Int J Mol Epidemiol Genet 2012;3(4):333-360
www.ijmeg.org /ISSN:1948-1756/IJMEG1107002

Original Article

The case-only independence assumption: associations between genetic polymorphisms and smoking among controls in two population-based studies

M Elizabeth Hodgson[1], Andrew F Olshan[1,4], Kari E North[1,5], Charles L Poole[1], Donglin Zeng[2], Chiu-Kit Tse[1], Tope O Keku[3], Joseph Galanko[1], Robert Sandler[4], Robert C Millikan[1,4,6]

[1]Departments of Epidemiology and [2]Biostatistics, Gillings School of Global Public Health, [3]Center for Gastrointestinal Biology and Disease, [4]Lineberger Comprehensive Cancer Center, School of Medicine, [5]Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, North Carolina 27599; [6]Deceased

Received July 16, 2012; Accepted October 19, 2012; Epub November 15, 2012; Published November 30, 2012

Abstract: The independence assumption for a case-only analysis of statistical interaction, i. e. that genetic (G) and environmental exposures (E) are not associated in the source population, is often checked in surrogate populations. Few studies have examined G-E association in empirical data, particularly in controls from population-based studies, the type of controls expected to provide the most valid surrogate estimates of G-E association. We used controls from two population-based case-control studies to evaluate G-E independence for 43 selected genetic polymorphisms and smoking behavior. The odds ratio (OR_z) was used to estimate G-E association and, therefore, the magnitude of bias introduced into the case-only odds ratio (COR). Odds ratios of moderate magnitude [mmORz], defined as ORz≤0.7 or ORz>=1.4, were found at least one of the six smoking measures (ever, former, current, cig/day, years smoked, pack-years) for 45% and 59% of the SNPs examined in the control groups of two independently conducted North Carolina studies, respectively. Consequently, case-only estimates of G-E interaction in the context of a multiplicative benchmark would be biased for these SNPs and smoking measures. $MmOR_z$s were found more often for smoking amount than smoking status. We recommend that a stand-alone case-only study should only be conducted when G-E independence can be verified for each polymorphism and exposure metric with population-specific data. Our results suggest that OR_z is specific to each underlying population rather than an estimate of a 'universal' OR_z for that SNP and smoking measure. Further, misspecification of smoking is likely to introduce bias into the COR.

Keywords: Case-only, controls, gene-environment interaction, genetic polymorphisms, smoking

Introduction

The case-only study design as proposed by Prentice et. al. and popularized by Piegorsch and Khoury et.al. [1, 2] been used increasingly over the last 20 years to estimate the magnitude of statistical interaction between two exposures, most often gene-environment interaction (GxE) in cancer studies. Provided the independence assumption is met (i.e. that the genetic and environmental factors are independent in the population that produced the cases), the case-only study estimates the synergy index (SIM), a measure of statistical interaction based upon a multiplicative benchmark. Correlation of two exposures among cases (the case-only odds ratio, COR) is interpretable as an estimate of SIM if there no correlation of the two exposures among controls (measured by the control-only odds ratio, OR_z) [3]. Data simulations have demonstrated that small violations of the independence assumption strongly bias the case-only interaction parameter [4]. Albert et. al. varied the magnitude of control group G-E association to explore the effect of independence assumption violation on the COR. As the magnitude of gene-environment association in controls (OR_z) increased above the null, the COR was increasingly and proportionally biased away from the SIM.

Investigators conducting stand-alone case-only studies (i.e. studies with no controls from the underlying population), rather than case-only analyses with at least partial controls, such as proposed by Mukherjee et. al. in 2008 [5, 6],

must utilize data from other sources to evaluate the independence assumption. There is scant literature on empirical G-E control group associations, and few published estimates of OR_z that can be used to guide investigators who wish to conduct these case-only analyses.

However, using data from a study of *XRCC1* genotype and lung cancer, Albert et. al. showed empirically that the magnitude of OR_z equaled the magnitude of bias introduced into the COR relative to the SIM [4]. In another example, an OR_z of 1.2, representing the association between genotype and alcohol drinking status, biased the COR by nearly 30%, exceeding one commonly used threshold for an acceptable level of bias due to confounding (10%). When OR_z has a similar magnitude but opposite direction to the SIM, a case-only study may fail to detect interaction effects (Type II error) [4, 7, 8]. The case-only study has been suggested as a screening tool to identify candidate genes that may interact with environmental exposures, however, appreciable bias in the COR would invalidate such an approach. Few studies have provided empirical evidence of the extent of the potential bias in the COR.

The purpose of the current study is to more fully examine the validity of the independence assumption of the case-only design, using two population-based case-control studies to explore specific gene-smoking control group associations (OR_z). The control groups are from the Carolina Breast Cancer Study (CBCS) and the North Carolina Colon Cancer Study (NCCCS). The SNPs chosen are often used to study gene-smoking interaction and/or smoking behavior. They include SNPs for genes in the DNA repair, xenobiotic metabolism, and cell cycle control pathways. Both studies oversampled African Americans, and the NCCCS has both male and female participants, so effect measure modification and/or confounding by age, race and gender could also be addressed. Finally, all genes were grouped by the function of the gene pathway they participated in, and patterns in OR_z according to biologic pathway were evaluated.

Materials and methods

Study populations

CBCS and NCCCS: The Carolina Breast Cancer Study and the North Carolina Colon Cancer Study are population-based case-control studies conducted in central North Carolina in the mid- to late 1990's (CBCS: N_{cases}=2311, $N_{controls}$=2022; NCCCS: N_{cases}=646, $N_{controls}$=1053) [9-13]. CBCS controls were pooled from Phase I (N=790), Phase II (N=774) and the Carcinoma in situ (N=458) study. For both studies, controls were selected from NC Division of Motor Vehicles lists (<65 years of age) and Health Care Financing Administration lists (≥65 years of age), using randomized recruitment and frequency matched on age, race and gender [14]. The CBCS and NCCCS used similar questionnaires and both have extensive data on tobacco smoking.

A panel of genetic polymorphisms was chosen from available genotype data in the CBCS and NCCCS based on relevance to smoking or smoking-related health effects. Genes selected from the CBCS included xenobiotic metabolism genes (*CYP1A1, GSTM1, GSTP1, GSTT1, NAT1, NAT2, COMT*), DNA repair genes (base excision repair: *APE 148, hOGG1, MYH, XRCC1*; double strand break repair: *BRCA2, NBS1, XRCC2, XRCC3, XRCC4*; mismatch repair: *MGMT*; nucleotide excision repair: *ERCC1, ERCC6, RAD23B, XPC, XPD, XPF, XPG*), and others (*MnSOD, MPO, NQO1, CDH1, TGFB1*). NCCCS genes included: xenobiotic metabolism genes (*GSTM1, GSTT1, MEH*), DNA repair genes (base excision repair: *ADPRT, ADPRTL2, APE 148, XRCC1*; double strand break repair: *NBS1, XRCC3*; mismatch repair: *MLH1, MSH3, MSH6*; nucleotide excision repair: *RAD23B, XPC, XPD, XPF, XPG*), and *MnSOD*. Methods of collection and genotyping have been described previously [15-25].

Statistical methods

Hardy Weinberg equilibrium was tested in race-specific control groups at α=0.05 for all polymorphisms except *GSTM1, GSTT1, NAT1* and *NAT2*. Population-based controls from the CBCS and the NCCCS were used to estimate OR_z and 95% confidence intervals (CI) for gene-smoking associations via logistic regression. A model of the general form logit (G+/G-) = $\alpha + \beta_{(1)} E_1 + \beta_{(2\text{-}i)} COV_{(2\text{-}i)}$ + error (where G+= positive for genetic variant, E+=positive for the smoking behavior, COV=any additional covariates) was used for all SNPs. The dominant model was chosen to preserve power; homozygotes for the

Table 1. Characteristics of CBCS and NCCCS control groups

	Full CBCS and NCCCS				Non-African American women, 40-74 y			
	CBCS		NCCCS		CBCS		NCCCS	
	N	%	N	%	N	%	N	%
Total N	2022		1053		1107		222	
Gender								
Female	2022	100	535	50.8	1107	100	222	100
Male	0		518	49.2	0		0	
Race								
White*	1234	61.0	616	58.5	1107	100	222	100
African American	788	39.0	437	41.5	0		0	
Age at selection (years)								
Mean +/-SD	52.6 +/-11.2		66.1+/-9.5		55.1+/- 10.0		63.5+/-8.2	
Median	50		68		53		66	
Range	21-74		40-81		40-74		41-74	
Smoking behavior								
Smoking Status								
Never	1087	53.8	450	42.9	558	50.4	119	53.6
Former	547	27.1	412	39.2	344	31.1	76	34.2
Current	388	19.2	188	17.9	205	18.5	27	12.2
	2022		1050		1107		222	
Duration (years)								
<10	271	29.1	128	21.4	143	15.0	30	29.4
11-20	235	25.3	130	21.7	265	27.8	18	17.6
>20	424	45.6	340	56.9	546	57.2	54	52.9
	930		598		954		102	
Intensity (pack/day)								
<1/2	329	35.4	188	31.6	161	29.5	31	30.1
1/2 - 1	324	34.8	223	37.5	189	34.7	42	40.8
>1	277	29.8	184	30.9	195	35.8	30	29.1
	930		595		545		103	
Pack-years†								
N	925		593		542		102	
Mean +/- SD	17.5 +/-17.3		27.1+/-27		20.7+/-18.3		26.3+/-27.4	
Median	11.6		18.8		19.1		21	
Range	0.1-80		0.1-137.5		79.8		124.8	
≤35 pack-years	783	84.6	424	71.5	431	79.5	71	69.6
>35 pack-years	142	15.4	169	28.5	111	20.5	31	30.4
	925		593		542		102	

*Participants reporting non-African American race (98% white for CBCS, 98.9% white in NCCCS); †Smokers only. NOTE: CBCS=Carolina Breast Cancer Study, NCCCS=North Carolina Colon Cancer Study, SD=standard deviation, N=number of controls.

most common allele ("no variant") formed the referent group (G-) and were compared to heterozygotes plus homozygotes for the less common allele (G+, "any variant").

Smoking status was categorized as ever, former or current smoker. Three measures of smoking amount were used: duration (<10 years, 11-20 years, >20 years), intensity (<1/2 pack/day, 1/2-1 pack/day, >1 pack/day) and pack-years (PY: ≤35 PY, >35 PY). Pack-years (PYs) were derived from categorical variables used for packs/day and years smoked (pack-years were equal to the midpoint of the category for number of years smoked multiplied by the midpoint of the category for number of packs smoked/day).

Each control group was further evaluated for effect measure modification of OR_z by stratifying on race (white, African American), age (CBCS: <50y, ≥ 50y; NCCCS: <65y, >65y) and gender (NCCCS only). Age cutpoints were based on the age distributions in each study. A likelihood ratio test was performed comparing models with and without a race*smoking interaction term. Strata were not pooled when the interaction term was significant (α=0.05) for a majority of smoking measures.

Table 2. Gene variants in CBCS and NCCCS

Gene & codon/ nucleotide position	rs#	Common* allele (amino acid)	Variant* allele (amino acid)	Nucleotide common/variant	Gene name and official abbreviation†	Study
ADPRT 762	rs1136410	Val	Ala	T/C	poly (ADP-ribose) polymerase 1 [*PARP1*]	NCCCS
ADPRTL2 328‡				C/T	poly (ADP-ribose) polymerase 2 [*PARP2*]	NCCCS
APE1 148	rs1130409	Asp	Glu	T/G	APEX nuclease (multifunctional DNA repair enzyme) 1 [*APEX1*]	Both
BRCA2 intron 24	rs206340	--	--	G/A	breast cancer 2, early onset [*BRCA2*]	CBCS
BRCA2 372	rs144848	Asn	His	A/C	breast cancer 2, early onset [*BRCA2*]	CBCS
CDH1 -160	rs16260	--	--	C/A	cadherin 1, type 1, E-cadherin (epithelial) [*CDH1*]	CBCS
COMT 158[4]	rs4680	Val	Met	G/A	catechol-O-methyltransferase [*COMT*]	CBCS
*CYPIA1 M1 (CYP1A1*2A)*	rs4646903	(*1A)	(*2A)	T/C	cytochrome P450, family 1, subfamily A, polypeptide 1 [*CYP1A1*]	CBCS
*CYPIA1 M2 (CYP1A1*2C)*	rs1048943	Ile (*1A)	Val	A/G	cytochrome P450, family 1, subfamily A, polypeptide 1 [*CYP1A1*]	CBCS
*CYPIA1 M3 (CYP1A1*3)*	rs4986882	(*1A)	(*3)	T/C	cytochrome P450, family 1, subfamily A, polypeptide 1 [*CYP1A1*]	CBCS
*CYPIA1 M4 (CYP1A1*4)*	rs1799814	Thr (*1A)	Asn	C/A	cytochrome P450, family 1, subfamily A, polypeptide 1 [*CYP1A1*]	CBCS
ERCC1 nt8092	*rs3212986*	Gln	Lys	C/A	excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense sequence) [*ERCC1*]	CBCS
ERCC6 1213	rs2228527	Arg	Gly	A/G	excision repair cross-complementing rodent repair deficiency, complementation group 6 [*ERCC6*]	CBCS
ERCC6 1230	rs4253211	Arg	Pro	G/C	excision repair cross-complementing rodent repair deficiency, complementation group 6 [*ERCC6*]	CBCS
GSTM1‖		present	null		glutathione S-transferase mu 1 [*GSTM1*]	Both
*GSTP1 105***	rs1695	Ile	Val	A/C	glutathione S-transferase pi 1 [*GSTP1*]	CBCS
GSTT1‖		present	null		glutathione S-transferase theta 1 [*GSTT1*]	Both
MEH 113	rs1051740	Tyr	His	T/C	epoxide hydrolase 1, microsomal (xenobiotic) [*EPHX1*]	NCCCS
MEH 139	rs55784606	His	Tyr	C/T	epoxide hydrolase 1, microsomal (xenobiotic) [*EPHX1*]	NCCCS
MGMT 84	rs12197	Leu	Phe	C/T	O-6-methylguanine-DNA methyltransferase [*MGMT*]	CBCS
MLH1 219	*rs1799977*	Ile	Val	A/G	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [*MLH1*]	NCCCS
MNSOD 16††	rs4880	Val	Ala	T/C	superoxide dismutase 2, mitochondrial [*SOD2*]	Both
MPO-463	rs2333227	--	--	G/A	myeloperoxidase [*MPO*]	CBCS
MSH3 1036	*rs26279*	Thr	Ala	A/G	mutS homolog 3 (E. coli) [*MSH3*]	NCCCS
MSH3 940	*rs184967*	Arg	Gln	G/A	mutS homolog 3 (E. coli) [*MSH3*]	NCCCS
MSH6 39	rs1042821	Gly	Glu	G/A	mutS homolog 6 (E. coli) [*MSH6*]	NCCCS
MYH 324	rs3219489	Gln	His	G/C	mutY homolog (E. coli) [*MUTYH*]	CBCS
NAT1	rs1057126	(*10, rapid)	(Non *10)	T/A	N-acetyltransferase 1 (arylamine N-acetyltransferase) [*NAT1*]	CBCS
NAT2	Reference	(*4, rapid)	(*5, *6, *7, *14, slow)		N-acetyltransferase 2 (arylamine N-acetyltransferase) [*NAT2*]	CBCS
NBS1 185	rs1805794	Glu	Gln	G/C	Nijmegen breakage syndrome 1 (nibrin) [*NIB*]	Both
NQO1 187	rs1800566	Pro	Ser	C/T	NAD(P)H dehydrogenase, quinone 1 [*NQO1*]	CBCS
OGG1 326	rs1052133	Ser	Cys	C/G	8-oxoguanine DNA glycosylase [*OGG1*]	CBCS

POLD1 119	rs1726801	Arg	His	G/A	polymerase (DNA directed), delta 1, catalytic subunit 125kDa [*POLD1*]	NCCCS
RAD23B	rs1805329	Ala	Val	C/T	*RAD23* homolog B (S. cerevisiae) [*RAD23B*]	Both
TGFB1	rs1800470	Leu	Pro	T/C	transforming growth factor, beta 1 [*TGFB1*]	CBCS
XPC 499	rs2228000	Ala	Val	C/T	xeroderma pigmentosum, complementation group C [*XPC*]	NCCCS
XPC 939	rs2228001	Lys	Gln	A/C	xeroderma pigmentosum, complementation group C [*XPC*]	Both
XPD 312	rs1799793	Asp	Asn	G/A	excision repair cross-complementing rodent repair deficiency, complementation group 2 [*ERCC2*]	Both
XPD 751	rs13181	Lys	Gln	A/C	excision repair cross-complementing rodent repair deficiency, complementation group 2 [*ERCC2*]	Both
XPF 415	rs1800067	Arg	Gln	G/A	excision repair cross-complementing rodent repair deficiency, complementation group 4 [*ERCC4*]	Both
XPF 662	rs2020955	Ser	Pro	T/C	excision repair cross-complementing rodent repair deficiency, complementation group 4 [*ERCC4*]	CBCS
XPG 1104	rs17655	Asp	His	G/C	excision repair cross-complementing rodent repair deficiency, complementation group 5 [*ERCC5*]	Both
XRCC1 194	rs1799782	Arg	Trp	C/T	X-ray repair complementing defective repair in Chinese hamster cells 1 [*XRCC1*]	Both
XRCC1 280	rs25489	Arg	His	G/A	X-ray repair complementing defective repair in Chinese hamster cells 1 [*XRCC1*]	Both
XRCC1 399	rs25487	Arg	Gln	G/A	X-ray repair complementing defective repair in Chinese hamster cells 1 [*XRCC1*]	Both
XRCC2 188	rs3218536	Arg	His	G/A	X-ray repair complementing defective repair in Chinese hamster cells 2 [*XRCC2*]	CBCS
XRCC3 241	rs 861539	Thr	Met	C/T	X-ray repair complementing defective repair in Chinese hamster cells 3 [*XRCC3*]	Both
XRCC4-28073‡‡	rs2075685	T	G	T/G	X-ray repair complementing defective repair in Chinese hamster cells 4 [*XRCC4*]	CBCS

*Analyzed as common and variant as defined by frequency in CBCS/NCCS datasets. The less frequent allele varied by race where noted; †http://www.ncbi.nlm.nih.gov/sites/entrez (accessed 5/13/2009); ‡*ADPRTL2* 328: Less frequent nucleotide was C in African Americans, T in non-African Americans; §*COMT*: less frequent allele was Met in African Americans, Val in non-African Americans; ‖Present (referent) or null; **GSTP1*: Less frequent allele was Ile in African Americans, Val in non-African Americans; ††*MnSOD* (CBCS & NCCCS): Less frequent allele was Ala in African Americans, Val in non-African Americans; ‡‡*XRCC4* -28073: Less frequent nucleotide was G in African Americans, T in non-African Americans. NOTE: CBCS=Carolina Breast Cancer Study, NCCCS=North Carolina Colon Cancer Study, SD=standard deviation, N=number of controls, SNP=single nucleotide polymorphism, Ala=alanine, Arg=arginine, Asp=aspartic acid, Asn=asparagine , Glu=glutamic acid, Gln=glutamine, Gly=glycine, His=histidine, Ile=isoleucine, Leu=leucine, Lys=lysine, Met=methionine, Pro=proline, Phe=phenylalanine, Thr=threonine, Trp=tryptophan, Tyr=tyrosine, Ser=serine, Val=valine; C=cytosine, A=adenine, G=guanine, T=thymine.

Table 3. Gene variant-smoking status associations in the CBCS, overall and by race[*,†]

Gene pathway/SNP[‖]	Ever smokers[‡]					Current smokers[§]				
	OR$_z$[†]	NAA	AA	<50y	≥50y	**OR$_z$**	NAA	AA	<50y	≥50y
Xenobiotic metabolism[**]										
CYPIA1 M1	**1.0**	0.8	1.1	1.3	0.7	**1.0**	0.8	1.1	1.0	1.0
CYPIA1 M2	**1.8**	1.6								
CYPIA1 M3	**0.9**		1.0							
CYPIA1 M4	**1.3**	1.5				**2.5**	2.9			
GSTM1	**1.0**	1.2	0.8	1.1	1.0	**1.1**	1.0	1.1	1.3	0.7
GSTP1	**1.2**	1.4	0.8	1.2	1.2	**0.7**	0.7		0.7	0.7
GSTT1	**1.0**	0.7	1.5	0.9	1.1	**1.1**	0.9		0.9	
NAT1	**0.9**	1.1	0.8	1.0	1.0	**1.2**	1.2		1.5	
NAT2	**0.9**	1.1	0.9	1.1	1.0	**1.3**	2.1		1.5	
COMT	**0.8**	0.6	1.2	1.0	0.7	**0.9**	0.7	1.3	0.9	0.9
DNA repair										
Base excision repair										
APE1 148	**1.1**	1.3	0.9	1.3	1.0	**1.2**	1.3	1.0	1.3	1.0
hOGG1	**1.0**	1.0	1.1	1.1	1.0	**0.9**	1.0	0.8	1.0	0.8
MYH 324	**1.0**	1.0	0.8	1.0	0.9	**0.8**	0.9	0.8	0.8	0.8
XRCC1 194	**1.1**	1.0	1.3	1.2	1.1	**1.1**	0.9	1.5	1.0	1.3
XRCC1 280	**0.9**	0.9	0.9	0.7	1.1	**0.9**	0.8		0.8	
XRCC1 399	**1.0**	1.1	1.1	1.1	1.1	**1.2**	1.2	1.4	1.2	1.3
Double strand break repair										
BRCA2 24	**0.9**	0.9	0.9	0.9	0.9	**0.9**	0.8	1.1	0.9	0.9
BRCA2 372	**1.2**	1.2	1.2	1.0	1.4	**1.2**	1.1	1.2	1.0	1.4
NBS1 185	**1.2**	1.3	1.0	1.4	1.1	**1.0**	1.1	1.1	1.2	0.9
XRCC2 188	**0.9**	0.8		1.0	0.9	**0.9**	0.9		1.1	
XRCC3 241	**0.9**	0.9	1.0	1.0	0.9	**1.2**	1.2	1.2	1.2	1.2
XRCC4 -28073	**1.2**	1.2	1.3	1.5	1.0	**1.2**	1.1	1.4	1.5	1.0
Mismatch repair										
MGMT 84	**0.9**	0.9	1.0	1.0	0.9	**0.8**	0.9	0.7	0.8	0.8
Nucleotide excision repair										
ERCC1 8092	**1.0**	0.9	1.1	1.0	0.9	**1.0**	0.8	1.1	0.8	1.1
ERCC6 1213	**1.2**	1.4	1.0	1.3	1.2	**1.6**	1.7	1.4	1.6	1.5
ERCC6 1230	**0.9**	0.9	1.2	1.0	0.8	**1.1**	1.1		1.6	0.8
HRAD23B	**1.1**	1.1	1.0	1.2	1.0	**1.2**	1.3		1.1	1.3
XPC 939	**0.9**	0.9	1.0	0.9	1.0	**1.0**	1.1	0.9	1.0	1.0
XPD 312	**1.0**	1.0	1.1	1.1	1.1	**1.1**	1.1	1.0	1.0	1.2
XPD 751	**1.2**	1.2	1.1	1.1	1.3	**1.2**	1.1	1.3	1.0	1.4
XPF 415	**1.0**	1.1		1.0	1.0	**1.0**	0.9		0.7	1.2
XPF 662	**1.1**		1.2	1.0	1.4	**1.4**		1.4	1.5	1.3
XPG 1104	**0.9**	1.0	0.7	0.9	0.9	**0.8**	0.8	0.9	0.9	0.8
Cell adhesion										
CDH1	**0.8**	0.8	0.8	0.9	0.8	**0.8**	0.8	0.9	0.9	0.8
Cell growth										
TGFB1	**1.1**	1.1	1.1	1.3	0.9	**0.8**	0.8	0.9	0.9	0.7
Oxidative stress defense										
MnSOD	**1.0**	0.9	1.0	1.1	0.8	**0.9**	0.8	0.9	0.8	0.9
MPO	**1.0**	1.2	0.9	0.8	1.4	**1.0**	1.2	0.8	0.9	1.3
NQO1[††]		1.3	0.8	1.2	1.0	**1.0**	1.3	0.7	1.2	1.0

Gene pathway/SNP[‖]	Former smokers[‡]					Current smokers[‡]				
	OR$_z$	NAA	AA	<50y	≥50y	**OR$_z$**	NAA	AA	<50y	≥50y
Xenobiotic metabolism[**]										
CYPIA1 M1	**1.0**	0.8	1.1	1.5	0.6	**1.0**	0.8	1.2	1.1	0.8
CYPIA1 M2	**2.1**									
CYPIA1 M3										
CYPIA1 M4										
GSTM1	**1.0**	1.3		0.9	1.1	**1.1**	1.1	1.0	1.3	0.8
GSTP1	**1.8**	1.9		1.9	1.5	**0.8**	0.9		0.8	
GSTT1	**0.9**	0.7				**1.1**				
NAT1	**0.8**	1.0		0.7	1.1	**1.1**	1.3		1.4	

NAT2	**0.8**	0.9		0.9	0.9	**1.2**			1.4	
COMT	**0.8**	0.6	1.1	1.0	0.7	**0.9**	0.6	1.3	0.9	0.8
DNA repair										
Base excision repair										
APE1 148	**1.1**	1.2	0.8	1.2	1.0	**1.2**	1.4	1.0	1.4	1.0
hOGG1	**1.1**	1.0	1.3	1.1	1.1	**1.0**	1.0	0.8	1.0	0.9
MYH 324	**1.1**	1.1	1.0	1.2	1.0	**0.9**	0.9	0.7	0.9	0.8
XRCC1 194	**1.1**	1.1	1.2	1.3	1.0	**1.2**	1.0	1.5	1.1	1.3
XRCC1 280	**0.9**	1.0		0.7	1.2	**0.9**	0.8		0.8	
XRCC1 399	**1.0**	1.0	0.9	0.9	1.0	**1.2**	1.2	1.4	1.2	1.3
Double strand break repair										
BRCA2 24	**1.0**	1.0	0.8	1.0	0.9	**0.9**	0.8	1.0	0.9	0.9
BRCA2 372	**1.2**	1.2	1.2	1.0	1.4	**1.3**	1.2	1.3	1.0	1.5
NBS1 185	**1.2**	1.4	0.9	1.4	1.1	**1.1**	1.2	1.0	1.4	0.9
XRCC2 188	**0.9**	0.9		0.9	1.0	**0.8**	0.8		1.0	
XRCC3 241	**0.8**	0.8	0.9	0.9	0.8	**1.1**	1.1	1.2	1.2	1.1
XRCC4 -28073	**1.1**	1.2	1.1	1.3	1.0	**1.3**	1.2	1.5	1.6	1.0
Mismatch repair										
MGMT 84	**1.1**	1.0	1.2	1.2	1.0	**0.8**	0.9	0.7	0.9	0.8
Nucleotide excision repair										
ERCC1 8092	**1.0**	0.9	1.0	1.2	0.8	**1.0**	0.8	1.1	0.9	1.0
ERCC6 1213	**1.0**	1.1	0.8	1.0	1.0	**1.6**	1.8	1.3	1.6	1.5
ERCC6 1230	**0.8**	0.8		0.7	0.8	**1.0**	1.0		1.4	0.7
HRAD23B	**1.0**	1.0	1.2	1.2	0.9	**1.2**	1.3		1.1	1.3
XPC 939	**0.9**	0.8	1.1	0.8	0.9	**1.0**	1.0	0.9	0.9	1.0
XPD 312	**1.0**	1.0	1.2	1.1	1.0	**1.1**	1.1	1.1	1.0	1.2
XPD 751	**1.1**	1.2	1.0	1.1	1.2	**1.2**	1.2	1.2	1.0	1.5
XPF 415	**1.1**	1.1		1.2	0.9	**1**	0.9		0.8	1.2
XPF 662	**1.0**		1.1	0.7	1.4	**1.3**		1.4	1.3	1.5
XPG 1104	**1.0**	1.1	0.7	0.9	1.0	**0.8**	0.9	0.8	0.8	0.8
Cell adhesion										
CDH1	**0.8**	0.9	0.7	0.9	0.8	**0.8**	0.8	0.9	0.9	0.7
Cell growth										
TGFB1	**1.2**	1.3	1.2	1.6	1.0	**0.9**	0.8	1.0	1.1	0.7
Oxidative stress defense										
MnSOD	**1.1**	1.0	1.0	1.3	0.8	**0.9**	0.8	0.9	0.9	0.8
MPO	**1.0**	1.2	1.0	0.8	1.3	**1.0**	1.2	0.8	0.8	1.4
NQO1††	**1.0**	1.2	0.8	1.2	1.0	**1.0**	1.4	0.7	1.2	1.0

*Odds ratios are race and age adjusted unless stratified by race or age, respectively; †Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4; ‡Referent is never smokers for all smoking categories unless otherwise noted; §Referent is not-current smokers (former + never); ||SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; **Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; ††Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05; **Bold** = Overall OR_z.

= $OR_z \leq 0.7$

= $OR_z \geq 1.4$

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, NAA=Non African-American (98% white), AA=African American, PY=pack-years, SNP=single nucleotide polymorphism, y=years.

Based on directed acyclic graphs [26], and their status as matching factors, age (continuous), race (white or African American) and gender (NCCCS only) were included as potential confounders of the gene-smoking relationship. We also evaluated first degree family history of any cancer and total family income as confounders. Neither variable met the criteria for confounding in this study (unit change of |0.15| or more in the β coefficient estimating OR_z). Associations were characterized by magnitude of OR_z and precision of the 95% CI. An $OR_z \geq 1.4$ or <0.7 was considered a moderate magnitude association ($mmOR_z$) and evidence of non-null association. Unacceptable imprecision was defined as odds ratio estimates with confidence limit ratios >4 (CLR, upper CI limit/lower CI limit) and were excluded unless otherwise stated. SAS 9.1 was used for all modeling [27].

Agreement between the CBCS and NCCCS was assessed using a weighted kappa statistic [28].

Table 4. Gene variant-smoking duration association in the CBCS, overall and by race*,†

Gene pathway/ SNP‖	≤10 years‡					11-20 years					>20 years				
	OR_z†	NAA	AA	<50y	≥50y	**OR_z**	NAA	AA	<50y	≥50y	**OR_z**	NAA	AA	<50y	≥50y
Xenobiotic metabolism[6]															
CYPIA1 M1	**1.0**	1.2		1.3		**1.3**			1.4		**0.8**	0.7	0.9		0.6
CYPIA1 M2															
CYPIA1 M3															
CYPIA1 M4															
GSTM1	**1.1**	1.3		1.0		**1.0**	0.9		1.4		**1.0**	1.3			1.1
GSTP1	**1.9**	1.8				**1.0**	1.2		0.9		**1.0**	1.3			1.2
GSTT1						**1.0**					**1.0**				1.3
NAT1	**0.6**					**1.1**					**1.1**	1.2			
NAT2	**0.8**					**0.9**					**1.1**	1.5			1.1
COMT	**1.0**					**0.8**	0.5		0.6		**0.8**	0.7			0.8
DNA repair															
Base excision repair															
APE1 148	**1.2**	1.2	1.0	1.4	0.9	**1.1**	1.3	0.9	1.1	1.1	**1.1**	1.3	0.9	1.4	1.0
hOGG1	**1.4**	1.1	1.9	1.3	1.3	**1.1**	1.1	1.0	1.2	0.8	**0.9**	0.9	0.8	0.7	1.0
MYH 324	**1.0**	1.1	0.8	1.0	1.0	**1.0**	0.9	1.1	1.0	0.8	**1.0**	1.1	0.7	1.1	0.9
XRCC1 194	**1.4**	1.4	1.4	1.5	1.5	**0.9**	0.7		1.4		**1.1**	1.0	1.3	0.9	1.2
XRCC1 280	**0.8**	0.8				**0.9**					**1.0**	1.1			1.1
XRCC1 399	**0.8**	0.8	0.8	0.8	1.0	**1.1**	1.2	0.9	1.0	1.3	**1.2**	1.1	1.5	1.7	1.1
Double strand break repair															
BRCA2 24	**0.9**	1.1	0.7	0.9	0.9	**1.0**	0.9	1.3	1.0	0.9	**0.9**	0.8	0.9	0.8	0.9
BRCA2 372	**1.0**	1.0	1.1	0.9	1.3	**1.5**	1.7	0.9	1.2	1.6	**1.3**	1.1	1.6	1.0	1.4
NBS1 185	**1.2**	1.2	1.0	1.3	0.9	**1.3**	1.3	1.5	1.8	1.0	**1.1**	1.4	0.7	1.1	1.1
XRCC2 188	**0.9**	0.7		1.0		**0.9**	0.9				**0.9**	0.9		1.2	0.8
XRCC3 241	**0.8**	0.8	0.9	0.8	0.8	**0.8**	0.9	0.8	1.1	0.7	**1.1**	0.9	1.2	1.2	0.9
XRCC4 1394	**1.1**	1.0	1.7	1.4	1.0	**1.0**	1.1	1.1	1.7	0.7	**1.3**	1.4	1.1	1.3	1.2
Mismatch repair															
MGMT 84	**1.3**	1.3	1.1	1.1	1.4	**1.0**	0.9	1.4	1.0	1.1	**0.7**	0.8	0.6	0.9	0.7
Nucleotide excision repair															
ERCC1 8092	**1.0**	1.0	0.9	1.1	0.7	**1.1**	0.8	1.7	1.2	0.8	**0.9**	0.9	0.9	0.8	1.0
ERCC6 1213	**1.2**	1.1	1.3	1.3	1.0	**1.2**	1.3	1.0	1.2	1.1	**1.2**	1.5	0.9	1.3	1.2
ERCC6 1230	**0.8**	0.8		1.0	0.8	**0.9**	0.8		0.8	1.0	**1.0**	0.9		1.5	0.8
HRAD23B	**1.2**	1.0		1.5	0.7	**0.9**	0.8		0.8	0.9	**1.2**	1.3	0.8	1.2	1.1
XPC 939	**1.0**	1.1	0.9	0.9	1.3	**0.9**	0.7	1.2	0.9	0.8	**0.9**	0.9	0.8	0.8	0.9
XPD 312	**0.9**	1.0	0.8	0.9	0.8	**1.2**	1.1	1.5	1.3	1.2	**1.1**	1.1	1.2	1.1	1.1
XPD 751	**0.9**	1.1	0.7	0.9	1.1	**1.3**	1.3	1.3	1.3	1.2	**1.3**	1.2	1.3	1.1	1.3
XPF 415	**1.1**	1.3		1.3		**0.9**	0.9				**1.0**	1.0		1.0	1.0
XPF 662	**1.1**		1.2	0.9		**1.4**		1.4	1.0		**1.0**		1.2		1.1
XPG 1104	**0.9**	1.1	0.7	0.9	1.0	**1.0**	1.2	0.8	1.0	1.1	**0.9**	1.0	0.7	0.8	0.9
Cell adhesion															
CDH1	**0.9**	0.9	0.7	0.9	0.8	**0.7**	0.7	0.7	0.7	0.8	**0.8**	0.8	0.9	1.0	0.8
Cell growth															
TGFB1	**1.3**	1.4	1.2	1.5	1.2	**1.1**	1.1	1.0	1.2	0.9	**1.0**	0.9	1.0	1.2	0.8
Oxidative stress defense															
MnSOD	**1.0**	1.0	0.8	1.1	0.8	**1.0**	0.8	1.3	1.1	0.7	**1.0**	0.9	1.0	1.0	0.9
MPO	**0.9**	1.2	0.7	0.6	1.8	**1.1**	1.0	1.0	1.0	1.0	**1.1**	1.3	1.0	0.9	1.4
NQO1††		1.2	0.8	1.1	1.0	**1.0**	1.3	1.1	1.2	1.2	**1.1**	1.4	0.6	1.4	0.9

*Odds ratios are race and age adjusted unless stratified by race or age, respectively; †Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4; ‡Referent is never smokers for all smoking categories unless otherwise noted; ‖SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; **Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; ††Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05; **Bold** = Overall OR_z.

= OR_z ≤0.7

= OR_z ≥1.4

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, NAA=Non African-American (98% white), AA=African American, PY=pack-years, SNP=single nucleotide polymorphism, y=years.

Table 5. Gene variant-smoking intensity association in the CBCS, overall and by race[*,†]

Gene pathway/SNP[‖]	<1/2 pack/day[‡]					1/2 - 1 pack/day					>1 pack/day				
	OR_z[2]	NAA	AA	<50y	≥50y	**OR_z**	NAA	AA	<50y	≥50y	**OR_z**	NAA	AA	<50y	≥50y
Xenobiotic metabolism[**]															
CYPIA1 M1	**0.9**	0.8	1.0	1.3		**1.0**		1.3	1.2		**1.0**	1.0			
CYPIA1 M2															
CYPIA1 M3															
CYPIA1 M4															
GSTM1	**0.9**	1.0		1.2		**0.9**	1.1		1.0		**1.3**	1.4		1.1	1.5
GSTP1	**1.2**	1.5		1.4		**1.8**	2.1				**0.9**	0.9			
GSTT1	**1.3**					**1.1**									
NAT1	**0.8**					**1.2**					**0.9**	1.1			
NAT2	**0.8**					**1.0**					**1.1**	1.5			
COMT	**0.9**	0.5	1.5	1.3	0.6	**1.2**			0.9		**0.6**	0.5			
DNA repair															
Base excision repair															
APE1 148	**1.3**	1.3	1.3	1.8	1.0	**0.9**	1.2	0.6	1.2	0.8	**1.2**	1.4	0.8	1.0	1.4
hOGG1	**0.9**	0.8	1.1	1.1	0.8	**1.1**	1.1	1.1	1.1	1.1	**1.1**	1.0	1.0	1.0	1.0
MYH 324	**1.0**	1.0	0.9	1.0	0.9	**0.9**	1.0	0.7	1.0	0.8	**1.0**	1.0	1.3	1.1	1.0
XRCC1 194	**1.2**	1.4	0.9	1.4	1.0	**1.0**	0.6	1.9	1.1	0.8	**1.2**	1.2		1.2	1.4
XRCC1 280	**0.9**	0.9			1.4	**1.0**	0.8		0.8		**0.8**	0.9			
XRCC1 399	**0.9**	1.0	0.9	0.8	1.1	**1.1**	1.2	1.3	1.2	1.2	**1.1**	1.1	1.6	1.4	1.0
Double strand break repair															
BRCA2 24	**0.8**	1.0	0.7	0.8	0.9	**1.0**	0.9	1.3	1.1	0.9	**0.9**	0.9	1.0	0.8	1.0
BRCA2 372	**1.1**	0.9	1.5	1.1	1.1	**1.2**	1.2	1.0	0.8	1.5	**1.6**	1.6		1.2	1.8
NBS1 185	**1.1**	1.2	1.0	1.2	1.0	**1.2**	1.4	1.0	1.2	1.2	**1.2**	1.4	1.0	1.7	1.0
XRCC2 188	**0.8**	0.7		1.0		**0.8**	0.7		0.8	0.8	**1.1**	1.1		1.1	1.2
XRCC3 241	**0.9**	0.7	1.2	1.0	0.8	**0.9**	1.1	0.8	1.1	0.8	**0.8**	0.9	0.8	0.8	0.9
XRCC4 1394	**1.1**	0.9	1.4	1.5	0.9	**1.2**	1.1	1.6	1.8	1.0	**1.2**	1.5	0.7	1.2	1.3
Mismatch repair															
MGMT 84	**1.1**	1.1	1.1	1.3	0.9	**0.8**	0.9	0.7	0.6	1.1	**0.9**	0.8	1.2	1.1	0.8
Nucleotide excision repair															
ERCC1 8092	**1.0**	0.9	1.0	1.2	0.7	**1.0**	1.1	1.0	1.0	1.1	**0.9**	0.7	1.5	0.8	0.9
ERCC6 1213	**1.3**	1.3	1.2	1.7	1.0	**1.3**	1.4	1.3	1.3	1.5	**1.0**	1.3		1.0	1.1
ERCC6 1230	**1.0**	1.0		1.1	1.0	**1.0**	0.9		1.2	0.8	**0.8**	0.8		0.9	0.7
HRAD23B	**1.0**	1.0	1.0	1.3	0.8	**1.1**	1.1		1.1	1.0	**1.2**	1.2		1.1	1.2
XPC 939	**1.1**	1.0	1.2	1.1	1.1	**0.8**	0.8	0.8	0.8	0.8	**0.9**	0.9	0.9	0.8	1.1
XPD 312	**1.0**	1.0	1.0	1.1	0.9	**1.1**	1.1	1.1	1.2	1.1	**1.1**	1.0	1.6	0.9	1.2
XPD 751	**1.1**	1.3	1.1	1.0	1.4	**1.2**	1.2	1.2	1.2	1.1	**1.2**	1.3	1.0	1.0	1.3
XPF 415	**0.9**	0.9		1.2		**1.0**	0.9		0.9	1.0	**1.1**	1.3		0.9	1.5
XPF 662	**1.4**		1.6	1.5	1.7	**0.9**		0.9	0.7	1.3	**0.9**		1.0		
XPG 1104	**0.9**	1.0	0.8	1.1	0.8	**0.9**	1.1	0.6	0.8	1.0	**0.9**	1.0	0.7	0.8	1.1
Cell adhesion															
CDH1	**0.9**	1.0	0.8	1.0	0.8	**0.8**	0.8	0.8	0.9	0.8	**0.7**	0.7		0.7	0.7
Cell growth															
TGFB1	**1.2**	1.2	1.1	1.4	1.1	**1.1**	1.1	1.0	1.3	0.9	**0.9**	0.9	1.0	1.2	0.8
Oxidative stress defense															
MnSOD	**1.0**	0.9	0.9	1.0	0.8	**1.0**	0.8	1.1	1.0	0.9	**1.0**	1.0	0.8	1.2	0.8
MPO	**1.0**	1.4	0.8	0.7	1.5	**1.0**	1.1	1.0	0.8	1.4	**1.1**	1.1	1.2	1.0	1.2
NQO1[††]		1.2	0.8	1.1	0.9	**1.2**	1.5	1.0	1.5	1.1	**0.9**	1.2		1.0	1.0

[*]Odds ratios are race and age adjusted unless stratified by race or age, respectively; [†]Odds ratio not displayed if 95% CI width (upper limit/lower limit) >4; [‡]Referent is never smokers for all smoking categories unless otherwise noted; [‖]SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; [**]Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; [††]Could not be pooled. LRT p-value for race*smoking interaction term <0.05; **Bold** = Overall OR_z.

= OR_z ≤0.7

= OR_z ≥1.4

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, NAA=Non African-American (98% white), AA=African American, PY=pack-years, SNP=single nucleotide polymorphism, y=years.

Table 6. Gene variant-PY association in the CBCS, overall and by race[*,†]

	≤35 PY					>35PY				
Gene pathway/NP[‖]	**OR$_z$**[†]	NAA	AA	<50y	≥50y	**OR$_z$**	NAA	AA	<50y	≥50y
Xenobiotic metabolism[**]										
CYPIA1 M1	1.0	0.8	1.1	1.3	0.6					
CYPIA1 M2	**1.6**									
CYPIA1 M3	**1.0**		1.0							
CYPIA1 M4							0.6			0.8
GSTM1	**0.9**	1.0	0.8	1.0	0.8	**1.7**				
GSTP1	**1.4**	1.6	0.8	1.2	1.4	**0.7**				
GSTT1	**1.0**	0.8	1.4	0.8	1.4		1.1			1.0
NAT1	**0.9**	1.1	0.9	1.1	1.0					
NAT2	**0.9**	1.0	1.0	1.0	0.9		1.2		1.2	1.0
COMT	**0.9**	0.6	1.3	1.0	0.9	**0.5**				
DNA repair										
Base excision repair										
APE1 148	**1.1**	1.2	1.0	1.3	1.0	**1.3**	1.8			1.1
hOGG1	**1.1**	1.0	1.2	1.1	1.0	**0.9**	1.1			1.0
MYH 324	**1.0**	1.0	0.8	1.0	0.9	**1.0**	1.1		1.3	1.0
XRCC1 194	**1.1**	0.9	1.3	1.3	0.9	**1.6**	1.5			1.7
XRCC1 280	**0.9**	0.8	1.0	0.6	1.3					
XRCC1 399	**1.0**	1.1	1.0	1.0	1.1	**1.0**	0.9			0.9
Double strand break repair										
BRCA2 24	**0.9**	1.0	0.9	0.9	0.9	**0.8**	0.7			0.9
BRCA2 372	**1.2**	1.2	1.2	1.0	1.3	**1.6**	1.5			2.0
NBS1 185	**1.2**	1.4	1.0	1.4	1.1	**1.0**	1.2		1.2	1.0
XRCC2 188	**0.9**	0.8		0.9	0.8	**1.1**	1.1			
XRCC3 241	**0.9**	0.9	1.0	1.1	0.8	**0.8**	0.8			0.9
XRCC4 1394	**1.1**	1.1	1.3	1.4	1.0	**1.5**	1.6			1.3
Mismatch repair										
MGMT 84	**1.0**	1.0	1.0	1.0	1.0	**0.5**	0.6			0.6
Nucleotide excision repair										
ERCC1 8092	**1.0**	1.0	1.1	1.1	0.9	**0.7**	0.6			0.8
ERCC6 1213	**1.3**	1.4	1.1	1.3	1.3	**0.8**	1.1			0.8
ERCC6 1230	**0.9**	0.9		1.1	0.9	**0.7**	0.7			0.6
HRAD23B	**1.1**	1.1	1.0	1.2	1.0	**1.2**	1.3			1.1
XPC 939	**0.9**	0.8	0.9	0.9	0.9	**1.1**	1.2			1.2
XPD 312	**1.1**	1.1	1.1	1.1	1.1	**0.9**	0.9			1.0
XPD 751	**1.2**	1.3	1.1	1.1	1.3	**1.1**	1.1		0.8	1.2
XPF 415	**1.0**	1.0		1.1	0.9	**1.0**	1.2			1.4
XPF 662	**1.2**		1.3	1.1	1.5					
XPG 1104	**0.9**	1.0	0.7	0.9	0.9	**1.2**	1.5			1.2
Cell adhesion										
CDH1	**0.8**	0.8	0.8	0.8	0.8	**0.9**	0.9		1.4	0.8
Cell growth										
TGFB1	**1.1**	1.1	1.0	1.3	1.0	**0.8**	0.9			0.8
Oxidative stress defense										
MnSOD	**1.0**	0.8	1.0	1.0	0.8	**1.4**	1.5			1.1
MPO	**1.0**	1.2	0.9	0.8	1.4	**1.1**	1.1		1.3	1.1
NQO1[††]		1.3	0.8	1.2	1.0	**1.1**	1.4			0.9

[*]Odds ratios are race and age adjusted unless stratified by race or age, respectively; [†]Odds ratio not displayed if 95% CI width (upper limit/lower limit) >4; [‡]Referent is never smokers for all smoking categories unless otherwise noted; [§]Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day; [‖]SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; [**]Primary functional category; gene may function in additional pathways e. g. *COMT* in estrogen metabolism; [††]Could not be pooled. LRT p-value for race[*]smoking interaction term <0.05; **Bold** = Overall OR$_z$.

= OR$_z$ ≤0.7

= OR$_z$ ≥1.4

Abbreviations: OR$_z$=odds ratio in controls, CI=confidence interval, NAA=Non African-American (98% white), AA=African American, PY=pack-years, SNP=single nucleotide polymorphism, y=years.

The weighted kappa measures agreement beyond what would be expected due to chance alone using a multi-level ordinal scale to categorize a series of subjects. The categories used for OR_z were: a) $OR_z<0.9$, b) $0.9\leq OR_z\leq 1.1$ and c) $OR_z>1.1$. Because race, age and gender distributions differed across the two studies, datasets restricted to white women 40-74 years of age were also compared. Due to reduced sample size in the restricted NCCCS dataset, precision requirements were relaxed (OR_z with CLR <5 were included) to provide sufficient estimates for comparison of most polymorphisms. SAS 9.1 was used to calculate weighted kappa statistics [27].

We examined the effects of exposure misspecification on OR_z using different metrics for smoking classification. If the exposure measurement used to evaluate the independence assumption was too crude (e.g. ever/never smoking rather than dose, duration or PY), a sizable OR_z could be missed. We also compared using the p-value for OR_z when evaluating the independence assumption to using magnitude and precision (95%CI) of the estimate as a decision tool.

Results

Characteristics of the two study populations are presented in **Table 1**. Controls from the NCCCS were older than the CBCS and included both men and women. Consistent with gender and age differences in smoking prevalence in the US [29, 30], there was a higher proportion of never smokers and a shorter average smoking duration in CBCS controls compared to the NCCCS.

In the CBCS, 38 polymorphisms in 29 genes were evaluated; 17 were DNA repair genes. In the NCCCS, 25 polymorphisms and four haplotypes from 19 genes were evaluated. Fifteen genes were DNA repair genes. **Table 2** provides the rs# and official name for each SNP.

Four SNPs (3%) were out of Hardy Weinberg equilibrium (α=0.05), two in CBCS controls (*CYP1A1* in non-African Americans, *XRCC3 241* in African Americans) and two in NCCCS controls (*RAD23B* and *XPF* 415 in non-African Americans) approximately what one would expect by chance alone. Percent 'any variant' for all loci, with the exception of *GSTT1,* was consistent between the CBCS and NCCCS within race. Allele frequencies and HWE p-values for CBCS and NCCCS controls, stratified by race, are available on request.

Tables 3-**6** and **Tables 7**-**11** present overall and race-, age- and gender-stratified OR_z for CBCS and NCCCS, respectively. All results are adjusted for race, age [continuous] and gender unless stratified by that same factor.

All models were adjusted for matching variables (race, age and gender) unless stratified or restricted by these factors. Approximately half of the polymorphisms in the CBCS showed joint confounding by race and age (difference of |>0.15| in β coefficients), almost entirely in former smoking and/or >35 PY. Confounding by race and age were more marked in measures of smoking amount than smoking status. Unadjusted and race-, age-, and gender-adjusted estimates did not vary substantially in the NCCCS. Family history of any cancer and family income were not confounders in either dataset.

CBCS

Xenobiotic metabolizing genes were slightly overrepresented among the SNPs showing moderate associations with smoking behavior (range: 0.5 - 2.5). DNA repair genes were overrepresented among the weaker associations (0.7-1.6). Among the metabolism genes, *CYP1A1* M2 was associated with smoking status and <35 PY (vs. never). *COMT* was inversely associated with high intensity and >35 PY of smoking, but not with any measures of smoking status. Among DNA repair genes, *XPF* 662, *XRCC1* 194, *BRCA2* 372, and *MGMT* 84 showed associations with smoking behavior, particularly for smoking amount (duration, dose or PY). Of the 21 evaluable DNA repair genes, six were associated with high PY, with four of them (*ERCC6* 1230, *ERCC1* 8092 and *XRCC4* -28073, *MnSOD*) associated only with PY but not smoking status, duration or intensity. In the CBCS, three SNPs showed consistency across smoking categories with $mmOR_z$s in at least one smoking status category and at least one smoking dose category: *CYP1A1* M2*, GSTP1*, and *XPF* 662 (**Tables 3**-**6**).

Within smoking categories, one SNP showed a $mmOR_z$ for ever smoking (*CYP1A1* M2); one

Table 7. Gene variant-smoking status associations in the NCCCS, overall and by gender and race[*,†]

Gene pathway/ Gene variant[‖]	Ever							Current vs. not current						
	OR_z[†]	W	M	NAA	AA	<65y	≥65y	OR_z	W	M	NAA	AA	<65y	≥65y
Xenobiotic metabolism[**]														
GST hap C[††]	**1.2**	1.3	1.1	1.0	1.7	1.4	1.1	**1.0**	1.4	0.9	0.7	1.5	1.0	1.1
GST hap A[‡‡]	**1.4**	1.6	1.2	0.9	2.2	1.8	1.1	**1.5**		1.4		2.3	1.5	1.4
GST hap B[‡‡]	**0.8**	0.8	0.6	0.7			0.7	**0.6**						
GST hap D[‡‡]	**1.3**	1.3	1.3	1.2	1.7	1.6	1.1	**0.9**		0.7	0.8		1.0	0.8
GSTM1	**1.0**	1.0	1.0	1.0	1.1	1.0	1.0	**0.7**	1.1	0.5	0.8	0.7	0.7	0.8
GSTT1[§§]		1.0	0.9	0.7	1.5	1.0	0.9		1.1	1.1	0.7	1.8	1.0	1.3
MEH 113	**0.7**	0.7	0.8	0.8	0.7	0.7	0.9	**0.8**	0.6	0.6	0.7	0.5	0.5	0.8
MEH 139	**0.8**	1.3	1.0	1.0	1.5	1.1	1.2	**0.6**	1.0	0.7	0.8		0.8	0.9
DNA repair														
POLD1 119	**0.7**	1.2	0.9	1.1		1.3	0.9	**0.8**	1.1	0.8	1.1		1.2	0.7
Base excision repair														
ADPRT 762	**1.1**	1.3	0.9	1.1		1.1	1.1	**1.2**	1.8	0.9	1.1		1.3	1.1
ADPRTL2 328	**1.1**	1.2	1.0	1.2		1.1	1.1	**1.1**		1.0	1.2		1.0	1.1
APE1 148	**1.1**	1.3	1.0	1.2	1.0	1.1	1.1	**1.0**	1.2	0.9	1.2	0.9	0.8	1.3
XRCC1 194	**0.8**	0.6		0.8	0.9		0.8	**0.9**						
XRCC1 280	**1.3**			1.2										
XRCC1 399	**1.1**	1.3	0.9	1.1	1.0	0.8	1.3	**1.0**	0.8	1.1	0.9	1.0	0.7	1.3
Double strand break repair														
NBS1 185	**0.9**	0.7	0.7	0.7	0.7	0.6	0.8	**0.8**	0.7	0.9	1.4	0.6	0.9	0.7
XRCC3 241	**0.9**	0.7	1.1	0.8	1.1	0.9	0.9	**1.1**	1.2	1.1	1.2	1.1	1.2	1.1
Mismatch repair														
MLH1 219	**1.1**	0.7	1.3	0.8	1.3	1.6	0.8	**0.8**	0.9	1.3	0.9	1.4	1.2	1.1
MSH3 1036	**1.1**	1.8	0.8	1.2	1.3	1.4	1.1	**1.0**	0.9	0.6	0.6		1.0	-1.0
MSH3 940	**1.2**	1.1	0.7	0.9	0.8	0.8	0.9	**0.7**	0.8	0.6	0.6	0.7	0.8	0.6
MSH6 39	**0.9**	0.8	1.0	0.9	0.9	1.0	0.8	**0.7**	0.8	0.8	0.6	1.2	0.9	0.8
Nucleotide excision repair														
RAD23B	**1.1**	0.6	0.9	0.7	0.9	0.9	0.7	**1.0**	0.4	1.0	0.8	0.7	0.8	0.7
XPC 499	**0.8**	0.9	0.8	0.8	1.3	1.1	0.8	**1.1**	1.0	1.1	0.9		1.1	1.0
XPC 939	**1.2**	1.3	1.1	1.3	1.1	1.0	1.3	**1.0**	1.2	0.8	1.1	0.8	0.8	1.1
XPD 312	**1.0**	1.1	0.9	1.0	1.0	1.1	1.0	**0.9**	0.9	0.8	0.9		1.1	0.7
XPD 751	**1.2**	1.4	1.1	1.2	1.2	1.4	1.2	**1.0**	1.1	0.9	0.8	1.3	1.1	0.9
XPF 415	**1.0**		1.7	1.1			1.1	**1.3**		1.5	1.7			
XPG 1104	**1.0**	1.1	0.8	0.9	1.0	1.4	0.8	**1.2**	1.2	1.2	1.1	1.4	1.7	0.9
Oxidative stress defense														
MNSOD	**1.0**	1.2	0.9	1.1	1.0	1.2	0.9	**1.1**	0.8	1.1	1.1	0.9	0.9	1.1

[*]Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively; [†]Odds ratio not displayed if confidence limit ratio >4; [‡]Referent is never smokers for all smoking categories unless otherwise noted; [§]Referent is not-current smokers (former + never); [‖]SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; [**]Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; [††]*GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined; [‡‡]*GST* hap A=*GSTT1* null & *GSTM1* present, *GST* hap B=*GSTT1* null & *GSTM1* null, *GST* hap D=*GSTT1* present & *GSTM1* null; *GST* hap C is referent; [***]Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day; [§§]Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05; **Bold** = Overall OR_z.

= OR_z ≤0.7

= OR_z ≥1.4

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, W=Women, M=Men, NAA=Non African-American (98% white), AA=African American, PY=pack-years, y=years.

Table 8. Gene variant-smoking status associations in the NCCCS, overall and by gender and race[*,†]

	Former							Current						
Gene pathway/ Gene variant[‖]	**OR_z**[†]	W	M	NAA	AA	<65y	≥65y	**OR_z**	W	M	NAA	AA	<65y	≥65y
Xenobiotic metabolism [**]														
GST hap C[††]	**1.3**	1.2	1.2	1.0	1.6	1.5	1.1	**1.2**	1.5	1.0	0.8	1.9	1.2	1.1
GST hap A[‡‡]	**1.3**	1.5	1.1	0.9	1.7		1.0	**1.7**						
GST hap B[‡‡]	**0.9**			0.8			0.7							
GST hap D[‡‡]	**1.4**	1.3	1.5	1.2		1.8	1.2	**1.1**		0.9	0.9			
GSTM1	**1.2**	1.0	1.2	1.1	1.3	1.2	1.1	**0.8**	1.1	0.6	0.8		0.8	0.8
GSTT1	**0.9**	1.0	0.8	0.8	1.2	1.0	0.8		1.1	1.0	0.6	2.0	1.0	1.1
MEH 113	**0.8**	0.8	0.9	0.8	0.9	0.8	0.9	**0.7**	0.6	0.6	0.6	0.5	0.4	0.7
MEH 139	**0.9**	1.3	1.1	1.1		1.3	1.2	**0.6**	1.1	0.8	0.8		0.9	1.0
DNA repair														
POLD1 119	**0.7**	1.2	1.0	1.1		1.3	1.0	**0.7**		0.8	1.2		1.3	
Base excision repair														
ADPRT 762	**1.0**	1.1	0.9	1.1		1.0	1.1	**1.2**		0.8	1.1		1.2	
ADPRTL2 328	**1.1**	1.1	1.0	1.2		1.1	1.1	**1.1**		1.0	1.3		1.1	1.2
APE1 148	**1.1**	1.3	1.0	1.2	1.0	1.3	1.0	**1.1**	1.3	0.9	1.3	0.9	0.9	1.3
XRCC1 194	**0.8**	0.6		0.7			0.8	**0.8**						
XRCC1 280	**1.3**			1.3										
XRCC1 399	**1.1**	1.5	0.8	1.2	0.9	0.9	1.3	**1.0**	0.9	1.0	1.0		0.6	1.5
Double strand break repair														
NBS1 185	**0.9**	0.7	0.8	0.6	0.8	0.6	0.8	**0.8**		0.7	1.1	0.5	0.7	0.7
XRCC3 241	**0.8**	0.7	1.1	0.7	1.1	0.8	0.9	**1.0**	1.0	1.1	1.0	1.2	1.1	1.0
Mismatch repair														
MLH1 219	**1.2**	0.7	1.3	0.8	1.2	1.6	0.7	**0.9**	0.8	1.5	0.8	1.5	1.5	0.9
MSH3 1036	**1.1**	2.1	0.9	1.4	1.5	1.5	1.3	**1.0**			0.7			
MSH3 940	**1.4**	1.2	0.8	1.0	0.8	0.8	1.0	**0.8**	0.8	0.6	0.7	0.7	0.7	0.6
MSH6 39	**1.0**	0.8	1.0	1.0	0.8	1.1	0.8	**0.7**	0.8	0.9	0.6	1.1	0.9	0.8
Nucleotide excision repair														
RAD23B	**1.1**	0.7	0.9	0.7	1.0	1.0	0.7	**1.0**	0.4	1.0	0.7	0.7	0.8	0.6
XPC 499	**0.8**	0.9	0.7	0.7		1.0	0.7	**1.0**	0.9	0.9	0.8		1.1	0.9
XPC 939	**1.3**	1.3	1.2	1.3	1.2	1.1	1.3	**1.1**	1.4	0.9	1.3	0.9	0.9	1.2
XPD 312	**1.1**	1.2	0.9	1.0	1.1	1.1	1.1	**0.9**	1.0	0.8	0.9		1.2	0.7
XPD 751	**1.3**	1.5	1.1	1.3	1.1	1.4	1.2	**1.1**	1.3	0.9	1.0	1.3	1.3	1.0
XPF 415	**0.9**			0.9			1.1	**1.2**			1.6			
XPG 1104	**0.9**	1.1	0.8	0.9	0.9	1.1	0.8	**1.2**	1.2	1.0	1.0	1.4	1.8	0.8
Oxidative stress defense														
MNSOD	**1.0**	1.4	0.8	1.1	1.0	1.4	0.9	**1.1**	0.9	1.0	1.1	0.9	1.0	1.0

[*]Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively; [†]Odds ratio not displayed if confidence limit ratio >4; [‡]Referent is never smokers for all smoking categories unless otherwise noted; [‖]SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; [**]Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; [††]*GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined; [‡‡]*GST* hap A=*GSTT1* null & *GSTM1* present, *GST* hap B=*GSTT1* null & *GSTM1* null, *GST* hap D=*GSTT1* present & *GSTM1* null; *GST* hap C is referent; [***]Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day; [§§]Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05; **Bold** = Overall OR_z.

= OR_z <=0.7

= OR_z >=1.4

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, W=Women, M=Men, NAA=Non African-American (98% white), AA=African American, PY=pack-years, y=years.

Table 9. Gene variant-smoking duration association in the NCCCS, overall and by gender and race[*,†]

Gene pathway/ Gene variant[‖]	<10y OR$_z$[†]	W	M	NAA	AA	Y	O	11-20y OR$_z$	W	M	NAA	AA	Y	O	>20 y OR$_z$	W	M	NAA	AA	Y	O
Xenobiotic metabolism[**]																					
GST hap C[††]	**1.5**	1.6	1.4	1.3		1.2	**1.1**	0.9	1.0		1.1	1.0	**1.2**	1.2	1.1	0.8		1.1	1.0
GST hap A[‡‡]	**1.6**						**1.3**		-1.0				**1.4**	1.6	1.2	0.7			
GST hap B[‡‡]									-1.0				**0.7**			0.6			
GST hap D[‡‡]	**1.7**			1.6			**1.1**		1.1				**1.2**	1.3	1.2	1.0			
GSTM1	**1.3**	1.2	1.3	1.2		1.1	**0.9**	0.8	1.0		0.8	1.0	**1.0**	1.0	1.0	1.0		0.8	1.0
GSTT1	**1.0**	1.3	0.8	0.8		0.9	**1.0**	1.0	0.9		1.1	0.9	**0.9**	1.0	0.9	0.6		1.1	0.9
MEH 113	**0.7**	1.1	1.0	1.0		1.4	**0.7**	0.9	0.8		0.9	0.7	**0.8**	0.6	0.7	0.7		0.9	0.7
MEH 139	**1.0**	1.2	1.4	1.4		1.3	**0.8**	0.8	1.0			1.5	**0.7**	1.3	0.9	0.9			1.5
DNA repair																					
POLD1 119	**0.7**			1.2			**0.9**		0.9				**0.7**	1.2	1.0	1.2			
Base excision repair																					
ADPRT 762	**1.2**			1.4			**0.7**		0.7				**1.2**	1.5	1.0	1.1			
ADPRTL2 328	**1.0**		0.7	1.1		0.9	**1.0**	1.2	1.2				**1.2**	1.3	1.0	1.3			
APE1 148	**1.4**		0.8	1.3		1.3	**1.1**	1.8	1.2		1.0	1.2	**1.0**	1.3	0.8	1.2		1.0	1.2
XRCC1 194													**0.9**			0.8			
XRCC1 280	**1.0**												**1.2**						
XRCC1 399	**1.3**	1.1	1.3	1.3		1.7	**1.2**	0.9	1.2			1.8	**1.0**	1.2	0.8	1.0			1.8
Double strand break repair																					
NBS1 185	**0.9**						**1.2**	0.9					**0.8**	0.6	0.7	0.8			
XRCC3 241	**0.7**	0.5	1.1	0.7		0.7	**0.7**	0.7	0.6		1.0	0.5	**1.0**	0.9	1.3	0.9		1.0	0.5
Mismatch repair																					
MLH1 219	**1.4**		1.1	0.8		0.9	**1.0**		1.0			1.0	**1.1**	0.6	1.2	0.8			1.0
MSH3 1036	**1.0**			1.2		1.2	**1.1**		1.2			1.5	**1.0**	1.4	0.8	1.2			1.5
MSH3 940	**1.2**	1.0		1.0		0.7	**1.6**	1.0	0.9			1.3	**1.1**	1.2	0.7	0.9			1.3
MSH6 39	**0.8**	0.9	1.0	1.0		.9	1.0	**0.9**	1.2	1.1		1.7	1.0	**0.9**	0.6	0.9	0.8		1.7	1.0
Nucleotide excision repair																					
RAD23B	**1.2**	0.5	1.0	0.8		1.1	0.5	**0.8**	1.0	0.7		0.9	0.7	**1.1**	0.7	0.9	0.7		0.9	0.7
XPC 499	**0.8**	1.0	0.7	0.7			0.6	**0.9**	0.7	0.7			1.3	**0.9**	0.8	0.9	0.8			1.3
XPC 939	**1.0**	0.9	1.1	0.9		0.7	1.3	**1.1**	1.2	1.2	1.0	1.0	1.1	**1.3**	1.8	1.1	1.6	1.0	1.0	1.1
XPD 312	**1.3**	1.1	1.3	1.2			1.3	**1.0**	0.9	1.0			0.8	**0.9**	1.1	0.8	0.9		-	0.8
XPD 751	**1.5**	1.9	1.2	1.6			1.4	**1.5**	1.3	1.8		1.3	1.8	**1.0**	1.2	0.9	1.0		1.3	1.8
XPF 415	**1.0**		1.0				1.0							**1.1**	1.0		1.2			
XPG 1104	**0.7**		0.8	0.8			0.7	**0.8**	0.6	0.6		1.3	0.5	**1.2**	1.5	1.0	1.1		1.3	0.5
Oxidative stress defense																					
MNSOD	**1.0**	1.2	0.8	1.1		1.2	0.9	**1.4**	0.8	1.1		1.3	0.9	**0.9**	1.1	0.9	1.1		1.3	0.9

[*]Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively; [†]Odds ratio not displayed if confidence limit ratio >4; [‡]Referent is never smokers for all smoking categories unless otherwise noted; [‖]SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; [**]Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; [††]*GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined; [‡‡]*GST* hap A=*GSTT1* null & *GSTM1* present, *GST* hap B=*GSTT1* null & *GSTM1* null, *GST* hap D=*GSTT1* present & *GSTM1* null; *GST* hap C is referent; [***]Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day; [§§]Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05; **Bold** = Overall OR$_z$.

= OR$_z$ ≤0.7

= OR$_z$ ≥1.4

Abbreviations: OR$_z$=odds ratio in controls, CI=confidence interval, W=Women, M=Men, NAA=Non African-American (98% white), AA=African American, PY=pack-years, Y: <65 years of age, O: ≥65 years of age.

Table 10. Gene variant-smoking intensity association in the NCCCS, overall and by gender and race[*,†]

Gene pathway/ Gene variant[||]	<1/2 pack/day							1/2 - 1 pack/day							>1 pack/day						
	OR_z[†]	W	M	NAA	AA	Y	O	**OR_z**	W	M	NAA	AA	Y	O	**OR_z**	W	M	NAA	AA	Y	O
Xenobiotic metabolism[**]																					
GST hap C[††]	**1.7**	2.5	1.3	1.3	2.3	2.0	1.6	**1.0**	0.8	1.0	0.7	1.3	1.5	0.7	**1.2**	1.1	1.2	1.1	1.3	1.5	0.7
GST hap A[‡‡]	**1.8**						1.7	**1.2**		1.2				0.7	**1.3**						0.7
GST hap B[‡‡]	**1.1**							**0.5**							**0.9**						
GST hap D[‡‡]	**1.9**			1.6			1.5	**1.0**	0.9	1.1	0.9			0.8	**1.3**		1.2	1.3			0.8
GSTM1	**1.4**	1.6	1.2	1.3	1.5	1.7	1.2	**0.8**	0.7	0.8	0.8		0.8	0.8	**1.1**	1.0	1.1	1.1		0.8	0.8
GSTT1	**1.1**	1.7	0.7	0.8	1.7	1.0	1.2	**0.8**	0.8	0.9	0.6	1.3	1.3	0.6	**0.9**		1.0	0.7	1.3	1.3	0.6
MEH 113	**0.9**	0.9	0.9	0.8	0.9	0.8	1.0	**0.7**	0.7	0.9	0.8	0.7	0.6	0.9	**0.6**		0.7	0.7	0.7	0.6	0.9
MEH 139	**0.9**	1.1	1.0	1.1			1.2	**0.8**	1.6	0.9	1.0		1.2	1.1	**0.6**		1.1	1.0		1.2	1.1
DNA repair																					
POLD1 119	**0.6**	1.1	1.3	1.2			1.0	**0.8**	1.3	0.8	1.1		1.1	1.0	**0.6**		0.8	1.0		1.1	1.0
Base excision repair																					
ADPRT 762	**1.1**	1.2		1.1			1.0	**1.1**	1.3	0.9	1.1		1.3	1.0	**1.1**		0.8	1.0		1.3	1.0
ADPRTL2 328	**0.9**	0.9	0.8	1.1			0.9	**1.2**	1.4	1.0	1.2		1.0	1.3	**1.2**		1.1	1.3		1.0	1.3
APE1 148	**1.1**	1.5	0.8	1.2	0.9	1.2	1.0	**1.2**	1.3	1.2	1.3	1.2	1.4	1.2	**1.1**		1.0	1.2	1.2	1.4	1.2
XRCC1 194	**0.7**							**0.7**							**1.1**			1.0			
XRCC1 280	**1.0**																				
XRCC1 399	**1.2**	1.5	1.0	1.4	1.1	0.8	1.6	**1.2**	1.2	1.0	1.2		1.0	1.3	**0.8**	1.0	0.7	0.9		1.0	1.3
Double strand break repair																					
NBS1 185	**1.0**	0.6	0.6	1.0	0.6	1.0	0.7	**0.8**	0.8	0.8	0.8	0.8	0.8	0.8	**1.0**		0.8	0.6	0.8	0.8	0.8
XRCC3 241	**0.9**	0.7	1.1	0.7	1.2	0.7	1.0	**0.8**	0.9	0.8	0.8	1.0	1.0	0.8	**1.1**	0.7	1.4	1.0	1.0	1.0	0.8
MLH1 219	**1.1**	0.7	1.9	0.9	1.4	-.0	0.8	**1.2**	0.6	1.4	0.9	1.2	1.2	0.9	**1.2**		1.0	0.8	1.2	1.2	0.9
MSH3 1036	**0.8**	1.4	0.7	1.2			0.9	**1.1**	1.6	0.7	1.0		1.6	0.9	**1.3**		0.9	1.5	-1.	1.6	0.9
MSH3 940	**1.1**	1.2	0.5	1.0	0.7	0.8	0.8	**1.2**	0.8	0.9	0.9	0.9	0.8	0.9	**1.6**		0.8	1.0	0.9	0.8	0.9
MSH6 39	**0.8**	1.0	1.0	0.9	1.0	1.2	0.9	**0.9**	0.6	1.0	0.8	0.8	0.9	0.8	**0.9**	1.1	1.0	1.0	0.8	0.9	0.8
Nucleotide excision repair																					
RAD23B	**1.2**	0.7	1.2	0.8	1.1	1.3	0.8	**1.0**	0.6	0.9	0.7	0.7	0.9	0.6	**0.9**		0.8	0.6	0.7	0.9	0.6
XPC 499	**1.0**	1.3	0.8	0.8			1.0	**0.8**	0.8	0.7	0.7			0.7	**0.8**		0.9	0.7			0.7
XPC 939	**1.2**	1.0	1.4	1.3	1.2	1.1	1.3	**1.4**	1.6	1.2	1.4	1.3	1.2	1.5	**1.0**		0.8	1.2	1.3	1.2	1.5
XPD 312	**1.1**	0.8	1.3	1.1	1.1		1.1	**0.9**	1.5	0.7	1.0		0.8	1.0	**1.0**		0.9	1.0		0.8	1.0
XPD 751	**1.4**	1.4	1.3	1.5	1.3	1.4	1.4	**1.0**	1.3	0.8	1.0	0.9	0.7	1.2	**1.4**		1.2	1.3	0.9	0.7	1.2
XPF 415	**-1.0**	-.0	-.0	-.0	-.0	-.0		**1.2**	1.0		1.3			1.4	**1.0**			0.9			1.4
XPG 1104	**0.9**	1.0	0.8	0.8	1.0	0.8	0.9	**1.1**	0.9	1.1	1.0	1.2	1.5	0.8	**1.0**	-.0	0.7	1.0	1.2	1.5	0.8
Oxidative stress defense																					
MNSOD	**1.1**	0.8	0.8	1.1	0.7	1.0	0.7	**1.0**	1.5	0.8	1.0	1.3	1.4	0.9	**0.9**	2.2	0.9	1.3	1.3	1.4	0.9

[*]Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively; Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4; [†]Odds ratio not displayed if confidence limit ratio >4; [‡]Referent is never smokers for all smoking categories unless otherwise noted; [||]SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; [**]Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; [††]*GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined; [‡‡]*GST* hap A=*GSTT1* null & *GSTM1* present, *GST* hap B=*GSTT1* null & *GSTM1* null, *GST* hap D=*GSTT1* present & *GSTM1* null; *GST* hap C is referent; [***]Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day; [§§]Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05; **Bold** = Overall OR_z.

= $OR_z \leq 0.7$

= $OR_z \geq 1.4$

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, W=Women, M=Men, NAA=Non African-American (98% white), AA=African American, Y=Young (<65y), O=Old (>=65y), PY=pack-years, y=years.

Table 11. Gene variant-pack-years of smoking association in the NCCCS, overall and by gender and race[*,†]

Gene pathway/ Gene variant[‖]	≤35 PY[***]							>35PY[***]						
	OR_z[†]	W	M	NAA	AA	<65y	≥65y	**OR_z**	W	M	NAA	AA	<65y	≥65y
Xenobiotic metabolism[**]														
GST hap C[††]	**1.3**	1.6	1.1	1.1	1.7	1.5	1.2	**1.0**		1.1	0.8			1.0
GST hap A[‡‡]	**1.5**	1.9	1.2	1.0	2.1		1.2	**1.2**						
GST hap B[‡‡]	**0.8**	1.2		0.8										
GST hap D[‡‡]	**1.4**	1.6	1.3	1.3	1.8	1.8	1.2	**1.0**		1.2	1.0			0.9
GSTM1	**1.1**	1.3	1.0	1.1	1.1	1.1	1.1	**0.9**		1.0	0.9			1.0
GSTT1		1.2	0.8	0.8	1.4	1.1	0.9			0.9	0.6			0.9
MEH 113	**0.8**	0.7	0.9	0.8	0.8	0.7	1.0	**0.7**		0.6	0.7			0.6
MEH 139	**0.8**	1.4	1.0	1.1	1.4	1.2	1.2	**0.6**		1.0	0.9			1.2
DNA repair														
POLD1 119	**0.7**	1.1	0.9	1.0		1.2	0.9	**0.6**		1.0	1.2			1.0
Base excision repair														
ADPRT 762	**1.1**	1.2	1.0	1.2		1.0	1.2	**1.0**			1.0			0.9
ADPRTL2 328	**1.1**	1.2	1.0	1.2		1.0	1.1	**1.2**		1.0	1.2			1.2
APE1 148	**1.1**	1.2	0.9	1.2	1.0	1.1	1.0	**1.2**		1.0	1.4			1.3
XRCC1 194	**0.8**	0.5		0.7	0.9		0.7	**0.8**						
XRCC1 280	**1.2**													
XRCC1 399	**1.2**	1.2	1.0	1.3	1.0	0.8	1.4	**0.9**		0.7	0.9			1.0
Double strand break repair														
NBS1 185	**1.0**	0.7	0.8	0.7	0.7	0.7	0.8	**0.8**		0.6				0.7
XRCC3 241	**0.8**	0.8	0.9	0.7	1.1	0.9	0.8	**1.1**		1.5	1.1			1.2
Mismatch repair														
MLH1 219	**1.2**	0.8	1.6	1.0	1.3	1.8	0.9	**1.0**		1.0	0.6			0.6
MSH3 1036	**1.0**	1.7	0.8	1.1	1.3	1.4	1.1	**1.5**		0.8	1.3			1.3
MSH3 940	**1.2**	0.9	0.7	0.9	0.8	0.8	0.9	**1.3**		0.7	1.0			1.1
MSH6 39	**0.8**	0.9	1.0	1.0	0.9	1.1	0.9	**1.0**		0.9	0.7			0.7
Nucleotide excision repair														
RAD23B	**1.0**	0.6	1.0	0.7	0.9	0.9	0.7	**1.1**		0.8	0.7			0.5
XPC 499	**0.9**	1.1	0.8	0.8	1.3	1.0	0.9	**0.7**		0.8	0.6			0.6
XPC 939	**1.1**	1.1	1.1	1.2	1.1	1.0	1.3	**1.5**		1.1	1.6			1.6
XPD 312	**1.0**	1.0	1.0	1.0	1.0	1.0	1.0	**1.0**		0.8	0.9			0.8
XPD 751	**1.2**	1.4	1.1	1.3	1.1	1.2	1.3	**1.2**		1.0	1.0			0.9
XPF 415	**0.9**	10		1.0	1.0		1.1	**1.1**			1.1	1.0		
XPG 1104	**0.9**	0.9	0.8	0.8	1.0	1.1	0.8	**1.3**		0.9	1.3			0.8
Oxidative stress defense														
MNSOD	**1.1**	1.1	0.8	1.0	0.9	1.2	0.8	**0.8**	1.0	1.1	1.5			1.4

[*]Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively; [†]Odds ratio not displayed if confidence limit ratio >4; [‡]Referent is never smokers for all smoking categories unless otherwise noted; [‖]SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; [**]Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; [††]*GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined; [‡‡]*GST* hap A=*GSTT1* null & *GSTM1* present, *GST* hap B=*GSTT1* null & *GSTM1* null, *GST* hap D=*GSTT1* present & *GSTM1* null; *GST* hap C is referent; [***] Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day; [§§]Could not be pooled for some measures of smoking. LRT p-value for race[*]smoking interaction term <0.05; **Bold** = Overall OR_z.

= OR_z ≤0.7

= OR_z ≥1.4

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, W=Women, M=Men, NAA=Non African-American (98% white), AA=African American, PY=pack-years, y=years.

other metabolism gene and two DNA repair genes (both NER) showed moderate associations with current smoking. For duration, two metabolism SNPs (*GSTP1* and *NAT1*) and two DNA repair SNPs (*OGG1* and *XRCC1* 194) had mmOR$_z$s for <10yrs. Only one SNP was associated with the longest duration of smoking (*MGMT* 84). Eleven SNPs were associated with either low or high PY, including five that were not associated with any other measure of smoking.

When the CBCS OR$_z$s were stratified by race, there was little evidence of heterogeneity. Using p-values to evaluate effect measure modification by race, approximately 6% of the likelihood ratio tests for a race-smoking interaction term were significant at α=0.05, about what would be expected by chance. There was no pattern of significant interaction by race for any given smoking measure. Only *NQO1* was significant for interaction for more than one smoking measure; OR$_z$ differed significantly for all smoking measures and was inverse for African Americans and positive for non-African Americans.

Misspecification of smoking exposure strongly affected the frequency of bias in the COR (**Table 12**). SNPs with mmOR$_z$s for smoking amount rarely showed equivalent mmOR$_z$s for smoking status.

NCCCS

In the NCCCS controls, five SNPs in four genes (*MEH* 113, *MEH* 139, *GSTM1, POLD1* 119, *MSH3* 940) and three haplotypes of *GST*, were moderately associated with smoking behavior (**Tables 7**-**11**). *MEH* 113 and *MEH* 139 were both inversely associated with smoking for at least one smoking status category and one smoking amount category. The bulk of moderate OR$_z$s in the metabolic genes can be attributed to the two SNPs in the *MEH* gene. *POLD1* 119, a DNA repair gene, was most consistently associated with smoking across categories. As in the CBCS, metabolism genes were overrepresented in the stronger associations and DNA repair genes in the weaker associations. Associations between metabolism SNPs and smoking were consistently inverse.

Within smoking categories, three of the four metabolism gene SNPs (*GSTM1, MEH* 113 & 139) were inversely associated with smoking status; three DNA repair SNPs showed inverse mmOR$_z$s for smoking status (*POLD* 119, *MSH3* 940 and *MSH6* 39). All measures of amount showed clustering of positive associations in MMR and NER DNA repair genes, and inverse associations for metabolic genes and BER DNA repair genes. Six SNPs were associated with high PY (*MEH* 113, *MEH* 139, *POLD* 119, *MSH3* 1036, *XPC* 499 and *XPC* 939), two of which had no association with any other smoking measure (*MSH3* 1036, *XPC* 499).

When stratified by gender, only *MSH3* 940 differed significantly across more than one smoking measure. OR$_z$s for ever smoking, duration and PY were higher among women than men when positive or closer to the null when inverse. *MSH3* 1036 showed the same pattern however the LRT for gender was not significant for any measure of smoking.

No strong patterns emerged with stratification by race, although estimates were often on opposite sides and close to the null. The exception was *GSTT1* where stratification by race produced moderate inverse associations in whites and moderate positive associations in African Americans for most evaluable smoking measures (ever, current, and intensity). Approximately 3% of the likelihood ratio tests for a race-smoking interaction term were statistically significant.

As in the CBCS, exposure misspecification frequently produced bias. For smoking amount, there were 20 SNPs or haplotypes with moderate magnitude associations; only 6 showed a similar result for smoking status.

CBCS and NCCCS

For the 15 SNPs measured in both studies (**Table 13**), the weighted kappa for agreement in OR$_z$ was -0.07 (95% CI: -0.19, 0.06), indicating slight disagreement (**Table 14**) [null: 0.9-1.1 (inclusive)] [31]. When CBCS and NCCCS datasets were restricted to white women 40-74 years of age to improve comparability (**Table 15**), results were only evaluable in the NCCCS for 13 or fewer SNPs, even with the limits for precision relaxed (CLR <5). The kappa for agreement in OR$_z$ was then 0.22 (95% CI: -0.01, 0.46), considered slight agreement [31]. No SNP exhibited an ORz ≥1.4 or ≤0.7 in both studies.

Table 12. Misspecification for Gene variant-smoking status associations (ORz*) in the CBCS and NCCCS

Gene pathway**/ Gene variant‖	Status			Duration				Intensity			Pack-years§§	
	Ever‡	Current§	Former	Current	≤10 yrs	11-20 yrs	>20 yrs	<1/2 pk/day	1/2 - 1 pk/day	>1 pk/ day	≤35 PY	>35 PY
CBCS*,†												
Xenobiotic metabolism**												
CYPIA1 M1	1.0	1.0	1.0	1.0	1.0	1.3	0.8	0.9	1.0	1.0	1.0	
CYPIA1 M2	1.8		2.1‖‖								1.6	
CYPIA1 M3	0.9										1.0	
CYPIA1 M4	1.3	2.5‖‖										
GSTM1	1.0	1.1	1.0	1.1	1.1	1.0	1.0	0.9	0.9	1.3	0.9	1.7
GSTP1	1.2	0.7	1.8‖‖	0.8	1.9‖‖	1.0	1.0	1.2	1.8	0.9	1.4	0.7
GSTT1	1.0	1.1	0.9	1.1		1.0	1.0	1.3	1.1		1.0	
NAT1	0.9	1.2	0.8	1.1	0.6	1.1	1.1	0.8	1.2	0.9	0.9	
NAT2	0.9	1.3	0.8	1.2	0.8	0.9	1.1	0.8	1.0	1.1	0.9	
COMT	0.8	0.9	0.8	0.9	1.0	0.8	0.8	0.9	1.2	0.6‖‖	0.9	0.5‖‖
DNA repair												
Base excision repair												
APE1 148	1.1	1.2	1.1	1.2	1.2	1.1	1.1	1.3	0.9	1.2	1.1	1.3
hOGG1	1.0	0.9	1.1	1.0	1.4	1.1	0.9	0.9	1.1	1.1	1.1	0.9
MYH 324	1.0	0.8	1.1	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
XRCC1 194	1.1	1.1	1.1	1.2	1.4	0.9	1.1	1.2	1.0	1.2	1.1	1.6
XRCC1 280	0.9	0.9	0.9	0.9	0.8	0.9	1.0	0.9	1.0	0.8	0.9	
XRCC1 399	1.0	1.2	1.0	1.2	0.8	1.1	1.2	0.9	1.1	1.1	1.0	1.0
Double strand break repair												
BRCA2 24	0.9	0.9	1.0	0.9	0.9	1.0	0.9	0.8	1.0	0.9	0.9	0.8
BRCA2 372	1.2‖‖	1.2	1.2	1.3	1.0	1.5‖‖	1.3‖‖	1.1	1.2	1.6‖‖	1.2	1.6‖‖
NBS1 185	1.2	1.0	1.2	1.1	1.2	1.3	1.1	1.1	1.2	1.2	1.2	1.0
XRCC2 188	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.8	0.8	1.1	0.9	1.1
XRCC3 241	0.9	1.2	0.8	1.1	0.8	0.8	1.1	0.9	0.9	0.8	0.9	0.8
XRCC4 -28073	1.2	1.2	1.1	1.3	1.1	1.0	1.3	1.1	1.2	1.2	1.1	1.5
Mismatch repair												
MGMT 84	0.9	0.8	1.1	0.8	1.3	1.0	0.7‖‖	1.1	0.8	0.9	1.0	0.5‖‖
Nucleotide excision repair												
ERCC1 8092	1.0	1.0	1.0	1.0	1.0	1.1	0.9	1.0	1.0	0.9	1.0	0.7
ERCC6 1213	1.2	1.6‖‖	1.0	1.6‖‖	1.2	1.2	1.2	1.3	1.3	1.0	1.3‖‖	0.8
ERCC6 1230	0.9	1.1	0.8	1.0	0.8	0.9	1.0	1.0	1.0	0.8	0.9	0.7
HRAD23B	1.1	1.2	1.0	1.2	1.2	0.9	1.2	1.0	1.1	1.2	1.1	1.2
XPC 939	0.9	1.0	0.9	1.0	1.0	0.9	0.9	1.1	0.8	0.9	0.9	1.1
XPD 312	1.0	1.1	1.0	1.1	0.9	1.2	1.1	1.0	1.1	1.1	1.1	0.9
XPD 751	1.2	1.2	1.1	1.2	0.9	1.3	1.3	1.1	1.2	1.2	1.2	1.1
XPF 415	1.0	1.0	1.1	1.0	1.1	0.9	1.0	0.9	1.0	1.1	1.0	1.0
XPF 662	1.1	1.4	1.0	1.3	1.1	1.4	1.0	1.4	0.9	0.9	1.2	
XPG 1104	0.9	0.8	1.0	0.8	0.9	1.0	0.9	0.9	0.9	0.9	0.9	1.2
Other												
CDH1	0.8	0.8	0.8	0.8	0.9	0.7‖‖	0.8	0.9	0.8	0.7‖‖	0.8‖‖	0.9
TGFB1	1.1	0.8	1.2	0.9	1.3	1.1	1.0	1.2	1.1	0.9	1.1	0.8
MnSOD	1.0	0.9	1.1	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.4
MPO	1.0	1.0	1.0	1.0	0.9	1.1	1.1	1.0	1.0	1.1	1.0	1.1
NQO1	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.0	1.2	0.9	1.0	1.1
NCCCS*,†												
Xenobiotic metabolism**												
GST hap C††	1.2	1.0	1.3	1.2	1.5	1.1	1.2	1.7‖‖	1.0	1.2	1.3	1.0
GST hap A‡‡	1.4	1.5	1.3	1.7	1.6	1.3	1.4	1.8‖‖	1.2	1.3	1.5	1.2
GST hap B‡‡	0.8	0.6	0.9				0.7	1.1	0.5‖‖	0.9	0.8	
GST hap D‡‡	1.3	0.9	1.4	1.1	1.7	1.1	1.2	1.9‖‖	1.0	1.3	1.4	1.0
GSTM1	1.0	0.7	1.2	0.8	1.3	0.9	1.0	1.4	0.8	1.1	1.1	0.9
GSTT1	1.0	1.1	0.9	1.1	1.0	1.0	0.9	1.1	0.8	0.9	1.0	0.9
Xenobiotic metabolism6												
MEH 113	0.7	0.8	0.8	0.7	0.7	0.7	0.8	0.9	0.7	0.6‖‖	0.8	0.7‖‖
MEH 139	0.8	0.6‖‖	0.9	0.6‖‖	1.0	0.8	0.7‖‖	0.9	0.8	0.6‖‖	0.8	0.6‖‖
DNA repair												

POLD1 119	0.7	0.8	0.7	0.7	0.7	0.9	0.7‖‖	0.6‖‖	0.8	0.6	0.7	0.6
Base excision repair												
ADPRT 762	1.1	1.2	1.0	1.2	1.2	0.7	1.2	1.1	1.1	1.1	1.1	1.0
ADPRTL2 328	1.1	1.1	1.1	1.1	1.0	1.0	1.2	0.9	1.2	1.2	1.1	1.2
APE1 148	1.1	1.0	1.1	1.1	1.4	1.1	1.0	1.1	1.2	1.1	1.1	1.2
XRCC1 194	0.8	0.9	0.8	0.8			0.9	0.7	0.7	1.1	0.8	0.8
XRCC1 280	1.3		1.3				1.2				1.2	
XRCC1 399	1.1	1.0	1.1	1.0	1.3	1.2	1.0	1.2	1.2	0.8	1.2	0.9
Double strand break repair												
NBS1 185	0.9	0.8	0.9	0.8	0.9	1.2	0.8	1.0	0.8	1.0	1.0	0.8
XRCC3 241	0.9	1.1	0.8	1.0	0.7	0.7	1.0	0.9	0.8	1.1	0.8	1.1
Mismatch repair												
MLH1 219	1.1	0.8	1.2	0.9	1.4	1.0	1.1	1.1	1.2	1.2	1.2	1.0
MSH3 1036	1.1	1.0	1.1	1.0	1.0	1.1	1.0	0.8	1.1	1.3	1.0	1.5
MSH3 940	1.2	0.7	1.4	0.8	1.2	1.6	1.1	1.1	1.2	1.6	1.2	1.3
MSH6 39	0.9	0.7	1.0	0.7	0.8	0.9	0.9	0.8	0.9	0.9	0.8	1.0
Nucleotide excision repair												
RAD23B	1.1	1.0	1.1	1.0	1.2	0.8	1.1	1.2	1.0	0.9	1.0	1.1
XPC 499	0.8	1.1	0.8	1.0	0.8	0.9	0.9	1.0	0.8	0.8	0.9	0.7
XPC 939	1.2	1.0	1.3	1.1	1.0	1.1	1.3	1.2	1.4	1.0	1.1	1.5
XPD 312	1.0	0.9	1.1	0.9	1.3	1.0	0.9	1.1	0.9	1.0	1.0	1.0
XPD 751	1.2	1.0	1.3	1.1	1.5	1.5	1.0	1.4	1.0	1.4	1.2	1.2
XPF 415	1.0	1.3	0.9	1.2	1.0		1.1		1.2	1.0	0.9	1.1
XPG 1104	1.0	1.2	0.9	1.2	0.7	0.8	1.2	0.9	1.1	1.0	0.9	1.3
Other												
MNSOD	1.0	1.1	1.0	1.1	1.0	1.4	0.9	1.1	1.0	0.9	1.1	0.8

*Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4; †Odds ratios are race and age-adjusted for CBCS; race, age and gender-adjusted for NCCCS; ‡Referent is never smokers for all smoking categories unless otherwise noted; §Referent is not-current smokers (former + never); ‖SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; **Primary functional category; gene may function in additional pathways eg *COMT* in estrogen metabolism; ††*GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined; ‡‡*GST* hap A=*GSTT1* null & *GSTM1* present, *GST* hap B=*GSTT1* null & *GSTM1* null, *GST* hap D=*GSTT1* present & *GSTM1* null; *GST* hap C is referent; §Pack-years= midpoint of category for number of years smoked x midpoint of category for numer of packs smoked/day; ‖‖Statistically significant at alpha=0.05.

= $OR_z \leq 0.7$
= $OR_z \geq 1.4$
= $0.7 < OR_z < 1.4$ and significant at alpha=0.05

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, PY=pack-years, y=years, pk/day=packs/day, CBCS=Carolina Breast Cancer Study, NCCCS=North Carolina Colon Cancer Study, SNP=single nucleotide polymorphism.

Results were unchanged when all data, regardless of CLR, were included.

Discussion

The primary aim of this study was to provide new empirical, population-based estimates of association in 2 control groups (OR_zs) for selected, frequently studied, gene-smoking combinations. This information is intended to assist investigators considering conducting a stand-alone case-only study (i.e. one with no controls) to evaluate the independence assumption with respect to these gene-smoking pairs. Without data on the underlying population the cases arose from, ancillary data, such as the current study, must be used to assess the potential for bias in the case-only estimate. Recently, Mukherjee and Chatterjee have proposed a method to reduce the bias introduced when case-only analyses are used to enhance precision, however these methods still require at least partial controls, and depend on empirical evidence from non-cases (e.g. controls) to assess G-E association in the population [5, 6]. In fact, the scarcity of empirical data on G-E independence has been identified as problematic by these and other authors [5, 6, 32, 33].

The secondary motivation was to provide information that could help establish whether specific commonly studied G-E associations might be considered 'universal' (i.e. having a consistent magnitude from study to study) or should more properly be considered population-specif-

Table 13. Gene variant-smoking associations in CBCS and NCCCS controls

Gene pathway[§§]/ gene variant[‡‡]	Smoking Status								Duration						Intensity						Pack-years			
	Ever smoking (OR_z)[‖**]		Current smokers (Ref: Not Current)[††]		Former smokers (OR_z)		Current smokers (OR_z)		≤10y (OR_z)		11-20y (OR_z)		>20y (OR_z)		<1/2pk (OR_z)		1/2 - 1 pk (OR_z)		>1 pk (OR_z)		≤35 PY [‖‖](OR_z))		>35PY (OR_z)	
	B[‡]	C[§]	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C
Xenobiotic metabolism[**]																								
GSTM1	1.0	1.0	1.1	0.7	1.0	1.2	1.1	0.8	1.1	1.3	1.0	0.9	1.0	1.0	0.9	1.4	0.9	0.8	1.3	1.1	0.9	1.1	1.7	0.9
GSTT1	1.0	1.0	1.1	1.1	0.9	0.9	1.1	1.1		1.0	1.0	1.0	1.0	0.9	1.3	1.1	1.1	0.8		0.9	1.0	1.0		0.9
DNA repair																								
Base excision repair																								
APE1 148	1.1	1.1	1.2	1.0	1.1	1.1	1.2	1.1	1.2	1.4	1.1	1.1	1.1	1.0	1.3	1.1	0.9	1.2	1.2	1.1	1.1	1.1	1.3	1.2
XRCC1 194	1.1	0.8	1.1	0.9	1.1	0.8	1.2	0.8	1.4		0.9		1.1	0.9	1.2	0.7	1.0	0.7	1.2	1.1	1.1	0.8	1.6	0.8
XRCC1 280	0.9	1.3	0.9		0.9	1.3	0.9		0.8		0.9		1.0	1.2	0.9		1.0		0.8		0.9	1.2		
XRCC1 399	1.0	1.1	1.2	1.0	1.0	1.1	1.2	1.0	0.8	1.3	1.1	1.2	1.2	1.0	0.9	1.2	1.1	1.2	1.1	0.8	1.0	1.2	1.0	0.9
Double strand break repair																								
NBS1 185	1.2	0.9	1.0	0.8	1.2	0.9	1.1	0.8	1.2	0.9	1.3	1.2	1.1	0.8	1.1	1.0	1.2	0.8	1.2	1.0	1.2	1.0	1.0	0.8
XRCC3 241	0.9	0.9	1.2	1.1	0.8	0.8	1.1	1.0	0.8	0.7	0.8	0.7	1.1	1.0	0.9	0.9	0.9	0.8	0.8	1.1	0.9	0.8	0.8	1.1
Nucleotide excision repair																								
HRAD23B	1.1	1.1	1.2	1.0	1.0	1.1	1.2	1.0	1.2	1.2	0.9	0.8	1.2	1.1	1.0	1.2	1.1	1.0	1.2	0.9	1.1	1.0	1.2	1.1
XPC 939	0.9	1.2	1.0	1.0	0.9	1.3	1.0	1.1	1.0	1.0	0.9	1.1	0.9	1.3	1.1	1.2	0.8	1.4	0.9	1.0	0.9	1.1	1.1	1.5
XPD 312	1.0	1.0	1.1	0.9	1.0	1.1	1.1	0.9	0.9	1.3	1.2	1.0	1.1	0.9	1.0	1.1	1.1	0.9	1.1	1.0	1.1	1.0	0.9	1.0
XPD 751	1.2	1.2	1.2	1.0	1.1	1.3	1.2	1.1	0.9	1.5	1.3	1.5	1.3	1.0	1.1	1.4	1.2	1.0	1.2	1.4	1.2	1.2	1.1	1.2
XPF 415	1.0	1.0	1.0	1.3	1.1	0.9	1.0	1.2	1.1	1.0	0.9		1.0	1.1	0.9		1.0	1.2	1.1	1.0	1.0	0.9	1.0	1.1
XPG 1104	0.9	1.0	0.8	1.2	1.0	0.9	0.8	1.2	0.9	0.7	1.0	0.8	0.9	1.2	0.9	0.9	0.9	1.1	0.9	1.0	0.9	0.9	1.2	1.3
Oxidative stress defense																								
MNSOD	1.0	1.0	0.9	1.1	1.1	1.0	0.9	1.1	1.0	1.0	1.0	1.4	1.0	0.9	1.0	1.1	1.0	1.0	1.0	0.9	1.0	1.1	1.4	0.8

[‡]All odds ratios from CBCS (B) are race and age adjusted; [§]All odds ratios from NCCCS (C) are race, age and gender adjusted; [‖]Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4; [**]Referent is never smokers for all smoking categories unless otherwise noted; [††]Referent is not-current smokers (former + never); [‡‡]SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present; [§§] Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism; [‖‖]Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day.

= OR_z ≤0.7

= OR_z ≥1.4

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, PY=pack-years, met=metabolism, Ph=Phase, CBCS=B=Carolina Breast Cancer Study, NCCCS=C=North Carolina Colon Cancer Study, y=years, pk=packs/day.

Table 14. Agreement between CBCS and NCCCS gene variant-smoking associations

		Kappa*	95% CI		N
Full CBCS and NCCCS					
Null=OR_z: 0.9-1.1	CLR<4	-0.07	-0.19	0.06	165
Restricted CBCS and NCCCS: white women 40-74 y					
Null=OR_z: 0.9-1.1	CLR<5	0.22	-0.01	0.46	52
Null=OR_z: 0.8-1.2	CLR<5	0.19	0.01	0.36	52
Null=OR_z: 0.9-1.1	CLR<20†	0.16	0.02	0.30	163
Null=OR_z: 0.8-1.2	CLR<20	0.20	0.09	0.31	163

*Weighted kappa statistic; †At CLR<20 all data was included except subgroups with empty cells. NOTE: CBCS=Carolina Breast Cancer Study, NCCCS=North Carolina Colon Study, CI=confidence interval, CLR=Confidence limit ratio (upper limit/lower limit), N=number of observations meeting stated CLR condition.

ic. Our recent meta-analysis of published control group data suggested that G-E associations in controls are likely to be population-specific, not universal. However, the analyses only 6 SNPs from 3 genes, and few of the included controls were groups on which the independence assumption is based [7].

Odds ratios for CBCS and NCCCS controls (OR_zs) were of moderate magnitude [$OR_z \geq 1.4$ or ≤ 0.7] for at least one of the six smoking measures for approximately half of the SNPs examined in each of these population-based control groups (CBCS: 45%, NCCCS: 59%). We focused on this magnitude of association because an OR_z of ≥1.4 would inflate the corresponding SIM (interaction term from a case-control study), if positive, by ~40%. This is a substantive degree of bias and could easily mislead researchers into incorrectly concluding G-E interaction exists when it does not. These mmOR_zs were found across all functional categories of genes. For most DNA repair SNPs, particularly BER and DSB genes, both studies showed a preponderance of mmOR_zs in categories of smoking amount rather than smoking status. In contrast, metabolic gene SNPs had mmOR_zs for both status and amount.

Smoking behavior in controls

Several SNPs in the CBCS and NCCCS were notable with mmOR_zs in at least one smoking status category and at least one level of a smoking amount. In the CBCS these were: *CYP1A1* M2 (positive), *GSTP1* (positive & inverse), and *XPF* 662 (positive). In the NCCCS five SNPs had comparable signals: *MEH* 113 and 139, *GSTM1*, *POLD1* 119, and *MSH3* 940. The metabolic genes generally exhibited inverse OR_zs, as did *POLD1* 119, a DNA repair gene. *POLD1* 119 showed the most consistent results across smoking measures. In the CBCS, *COMT, CDH1, XRCC1* 194, *BRCA2* 372, and *MGMT* 84 showed moderate associations in more than one smoking measure; *CYP1A1* M4 and *ERCC6* 1213 showed association in more than one level of a single smoking measure.

Published results from only a few population-based or control group studies are available to compare with our study [34-36]. Smits et. al. used pooled control group data from the International Collaborative Study on Genetic Susceptibility to Environmental Carcinogens (GSEC) to estimate OR_zs between polymorphisms in five metabolic genes (*CYP 1A1*, *GSTT1*, *GSTM1*, *GSTP1* and *NAT2*) and six measures of smoking (ever, former, current, cig/day, years smoked and PY). Total sample size for each gene varied (*GSTM1*: N=10,719 to *GSTP1*: N=2,792); however, less than half of controls had information on smoking amount. Odds ratios were adjusted for study, age, sex and ethnicity. Results for these five genes and smoking status were usually at or near the null. Overall, the results were broadly similar to the CBCS and NCCCS, but there were several differences that have implications for the validity and interpretation of case-only interaction estimates. For example, in GSEC controls the overall OR_z for *GSTP1* and current smoking was above the null; but below the null in the CBCS. The CBCS OR_z was similar to female GSEC controls, but different than the GSEC OR_z for non-hospital controls. GSEC and CBCS OR_zs for *GSTP1* were even further apart for former smokers. This variation in OR_z between pooled controls from

Table 15. Gene-variant - smoking associations * in CBCS & NCCCS: Non-African American female controls 40-74 years of age

	Smoking † Status								Duration (years)						Intensity (pack/day)						Pack-years‡			
	Ever smoking (OR_z)*,§		Current smokers (Ref: Not Current)‖		Former smokers (OR_z)		Current smokers (OR_z)		≤10y (OR_z)		11-20y (OR_z)		>20y (OR_z)		<1/2pk (OR_z)		1/2 - 1 pk (OR_z)		>1 pk (OR_z)		≤35 PY(OR_z)		>35PY (OR_z)	
SNP††	B‡‡	C§§	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C
CBCS Only																								
Xenobiotic metabolism**																								
CYPIA1 M1	0.8		0.9		0.7		0.8		1.0				0.7		0.7		0.6		1.0		0.7		1.1	
CYPIA1 M2	1.5				2.0																			
CYPIA1 M3																								
CYPIA1 M4	1.6		3.1				2.8														1.2			
GSTP1	1.5		0.8		2.0		1.0		2.0		1.5		1.3		1.8		2.2		0.9		1.9		0.7	
NAT1	1.2		1.3		1.1		1.4				1.3		1.2				1.7		1.0		1.2			
NAT2	1.1		2.0		0.9		1.9				1.1		1.5		0.8		1.2		1.3		0.9			
COMT	0.6		0.7		0.6		0.6				0.4		0.7		0.5		0.8		0.5		0.7		0.5	
DNA repair																								
hOGG1	1.0		0.9		1.0		0.9		1.1		1.1		0.9		0.8		1.1		1.1		1.0		1.0	
MYH 324	1.0		0.9		1.1		0.9		1.2		0.8		1.1		1.0		1.0		1.0		1.0		1.1	
BRCA2 24	0.9		0.8		1.0		0.8		1.0		0.8		0.9		1.0		0.8		0.9		1.0		0.7	
BRCA2 372	1.3		1.2		1.3		1.3		1.2		1.8		1.2		0.9		1.3		1.7		1.2		1.6	
XRCC2 188	0.8		0.7		0.8		0.7		0.5		0.9		0.9		0.6		0.6		1.0		0.7		1.0	
XRCC4 -28073	1.1		1.1		1.1		1.1		0.7		1.0		1.5		0.9		1.1		1.4		1.0		1.7	
MGMT 84	1.0		1.0		0.9		1.0		1.3		0.9		0.8		1.0		1.0		0.8		1.0		0.6	
ERCC1 8092	0.9		0.8		0.9		0.8		0.9		0.8		0.9		0.9		1.1		0.8		1.0		0.7	
ERCC6 1213	1.4		1.8		1.2		1.9		1.2		1.4		1.5		1.3		1.4		1.4		1.5		1.1	
ERCC6 1230	0.8		1.0		0.7		0.9		0.7		0.8		0.9		0.8		0.9		0.7		0.8		0.6	
Other																								
CDH1	0.8		0.9		0.8		0.8		0.9		0.8		0.8		1.0		0.8		0.8		0.8		0.9	
TGFB1	0.9		0.7		1.1		0.7		1.3		0.9		0.8		1.1		1.0		0.8		1.0		0.8	
MPO	1.3		1.2		1.3		1.3		1.4		1.0		1.4		1.5		1.2		1.2		1.3		1.2	
NQO1	1.4		1.4		1.3		1.6		1.2		1.4		1.5		1.2		1.6		1.3		1.4		1.4	
NCCCS Only																								
Xenobiotic metabolism**																								
MEH 113		0.6				0.6								0.4				0.5				0.5		
MEH 139		1.0				1.0								1.1				1.3				1.2		
GST hap C‖‖		0.7				0.7								0.6				0.4				1.0		
*GST hap A****																								
GST hap B†††																								
GST hap D‡‡‡		1.1				1.0								0.9								1.4		
DNA repair																								
POLD1 119		1.5				1.4								1.5				1.7				1.6		
ADPRT 762		1.7				1.4								1.7				1.9				1.7		

ADPRTL2 328	1.1			1.2						1.0		1.0		1.2		
MLH1 219	0.6			0.6						0.5				0.7		
MSH3 1036	2.2			2.7										2.1		
MSH3 940	1.5			1.6						1.7				1.1		
MSH6 39	0.7			0.7						0.6		0.5		0.8		
XPC 499	0.7			0.7						0.6		0.6		0.8		
CBCS and NCCCS																
Xenobiotic metabolism**																
GSTM1	1.1	0.8	0.9	1.3	0.8	1.0	1.4	0.9	1.2	0.7	1.0	1.0	1.4	1.0	1.1	1.7
GSTT1	0.7	0.5	0.9	0.7	0.6	0.8			0.8			1.1		0.8	0.6	
DNA repair																
APE1 148	1.3	1.3	1.3	1.2	1.2	1.4	1.2	1.3	1.3		1.3	1.2	1.5	1.2	1.4	1.8
XRCC1 194	1.0		0.9	1.1		1.0	1.4	0.7	1.0		1.2	0.6	1.4	0.9		1.5
XRCC1 280	0.9		0.8	1.0		0.8	0.9		1.1		0.8	0.9	0.9	0.8		1.2
XRCC1 399	1.1	1.6	1.3	1.0	1.8	1.3	0.9	1.3	1.2	1.8	1.1	1.3	1.0	1.2	1.6	0.9
NBS1 185	1.3	0.7	1.0	1.4	0.7	1.2	1.3	1.3	1.4	0.6	1.1	1.4	1.4	1.4	0.8	1.1
XRCC3 241	0.9	0.5	1.2	0.8	0.4	1.1	0.8	1.0	0.9	0.6	0.8	1.1	0.8	0.9	0.5	0.8
HRAD23B	1.1	0.6	1.5	0.9	0.8	1.4	0.8	0.8	1.3	0.6	0.9	1.1	1.2	1.0	0.6	1.2
XPC 939	0.9	1.7	1.1	0.9	1.6	1.0	1.2	0.8	0.9	2.4	1.2	0.8	0.9	0.9	1.4	1.2
XPD 312	1.1	1.7	1.1	1.0	1.7	1.1	1.0	1.1	1.1	1.8	1.0	1.1	1.1	1.1	1.4	0.9
XPD 751	1.2	1.9	1.1	1.3	2.1	1.2	1.1	1.3	1.3	1.5	1.3	1.1	1.3	1.3	1.6	1.1
XPF 415	1.2		1.1	1.1		1.2	1.4	1.1	1.0		0.9	1.1	1.5	1.1		1.3
XPG 1104	1.0	1.1	0.8	1.1	1.0	0.9	1.1	1.1	1.0	1.6	1.0	1.1	1.0	0.9	0.6	1.4
Other																
MnSOD	0.9	1.5	0.9	1.0	1.6	0.9	1.1	0.7	1.0	1.2	0.9	0.8	1.1	0.8	1.3	1.6

‡‡B=CBCS (breast cancer study); §§C=NCCCS (colon cancer study); *OR=Odds ratio; OR not presented if 95% confidence limit ratio >5 (upper limit/lower limit>5); †Referent is never smokers for all OR_z unless otherwise noted; ‡Pack-years= number of years smoked x packs smoked/day [20cigarettes=1 pack]; §All OR_z are age adjusted (continuous); ||Referent is not-current smokers (former + never); **Primary functional category, gene may function in additional pathways e. g. *COMT* in estrogen metabolism; ††SNP referent = homozygous for common allele (compared to heterozygotes + homozygous for less common alleles); ref=present for *GSTM1* & *GSTT1*; ref=rapid for *NAT1* and *NAT2*.; ||||*GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null; ***GST* hap A=haplotype of *GSTT1* null & *GSTM1* present ; *GST* hap C is referent; †††*GST* hap B=haplotype of *GSTT1* null & *GSTM1* null; *GST* hap C is referent; ‡‡‡*GST* hap D=haplotype of *GSTT1* present & *GSTM1* null; *GST* hap C is referent; **Bold** = Overall OR_z.

= OR_z ≤0.7

= OR_z ≥1.4

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, PY=pack-years, CBCS=Carolina Breast Cancer Study, NCCCS=North Carolina Colon Cancer Study, y=years, pk=packs/day, SNP=single nucleotide polymorphism, B=Breast cancer (CBCS), C=colon cancer (NCCCS).

the GSEC and these two population-based control groups, as well as the variation between subgroups in the pooled controls, suggest that OR_z is specific to each underlying population and subgroup rather than an estimate of some 'universal' OR_z for that SNP and smoking measure. Furthermore, the *GSTP1* results, in particular, imply that increasing sample size by pooling is not sufficient to compensate for lack of controls from the relevant underlying population.

Finally, in the largest population-based candidate gene study of smoking to date, Liu et. al. examined a panel of 153 SNPs in 40 smoking-related genes in a sample of Japanese men 40-49 years of age (N=339) [36]. Liu et. al. found significant associations for 14 SNPs and current smoking (referent=not current smoker). OR_zs were presented only when statistically significant. The OR_zs for *MEH* SNPs were consistent with NCCCS results: 1) Liu et. al. study: MEH rs2292566: OR_z=0.4 (0.2, 0.8) and 2) NCCCS -*MEH* 113 & 139: OR_z=0.8 (0.5, 1.1) and 0.6 (0.4, 0.9) respectively.

In an evaluation of the independence assumption for gene-smoking associations in controls, Hamajima et. al. [35] calculated OR_z(95%CI) using four published control groups [37-40] for ever smoking and SNPs in *CYP2E1, NAT2*, and *CYP1A1*. None of the OR_zs were significant at α=0.10, however, the magnitude of OR_zs ranged from 2.3 (*CYP2E1*) to 0.6 (*CYP2E1*); OR_zs for *NAT2* (slow) and *CYP1A1* (M2) were 0.6 and 0.7, respectively. Although the authors noted that the magnitude of the OR_zcould have introduced bias into the COR, they concluded, on the basis of statistical significance alone, that these SNPs could be used with smoking in a case-only interaction study. A similar approach was used by Egan et. al., where the magnitude of gene-environment associations varied from 0.5 to 1.1, and in Marcus et. al., where associations ranged from 0.5 to 1.8 [41, 42]. In each case, the only associations considered problematic were the statistically significant ones. This is in contrast to methods of assessing bias in common practice, where the magnitude of the change in the estimate of interest is of primary concern [43].

Metabolic genes and smoking behavior

A more extensive literature exists on smoking and specific metabolic genes, (e.g. those coding for nicotine-metabolizing enzymes) [44, 45]. Variation in these genes can alter enzyme activity, regulation or [46] plausibly increase or decrease disease risk or influence smoking behaviors. Of the seven metabolic genes included the CBCS data, five (*CYP1A1, GSTM1, GSTP1, NAT1* and *COMT*) had mmOR_zs in at least one measure of smoking. In the NCCCS, all three metabolic genes (*GSTM1*, *GSTT1*, and *MEH*) showed moderate association with at least one measure of smoking.

Of these, the *COMT* Val158Met SNP (rs4680) is the only SNP that has been extensively studied with respect to its possible influence on smoking behavior [47]. Results have been equivocal with two large population-based European studies coming to different conclusions [48, 49]. Omdivar et. al. found a 20% reduction in incident smoking cessation for carriers of the low activity form of the *COMT* allele (Met carriers) whereas Breitling et. al. found no association [OR=0.97 (0.83, 1.12)]. Results from the CBCS were consistent with Met carriers having slightly reduced duration and PY of smoking (OR_z= 0.5 for >35PY; OR_z= 0.8 and 0.8 for 11-20y and >20y smoking, respectively).

For *CYP1A1*, Chen et. al. demonstrated that having at least one *CYP1A1**2A allele was associated with smoking reduction and increased quitting during pregnancy [2.2 (1.0, 4.6) and 1.7 (1.0, 2.9), respectively] [50]. CBCS results for women <50y were consistent with higher quitting for those with an M1 allele. (OR_z=1.5 and 1.1, former and current smoking, respectively). For *GSTM1*, Chen et. al. found no association between *GSTM1* null and less smoking, whereas results from the NCCCS showed an association with less smoking (OR_z for women=1.6 for <1/2 pack/day). Findings for *GSTP1, GSTT1* and *MEH* have not been reported previously.

DNA repair genes and smoking

Studies that examined DNA repair genes and smoking behavior are scarce. The population-based candidate gene study of habitual smoking by Liu et. al. included several DNA repair genes in addition to the metabolic genes discussed earlier [36]. Of the statistically significant DNA repair gene SNPs reported, only *OGG1* was in the current study [OR_z =0.6 (0.4, 1.0), and OR_z =1.0 (0.9, 1.3) for ever smoking in

Liu et. al. and CBCS, respectively]. Findings for the other DNA repair SNPs in the CBCS and NCCCS [*XPF, MSH3* (stratified by gender), and *POLD1*] have not been reported previously.

Implications for case-only studies

Based on the magnitude of the gene-smoking associations observed in the CBCS and NCCCS ($OR_z \geq 1.4$ or ≤ 0.7), a case-only interaction estimate would be biased for at least one level of smoking behavior in at least one of the six measures examined for approximately half of the SNPs (CBCS: 45%, NCCCS: 59%). Moderate magnitude OR_zs were most often found for measures of smoking amount rather than smoking status, consistent with the finding that a gene or genes in the chromosome 15q24-25 region are associated with nicotine dependence, lung cancer and chronic obstructive pulmonary disease (COPD), but not smoking initiation or smoking status [51]. Although our specific results need to be replicated in other population-based control series, the implications for the conduct of case-only studies are clear. The magnitude of OR_z is not reliably close to the null for many of these SNPs, making them unsuitable for a stand-alone case-only interaction analysis. These results also show that evaluating the independence assumption using smoking status alone is not sufficient evidence of G-E independence for smoking amount measures such as duration, intensity and PY, the measures of interest for many case-only analyses. Few SNPs with mmOR_zs in any category of smoking amount had OR_zs of comparable magnitude for measures of smoking status in either control group (CBCS: 25%, NCCCS: 13%). Similarly, making a decision based solely on the p-value of OR_z would result in approximately half of the moderate magnitude associations in the CBCS controls being missed and around 80% of the mmOR_zs in the NCCCS being missed. This was observed across all gene categories in both control groups.

Strengths and limitations

Two major strengths of this study are the population-based design and sample size. The independence assumption for case-only analyses is a large sample assumption that pertains specifically to G-E associations in the population underlying the case series. With a control group rather than a population sample, the true parameter (RR_z) could only be estimated by OR_z. However, OR_z is the most easily available parameter in the literature, and is usually used to evaluate the independence assumption, making it the more relevant measure for this study.

Both studies had information on smoking amount, often the exposures of interest in a case-only interaction analysis, and rarely available in the published literature. The CBCS and NCCCS are drawn from overlapping underlying populations, using the same sampling methods, enhancing comparability of the two control groups. Because the current study was a convenience sample of SNPs originally chosen for their relevance to different cancers, there were a limited number of SNPs included in both studies. Further, for African American women 40-74 years of age in the NCCCS, very few SNPs and smoking measures met our precision criteria so we were not able to assess agreement over all SNPs for this restricted group.

Selection bias could have distorted the true gene-smoking relationship in the controls if joint smoking and genetic status are associated with reduced or increased participation rates. If participation rates varied by family history (or any proxy for G+), OR_z would be driven away from the true OR_z in an unpredictable direction. However, the population prevalence of current smoking in the CBCS (20%) was similar to NC women in the 2001 Behavioral Risk Factor Surveillance System (BRFSS) (23%), while former smokers and never smokers, respectively, are only slightly over- and under-represented in the CBCS (CBCS: 29%, BRFSS: 20%, CBCS: 51%, BRFSS: 57%) [18, 30].

The precise biological functions of most of the SNPs in this study were unknown, limiting causal interpretations. Associations could have been due to chance or to polymorphisms in linkage disequilibrium with the assayed polymorphisms. Linkage disequilibrium can vary across ethnicities; however, with one exception (*NQO1*), results did not vary substantively by race. Additionally, agreement was substantially enhanced when the CBCS and NCCCS datasets were restricted by gender, race and age, which would not be expected if the SNP-smoking associations were due solely to chance.

Conclusions

Our study, based on an expanded study resource, extended previous analyses and showed that the gene-smoking OR_zs in population controls are often of sufficient magnitude to introduce appreciable bias into a case-only study estimate of multiplicative interaction. We therefore recommend that a stand-alone case-only study should be conducted only when the independence assumption can be verified with appropriate empirical data, either through a direct evaluation of population-specific data or, if sufficient published data are available, use of OR_zs within a narrow, pre-specified range of acceptable bias, across a wide variety of population-based studies. These data are needed for every smoking metric proposed for the case-only analyses. In the short term, it would be extremely useful to have more detailed control group information available from large population-based studies for a variety of genes. Specifically, it would be useful to have more detailed data on smoking amount, ideally stratified by race and gender. Given that many studies already collect more detailed information on smoking behavior in controls than is actually presented in a paper, these data could relatively easily be archived as supplemental tables online or presented from multiple studies in a collaborative report. Other exposures whose effect might be modified by genetic variation (e.g. air pollution, infectious diseases, alcohol consumption, chemotherapeutics) should also be examined.

Conflicts of interest statement

None. Corresponding author is currently employed at GlaxoSmithKline, 5 Moore Drive, RTP, NC 27709.

Address correspondence to: M. Elizabeth Hodgson, 1310 Crabapple Lane, Raleigh NC 27607. E-mail: m.elizabeth.hodgson@gmail.com

References

[1] Prentice RL, Vollmer WM and Kalbfleisch JD. On the use of case series to identify disease risk factors. Biometrics 1984; 40: 445-458.

[2] Piegorsch WW, Weinberg CR and Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994; 13: 153-162.

[3] Khoury MJ and Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene- environment interaction: Case-control studies with no controls! Am J Epidemiol 1996; 144: 207-213.

[4] Albert PS, Ratnasinghe D Tangrea J, and Wacholder S. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 2001; 154: 687-693.

[5] Mukherjee B, Ahn J, Gruber SB, Rennert G, Moreno V and Chatterjee N. Tests for gene-environment interaction from case-control data: A novel study of type I error, power and designs. Genet Epidemiol 2008; 32: 615-626.

[6] Mukherjee B and Chatterjee N. Exploiting gene-environment independence for analysis of case-control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. Biometrics 2008; 64: 685-694.

[7] Hodgson ME, Poole C, Olshan AF, North KE, Zeng D and Millikan RC. Smoking and selected DNA repair gene polymorphisms in controls: Systematic review and meta-analysis. Cancer Epidemiol Biomarkers Prev 2010; 19: 3055-3086.

[8] Liu X, Fallin MD and Kao WHL. Genetic dissection methods: Designs used for tests of gene-environment interaction. Curr Opin Genet Dev 2004; 14: 241-245.

[9] Hall IJ, Moorman PG, Millikan RC and Newman B. Comparative analysis of breast cancer risk factors among African-American women and white women. Am J Epidemiol 2005; 161: 40-51.

[10] Il'yasova D, Martin C and Sandler RS. Tea intake and risk of colon cancer in African-Americans and Whites: North Carolina colon cancer study. Cancer Causes Control 2003; 14: 767-772.

[11] Millikan R, Eaton A, Worley K, Biscocho L, Hodgson E, Huang WY, Geradts J, Iacocca M, Cowan D, Conway K and Dressler L. HER2 codon 655 polymorphism and risk of breast cancer in African Americans and whites. Breast Cancer Res Treat 2003; 79: 355-364.

[12] Newman B, Moorman PG, Millikan R, Qaqish BF, Geradts J, Aldrich TE and Liu ET. The Carolina Breast Cancer Study: Integrating population-based epidemiology and molecular biology. Breast Cancer Res Treat 1995; 35: 51-60.

[13] Satia-Abouta J, Galanko JA, Potter JD, Ammerman A, Martin CF and Sandler RS. Associations of Total Energy and Macronutrients with Colon Cancer Risk in African Americans and Whites: Results from the North Carolina Colon Cancer Study. Am J Epidemiol 2003; 158: 951-962.

[14] Weinberg CR and Sandler DP. Randomized recruitment in case-control studies. Am J Epidemiol 1991; 134: 421-432.
[15] Duell EJ, Wiencke JK, Cheng TJ, Varkonyi A, Zuo ZF, Ashok TDS, Mark EJ, Wain JC, Christiani DC and Kelsey KT. Polymorphisms in the DNA repair genes XRCC1 and ERCC2 and biomarkers of DNA damage in human blood mononuclear cells. Carcinogenesis 2000; 21: 965-971.
[16] Li Y, Millikan RC, Bell DA, Cui L, Tse CK, Newman B and Conway K. Cigarette smoking, cytochrome P4501A1 polymorphisms, and breast cancer among African-American and white women. Breast Cancer Res 2004; 6: R460-R473.
[17] Mechanic LE, Millikan RC, Player J, de Cotret AR, Winkel S, Worley K, Heard K, Heard K, Tse CK and Keku T. Polymorphisms in nucleotide excision repair genes, smoking and breast cancer in African Americans and whites: A population-based case-control study. Carcinogenesis 2006; 27: 1377-1385.
[18] Millikan RC, Pittman GS, Newman B, Tse CKJ, Selmin O, Rockhill B, Savitz D, Moorman PG and Bell DA. Cigarette smoking, N-acetyltransferases 1 and 2, and breast cancer risk. Cancer Epidemiol Biomarkers Prev 1998; 7: 371-378.
[19] Millikan R, Pittman G, Tse CK, Savitz DA, Newman B and Bell D. Glutathione S-transferases M1, T1, and P1 and breast cancer. Cancer Epidemiol Biomarkers Prev 2000; 9: 567-573.
[20] Millikan RC. NAT1 *10 and NAT1 *11 polymorphisms and breast cancer risk. Cancer Epidemiol Biomarkers Prev 2000; 9: 217-219.
[21] Millikan RC, Player J, de Cotret AR, Moorman P, Pittman G, Vannappagari V, Tse CK and Keku T. Manganese superoxide dismutase Ala-9Val polymorphism and risk of breast cancer in a population-based case-control study of African Americans and whites. Breast Cancer Res 2004; 6: R264-R274.
[22] Millikan RC, Player JS, DeCotret AR, Tse CK and Keku T. Polymorphisms in DNA repair genes, medical exposure to ionizing radiation, and breast cancer risk. Cancer Epidemiol Biomarkers Prev 2005; 14: 2326-2334.
[23] Pachkowski BF, Winkel S, Kubota Y, Swenberg JA, Millikan RC and Nakamura J. XRCC1 genotype and breast cancer: Functional studies and epidemiologic data show interactions between XRCC1 codon 280 his and smoking. Cancer Res 2006; 66: 2860-2868.
[24] Butler LM, Millikan RC, Sinha R, Keku TO, Winkel S, Harlan B, Eaton A, Gammon MD and Sandler RS. Modification by N-acetyltransferase 1 genotype on the association between dietary heterocyclic amines and colon cancer in a multiethnic study. Mutat Res Fundam Mol Mech Mutagen 2008; 638: 162-174.
[25] Butler LM, Sinha R, Millikan RC, Martin CF, Newman B, Gammon MD, Ammerman AS and Sandler RS. Heterocyclic amines, meat intake, and association with colon cancer in a population-based study. Am J Epidemiol 2003; 157: 434-445.
[26] Hernan MA, Hernandez-Diaz S, Werler MM and Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. Am J Epidemiol 2002; 155: 176-184.
[27] SAS Institute Inc.Cary NC. SAS 9.1.3. 2002.
[28] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70: 213-220.
[29] Flegal KM. The effects of changes in smoking prevalence on obesity prevalence in the United States. Am J Public Health 2007; 97: 1510-1514.
[30] Centers for Disease Control and Prevention NCfCDPaHP. Behavioral Risk Factor Surveillance System. Tobacco Use Data 2001 [cited 2009/10/03]. Centers for Disease Control and Prevention NCfCDPaHP 2009.
[31] Landis J and Koch G. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.
[32] Smith GD, Lawlor DA, Harbord R, Timpson N, Day I and Ebrahim S. Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology. PLoS Med 2007; 4: 1985-1992.
[33] Chatterjee N and Wacholder S. Invited commentary: Efficient testing of gene-environment interaction. Am J Epidemiol 2009; 169: 231-233.
[34] Smits KM, Benhamou S, Garte S, Weijenberg MP, Alamanos Y, Ambrosone C, Autrup H, Autrup JL, Baranova H, Bathum L, Boffetta P, Bouchardy C, Brockmoller J, Butkiewicz D, Cascorbi I, Clapper ML, Coutelle C, Daly AK, Muzi G, Dolzan V, Duzhak TG, Farker K, Golka K, Haugen A, Hein DW, Hildesheim A, Hirvonen A, Hsieh LL, Ingelman-Sundberg M, Kalina I, Kang D, Katoh T, Kihara M, Ono-Kihara M, Kim H, Kiyohara C, Kremers P, Lazarus P, Le Marchand L, Lechner MC, London S, Manni JJ, Maugard CM, Morgan GJ, Morita S, Nazar-Stewart V, Kristensen VN, Oda Y, Parl FF, Peters WHM, Rannug A, Rebbeck T, Pinto LFR, Risch A, Romkes M, Salagovic J, Schoket B, Seidegard J, Shields PG, Sim E, Sinnett D, Strange RC, Stucker I, Sugimura H, To-Figueras J, Vineis P, Yu MC, Zheng W, Pedotti P and Taioli E. Association of metabolic gene polymorphisms with

tobacco consumption in healthy controls. Int J Cancer 2004; 110: 266-270.
[35] Hamajima N, Yuasa H, Matsuo K and Kurobe Y. Detection of gene-environment interaction by case-only studies. Jpn J Clin Oncol 1999; 29: 490-493.
[36] Liu Y, Yoshimura K, Hanaoka T, Ohnami S, Ohnami S, Kohno T, Yoshida T, Sakamoto H, Sobue T and Tsugane S. Association of habitual smoking and drinking with single nucleotide polymorphism (SNP) in 40 candidate genes: Data from random population-based Japanese samples. J Hum Genet 2005; 50: 62-68.
[37] Hildesheim A, Anderson LM, Chen CJ, Cheng YJ, Brinton LA, Daly AK, Reed CD, Chen IH, Caporaso NE, Hsu MM, Chen JY, Idle JR, Hoover RN, Yang CS and Chhabra SK. CYP2E1 genetic polymorphisms and risk of nasopharyngeal carcinoma in taiwan. J Natl Cancer Inst 1997; 89: 1207-1212.
[38] Taylor JA, Umbach DM, Stephens E, Castranio T, Paulson D, Robertson C, Mohler JL and Bell DA. The role of N-acetylation polymorphisms in smoking-associated bladder cancer: Evidence of a gene-gene-exposure three-way interaction. Cancer Res 1998; 58: 3603-3610.
[39] Sugimura H, Wakai K, Genka K, Nagura K, Igarashi H, Nagayama K, Ohkawa A, Baba S, Morris BJ, Tsugane S, Ohno Y, Gao C, Li Z, Takezaki T, Tajima K and Iwamasa T. Association of Ile462Val (exon 7) polymorphism of cytochrome P450 IA1 with lung cancer in the Asian population: Further evidence from a case- control study in Okinawa. Cancer Epidemiol Biomarkers Prev 1998; 7: 413-417.
[40] Wu X, Shi H, Jiang H, Kemp B, Hong WK, Delclos GL and Spitz MR. Associations between cytochrome P4502E1 genotype, mutagen sensitivity, cigarette smoking and susceptibility to lung cancer. Carcinogenesis 1997; 18: 967-973.
[41] Egan KM, Newcomb PA, Titus-Ernstoff L, Trentham-Dietz A, Mignone LI, Farin F and Hunter DJ. Association of NAT2 and smoking in relation to breast cancer incidence in a population-based case-control study (United States). Cancer Causes Control 2003; 14: 43-51.
[42] Marcus PM, Hayes RB, Vineis P, Garcia-Closas M, Caporaso NE, Autrup H, Branch RA, Brockmoller J, Ishizaki T, Karakaya AE, Ladero JM, Mommsen S, Okkels H, Romkes M, Roots I and Rothman N. Cigarette smoking, N-acetyltransferase 2 acetylation status, and bladder cancer risk: A case-series meta-analysis of a gene-environment interaction. Cancer Epidemiol Biomarkers Prev 2000; 9: 461-467.
[43] Rothman K. Modern Epidemiology, 3rd edition. 2008.
[44] Gresner P, Gromadzinska J and Wasowicz W. Polymorphism of selected enzymes involved in detoxification and biotransformation in relation to lung cancer. Lung Cancer 2007; 57: 1-25.
[45] Nishikawa A, Mori Y, Lee IS, Tanaka T and Hirose M. Cigarette smoking, metabolic activation and carcinogenesis. Curr Drug Metab 2004; 5: 363-373.
[46] Hecht SS, Carmella SG, Yoder A, Chen M, Li ZZ, Le C, Dayton R, Jensen J and Hatsukami DK. Comparison of polymorphisms in genes involved in polycyclic aromatic hydrocarbon metabolism with urinary phenanthrene metabolite ratios in smokers. Cancer Epidemiol Biomarkers Prev 2006; 15: 1805-1811.
[47] David SP and Munafo MR. Genetic variation in the dopamine pathway and smoking cessation. Pharmacogenomics 2008; 9: 1307-1321.
[48] Breitling LP, Dahmen N, Illig T, Rujescu D, Nitz B, Raum E, Winterer G, Rothenbacher D and Brenner H. Variants in COMT and spontaneous smoking cessation: Retrospective cohort analysis of 925 cessation events. Pharmacogenet Genomics 2009; 19: 657-659.
[49] Omidvar M, Stolk L, Uitterlinden AG, Hofman A, Van Duijn CM and Tiemeier H. The effect of catechol-O-methyltransferase Met/Val functional polymorphism on smoking cessation: retrospective and prospective analyses in a cohort study. Pharmacogenet Genomics 2009; 19: 45-51.
[50] Chen X and Woodcroft KJ. Polymorphisms in metabolic genes CYP1A1 and GSTM1 and changes in maternal smoking during pregnancy. Nicotine Tob Res 2009; 11: 225-233.
[51] Bierut LJ. Convergence of genetic findings for nicotine dependence and smoking related diseases with chromosome 15q24-25. Trends Pharmacol Sci 2010; 31: 46-51.