women's CARE) Study is a large multi-center population-based case-control study that was designed to examine the effects of oral contraceptive (OC) use on breast cancer risk among African American women ages 35-64 years in five U.S. locations [5]. One of the five sites was Los Angeles County. In person interviews were conducted to collect detailed information about breast cancer risk factors (including OC use) and blood specimens were collected from subsets of cases and controls for genetic analyses. Cases from Los Angeles County were identified through the National Cancer Institute's local Surveillance, Epidemiology, and End Results (SEER) registry using rapid-reporting techniques. Eligibility criteria included: 1) they must have been alive and a resident of Los Angeles County at the reference date; 2) between 35 and 64 years of age; 3) self-reported African American race; 3) histologically-confirmed invasive breast cancer and no prior breast cancer prior to the reference date; 4) breast cancer diagnosis after July 1, 1994 and on or before April 30, 1998; 5) born in the U.S. and able to speak English and; 6) physically and mentally capable of completing the interview (e.g. could see the interview materials, hear and comprehend interviewer's questions, and verbally respond). Blood specimens were collected from 82% of invasive cases that were asked to donate a blood specimen. Controls were identified by random digit dialing from the same geographic area (i.e. Los Angeles County) and were women who had participated in the Women's CARE Study. All participants provided written informed consent. CARE contributed 380 African American cases and 224 African American controls to stage 1 of the scan.

Self-reported information on complete smoking history was collected during the interview up to a reference date that was the date of diagnosis for the breast cancer patient and the date of first contact with the control participant's household during the random digit dialing procedure. Ever smoking was defined as having smoked at least 100 cigarettes in the woman's lifetime.

The Women's Circle of Health Study (WCHS): The WCHS is an ongoing case-control study of breast cancer among European American (EA) and African American (AA) women 3). Originally initiated in the New York City (NYC) boroughs (Manhattan, the Bronx, Brooklyn and Queens), eligible cases were identified through collaborations with physicians at each of the hospitals. Currently, the study is limited to seven counties in New Jersey (NJ) (Bergen, Essex, Hudson, Mercer, Middlesex, Passaic, and Union), where eligible women newly diagnosed with breast cancer are identified through the New Jersey State Cancer Registry in collaboration with researchers at Cancer Epidemiology Services (CES) of the New Jersey Department of Health and Senior Services (NJDHSS) and the Cancer Institute of New Jersey (CINJ) through rapid case ascertainment. Samples and data for the AABC include AA women, 20 to 75 years of age, newly diagnosed with primary, histologically confirmed invasive breast cancer who speak English. Controls were identified through Random Digit Dialing (RDD), frequency matched to cases by 5-year age groups and race. The data collection for the study consists of an in-depth in person interview (with collection of reproductive and hormonal factors, diet, medical and family histories of cancer, and health lifestyle factors). Behavioral questionnaires and a Food Frequency Questionnaire as well as body measurements are also collected. A saliva sample is collected using Oragene® Kits and DNA is extracted

in batches, using the DNA Genotek protocol for DNA extraction from saliva. Enrollment and collection of specimens in the WCHS is currently ongoing. The WCHS contributed 272 invasive African American cases and 240 African American controls to stage 1 of the GWAS.

Self-reported information on smoking was collected via interview. For cases, smoking information was requested at the date of diagnosis while for controls smoking information was requested 3 months prior to interview. Ever smoking was defined as ≥1 cigarette smoked per day for a least 1 year.

The Northern California Breast Cancer Family Registry (NC-BCFR): The NC-BCFR is a population-based family study conducted in the Greater San Francisco Bay area, and is one of 6 sites collaborating in the Breast Cancer Family Registry (BCFR), an international consortium funded by NCI (described in detail in John et al. (6)). Women aged 18-64 years, residing in the 9 counties of the Greater San Francisco Bay Area, and newly diagnosed with invasive or in situ breast cancer from 1995-2009 were identified through the Greater San Francisco Bay Area Cancer Registry. Cases were eligible to enroll in the NC-BCFR if they had indicators of increased genetic risk (i.e., diagnosed before age 35 years, prior diagnosis of ovarian or childhood cancer, bilateral breast cancer with the first diagnosis before age 50 years, or a family history of breast, ovarian or childhood cancer in first-degree relatives). Cases not meeting these criteria were randomly sampled (2.5% of non-Hispanic whites, 32% of other race/ethnicities). Population controls aged 18-64 years were identified through random-digit dialing from 1999-2000 and frequency-matched to cases diagnosed from 1995-1998 on race/ethnicity and five-year age group at a case:control ratio of 2:1. Cases and controls provided information on cancer family history and breast cancer risk factors by interview, and provided a blood or mouthwash sample. All study participants provided written informed consent. The NC-BCFR contributed for 440 invasive African American cases and 53 African American controls to stage 1 of the GWAS.

Self-reported information on smoking was collected via interview. Smoking information was collected up to one year prior to diagnosis for cases and up to the time of interview for controls. Ever smoking was defined as ≥1 cigarette smoked per day for 3 months or longer.

The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) Cohort: PLCO, coordinated by the U.S. National Cancer Institute (NCI) in 10 U.S. centers, enrolled during 1993 - 2001 approximately 155,000 men and women, aged 55-74, in a randomized, two-arm trial to determine if screening reduced the mortality from these cancers. Approximately 37,500 women were assigned to each arm. At entry, demographic, medical, and risk factor information was collected from all participants. A general risk factor questionnaire was distributed during 2006-7 to update information. Sequential blood samples, including plasma, serum, buffy coat, and whole blood, were collected from participants assigned to the screening arm; participation was 93% at the baseline blood draw (1993 – 2001). Buccal cells were collected from the participants assigned to the control arm. All incident cancers are ascertained by annual mailed questionnaires. Hospital confirmation of diagnosis, medical records, and pathology reports are requested for all cancers reported. [7]. A total of 1642 women (61 Black, non-Hispanic) in the screening arm and 1649 women (60 Black, non-Hispanic) in the control arm, with no history of breast cancer at baseline and a completed baseline questionnaire were diagnosed with incident breast cancer by December, 2008. Of the Black women with breast cancer, 44 in the screening arm and 28 in the control arm had DNA available for genotyping and had provided informed consent. Using incidence density sampling, two controls who were breast cancer-free at the age of diagnosis of the case (5-year categories) were identified for each case. Controls were matched on race (Black, non-Hispanic), study arm (screening/control), date at cohort entry (<, ≥ median entry time for all female Black participants), and age at cohort entry. Thus, a total of 64 cases and 133 controls were selected. Genotyping of the PLCO samples was performed at the NCI Core Genotyping Facility (CGF).

Self-reported information on smoking was collected via questionnaire at baseline. Ever smoking was defined as smoking regularly for >6 months. Information on the number of cigarettes smoked per day was collected as categorical data (1-10, 11-20, 21-30, 31-40, 41-60, 61-80), and was converted to a continuous variable for analysis.

The Nashville Breast Health Study (NBHS): The NBHS is a population-based case-control study of incident breast cancer conducted in the Nashville, TN metropolitan area [8]. Through a rapid case-ascertainment system, we identified cases in conjunction with the Vanderbilt University Medical Center, Meharry Medical College/Metropolitan Nashville Hospital, Baptist Hospital, Saint Thomas Hospital, Centennial Hospital and The Tennessee State Cancer Registry. Eligible cases were Caucasian women diagnosed between April 1, 2001, and March 31, 2008, and African American women diagnosed after April 1, 2001, with invasive breast cancer or ductal carcinoma *in situ* who were between the ages of 25 and 75, had no prior history of cancer other than non-melanoma skin cancer, had a resident telephone, spoke English and who were able to provide consent to the study. In an effort to increase the numbers of African American women included in our study, additional cases were identified from Hamilton, TN, and Shelby, TN, counties after Nov. 17, 2006, and from the entire state of Tennessee following May 28, 2008. Controls were identified via random digit dialing (RDD) of households in the eight counties including and surrounding Nashville. Additionally, African American neighbor and non-blood relatives were identified by African American cases as potential controls. Eligibility criteria for controls were the same as cases with the exception that controls could not have a prior cancer diagnosis other than simple skin cancer. Controls were frequency matched to cases on 5-year age group, race and county of residence. Approval for this study was garnered from the Institutional Review Board of Vanderbilt University Medical Center and those of the individual collaborating institutions. All participants provided informed consent prior to enrollment in this study. Information on demographic factors, known and potential risk factors for breast cancer, lifestyle factors and dietary history were ascertained through a structured telephone interview and a self-administered food frequency questionnaire. Buccal cell samples were collected via two methods: Oragene saliva collection kits (DNA Genotek, Ottawa, Canada) and mouthwash samples. A total of 325 eligible, consented African American cases and 197 eligible, consented African American controls were available for inclusion in this study. Sufficient DNA for genotyping was not available for 15 cases and 11 controls, leaving a total of 310 cases and 186 for the analysis.

Self-reported information on smoking was collected via questionnaire at study entry for both controls and cases (after diagnosis). Ever smoking was defined as ≥100 cigarettes smoked in a lifetime.

Wake Forest University (WFBC): Study participants were recruited at Wake Forest University Health Sciences beginning in November 1998 (samples collected up to December 2008 were used for this study) (11). Newly-diagnosed breast cancer cases, prior to any therapy, were enrolled at the Wake Forest University Breast Care Center. Histopathology and medical records were reviewed to confirm diagnosis. Controls were recruited from the patient population receiving routine mammography at the Breast Screening and Diagnostic Center. Eligibility criteria for controls included normal mammography results and no prior cancer history. Study participants reviewed a brief description of the protocol with a research coordinator and provided their signed, informed consent, as approved by the medical center's Institutional Review Board. Whole blood (20 ml) was collected from enrolled subjects and processed within 2 hours after phlebotomy. Every study participant completed a self-administered baseline questionnaire, which included information on demographics, reproductive history, medical conditions, and family history of cancer. Genomic DNA was extracted from frozen whole blood using the QIAamp DNA Blood Mini kit (Qiagen, Valencia, CA). This study contributed 125 cases (9 in situ) and 153 controls to the current GWAS.

Self-reported information on smoking was collected via questionnaire at study entry for both controls and cases (after diagnosis). Ever smoking was defined as ≥100 cigarettes smoked in a lifetime. The number of cigarettes per day was collected as continuous variable for analysis.

Genotyping and QC for AABC

The AABC scan includes 5,984 samples from 9 studies (3,153 cases and 2,831 controls). Eight of these studies are included in this meta-analysis of smoking traits. Genotyping was conducted using the Illumina Human1M-Duo BeadChip.(43) We attempted genotyping of 5,932, removing samples (n=52) with DNA concentrations <20 ng/ul. Following genotyping, we removed samples based on the following exclusion criteria: 1) unknown replicates (≥98.9% genetically identical) that we were able to confirm (only one of each duplicate was removed, n=15); 2) unknown replicates that we were not able to confirm through discussions with study investigators (pair or triplicate removed, n=14); 3) samples with call rates <95% after a second attempt (n=100); 4) samples with ≤ 5% African ancestry (n=36) (discussed below); and, 5) samples with <15% mean heterozygosity of SNPs in the X chromosome and/or similar mean allele intensities of SNPs on the X and Y chromosomes (n=6) (these are likely to be males).

In the analysis, we removed SNPs with <95% call rate (n=21,732) or minor allele frequencies (MAFs) <1% (n=80,193). To assess genotyping reproducibility we included 138 replicate samples; the average concordance rate was 99.95% (>99.93% for all pairs). We also eliminated SNPs with genotyping concordance rates <98% based on the replicates (n=11,701). The final analysis dataset included 1,043,036 SNPs genotyped on 3,016 cases and 2,745 controls, with an average SNP call rate of 99.7% and average sample call rate of 99.8%.

Global Ancestry Estimation. We also applied principal components analysis (PCA) [9] to estimate axes of variation among the 5,761 individuals using 2,546 ancestry informative markers. The first eigenvector accounted for 10.1% of the variation between subjects, and subsequent eigenvectors accounted for no more than 0.5%.

2. Genome-wide Association Study of Prostate Cancer in African Americans (AAPC)

The Multiethnic Cohort (MEC), described above: Through January 1, 2008 the African American prostate cancer case-control study in the MEC included 1,094 cases and 1,096 controls (1).

The Southern Community Cohort Study (SCCS): The SCCS is a prospective cohort of African and non-African Americans which during 2002-2009 enrolled approximately 86,000 residents aged 40-79 years across 12 southern states [10]. Recruitment occurred mainly at community health centers, institutions providing basic health services primarily to the medically uninsured, so that the cohort includes many adults of lower income and educational status. Each study participant completed a detailed baseline questionnaire, and nearly 90% provided a biologic specimen (approximately 45% a blood sample and 45% buccal cells). Follow-up of the cohort is conducted by linkage to national mortality registers and to state cancer registries. Included in this study are 212 incident African American prostate cancer cases and a matched stratified random sample of 419 African American male cohort members without prostate cancer at the index date selected by incidence density sampling.

Self-reported information on smoking was collected via questionnaire at baseline. Ever smoking was defined as ≥100 cigarettes smoked in a lifetime.

The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO), described above. Included in this study are 286 African American prostate cancer cases and 269 controls without a history of prostate cancer, matched on age at randomization and study year of the trial (9).

Self-reported information on smoking was collected via questionnaire at baseline. Ever smoking was defined as smoking regularly for >6 months.

The Cancer Prevention Study II Nutrition Cohort (CPS-II). The CPS-II Nutrition Cohort includes over 86,000 men and 97,000 women from 21 US states who completed a mailed questionnaire in 1992 (aged 40-92 years at baseline) [11]. Starting in 1997, follow-up questionnaires were sent to surviving cohort members every other year to update exposure information and to ascertain occurrence of new cases of cancer; a >90% response rate has been achieved for each follow-up questionnaire. From 1998-2001, blood samples were collected in a subgroup of 39,376 cohort members. To further supplement the DNA resources, during 2000-2001, buccal cell samples were collected by mail from an additional 70,000 cohort members. Incident cancers are verified through medical records, or through state cancer registries or death certificates when the medical record can not be obtained. Genomic DNA from 76 African American prostate cancer cases and 152 age-matched controls were included in stage 1 of the scan. Self-reported information on smoking was collected via questionnaire at baseline. Ever smoking was defined as ≥100 cigarettes smoked in a lifetime.

Prostate Cancer Case-Control Studies at MD Anderson (MDA): Participants in this study were identified from epidemiological prostate cancer studies conducted at the University of Texas M.D. Anderson Cancer Center in the Houston Metropolitan area since 1996 Cases were accrued from six institutions in the Houston Medical Center and were no restricted with respect to Gleason score, stage or PSA. Controls were identified via random-digit-dialing or among hospital visitors and they were frequency matched to cases on age and race. Lifestyle, demographic, and family history data were collected using a standardized questionnaire. These studies contributed 543 African American cases and 474 controls to this study. [35].

Self-reported smoking status was determined from data collected via interviewer-administered questionnaires. Ever-smokers were defined as those participants who reported smoking more than 100 cigarettes in their lifetimes. They were further categorized as “current smokers” or “former smokers”; former smokers were defined as those who had quit more than a year prior to diagnosis for the cases or the interview date for controls.

The Los Angeles Study of Aggressive Prostate Cancer (LAAPC): The LAAPC is a population-based case-control study of aggressive prostate among African Americans in Los Angeles County [12]. Cases were identified through the Los Angeles County Cancer Surveillance Program rapid case ascertainment system and eligible cases included African American men diagnosed with a first primary prostate cancer between January 1, 1999 and December 31, 2003. Eligible cases also had either tumor extension outside the prostate, metastatic prostate cancer in sites other than prostate, or needle biopsy of the prostate with Gleason grade 8 or higher, or Gleason grade 7 and tumor in more than 2/3 of the biopsy cores. Controls were identified by a neighborhood walk algorithm and were men never diagnosed with prostate cancer, and were frequency matched to cases on age (±5 years). For this study, genomic DNA was included for 296 cases and 140 controls. We also included an additional 163 African American controls from the MEC that were frequency matched to cases on age.

Self-reported information on smoking was collected via questionnaire at the interview date for cases (after diagnosis) and controls. Ever smoking was defined as ≥1 cigarette smoked per day for a ≥6 months.

Prostate Cancer Genetics Study (CaP Genes): The African-American component of this study population comprised 160 men: 75 cases diagnosed with more aggressive prostate cancer and 85 age-matched controls [13]. All subjects were recruited and frequency-matched on the major medical institutions in Cleveland, Ohio (i.e., the Cleveland Clinic, University Hospitals of Cleveland, and their affiliates) between 2001 and 2004. The cases were newly diagnosed with histologically confirmed disease: Gleason score 7; tumor stage T2c; or a prostate-specific antigen level >10 ng/ml at diagnosis. Controls were men without a prostate cancer diagnosis who underwent standard annual medical examinations at the collaborating medical institutions.

Self-reported information on smoking was collected via questionnaire at the reference date for controls and 1 year before case diagnosis for cases. The number of cigarettes per day was collected as categorical data (number of packs per day: (<0.5, 0.5, 1, 1.5, 2), and was converted to a continuous variable for analysis.

Case-Control Study of Prostate Cancer among African Americans in Washington, DC (DCPC): Unrelated men self-described as African American were recruited for several case-control studies on genetic risk factors for prostate cancer between the years 2001 and 2005 from the Division of Urology at Howard University Hospital (HUH) in Washington, DC. Control subjects unrelated to the cases and matched for age (± 5 years) were also ascertained from the prostate cancer screening population of the Division of Urology at HUH [14]. These studies included 292 cases and 359 controls.

Self-reported information on smoking was collected via questionnaire at the reference date for controls and 1 year before case diagnosis for cases. Ever smoking was defined as ≥100 cigarettes smoked in a lifetime. Information on the number of cigarettes smoked per day was collected as categorical data (1-5, 6-14, 15-24, 24-34, 35+), and was converted to a continuous variable for analysis. Information on age at smoking initiation was not available.

King County (Washington) Prostate Cancer Studies (KCPCS): The study population consists of participants from one of two population-based case-control studies among residents of King County, Washington [15, 16]. Incident Caucasian and African American cases with histologically confirmed prostate cancer were ascertained from the Seattle-Puget Sound SEER cancer registry during two time periods, 1993-1996 and 2002-2005. Age-matched (5-year age groups) controls were men without a self-reported history of being diagnosed with prostate cancer and were identified using one-step random digit telephone dialing. Controls were ascertained during the same time periods as the cases. A total of 145 incident African American cases and 81 African American controls were included from these studies.

Smoking status was determined for cases and controls by asking about their smoking history "prior to reference date," which was the date of diagnosis (month/year) for cases and a randomly assigned date for controls that matched the distribution of cases' diagnosis dates. Ever smoking was defined as ≥1 cigarette smoked per day for ≥6 months.

The Gene-Environment Interaction in Prostate Cancer Study (GECAP): The Henry Ford Health System (HFHS) recruited cases diagnosed with adenocarcinoma of the prostate of Caucasian or African-American race, less than 75 years of age, and living in the metropolitan Detroit tri-county area [17]. Controls were randomly selected from the same HFHS population base from which cases were drawn. The control sample was frequency matched at a ratio of 3 enrolled cases to 1 control based on race and five-year age stratum. In total, 637 cases and 244 controls were enrolled between January 2002 and December 2004. Of study enrollees, DNA for 234 African Americans cases and 92 controls were included in stage 1 of the scan.

Self-reported information on smoking was collected via questionnaire at the reference date for controls and 1 year before case diagnosis for cases. Ever smoking was defined as smoking regularly for

>6 months. The number of cigarettes per day was collected as categorical data (number of packs per day: (<0.5, 0.5-1, 1-1.5, 1.5-2, 2+), and was converted to a continuous variable for analysis.

Genotyping and Quality Control for AAPC

The AAPC scan includes 7,123 samples from 11 studies (3,621 cases and 3,502 controls). Ten of these studies are included in this meta-analysis of smoking traits. Genotyping was conducted using the Illumina Infinium 1M-Duo bead array at the University of Southern California and the NCI Genotyping Core Facility (PLCO study). Following genotyping samples were removed based on the following exclusion criteria: 1) unknown replicates across studies (n=24, none within studies); 2) call rates <95% (n=126); 3) samples with >10% mean heterozygosity on the X chromosome and/or <10% mean intensity on the Y chromosome - we inferred 3 samples to be XX and 6 to be XXY; 4) ancestry outliers (n=108, discussed below), and; 5) samples that were related (n=141). To assess genotyping reproducibility we included 158 replicate samples; the average concordance rate was 99.99% (≥99.3% for all pairs). Starting with 1,153,397 SNPs, we removed SNPs with <95% call rate, MAFs <1%, or >1 QC mismatch based on sample replicates (n=105,411). The analysis included 1,047,986 SNPs among 3,425 cases and 3,290 controls.(45,46)

Global Ancestry Estimation.

The EIGENSTRAT software was used to calculate eigenvectors that explained genetic differences in ancestry among samples in the study (29). The program included data from both HapMap Phase 3 populations and our study, so that comparisons to reference populations of known ethnicity could be made. A total of 2,546 ancestry-informative SNPs from the Illumina array were selected based on low inter-marker correlation and ability to differentiate between samples of African and European descent. An individual was subject to filtering from the analysis if his value along eigenvector 1 or 2 was outside of 4 SDs of the mean of each respective eigenvector. We identified 108 individuals who met this criterion. Together the top 10 eigenvectors (used in the analysis) explain 21% of the global genetic variability among subjects.

3. Cardiovascular Health Study

The CHS is a population-based cohort study of risk factors for CHD and stroke in adults ≥65 years conducted across four field centers.[18] The original predominantly Caucasian cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists; subsequently, an additional predominantly African-American cohort of 687 persons were enrolled subsequently for a total sample of 5,888. DNA was extracted from blood samples drawn on all participants at their baseline examination in 1989-90 (original cohort) or 1992-93 (African American cohort). In 2010, genotyping was performed at the General Clinical Research Center's Phenotyping/Genotyping Laboratory at Cedars-Sinai using the Illumina HumanOmni1-Quad_v1 BeadChip system on 844 African-American CHS participants who consented to genetic testing, and had DNA available for genotyping. Genotyping was attempted in 844 participants, and was successful in 823 persons; the latter constitute the CHS sample for African-American genome-wide association studies. Of the 823, we excluded five who did not give consent for their DNA to be used for non-cardiovascular analyses, six who reported Hispanic ethnicity, and 11 whose baseline smoking status was unknown, leaving 801 for this analysis.

4. Candidate Gene Association Resource (CARe) CARe samples were collected from five NHLBI-funded cohort studies:

A. Atherosclerosis Risk Communities Study (ARIC): The ARIC study is a population-based, prospective cohort study of cardiovascular disease and its risk factors sponsored by National Heart, Lung and Blood Institute (NHLBI) [19]. ARIC included 15,792 individuals aged 45-64 years at baseline (1987-89), chosen by probability sampling from four US communities [20]. Cohort members completed four clinic examinations, conducted three years apart between 1987 and 1998. Follow-up for clinical events was annual. The current analysis included smoking data that was measured at baseline and the sample included 1,832 African American male and female ever smokers.
B. The Coronary Artery Risk Development in Young Adults (CARDIA) Study: The CARDIA study is a population based, prospective cohort examining the development and determinants of clinical and subclinical cardiovascular disease and its risk factors. It began in 1985-6 with a group of 5115 black and white men and women aged 18-30 years. The participants were selected so that there would be approximately the same number of people in subgroups of race, gender, education (high school or less and more than high school) and age (18-24 and 25-30) in each of 4 centers: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. These same participants were asked to participate in follow-up examinations during 1987-1988 (Year 2), 1990-1991 (Year 5), 1992-1993 (Year 7), 1995-1996 (Year 10), 2000-2001 (Year 15), and 2005-2006 (Year 20). The current analysis included smoking data that was measured at Year 15 (2000-2001) and the sample included 646 African American male and female ever smokers.
C. The Cleveland Family Study (CFS): The CFS is a family-based, longitudinal study designed to characterize the genetic and non-genetic risk factors for sleep apnea. In total, 2,534 individuals (46% African American) from 352 families were studied on up to 4 occasions over a period of 16 years (1990-2006). The initial aim of the study was to quantify the familial aggregation of sleep apnea. Over time, the aims were expanded to characterize the natural history of sleep apnea, sleep apnea outcomes, and to identify the genetic basis for sleep apnea. Data were collected over 4 exam cycles, each occurring ~every 4 years over a 16 year time interval. Data for this analysis included unrelated African American male and female ever smokers 325 at baseline.
D. Jackson Heart Study (JHS): is the largest study in history to investigate the genetic factors that affect high blood pressure, heart disease, strokes and diabetes in African Americans. The JHS is an outgrowth of the ARIC Study. In this analysis we included 1,092 male and female ever smokers from the first exam (September 2000 to March 2004).
E. The Multi-Ethnic Study of Atherosclerosis (MESA): The MESA is a study of the characteristics of subclinical cardiovascular disease and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease in a population-based sample of 6,814 asymptomatic men and women aged 45-84. Approximately 38 percent of the recruited participants are white, 28 percent African-American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. The first examination took place over two years, from July 2000-July 2002 and was followed by two 18-month examination periods and an additional two-year examination period. In this analysis we included 938 individuals from the first exam on which smoking data was available.

5. GENESTAR

GeneSTAR is a 27 year prospective family-based study of incident CAD, diabetes, stroke, and other vascular diseases in initially healthy African American and European American adult relatives of probands with documented coronary disease prior to age 60. The genotyped sample size is 3232, with 38% African American. Participants are probands and siblings of the probands, offspring of the siblings and probands, and coparents of the offspring. Persons were enrolled from 1983 to 2006 and followed at regular 5-year intervals. All are phenotyped for risk factor covariables and have biochemically validated smoking data (exhaled CO). The smoking rates among our families are higher than the general

population. All were genotyped using the Illumina 1 Mv1_C at deCODE Genetics and imputed to 2.5M snps using MACH (version 1.0.16) using the combined CEU+YRI haplotypes from MACH's website (release 21, build 36) as a reference panel. In this analysis, we included data from the most recent visit for 1175 African Americans.

6. HANDLS

The Healthy Aging in Neighborhoods of Diversity across the Life Span study (HANDLS) is an interdisciplinary, community-based, prospective longitudinal epidemiologic study examining the influences of race and socioeconomic status (SES) on the development of age-related health disparities among socioeconomically diverse African Americans and whites in Baltimore. This study investigates whether health disparities develop or persist due to differences in SES, differences in race, or their interaction. This study is unique because it will assess over 20-year period physical parameters as well as evaluate genetic, biologic, demographic, and psychosocial, parameters of African American and white participants in higher and lower SES. The study domains include: nutrition, cognition, biologic biomarkers, body composition and bone quality, psychophysiology, physical function and performance, sociodemographics, psychosocial, neighborhood environment and cardiovascular disease. Utilizing data from these study domains will facilitate understand the driving factors behind persistent black-white health disparities in overall longevity, cardiovascular disease, and cognitive decline. The mechanisms or biologic and molecular pathways through which the health and longevity trajectories of individuals in American society are influenced are unknown at this time.

The HANDLS design is an area probability sample of Baltimore based on the 2000 Census. The study protocol facilitated our ability to recruit 3722 participants from Baltimore, MD with mean age 47.7 (range 30-64) years, 54.5% males/female, 2200 African Americans (59%) and 1522 whites (41%); 41% reported household incomes below the 125% poverty delimiter. There were no significant age differences associated with sex or race. Participants below the 125% poverty delimiter were slightly younger than those above the delimiter. Age, race, and sex, but not poverty status, were associated with the likelihood of an examination. Older participants, women, and whites were more likely to complete their examinations. Among those who completed their examinations, there were no age differences associated with sex and poverty status, but African Americans were negligibly younger than whites. The study is currently conducting wave 3 designed as a re-examination wave of all participants seen between 2004-2009. This wave began in July of 2009 and will conclude in 2012.

Genotyping was focused on a subset of participants self-reporting as African American was undertaken at the Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health. In the larger genotyping effort, a small set of self-reported European ancestry samples were included. This research was supported by the Intramural Research Program of the NIH, National Institute on Aging and the National Center on Minority Health and Health Disparities.

HANDLS Genetic Data: 1024 participants were successfully genotyped to 907763 SNPS at the equivalent of Illumina 1M SNP coverage (709 samples using Illumina 1M and 1Mduo arrays, the remainder using a combination of 550K, 370K, 510S and 240S to equate the million SNP level of coverage), passing inclusion criteria into the genetic component of the study. Initial inclusion criteria for genetic data in HANDLS includes concordance between self reported sex and sex estimated from X chromosome heterogeneity, > 95% call rate per participant (across all equivalent arrays), concordance between self-reported African ancestry and ancestry confirmed by analyses of genotyped SNPs, and no cryptic relatedness to any other samples at a level of proportional sharing of genotypes > 15% (effectively

excluding 1st cousins and closer relatives from the set of probands used in analyses). In addition, SNPs were filtered for HWE p-value > 1e-7, missing by haplotype p-values > 1e-7, minor allele frequency > 0.01, and call rate > 95%. Basic genotype quality control and data management was conducted using PLINKv1.06 (PMID: 17701901). Cryptic relatedness was estimated via pairwise identity by descent analyses in PLINK and confirmed using RELPAIR (PMID: 11032786).

Ancestry estimates were assessed using both STRUCTUREv2.3 (PMID: 10835412, PMID: 12930761, PMID: 18784791) and the multidimensional scaling (MDS) function in PLINKv1.06. In the multidimensional scaling analysis, HANDLS participants were clustered with data made available from HapMap Phase 3 for the YRI, ASW, CEU, TSI, JPT and CHB populations, using a set of 36892 linkage-disequilibrium-pruned SNPs common to each population. This set of SNPs were chosen as they are not in $r^2 > 0.20$ with another SNP in overlapping sliding windows of 100 SNPs in the ASW samples. HANDLS participants with component vector estimates consistent with the HapMap ASW samples for the first 4 component vectors were included. In addition, the 1024 quality controlled HANDLS samples were later clustered among themselves using MDS to generate 10 component vectors estimating internal population structure within the HANDLS study. Of the SNPs utilized for MDS clustering, the 2000 SNPs with the most divergent allele frequency estimates between African populations (frequency estimates based on YRI samples) and European populations (frequency estimates based on combined CEU and TSI samples) were utilized as ancestry informative markers (AIMs). These 2000 AIMs were associated with frequency differences on the level of p-values < 1e-3 based on chi-squared tests. A two population model in STRUCTURE was used to estimate percent African and percent European ancestry in the HANDLS samples, for a 10000 iteration burn-in period, and a10000 iteration follow-up of the Markov Chain Monte Carlo model utilized by STRUCTURE. The ancestry estimates from STRUCTURE were highly concordant with the first component vector of the MDS clustering of HANDLS samples, with an r^2 of > 0.82.

HANDLS participant genotypes were imputed using MACHv1.0.16 (http://www.sph.umich.edu/csg/abecasis/mach/) based on combined haplotype data for HapMap Phase 2 YRI and CEU samples that includes monomorphic SNPs in either of the two constituent populations (release 22, build 36). This process followed two stages, first estimating recombination and crossover events in a random sample of 200 participants, then based on this data and the reference haplotypes, 200 iterations of the maximum likelihood model were used to estimate genotype dosages for imputed SNPs. After filtering based on a minimum imputation quality of 0.30, indicated by the RSQR estimate in MACH, with a total yield of 2939993 SNPs. Genotype clusters are available for SNPs genotyped in HANDLS upon request.

Phenotypic Data and Analysis: Missing data was an initial exclusion factor. All analyses conducted as per analysis plan. Data was analyzed using MACH2QTL v 1.08, MACH2DAT v 1.08 and R v 2.1.10. Based on analysis plan, outliers for rate and age at initiation were removed including 6 for the rate phenotype and 31 for the age phenotype. For comparisons of current versus former smokers, former smokers were treated as the control group. For comparisons of ever versus never smokers, never smokers were treated as controls.

7. HealthABC

Participants were part of the Health, Aging and Body Composition (Health ABC) study, a prospective cohort study of 3,075 community-dwelling black and white men and women living in Memphis, TN, or Pittsburgh, PA, and aged 70–79 years at recruitment in 1997. To identify potential participants, a random sample of white and all black Medicare-eligible elders, within designated zip code areas, were

contacted. To be eligible, participants had to report no difficulty with activities of daily living, walking a quarter of a mile, or climbing 10 steps without resting. They also had to be free of life-threatening cancer diagnoses and have no plans to move out of the study area for at least 3 years. The sample was approximately balanced for sex (51% women) and 42% of participants were black. Participants self-designated race/ethnicity from a fixed set of options (Asian/Pacific Islander, black/African American, white/Caucasian, Latino/Hispanic, do not know, other). The study was designed to have sufficient numbers of Blacks to allow separate estimates of the relationship of body composition to functional decline. All eligible participants signed a written informed consent, approved by the institutional review boards at the clinical sites. This study was approved by the institutional review boards of the clinical sites and the coordinating center (University of California, San Francisco). In this analysis we included 1,137 individuals with adequate quality DNA samples from the baseline assessment during which tobacco history was assessed.

Genomic DNA was extracted from buffy coat collected using PUREGENE® DNA Purification Kit during the baseline exam. Genotyping was performed by the Center for Inherited Disease Research (CIDR) using the Illumina Human1M-Duo BeadChip system. Samples were excluded from the dataset for the reasons of sample failure, genotypic sex mismatch, and first-degree relative of an included individual based on genotype data. Genotyping was successful for 1,151,215 SNPs in 1,139 unrelate African Americans. Imputation was done for the autosomes using the MACH software version 1.0.16. SNPs with minor allele frequency ≥ 1%, call rate ≥97% and HWE p≥10-6 were used for imputation. HapMap II phased haplotypes were used as reference panels. For African Americans, genotypes were available on 1,007,948 high quality SNPS for imputation based on a 1:1 mixture of the CEPH:Yoruban reference panel (release 22, build 36). A total of 1,958,375 SNPs in African Americans are available for analysis.

8. HYPERGEN

The Hypertension Genetic Epidemiology Network is a multicenter family-based study, which was created to research the genetic causes of hypertension and related conditions. HyperGEN recruited African American and Caucasian participants at five field centers, with recruitment based largely on ongoing population-based studies. The final African American GWAS dataset contained 1,258 subjects with 847,008 autosomal SNPs, and the final Caucasian GWAS dataset contained 1,270 subjects with 358,327 SNPs. In this analysis we included 1241 African Americans with adequate quality DNA samples from the baseline assessment.

9. Women's Health Initiative SNP Health Association Resource (WHI)

The Women's Health Initiative was an NHLBI-funded observational study and clinical trial focused on prevention of heart disease, breast and colorectal cancer, and osteoporosis in post-menopausal women, and enrolled 161,808 women across the U.S. between 1993-98 and followed through 2005, described in detail [21, 22]. The WHI extension study enrolled 115,400 women from the original group for follow-up through 2010. SHARe participants are those women self-reporting ethnicity as African American and provided genetic samples (n = 8,395). In this analysis we included 8,208 individuals with adequate quality DNA samples from the baseline assessment during which tobacco history was assessed.

Acknowledgements

1. AABC

This scan was supported by a Department of Defense Breast Cancer Research Program Era of Hope Scholar Award to CAH and the Norris Foundation. Each of the participating studies was supported by the following grants: MEC (National Institutes of Health grants R01-CA63464 and R37-CA54281); CARE (National Institute for Child Health and Development grant NO1-HD-3-3175), WCHS (U.S. Army Medical Research and Material Command (USAMRMC) grant DAMD-17-01-0-0334, the National Institutes of Health grant R01-CA100598, and the Breast Cancer Research Foundation), SFBC (National Institutes of Health grant R01-CA77305 and United States Army Medical Research Program grant DAMD17-96-6071), CBCS (National Institutes of Health Specialized Program of Research Excellence in Breast Cancer, grant number P50-CA58223, and Center for Environmental Health and Susceptibility, National Institute of Environmental Health Sciences, National Institutes of Health, grant number P30-ES10126), PLCO (Intramural Research Program, National Cancer Institute, National Institutes of Health), NHBS (National Institutes of Health grant R01-CA100374), and WFBC (National Institutes of Health grant R01-CA73629). NC-BCFR was supported by the National Cancer Institute, National Institutes of Health under RFA # CA-95-011 and through cooperative agreements with members of the Breast Cancer Family Registry (BCFR) and Principal Investigators, including the Cancer Prevention Institute of California (U01 CA69417); the content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating BCFR centers, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the BCFR.

2. AAPC

The MEC and the genotyping in this study were supported by NIH grants CA63464, CA54281, CA1326792, CA148085 and HG004726. Genotyping of the PLCO samples was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH. LAAPC was funded by grant 99-00524V-10258 from the Cancer Research Fund, under Interagency Agreement #97-12013 (University of California contract #98-00924V) with the Department of Health Services Cancer Research Program. Cancer incidence data for the MEC and LAAPC studies have been collected by the Los Angeles Cancer Surveillance Program of the University of Southern California with Federal funds from the NCI, NIH, Department of Health and Human Services, under Contract No. N01-PC-35139, and the California Department of Health Services as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885, and grant number 1U58DP000807-3 from the Centers for Disease Control and Prevention. MDA was support by grants, R01CA68578, DAMD W81XWH-07-1-0645, and P50-CA140388. GECAP was supported by NIH grant ES011126. CaP Genes was supported by CA88164. IPCG was support by DOD grant W81XWH-07-1-0122. DCPC was supported by NIH grant S06GM08016 and DOD grants DAMD W81XWH-07-1-0203 and DAMD W81XWH-06-1-0066. SCCS is funded by NIH grant CA092447, and SCCS sample preparation was conducted at the Epidemiology Biospecimen Core Lab that is supported in part by the Vanderbilt-Ingram Cancer Center (P30 CA68485).

3. Cardiovascular Health Study

This CHS research was supported by NHLBI contracts N01-HC-85239, N01-HC-85079 through N01-HC-85086; N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133 and NHLBI grants HL080295, HL075366, HL087652, HL105756 with additional contribution from NINDS. Additional support was provided through AG-023629, AG-15928, AG-20098, and AG-027058 from the NIA. See also http://www.chs-nhlbi.org/pi.htm. DNA handling and genotyping was supported in part by National Center for Research Resources CTSI grant UL 1RR033176 and National Institute of Diabetes and Digestive and Kidney Diseases grant DK063491 to the Southern California Diabetes Endocrinology

Research Center. Evan L. Thacker was supported by NHLBI training grant T32-HL007902.

4. Candidate Gene Association Resource (CARe)

MD Scientist Fellowship in Genetic Medicine (Northwestern Memorial Foundation; Principal Investigator: A. Hamidovic), National Research Service Award F32DA024920 (NIH/NIDA; Principal Investigator: A. Hamidovic), Dr. Bonnie Spring's Professional Account at Northwestern Feinberg School of Medicine, KL2 RR024130-02 (support for E. Jorgenson). The Candidate gene Association Resource (CARe) wishes to acknowledge the support of the National Heart, Lung and Blood Institute and the contributions of the research institutions, study investigators, field staff and study participants in creating this resource for biomedical research (NHLBI contract number HHSN268200960009C). The following eight parent studies have contributed parent study data, ancillary study data, and DNA samples through the Broad Institute (N01-HC-65226) to create this genotype/phenotype database for wide dissemination to the biomedical research community: the Atherosclerosis Risk in Communities (ARIC) study, the Cardiovascular Health Study (CHS), the Cleveland Family Study (CFS), the Coronary Artery Risk Development in Young Adults (CARDIA) study, the Framingham Heart Study (FHS), the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Sleep Heart Health Study (SHHS). The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

5. GENESTAR

GeneSTAR's work was supported by grants from the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL097698, HL59684, HL58625-01A1, HL071025-01A1, HL089474-01A1, HL092165-01A1, HL099747); by grants from the National Institutes of Health/National Institute of Nursing Research (NR0224103**,** NR008153-01); by a grant from the National Institutes of Health/National Institute of Neurological Disorders and Stroke; by a grant from the National Institutes of Health/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center; and by the Intramural Research Program of the National Institutes of Health/National Human Genome Research Institute.

6. HANDLS

HANDLS research was supported by the Intramural Research Program of the NIH, National Institute on Aging and the National Center on Minority Health and Health Disparities (contract # Z01-AG000513 and human subjects protocol # 2009-149). Data analyses for the HANDLS study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, Md. (http://biowulf.nih.gov).

7. HealthABC

This research is supported in part by the Intramural Research Program of the NIH, National Institute on Aging. This research was supported by NIA contracts N01AG62101, N01AG62103, N01AG62106 and NIA grant 1R03AG032498-01. The genome-wide association study was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences and genotyping services were provided by the Center

for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C.

8. HYPERGEN

This hypertension network is funded by cooperative agreements (U10) with NHLBI: HL54471, HL54472, HL54473, HL54495, HL54496, HL54497, HL54509, HL54515 and HL55673 (Echo). Primary Centers and Investigators of 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.; Stephen Mandel; Shamika Ketkar; Matthew Brown; Michael A. Province, Ph.D.; Ingrid Borecki, Ph.D.; Derek Morgan. 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.

9. Women's Health Initiative SNP Health Association Resource (WHI)

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. Personal funding to Sean P. David from National Institute on Drug Abuse grants DA-027331 and DA-017441 and National Institute of General Medical Sciences grant GM-061374. Andrew W. Bergen and Gary E. Swan are supported by DA-020830.

Program Office: (National Heart, Lung, and Blood Institute, Bethesda, Maryland) Jacques Rossouw, Shari Ludlam, Joan McGowan, Leslie Ford, and Nancy Geller. Clinical Coordinating Center: Clinical Coordinating Center: (Fred Hutchinson Cancer Research Center, Seattle, WA) Garnet Anderson, Ross Prentice, Andrea LaCroix, and Charles Kooperberg. Investigators and Academic Centers: (Brigham and Women's Hospital, Harvard Medical School, Boston, MA) JoAnn E. Manson; (MedStar Health Research Institute/Howard University, Washington, DC) Barbara V. Howard; (Stanford Prevention Research Center, Stanford, CA) Marcia L. Stefanick; (The Ohio State University, Columbus, OH) Rebecca Jackson; (University of Arizona, Tucson/Phoenix, AZ) Cynthia A. Thomson; (University at Buffalo, Buffalo, NY) Jean Wactawski-Wende; (University of Florida, Gainesville/Jacksonville, FL) Marian Limacher; (University of Iowa, Iowa City/Davenport, IA) Robert Wallace; (University of Pittsburgh, Pittsburgh, PA) Lewis Kuller; (Wake Forest University School of Medicine, Winston-Salem, NC) Sally Shumaker. Women's Health Initiative Memory Study: (Wake Forest University School of Medicine, Winston-Salem, NC) Sally Shumaker.

References

1. Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, Ardissino D *et al.* Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; **42**(5):441-7.

2. Kolonel LN, Henderson BE, Hankin JH, Nomura AM, Wilkens LR, Pike MC *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *American journal of epidemiology* 2000; **151**(4)**:** 346-357.

3. Newman B, Moorman PG, Millikan R, Qaqish BF, Geradts J, Aldrich TE *et al.* The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. *Breast Cancer Res Treat* 1995; **35**(1)**:** 51-60.

4. Moorman PG, Newman B, Millikan RC, Tse CK, Sandler DP. Participation rates in a case-control study: the impact of age, race, and race of interviewer. *Annals of epidemiology* 1999; **9**(3)**:** 188-195.

5. Marchbanks PA, McDonald JA, Wilson HG, Burnett NM, Daling JR, Bernstein L *et al.* The NICHD Women's Contraceptive and Reproductive Experiences Study: methods and operational results. *Annals of epidemiology* 2002; **12**(4)**:** 213-221.

6. John EM, Hopper JL, Beck JC, Knight JA, Neuhausen SL, Senie RT *et al.* The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res* 2004; **6**(4)**:** R375-389.

7. Prorok PC, Andriole GL, Bresalier RS, Buys SS, Chia D, Crawford ED *et al.* Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin Trials* 2000; **21**(6 Suppl)**:** 273S-309S.

8. Zheng W, Cai Q, Signorello LB, Long J, Hargreaves MK, Deming SL *et al.* Evaluation of 11 breast cancer susceptibility loci in African-American women. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2009; **18**(10)**:** 2761-2764.

9. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**(8)**:** 904-909.

10. Signorello LB, Hargreaves MK, Steinwandel MD, Zheng W, Cai Q, Schlundt DG *et al.* Southern community cohort study: establishing a cohort to investigate health disparities. *Journal of the National Medical Association* 2005; **97**(7)**:** 972-979.

11. Calle EE, Rodriguez C, Jacobs EJ, Almon ML, Chao A, McCullough ML *et al.* The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* 2002; **94**(9)**:** 2490-2501.

12. Ingles SA, Coetzee GA, Ross RK, Henderson BE, Kolonel LN, Crocitto L *et al.* Association of prostate cancer with vitamin D receptor haplotypes in African-Americans. *Cancer research* 1998; **58**(8)**:** 1620-1623.

13. Liu X, Plummer SJ, Nock NL, Casey G, Witte JS. Nonsteroidal antiinflammatory drugs and decreased risk of advanced prostate cancer: modification by lymphotoxin alpha. *American journal of epidemiology* 2006; **164**(10)**:** 984-989.

14. Robbins C, Torres JB, Hooker S, Bonilla C, Hernandez W, Candreva A *et al.* Confirmation study of prostate cancer risk variants at 8q24 in African Americans identifies a novel risk locus. *Genome research* 2007; **17**(12)**:** 1717-1722.

15. Agalliu I, Salinas CA, Hansten PD, Ostrander EA, Stanford JL. Statin use and risk of prostate cancer: results from a population-based epidemiologic study. *American journal of epidemiology* 2008; **168**(3)**:** 250-260.

16. Stanford JL, Wicklund KG, McKnight B, Daling JR, Brawer MK. Vasectomy and risk of prostate cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 1999; **8**(10)**:** 881-886.

17. Rybicki BA, Neslund-Dudas C, Nock NL, Schultz LR, Eklund L, Rosbolt J *et al.* Prostate cancer risk from occupational exposure to polycyclic aromatic hydrocarbons interacting with the GSTP1 Ile105Val polymorphism. *Cancer detection and prevention* 2006; **30**(5)**:** 412-422.

18. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA *et al.* The Cardiovascular Health Study: design and rationale. *Annals of epidemiology* 1991; **1**(3)**:** 263-276.

19. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *American journal of epidemiology* 1989; **129**(4)**:** 687-702.

20. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 2009; **41**(6):666-76

21. Stefanick ML, Cochrane BB, Hsia J, Barad DH, Liu JH, Johnson SR. The Women's Health Initiative postmenopausal hormone trials: overview and baseline characteristics of participants. *Annals of epidemiology* 2003; **13**(9 Suppl)**:** S78-86.

22. Hays J, Hunt JR, Hubbell FA, Anderson GL, Limacher M, Allen C *et al.* The Women's Health Initiative recruitment methods and results. *Annals of epidemiology* 2003; **13**(9 Suppl)**:** S18-77.