

THE CASE-ONLY METHOD FOR GENE-ENVIRONMENT INTERACTION STUDIES:
THE INDEPENDENCE ASSUMPTION ILLUSTRATED WITH EMPIRICAL DATA
FROM THE PUBLISHED LITERATURE AND TWO POPULATION-BASED CONTROL
GROUPS, THE CAROLINA BREAST CANCER STUDY AND THE NORTH
CAROLINA COLON CANCER STUDY

M. Elizabeth Hodgson

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in
partial fulfillment of the requirements for the degree of Doctor of Philosophy in the
Department of Epidemiology, School of Public Health

Chapel Hill
2009

Submitted to:

Robert C. Millikan, PhD, DVM

Charles L. Poole III, ScD

Andrew F. Olshan, PhD

Kari E. North, PhD

Donglin Zeng, PhD

This dissertation is dedicated to my husband and sons
Rob, Carl and Max Larson
for their unwavering support and encouragement
and to my parents
Ernest and Mary K. Hodgson
for their unfailing love and inspiration.

ACKNOWLEDGEMENTS

Above all, I would like to thank my thesis advisor, Bob Millikan, for his insight, support and patience over the many years I have been his advisee. I may not have taken the road less traveled but I certainly have taken the long way around. And he has been there for me each step along the way, guiding and encouraging. I would like to thank Charlie Poole for many insightful conversations, both on the mysteries of meta-analysis and DAGs, and on the mysteries of the history and practice of epidemiology ‘in the trenches’. He is an impressive methodologist and a great storyteller, a wonderful combination. I would like to thank my other committee members, Andy Olshan, Kari North and Donglin Zeng. They have all contributed their unique expertise to my training. Their contributions run the gamut from helping me see the context and implications of epidemiology, to helping me see how the smallest details can contribute to excellence. I thank them all and will carry these lessons forward.

I would like to thank Jessica Tse, without whose programming support I would still be wrestling with SAS. Her unfailing generosity and helpfulness have been absolutely crucial to the completion of this dissertation. Rob Larson provided administrative support and was my personal Excel guru. No one can get more work out of a spreadsheet than he can and I benefitted greatly from his expertise.

Thao Vo, Vani Vannapagari and Sandy Deming have been invaluable colleagues and friends. Like Bob, they have been there for the long road, and provided every kind of

support an epidemiologist in training could want – insightful methodological and substantive discussions, meticulous editing, constant encouragement, as well as counsel and hand-holding at crucial moments.

Finally, I would like to thank all the study participants. Without them, none of this would be possible. I hope to be worthy of their trust.

ABSTRACT

M. ELIZABETH HODGSON: The Case-only Method for Gene-Environment Interaction Studies: The Independence Assumption Illustrated with Empirical Data from the Published literature and Two Population-based Control groups, the Carolina Breast Cancer Study and the North Carolina Colon Cancer Study

(Under the direction of Robert C. Millikan)

Gene-environment interaction in the etiology of disease is a topic of on-going interest. While there has been increasing use of the case-only study design to investigate gene-environment interaction in cancer, as well as other disease areas, concerns about the underlying assumption that the genetic and environmental exposures are independent in the underlying population (the independence assumption) have not been adequately addressed. The case-only study design requires only cases, no population controls or cohort, to estimate statistical interaction. This design has obvious cost advantages, as well as some methodological and ethical advantages. However, for results to be valid the independence assumption must be met. There has been little investigation into the frequency and magnitude of independence assumption violation for DNA repair genes and smoking, an interaction of particular interest in cancer. Nor have optimal methods for validating the independence assumption received much attention.

Empirical data of two types were used to evaluate the independence assumption for selected genetic variants and smoking behavior. A systematic review of the literature identified 55 studies that presented the joint distribution of smoking and SNPs in 3 DNA repair genes (*XRCC1* Arg399Gln, Arg194Trp, or Arg280His, *XPB* Lys751Gln, and

Asp312Asn, and *XRCC3* Thr241Met). Measures of smoking were ever/never smoking, current/not current smoker, duration of smoking (≤ 10 years, 11-20 years, > 20 years), intensity ($< 1/2$ pack/day, $1/2$ -1 pack/day, > 1 pack/day), and pack-years (≤ 35 pack-years, > 35 pack-years). The odds ratio for SNP-smoking association in controls (OR_z) was used to estimate the gene-environment association in the underlying population. Results showed that OR_z was not reliably null for any of the SNP-smoking combinations. Studies with *XRCC1* 399 / ever-never smoking and *XPB* 751 / pack-years were too heterogeneous for summary estimates [ranges, OR_z (95% confidence interval (CI)): 0.7 (0.4, 1.2) – 1.9 (1.2, 2.8) and 0.8 (0.5, 1.3) – 2.3 (0.8, 6.1), respectively]. In addition, estimates for studies considered homogeneous (Cochran's Q p-value < 0.10) varied 2- to 5-fold within meta-analysis. No study characteristics were identified that could explain heterogeneity.

Data from two population-based control groups, the Carolina Breast Cancer Study and the North Carolina Colon Cancer Study, were used to evaluate the independence assumption for smoking and a panel of eight metabolic and 26 DNA repair genes plausibly related to smoking behavior. OR_z was not consistent across smoking measures precluding the use of one smoking measure (e.g. ever-never) as a substitute for evaluating other measures such as duration and dose. In particular, results for smoking status were most often near the null, while measures of smoking amount for the same SNPs were of sufficient magnitude to cause appreciable bias in the case-only estimates ($OR_z \leq 0.7$ or ≥ 1.4) approximately half of the time. There were no strong patterns of the magnitude or direction of OR_z differing by race, age, gender or biological pathway (xenobiotic metabolism, DNA repair).

Taken together, results suggest that OR_z should be considered population-specific. Therefore, the independence assumption should be evaluated in the population underlying a case-only study, rather than in a proxy control group(s) or pooled controls. A systematic search for relevant literature and control data, in addition to a comprehensive evaluation of all smoking measures used in the case-only analysis are essential for evaluation of the independence assumption.

TABLE OF CONTENTS

ABSTRACT.....	v
TABLE OF CONTENTS.....	viii
LIST OF TABLES.....	ix
LIST OF FIGURES	xi
LIST OF ABBREVIATIONS.....	xii

Chapter

I. INTRODUCTION.....	1
II. REVIEW OF THE LITERATURE	5
A. Published case-only analyses.....	5
B. Validity of independence assumption.....	9
C. The independence assumption in empirical data.....	17
D. Effect of independence assumption violation in data simulation	22
E. Independence assumption verification methods.....	23
III. STATEMENT OF SPECIFIC AIMS	26
A. Specific aims.....	26
B. Hypotheses.....	28
C. Rationale.....	29
1. Smoking and genes	30
2. Meta-analyses	39
3. CBCS and NCCCS control groups	40
IV. METHODS.....	45
A. Overview	45
B. Literature-based analysis of the independence assumption.....	46
1. Overview.....	46
2. Literature Search.....	49
3. Data abstraction	53
4. Meta-analyses	54
5. Study characteristic analysis	57
C. Gene-smoking association in CBCS and NCCCS controls.....	59
1. Overview.....	59
2. Study populations.....	59
3. Statistical methods: Estimation and evaluation of OR_z	60
4. Agreement.....	68
5. Misspecification of smoking.....	69

V.	RESULTS	71
A.	MANUSCRIPT 1: Smoking and selected DNA repair gene polymorphisms in control groups: systematic review and meta-analysis.....	71
1.	Introduction.....	71
2.	Methods.....	76
3.	Results.....	82
4.	Discussion.....	87
5.	Tables and Figures	95
B.	MANUSCRIPT 2: Association of DNA repair and metabolic gene polymorphisms with tobacco smoking in controls from two population-based case-control studies: Carolina Breast Cancer Study and North Carolina Colon Cancer Study	147
1.	Introduction.....	147
2.	Methods.....	150
3.	Results.....	154
4.	Discussion.....	162
5.	Tables and Figures	173
VI.	CONCLUSIONS AND DISCUSSION	213
A.	Findings and implications for stand-alone case-only studies	213
B.	Strengths of the systematic review and meta-analysis	217
C.	Limitations of the systematic review and meta-analysis	218
D.	Strengths of the control group analyses.....	219
E.	Limitations of the control group analyses	220
F.	Future directions	221
VII.	APPENDICES	224
A.	Informed consent	224
B.	Supplementary tables and figures: Manuscript 1	225
C.	Supplementary tables and figures: Manuscript 2	232
VIII.	REFERENCES	236

LIST OF TABLES

IV.1	SNP Designations for Data Abstraction.....	52
V.A.1a.	Characteristics of 50 Studies Included in Meta-analyses (46 study populations).....	95
V.A.1b.	Characteristics of 5 Additional Study Populations Not in Any Meta-analyses.....	110
V.A.2	DNA Repair Gene Variation and Smoking: Summary Estimates	112
V.A.3	Genotype-Smoking Associations for Selected Specifications of Current-Not current Smoking and Pack-Years of smoking	114
V.A.4.	Genotype-Smoking Associations Stratified by Study Design	116
V.A.5.	<i>XRCC1</i> Arg399Gln and Smoking: Overall and by Study Characteristics	123
V.A.6.	<i>XPB</i> Lys751Gln and Smoking: Overall and by Study Characteristics	130
V.A.7.	<i>XRCC3</i> Thr241Met and Smoking: Overall and by Study Characteristics	137
V.B.1.	Characteristics of CBCS and NCCCS control groups	173
V.B.2.	Gene variants in CBCS and NCCCS.....	175
V.B.3a.	Gene variant-smoking status associations in the CBCS, overall and by race	179
V.B.3b.	Gene variant-smoking duration association in the CBCS, overall and by race....	183
V.B.3c.	Gene variant-smoking intensity association in the CBCS, overall and by race....	185
V.B.3d.	Gene variant-PY association in the CBCS, overall and by race.....	187
V.B.4a.	Gene variant-smoking status associations in the NCCCS, overall, by gender and by race.....	189
V.B.4b.	Gene variant-smoking status associations in the NCCCS, overall, by gender and by race.....	191

V.B.4c. Gene variant-smoking duration association in the NCCCS, overall and by gender and race	193
V.B.4d. Gene variant-smoking intensity association in the NCCCS, overall and by gender and race.....	195
V.B.4e. Gene variant-pack-years of smoking association in the NCCCS, overall and By gender and race.....	198
V.B.5: Gene variant-smoking associations in CBCS and NCCCS.....	200
V.B.6. Gene-variant - smoking associations in CBCS & NCCCS: Non-African American female controls 40-74 years of age	202
V.B.7. Gene variant-smoking status association in GSEC, CBCS and NCCCS Controls.....	206
V.B.8. Misspecification for gene variant-smoking status associations (OR_z) in the CBCS and NCCCS.....	208
V.B.9. Agreement between CBCS and NCCCS gene variant-smoking associations.....	212
VIII.B.1. Association between <i>XRCC1</i> Arg399His and smoking: Individual study results.....	225
VIII.B.2. Association between <i>XRCC1</i> Arg194Trp and smoking: Individual study results.....	227
VIII.B.3. Association between <i>XRCC1</i> Arg280His and smoking: Individual study results	228
VIII.B.4. Association between <i>XPB</i> Lys751Gln and smoking: Individual study results.....	229
VIII.B.5. Association between <i>XPB</i> Asp312Asn and smoking: Individual study results	230
VIII.B.6. Association between <i>XRCC3</i> Thr41Met and smoking: Individual study results	231
VIII.C.1. Genotype prevalence and Hardy-Weinberg equilibrium in CBCS and NCCCS.....	232

LIST OF FIGURES

V.A.1. Weighted forest plot for OR_z for <i>XRCCI</i> 399 and ever-never smoking	145
V.A.2. Funnel plot for OR_z for <i>XRCCI</i> 399 and ever-never smoking.....	146
VIII.C.3. Directed Acyclic Graph	232

LIST OF ABBREVIATIONS

Gene abbreviations

<i>5-HTT</i>	solute carrier family 6 (neurotransmitter transporter, serotonin) (also known as <i>SLC6A4</i>)
<i>5-HTTLPR</i>	serotonin-transporter-linked polymorphic region
<i>ADH1A</i>	alcohol dehydrogenase 1A (class I), alpha polypeptide
<i>ADH1B</i>	alcohol dehydrogenase 1B (class I), beta polypeptide
<i>ADH1C</i>	alcohol dehydrogenase 1C (class I), gamma polypeptide
<i>ADPRT</i>	poly (ADP-ribose) polymerase 1 (now <i>PARP1</i>)
<i>ADPRTL2</i>	poly (ADP-ribose) polymerase 2 (now <i>PARP2</i>)
<i>ALDH2</i>	aldehyde dehydrogenase 2 family (mitochondrial)
<i>APE1</i>	apurinic / apyrimidinic endonuclease
<i>ATM</i>	ataxia telangiectasia mutated
<i>BDNF</i>	brain-derived neurotrophic factor
<i>BRCA1</i>	breast cancer 1, early onset
<i>BRCA2</i>	breast cancer 2, early onset
<i>CDH1</i>	cadherin 1, type 1, E-cadherin (epithelial)
<i>CHRNA4</i>	cholinergic receptor, nicotinic, alpha 4
<i>COMT</i>	catechol-O-methyltransferase
<i>CSB</i>	chorionic somatomammotropin hormone 2 (also known as <i>CSH2</i>)
<i>CYP1A1</i>	cytochrome P450, family 1, subfamily A, polypeptide 1
<i>CYP1B1</i>	cytochrome P450, family 1, subfamily B, polypeptide 1
<i>CYP2A6</i>	cytochrome P450, family 2, subfamily A, polypeptide 6
<i>CYP2C9</i>	cytochrome P450, family 2, subfamily C, polypeptide 9

<i>CYP2C19</i>	cytochrome P450, family 2, subfamily C, polypeptide 19
<i>CYP2E1</i>	cytochrome P450, family 2, subfamily E, polypeptide 1
<i>CYP17A</i>	cytochrome P450, family 17, subfamily A, polypeptide 1
<i>CYP19A1</i>	cytochrome P450, family 19, subfamily A, polypeptide 1
<i>DRD2</i>	dopamine receptor D2
<i>DRD3</i>	dopamine receptor D3
<i>DRD4</i>	dopamine receptor D4
<i>EPHX1</i>	epoxide hydrolase 1, microsomal (xenobiotic)
<i>EPHX2</i>	epoxide hydrolase 2, cytoplasmic
<i>ERCC1</i>	excision repair complementing defective 1
<i>ERCC2</i>	excision repair complementing defective 2 (formerly <i>XPD</i>)
<i>ERCC6</i>	excision repair complementing defective 6
<i>ESR1</i>	estrogen receptor 1
<i>ESR2</i>	estrogen receptor 2 (ER beta)
<i>GST</i>	glutathione S-transferase
<i>GSTM1</i>	glutathione S-transferase mu 1
<i>GSTM2</i>	glutathione S-transferase mu 2 (muscle)
<i>GSTM3</i>	glutathione S-transferase mu 3(brain)
<i>GSTP1</i>	glutathione S-transferase pi 1
<i>GSTT1</i>	glutathione S-transferase theta 1
<i>GSTT2</i>	glutathione S-transferase theta 2
<i>hOGG1</i>	human 8-oxoguanine DNA glycosylase (now <i>OGG1</i>)
<i>HSP</i>	heat shock protein

<i>MAO-A</i>	monoamine oxidase A
<i>MAO-B</i>	monoamine oxidase B
<i>MEH</i>	microsomal epoxide hydrolase
<i>MGMT</i>	O-6-methylguanine-DNA methyltransferase
<i>MLH1</i>	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
<i>MSH3</i>	mutS homolog 3 (E. coli)
<i>MSH5</i>	mutS homolog 5 (E. coli)
<i>MnSOD</i>	manganese superoxide dismutase
<i>MPO</i>	myeloperoxidase
<i>MTHFR</i>	5,10-methylenetetrahydrofolate reductase (NADPH)
<i>MYH</i>	mutY homolog (E. coli) (also known as <i>MUTYH</i>)
<i>NAT1</i>	N-acetyltransferase 1 (arylamine N-acetyltransferase)
<i>NAT2</i>	N-acetyltransferase 2 (arylamine N-acetyltransferase)
<i>NBS1</i>	Nijmegen breakage syndrome; nibrin (now <i>NBN</i>)
<i>NOS2A</i>	nitric oxide synthase 2, inducible (now <i>NOS2</i>)
<i>NOS3</i>	nitric oxide synthase 3 (endothelial cell)
<i>NQO1</i>	NAD(P)H dehydrogenase, quinone 1
<i>NR3C1</i>	nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor)
<i>NUDT1</i>	nudix (nucleoside diphosphate linked moiety X)-type motif 1
<i>OGG1</i>	8-oxoguanine DNA glycosylase
<i>PGR</i>	progesterone receptor
<i>POLD</i>	polymerase (DNA directed), delta 1, catalytic subunit 125kDa (now <i>POLD1</i>)
<i>PPARG2</i>	peroxisome proliferator-activated receptor gamma

<i>RAD23B</i>	<i>RAD23</i> homolog B (<i>S. cerevisiae</i>)
<i>SLC6A4</i>	solute carrier family 6 (neurotransmitter transporter, serotonin) (also known as <i>5-HTT</i>)
<i>SULT1A1</i>	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1
<i>XPC</i>	xeroderma pigmentosum, complementation group C
<i>XPB</i>	xeroderma pigmentosum complementation group D (now <i>ERCC2</i>)
<i>XPF</i>	excision repair cross-complementing rodent repair deficiency, complementation group 4 (now <i>ERCC4</i>)
<i>XPB</i>	excision repair cross-complementing rodent repair deficiency, complementation group 5 (now <i>ERCC5</i>)
<i>XRCC1</i>	X-ray cross complementing gene 1
<i>XRCC2</i>	X-ray cross complementing gene 2
<i>XRCC3</i>	X-ray cross complementing gene 3
<i>XRCC4</i>	X-ray cross complementing gene 4

Other abbreviations

BER	Base excision repair
BMI	Body mass index
BRFSS	Behavioral Risk Factor Surveillance System
CBCS	Carolina Breast Cancer Study
CDC	Centers for Disease Control
CI	Confidence interval
CIS	Carcinoma <i>in situ</i>
COR	Case-only odds ratio
COV	Covariate
CS	Cockayne syndrome
DAG	Directed acyclic graph
DMV	Department of Motor Vehicles
DNA	Deoxyribonucleic acid
DSB	Double strand break
E	Environmental exposure (E+ =exposed, E- =unexposed)
ETS	Environmental tobacco smoke (“passive” smoking)
G	Genetic exposure (G+ =exposed, G- =unexposed)
G-E	Gene-Environment association
GSEC	Collaborative Study on Genetic Susceptibility to Environmental Carcinogens
GWAS	Genome wide association scan
GxE	Gene-Environment interaction

HCFA	Health Care Finance Administration
HDL	High-density lipoprotein
HuGE	Human Genome Epidemiology
HWE	Hardy Weinberg equilibrium
IBC	Invasive breast cancer
IOR	Interaction odds ratio
LD	Linkage disequilibrium
MAF	Minor allele frequency
N	Number of observations
NCCCS	North Carolina Colon Cancer Study
NER	Nucleotide excision repair
NNK	4-(methylnitrosamino)- 1-(3-pyridyl)-1-butanone
NNAL	4-(methylnitrosamino)-1-(3-pyridyl)-1-butanol
OR	Odds ratio
OR _z	Odds ratio in controls
PAH	Polycyclic aromatic hydrocarbons
PSA	Prostate specific antigen
PY	Pack-years
Q (Cochran's)	Statistic for Cochran's test of homogeneity
ROR	Ratio of ratios
ROS	Reactive oxygen species
RR	Relative risk
SES	Socioeconomic status

SIM	Synergy index on a multiplicative scale
SNP	Single nucleotide polymorphism
XP	Xeroderma pigmentosum
Z	Gene-environment association in the underlying population

I. INTRODUCTION

The case-only study design as proposed by Prentice (1984) and popularized by Piegorsch (1994) and Khoury (1996) has become increasingly popular over the last decade, especially for studies of gene-environment interaction (GxE) in cancer. It is used in epidemiologic studies to estimate the magnitude of statistical interaction between two measured exposures with respect to a given outcome [1-3]. This method requires only cases, no population controls or cohort. Provided the design assumptions are met, the case-only study can estimate statistical interactions that deviate from the multiplicative null. The relationship between gene-environment interaction estimated by the case-only odds ratio (COR) and the same gene-environment interaction estimated by a case-control study can be expressed as follows (OR=odds ratio):

$$OR_{\text{gene*env, case-only}} = OR_{\text{gene*envr, case-control}} / (OR_{\text{gene, case-control}} * OR_{\text{envr, case-control}}) * Z$$

where Z (estimated by OR_z) is the association between the gene and the environmental exposure in the control group of a case-control study [3]. The quantity $[OR_{\text{gene*envr, case-control}} / (OR_{\text{gene, case-control}} * OR_{\text{envr, case-control}})]$ is sometimes referred to as the synergy index on a multiplicative scale, or synergy index on a multiple scale (SIM). When there is no association between the genetic exposure and the environmental exposure in the population (i.e. $Z=1$), the case-only OR is equivalent to the (multiplicative) deviation from a (perfectly) multiplicative relationship between the genetic and environmental exposures (i.e. $COR = SIM$). Using these abbreviations, the relationship can be expressed succinctly as:

$$COR = SIM * OR_z.$$

There are a number of possible causal and non-causal reasons for Z to take on values other than one. Values of Z greater or less than 1 can be due to a biological relationship between the gene and the exposure, either because the polymorphism itself is a causal variant of the gene or because it is in linkage disequilibrium (LD) with a causal variant. One non-causal reason for Z to vary from the null is that environmental and genetic exposures may have been non-differentially misclassified with respect to each other, as can happen when population stratification is present [4]. Non-random misclassification of either the genetic exposure (e.g. through linkage disequilibrium) or the environmental exposure (e.g. heavy smokers underreport smoking more than light smokers) can also create apparent association in a study population. Selection bias can also cause association between two exposures in a study population. For instance, if smokers with a family history of the outcome are less likely to participate as controls than smokers without a family history or non-smokers with a family history, a spurious inverse control group association could be created between smoking and any genetic exposure related to family history.

Conversely, a positive association could be seen in the study, even though there is no association in the underlying population, if smokers with a family history are more likely to participate than non-smokers with a family history or smokers without a family history. Cohort effects could affect control group associations if, for instance, a genetic exposure is associated with longevity, and the environmental exposure is one that has changed prevalence in the population over time, such as smoking or dietary patterns. Chance can also play a role. Since the expectation that $Z=1$ is a large sample asymptotic approximation, as sample size decreases, Z will deviate from the null with increasing frequency through random error alone [5]. Consequently, as Z is evaluated in subgroups, and sample size drops, Z can deviate from

unity by chance alone. Further, as sample size decreases, the power to detect interaction also drops sharply [6]. The assumption that $Z=1$ can only be evaluated if both exposures have been measured in the population at risk or, in the context of a case-control study, in an appropriate control population.

There are clear advantages to the case-only method in several settings. The lack of requirement for a control group reduces costs, but there are methodological and ethical advantages as well. Differential recruiting success between cases and controls raises questions of selection bias and the difficulty of establishing an appropriate control group for hospital-based studies of rare diseases is well known [7]. Because only cases are used in the analysis, recall bias generated by differential recall between cases and controls cannot affect case-only studies, although differential recall among cases by genetic and/or environmental factors is still possible. Estimation of the interaction parameter from case-only analyses is more efficient than for a traditional case-control study (i.e. fewer cases are required for similar precision of estimate) [8]. Invasive procedures that are part of cases' diagnosis or treatment often cannot be done ethically in healthy volunteers, especially with vulnerable populations such as children [9]. Additionally, the cost/benefit balance for controls in a study that collects genetic information is different than for a study that does not collect genetic information. That is, there are potentially greater costs (e.g. potential misuse of information, potential for unwanted information about genetically related individuals to be revealed etc.) for the same benefits.

Further, questions of sample size have a strong albeit controversial ethical dimension, with some arguing that smaller studies are more ethical due to a more favorable risk/benefit ratio [10-11], and others arguing that under-powered studies are unethical [12-13]. This is

particularly relevant for gene-environment interaction research, where the traditional case-control approach requires large sample sizes, and there is ongoing interest in exploring valid alternative methodologies such as sequential testing, or case-only studies [3, 14-15].

But these advantages come at a cost. A case-only study only estimates interaction on a multiplicative scale, and cannot estimate the independent effect of either exposure, or additive joint effects, limiting its use to situations where the independent or additive effects of the two exposures are not of interest. However, where independent effects are already well described (e.g. smoking and lung cancer) or thought to be negligible (e.g. low penetrance polymorphic genes) this may still be an attractive design [2]. It has been proposed as a screening method to identify candidate genes, or gene-environment or gene-gene interactions that may be etiologically important for further investigation [5, 16-17].

In addition to limits on the estimates that can be obtained, the validity of case-only studies is limited by multiple design assumptions. Many are common to all epidemiologic study designs; no misclassification of exposure or disease, no selection bias, no uncontrolled confounding and a sufficient sample size are examples. In addition, however, the validity of the case-only estimate of interaction rests on the assumption that the two exposures are independent in the population from which the cases arose [2], referred to henceforth as the independence assumption.

II. REVIEW OF THE LITERATURE

A. Published case-only analyses

To date, numerous studies have been published using the case-only method of assessing interaction. Although the case-only study design is theoretically applicable to studying statistical interaction between any two exposures for a given case definition, it has been proposed as particularly useful for gene-environment interaction or gene-gene interaction [3]. The distribution of published reports bears this out. Of the 27 case-only interaction studies found in PubMed from Jan 1, 2007- Sept 21, 2009 [search term: “case-only” , Limits: English, Human], 25 contained assessments of gene-environment interaction, with strongest interest in cancer outcomes (18 publications). Two included gene-gene interaction; four included environment-environment interaction where environment was broadly defined as any non-genotypic factor.

Of the 15 most recent case-only interaction analyses published (2008-2009) 11 were nested in existing case-control studies, one was nested in a cohort although genetic data were not available for non-cases and three were stand-alone case-only studies (no controls or cohort). For approximately half of the nested analyses, both the case-only estimates of interaction (COR) and case-control estimates of interaction [SIM or interaction odds ratio (IOR)] were presented. Case-only analyses fully nested in case-control studies (also called adjunct case-only analyses) are generally performed to take advantage of the increased precision afforded by the case-only approach, or address potential shortcomings in the controls, such as differential recall between cases and controls [18] or a low response rate in controls [19]. In fact, over the last decade, at least two variations on the case-only approach

have been implemented: partially nested extensions of case-control studies, (additional cases added to the cases in a case-control study) [20-22], interaction of time-varying population-level factors and fixed individual-level factors [23-24].

Smoking behavior and/or tobacco use is the single most frequently examined environmental exposure in case-only interaction analyses. It was assessed in approximately half of the case-only analyses in the last two years, most often in conjunction with xenobiotic metabolizing genes [e.g. *GSTs* (glutathione S-transferases), *CYP1A1* (cytochrome P450, family 1, subfamily A, polypeptide 1)]. DNA repair genes are often examined in traditional interaction studies, frequently with smoking, however, there have only been two case-only interaction analyses of DNA repair genes between Jan 1, 2007 and Sept 23, 2009 and neither examined smoking [25-26].

In the recent literature (2008-2009), in addition to nested case-only studies, there have been three stand-alone case-only analyses (i.e. no controls, no relevant data for controls and/or no case-control estimates presented [25, 27-28]. Studies designed as case-only (no controls) have been employed to address a range of issues beyond increased precision or reduced cost. Case-only studies can address the ethical problem of carrying out invasive or frightening exposure measurements on healthy participants, particularly children [9, 25, 29-32]. They have been used to examine interaction when a control group is not easily identifiable, as in the case of very rare diseases where cases are collected over several population [33-34], where appropriate controls are prohibitively expensive to identify [35] or for special populations such as centenarians [16, 36]. When the genetic exposure is both rare and highly penetrant, such as *BRCA1/2*, it may be prohibitively difficult to collect sufficient controls for interaction analyses [34, 37-39].

An important distinction among the case-only studies of gene-environment interaction is their approach to verifying the source population assumption of independence. At the extremes, approaches range from no explicit mention of the independence assumption [19, 37, 40] to assessment in a sample of a geographically and demographically similar population [25, 41]. Justification is often based on the plausibility of the independence assumption alone [9, 18, 27, 34, 42-48]. However, a number of studies have undertaken more quantitative evaluations of the independence assumption [20-22, 25-26, 35, 41, 49-53]. Among the case-only studies published in 2008-2009, only one presented the control-only estimates (OR_z) for the relevant analyses [35]. Two of the three stand-alone case studies justified the independence assumption: 1) Smits (2008) referenced a large study of pooled GSEC controls [28, 54] and 2) Yang (2008) used subjects “randomly selected from the same population” [25]. Not surprisingly, most case-only studies that presented a quantitative or semi-quantitative assessment of the independence assumption were at least partially nested within case-control studies. Although nested case-only studies can assess the independence assumption most rigorously, assuming the control group adequately represents the underlying population, the fact that they have a control group means that cannot realize the full cost or ethical advantages that help make the case-only study design attractive.

A number of approaches have been taken for quantitative assessment of the independence assumption, whether the independence assumption is being evaluated in study control groups and/or in ancillary data (i.e. data external to the published study), although all approaches have ultimately relied almost exclusively on statistical significance. Some studies assessed the independence assumption using the χ^2 test for categorical variables at $\alpha=0.05$, while others simply stated that no significant associations were found, without specifying

method of assessment. Few studies provided information on the magnitude of any associations between genotype and environmental exposure from control groups or ancillary data. However, even in the most thorough presentations of control-only data, statistical significance is the paramount concern. For example, in Egan (2003), where equivalently adjusted case-only and control-only analyses of each categorization of environmental exposure and subgroup examined were presented side by side [52], the magnitude of the OR_z s in controls varied from 0.5 to 1.1, yet only the sole statistically significant association was considered problematic. Similarly, data presented in Marcus et. al. (2000) allowed calculation of control group OR_z s for each study included in the pooled analysis, which showed wide variation in the magnitude of OR_z s (0.5 - 1.8).

Stand-alone case-only studies often do not present any quantitative assessment of the independence assumption. One stand-alone study to do so calculated unadjusted OR_z (95%CI)s from published control group cross-classifications of genotype [*CYP1B1* Val432Leu, catechol-O-methyltransferase (*COMT*) Val108Met and sulfotransferase 1A, member 1 (*SULT1A1*) Arg213His] and environmental exposure (smoking, ever/never) and presented an OR_z (95% CI) for each of the 3 associations studied [53]. All associations between ever smoking and variant genotype were weakly inverse and none were statistically significant: OR_z (95%CI) = 0.77 (0.19, 3.10) for *CYP1B1* (Leu/Leu vs. any Val), 0.90 (0.45, 1.81) for *COMT* (Val/Val vs. any Met), and 0.72 (0.38, 1.34) for *SULT1A1* (Arg/Arg vs. any His). The control group associations for *CYP1B1* and *COMT* were from a population-based study of ovarian cancer conducted in Hawaii among women of 3 different ethnicities ($N_{\text{control}}=144$) [55]; the *SULT1A1* control group association examined was from a study of

lung cancer conducted in Texas that used managed care enrollees as controls ($N_{\text{control}}=444$) [56].

Since publication of the Saintot et. al. (2003) study, an Italian hospital-based study of male bladder cancer ($N=214$) has been published with appropriate control group data to assess the independence assumption for *CYP1B1* Val432Leu and smoking [57]. When control group associations were calculated for *CYP1B1* genotype as Val/Val (ref) vs. any Leu, however, the $OR_z(95\%CI)$ for *CYP1B1* and smoking are 1.4(0.6, 3.3), 1.2(0.5, 3.2) and 1.7(0.6, 4.7) for never/ever, never/light smoking, and never/heavy smoking respectively. With statistical significance as the sole criterion, the independence assumption would be met in all cases. However, insofar as the associations in these ancillary data accurately represent associations in the underlying population in the Saintot et. al. study, this shows that the COR for smoking and *CYP1B1* Val432Leu genotype in Saintot (2003) could be biased to a greater or lesser degree, and in either direction, depending on the specific case-only analysis done (genotype category, smoking categories etc.). In addition, other criteria may need to be examined for control populations including, but not limited to, existence of Hardy-Weinberg equilibrium in controls, demographic similarity to case-only population (gender, age, ethnicity etc.) and study design.

B. Validity of independence assumption

Gene-environment associations in populations can be causal or non-causal. When the ‘implausibility’ of specific G-E association is argued in published case-only analyses, however, it is generally considered only within the framework of causality. Interest in genetic influences on behavioral traits is long-standing and there is an extensive and increasing literature in behavioral genetics. One essay on the future of behavioral genetics in the era of

genomics states that ‘nearly all behavioral variation reflects some genetic influence’ [58]. While there are currently only a limited number of established associations between genotype and behavior, research in the areas of personality, psychiatric disorders and addiction is flourishing [59-61]. Genes being investigated for impact on behaviors that influence health outcomes include: monoamine oxidase A (*MAO-A*) and antisocial behavior, serotonin transporter (*SLC6A4*) and anxiety and depression, *COMT* and frontal lobe function, brain-derived neurotrophic factor (*BDNF*) and long-term memory, and dopamine receptor D2 (*DRD2*) and substance abuse, gambling and alcoholism, among many others [60-61]. Most of these genes, if not all, function in multiple pathways with poorly characterized effects. *COMT*, for instance, because of its role in the dopamine pathway and hormone metabolism has been studied in conjunction with schizophrenia [62], attention deficit hyperactivity disorder [63], smoking [64], alcoholism [65], cataract [66], Alzheimer disease [67], and breast [68-69], ovarian [55], hepatocellular [70], and bladder cancer [71]. These are widely divergent health outcomes and many have strong behavioral and exposure-related components.

Given the current limited state of knowledge of genetic influences on health-related behaviors and exposures, and the wide variety of gene-behavior associations considered plausible enough to be under investigation, it seems unwise to argue the validity of the independence assumption based entirely on the implausibility of a causal association. A more prudent approach would be to thoroughly examine any empirical evidence for or against causal association between the relevant gene and exposure before proceeding. In this section, I will discuss two examples of gene-exposure association that there is empirical evidence for,

in order to illustrate various ways gene-exposure association can be problematic for case-only analysis, and for the process of evaluating the independence assumption.

The strongest example of an independence assumption violation is the association between aldehyde dehydrogenase 2 (*ALDH2*) genotype and alcohol consumption [72]. This association in the underlying population means that a case-only study of the interaction of *ALDH2* and alcohol consumption would be invalid. The *ALDH2*1* allele codes for normally functioning aldehyde dehydrogenase, a rate-limiting enzyme in the ethanol metabolism pathway. *ALDH2*2* is a variant allele coding for a much lower activity form of aldehyde dehydrogenase. Individuals homozygous for the *ALDH2*2* allele experience flushing, tachycardia, headache and nausea after consuming alcohol. Consequently, these individuals tend not to consume alcohol, and have virtually no risk of alcoholism [73-74]. Heterozygotes (*ALDH2*1/2*) can also experience aversive reactions, although with widely varying severity. Consequently, individuals heterozygous for *ALDH2* tend to consume less alcohol overall [75], consume fewer drinks at one sitting and engage in binge drinking less often than *ALDH2*1/1* homozygotes [76]. Additionally, research on the subjective experience of alcohol consumption demonstrates that reactions vary by *ALDH2* genotype. In a sample of college-age Asian men and women with equivalent blood alcohol levels, heterozygotes (*ALDH2*1/2*) reported a more intense subjective reaction to alcohol, as well as more flushing and higher cortisol levels, than those homozygous for the wild type allele (*ALDH2*1/1*) [72, 77]. In another study, participants rated a panel of subjective responses to alcohol consumption with heterozygous individuals reporting more dizziness, higher intensity of effect and more facial warming than *ALDH2*1/1* individuals [78]. Despite the protection from alcoholism that aversion to excess and/or habitual consumption of alcohol provides, it has been shown that

individuals with the *ALDH2**2 allele who do consume alcohol are at higher, rather than lower, risk of many of the negative consequences of alcohol consumption. These can include neurocognitive impairments [79] and esophageal cancer [80]. It has been hypothesized that this occurs through the high levels of acetaldehyde built up after drinking [78, 81], leading to increased oxidative stress[82].

This *ALDH2* polymorphism has also been variously associated with poor glycemic control in Type II diabetics who are light to moderate drinkers [83], gout [84], and HDL [85], and cortisol [86] and lipid peroxide responses [82] to alcohol consumption. Any one of these environmental exposures could be a plausible candidate for a future gene-environment interaction study with *ALDH2*, either directly or by proxy, for a number of common disease outcomes (e.g. cardiovascular disease, cancers). For example, the interaction of alcohol metabolizing genes and stress (whether measured by blood cortisol levels, life events questionnaires or other means) would be of interest for studies of cardiovascular disease, breast cancer and esophageal cancer. The interaction of alcohol metabolizing genes and glycemic control would be of interest in studies of insulin resistance, time to initiation of insulin use or severity of Type II diabetes. Association in the control group between the *ALDH2* polymorphism and any of these exposures has the potential to violate the independence assumption and invalidate or bias a case-only analysis of that interaction.

Although the *ALDH2*-alcohol consumption association is well documented and the non-independence of these two exposures clearly makes a case-only analysis of interaction inappropriate, researchers more often find the association of interest less well understood. For example, the association (or lack thereof) between *CYP2A6* and various aspects of smoking behavior has received considerable attention in the last 15 years [87-90]. *CYP2A6* is a

polymorphic gene coding for the enzyme (cytochrome P450 2A6) that is responsible for the bulk (~80%) of the biotransformation of nicotine to cotinine (the major breakdown product of nicotine), then on to 3'-hydroxycotinine [91-92]. As of 2002, there were 10 known variants of *CYP2A6*, coding for enzymes with no activity (or deletions), reduced activity or enhanced activity [93].

Because people usually smoke to raise nicotine levels in the blood and brain, it has been proposed that fast metabolizers of nicotine need to smoke more than slow metabolizers to achieve the same steady-state nicotine levels [94-96]. The relationship between *CYP2A6* polymorphisms and altered nicotine metabolism has been demonstrated in experimental studies that followed similar protocols [93, 97-98]. In these studies, it was shown that *CYP2A6* genotype was closely and consistently correlated with enzyme activity as measured by nicotine and cotinine levels after nicotine administration. For *CYP2A6**1/*1 individuals ('normal' wild-type metabolizers), plasma nicotine levels were higher and cotinine levels lower than for all other genotypes, except gene duplications.

The relationship between *CYP2A6* activity and altered smoking behavior has also been investigated *in vivo*. A double-blind placebo-controlled experiment reported by Sellers demonstrated that administration of a *CYP2A6* inhibitor (oral methoxsalen) together with oral nicotine caused a consistent and stepwise reduction in smoking behavior when participants were given a 'free smoking period' after drug administration [95]. All smoking indices tested (breath carbon monoxide increase, number of cigarettes smoked, time to next cigarette, number of total puffs, nicotine/cotinine ratio, carbon monoxide increase/puff, and self-rated desire to smoke) were consistent with a reduction in smoking behavior for participants with impaired *CYP2A6* function [95].

Despite experimental evidence for a strong *CYP2A6* genetic component to smoking behavior, epidemiological evidence has been more equivocal, although consensus has begun to emerge that at least some aspects of adult smoking behavior are influenced by *CYP2A6* variation (recently reviewed in Ray (2009) [89]). However, as recently as 2003, there was no firm consensus as to whether *CYP2A6* influenced smoking behavior [88]. In a 2003 review by Tricker et.al., eight out of 12 studies showed results consistent with the hypothesis that individuals with variant genotypes would score lower on measures of smoking behavior [99-105]. However, few were able to demonstrate statistical significance [99, 101-102] and one of the three had technical difficulties with genotyping [101]. Only two of the 12 [102, 106] were able to use biomarkers of cigarette consumption rather than self-reported measures of smoking status and behavior, one positive [102] and the other null [106]. Further, a meta-analysis of *CYP2A6* genotype and smoking behavior, which included 11 of the 12 studies reviewed by Tricker (Pianzella 1998 was excluded), failed to provide evidence of an association between *CYP2A6* genotype and smoking status [87]. Unfortunately, the authors were only able to categorize smoking into crude categories of SMOKE (higher tobacco use or dependence) vs. NO SMOKE (no tobacco use or non-dependent smoking) and SMOKE more vs. SMOKE less. In contrast to the overall results, the most methodologically rigorous study [102], showed a clear trend for those with lower activity level genotypes to have lower breath carbon monoxide levels, lower cotinine levels, and a higher nicotine/cotinine ratio than smokers with a more active genotype. This supports the hypothesis that slow nicotine metabolizers require less cigarette consumption to maintain nicotine levels than faster metabolizers. Given a behavior as complex as smoking, and the level of detail available, it is not surprising that subtle or specific effects were not evident in this meta-analysis [87].

Clearly, as of 2003, there was more to be done before the putative association between indices of smoking behavior and variation in the *CYP2A6* gene could be convincingly demonstrated or a determination made about whether the independence assumption would be violated in a case-only analysis of *CYP2A6* variation and smoking behavior. However, if the only the criterion for verifying the independence assumption is the statistical significance of the putative relationship between *CYP2A6* and smoking in controls, an investigator would have been justified proceeding with a case-only study of *CYP2A6*, smoking and lung cancer at this point in time.

If the independence assumption were verified, a case-only study of the interaction of *CYP2A6* and smoking would be attractive for a number of reasons. It is hypothesized that, in addition to the protective effect of smoking less, individuals with lower *CYP2A6* activity are at lower risk of lung cancer from smoking than those with higher activity enzymes because of reduced procarginogen activation. *CYP2A6* is able to metabolize 4-(methylnitrosamino)-1-(3-pyridyl)-1-buanone (NNK), a component of cigarette smoke, to a reactive mutagenic compound. When *CYP2A6* is inhibited by methoxsalen in *CYP2A6**1/*1 individuals, production of the reactive metabolite is reduced and NNK metabolism is shifted to NNAL and NNAL-glucuronide, non-mutagenic and readily excretable compounds [89, 107-108], decreasing exposure to carcinogenic intermediates. There have also been investigations of possible behavioral mechanisms by which cancer risk is reduced in slow metabolizers [109]. Consequently, the interaction of *CYP2A6* variants and smoking is of high interest, both for public health and clinical practice, and a number of traditional case-control studies have examined this [103, 110-112].

The case-only study has several advantages over a traditional case-control study in this context. No controls need to be recruited, interviewed, genotyped or phenotyped. This is especially appealing for a study in a Caucasian or African-American population because the frequency of variant alleles for *CYP2A6* is generally quite low [$<5\%$ [93, 113]] and large numbers of controls would be needed for an interaction analysis. A case-only study would not generate ORs for the main effects of smoking or *CYP2A6* variation but this might not be seen as a severe limitation. For many cancers, particularly lung cancer, the main effects of smoking are well established. The *CYP2A6* enzyme has a limited number of substrates, primarily exogenous [88], and is therefore unlikely to have a substantial main effect in the absence of environmental exposure. Selection bias due to differential non-participation of smokers as controls would not be a factor. Although selection bias related to case recruitment would still remain a consideration, recruitment of cases is generally more successful than for controls and the potential for selection bias should be lower. Lastly, a case-only study would eliminate the potential for differential recall between cases and controls, a concern with behavioral risk factors such as smoking where participants believe that exposure have could affected their case status. If the only the criterion for verifying the independence assumption is the statistical significance of the putative relationship between *CYP2A6* and smoking in controls, an investigator in 2003 would be justified proceeding with a case-only study of *CYP2A6*, smoking and lung cancer, despite the gathering experimental and epidemiological evidence of association. However, from the current vantage point, it is clear that this would have led to biased estimates of interaction. It is also clear that more than statistical significance is needed to guide evaluation of the independence assumption.

C. The independence assumption in empirical data

Outside of investigators conducting specific case-only analyses, as discussed previously, relatively little work has been done directly assessing particular control group G-E associations likely to be important in interaction studies. Two notable exceptions are the recent large analyses of smoking and metabolic gene polymorphisms by Smits et. al. [54] and a similar study of metabolic gene polymorphisms and alcohol consumption by Raimondi et. al. [114]. The former study examined associations between polymorphisms in five xenobiotic metabolizing genes, *CYP1A1*, *GSTM1*, *GSTT1*, *GSTP1* and *NAT2* and tobacco consumption in pooled controls from case-control studies included in the International Collaborative Study on Genetic Susceptibility to Environmental Carcinogens (GSEC), and the latter examined associations between polymorphisms in *CYP2E1* RsaI, *CYP2E1* DraI, *ADH1C* and *NQO1* and alcohol consumption in the same data. The GSEC includes individual level data from both published and unpublished case-control studies of gene-environment interaction in cancer [115-116].

In the study of metabolic polymorphisms and smoking, the number of subjects included in each analysis varied from 2,792 for *GSTP1* to 10,719 for *CYP1A1*. Although the sample size was large, the study had several important limitations. The authors were only able to categorize smoking crudely, as never/current/former for the bulk of their data, and had information on dose for less than half of those (35.5%-47.3%). Smokers may refuse to be controls more often than non-smokers, and refusal could also be associated with family history and therefore genetic factors. Although this would be more problematic for high penetrance genes that track more closely with family history than is likely for single

nucleotide polymorphisms (SNPs), selection bias may have been present, a limitation not addressed by the authors.

Bias due to non-participation by smokers can vary between populations. For example, studies done in Canada, the US and Europe have shown both over-participation and under-participation by smokers. As an example of over-participation by smokers, Morabia et. al reported a higher prevalence of current smoking among male and female neighborhood controls (42% and 38%, respectively) than hospital-based controls (29% and 24%) or in the US overall (30% and 25%) [117]. Similarly, Ramos et. al. found higher proportions of current and former smokers in women who participated in a population-based study of myocardial infarction (12% and 10%, respectively) than in those who did not fully participate (7% and 4%, respectively) [118]. Conversely, Holt et. al (1997) found that women who smoked during pregnancy were more likely (24%) than non-smokers (13%) to refuse participation in a post-partum survey (24% and 13% refusals, respectively) [119]. Heilbrun et. al. (1982) found 57% of participants in a prospective study of cancer in Hawaii were current smokers while 61% of the men who refused were current smokers [120]. In a Canadian study of mammography, current smokers were underrepresented among women having time-appropriate mammograms [121].

Probably most important for the evaluation of the independence assumption, the controls in a pooled study do not represent any particular population at risk from which cases might arise. Identifying a relevant population base in which to assess the independence assumption is of fundamental importance to the validity of case-only studies. It is not often appreciated that populations might vary in ways that affect the independence assumption and thus the validity of case-only studies conducted in those populations. The independence

assumption pertains to unconfounded G-E association and different populations may well have different constellations of G-E confounders. In this context, it is extremely difficult, if not impossible, to use results from pooled studies to answer to the question of whether a specific case-only study might be valid in one or more of these populations. In Smits et. al. (2004) non-hospital (“healthy”) and hospital controls from GSEC studies were pooled (66% and 38%, respectively) for overall analyses. Overall estimates were adjusted for study, gender, age, and ethnicity. This could be problematic for at least gender, age and ethnicity if these variables are proxies for different exposures in different populations. Participants were primarily Caucasian (72.6%), with smaller proportions of Asians (11.6%) and African-Americans (5.2%). However, given that race/ethnicity is largely a socially constructed variable [122-123], it is unclear what meaning this has when taken out of the appropriate social context. Nonetheless, for this pooled analysis the authors conclude that “The use of the case-only design for epidemiologic studies including these polymorphisms is therefore justified, at least when studying smoking habits.” This conclusion was based on the paucity of statistical significance and lack of strong associations (all ORs were <1.3 or >0.8 for healthy controls, < 1.4 or >0.6 for hospital controls). In view of the study limitations discussed, however, this conclusion does not seem fully justified.

In a smaller (N=339) population-based study of Japanese males 40-49 years of age, the authors assessed association between ‘habitual smoking’ (ever/never) and drinking (drinker/non-drinker), and a panel of 153 single nucleotide polymorphisms (SNPs) in 40 candidate genes [64]. Genes chosen were those coding for xenobiotic metabolizing enzymes, DNA repair enzymes and ‘other stress-related proteins’. The xenobiotic metabolizing enzymes included the cytochrome P-450s *CYP1A1*, *CYP1B1*, *CYP2C9*, *CYP2C19*, *CYP2E1*,

CYP17A1, and *CYP19A1*, glutathione transferases *GSTM2*, *GSTM3*, *GSTT2*, and *GSTP1*, N-acetyl transferases *NAT1* and *NAT2*, alcohol dehydrogenases *ADH1A*, *ADH1B*, and *ADH1C*, aldehyde dehydrogenase *ALDH2*, and epoxide hydrolases *EPHX1* and *EPHX2*. The DNA repair enzymes were *OGG1* and *NUDT1* (*MTH1*). The other genes included, but were not limited to: the estrogen and progesterone metabolism genes, *ESR1*, *ESR2*, *ERRRG*, *PGR*, *COMT*, *HSP17B2*, and *HSP17B3*, serotonin transporter gene *SLC6A4*, glucocorticoid receptor *NR3C1*, nitric oxide synthase *NOS2A* and *NOS3* and dopamine receptor genes *DRD2*, *DRD3*, and *DRD4*. The SNPs analyzed were chosen from a larger pool of SNPs (N=289) after elimination of SNPs with a minor allele frequency of <1% and SNPs not in Hardy Weinberg equilibrium (HWE). Consistent with study goals, all genes were chosen because they were considered important candidates for future interaction studies, rather than because they were particularly likely candidates for gene-smoking/drinking association in a healthy population. Plausibility of individual gene-environment associations was discussed only for SNPs with statistically significant results.

For the DNA repair genes examined, *OGG1* and *NUDT1*, associations were found between smoking and three of four of the SNPs in *OGG1* (0.4-0.6, borderline statistical significance, variant carrier vs. variant non-carrier) but no statistically significant associations were found for either SNP of *NUDT1*. After adjustment for drinking status (never/former/current drinker), significant associations were reported for smoking and at least one SNP in five of the 40 genes tested [*OGG1* (DNA repair), *SLC6A4* (serotonin transport), *CYP17A1* (xenobiotic metabolism), *EPHX1* (xenobiotic metabolism) and *ESR1* (estrogen metabolism)]. The associations with smoking and *OGG1*, *CYP17A1* and *EPHX1* were novel findings with uncertain plausibility that must be replicated.

This study has several important limitations [64]. Because of sample size, smoking was dichotomous in all analyses: ever/never for the unadjusted estimates for each SNP, current/non-smokers and current vs. never smokers for adjusted estimates. Data were insufficient for estimates of effect for smoking dose or duration and dose-response relationships could not be explored. Functional data was lacking for most of the SNPs examined. The DNA samples included in this study were the subset of samples from a previous study with enough DNA remaining for further testing (53.5%), raising the possibility of selection bias. Overall, before calculating the adjusted estimates, the authors examined 153 SNPs for association with 2 exposures (smoking, drinking) using 2 models for each comparison (variant dominant, variant recessive) for a minimum of 612 comparisons. Statistically significant association was found for 29 (4.7%) of the comparisons. Given the limitations (high number of comparisons, limited sample size, gender- and age-restricted population and crude environmental exposure measurements), results for this particular panel of ‘stress-related proteins’ must be replicated before the associations are considered robust. Although these limitations were not discussed (other than sample size), the authors state that the study provides basic but essential information for future case-only studies (i.e. that some particular SNPs may be associated with smoking and others likely are not), a conclusion which seems warranted if results are considered with appropriate caution.

Finally, in a more limited exploration of G-E control group associations, a different group of investigators examined associations in four studies from Johns Hopkins University [124]. They found ‘very few’ statistically significant G-E associations among controls, though they specified neither the particular associations examined nor the magnitude of the associations. They did note, however, that 5 of the 7 significant interactions that were found

in case-control analyses were not found in the corresponding case-only analyses, primarily due to non-significant G-E control group associations that were in the opposite direction to the interaction effect. This phenomenon was also demonstrated in a 1999 study by Hamajima et. al., using data from 4 published studies of gene-smoking interaction [125]. In this paper, OR_z s from all 4 studies (range: 0.6-2.3) were non-significant and in the opposite direction from the SIM, causing all of the CORs to be closer to the null than the case-control estimates of interaction. In these studies, the tests of statistical significance (at $\alpha=0.05$) of the COR and SIM were concordant, although the magnitude of the COR and SIM were different. The most extreme example was from the study of *NAT2* and smoking in bladder cancer, where OR_z was 0.69 (95% confidence interval (CI) 0.37, 1.29), the COR was 0.83 (95% confidence interval (CI) 0.42, 1.66) and the SIM was 1.20 (95% CI 0.49, 2.96). Both of these studies of G-E association in control groups serve to illustrate the necessity of further exploration of empirical evidence of independence assumption violation and its effects on interaction studies and the interpretation of interaction estimates from different study designs.

D. Effect of independence assumption violation in data simulation

Data simulations have demonstrated that even small violations of the independence assumption can strongly bias the case-only interaction parameter [5]. Using logistic models, Albert et. al. (2001) varied the magnitude of control group G-E association to explore the effect of independence assumption violation on case-only interaction estimates. As expected from the previously presented equation ($COR = SIM * OR_z$), as values of OR_z above the null increased, the COR was biased away from the SIM in a multiplicative fashion. Using data from a study of *XRCC1* genotype and lung cancer by Ratnasinghe et. al. (2001), they showed that a control group association between genotype and pack-years of tobacco use of $OR_z=2.03$

created a bias in the COR of 105% [COR=0.90 (95% CI 0.41, 1.94), SIM= 0.44 (95% CI 0.17, 1.16)] [126]. An OR_z of 1.2, the association between genotype and alcohol drinking status (ever/never), biased the COR by nearly 30% in another example.

Further, violations of the independence assumption may cause the Type II error rate (false negatives) to be high. When control-group G-E associations are of similar magnitude but opposite in direction to the interaction effect, a case-only study may not detect interaction effects [5, 124]. Type II error when evaluating the independence assumption means that true Z_s of sufficient magnitude to bias the COR are not detected. Depending on the magnitude of bias in the COR from this source, these CORs could be extremely misleading in the context of screening for interaction or candidate genes for further investigation. Little work has been done to explore this possibility.

E. Independence assumption verification methods

Although the validity of case-only estimates rests heavily on the independence assumption, and case-only studies, particularly stand-alone case-only studies, have many advantages over traditional study designs for interaction analysis, there is relatively little the literature on methods of independence assumption verification. The previously discussed work by Albert et. al. (2001) which partially quantified, largely through data simulation, the magnitude of bias and effect on the Type I error rate of even modest G-E association in controls is one example. Another notable paper to focus on independence assumption evaluation is a recent paper by Gatto et. al. which elucidates conditions under which a control group is an appropriate proxy for the underlying study population when validating the independence assumption [127]. They conclude, also primarily from data simulations, that a control group can be used for this purpose only when 1) the baseline risk of disease is very

low [$p(\text{Disease} \mid \text{G-E-}) < 0.1\%$] or 2) the baseline risk of disease is low [$p(\text{Disease} \mid \text{G-E-}) \leq 1.0\%$] *and* the independent effect of the gene is weak ($\text{RR}_{\text{gene}} \leq 2.0$). If this is not the case, the magnitude of the control group G-E OR (OR_z) diverges substantially from the G-E RR in the underlying cohort (RR_z), which is what the independence assumption is based on. Although the data simulations in this study are framed in terms of gene-environment interaction, the conclusions apply to any two exposures examined in a case-only interaction study. In practice, because the case-only design cannot estimate main effects, it would not be an optimal choice for investigation of a gene expected to have an appreciable main effect, an important consideration.

Some empirical work on independence assumption evaluation has been done with pooled data, and with existing case-control studies, but has focused on quantifying specific independence assumption associations [54, 114], or on assessing the frequency of independence assumption violation [124], rather than on methods of independence assumption assessment. In practice, some published case-only studies have used only arguments about the plausibility of the independence assumption; others have gone further and attempted quantitative assessment of the independence assumption in control groups or other ancillary data. Quantitative assessments have relied on statistical significance of OR_z as the sole criterion of the validity of the independence assumption regardless of the method of assessing the G-E association. However, the practical effects of relying on statistical significance to evaluate estimates of Z , in particular the effects on bias in the COR remain to be elucidated.

A further difficulty is that methodological work that has been done generally assumes case-only analyses are fully nested within case-control studies. Although nested case-only

analyses of interaction, if properly conducted, do produce more precise estimates than case-control analyses of interaction, they cannot realize other important advantages of the case-only design such as smaller sample size, no risk to controls and cost reduction. By necessity, evaluating ancillary data for use in independence assumption validation must include considerations of the appropriateness of the control data to the case-series. Because the two most prominent studies to date that have explicitly considered ancillary control data have pooled controls (GSEC controls) [54], or consider a very limited population (Japanese men, 40-49 years of age) [64], very little light has been shed on what study characteristics a control group should possess in order to be a valid proxy for population controls for a given case-series. While Albert et. al. (2001) have proposed, and partially evaluated, a method for using a sample of controls for independence assumption validation, no analogous work has been done on the optimal method(s) of evaluating the independence assumption in ancillary data. So, although many aspects of independence assumption evaluation can presumably be generalized from nested case-only studies, more work needs to be done to ascertain which aspects can be generalized, and what additional practices are needed for independence assumption verification in stand-alone case-only studies.

III. STATEMENT OF SPECIFIC AIMS

A. Specific aims

Aim 1: Meta-analysis

To characterize specific gene-smoking associations in control groups found by systematic review of the published literature using meta-analytic techniques when appropriate.

1. To estimate the control group/population associations for a set of DNA repair gene variant-smoking pairs (OR_z) using published data. To estimate a summary OR_z where appropriate.

Association was estimated by calculating an unadjusted odds ratio and 95% confidence interval [OR (95% CI)] from published data. Genetic exposures were *XRCC1* [Arg399Gln (rs25487), Arg194Trp (rs179872), Arg280His(rs25489)], *XPB* [Lys751Gln (rs13181), Asp312Asn (rs1799793)] and *XRCC3* [Thr241Met(rs861539)]. Smoking exposures included, wherever possible, smoking status: current, former, never, not current smokers, and smoking amount: duration, intensity and pack-years of smoking.

2. To use meta-regression to evaluate study characteristics as potential predictors of heterogeneity.

Study characteristics included study-level characteristics such as: Hardy-Weinberg equilibrium in controls, geographic study area, study design, mean/median age of study population, proportion male gender and ethnicity.

Aim 2: Carolina Breast Cancer Study (CBCS) and North Carolina Colon Cancer Study (NCCCS) control group associations

To estimate gene variant-smoking associations in CBCS and NCCCS control groups overall, where appropriate, and within race, categorical age and gender, for all SNPs in CBCS or NCCCS plausibly biologically related to smoking behavior.

1. By race, age, gender and overall: To estimate gene variant-smoking associations [unadjusted OR_z (95% CI)] in CBCS (Phase I and II, CIS) and NCCCS control groups using unconditional logistic regression.
2. To evaluate effect measure modification by race, age or gender (NCCCS only) using the likelihood ratio test (at $\alpha=0.05$) for models with and without a race/age/gender x smoking term. To estimate OR_z s adjusted for the sampling variables race, age and gender when there was no appreciable effect measure modification. To test for Hardy Weinberg equilibrium within race for each polymorphism.
3. To assess potential patterns in independence assumption violation across gene pathway groups: SNPs were grouped by gene pathways (e.g. DNA repair, xenobiotic metabolism) for analysis as above, by race, age or overall.

4. To evaluate potential confounders of OR_z : To evaluate sampling variables [age at selection, race, gender (NCCCS only)], all individual level factors identified in the meta-analysis as potentially important [age, race], and other variables identified using directed acyclic graphs [family history of any cancer, family income] as potential confounders.

Aim 3: Environmental misspecification

To evaluate the impact of environmental exposure misspecification on independence assumption evaluation using CBCS/NCCCS control group data.

1. To evaluate the effect of error due to smoking misspecification: To describe the frequency, magnitude and direction of undetected bias in the COR when significance testing or the magnitude of OR_z were used to assess independence for a given specification of smoking such as ever smoking, but the COR would have been calculated for a different specification of smoking, such as intensity (packs/day).

B. Hypotheses

The primary hypothesis was that the independence assumption would be violated (i.e. $OR_z \neq 1$) for smoking behavior and a proportion of the genetic variants greater than would be expected by chance alone both in the meta-analysis and in the CBCS and NCCCS control groups. Further, that the violation(s) will be of sufficient magnitude to cause appreciable bias (>15%) in the COR. Secondary hypotheses are that 1) violations of the independence assumption will occur more frequently for measures of smoking amount (duration, dose or PY) than for smoking status (ever or current smoking), 2) misspecification of smoking

exposure (using a different smoking measure to evaluate the independence assumption than the smoking measure used in the case-only analysis) will lead to undetected bias in the COR and 3) the magnitude and direction of OR_z for SNPs assayed in the CBCS and NCCCS controls will agree more often than expected by chance.

The study characteristics that were expected to be influential predictors of the magnitude of OR_z were HWE status, population- vs. hospital/patient-based controls and older age. Assessment of Hardy Weinberg equilibrium provides some indication of possible population stratification by ethnicity, misclassification of genotype and/or selection bias in the control group [128] and these may induce G-E association. Since smoking can increase risk for numerous diseases, hospital-based controls are likely to have G-E associations not found in the general population. Similarly, for any genes associated with longevity, average age of the study population was expected to influence G-E association, particularly for an exposure such as smoking where patterns of use have changed over time, and continue to change. There is an extensive and rapidly expanding literature on the genetics of longevity [129-132]. Many different functional categories of genes are being examined including DNA repair genes, xenobiotic-metabolizing genes and genes involved in defense against reactive oxygen species (ROS). Not surprisingly these are many of the same genes being investigated for their potential as cancer susceptibility genes.

C. Rationale

The primary aim of the studies described in Chapters V-A and V-B was to enable investigators considering a stand-alone case-only study of gene-environment interaction to evaluate the independence assumption more rigorously than has been done previously and identify situations where case-only estimates are not valid. Although the case-only design can

be used to evaluate statistical (multiplicative) interaction between any two exposures, it is most commonly used for gene-environment interaction. In order to explore the independence assumption in empirical data, a model environmental exposure (smoking behavior) and a panel of genetic variants were chosen. Briefly, smoking behavior was chosen because of its importance in public health, and because smoking measures are very commonly collected. Polymorphic genes plausibly biologically related to smoking, primarily DNA repair genes, were the genetic exposures of interest. The study described in Chapter V-A was a systematic review including a series of meta-analyses of six DNA repair SNPs and smoking. The second study (Chapter V-B) was an exploration of control group gene-smoking associations (OR_z), including but not limited to the DNA repair genes in the systematic review, in two population-based control groups.

1. Smoking and genes

Interest in genetic influences on behavioral traits is long-standing and there is an extensive and burgeoning literature in behavioral genetics. While there are currently only a limited number of established associations between genotype and behavior, research in areas related to smoking behavior, including addiction, personality, and psychiatric disorders, is flourishing [59-61]. Most of these genes function in multiple pathways with poorly characterized effects. *COMT*, for instance, because of its role in the dopamine pathway and hormone metabolism has been studied in conjunction with widely divergent outcomes, many with behavioral components, such as schizophrenia, attention deficit hyperactivity disorder, alcoholism, cataract, blood pressure, Alzheimer disease and breast, ovarian, hepatocellular, and bladder cancer as well as smoking behavior [55, 62-71, 133].

Strong interest in examining smoking in gene-environment studies comes from a number of sources. Smoking is a highly prevalent exposure. Data from the 2003 Behavioral Risk Factor Surveillance System (BRFSS) survey indicated a median prevalence of 22.1% for current smoking among US adults with a range of 12%-31% by state [134]. In China, the prevalence of smoking and tobacco-related deaths has risen dramatically over the last decades, and is projected to increase in other developing countries in the coming decades, perhaps causing as many as 10 million deaths globally (out of a projected 60 million) by the year 2030 [135]. Tobacco smoking has a well-documented causal relationship with many cancers (e.g. lung and bladder), and is believed to contribute to other cancers (e.g. colon, kidney and prostate). However, variation in disease outcome with similar exposure does exist and the development of cancer is not an inevitable outcome even for heavy smokers. This is not surprising, given that tobacco smoke constituents (including carcinogens) are metabolized by xenobiotic-metabolizing enzymes encoded by highly polymorphic genes and that the genotoxic effects of tobacco smoke constituents [136] are modified by highly polymorphic DNA repair enzyme genes [137]. Consequently, gene-smoking interaction studies are of significant public health importance. However, they may be problematic for case-only studies if the genetic exposures under study, or genes in linkage disequilibrium with them, are causally or non-causally associated with aspects of smoking behavior.

Twin, adoption and linkage studies all demonstrate that there is a heritable (i.e. genetic) component to smoking behavior. Evidence from twin studies and association studies suggest that there are genetic influences on at least three aspects of smoking history: smoking initiation, nicotine addiction and success of smoking cessation [90, 138]. There is, however, substantial variation among populations with respect to the relative contributions of genetic

vs. environmental influences [139]. Interest has focused on variation in constituents of the dopamine pathway (e.g. *DRD2*, *COMT*), nicotine metabolism (e.g. *CYP2A6*), the serotonin pathway (*5HTT*), xenobiotic metabolizing pathways (e.g. *CYPs*, *GSTs*) and nicotinic acetylcholine receptors (e.g. *CHRNA4*) [139]. Despite the many candidate gene studies, few robust gene-smoking associations have been found. In fact, in a recent study designed to evaluate genetic screening for risk of smoking initiation, the criteria for choosing polymorphisms to include in the screening panel (positive results in at least three independent samples and a pooled OR of >1.1 for ever smoking) identified only five gene variants: *DRD2* TAQ1A, *TPH C779A*, *5-HTTLPR*, *MAO-B A644G* intron 13, and *COMT Val158Met* [140].

Metabolic genes and smoking: There is an extensive epidemiologic literature on smoking and metabolic genes, [i.e. those coding for enzymes that metabolize nicotine or other tobacco smoke constituents such as polycyclic aromatic hydrocarbons (PAH) or aromatic amines] [108, 141]. These are largely case-control studies focused on the potential of the genes to modify risk of disease for smokers. In addition to cancer, this has been an especially active area of research for cardiovascular disease and birth outcomes [142-143]. Smoking directly exposes the lungs to a range of toxic xenobiotics and is addictive; exposure to tobacco smoke constituents can last for decades, even through the entire lifespan via exposure to maternal smoking. Tobacco smoke constituents, including nicotine and PAHs are metabolized to toxic intermediates and/or carcinogens by phase I (activation) and phase II (conjugation) enzymes [141]. Variation in these polymorphic genes can alter enzyme activity, regulation or expression, [144-146] plausibly increasing or decreasing risk of disease or influencing smoking behaviors, such as the number cigarettes consumed daily or years as a smoker.

Among the gene variants included in the current project, *COMT* Val158Met SNP (rs4680) is the only SNP that has been extensively studied with respect to its possible influence on smoking behavior, most often smoking cessation [147]. Results have been equivocal with two recent large population-based European studies coming to different conclusion [148-149]. Omidvar et. al. found a 20% reduction in incident smoking cessation, and 30% lower odds of prevalent quitting for the carriers of the low activity form of the allele (Met carriers) whereas Breitling et. al. found no association [$OR_z=0.97$ (95% CI 0.83, 1.12)]. For other included xenobiotic metabolism genes, there is little research. For *CYP1A1*, Chen et. al. demonstrated, in a population of pregnant women (N=165), that having at least one *CYP1A1**2A allele was associated with smoking reduction [$OR(95\% CI)=2.2(1.0-4.6)$] and increased quitting [$OR(95\% CI)=1.7(1.0,2.9)$] during pregnancy [150]. There was no association between *GSTM1* and reducing or quitting smoking found by Chen et. al. [150].

DNA repair genes and smoking: Analogous work has not yet been done for DNA repair genes. Although polymorphisms in the xenobiotic-metabolizing genes can influence the level of reactive metabolites and hence the amount of DNA damage, ultimately it is only DNA damage that is left unrepaired and allowed to continue through the cell cycle that can contribute to the genomic instability necessary for the development of cancer. The carcinogenic ability of an environmental agent may therefore be mediated by an individual's DNA repair capacity. DNA repair enzymes are a group of proteins largely responsible for maintaining genomic integrity by repairing damage to DNA caused by endogenous metabolic intermediates and by-products, reactive intermediates of xenobiotic metabolism, including pharmaceuticals, and ionizing radiation.

Proteins in the DNA repair system, and their genes, fall into 4 broad functional categories, defined by the major type of damage each repairs. They are the nucleotide excision repair (NER) pathway genes, base excision repair (BER) pathway genes, double strand break (DSB) pathway genes and mismatch repair pathway genes. The base excision repair (BER) pathway is responsible for the removal of individual damaged bases and restoration of the sugar-phosphate backbone. It uses the undamaged strand as a template for restoring the correct base. The nucleotide excision repair (NER), in contrast to the BER, recognizes and removes bulky lesions from one strand of the DNA and restores stretches of DNA 25 nucleotides or more in length using the intact strand as a template as in BER. The double-strand break (DSB), as the name implies, is responsible for rejoining sections of DNA that have been broken across both strands, leaving no intact template for repair. DSB repair can be accomplished either through non-homologous end-joining, where the ends of two unrelated chromosomes are joined and some genetic material is lost, or homologous end joining, where the homologous chromosome is used as a template for repair. The BER pathway is largely responsible for repair of oxidative damage and the NER pathway for repair of bulky DNA adducts, both types of damage produced by constituents of tobacco smoke [151]. Cigarette smoke is genotoxic, with multiple studies showing smokers have increased rates of sister chromatid exchange and micronuclei formation in lymphocytes, increased DNA strand breaks in lymphocytes, buccal cells and urothelial cells, and for heavy smokers, oxidative damage to DNA in germ cells [136]. The DNA repair genes in the BER, NER and DSB pathways are highly polymorphic [151-152], and this variation is thought to contribute to cancer risk.

In addition to the importance of DNA repair gene polymorphisms, smoking, and their likely interaction in cancer etiology and public health, for a case-only study it is also important to consider whether there could be a causal association between DNA repair genes and smoking in healthy individuals. While there are no firm data in this area, and this is not currently an area of active investigation, plausible mechanisms exist for variability in DNA repair capacity (as measured by genotype) to be associated with smoking behavior, especially smoking dose. Smoking induces DNA repair, presumably through DNA damage caused by smoking. When there is reduced capacity for DNA repair, as may occur when an individual has an allele for a lower activity form of a DNA repair enzyme, any physiological effects of non-repaired DNA will be exhibited at a lower level of the damaging exposure (smoking in this case) than when DNA repair is optimal. Although no detailed work on the physiological effects of non-repaired DNA produced by smoking and the DNA repair genes that will be investigated in this project has been done, there is related evidence that there are physiological processes that could affect smoking behavior.

Patients with xeroderma pigmentosum (XP), who lack NER DNA repair, suffer from high cancer rates but can also show neurodegenerative effects, which are believed to be the result of accumulation of unrepaired DNA lesions and cell death [153], possibly due to oxidative damage [154]. Patients with Cockayne syndrome (CS), another inherited disorder of DNA repair, can exhibit symptoms of neurological degeneration in addition to premature aging and patients while patients with ataxia telangiectasia, an inherited neurological disorder, have increased cancer susceptibility and impaired DNA repair [155].

The pleasurable aspects of smoking, as well as smoking addiction, operate through neurological pathways, particularly through the nicotinic acetylcholine receptors, and

perturbations in this system could easily affect smoking dose and/or smoking cessation rates. Smoking is being actively investigated with respect to possible protective effects on risk of Parkinson's and Alzheimer's disease [156] and it has been proposed that pharmacologic stimulation of DNA repair via the ataxia telangiectasia-mutated (*ATM*) gene may be beneficial in Parkinson's [157].

In addition to possible interaction between smoking behavior and DNA repair through neurological mechanisms, more direct mechanisms are possible. It has recently been shown that mice without the *CSB* gene (the defect in Cockayne syndrome) were especially vulnerable to oxidative damage from paraquat exposure [158]. Consistent with this observation, it has also been shown, using cells from XP patients, that unrepaired oxidative damage can appreciably affect transcription and reduce gene expression [159]. Exposure to cigarette smoke is a source of oxidative damage. In patients with severe neurological manifestations of XP, death is often due to respiratory complications in childhood [160], although subtle manifestations of neurological effects are possible in adulthood [161-162]. Further, it has been hypothesized that DNA repair may play a role in idiopathic pulmonary fibrosis [163-164], and that accumulated DNA damage may play a role in chronic obstructive pulmonary disease [165-167]. Development of respiratory effects related to poor repair of oxidative DNA damage could plausibly influence smoking dose, smoking duration and/or smoking cessation rates.

A population-based study examining multiple SNPs (single nucleotide polymorphisms) for a panel of 'lifestyle-associated' genes in Japanese men (N=339) reported inverse associations (range of ORs= 0.4-0.6) between smoking status (smoker/non-smoker) and 3 different polymorphisms in *OGG1*, a DNA repair enzyme active in the BER pathway

[64]. Consistent with this observation, an *in vitro* study has shown that transcription-linked subcellular localization and expression of *OGGI* during the cell cycle was markedly different for wild-type *hOGGI* and *hOGGI* Ser326cys [168], one of the polymorphisms assessed by Liu et al. (2005). The other DNA repair gene measured in the Liu et al. study was *NUDT1*. For *NUDT1*, homozygosity at one SNP was inversely associated with drinking status [drinker/non-drinker: 0.1(0.0-0.8)], although sample size was small (four homozygotes). Although the authors had insufficient data to assess the possible relationship between more complex aspects of smoking (e. g. dose) and DNA repair genes their results are consistent with a possible causal relationship between DNA repair genes and smoking behavior.

As discussed previously, very little systematic work has been done aimed specifically at improving the conduct of stand-alone case-only studies. Briefly, a pooled analysis has been conducted for several xenobiotic-metabolizing genes and smoking, but was limited to crude categorizations of smoking pooled across possibly disparate populations [54]. A similar study was conducted with xenobiotic-metabolizing genes and alcohol consumption and had similar limitations [114]. The small population-based Japanese study discussed earlier reported on a panel of 40 genes plausibly related to smoking or alcohol consumption and their association with smoking and/or drinking status, respectively, but was limited to males [64]. No attempt was made in any of these studies to evaluate characteristics of the data that might make it appropriate (or not) for evaluation of the independence assumption.

Consequently, two studies were undertaken to address G-E association in empirical data: 1) a systematic review and meta-analysis of gene-smoking OR_z s in published control data, and 2) an analysis of gene-smoking OR_z s in two population-based control groups. The studies are described in Chapters V-A and V-B, respectively. The particular exposures

evaluated in these two studies (smoking and genes relevant to smoking and cancer) were chosen to provide immediately useful information for public health research, particularly cancer research, as well as provide a context for further exploration of associations that, if found in one or more populations, would prove problematic for gene-environment interaction studies in other populations.

The six SNPs that were chosen for the systematic review are in three DNA repair genes that operate in different genetic pathways. The genes were X-ray cross complementing gene 1 (*XRCC1*), xeroderma pigmentosum complementation group D [*XPB*, previously excision repair complementing defective 2 (*ERCC2*)] and X-ray cross-complementing gene 3 (*XRCC3*). Each gene is polymorphic, and each SNP has a minor allele frequency > 10% in most studied populations. *XRCC1* participates in the base excision repair (BER) pathway. Three important non-synonymous single nucleotide changes (Arg194Trp, Arg280His and Arg399Gln) have been identified in *XRCC1*. *XPB* is active in the nucleotide excision repair (NER) pathway. A single nucleotide change (SNP) in *XPB* exon 10 (Asp312Asn) and another in exon 23 (Lys751Gln) have been studied. *XRCC3* is in the double strand break (DSB) pathway. *XRCC3* is believed to code for an accessory protein in the process of homologous joining of broken double stranded DNA. In this case, the appropriate stretch of DNA on the homologous chromosome serves as a template for repair. *XRCC3* has one studied variant, a Thr241Met variant [169]. With the exception of the *XRCC1* Arg194Trp [170], the variants are thought to code for reduced DNA repair capacity, although in no case has this been definitely established [170-174].

For the control group analyses, a convenience panel of gene variants plausibly related to smoking was chosen. These included variants in xenobiotic metabolizing genes, DNA

repair genes, and variants in genes that respond to oxidative stress. All genes were relevant to gene-environment interaction in cancer; the parent studies are case-control studies of breast cancer and colorectal cancer. Most of the genes have a minor allele frequency >10%.

2. Meta-analyses

Systematic review of control group G-E associations can assist in evaluation of the independence assumption at multiple levels. Rigorous evaluation of the independence assumption for a proposed case-only study should include a thorough search of the published literature for relevant control group or cohort data, then evaluation of the magnitude of independence assumption violation in individual studies, both overall and in relevant subgroups. A systematic search, at a minimum, can inform the investigator that there are no control data on the G-E association of interest in the literature, and therefore no empirical data available to evaluate the independence assumption. This should preclude conducting a stand-alone case-only study. If the literature is scant and/or heterogeneous, one can nonetheless assess the potential range of bias that may be introduced into a case-only study. Specifically, the meta-analysis includes assessing the magnitude, direction, precision and statistical significance of each study-specific gene-smoking OR_z , as well as assessing heterogeneity of the OR_z s across studies. Since the magnitude of OR_z is an estimate of the magnitude of bias in the COR, it is the key parameter. Finally, if OR_z s from published studies are both homogeneous and within an acceptably narrow range for the purposes of the case-only study, meta-analysis provides an estimate of the magnitude of bias likely to be introduced into the COR. When studies are heterogeneous, or the range of OR_z s is wide enough to cause a substantive difference in the COR, stratifying studies by design or study population characteristics can illuminate the sources of heterogeneity. Lastly, meta-regression can be

used to formally estimate the strength of association between specific predictors, such as control group Hardy Weinberg equilibrium (HWE) status, population ethnicity, average age etc. and the magnitude of OR_z s. Understanding the relationship between important study characteristics and the magnitude of OR_z can help determine what, if any, ancillary data are appropriate to evaluate the independence assumption. Using these tools, meta-analysis provides context for evaluation of the independence assumption not currently available.

3. CBCS and NCCCS control groups

The second study (see Chapter V-B) was an exploration of gene-environment association in two population-based control groups. The studies were the Carolina Breast Cancer Study (CBCS) and the North Carolina Colon Cancer Study (NCCCS). As in the meta-analysis (Chapter V-A) smoking was the model environmental exposure. OR_z was estimated for all genetic variants related to smoking that had been assayed for these control groups. This included all of the SNPs in the meta-analysis. Genes were grouped *a priori* based on the biologic function of the gene (e.g. xenobiotic-metabolizing genes, DNA repair genes etc.) and associations were examined by group.

The CBCS and the NCCCS are large population-based case-control studies done in central North Carolina during the mid- to late 1990's which included urban, suburban and rural areas (CBCS: $N_{cases}=2311$, $N_{controls}=2022$; NCCCS: $N_{cases}=646$, $N_{controls}=1053$ [175-180]). Both studies over-sampled African Americans to increase power for subgroup analyses. The NCCCS included male and female participants. Potential participants were selected from NC Division of Motor Vehicles lists (<65 years of age) and Health Care Financing Administration lists (≥ 65 years of age), and randomized recruitment was used to select to potential participants to frequency match on relevant characteristics in each study [181]. CBCS

participants were all female, with approximately half African American and half <50 years of age. NCCCS participants were sampled such that the race, age and gender distribution of randomly selected cases is similar for controls (approximately 1:1 for gender and race).

The CBCS and NCCCS data offered a rich resource for this project for several reasons. First, both are population-based case-control studies. Conceptually, the independence assumption is an assumption about the underlying population from which cases arose (control groups are used as surrogates for the underlying population); population-based control groups should be a better approximation of the underlying population than hospital-based control groups or convenience samples. Ideally, a large population survey or population-based cohort would be preferable, but in practice there are few of these available. Investigators rely on control groups from case-control studies such as the CBCS and NCCCS for independence evaluation. In addition, the CBCS and NCCCS draw from approximately the same underlying population, using the same sampling method, so results could be compared for the 15 polymorphisms that were assayed in both control groups. Ascertaining the level of agreement between two studies using the same sampling methods, sampling from an underlying population in largely overlapping geographic area, during overlapping time periods, but with different study outcomes (breast and colorectal cancer) provides a window into the population-specific nature of the independence assumption in practice.

Second, both the CBCS and NCCCS collected extensive data on tobacco smoking behaviors. In the CBCS there were data such that smoking could be categorized as ever/never, former/current/never, total duration of smoking, average amount smoked per day during periods of smoking and pack-years. In the NCCCS there are data on smoking status (never, current and former), duration, and intensity as well. This level of detail is much more

than what is usually published in the ancillary studies that must be used for independence assumption evaluation.

Not only did this level of detail allow for estimation of OR_z for measures of smoking amount (duration, dose and PY), it allowed investigation into the effect of smoking misspecification on independence assumption evaluation (see Chapter V-B). The term ‘misspecification’ was used in this study, rather than the term ‘misclassification’, because the underlying conceptualization of this problem was that of variable misspecification in modeling, rather than measurement error. When a stand-alone case-only study is undertaken, the independence assumption must be examined in ancillary data. As would be expected from the differing study goals, in ancillary data the categorizations used in tables for the joint distribution of genotype and smoking are generally more crude than those that will be analyzed in the proposed case-only study. In the case of a polymorphic gene with 2 alleles at the locus of interest, this means that published data is often collapsed to 2 categories, carriers and non-carriers of the allele of interest, rather than published as 3 categories, 1 each for homozygotes of each allele and another for heterozygotes. For smoking, the only categorization that can be consistently found across studies is the dichotomous ‘ever/never’ smoker. Using one specification of exposure for independence assumption evaluation (e.g. ever/never) and a different one for case-only interaction analysis (e.g. packs/day) is precisely analogous to doing a case-control interaction study with a model that has the same exposure specified as binary for controls and continuous for cases. No work has been published to date on the validity of using different exposure specifications for independence assumption evaluation and case-only analysis.

The exposure misspecification results will be useful to researchers considering specific case-only studies, particularly with genes expected to interact with aspects of smoking amount. If case-only studies are to be used to ‘screen’ for genes that interact with smoking, these results can help identify ancillary studies with appropriate data for valid independence assumption assessment. Again, because of the wide range of research on smoking behavior, both in terms of smoking cessation and other modifications of smoking behavior, and on health outcomes, the utility of these results will not be limited to cancer research. Results from these studies should raise awareness of this shortcoming in the ancillary data currently available to researchers. Ideally, it will encourage research practices that make more detailed control data available for independence assumption evaluation.

Additionally, a wide range of genetic polymorphisms have been examined in both studies. The CBCS had data on polymorphisms in 17 DNA repair genes, 7 xenobiotic-metabolizing genes, and 5 other genes related to cell growth and oxidative damage defense. The NCCCS had genotype data for 15 DNA repair genes, 3 xenobiotic metabolism genes and 1 oxidative stress gene. (For a complete list of gene see Chapter V-B.)

Because both studies over-sampled African Americans there were sufficient white and African American participants we were able to perform subgroup analyses by stratified by race. Stratification by race allows population stratification to be addressed, at least crudely. Population stratification is a potentially important source of bias in genetic association studies [182].

Finally, these studies provided a large sample size for evaluating the independence assumption (CBCS controls: N=2022, NCCCS: N=1053). Since the independence assumption is a large sample approximation, it is crucial to evaluate it in samples with

sufficient power to detect relevant effect sizes. Across the range of smoking prevalences in CBCS and NCCCS controls and several control subgroups (African American in CBCS, ≥ 50 years of age in CBCS and NCCCS, males in NCCCS), and at gene carrier prevalences of 20% or more, there was good power to detect OR_z s of 1.6-1.7 and above, the magnitude of associations that previous data simulations have indicated to be problematic for case-only studies [5]. This is true for many of the measured genetic polymorphisms in these studies, which generally have carrier prevalences of $\geq 10\%$. In particular, there was excellent power to detect OR_z s of 1.6-1.7 and above for the 6 polymorphisms examined in the meta-analysis, either overall or in subgroups.

IV. METHODS

A. Overview

The methods used in the two studies that comprise the dissertation complement and support each other. Both studies were an examination of empirical data used to evaluate the independence assumption. The systematic review and meta-analysis gave a broad overview of the range of OR_z s to be found for the selected DNA repair gene SNPs. A systematic review shows the broad range of study types that have collected, but not necessarily presented, data that can be used for independence assumption evaluation. The magnitude of OR_z can be compared across numerous study-level characteristics such as design and HWE status. Methods of individual-level data analysis used for the CBCS and NCCCS control groups allow exploration of effect measure modification and confounding, information not available in a meta-analysis. Further, using two population-based control groups with detailed data on smoking provides a way to examine the effects of misspecification on independence assumption evaluation and level of agreement across studies, using a different method than comparison in a meta-analysis. Because the 6 SNPs in the meta-analysis are also assayed in the CBCS and NCCCS control groups results can be compared for the two methodologies. Use of these two approaches to explore independence assumption evaluation from different viewpoints enhances understanding of the methods necessary for valid evaluation of the assumption.

B. Literature-based analysis of the independence assumption

1. Overview

The first phase of this dissertation project was a systematic review and literature-based series of meta-analyses of specific gene-environment associations in control groups, cohorts, cross-sectional and convenience studies, with the goal of better understanding heterogeneity within specific gene-smoking associations. The environmental exposure was smoking and aspects of smoking behavior, such as dose and duration. The genetic exposures were polymorphisms in 3 genes coding for DNA repair enzymes: *XRCC1*, *XRCC3* and *XPB*.

Meta-analytic techniques can be used to quantify the magnitude and heterogeneity of the multiple gene-environment associations found in the published literature. It is a quantitative technique that can be used as part of a systematic review. In a meta-analysis, data from multiple studies addressing the same question undergo formal qualitative and quantitative assessment. It differs from a traditional narrative review in several important respects. First, the literature on the topic of interest is searched in a more systematic and explicit manner, with *a priori* inclusion and exclusion criteria for studies. Second, data from multiple studies is both quantitatively and qualitatively compared, and can be combined under certain conditions, unlike a traditional review, which is primarily qualitative. Where appropriate, summary estimates of effect that take sample size of the individual studies into account can be calculated. Meta-regression can be used to formally explore sources of heterogeneity between studies, usually an informal subjective process in the narrative review. Meta-analyses have the advantage of making the literature search process, at least, more explicit and, at best, more thorough as well. Additionally, meta-analysis is a more explicit

and objective way to compare study characteristics, and the only way to do it quantitatively [183].

Meta-analytic techniques are used for two broad purposes. One possible goal of meta-analysis is to produce a summary estimate of effect across studies. Provided certain criteria are met, this technique can be used to combine results from underpowered randomized clinical trials to produce a more precise estimate than otherwise available. Summary estimates can be generated either from weighted pooling of the raw data from multiple studies or, less directly, from combining the individual study estimates with appropriate weighting. These techniques can also be used with observational studies, however the criteria for producing a valid summary estimate are more difficult to meet and a summary estimate is often not of primary interest.

When a summary estimate is inappropriate (e.g. the studies are too heterogeneous to combine), or not desired, another possible goal of meta-analysis is to explore the sources of study heterogeneity [184-185]. This is most often done when there are multiple observational studies attempting to answer essentially the same question but results are inconsistent. In this situation the primary goal is to understand differences between the studies. Heterogeneity among studies can arise from both methodological and population factors [186-187]. Methodological factors that can produce heterogeneity include study design, method of exposure ascertainment, and outcome definition. As a hypothetical example, before widespread PSA (prostate specific antigen) screening, a study of race and prostate cancer including only screen-detected cases would be expected to yield a different (lower) magnitude of association than one that included only symptomatic cases, because race was associated with screening behavior. Another possible source of study heterogeneity is differences

between the study populations. A hypothetical example of this scenario would be if results from observational studies looking at the association of body mass index (BMI) and all-cause mortality done in sub-Saharan Africa differed from results from similar studies conducted in Western Europe, with higher BMI inversely associated with mortality in the former and positively associated in the latter. It would clearly be inappropriate to combine these studies to derive a summary estimate.

Meta-regression, which is analogous to standard regression techniques in most respects, is used to explore the influence of, and estimate the strength of, potential sources of heterogeneity [187]. In meta-regression, the unit of observation is the individual study, the outcome for each study is its effect estimate, and the independent predictors consist of defined values for each of the potential sources of variability. For instance, a meta-regression performed on 10 studies of lead exposure and kidney cancer would have an N of 10, the outcome for each study would be the magnitude of the association between lead and kidney cancer, and some independent variables could be study design (case-control v. cohort), study site, year study was performed, and method of exposure measurement (blood lead levels/bone lead levels). In meta-regression, outcome data on each observation (study) is weighted by the inverse of its variance, and can be further weighted by any additional factors the investigator considers important. The variance component can be deconstructed into two components, within study variance and residual variance. The residual variance is the variance not accounted for by the independent variables, called the between-study variance. Between-study variance can be considered equal across studies if the given group of studies can be considered repeated trials of the same study. In this case a 'fixed effects' model, which sets the between-study variance to 0, may be used. Although it has been argued that studies of

genetic exposures may be a special case [188], this approach is often not considered appropriate for groups of observational studies as it can be difficult to conceptualize them as repeated trials of the same study. Without this assumption, the between-study variance is allowed to vary, with a concomitant decrease in the precision of the summary estimate. This is known as a random-effects model, and is often used with observational studies. It assumes that each study comes from a distribution of studies, each with their own 'true' OR_z . This is in contrast to the fixed-effects model where each study is assumed to be an estimate of one underlying 'true' OR_z . The choice of random- or fixed-effects models for meta-analysis is a conceptual, rather than mechanistic, choice.

Stratified analysis, in conjunction with meta-regression, can help the investigator begin to identify which of the hypothesized sources of heterogeneity contribute to the variation in study results, as well as begin to quantify the relative strength and direction of those contributions. Stratified analysis provides separate summary estimates by study characteristic and meta-regression provides a measure of each stratum relative to the reference stratum (e.g. hospital-based case control studies relative to population-based case-control studies).

2. Literature Search

After consultation with Lynne Morris, a reference librarian at UNC's Health Sciences Library, I identified appropriate keywords and databases for a thorough search of the published genetic literature. I used the CDC Genomics and Disease Prevention database (GDPIInfo), PubMed, and the ISI Web of Science as the primary databases. Keywords for the CDC database were from the 'Factor Menu' keyword list and were as follows: smoking behavior; smoke (tobacco), passive; smoking (tobacco), bidi; smoking (tobacco); smoking

(tobacco), maternal; tobacco. I used the CDC Genomics and Disease Prevention database to determine the frequency with which various polymorphic genes are examined in conjunction with smoking. Because the focus of the meta-analysis was association in the control groups rather than the interaction, the search included studies with any outcome. The CDC database is limited to human studies, and has a strong focus on gene-environment interaction literature. After assessing the relative contribution of different genes to the gene-environment literature for smoking, a panel of genes for further searches was developed, emphasizing genes highly relevant to interaction analyses.

It was not expected that the CDC database would produce an exhaustive list of the relevant gene-environment literature, but instead would provide a guide to the polymorphic genes most frequently studied in conjunction with smoking. The preliminary CDC database search for *XRCC1* and smoking produced 25 references. The same search on PubMed, with the appropriate keywords for that database, produced 47 references and an analogous ISI Web of Science search found 64 publications. There were 73 distinct publications in the combined list of 136 references, but only 21 papers referenced by all 3 databases. Two papers were found only in the CDC database, 7 papers only in PubMed and 23 only in the Web of Science. Although each database has its strengths (CDC is appropriately focused for the current project, PubMed is more comprehensive but easily limited to relevant publications, and the Web of Science is the most comprehensive and current), it was clear that none alone would be sufficient for a thorough search of the literature. Each database captured a different subset of publications within the relevant time period (1999 forward) for the genetic polymorphisms of interest.

The goal for the final searches was to identify studies that presented the joint distribution of the polymorphisms of interest and smoking in non-cases (hereafter referred to as controls) in a form that allowed estimation of the gene-smoking $OR_z(95\%CI)$. Final searches were done in PubMed, ISI Web of Science and CDC databases to capture as much of the relevant literature on each gene of interest and smoking as possible. PubMed, ISI Web of Science and the CDC Genomics and Disease Prevention databases were searched up to March 6, 2007 for peer-reviewed literature likely to contain data on the joint distribution of any of the polymorphisms of interest [single nucleotide polymorphisms *XRCC1* Arg399Gln, Arg194Trp, and Arg280His, *ERCC2(XPD)* Lys751Gln and Asp312Asn, and *XRCC3* Thr241Met] and smoking behavior in non-case groups. Non-case groups were defined as any group not selected as the index group based on disease status (e.g. cohorts, convenience samples and control groups from case-control studies). For simplicity all non-case groups will be referred to as controls throughout this document.

For PubMed searches the terms ‘(smoking OR tobacco OR tobacco smoke OR tobacco smoke pollution)’, ‘(polymorphism OR polymorphism, genetic)’ and a gene-specific keyword (e.g. *XRCC1*) were used together to identify papers that included the polymorphisms of interest and smoking. ISI keywords were (‘smok*’ OR ‘tobacco’) and the gene-specific keyword. The searches used the ‘general search’ and ‘by topic’ options that search the abstract and title text, not just the title. Searches included all document types and in all languages. PubMed, ISI Web of Science and GDPInfo databases were searched up to March 6, 2007 for relevant peer-reviewed literature.

Two types of studies contributed information on gene-smoking associations. The first were ‘main effect’ studies, that is, studies that focused on the association between the given

genetic exposure and smoking behavior. For example, this included studies of DNA repair genotypes and degree of genetic damage in blood cells of healthy volunteers, smokers and nonsmokers. The second type of study that provided information for calculating OR_z was gene-environment interaction studies of the polymorphism of interest and smoking. These gene-environment interaction studies often contained tables of the joint distribution (in cases and controls) of the polymorphism and some aspect of smoking behavior. In order to further the goal of examining heterogeneity between studies, the inclusion criteria were broad. To be included a study had to present either 1) control data on the joint distribution of any of the genotypes of interest and any measure of smoking such that OR_z (95% CI) could be calculated or 2) an estimate of the OR_z and 95% CI in controls. Further, each study had to provide enough information on the specific polymorphism and genotyping method to be certain the same polymorphisms had been assayed across studies. SNP designations considered equivalent are in shown below (Table IV.1).

Table IV.1 SNP designations for data abstraction

<i>XRCC1</i> Arg 399 Gln	<i>XRCC1</i> Arg 194 Trp	<i>XRCC1</i> Arg 280 His	<i>XPB (ERCC2)</i> Lys 751 Gln	<i>XPB (ERCC2)</i> Asp 312 Asn	<i>XRCC3</i> Thr 241 Met
G28152A	C26304T	G27466A	A35931C	G23591A	C18067T
exon 10	exon 6	exon 9	exon 23	exon 10	exon 7
rs25487	rs1799782	rs25489	rs13181	rs1799793	rs861539
R399Q	R194W	R280H	K751Q	D312N	T241M
			<i>ERCC2_18880_A>C</i>	<i>ERCC2_6540_G>A</i>	

Data used in more than one published analysis was only included once in any given meta-analysis. Abstracts were excluded. There were no language restrictions on the searches. After searches were complete, abstracts were screened to ascertain that the study had smoking exposure, one of the SNPs of interest and non-cases. Full text versions were obtained for all

articles meeting these criteria except the one non-English paper. Non-English articles were included in the search to ascertain the proportion of missing studies caused by the language exclusion. Full text articles were evaluated for appropriate control data. All articles were screened and abstracted by E. Hodgson.

3. Data abstraction

If an abstract was not excluded by the initial screening, the paper was retrieved and reviewed for data appropriate for construction of, at minimum, a 2x2 table for genotype-smoking association in controls. If an unadjusted OR for any genotype-smoking association could be calculated, the as much of the following information as possible was abstracted: SNP, genotype categories (3 genotypes, dominant and/or recessive models), smoking status and dose categories [ever/never, current/not current, smoker/non-smoker, ever/former/current, pack-years (PY), duration and/or intensity], and cell counts for all genotype and smoking categories. Cell counts and smoking and genotype categories were abstracted onto preprinted forms to reduce data entry errors.

The following study characteristics were abstracted directly into an Excel spreadsheet for coding: year of publication, study design (case-control, cohort, cross-sectional, convenience, other), source of control group (for case-control: population, hospital, friends and non-blood related family, convenience, community, neighborhood, other; for cohorts: population, occupational, convenience, other), type of clinic that hospital- or clinic-based control groups were from (disease clinics, checkup clinics), matching characteristics (none, frequency-match, individual-match, matched on age, gender, ethnicity or other), study outcome (cancer [type], non-cancer disease, non-disease), full study/control group size (N), size of SNPxSmoking subset (N), country, percent male participants, ethnicity (% white, %

African American, % Asian American, % Han Chinese, % other), Hardy Weinberg equilibrium p-value, minor allele frequency, health exclusion criteria for study/control group (no exclusions, history of case cancer, history of any cancer, “cancer-free” only, other health conditions, “healthy” only, other), number of genes/SNPs assayed, and number of smoking categories. If the HWE p-value or MAF was not given but could be calculated, it was calculated and included. An estimate of central tendency for participants’ age in years (designated “average age”) was derived for each study using, in order of preference: median age, mean age, weighted average across study age categories, midpoint of age range. No individual-level characteristics were abstracted.

4. Meta-analyses

Environmental exposure: tobacco smoking behavior, including dose and duration, was the environmental factor investigated. Smoking was chosen due to public health importance conveyed by the high prevalence of the smoking and the strong justification for further gene-environment interaction studies. Smoking status was categorized as (1) ever/never (referent), and (2) current/not current (referent). Smoking amount was analyzed as (1) pack-years [PY, lowest non-zero category (referent) compared to highest], (2) duration [years, shortest non-zero category (referent) compared to longest] and (3) smoking intensity [cigarettes/day, lightest non-zero category (referent) compared to heaviest]. Original study categorizations were used when possible. The categories “passive smoking only” and “never active or passive smoking” were combined into “never” for our analyses. For analyses of current/not current smoking, never+former smokers and “non-smokers” (if identified as not current smokers) were considered not current smokers. Pack-years of smoking (pack-years = number of packs smoked per day multiplied by years smoked; 20 cigarettes=1pack) were

analyzed as relative categories [lightest smokers vs. heaviest smokers regardless of PY cutpoints] and absolute categories [< specified range of PY cutpoints vs. \geq the same range of PY cutpoints; ranges varied by SNP]. The ranges were chosen to maximize the number of included studies while keeping the range small enough that no study would have >1 cutpoint in a specified range. Similar to PY, smoking intensity was categorized by relative (lightest vs. heaviest smokers regardless of cigarette/day cutpoints) and absolute (<20 vs. \geq 20 cigarettes/day) measures. Smoking duration cutpoint range was 20-40 years, inclusive, for all SNPs.

Genetic exposures: Six polymorphisms in 3 DNA repair genes. The six SNPs were *XRCC1* [Arg399Gln (rs25487), Arg194Trp (rs179872), Arg280His (rs25489)], *XPB* [Lys751Gln (rs13181), Asp312Asn (rs1799793)] and *XRCC3* [Thr241Met (rs861539)]. The three genes participate in 3 different DNA repair pathways, the BER pathway (*XRCC1*), the NER pathway (*XPB*), and the DSB pathway (*XRCC3*). These genes were chosen for their relevance to cancer risk, the relatively high prevalence of several SNPs in these genes, the availability of published control group data, and their relevance to further interaction studies with tobacco smoking. Because there were too few studies that provided the joint distribution of genotype and smoking using all three genotypes (homozygous common allele, heterozygous, homozygous for the variant allele), all analyses were done using the dominant model (homozygous for the common allele as the referent vs. genotypes with any variant allele). Studies where only the $OR_z(95\% CI)$ for the recessive model were presented were included in the systematic review, but not in any of the meta-analyses.

Outcome measure: Unadjusted odds ratios (OR_z) and 95% confidence intervals for each SNP-smoking association were calculated from cell counts (Stata 9.2, using metan STB-44: sbe24).

Summary estimates: Forest plots with visually weighted point estimates were used to graphically display individual OR_z (95% CI) for each study and the summary OR_z (95% CI) for each SNP-smoking pair. To reduce the possibility of results being confounded by ethnicity (population stratification) in overall analyses and when examining the study characteristics likely to vary strongly by ethnicity (Hardy Weinberg equilibrium p-values and minor allele frequency) studies were stratified by ethnicity and treated as separate studies if possible. Fixed effects models were used for summary estimates unless studies were too heterogeneous to combine. Cochran's Q two-sided homogeneity p-values ($\alpha=0.10$) were used to assess overall heterogeneity in odds ratios [189].

Funnel plots were generated for inspection and testing. Two tests for statistical significance of funnel plot asymmetry were used were used to assess the potential for publication bias [190]. When data were sufficient ($N_{\text{studies}} \geq 5$), asymmetry was formally assessed using Begg and Mazumdar's test [191] and Egger's test [192] at $\alpha=0.10$. Funnel plots graph the effect size of each study against its variance (or other measure associated with sample size, inverse variance etc.). Generally, as variance increases effect sizes also increase if all studies, both positive and negative, have been published, creating a funnel shaped plot. If there is an area of the plot where studies that have similar effect sizes and variance are missing, for instance, small studies with null or negative findings, the 'funnel' will be asymmetrical. This could indicate a publication preference for large studies, regardless of results, and small studies only when the results are positive. There are other factors that can

cause funnel plot asymmetry but a symmetrical plot argues against publication bias. It is unlikely that publication bias based entirely on the magnitude of interaction effects is a common occurrence. However, it is possible that null results for main effects in small studies may be correlated with positive results in interaction analyses and produce some degree of publication bias.

5. Study characteristic analysis

Stratified analysis: To further the goal of understanding the heterogeneity of the included studies, multiple study characteristics were abstracted for stratified and meta-regression analysis. In study characteristic analysis, the goal is to see if effect size or homogeneity varies, on average, by strata of a given study characteristic. For instance, stratified analysis will show whether studies with controls out of HWE have a summary OR_z different than the summary OR_z of studies with controls in HWE. Cochran's Q statistic two-sided homogeneity p-values were used to assess heterogeneity within strata, as it is also of interest to see the SNP-smoking OR_z s with a subgroup of studies defined by a study characteristic are more homogeneous than for that SNP-smoking OR_z overall. Consistent with the goal of study characteristic analysis, stratum-specific estimates for each subgroup of studies were calculated regardless of homogeneity test results.

Stratified analysis of study design was performed for all SNP-smoking combinations, because this study characteristic was considered *a priori* the most important, given that the independence assumption applies population-based controls or cohorts. However, due to sample size considerations, the remaining study characteristics were examined only for *XRCC1* 399, *XPD* 751 and *XRCC3* 241 and only for ever-never smoking, current-not current smoking and pack-years of smoking. Forest plots were done for each stratified gene-smoking

association examined. Stratum-specific funnel plots were constructed for different strata of study characteristics where the data allowed. Stratified random-effects meta-analyses were used when the overall SNP-smoking association had a Cochran's Q p-value $\alpha < 0.10$, otherwise fixed effects meta-analysis was used, regardless of the homogeneity p-values of individual strata. For purposes of comparing strata either random- or fixed-effect models will allow comparison. The same method was used for all study characteristics of a given SNP-smoking combination to ensure comparability within SNP-smoking study characteristic analysis.

Meta-regression: Consistent with the goal of exploring study heterogeneity, meta-regression was performed regardless of homogeneity of the studies included in the summary OR_z s. Meta-regression provides a formal comparison of the stratum-specific OR_z s by study characteristic. It produces a ratio of the stratum-specific OR_z compared to the reference stratum for that study characteristic, therefore a ratio of ratios (ROR) with corresponding 95% CI. Therefore, regression coefficients for each study characteristic in the meta-regression model indicate the direction and magnitude of the association between that study characteristic and the magnitude of the OR_z . Meta-regression was performed for all SNP-smoking combinations where sample size allowed. The minimum conditions for generating meta-regression estimates (RORs) was that there were at least two studies in each of at least two strata. Because the sample size was generally small within strata multivariable regression (including multiple study characteristics in a single model) was not a viable modeling strategy and was not performed.

C. Gene-smoking association in CBCS and NCCCS controls

1. Overview

This portion of the project consisted of an empirical analysis of control group data from two population-based epidemiologic studies to estimate OR_z for measures of smoking and a panel of polymorphic genes plausibly related to smoking behavior. OR_z is of interest because it is itself a measure of bias in the COR, an estimate of the interaction parameter from a case-control study (SIM). The purpose in estimating OR_z is to estimate the degree of bias in the COR, relative to the SIM.

2. Study populations

As described previously, both the CBCS and NCCCS are large population-based case-control studies conducted in central North Carolina during the mid- to late 1990's that have collected extensive data on smoking and genetic exposures. Both studies over-sampled African American participants.

Both studies included urban, suburban and rural areas (CBCS: $N_{cases}=2311$, $N_{controls}=2022$; NCCCS: $N_{cases}=646$, $N_{controls}=1053$) [175, 177-180]. CBCS controls were pooled controls from Phase I ($N=790$), Phase II ($N=774$) and the Carcinoma in situ ($N=458$) study. Because the underlying study populations in the CBCS and NCCCS were similar but not identical, and agreement was of interest, controls were not pooled across studies. Both studies over-sampled African Americans. CBCS controls are all female; NCCCS controls also include male participants. Potential 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 [181].

CBCS participants were approximately half African American and half <50 years of age.

NCCCS participants were sampled such that the race, age and gender distribution of randomly selected cases is approximately 1:1 for gender and race. The CBCS and NCCCS used similar questionnaires and both have extensive data on tobacco smoking history.

3. Statistical methods: Estimation and evaluation of OR_z

Gene-smoking association was estimated for all genetic variants in the CBCS and NCCCS plausibly related to smoking behavior, including each of the DNA repair SNPs previously assessed in the meta-analysis.

Environmental exposure: In the CBCS and NCCCS smoking status was categorized as ever, former or current smoker. Four different comparisons of smoking status were derived from these: 1) ever (current + former smokers) vs. never smokers, 2) current vs. not current (never + former smokers) smokers, 3) current smokers vs. never smokers and 4) former vs. never smokers. 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 were derived from categorical variables used for packs/day and years smoked (pack-years are 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).

Genetic exposures: A sample of polymorphisms was chosen from available genotype data in the CBCS and NCCCS based on potential relevance to smoking behavior and/or other smoking-related health effects (convenience sample). Genes selected from the CBCS were 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*, *XPB*, *XPD*, *XPF*, *XPG*), oxidative stress defense genes (*MnSOD*, *MPO*, *NQO1*), a cell adhesion gene (*CDH1*) and a growth factor gene (*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*, *XPB*, *XPD*, *XPF*, *XPG*), and an oxidative stress defense gene (*MnSOD*). Methods of collection and most genotyping have been described previously [68, 171, 193-203]. Those homozygous for the most common allele (“no variant”) were the referent group (G-) and were compared to heterozygotes plus homozygotes for the less common allele (G+, “any variant”).

Hardy Weinberg equilibrium was tested at $\alpha=0.05$ for all polymorphisms except *GSTM1*, *GSTT1*, *NAT1* and *NAT2*. These genes were not categorized in such a way the HWE could be calculated. Alleles for each of these genes were grouped into 2 functional categories (null/present activity for *GSTM1* and *GSTT1*, fast/slow metabolizers for *NAT1* and *NAT2*) rather than as genotypes with the actual alleles that assort under HWE.

Gene pathways: Genes were classified *a priori* by their primary metabolic pathways (e.g. xenobiotic-metabolizing genes, DNA repair genes, oxidative stress defense genes etc.) although some genes function in multiple or overlapping pathways. For instance, some Phase I and Phase II xenobiotic metabolism enzymes have both exogenous substrates and endogenous substrates (e.g. *COMT*). All analyses considered whether there appeared to be patterns of association within and among gene pathways.

Outcome measure: Estimates of OR_z and 95% confidence intervals were generated using logistic regression. Because the goal of this project was examination of the independence assumption and potential bias introduced into case-only estimates of interaction by its violation, the magnitude, not the statistical significance, of OR_z was our primary concern. OR_z s of moderate magnitude (≥ 1.4 or ≤ 0.7) were judged, for the purposes of this project, to be of sufficient magnitude as to cause unacceptable bias in the COR in nearly all research contexts. Of course, each OR_z is only one estimate of the true underlying association and the true level of bias could be much larger or smaller, as suggested by the width of the 95% CI. Consequently, estimates with very wide confidence intervals were excluded from consideration. Confidence interval width was the ratio of the upper bound of the 95% CI to its lower bound. Only OR_z s with confidence limit ratios (CLR, upper bound of CI/lower bound of CI) less than four were included, with the exception of calculation of kappa for agreement between CBCS and NCCCS data (see section below on agreement) where OR_z s with CLR of up to 5 were used.

Population stratification: Because the CBCS and NCCCS have study participants from different ethnic groups (white and African American) it is important to consider whether population stratification could have a substantial impact on analyses. Population stratification, and the magnitude of the bias it may introduce into association studies, has been debated in the literature, most prominently by Thomas and Wacholder in 2002 [182, 204]. The term population stratification refers the fact that alleles at polymorphic sites in the genome can have different distributions in different populations, causing strata (subgroups) based on membership in these genetically distinct populations, within study populations. This can cause confounding in genetic studies under certain conditions. Specifically, if a study is

analyzed without regard to ethnic group and/or ancestry, and the allele of interest is more prevalent in some ancestrally related groups than others, and the baseline risk of the outcome varies among the same groups, confounding may occur. The term ‘population stratification’ will refer to this type of confounding from here on. Specifically, Wacholder et. al. outline 3 conditions necessary for population stratification to occur: 1. variation in genotype across ethnicities 2. variation in disease rates across ethnicities (after adjustment for known risk factors) 3. allele frequencies that track with disease rates across ethnicities (for reasons other than genotype of interest). The 4th condition they outline is actually a requirement for the confounding to substantially affect study results, rather than a requirement for confounding itself. It is that there be insufficient information on ancestry or ethnicity from study participants to reduce bias to an acceptable level [204].

Population stratification could cause bias in the estimate of control group associations between genetic exposures and smoking behavior (OR_z) if conditions 1-3 (above) were met and the self-reported racial categories in the CBCS and NCCCS were insufficient to control for this bias. Considering first the necessary conditions for confounding of the OR_z by population stratification, it is clear that it was possible. A number of the allele frequencies measured in the CBCS and NCCCS varied by race (e.g. *XRCC1* Arg399Gln [6.8% and 16.7% MAF in whites and African Americans respectively] and *XPD* Asp312Asn [6.0 and 16.1 MAF in whites and African Americans, respectively]). Smoking behavior also varied by race in this population. Although 19% of both white and African American CBCS participants reported being current smokers, never and former smoking are reported by 50% and 31% of whites and 60% and 21% of African Americans, respectively. This satisfied, at least in broad terms, the first two conditions. When there are only two subgroups in the data, the third

condition (that allele prevalence ‘track’ with outcome prevalence) is necessarily satisfied when the first two conditions are met.

The fourth condition, that the available information is not sufficient to adequately control for this bias, was much harder to evaluate. Primarily, it hinged on whether the self-reported racial categories collected for this study were an adequate proxy for 1. relevant genetically defined subpopulations and/or 2. the true unmeasured risk factor that causes differences in baseline smoking behavior. Though not conclusive, there has been some empirical evidence available to evaluate whether self-reported race/ethnicity is an adequate proxy for genetic subgroups for whites and African Americans (the two groups in the CBCS and NCCCS). In a large study (N=3636) that included participants from 4 self-reported ethnic groups, genetic cluster analysis using a panel of 326 microsatellite markers to identify clusters showed nearly perfect correspondence (99.86%) with self-reported ethnic group and genetic cluster [205]. The four ethnicities included were Caucasian, African American, East Asian and Hispanic. Multiple study sites were included and there was only minimal variation by study site within self-reported racial group. Analysis at a finer level was able to distinguish separate clusters for Chinese and Japanese participants but no reliable subgroups could be formed within Caucasians, African Americans or Hispanics [205].

While the Tang et. al. (2005) study provided evidence that broad self-reported racial and ethnic categories correspond well to genetic subgroups, at least in these study populations and for these markers, it did not address the possibility of varying degrees of admixture within these ethnic categories. An empirical analysis of the degree of bias possible in a study of N-Acetyltransferase 2 (*NAT2*) polymorphism and either male bladder cancer or female breast cancer using data that mimicked the allele frequencies and disease rates of the US population

of non-Hispanic European origin, demonstrated that bias from population substructure was of minimal importance for non-Hispanic whites in studies of cancer [206]. Taken together, these studies [205-206] indicate that self-reported information on race, as categorized in the CBCS and NCCCS as white or African American, is sufficient to appropriately stratify study participants by race and adjust for race.

Similar work has been done for the African American population, however, the historic circumstances of enforced immigration and subsequent admixture have made such work much more difficult and complex [207]. On an individual level, African American ancestry demonstrates a high degree of admixture with European ancestry as well as admixture of various ancestral populations in Africa. The degree of European admixture has been estimated at 12%-23% [208]. This admixture could be problematic if the degree of admixture tracks with smoking prevalence and with a true risk factor for smoking behavior [204-206]. In a simulation study, Wang et. al. (2004) modeled admixture from 2-10 ethnicities (subpopulations) to a maximum of gene prevalence differences between subpopulations of 5-95%, at OR=1 and OR=2. Their results showed that bias was acceptably low (<10%) at most of the scenarios presented. Bias was maximal (~20%) with only 2 ethnicities, a gene prevalence difference of 90% (5% in one group, 95% in the other group) and a true OR of 2.0. The 95% percentile of bias under this scenario, however, was a more modest 4% (OR=2.08 vs. 2.0). In CBCS and NCCCS data, the assumption of admixture from only 2 ethnicities was conservative, as it ignores admixture between African ancestral groups and Native Americans, but plausible on a broad scale. There were no gene prevalence differences as extreme as 5% & 95% in the CBCS or NCCCS controls. For alleles with MAF differences of 20% or less, which were typical in the CBCS and NCCCS, any potential bias

should be well below the maximum 9% (at-risk genotype differences 5-40% have bias of 2% at the 95th percentile).

In conclusion, although the potential for bias by population stratification, even after adjustment for race, certainly exists in the CBCS and NCCCS data, the magnitude should be small given the self-reported racial categories. Clearly some residual confounding by population stratification may have remained after controlling for race in overall analyses, but the information available was sufficient to stratify by race and examine race as an effect measure modifier before proceeding to any combined analyses.

Consequently, race-specific analyses were done for each gene variant and smoking measure. Effect measure modification by race was assessed by performing the likelihood ratio test comparing models with and without a race*smoking interaction term. Significant results for the interaction term ($\alpha=0.05$) in a majority of smoking measures precluded pooling African American and non-African American participants for that gene variant.

Modeling: Unconditional logistic regression with a dichotomous representation of the genetic variable (homozygous for common allele=referent [G-], heterozygous + homozygous for less common allele=exposed [G+]) as the dependent variable was used for all modeling. The dominant model was used to preserve power and precision of estimates of OR_z , especially for stratified analyses where cell sizes tended to be small for some subgroups (e.g. African American women in the NCCCS). A single model of the general form $\text{logit}(G+/G-) = \alpha + \beta_{(1)} E_1 + \beta_{(2-i)} COV_{(2-i)} + \text{error}$ (where G+= positive for genetic variant, E+=positive for the smoking behavior, COV=any additional covariates) was used for all SNPs.

Each dataset was evaluated for effect measure modification by stratification on race (white, African American), age (CBCS: <50y, >= 50y; NCCCS: <65y, >=65y) and gender

(NCCCS only), respectively. Effect measure modification would have also been assessed for any individual level factors identified in the meta-analysis as strong predictors of effect size, had there been any. Average age was a mildly suggestive as an influential study characteristic but would have been evaluated as a potential effect measure modifier regardless. Age distributions in the CBCS and NCCCS were different, necessitating different cutpoints for the binary age variables.

In order to decide whether to stratify analyses on race or gender a likelihood ratio test was performed comparing models with and without a race*smoking interaction term (or gender*smoking interaction term). Significant results for the interaction term ($\alpha=0.05$) in three or more smoking measures precluded pooling African American and non-African American (or male and female) participants for that SNP.

Although frequency matching procedures using randomized recruitment in the CBCS and NCCCS were based on projected case incidence, and no cases were used in the current analysis, the matching process distorted the prevalence of these factors in the underlying population, potentially affecting gene-smoking estimates. Based on directed acyclic graphs (DAGs) [209], and their status as matching factors, age (continuous), race (self-report: white or African American) and gender (NCCCS only) were assessed as potential confounders of the gene-smoking relationship. Based on the DAG (Figure V.B.1.), two additional variables were evaluated as potential confounders: first degree family history of any cancer, excluding non-melanoma skin cancer (Y/N) and total family income (<15K, 15-<30K, 30-<50K, $\geq 50K$). To order to address the possibility of missing data for family history or income introducing bias, OR_{zs} from the full dataset [adjusted for age, race and gender (NCCCS)] were compared with identically adjusted OR_{zs} in a dataset restricted to those with no missing

data. Estimates were not appreciably different, therefore the full dataset was used to assess confounding by family history of any cancer or total family income.

Percent change in β coefficients was calculated but not used to determine whether a covariate would be retained in the model; because of the high proportion of estimates close to the null (Range in CBCS: 0.5-2.5, NCCCS: 0.6-1.6) this commonly used criterion was not sufficiently informative. A potential confounder was retained if the absolute value of difference between smoking variable β coefficients from models with and without the potential confounder was > 0.15 (i.e. when $|\beta \text{ coefficient for smoking from model with potential confounder} - \beta \text{ coefficient for smoking from model without potential confounder}| > 0.15$ the covariate was retained). Rather than generating potentially dozens of gene variant-smoking-specific models, the same set of confounders was used for all gene variants for comparability across gene variants. If a covariate met this criterion for any gene variant-smoking estimate, it was retained in all models.

After assessment of effect measure modification and confounding, an association was characterized by the magnitude of OR_z (odds ratios ≥ 1.4 or < 0.7 were considered evidence of non-null association) and precision of the accompanying confidence interval. Estimates with $CLR > 4$ (upper limit/lower limit) were excluded from consideration unless otherwise stated. SAS 9.1 was used for all modeling [210].

4. Agreement

After assessing gene-variant-smoking OR_z s in the CBCS and NCCCS datasets separately, agreement between the two studies was assessed for the 15 polymorphisms included in both studies using a weighted kappa statistic [211]. The weighted kappa measures the degree of agreement between two or more raters that are using a multi-level ordinal scale

to categorize a series of subjects, beyond what would be expected due to chance alone. The raters were the CBCS and NCCCS, and the “subjects” of agreement were the 15 gene-smoking associations measured by both studies. OR_z was categorized into three categories: 1) below the null, $OR_z < 0.9$, 2) null, $0.9 \leq OR_z \leq 1.1$ and 3) above the null, $OR_z > 1.1$. Only OR_z with confidence interval widths of < 4 were assessed. With the weighted statistic disagreement between 2 adjacent categories has less influence on the statistic than disagreement between ratings further apart on the ordinal scale. The categories of Landis (1997) were used to describe strength of agreement or disagreement [212]. As a sensitivity analyses, agreement was also assessed with the definition of the null changed to 0.8-1.2 (inclusive) and including all data regardless of CLR.

5. Misspecification of smoking

This study also explored the issue of undetected bias in the COR introduced by misspecification of smoking during independence assumption evaluation. When a case-control analysis of interaction is done, all exposures are specified identically for cases and controls. For example, if smoking is categorized as ever-never for cases in a given model it is also ever-never for controls in the same model. However, when a stand-alone case-only study is considered, the independence assumption must be evaluated in ancillary data, and exposures may be specified differently than they will be in the case-only analyses. The most common specification of smoking available in the literature, ever-never, is unlikely to be the only measure of smoking assessed in a case-only analysis including smoking. When the OR_z s differ across different measures of smoking, additional bias, over and above the bias introduced when the measures are identical, will be introduced into the COR. Additionally, the decision whether or not to proceed with a case-only study may be affected.

The occurrence of moderate magnitude OR_z s for smoking status (ever-never or current-not current smoking) was compared to the occurrence of moderate magnitude OR_z s for any measure of smoking amount (duration, intensity or PY) for each genetic variant in the CBCS or NCCCS. Because the CBCS and NCCCS control data allowed estimation of OR_z across 5 different specifications of smoking for all genetic variants, particular genetic variants were identified that had discrepant OR_z s across different smoking measures. Further, since evaluation of the independence assumption in the literature is almost exclusively done by significance testing, we also compared significance testing of OR_z to the method used in this study, a method based on the precision and magnitude of OR_z (95% CI).

V. RESULTS

A. MANUSCRIPT 1: Smoking and selected DNA repair gene polymorphisms in control groups: systematic review and meta-analysis

1. Introduction

The case-only study design as proposed by Prentice et. al and promoted by Piegorsch et. al. [1-2] has been increasingly used to estimate the magnitude of statistical interaction between two measured exposures with respect to a given outcome, most commonly a genetic and an environmental exposure. This method requires only cases, no controls or defined cohort. Provided the design assumptions are met, the case-only study can estimate a specific form of statistical interaction, departure from constancy of rate ratios in the underlying population, but not main effects of the two exposures. The design assumption of interest is that the relevant exposures are independent in the underlying source population. Although the constancy of rate ratios between different strata of exposure in the underlying source population is the true parameter of interest (often represented as $Z=1$), case-control control groups are frequently used to estimate Z using OR_z . OR_z is ideally the odds ratio from an unmatched density-sampled control group of a case-control study but in practice many types of control groups have been used.

There are potential advantages to the case-only method in several settings. The lack of requirement for a control group has obvious cost advantages, but there are methodological and ethical advantages as well. Estimation of the interaction parameter from case-only analyses is more efficient than for a traditional case-control study (i.e. fewer cases are

required for similar precision of estimate) and with no need for controls, there are fewer participants overall [8]. Not having controls may mitigate selection biases due to, for example, differential recruiting success between cases and controls, or differential recall of environmental exposures by case-control status. Invasive procedures that are part of cases' diagnosis or treatment often cannot be done ethically in healthy volunteers, especially vulnerable groups such as pediatric populations [9]. But these advantages come at a cost. A case-only study only estimates interaction on a multiplicative scale (deviation of the rate ratio for those having both the genetic and environmental exposures from the product of rate ratios for those with either the genetic or the environmental exposure, but not both). It cannot estimate the independent effect of either exposure, or interaction on the additive scale (deviation of the rate ratio for those having both the genetic and environmental exposures from the sum of rate ratios for those with either the genetic or the environmental exposure). This limits its use to situations in which the independent effects of the two exposures are not of interest, nor are synergism or antagonism of the exposures [213-214]. Control-selection bias is the only validity threat the case-only design avoids, in comparison with the case-control design. Consequently, case-only studies have been proposed by several investigators as a mere screening method to identify candidate gene-environment or gene-gene interactions [5, 16-17].

However, the increase in precision and avoidance of control-selection bias in the case-only method requires a major assumption: that the two exposures are independent in the source population ($Z=1$) [1-2]. Data simulations have demonstrated that violations of the independence assumption that have a small magnitude can strongly bias the case-only interaction parameter, increase the mean-squared error (MSE), inflate size of Neyman-

Pearson hypothesis tests (i.e. the actual probability of rejecting a true null hypothesis) above the maximum tolerable level (e.g. $\alpha=0.05$)[5], and thereby reduce confidence interval coverage probabilities below their specified values (e.g. 95%). As Z grows further from 1 in either direction, these problems also increase appreciably. When control-group gene-environment (G-E) associations are of similar magnitude but opposite in direction to the interaction effect, a case-only study may not detect interaction effects, a Type II error [5, 124].

Generally, when the ‘implausibility’ of specific G-E associations is argued in published case-only analyses, it is considered only within the framework of causality, and non-causal scenarios are rarely invoked. Arguing from the causal perspective seems unwise for many, if not most, of the relevant gene-environment associations in the face of the wide variety of gene-behavior associations considered plausible enough for investigation (e.g. smoking behavior or diet), and our incomplete knowledge of genetic influences on health-related behaviors. The strongest example of a causal independence assumption violation is the well-known association between aldehyde dehydrogenase 2 (*ALDH2*) genotype and alcohol consumption, in which the variant allele produces unpleasant physical reaction when alcohol is consumed, greatly reducing alcohol consumption in carriers [72, 74]. Similarly, skin pigmentation is strongly associated with sun exposure, rendering a case-only study of the interaction of skin pigmentation and UV exposure in cancer invalid. Genetic influences on smoking behavior are also an active area of research, for example the association (or lack thereof) between *CYP2A6* and smoking behavior has received considerable attention in the last 15 years [87-88].

Gene-environment correlations in populations (sources of non-independence) can be non-causal as well. For example, Z can vary from the null if environmental and genetic exposures have been non-differentially misclassified with respect to each other, as can happen when population subgroups, often ethnic groups, have different gene variant prevalences and different patterns of environmental exposure (population stratification [4]. Non-random misclassification of either the genetic exposure (e.g. genotype is measured perfectly but is in linkage disequilibrium with the causal variant rather than actually being the etiologically active genotype) and the environmental exposure (e.g. heavy smokers underreport smoking more than light smokers) can also create an apparent association in a study population.

Selection bias could also cause association between two exposures in a control group being used to estimate Z . For instance, if smokers with a family history of the outcome are less likely to participate as controls than smokers without a family history or non-smokers with a family history, a spurious inverse control group association could be created between smoking and any genetic exposure related to family history. Cohort effects could affect control group associations if, for instance, a genetic exposure is associated with longevity, and the environmental exposure is one that has changed prevalence in the population over time, such as smoking or dietary patterns. When OR_z is being used to estimate Z , such distortions of OR_z can mislead investigators about the true magnitude and direction of Z in the source population and lead to incorrect interpretation of the COR. Chance can also play a role. Since the expectation that $OR_z=1$ when $Z=1$ is a large sample asymptotic approximation, as sample size decreases, OR_z will deviate from the null with increasing frequency through random error alone [5]. Consequently, as Z is estimated by OR_z in subgroups, and sample size drops, OR_z has a higher and higher probability of deviating from Z by chance alone. A prudent approach

to the independence assumption would be to thoroughly examine any empirical evidence and biologic theory for or against causal association between the relevant gene variants and exposure before proceeding with a case-only study.

Smoking is an environmental exposure that is commonly measured, can be quantified, and is important both in gene-environment interaction research and for public health overall. Variation in DNA repair is thought to be important in cancer [152, 215]. Three polymorphic DNA repair genes, X-ray cross complementing gene 1 (*XRCC1*), xeroderma pigmentosum complementation group D [*XPD*, previously excision repair complementing defective 2 (*ERCC2*)], and X-ray cross-complementing gene 3 (*XRCC3*) which participate in the base excision repair (BER) pathway, the nucleotide excision repair (NER) pathway, and the double strand break (DSB) pathway, respectively, have single nucleotide polymorphisms that have been investigated in numerous studies, particularly cancer studies. Three important non-synonymous single nucleotide changes (SNPs) have been studied for *XRCC1*: Arg399Gln (rs25487), Arg194Trp (rs1799782), and Arg280His (rs25489) [216-217]. A single nucleotide change in *XPD* exon 10 (Asp312Asn, rs1799793) and another in exon 23 (Lys751Gln, rs13181) have been studied [218-219]. *XRCC3* has one studied amino acid-changing variant, a Thr241Met variant (rs861539) [137, 219]. The BER pathway is largely responsible for repair of oxidative damage and the NER pathway for repair of bulky DNA adducts, both types of damage produced by constituents of tobacco smoke [151]. With the exception of the *XRCC1* Arg194Trp [170] the variants are thought to code for reduced DNA repair capacity, particularly *XRCC1* Arg280His [200], although functionality for some SNPs has not been definitely established [170, 172-174, 220]. Cigarette smoke is clearly genotoxic, with multiple studies showing smokers have increased rates of sister chromatid exchange and

micronuclei formation in lymphocytes, increased DNA strand breaks in lymphocytes, buccal cells and urothelial cells, and for heavy smokers, oxidative damage to DNA in germ cells [136].

We undertook a systematic a systematic review and meta-analysis of DNA repair variation and smoking behavior in control groups, using OR_z to estimate Z. The purpose in estimating OR_z was to estimate the degree of bias in the COR, relative to the interaction estimate from a case-control analysis, assuming no control-selection bias. Heterogeneity was explored using stratified analysis and meta-regression of study characteristics. The primary aim of this project is to evaluate the importance of the independence assumption for these SNPs and smoking behavior and enable investigators considering a stand-alone case-only study of gene-environment interaction to evaluate the independence assumption in a range of relevant conditions (e.g. cancer, cardiovascular disease, neurological diseases) more rigorously than has been done previously, potentially identifying situations in which case-only estimates may be more or less valid.

2. Methods

Data Abstraction

PubMed, ISI Web of Science and the CDC Genomics and Disease Prevention (GDPInfo: <http://apps.nccd.cdc.gov/genomics/GDPQueryTool/frmQueryAdvPage.asp>) databases were searched up to March 6, 2007 for peer-reviewed literature likely to contain data on the joint distribution of any of the polymorphisms of interest [single nucleotide polymorphisms *XRCC1* Arg399Gln, Arg194Trp, and Arg280His, *ERCC2(XPD)* Lys751Gln and Asp312Asn, and *XRCC3* Thr241Met] and smoking behavior in non-case groups. Non-case groups were defined as any group not selected as the index group based on disease status

(e.g. cohorts, convenience samples and control groups from case-control studies). For simplification non-case groups will be referred to as controls throughout this article. There were no language restrictions on searches. A list of keywords for PubMed and the ISI Web of Science was developed in consultation with an information specialist from UNC Health Science Library to ensure that searches would be as inclusive as possible. Keywords for PubMed were included as MeSH terms and text words whenever possible. Keywords for smoking were “smoking”, “tobacco”, “tobacco smoke”, “tobacco smoke pollution”, and “smoker”. The SNPs were searched by combining “polymorphism” and “polymorphism, genetic” with the SNP-specific keywords “*XRCC1*”, “*XPB*”, “xeroderma pigmentosum group d protein”, “*ERCC2*” and “*XRCC3*”. ISI Web of Science keywords were “smok*” and “tobacco,” and “*XRCC1*”, “*XPB*”, “*ERCC2*” and “*XRCC3*”. GDPInfo was searched using the advanced query and limiting by factor menu terms: “smoking behavior”, “smoking (tobacco) passive”, “smoking (tobacco) bidi”, “smoking (tobacco)”, “smoking (tobacco) maternal”, “tobacco”, “indoor air pollution”, “nicotine (nasal spray)”, and “nicotine (transdermal)”, and gene menu terms: “*XRCC1*”, “*XPB*”, “*ERCC2*” and “*XRCC3*”. No disease limits were used.

Inclusion criteria were deliberately broad. To be included, an article had to contain original control group data on the joint distribution of any genotype of interest (listed above) and any aspect of tobacco smoking behavior. This was most often a table with counts of participants cross-classifying the specific genotypes and smoking behaviors. Textual data that could be converted to an analogous table was also included. Reviews, animal studies, cell culture studies, case reports, case-only studies, abstracts, letters and editorials were excluded. The articles were reviewed by the first author of this paper (EH).

Abstracts were screened to determine whether the study included control participants and relevant genotype and smoking data. SNP designations considered equivalent are shown below (Table V.A.1).

Table V.A.1 SNP designations for data abstraction

<i>XRCC1</i>	<i>XRCC1</i>	<i>XRCC1</i>	<i>XPB (ERCC2)</i>	<i>XPB (ERCC2)</i>	<i>XRCC3</i>
Arg 399 Gln	Arg 194 Trp	Arg 280 His	Lys 751 Gln	Asp 312 Asn	Thr 241 Met
G28152A	C26304T	G27466A	A35931C	G23591A	C18067T
exon 10	exon 6	exon 9	exon 23	exon 10	exon 7
rs25487	rs1799782	rs25489	rs13181	rs1799793	rs861539
R399Q	R194W	R280H	K751Q	D312N	T241M
			<i>ERCC2_18880_A>C</i>	<i>ERCC2_6540_G>A</i>	

If an abstract was not excluded by the initial screening, the paper was retrieved and reviewed for data appropriate for construction of, at minimum, a 2x2 table for genotype-smoking association in controls. If an unadjusted OR for any genotype-smoking association could be calculated, the following data were abstracted: SNP, genotype categories (3 level additive, dominant and/or recessive models), smoking status and dose categories [ever/never, current/not current, smoker/non-smoker, ever/former/current, pack-years (PY), duration and/or intensity], and cell counts for all genotype and smoking categories. Cell counts and smoking and genotype categories were abstracted onto preprinted forms to reduce data entry errors. The following study characteristics were also abstracted: year of publication, study design (case-control, cohort, cross-sectional, convenience, other), source of control group (for case-control: population, hospital, friends and non-blood related family, convenience, community, neighborhood, other; for cohorts: population, occupational, convenience, other),

type of clinic that hospital- or clinic-based control groups were from (disease clinics, checkup clinics), matching characteristics (none, frequency-match, individual-match, matched on age, gender, ethnicity or other), study outcome (cancer [type], non-cancer disease, non-disease), full study/control group size (N), size of SNPxSmoking subset (N), country, percent male participants, ethnicity (% white, % African American, % Asian American, % Han Chinese, % other), Hardy Weinberg equilibrium p-value, minor allele frequency, health exclusion criteria for study/control group (no exclusions, history of case cancer, history of any cancer, “cancer-free” only, other health conditions, “healthy” only, other), number of genes/SNPs assayed, and number of smoking categories.

An estimate of central tendency for participants’ age in years (designated “average age”) was derived for each study using, in order of preference: median age, mean age, weighted average across study age categories, midpoint of age range. No individual-level characteristics were abstracted. One non-English language article could not be evaluated.

Selection of Study Comparisons

Three of the included study populations had control data in more than one article; however different SNPs were studied [221-226]. One study population had control data for *XRCC1* 399 and smoking stratified by ethnicity in one article [227] and not stratified by ethnicity in another [200]. Preference was given to the larger N unless a given study characteristic could be best examined using OR_z estimates stratified by ethnicity. No study population contributed to any analysis more than once maintaining independence of observations. Analyses focused on associations with genotype categorized using a dominant model (i.e. homozygotes of the most common allele were the referent group, compared to heterozygotes plus homozygotes of the minor allele) due to the small number of studies that

provided sufficient information to assess recessive or additive models. Smoking behavior was characterized by constructing five metrics from the available control data. Smoking status was categorized as (1) ever/never (referent), and (2) current/not current (referent). Smoking dose was analyzed as (1) pack-years [PY, lowest non-zero category (referent) compared to highest], (2) duration [years, shortest non-zero category (referent) compared to longest] and (3) smoking intensity [cigarettes/day, lightest non-zero category (referent) compared to heaviest]. Original study categorizations were used when possible. “0 PY of smoking” and “0 years of smoking” were considered equivalent to never smoking. The categories “passive smoking only” and “never active or passive smoking” were combined into “never” for our analyses. Studies that did not provide sufficient data to include ‘passive only’ smoking in the never smoking group were excluded. For analyses of current/not current smoking, never+former smokers and “non-smokers” (if identified as not current smokers) were considered not current smokers. Pack-years of smoking (pack-years = number of packs smoked per day multiplied by years smoked; 20 cigarettes=1pack) were analyzed as relative categories [lightest smokers vs. heaviest smokers regardless of PY cutpoints] and absolute categories [< specified range of PY cutpoints vs. >= the same range of PY cutpoints; ranges varied by SNP]. The ranges were chosen to maximize the number of included studies while keeping the range small enough that no study would have >1 cutpoint in a specified range. Similar to PY, smoking intensity was categorized by relative (lightest vs. heaviest smokers regardless of cigarette/day cutpoints) and absolute (<20 vs. >= 20 cigarettes/day) measures. Smoking duration cutpoint range was 20-40 years, inclusive, for all SNPs.

Statistical analyses

For all SNP-smoking analyses, crude ORs and 95% confidence limits were calculated from cell counts (Stata 9.2, using metan STB-44: sbe24). Funnel plot asymmetry, an indicator of possible publication bias [190], was considered suggestive of study characteristics associated with variance and Z. When data were sufficient ($N_{\text{studies}} \geq 5$), asymmetry was formally assessed using Begg and Mazumdar's test [191] and Egger's test [192] at $\alpha=0.10$. Cochran's Q two-sided homogeneity p-values ($\alpha=0.10$ due to low power of the test) were used to assess overall heterogeneity in odds ratios [189]. Where appropriate, summary odds ratios were estimated using Mantel Haenszel methods with fixed effects.

Study characteristic analyses: Key study characteristics hypothesized to influence variation in the strength of SNP-smoking associations among controls across studies were assessed using stratified meta-analysis and random-effects meta-regression, with the among-study variance estimated by restricted maximum likelihood [228]. Stratified meta-analysis produces a summary OR_z estimate for each stratum of a study characteristic. Meta-regression provides a formal comparison of the stratified estimates in the form of an estimated ratio of odds ratios.

Study characteristics were selected *a priori*. They included (1) study design (case-control, cohort, or convenience; patient-based control groups, healthy control groups), (2) continent, (3) ethnicity, (4) Hardy-Weinberg equilibrium p-value, (5) average age, (6) gender (% male), (7) study outcome (lung cancer, other cancer, non-cancer disease, non-disease), (8) minor allele frequency and (9) smoking prevalence. Study design was examined for all SNP-smoking combinations; additional study characteristics were examined for *XRCC1* 399, *XPB* 751 and *XRCC3* 241. Stratified random-effects meta-analyses were used when the overall SNP-smoking association had a Cochran's Q p-value $\alpha < 0.10$, otherwise fixed effects meta-

analysis was used, regardless of the homogeneity p-values of individual strata. To reduce the possibility of results being confounded by ethnicity (population stratification) in overall analyses and when examining the study characteristics likely to vary strongly by ethnicity (Hardy Weinberg equilibrium p-values and minor allele frequency) studies were stratified by ethnicity and treated as separate studies if possible. Stata 9.2 was used for all analyses. Results for study characteristics were assessed for consistency across smoking categories and across SNPs.

3. Results

Eligible studies: The literature searches identified 228 articles for evaluation. Of these, 55 articles were eligible for inclusion. The primary reason for exclusion was that an article did not present the genotype-smoking distribution in controls (N=98, 57% of exclusions). Exclusion reasons for the remainder included: review article or abstract only (13%), did not assess any relevant SNPs (9%), and did not have any non-cases (10%). Finally, of the 55 studies eligible for inclusion, five were not included in final summary estimates because no data were presented for dominant genetic model [46, 229], no measure of adult smoking behavior [230], former smokers excluded [231], or never smokers were included in lowest PY category [232]. No studies presented all five measures of interest for smoking behavior. Fifty articles representing 46 distinct study populations were included in the final meta-analyses (brief study descriptions in Table V.A.1a). Table V.A.1b presents five studies that were included in the systematic review but not in any meta-analyses. The number of individual controls included in each summary estimate ranged from 11,789 (*XRCC1* 399 ever/never smoking) to 305 (*XPD* 312 current/not current smoking). Generally, compared to the total N of observations in ever-never analyses, there were ~40% fewer observations in

current/not current analyses, and ~20% as many in PY analyses. The number of study populations included for each polymorphism was as follows: *XRCC1* Arg399Gln (N=32), *XRCC1* Arg194Trp (N=16), *XRCC1* Arg280His (N=8), *XPD* Lys751Gln (N=16), *XPD* Asp312Asn (N=9), and *XRCC3* Arg241Gln (N=13). Thirty-seven studies presented the control distribution of genotype and ever/never smoking, 16 for current/ not current smoking and 14 for PY. Far fewer presented duration (N=4) and/or intensity (N=4). Case-control studies predominated with 12 population-based [200, 222-223, 226-227, 233-239] and 23 hospital-based [221, 224-225, 240-259], four nested [260-263] and two other case-control studies. Most control groups were from cancer case-control studies (N=39), one was from a case-control study of rheumatoid arthritis. Nine cohort or convenience sample studies examined non-cancer outcomes, predominantly measures of DNA damage (8 of 9), one measured genotype frequency.

Association between DNA repair gene variants and smoking behavior. Across SNPs there was more variation in ORs assessing control-only G-E associations (OR_{zs}) for measures of smoking amount (PY, duration, intensity) than for measures of smoking status (ever-never, current-not current) (Table V.A.2). Ten of 11 summary estimates of smoking status fell between 0.9-1.1. Summary estimates for smoking amounts were distributed more broadly, with only five of 10 summary estimates between 0.9-1.1; the most extreme measures were found for duration and intensity. Although only two of 18 genotype-smoking groups were too heterogeneous for a fixed effects summary estimate, nearly all groups had study estimates above and below the null, generally varying 2-3 fold.

For *XRCC1* 399, three measures of smoking behavior were homogeneous enough for a summary estimate of the association between variant allele and smoking: current smoker/not

current (N=11), PY (N=9) and intensity (N=4). Higher PY and heavier smoking intensity, but not current vs. not current smoking, were associated with *XRCC1* Arg399Gln (any Gln) [OR (95%CI): 1.2 (1.0, 1.5) and 1.5(1.2, 1.9), respectively]. Odds ratios for *XRCC1* 399 and ever-never smoking ranged from 0.7 (95% CI: 0.3, 1.7) [250] to 1.9 (95% CI: 1.0, 3.7) [238] (Table VIII.B.1). After two studies were stratified by ethnicity and treated as separate studies, 13 studies had a genotype-smoking OR >1 and 10 had an OR<1. For the other two SNPs in *XRCC1* (194 and 280), having the variant allele was associated with longer smoking duration [*XRCC1* 194: 0.7 (0.5, 0.9), *XRCC1* 280: 1.2 (0.6, 2.3)] and current smoking [*XRCC1* 280: 1.2 (0.6, 2.3)] though confidence intervals were wide. For the two *XPD* SNPs (751, 312) there was considerable variation in the association between *XPD* 751 variant allele and higher PY. Study estimates ranged from 1.4 (0.8, 2.6) [221] to 0.5 (0.3, 1.0) [254] (Table VIII.B.4). Higher PY were associated with the variant allele for *XRCC3* 241 although the number of studies was small (N=4).

Sensitivity analyses. Among the studies that were assessed for current-not current smoking, a subset could also be assessed for never, former or current smoking (Table V.A.3). No consistent pattern emerged for comparisons of never smoking with former or current smoking. Absolute measures of PY, intensity and duration were calculated and compared to relative measures for consistency. Genotype-PY estimates for absolute cutpoints (i.e. all PY categories below specified cutpoint range vs. all categories above that cutpoint) were comparable to estimates using relative categories (lowest non-zero category vs. highest) although strata were sparse (Table V.A.3). Additionally, when studies with only smokers were dropped and never smoking was used as the reference category, results were essentially the same for relative and absolute measures of PY. Genotype-smoking association between

XRCC1 Arg399Gln and smoking intensity (cigarettes/day) could be estimated in four studies. There was an association between *XRCC1* 399 any Gln and greater smoking intensity [1.5(1.2, 1.9)]. As with PY, this association was consistent across both methods of smoking intensity categorization (lowest study-defined category vs. highest study-defined category, and <20 cigarettes/day vs. >20 cigarettes/day). Estimates for the two excluded studies that had referent groups roughly comparable to ever/never smoking and PY [231-232] were similar to those for included studies (Table V.A.1b).

Ever-never analyses included studies that did not present ever-never smoking as such, but had smoking amount data, usually PY, that was used to derive ever-never smoking. To see whether these studies differed from studies presenting only ever-never data, we excluded studies that did not also present smoking amount information. There was no difference in the distribution of study estimates or summary estimates [*XRCC1* 399: range of ORs 0.8 – 1.9, Cochran's Q p-value 0.02; *XPD* 751 summary estimate: 1.0 (0.8, 1.1)].

Funnel plot asymmetry. There was no evidence of funnel plot asymmetry for overall genotype-smoking associations (data not shown). In formal testing, the majority of p-values (75%) were ≥ 0.3 . The lowest p-value was $p=0.14$ for *XRCC1* 280 ever/never for both Begg and Egger tests.

Study characteristics. For study characteristics, stratified associations and univariate meta-regression were evaluated across SNPs and smoking categories. Associations were evaluated primarily on the basis of consistency and direction. Study design was examined for all six SNPs for ever/never, current/not current smoking and PY. For smoking status, genotype-smoking associations for *XRCC1* 399 and 194 and *XPD* 751 and 312 were generally stronger for population-based case-control studies than for hospital-based or patient-based

control groups, although the magnitude of the differences was small; the range of RORs was 0.7 to 0.9 for hospital/patient-based compared to population-based controls (referent) (Table V.A.4). However, for smoking dose as measured by PY (2 evaluable SNPs, *XRCC1* 399 and *XPB* 751) the hospital-based/patient-based control groups showed stronger genotype-smoking associations than population-based control groups (range of RORs: 1.2-1.5). When examining PY, for all SNPs, the genotype-smoking association for population-based control groups was below the null. The remaining study characteristics were examined only for *XRCC1* 399, *XPB* 751 and *XRCC3* 241 (Tables 5, 6, and 7 respectively) due to sparse data for the other SNPs.

For PY lung cancer studies were above the null for all three SNPs. Further, when compared to studies of other cancers the genotype-smoking association was stronger for lung cancer studies (referent) compared to other cancer studies [ROR= 0.8(0.5, 1.2) and 0.5(0.3, 0.9) for *XRCC1* 399 and *XPB* 751, respectively]. All studies with PY were cancer studies. Older average age of study participants (>63y vs. ≤59y and >median age) weakly but consistently showed stronger associations between ever smoking and variant allele for *XRCC1* 399, *XPB* 751 and *XRCC3* 241 than did younger age. For *XRCC1* 399 only, this was evident across all three smoking categories. Also, for *XRCC1* 399 current-not current smokers and PY only, studies with lower minor allele frequencies (N=3) showed stronger associations (~2.0) than those with higher MAF. These three studies had only African-American or Asian participants. No strong and/or consistent patterns emerged for the other study characteristics examined: continent (North America, Europe, and Asia), ethnicity (White, African American, Han, multi-ethnic), HWE p-value (<0.1, ≥0.1), gender (% male; all male, mixed gender, all

female), minor allele frequency (tertiles for each SNP), and smoking prevalence proportion (≤ 0.507 , > 0.507).

4. Discussion

This systematic review of DNA repair genotypes and smoking behavior in control data from 46 study populations was conducted with the goal of informing the practice of case-only analyses of gene-environment interaction. Results from this systematic review and meta-analysis show considerable variation in estimates of Z for *XRCC1* 399 ever-never smoking and *XPB* 751 PY (Cochran Q p -values < 0.1 , ~ 3 fold range in ORs, ORs on both sides of the null). Even when studies were homogeneous enough for a summary estimate, point estimates of OR_z varied as much as 5-fold. Summary estimates for individual SNPs varied across smoking categorizations, with larger magnitudes of association generally found for measures of smoking dose (PY, intensity, duration) than for smoking status (ever-never, current-not current). There was a weak association between *XRCC1* 399 and higher smoking dose (PY, intensity). No study characteristics examined strongly predicted the magnitude of association although study outcome (lung cancer vs. other cancer for PY), study design (population-based vs. hospital/patient-based), and age warrant further investigation.

A key assumption of the case-only study design for interaction is that the genotype and environmental exposure are independent in the underlying population [3]. Descriptively, any deviation from the $Z=1$ (estimated by OR_z) introduces bias in the case-only interaction estimate [5]. Further, when $Z \neq 1$ in population subgroups, the COR for those subgroups will be biased as well. The bias introduced into the COR is in addition to other sources of bias, such as selection bias, information bias etc.

Although the validity of case-only estimates rests heavily on the independence assumption, literature on independence assumption verification is scant. Data simulations have demonstrated that even small violations of the independence assumption can strongly bias the case-only interaction parameter [5]. Using data from a study of *XRCC1* genotype and lung cancer by Ratnasinghe et. al., Albert et. al. showed that a control group association between genotype and pack-years of tobacco use of $OR_z=2.0$ created bias in the case-only odds ratio (COR) of 105% [$COR=0.9(0.4, 1.9)$], synergy index from case-control analysis of the same data (SIM)= $0.4(0.2, 1.2)$] [5, 126]. Even an OR_z of 1.2 biased the COR by nearly 30%. Another notable paper to focus on independence assumption evaluation using data simulation is by Gatto et. al. who elucidate conditions under which a control group is an appropriate proxy for the underlying study population when validating the independence assumption [127].

However, little empirical work has been done on the magnitude of control-only associations (OR_z) between DNA repair gene variations and smoking that quantitatively assesses this additional bias. A population-based study (N=339) of Japanese males assessed association between ‘habitual smoking’ (ever/never) and a panel of 153 SNPs in 40 candidate genes, including the DNA repair genes *OGG1* and *NUDT1*(*MTH1*) [64]. Association was found between smoking and 3 of 4 of the SNPs in *OGG1* (0.4-0.6, borderline statistical significance, variant carrier vs. variant non-carrier) but no statistically significant associations were found for *NUDT1*.

Smoking dose (PY and/or intensity) could be causally associated with variation in *XRCC1* 399, or with a polymorphism in linkage disequilibrium with *XRCC1* 399. There is evidence that the *XRCC1* 399 and *XPD* 751 variants are functional [174, 264-265]. Different

aspects of smoking behavior (smoking initiation, smoking cessation, intensity etc.) operate through multiple overlapping pathways [266] therefore would not be expected to be identically affected by DNA repair variation, which is borne out by the differing results for smoking status and smoking dose for several SNPs (*XRCC1* 399, *XRCC1* 280, *XPB* 751, *XRCC3* 241). Although speculative, there is some evidence that variation in DNA repair activity may affect neurological and/or respiratory outcomes, which could in turn affect smoking behavior [153-155, 157, 163]. If the variants are functional, or linked to functional variants, heterogeneity could be due to gene-environment interaction in specific populations or to differing linkage disequilibrium patterns across populations.

There are also several possible non-causal explanations for these findings. Although publication bias is always a concern with meta-analyses, the study goals of the contributing studies, visual inspection of funnel plots and formal tests of asymmetry argue against this. There could be similar selection bias in individual control groups, or strong selection bias in a subsample for studies leading to spurious results for *XRCC1* 399 dose estimates. This is possible, since just over half of the studies with dose information for *XRCC1* 399 were lung cancer studies (8 of 14) and lung cancer studies had on average higher OR_z s than other cancer studies for *XRCC1* 399, *XPB* 751 and *XRCC3* 241 PY analyses. Control groups from non-lung cancer studies, compared to lung cancer controls (referent), showed ROR (95%CI) of 0.8 (0.5, 1.2) and 0.5 (0.3, 0.9) for *XRCC1* 399 and *XPB* 751, respectively. For *XRCC3* 241 the OR_z (95%CI) were 0.8(0.5, 1.1) and 1.1(0.4, 2.7) for non-lung cancer controls and lung cancer controls, respectively. The connection between smoking and lung cancer is well known, possibly leading to more variation in response rates or recall by smoking history and/or family history of cancer, but the direction of possible bias is unpredictable. The OR_z

for the one *XRCC1* 399 study that explicitly excluded participants with smoking-related diseases was essentially the same as the summary estimate [243]. If heavier smokers with diseases/family histories related to DNA repair were overrepresented in hospital/patient-based controls and/or under represented in population-based control groups, OR_z would likely be biased upward for hospital-based studies and downward for population-based studies. This could conceivably cause estimates of OR_z for hospital-based studies to be higher than OR_z for population-based studies when the reverse is true in the underlying populations. In our analysis, for smoking status measures, average OR_z s tended to be above the null for population-based studies and below the null for hospital-based controls. For PY, the reverse was true, with hospital-based controls having average OR_z s around 1.3 and population-based controls slightly below the null. (Table V.A.4). Population stratification could have contributed to the heterogeneity in *XRCC1* 399 ever-never and *XPD* 751 PY estimates since the variant alleles are found at different frequencies in different ethnic groups within the same study, and smoking behavior may also differ by ethnicity. Although this cannot be rigorously addressed without individual level data, there were no clear patterns in OR_z for any SNP for study-level ethnicity, either by stated ethnicity, when stratified by single-ethnicity vs. multi-ethnicity studies, or when MAF was used as a crude proxy to assign ethnicity for studies with unknown ethnic makeup. Finally, chance could play a role, particularly given the large number of associations examined and sparse data for many analyses.

Implications for stand-alone case-only studies

Z is a measure of the magnitude of bias in the COR. If $Z=1$, the case-only estimate of interaction is not biased by genotype-environment association in the underlying population [3]. Commonly, this assumption is assessed in control data from a small number of outside

studies, using significance testing. Significance testing alone is not sufficient for assessment of potential bias [7]. Rarely is Z estimated and/or adjusted for, analogous to other forms of bias such as confounding. Results from this project illustrate some of the pitfalls of this approach. For instance, for *XRCC1* 399 ever-never smoking, 18 of the 21 included studies have estimates that are not statistically significantly different than the null; considering any or all of these in a testing framework would lead to the conclusion that the independence assumption was valid and a case-only study would give an unbiased estimate of multiplicative interaction. However, the range of OR_z s for these 18 studies is 0.7-1.6, many with wide CIs, indicating the potential for substantial bias. Similarly, although the summary OR_z s for PY are close to the null, the upper limits of the CIs were approximately 1.5 for SNPs in *XRCC1* and *XPB* and the lowest CI limit was 0.6; the range of potential bias is larger than obvious from examining only the magnitude of OR_z . Given that less than half of the studies that collect control genotype and smoking information present it in publications, and that very different conclusions can be drawn from different subsets of studies, this common approach seems inappropriate.

In the estimation framework, results from this project demonstrate the difficulty of using ancillary data to assess the independence assumption. Even when the Cochran's Q p-value is high, such as for *XRCC1* 399 current-not current smoking ($p=0.4$), point estimates of OR_z can vary as much as 5-fold [2.1(1.1, 3.9) for African Americans [227] to 0.4(0.1, 1.2) [267]]. Without further information that certain study characteristics might be influential, there is no good way to decide which of the available ancillary control groups might best represent the underlying (unmeasured) population for a proposed case-only study. Further, it is necessary to do a broad literature search to even to be aware of the possible values of OR_z

and range of bias in the COR. Additionally, since both summary estimates and individual study estimates vary across smoking categories, it is important that the independence assumption be evaluated for all smoking categories that will be used in the case-only analyses. For investigations of dose, it will be difficult for many SNPs to locate enough published control group data to even assess the possible range of the magnitude of bias.

This study has several strengths. Using a broad comprehensive search strategy in collaboration with information specialists increased power to detect and investigate heterogeneity between studies. Sample size was large for smoking status analyses and relatively large for *XRCC1* 399 and *XPB* 751 PY analyses. There was sufficient data for many studies to compare OR_z for smoking status and smoking dose both within studies, and by smoking category across multiple SNPs. The fact that none of the studies was conducted with the goal of assessing control group associations is both a strength and a weakness of this systematic review. Of the studies that collected the appropriate information only about 1/3 presented it in such a way that it could be abstracted for this meta-analysis, limiting sample size, especially for measures of smoking dose. However, publication bias was expected to be minimal since gene-environment interaction studies are typically not evaluated on the basis of control group associations. This was as supported by the formal tests of funnel plot asymmetry.

Only unadjusted odds ratios could be calculated so study estimates may be confounded. No individual level data was collected. This could be problematic for age and ethnicity in particular. Although some study characteristics could be determined accurately from articles, others were more likely to be misclassified. In particular, average age of study participants was difficult to determine. However, the fact that age was not a central study

feature for any of the studies makes it likely that misclassification is non-differential with respect to smoking and genotype. Several potentially informative study characteristics could not be examined because too few articles presented information that could be assessed using the same metric. In particular, response rates, which may vary by smoking behavior and family/personal history of cancer [117-120], and control group exclusion criteria, were presented using very different levels and types of detail from study to study. Only two of the 12 articles with multi-ethnic study populations presented data stratified by ethnicity, complicating interpretation of HWE p-value, ethnicity and MAF as study characteristics. Few studies presented enough control group information to examine multiple measures of smoking in the same study population, with the exception of studies that presented PY, since smoking status (ever-never) could nearly always be derived. Results did not change appreciably when studies without dose were excluded from ever-never analyses, indicating that articles that presented dose were not driving estimates of smoking status.

This systematic review of control-group associations between smoking and selected polymorphic genes commonly used in interaction studies was conducted to accomplish several objectives. The overarching goal was to enable investigators to make more effective use of ancillary data to evaluate the independence assumption prior to launching a stand-alone case-only study. Results from this study suggest that the independence assumption is frequently violated and caution is warranted before proceeding with any case-only interaction analysis. At a minimum, the independence assumption should be more rigorously evaluated than is often done. For a case-only analysis of a case-control study, separate OR_{2s} should be calculated for each anticipated COR in the relevant subgroup before proceeding. Evaluation of the independence assumption for a proposed stand-alone case-only study should include,

whenever possible, results from studies similar to the current study, relevant literature reviews, and a thorough search for individual studies with control or cohort data to ascertain at least the range of OR_{zs} , both overall and in relevant subgroups.

Evaluation of the independence assumption for case-only interaction studies would be greatly improved with more transparency and finer detail and in published articles, accomplished perhaps by expanding supplementary online tables to include selected joint genotype-smoking distributions in non-case groups. With the current emphasis on pooling controls, our results indicate that investigators should remain cautious about proceeding with case-only studies without further examination of the independence assumption in individual studies. If it could reliably be shown that $Z=1$ across individual studies, more use could be made of data pooling from control groups and cohorts for selected SNPs and exposures, especially where individual level data on potential confounders can be provided.

5. Tables and Figures

Table V.A.1a. Characteristics of 50 studies included in summary estimates (46 study populations) ¹

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses ²
Butkiewicz 2001	Poland (Upper Silesia)	Case-control, hospital-based	56.3 (8.8)y [Mean(SD)] 39-79y [Range]	Controls partially selected from healthy males from groups previously recruited for occupational studies. Rest of controls are 52 members of 4 families in Utah (CEPH reference families). <i>Matching:</i> Frequency-matched to case group on age, smoking habit and occupational exposure.	Lung cancer	<u>XPDAsp312Asn</u> Ever/never, PY
Cao 2006	Southern China	Case-control	45.7y (15.6y) [Mean (SD)]	Controls were "cancer-free" participants from a community cancer screening program. <i>Matching:</i> Matched to cases on age & ethnicity.	Nasopharyngeal cancer	<u>XRCCI Arg399Gln</u> Current sm/ not <u>XRCCI Arg194Trp</u> Current sm/ not
David-Beabes 2001	US (Los Angeles)	Case-control, population-based	62.9y (7.9y) [Mean (SD)]	Controls selected from Drivers License lists (<65y) or Medicare lists (>=65y). <i>Matching:</i> Frequency-matched to cases on age, gender & ethnicity.	Lung cancer	<u>XRCCI Arg399Gln</u> Ever/never, Intensity <u>XRCCI Arg194Trp</u> Ever/never, Intensity

Table V.A.1a. Studies included in meta-analyses (continued)

96

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses ²
Duell 2001 Carolina Breast Cancer Study (CBCS) (Same study population as Pachkowski 2006) ³	US (North Carolina)	Case- control, population- based	51.6y [Mean]	Controls were women selected from Drivers License (<65y) or Medicare (>=65y) lists. African American & younger (<50y) cases oversampled. <i>Matching:</i> Frequency matched to cases on age & ethnicity.	Breast cancer	<u>XRCC1</u> <u>Arg399Gln</u> Ever/never, Current sm/ not
Duell 2002 & Duell 2002 (parent study)	US (Northern California)	Case- control, population- based	24-54y (24%) 55-66y (26%) 67-73y (26%) 74-85y (23%) * [Frequency distribution]	Controls were identified by random digit dialing & Medicare lists (>=65y) & resided in any of 6 San Francisco Bay area counties . <i>Matching:</i> Frequency- matched to case group on age & gender.	Pancreatic cancer	<u>XRCC1</u> <u>Arg399Gln</u> ⁴ Duration
Garcia-Closas 2006	Spain	Case- control, hospital- based	66 (10)y [Mean(SD)] 21-80y [Range]	Controls were from participating hospitals with diagnoses unrelated to exposure(s) of interest (includes smoking). <i>Matching:</i> Individually matched to cases on age, gender, ethnicity and region.	Bladder cancer	<u>XPD Asp312Asn</u> Ever-never

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year	Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
	Han 2003 (Nurse's Health Study)	US	Case-control, nested	57.8 y [Mean]	Controls were a random selection from the subcohort of the Nurses Health Study that gave blood in 1989-90. No diagnosed cancer other than NMSC. <i>Matching:</i> Individually matched to cases on year of birth, menopausal status, HRT at blood collection, month of blood return, time of day of blood collection, fasting status at blood draw.	Breast cancer	<u>XRCCI Arg194Trp</u> Ever-never Duration
	Harms 2004	US (Texas)	Case-control, hospital-based	57.2 (9.3) [Mean (SD)]	Controls were current smokers who were non-case patients at the University of Texas Medical Branch in Galveston plus population from surrounding area. Meta-analysis included only non-Hispanic whites. <i>Matching:</i> Frequency-matched to case group on age, ethnicity and gender.	Lung cancer, Subset of controls: DNA damage (chromosomal aberrations)	<u>XRCCI Arg399Gln</u> PY <u>XPD Lys751Gln</u> PY <u>XRCC3 Thr241Met</u> PY
	Hoffmann 2005	Germany ²	Convenience sample	27.0 (5.7) [smokers] 26.3y (3.9y) [nonsmokers] [Mean (SD)]	Healthy male smokers & nonsmokers (1:1)	DNA damage (comet assay)	<u>XRCCI Arg399Gln</u> Current sm/ not <u>XPD Lys751Gln</u> Current sm/ not <u>XRCC3 Thr241Met</u> Current sm/not

Table 1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Hou 2002 (Same study population as Ryk2006b)	Sweden (Stockholm)	Case-control, hospital-based	68y [Median] 65y (14.5y) [Mean (SD)] 30-89y [Range]	Healthy controls were recruited from Stockholm residence files every 6 months. <i>Matching:</i> Frequency-matched to case group on age, gender, catchment area & smoking status (never/former/current). Never-smoking cases were over-sampled (50% of case group).	Lung cancer	<u><i>XPD</i> Lys751Gln</u> Ever/never <u><i>XPD</i> Asp312Asn</u> Ever/never
Huang 2005a	Poland (Warsaw)	Case-control, population-based	<50y (12%) 50-59y (17%) 60-69y (39%) >=70y (32%) [Frequency distribution]	Controls randomly chosen from Warsaw population registry. <i>Matching:</i> Frequency-matched to case group on age & gender	Gastric cancer	<u><i>XRCC1</i> Arg399Gln</u> Ever/never PY <u><i>XPD</i> Lys751Gln</u> Ever/never PY <u><i>XRCC3</i> Thr241Met</u> Ever/never PY
Hung 2005	Eastern Europe	Case-control, hospital-based	<=40y (3%) 41-50y (15%) 51-60y (31%) 61-70y (36%) >70y (16%) [Frequency distribution]	Controls were patients in same hospitals as cases (15 centers in 6 Eastern European countries) except controls from Warsaw who were randomly sampled from the population register. Patients with tobacco-related diseases were excluded from control group. <i>Matching:</i> Frequency-matched to case group on age, gender, center & referral area.	Lung cancer	<u><i>XRCC1</i> Arg399Gln</u> Ever/never PY <u><i>XRCC1</i> Arg194Trp</u> Ever/never PY <u><i>XRCC1</i> Arg280His</u> Ever/never PY

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Ito 2004	Japan	Case-control, hospital-based	62.6y (9.1y) [Mean (SD)] 35-79y [Range]	Controls were random sample of cancer-free visitors to Aichi Cancer Center Hospital who provided blood. <i>Matching:</i> Frequency-matched to case group on age & gender.	Lung cancer	<u>XRCC1 Arg399Gln</u> Ever/never Current sm/not PY
Jiao 2007	US (Texas)	Case-control, hospital-based (friends & family)	<50y (15%) 50-59y (27%) 60-69y (34%) ≥70y (23%) [Frequency distribution]	Controls were friends & non-genetically related family of non-pancreatic cancer patients. <i>Matching:</i> Frequency-matched to case group on age, gender & ethnicity	Pancreatic cancer	<u>XPD Lys751Gln</u> Ever/never <u>XPD Asp312Asn</u> Ever/never
Jin 2005	China (Zhejiang)	Case-control, nested	62.2 (10.3)y [Mean (SD)] 40-49y (14%) 50-59y (32%) 60-69y (28%) 70+y (26%) [Frequency distribution]	Controls were randomly chosen from colorectal cancer screening trial with individually matched communities. No previously diagnosed malignancy. <i>Matching:</i> Frequency matched to case group on age, gender and habitation.	Colorectal cancer	<u>XRCC3 Thr241Met</u> Current sm/not
Justenhoven 2004 (GENICA)	Germany (Bonn)	Case-control, population-based	<50 y (23%) ≥50 y (77%) [Frequency distribution]	Controls were population-based from Interdisciplinary Study Group on Gene Environment Interactions and Breast Cancer in Germany (GENICA). <i>Matching:</i> Individually matched to cases on age	Breast cancer	<u>XPD Asp312Asn</u> Ever/never

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses <input type="checkbox"/>
Kelsey 2004	US (New Hampshire)	Case-control, population-based	62 (10) [Mean (SD)]	Controls randomly selected from driver's license (<65 y) or Medicare (>=65y) records. Some controls were shared with non-melanoma skin cancer study. <i>Matching:</i> Frequency-matched to cases on age & gender	Bladder cancer	<u><i>XRCCI</i> Arg399Gln</u> Ever/never
Kocabas 2006	Turkey	Convenience sample	26-78y [Range]	Healthy volunteers	Genotype	<u><i>XRCCI</i> Arg399Gln</u> Current sm/not
Koyama 2005	Japan	Case-control	23.6y (4.7y) [Mean (SD)]	Controls were healthy individuals with no autoimmune disease <i>Matching:</i> Matched on ethnicity	Rheumatoid arthritis	<u><i>XRCCI</i> Arg399Gln</u> Current sm/not <u><i>XRCCI</i> Arg194Trp</u> Current sm/ not <u><i>XRCCI</i> Arg280His</u> Current sm/ not
Lei 2002	Taiwan ⁵	Cohort, occupational	32.4y (5.2y)[Mean (SD)]	Controls were a subcohort of male resin synthesis plant workers unexposed to epichlorohydrin >1 ppm.	DNA damage(Sister chromatid exchange)	<u><i>XRCCI</i> Arg399Gln</u> Current sm/not Intensity
Lunn 1999	US (North Carolina)	Convenience	not given	Controls were participants in a community-based study of African Americans and whites who were heterozygous for the glycoporphin A antigen. Additional white & African Americans from the same community sample were added for genotype frequency estimation only.	DNA damage (Glycophorin A somatic mutations)	<u><i>XRCCI</i> Arg399Gln</u> Current sm/not <u><i>XRCCI</i> Arg194Trp</u> Current sm/ not <u><i>XRCCI</i> Arg280His</u> Current sm/ not

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses <input type="checkbox"/>
Matullo 2001 [European Prospective Investigation into Cancer & Nutrition in Italy (EPIC-Italy)] Palli 2000 (parent study)	Italy	Cohort, prospective	49.8y [Mean] <=44y (33%) 45-54y (33%) >54y (34%) [Frequency distribution]	Controls were a random selection of EPIC participants from Northern Italy (Varese and Turin), Central Italy (Florence) and Southern Italy (Ragusa and Naples). Recruitment criteria varied by site and included blood donors, women being screened for breast cancer, population-based recruitment etc.	DNA damage (DNA adducts)	<u>XRCC1 Arg399Gln</u> Ever/never Current sm/not <u>XPD Lys751Gln</u> Ever/never Current sm/not <u>XRCC3 Thr241Met</u> Ever/never Current sm/not
Matullo 2005	Italy (Turin)	Case- control, hospital- based	34-76y [Range]	Controls were a random selection of male patients at 2 urology clinics (benign diseases only) and at medical and surgical clinics. Patients with cancer, liver or renal diseases or smoking-related conditions were excluded. <i>Matching: none</i>	Bladder cancer	<u>XRCC1 Arg399Gln</u> Ever/never Current sm/not <u>XPD Lys751Gln</u> Ever/never Current sm/not <u>XRCC1 Arg194Trp</u> Ever/never: Current sm/not <u>XPD Asp312Asn</u> Ever/never Current sm/not <u>XRCC3 Thr241Met</u> Ever/never Current sm/not

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Metsola 2005	Finland	Case-control, hospital-based	53.5y [Mean] 37-77y [Range]	Controls resided in same area as cases that attended Kuopio University Hospital and were randomly selected from the Finnish National Population Register. <i>Matching:</i> none	Breast cancer	<u>XRCCI Arg399Gln</u> Ever/never PY <u>XPD Lys751Gln</u> Ever/never PY <u>XRCCI Arg280His</u> Ever/never PY
Misra 2003 [Alpha-tocopherol Beta-carotene Cancer Prevention Study (ATBC Finland)]	Finland	Case-control, nested	59y [Median] 55-63y [Range]	Controls were from a case-control study nested in the ATBC Trial cohort. All were male smokers from southwestern Finland. Intervention group received alpha-tocopherol &/or beta-carotene supplements. Sampling from cohort based on incidence density sampling & availability of blood samples. <i>Matching:</i> Individually-matched on age, intervention group, study clinic & date of blood draw.	Lung cancer	<u>XRCCI Arg399Gln</u> Intensity
Olshan 2002 Olshan 2000 (parent study)	US (North Carolina)	Case-control, hospital-based	20-49y (27%) 50-59y (22%) 60-69y (33%) ≥70y (17%) [Frequency distribution]	Controls were surgical patients attending the same clinic as cases. Controls with aspirin triad were excluded. Meta-analysis included only white controls. <i>Matching:</i> Frequency-matched to case group on age & gender.	Head & neck cancer	<u>XRCCI Arg399Gln</u> Ever/never <u>XRCCI Arg194Trp</u> Ever/never

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Pachkowski 2006 ¹ Carolina Breast Cancer Study (CBCS)	US (North Carolina)	Case-control, population- based	<=45y (30%) >45y (70%) [Frequency distribution]	Controls were women selected from Drivers License (<65y) or Medicare (>=65y) lists. African American & younger (<50y) cases oversampled. <i>Matching:</i> Frequency matched to cases on age & ethnicity.	Breast cancer	<u>XRCCI Arg399Gln</u> Ever/never Current sm/not Intensity Duration <u>XRCCI Arg194Trp</u> Ever/never Current sm/not Duration Intensity <u>XRCCI Arg280His</u> Ever/never Current sm/not Duration Intensity
Park 2002	Korea	Case-control, hospital-based	60.7y (8.9y) [Mean (SD)] 38-86y [Range]	Controls were randomly selected from healthy male volunteers at a hospital check-up clinic. <i>Matching:</i> Frequency matched to case group on age.	Lung cancer	<u>XRCCI Arg399Gln</u> PY
Patel 2005 [Cancer Prevention Study II (CPS- II) Nutrition cohort]	US	Case-control, nested	(Combined cases & controls) 62y [Median] 43-75y [Range]	Controls were women from the CPS-II Nutrition Cohort, a subgroup of the CPS-II baseline mortality cohort. Controls were randomly selected cancer-free participants meeting case-matching criteria. <i>Matching:</i> Individually-matched to cases on age, ethnicity & date of blood collection.	Breast cancer	<u>XRCCI Arg399Gln</u> Ever/never <u>XRCCI Arg194Trp</u> Ever/never

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Ramachandran 2006	India	Case- control, hospital- based (friends & family)	not given	Controls were visitors & family to Head and Neck Clinic of the Thiruvananthapuram Regional cancer center. <i>Matching:</i> Frequency-matched to cases on age, gender & "habits" ⁶	Head & neck cancer	<u>XRCCI Arg399Gln</u> Ever/never
Ryk 2006 (Same study population as Hou 2002)	Sweden(S tockholm)	Case- control, population- based	68y [Median] 30-89y [Range]	Healthy controls were recruited from Stockholm residence files every 6 months. <i>Matching:</i> Frequency-matched to case group on age, gender, catchment area & smoking status (never/former/current). Never-smoking cases were over-sampled (50% of case group).	Lung cancer	<u>XRCCI Arg399Gln</u> Ever/never
Schabath 2005 (Study population may overlap Shen 2002)	US (Texas)	Case- control, hospital- based	62y [Mean]	Healthy control subjects were recruited from Kelsey-Seybold Clinics (a large private multispecialty physicians group, Houston). <i>Matching:</i> Frequency-matched to case group on age, gender & ethnicity	Bladder cancer	<u>XPD Lys751Gln</u> Ever/never PY <u>XPD Asp312Asn</u> Ever/never PY
Schneider 2005	Germany	Case- control, hospital- based		Controls were from outpatient clinics free of any benign or malignant tumors & unrelated to cases <i>Matching:</i> none	Lung cancer	<u>XRCCI Arg399Gln</u> Ever/never PY <u>XRCCI Arg194Trp</u> Ever/never PY <u>XRCCI Arg280His</u> Ever-never PY

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Shen 2000	China (Jiangsu Province)	Case- control, population- based	61.6y [Mean] 62y [Median] 32-67y [Range]	Controls were selected from closest unrelated neighbors in same village as cases. Controls were healthy & cancer-free. <i>Matching</i> : Individually-matched to cases on age, gender & village.	Gastric cancer	<u>XRCC1 Arg399Gln</u> Ever/never <u>XRCC1 Arg194Trp</u> Ever/never
Shen 2002 (Study population may overlap Schabath 2005)	US (Texas)	Case- control, hospital- based	56y [Median] 55.8y [Mean] 19-84y [Range]	Controls were from a local managed care organization (Houston, Kelsey-Seybold). "Cancer-free". Meta-analysis included only non-Hispanic whites. <i>Matching</i> : Frequency matched to case group on age, gender, ethnicity, smoking status and alcohol consumption.	Head & neck cancer	<u>XRCC3 Thr241Met</u> Ever/never Current sm/not
Shen 2003	Italy	Case- control, hospital- based	(Cases & controls) 63y [Mean] (Controls only) <=40y (5%) 41-50y (8%) 51-60y (23%) 61-70y (38%) >70y (26%) [Freq distribution]	Controls were male patients from urology depts of 2 main hospitals in Brescia Italy with non-neoplastic diseases. <i>Matching</i> : Frequency-matched to case group on age, period of recruitment & hospital	Bladder cancer	<u>XRCC1 Arg399Gln</u> Ever/never PY <u>XPD Lys751Gln</u> Ever/never PY <u>XRCC3 Thr241Met</u> Ever/never PY
Shen 2005 [Same study population as Terry 2004, Long Island Breast Cancer Study Project (LIBCSP)]	US (New York)	Case- control, population- based	<35y (3%) 35-44y (16%) 45-54y (27%) 55-64y (26%) 65-74y (20%) 75-84y (7%) >=85y (1%) [Frequency distribution]	Controls were women identified by random digit dialing (<65y) & Medicare records (>=65 y) residing in Long Island NY in Nassau and Suffolk counties. <i>Matching</i> : Frequency-matched to case group by age & ethnicity	Breast cancer	<u>XRCC1 Arg399Gln</u> Ever/never <u>XRCC1 Arg194Trp</u> Ever/never

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Skjelbred 2006a [European Study Group on Cytogenetic Biomarkers and Health (ESCH)]	Norway	Cohort, occupational	18-71y [Range]	Controls were males from ESCH cohort (European Study Group on Cytogenetic Biomarkers and Health; combined Nordic & Italian cohorts) + additional males. This study includes only the Norwegian Caucasian males in the Cancer Risk Biomarker group. Sample enriched for occupational exposures likely to cause chromosomal aberrations (~47% w possible occupational exposure to clastogenic/carcinogenic agents).	DNA damage (chromosomal aberrations)	<u>XRCC1 Arg399Gln</u> Current sm/not <u>XPD Lys751Gln</u> Current sm/not <u>XRCC1 Arg194Trp</u> Current sm/not <u>XRCC1 Arg280His</u> Current sm/not <u>XRCC3 Thr241Met</u> Current sm/not
Smedby 2006 (Scandinavian Lymphoma Etiology Study)	Sweden & Denmark	Case-control, population-based	59y [Mean] 19-74y [Range]	Controls randomly selected from population registries of Denmark & Sweden + regional pilot in Denmark. No hematologic malignancies. <i>Matching:</i> Frequency-matched to case group on age & gender.	Lymphoma	<u>XRCC3 Thr241Met</u> Ever/never Current sm/not
Stern 2001 [Same study population as Stern 2002a, 2002b(excluded)]	US (North Carolina)	Case-control, hospital-based	63.3 (10.4) y [Mean(SD)] <60 y (31.5%) 60-70 y (41.8%) >70y (26.8%) [Frequency distribution]	Controls were male urology clinic patients w no history of any cancer other than NMSC <i>Matching:</i> Frequency matched to case group on age, ethnicity, and gender	Bladder cancer	<u>XRCC1 Arg194Trp</u> Ever/never Duration
Stern 2002a [Same study population as Stern 2001, 2002b(excluded)]	US (North Carolina)	Case-control, hospital-based	63.3 (10.4) y [Mean(SD)] <60 y (31.5%) 60-70 y (41.8%) >70y (26.8%) [Frequency distribution]	Controls were male urology clinic patients w no history of any cancer other than NMSC <i>Matching:</i> Frequency matched to case group on age, ethnicity, and gender	Bladder cancer	<u>XRCC3 Thr241Met</u> Ever/never PY

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Terry 2004	US (New York)	Case-control, population-based	<35y (3%) 35-44y (16%) 45-54y (27%) 55-64y (26%) 65-74y (20%) 75-84y (7%) >=85y (1%) [Frequency distribution]	Controls were women identified by random digit dialing (<65y) & Medicare records (>=65 y) residing in Long Island NY in Nassau and Suffolk counties. <i>Matching:</i> Frequency-matched to case group by age & ethnicity	Breast cancer	<u><i>XPD</i> Lys751Gln</u> Ever/never Current sm/not
[Same study population as Shen 2005a, Long Island Breast Cancer Study Project (LIBCSP)] Tuimala 2004	Finland & Hungary	Convenience sample	41.0y [Mean(SD)] 21-64y [Range]	Controls were from 2 parent case-control studies: 1. Finnish office workers from case-control study of isocyanate asthma 2. Hungarian healthy blood donors and clerks attending pre-employment physicals from case-control study of head and neck cancer (nonsmoking drinkers excluded). Control groups were pooled for analysis [Finns(N=61) + Hungarians(N=84)].	DNA damage(Chromosomal abberations &Sister chromatid exchange)	<u><i>XRCC1</i> Arg399Gln</u> Ever/never <u><i>XRCC1</i> Arg280His</u> Ever/never <u><i>XRCC3</i> Thr241Met</u> Ever/never
Wilding 2005	UK	Cohort, occupational	~69y [Mean] All >50y	Cohort consisted of retired male workers from nuclear facility.	DNA damage (chromosomal aberrations)	<u><i>XRCC1</i> Arg399Gln</u> Ever/never <u><i>XRCC1</i> Arg194Trp</u> Ever/never <u><i>XRCC3</i> Thr241Met</u> Ever/never
Xing 2002	China (Beijing)	Case-control, hospital-based	58y (6.8) [Mean (SD)]	Population controls randomly selected from a nutritional survey having participants from the same region as case patients. <i>Matching:</i> Frequency-matched to case group on age & gender.	Lung cancer	<u><i>XPD</i> Lys751Gln</u> Ever/never PY <u><i>XPD</i> Asp312Asn</u> Ever/never PY

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
Yu 2004a (Same population as Yu 2004b)	China	Case- control, hospital- based	57.4y (9.4y) [Mean(SD)] <60y (66%) >60 y (34%) [Frequency distribution]	Controls were randomly selected from a pool of volunteers who visited the general health check-up division at Tongji Hospital of Huazhong University of Science & Technology clinics. <i>Matching:</i> Frequency-matched to case group on age & gender	Esophageal cancer	<u>XRCCI Arg399Gln</u> Ever/never
Yu 2004b (Same population as Yu 2004a)	China	Case- control, hospital- based	57.4y (9.4y) [Mean(SD)] <60y (66%) >60 y (34%) [Frequency distribution]	Controls were randomly selected from a pool of volunteers who visited the general health check-up division at Tongji Hospital of Huazhong University of Science & Technology clinics. <i>Matching:</i> Frequency-matched to case group on age & gender	Esophageal cancer	<u>XPD Lys751Gln</u> Ever/never
Zhou 2002 (Same study population as Zhou 2003)	US (Massachus- etts)	Case- control, hospital- based (friends & family)	58.5y (12.4y) [Mean (SD)] 19-100y [Range] <55y (38%) 55-64y (26%) ≥65 y (36%) [Frequency distribution]	Controls were friends & non-genetically related family of lung cancer cases at Massachusetts General Hospital. Also friends & family of non-lung cancer pts in cardiothoracic wards (<10%). <i>Matching:</i> none	Lung cancer	<u>XPD Lys751Gln</u> Ever/never PY <u>XPD Asp312Asn</u> Ever/never PY

Table V.A.1a. Studies included in meta-analyses (continued)

Author & year	Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment	Study outcome	Analyses □
	Zhou 2003 (Same study population as Zhou 2002)	US (Massachusetts)	Case-control, hospital-based (friends & family)	58.5y (12.4y) [Mean (SD)] 19-100y [Range] <55y (38%) 55-64y (26%) >=65 y (36%) [Frequency distribution]	Controls were friends & non-genetically related family of lung cancer cases at Massachusetts General Hospital. Also friends & family of non-lung cancer pts in cardiothoracic wards (<10%). <i>Matching: none</i>	Lung cancer	<u><i>XRCC1</i> Arg399Gln</u> Ever/never PY
	Zijno 2006	Italy(Rome)	Cohort, occupational	43.4 y [Mean]	Traffic wardens [N~133] & office workers [N~57]) in the municipality of Rome. Study of urban air pollutants and genotoxic endpoints. <i>Matching: none</i>	DNA damage (Sister chromatid exchange)	<u><i>XPD</i> Lys751Gln</u> Current sm/not

Table V.A.1b. Characteristics of 5 additional studies (5 study populations) not included in any summary estimates.

Author & year	Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment & reason for exclusion	Study outcome	Analyses [□]
	Affatato 2004	US (Texas)	Convenience sample	48y [Median]	Volunteers from University of Texas Medical Branch in Galveston answering notices for smokers & nonsmokers for genetic studies. Excluded cancer history and exposure to potential mutagens (radiation etc.). <i>Matching:</i> none <i>Exclusion:</i> Included only never smokers and current smokers. Former smokers excluded.	DNA damage (chromosomal aberrations)	<u><i>XPDLys751Gln</i></u> Never/ current smoker only: 1.05 (0.39, 2.88) <u><i>XPDLys312Asn</i></u> Never/ current smoker only: 1.31 (0.47, 3.62)
	Figueiredo 2004	Canada (Ontario)	Case-control, population-based	45.2 (6.5) y [Mean (SD)]	Controls were women identified by random digit dialing. No history of breast cancer. <i>Matching:</i> Frequency-matched to case group on age. <i>Exclusion:</i> No data presented on adult smoking.	Breast cancer	<u><i>XRCC1 Arg399Gln</i></u> Adolescent smoking yes/no: 1.02 (0.58, 1.79) <u><i>XRCC3 Thr241Met</i></u> Adolescent smoking yes/no: 1.42 (0.80, 2.50)
	Stern 2002b (Same study population as Stern 2001 & 2002a)	US (North Carolina)	Case-control, hospital-based	63.4 (10.3) [Mean (SD)] <=60 y (31.1%) 60-70 y (42.1%) >=70 y (26.8%) [Frequency distribution]	Controls were males from urology clinics without a history of cancer except NMSC. <i>Matching:</i> Frequency matched to case group on age, gender and ethnicity. Only blacks and whites included in analysis. <i>Exclusion:</i> Data presented for recessive model only.	Bladder cancer	<u><i>XPDLys751Gln</i></u> Ever/never (Any Lys v. Gln): 1.76 (0.75, 4.16) Duration (Any Lys v. Gln): 0.91 (0.34, 2.41)
	Stern 2006a	US (California)	Case-control	61 (7) y [Mean (SD)]	Controls were from screening study for colorectal adenomas. No history of invasive cancer, past polyps. <i>Matching:</i> Individually matched to cases on age, gender, date and center of procedure. <i>Exclusion:</i> Data presented for recessive model only.	Colorectal adenoma	<u><i>XPDLys751Gln</i></u> Ever/never (Any Lys v. Gln): 0.72 (0.45, 1.15)

Table V.A.1b. Characteristics of 5 additional studies (5 study populations) but not included in any summary estimates (continued)

Author & year	Study name (abbreviation)	Location	Design	Age in controls [metric]	Control ascertainment & reason for exclusion	Study outcome	Analyses [□]
	Wang 2003b	US (Texas)	Case-control, hospital-based	~ 62 y [Mean]	Controls were from community centers. No previous cancer history except NMSC. African-American and Mexican-American controls only. <i>Matching:</i> Frequency matched to case group on age, gender, ethnicity & city of residence. <i>Exclusion:</i> Never smokers included in lowest PY category	Lung cancer	<u>XRCC3 Thr241Met</u> PY (Never+low PY v. hi PY): 0.21 (0.07, 0.64)

Abbreviations: PY = pack-years, NMSC = non-melanoma skin cancer, y=years of age. Sm=smoker, not=not a current smoker

¹ Fifty-five studies met overall inclusion criteria. 50 studies, representing 46 study populations, could be included in at least 1 genotype-smoking summary estimate. Five studies met inclusion criteria but could not be included in any genotype-smoking summary analyses.

² Reference groups for smoking analyses: never smoker, not a current smoker (never+former smoking), lowest non-zero PY category (v. highest), shortest non-zero smoking duration category (v. highest) & lowest non-zero category of smoking intensity (v. highest). All genotype contrasts: homozygous for the more common allele (ref) v. 1 or more copies of the less common variant

³ Duell 2001 & Pachkowski 2006 are from the CBCS. Duell 2001 was used for analyses stratified on ethnicity, Pachkowski 2006 used for all others.

⁴ Excluded from Ever/never and Current sm/not because "never smoking" was not comparable to other studies Never/former/current categories were 1. never active or passive and 2. passive+cigar/pipe smoking)

⁵ Assumed because of lead author's institutional affiliation

⁶ Appears to include to smoking behavior, alcohol consumption and/or betel quid chewing.

Table V.A.2: DNA repair gene variation and smoking summary estimates

	Ever-never			Current-Not current			Pack-years			Intensity			Duration		
	N	Q p- value	OR _z (95% CI) ¹	N	Q p- value	OR _z (95% CI) ²	N	Q p-value	OR _z (95% CI) ³	N	Q p- value	OR _z (95% CI) ⁴	N	Q p- value	OR _z (95% CI) ⁵
Gene and SNP															
<i>XRCC1</i>															
Arg399						1.0			1.2			1.5			
Gln ⁶	21	0.01	--- ¹²	11	0.40	(0.9, 1.1)	9	0.30	(1.0, 1.5)	4	0.49	(1.2, 1.9)	2	0.03	--- ¹²
Arg194			1.0			1.1			1.1			1.1			0.7
Trp ⁷	12	0.62	(0.9, 1.1)	6	0.68	(0.9, 1.3)	2	0.73	(0.7, 1.6)	2	0.89	(0.8, 1.6)	3	0.47	(0.5, 0.9)
Arg280			1.0			0.7			1.0			0.9			1.2
His ⁸	5	0.47	(0.8, 1.2)	4	0.51	(0.5, 1.1)	3	0.32	(0.6, 1.5)	1	--	(0.5, 1.8)	1	--	(0.6, 2.3)
<i>XPB</i>															
Lys751			0.9			1.1			---						
Gln ⁹	12	0.46	(0.8, 1.1)	6	0.25	(0.9, 1.3)	7	0.02	---	0			0		
Asp312			1.1			1.1			1.1						
Asn ¹⁰	9	0.79	(1.0, 1.2)	1	---	(0.7, 1.9)	4	0.11	(0.8, 1.5)	0			0		
<i>XRCC3</i>															
Thr241			1.0			0.9 (0.8,			0.8						
Met ¹¹	9	0.52	(0.9, 1.2)	7	0.73	1.1)	4	0.67	(0.6, 1.2)	0			0		

Abbreviations: CI=Confidence interval, na=not applicable, PY=pack-years, OR_z=control-only genotype-smoking odds ratio, Q=Cochran's test of heterogeneity, N=number of studies, G+=gene is positive for any variant allele G-=negative for variant allele (referent), E+=positive for smoking measures, E-=negative for smoking measure (referent), N=number of studies, SNP=single nucleotide polymorphism, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ Fixed effects summary estimates for G+/G- vs. E+/E- . G- is homozygous for the common allele (ref), G+ is the genotype with 1 or 2 variant alleles, E- (ref) is never smoker and E+ is ever smoking

² Fixed effects summary estimates for G+/G- vs. E+/E- . G- is homozygous for the common allele (ref), G+ is the genotype with 1 or 2 variant alleles, E- (ref) is not current smoker and E+ is current smoking

³ Fixed effects summary estimates for G+/G- vs. E+/E- . G- is homozygous for the common allele (ref), G+ is the genotype with 1 or 2 variant alleles, E- (ref) is lowest non-zero PY category and E+ is highest category of PY

⁴ Fixed effects summary estimates for G+/G- vs. E+/E- . G- is homozygous for the common allele (ref), G+ is the genotype with 1 or 2 variant alleles, E- (ref) is lowest non-zero category of intensity (cig/day) and E+ is highest category of smoking intensity.

⁵ Fixed effects summary estimates for G+/G- vs. E+/E- . G- is homozygous for the common allele (ref), G+ is the genotype with 1 or 2 variant alleles, E- (ref) is lowest non-zero category of duration(yrs) and E+ is highest category of smoking duration.

Table V.A.2: DNA repair gene variation and smoking summary estimates (continued)

⁶ Arg/Arg v. any Gln

⁷ Arg/Arg v. any Trp

⁸ Arg/Arg v. any His

⁹ Lys/lys v. any Gln

¹⁰ Asp/Asp v. any Asn

¹¹ Thr/Thr v. any Met

¹² Studies too heterogeneous for fixed effects summary estimate

Table V.A.3. Genotype-smoking associations for selected specification of current-not current smoking and pack-years of smoking

	XRCC1									XPD						XRCC3		
	Arg399Gln			Arg194Trp			Arg280His			Lys751Gln			Asp312Asn			Thr241Met		
	N	OR _z	(95% CI)	N	OR _z	(95% CI)	N	OR _z	(95% CI)	N	OR _z	(95% CI)	N	OR _z	(95% CI)	N	OR _z	(95% CI)
Current / not current smoking																		
Not current (never + former)(ref) vs. current smoker	11	1.0	0.9, 1.1	6	1.1	0.9, 1.3	4	0.7	0.5, 1.1	6	1.1	0.9, 1.3	1	1.1	0.7, 1.9	7	0.9	0.8, 1.1
Never (ref) vs. former smoker	4	1.1	0.9, 1.3	2	1.1	0.8, 1.4	1	1.0	0.7, 1.5	3	0.8	0.6, 1.0	1	1.3	0.7, 2.4	4	1.0	0.8, 1.3
Never (ref) vs. current smoker	4	1.1	0.9, 1.4	2	1.2	0.8, 1.6	1	0.9	0.5, 1.4	3	0.9	0.7, 1.1	1	1.3	0.7, 2.3	4	0.9	0.7, 1.2
Pack-years of smoking																		
Relative PY ¹																		
Lightest (ref) vs. heaviest smokers ²	9	1.2	1.0, 1.5	2	1.1	0.7, 1.6	3	1.0	0.6, 1.5	7	1.2	1.0, 1.5	4	1.1	0.8, 1.5	4	0.8	0.6, 1.2
Never (0 PY)(ref) vs. lightest smokers ³	7	-- ⁶		2	0.9	0.7, 1.2	3	1.1	0.8, 1.5	6	0.9	0.8, 1.1	4	1.0	0.8, 1.3	3	1.3	0.9, 1.8
Never (0 PY)(ref) vs. heaviest smokers	7	1.2	1.0, 1.4	2	1.1	0.8, 1.6	3	1.1	0.7, 1.5	6	1.0	0.8, 1.3	4	1.1	0.8, 1.4	3	1.0	0.7, 1.4
Absolute PY ⁴																		
Light (<cutpoint range, ref) vs. heavy smokers (>=cutpoint range) ⁵	6	1.2	1.0, 1.5	2	0.9	0.7, 1.3	2	1.2	0.8, 1.7	4	1.1	0.9, 1.3	3	1.0	0.8, 1.2	3	0.8	0.5, 1.1
Never (0 PY) (ref) vs. (<cutpoint range, ref)	4	1.1	0.9, 1.2	2	0.8	0.7, 1.1	2	1.0	0.8, 1.4	4	0.9	0.7, 1.1	3	1.0	0.8, 1.2	2	1.4	1.0, 2.0
Never (0 PY) (ref) vs. (>=cutpoint range)	4	1.3	1.0, 1.6	2	1.1	0.8, 1.5	2	1.3	0.8, 1.9	4	0.9	0.7, 1.1	3	1.0	0.8, 1.2	2	1.0	0.6, 1.6

Table V.A.3. Genotype-smoking associations for selected specifications of current-not current smoking and pack-years of smoking (continued)

Abbreviations: ref=referent, PY=pack-year, OR_z =control-only genotype-smoking odds ratio, N=number of studies, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ Lowest and highest study-defined PY categories, regardless of PY cutpoints.

² Lowest study-defined non-zero PY category (reference) compared to highest study-defined PY category, regardless of PY cutpoints (i.e. smokers only).

³ Never smokers (0PY) as common referent for lowest study-defined PY category and highest study-defined PY category, regardless of PY cutpoints.

⁴ Light and heavy smokers are defined as smoking less or more, respectively, than the study-defined cutpoint within the cutpoint range for the specified SNP; Studies w no cutpoint in this range are excluded, no studies included multiple cutpoints in this range; Ranges for absolute PY: *XRCC1* Arg399Gln (32-42PY)

⁵ Light includes all study-defined categories w a lower bound less than the cutpoint range for that SNP (excluding 0 PY); Heavy includes all study-defined categories w an upper bound greater than the cutpoint range for that SNP

⁶ Studies too heterogeneous for fixed effects summary estimate

Table V.A.4: Genotype-smoking associations stratified by study design

<i>XRCCI</i>									
Arg399Gln			Arg194Trp			Arg280His			
N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	
Ever-never smoking									
Case-based categories									
Case-control									
Population									
- based	8	1.1 (0.9, 1.5)	Ref	4	1.0 (0.9, 1.3)	Ref	2	0.9 (0.7, 1.3)	Ref
Hospital-									
based	9	1.0 (0.9, 1.2)	0.9 (0.7, 1.2)	5	0.9 (0.8, 1.1)	0.9 (0.7, 1.2)	2	1.1 (0.8, 1.4)	1.2 (0.8, 1.8)
Other	1	0.9 (0.6, 1.3)	0.8 (0.4, 1.3)	2	1.1 (0.9, 1.5)	1.1 (0.8, 1.5)	0		
Cohort/ convenience	3	1.0 (0.8, 1.4)	0.9 (0.6, 1.4)	1	0.7 (0.3, 1.4)	0.7 (0.3, 1.4)	1	0.5 (0.2, 1.1)	0.5 (0.2, 1.3)
Case-control	18	1.1 (0.9, 1.2)	Ref	11	1.0 (0.9, 1.2)	NA	4	1.0 (0.8, 1.2)	NA
Other	3	1.0 (0.8, 1.4)	1.0 (0.7, 1.4)	1	0.7 (0.3, 1.4)		1	0.4 (0.2, 1.1)	
Non-case-based categories									
Case-control									
Population									
controls	8	1.1 (0.9, 1.5)	Ref	4	1.0 (0.9, 1.3)	Ref	2	0.9 (0.7, 1.3)	Ref
Patient									
controls ³	6	1.1 (0.9, 1.2)	0.9 (0.7, 1.3)	5	0.9 (0.8, 1.1)	0.9 (0.7, 1.2)	2	1.1 (0.8, 1.4)	1.2 (0.8, 1.8)
Non-patient									
controls ⁴	4	0.9 (0.8, 1.1)	0.8 (0.6, 1.1)	2	1.1 (0.9, 1.5)	1.1 (0.8, 1.5)	0		
Cohort/ convenience	3	1.0 (0.8, 1.4)	0.9 (0.6, 1.4)	1	0.7 (0.3, 1.4)	0.7 (0.3, 1.4)	1	0.5 (0.2, 1.1)	0.5 (0.2, 1.3)
Health status of non-cases									
Not patients ⁵	15	1.1 (0.9, 1.3)	Ref	7	1.1 (0.9, 1.2)	Ref	3	0.9 (0.7, 1.1)	Ref
Patients	6	1.1 (0.9, 1.2)	1.0 (0.8, 1.3)	5	0.9 (0.8, 1.1)	0.9 (0.7, 1.1)	2	1.1 (0.8, 1.4)	1.3 (0.8, 1.9)
Unknown									
patient status	0			0			0		

Table V.A.4: Genotype-smoking associations stratified by study design (continued)

XRCCI									
	Arg399Gln			Arg194Trp			Arg280His		
	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²
	Current-not current smoking								
Case-based categories									
Case-control									
Population-based	1	1.2 (1.0, 1.5)	Ref	1	1.1 (0.8, 1.6)	NA	1	0.9 (0.6, 1.4)	NA
Hospital-based	3	0.9 (0.7, 1.2)	0.7 (0.5, 1.0)	2	0.9 (0.7, 1.3)		0		
Other	0			0			0		
Unknown	1	1.6 (0.7, 3.7)	1.3 (0.5, 3.2)	1	0.9 (0.4, 2.2)		1	1.1 (0.2, 5.8)	
Cohort/convenience	6	0.8 (0.6, 1.1)	0.7 (0.5, 1.0)	2	1.5 (0.8, 2.7)		2	0.5 (0.2, 1.0)	
Case-control	5	1.1 (0.9, 1.3)	Ref	4	1.0 (0.8, 1.3)	Ref	2	0.9 (0.6, 1.4)	Ref
Other	6	0.8 (0.6, 1.1)	0.8 (0.6, 1.1)	2	1.5 (0.8, 2.7)	1.5 (0.8, 2.7)	2	0.5 (0.2, 1.0)	0.5 (0.2, 1.2)
Non-case-based categories									
Case-control									
Population controls	1	1.2 (1.0, 1.5)	Ref	1	1.1 (0.8, 1.6)	NA	1	0.9 (0.6, 1.4)	NA
Patient controls ³	2	0.8 (0.6, 1.2)	0.7 (0.5, 1.0)	1	1.1 (0.6, 2.1)		0		
Non-patient controls ⁴	1	1.0 (0.7, 1.5)	0.8 (0.5, 1.3)	1	0.8 (0.6, 1.2)		0		
Unknown	1	1.6 (0.7, 3.7)	1.3 (0.5, 3.2)	1	0.9 (0.4, 2.2)		1	1.1 (0.2, 5.8)	
Cohort/convenience	6	0.8 (0.6, 1.1)	0.7 (0.5, 1.0)	2	1.5 (0.8, 2.7)		2	0.5 (0.2, 1.0)	
Health status of non-cases									
Not patients ⁵	8	1.0 (0.9, 1.2)	Ref	4	1.1 (0.8, 1.3)	NA	3	0.7 (0.5, 1.1)	NA
Patients	2	0.8 (0.6, 1.2)	0.9 (0.6, 1.3)	1	1.1 (0.6, 2.1)		0		
Unknown patient status	1	1.6 (0.7, 3.7)	1.6 (0.7, 3.7)	1	0.9 (0.4, 2.2)		1	1.1 (0.2, 5.8)	

Table V.A.4: Genotype-smoking associations stratified by study design (continued)

XRCC1									
Arg399Gln				Arg194Trp			Arg280His		
	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²
Pack-years									
Case-based categories									
Case-control									
Population-based	2	0.9 (0.6, 1.4)	Ref	0		NA	1	0.5 (0.2, 1.5)	NA
Hospital-based	7	1.3 (1.1, 1.6)	1.5 (0.9, 2.4)	2	1.1 (0.7, 1.6)		2	1.1 (0.7, 1.8)	
Other	0			0			0		
Unknown	0			0			0		
Cohort/convenience	0			0			0		
Case-control	9	1.2 (1.0, 1.5)	NA	2	1.1 (0.7, 1.6)	NA	3	1.0 (0.6, 1.5)	NA
Other	0			0			0		
Non-case-based categories									
Case-control									
Population controls	2	0.9 (0.6, 1.4)	Ref	0		NA	1	0.5 (0.2, 1.5)	NA
Patient controls ³	4	1.4 (1.1, 1.7)	1.5 (0.9, 2.5)	2	1.1 (0.7, 1.6)		2	1.1 (0.7, 1.8)	
Non-patient controls ⁴	3	1.3 (0.9, 1.8)	1.4 (0.8, 2.5)	0			0		
Unknown	0			0			0		
Cohort/convenience	0			0			0		
Health status of non-cases									
Not patients ⁵	5	1.1 (0.9, 1.5)	Ref	0		NA	1	0.5 (0.2, 1.5)	NA
Patients	4	1.4 (1.1, 1.7)	1.2 (0.8, 1.8)	2	1.1 (0.7, 1.6)		2	1.1 (0.7, 1.8)	
Unknown patient status	0			0			0		

Table V.A.4: Genotype-smoking associations stratified by study design (continued)

Case-control Genotype Smoking Associations Stratified by Study Design (Continued)									
XPD						XRCC3			
Lys751Gln				Asp312Asn			Thr241Met		
	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²
Ever-never smoking									
Case-based categories									
Case-control									
Population-based	4	1.0 (0.8 , 1.1)	Ref	2	1.3 (1.0, 1.8)	Ref	2	1.0 (0.8, 1.3)	Ref
Hospital-based	7	0.9 (0.8, 1.1)	0.9 (0.7, 1.2)	7	1.0 (0.9, 1.2)	0.8 (0.6, 1.1)	4	1.0 (0.7, 1.3)	1.0 (0.6, 1.4)
Other	0			0			0		
Cohort/ convenience	1	1.0 (0.6, 1.6)	1.0 (0.6, 1.7)	0			3	1.2 (0.9, 1.6)	1.2 (0.8, 1.8)
Case-control	11	0.9 (0.8, 1.1)	NA	9	1.1 (1.0, 1.2)	NA	6	1.0 (0.8, 1.2)	Ref
Other	1	1.0 (0.6, 1.6)		0			3	1.2 (0.9, 1.6)	1.2 (0.8, 1.7)
Non-case-based categories									
Case-control									
Population controls	4	1.0 (0.8 , 1.1)	Ref	2	1.3 (1.0, 1.8)	Ref	2	1.0 (0.8, 1.3)	Ref
Patient controls ³	3	0.8 (0.6, 1.1)	0.8 (0.5, 1.2)	3	1.1 (0.9, 1.3)	0.8 (0.6, 1.1)	4	1.0 (0.7, 1.3)	1.0 (0.6, 1.4)
Non-patient controls ⁴	4	1.0 (0.8, 1.2)	1.0 (0.7, 1.4)	4	1.0 (0.8, 1.2)	0.7 (0.5, 1.1)	0		
Cohort/ convenience	1	1.0 (0.6, 1.6)	1.0 (0.5, 1.7)	0			3	1.2 (0.9, 1.6)	1.2 (0.8, 1.8)
Health status of non-cases									
Not patients ⁵	9	1.0 (0.9, 1.1)	Ref	6	1.1 (0.9, 1.3)	Ref	5	1.1 (0.9, 1.3)	Ref
Patients	3	0.8 (0.6, 1.1)	0.8 (0.6, 1.1)	3	1.1 (0.9, 1.3)	1 (0.8, 1.2)	4	1.0 (0.7, 1.3)	0.9 (0.6, 1.3)
Unknown patient status	0			0			0		

Table V.A.4: Genotype-smoking associations stratified by study design (continued)

XPD						XRCC3			
Lys751Gln			Asp312Asn			Thr241Met			
	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²
Current-not current smoking									
Case-based categories									
Case-control									
Population-based	1	1.0 (0.7, 1.3)	NA	0			1	1.0 (0.7, 1.5)	Ref
Hospital-based	1	0.7 (0.5, 1.2)		1	1.1 (0.7, 1.9)	NA	2	0.9 (0.6, 1.2)	0.8 (0.5, 1.4)
Other	0			0			1	2.0 (0.6, 6.6)	2.0 (0.6, 6.8)
Unknown	0			0			0		
Cohort/convenience	4	1.4 (1.0, 1.8)		0			3	0.9 (0.6, 1.2)	0.8 (0.5, 1.4)
Case-control	2	0.9 (0.7, 1.2)	Ref	1	1.1 (0.7, 1.9)	NA	4	1.0 (0.8, 1.2)	Ref
Other	4	1.4 (1.0, 1.8)	1.5 (1.0, 2.2)	0			3	0.9 (0.6, 1.2)	0.9 (0.6, 1.3)
Non-case-based categories									
Case-control									
Population controls	1	1.0 (0.7, 1.3)	NA	0			1	1.0 (0.7, 1.5)	Ref
Patient controls ³	1	0.7 (0.5, 1.2)		1	1.1 (0.7, 1.9)	NA	2	0.9 (0.6, 1.2)	0.8 (0.5, 1.4)
Non-patient controls ⁴	0			0			1	2.0 (0.6, 6.6)	2.0 (0.6, 6.8)
Unknown	0			0			0		
Cohort/convenience	4	1.4 (1.0, 1.8)		0			3	0.9 (0.6, 1.2)	0.8 (0.5, 1.4)
Health status of non-cases									
Not patients ⁵	5	1.2 (0.9, 1.4)	NA	0			5	1.0 (0.8, 1.2)	Ref
Patients	1	0.7 (0.5, 1.2)		1	1.1 (0.7, 1.9)	NA	2	0.9 (0.6, 1.2)	0.9 (0.6, 1.3)
Unknown patient status	0			0					

Table V.A.4: Genotype-smoking associations stratified by study design (continued)

XPD						XRCC3		
Lys751Gln			Asp312Asn			Thr241Met		
N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²	N	OR _z (95% CI) ¹	Ratio of ORs (95%CI) ²
Pack-years								
Case-based categories								
Case-control								
Population-based	2	0.8 (0.5, 1.3)	Ref	0		1	0.8 (0.4, 1.4)	NA
Hospital-based	5	1.3 (0.8, 2.1)	1.5 (0.7, 3.5)	4	1.1 (0.8, 1.5)	NA	3	0.8 (0.5, 1.3)
Other	0		0	0		0		
Unknown	0		0	0		0		
Cohort/convenience	0		0	0		0		
Case-control	7	1.1 (0.8, 1.7)	NA	4	1.1 (0.8, 1.58)	NA	4	NA
Other	0		0	0		0		
Non-case-based categories								
Case-control								
Population controls	2	0.8 (0.5, 1.3)	Ref	0		1	0.8 (0.4, 1.4)	NA
Patient controls ³	2	1.0 (0.6, 1.6)	1.2 (0.5, 2.6)	1	0.9 (0.5, 1.4)	NA	2	0.8 (0.5, 1.3)
Non-patient controls ⁴	3	1.6 (0.8, 3.1)	2.0 (0.9, 4.3)	3	1.3 (0.9, 1.8)		1	1.1 (0.4, 2.7)
Unknown	0		0	0		0		
Cohort/convenience	0		0	0		0		
Health status of non-cases								
Not patients ⁵	5	1.2 (0.7, 2.0)	Ref	3	1.3 (0.9, 1.8)	NA	2	0.9 (0.5, 1.4)
Patients	2	1.0 (0.6, 1.6)	0.8 (0.3, 1.8)	1	0.9 (0.5, 1.4)		2	0.8 (0.5, 1.3)
Unknown patient status	0			0				0.9 (0.5, 1.8)

Table V.A.4: Genotype-smoking associations stratified by study design (continued)

Abbreviations: CI=Confidence interval, HWE = Hardy Weinberg equilibrium, MAF=minor allele frequency, na=not applicable, OR_c=control-only genotype-smoking odds ratio, PY=pack-years, N=number of studies, Ref=referent, Q=Cochran's test of homogeneity, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ Unadjusted OR (95% CI): Genotype contrast for all SNPs is A/A (ref) vs. any a, where A is the more common allele and a is the less common; random effects for *XRCC1* Arg399Gln ever-never & *XPD* Lys751Gln, fixed effects for others

² Ratio of Odds Ratios: Meta-regression used to compare odds ratios in given study design stratum to the odds ratio in the designated reference stratum

³ Patient controls: controls are persons attending a hospital or disease clinic for treatment or diagnosis, does not include patients at wellness or check-up clinics

⁴ Non-patient controls: Case-control study participants who are not patients (i.e. not treated at hospital or disease clinic); they may be friend and family controls, cohort members in a nested case-control etc.; also excludes population-based controls

⁵ Non-cases that are not patients: population-based controls, friends and/or non-blood-related family of patients, convenience, community samples or cohort members.

Table V.A.5. *XRCCI* Arg399Gln and Smoking: Overall and by study characteristics

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI)
Overall												
Not stratified by ethnicity w/in study	21	0.01			11	0.4			9	0.3		
Stratified by ethnicity within study ⁶	23	0.02			12	0.3			na			
By Study Characteristic												
Continent												
North America	7	0.01	1.1 (0.9, 1.3)	Ref 1.0	2	0.6	1.2 (1.0 , 1.5)	Ref 0.7	2	0.1	1.2 (0.8, 1.7)	1.1 (0.7, 1.7)
Europe	11	0.2	1.1 (0.9, 1.3)	(0.8, 1.3)	5	0.5	0.8 (0.6, 1.0)	0.7 (0.5, 0.9)	5	0.8	1.1 (0.9, 1.4)	Ref 1.7
Asia	3	0.1	1.1 (0.7, 1.9)	1.0 (0.7, 1.6)	4	0.7	1.0 (0.8, 1.3)	0.8 (0.6, 1.1)	2	0.6	1.9 (1.2, 2.9)	1.7 (1.0 , 2.7)
Ethnicity/nationality ⁶												
Single-ethnicity studies ^{6,7}	14	0.3	1.0 (0.9, 1.2)	Ref 1.2	4	0.1	1.0 (0.8, 1.2)	Ref 0.9	5	0.4	1.1 (0.9, 1.5)	Ref
Multi-ethnic studies ⁷	2	0.01	1.2 (0.7, 2.1)	(0.9, 1.7)	1	na	0.9 (0.2, 3.4)	0.9 (0.2, 3.5)	0			
Unknown ethnicity	7	0.1	1.1 (0.9, 1.3)	1.1 (0.8, 1.4)	7	0.6	0.9 (0.7, 1.1)	0.9 (0.6, 1.3)	4	0.2	1.3 (1.1, 1.7)	1.2 (0.8, 1.8)
White >=99% ⁸	10	0.6	1.0 (0.9, 1.1)	Ref 1.1	2	0.2	0.9 (0.6, 1.2)	na	5	0.4	1.1 (0.9, 1.5)	na
African American >=99%	2	0.03	1.1 (0.5, 2.5)	1.1 (0.7, 1.8)	1	na	2.1 (1.1, 3.9)		0			
Han >=99%	2	0.2	1.4 (0.8, 2.5)	1.5 (0.8, 2.5)	1	na	1.0 (0.7, 1.4)		0			
Multi-ethnic studies	2	0.01	1.2 (0.7, 2.1)	1.3 (0.9, 1.8)	1	na	0.9 (0.2, 3.4)		0			

Table V.A.5. *XRCCI* Arg399Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
HWE p-value ⁶												
Single-ethnicity (continuous)	13	0.2	1.0 (0.9, 1.2)	0.8 (0.5, 1.3)	4	0.1	1.0 (0.8, 1.2)	1.3 (0.2, 7.2)	5	0.4	1.1 (0.9, 1.5)	16.9 (0.4, 713)
p < 0.05	0											
p ≥ 0.05, < 0.50	6	0.1	1.2 (1.0, 1.4)	Ref 0.9	2	0.5	1.2 (0.8, 1.8)	Ref 0.8	1	---	1.2 (0.9, 1.7)	na
p ≥ 0.50	16	0.05	1.1 (0.9, 1.2)	(0.7, 1.2)	10	0.3	0.9 (0.8, 1.1)	(0.5, 1.2)	8	0.2	1.2 (1.0, 1.6)	
p < 0.50	6	0.1	1.2 (1.0, 1.4)	Ref 0.9	2	0.5	1.2 (0.8, 1.8)	Ref 0.8	1	---	1.2 (0.9, 1.7)	na
p ≥ 0.50	16	0.05	1.1 (0.9, 1.2)	(0.7, 1.2)	10	0.3	0.9 (0.8, 1.1)	(0.5, 1.2)	8	0.2	1.2 (1.0, 1.6)	
p < 0.10	1	---	0.9 (0.6, 1.3)	na	0				0			
p ≥ 0.10	21	0.02	1.1 (1.0, 1.2)		12				9			
Age												
Age non-missing	20	0.01	1.1 (1.0, 1.2)	1.0 (1.0, 1.0)	10	0.3	1.0 (0.9, 1.1)	1.0 (1.0, 1.0)	9	0.3	1.2 (1.0, 1.5)	1.0 (1.0, 1.1)
≤ 47.9y ⁹	1	---	0.8 (0.4, 1.6)	na	5	0.3	0.9 (0.7, 1.1)	Ref	0			na
> 47.9 y	19	0.01	1.1 (1.0, 1.2)		5	0.4	1.1 (0.9, 1.3)	1.2 (0.9, 1.7)	9	0.3	1.2 (1.0, 1.5)	

Table V.A.5. *XRCCI* Arg399Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
Age (continued)												
<=59y ¹⁰	8	0.1	1.1 (0.9, 1.3)	Ref 0.9	8	0.6	0.9 (0.7, 1.1)	na	3	0.1	1.1 (0.8, 1.6)	Ref 1.3
>59, <=63y	7	0.02	1.0 (0.8, 1.3)	(0.7, 1.2)	1	---	0.9 (0.6, 1.3)		5	0.4	1.3 (1.1, 1.6)	(0.8, 2.1)
>63y	5	0.5	1.2 (1.0, 1.4)	1.1 (0.8, 1.5)	1	---	1.2 (1.0, 1.5)		1	---	0.8 (0.2, 2.6)	0.7 (0.2, 2.7)
<=59y ¹¹	8	0.1	1.1 (0.9, 1.3)	Ref 1.0	8	0.6	0.9 (0.7, 1.1)	Ref 1.3	3	0.1	1.1 (0.8, 1.6)	Ref 1.2
>59y	12	0.02	1.1 (0.9, 1.3)	(0.8, 1.3)	2	0.2	1.1 (0.9, 1.4)	(1.0, 1.7)	6	0.4	1.3 (1.0, 1.6)	(0.8, 1.8)
At or below median	12	0.1	1 (0.9, 1.2)	Ref	5	0.3	0.9 (0.7, 1.1)	Ref 1.2	5	0.4	1.2 (1.0, 1.5)	Ref 1.1
Above median	8	0.02	1.2 (0.9, 1.4)	1.1 (0.9, 1.4)	5	0.4	1.1 (0.9, 1.3)	(0.9, 1.7)	4	0.2	1.3 (1.0, 1.8)	(0.7, 1.6)
Gender												
Percent male (continuous)	21			1.1 (0.7, 1.5)	10	0.3	1.0 (0.9, 1.1)	0.7 (0.5, 0.9)	9	0.3	1.2 (1.0, 1.5)	1.3 (0.6, 2.8)
Percent male (mixed gender only)	13	0.1	1.0 (0.9, 1.2)	1.4 (0.6, 3.5)	4	0.7	1.0 (0.8, 1.3)	0.9 (0.1, 8)	6	0.1	1.3 (1.0, 1.5)	1 (0.1, 11.4)

Table V.A.5. *XRCCI* Arg399Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
Gender (continued)												
100% female participants	5	0.01	1.1 (0.8, 1.4)	Ref	2	0.3	1.2 (0.9, 1.5)	Ref	1	--	0.9 (0.4, 1.8)	Ref
<= 69% male participants ¹²	7	0.1	1.0 (0.8, 1.3)	1.0 (0.7, 1.3)	2	0.7	1.0 (0.8, 1.4)	0.9 (0.6, 1.3)	3	0.1	1.1 (0.8, 1.5)	1.1 (0.5, 2.8)
> 69% male participants	6	0.3	1.1 (0.9, 1.3)	1.0 (0.7, 1.4)	2	0.2	1.0 (0.7, 1.4)	0.8 (0.5, 1.3)	3	0.2	1.4 (1.1, 1.8)	1.5 (0.6, 3.8)
100% male	3	0.7	1.2 (0.9, 1.6)	1.1 (0.7, 1.7)	4	0.7	0.7 (0.6, 1.0)	0.6 (0.4, 0.9)	2	0.7	1.3 (0.8, 2.2)	1.5 (0.5, 4)
All female	5	0.01	1.1 (0.8, 1.4)	Ref	2	0.3	1.2 (0.9, 1.5)	Ref	1	--	0.9 (0.4, 1.8)	Ref
Mixed gender	13	0.1	1.0 (0.9, 1.2)	1.0 (0.7, 1.3)	4	0.7	1.0 (0.8, 1.3)	0.9 (0.6, 1.2)	6	0.1	1.3 (1.0, 1.5)	1.4 (0.7, 2.9)
All male	3	0.7	1.2 (0.9, 1.6)	1.1 (0.7, 1.7)	4	0.7	0.7 (0.6, 1.0)	0.6 (0.4, 0.9)	2	0.7	1.3 (0.8, 2.2)	1.5 (0.6, 3.6)
Study outcome												
Lung cancer	6	0.4	1.0 (0.9, 1.1)	Ref	1	---	0.9 (0.6, 1.3)	Ref	6	0.2	1.3 (1.1, 1.6)	Ref
Other cancer	12	0.02	1.2 (1.0, 1.4)	1.2 (1.0, 1.6)	3	0.3	1.1 (0.9, 1.3)	1.2 (0.8, 2)	3	0.7	1.0 (0.7, 1.5)	0.8 (0.5, 1.2)
Non-cancer disease	0				1	---	1.6 (0.6, 3.7)	1.8 (0.7, 4.7)	0			
Non-disease	3	0.7	1.0 (0.8, 1.4)	1.1 (0.7, 1.6)	6	0.6	0.8 (0.6, 1.1)	0.9 (0.6, 1.5)	0			

Table V.A.5. *XRCCI* Arg399Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
Study Outcome (continued)												
Cancer	18	0.01	1.1 (0.9, 1.2)	Ref	4	0.3	1.1 (0.9, 1.2)	Ref	9	0.3	1.2 (1.0, 1.5)	Ref
Non-cancer disease	0				1	---	1.6 (0.6, 3.7)	1.5 (0.6, 3.7)	0			
Non-disease	3	0.7	1.0 (0.8, 1.4)	0.9 (0.6, 1.4)	6	0.6	0.8 (0.6, 1.1)	0.8 (0.6, 1.1)	0			
Lung cancer	6	0.4	1.0 (0.9, 1.1)	Ref	1	---	0.9 (0.6, 1.3)	na	6	0.2	1.3 (1.1, 1.6)	Ref
All other	15	0.04	1.2 (1.0, 1.3)	1.2 (1.0, 1.5)	10	0.3	1 (0.9, 1.2)		3	0.7	1 (0.7, 1.5)	0.8 (0.5, 1.2)
MAF ²												
MAF (cutpoints assigned by tertiles across SNP)												
0.10-0.27	6	0.1	1.1 (0.7, 1.5)	Ref	4	0.1	1.1 (0.9, 1.4)	Ref	2	0.6	1.9 (1.2, 2.9)	Ref
>0.27-0.36	10	0.01	1.1 (1.0, 1.3)	1.1 (0.8, 1.5)	4	0.4	0.9 (0.6, 1.4)	0.8 (0.5, 1.4)	7	0.5	1.1 (0.9, 1.4)	0.6 (0.4, 1.0)
>0.36-0.50	6	0.9	1 (0.8, 1.2)	1 (0.7, 1.4)	4	0.7	0.8 (0.7, 1.1)	0.8 (0.5, 1.1)	0			
MAF (cutpoints proxy for ethnicity)												
10%-20%	3	0.1	1 (0.5, 1.8)	Ref	1	---	2.1 (1.1, 3.9)	Ref	0			
>20%-27%	3	0.1	1.1 (0.7, 1.9)	1.1 (0.6, 1.9)	4	0.7	1.0 (0.8, 1.3)	0.5 (0.2, 0.9)	2	0.6	1.9 (1.2, 2.9)	Ref
>27%-50%	16	0.04	1.1 (1.0, 1.2)	1.1 (0.7, 1.7)	7	0.6	0.9 (0.7, 1.1)	0.4 (0.2, 0.8)	7	0.5	1.1 (0.9, 1.4)	0.6 (0.4, 1.0)

Table V.A.5. *XRCCI* Arg399Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
MAF ² (continued)												
MAF-assigned ethnicity												
White	15	0.3	1.1 (0.9, 1.2)	Ref 0.9	6	0.5	0.9 (0.7, 1.1)	Ref 2.4	7	0.5	1.1 (0.9, 1.4)	Ref
African American ¹³	3	0.1	1 (0.5, 1.8)	(0.6, 1.5)	1		2.1 (1.1, 3.9)	(1.2, 4.7)	0			
Han	3	0.1	1.1 (0.7, 1.9)	1.0 (0.7, 1.5)	4	0.7	1.0 (0.8, 1.3)	1.1 (0.8, 1.6)	2	0.6	1.9 (1.2, 2.9)	1.6 (1.0, 2.6)
Multi-ethnic studies	2	0.01	1.2 (0.7, 2.1)	1.2 (0.8, 1.7)	1	--	0.9 (0.2, 3.4)	1 (0.3, 4.1)	0			
Smoking prevalence ¹⁴												
Continuous	21	0.01		1.7 (0.6, 4.7)	11	0.4	1.0 (0.9, 1.1)	0.2 (0.1, 0.7)	9	0.3	1.2 (1.0, 1.5)	0.7 (0.2, 2.6)
0-0.365	2	0.9	0.7 (0.5, 1.1)	Ref 1.5	7	0.6	1.1 (0.9, 1.2)	Ref 0.6	5	0.3	1.4 (1.1, 1.8)	Ref 0.9
>0.365-507	3	0.3	1.1 (0.9, 1.3)	1.5 (0.9, 2.4)	2	0.3	0.6 (0.3, 1.3)	0.6 (0.3, 1.3)	1	---	1.2 (0.9, 1.7)	0.9 (0.6, 1.3)
>0.507-0.602	7	0.02	1.2 (0.9, 1.5)	1.6 (1.0, 2.6)	1	---	0.7 (0.5, 1.1)	0.7 (0.5, 1.1)	3	0.3	0.9 (0.6, 1.4)	0.7 (0.4, 1.1)
>0.602-1	9	0.3	1 (0.9, 1.2)	1.4 (0.9, 2.3)	1	---	0.9 (0.2, 3.4)	0.8 (0.2, 3.3)	0			
>0-0.507	5	0.2	1 (0.8, 1.2)	Ref 1.1	9	0.5	1.0 (0.9, 1.2)	Ref 0.7	6	0.4	1.3 (1.1, 1.6)	Ref 0.7
>0.507-1	16	0.01	1.1 (1.0, 1.3)	1.1 (0.9, 1.5)	2	0.8	0.7 (0.5, 1.1)	0.7 (0.5, 1.1)	3	0.3	0.9 (0.6, 1.4)	0.7 (0.4, 1.1)

Abbreviations: CI=Confidence interval, na=not applicable, HWE = Hardy Weinberg equilibrium PY=pack-years, OR_z=control-only genotype-smoking odds ratio (bolded), N=number of studies, Ref=referent, Q=Cochran's test of homogeneity, SNP=single nucleotide polymorphism, MAF=minor allele frequency, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

Table V.A.5. *XRCCI* Arg399Gln and Smoking: Overall and by study characteristics (continued)

- ¹ OR=Unadjusted odds ratio for *XRCCI* Arg399Gln: *XRCCI* Arg/Arg (ref) vs. any Gln, never smoking (ref) vs. ever smoking, random effects estimates
- ² OR=Unadjusted odds ratio for *XRCCI* Arg399Gln: *XRCCI* Arg/Arg (ref) vs. any Gln, not current smoker (ref) vs. current smoker, fixed effects estimates
- ³ OR=Unadjusted odds ratio for *XRCCI* Arg399Gln: *XRCCI* Arg/Arg (ref) vs. any Gln, lightest smokers (ref) vs. heaviest smokers [lightest excludes never smokers], fixed effects
- ⁴ PY contrast is between lightest non-zero category of pack-years (ref) vs. heaviest category of PY
- ⁵ Ratio of Odds Ratios: Compares odds ratio in given study characteristic stratum to the odds ratio in the designated reference stratum for that study characteristic by meta-regression
- ⁶ Studies that can be stratified by ethnicity are included as separate single-ethnicity studies; studies w 99%-100% of 1 ethnicity are classified as single-ethnicity
- ⁷ Only includes studies with explicitly stated ethnic makeup
- ⁸ White = Caucasian, white or non-Hispanic white; African American = African American or black;
Han = Han, Han Chinese or ethnic Chinese; Japan, Korea and China may include ethnic minorities.
- ⁹ Median of studies included in *XRCCI* 399 current/not current smoker analyses.
- ¹⁰ Categories based on thirds from studies included in *XRCCI* 399 PY analyses ($\leq 59y$, $>59y-63y$, $>63y$)
- ¹¹ Median of all studies w age info (all SNPs) w age info [range:23.6-69y, mean: 56.51y SD: 9.84y]
- ¹² Median proportion male in all studies (all SNPs, all smoking exposures): 0.69
- ¹³ For 399 ever-never 1 study from Hungary is included using MAF as proxy for ethnicity
- ¹⁴ Smoking prevalence is contrast-specific (defined as "ever", "current" or "heavier PY" as appropriate)

Table V.A.6. *XPD* Lys751Gln and Smoking: Overall and by study characteristics

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)
Overall	12	0.5	0.9 (0.8, 1.1)		6	0.2	1.1 (0.9, 1.3)		7	0.019	1.2 (1.0, 1.5)	
By Study Characteristic												
Continent												
North America	4	0.4	0.9 (0.8, 1)	Ref 1.2	1	---	0.9 (0.7, 1.3)	Ref 1.2	3	0.006	1.5 (0.7, 3.2)	Ref 0.7
Europe	6	0.3	1.0 (0.8, 1.3)	1.1 (0.9, 1.5)	5	0.2	1.2 (0.9, 1.5)	1.2 (0.8, 1.8)	3	0.503	1.0 (0.7, 1.4)	0.5 (0.3, 1.5)
Asia	2	0.5	1.0 (0.6, 1.7)	1.1 (0.7, 1.9)	0				1	na	0.8 (0.3, 1.7)	0.5 (0.1, 1.9)
Ethnicity/nationality												
Single-ethnicity studies ⁶	6	0.3	1.0 (0.8, 1.2)	Ref 0.8	1	---	1.2 (0.8, 1.8)	na	5	0.097	1.4 (0.9, 2.1)	na
Multi-ethnic studies ⁷	2	0.6	0.8 (0.7, 1.0)	0.8 (0.6, 1.1)	1	---	0.9 (0.7, 1.3)		1	---	0.8 (0.5, 1.3)	
Unknown ethnicity	4	0.7	1.0 (0.8, 1.3)	1.1 (0.8, 1.4)	4	0.1	1.1 (0.8, 1.5)		1	---	0.8 (0.4, 1.4)	
White >=99% ⁸	4	0.2	1.0 (0.8, 1.2)	Ref 1.0	1	---	1.2 (0.8, 1.8)	na	4	0.181	1.6 (1.0, 2.4)	na
African American >=99%	0			0.8 (0.6, 1.8)	0				0			
Han >=99%	2	0.5	1.0 (0.6, 1.7)	0.8 (0.6, 1.1)	0				1	---	0.8 (0.3, 1.7)	

Table V.A.6. *XPD* Lys751Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
HWE p-value												
Continuous (single ethnicity)	6	0.3	1.0 (0.8, 1.2)	0.8 (0.5, 1.2)	1	---	1.2 (0.8, 1.8)	na	5	0.097	1.4 (0.9, 2.1)	2.2 (0, 93.7)
HWE p <0.05	1	---	1.1 (0.8, 1.7)	Ref	2	0.6	1.3 (0.9, 1.8)	Ref	0			na
HWE p >=0.05, <0.50	4	0.6	1.1 (0.8, 1.4)	1.0 (0.6, 1.6)	1	---	0.7 (0.5, 1.2)	0.6 (0.3, 1.2)	1	---	0.8 (0.4, 1.4)	
HWE p >=0.50	7	0.4	0.9 (0.8, 1.0)	0.8 (0.5, 1.2)	3	0.2	1.1 (0.9, 1.4)	0.9 (0.5, 1.5)	6	0.022	1.2 (0.8, 1.9)	
HWE p <0.50	5	0.8	1.1 (0.9, 1.4)	Ref	3	0.2	1.0 (0.8, 1.4)	Ref	1	---	0.8 (0.4, 1.4)	Ref
HWE p >=0.50	7	0.4	0.9 (0.8, 1.0)	0.8 (0.6, 1.0)	3	0.2	1.1 (0.9, 1.4)	1.1 (0.7, 1.9)	6	0.022	1.2 (0.8, 1.9)	1.6 (0.5, 4.4)
HWE p <0.10	2	0.9	1.1 (0.8, 1.5)	Ref	2	0.6	1.3 (0.9, 1.8)	Ref	1	---	0.8 (0.4, 1.4)	na
HWE p >=0.10	10	0.4	0.9 (0.8, 1.0)	0.8 (0.6, 1.1)	4	0.1	1.0 (0.8, 1.3)	0.8 (0.5, 1.2)	6	0.022	1.2 (0.8, 1.9)	
Age												
Age non-missing (continuous)	12	0.5	0.9 (0.8, 1.1)	1.0 (1.0, 1.0)	6	0.2	1.1 (0.9, 1.3)	1.0 (1.0, 1.0)	7	0.019	1.1 (0.8, 1.7)	1.0 (0.9, 1.1)
<= 47.9y ⁹	0			na	3	0.5	1.4 (1.0, 1.9)	Ref	0			na
> 47.9 y	12	0.5	0.9 (0.8, 1.1)	0 (0, 0)	3	0.4	0.9 (0.7, 1.2)	0.7 (0.5, 1.0)	7	0.019	1.1 (0.8, 1.7)	

Table V.A.6. *XPD* Lys751Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
Age (continued)												
<=59y ¹⁰	7	0.6	0.9 (0.8, 1.0)	Ref 1.1	6	0.2	1.1 (0.9, 1.3)	na	4	0.054	1.4 (0.8, 2.4)	Ref 0.6
>59, <=63y	4	0.2	1.0 (0.8, 1.2)	(0.8, 1.4)	0				3	0.362	0.9 (0.6, 1.2)	(0.3, 1.2)
>63y	1	---	1.4 (0.8, 2.6)	1.5 (0.8, 2.9)	0				0			
<=59y ¹¹	7	0.6	0.9 (0.8, 1.0)	Ref 1.1	6	0.2	1.1 (0.9, 1.3)	na	4	0.054	1.4 (0.8, 2.4)	Ref 0.6
>59y	5	0.2	1.0 (0.8, 1.2)	(0.9, 1.4)	0				3	0.362	0.9 (0.6, 1.2)	(0.3, 1.2)
At or below median	6	0.5	0.9 (0.8, 1.1)	Ref 1.1	3	0.5	1.4 (1.0, 1.9)	Ref 0.7	4	0.054	1.4 (0.8, 2.4)	Ref 0.6
Above median	6	0.3	1.0 (0.8, 1.1)	(0.9, 1.4)	3	0.4	0.9 (0.7, 1.2)	(0.5, 1.0)	3	0.362	0.9 (0.6, 1.2)	(0.3, 1.2)
Gender												
Percent male (continuous)	12	0.5	0.9 (0.8, 1.1)	1.0 (0.7, 1.4)	6	0.2	1.1 (0.9, 1.3)	1.1 (0.7, 2)	7		0 (0, 0)	0.8 (0.2, 3.4)
Percent male, mixed gender only	8	0.9	1.0 (0.9, 1.2)	0.8 (0.3, 2.4)	2	0.3	1.5 (1.0, 2.3)	na	5	0.006	1.2 (0.7, 2)	0 (0, 0.1)
100% female participants	2	0.1	0.9 (0.7, 1.1)	Ref 1.2	1	---	0.9 (0.7, 1.3)	na	1	---	0.9 (0.4, 1.9)	Ref 1.7
<= 69% male participants ¹²	5	0.7	1.0 (0.9, 1.2)	(0.9, 1.5)	1	---	1.3 (0.7, 2.2)		3	0.024	1.5 (0.7, 3.1)	(0.5, 5.6)
> 69% male participants	3	0.7	0.9 (0.7, 1.3)	1.1 (0.7, 1.5)	1	---	1.9 (1.0, 3.8)		2	0.970	0.8 (0.5, 1.2)	0.8 (0.2, 2.9)
100% male	2	0.2	0.7 (0.5, 1.1)	0.8 (0.5, 1.3)	3	0.2	1.0 (0.8, 1.4)		1	---	1.3 (0.7, 2.5)	1.4 (0.3, 6.1)

Table V.A.6. *XPD* Lys751Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
Gender (continued)												
All female	2	0.1	0.9 (0.7, 1.1)	Ref 1.1	1	---	0.9 (0.7, 1.3)	Ref 1.6	1	na	0.9 (0.4, 1.9)	na
Mixed gender	8	0.9	1.0 (0.9, 1.2)	(0.9, 1.5)	2	0.3	1.5 (1.0, 2.3)	(0.9, 2.7)	5	0.006	1.2 (0.7, 2)	
All male	2	0.2	0.7 (0.5, 1.1)	0.8 (0.5, 1.3)	3	0.2	1.0 (0.8, 1.4)	1.1 (0.7, 1.7)	1	na	1.3 (0.7, 2.5)	
Study outcome												
Lung cancer	3	0.5	1.0 (0.8, 1.2)	Ref 0.9	0		0.9 (0.7, 1.2)	Ref	3	0.080	1.6 (0.8, 3.1)	Ref 0.5
Other cancer	8	0.2	0.9 (0.8, 1.1)	(0.7, 1.2)	2	0.4	0.9 (0.7, 1.2)	Ref	4	0.565	0.9 (0.7, 1.2)	(0.3, 0.9)
Non-cancer disease	0				0				0			
Non-disease	1	---	1.0 (0.6, 1.6)	1.0 (0.6, 1.7)	4	0.7	1.3 (1.0, 1.8)	1.5 (1.0, 2.2)	0			
Cancer	11	0.4	0.9 (0.8, 1.1)	na	2	0.4	0.9 (0.7, 1.2)	Ref	7	0.019	1.1 (0.8, 1.7)	na
Non-cancer disease	0				0				0			
Non-disease	1	na	1.0 (0.6, 1.6)		4	0.7	1.3 (1.0, 1.8)	1.5 (1.0, 2.2)	0			
Lung cancer	3	0.5	1.0 (0.8, 1.2)	Ref 0.9	0		1.1 (0.9, 1.3)	na	3	0.080	1.6 (0.8, 3.1)	Ref 0.5
All other	9	0.3	0.9 (0.8, 1.1)	(0.7, 1.2)	6	0.2	1.1 (0.9, 1.3)		4	0.565	0.9 (0.7, 1.2)	(0.3, 0.9)

Table V.A.6. *XPD* Lys751Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
MAF												
MAF non-missing	12	0.5	0.9 (0.8, 1.1)	1.3 (0.3, 6.2)	6	0.2	1.1 (0.9, 1.3)	0 (0, 3772.6)	7			1.2 (0, 45.7)
MAF (cutpoints assigned by median across SNP)												
0.01-0.37	7	0.5	0.9 (0.8, 1.0)	Ref	2	0.3	1.0 (0.7, 1.3)	Ref	5	0.016	1.3 (0.8, 2.1)	Ref
>0.37-0.50	5	0.7	1.1 (0.9, 1.4)	1.2 (1.0, 1.6)	4	0.2	1.1 (0.9, 1.5)	1.1 (0.7, 1.8)	2	0.768	0.8 (0.5, 1.3)	0.7 (0.3, 1.5)
MAF (cutpoints proxy for ethnicity)												
1%-15%	2	0.5	1.0 (0.6, 1.7)	Ref	0			na	1	---	0.8 (0.3, 1.7)	na
>15%-50%	10	0.3	0.9 (0.8, 1.0)	0.9 (0.6, 1.6)	6	0.2	1.1 (0.9, 1.3)		6	0.015	1.2 (0.8, 1.8)	
MAF-assigned ethnicity												
White	8	0.4	1.0 (0.9, 1.2)	Ref	5	0.2	1.2 (0.9, 1.5)	na	5	0.052	1.3 (0.9, 2.1)	na
African American	0				0				0			
Han	2	0.5	1.0 (0.6, 1.7)	1.0 (0.6, 1.7)	0			na	1	---	0.8 (0.3, 1.7)	na
Multi-ethnic studies	2	0.6	0.8 (0.7, 1.0)	0.8 (0.6, 1.0)	1	---	0.9 (0.7, 1.3)		1	---	0.8 (0.5, 1.3)	
Smoking prevalence ¹³												
Continuous	12	0.5	0.9 (0.8, 1.1)	0.3 (0.1, 1.1)	6	0.2	1.1 (0.9, 1.3)	2 (0.5, 8.6)	7			0.3 (0, 7.8)

Table V.A.6. *XP*D Lys751Gln and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)	N	Q p-value	OR _z (95% CI)	Ratio of ORs (95% CI)
By Study Characteristic												
Smoking prevalence (continued) ¹³												
0-0.365	1	---	1.3 (0.8, 1.9)	Ref 0.9	4	0.1	1.0 (0.8, 1.3)	Ref 1.5	2	0.009	1.3 (0.5, 3.6)	Ref 0.6
>0.365-507	2	0.7	1.2 (0.8, 1.7)	(0.5, 1.6)	1	---	1.6 (0.6, 4.2)	(0.5, 4.6)	3	0.918	0.8 (0.5, 1.1)	(0.3, 1.3)
>0.507-0.602	6	0.5	0.9 (0.8, 1.1)	0.7 (0.5, 1.1)	1	---	1.2 (0.8, 1.8)	1.1 (0.6, 2.2)	2	0.359	1.5 (0.9, 2.6)	1.2 (0.5, 2.9)
>0.602-1	3	0.3	0.9 (0.7, 1.1)	0.7 (0.4, 1.1)	0				0			
>0-0.507	3	0.9	1.2 (0.9, 1.6)	Ref 0.7	5	0.2	1.0 (0.8, 1.3)	na	5	0.010	1.0 (0.6, 1.6)	Ref 1.6
>0.507-1	9	0.5	0.9 (0.8, 1.0)	0.7 (0.5, 1.0)	1	---	1.2 (0.8, 1.8)		2	0.359	1.5 (0.9, 2.6)	1.6 (0.7, 3.9)

Abbreviations: CI=Confidence interval, na=not applicable, HWE = Hardy Weinberg equilibrium, PY=pack-years, OR_z=control-only genotype-smoking odds ratio,), N=number of studies, Ref=referent, Q=Cochran's test of homogeneity, SNP=single nucleotide polymorphism, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ OR=Unadjusted odds ratio for *XP*D Lys751Gln: Lys/Lys (ref) vs. any Gln, never smoking (ref) vs. ever smoking, fixed effects

² OR=Unadjusted odds ratio for *XP*D Lys751Gln: Lys/Lys (ref) vs. any Gln, not current smoker (ref) vs. current smoker, fixed effects

³ OR=Unadjusted odds ratio for *XP*D Lys751Gln: Lys/Lys (ref) vs. any Gln, lightest non-zero smokers (ref) vs. heaviest smokers; stratified random effects

⁴ PY contrast is between lightest non-zero category of pack-years (ref) vs. heaviest category of PY

⁵ Ratio of Odds Ratios: Compares OR in given study characteristic stratum to the OR in the designated reference stratum for that study characteristic by meta-regression

⁶ Studies w 99%-100% of 1 ethnicity are classified as single-ethnicity

⁷ Only includes studies with explicitly stated ethnic makeup

⁸ White = Caucasian, white or non-Hispanic white; African American = African American or black; Han = Han, Han Chinese or ethnic Chinese

Table V.A.6. *XPD* Lys751Gln and Smoking: Overall and by study characteristics (continued)

⁹ Median of studies included in *XRCC1* 751 current/not current smoker analyses.

¹⁰ Categories based on thirds from studies included in *XRCC1* 399 PY analyses (≤ 59 y, >59 y-63y, >63 y)

¹¹ Median of all studies w age info (all SNPs) w age info [range:23.6-69y, mean: 56.7y SD: 9.8y]

¹² Median proportion male in all studies (all SNPs, all smoking exposures): 0.69

¹³ Smoking prevalence is contrast-specific (defined as "ever", "current" or "heavier PY" as appropriate)

Table V.A.7. *XRCC3* Thr241Met and Smoking: Overall and by study characteristics

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵
Overall	9	0.5	1.0 (0.9, 1.2)		7	0.7	0.9 (0.8, 1.1)		4	0.7	0.8 (0.6, 1.2)	
By Study Characteristic												
Continent												
North America	2	0.1	0.9 (0.6, 1.3)	Ref	1	---	0.8 (0.5, 1.2)	na	2	0.3	0.7 (0.4, 1.3)	Ref
Europe	7	0.7	1.1 (0.9, 1.3)	1.1 (0.8, 1.7)	5	0.9	0.9 (0.7, 1.1)		2	0.6	0.9 (0.6, 1.4)	1.2 (0.6, 2.5)
Asia	0				1	---	2.0 (0.6, 6.6)		0			
Ethnicity/nationality												
Single-ethnicity studies ⁶	3	0.7	0.8 (0.6, 1.1)	Ref	3	0.3	0.8 (0.6, 1.1)	Ref	2	0.9	1.0 (0.6, 1.7)	na
Multi-ethnic studies ⁷	2	0.2	1.0 (0.7, 1.3)	1.2 (0.8, 1.9)	1	---	1.0 (0.7, 1.5)	1.2 (0.7, 2)	1	---	0.6 (0.3, 1.2)	
Unknown ethnicity	4	0.9	1.2 (1.0, 1.5)	1.5 (1.0, 2.2)	3	0.7	0.9 (0.7, 1.3)	1.1 (0.7, 1.7)	1	---	0.8 (0.4, 1.4)	
White >=99% ⁸	3	0.7	0.8 (0.6, 1.1)	Ref	2	0.9	0.8 (0.6, 1.1)	Ref	2	0.9	1.0 (0.6, 1.7)	na
African American >=99%	0				0				0			
Han >=99%	0				1	---	2.0 (0.6, 6.6)		0			

Table V.A.7. *XRCC3* Thr241Met and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵
By Study Characteristic												
HWE p-value												
Continuous (single ethnicity)	3	0.7	0.8 (0.6, 1.1)	1.0 (0.3, 3.4)	3	0.3	0.8 (0.6, 1.1)	0.4 (0.1, 1.6)	2	0.9	1.0 (0.6, 1.7)	
HWE p <0.05	0				1	---	2.0 (0.6, 6.6)	na	0			na
HWE p >=0.05, <0.50	2	0.3	1.1 (0.8, 1.6)	Ref	0				2	0.6	0.9 (0.6, 1.4)	
HWE p >=0.50	7	0.4	1.0 (0.8, 1.2)	0.9 (0.6, 1.4)	6	0.9	0.9 (0.7, 1.1)		2	0.3	0.7 (0.4, 1.3)	
HWE p <0.50	2	0.3	1.1 (0.8, 1.6)	Ref	1	---	2.0 (0.6, 6.6)	na	2	0.6	0.9 (0.6, 1.4)	na
HWE p >=0.50	7	0.4	1.0 (0.8, 1.2)	0.9 (0.6, 1.4)	6	0.9	0.9 (0.7, 1.1)		2	0.3	0.7 (0.4, 1.3)	
HWE p <0.10	0			na	1	---	2.0 (0.6, 6.6)	na	0			na
HWE p >=0.10	9	0.5	1.0 (0.9, 1.2)		6	0.9	0.9 (0.7, 1.1)		4	0.7	0.8 (0.6, 1.2)	
Age												
Age non- missing	9	0.5	1.0 (0.9, 1.2)	1.0 (1.0, 1.0)	7	0.7	0.9 (0.8, 1.1)	1.0 (1.0, 1.0)	4	0.7	0.8 (0.6, 1.2)	0.9 (0.8, 1.1)

Table V.A.7. *XRCC3* Thr241Met and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵
By Study Characteristic												
Age												
<= 47.9y ⁹	1	---		na	2	0.6	0.8 (0.6, 1.1)	Ref	0			na
> 47.9 y	8	0.4	1.0 (0.9, 1.2)		5	0.6	1.0 (0.8, 1.2)	1.2 (0.8, 1.9)	4	0.7	0.8 (0.6, 1.2)	
<=59y ¹⁰	5	0.3	1.0 (0.8, 1.2)	Ref	6	0.9	0.9 (0.7, 1.1)	na	1	---	1.1 (0.4, 2.7)	na
>59, <=63y	2	0.3	1.1 (0.8, 1.6)	1.1 (0.8, 1.7)	1	---	2.0 (0.6, 6.6)		2	0.6	0.9 (0.6, 1.4)	
>63y	2	0.5	1.2 (0.8, 1.7)	1.2 (0.8, 1.9)	0				1	---	0.6 (0.3, 1.2)	
<=59y ¹¹	5	0.3	1 (0.8, 1.2)	Ref	6	0.9	0.9 (0.7, 1.1)	na	1	---	1.1 (0.4, 2.7)	na
>59y	4	0.7	1.1 (0.9, 1.5)	1.2 (0.8, 1.6)	1	---	2.0 (0.6, 6.6)		3	0.5	0.8 (0.5, 1.1)	
At or below median	5	0.3	1.0 (0.8, 1.2)	Ref	4	0.8	0.8 (0.7, 1.1)	Ref	3	0.8	0.9 (0.6, 1.4)	na
Above median	4	0.7	1.1 (0.9, 1.5)	1.2 (0.8, 1.6)	3	0.5	1.0 (0.8, 1.4)	1.2 (0.8, 1.8)	1	---	0.6 (0.3, 1.2)	

Table V.A.7. *XRCC3* Thr241Met and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵
By Study Characteristic												
Gender												
Percent male (continuous)	9	0.5	1.0 (0.9, 1.2)	1.0 (0.4, 2.4)	7	0.7	0.9 (0.8, 1.1)	0.6 (0.3, 1.5)	4	0.7	0.8 (0.6, 1.2)	1.1 (0.1, 8.6)
Percent male, mixed gender only	6	0.3	1 (0.9, 1.3)	1.1 (0.1, 12.6)	4	0.5	1.0 (0.8, 1.3)	0.1 (0, 3.4)				
100% female participants	0				0				0			na
<= 69% male participants ¹²	4	0.1	1.0 (0.8, 1.2)	Ref	4	0.5	1.0 (0.8, 1.3)	Ref	2	0.6	0.9 (0.5, 1.4)	
> 69% male participants	2	0.5	1.2 (0.8, 1.9)	1.2 (0.7, 2)	0				1	---	0.6 (0.3, 1.2)	
100% male	3	0.8	1.0 (0.8, 1.4)	1.0 (0.7, 1.5)	3	0.8	0.8 (0.6, 1.1)	0.9 (0.6, 1.3)	1	---	1.0 (0.5, 1.9)	
All female	0				0				0			
Mixed gender	6	0.3	1.0 (0.9, 1.3)	Ref	4	0.5	1.0 (0.8, 1.3)	Ref	3	0.6	0.8 (0.5, 1.1)	na
All male	3	0.8	1.0 (0.8, 1.4)	1.0 (0.7, 1.4)	3	0.8	0.8 (0.6, 1.1)	0.9 (0.6, 1.3)	1	---	1.0 (0.5, 1.9)	

Table V.A.7. *XRCC3* Thr241Met and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵
By Study Characteristic												
Study outcome												
Lung cancer	0				0				1	---	1.1 (0.4, 2.7)	na
Other cancer	5	0.4	1 (0.8, 1.2)	Ref	4	0.5	1.0 (0.7, 1.2)	Ref	3	0.5	0.8 (0.5, 1.1)	
Non-cancer disease	0				0				0			
Non-disease	4	0.7	1.2 (0.9, 1.6)	1.2 (0.8, 1.7)	3	0.7	0.9 (0.6, 1.2)	0.9 (0.6, 1.3)	0			
Cancer	5	0.4	1.0 (0.8, 1.2)	na	4	0.5	1.0 (0.7, 1.2)	Ref	4	0.7	0.8 (0.6, 1.2)	na
Non-cancer disease	0				0				0			
Non-disease	4	0.7	1.2 (0.9, 1.6)		3	0.7	0.9 (0.6, 1.2)	0.9 (0.6, 1.3)	0			
Lung cancer	0				0			na	1	---	1.1 (0.4, 2.7)	na
All other	9	0.5	1.0 (0.9, 1.2)	na	6	0.7	0.9 (0.8, 1.1)		3	0.5	0.8 (0.5, 1.1)	
MAF												
MAF non- missing	9	0.5	1.0 (0.9, 1.2)	0.9 (0, 235.8)	7	0.7	0.9 (0.8, 1.1)	0.2 (0, 4.6)	4	0.7	0.8 (0.6, 1.2)	0 (0, 45050)

Table V.A.7. *XRCC3* Thr241Met and Smoking: Overall and by study characteristics (continued)

	Ever-never ¹				Current-Not current ²				PY ^{3,4}			
	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N	Q p- value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵
By Study Characteristic												
MAF (continued)												
MAF (cutpoints assigned by median across SNP)												
0.01-0.37	5	0.3	1.0 (0.8, 1.3)	Ref	2	0.1	0.9 (0.6, 1.4)	Ref	4	0.7	0.8 (0.6, 1.2)	na
>0.37-0.50	4	0.5	1.0 (0.8, 1.3)	1.0 (0.7, 1.4)	5	0.9	0.9 (0.7, 1.1)	1.0 (0.6, 1.7)	0			
MAF (cutpoints proxy for ethnicity)												
1%-15%	0			Ref	1	---	2.0 (0.6, 6.6)	na	0			na
>15%-50%	9	0.5	1.0 (0.9, 1.2)		6	0.9	0.9 (0.7, 1.1)		4	0.7	0.8 (0.6, 1.2)	
MAF-assigned ethnicity												
White	7	0.5	1.1 (0.9, 1.3)	Ref	5	0.9	0.9 (0.7, 1.1)	na	3	0.8	0.9 (0.6, 1.4)	na
African American	0				0				0			
Han	0				1	---	2.0 (0.6, 6.6)		0			
Multi-ethnic studies	2	0.2	1.0 (0.7, 1.3)	0.9 (0.7, 1.3)	1	---	1.0 (0.7, 1.5)		1	---	0.6 (0.3, 1.2)	
Smoking prevalence ¹³												
Continuous	9	0.5	1.0 (0.9, 1.2)	0.1 (0, 1.8)	7	0.7	0.9 (0.8, 1.1)	0.5 (0.1, 3.1)	4	0.7	0.8 (0.6, 1.2)	2.4 (0.1, 56.6)

Table V.A.7. *XRCC3* Thr241Met and Smoking: Overall and by study characteristics (continued)

	Table 4.11.1. ORs of FPE, FPE and Smoking: Overall and by Study Characteristics (continued)											
	Ever-never ¹					Current-Not current ²				PY ^{3,4}		
	Q				N	Q				N	Q	
N	p-value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N		p-value	OR _z (95% CI)	Ratio of ORs (95% CI) ⁵	N		p-value	OR _z (95% CI)
By Study Characteristic												
Smoking prevalence (continued) ¹³												
0-0.365	0				4	0.8	0.9 (0.7, 1.2)	Ref	1	---	0.8 (0.4, 1.4)	na
>0.365-507	0				2	0.1	1.1 (0.5, 2.3)	1.1 (0.5, 2.5)	1	---	0.6 (0.3, 1.2)	
>0.507-0.602	4	0.8	1.2 (0.9, 1.5)	Ref	1	---	0.8 (0.6, 1.2)	0.9 (0.6, 1.4)	2	0.9	1.0 (0.6, 1.7)	
>0.602-1	5	0.4	0.9 (0.7, 1.1)	0.8 (0.6, 1.1)	0				0			
>0-0.507	0			na	6	0.7	0.9 (0.8, 1.2)	na	2	0.5	0.7 (0.4, 1.1)	na
>0.507-1	9	0.5	1.0 (0.9, 1.2)		1	---	0.8 (0.6, 1.2)		2	0.9	1.0 (0.6, 1.7)	

Abbreviations: CI=Confidence interval, na=not applicable, HWE = Hardy Weinberg equilibrium, PY=pack-years, OR_z=control-only genotype-smoking odds ratio,), N=number of studies, Ref=referent, Q=Cochran's test of homogeneity, SNP=single nucleotide polymorphism, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ OR=Unadjusted odds ratio for *XRCC3* Thr241Met: Thr/Thr (ref) vs. any Met, never smoking (ref) vs. ever smoking, fixed effects

² OR=Unadjusted odds ratio for *XRCC3* Thr241Met: Thr/Thr (ref) vs. any Met, not current smoker (ref) vs. current smoker, fixed effects

³ OR=Unadjusted odds ratio for *XRCC3* Thr241Met: Thr/Thr (ref) vs. any Met, lightest non-zero smokers (ref) vs. heaviest smokers; stratified random effects

Table V.A.7. *XRCC3* Thr241Met and Smoking: Overall and by study characteristics (continued)

⁴ PY contrast is between lightest non-zero category of pack-years (ref) vs. heaviest category of PY

⁵ Ratio of Odds Ratios: Compares odds ratio in given study characteristic stratum to the odds ratio in the designated reference stratum for that study characteristic by meta-regression

⁶ Studies w 99%-100% of 1 ethnicity are classified as single-ethnicity

⁷ Only includes studies with explicitly stated ethnic makeup

⁸ White = Caucasian, white or non-Hispanic white; African American = African American or black; Han = Han, Han Chinese or ethnic Chinese

⁹ Median of studies included in *XRCC1* 751 current/not current smoker analyses.

¹⁰ Categories based on thirds from studies included in *XRCC1* 399 PY analyses ($\leq 59y$, $>59y-63y$, $>63y$)

¹¹ Median of all studies w age info (all SNPs) w age info [range:23.6-69y, mean: 56.7y SD: 9.8y]

¹² Median proportion male in all studies (all SNPs, all smoking exposures): 0.69

¹³ Smoking prevalence is contrast-specific (defined as "ever", "current" or "heavier PY" as appropriate)

Figure V.A.1: Weighted Forest Plot for *XRCC1* 399 and ever-never smoking

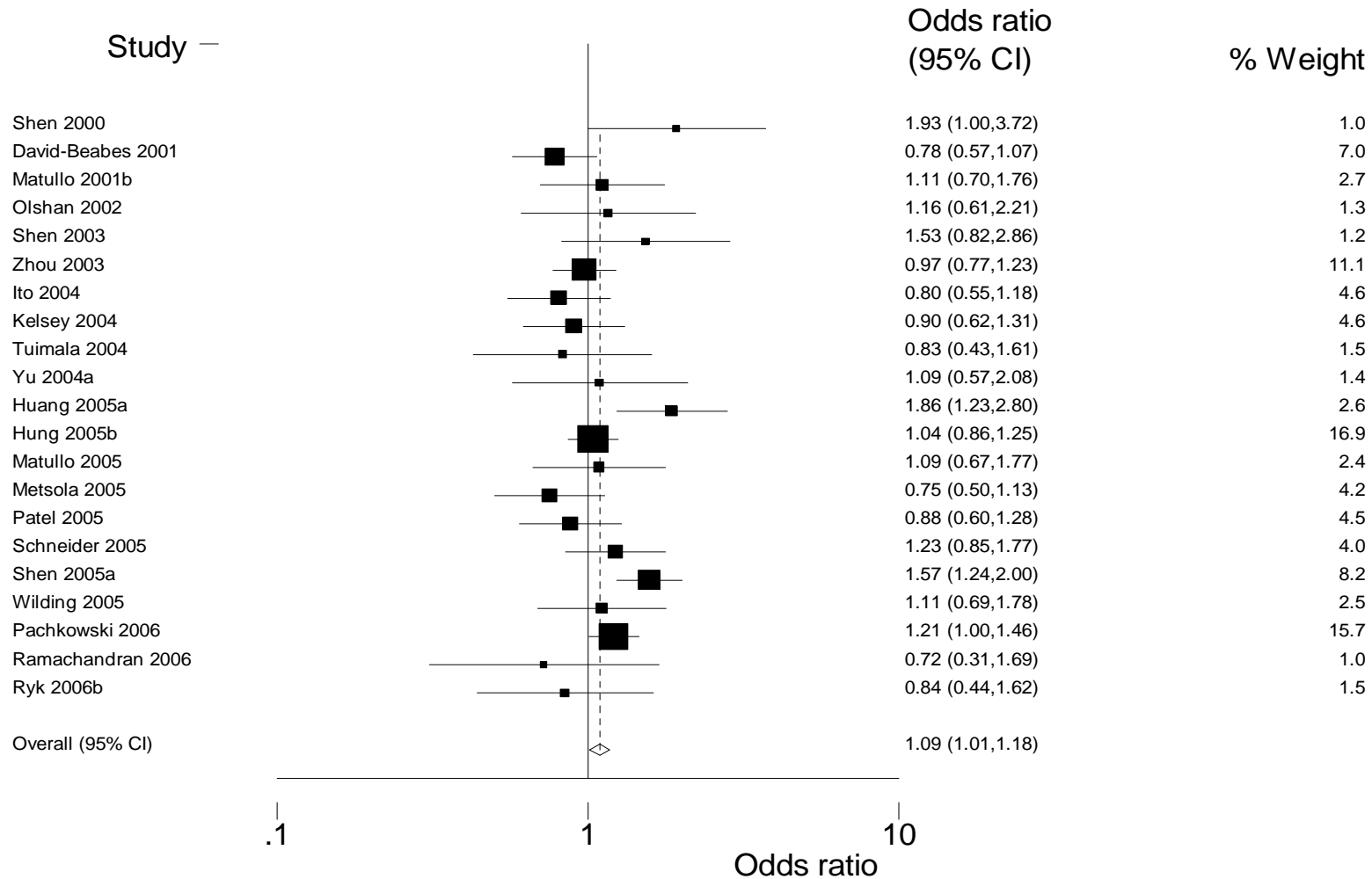
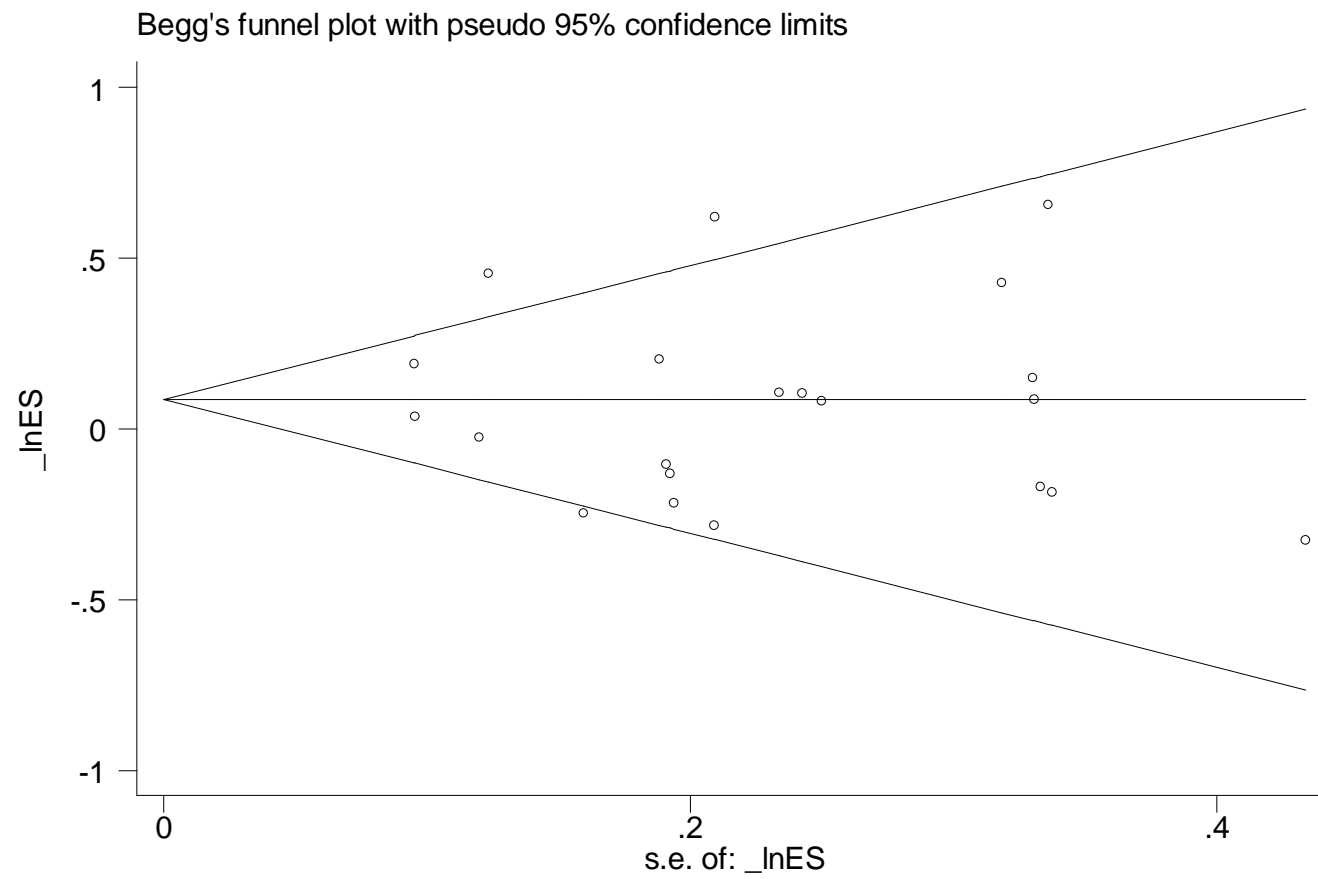


Figure V.A.2. Funnel plot for *XRCCI* 399 and ever-never smoking



B. MANUSCRIPT 2: Association of DNA repair and metabolic gene polymorphisms with tobacco smoking in controls from two population-based case-control studies: Carolina Breast Cancer Study and North Carolina Colon Cancer Study

1. Introduction

The case-only study design as proposed by Prentice et. al [1] and popularized by Piegorsch et. al. and Khoury et. al . [2-3] has 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. This method requires only cases, no population controls or defined cohort. Potential advantages of the design are reduced cost and increased precision [8]. Also, no invasive procedures are needed for healthy volunteers, especially in vulnerable populations (e.g. children) [9]. It has been proposed as a screening method to identify candidate gene-environment or gene-gene interactions and/or genes that may be etiologically important for further investigation [5, 16-17]. Because further investigation in more rigorous full-scale studies of genes identified in case-only studies requires significant additional money and time, it is important to evaluate the assumptions of case-only method.

Provided the design assumptions are met, in particular the independence assumption (i.e. that the genetic and environmental factors are independent in the population that produced the cases), the case-only study estimates statistical interaction that deviates from the null in a multiplicative model but not the independent effects of the genetic or environmental factors or their joint effects on the additive scale. When this design assumption is not met, bias is introduced into the case-only estimate of interaction (COR) [5]. Traditionally, case-control studies have been used to detect statistical interaction. The relationship between gene-environment interaction

estimated by the case-only odds ratio and the same gene-environment interaction estimated by a case-control study can be expressed as follows (*Equation 1*):

$$OR_{\text{gene*env, case-only}} = OR_{\text{gene*envr, case-control}} / (OR_{\text{gene, case-control}} * OR_{\text{envr, case-control}}) * Z$$

where Z is the association between the gene and the environmental exposure in the control group of a case-control study [3]. The quantity $[OR_{\text{gene*envr, case-control}} / (OR_{\text{gene, case-control}} * OR_{\text{envr, case-control}})]$ is sometimes referred to as the synergy index on a multiplicative scale, or SIM. When there is no association between the genetic exposure and the environmental exposure in the population (i.e. $Z=1$), the COR is equivalent to the deviation from a multiplicative relationship between the genetic and environmental exposures (i.e. $COR = SIM$). Using these abbreviations, the relationship can be expressed succinctly as (*Equation 2*):

$$COR = SIM * OR_z$$

where OR_z is the control only G-E odds ratio used to estimate Z, the underlying population G-E association.

Data simulations have demonstrated that even small violations of the independence assumption can strongly bias the case-only interaction parameter [5]. Using logistic models, Albert et. al. varied the magnitude of control group G-E association to explore the effect of independence assumption violation on case-only interaction estimates. As expected from *Equation 2*, as values of OR_z increased above the null, the COR was increasingly and proportionally biased away from the SIM. Using data from a study of *XRCC1* genotype and lung cancer by Ratnasinghe et. al., Albert et. al. showed empirically that the magnitude of OR_z equaled the magnitude of bias introduced into the COR relative to the SIM due to violation of the

independence assumption violation: $OR_z=2.03$ for genotype and pack-years of tobacco use created a bias in the COR of 105% relative to the SIM [$COR=0.90$ (0.41,1.94), $SIM=0.44$ (0.17, 1.16)] [126]. In another example, an OR_z of 1.2 representing the association between genotype and alcohol drinking status (ever/never) biased the COR by nearly 30%, which exceeds a commonly used threshold for an acceptable level of confounding bias (10%). Further, violations of the independence assumption may cause the Type II error rate (false negative) to be high. When control-group G-E associations are of similar magnitude but opposite in direction to the interaction effect, a case-only study may fail to detect interaction effects [5, 124]. Because the case-only study has been suggested as a useful screening tool to identify candidate genes for further investigation, a high Type II error rate would be problematic. Little work has been done to explore this possibility.

Although the validity of case-only estimates rests heavily on the independence assumption, and case-only studies, particularly stand-alone case-only studies, have some advantages over traditional study designs for interaction analysis, the literature specific to control group associations of interest for interaction studies is scant. In the traditional population-based case-control study cases and controls are sampled from the same underlying population. However, many investigators use data from a different population than the cases came from to evaluate the independence assumption. The assumption that $Z=1$ can only be evaluated if both exposures have been measured in the same underlying population at risk or, in the context of a case-control study drawn from the population at risk, estimated by OR_z in the controls, [2] or finally, by a suitable proxy for either of the preceding groups.

The current study aims to address at least one of the gaps in the existing literature, by contributing results from large population-based control associations. We explored gene-smoking

control group associations in two population-based case-control studies, 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 in DNA repair genes (repair of genetic damage from smoking), xenobiotic metabolism genes (activation of procarcinogens and excretion of toxic intermediates), and cell cycle control genes. Both studies oversampled African Americans, and the NCCCS has both male and female participants, so issues of effect modification and/or confounding by age, race and gender were addressed. Finally, all genes were grouped by the function of the gene pathway they participated in, and any patterns by pathway were noted.

2. Methods

Study populations

CBCS and NCCCS

Population-based controls from the CBCS and the NCCCS were used to estimate OR_z for gene-smoking associations. The Carolina Breast Cancer Study and the North Carolina Colon Cancer Study are population-based case-control studies conducted in central North Carolina during the mid- to late 1990's which included urban, suburban and rural areas (CBCS: $N_{cases}=2311$, $N_{controls}=2022$; NCCCS: $N_{cases}=646$, $N_{controls}=1053$) [175, 177-180]. CBCS controls were pooled controls from Phase I ($N=790$), Phase II ($N=774$) and the Carcinoma in situ ($N=458$) study. The CBCS controls were not pooled with NCCCS controls. Both studies over-sampled African Americans. CBCS controls are all female; NCCCS controls also include male participants. Potential 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 [181]. CBCS participants were

approximately half African American and half <50 years of age. NCCCS participants were sampled such that the race, age and gender distribution of randomly selected cases is approximately 1:1 for gender and race. The CBCS and NCCCS used similar questionnaires and both have extensive data on tobacco smoking history.

A sample of polymorphisms was chosen from available genotype data in the CBCS and NCCCS based on potential relevance to smoking behavior and/or other smoking-related health effects. Genes selected from the CBCS were 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*, *XPB*, *XPD*, *XPF*, *XPG*), oxidative stress defense genes (*MnSOD*, *MPO*, *NQO1*), a cell adhesion gene (*CDH1*) and a growth factors gene (*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 an oxidative stress defense gene (*MnSOD*). Methods of collection and genotyping have been described previously [68, 171, 193-203, 268].

Statistical methods

Hardy Weinberg equilibrium was tested at $\alpha=0.05$ for all polymorphisms except *GSTM1*, *GSTT1*, *NAT1* and *NAT2*. Estimates of OR_z and 95% confidence intervals were generated using logistic regression with a dichotomous representation of the genetic variable (homozygous for common allele=referent [G-], heterozygous + homozygous for less common allele=exposed [G+]) as the dependent variable. A single model of the general form $\text{logit}(G+/G-) = \alpha + \beta_{(1)}E_1 + \beta_{(2-i)}$

$COV_{(2-i)} + \text{error}$ (where G+= positive for genetic variant, E+=positive for the smoking behavior, COV=any additional covariates) was used for all SNPs. Those homozygous for the most common allele (“no variant”) were the referent group (G-) and were compared to heterozygotes plus homozygotes for the less common allele (G+, “any variant”).

In the CBCS and NCCCS smoking status was categorized as ever, former or current smoker. Three measures of smoking dose 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 were derived from categorical variables used for packs/day and years smoked (pack-years are 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 dataset was evaluated for OR_z effect measure modification using stratification on race (white, African American), age (CBCS: <50y, ≥ 50y; NCCCS: <65y, ≥65y) and gender (NCCCS only), respectively. Based on directed acyclic graphs [209], 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. In order to decide whether to stratify analyses on race, a likelihood ratio test was performed comparing models with and without a race*smoking interaction term. Significant results for the interaction term ($\alpha=0.05$) in a majority of smoking measures precluded pooling African American and non-African American participants for that SNP. Because sample size was often low for African Americans, crossover (OR_z s on opposite sides of the null) was also examined to better characterize any differences by race.

Although matching procedures were based on projected case incidence, and no cases were used in the current analysis, the matching process distorted the prevalence of these factors in the underlying population, potentially affecting gene-smoking estimates; consequently we adjusted for

all matching factors (race, age and gender). Based on DAGs, two additional variables were evaluated as potential confounders: first degree family history of any cancer, excluding non-melanoma skin cancer (Y/N) and total family income (<\$15K, >=\$15K-<\$30K, >=\$30K-<\$50K, >=\$50K). Percent change in β coefficients was calculated but not used to determine whether a covariate would be retained in the model; because of the high proportion of estimates close to the null (Range in CBCS: 0.5-2.5, NCCCS: 0.6-1.6) this commonly used criterion was not sufficiently informative. A potential confounder was retained if the absolute value of difference between smoking variable β coefficients from models with and without the potential confounder was > 0.15 (i.e. when $|\beta \text{ coefficient for smoking from model with potential confounder} - \beta \text{ coefficient for smoking from model without potential confounder}| > 0.15$ covariate is retained). For consistency, if a covariate met this criterion for any polymorphism-smoking estimate, it was retained in all models. After assessment of effect measure modification and confounding, an association was characterized by magnitude of OR_z (odds ratios ≥ 1.4 or < 0.7 were considered evidence of non-null association) and precision of the accompanying confidence interval. Estimates with confidence limit ratios > 4 (CLR, upper CI limit/lower CI limit) were excluded from consideration unless otherwise stated. SAS 9.1 was used for all modeling [210].

After assessing the CBCS and NCCCS datasets separately, agreement between the two studies was assessed for the 15 polymorphisms included in both studies using a weighted kappa statistic [211]. The weighted kappa measures the degree of agreement between two or more raters that are using a multi-level ordinal scale to categorize a series of subjects, beyond what would be expected due to chance alone. Here the raters were the CBCS and NCCCS, and the “subjects” of agreement were the 15 gene-smoking associations measured by both studies. OR_z was categorized into three categories: 1) below the null, $OR_z < 0.9$, 2) null, $0.9 \leq OR_z \leq 1.1$ and 3) above the null,

$OR_z > 1.1$. With the weighted statistic disagreement between 2 adjacent categories is considered less important than disagreement between ratings further apart on the ordinal scale. Because distributions of race, age and gender differed across the two studies, restricted datasets were created and compared. These datasets were restricted to white women 40-74 years of age. Due to reduced sample size in the NCCCS restricted dataset precision requirements were relaxed (estimates with CLR < 5 were included) to provide sufficient estimates for a comparison of most polymorphisms and several major smoking behaviors. SAS 9.1 was used to calculate weighted kappa statistics [210].

Misspecification of smoking exposure occurs when the independence assumption is evaluated with one measure of smoking (e.g. ever-never) but the case-only analysis is performed on a different measure of smoking (e.g. duration of smoking). Any difference in OR_z between smoking measures leads to undetected bias in the COR. We examined the frequency of differences in OR_z s for smoking status (ever-never, current-not current smoking) and measures of smoking amount (duration, intensity and PY). We also compared the consequence of using the p-value for OR_z (95%CI) vs. using the magnitude of OR_z as a decision tool when evaluating the independence assumption.

3. Results

The study populations used in this analysis were drawn from largely overlapping source populations of white and African American residents of central and eastern North Carolina. Although the underlying source population is essentially the same, the two population-based study populations varied substantially by gender and sample size due to sampling criteria (Table V.B.1). CBCS cases in Phases I and II were diagnosed with invasive breast cancer and CIS cases had breast carcinoma in situ. The invasive study oversampled women < 50 years of age and African

American women; CIS did not. Controls were frequency-matched by race and age (+/- 5 years) to the respective case groups. Response rates were 55% overall for Phases I (1993-1996) and II (1996-2001), and 65.2% for CIS (1996-2001) [178, 193, 269]. The overall response rate for the pooled CBCS controls was 57% (N=2022). The response rate for DNA samples was 90%. Prevalence of current smoking was similar across CBCS control subgroups (Phase I: 21%, Phase II: 19%) and the CIS (17%). CIS controls were slightly older than invasive study controls (Median age: Phase I 49y, Phase II 50y, CIS 53y). Controls from the NCCCS were older than CBCS controls and included both men and women. Consistent with gender and age differences in smoking prevalence in the US [270-271], there were a higher proportion of never smokers and a shorter average smoking duration in CBCS controls compared to the NCCCS.

Table V.B.2 provides the rs# and official name for each SNP included in the analysis, as well noting the most common allele for each in the CBCS and NCCCS datasets. The common allele for the full dataset was used as the referent even when the common allele differed by race. SNPs where the common allele differed by race are noted in Table V.B.2. In the CBCS, 38 polymorphisms in 29 genes were evaluated; 17 genes were DNA repair genes. In the NCCCS, 25 polymorphisms and four haplotypes from 19 genes were evaluated. Fifteen genes were DNA repair genes. For the 15 polymorphisms included in both studies, two were in metabolic genes, 12 were in nine DNA repair genes, and one was in an oxidative stress gene.

Allele frequencies and HWE p-values for CBCS and NCCCS controls, stratified by race, are presented in Table VIII.B.1. Within race only four SNPs (3%) were out of Hardy Weinberg equilibrium ($\alpha=0.05$), 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. HWE can not be calculated

for *GSTM1*, *GSTT1*, *NAT 1* and *NAT2* because these polymorphisms are categorized by enzymatic activity (present or absent [null] for *GSTs*; rapid or slow for *NATs*) rather than as discrete alleles.

Percent ‘any variant’ was consistent between the CBCS and NCCCS within race; the sole exception was *GSTT1* (CBCS: 16.4% and 16.6% null, in non-African Americans and African Americans, respectively; NCCCS: 29.6% and 33.3% null, in non-African Americans and African Americans, respectively). Tables 4a-d and 5a-d 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 same. All OR_z sufficiently precise for evaluation ($CLR < 4$) were between 0.4 and 2.5.

All models were adjusted for matching variables (race, age and gender) unless stratified or restricted by same. Approximately half of the polymorphisms showed joint confounding by race and age (difference of $|>0.15|$ in β coefficients), almost entirely former smoking and/or >35 PY in the CBCS, although no absolute difference in β coefficients exceeded 0.4 for OR_z s w $CLR < 4$. Unadjusted and race-, age-, and gender-adjusted estimates did not vary substantially in the NCCCS. Confounding by race and age were more marked in measures of smoking dose than smoking status, but did not vary by functional gene category. Based on directed acyclic graphs [209], family history of any cancer and family income were identified as potential confounders, but neither changed estimates substantially in either dataset. They were not included in any models.

CBCS

In the CBCS overall, three SNPs showed consistency across smoking categories with moderate OR_z s (defined as an $OR_z \geq 1.4$ or ≤ 0.7) in at least one smoking status category (ever, former or current) and at least one smoking dose category (duration, intensity or pack-years):

CYP1A1 M2, *GSTP1*, and *XPF* 662 (Tables 4a-d). An additional five SNPs had two or more moderate OR_z s in any smoking category: *COMT*, *CDH1*, *XRCC1* 194, *BRCA2* 372, and *MGMT* 84. Finally, two SNPs, *CYP1A1* M4 and *ERCC6* 1213, showed moderate OR_z s in more than one level of a single measure of smoking behavior (e.g. moderate OR_z s in two levels of smoking duration but no other measures of smoking). Only estimates with a $CLR < 4$ were considered precise enough for evaluation.

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 positively associated with smoking status and <35 PY (vs. never). No other smoking categories were evaluable for *CYP1A1* M2 due to low precision. *GSTP1* was positively associated with former, short duration, moderate intensity and low PY of smoking but inversely associated with current smoking and high PY of smoking. *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 measures of 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.

Within smoking categories, one SNP showed a moderate magnitude OR_z for ever smoking (*CYP1A1* M2); one 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 moderate magnitude OR_z s for

<10yrs and three others had moderate magnitude OR_z s for 11-20 years (*BRCA2* 372, *XPF* 662, and *CDH1*). 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 nor did any strong patterns by race emerge. 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 $\alpha=0.05$, 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.

To further highlight more extreme differences, we examined crossover (OR_z s on opposite sides of the null) to evaluate effect measure modification by race. Seven SNPs showed substantial variation by race in at least one smoking category (*GSTT1*, *COMT*, *XRCC1* 194, *NBS1* 185, *XRCC4* -28073, *ERCC1* 8092 and *NQO1*). *GSTT1* and *COMT* were inverse in whites and positive in African Americans. For SNPs that varied by race, the direction of association for each stratum was consistent across smoking categories. When estimates were stratified by age (< 50 yr, ≥ 50 yrs) there was minor variation; it was consistently less than variation by race.

Misspecification of smoking exposure (i.e. using status to evaluate the independence assumption then conducting a case-only analysis of a smoking amount measure) strongly affected the frequency that bias would be introduced into the COR. In the CBCS, for smoking status, there were four SNPs with positive moderate magnitude OR_z s (*CYP1A1* M2 for ever smoking; *CYP1A1* M4, *ERCC6* 1213, and *XPF* 662 for current smoking). For smoking amounts (duration, intensity

or PY), nine had positive moderate magnitude OR_z s (Table V.B.9). However, of these nine SNPs, only one also had a positive moderate magnitude OR_z s for smoking status (*CYP1A1* M2). One SNP with an inverse moderate magnitude OR_z for smoking status (*GSTP1* and current smoking) showed largely inverse moderate magnitude OR_z s for smoking amounts. None of the other six SNPs with inverse moderate magnitude OR_z s for any measure of smoking amount had moderate magnitude OR_z s for smoking status.

Using the magnitude of OR_z as an indicator of independence assumption violation identified more instances that bias would be introduced into the COR than using significance testing. Of the 22 positive moderate magnitude OR_z s in the CBCS, nine were statistically significant at $\alpha=0.05$. There were three statistically significant positive OR_z s of smaller magnitude. There were 11 inverse OR_z s of moderate magnitude, six of which were statistically significant. One smaller magnitude OR_z was statistically significant.

NCCCS

In the NCCCS controls, using the same criteria for moderate magnitude association as listed for the CBCS, 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 5a-d). *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. Weak and/or suggestive associations were found for *XPC* 939, *XRCC1* 194, *XRCC3* 241, *XPB* 751 and *MSH6* 39. As in the CBCS, metabolism genes were overrepresented in the stronger associations and DNA repair genes in the weaker associations. Associations between minor alleles for metabolism SNPs and smoking were

consistently inverse whereas association between smoking and DNA repair SNPs were both positive and inverse.

Within smoking categories, three of the four metabolism gene SNPs (*GSTM1*, *MEH* 113 & 139) were inversely associated with smoking status (ever, former or current smoking); three DNA repair SNPs showed moderate magnitude inverse OR_z s for smoking status (ever & current smoking: *POLD* 119, current smoking: *MSH3* 940 and *MSH6* 39). Short and moderate duration smoking showed both positive and inverse OR_z s, whereas the OR_z for smoking >20 yrs was near the null for all SNPs except *MEH* 139. All measures of amount (duration, intensity and PY) showed some clustering of positive associations in MMR and NER DNA repair genes, and inverse associations for metabolic genes and BER DNA repair genes. Only one SNP (*POLD* 119) showed an association with low PY, although there were six 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 the data were stratified by gender, estimates for ever smoking were slightly more likely to be positive or more strongly positive for women than for men, although this was not true for other measures of smoking status. Results were similar for smoking duration >20y, however low precision in estimates for short and moderate duration meant few comparisons across gender could be made. For smoking intensity, stratification by gender produced associations on opposite sides of the null more often than for other smoking categories although precision limited comparisons for the heaviest smokers. For low PY, estimates for women were again more likely to be positive, or more strongly positive, than estimates for men. High PY could not be evaluated. Only *MSH3* 940 differed significantly by gender across more than one smoking measure. OR_z s for ever, 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 at $\alpha=0.05$.

No strong patterns emerged after stratification by race, although estimates were often on opposite sides but still 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). Results for duration and high PY were generally not evaluable by race due to poor precision. Approximately 3% of the likelihood ratio tests for a race-smoking interaction term were significant at $\alpha=0.05$, about what would be expected by chance. There was no pattern of significant interaction by race for any given smoking measure. *GSTT1* was the only gene with more than one statistically significant (at $\alpha=0.05$) race*smoking interaction term: ever smoking, never/former/current smoking and PY; it was generally positive for African Americans and inverse for non-African Americans. Where evaluable, stratification by age (<65y, ≥ 65 y) yielded OR_z s that were more similar across strata than gender- or race-stratified OR_z s.

The effect of misspecification of the smoking variable in the independence assumption was assessed as in the CBCS (Table V.B.9). In the NCCCS, for smoking status, there were six SNPs with moderate magnitude OR_z s, all inverse (*MEH* 113, *POLD* 119 with ever smoking, *GSTM1* null, *MEH* 139, *MSH3* 940, and *MSH6* 39 with current smoking). Three of the four *GST* haplotypes showed moderate magnitude OR_z s with smoking status. For smoking amount, there were 11 SNPs or haplotypes with positive moderate magnitude associations; only 1 showed a similar result for smoking status. Only half of the SNPs with inverse moderate magnitude OR_z s for any smoking amount measure category also had inverse moderate magnitude OR_z s for smoking status (5 of 9).

In the NCCCS, there were 22 positive moderate magnitude OR_z s; three were statistically significant. Of the 35 inverse moderate magnitude OR_z s, six were statistically significant. There were no statistically significant OR_z s between 0.7 and 1.4.

CBCS and NCCCS

Comparing CBCS and NCCCS results for the 15 SNPs measured in both studies (Table V.B.6), no SNP had $OR_z \geq 1.4$ or ≤ 0.7 in both studies. With the null defined as between 0.9-1.1 (inclusive), the weighted kappa for agreement was -0.07 (95% CI: -0.19, 0.06), indicating slight disagreement (Table V.B.10) [212]. When CBCS and NCCCS datasets were restricted to white women 40-74 years of age to improve comparability (Table V.B.7), results were only evaluable in the NCCCS for ever, former, long duration, moderate intensity and low PY of smoking, for 13 or fewer SNPs, even with the limits for precision relaxed to include estimates with $CLR < 5$. Under these conditions, the kappa for agreement was 0.22 (95% CI: -0.01, 0.46), usually considered slight agreement [212]. Changing the definition of the null to 0.8-1.2 (inclusive) or including all data regardless of CLR did not change results.

4. Discussion

The aim of the current project was to assess the magnitude of associations between a convenience sample of SNPs in two population-based control groups and multiple measures of smoking behavior. The primary motivation was to evaluate any gene-smoking associations in light of the bias that would be introduced into a case-only analysis of gene-environment interaction when the independence assumption is violated.

Odds ratios for the control groups (OR_z s) in the current study were of moderate magnitude [≥ 1.4 or ≤ 0.7] in at least one of the six smoking behavior measures for approximately half of the SNPs examined in each of these population-based control groups (CBCS: 45%, NCCCS:

59%). This analysis 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 $\geq 40\%$. This is a substantive degree of bias in most contexts and could easily mislead researchers into concluding G-E interaction exists when it doesn't or that G-E interaction is much stronger than it actually is. Alternatively, G-E interaction may be missed completely when the SIM is inverse and the OR_z is positive. The converse is true for $OR_z \leq 0.7$. These moderate magnitude OR_z s were found across all functional categories of putative gene function. For most DNA repair gene SNPs, particularly BER and DSB genes, both studies showed a preponderance of moderate magnitude OR_z s in categories of smoking dose (pack/day, years smoked or PY) rather than smoking status (ever, former, current). In contrast, metabolic gene SNPs had moderate magnitude OR_z s in both status and dose measures. There were too few SNPs in other functional categories to observe any patterns.

Metabolic genes and smoking behavior

There is an extensive epidemiologic literature on smoking and metabolic genes, [i.e. those coding for enzymes that metabolize nicotine or other tobacco smoke constituents such as polycyclic aromatic hydrocarbons (PAH)] [145-146]. Variation in these genes can alter enzyme activity, regulation or expression [144] plausibly increasing or decreasing risk of disease or influencing smoking behaviors, such as the number cigarettes consumed daily or years as a smoker. Of the seven metabolic genes included the CBCS data, five (*CYP1A1*, *GSTM1*, *GSTP1*, *NAT1* and *COMT*) showed moderate association in at least one measure of smoking. Only *CYP1A1* was moderately associated with ever smoking. In the NCCCS, all three metabolic genes (*GSTM1*, *GSTT1*, and *MEH*) showed moderate association with at least one measure of smoking.

The *COMT* Val158Met SNP (rs4680) is the only SNP in the current study that has been extensively studied with respect to its possible influence on smoking behavior, [147]. Results have been equivocal with two recent large population-based European studies coming to different conclusions [148-149]. Omdivar et. al. found a 20% reduction in incident smoking cessation for carriers of the low activity form of the 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=0.9 and 0.5 for ≤35PY and >35PY; OR_z=1.0, 0.8 and 0.8 for <10y, 11-20y and >20y, 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] [150]. 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 less smoking (OR_z for women=1.6, 0.7 and 1.0 for <1/2 pack/day, 1/2-1 pk/day and >1 pk/day, respectively). Findings for *GSTP1*, *GSTT1* and *MEH* have not been reported previously.

DNA repair genes and smoking

Studies that have examined DNA repair genes and smoking behavior are scarce. The population-based candidate gene study of habitual smoking by Lui et. al included several DNA repair genes in addition to the metabolic genes discussed earlier [64]. Again, OR_zs were presented only for statistically significant DNA repair gene SNPs. Only one was in the current study, *OGGI* [OR_z =0.6 (0.4, 1.0) for ever smoking]. There was no association with ever smoking for the *OGGI* SNP in the CBCS [OR_z =1.0 (0.9, 1.3)].

A recent meta-analysis of gene-smoking association assessed *XRCC1* 399, 194 and 280. Several of the summary OR_z s were of moderate magnitude ($OR_z \geq 1.4$ or ≤ 0.7): *XRCC1* 194 and longer duration $OR_z = 0.7$ (0.5, 0.9), *XRCC1* 280 and current smoking $OR_z = 0.7$ (0.5, 1.1), and *XRCC1* 399 and greater intensity $OR_z = 1.5$ (1.2, 1.9). Summary estimates included CBCS data; however, the control group data from the other included studies were consistent with CBCS results [224, 233, 260, 262, 272-275]. Findings for the other DNA repair SNPs in the CBCS and NCCCS [*XPF*, *MSH3* (stratified by gender), and *POLD1*] have not been reported previously.

Two SNPs in the current project varied strongly by race, *NQO1* and *GSTT1*. For *NQO1*, an oxidative stress response gene, OR_z was consistently positive in non-African Americans, and inverse among African Americans, notably for current smoking (OR_z for non-African American=1.4, African American=0.7) and smoking >20 y (OR_z for non-African American=1.4, African American=0.6). *NQO1* is thought to be a susceptibility factor for coronary heart disease, and cancer, particularly with environmental exposures such as smoking and benzene, respectively [276-277]. In contrast, *GSTT1*, a Phase 2 metabolic gene, was generally inverse for non-African Americans and positive for African Americans. A number of the relevant exposures (e.g. benzene, pesticides, quinone-based chemotherapy) may differ by race or SES, and could plausibly be related to changes in smoking behavior that vary by race.

Smoking behavior in controls

Several SNPs in the CBCS and NCCCS stood out with moderate magnitude OR_z s in at least one status category and at least one level of a dose measure (pack/day, years smoked, PY). 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 had inverse OR_z s, 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. In the NCCCS several genes also showed weaker signals: *XPC* 939, *XRCC1* 194, *XRCC3* 241, *XPD* 751 and *MSH6* 39.

Even given the wealth of smoking behaviors, genetic variations and populations studied, and the biological plausibility of smoking behavior being influenced by toxic intermediates in xenobiotic metabolism pathways, it is difficult to find studies of smoking in controls or population samples to compare with the current study [54, 64, 125] (Table V.B.8). Smits et. al. used pooled control group data from the International Collaborative Study on Genetic Susceptibility to Environmental Carcinogens (GSEC) to estimate OR_z s 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) (Table V.B.8). 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. Results were adjusted for study, age, sex and ethnicity. Results for these five genes and smoking status were most often at or near the null. Overall, they were broadly similar to CBCS and NCCCS results, even though controls pooled across multiple studies would not necessarily be expected to have an OR_z similar to that of any given study. Despite this, there were 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 just above the null, but in the CBCS it was below the null. For female GSEC controls, the OR_z was similar to the CBCS, but the OR_z for non-hospital controls was above the null. GSEC and CBCS

OR_zs for *GSTP1* were even further apart for former smokers (Table V.B.8). Variation in OR_z between pooled controls from small to moderate sized studies (GSEC) and the two relatively large population-based control groups in the current study, as well as the variation between subgroups in the pooled controls, suggest that the OR_z should be considered specific to each underlying population rather than an estimate of some ‘universal’ OR_z for that SNP and smoking measure. 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 (N=339), Lui et. al. examined a panel of 153 SNPs in 40 candidate genes potentially involved in tobacco consumption in a sample of Japanese men 40-49 years of age [64]. Lui 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_z for *MEH* was consistent with NCCCS although the specific SNPs were different: *MEH* rs2292566 [64]: OR_z=0.4 (0.2, 0.8)]; *MEH* 113 & 139 (NCCCS): 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. el. [125] calculated OR_z(95%CI) in four published control groups [278-281] for ever smoking and SNPs in *CYP2E1*, *NAT2*, and *CYP1A1*. None of the OR_zs were significant at $\alpha=0.10$, 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_z was the amount of bias introduced into the COR, they concluded, on the basis of statistical significance, that these SNPs could be used with smoking in a case-only study of interaction.

In the literature, quantitative approaches have ultimately relied almost exclusively on statistical significance regardless of whether the independence assumption was being evaluated in a control group from the same study population as the cases or in ancillary data (i.e. data external to the published study). For instance, in Egan et. al. (2003) the magnitude of gene-environment associations varied from 0.5 to 1.1, and in Marcus et. al. (2000), it was 0.5 to 1.8, nonetheless the only associations considered problematic were the statistically significant ones [21, 52]. 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 [214].

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 (ever, former, current, cig/day [3 level], years smoked [3 level], PY [2 level]) for approximately half of the SNPs examined in these population-based control groups (CBCS: 45%, NCCCS: 59%). For most functional categories except metabolism gene SNPs, moderate magnitude OR_z s were most often found for measures of smoking dose (cig/day, years smoked, PY) rather than smoking status (ever, former, current). These results need to be replicated in other population-based control series or other relevant samples.

Nonetheless, some implications for the conduct of case-only studies are clear. Smoking status measures are more easily extracted from the published literature than measures of smoking amount. Consequently, ever-never and current-not current smoker are most often used to check the independence assumption (Hodgson in preparation). Results from the current study show that the magnitude of OR_z is not reliably close to the null for many of these SNPs, making them

unsuitable for a case-only interaction analysis. These results also clearly show that for many SNPs evaluating the independence assumption using smoking status is insufficient evidence of no association for measures of smoking amount such as duration, intensity and PY, the measures of interest for many case-only analyses. Very few SNPs with moderate magnitude OR_z s in any category of smoking amount had comparable magnitude OR_z s 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 association in the CBCS controls being missed and around 80% of the moderate magnitude OR_z s in the NCCCS being missed. This was observed across all gene categories in both control groups.

Strengths and Limitations

The primary 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 that underlies the sample of cases. Using a control group rather than a population sample meant that the true parameter (RR_z) could only be estimated; OR_z was a proxy for RR_z . However, OR_z is the information most easily available in the literature, and most often used to evaluate the independence assumption, making it the most relevant measure to examine to inform the practice of case-only study design.

We were able to use individual level data such as race, gender and age to check for potential effect measure modification and confounding, something not generally possible when checking the independence assumption using the published literature. Genotype prevalence varies by race for many metabolic and DNA repair genes; smoking behaviors vary by race, gender and age [270-271, 282-283]. Consequently it is important to be able to address the effect of race, gender and age on the gene-smoking association when evaluating the independence assumption.

Both studies had information on smoking intensity, duration and PY, often the exposures of interest in a case-only interaction analysis, but not often available in the published literature, at least for controls. Both the CBCS and NCCCS oversampled African Americans making subgroup analyses by race feasible for most SNPs. The CBCS and NCCCS are drawn from essentially the same underlying population: largely overlapping geographic areas, during approximately the same time period, 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 two different cancers, there were a limited number of SNPs included in both studies. A further limitation was that for African American women 40-74 years of age in the NCCCS, very few SNPs and smoking measures meet our precision criteria thus it was not possible to assess agreement between the two studies 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. Bias due to nonparticipation by smoking status alone may be non-differential with respect to the gene-smoking association because potential participants are unaware of their gene status. However if participation rates also vary by family history (or any proxy for $G+$), OR_z would be driven away from the true OR_z in an unpredictable direction, depending on the participation rates of smokers with or without a family history (e.g. if smokers with a family history of cancer refuse participation more often than other groups, a true positive OR_z could be driven downward, even below the null, but if the non-participation rate in smokers with no family history is even higher the OR_z will increase away from the null). However, the population prevalence of current smoking in the CBCS (20%) was similar to NC women in the 2001 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%) [195, 271], arguing that selection bias due to the joint distribution of smoking and gene status is likely to be small.

The precise biological functions of most of the SNPs in this study were unknown, limiting causal interpretations of any associations found. Population stratification could have caused some residual confounding despite adjustment for self-reported race. Any 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 the one exception noted (*NQOI*), 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. If the SNP-smoking associations in the control groups were due entirely to chance, agreement would not be expected to improve solely due to restriction by race, age and gender.

Conclusions

Our findings show that the gene-smoking OR_{2s} in population controls are often of sufficient magnitude that these associations would produce unacceptable bias in the COR in a case-only study of GxE interaction. Thus, caution is warranted when using the case-only method. A stand-alone case-only study should be conducted only when the independence assumption can be verified with appropriate empirical data. Appropriate data means either population-specific data or, if sufficient published data are available, OR_{2s} within a narrow, pre-specified range of acceptable bias, across a wide variety of population-based studies. This data is needed for every smoking metric that 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 metrics (duration, intensity, etc.) than is usually presented, ideally stratified by race and

gender. Given that many studies already collect much 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. Other exposures whose effect might be modified by genetic variation (e.g. air pollution, infectious diseases, alcohol consumption, chemotherapeutics) should also be examined.

5. Tables and Figures

Table V.B.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 ¹	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	

Table V.B.1. Characteristics of CBCS and NCCCS control groups (continued)

	Full CBCS and NCCCS				Non-African American women, 40-74 y			
	CBCS		NCCCS		CBCS		NCCCS	
	N	%	N	%	N	%	N	%
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-years²								
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	

Abbreviations: CBCS=Carolina Breast Cancer Study, NCCCS=North Carolina Colon Cancer Study, SD=standard deviation,

N=number of controls

¹ Participants reporting non-African American race (98% white for CBCS, 98.9% white in NCCCS)

² Smokers only

Table V.B.2. Gene variants in CBCS and NCCCS

Gene & codon/ nucleotide position	rs#	Common ¹ allele (amino acid)	Variant ¹ allele (amino acid)	Nucleotide common/ variant	Gene name and official abbreviation ²	Study
<i>ADPRT</i> 762	rs1136410	Val	Ala	T/C	poly (ADP-ribose) polymerase 1 [<i>PARP1</i>]	NCCCS
<i>ADPRTL2</i> 328 ³				C/T	poly (ADP-ribose) polymerase 2 [<i>PARP2</i>] APEX nuclease (multifunctional DNA repair enzyme) 1 [<i>APEX1</i>]	NCCCS
<i>APE1</i> 148	rs1130409	Asp	Glu	T/G		Both
<i>BRCA2</i> intron 24	rs206340	--	--	G/A	breast cancer 2, early onset [<i>BRCA2</i>]	CBCS
<i>BRCA2</i> 372	rs144848	Asn	His	A/C	breast cancer 2, early onset [<i>BRCA2</i>]	CBCS
<i>CDH1</i> -160	rs16260	--	--	C/A	cadherin 1, type 1, E-cadherin (epithelial) [<i>CDH1</i>]	CBCS
<i>COMT</i> 158 ⁴	rs4680	Val	Met	G/A	catechol-O-methyltransferase [<i>COMT</i>]	CBCS
<i>CYP1A1 M1</i> (<i>CYP1A1</i> *2A)	rs4646903	(*1A)	(*2A)	T/C	cytochrome P450, family 1, subfamily A, polypeptide 1 [<i>CYP1A1</i>]	CBCS
<i>CYP1A1 M2</i> (<i>CYP1A1</i> *2C)	rs1048943	Ile (*1A)	Val	A/G	cytochrome P450, family 1, subfamily A, polypeptide 1 [<i>CYP1A1</i>]	CBCS
<i>CYP1A1 M3</i> (<i>CYP1A1</i> *3)	rs4986882	(*1A)	(*3)	T/C	cytochrome P450, family 1, subfamily A, polypeptide 1 [<i>CYP1A1</i>]	CBCS
<i>CYP1A1 M4</i> (<i>CYP1A1</i> *4)	rs1799814	Thr (*1A)	Asn	C/A	cytochrome P450, family 1, subfamily A, polypeptide 1 [<i>CYP1A1</i>]	CBCS
<i>ERCC1</i> nt8092	rs3212986	Gln	Lys	C/A	excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense sequence) [<i>ERCC1</i>]	CBCS

Table V.B.2. Gene variants in CBCS and NCCCS (continued)

Gene & codon/ nucleotide position	rs#	Common ¹ allele (amino acid)	Variant ¹ allele (amino acid)	Nucleotide common/ variant	Gene name and official abbreviation ²	Study
<i>ERCC6 1213</i>	rs2228527	Arg	Gly	A/G	excision repair cross-complementing rodent repair deficiency, complementation group 6 [<i>ERCC6</i>]	CBCS
<i>ERCC6 1230</i>	rs4253211	Arg	Pro	G/C	excision repair cross-complementing rodent repair deficiency, complementation group 6 [<i>ERCC6</i>]	CBCS
<i>GSTM1</i> ⁵		present	null		glutathione S-transferase mu 1 [<i>GSTM1</i>]	Both
<i>GSTP1 105</i> ⁶	rs1695	Ile	Val	A/C	glutathione S-transferase pi 1 [<i>GSTP1</i>]	CBCS
<i>GSTT1</i> ⁵		present	null		glutathione S-transferase theta 1 [<i>GSTT1</i>]	Both
<i>MEH 113</i>	rs1051740	Tyr	His	T/C	epoxide hydrolase 1, microsomal (xenobiotic) [<i>EPHX1</i>]	NCCCS
<i>MEH 139</i>	rs55784606	His	Tyr	C/T	epoxide hydrolase 1, microsomal (xenobiotic) [<i>EPHX1</i>]	NCCCS
<i>MGMT 84</i>	rs12197	Leu	Phe	C/T	O-6-methylguanine-DNA methyltransferase [<i>MGMT</i>]	CBCS
<i>MLH1 219</i>	rs1799977	Ile	Val	A/G	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli) [<i>MLH1</i>]	NCCCS
<i>MNSOD 16</i> ⁷	rs4880	Val	Ala	T/C	superoxide dismutase 2, mitochondrial [<i>SOD2</i>]	Both
<i>MPO -463</i>	rs2333227	--	--	G/A	myeloperoxidase [<i>MPO</i>]	CBCS
<i>MSH3 1036</i>	rs26279	Thr	Ala	A/G	mutS homolog 3 (E. coli) [<i>MSH3</i>]	NCCCS
<i>MSH3 940</i>	rs184967	Arg	Gln	G/A	mutS homolog 3 (E. coli) [<i>MSH3</i>]	NCCCS
<i>MSH6 39</i>	rs1042821	Gly	Glu	G/A	mutS homolog 6 (E. coli) [<i>MSH6</i>]	NCCCS

Table V.B.2. Gene variants in CBCS and NCCCS (continued)

Gene & codon/ nucleotide position	rs#	Common ¹ allele (amino acid)	Variant ¹ allele (amino acid)	Nucleotide common/ variant	Gene name and official abbreviation ²	Study
<i>MYH 324</i>	rs3219489	Gln	His	G/C	mutY homolog (E. coli) [<i>MUTYH</i>]	CBCS
<i>NAT1</i>	rs1057126	(*10, rapid)	(Non *10)	T/A	N-acetyltransferase 1 (arylamine N-acetyltransferase) [<i>NAT1</i>]	CBCS
<i>NAT2</i>	Reference	(*4, rapid)	(*5,*6,*7,*14,slow)		N-acetyltransferase 2 (arylamine N-acetyltransferase) [<i>NAT2</i>]	CBCS
<i>NBS1 185</i>	rs1805794	Glu	Gln	G/C	Nijmegen breakage syndrome 1 (nibrin) [<i>NIB</i>]	Both
<i>NQO1 187</i>	rs1800566	Pro	Ser	C/T	NAD(P)H dehydrogenase, quinone 1 [<i>NQO1</i>]	CBCS
<i>OGG1 326</i>	rs1052133	Ser	Cys	C/G	8-oxoguanine DNA glycosylase [<i>OGG1</i>]	CBCS
<i>POLD1 119</i>	rs1726801	Arg	His	G/A	polymerase (DNA directed), delta 1, catalytic subunit 125kDa [<i>POLD1</i>]	NCCCS
<i>RAD23B</i>	rs1805329	Ala	Val	C/T	<i>RAD23</i> homolog B (<i>S. cerevisiae</i>) [<i>RAD23B</i>]	Both
<i>TGFB1</i>	rs1800470	Leu	Pro	T/C	transforming growth factor, beta 1 [<i>TGFB1</i>]	CBCS
<i>XPC 499</i>	rs2228000	Ala	Val	C/T	xeroderma pigmentosum, complementation group C [<i>XPC</i>]	NCCCS
<i>XPC 939</i>	rs2228001	Lys	Gln	A/C	xeroderma pigmentosum, complementation group C [<i>XPC</i>]	Both
<i>XPD 312</i>	rs1799793	Asp	Asn	G/A	excision repair cross-complementing rodent repair deficiency, complementation group 2 [<i>ERCC2</i>]	Both
<i>XPD 751</i>	rs13181	Lys	Gln	A/C	excision repair cross-complementing rodent repair deficiency, complementation group 2 [<i>ERCC2</i>]	Both
<i>XPF 415</i>	rs1800067	Arg	Gln	G/A	excision repair cross-complementing rodent repair deficiency, complementation group 4 [<i>ERCC4</i>]	Both
<i>XPF 662</i>	rs2020955	Ser	Pro	T/C	excision repair cross-complementing rodent repair deficiency, complementation group 4 [<i>ERCC4</i>]	CBCS
<i>XPG 1104</i>	rs17655	Asp	His	G/C	excision repair cross-complementing rodent repair deficiency, complementation group 5 [<i>ERCC5</i>]	Both
<i>XRCC1 194</i>	rs1799782	Arg	Trp	C/T	X-ray repair complementing defective repair in Chinese hamster cells 1 [<i>XRCC1</i>]	Both

Table V.B.2. Gene variants in CBCS and NCCCS (continued)

Gene & codon/ nucleotide position	rs#	Common ¹ allele (amino acid)	Variant ¹ allele (amino acid)	Nucleotide common/ variant	Gene name and official abbreviation ²	Study
<i>XRCC1</i> 280	rs25489	Arg	His	G/A	X-ray repair complementing defective repair in Chinese hamster cells 1 [<i>XRCC1</i>]	Both
<i>XRCC1</i> 399	rs25487	Arg	Gln	G/A	X-ray repair complementing defective repair in Chinese hamster cells 1 [<i>XRCC1</i>]	Both
<i>XRCC2</i> 188	rs3218536	Arg	His	G/A	X-ray repair complementing defective repair in Chinese hamster cells 2 [<i>XRCC2</i>]	CBCS
<i>XRCC3</i> 241	rs 861539	Thr	Met	C/T	X-ray repair complementing defective repair in Chinese hamster cells 3 [<i>XRCC3</i>]	Both
<i>XRCC4</i> - 28073 ⁸	rs2075685	T	G	T/G	X-ray repair complementing defective repair in Chinese hamster cells 4 [<i>XRCC4</i>]	CBCS

Abbreviations: 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

¹ Analyzed as common and variant as defined by frequency in CBCS/NCCS datasets. The less frequent allele varied by race where noted.

² <http://www.ncbi.nlm.nih.gov/sites/entrez> (accessed 5/13/2009)

³ *ADPRTL2* 328: Less frequent nucleotide was C in African Americans, T in non-African Americans

⁴ *COMT*: less frequent allele was Met in African Americans, Val in non-African Americans

⁵ Present (referent) or null

⁶ *GSTP1*: Less frequent allele was Ile in African Americans, Val in non-African Americans

⁷ *MnSOD* (CBCS & NCCCS): Less frequent allele was Ala in African Americans, Val in non-African Americans

⁸ *XRCC4* -28073: Less frequent nucleotide was G in African Americans, T in non-African Americans

Table V.B.3a. Gene variant-smoking status associations in the CBCS, overall and by race ^{1,2}										
Gene pathway/ SNP ⁵	Ever smokers ³					Current smokers ⁴				
	OR _z ²	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y
Xenobiotic metabolism⁶										
<i>CYP1A1 M1</i>	1.0	0.8	1.1	1.3	0.7	1.0	0.8	1.1	1.0	1.0
<i>CYP1A1 M2</i>	1.8	1.6								
<i>CYP1A1 M3</i>	0.9		1.0							
<i>CYP1A1 M4</i>	1.3	1.5				2.5	2.9			
<i>GSTM1</i>	1.0	1.2	0.8	1.1	1.0	1.1	1.0	1.1	1.3	0.7
<i>GSTP1</i>	1.2	1.4	0.8	1.2	1.2	0.7	0.7		0.7	0.7
<i>GSTT1</i>	1.0	0.7	1.5	0.9	1.1	1.1	0.9		0.9	
<i>NAT1</i>	0.9	1.1	0.8	1.0	1.0	1.2	1.2		1.5	
<i>NAT2</i>	0.9	1.1	0.9	1.1	1.0	1.3	2.1		1.5	
<i>COMT</i>	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										
<i>APE1 148</i>	1.1	1.3	0.9	1.3	1.0	1.2	1.3	1.0	1.3	1.0
<i>hOGG1</i>	1.0	1.0	1.1	1.1	1.0	0.9	1.0	0.8	1.0	0.8
<i>MYH 324</i>	1.0	1.0	0.8	1.0	0.9	0.8	0.9	0.8	0.8	0.8
<i>XRCC1 194</i>	1.1	1.0	1.3	1.2	1.1	1.1	0.9	1.5	1.0	1.3
<i>XRCC1 280</i>	0.9	0.9	0.9	0.7	1.1	0.9	0.8		0.8	
<i>XRCC1 399</i>	1.0	1.1	1.1	1.1	1.1	1.2	1.2	1.4	1.2	1.3
Double strand break repair										
<i>BRCA2 24</i>	0.9	0.9	0.9	0.9	0.9	0.9	0.8	1.1	0.9	0.9
<i>BRCA2 372</i>	1.2	1.2	1.2	1.0	1.4	1.2	1.1	1.2	1.0	1.4
<i>NBS1 185</i>	1.2	1.3	1.0	1.4	1.1	1.0	1.1	1.1	1.2	0.9
<i>XRCC2 188</i>	0.9	0.8		1.0	0.9	0.9	0.9		1.1	
<i>XRCC3 241</i>	0.9	0.9	1.0	1.0	0.9	1.2	1.2	1.2	1.2	1.2
<i>XRCC4 -28073</i>	1.2	1.2	1.3	1.5	1.0	1.2	1.1	1.4	1.5	1.0
Mismatch repair										
<i>MGMT 84</i>	0.9	0.9	1.0	1.0	0.9	0.8	0.9	0.7	0.8	0.8
Nucleotide excision repair										
<i>ERCC1 8092</i>	1.0	0.9	1.1	1.0	0.9	1.0	0.8	1.1	0.8	1.1
<i>ERCC6 1213</i>	1.2	1.4	1.0	1.3	1.2	1.6	1.7	1.4	1.6	1.5

Table V.B.3a. Gene variant-smoking status associations in the CBCS, overall and by race ^{1,2} (continued)										
Gene pathway/ SNP ⁵	Ever smokers ³					Current smokers ⁴				
	OR _z ²	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y
Nucleotide excision repair (continued)										
<i>ERCC6 1230</i>	0.9	0.9	1.2	1.0	0.8	1.1	1.1		1.6	0.8
<i>HRAD23B</i>	1.1	1.1	1.0	1.2	1.0	1.2	1.3		1.1	1.3
<i>XPC 939</i>	0.9	0.9	1.0	0.9	1.0	1.0	1.1	0.9	1.0	1.0
<i>XPD 312</i>	1.0	1.0	1.1	1.1	1.1	1.1	1.1	1.0	1.0	1.2
<i>XPD 751</i>	1.2	1.2	1.1	1.1	1.3	1.2	1.1	1.3	1.0	1.4
<i>XPF 415</i>	1.0	1.1		1.0	1.0	1.0	0.9		0.7	1.2
<i>XPF 662</i>	1.1		1.2	1.0	1.4	1.4		1.4	1.5	1.3
<i>XPG 1104</i>	0.9	1.0	0.7	0.9	0.9	0.8	0.8	0.9	0.9	0.8
Cell adhesion										
<i>CDH1</i>	0.8	0.8	0.8	0.9	0.8	0.8	0.8	0.9	0.9	0.8
Cell growth										
<i>TGFB1</i>	1.1	1.1	1.1	1.3	0.9	0.8	0.8	0.9	0.9	0.7
Oxidative stress defense										
<i>MnSOD</i>	1.0	0.9	1.0	1.1	0.8	0.9	0.8	0.9	0.8	0.9
<i>MPO</i>	1.0	1.2	0.9	0.8	1.4	1.0	1.2	0.8	0.9	1.3
<i>NQO1</i> ⁷		1.3	0.8	1.2	1.0	1.0	1.3	0.7	1.2	1.0

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

¹ Odds ratios are race and age adjusted unless stratified by race or age, respectively

² Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁴ Referent is not-current smokers (former + never)

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

⁷ Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05

Bold = Overall OR_z

<=0.7

>=1.4

Table V.B.3a. Gene variant-smoking status associations in the CBCS, overall and by race ^{1,2} (continued)										
Gene pathway/ SNP ⁵	Former smokers ³					Current smokers ³				
	OR _z	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y
Xenobiotic metabolism ⁶										
<i>CYP1A1 M1</i>	1.0	0.8	1.1	1.5	0.6	1.0	0.8	1.2	1.1	0.8
<i>CYP1A1 M2</i>	2.1									
<i>CYP1A1 M3</i>										
<i>CYP1A1 M4</i>										
<i>GSTM1</i>	1.0	1.3		0.9	1.1	1.1	1.1	1.0	1.3	0.8
<i>GSTP1</i>	1.8	1.9		1.9	1.5	0.8	0.9		0.8	
<i>GSTT1</i>	0.9	0.7				1.1				
<i>NAT1</i>	0.8	1.0		0.7	1.1	1.1	1.3		1.4	
<i>NAT2</i>	0.8	0.9		0.9	0.9	1.2			1.4	
<i>COMT</i>	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										
<i>APE1 148</i>	1.1	1.2	0.8	1.2	1.0	1.2	1.4	1.0	1.4	1.0
<i>hOGG1</i>	1.1	1.0	1.3	1.1	1.1	1.0	1.0	0.8	1.0	0.9
<i>MYH 324</i>	1.1	1.1	1.0	1.2	1.0	0.9	0.9	0.7	0.9	0.8
<i>XRCC1 194</i>	1.1	1.1	1.2	1.3	1.0	1.2	1.0	1.5	1.1	1.3
<i>XRCC1 280</i>	0.9	1.0		0.7	1.2	0.9	0.8		0.8	
<i>XRCC1 399</i>	1.0	1.0	0.9	0.9	1.0	1.2	1.2	1.4	1.2	1.3
Double strand break repair										
<i>BRCA2 24</i>	1.0	1.0	0.8	1.0	0.9	0.9	0.8	1.0	0.9	0.9
<i>BRCA2 372</i>	1.2	1.2	1.2	1.0	1.4	1.3	1.2	1.3	1.0	1.5
<i>NBS1 185</i>	1.2	1.4	0.9	1.4	1.1	1.1	1.2	1.0	1.4	0.9
<i>XRCC2 188</i>	0.9	0.9		0.9	1.0	0.8	0.8		1.0	
<i>XRCC3 241</i>	0.8	0.8	0.9	0.9	0.8	1.1	1.1	1.2	1.2	1.1
<i>XRCC4 -28073</i>	1.1	1.2	1.1	1.3	1.0	1.3	1.2	1.5	1.6	1.0
Mismatch repair										
<i>MGMT 84</i>	1.1	1.0	1.2	1.2	1.0	0.8	0.9	0.7	0.9	0.8
Nucleotide excision repair										
<i>ERCC1 8092</i>	1.0	0.9	1.0	1.2	0.8	1.0	0.8	1.1	0.9	1.0
<i>ERCC6 1213</i>	1.0	1.1	0.8	1.0	1.0	1.6	1.8	1.3	1.6	1.5

Table V.B.3a. Gene variant-smoking status associations in the CBCS, overall and by race ^{1,2}										
Gene pathway/ SNP ⁵	Former smokers ³					Current smokers ³				
	OR _z	NAA	AA	<50y	≥50y	OR _z	NAA	AA	<50y	≥50y
Nucleotide excision repair (continued)										
<i>ERCC6 1230</i>	0.8	0.8		0.7	0.8	1.0	1.0		1.4	0.7
<i>HRAD23B</i>	1.0	1.0	1.2	1.2	0.9	1.2	1.3		1.1	1.3
<i>XPC 939</i>	0.9	0.8	1.1	0.8	0.9	1.0	1.0	0.9	0.9	1.0
<i>XPD 312</i>	1.0	1.0	1.2	1.1	1.0	1.1	1.1	1.1	1.0	1.2
<i>XPD 751</i>	1.1	1.2	1.0	1.1	1.2	1.2	1.2	1.2	1.0	1.5
<i>XPF 415</i>	1.1	1.1		1.2	0.9	1	0.9		0.8	1.2
<i>XPF 662</i>	1.0		1.1	0.7	1.4	1.3		1.4	1.3	1.5
<i>XPG 1104</i>	1.0	1.1	0.7	0.9	1.0	0.8	0.9	0.8	0.8	0.8
Cell adhesion										
<i>CDH1</i>	0.8	0.9	0.7	0.9	0.8	0.8	0.8	0.9	0.9	0.7
Cell growth										
<i>TGFBI</i>	1.2	1.3	1.2	1.6	1.0	0.9	0.8	1.0	1.1	0.7
Oxidative stress defense										
<i>MnSOD</i>	1.1	1.0	1.0	1.3	0.8	0.9	0.8	0.9	0.9	0.8
<i>MPO</i>	1.0	1.2	1.0	0.8	1.3	1.0	1.2	0.8	0.8	1.4
<i>NQO1</i> ⁷	1.0	1.2	0.8	1.2	1.0	1.0	1.4	0.7	1.2	1.0

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

¹ Odds ratios are race and age adjusted unless stratified by race or age, respectively

² Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁴ Referent is not-current smokers (former + never)

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

⁷ Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05

Bold = Overall OR_z

<=0.7 >=1.4

Table V.B.3b. Gene variant-smoking duration association in the CBCS, overall and by race ^{1,2}															
	<=10 years ³					11-20 years					>20 years				
Gene pathway/ SNP ⁵	OR _z ²	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y
Xenobiotic metabolism ⁶															
CYP1A1 M1	1.0	1.2		1.3		1.3			1.4		0.8	0.7	0.9		0.6
CYP1A1 M2															
CYP1A1 M3															
CYP1A1 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

Table V.B.3b. Gene variant-smoking duration association in the CBCS, overall and by race ^{1,2} (continued)															
	<=10 years ³					11-20 years					>20 years				
Gene pathway/ SNP ⁵	OR _z ²	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y
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
NOO1 ⁷		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

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

¹ Odds ratios are race and age adjusted unless stratified by race or age, respectively

² Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

⁷ Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05

Bold = Overall OR_z

<=0.7 >=1.4

Table V.B.3c. Gene variant-smoking intensity association in the CBCS, overall and by race ^{1,2}															
	<1/2 pack/day ³					1/2 - 1 pack/day					>1 pack/day				
Gene pathway/ SNP ⁵	OR _z ²	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y
Xenobiotic metabolism ⁶															
CYP1A1 M1	0.9	0.8	1.0	1.3		1.0		1.3	1.2		1.0	1.0			
CYP1A1 M2															
CYP1A1 M3															
CYP1A1 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

Table V.B.3c. Gene variant-smoking intensity association in the CBCS, overall and by race ^{1,2} (continued)															
Gene pathway/ SNP ⁵	<1/2 pack/day ³					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
Nucleotide excision repair															
<i>ERCC1 8092</i>	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
<i>ERCC6 1213</i>	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
<i>ERCC6 1230</i>	1.0	1.0		1.1	1.0	1.0	0.9		1.2	0.8	0.8	0.8		0.9	0.7
<i>HRAD23B</i>	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
<i>XPC 939</i>	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
<i>XPD 312</i>	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
<i>XPD 751</i>	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
<i>XPF 415</i>	0.9	0.9		1.2		1.0	0.9		0.9	1.0	1.1	1.3		0.9	1.5
<i>XPF 662</i>	1.4		1.6	1.5	1.7	0.9		0.9	0.7	1.3	0.9		1.0		
<i>XPG 1104</i>	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															
<i>CDH1</i>	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															
<i>TGFB1</i>	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															
<i>MnSOD</i>	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
<i>MPO</i>	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
<i>NQO1</i> ⁷		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

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

¹ Odds ratios are race and age adjusted unless stratified by race or age, respectively

² Odds ratio not displayed if 95% CI width (upper limit/lower limit) >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

⁷ Could not be pooled. LRT p-value for race*smoking interaction term <0.05

Bold = Overall OR_z

<=0.7 >=1.4

Table V.B.3d. Gene variant-PY association in the CBCS, overall and by race ^{1,2}

	<=35 PY					>35PY				
Gene pathway/ SNP ⁵	OR _z ²	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y
Xenobiotic metabolism ⁶										
<i>CYP1A1 M1</i>	1.0	0.8	1.1	1.3	0.6					
<i>CYP1A1 M2</i>	1.6									
<i>CYP1A1 M3</i>	1.0		1.0							
<i>CYP1A1 M4</i>							0.6			0.8
<i>GSTM1</i>	0.9	1.0	0.8	1.0	0.8	1.7				
<i>GSTP1</i>	1.4	1.6	0.8	1.2	1.4	0.7				
<i>GSTT1</i>	1.0	0.8	1.4	0.8	1.4		1.1			1.0
<i>NAT1</i>	0.9	1.1	0.9	1.1	1.0					
<i>NAT2</i>	0.9	1.0	1.0	1.0	0.9		1.2		1.2	1.0
<i>COMT</i>	0.9	0.6	1.3	1.0	0.9	0.5				
DNA repair										
Base excision repair										
<i>APE1 148</i>	1.1	1.2	1.0	1.3	1.0	1.3	1.8			1.1
<i>hOGG1</i>	1.1	1.0	1.2	1.1	1.0	0.9	1.1			1.0
<i>MYH 324</i>	1.0	1.0	0.8	1.0	0.9	1.0	1.1		1.3	1.0
<i>XRCC1 194</i>	1.1	0.9	1.3	1.3	0.9	1.6	1.5			1.7
<i>XRCC1 280</i>	0.9	0.8	1.0	0.6	1.3					
<i>XRCC1 399</i>	1.0	1.1	1.0	1.0	1.1	1.0	0.9			0.9
Double strand break repair										
<i>BRCA2 24</i>	0.9	1.0	0.9	0.9	0.9	0.8	0.7			0.9
<i>BRCA2 372</i>	1.2	1.2	1.2	1.0	1.3	1.6	1.5			2.0
<i>NBS1 185</i>	1.2	1.4	1.0	1.4	1.1	1.0	1.2		1.2	1.0
<i>XRCC2 188</i>	0.9	0.8		0.9	0.8	1.1	1.1			
<i>XRCC3 241</i>	0.9	0.9	1.0	1.1	0.8	0.8	0.8			0.9
<i>XRCC4 1394</i>	1.1	1.1	1.3	1.4	1.0	1.5	1.6			1.3
Mismatch repair										
<i>MGMT 84</i>	1.0	1.0	1.0	1.0	1.0	0.5	0.6			0.6

Table V.B.3d. Gene variant-PY association in the CBCS, overall and by race ^{1,2} (continued)

Gene pathway/ SNP ⁵	<=35 PY					>35PY				
	OR _z ²	NAA	AA	<50y	>=50y	OR _z	NAA	AA	<50y	>=50y
Nucleotide excision repair										
<i>ERCC1 8092</i>	1.0	1.0	1.1	1.1	0.9	0.7	0.6			0.8
<i>ERCC6 1213</i>	1.3	1.4	1.1	1.3	1.3	0.8	1.1			0.8
<i>ERCC6 1230</i>	0.9	0.9		1.1	0.9	0.7	0.7			0.6
<i>HRAD23B</i>	1.1	1.1	1.0	1.2	1.0	1.2	1.3			1.1
<i>XPC 939</i>	0.9	0.8	0.9	0.9	0.9	1.1	1.2			1.2
<i>XPD 312</i>	1.1	1.1	1.1	1.1	1.1	0.9	0.9			1.0
<i>XPD 751</i>	1.2	1.3	1.1	1.1	1.3	1.1	1.1		0.8	1.2
<i>XPF 415</i>	1.0	1.0		1.1	0.9	1.0	1.2			1.4
<i>XPF 662</i>	1.2		1.3	1.1	1.5					
<i>XPG 1104</i>	0.9	1.0	0.7	0.9	0.9	1.2	1.5			1.2
Cell adhesion										
<i>CDH1</i>	0.8	0.8	0.8	0.8	0.8	0.9	0.9		1.4	0.8
Cell growth										
<i>TGFB1</i>	1.1	1.1	1.0	1.3	1.0	0.8	0.9			0.8
Oxidative stress defense										
<i>MnSOD</i>	1.0	0.8	1.0	1.0	0.8	1.4	1.5			1.1
<i>MPO</i>	1.0	1.2	0.9	0.8	1.4	1.1	1.1		1.3	1.1
<i>NQO1</i> ⁷		1.3	0.8	1.2	1.0	1.1	1.4			0.9

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

¹ Odds ratios are race and age adjusted unless stratified by race or age, respectively

² Odds ratio not displayed if 95% CI width (upper limit/lower limit) >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁴ Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e. g. *COMT* in estrogen metabolism

⁷ 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

Table V.B.4a. Gene variant-smoking status associations in the NCCCS, overall and by gender and race ^{1,2}

Gene pathway/ Gene variant ⁵	Ever							Current vs. not current						
	OR _z ²	W	M	NAA	AA	<65y	>=65y	OR _z	W	M	NAA	AA	<65y	>=65y
Xenobiotic metabolism ⁶														
<i>GST hap C</i> ⁷	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
<i>GST hap A</i> ⁸	1.4	1.6	1.2	0.9	2.2	1.8	1.1	1.5		1.4		2.3	1.5	1.4
<i>GST hap B</i> ⁸	0.8	0.8	0.6	0.7			0.7	0.6						
<i>GST hap D</i> ⁸	1.3	1.3	1.3	1.2	1.7	1.6	1.1	0.9		0.7	0.8		1.0	0.8
<i>GSTM1</i>	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
<i>GSTT1</i> ⁹		1.0	0.9	0.7	1.5	1.0	0.9		1.1	1.1	0.7	1.8	1.0	1.3
<i>MEH 113</i>	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
<i>MEH 139</i>	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														
<i>POLD1 119</i>	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														
<i>ADPRT 762</i>	1.1	1.3	0.9	1.1		1.1	1.1	1.2	1.8	0.9	1.1		1.3	1.1
<i>ADPRTL2 328</i>	1.1	1.2	1.0	1.2		1.1	1.1	1.1		1.0	1.2		1.0	1.1
<i>APE1 148</i>	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
<i>XRCC1 194</i>	0.8	0.6		0.8	0.9		0.8	0.9						
<i>XRCC1 280</i>	1.3			1.2										
<i>XRCC1 399</i>	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														
<i>NBS1 185</i>	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
<i>XRCC3 241</i>	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														
<i>MLH1 219</i>	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
<i>MSH3 1036</i>	1.1	1.8	0.8	1.2	1.3	1.4	1.1	1.0	0.9	0.6	0.6		1.0	
<i>MSH3 940</i>	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
<i>MSH6 39</i>	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

Table V.B.4a. Gene variant-smoking status associations in the NCCCS, overall and by gender and race^{1,2} (continued)

Gene pathway/ Gene variant ⁵	Ever							Current vs. not current						
	OR _z ²	W	M	NAA	AA	<65y	>=65y	OR _z	W	M	NAA	AA	<65y	>=65y
Nucleotide excision repair														
<i>RAD23B</i>	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
<i>XPC 499</i>	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
<i>XPC 939</i>	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
<i>XPB 312</i>	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
<i>XPB 751</i>	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
<i>XPF 415</i>	1.0		1.7	1.1			1.1	1.3		1.5	1.7			
<i>XPG 1104</i>	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														
<i>MNSOD</i>	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

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

¹ Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively

² Odds ratio not displayed if confidence limit ratio >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁴ Referent is not-current smokers (former + never)

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

⁷ *GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined

⁸ *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

¹¹ Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day

⁹ 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

Table V.B.4b. Gene variant-smoking status associations in the NCCCS, overall and by gender and race ^{1,2}

Gene pathway/ Gene variant ⁵	Former							Current						
	OR _z ²	W	M	NAA	AA	<65y	>=65y	OR _z	W	M	NAA	AA	<65y	>=65y
Xenobiotic metabolism ⁶														
<i>GST hap C</i> ⁷	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
<i>GST hap A</i> ⁸	1.3	1.5	1.1	0.9	1.7		1.0	1.7						
<i>GST hap B</i> ⁸	0.9			0.8			0.7							
<i>GST hap D</i> ⁸	1.4	1.3	1.5	1.2		1.8	1.2	1.1		0.9	0.9			
<i>GSTM1</i>	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
<i>GSTT1</i>	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
<i>MEH 113</i>	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
<i>MEH 139</i>	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														
<i>POLD1 119</i>	0.7	1.2	1.0	1.1		1.3	1.0	0.7		0.8	1.2		1.3	
Base excision repair														
<i>ADPRT 762</i>	1.0	1.1	0.9	1.1		1.0	1.1	1.2		0.8	1.1		1.2	
<i>ADPRTL2 328</i>	1.1	1.1	1.0	1.2		1.1	1.1	1.1		1.0	1.3		1.1	1.2
<i>APE1 148</i>	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
<i>XRCC1 194</i>	0.8	0.6		0.7			0.8	0.8						
<i>XRCC1 280</i>	1.3			1.3										
<i>XRCC1 399</i>	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														
<i>NBS1 185</i>	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
<i>XRCC3 241</i>	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														
<i>MLH1 219</i>	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
<i>MSH3 1036</i>	1.1	2.1	0.9	1.4	1.5	1.5	1.3	1.0			0.7			
<i>MSH3 940</i>	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
<i>MSH6 39</i>	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

Table V.B.4b. Gene variant-smoking status associations in the NCCCS, overall and by gender and race ^{1,2} (continued)														
Gene pathway/ Gene variant ⁵	Former							Current						
	OR _z ²	W	M	NAA	AA	<65y	>=65y	OR _z	W	M	NAA	AA	<65y	>=65y
Nucleotide excision repair														
<i>RAD23B</i>	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
<i>XPC 499</i>	0.8	0.9	0.7	0.7		1.0	0.7	1.0	0.9	0.9	0.8		1.1	0.9
<i>XPC 939</i>	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
<i>XPD 312</i>	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
<i>XPD 751</i>	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
<i>XPF 415</i>	0.9			0.9			1.1	1.2			1.6			
<i>XPG 1104</i>	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														
<i>MNSOD</i>	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

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

¹ Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively

² Odds ratio not displayed if confidence limit ratio >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

⁷ *GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined

⁸ *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

¹¹ Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day

⁹ 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

Table V.B.4c. Gene variant-smoking duration association in the NCCCS, overall and by gender and race ^{1,2}

Gene pathway/ Gene variant ⁵	<10y							11-20y							>20 y						
	OR _z ²	W	M	NAA	AA	Y	O	OR _z	W	M	NAA	AA	Y	O	OR _z	W	M	NAA	AA	Y	O
Xenobiotic metabolism ⁶																					
<i>GST hap C</i> ⁷	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
<i>GST hap A</i> ⁸	1.6							1.3							1.4	1.6	1.2	0.7			
<i>GST hap B</i> ⁸															0.7			0.6			
<i>GST hap D</i> ⁸	1.7			1.6				1.1			1.1				1.2	1.3	1.2	1.0			
<i>GSTM1</i>	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
<i>GSTT1</i>	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
<i>MEH 113</i>	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
<i>MEH 139</i>	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																					
<i>POLD1 119</i>	0.7			1.2				0.9			0.9				0.7	1.2	1.0	1.2			
Base excision repair																					
<i>ADPRT 762</i>	1.2			1.4				0.7			0.7				1.2	1.5	1.0	1.1			
<i>ADPRTL2 328</i>	1.0		0.7	1.1			0.9	1.0	1.2		1.2				1.2	1.3	1.0	1.3			
<i>APE1 148</i>	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
<i>XRCC1 194</i>															0.9			0.8			
<i>XRCC1 280</i>															1.2						
<i>XRCC1 399</i>	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																					
<i>NBS1 185</i>	0.9							1.2	0.9						0.8	0.6	0.7	0.8			
<i>XRCC3 241</i>	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
<i>MLH1 219</i>	1.4		1.1	0.8			0.9	1.0			1.0			1.0	1.1	0.6	1.2	0.8			1.0
<i>MSH3 1036</i>	1.0			1.2			1.2	1.1			1.2			1.5	1.0	1.4	0.8	1.2			1.5
<i>MSH3 940</i>	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
<i>MSH6 39</i>	0.8	0.9	1.0	1.0		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

Table V.B.4c. Gene variant-smoking duration association in the NCCCS, overall and by gender and race ^{1,2} (continued)

Gene pathway/ Gene variant ⁵	<10y							11-20y							>20 y						
	OR _z ²	W	M	NAA	A	Y	O	OR _z	W	M	NAA	AA	Y	O	OR _z	W	M	NAA	AA	Y	O
Nucleotide excision repair																					
<i>RAD23B</i>	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
<i>XPC 499</i>	0.8		0.7	0.7			0.6	0.9	0.7		0.7			1.3	0.9	0.8	0.9	0.8			1.3
<i>XPC 939</i>	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
<i>XPD 312</i>	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
<i>XPD 751</i>	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
<i>XPF 415</i>	1.0														1.1			1.2			
<i>XPG 1104</i>	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																					
<i>MNSOD</i>	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

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

¹ Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively

² Odds ratio not displayed if confidence limit ratio >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

⁷ *GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined

⁸ *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

¹¹ Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day

⁹ Could not be pooled for some measures of smoking. LRT p-value for race*smoking interaction term <0.05

Bold = Overall OR_z

Light blue = OR_z ≤0.7

Orange = OR_z ≥1.4

Table V.B.4d. Gene variant-smoking intensity association in the NCCCS, overall and by gender and race^{1,2}

Gene pathway/ Gene variant ⁵	<1/2 pack/day							1/2 - 1 pack/day							>1 pack/day						
	OR _z ²	W	M	NAA	A	Y	O	OR _z	W	M	NAA	AA	Y	O	OR _z	W	M	NAA	AA	Y	O
Xenobiotic metabolism⁶																					
<i>GST</i> hap C ⁷	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
<i>GST</i> hap A ⁸	1.8						1.7	1.2		1.2				0.7	1.3						0.7
<i>GST</i> hap B ⁸	1.1							0.5							0.9						
<i>GST</i> hap D ⁸	1.9			1.6			1.5	1.0	0.9	1.1	0.9			0.8	1.3		1.2	1.3			0.8
<i>GSTM1</i>	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
<i>GSTT1</i>	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
<i>MEH 113</i>	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
<i>MEH 139</i>	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																					
<i>POLD1</i> 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																					
<i>ADPRT</i> 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
<i>ADPRTL2</i> 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
<i>APE1</i> 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
<i>XRCC1</i> 194	0.7							0.7							1.1			1.0			
<i>XRCC1</i> 280																					
<i>XRCC1</i> 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

Table V.B.4d. Gene variant-smoking intensity association in the NCCCS, overall and by gender and race ^{1,2} (continued)

Gene pathway/ Gene variant ⁵	<1/2 pack/day							1/2 - 1 pack/day							>1 pack/day						
	OR _z ²	W	M	NAA	AA	Y	O	OR _z	W	M	NAA	AA	Y	O	OR _z	W	M	NAA	AA	Y	O
Double strand break repair																					
<i>NBS1 185</i>	1.0	0.6	0.6		0.6		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
<i>XRCC3 241</i>	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
<i>MLH1 219</i>	1.1	0.7	1.9	0.9	1.4		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
<i>MSH3 1036</i>	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		1.6	0.9
<i>MSH3 940</i>	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
<i>MSH6 39</i>	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																					
<i>RAD23B</i>	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
<i>XPC 499</i>	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
<i>XPC 939</i>	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
<i>XPD 312</i>	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
<i>XPD 751</i>	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
<i>XPF 415</i>								1.2			1.3			1.4	1.0			0.9			1.4
<i>XPG 1104</i>	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		0.7	1.0	1.2	1.5	0.8
Oxidative stress defense																					
<i>MNSOD</i>	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

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

1 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

2 Odds ratio not displayed if confidence limit ratio >4

3 Referent is never smokers for all smoking categories unless otherwise noted

5 SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

6 Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

7 *GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined

Table V.B.4d. Gene variant-smoking intensity association in the NCCCS, overall and by gender and race ^{1,2} (continued)

8 *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

11 Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day

⁹ 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

Table V.B.4e. Gene variant-pack-years of smoking association in the NCCCS, overall and by gender and race ^{1,2}

<=35 PY								>35PY							
Gene pathway/ Gene variant ⁵	OR _z ²	W	M	NAA	AA	<65y	>=65y	OR _z	W	M	NAA	AA	<65y	>=65y	
Xenobiotic metabolism ⁶															
GST hap C ⁷	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 ⁸	1.5	1.9	1.2	1.0	2.1		1.2	1.2							
GST hap B ⁸	0.8	1.2		0.8											
GST hap D ⁸	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	
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	

Table V.B.4e.

Gene variant-pack-years of smoking association in the NCCCS, overall and by gender and race^{1,2} (continued)

<=35 PY								>35PY							
Gene pathway/ Gene variant ⁵	OR _z ²	W	M	NAA	AA	<65y	>=65y	OR _z	W	M	NAA	AA	<65y	>=65y	
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			1.0			1.1	1.1			1.1				
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		1.1	1.5			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

¹ Odds ratios are race, age and gender adjusted unless stratified by race, age or gender, respectively

² Odds ratio not displayed if confidence limit ratio >4

³ Referent is never smokers for all smoking categories unless otherwise noted

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

⁷ *GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined

⁸ *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

¹¹ Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day

⁹ 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

Table V.B.5: Gene variant-smoking associations in CBCS and NCCCS

	Smoking Status								Duration						Intensity						Pack-years			
	Ever smoking (OR _z) ^{5,6}		Current smokers (Ref: Not Current) ⁷		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 ¹⁰ (OR _z))		>35PY (OR _z)	
Gene pathway ⁹ / gene variant ⁸	B ³	C ⁴	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C
Xenobiotic metabolism⁶																								
<i>GSTM1</i>	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
<i>GSTT1</i>	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																								
<i>APE1 148</i>	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
<i>XRCC1 194</i>	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
<i>XRCC1 280</i>	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		
<i>XRCC1 399</i>	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																								
<i>NBS1 185</i>	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
<i>XRCC3 241</i>	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																								
<i>HRAD23B</i>	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
<i>XPC 939</i>	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
<i>XPD 312</i>	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
<i>XPD 751</i>	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
<i>XPF 415</i>	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
<i>XPG 1104</i>	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																								
<i>MNSOD</i>	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

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 V.B.5: Gene variant-smoking associations in CBCS and NCCCS (continued)

³ All odds ratios from CBCS (B) are race and age adjusted

⁴ All odds ratios from NCCCS (C) are race, age and gender adjusted

⁵ Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4

⁶ Referent is never smokers for all smoking categories unless otherwise noted

⁷ Referent is not-current smokers (former + never)

⁸ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁹ Primary functional category; gene may function in additional pathways e.g. *COMT* in estrogen metabolism

¹⁰ Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day

Bold = Overall OR_z

 = OR_z ≤ 0.7


 = OR_z ≥ 1.4

Table V.B.6. Gene-variant - smoking associations ¹ in CBCS & NCCCS: Non-African American female controls 40-74 years of age

	Smoking ² Status				Duration (years)			Intensity (pack/day)			Pack-years ³	
	Ever smoking (OR _z) ^{1,4}	Current smokers (Ref: Not Current) ⁵	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 ⁷	B ⁸ C ⁹	B C	B C	B C	B C	B C	B C	B C	B C	B C	B C	B C
CBCS												
Xenobiotic metabolism ⁶												
<i>CYP1A1 M1</i>	0.8	0.9	0.7	0.8	1.0		0.7	0.7	0.6	1.0	0.7	1.1
<i>CYP1A1 M2</i>	1.5		2.0									
<i>CYP1A1 M3</i>												
<i>CYP1A1 M4</i>	1.6	3.1		2.8							1.2	
<i>GSTP1</i>	1.5	0.8	2.0	1.0	2.0	1.5	1.3	1.8	2.2	0.9	1.9	0.7
<i>NAT1</i>	1.2	1.3	1.1	1.4		1.3	1.2		1.7	1.0	1.2	
<i>NAT2</i>	1.1	2.0	0.9	1.9		1.1	1.5	0.8	1.2	1.3	0.9	
<i>COMT</i>	0.6	0.7	0.6	0.6		0.4	0.7	0.5	0.8	0.5	0.7	0.5
DNA repair												
<i>hOGG1</i>	1.0	0.9	1.0	0.9	1.1	1.1	0.9	0.8	1.1	1.1	1.0	1.0
<i>MYH 324</i>	1.0	0.9	1.1	0.9	1.2	0.8	1.1	1.0	1.0	1.0	1.0	1.1
<i>BRCA2 24</i>	0.9	0.8	1.0	0.8	1.0	0.8	0.9	1.0	0.8	0.9	1.0	0.7
<i>BRCA2 372</i>	1.3	1.2	1.3	1.3	1.2	1.8	1.2	0.9	1.3	1.7	1.2	1.6
<i>XRCC2 188</i>	0.8	0.7	0.8	0.7	0.5	0.9	0.9	0.6	0.6	1.0	0.7	1.0
<i>XRCC4 -</i>												
<i>28073</i>	1.1	1.1	1.1	1.1	0.7	1.0	1.5	0.9	1.1	1.4	1.0	1.7
<i>MGMT 84</i>	1.0	1.0	0.9	1.0	1.3	0.9	0.8	1.0	1.0	0.8	1.0	0.6
<i>ERCC1</i>												
<i>8092</i>	0.9	0.8	0.9	0.8	0.9	0.8	0.9	0.9	1.1	0.8	1.0	0.7
<i>ERCC6</i>												
<i>1213</i>	1.4	1.8	1.2	1.9	1.2	1.4	1.5	1.3	1.4	1.4	1.5	1.1
<i>ERCC6</i>												
<i>1230</i>	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												
<i>CDH1</i>	0.8	0.9	0.8	0.8	0.9	0.8	0.8	1.0	0.8	0.8	0.8	0.9
<i>TGFB1</i>	0.9	0.7	1.1	0.7	1.3	0.9	0.8	1.1	1.0	0.8	1.0	0.8
<i>MPO</i>	1.3	1.2	1.3	1.3	1.4	1.0	1.4	1.5	1.2	1.2	1.3	1.2
<i>NQO1</i>	1.4	1.4	1.3	1.6	1.2	1.4	1.5	1.2	1.6	1.3	1.4	1.4

Table V.B.6. Gene-variant - smoking associations ¹ in CBCS & NCCCS: Non-African American female controls 40-74 years of age (continued)

	Smoking ² Status				Duration (years)			Intensity (pack/day)			Pack-years ³			
	Ever smoking (OR _z) ^{1,4}	Current smokers (Ref: Not Current) ⁵	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 ⁷	B ⁸ C ⁹	B C	B C	B C	B C	B C	B C	B C	B C	B C	B C	B C	B C	
NCCCS														
Xenobiotic metabolism ⁶														
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 ¹⁰	0.7		0.7				0.6		0.4		1.0			
GST hap A ¹¹														
GST hap B ¹²														
GST hap D ¹³	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			

Table V.B.6. Gene-variant - smoking associations ¹ in CBCS & NCCCS: Non-African American female controls 40-74 years of age (continued)

	Smoking ² Status								Duration (years)			Intensity (pack/day)			Pack-years ³						
	Ever smoking (OR _z) ^{1,4}		Current smokers (Ref: Not Current) ⁵	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 ⁷	B ⁸	C ⁹	B	C	B	C	B	C	B	C	B	C	B	C	B	C	B	C			
CBCS and NCCCS																					
Xenobiotic metabolism ⁶																					
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	1.2		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
XPB 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
XPB 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

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)

Table V.B.6. Gene-variant - smoking associations ¹ in CBCS & NCCCS: Non-African American female controls 40-74 years of age (continued)

¹ OR=Odds ratio; OR not presented if 95% confidence limit ratio >5 (upper limit/lower limit>5)

² Referent is never smokers for all OR_z unless otherwise noted

³ Pack-years= number of years smoked x packs smoked/day [20cigarettes=1 pack]

⁴ All OR_z are age adjusted (continuous)

⁵ Referent is not-current smokers (former + never)

⁶ Primary functional category, gene may function in additional pathways e. g. *COMT* in estrogen metabolism

⁷ 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*.

⁸ B=CBCS (breast cancer study)

⁹ CO=NCCCS (colon cancer study)

¹⁰ *GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null

¹¹ *GST* hap A=haplotype of *GSTT1* null & *GSTM1* present ; *GST* hap C is referent

¹² *GST* hap B=haplotype of *GSTT1* null & *GSTM1* null; *GST* hap C is referent

¹³ *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

Table V.B.7. Gene variant-smoking status association in GSEC, CBCS and NCCCS controls

Gene ⁴	Ever						Former											
	GSEC ¹			CBCS ²			GSEC ¹			CBCS ²			NCCCS ³					
	OR _z	95% CI		OR _z	95% CI		OR _z	95% CI		OR _z	95% CI		OR _z	95% CI				
CYP1A1																		
All	0.9	0.8	1.1				--			0.9	0.7	1.1			--			
Women	1.1	0.9	1.4	1.0	0.7	1.3	--			1.1	0.8	1.6	1.0	0.6	1.4			
Non-hospital	0.9	0.7	1.1	1.0	0.7	1.3	--			0.8	0.7	1.1	1.0	0.6	1.4			
GSTM1																		
All	0.9	0.9	1.0				1.0	0.8	1.4	0.9	0.8	1.1			1.2	0.8	1.6	
Women	0.9	0.8	1.1	1.0	0.7	1.4	1.0	0.7	1.6	0.8	0.7	1.0	1.0	0.7	1.5	1.0	0.6	1.6
Non-hospital	0.9	0.8	1.0	1.0	0.7	1.4	1.0	0.8	1.4	1.0	0.8	1.1	1.0	0.7	1.5	1.2	0.8	1.6
GSTT1																		
All	1.0	0.9	1.2				1.0	0.7	1.3	1.1	0.9	1.3				0.9	0.7	1.3
Women	0.8	0.7	1.0	1.0	0.6	1.5	1.0	0.7	1.6	0.9	0.7	1.3	0.9	0.6	1.6	1.0	0.6	1.6
Non-hospital	1.3	1.1	1.5	1.0	0.6	1.5	1.0	0.7	1.3	1.3	1.0	1.7	0.9	0.6	1.6	0.9	0.7	1.3
GSTP1																		
All	1.1	0.9	1.2				--			1.0	0.8	1.2				--		
Women	0.8	0.6	1.1	1.2	0.8	1.7	--			0.8	0.6	1.2	1.8	1.1	2.6	--		
Non-hospital	1.1	0.9	1.3	1.2	0.8	1.7	--			1.0	0.8	1.3	1.8	1.1	2.6	--		
NAT2																		
All	1.0	0.9	1.2				--			1.1	0.9	1.3				--		
Women	1.1	0.9	1.3	0.9	0.7	1.5	--			1.3	1.0	1.8	0.8	0.6	1.4	--		
Non-hospital	0.9	0.7	1.0	0.9	0.7	1.5	--			1.0	0.8	1.2	0.8	0.6	1.4	--		

Table V.B.7. Gene variant-smoking status association in GSEC, CBCS and NCCCS controls (continued)

Gene ⁴	Current								
	GSEC ¹			CBCS ²			NCCCS ³		
	OR _z	95% CI		OR _z	95% CI		OR _z	95% CI	
<i>CYP1A1</i>									
All	1.0	0.9	1.2				--		
Women	1.3	0.98	1.7	1.0	0.6	1.4	--		
Non-hospital	1.0	0.8	1.2	1.0	0.6	1.4	--		
<i>GSTM1</i>									
All	0.9	0.8	1.0				0.8	0.5	1.2
Women	1.0	0.8	1.2	1.1	0.7	1.6	1.1	0.6	2.1
Non-hospital	0.9	0.8	1.0	1.1	0.7	1.6	0.8	0.5	1.2
<i>GSTT1</i>									
All	1.0	0.8	1.2				1.1	0.7	1.6
Women	--			1.1	0.6	1.7	1.1	0.6	2.1
Non-hospital	1.2	1.0	1.6	1.1	0.6	1.7	1.1	0.7	1.6
<i>GSTP1</i>									
All	1.1	0.9	1.3				--		
Women	0.7	0.4	1.0	0.8	0.5	1.2	--		
Non-hospital	1.1	0.9	1.4	0.8	0.5	1.2	--		
<i>NAT2</i>									
All	0.9	0.8	1.1				--		
Women	1.0	0.8	1.3	1.2	0.8	2.1	--		
Non-hospital	0.8	0.6	0.9	1.2	0.8	2.1	--		

Abbreviations: OR_z=odds ratio in controls, CI=confidence interval, CBCS=Carolina Breast Cancer Study, NCCCS=North Carolina Colon Cancer Study, GSEC= Collaborative Study on Genetic Susceptibility to Environmental Carcinogens

¹ Adjusted for study, gender, age and ethnicity (Smits 2004)

² Adjusted for age and race

³ Adjusted for gender, age and race unless stratified by gender

⁴ Smits 2004: Referent is "wild-type" (WT) (i.e. med type homozygotes) vs. having >=1 variant allele; *CYP1A1*: M1 is the variant allele, *NAT2*: *4 allele is variant allele (rapid acetylator); *GST* and *GSTM* referents are genotypes with >= 1 allele vs. deletion of both alleles (variant=null);

Bold = OR_z

Table V.B.8. Misspecification for Gene variant-smoking status associations (OR_z^1) in the CBCS and NCCCS

Gene pathway ⁶ / Gene variant ⁵	Status				Duration			Intensity			Pack-years ⁹	
	Ever ³	Current ⁴	Former	Current	<=10 years	11-20 years	>20 years	<1/2 pk/day	1/2 - 1 pk/day	>1 pk/day	<=35 PY	>35 PY
CBCS ^{1,2}												
Xenobiotic metabolism⁶												
<i>CYP1A1 M1</i>	1.0	1.0	1.0	1.0	1.0	1.3	0.8	0.9	1.0	1.0	1.0	
<i>CYP1A1 M2</i>	1.8		2.1 ¹⁰								1.6	
<i>CYP1A1 M3</i>	0.9										1.0	
<i>CYP1A1 M4</i>	1.3	2.5 ¹⁰										
<i>GSTM1</i>	1.0	1.1	1.0	1.1	1.1	1.0	1.0	0.9	0.9	1.3	0.9	1.7
<i>GSTP1</i>	1.2	0.7	1.8 ¹⁰	0.8	1.9 ¹⁰	1.0	1.0	1.2	1.8	0.9	1.4	0.7
<i>GSTT1</i>	1.0	1.1	0.9	1.1		1.0	1.0	1.3	1.1		1.0	
<i>NAT1</i>	0.9	1.2	0.8	1.1	0.6	1.1	1.1	0.8	1.2	0.9	0.9	
<i>NAT2</i>	0.9	1.3	0.8	1.2	0.8	0.9	1.1	0.8	1.0	1.1	0.9	
<i>COMT</i>	0.8	0.9	0.8	0.9	1.0	0.8	0.8	0.9	1.2	0.6 ¹⁰	0.9	0.5 ¹⁰
DNA repair												
Base excision repair												
<i>APE1 148</i>	1.1	1.2	1.1	1.2	1.2	1.1	1.1	1.3	0.9	1.2	1.1	1.3
<i>hOGG1</i>	1.0	0.9	1.1	1.0	1.4	1.1	0.9	0.9	1.1	1.1	1.1	0.9
<i>MYH 324</i>	1.0	0.8	1.1	0.9	1.0	1.0	1.0	1.0	0.9	1.0	1.0	1.0
<i>XRCC1 194</i>	1.1	1.1	1.1	1.2	1.4	0.9	1.1	1.2	1.0	1.2	1.1	1.6
<i>XRCC1 280</i>	0.9	0.9	0.9	0.9	0.8	0.9	1.0	0.9	1.0	0.8	0.9	
<i>XRCC1 399</i>	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												
<i>BRCA2 24</i>	0.9	0.9	1.0	0.9	0.9	1.0	0.9	0.8	1.0	0.9	0.9	0.8
<i>BRCA2 372</i>	1.2 ¹⁰	1.2	1.2	1.3	1.0	1.5 ¹⁰	1.3 ¹⁰	1.1	1.2	1.6 ¹⁰	1.2	1.6 ¹⁰
<i>NBS1 185</i>	1.2	1.0	1.2	1.1	1.2	1.3	1.1	1.1	1.2	1.2	1.2	1.0
<i>XRCC2 188</i>	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.8	0.8	1.1	0.9	1.1
<i>XRCC3 241</i>	0.9	1.2	0.8	1.1	0.8	0.8	1.1	0.9	0.9	0.8	0.9	0.8
<i>XRCC4 -28073</i>	1.2	1.2	1.1	1.3	1.1	1.0	1.3	1.1	1.2	1.2	1.1	1.5

Table V.B.8. Misspecification for Gene variant-smoking status associations (OR_z^1) in the CBCS and NCCCS (continued)

Gene pathway ⁶ / Gene variant ⁵	Status				Duration			Intensity			Pack-years ⁹	
	Ever ³	Current ⁴	Former	Current	<=10 years	11-20 years	>20 years	<1/2 pk/day	1/2 - 1 pk/day	>1 pk/day	<=35 PY	>35 PY
CBCS ^{1,2}												
Mismatch repair												
<i>MGMT</i> 84	0.9	0.8	1.1	0.8	1.3	1.0	0.7 ¹⁰	1.1	0.8	0.9	1.0	0.5 ¹⁰
Nucleotide excision repair												
<i>ERCC1</i> 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
<i>ERCC6</i> 1213	1.2	1.6 ¹⁰	1.0	1.6 ¹⁰	1.2	1.2	1.2	1.3	1.3	1.0	1.3 ¹⁰	0.8
<i>ERCC6</i> 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
<i>HRAD23B</i>	1.1	1.2	1.0	1.2	1.2	0.9	1.2	1.0	1.1	1.2	1.1	1.2
<i>XPC</i> 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
<i>XPD</i> 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
<i>XPD</i> 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
<i>XPF</i> 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
<i>XPF</i> 662	1.1	1.4	1.0	1.3	1.1	1.4	1.0	1.4	0.9	0.9	1.2	
<i>XPG</i> 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												
<i>CDH1</i>	0.8	0.8	0.8	0.8	0.9	0.7 ¹⁰	0.8	0.9	0.8	0.7 ¹⁰	0.8 ¹⁰	0.9
<i>TGFB1</i>	1.1	0.8	1.2	0.9	1.3	1.1	1.0	1.2	1.1	0.9	1.1	0.8
<i>MnSOD</i>	1.0	0.9	1.1	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.4
<i>MPO</i>	1.0	1.0	1.0	1.0	0.9	1.1	1.1	1.0	1.0	1.1	1.0	1.1
<i>NQO1</i>	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 ^{1,2}												
Xenobiotic metabolism ⁶												
<i>GST</i> hap C ⁷	1.2	1.0	1.3	1.2	1.5	1.1	1.2	1.7 ¹⁰	1.0	1.2	1.3	1.0
<i>GST</i> hap A ⁸	1.4	1.5	1.3	1.7	1.6	1.3	1.4	1.8 ¹⁰	1.2	1.3	1.5	1.2
<i>GST</i> hap B ⁸	0.8	0.6	0.9				0.7	1.1	0.5 ¹⁰	0.9	0.8	
<i>GST</i> hap D ⁸	1.3	0.9	1.4	1.1	1.7	1.1	1.2	1.9 ¹⁰	1.0	1.3	1.4	1.0
<i>GSTM1</i>	1.0	0.7	1.2	0.8	1.3	0.9	1.0	1.4	0.8	1.1	1.1	0.9
<i>GSTT1</i>	1.0	1.1	0.9	1.1	1.0	1.0	0.9	1.1	0.8	0.9	1.0	0.9

Table V.B.8. Misspecification for Gene variant-smoking status associations (OR_z¹) in the CBCS and NCCCS (continued)

Gene pathway ⁶ / Gene variant ⁵	Status				Duration			Intensity			Pack-years ⁹	
	Ever ³	Current ⁴	Former	Current	<=10 years	11-20 years	>20 years	<1/2 pk/day	1/2 - 1 pk/day	>1 pk/day	<=35 PY	>35 PY
NCCCS ^{1,2}												
Xenobiotic metabolism ⁶ (continued)												
MEH 113	0.7	0.8	0.8	0.7	0.7	0.7	0.8	0.9	0.7	0.6 ¹⁰	0.8	0.7 ¹⁰
MEH 139	0.8	0.6 ¹⁰	0.9	0.6 ¹⁰	1.0	0.8	0.7 ¹⁰	0.9	0.8	0.6 ¹⁰	0.8	0.6 ¹⁰
DNA repair												
POLD1 119	0.7	0.8	0.7	0.7	0.7	0.9	0.7 ¹⁰	0.6 ¹⁰	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

Table V.B.8. Misspecification for Gene variant-smoking status associations (OR_z¹) in the CBCS and NCCCS (continued)

Gene pathway ⁶ / Gene variant ⁵	Status				Duration			Intensity			Pack-years ⁹	
	Ever ³	Current ⁴	Former	Current	<=10 years	11-20 years	>20 years	<1/2 pk/day	1/2 - 1 pk/day	>1 pk/day	<=35 PY	>35 PY
NCCCS ^{1,2}												
Other												
<i>MNSOD</i>	1.0	1.1	1.0	1.1	1.0	1.4	0.9	1.1	1.0	0.9	1.1	0.8

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

¹ Odds ratio not displayed if 95% confidence limit ratio (upper limit/lower limit) >4

² Odds ratios are race and age-adjusted for CBCS; race, age and gender-adjusted for NCCCS

³ Referent is never smokers for all smoking categories unless otherwise noted

⁴ Referent is not-current smokers (former + never)

⁵ SNP referent = homozygous for common allele, compared to heterozygotes + homozygous for less common alleles, *GSTM1* & *GSTT1* referent=present

⁶ Primary functional category; gene may function in additional pathways eg *COMT* in estrogen metabolism

⁷ *GST* hap C = haplotype of *GSTT1* present & *GSTM1* present (referent) vs. all other *GSTT1* & *GSTM1* combinations of present and null combined

⁸ *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

⁹ Pack-years= midpoint of category for number of years smoked x midpoint of category for number of packs smoked/day

¹⁰ Statistically significant at alpha=0.05

Bold = OR_z significant at alpha=0.05

= OR_z <=0.7

= OR_z >=1.4

= 0.7<OR_z<1.4 and significant at alpha=0.05

Table V.B.9.
Agreement between CBCS and NCCCS gene variant-smoking associations

		Kappa ¹	95% CI		N
Full CBCS and NCCCCS					
Null=OR _z : 0.9-1.1	CLR<4	-0.07	-0.19	0.06	165
Restricted CBCS and NCCCCS: 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 ²	0.16	0.02	0.30	163
Null=OR _z : 0.8-1.2	CLR<20	0.20	0.09	0.31	163

Abbreviations: 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 with CLR<5

¹ Weighted kappa statistic

² At CLR>20 all data was included except subgroups with empty cells

VI. CONCLUSIONS AND DISCUSSION

A. Findings and implications for stand-alone case-only studies

For the interaction estimate from a case-only study (COR) to be equal to the interaction estimate from a case-control study (SIM) in the same population, there must be no association between the relevant exposures in the source population ($Z=1$). The COR will be biased to the degree that Z , or a proxy for Z such as the odds ratio from a control group (OR_z), is not equal to one. This assumption, that the interacting exposures analyzed in a case-only study are independent in the population at risk ($Z=1$), is commonly called the independence assumption. The overall goal of the dissertation was to examine the case-only independence assumption in two different types of empirical control data, study-level data found in the literature and individual-level data from two population-based control groups. Three main conclusions emerged from the results of the studies detailed in Chapters V-A and V-B. First, the heterogeneity in OR_z across studies is too great to warrant the assumption that $Z = 1$ for the studied DNA repair gene variants and smoking measures.

Results from the systematic review and meta-analysis of DNA repair SNPs and smoking behavior (Chapter V-A) showed substantial variation in OR_z across the 55 included studies for all SNP-smoking OR_z s. The magnitude of many individual study OR_z s was sufficient to bias the COR to an unacceptable degree (moderate magnitude OR_z defined as $OR_z \leq 0.7$ or ≥ 1.4). The proportion of studies with at least one moderate magnitude gene-smoking OR_z ranged from 0.38 (*XRCC1* 280) to 0.63 (*XRCC3* 241). In addition, *XRCC1* 399 / ever-never smoking and *XPB* 751 / PY were too heterogeneous for summary estimates [ranges, OR (95% CI): 0.7 (0.4, 1.2) – 1.9 (1.2,

2.8) and 0.8 (0.5, 1.3) – 2.3 (0.8, 6.1), respectively). Even when studies were relatively homogeneous (p-value for Cochran's $Q > .10$), OR_z s for the studies in the meta-analysis varied from 2- to 5-fold. Further, nearly all SNP-smoking combinations had OR_z s both above and below the null. These results show that it is insufficient to look at a one or a small number of control groups to assess the potential magnitude or direction of bias that may be introduced into a case-only interaction estimate (COR) when the independence assumption is violated.

Our analysis of study characteristics in Chapter V-A suggests that the independence assumption must be evaluated in a population-specific manner unless there is clear evidence that Z is reliably close to the null across multiple populations. An important step prior to conducting a case-only study is deciding what ancillary control data is most appropriate for evaluating the independence assumption. For instance, if certain study design characteristics can be identified *a priori* that are more valid for evaluating the independence assumption (e.g. population-based *versus* hospital controls), then only studies with that characteristic should be used to evaluate the independence assumption. Analysis of study characteristics can clarify whether or not population-based studies are homogeneous within specific strata even when there is overall heterogeneity or heterogeneity across other strata. Study characteristics chosen *a priori* as potentially influencing the magnitude or heterogeneity of OR_z may also be used to identify situations where the OR_z s vary across strata (e.g. male participants, female participants or mixed gender studies). However, in our data, no study characteristic was identified as a major source of heterogeneity. Study outcome (lung vs. other cancer), study design (population-based vs. hospital/patient-based controls), and average age of study participants were suggestive but did not show consistent correlations with OR_z values. Therefore, there were no study characteristics that stood out strongly enough to be a reliable guide for decision-making.

Subgroup analysis of the CBCS and NCCCS control groups in Chapter V-B supported the meta-analysis study characteristics analysis. There was little variation in OR_z across strata of age, race or gender, consistent with the uninformative nature of the study-level variables average age, ethnicity or gender proportion in Chapter V-A. Results from the meta-analysis suggest it is inappropriate to estimate Z using a limited number of control groups with similar study characteristics. Unless further research identifies new study characteristics that may be influential, a broad sample of studies is necessary to determine the likely range of bias that may be introduced by the unmeasured Z .

The second conclusion is that heterogeneity of OR_z s across smoking measures precludes the use of one measure of smoking (e.g. ever-never smoking) to evaluate the independence assumption, particularly when analyses of multiple smoking measures (e.g. dose, duration) are planned in a case-only study. Results from Chapters V-A and V-B support this conclusion. The CBCS/NCCCS control group analyses confirmed the variability in OR_z values across measures of smoking within two population-based control groups. Consequently, the independence assumption needs to be assessed for each exposure measure that will be used in the case-only analysis. Taken together, our results do not support the independence of DNA repair SNPs and smoking behavior, either across studies or smoking measures, nor do they support the independence of the xenobiotic metabolism genes *CYP1A1*, *GSTM1*, *GSTP1*, *NAT1*, *COMT* or *MEH* and smoking across smoking measures.

The third conclusion is that no strong patterns were apparent when genes were categorized by the biological pathways in which they participate. Neither study showed substantial clustering of moderate magnitude OR_z s by gene category, with 25% and 18% of the xenobiotic metabolizing genes and DNA repair genes, respectively, consistently null across all categories and both studies.

There was, however, a suggestion that xenobiotic metabolism genes were more likely than DNA repair genes to exhibit variants with multiple moderate magnitude gene-smoking OR_{zs} . Our results clearly indicate that the independence assumption must be assessed for each gene variant, rather than by gene category, for the proposed case-only analysis to be valid.

Taken together, these studies show that a systematic approach to assessing the independence association is essential prior to conducting a case-only analysis of gene-smoking interaction. Results from both studies showed that moderate magnitude OR_{zs} (≥ 1.4 or ≤ 0.7), sufficient to cause bias of $>10\%$ in the COR, occur in numerous control groups and for multiple measures of smoking, particularly measures of smoking duration, intensity or PY. This systematic approach should include conducting a thorough literature search for published systematic reviews, studies of appropriately pooled data, and studies with relevant control data as this information is necessary to establish at least the likely range of OR_{zs} . A sensitivity analysis should be conducted with available data. In addition, searching for information on the target population of interest is critical. In the absence of appropriate population-specific data, a validation study should be considered. Further, results show that assessing the independence assumption only for ever-never smoking is insufficient, if the proposed case-only analysis is to include any other measures of smoking behavior. At present, control group data on other smoking measures can be difficult or impossible to find in the published literature.

Finally, smoking interaction is an area of interest for many non-cancer outcomes (e.g. cardiovascular disease) and many of the genes assessed in the current project are not cancer-specific (e.g. *COMT* and *CYP1A1*). The utility and implications of these results are not limited to cancer research.

B. Strengths of the systematic review and meta-analysis

There are several notable strengths to the systematic review and meta-analysis of DNA repair genes and smoking presented in Chapter V-A. The literature search was extensive, using three publicly available databases, and facilitated by consultation with an information specialist. Sample size was relatively large (N=55 studies overall), despite the paucity of published control group data on the joint distribution of DNA repair genotypes and smoking. For some SNPs there were sufficient data to examine smoking duration and intensity, the components of PY. This is important since the smoking measure of interest for a proposed case-only study is likely be something other than ever-never smoking, the crudest and most commonly presented measure of smoking for control groups in the literature. Further, for most studies that presented data on smoking amount, we were able to construct at least one measure of smoking status and compare results within the same study, confirming that the direction and magnitude of OR_z for smoking status and smoking amount often differ within the same study population.

The large number of studies increased our ability to detect and investigate heterogeneity between studies and improved precision when study results were sufficiently homogeneous to warrant summary estimates. We were able to examine numerous study characteristics (continent, ethnicity, average age, proportion male, HWE p-value, study outcome, minor allele frequency, and smoking prevalence) using a variety of metrics, in an effort to discover the source(s) of heterogeneity. Although no characteristic was shown to be a major source of heterogeneity, this in itself is useful information for investigators with only published or ancillary data for evaluation of the independence assumption.

Visual inspection and tests of funnel plot asymmetry supported our expectation that publication bias would be minimal, given that control group associations are not generally considered when publication decisions are made.

C. Limitations of the systematic review and meta-analysis

Because no individual level data on controls was available in the literature, only unadjusted estimates of OR_z could be calculated from published control group data using the joint distribution of genotype and smoking. Consequently, OR_z s could be confounded. Additionally, results only apply at the level of the study, even when a characteristic is an individual level variable at collection, such as age or gender. This is analogous to the ecologic fallacy. For instance, conclusions that apply to studies with a higher average age for participants do not necessarily apply to the older participants within that study, and in fact could be due entirely to the younger individuals in the study.

Some study characteristics were difficult to determine accurately from published reports, in particular age and ethnicity. Although a consistent rubric for central tendency of age (“average” age) was applied, and all studies but one gave some indication of participant age, the level of detail varied widely and some studies are likely to be misclassified. Participants’ ethnicity was often not reported. However, results did not differ according to study-reported ethnicity, ethnicity assigned by continent and ethnicity assigned by MAF. Since participation can vary by smoking status [117-120], it would have been informative to have included response rates as a study characteristic. However, too few studies presented comparable data on response rates to assess this characteristic.

D. Strengths of the control group analyses

The strengths of the control group analysis of genetic variation and smoking using CBCS and NCCCS data presented in Chapter V-B complemented the strengths of the systematic review and offset some of its limitations. Both the CBCS and NCCCS control groups are large and population-based. The target population for the independence assumption is the population from which the cases arise. In these two studies, the control groups come from essentially the same underlying population, albeit from partially overlapping geographic areas and time periods. The studies used the same sampling scheme, oversampling African Americans using DMV records and HCFA lists, so that there was adequate sample size to perform subgroup analysis by race.

Because gene-smoking interaction is of interest in breast and colon cancer, there were sufficient data to examine multiple genes in different metabolic pathways plausibly related to smoking behavior. The CBCS and NCCCS had data for 38 and 25 relevant gene variants, respectively. Fifteen of the variants were assayed in both studies so agreement between studies with essentially the same underlying population could be examined. Additionally, the gene variants were in several different pathways, primarily xenobiotic metabolizing genes and DNA repair genes, giving additional insight into whether genes in a common pathway might have similar associations with smoking.

The sample size, level of detail in the smoking information and inclusion of more than one ethnic group are notable strengths of this study. Confounding by race, age, gender, family history of any cancer and family income were also evaluated. Confounder evaluation is important since only unadjusted OR_z s are available from the published literature and the independence assumption applies to the unconfounded G-E association. Both studies had detailed data on smoking behavior so OR_z could be calculated for multiple metrics for smoking status (ever-never and current-not

current smoking) and amount (duration, intensity and PY). In contrast, for the meta-analysis no individual level data was available, nor could OR_z be calculated for all five smoking measures for any single study.

E. Limitations of the control group analyses

Selection bias in the CBCS or NCCCS could have biased OR_z s. If non-response by eligible controls was associated with smoking behavior and with gene status, OR_z s could be biased in an unpredictable direction; however, prevalence of current smoking in the CBCS was similar to that in North Carolina during this time period. Participants are very unlikely to know their genotype, although it is possible that knowing one's family history of cancer (a crude proxy for genotype) could affect participation.

Some misclassification of smoking behavior is likely and could have affected results. Smoking data is self-reported; there are no biological measures of smoking behavior in the CBCS or NCCCS. Given that the negative health effects of smoking are well known, it is unlikely that controls would over-estimate their tobacco consumption. If a proportion of current smokers were misclassified as not current smokers (former+never), and misclassification was non-differential by genotype, OR_z s for current smoking would be biased away from the null. However, if smoking misclassification is similar to estimates in the literature for the general population (<2%-13%) [284-287], the magnitude of bias was small at the OR_z s, smoking and genotype prevalences typical in CBCS and NCCCS control groups. For measures of smoking with more than two categories, such as current/former/never, duration or intensity, the direction of bias is unpredictable. If misclassification was differential by genotype, the direction of bias is unpredictable. However, participants are unaware of their genotypes; therefore it is unlikely that any misclassification is differential by genotype.

Population stratification could have caused some residual confounding despite adjustment for race. Stratification by race did not reveal any systematic differences in OR_{zs} however.

The gene variants included in this study were a convenience sample of genetic data from parent studies of cancer rather than candidate genes from a study designed to investigate smoking behavior. As such, genes thought to be important in smoking behavior, but not in breast or colon cancer, could not be assessed.

Precise gene function is unknown for the majority of gene variants. This limits interpretation of gene-smoking associations but is a limitation common to many studies of genetic exposures at this point in time. Further, gene variants could be in linkage disequilibrium with causal variants, rather than being the true causal variant. Finally, chance could have played a role in the associations found. However, using strict criteria for confidence limit ratios reduced the number of imprecise estimates considered, therefore reducing the role of chance in these results.

F. Future directions

In the short term, it would be extremely useful to have more detailed control group information publicly available from large population-based studies for a variety of genes and exposures. Any environmental factor (i.e. non-genetic factor) whose effect might be modified by genetic variation should be included in the accessible databases. This should include, but not be limited to smoking, alcohol consumption, BMI, air pollution levels, occupational exposures and birth/prenatal exposures such as birthweight, and common medication use such as NSAIDS. Just as ever-never smoking OR_{zs} did not predict OR_{zs} for smoking amounts, this data must be available at approximately the same level of detail as the proposed case-only analyses to be a valid test of the independence assumption.

At present, control group data on smoking measures other than ever-never can be difficult or impossible to find in the published literature. The results from the CBCS and NCCCS analysis confirm that this information often differs from OR_z for smoking status. It would be useful to have more detailed data on smoking amounts (duration, intensity, time since cessation for former smokers, age of initiation, nicotine dependence indices, etc.) than is usually presented, ideally stratified by race, age and gender. Given that many studies already collect much 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. Other important information (hospital-based vs. population-based, inclusion or exclusion criteria, response rates) is already given in most papers, although the data on response rates would need to be presented in some standard format to be useful. As mentioned, there are numerous other potentially useful data that could be made available for more rigorous evaluation of the independence assumption for case-only studies.

Long term, population-based studies specifically designed to address gene-smoking associations are needed. To be most useful for evaluating the independence assumption, additional smoking phenotypes should be included, ideally including biological measures of current smoking status, and a broader panel of SNPs. SNPs chosen for this purpose should plausibly be of interest for gene-smoking interaction in disease, rather than only those genes currently being studied for their likely influence on smoking behavior. For instance, well-designed studies focused on specific aspects of smoking behavior abound but rarely include DNA repair genes, a class of genes of great interest for gene-smoking interaction in cancer (etiology and treatment), heart disease and neurological diseases. Recent genome-wide association studies (GWAS) have raised the possibility that the search for the genetic underpinnings of all parts of the smoking trajectory, from initiation to dependence to cessation, may need to be broadened [288].

Designing gene-smoking studies to accommodate genes of interest in additional fields of wide public health impact such as cancer and heart disease would be an efficient use of scientific resources. As suggested by [289] genotyping chips would facilitate such multi-purpose studies if SNPs relevant to gene-smoking interaction were routinely included. These studies would serve the purposes of elucidating the etiology of tobacco dependence and cessation, while decreasing the number of case-only studies with unacceptable levels of bias and improving the accuracy of estimates of interaction from case-only studies.

VII. APPENDICES

A. Informed consent

Informed consent was obtained for the CBCS and NCCCS parent studies. There was no further participant contact or information gathering. This analysis was exempted as “not human subjects research” by the Public Health Institutional Review Board (IRB) Dec 22, 2005. IRB number: 05-2821.

B. Supplementary tables and figures: Manuscript 1

Table VIII.B.1. Association between *XRCC1* Arg399His and smoking : Individual study results

Author and Year	Never/ever	Current/Not current	Pack-years ¹	Intensity ²	Duration ³
	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)
Cao 2006		0.99 (0.68, 1.45)			
David-Beabes 2001	All: 0.78 (0.57,1.07) Wh: 0.87 (0.59, 1.28) AA: 0.69 (0.39, 1.22)			1.52 (1.05, 2.22)	
Duell 2001 ⁴	Wh: 1.16 (0.77, 1.73) AA: 1.65 (0.95, 2.88) All: 1.48 (1.08, 2.02)	Wh: 1.10 (0.67, 1.81) AA: 2.08 (1.11, 3.91) All: 1.35 (0.93, 1.98)			---
Duell 2002					0.86 (0.59, 1.24)
Harms 2004			0.53 (0.20, 1.37)		
Hoffmann 2005		0.43 (0.15, 1.19)			
Huang 2005a	1.86 (1.23, 2.80)		0.90 (0.49, 1.67)		
Hung 2005b	1.04 (0.86, 1.25)		1.24 (0.92, 1.69)		
Ito 2004	0.80 (0.55, 1.18)	0.88 (0.58, 1.32)	2.02 (1.21, 3.40)		
Kelsey 2004	0.90 (0.62, 1.31)				
Kocabas 2006		0.87 (0.46, 1.67)			
Koyama 2005		1.55 (0.65, 3.72)			
Lei 2002		0.93 (0.34, 2.55)		0.36 (0.05, 2.34)	
Lunn 1999		0.88 (0.23, 3.49)			

Table VIII.B.1. Association between *XRCC1* Arg399His and smoking : Individual study results (continued)

Author and Year	Never/ever	Current/Not current	Pack-years¹	Intensity²	Duration³
	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)
Matullo 2001b	1.11 (0.70, 1.76)	1.14 (0.68, 1.91)			
Matullo 2005	1.09 (0.67, 1.77)	0.80 (0.50, 1.29)			
Metsola 2005	0.75 (0.50, 1.13)		0.91 (0.45, 1.84)		
Misra 2003				1.41 (0.76, 2.62)	
Olshan 2002	1.16 (0.61, 2.21)				
Pachkowski 2006 ⁴	1.21 (1.00, 1.46)	1.22 (0.96, 1.54)		1.61 (1.15, 2.27)	1.48 (1.07, 2.05)
Park 2002			1.52 (0.65, 3.55)		
Patel 2005	0.88 (0.60, 1.28)				
Ramachandran 2006	0.72 (0.31, 1.69)				
Ryk 2006b	0.84 (0.44, 1.62)				
Schneider 2005	1.23 (0.85, 1.77)		0.77 (0.23, 2.60)		
Shen 2000	1.93 (1.00, 3.72)				
Shen 2003	1.53 (0.82, 2.86)		1.24 (0.66, 2.34)		
Shen 2005a	1.57 (1.24, 2.00)				
Skelbred 2006a		0.74 (0.50, 1.08)			
Tuimala 2004	0.83 (0.43, 1.61)				
Wilding 2005	1.11 (0.69, 1.78)				
Yu 2004a	1.09 (0.57, 2.08)				
Zhou 2003	0.97 (0.77, 1.23)		1.40 (0.93, 2.10)		

Abbreviations: OR_z=control-only genotype-smoking odds ratio, 95% CI= 95% confidence interval, ref=referent, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ Pack-years: lightest non-zero category (ref) vs. heaviest

² Intensity (packs/day): lightest non-zero category (ref) vs. heaviest

³ Duration (years): shortest non-zero category (ref) vs. longest

⁴ Duell 2001 used in stratified analyses; Pachkowski 2006 used for non-stratified analyses

Table VIII.B.2.**Association between *XRCC1* Arg194Trp and smoking : Individual study results**

Author and Year	Never/ever	Current/Not current	Pack-years ¹	Intensity ²	Duration ³
	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)
Cao 2006		0.83 (0.57, 1.22)			
David-Beabes 2001	All: 1.28 (0.79, 2.06) Wh: 1.37 (0.74, 2.54) AA: 1.14 (0.53, 2.44)			1.13 (0.66, 1.93)	
Han 2003	1.11 (0.81, 1.53)				0.68 (0.40, 1.18)
Hung 2005b	0.97 (0.75, 1.26)		1.1 (0.73, 1.64)		
Koyama 2005		0.91 (0.38, 2.19)			
Lunn 1999		2.08 (0.10, 41.62)			
Matullo 2005	0.98 (0.52, 1.84)	1.14 (0.62, 2.11)			
Olshan 2002	1.07 (0.45, 2.53)				
Pachkowski 2006	1.13 (0.86, 1.48)	1.14 (0.81, 1.60)		1.08 (0.67, 1.73)	0.74 (0.47, 1.15)
Patel 2005	1.20 (0.72, 2.01)				
Schneider 2005	0.60 (0.36, 1.00)		0.76 (0.09, 6.12)		
Shen 2000	0.87 (0.46, 1.66)				
Shen 2005a	0.85 (0.59, 1.21)				
Skelbred 2006a		1.45 (0.80, 2.64)			
Stern 2001	1.20 (0.57, 2.56)				0.34 (0.11, 1.07)
Wilding 2005	0.68 (0.33, 1.39)				

Abbreviations: OR_z=control-only genotype-smoking odds ratio, 95% CI= 95% confidence interval, ref=referent, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ Pack-years: lightest non-zero category (ref) vs. heaviest

² Intensity (packs/day): lightest non-zero category (ref) vs. heaviest

³ Duration (years): shortest non-zero category (ref) vs. longest

Table VIII.B.3.**Association between *XRCC1* Arg280His and smoking: Individual study results**

Author and Year	Never/ever	Current/Not current	Pack-years¹	Intensity²	Duration³
	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)
Hung 2005b	1.09 (0.79, 1.50)		1.19 (0.73, 1.93)		
Koyama 2005		1.06 (0.19, 5.81)			
Lunn 1999		0.38 (0.03, 4.69)			
Metsola 2005	0.87 (0.49, 1.54)		0.54 (0.20, 1.50)		
Pachkowski 2006	0.95 (0.67, 1.35)	0.88 (0.55, 1.38)		0.93 (0.48, 1.81)	1.20 (0.64, 2.26)
Schneider 2005	1.06 (0.57, 2.00)		0.40 (0.02, 6.97)		
Skelbred 2006a		0.46 (0.21, 1.02)			
Tuimala 2004	0.45 (0.18, 1.11)				

Abbreviations: OR_z=control-only genotype-smoking odds ratio, 95% CI= 95% confidence interval, ref=referent, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ Pack-years: lightest non-zero category (ref) vs. heaviest

² Intensity (packs/day): lightest non-zero category (ref) vs. heaviest

³ Duration (years): shortest non-zero category (ref) vs. longest

Table VIII.B.4.**Association between *XPD* Lys751Gln and smoking : Individual study results**

Author and Year	Never/ever	Current/Not current	Pack-years ¹
	OR _z (95% CI)	OR _z (95% CI)	OR _z (95% CI)
Harms 2004			2.27 (0.84, 6.11)
Hoffmann 2005		1.57 (0.58, 4.25)	
Hou 2002	1.40 (0.75, 2.62)		
Huang 2005a	1.11 (0.73, 1.69)		0.78 (0.42, 1.44)
Jiao 2007b	1.13 (0.75, 1.71)		
Matullo 2001b	0.97 (0.60, 1.57)	1.25 (0.72, 2.17)	
Matullo 2005	0.86 (0.51, 1.44)	0.74 (0.46, 1.22)	
Metsola 2005	1.26 (0.82, 1.93)		0.90 (0.43, 1.88)
Schabath 2005	0.89 (0.62, 1.29)		0.75 (0.45, 1.25)
Shen 2003	0.52 (0.26, 1.03)		1.31 (0.70, 2.46)
Skelbred 2006a		1.21 (0.82, 1.80)	
Terry 2004	0.78 (0.61, 0.99)	0.95 (0.70, 1.29)	
Xing 2002a	0.91 (0.50, 1.63)		0.77 (0.34, 1.73)
Yu 2004b	1.35 (0.52, 3.54)		
Zhou 2002	0.93 (0.74, 1.19)		2.15 (1.38, 3.33)
Zinjo 2006		1.91 (0.96, 3.81)	

Abbreviations: OR_z=control-only genotype-smoking odds ratio, 95% CI= 95% confidence interval,
 ref=referent, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine,
 Asp=Aspartic acid, Asn=Asparagine

¹ Pack-years: lightest non-zero category (ref) vs. heaviest

Table VIII.B.5.**Association between *XPD* Asp312Asn and smoking : Individual study results**

Author and Year	Never/ever	Current/Not current	Pack-years ¹
	OR _z (95% CI)	OR _z (95% CI)	OR _z (95% CI)
Butkiewicz 2001	1.04 (0.42, 2.58)		0.63 (0.21, 1.91)
Garcia-Closas 2006	1.04 (0.80, 1.34)		
Hou 2002	1.48 (0.79, 2.78)		
Jiao 2007b	1.16 (0.77, 1.74)		
Justenhoven 2004	1.31 (0.95, 1.81)		
Matullo 2005	1.30 (0.79, 2.16)	1.12 (0.68, 1.86)	
Schabath 2005	1.01 (0.71, 1.45)		0.86 (0.52, 1.42)
Xing 2002a	0.79 (0.42, 1.47)		0.66 (0.27, 1.61)
Zhou 2002	0.98 (0.78, 1.25)		1.58 (1.05, 2.40)

Abbreviations: OR_z=control-only genotype-smoking odds ratio, 95% CI= 95% confidence interval,
 ref=referent, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine,
 Asp=Aspartic acid, Asn=Asparagine

¹ Pack-years: lightest non-zero category (ref) vs. heaviest

Table VIII.B.6.**Association between *XRCC3* Thr41Met and smoking : Individual study results**

Author and Year	Never/ever	Current/Not current	Pack-years¹
	OR _z (95%CI)	OR _z (95%CI)	OR _z (95%CI)
Harms 2004			1.06 (0.41, 2.73)
Hoffmann 2005		0.63 (0.22, 1.77)	
Huang 2005a	1.24 (0.82, 1.85)		0.79 (0.43, 1.44)
Jin 2005		2.04 (0.63, 6.60)	
Matullo 2001b	1.40 (0.87, 2.25)	1.03 (0.60, 1.77)	
Matullo 2005	1.11 (0.68, 1.81)	0.94 (0.58, 1.52)	
Shen 2002	0.70 (0.44, 1.14)	0.78 (0.49, 1.23)	
Shen 2003	0.83 (0.43, 1.63)		1.00 (0.52, 1.92)
Skelbred 2006a		0.82 (0.56, 1.20)	
Smedby 2006	0.87 (0.61, 1.23)	1.03 (0.70, 1.51)	
Stern 2002a	1.36 (0.78, 2.40)		0.58 (0.29, 1.18)
Tuimala 2004	1.03 (0.51, 2.08)		
Wilding 2005	1.06 (0.66, 1.70)		

Abbreviations: OR_z=control-only genotype-smoking odds ratio, 95% CI= 95% confidence interval, ref=referent, Arg=Arginine, Gln=Glutamine, Trp=Tryptophan, His=Histidine, Met=methionine, Asp=Aspartic acid, Asn=Asparagine

¹ Pack-years: lightest non-zero category (ref) vs. heaviest

C. Supplementary tables and figures: Manuscript 2

Table VIII.C.1. Genotype prevalence and Hardy-Weinberg equilibrium in CBCS and NCCCS

SNP	CBCS								NCCCS							
	Non-African American				African American				Non-African American				African American			
	No var ¹	Any var ¹	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value
<i>ADPRT</i> 762	--	--	--	--	--	--	--	--	386	153	28.4%	0.94	295	30	9.2%	0.15
<i>ADPRTL2</i> 328	--	--	--	--	--	--	--	--	226	302	57.2%	0.99	29	285	90.8%	0.68
<i>APEI</i> 148	300	836	73.6%	0.41	251	426	62.9%	0.89	153	387	71.7%	0.21	116	208	64.2%	0.71
<i>BRCA2</i> 24	695	439	38.7%	0.08	408	268	39.6%	0.28	--	--	--	--	--	--	--	--
<i>BRCA2</i> 372	579	556	49.0%	0.7	510	165	24.4%	0.89	--	--	--	--	--	--	--	--
<i>CDH1</i>	610	522	46.1%	0.91	492	182	27.0%	0.28	--	--	--	--	--	--	--	--
<i>COMT</i>	86	293	77.3%	0.92	110	153	58.2%	0.84	--	--	--	--	--	--	--	--
<i>CYP1A1</i> M1	325	90	21.7%	0.53	165	115	41.1%	0.58	--	--	--	--	--	--	--	--
<i>CYP1A1</i> M2	378	39	9.4%	0.3	274	11	3.9%	0.74	--	--	--	--	--	--	--	--
<i>CYP1A1</i> M3	413	2	0.5%	<.001	227	5	2.2%	0.07	--	--	--	--	--	--	--	--
<i>CYP1A1</i> M4	377	40	9.6%	0.3	278	7	2.5%	0.83	--	--	--	--	--	--	--	--
<i>ERCC1</i> 8092	656	478	42.2%	0.71	342	340	49.9%	0.99	--	--	--	--	--	--	--	--
<i>ERCC6</i> 1213	713	417	36.9%	0.75	465	213	31.4%	0.83	--	--	--	--	--	--	--	--
<i>ERCC6</i> 1230	887	244	21.6%	0.41	643	35	5.2%	0.49	--	--	--	--	--	--	--	--
<i>GSTM1</i> ²	177	192	52.0%	--	187	72	27.8%	--	258	289	52.8%	--	245	82	25.1%	--
<i>GSTP1</i>	141	207	59.5%	0.23	54	193	78.1%	0.14	--	--	--	--	--	--	--	--
<i>GSTT1</i> ²	312	61	16.4%	--	216	43	16.6%	--	385	162	29.6%	--	218	109	33.3%	--

Table VIII.C.1. Genotype prevalence and Hardy-Weinberg equilibrium in CBCS and NCCCS (continued)

SNP	CBCS								NCCCS							
	Non-African American				African American				Non-African American				African American			
	No var ¹	Any var ¹	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value
<i>GST</i>																
<i>hap C</i> ^{2,3}	--	--	--	--	--	--	--	--	179	368	67.3%	--	163	164	50.2%	--
<i>GST</i>																
<i>hap A</i> ^{2,3}	--	--	--	--	--	--	--	--	79	179	69.4%	--	82	163	66.5%	--
<i>GST</i>																
<i>hap B</i> ^{2,3}	--	--	--	--	--	--	--	--	83	179	68.3%	--	27	163	85.8%	--
<i>GST</i>																
<i>hap D</i> ^{2,3}	--	--	--	--	--	--	--	--	206	179	46.5%	--	55	163	74.8%	--
<i>hOGG1</i>	652	483	42.6%	0.42	474	204	30.1%	0.45	--	--	--	--	--	--	--	--
<i>MEH 113</i>	--	--	--	--	--	--	--	--	258	288	52.7%	0.37	198	127	39.1%	0.12
<i>MEH 139</i>	--	--	--	--	--	--	--	--	343	203	37.2%	0.17	135	191	58.6%	0.95
<i>MGMT 84</i>	867	269	23.7%	0.12	504	174	25.7%	0.1	--	--	--	--	--	--	--	--
<i>MLH1</i>																
<i>219</i>	--	--	--	--	--	--	--	--	253	286	53.1%	0.83	274	51	15.7%	0.58
<i>MNSOD</i>	266	869	76.6%	0.27	196	481	71.0%	0.08	138	408	74.7%	0.55	105	220	67.7%	0.15
<i>MPO</i>	699	435	38.4%	0.93	296	382	56.3%	0.37	--	--	--	--	--	--	--	--
<i>MSH3</i>																
<i>1036</i>	--	--	--	--	--	--	--	--	287	256	47.1%	0.48	139	183	56.8%	0.73
<i>MSH3 940</i>	--	--	--	--	--	--	--	--	402	145	26.5%	0.26	264	59	18.3%	0.99
<i>MSH6 39</i>	--	--	--	--	--	--	--	--	393	149	27.5%	0.78	207	113	35.3%	0.34
<i>MYH 324</i>	627	505	44.6%	0.19	367	306	45.5%	0.65	--	--	--	--	--	--	--	--
<i>NAT1</i> ²	103	170	62.3%	--	145	47	24.5%	--	--	--	--	--	--	--	--	--
<i>NAT2</i> ²	109	165	60.2%	--	116	79	40.5%	--	--	--	--	--	--	--	--	--
<i>NBS1 185</i>	518	618	54.4%	0.66	400	281	41.3%	0.53	242	293	54.8%	0.39	183	140	43.3%	0.86
<i>NQO1</i>	742	389	34.4%	0.28	457	217	32.2%	0.83	--	--	--	--	--	--	--	--
<i>POLD1</i>																
<i>119</i>	--	--	--	--	--	--	--	--	452	77	14.6%	0.58	171	149	46.6%	0.56
<i>RAD23B</i>	756	377	33.3%	0.52	604	75	11.0%	0.41	335	193	36.6%	<0.01	293	29	9.0%	0.4
<i>TGFB1</i>	457	673	59.6%	0.79	224	451	66.8%	0.19	--	--	--	--	--	--	--	--

Table VIII.C.1. Genotype prevalence and Hardy-Weinberg equilibrium in CBCS and NCCCS (continued)

SNP	CBCS								NCCCS							
	Non-African American				African American				Non-African American				African American			
	No var ¹	Any var ¹	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value	No var	Any var	% with any var	HWE p-value
<i>XPC 499</i>	--	--	--	--	--	--	--	--	300	246	45.1%	0.8	278	47	14.5%	0.16
<i>XPC 939</i>	400	723	64.4%	0.97	338	341	50.2%	0.19	192	350	64.6%	0.74	150	174	53.7%	0.71
<i>XPB 312</i>	489	644	56.8%	0.64	517	158	23.4%	0.45	233	302	56.4%	0.62	259	63	19.6%	0.82
<i>XPB 751</i>	445	688	60.7%	0.53	393	286	42.1%	0.85	212	324	60.4%	0.42	187	135	41.9%	0.86
<i>XPF 415</i>	980	153	13.5%	0.27	642	31	4.6%	0.54	466	81	14.8%	0.046	309	16	4.9%	0.65
<i>XPF 662</i>	249	1	0.4%	0.97	434	240	35.6%	0.2	--	--	--	--	--	--	--	--
<i>XPG 1104</i>	661	472	41.7%	0.69	231	443	65.7%	0.51	341	202	37.2%	0.26	101	218	68.3%	0.53
<i>XRCC1 194</i>	987	148	13.0%	0.43	593	89	13.0%	0.95	477	61	11.3%	0.53	277	43	13.4%	0.2
<i>XRCC1 280</i>	1030	99	8.8%	0.86	642	39	5.7%	0.58	503	44	8.0%	0.27	310	15	4.6%	0.67
<i>XRCC1 399</i>	480	642	57.2%	0.24	493	183	27.1%	0.36	222	318	58.9%	0.95	251	74	22.8%	0.07
<i>XRCC2 188</i>	982	152	13.4%	0.52	653	25	3.7%	0.62	--	--	--	--	--	--	--	--
<i>XRCC3 241</i>	435	697	61.6%	0.09	421	255	37.7%	0.01	206	332	61.7%	0.89	204	120	37.0%	0.96
<i>XRCC4 -28073</i>	244	889	78.5%	0.56	212	463	68.6%	0.2	--	--	--	--	--	--	--	--

Abbreviations: CBCS=Carolina Breast Cancer Study, NCCCS=North Carolina Colon Study, HWE=Hardy Weinberg Equilibrium, var=variant

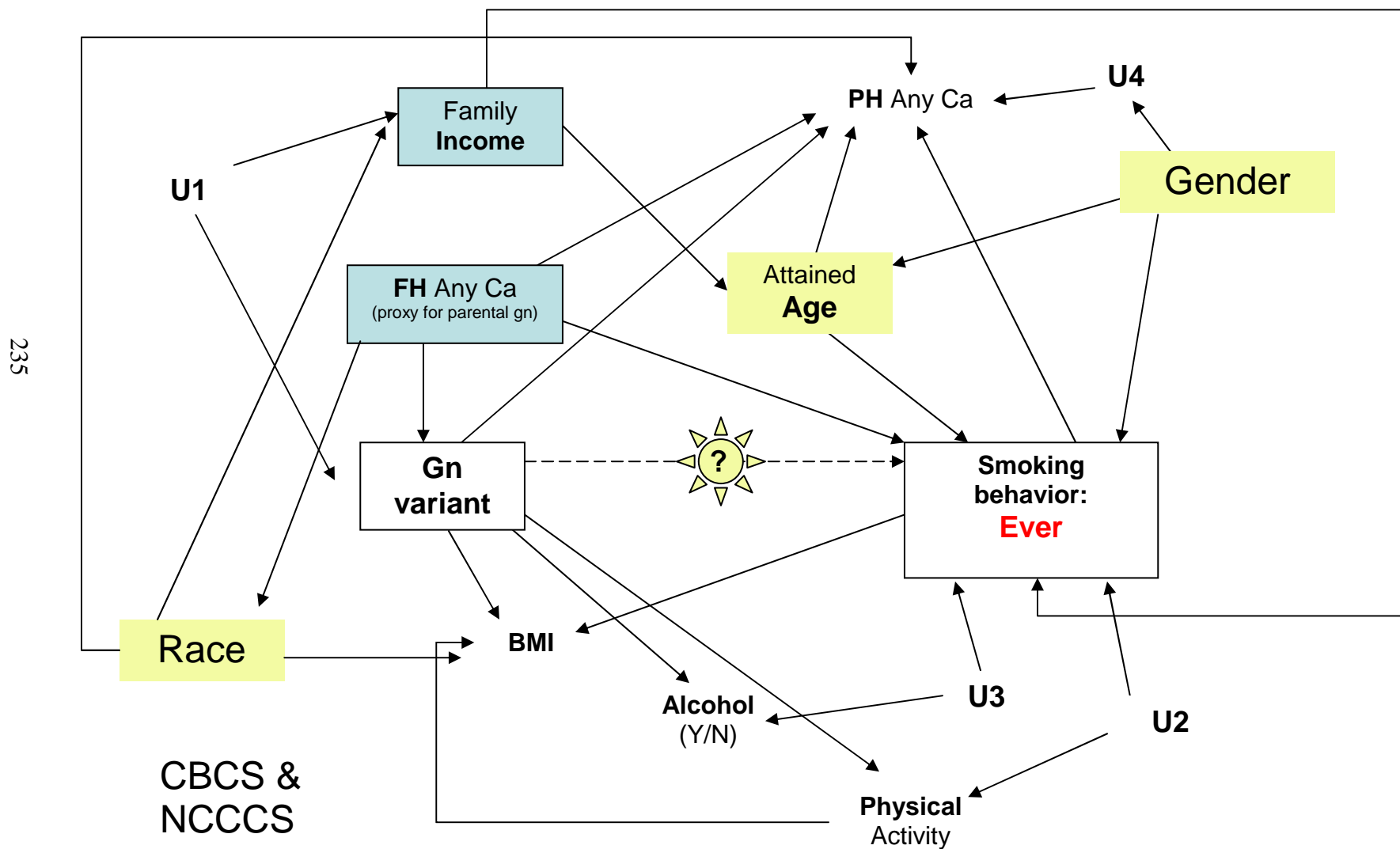
¹ See Table 2 for definitions of common vs. variant alleles for each SNP, common=higher frequency in overall dataset.

² Percent with allele present (present=referent) or not present (null) instead of % no variant allele & % any variant allele

³ p=present, n=null; Haplotypes for *GSTT1* and *GSTM1*: *GST* hap A=*GSTT1*(n)/*GSTM1*(p), *GST* hap B=*GSTT1*(n)/*GSTM1*(n), *GST* hap C = *GSTT1*(p)/*GSTM1*(p) [referent], *GST* hap D=*GSTT1*(p)/*GSTM1*(n)

Bold = p-value <0.05

Figure VIII.B.1. Directed acyclic graph of variable relationships in controls



VIII. REFERENCES

1. Prentice, R.L., W.M. Vollmer, and J.D. Kalbfleisch, *On the use of case series to identify disease risk factors*. Biometrics, 1984. **40**(2): p. 445-58.
2. Piegorsch, W.W., C.R. Weinberg, and J.A. Taylor, *Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies*. Stat Med, 1994. **13**(2): p. 153-62.
3. Khoury, M.J. and W.D. Flanders, *Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls!* Am.J.Epidemiol., 1996. **144**(3): p. 207-213.
4. Garcia-Closas, M., W.D. Thompson, and J.M. Robins, *Differential misclassification and the assessment of gene-environment interactions in case-control studies*. Am.J.Epidemiol., 1998. **147**(5): p. 426-433.
5. Albert, P.S., et al., *Limitations of the case-only design for identifying gene-environment interactions*. American Journal of Epidemiology., 2001. **154**(8): p. 687-693.
6. Smith, P.G. and N.E. Day, *The design of case-control studies: the influence of confounding and interaction effects*. Int.J.Epidemiol., 1984. **13**(3): p. 356-365.
7. Rothman, K.J. and S. Greenland, *Modern Epidemiology*. Vol. Second edition. 1998, Philadelphia: Lippincott-Raven Publishers.
8. Yang, Q., M.J. Khoury, and W.D. Flanders, *Sample size requirements in case-only designs to detect gene-environment interaction*. Am.J.Epidemiol., 1997. **146**(9): p. 713-720.
9. Infante-Rivard, C., et al., *Risk of childhood leukemia associated with exposure to pesticides and with gene polymorphisms*. Epidemiology, 1999. **10**(5): p. 481-487.
10. Bacchetti, P., et al., *Bacchetti et Al. Respond to "ethics and sample size--another view"*. Am.J.Epidemiol., 2005. **161**(2): p. 113.
11. Bacchetti, P., et al., *Ethics and sample size*. Am.J.Epidemiol., 2005. **161**(2): p. 105-110.
12. Goldman, L.R. and J.M. Links, *Testing toxic compounds in human subjects: ethical standards and good science*. Environ.Health Perspect., 2004. **112**(8): p. A458-A459.
13. Prentice, R., *Invited commentary: ethics and sample size--another view*. Am.J.Epidemiol., 2005. **161**(2): p. 111-112.
14. Baksh, M.F., et al., *Design considerations in the sequential analysis of matched case-control data*. Stat.Med., 2004.

15. Boddeker, I.R. and A. Ziegler, *Sequential designs for genetic epidemiological linkage or association studies - A review of the literature*. Biometrical Journal, 2001. **43**(4): p. 501-525.
16. Tan, Q., et al., *A case-only approach for assessing gene by sex interaction in human longevity*. J.Gerontol.A Biol.Sci.Med.Sci., 2002. **57**(4): p. B129-B133.
17. Weinberg, C.R. and D.M. Umbach, *Choosing a retrospective design to assess joint genetic and environmental contributions to risk. [see comments.]. [Review] [32 refs]*. American Journal of Epidemiology., 2000. **152**(3): p. 197-203.
18. Theodoratou, E., et al., *Modification of the inverse association between dietary vitamin D intake and colorectal cancer risk by a FokI variant supports a chemoprotective action of Vitamin D intake mediated through VDR binding*. Int J Cancer, 2008. **123**(9): p. 2170-9.
19. Upadhyay, R., et al., *Functional polymorphisms of cyclooxygenase-2 (COX-2) gene and risk for esophageal squamous cell carcinoma*. Mutat Res, 2009. **663**(1-2): p. 52-9.
20. Chang-Claude, J., et al., *The patched polymorphism Pro1315Leu (C3944T) may modulate the association between use of oral contraceptives and breast cancer risk*. Int.J.Cancer, 2003. **103**(6): p. 779-783.
21. Marcus, P.M., et al., *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**(5): p. 461-467.
22. Stucker, I., et al., *Lack of interaction between asbestos exposure and glutathione S-transferase M1 and T1 genotypes in lung carcinogenesis*. Cancer Epidemiol.Biomarkers Prev., 2001. **10**(12): p. 1253-1258.
23. Armstrong, B.G., *Fixed factors that modify the effects of time-varying factors: applying the case-only approach*. Epidemiology, 2003. **14**(4): p. 467-472.
24. Wong, C.M., et al., *The effects of air pollution on mortality in socially deprived urban areas in Hong Kong, China*. Environ Health Perspect, 2008. **116**(9): p. 1189-94.
25. Yang, Y., et al., *Case-only study of interactions between DNA repair genes (hMLH1, APEX1, MGMT, XRCC1 and XPD) and low-frequency electromagnetic fields in childhood acute leukemia*. Leuk Lymphoma, 2008. **49**(12): p. 2344-50.
26. Joshi, A.D., et al., *Red meat and poultry intake, polymorphisms in the nucleotide excision repair and mismatch repair pathways and colorectal cancer risk*. Carcinogenesis, 2009. **30**(3): p. 472-9.
27. Diekstra, F.P., et al., *Interaction between PON1 and population density in amyotrophic lateral sclerosis*. Neuroreport, 2009. **20**(2): p. 186-90.

28. Smits, K.M., et al., *Polymorphisms in genes related to activation or detoxification of carcinogens might interact with smoking to increase renal cancer risk: results from The Netherlands Cohort Study on diet and cancer*. World J Urol, 2008. **26**(1): p. 103-10.
29. Craig, M.E., et al., *Reduced frequency of HLA DRB1*03-DQB1*02 in children with type 1 diabetes associated with enterovirus RNA*. J.Infect.Dis., 2003. **187**(10): p. 1562-1570.
30. Infante-Rivard, C., D. Amre, and D. Sinnett, *GSTT1 and CYP2E1 polymorphisms and trihalomethanes in drinking water: effect on childhood leukemia*. Environ.Health Perspect., 2002. **110**(6): p. 591-593.
31. Infante-Rivard, C., et al., *Childhood acute lymphoblastic leukemia associated with parental alcohol consumption and polymorphisms of carcinogen-metabolizing genes*. Epidemiology, 2002. **13**(3): p. 277-281.
32. Infante-Rivard, C., *Diagnostic x rays, DNA repair genes and childhood acute lymphoblastic leukemia*. Health Phys., 2003. **85**(1): p. 60-64.
33. Norum, P.L., et al., *Glutathione S-transferase genotype and p53 mutations in adenocarcinoma of the small intestine*. Scand.J.Gastroenterol., 2003. **38**(8): p. 845-849.
34. Palli, D., et al., *A gene-environment interaction between occupation and BRCA1/BRCA2 mutations in male breast cancer?* Eur.J.Cancer, 2004. **40**(16): p. 2474-2479.
35. Hussain, S.K., et al., *Cervical and vulvar cancer risk in relation to the joint effects of cigarette smoking and genetic variation in interleukin 2*. Cancer Epidemiol Biomarkers Prev, 2008. **17**(7): p. 1790-9.
36. Tan, Q., et al., *A centenarian-only approach for assessing gene-gene interaction in human longevity*. Eur.J.Hum.Genet., 2002. **10**(2): p. 119-124.
37. Lubin, F., et al., *Body mass index at age 18 years and during adult life and ovarian cancer risk*. Am.J.Epidemiol., 2003. **157**(2): p. 113-120.
38. Modan, B., et al., *Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation*. N.Engl.J.Med., 2001. **345**(4): p. 235-240.
39. Modugno, F., et al., *Reproductive factors and ovarian cancer risk in Jewish BRCA1 and BRCA2 mutation carriers (United States)*. Cancer Causes Control, 2003. **14**(5): p. 439-446.
40. Basham, V.M., et al., *Polymorphisms in CYP1A1 and smoking: no association with breast cancer risk*. Carcinogenesis, 2001. **22**(11): p. 1797-1800.
41. Sturmer, T., et al., *Interaction between alcohol dehydrogenase II gene, alcohol consumption, and risk for breast cancer*. Br.J.Cancer, 2002. **87**(5): p. 519-523.

42. Bennett, W.P., et al., *Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women*. J.Natl.Cancer Inst., 1999. **91**(23): p. 2009-2014.
43. Eekhoff, E.M., F.R. Rosendaal, and J.P. Vandenbroucke, *Minor events and the risk of deep venous thrombosis*. Thromb.Haemost., 2000. **83**(3): p. 408-411.
44. Fallin, M.D., et al., *Family-based analysis of MSX1 haplotypes for association with oral clefts*. Genet.Epidemiol., 2003. **25**(2): p. 168-175.
45. Infante-Rivard, C., G. Mathonnet, and D. Sinnett, *Risk of childhood leukemia associated with diagnostic irradiation and polymorphisms in DNA repair genes*. Environ.Health Perspect., 2000. **108**(6): p. 495-498.
46. Stern, M.C., et al., *XPD codon 751 polymorphism, metabolism genes, smoking, and bladder cancer risk*. Cancer Epidemiology Biomarkers & Prevention, 2002. **11**(10): p. 1004-1011.
47. Theodoratou, E., et al., *Modification of the associations between lifestyle, dietary factors and colorectal cancer risk by APC variants*. Carcinogenesis, 2008. **29**(9): p. 1774-80.
48. Jain, M., et al., *Microsomal epoxide hydrolase (EPHX1), slow (exon 3, 113His) and fast (exon 4, 139Arg) alleles confer susceptibility to squamous cell esophageal cancer*. Toxicol Appl Pharmacol, 2008. **230**(2): p. 247-51.
49. Caceres, D.D., et al., *Association between p53 codon 72 genetic polymorphism and tobacco use and lung cancer risk*. Lung, 2009. **187**(2): p. 110-5.
50. Becher, H., S. Schmidt, and J. Chang-Claude, *Reproductive factors and familial predisposition for breast cancer by age 50 years. A case-control-family study for assessing main effects and possible gene-environment interaction*. Int.J.Epidemiol., 2003. **32**(1): p. 38-48.
51. Deng, Y., et al., *Case-only study of interactions between genetic polymorphisms of GSTM1, P1, T1 and Z1 and smoking in Parkinson's disease*. Neurosci.Lett., 2004. **366**(3): p. 326-331.
52. Egan, K.M., et al., *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**(1): p. 43-51.
53. Saintot, M., et al., *Interactions between genetic polymorphism of cytochrome P450-1B1, sulfotransferase 1A1, catechol-o-methyltransferase and tobacco exposure in breast cancer risk*. Int.J.Cancer, 2003. **107**(4): p. 652-657.
54. Smits, K.M., et al., *Association of metabolic gene polymorphisms with tobacco consumption in healthy controls*. Int.J.Cancer, 2004. **110**(2): p. 266-270.

55. Goodman, M.T., et al., *Case-control study of ovarian cancer and polymorphisms in genes involved in catecholestrogen formation and metabolism*. Cancer Epidemiol.Biomarkers Prev., 2001. **10**(3): p. 209-216.
56. Wang, Y., et al., *Sulfotransferase (SULT) 1A1 polymorphism as a predisposition factor for lung cancer: a case-control analysis*. Lung Cancer, 2002. **35**(2): p. 137-142.
57. Hung, R.J., et al., *GST, NAT, SULT1A1, CYP1B1 genetic polymorphisms, interactions with environmental exposures and bladder cancer risk in a high-risk population*. Int.J.Cancer, 2004. **110**(4): p. 598-604.
58. Dick, D.M. and R.J. Rose, *Behavior Genetics: What's New? What's Next? Current Directions in Psychological Science*, 2002. **11**(2): p. 70-74.
59. Bouchard, T.J., Jr. and M. McGue, *Genetic and environmental influences on human psychological differences*. J.Neurobiol., 2003. **54**(1): p. 4-45.
60. Inoue, K. and J.R. Lupski, *Genetics and genomics of behavioral and psychiatric disorders*. Curr.Opin.Genet.Dev., 2003. **13**(3): p. 303-309.
61. Reif, A. and K.P. Lesch, *Toward a molecular architecture of personality*. Behav.Brain Res., 2003. **139**(1-2): p. 1-20.
62. Fan, J.B., et al., *Catechol-O-methyltransferase gene Val/Met functional polymorphism and risk of schizophrenia: a large-scale association study plus meta-analysis*. Biol.Psychiatry, 2005. **57**(2): p. 139-144.
63. Bellgrove, M.A., et al., *The methionine allele of the COMT polymorphism impairs prefrontal cognition in children and adolescents with ADHD*. Exp.Brain Res., 2005.
64. Liu, Y., et al., *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.
65. Oroszi, G. and D. Goldman, *Alcoholism: genes and mechanisms*. Pharmacogenomics., 2004. **5**(8): p. 1037-1048.
66. Lee, S.M., et al., *Polymorphism of estrogen metabolism genes and cataract*. Med.Hypotheses, 2004. **63**(3): p. 494-497.
67. Wang, P.N., et al., *Estrogen-Metabolizing Gene COMT Polymorphism Synergistic APOE epsilon4 Allele Increases the Risk of Alzheimer Disease*. Dement.Geriatr.Cogn Disord., 2005. **19**(2-3): p. 120-125.
68. Millikan, R.C., et al., *Catechol-O-methyltransferase and breast cancer risk*. Carcinogenesis, 1998. **19**(11): p. 1943-7.

69. Wu, A.H., et al., *Tea and circulating estrogen levels in postmenopausal Chinese women in singapore*. Carcinogenesis, 2005.
70. Yin, P.H., et al., *Polymorphisms of estrogen-metabolizing genes and risk of hepatocellular carcinoma in Taiwan females*. Cancer Lett., 2004. **212**(2): p. 195-201.
71. Hung, R.J., et al., *Genetic polymorphisms of MPO, COMT, MnSOD, NQO1, interactions with environmental exposures and bladder cancer risk*. Carcinogenesis, 2004. **25**(6): p. 973-978.
72. Fromme, K., et al., *Biological and behavioral markers of alcohol sensitivity*. Alcohol Clin.Exp.Res., 2004. **28**(2): p. 247-256.
73. Harada, S., et al., *Possible protective role against alcoholism for aldehyde dehydrogenase isozyme deficiency in Japan*. Lancet, 1982. **2**(8302): p. 827.
74. Harada, S., et al., *Metabolic and ethnic determinants of alcohol drinking habits and vulnerability to alcohol-related disorder*. Alcohol Clin.Exp.Res., 2001. **25**(5 Suppl ISBRA): p. 71S-75S.
75. Higuchi, S., et al., *Alcohol and aldehyde dehydrogenase genotypes and drinking behavior in Japanese*. Alcohol Clin.Exp.Res., 1996. **20**(3): p. 493-497.
76. Takeshita, T. and K. Morimoto, *Self-reported alcohol-associated symptoms and drinking behavior in three ALDH2 genotypes among Japanese university students*. Alcohol Clin.Exp.Res., 1999. **23**(6): p. 1065-1069.
77. Wall, T.L., et al., *Subjective feelings of alcohol intoxication in Asians with genetic variations of ALDH2 alleles*. Alcohol Clin.Exp.Res., 1992. **16**(5): p. 991-995.
78. Luu, S.U., et al., *Ethanol and acetaldehyde metabolism in chinese with different aldehyde dehydrogenase-2 genotypes*. Proc.Natl.Sci.Counc.Repub.China B, 1995. **19**(3): p. 129-136.
79. Wall, T.L. and C.L. Ehlers, *Acute effects of alcohol on P300 in Asians with different ALDH2 genotypes*. Alcohol Clin.Exp.Res., 1995. **19**(3): p. 617-622.
80. Yokoyama, A. and T. Omori, *Genetic polymorphisms of alcohol and aldehyde dehydrogenases and risk for esophageal and head and neck cancers*. Japanese Journal of Clinical Oncology, 2003. **33**(3): p. 111-121.
81. Nishimura, F.T., et al., *Effects of aldehyde dehydrogenase-2 genotype on cardiovascular and endocrine responses to alcohol in young Japanese subjects*. Auton.Neurosci., 2002. **102**(1-2): p. 60-70.
82. Ohsawa, I., et al., *Genetic deficiency of a mitochondrial aldehyde dehydrogenase increases serum lipid peroxides in community-dwelling females*. J.Hum.Genet., 2003. **48**(8): p. 404-409.

83. Murata, C., et al., *Inactive aldehyde dehydrogenase 2 worsens glycemic control in patients with type 2 diabetes mellitus who drink low to moderate amounts of alcohol*. Alcohol Clin.Exp.Res., 2000. **24**(4 Suppl): p. 5S-11S.
84. Yamanaka, H., et al., *Analysis of the genotypes for aldehyde dehydrogenase 2 in Japanese patients with primary gout*. Adv.Exp.Med.Biol., 1994. **370**: p. 53-56.
85. Nakamura, Y., et al., *Genetic variation in aldehyde dehydrogenase 2 and the effect of alcohol consumption on cholesterol levels*. Atherosclerosis, 2002. **164**(1): p. 171-177.
86. Wall, T.L., et al., *Cortisol responses following placebo and alcohol in Asians with different ALDH2 genotypes*. J.Stud.Alcohol, 1994. **55**(2): p. 207-213.
87. Carter, B., T. Long, and P. Cinciripini, *A meta-analytic review of the CYP2A6 genotype and smoking behavior*. Nicotine.Tob.Res., 2004. **6**(2): p. 221-227.
88. Tricker, A.R., *Nicotine metabolism, human drug metabolism polymorphisms, and smoking behaviour*. Toxicology, 2003. **183**(1-3): p. 151-173.
89. Ray, R., R.F. Tyndale, and C. Lerman, *Nicotine Dependence Pharmacogenetics: Role of Genetic Variation in Nicotine-Metabolizing Enzymes*. J Neurogenet, 2009: p. 1-10.
90. Schnoll, R.A., T.A. Johnson, and C. Lerman, *Genetics and smoking behavior*. Curr Psychiatry Rep, 2007. **9**(5): p. 349-57.
91. Messina, E.S., R.F. Tyndale, and E.M. Sellers, *A major role for CYP2A6 in nicotine C-oxidation by human liver microsomes*. J.Pharmacol.Exp.Ther., 1997. **282**(3): p. 1608-1614.
92. Xu, C., et al., *CYP2A6 genetic variation and potential consequences*. Adv.Drug Deliv.Rev., 2002. **54**(10): p. 1245-1256.
93. Xu, C., et al., *An in vivo pilot study characterizing the new CYP2A6*7, *8, and *10 alleles*. Biochem.Biophys.Res.Comm., 2002. **290**(1): p. 318-324.
94. Benowitz, N.L., *Drug therapy. Pharmacologic aspects of cigarette smoking and nicotine addiction*. N.Engl.J.Med., 1988. **319**(20): p. 1318-1330.
95. Sellers, E.M., H.L. Kaplan, and R.F. Tyndale, *Inhibition of cytochrome P450 2A6 increases nicotine's oral bioavailability and decreases smoking*. Clin.Pharmacol.Ther., 2000. **68**(1): p. 35-43.
96. Audrain-McGovern, J., et al., *The role of CYP2A6 in the emergence of nicotine dependence in adolescents*. Pediatrics, 2007. **119**(1): p. e264-74.
97. Kwon, J.T., et al., *Nicotine metabolism and CYP2A6 allele frequencies in Koreans*. Pharmacogenetics, 2001. **11**(4): p. 317-323.

98. Nakajima, M., et al., *Relationship between interindividual differences in nicotine metabolism and CYP2A6 genetic polymorphism in humans*. Clin.Pharmacol.Ther., 2001. **69**(1): p. 72-78.
99. Gu, D.F., et al., *The use of long PCR to confirm three common alleles at the CYP2A6 locus and the relationship between genotype and smoking habit*. Ann.Hum.Genet., 2000. **64**(Pt 5): p. 383-390.
100. Lorient, M.A., et al., *Genetic polymorphisms of cytochrome P450 2A6 in a case-control study on lung cancer in a French population*. Pharmacogenetics, 2001. **11**(1): p. 39-44.
101. Pianezza, M.L., E.M. Sellers, and R.F. Tyndale, *Nicotine metabolism defect reduces smoking*. Nature, 1998. **393**(6687): p. 750.
102. Rao, Y., et al., *Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking*. Mol.Pharmacol., 2000. **58**(4): p. 747-755.
103. Tan, W., et al., *Frequency of CYP2A6 gene deletion and its relation to risk of lung and esophageal cancer in the Chinese population*. Int.J.Cancer, 2001. **95**(2): p. 96-101.
104. Tiihonen, J., et al., *CYP2A6 genotype and smoking*. Mol.Psychiatry, 2000. **5**(4): p. 347-348.
105. Yang, M., et al., *Individual differences in urinary cotinine levels in Japanese smokers: relation to genetic polymorphism of drug-metabolizing enzymes*. Cancer Epidemiol.Biomarkers Prev., 2001. **10**(6): p. 589-593.
106. Schulz, T.G., P. Ruhnau, and E. Hallier, *Lack of correlation between CYP2A6 genotype and smoking habits*. Adv.Exp.Med.Biol., 2001. **500**: p. 213-215.
107. Sellers, E.M., et al., *The effect of methoxsalen on nicotine and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) metabolism in vivo*. Nicotine.Tob.Res., 2003. **5**(6): p. 891-899.
108. Rossini, A., et al., *CYP2A6 polymorphisms and risk for tobacco-related cancers*. Pharmacogenomics, 2008. **9**(11): p. 1737-52.
109. Strasser, A.A., et al., *An association of CYP2A6 genotype and smoking topography*. Nicotine Tob Res, 2007. **9**(4): p. 511-8.
110. Ariyoshi, N., et al., *Genetic polymorphism of CYP2A6 gene and tobacco-induced lung cancer risk in male smokers*. Cancer Epidemiol.Biomarkers Prev., 2002. **11**(9): p. 890-894.
111. Fujieda, M., et al., *Evaluation of CYP2A6 genetic polymorphisms as determinants of smoking behavior and tobacco-related lung cancer risk in male Japanese smokers*. Carcinogenesis, 2004.
112. Rotunno, M., et al., *Phase I metabolic genes and risk of lung cancer: multiple polymorphisms and mRNA expression*. PLoS One, 2009. **4**(5): p. e5652.

113. Paschke, T., et al., *Comparison of cytochrome P450 2A6 polymorphism frequencies in Caucasians and African-Americans using a new one-step PCR-RFLP genotyping method*. Toxicology, 2001. **168**(3): p. 259-268.
114. Raimondi, S., et al., *Association of metabolic gene polymorphisms with alcohol consumption in controls*. Biomarkers, 2004. **9**(2): p. 180-189.
115. Gaspari, L., D. Marinelli, and E. Taioli, *International collaborative study on genetic susceptibility to environmental carcinogens (GSEC): an update*. Int.J.Hyg.Environ.Health, 2001. **204**(1): p. 39-42.
116. Taioli, E., *International collaborative study on genetic susceptibility to environmental carcinogens*. Cancer Epidemiol.Biomarkers Prev., 1999. **8**(8): p. 727-728.
117. Morabia, A., S.D. Stellman, and E.L. Wynder, *Smoking prevalence in neighborhood and hospital controls: implications for hospital-based case-control studies*. J.Clin.Epidemiol., 1996. **49**(8): p. 885-889.
118. Ramos, E., C. Lopes, and H. Barros, *Investigating the effect of nonparticipation using a population-based case-control study on myocardial infarction*. Ann Epidemiol, 2004. **14**(6): p. 437-41.
119. Holt, V.L., D.P. Martin, and J.P. LoGerfo, *Correlates and effect of non-response in a postpartum survey of obstetrical care quality*. J Clin Epidemiol, 1997. **50**(10): p. 1117-22.
120. Heilbrun, L.K., A. Nomura, and G.N. Stemmermann, *The effects of nonresponse in a prospective study of cancer*. Am J Epidemiol, 1982. **116**(2): p. 353-63.
121. Maxwell, C.J., C.M. Bancej, and J. Snider, *Predictors of mammography use among Canadian women aged 50-69: findings from the 1996/97 National Population Health Survey*. CMAJ., 2001. **164**(3): p. 329-334.
122. Foster, M.W. and R.R. Sharp, *Beyond race: towards a whole-genome perspective on human populations and genetic variation*. Nat.Rev.Genet., 2004. **5**(10): p. 790-796.
123. Keita, S.O., et al., *Conceptualizing human variation*. Nat.Genet., 2004. **36**(11 Suppl): p. S17-S20.
124. Liu, X., M.D. Fallin, and W.H. Kao, *Genetic dissection methods: designs used for tests of gene-environment interaction*. Curr.Opin.Genet.Dev., 2004. **14**(3): p. 241-245.
125. Hamajima, N., et al., *Detection of gene-environment interaction by case-only studies*. Japanese Journal of Clinical Oncology, 1999. **29**(10): p. 490-493.
126. Ratnasinghe, D., et al., *Polymorphisms of the DNA repair gene XRCC1 and lung cancer risk*. Cancer Epidemiol.Biomarkers Prev., 2001. **10**(2): p. 119-123.

127. Gatto, N.M., et al., *Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias*. Int.J.Epidemiol., 2004. **33**(5): p. 1014-1024.
128. Koushik, A., R.W. Platt, and E.L. Franco, *p53 codon 72 polymorphism and cervical neoplasia: a meta-analysis review*. Cancer Epidemiol.Biomarkers Prev., 2004. **13**(1): p. 11-22.
129. Browner, W.S., et al., *The genetics of human longevity*. Am.J.Med., 2004. **117**(11): p. 851-860.
130. Gaspari, L., et al., *Metabolic gene polymorphisms and p53 mutations in healthy centenarians and younger controls*. Biomarkers, 2003. **8**(6): p. 522-528.
131. Hasty, P., *The impact of DNA damage, genetic mutation and cellular responses on cancer prevention, longevity and aging: observations in humans and mice*. Mech.Ageing Dev., 2005. **126**(1): p. 71-77.
132. Lombard, D.B., et al., *DNA repair, genome stability, and aging*. Cell, 2005. **120**(4): p. 497-512.
133. Stewart, S.H., et al., *COMT genotype influences the effect of alcohol on blood pressure: results from the COMBINE study*. Am J Hypertens, 2009. **22**(1): p. 87-91.
134. Bombard, J., et al., *State-specific prevalence of current cigarette smoking among adults--United States, 2003*. MMWR Morb.Mortal.Wkly.Rep., 2004. **53**(44): p. 1035-1037.
135. Peto, R., Z.M. Chen, and J. Boreham, *Tobacco--the growing epidemic*. Nat.Med., 1999. **5**(1): p. 15-17.
136. DeMarini, D.M., *Genotoxicity of tobacco smoke and tobacco smoke condensate: a review*. Mutation Research-Reviews in Mutation Research, 2004. **567**(2-3): p. 447-474.
137. Wu, X.F., et al., *Genetic susceptibility to tobacco-related cancer*. Oncogene, 2004. **23**(38): p. 6500-6523.
138. Munafo, M., et al., *The genetic basis for smoking behavior: a systematic review and meta-analysis*. Nicotine Tob Res, 2004. **6**(4): p. 583-97.
139. Munafo, M.R. and E.C. Johnstone, *Genes and cigarette smoking*. Addiction, 2008. **103**(6): p. 893-904.
140. Gartner, C.E., J.J. Barendregt, and W.D. Hall, *Multiple genetic tests for susceptibility to smoking do not outperform simple family history*. Addiction, 2009. **104**(1): p. 118-26.
141. Taioli, E., *Gene-environment interaction in tobacco-related cancers*. Carcinogenesis, 2008. **29**(8): p. 1467-74.
142. Andreassi, M.G., *Metabolic syndrome, diabetes and atherosclerosis: influence of gene-environment interaction*. Mutat Res, 2009. **667**(1-2): p. 35-43.

143. Godschalk, R.W. and J.C. Kleinjans, *Characterization of the exposure-disease continuum in neonates of mothers exposed to carcinogens during pregnancy*. Basic Clin Pharmacol Toxicol, 2008. **102**(2): p. 109-17.
144. Hecht, S.S., et al., *Comparison of polymorphisms in genes involved in polycyclic aromatic hydrocarbon metabolism with urinary phenanthrene metabolite ratios in smokers*. Cancer Epidemiol Biomarkers Prev, 2006. **15**(10): p. 1805-11.
145. Nishikawa, A., et al., *Cigarette smoking, metabolic activation and carcinogenesis*. Curr. Drug Metab, 2004. **5**(5): p. 363-373.
146. Gresner, P., J. Gromadzinska, and W. Wasowicz, *Polymorphism of selected enzymes involved in detoxification and biotransformation in relation to lung cancer*. Lung Cancer, 2007. **57**(1): p. 1-25.
147. David, S.P. and M.R. Munafo, *Genetic variation in the dopamine pathway and smoking cessation*. Pharmacogenomics, 2008. **9**(9): p. 1307-21.
148. Breitling, L.P., et al., *Variants in COMT and spontaneous smoking cessation: retrospective cohort analysis of 925 cessation events*. Pharmacogenet Genomics, 2009.
149. Omidvar, M., et al., *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**(1): p. 45-51.
150. Chen, X. and K.J. Woodcroft, *Polymorphisms in metabolic genes CYP1A1 and GSTM1 and changes in maternal smoking during pregnancy*. Nicotine Tob Res, 2009. **11**(3): p. 225-33.
151. Mohrenweiser, H.W., D.M. Wilson, III, and I.M. Jones, *Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes*. Mutat. Res., 2003. **526**(1-2): p. 93-125.
152. Goode, E.L., C.M. Ulrich, and J.D. Potter, *Polymorphisms in DNA repair genes and associations with cancer risk*. Cancer Epidemiology Biomarkers & Prevention, 2002. **11**(12): p. 1513-1530.
153. Andrews, A.D., S.F. Barrett, and J.H. Robbins, *Xeroderma pigmentosum neurological abnormalities correlate with colony-forming ability after ultraviolet radiation*. Proc. Natl. Acad. Sci U.S.A., 1978. **75**(4): p. 1984-1988.
154. Reardon, J.T., et al., *In vitro repair of oxidative DNA damage by human nucleotide excision repair system: possible explanation for neurodegeneration in xeroderma pigmentosum patients*. Proc. Natl. Acad. Sci U.S.A., 1997. **94**(17): p. 9463-9468.
155. Kurz, E.U. and S.P. Lees-Miller, *DNA damage-induced activation of ATM and ATM-dependent signaling pathways*. DNA Repair (Amst), 2004. **3**(8-9): p. 889-900.

156. Picciotto, M.R. and M. Zoli, *Nicotinic receptors in aging and dementia*. J.Neurobiol., 2002. **53**(4): p. 641-655.
157. Shackelford, R.E., et al., *Pharmacological manipulation of ataxia-telangiectasia kinase activity as a treatment for Parkinson's disease*. Med.Hypotheses, 2005. **64**(4): p. 736-741.
158. de Waard, H., et al., *Cell type-specific hypersensitivity to oxidative damage in CSB and XPA mice*. DNA Repair (Amst), 2003. **2**(1): p. 13-25.
159. Marietta, C., H. Gulam, and P.J. Brooks, *A single 8,5'-cyclo-2'-deoxyadenosine lesion in a TATA box prevents binding of the TATA binding protein and strongly reduces transcription in vivo*. DNA Repair (Amst), 2002. **1**(11): p. 967-975.
160. Graham, J.M., Jr., et al., *Cerebro-oculo-facio-skeletal syndrome with a nucleotide excision-repair defect and a mutated XPD gene, with prenatal diagnosis in a triplet pregnancy*. Am.J.Hum.Genet., 2001. **69**(2): p. 291-300.
161. Robbins, J.H., et al., *Neurological disease in xeroderma pigmentosum. Documentation of a late onset type of the juvenile onset form*. Brain, 1991. **114** (Pt 3): p. 1335-1361.
162. Robbins, J.H., R.A. Brumback, and A.N. Moshell, *Clinically asymptomatic xeroderma pigmentosum neurological disease in an adult: evidence for a neurodegeneration in later life caused by defective DNA repair*. Eur.Neurol., 1993. **33**(3): p. 188-190.
163. Inoshima, I., et al., *Induction of CDK inhibitor p21 gene as a new therapeutic strategy against pulmonary fibrosis*. Am.J.Physiol Lung Cell Mol.Physiol, 2004. **286**(4): p. L727-L733.
164. Kuwano, K., et al., *P21Waf1/Cip1/Sdi1 and p53 expression in association with DNA strand breaks in idiopathic pulmonary fibrosis*. Am.J.Respir.Crit Care Med., 1996. **154**(2 Pt 1): p. 477-483.
165. Ceylan, E., et al., *Increased DNA damage in patients with chronic obstructive pulmonary disease who had once smoked or been exposed to biomass*. Respir Med, 2006. **100**(7): p. 1270-6.
166. Tzortzaki, E.G. and N.M. Siafakas, *A hypothesis for the initiation of COPD*. Eur Respir J, 2009. **34**(2): p. 310-5.
167. Casella, M., et al., *No evidence of chromosome damage in chronic obstructive pulmonary disease (COPD)*. Mutagenesis, 2006. **21**(2): p. 167-71.
168. Luna, L., et al., *Dynamic relocation of hOGG1 during the cell cycle is disrupted in cells harbouring the hOGG1-Cys326 polymorphic variant*. Nucleic Acids Res., 2005. **33**(6): p. 1813-1824.
169. Wu, M.T., et al., *Genetic polymorphism of p53 and XRCC1 in cervical intraepithelial neoplasm in Taiwanese women*. Journal of the Formosan Medical Association, 2004. **103**(5): p. 337-343.

170. Godderis, L., et al., *Influence of genetic polymorphisms on biomarkers of exposure and genotoxic effects in styrene-exposed workers*. Environ.Mol.Mutagen., 2004. **44**(4): p. 293-303.
171. Duell, E.J., et al., *Polymorphisms in the DNA repair genes XRCC1 and ERCC2 and biomarkers of DNA damage in human blood mononuclear cells. [erratum appears in Carcinogenesis 2000 Jul;21(7):1457.]*. Carcinogenesis., 2000. **21**(5): p. 965-971.
172. Hu, Z., et al., *DNA repair gene XPD polymorphism and lung cancer risk: a meta-analysis*. Lung Cancer, 2004. **46**(1): p. 1-10.
173. Tuimala, J., et al., *Genetic polymorphisms of DNA repair and xenobiotic-metabolizing enzymes: effects on levels of sister chromatid exchanges and chromosomal aberrations*. Mutat.Res., 2004. **554**(1-2): p. 319-333.
174. Vodicka, P., et al., *Genetic polymorphisms in DNA repair genes and possible links with DNA repair rates, chromosomal aberrations and single-strand breaks in DNA*. Carcinogenesis, 2004. **25**(5): p. 757-763.
175. Hall, I.J., et al., *Comparative analysis of breast cancer risk factors among African-American women and White women*. Am J Epidemiol, 2005. **161**(1): p. 40-51.
176. Hall, S.A., et al., *Urbanization and breast cancer incidence in North Carolina, 1995-1999*. Ann Epidemiol, 2005. **15**(10): p. 796-803.
177. Il'yasova, D., C. Martin, and R.S. Sandler, *Tea intake and risk of colon cancer in African-Americans and whites: North Carolina colon cancer study*. Cancer Causes Control, 2003. **14**(8): p. 767-772.
178. Millikan, R., et al., *HER2 codon 655 polymorphism and risk of breast cancer in African Americans and whites*. Breast Cancer Research and Treatment, 2003. **79**(3): p. 355-364.
179. Newman, B., et al., *The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology*. Breast Cancer Research & Treatment, 1995. **35**(1): p. 51-60.
180. Satia-Abouta, J., et al., *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**(10): p. 951-962.
181. Weinberg, C.R. and D.P. Sandler, *Randomized recruitment in case-control studies*. American Journal of Epidemiology, 1991. **134**(4): p. 421-432.
182. Thomas, D.C. and J.S. Witte, *Point: population stratification: a problem for case-control studies of candidate-gene associations?* Cancer Epidemiol.Biomarkers Prev., 2002. **11**(6): p. 505-512.
183. Greenland, S., *Quantitative methods in the review of epidemiologic literature*. Epidemiol.Rev., 1987. **9**: p. 1-30.

184. Egger, M. and G.D. Smith, *Principles of and procedures for systematic reviews*, in *Systematic Reviews in Health Care, Meta-analysis in Context*, M. Egger, G.D. Smith, and D.G. Altman, Editors. 2001, BMJ Publishing Group, BMA House: London. p. 23-42.
185. Garbe, E., L. Levesque, and S. Suissa, *Variability of breast cancer risk in observational studies of hormone replacement therapy: a meta-regression analysis*. *Maturitas*, 2004. **47**(3): p. 175-183.
186. Salanti, G., S. Sanderson, and J.P. Higgins, *Obstacles and opportunities in meta-analysis of genetic association studies*. *Genet.Med.*, 2005. **7**(1): p. 13-20.
187. Thompson, S.G., *Why and how sources of heterogeneity should be investigated*, in *Systematic Reviews in Health Care, Meta-analysis in Context*, M. Egger, G.D. Smith, and D.G. Altman, Editors. 2001, BMJ Publishing Group, BMA House: London. p. 157-175.
188. Davey, S.G. and S. Ebrahim, 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int.J.Epidemiol.*, 2003. **32**(1): p. 1-22.
189. Hardy, R.J. and S.G. Thompson, *Detecting and describing heterogeneity in meta-analysis*. *Stat.Med.*, 1998. **17**(8): p. 841-856.
190. Sterne, J.A., D. Gavaghan, and M. Egger, *Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature*. *J Clin Epidemiol*, 2000. **53**(11): p. 1119-29.
191. Begg, C.B. and M. Mazumdar, *Operating characteristics of a rank correlation test for publication bias*. *Biometrics*, 1994. **50**(4): p. 1088-101.
192. Egger, M., et al., *Bias in meta-analysis detected by a simple, graphical test*. *BMJ*, 1997. **315**(7109): p. 629-34.
193. Li, Y., et al., *Cigarette smoking, cytochrome P4501A1 polymorphisms, and breast cancer among African-American and white women*. *Breast Cancer Res*, 2004. **6**(4): p. R460-73.
194. Mechanic, L.E., et al., *Polymorphisms in nucleotide excision repair genes, smoking and breast cancer in African Americans and whites: a population-based case-control study*. *Carcinogenesis*, 2006. **27**(7): p. 1377-1385.
195. Millikan, R.C., et al., *Cigarette smoking, N-acetyltransferases 1 and 2, and breast cancer risk*. *Cancer Epidemiol Biomarkers Prev*, 1998. **7**(5): p. 371-8.
196. Millikan, R., et al., *Glutathione S-transferases M1, T1, and P1 and breast cancer*. *Cancer Epidemiol Biomarkers Prev*, 2000. **9**(6): p. 567-73.
197. Millikan, R.C., *NAT1*10 and NAT1*11 polymorphisms and breast cancer risk*. *Cancer Epidemiol Biomarkers Prev*, 2000. **9**(2): p. 217-9.

98. Millikan, R.C., et al., *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 Research, 2004. **6**(4): p. R264-R274.
199. Millikan, R.C., et al., *Polymorphisms in DNA repair genes, medical exposure to ionizing radiation, and breast cancer risk*. Cancer Epidemiology Biomarkers & Prevention, 2005. **14**(10): p. 2326-2334.
200. Pachkowski, B.F., et al., *XRCC1 genotype and breast cancer: functional studies and epidemiologic data show interactions between XRCC1 codon 280 His and smoking*. Cancer Res., 2006. **66**(5): p. 2860-2868.
201. Huang, K., et al., *GSTM1 and GSTT1 polymorphisms, cigarette smoking, and risk of colon cancer: a population-based case-control study in North Carolina (United States)*. Cancer Causes Control, 2006. **17**(4): p. 385-94.
202. Butler, L.M., et al., *Modification by N-acetyltransferase 1 genotype on the association between dietary heterocyclic amines and colon cancer in a multiethnic study*. Mutat Res, 2008. **638**(1-2): p. 162-74.
203. Kinney, A.Y., et al., *Roles of religious involvement and social support in the risk of colon cancer among Blacks and Whites*. Am J Epidemiol, 2003. **158**(11): p. 1097-107.
204. Wacholder, S., N. Rothman, and N. Caporaso, *Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. [letter; comment.]*. Cancer Epidemiology, Biomarkers & Prevention., 2002. **11**(6): p. 513-520.
205. Tang, H., et al., *Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies*. Am.J.Hum.Genet., 2005. **76**(2): p. 268-275.
206. Wacholder, S., N. Rothman, and N. Caporaso, *Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias*. J.Natl.Cancer Inst., 2000. **92**(14): p. 1151-1158.
207. Wang, Y., R. Localio, and T.R. Rebbeck, *Evaluating bias due to population stratification in case-control association studies of admixed populations*. Genet.Epidemiol., 2004. **27**(1): p. 14-20.
208. Parra, E.J., et al., *Estimating African American admixture proportions by use of population-specific alleles*. Am.J.Hum.Genet., 1998. **63**(6): p. 1839-1851.
209. Hernan, M.A., et al., *Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology*. Am J Epidemiol, 2002. **155**(2): p. 176-84.
210. SAS Institute Inc., C., NC, USA, SAS 9.1.3. 2002.

211. Cohen, J., *Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit*. Psychol Bull, 1968. **70**(4): p. 213-20.
212. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data*. Biometrics, 1977. **33**(1): p. 159-74.
213. Greenland, S. and C. Poole, *Invariants and noninvariants in the concept of interdependent effects*. Scand J Work Environ Health, 1988. **14**(2): p. 125-9.
214. Rothman, K.J., *Modern Epidemiology, 3rd edition*. Vol. 3rd. 2008: Lippincott Williams & Wilkins.
215. Berwick, M. and P. Vineis, *Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review*. J Natl Cancer Inst, 2000. **92**(11): p. 874-97.
216. Kiyohara, C., K. Takayama, and Y. Nakanishi, *Association of genetic polymorphisms in the base excision repair pathway with lung cancer risk: a meta-analysis*. Lung Cancer, 2006. **54**(3): p. 267-283.
217. Qu, T. and K. Morimoto, *X-ray repair cross-complementing group 1 polymorphisms and cancer risks in Asian populations: a mini review*. Cancer Detect.Prev., 2005. **29**(3): p. 215-220.
218. Kiyohara, C. and K. Yoshimasu, *Genetic polymorphisms in the nucleotide excision repair pathway and lung cancer risk: a meta-analysis*. Int J Med Sci, 2007. **4**(2): p. 59-71.
219. Manuguerra, M., et al., *XRCC3 and XPD/ERCC2 single nucleotide polymorphisms and the risk of cancer: a HuGE review*. Am J Epidemiol, 2006. **164**(4): p. 297-302.
220. Stoll, B.A., *Premalignant breast lesions: role for biological markers in predicting progression to cancer. [Review] [70 refs]*. European Journal of Cancer., 1999. **35**(5): p. 693-697.
221. Hou, S.M., et al., *The XPD variant alleles are associated with increased aromatic DNA adduct level and lung cancer risk*. Carcinogenesis, 2002. **23**(4): p. 599-603.
222. Ryk, C., et al., *Polymorphisms in the DNA repair genes XRCC1, APEX1, XRCC3 and NBS1, and the risk for lung cancer in never- and ever-smokers*. Lung Cancer, 2006. **54**(3): p. 285-292.
223. Shen, J., et al., *Polymorphisms in XRCC1 modify the association between polycyclic aromatic hydrocarbon-DNA adducts, cigarette smoking, dietary antioxidants, and breast cancer risk*. Cancer Epidemiology Biomarkers & Prevention, 2005. **14**(2): p. 336-342.
224. Stern, M.C., et al., *DNA repair gene XRCC1 polymorphisms, smoking, and bladder cancer risk*. Cancer Epidemiol.Biomarkers Prev., 2001. **10**(2): p. 125-131.

225. Stern, M.C., et al., *DNA repair gene XRCC3 codon 241 polymorphism, its interaction with smoking and XRCC1 polymorphisms, and bladder cancer risk*. Cancer Epidemiol.Biomarkers Prev., 2002. **11**(9): p. 939-943.
226. Terry, M.B., et al., *Polymorphism in the DNA repair gene XPD, polycyclic aromatic hydrocarbon-DNA adducts, cigarette smoking, and breast cancer risk*. Cancer Epidemiology Biomarkers & Prevention, 2004. **13**(12): p. 2053-2058.
227. Duell, E.J., et al., *Polymorphisms in the DNA repair gene XRCC1 and breast cancer*. Cancer Epidemiol.Biomarkers Prev., 2001. **10**(3): p. 217-222.
228. Thompson, S.G. and S.J. Sharp, *Explaining heterogeneity in meta-analysis: a comparison of methods*. Stat.Med., 1999. **18**(20): p. 2693-2708.
229. Stern, M.C., et al., *XRCC1, XRCC3, and XPD polymorphisms as modifiers of the effect of smoking and alcohol on colorectal adenoma risk*. Cancer Epidemiol.Biomarkers Prev., 2006. **15**(12): p. 2384-2390.
230. Figueiredo, J.C., et al., *Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario Site of the Breast Cancer Family Registry*. Cancer Epidemiology Biomarkers & Prevention, 2004. **13**(4): p. 583-591.
231. Affatato, A.A., et al., *Effect of XPD/ERCC2 polymorphisms on chromosome aberration frequencies in smokers and on sensitivity to the mutagenic tobacco-specific nitrosamine NNK*. Environmental and Molecular Mutagenesis, 2004. **44**(1): p. 65-73.
232. Wang, Y., et al., *XRCC3 genetic polymorphism, smoking, and lung carcinoma risk in minority populations*. Cancer, 2003. **98**(8): p. 1701-1706.
233. David-Beabes, G.L. and S.J. London, *Genetic polymorphism of XRCC1 and lung cancer risk among African-Americans and Caucasians*. Lung Cancer, 2001. **34**(3): p. 333-339.
234. Duell, E.J., et al., *A population-based study of the Arg399Gln polymorphism in X-ray repair cross- complementing group 1 (XRCC1) and risk of pancreatic adenocarcinoma*. Cancer Res., 2002. **62**(16): p. 4630-4636.
235. Huang, W.Y., et al., *Selected DNA repair polymorphisms and gastric cancer in Poland*. Carcinogenesis, 2005. **26**(8): p. 1354-1359.
236. Justenhoven, C., et al., *ERCC2 genotypes and a corresponding haplotype are linked with breast cancer risk in a German population*. Cancer Epidemiology Biomarkers & Prevention, 2004. **13**(12): p. 2059-2064.
237. Kelsey, K.T., et al., *A population-based case-control study of the XRCC1 Arg399Gln polymorphism and susceptibility to bladder cancer*. Cancer Epidemiol.Biomarkers Prev., 2004. **13**(8): p. 1337-1341.

238. Shen, H.B., et al., *Polymorphisms of the DNA repair gene XRCC1 and risk of gastric cancer in a Chinese population*. International Journal of Cancer, 2000. **88**(4): p. 601-606.
239. Smedby, K.E., et al., *Variation in DNA repair genes ERCC2, XRCC1, and XRCC3 and risk of follicular lymphoma*. Cancer Epidemiol.Biomarkers Prev., 2006. **15**(2): p. 258-265.
240. Butkiewicz, D., et al., *Genetic polymorphisms in DNA repair genes and risk of lung cancer*. Carcinogenesis, 2001. **22**(4): p. 593-597.
241. Garcia-Closas, M., et al., *Genetic variation in the nucleotide excision repair pathway and bladder cancer risk*. Cancer Epidemiology Biomarkers & Prevention, 2006. **15**(3): p. 536-542.
242. Harms, C., et al., *Polymorphisms in DNA repair genes, chromosome aberrations, and lung cancer*. Environ.Mol.Mutagen., 2004. **44**(1): p. 74-82.
243. Hung, R.J., et al., *Large-scale investigation of base excision repair genetic polymorphisms and lung cancer risk in a multicenter study*. Journal of the National Cancer Institute, 2005. **97**(8): p. 567-576.
244. Ito, H., et al., *Gene-environment interactions between the smoking habit and polymorphisms in the DNA repair genes, APE1 Asp148Glu and XRCC1 Arg399Gln, in Japanese lung cancer risk*. Carcinogenesis, 2004. **25**(8): p. 1395-1401.
245. Jiao, L., et al., *The XPD Asp312Asn and Lys751Gln polymorphisms, corresponding haplotype, and pancreatic cancer risk*. Cancer Lett., 2007. **245**(1-2): p. 61-68.
246. Matullo, G., et al., *Polymorphisms/haplotypes in DNA repair genes and smoking: a bladder cancer case-control study*. Cancer Epidemiol.Biomarkers Prev., 2005. **14**(11 Pt 1): p. 2569-2578.
247. Metsola, K., et al., *XRCC1 and XPD genetic polymorphisms, smoking and breast cancer risk in a Finnish case-control study*. Breast Cancer Res., 2005. **7**(6): p. R987-R997.
248. Olshan, A.F., et al., *XRCC1 polymorphisms and head and neck cancer*. Cancer Lett., 2002. **178**(2): p. 181-186.
249. Park, J.Y., et al., *Polymorphism of the DNA repair gene XRCC1 and risk of primary lung cancer*. Cancer Epidemiol.Biomarkers Prev., 2002. **11**(1): p. 23-27.
250. Ramachandran, S., et al., *Single nucleotide polymorphisms of DNA repair genes XRCC1 and XPD and its molecular mapping in Indian oral cancer*. Oral Oncology, 2006. **42**(4): p. 350-362.
251. Schabath, M.B., et al., *Polymorphisms in XPD Exons 10 and 23 and bladder cancer risk*. Cancer Epidemiology Biomarkers & Prevention, 2005. **14**(4): p. 878-884.
252. Schneider, J., et al., *XRCC1 polymorphism and lung cancer risk in relation to tobacco smoking*. Int.J.Mol.Med., 2005. **16**(4): p. 709-716.

253. Shen, H., et al., *A variant of the DNA repair gene XRCC3 and risk of squamous cell carcinoma of the head and neck: a case-control analysis*. Int.J.Cancer, 2002. **99**(6): p. 869-872.
254. Shen, M., et al., *Polymorphisms of the DNA repair genes XRCC1, XRCC3, XPD, interaction with environmental exposures, and bladder cancer risk in a case-control study in northern Italy*. Cancer Epidemiol.Biomarkers Prev., 2003. **12**(11 Pt 1): p. 1234-1240.
255. Xing, D., et al., *Polymorphisms of the DNA repair gene XPD and risk of lung cancer in a Chinese population*. Lung Cancer, 2002. **38**(2): p. 123-129.
256. Yu, H.P., et al., *Polymorphisms in the DNA repair gene XPD and susceptibility to esophageal squamous cell carcinoma*. Cancer Genet.Cytogenet., 2004. **154**(1): p. 10-15.
257. Yu, H.P., et al., *DNA repair gene XRCC1 polymorphisms, smoking, and esophageal cancer risk*. Cancer Detect.Prev., 2004. **28**(3): p. 194-199.
258. Zhou, W., et al., *Gene-environment interaction for the ERCC2 polymorphisms and cumulative cigarette smoking exposure in lung cancer*. Cancer Res., 2002. **62**(5): p. 1377-1381.
259. Zhou, W., et al., *Polymorphisms in the DNA repair genes XRCC1 and ERCC2, smoking, and lung cancer risk*. Cancer Epidemiol.Biomarkers Prev., 2003. **12**(4): p. 359-365.
260. Han, J.L., et al., *A prospective study of XRCC1 haplotypes and their interaction with plasma carotenoids on breast cancer risk*. Cancer Research, 2003. **63**(23): p. 8536-8541.
261. Jin, M.J., et al., *The association of the DNA repair gene XRCC3 Thr241Met polymorphism with susceptibility to colorectal cancer in a Chinese population*. Cancer Genet.Cytogenet., 2005. **163**(1): p. 38-43.
262. Misra, R.R., et al., *Polymorphisms in the DNA repair genes XPD, XRCC1, XRCC3, and APE/ref-1, and the risk of lung cancer among male smokers in Finland*. Cancer Lett., 2003. **191**(2): p. 171-178.
263. Patel, A.V., et al., *A prospective study of XRCC1 (X-ray cross-complementing group 1) polymorphisms and breast cancer risk*. Breast Cancer Res., 2005. **7**(6): p. R1168-R1173.
264. Naccarati, A., et al., *Genetic polymorphisms and possible gene-gene interactions in metabolic and DNA repair genes: Effects on DNA damage*. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis, 2006. **593**(1-2): p. 22-31.
265. Vodicka, P., et al., *Association of DNA repair polymorphisms with DNA repair functional outcomes in healthy human subjects*. Carcinogenesis, 2007. **28**(3): p. 657-64.
266. Tyndale, R.F., *Genetics of alcohol and tobacco use in humans*. Ann Med, 2003. **35**(2): p. 94-121.

267. Hoffmann, H., et al., *Genetic polymorphisms and the effect of cigarette smoking in the comet assay*. Mutagenesis, 2005. **20**(5): p. 359-364.
268. Butler, L.M., et al., *Heterocyclic amines, meat intake, and association with colon cancer in a population-based study*. Am J Epidemiol, 2003. **157**(5): p. 434-45.
269. Moorman, P.G., et al., *Participation rates in a case-control study: the impact of age, race, and race of interviewer*. Ann Epidemiol, 1999. **9**(3): p. 188-95.
270. Flegal, K.M., *The effects of changes in smoking prevalence on obesity prevalence in the United States*. Am J Public Health, 2007. **97**(8): p. 1510-4.
271. Centers for Disease Control and Prevention, N.C.f.C.D.P.a.H.P. *Behavioral Risk Factor Surveillance System*. Tobacco Use Data 2001 [cited 2009 10/03/2009]; Available from: <http://apps.nccd.cdc.gov/brfss/page.asp?yr=2001&state=NC&cat=TU#TU>.
272. Koyama, A., et al., *Possible association of the X-ray cross complementing gene 1 (XRCC1) Arg280His polymorphism as a risk for rheumatoid arthritis*. Rheumatol.Int., 2006. **26**(8): p. 749-751.
273. Lunn, R.M., et al., *XRCC1 polymorphisms: Effects on aflatoxin B-1-DNA adducts and glycophorin A variant frequency*. Cancer Research, 1999. **59**(11): p. 2557-2561.
274. Lei, Y.C., et al., *Effects on sister chromatid exchange frequency of polymorphisms in DNA repair gene XRCC1 in smokers*. Mutat.Res., 2002. **519**(1-2): p. 93-101.
275. Skjelbred, C.F., et al., *Polymorphisms of the XRCC1, XRCC3 and XPD genes and risk of colorectal adenoma and carcinoma, in a Norwegian cohort: a case control study*. BMC Cancer, 2006. **6**: p. 67.
276. Martin, N.J., et al., *Polymorphisms in the NQO1, GSTT and GSTM genes are associated with coronary heart disease and biomarkers of oxidative stress*. Mutat Res, 2009. **674**(1-2): p. 93-100.
277. Nebert, D.W., et al., *NAD(P)H:quinone oxidoreductase (NQO1) polymorphism, exposure to benzene, and predisposition to disease: a HuGE review*. Genet Med, 2002. **4**(2): p. 62-70.
278. Hildesheim, A., et al., *CYP2E1 genetic polymorphisms and risk of nasopharyngeal carcinoma in Taiwan*. J Natl Cancer Inst, 1997. **89**(16): p. 1207-12.
279. Wu, X., et al., *Associations between cytochrome P4502E1 genotype, mutagen sensitivity, cigarette smoking and susceptibility to lung cancer*. Carcinogenesis, 1997. **18**(5): p. 967-73.
280. Taylor, J.A., et al., *The role of N-acetylation polymorphisms in smoking-associated bladder cancer: evidence of a gene-gene-exposure three-way interaction*. Cancer Res, 1998. **58**(16): p. 3603-10.

281. Sugimura, H., et al., *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**(5): p. 413-7.
282. Pleis, J.R. and M. Lethbridge-Cejku, *Summary health statistics for U.S. adults: National Health Interview Survey, 2006.* Vital Health Stat 10, 2007(235): p. 1-153.
283. Tauras, J.A., *Differential impact of state tobacco control policies among race and ethnic groups.* Addiction, 2007. **102 Suppl 2**: p. 95-103.
284. Caraballo, R.S., et al., *Factors associated with discrepancies between self-reports on cigarette smoking and measured serum cotinine levels among persons aged 17 years or older: Third National Health and Nutrition Examination Survey, 1988-1994.* Am J Epidemiol, 2001. **153**(8): p. 807-14.
285. Arheart, K.L., et al., *Accuracy of self-reported smoking and secondhand smoke exposure in the US workforce: the National Health and Nutrition Examination Surveys.* J Occup Environ Med, 2008. **50**(12): p. 1414-20.
286. Everhart, J., et al., *Acculturation and misclassification of tobacco use status among Hispanic men and women in the United States.* Nicotine Tob Res, 2009. **11**(3): p. 240-7.
287. Pell, J.P., et al., *Validity of self-reported smoking status: Comparison of patients admitted to hospital with acute coronary syndrome and the general population.* Nicotine Tob Res, 2008. **10**(5): p. 861-6.
288. Bierut, L.J., et al., *Novel genes identified in a high-density genome wide association study for nicotine dependence.* Hum Mol Genet, 2007. **16**(1): p. 24-35.
289. Saccone, S.F., et al., *Supplementing high-density SNP microarrays for additional coverage of disease-related genes: addiction as a paradigm.* PLoS One, 2009. **4**(4): p. e5225.