ASSOCIATIONS BETWEEN GENETIC POLYMORPHISMS IN DNA BYPASS POLYMERASES AND BASE EXCISION REPAIR GENES WITH THE RISK OF BREAST CANCER

Leila Family

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.

Chapel Hill
2014

Approved by:

Andrew F. Olshan

Jeanette T. Bensen

Melissa A. Troester

Michael C. Wu

Carey K. Anders

ABSTRACT

LEILA FAMILY: Associations between genetic polymorphisms in DNA bypass polymerases and base excision repair genes with the risk of breast cancer
(Under the direction of Andrew F. Olshan)

Mutations in *BRCA1*, a DNA repair gene, have been associated with a lifetime increased risk of breast cancer (1). Therefore, researchers hypothesized there may be other DNA repair genes associated with breast cancer risk. However thus far, studies of common low-penetrant DNA repair SNPs have not yielded consistent results. In this proposed study, we hypothesized one or more of the following mechanisms may explain the lack of main SNP effects: combined SNP effects, modification by race or breast cancer subtype, and functional redundancy. To evaluate these hypotheses, we used genotype data from the Carolina Breast Cancer Study (1,972 cases and 1,776 controls) to investigate race-specific, subtype-specific, and combined SNP associations using unconditional logistic regression in two DNA damage pathways, base excision repair (BER) and translesion synthesis (TLS). For BER, we evaluated the association between 31 single-nucleotide polymorphisms (SNPs) in 15 genes and breast cancer risk. SKAT, a pathway-based analytic method, was used to evaluate the combined SNP effects within the BER pathway. Among Whites, our results showed a significant positive association for *NEIL2* rs1534862 and a significant inverse association for *PCNA* rs17352. Among African Americans, we found a suggestive positive association for *UNG* rs3219275 and an inverse association for *NEIL2* rs8191613. Tumor subtype analysis showed that *NEIL2* rs1534862 was associated with luminal and HER2+/ER- subtypes. SKAT analysis showed no significant combined effects between SNPs. For DNA bypass polymerases, we evaluated the association between 22 single-

nucleotide polymorphisms (SNPs) in 7 bypass polymerase genes and breast cancer risk. We found similar increased odds ratios for breast cancer with three *POLQ* SNPs (rs487848, rs532411, rs3218634), which were also in high LD in both races. Furthermore, analysis by specific tumor subtypes showed all three SNPs were associated with increased risk of luminal breast cancer. These significant findings need to be replicated independently in other studies. Overall, our results did not indicate associations with breast cancer, which may concur with the theory that our cells possess an intricate system of functionally redundant DNA repair mechanisms in order to avoid the catastrophic effects of genomic instability.

To Dr. Robert C. Millikan

Who always believed in me

"*The thing always happens that you really believe in,*

and the belief in a thing makes it happen."

-Frank Lloyd Wright

ACKNOWLEDGEMENTS

First, I would like to thank God for giving the strength and endurance to continue and finish my dissertation work. Also, I would like to thank my dissertation committee members, Dr. Andrew Olshan, Dr. Jeannette Bensen, Dr. Melissa Troester, Dr. Michael Wu, and Dr. Carey Anders, for providing me with their expertise and valuable feedback. I would like to extend a special thanks to my chair, Dr. Andrew Olshan, for his continuous guidance and support and for navigating me through the dissertation process and dealing with unforeseen circumstances.

I would also like to thank everyone who has contributed their time and energy to the Carolina Breast Cancer Study. Without their hard work and dedication, this work would not be possible. I would especially like to thank Mary Beth Bell, CBCS Project Manager, for her strong commitment to the study. Of course, I would also like to acknowledge the selfless contributions of all the women that participated in the Carolina Breast Cancer Study.

I would also like to thank my vast circle of support, including the helpful staff in the Epidemiology Department. I especially want to thank Nancy Colvin for her unwavering support and dedication to my success. I would also like to thank my awesome friends and family. In particular, I would like to give special thanks to Katie O'Brien and Lauren McCullough. These ladies went out of their way to make sure I was successful, including volunteering countless hours of their time to provide advice and feedback.

I would also like to take a moment to acknowledge and honor the work of Dr. Robert Millikan and Dr. Keith Amos, whose impact in the field of breast cancer will be felt for years to come.

Last, but certainly not least, I would like to thank my amazing parents, Gity and Siamak Family, for their unwavering support and unconditional love throughout my whole life, and the many selfless sacrifices they have made for me. I am truly blessed and honored to know such wonderful and genuinely humble people.

I would also like to acknowledge the sources of financial support that enabled me to complete this research: the University Cancer Research Fund of North Carolina, the National Cancer Institute Specialized Program of Research Excellence (SPORE) in Breast Cancer (NIH/ NCI P50-CA58223) and NRSA Pre Doctoral Training Grant (2007-2009), University of North Carolina at Chapel Hill Grant # 2-T32-CA09330.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AA	African-American
ACS	American Cancer Society
ADP	Adenosine diphosphate
AIMs	Ancestry informative markers
AP	Apurinic/Apyrimindinic site
APE1	AP endonuclease 1
ATM	Ataxia telangiectasia mutated homolog
ATP	Adenosine triphosphate
ATR	Ataxia telangiectasia and Rad3 related
BER	Base excision repair
BMI	Body mass index
BRCA1	Breast Cancer Gene 1
BRCA2	Breast Cancer Gene 2
BRIP1	*BRCA1* interacting protein C-terminal helicase 1
CBCS	Carolina Breast Cancer Study
CEU	HapMap population of individuals of northern and western European ancestry living in Utah
CGEMS	Cancer Genetic Markers of Susceptibility
CGHFBC	Collaborative Group on Hormonal Factors in Breast Cancer
CHEK2	Checkpoint kinase 2
CI	Confidence interval
CIMBA	Consortium of Investigators of Modifiers of *BRCA1*/2

CIS	Carcinoma in situ
CK 5/6	Cytokeratin 5/6
CLR	Confidence Limit Ratio
DAG	Directed acyclic graph
DCIS	Ductal carcinoma *in situ*
Df	Degrees of freedom
DNA	Deoxyribonucleic acid
dNMP	Deo-nucleotide monophosphate
dNTP	Deo-nucleotide triphosphate
DRC	DNA repair capacity
DSB	Double strand break
EGFR	Epidermal growth factor receptor
ER	Estrogen receptor
FEN1	Flap-structure-specific endonuclease 1
GWAS	Genome-wide association study
HER2	Human Epidermal Growth Factor Response 2
HR	Homologous recombination
HRT	Hormone replacement therapy
HWE	Hardy Weinberg Equilibrium
IHC	Immunohistochemistry
IR	Ionizing radiation
LCIS	Lobular carcinoma *in situ*
LD	Linkage disequilibrium

LIG3	Ligase III, DNA, ATP-dependent
LOH	Loss of heterozygosity
MAF	Minor allele frequency
MDB4	Methyl-CpG binding domain protein 4
MMR	Mismatch repair
MPG	N-methylpurine-DNA-glycosylase
MYH	A/G specific adenine DNA glycosylase
NAACCR	North American Association of Central Cancer Registries
NC	North Carolina
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NEIL1	Nei endonuclease VIII-like 1
NEIL2	Nei endonuclease VIII-like 2
NHEJ	Non-homologous end-joining
NHS	Nurses' Health Study
OC	Oral contraceptive
OGG1	8-oxoguanine DNA glycosylase
OR	Odds ratio
PAH	Polycyclic aromatic hydrocardon
PALB2	Partner and localizer of BRCA2
PARP1	Poly(ADP-ribose) polymerase 1
PARP3	Poly(ADP-ribose) polymerase family, member 3
PCNA	Proliferating cell nuclear antigen

POLB	DNA polymerase beta
POLH	DNA polymerase eta
POLI	DNA polymerase iota
POLL	DNA polymerase lambda
PR	Progesterone receptor
PTEN	Phosphatase and tensin homolog
R^2	Pairwise correlation coefficient
RFC1	Replication factor C (activator 1)
SCCOOP	Squamous cell carcinomas of the oral cavity and oropharynx
SEER	Surveillance, Epidemiology and End Results Program
SMUG1	Single-strand-selective monofunctional uracil-DNA glycosylase 1
SNP	Single nucleotide polymorphism
SSB	Single strand break
TDG	Thymine-DNA glycosylase
TLS	Translesion synthesis
TP53	Tumor protein p53
UNC	University of North Carolina at Chapel Hill
UNG	Uracil-DNA glycosylase
US	United States
UTR	Untranslated region
UV	Ultraviolet radiation
WEB	Western New York Exposures and Breast Cancer
WHR	Waist to hip ratio

XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1
YRI	HapMap population of Yorubans from Nigeria

CHAPTER 1. REVIEW OF THE LITERATURE

1.1 Introduction

In the past few decades, there have been significant strides in breast cancer research in both etiology and treatment. Breast cancer is one of the most investigated types of cancer in the US, garnering a large proportion of both private and public cancer research funding as well as media attention. Consequently, the results spawned from these research efforts have afforded many women new treatment and prevention options improving overall survival from breast cancer, with an estimated 3 million survivors in the United States (2).

1.2 Definition of breast cancer

The term "breast cancer" refers to a malignant tumor that has developed from cells in the breast. There are two main types of breast cancer: ductal carcinoma and lobular carcinoma. Most breast cancers are categorized as ductal carcinomas; that is they originate from the ducts, the passages that transfer milk from the lobule to the nipple. Lobular carcinoma originates from the lobules, the milk-producing glands. A less frequent type of breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast (2) (Figure 1)

Breast cancer can be further classified as invasive or noninvasive. Invasive cancer has spread from the milk duct or lobule to other tissues in the breast. Noninvasive or *in situ* cancers are confined within the ducts or lobules and named accordingly, ductal carcinoma *in situ* (DCIS) and lobular carcinoma *in situ* (LCIS).The majority of *in situ* breast cancers are DCIS, which

accounted for about 83% of *in situ* cases diagnosed during 2004-2008. Lobular carcinoma *in situ* (LCIS) is a marker for an increased risk of invasive cancer in the same or both breasts (2).

1.3 Epidemiology of breast cancer

1.3.1 Breast cancer incidence

Breast cancer is the most commonly diagnosed (non-skin) cancer in women in the United States, representing 29% of all female cancer cases (3). According to the American Cancer Society (ACS) 2013 Cancer Statistics Report, there will be an estimated 232,340 new cases of invasive and 64,640 new cases of carcinoma *in situ* this year (3). In North Carolina, there will be an estimated 7,090 new cases of female breast cancer in 2013 (3). In 2012, the age-adjusted incidence rate for breast cancer in the United States was 124.3 per 100,000 women per year. These rates are based on cases diagnosed during 2005-2009 from 18 SEER geographic areas (4). At a population level, women have a 12% risk (1 in 8) of developing breast cancer in the course of their lifetime.

After increasing for more than two decades, female breast cancer incidence rates began to slowly decrease in 2000, and then dropped abruptly by about 7% from 2002 to 2003. This large decrease was thought to be due to the decline in use of hormone replacement therapy (HRT) after menopause that occurred after the Women's Health Initiative ended their clinical trial early. The preliminary data showed that the risks of HRT may outweigh their benefits. The study linked the use of hormone therapy to an increased risk of breast cancer and heart disease (5). In the past couple of years, breast cancer incidence rates have stabilized (3).

Age-adjusted breast cancer incidence rates (per 100,000) also vary by race. White women are at the highest risk of breast cancer (122.3) followed by African Americans (116.1). Asians and American Indian women have the lowest risks, 84.9 and 89.2 respectively (3). In North

Carolina, the incidence rates (per 100,000) are slightly higher for Whites and moderately higher for African-Americans compared to the national average (124.5 and 122.3 respectively) (3). According to data from 18 SEER geographic regions, while incidence is similar for premenopausal White and African-African women, after menopause the incidence rates diverge, and postmenopausal White women have substantially higher incidence compared to postmenopausal African-American women (4) (Figure 2).

1.3.2 Breast cancer mortality

Although breast cancer mortality has declined by 30% in the past 25 years, it is still the second leading cause of cancer mortality in the United States after lung cancer for women (2).The ACS estimates 39,620 deaths due to breast cancer in 2013. Breast cancer accounts for about 3% of all-cause mortality and 14% of all cancer deaths in the US (2). At the state level, there will be an estimated 1,260 deaths from breast cancer in North Carolina in the year 2013. Death rates from breast cancer have been declining in the past few decades especially in premenopausal women. These decreases are believed to be the result of earlier detection through screening and increased awareness, as well as improved treatment.

There are differences in survival by both age and race. Although the mortality gap has lessened over the past several years, African-American women have higher mortality rates from breast cancer compared to White women, especially for younger women, despite the fact that Whites have a higher incidence (4) (Figure 2). Understanding this survival paradox is the first step in helping to improve breast cancer survival among younger African American women.

1.3.3 Non-genetic risk factors of breast cancer

In the past two decades, there have been a multitude of epidemiological studies evaluating the risk factors for breast cancer, primarily among postmenopausal White women.

Many of these individual studies yielded inconsistent results due to small sample size. The Collaborative Group on Hormonal Factors in Breast Cancer (CGHFBC) was established to aggregate data from multiple studies (10,000s of cases and controls) for a number of putative risk factors such as menarche and menopause, abortion, breastfeeding, alcohol and tobacco, and family history (6-10). Results from the Collaborative Group studies have provided conclusive evidence for several hormone-related factors such as nulliparity, older age at first birth, younger age at menarche and older age at menopause, long-term use of HRT being associated with increased risk of breast cancer (11-14). Lifestyle factors such as moderate alcohol use and postmenopausal weight gain have also been established to be positively associated with risk while physical activity has been associated with an inverse association (9, 11-19).The strongest risk factors are older age and personal family history; the latter which is correlated with genetic factors.

1.3.3.1 Non-genetic risk factors of breast cancer by race

Results from past breast cancer research studies may not provide a fully representative story. Most of this research relied on data collected from postmenopausal White women. In the past, it was difficult to examine risk factors in other racial/ethnic groups, since many studies did have enough power to evaluate breast cancer by race. Recent studies consisting of larger, more diverse cohorts of women have allowed researchers to re-evaluate these "well-established" risk factors by race (20, 21). These studies have identified several risk factors that may differ by race. For example, the effects of body mass index (BMI) may vary by race. While higher BMI in post-menopausal White women may increase breast cancer risk, in African-American women of any age BMI may act as a protective factor (22). While increasing parity and early age at first birth are considered protective factors in White women, these factors may have the opposite effect in

African-American women. Multi-parity was associated with increased risk of breast cancer among younger African-American women (for 3 or 4 pregnancies: OR =1.5, 95% (CI): 0.9, 2.6; for 5 or more pregnancies: OR = 1.4, 95% CI: 0.6, 3.1), but not among younger White women with the same number of pregnancies (20).

1.3.4 Genetic risk factors of breast cancer

Family history, one of the strongest risk factors for breast cancer, is linked to inherited genetic susceptibility. About 20 - 30% of women with breast cancer have a family history of the disease. Having one first-degree relative (i.e. mother, sister) may increase risk by two-fold while having two first-degree relative may increase risk by as much as three-fold (10, 23, 24). In the CGHFBC cohort, risk ratios for breast cancer increased significantly with increasing numbers of affected first-degree relatives compared with women who had no affected relatives ($p<0.0001$) (10). However a recent study reported no significant differences by the number of affected first-degree or second-degree family history (24). Overall, only 2.5% of breast cancer cases were found to be attributable to a positive family history (23).

In 1990, *BRCA1* was identified as one of the first breast cancer susceptibility genes followed by the discovery of *BRCA2* (1, 25). *BRCA1* and *BRCA2* genes are frequently mutated in familial breast and ovarian cancers. Women who carry these mutations have a lifetime increased risk of developing breast cancer (10). The average cumulative risks of breast cancer among *BRCA1* and *BRCA 2* carriers by age 70 are 65% (95% CI: 44-78%) and 45% (95% CI: 31-56%) respectively (26). However, *BRCA1* and *BRCA2* mutations are estimated to account for only 5-10% of all breast cancers and 15-20% of familial cases (27, 28). Several moderate penetrant genes such as *ATM*, *CHEK2*, *PTEN*, and *TP53* that predispose patients to genetic syndromes such ataxia telangiectasia, Li-Fraumeni, and Cowden's syndrome have also been

consistently associated with higher risk of breast cancer (28, 29). Recent studies have identified *BRIP1* and *PALB2* as two novel breast cancer susceptibility markers involved in DNA repair (30, 31). Many of these genes are involved in the regulation of DNA repair and checkpoint signaling (28).

It has been proposed that there are other common low-penetrant genes that may modify breast cancer risk in *BRCA1*/2 carriers. Antoniou and others established the Consortium of Investigators on Modifiers of *BRCA1*/2 (CIMBA), a large research collaborative between 40 study centers in 22 countries to investigate potential modifiers in this high risk subgroup. Results from these studies demonstrated that common variants in *LSP1* and *ZNF365* as well as several susceptibility loci, 2q35, 8q24, 12p11, 12q24, 9p21, 9q31.2, were associated with breast cancer in *BRCA1* and/or *BRCA2* carriers (32-40)

With the completion of the Human Genome Project in 2003 (41), researchers were enabled to use this cache of comprehensive genome-wide data to evaluate associations that were not possible with candidate gene studies. Large collaboration efforts such as Cancer Genetic Markers of Susceptibility (CGEMS) and Breast Cancer Association Consortium (BCAC) increased sample size and hence power in genome-wide association studies (GWAS). As a result, since the advent of GWAS, several breast cancer susceptibility markers from genetic association studies have been replicated and several novel susceptibility markers have been identified (42-47). While GWAS have enhanced our current understanding of breast cancer susceptibility genes and loci, these known genetic factors still only account for about 28% of the inherited causes of the disease (28).

1.3.4.1 Genetic risk factors of breast cancer by race

In addition, genetic risk factors may vary by race. Compared to White women with breast cancer, African American cases are less likely to have a *BRCA1* mutation (48). Furthermore, recent results from GWAS suggest that susceptibility loci may differ by race. Initially, many GWAS loci such as TERT-CLPM1L were identified in populations of European descent (33, 47, 49). However, Zheng 2012 failed to replicate this positive association in women of African descent (50). In addition, GWAS studies have shown modification by race for various breast cancer susceptibility loci (51).

1.4 Heterogeneity of breast cancer

Research from the past decade has shown that breast cancer is a complex and heterogenous disease involving multiple pathways and a combination of genetic and non-genetic risk factors. There is evidence of heterogeneity by both hormone receptor status and more recently by intrinsic tumor subtype. Historically, breast cancer tumors are classified based on their hormone receptor status (i.e. estrogen receptor (ER) and progesterone receptor (PR)) and *HER2* status mainly to guide clinical treatment options (52). Endocrine therapies such as aromatase inhibitors were used for hormone receptor positive (ER+ and PR+) tumors, while therapeutics such as Herceptin were used for tumors overexpressing *HER2* (52). Tumors that were negative for all three of these markers were classified as triple-negative and were not candidates for endocrine therapy targeted treatments. Several studies show risk profile differences between hormone-positive vs. hormone-negative tumors (53, 54). A Carolina Breast Cancer Study (CBCS) report showed that several hormone-related factors were associated with stronger increased risks for ER+PR+ than for ER-PR- breast cancer; including early age at menarche, nulliparity/late age at first full-term pregnancy or a high body mass index (BMI)

among postmenopausal women and high waist to hip ratio (WHR) among premenopausal women. Conversely, family history and medical radiation exposure were associated with ER-/PR- tumors (53). Prospective data from the Nurses' Health Study confirmed significant differences by ER/PR status for age, menopausal status, postmenopausal BMI, adverse effect of first pregnancy, and past use of postmenopausal hormones (54). Additionally, risk factors profiles may vary by *HER2/neu* status. A CBCS report showed that early age at menarche, higher WHR, and family history of breast or ovarian cancer were associated with increased odds ratios (ORs) for both *HER2/neu+* and *HER2/neu-* breast cancers while breastfeeding for more than a year was inversely associated (53).

Recent technological developments in microarray analysis have led to the molecular subtyping of breast cancer tumors to further discern breast cancer heterogeneity. Perou et al. used cDNA microarrays to measure the gene expression of more than 1700 genes. A hierarchical clustering algorithm identified four different patterns of gene expression in *in vitro* human breast cells and tumors: ER+/luminal, basal-like, *HER2*+/ER-, and normal (55, 56). In a second study with more samples, Perou et al. further dichotomized luminal tumors into luminal A and luminal B (56). In addition, Sorlie et al. was able to define these molecular types using a set of only 534 genes (57). Nielsen et al. used immunohistochemistry to categorize molecular profiles based on a panel of four antibodies (ER, *EGRF*/*HER1*, *HER2*, and cytokeratin 5/6) and found that this method was equivalent to the gene expression technique (58). Carey further updated IHC subtype definitions to include PR status as well as dichotomize *HER2* status into *HER2+* or *HER2-* (59). Therefore, breast cancer tumors were classified into 4 distinct molecular subtypes: luminal A (ER+ and/or PR+, *HER2-*), luminal B (ER+ and/or PR+/*HER2+*), *HER2+*/ER- (ER-,

PR-, *HER2+*), basal-like (ER-, PR-, *HER2-*, *CK 5/6+* and/or *EGFR+*) and an unclassified category (59).

1.4.1 Non-genetic risk factors of breast cancer by subtype

More recently, molecular profiling of breast cancer tumors has enabled researchers to evaluate risk factors based on these "intrinsic" subtypes. Several studies have observed differences in the associations between breast cancer risk factors and subtypes of breast cancer. Basal-like subtype may have a different risk factor profiles compared to the luminal A subtype (22, 60-65). Compared to the luminal A subtype, basal-like cases were also more likely to have younger age at menarche (62), younger age at first full-term pregnancy(22, 62, 66), higher parity (22, 62, 66, 67) were less likely to breastfeed (22, 62, 64, 66, 67). While long term breastfeeding (>6 months) was inversely associated with breast cancer across subtypes, the protective effect was strongest for basal-like tumors (68). In addition to reproductive factors, obesity or elevated BMI was also associated with increased risk of basal-like breast cancer and luminal B cancers compared to luminal A cases, especially for premenopausal women (22, 62, 64, 66, 68). Many studies also reported that family history may play a bigger role for women with basal-like tumors compared to other subtypes, especially for premenopausal women (24, 61, 64, 69).

Reports from the CBCS have suggested heterogeneity among *in situ* tumors, Phillips et al. showed that many risk factors for invasive and high grade *in situ* tumors were similar, but differed from risk factors for low or medium grade *in situ* tumors in both strength and magnitude of effects. For example, higher parity showed a strong inverse association with high grade DCIS but had a weaker inverse association for low to medium grade DCIS. In addition, ten or more years of oral contraceptive showed a positive association with high-grade DCIS and IBC but an

inverse association for low to medium DCIS (70). In summary, there may be heterogeneity within *in situ* tumors which needs to be further investigated.

In addition, basal-like tumors have poorer prognoses compared to luminal tumors (57, 59, 71). Basal-like tumors showed more aggressive features compared to Luminal A, including higher mitotic index ($P < 0.0001$), higher grade ($P < 0.0001$), and a higher frequency of p53 mutations ($P < 0.001$)(57). *In situ* basal-like cancers also shared similar poor clinical outcomes with invasive cases (72)

Several studies have shown that basal-like tumors occur at a higher incidence among African-Americans compared to whites (71, 73, 74). In a study of Ghanaian women, African-American and white women from the US, proportion of African ancestry was significantly associated with triple-negative tumors. Ghanaians had the highest prevalence of triple-negative tumors (82.2%), followed by African Americans (32.8%) and White Americans(10.2%) (75).

1.4.2 Genetic risk factors of breast cancer by subtype

Carriers of *BRCA1* mutations are at higher risk of developing basal-like tumors (76-81). It has been estimated that 80-90% of cancers in *BRCA1* mutation carriers are of basal-like subtype (81). Furthermore, the prevalence of *BRCA1* mutation carriers in triple-negative tumors was approximately 20% and 11%, respectively, in two studies (79, 80). Therefore, loss of *BRCA1* function may have an etiological role in the development of the basal-like phenotype. Expression microarray analyses have also indicated similarities in gene expression between *BRCA1* cancers and sporadic basal-like cancers (82). *BRCA1*-mutated and basal-like tumors share many similar characteristics including higher levels of genomic instability compared to ER+ or luminal tumors. Compared to other subtypes, basal-like tumors have the highest levels of genomic instability as represented by greater number of insertions, deletions and copy number

alterations. At least three studies reported loss of 5q and gain of 10p in basal-like cancers (83-85). Van Loo et al. showed that basal-like tumors were associated with low ploidy, high frequency of loss of heterozygosity (LOH), and the highest frequency of copy number events when compared to the other subtypes (86). There are also a higher number of chromosomal rearrangements and aneuploidy increase in defective DNA repair genes. These results provide evidence for both deficient DNA repair genes and basal-like breast cancer being associated with genomic instability.

Recently identified breast cancer susceptibility loci in GWAS (*CASP8, FGFR2, TNRC9, MAP3K1, LSP1, 8q24, 2q35, 5p12, 16q12*) may also vary by tumor characteristics such as hormonal status or intrinsic subtype. Susceptibility loci in *FGFR2, TNRC9, 8q24, 2q35, and 5p12, 9q13.2* had stronger associations for estrogen receptor-positive (ER+) disease than estrogen receptor-negative (ER-) disease (51, 87-89). Two candidate loci in *CASP8* (rs1045485, rs17468277) and *TGFB1* (rs1982073), were strongly associated with the risk of PR- tumors and 16q12 and 2q35 were associated with basal-like subtype (87).

A common variant at the *TERT-CLPTM1L* locus was also found to be associated with estrogen receptor-negative breast cancer (90). Of note, this locus was not replicated in women with African ancestry (91). In a recent population-based case-control study, Domagala et al. found several *CHEK2* mutations associated with different molecular subtypes of breast cancer. Truncating mutations were associated with luminal B, and I157T *CHEK2* mutation was associated with luminal A. (92, 93). The GWAS discovery of a novel locus 19p13 was shown to both modify risk of breast cancer in *BRCA1* mutation carriers as well as in hormone receptor negative and triple negative cases in the general population (32, 33, 43). In the CIMBA study, analyses based on tumor histopathology showed that 19p13 variants were associated with ER-

breast cancer for both *BRCA1*/2 mutation carriers (32, 33). Results from the Breast Cancer Association Consortium (BCAC) further showed that 19p13 was associated with triple negative subtype (43). MERIT40 interacts with *BRCA1* and plays a role in the repair of double-strand break in the HR pathway. These results provide evidence that DNA repair may vary by breast cancer tumor subtype.

1.4.3 Summary of breast cancer risk factors

Breast cancer is a multifactorial disease, which results from the combined effect of genetic and non-genetic risk factors that can vary by both race and subtype. Linkage association studies were the first to hint at a genetic component underlying familial breast cancer. Furthermore, women with a first degree family history of breast cancer were found to be at almost twice the risk as women without a family history (10). Therefore, a positive family history of breast cancer may serve as a surrogate for shared genetic variation (94).

1.5 Variation in DNA repair capacity

The discovery of mutations in *BRCA1* in the early 1990s offered insight into the genetic etiology of familial breast cancer. Carriers of *BRCA* 1/2 mutations were found to have deficient DNA repair capacity (DRC) compared to normal controls (95). Experimental studies showed that the near complete loss of DNA repair capacity can lead to genetic instability and a high risk of developing cancer (96). However the prevalence of *BRCA1* mutations and mutations in other moderate to high penetrant DNA repair genes such as *BRCA2, ATM, CHEK2, PTEN, BRIP1, and PALP2* are rare in the general population and only explain 15-20% of genetic susceptibility to breast cancer (25, 97-99).

1.5.1 Low-penetrant common DNA repair variation in breast cancer

The polygenic model of cancer was proposed to explain the missing heritability (100, 101). Under this model, a combination of multiple low penetrance genes would contribute to overall genetic risk. There is increasing evidence that mild reductions in DNA repair capacity, assumed to be the consequence of common genetic variation, can also affect breast cancer susceptibility. The extensive variation in the coding regions of DNA repair genes and the large number of genes in each DNA pathway results in complex genotypes with potential to impact cancer risk in the general population (94, 102). In our proposed study, we focused our investigation on these variants in common, low-penetrant DNA repair variants to evaluate their potential association with breast cancer.

1.6 DNA damage responses

The combination of endogenous and exogenous sources of DNA damage can result in as many as one million DNA lesions per cell per day (103). Endogenous sources include replication errors and spontaneous reactions, while exogenous sources of DNA damage include X-rays, oxygen radicals, alkylating agents, UV light, polycyclic aromatic hydrocarbons (PAHs), IR, and anti-tumor agents. Unrepaired DNA damage can result in gene mutations such as point mutations, deletions and insertions as well as chromosomal alterations such as chromosomal rearrangements and aneuploidy. In order to maintain genomic stability, organisms have evolved a complex series of damage repair responses to process DNA damage in a timely and efficient manner including 1) apoptosis, 2) checkpoint signaling 3) DNA repair and 4) damage tolerance (Figure 3). The proposed study will focus on the latter two mechanisms, specifically base excision repair and translesion synthesis.

1.6.1 Overview of DNA repair

DNA repair mechanisms protect somatic cells from mutations in tumor suppressor genes and oncogenes that can lead to cancer initiation and progression. Over the course of evolution, cells have evolved several DNA repair pathways for repairing distinct types of DNA damage. Specialized DNA repair pathways include direct reversal repair, mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER), and recombinational repair (homologous recombination (HR) and non-homologous end-joining (NHEJ) (104). Figure 4 summarizes the source of DNA damage, the ensuing DNA lesion, and the DNA repair pathway used to repair the lesion. Functional DNA repair plays an important role in tumor suppression. There have been dozens of epidemiological studies examining common variation in multiple DNA repair pathway. The focus of this study will be on single nucleotide polymorphisms (SNPs) in the BER pathway.

1.6.2 Overview of base excision repair (BER)

Base excision repair (BER) is the fundamental pathway responsible for the repair of damaged DNA bases induced by various sources of endogenous and exogenous damage. BER is specialized to repair non-bulky DNA base lesions such as base adducts and abasic sites caused by deamination, alkylation, or oxidation. The repair process consists of five enzymatic steps: 1) cleavage of the sugar-phosphate chain, 2) excision of the abasic (AP) site, 3) removal of the remaining sugar-phosphate chain, 4) DNA synthesis, and 5) ligation (105). Table 1 summarizes BER genes and their functions in DNA repair.

To date, there are at least 11 known human DNA glycosylases. DNA glycosylases play an important role in the initial recognition of a lesion and recruitment to the site of the damage (106). Different glycosylases are specialized for different lesions and some glycosylases may

recognize more than one substrate (Table 2). DNA glycosylases initiate repair by releasing the modified/damaged base out of the double helix and cleaving the N-glycosidic bond of the damaged base, resulting in an apurinic/ apyrimidinic (AP) site. The location and type of the AP site can also be determining factors on which glycosylase is recruited to the site (107, 108). If the AP site was created by a glycosylase that does not possess AP lyase activity (*UNG*, *SMUG1*, *TDG*, *MPG*, *MDB4*, *MYH*), or *NTH1* and *OGG1*, repair of the AP site is *APE1*-dependent. A newly discovered family of glycosylases (*NEIL1*, *NEIL2*, and *NEIL3)* was shown to efficiently repair AP site independently from *APE1* (107, 109, 110).

The repair of AP sites is crucial since they can interrupt normal DNA replication, and become a threat to genomic integrity. *APE1* or a member of the *NEIL* family converts the lesion into a single-strand break (SSB). The SSB requires removal of the altered 3′-terminal groups prior to ligation. After removal of obstructive termini, replacement of the excised nucleotide can be completed either via short-patch where a single nucleotide is replaced or long-patch BER where 2-10 new nucleotides are synthesized (111). Choice of pathway depends on several different factors including the type of lesion, the cell cycle stage, and whether the cell is terminally differentiated or actively dividing (112). The short-patch pathway requires a different set of genes from the long-patch pathway (111, 113). The main distinction is whether the abasic sugar is oxidized or reduced, which dictates if *POLB* is involved (short-patch) or not (long-patch) (114).

Finally, the posttranslational modification of proteins is mediated by poly (ADP ribose) polymerases (PARPs). Members of the PARP family *(PARP1*, *PARP2*, and *PARP3)* catalyze the transfer and polymerization of ADP ribose (115). *RFC1* loads the *PCNA* clamp onto DNA,

thereby recruiting DNA polymerases to the site of DNA synthesis to the 3' end of primer, promoting DNA synthesis (116, 117).

1.6.2.1 Base excision repair and breast cancer

It has been proposed that base excision repair may be involved in tumor suppression (118, 119). Experimental studies have demonstrated that deletion of certain BER genes is associated with an increased mutation rate in a variety of organisms, and hypothesize that this loss could contribute to the development of cancer in humans (105). In addition, several dozen case-control genetic association studies have been conducted (Table 3).The following section summarizes the experimental and epidemiologic literature for the association between base excision repair and breast cancer.

1.6.2.1.1 *UNG*

While no variants have been associated with breast cancer, two novel SNPs (*UNG* Arg88Cys and *UNG* Gly143Arg) have been identified using mutational analysis in colorectal cancer and glioblastoma cell lines, respectively (106, 120). In an *in vivo* study, knockout of the *UNG* gene led to carcinogenesis in mice. Older (>18 months) *UNG* knockout mice developed B cell lymphomas compared with only 1.3% of control animals (121).

1.6.2.1.2 *SMUG1*

In a 2011 Western New York Exposures and Breast Cancer (WEB) report (1,077 cases, 1,910 matched controls), two polymorphisms in the *SMUG1* promoter region (rs2029166 and rs7296239) were found to moderately effect the risk of breast cancer in heterozygotes (OR=1.3,

95% CI: 1.1-1.5)(122). Another study examined the association between *SMUG1* variants and uracil blood concentration in 431 participants from the Boston Puerto Rican Health Study and found a significant association with the SNPs studied. Increased level of uracil misincorporation may induce mutagenic lesions and possibly lead to increased cancer risk (123).

1.6.2.1.3 *MBD*

Frameshift mutations in *MBD4* have been associated with gastrointestinal cancers in two Asian study populations (124, 125). The Glu346Lys polymorphism has been associated with lung, esophageal, and gastrointestinal cancers (125-127). To our knowledge, there are no experimental or epidemiologic studies investigating genetic variants of *MBD4/MED1* with breast cancer risk.

1.6.2.1.4 *MPG*

Based on the literature, there are no known experimental or epidemiologic studies associating genetic variants of *MPG* with breast cancer risk. Three laboratory studies reported altered expression of *MPG* in human gonad cells and astrocytic tumors (128-130).

1.6.2.1.5 *MYH/MUYTH*

Mutations in the *MUYTH* gene result in MAP *(MUTYH*-associated polyposis) a heritable predisposition to colorectal tumors (131, 132). While a Dutch study initially reported increased mutation frequency of several *MUTYH* SNPs among women of families with HBCC (Hereditary Breast and Colon Cancer) (133), a validation study failed to replicate these results in a larger case-control study (132). In a Chinese case-control study (545 cases, 545 controls), there were no associations with breast cancer overall, but the dominant model for AluYb8 insertion was

significantly associated with increased risk of early-onset breast cancer (<55 years old) OR=1.51: 95% CI: 1.09-208) (134).

1.6.2.1.6 *TDG*

Polymorphisms G199S and V367M are the most common genetic polymorphisms in human *TDG*. A recent Polish study revealed a possible association with these *TDG* polymorphisms and lung cancer however these results may be biased due to small sample size (135). Further studies are needed to fully understand the relationship between *TDG* and cancer.

1.6.2.1.7 *OGG1*

Functional lab evidence has suggested that rs1052133 (S326C) in *OGG1* may be associated with decreased DNA glycosylase activity in the repair of 8-oxoG, a mutagenic byproduct of exposure to reactive oxygen (136). However, the results from epidemiological studies have been less conclusive. At least six independent epidemiologic studies have evaluated the association between the S326C polymorphism with breast cancer risk (137-142). Two reports suggested an increased risk for S326C (137, 138) while two reports failed to find any significant association (141, 143). An earlier review by Goode had identified S326C as being associated with increased breast risk (144), however two recent meta-analyses were not able to replicate this finding (145, 146). In a review of 14 functional studies and 19 epidemiological studies, Weiss et al. found no significant association between the *OGG1* polymorphism and breast cancer (146). A recent case-control study in China (518 cases, 777 controls) showed two functional variations in 5'-UTR of *OGG1* gene were significantly associated with the risk of breast cancer (OR=2.0 95% CI: 1.0-3.9 and OR=2.4 95% CI: 1.2-5.0) (147).

1.6.2.1.8 *NEIL1*

In experimental studies, *NEIL1* have been shown to interact with other BER genes including *POLB*, *LIG3*, and *PCNA* (109, 148). In an *in vitro* study, *NEIL1* downregulation enhanced spontaneous mutation by three-fold in Chinese hamster and human cell lines (149). To our knowledge, no epidemiologic studies have been conducted.

1.6.2.1.9 *NEIL2*

NEIL2 was shown to interact with *POLB* and *LIG3* (109, 110, 150). Variant risk genotypes in *NEIL2* have been associated with increased risk in SCCOOP (head and neck cancers) and lung cancers (151, 152). In an *in vivo* study, *NEIL2* expression was significantly reduced by over 50% in the presence of 2 SNPs (rs74800505 and rs8191518) which were in significant LD (153). *NEIL2* rs6982453 was associated with a significantly protective effect in breast cancer in the Multiethnic Cohort Study (OR=0.86, 95% CI: 0.79-0.94, $p<0.001$) (154).

1.6.2.1.10 *APE1*

A number of functional polymorphisms have been identified in *APE1* with the most commonly studied SNP being *APE1* Asp148Glu (155). This polymorphism has been associated with risk of bladder, lung, prostate and gastric cancers (156-159). Overexpression of *APE1* has been linked to chemotherapy and radiation therapy resistance (160). However the epidemiologic evidence for *APE1* and risk of breast cancer is inconclusive. While two case-control studies reported a borderline significant protective association for carriers of heterozygous variant (Asp/Glu) in Thai and White American women respectively (138, 161), two other reports found null associations for African American and White American women respectively (143, 161). A meta-analysis of 8 studies did not reveal any significant association for *APE1* Asp148Glu for any

genetic models (162). However, a recent lab study has linked deregulation of *APE1* acetylation to triple negative breast cancer (163).

1.6.2.1.11 *POLB*

DNA polymerase beta or *POLB* has been shown to be overexpressed in several cancers (164-166). Seven germline mutations (P242R, E295K, G231D, K289M, E232K, T233I) have been identified in *POLB* (167). In an *in vitro* study, Yamtich et al. 2012 found that expression of these variant germline SNPs could be related to increased cancer susceptibility following treatment with an alkylating agent (165). Gieseking also identified two *POLB* SNPs (E232K and T233I) to be associated with lower fidelity when processing undamaged DNA, which may lead to mutagenesis (168).Estimates of somatic mutations in Pol β range from 15% to 75% of tumors in various types of cancer (169, 170). Functional analyses have implicated many of these variants in cancer etiology and/or progression (170-173). An *in vivo* study showed that overexpression of *POLB* variants in mouse cells resulted in cellular transformation. Furthermore, knockout of *POLB* caused embryonic lethality. While there have been no positive associations between *POLB* and breast cancer, there have been multiple reports suggesting *POLB* involvement in lung and colorectal cancers in other epidemiological studies (174-176)

1.6.2.1.12 *XRCC1*

While the majority of studies did not find any significant associations (177-183), the XRCC1 Arg399Gln polymorphism was associated with a protective effect in one report (184) and an increased risk in seven other reports (105, 138, 139, 185-188). We suspect that these significant positive findings were mostly false positives due to study design and low power issues. Several of these studies had smaller sample sizes which may not have had adequate power to detect modest SNP effects. This was evidenced by wide confidence intervals or high

CLR indicating imprecise estimates in several studies (138, 139, 186, 187, 189, 190). In addition, results may be biased due to selection of controls (i.e. hospital-based controls (189) or cases (i.e. cases selected for family history) (184). Alternatively, since the majority of significant findings were from studies in Asian populations, there is the possibility of effect modification by Asian race/ethnicity. At least four independent meta-analyses of *XRCC1* Arg399Gln have provided evidence for this theory (143, 191-193). Additionally, two U.S.-based population-based case-controls studies found no overall associations, but showed subgroup effects for African-American and postmenopausal women in the CBCS and WEB study, respectively (92, 137). While no significant associations were observed in premenopausal women, postmenopausal women with any Gln variant had increased risk of breast cancer (OR = 1.24; 95% CI: 1.01-1.51) (137). Duell found a similar increased risk for African-Americans in the CBCS (OR=1.5 95% CI: 1.1-2.3) (92).

Two reports found an increased risk for at least one variant of *XRCC1* Arg194Trp (161, 189), while another report did not (92). In a meta-analysis of 11 studies including both White and Asian populations, Zhang found no association between Arg194Trp and breast cancer risk (143).

The majority of these studies reported no association with *XRCC1* Arg280His with the exception of one population-based case-control study of women from Cyprus. Loizidou et al. found homozygous carriers of *XRCC1* 280His to have an increased risk of breast cancer (OR=4.7; 95% CI: 1.0-21.7, P=0.03). Although this study contained 1,109 cases and 1,177 controls, a highly imprecise estimate was reported (194). The authors reported that this SNP failed HWE ($p<0.05$) which may indicate genotyping error. Meta-analyses of *XRCC1* Arg280His have yielded conflicting results. While Hung did not find any association between cancer risk

and the *XRCC1* Arg280His SNP, two other meta-analyses reported an overall increase risk of cancer for the variant genotypes (His/His + Arg/His) compared with the wild-type homozygote genotype (Arg/Arg) (191, 195).

1.6.2.1.13 *LIG3*

Knockout of *LIG3* are embryonic lethal in mice (196). However, to our knowledge, there are no known *LIG3* SNPs that have been studied for association with cancer in the epidemiologic study literature (114).

1.6.2.1.14 *FEN1*

FEN1 was significantly upregulated and aberrant expression was associated with promoter hypomethylation in breast cancer cells in a gene expression study of 241 matched pairs of cancer and normal tissues (197).

1.6.2.1.15 *PARP1*

PARP1 has been shown to inhibit DNA repair in both the short and long patch pathways (198, 199). Conversely, cells deficient in *PARP1* show increased rates of repair (198). Bieche and colleagues reported overexpression of *PARP1* and low genomic instability in a study of breast cancer cells (200). In another study, inhibition of *PARP1* was shown in tumors from *BRCA* mutation carriers (201). However, a recent meta-analysis of 8 studies did not show an association between *PARP1* V762A and breast cancer (162).

In a lab-based study, *PARP1*-deficient cells were assessed for their capacity to repair AP sites induced by uracil or 8-oxoguanine. For both DNA lesions, *PARP1*-deficient cells were about half as efficient as wild-type cells for short-patch repair synthesis, and were highly

inefficient in the long-patch repair pathway. Inefficient BER occurred when both *PARP1* and *POLB* were absent (199).

In a subset of Nurses' Health Study II cohort (NHS II), Han 2009 found *PARP1* rs10915985 to be significantly associated with premenopausal breast cancer in the additive model (OR=1.31, 95% CI: 1.04-1.64), however this SNP was not genotyped in CBCS (202).

1.6.2.1.16 *PCNA*

Several yeast models have associated *PCNA* mutations with cancer and genomic instability (203). In addition, Ma and colleagues sequenced the coding region and adjacent noncoding region of *PCNA* in 60 individuals and identified 9 sequence variants, including 7 SNPs which were located in introns involved in the control of *PCNA* gene expression. Results from the analyses showed no associations with melanoma, breast cancer or lung cancer compared with healthy controls (204).

1.6.2.1.17 *RFC1*

Experimental studies in have shown *RFC1* to function in both DNA replication and repair, specifically NER (116, 205). Replication factor C (*RFC*) is a five-subunit DNA polymerase accessory protein that functions as a structure-specific, DNA-dependent ATPase. *RFC* acts as a sensor in the DNA damage checkpoint pathway and plays a role in DNA synthesis. To our knowledge, we are not aware of any epidemiologic studies examing *RFC1* variants.

1.6.2.2 Critique and Summary of BER literature

Despite the strong associations of *BRCA1* and *BRCA2* and moderate penetrant genes such as *CHEK*, *PALB*, and *ATM* with breast cancer risk, the risks conferred by individual low

penetrant BER genes for breast cancer have been underwhelming and attempts to understand the contribution of low penetrant SNPs has been challenging. To date, there have been dozens of population-based case-control genetic studies, including the Carolina Breast Cancer Study (CBCS), that have investigated the association between common genetic variation in BER genes (*XRCC1*, *OGG1*, *APE1*, *NEIL1* and *NEIL2*) and breast cancer risk (92, 143, 182, 184, 189, 206-208). Most studies examined BER SNPs in the *XRCC1*, *APE1*, and *OGG1* genes. While the majority of studies of White women showed no significant associations with *XRCC1* SNPs (rs1799782, rs25489, and rs25487), several studies in non-White populations indicated potential effect modification by race/ethnicity for rs25487, Arg399Gln.The evidence for *OGG1* Ser326Cys and *APE1* Asp148Glu polymorphism and breast cancer risk was null to weak (137, 138, 141-143). Additionally, findings from other BER SNPs studies have been inconclusive.

This failure to reveal significant associations between individual BER SNPs and breast cancer is not surprising, given that carcinogenesis is a multistep, multi-genic process. Therefore it is plausible that any one single genetic polymorphism would not have a dramatic effect on cancer risk. Interaction between multiple common low-penetrant SNPs may be needed to produce a significant effect. The polygenic model of cancer posits that although the risks conferred by an individual locus are small, some risks may act multiplicatively or additively. In this model, each variant is only one of the many genetic and environmental causal factors, each of which are neither necessary nor sufficient to individually cause the disease. Therefore, accumulation of mutations may be more important than a single SNP mutation (209).

Supporting evidence for the polygenic or multi-SNP effect in DNA repair is abundant. Despite not finding main effects, many DNA repair studies have found significant multi-SNP effects. As an example, Harlid et al. 2012 examined the individual and joint effects between 10

GWAS-validated breast cancer SNPs in a large European biobank-based study (3,584 cases, 5063 controls) and found a highly significant trend for increasing breast cancer risk with increasing number of previously validated risk alleles (p-trend 5.6×10^{-20}) and for the maximum versus the minimum number of risk alleles (OR=1.84, 95% CI: 1.59-2.14) (210).

Recent studies have used hierarchical modeling and other multi-SNP methods to evaluate cancer risk at a gene or pathway level in various cancers (158, 211, 212). For breast cancer, two reports from the Cancer Genetic Markers of Susceptibility (CGEMS) Project, a study nested within the Nurses' Health Study, evaluated the combined effects of low-penetrant SNPs in multiple DNA repair pathways using Admixture Maximum Likelihood (AML) and Kernel machine tests (202, 213). Han 2009 found several significant main effects for SNPs in *PARP1*, *NEIL2*, *APE1*, and *POLD* for premenopausal women ($p<0.05$)(202), while a second report failed to replicate any of this findings in postmenopausal women (213).

Another potential theory relates to the functional redundancy of genes to maintain genomic stability. For example, in mouse models, knockouts of core BER proteins such as *XRCC1*, *POLB*, *APE1*, and *FEN1* all result in embryonic lethality (214-217). Furthermore, the coding regions of *PCNA* and *FEN1* are highly conserved (204). On the other hand, for DNA glycosylases with multiple redundant pathways, there are no obvious phenotypes in nullizygous mice lacking a single oxidative DNA glycosylase. Studies of double knockout mice have shown they are prone to tumorigenesis. Chan et al. showed that targeted deletion of *NTH* and *NEIL1* resulted in mice with a higher frequency of lung and liver tumors compared to single knockout mice (218). In another experimental study, knockout of *MYH* or *OGG1* individually showed very little effect, however *MYH*/*OGG1* double mutant mice showed high susceptibility to tumor

formation (219). These studies suggest functional redundancy of DNA glycosylases and highlight the integral role of BER genes to preserve genomic integrity.

1.6.3 Overview of DNA tolerance

The process of maintaining accurate DNA replication is essential to the genomic stability of all cells. In the event that DNA damage should escape repair surveillance prior to initiation of DNA replication, organisms have evolved a series of tolerance mechanisms for allowing replication and cell division to process.

The first step in DNA replication involves the unwinding of DNA at the origin. The replication fork is a structure that forms within the nucleus during DNA replication. It is created by helicases, which break the hydrogen bonds holding the two DNA strands together. The resulting structure has two branches, each one made up of a single strand of DNA. These two strands serve as the template for the leading and lagging strands, which will be created as DNA polymerases match complementary nucleotides to the templates. The leading strand is synthesized continuously in the direction of replication fork, 5' to 3', while the lagging strand is synthesized in small pieces (Okazaki fragments) backward from the overall direction of replication (220, 221). Several DNA polymerases are involved in DNA replication. DNA polymerase alpha initiates DNA synthesis on both the leading and lagging strands providing an RNA primer and synthesizing approximately 20-30 bases of DNA. Pol epsilon (*POLE)* and pol delta (*POLD2*) elongate these primers created by pol alpha (222). *PCNA* is the sliding clamp for *POLD1* and *POLE* (223). *POLD1* and *POLE* also possess proofreading 3'-5' exonuclease activity that is important in preventing mutations.

DNA replicative polymerases, such as pol alpha, pol epsilon (*POLE*), and pol delta (*POLD*) which carry out the bulk of DNA synthesis, have evolved to be very precise and

efficient, with an estimated error rate of 1 in 10 billion base pairs (224). Despite this high fidelity, a replication error may generate a one-sided double-strand break (DSB) or degrade to a full DSB if it not repaired prior to initiation of DNA replication (225, 226). In order to resume DNA replication at a stalled replication fork, two damage tolerance mechanisms have been proposed; template switching in homologous recombination (HR) and translesion synthesis (TLS) (227). Posttranslational modification of *PCNA* by ubiquitin may play a role in determining which DNA repair tolerance mechanism to employ. Studies showed that the mono-ubiquitylation of *PCNA* may activate translesion synthesis by damage-tolerant DNA polymerases, while poly-ubiquitylation of *PCNA* may activate error-free pathway involving template switching in HR (228-231). During template switching in HR, although normal synthesis of DNA is blocked by a lesion on one of the template strands, synthesis on the undamaged template strands can continue to a limited extent. The newly synthesized daughter strand is used as the template, hence the term "template switching". If template switching is unsuccessful, translesion synthesis is activated to bypass the lesion (119, 222, 227).

1.6.4 Overview of translesion synthesis (TLS)

The focus of this study will be on the second DNA tolerance mechanism: translesion synthesis (TLS). Translesion synthesis is conducted by a specialized type of DNA polymerases. Aptly named, bypass polymerases do not directly repair the damage, but rather bypass or tolerate the damage to prevent replication fork stalling. Unlike replicative polymerases, bypass polymerases lack 3' to 5' exonuclease (proofreading) activity and are able resume replication without an undamaged template (232, 233). However, this also contributes to their low fidelity and potential mis-incorporation of nucleotides (234).

Evidence from experimental studies shows bypass polymerases as being both efficient and mutagenic. The ability of DNA bypass polymerases to bypass DNA lesions was first described in yeast. Nelson and colleagues found that *REV3L* (pol zeta) successfully mediated the bypass of UV-induced thymine-thymine cyclobutane pyrimidine dimers (TT-CPDs) and *REV1* was able to insert deoxycytidine monophosphate (dCMPs) opposite abasic sites (235, 236). These findings were subsequently followed by the discovery of UV lesion bypass activity of human pol-eta, which was shown to be defective in a group of xeroderma pigmentosum (XP) patients (237, 238).

Reduced fidelity of DNA bypass polymerases may be dependent on a number of factors including physical structure, location within the cell cycle, and type of lesion. All DNA polymerases possess a "right hand-like" structure, which share three common domains (palm, thumb, and little fingers). However, differences in the active sites between family members may contribute to the low fidelity of these proteins (239-241). The discovery of the crystal structures of several Y-family DNA polymerases have implicated that more open active sites may be the reason for the error propensity of low-fidelity polymerases (242).

The level of fidelity of bypass polymerases has also been shown to be lesion specific. Different lesions are bypassed in an efficient or mutagenic manner depending on the bypass polymerases involved (Table 4). Experimental studies have suggested the "two-step two-polymerase model", in which one bypass polymerase initiates insertion while a second extends past the lesion (243) For example, members of the Y family DNA bypass polymerases (*POLH*, *POLI*, *POLK*) bypasses the lesion while pol zeta (*REV3L*) allows the cell to continue replication past the lesion (233, 243, 244). In a yeast cell line study, AP sites were bypassed by *POLH* with assistance from *REV3* for DNA extension (245). Another study provided evidence for a similar

process between *POLI* and *REV3L* (233). In an *in vitro* study conducted by Seki et al., although *POLQ* was unable to bypass a cyclobutane pyrimidine dimer or a (6-4) photoproduct alone, when combined with *POLI* it could successfully insert a base opposite a UV-induced (6-4) photoproduct and complete bypass (246). The experimental evidence reveals a comprehensive system of functionally redundant genes that are specialized to bypass several types of DNA lesions.

1.6.4.1 DNA bypass polymerases and cancer

As opposed to the extensive literature on DNA repair genes and mechanisms and cancer, less is known about bypass polymerases and their potential role in cancer. To date, there have been at least 15 different DNA polymerases identified in humans, which are specialized for replication, repair or the tolerance of DNA damage. The focus of the second aim of this study will be on DNA bypass polymerases, *POLI, POLH, REV1, POLL, POLM* and *REV3L*. Given their intrinsic nature of reduced fidelity and mutagenic potential in the repair of certain DNA lesions, several DNA bypass polymerases are suspected to be involved in cancer risk. It has been proposed that point mutations may arise from the error-generating activities of DNA bypass polymerases which may lead to carcinogenesis. However, a second perspective considers efficient bypass polymerases as maintaining genomic integrity. That is, DNA bypass polymerases may defend against chromosome instability in cells. At least one DNA bypass polymerase, *REV3L* (pol zeta), has been identified as a suppressor of spontaneous tumorigenesis (247).

1.6.4.1.1 *POLH*

POLH is a member of the Y Family that encodes the protein pol eta. The identification of mutations in *POLH* in xeroderma pigmentosum (XP) cells marked one of the first links between bypass polymerases and cancer (248, 249). Other studies have also confirmed that the loss of functional *POLH* increases sensitivity to UV radiation and also increases the risk of xeroderma pigmentosum, a rare type of skin cancer (225, 250). McGregor and colleagues reported that the frequency of UV-radiation induced mutations in the XPV cells was significantly higher than those in normal cells (251). Using a knockdown approach, Albertella et al. showed that inhibited expression of *POLH* was associated with a 3.6 fold increased mutation frequency when compared to control cells (250). Glick et al. failed to find *POLH* mutations in XP patients (252). However, as a result of functional redundancy in TLS, XPV patients that are unable to bypass across CPD due to a mutated *POLH* gene may be able to bypass the lesion through an alternate but more error-prone mechanism using *POLI* or *POLK* for insertion and *REV3* for extension around the CPD (243).

Mutations in *POLH* may also cause arrest of DNA replication at sites of DNA damage (225, 233, 238). Cleaver et al. demonstrated that XP variants cells lacking *POLH* exhibited stalling at the S phase checkpoint following UV damage (225). *POLH* physically interacts with *PCNA*-binding motifs at oxidative DNA damage sites (230).Two independent studies found somatic *POLH* mutation (G153D) in 2-9% of breast tumors (253)

1.6.4.1.2 *POLI*

POLI is another member of the Y family DNA bypass polymerases that encodes the protein pol iota. It may be associated with increased spontaneous mutagenesis during DNA replication. In an *in vitro* study of breast cancer cells, Yang and colleagues found that *POLI* expression was elevated and correlated with a significant decrease in DNA synthesis fidelity (254).

1.6.4.1.3 *REV1*

While *REV1* has restricted DNA polymerase activity, its main function is to serve as a scaffolding protein that recruits and coordinates other DNA bypass polymerases (*POLI*, *POLH*, *POLK*, *REV3L*) to the site of the lesion (255-257). In addition, *REV1* is able to insert deoxycytidine monophosphate (dCMPs) opposite abasic sites (255).

REV1 has been implicated in cancer in several experimental and epidemiologic studies. Lawrence et al. showed that *REV1* contributed to 98% of all base pair substitution errors and 90% of frameshift mutations induced by UV damage in yeast cells (258). A few studies have also linked mutations in *REV1* to lung and cervical cancer (259). *REV1* mutants show decreased spontaneous and induced mutagenesis by DNA-damaging agents. In an *in vitro* study, Clark et al. demonstrated that reduced levels of *REV1* were associated with a 75% reduction in UV-induced mutations (260).

1.6.4.1.4 *POLQ*

POLQ is a member of the A Family that encodes the protein polymerase theta. The high error rate for *POLQ* is closely related to Family Y polymerases (238). Most recently, *POLQ* has

been implicated in breast cancer. In a study conducted by Lemee 2010, levels of *POLQ* were upregulated in breast cancer cells (261). Another study also found elevated levels of *POLQ* expression compared to normal cells (222). Higgins et al. linked this overexpression of *POLQ* to poor prognosis in early breast cancer patients (262). Recently, *POLQ* was implicated as being involved in BER. *POLQ*-deficient mutants exhibit hypersensitivity to oxidative base damage induced by H_2O_2 (263).

1.6.4.1.5 *REV3L*

*REV3*L or pol zeta is a member of the B family. The ability to bypass DNA lesions was first discovered in yeast when *REV3* was shown to bypass UV-induced thymine-thymine cyclobutane pyrimidine dimers (TT-CPD)(264). Deletion mutation or loss of *REV3* may enhance spontaneous tumorigenesis (247, 265). In a mouse model, Wittschieben and colleagues showed that *REV3*L-deficient cells had enhanced tumorigenesis in mammary cells (247). In another lab-based study, Stone et al. compared wildtype and mutated *REV3* and found that yeast strains with the variant allele were more prone to mutagenic bypass (266). These results corroborate the role of *REV3L* as an inhibitor of spontaneous tumorigenesis.

1.6.4.1.6 *POLL*

POLL is a member of the X Family that encodes the protein polymerase lambda. *POLL* is thought to have dual functions in TLS and BER (267-269). *POLL* shares homology with *POLB* (270, 271) which may explain its role as a backup polymerase for *POLB*. Auofouchi 2000 showed that mRNA expression of *POLL* is downregulated when treated with DNA damaging agents such as UV light or H_2O_2. A novel nonsynonymous SNP (Arg438Trp) was shown to have

reduced base substitution fidelity in *in vitro* activity assays and increased mutation frequency in mammalian cells (272).

1.6.4.2 Critique and summary of bypass polymerase literature

While other DNA repair pathway genes have been studied extensively in breast cancer, the focus on DNA bypass polymerases is relatively recent. The discovery of germline mutations in *POLH* in patients with Xeroderma pigmentosum (XP), a rare form of skin cancer, was the first evidence that bypass polymerases may be involved in human cancer (249). However, the literature on bypass polymerases and breast cancer is sparse. We identified four experimental studies (166, 254, 261, 273) and two epidemiologic reports from the NHS (Nurses' Health Study) (202, 213). In an *in vitro* study of breast cancer cells, Yang et al. reported elevated *POLI* expression (254). Wang et al. found *POLB* overexpression in several cancer cell lines and tumors (166). Finally, *POLQ* overexpression in tumors was associated with poor prognosis of breast cancer (261, 273).

1.6 Conclusions

The overall BER DNA repair literature does not provide conclusive evidence for *single* common genetic polymorphisms (SNPs) as contributing to breast cancer risk. However, we propose there are several potential explanations for the observed lack of significant main SNP effects in BER. First, many genetic association studies may have been underpowered to detect modest effects in common low-penetrant SNPs, yielding false positive results. Second, several studies showed increasing risk with increasing number of SNPs or combined SNP effects, which may concur with the polygenic model. Third, several studies suggested subgroup effects by

race/ethnicity. Finally, other redundant DNA damage response mechanisms may be involved in maintaining genomic stability.

Researchers have identified at least 15 different DNA polymerases in humans which are essential for DNA replication, DNA repair and the tolerance of DNA damage. DNA bypass polymerases are key players in translesion synthesis (TLS) that serve as a backup if other DNA repair mechanisms fail. While bypass polymerases do not directly repair the damage, they tolerate or bypass the damage and prevent replication fork stalling, sparing the cell from going into apoptosis or DNA damage induced mutagenesis. Both *in vitro* and *in vivo* studies have shown that DNA bypass polymerases can efficiently bypass lesions that were not properly repaired by the classical DNA repair pathways. However, these bypass polymerases may also induce mutations due to their low fidelity. To further clarify their roles, we propose investigating the role of these bypass polymerases in breast cancer.

Although we have identified many of the genetic and environmental risk factors of breast cancer, there are still other (genetic) factors yet to be identified to account for the missing heritability of the disease. In this proposed study, we seek to identify SNPs in DNA damage response pathways that may be associated with breast cancer and breast cancer subtype. We propose a candidate pathway approach to evaluating the SNPs effects of bypass polymerases in breast cancer. These bypass polymerase genes have yet to be fully explored in epidemiological studies of breast cancer. Only two recent reports from the NHS II have explored the association between DNA bypass polymerases SNPs within breast cancer (202, 213). Therefore, additional studies exploring these associations are needed. This current study proposes using data from the Carolina Breast Cancer Study, a large racially diverse study population of women with breast cancer, to further elucidate the role of bypass polymerases genes and base excision repair genes

in breast cancer by race and subtype. We will also conduct pathway-based analyses to assess combined SNP effects.

Table 1. Functions of BER genes

DNA glycolyases	Type of base damage	Function	References
Monofunctional glycolyases			
UNG	Deamination	removes uracil from DNA	Broderick 2006, Moon 1998, Nilsen 2003
TDG	Deamination	removes thymine moieties from G/T mismatches, C/T and T/T mismatches, removes uracil and 5-bromouracil from mismatches with G	Visnas 2008, Hardeland 2001
SMUG1	Deamination	removes uracil from DNA	Marian 2011, Chanson 2009
MBD4	Deamination	U or T opposite G at CpG sequence	Yamada 2002, Song 2009, Hao 2004, Miao 2008
MPG	Alkylation	alkylated bases, 3-methyadenine (3-meA), methylguanine, etheno A and guanine, 8-oxoG	Kim 2002, Kim 2003
MYH	Oxidation	removes As that are mispaired with G, C, or oxo-G	Dallosso 2008, Out 2012, Wasielewski 2010, Rennert 2012, Beiner 2009, Zhu 2011
Bifunctional Glycolyases			
OGG1	Oxidation	8-oxoG opposite C	Tani 1998. Roberts 2011, Sangrarang 2009, Sterpone 2010, Rossner 2006, Goode 2002
NEIL1	Oxidation	removes oxidized pymidines, 8-oxoG	Das 2006, Dou 2008, Maiti 2008
NEIL2	Oxidation	removes oxidized pymidines, oxidized cytosine	Das 2006, Conlon 2005, Dey 2012, Zhai 2008, Kinslow 2008, Haiman 2008
Other BER genes	**Class/Type**		
APE1 (APEX1)	AP endonuclease	Recognizes and cleaves the phosphodiester bond 5' attached to the AP site	Agachan 2009, Kuasne 2011, Popanda 2004, Canbay 2010, Zawahit 2009, Smith 2008, Sangrarang 2008, Zhang 2006, Poletto 2012
PARP1 (ADRPT1)		Modifies nuclear protein by poly-ADP-ribosylation	Allinson 2003, Dantzer 2002, Bieche 1996, Fong 2009
PARP3 (ADRPT2)		Modifies nuclear protein by poly-ADP-ribosylation	Matsutani 2002
POLB	DNA polymerase	Gap filling enzyme in short-patch BER	Lang 2007, Yamtich 2012, Wang 1995, Makridakis 2012, Starcevic 2004, Dalal 2005, Donigan 2012, Kazma 2012, Nemec 2012
LIG3	Ligase	Catalyzes the nick-sealing step in short-patch BER along with cofactor XRCCI	Puebla-Osorio 2006
XRCC1	Ligase	central scaffolding protein binding LIG3, DNA polymerase B, and PARP	Chacko 2005, Smith 2008, Sangrarang 2008, Silva 2007, Sterpone 2010, Mitra 2008, Hussein 2012 Ali 2008
PCNA	scaffolding protein	senses DNA strand breaks and initiates DNA damage signaling (Scheiber), posttranslational modification by ubquitin	Malkas 2006, Ma 2000
RFC1	large subunit of replication factor C	binds to the 3' end of primers and promotes synthesis of both strands	Overmeer 2010, Fotedar 1996
FEN1	endonuclease	removes 5' flap in long patch BER	Singh 2008

Table 2. DNA Glycosylases

List of BER Glycosylases and associated substrate(s)		
Glycosylase	*Damaged base type*	*Substrates, Base released*
UNG	Deamination	Uracil, U:G, U:A, 5-FU
TDG	Deamination	U:G,, Etheno C:G, T: G
SMUG1	Deamination	Uracil, U: A, U: G
MBD4	Deamination	Uracil or T
MPG	Alkylation	3-MeA, 7-MeA, 3-MeG, 7-MeG
MYH	Oxidation	A:G, A: 8-oxoG
OGG1	Oxidation	8-oxoG: C, faPyG
NTH1	Oxidation	Tg, Cg, 5ohC
NEIL1	Oxidation	8-oxoG
NEIL2	Oxidation	8-oxoG
NEIL3*	Oxidation	oxidized purines

`

Table 3. Associations between BER genes and breast cancer risk

Gene/SNP	Study	Year	Country	Study Population	Cases	Controls	OR (95% CI) Arg/Gln	OR (95% CI) Gln/Gln
XRCC1	Brewster	2006	United States	All	321	321	1.5 (0.9-2.0)	1.1 (0.7-1.8)
rs25487	Chacko	2005	India	Asian	123	123	2.0 (1.2-3.6)	2.7 (1.1-6.6)
	Costa	2007	Portugal	European	285	442	0.5 (0.4-0.8)	
	Deligezer	2004	Turkey	Asian	151	133	0.9 (0.5-1.5)	1.3 (0.6-2.6)
	Duell	2001	United States	Black	253	266	1.5 (1.1-2.3)	2.1 (0.6-7.3)
	Duell	2001	United States	Caucasian	386	381	1.1 (0.8-1.5)	0.8 (0.5-1.3)
	Dulfloth	2005	Brazil	Mixed	86	120	1.1 (0.7-1.8)	1.7 (0.7-4.2)
	Figueiredo	2004	Canada	Caucasian	402	402	0.9 (0.7-1.2)	0.9 (0.6-1.4)
	Forsti	2004	Finland	All	223	298	1.1 (0.8-1.6)	0.9 (0.5-1.7)
	Hussein	2012	Egypt	Caucasian	100	100	1.7 (0.9-3.1)	1.6 (0.6-4.1)
	Kim	2002	Korea	Asian	205	205	0.8 (0.5-1.2)	2.4 (1.2-4.7)
	Lozidou	2008	Greece	Caucasian	1,109	1177	0.6 (0.8-1.1)	0.9 (0.7-1.2)
	Metsola	2005	Finland	Caucasian	483	482	1.2 (0.9-1.7)	1.4 (0.8-2.3)
	Mitra	2008	India	Asian	155	235	0.9 (0.6-1.5)	2.9 (1.7-5.1)
	Moullan	2003	France	Caucasian	254	312	0.9 (0.6-1.3)	1.0 (0.6-1.6)
	Pachkowski	2006	United States	Caucasian	1,281	1,137	1.1 (0.9-1.3)	1.0 (0.8-1.3)
	Pachkowski	2006	United States	Black	786	681	1.1 (0.9-1.5)	1.8 (0.8-3.8)
	Patel	2005	United States	All	502	502	1.0 (0.7-1.4)	1.3 (0.8-2.0)
	Roberts	2011	United States	premenopausal	1,099	1,945	0.9 (0.6-1.2)	0.8 (0.6-1.2)
	Roberts	2011	United States	postmenopausal	1,099	1,945	1.2 (1.0-1.5)	1.3 (0.9-1.8)
	Sangrajrang	2008	Thailand	Asian	507	425	1.2 (0.9-1.6)	1.8 (0.9-3.3)
	Shen	2005	United States	All	1,067	1110	1.1 (0.9-1.3)	1.0 (0.7-1.3)
	Shu	2003	China	Asian	1,088	1182	1.0 (0.8-1.1)	1.2 (0.9-1.7)
	$Smith_a$	2003	United States	Caucasian	253	268	1.0 (0.7-1.5)	1.1 (0.6-2.0)
	$Smith_b$	2003	United States	Caucasian	162	302	0.7 (0.4-1.2)	1.1 (0.5-2.7)
	Smith	2008	United States	Caucasian	336	416	1.0 (0.7-1.4)	0.9 (0.6-1.5)
	Smith	2008	United States	Black	63	78	1.1 (0.4-2.9)	2.1 (0.09-52.2)
	Sterpone	2010	Italy	Caucasian	43	31	4.8 (1.6-14.8)	4.4 (1.1-17.1)
	Thyagarajan	2006	United States	Caucasian	460	324	1.3 (0.8-2.0)	0.9 (0.5-1.7)
	Zhai	2006	China	Asian	523	639	0.8 (0.6-1.1)	1.0 (0.6-1.7)
	Zhang	2006	Poland	Caucasian	1995	2296	1.1 (1.0-1.3)	1.1 (0.9-1.4)
	Zhang	2006	United States	Caucasian	3368	2880	1.1 (0.9-1.2)	0.9 (0.8-1.1)
							His/His + Arg/His**	
XRCC1	Chacko	2005	India	Asian	123	123	0.6 (0.3-1.0)	
rs25489	Lozidou	2008	Greece	Caucasian	1109	1177	4.7 (1.0-21.7)	
	Metsola	2005	Finland	Caucasian	483	482	1.2 (0.8-1.7)	
	Pachkowski	2006	United States	Caucasian	1281	1137	1.2 (0.9-1.6)	
	Pachkowski	2006	United States	Black	786	681	1.3 (0.8-2.0)	
	Sangrajrang	2008	Thailand	Asian	507	425	1.3 (0.9-1.9)	
	Smith	2008	United States	Caucasian	336	416	0.7 (0.4-1.2)	
	Smith	2008	United States	Black	63	78	0.7 (0.1-3.0)	
	Zhang	2006	United States	Caucasian	1898	1514	1.1 (0.8-1.4)	

**dominant model, not enough homozygous variants to do general/codominant model, referent genotype: Arg/Arg

	Study	Year	Country	Study Population	Cases	Controls	Arg/Trp + Trp/Trp**	
XRCC1	Brewster	2006	United States	All	321	321	1.2 (0.7-1.8)	
rs1799782	Chacko	2005	India	Asian	123	123	2.0 (1.1-3.5)	
	Deligezer	2004	Turkish	Asian	151	133	0.5 (0.2-1.2)	
	Duell	2001	United States	Black	161	166	0.7 (0.4-1.2)	
	Duell	2001	United States	Caucasian	251	234	0.7 (0.3-1.4)	
	Forsti	2004	Finland	All	223	298	1.3 (0.6-2.6)	
	Kim	2002	Korea	Asian	205	205	1.1 (0.7-1.6)	
	Lozidou	2008	Greece	Caucasian	1109	1177	1.0 (0.8-1.3)	
	Mitra	2008	India	Asian	155	235	0.4 (0.2-0.7)	
	Moullan	2003	France	Caucasian	254	312	1.0 (0.6-1.7)	
	Pachkows	2006	United States	Caucasian	1281	1137	0.9 (0.7-1.2)	
	Pachkows	2006	United States	Black	786	681	1.0 (0.7-1.3)	
	Patel	2005	United States	All	502	502	0.6 (0.4-0.1.0)	
	Roberts	2011	United States	premenopausal	1099	1945	0.9 (0.7-1.3)	
	Roberts	2011	United States	postmenopausal	1099	1945	1.2 (1.0-1.5)	
	Sangrajran	2008	Thailand	Asian	507	425	1.1 (0.8-1.4)	
	Shen	2005	United States	All	1067	1110	0.9 (0.7-1.2)	
	Silva	2007	Portugal	Caucasian	241	457	2.0 (1.2-3.3)	
	$Smith_a$	2003	United States	Caucasian	253	268	1.6 (0.9-2.9)	
	$Smith_b$	2003	United States	Caucasian	162	302	2.0 (0.9-4.6)	
	Smith	2008	United States	Caucasian	336	416	1.2 (0.7-2.0)	
	Smith	2008	United States	Black	63	78	0.4 (0.1-1.7)	
	Sterpone	2010	Italy	Caucasian	43	31	1.8 (0.4-7.7)	
	Thyagaraja	2006	United States	Caucasian	460	324	1.2 (0.8-1.9)	
	Zhang	2006	United States	Caucasian	1898	1514	0.9 (0.8-1.2)	
	**dominant model							
							Asp/Glu	Glu/Glu
APE1	Sangrajran	2008	Thailand	Asian	507	425	0.6 (0.4-0.9)	0.9 (0.7-1.3)
rs1130409	Smith	2008	United States	Caucasian	336	416	0.7 (0.5-0.9)	0.8 (0.5-1.2)
	Smith	2008	United States	Black	63	78	1.0 (0.4-2.4)	0.9 (0.3-3.0)
	Zhang	2006	United States	Caucasian	1898	1514	1.0 (0.9-1.2)	1.0 (0.8-1.3)
							Cys/Cys	Ser/Cys
OGG1	Choi	2003	Korea	Asian	475	500	1.0 (0.8-1.4)	1.3 (0.9-1.9)
rs1052133	Roberts	2011	United States	premenopausal	1099	1945	1.0 (0.7-1.4)	1.2 (0.5-2.5)
	Roberts	2011	United States	postmenopausal	1099	1945	1.0 (0.8-1.2)	1.2 (0.8-1.9)
	Sangrajran	2008	Thailand	Asian	507	425	1.4 (1.0-2.1)	1.0 (0.7-1.3)
	Sterpone	2010	Italy	Caucasian	43	31	1.4 (0.5-3.6)	0.8 (0.1-6.7)
	Vogel	2003	Denmark	Caucasian	452	434	0.8 (0.6-1.1)	1.0 (0.5-1.9)
	Zhang	2006	United States	Caucasian	1898	1514	1.0 (0.8-1.2)	1.0 (0.7-1.4)
							Val/Ala	Ala/Ala
PARP1	Zhai	2006	China	Asian	523	639	0.9 (0.6-1.2)	0.9 (0.6-1.3)
rs1136410	Smith	2008	United States	Caucasian	336	416	0.7 (0.5-0.9)	0.7 (0.3-1.9)
	Smith	2008	United States	Black	63	78	4.6 (0.9-23.1)	
	Zhang	2006	United States	Caucasian	1898	1514	1.0 (0.9-1.2)	0.9 (0.6-1.4)

Table 4. Efficient and Mutagenic Bypass of DNA Lesions

Type of Lesion	Mode of formation	Efficient Bypass	Mutagenic Bypass
	Endogenous		
AP site	Hydrolytic depurination	POL alpha (Avkin 2002)	POLB (Blanca 2004, Efrati 1997)
		POLD (Avkin 2002)	POLK (Ohashi 2000)
		POLE (Avkin 2002)	POLH (Masutani 2000)
		POLH (Choi 2010)	POLL (Maga 2002, Blanca 2004)
		POLD/PNCA (Choi 2010)	POLM (Zhang 2002)
		POLB (Gieseking 2011)	REV1 + POLH (Choi 2010)
		POLQ (Seki 2004)	POLI + POLH (Choi 2010)
8-oxo-G	Guanine oxidation	POLK (Haracska 2002)	POLK (Irimia 2009, Zhang 2000)
		POLH (Maga 2007)	
		POLD (Avkin 2002)	
		POLM (Zhang 2002)	
		POLI (Zhang 2001, Vaisman 2001)	
		POLL (Maga 2007, vanLoon 2010)	
Thymine Glycol	pyrimidine oxidation	POLK (Fischhaber 2002)	POLQ (Seki 2004, Arana 2008)
		POLN (Takata 2006)	POLN (Takata 2006)
		POLB (Belousova 2010)	POLM (Kusumoto 2002)
		POLL (Belousova 2010)	
		POLK + REV3 (Yoon 2010)	
	Exogenous		
[6-4] photoproduct	UV light	POLB (Servant 2002)	POLB (Servant 2002)
		POLH + REV3 (Johnson 2001)	POLI + POLQ (Seki 2008)
		POLI + REV3 (Johnson 2000)	REV1 (Zhang 2002)
		X+REV3 (Yoon 2010)	REV3 (Guo 2001)
		POLQ (Seki 2008)	
cyclobutane pyrimidine dimer (CPD)		POLB (Servant 2002)	POLI+REV3 (Ziv 2009, Vaisman 2003)
		POLH (McCulloch 2008, Masutani 1999, Albertella 2005, Hendel 2008, Broyde 2010)	POLK+REV3 (Ziv 2009)
		POLM + REV3 (Zhang 2002)	
platinum DNA adducts		POLH + REV3 (Bassett 2002, Sharma 2012, Chaney 2005)	POLB (Bassett 2002)
Benzo[a]pyrene-guanine (BP-G)		POLK (Zhang 2002, Ohashi 2000, Avkin 2004, Suzuki 2002)	POLK + REV3 (Sharma 2012, Lin 2006)
		REV3 (Johnson 2000)	POLH + REV3 (Shachar 2009, Goodman 2002)
		POLM (Zhang 2002)	
cis-syn TT dimer		POLK + REV3 (Haracska 2002)	
		POLM (Zhang 2002)	
		POLH (Broyde 2010)	POLH (McCulloch 2008)

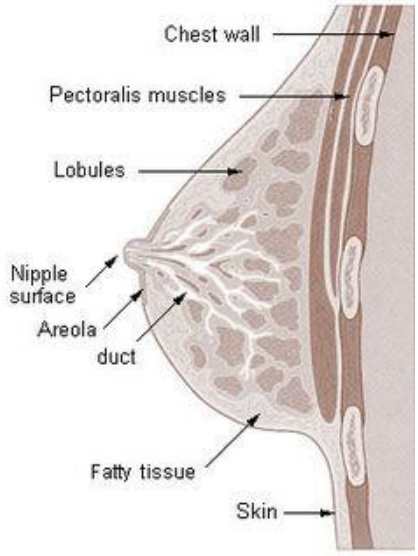


Figure 1. Breast Anatomy
Source: www.homeopathynow.com

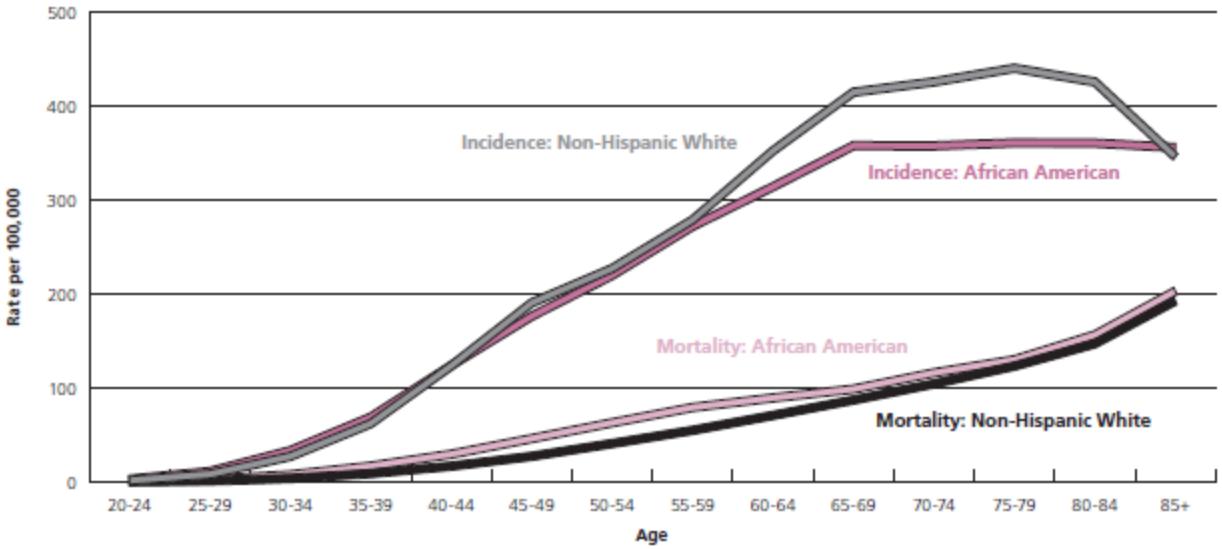


Figure 2. Breast cancer incidence and mortality by race and age

Sources: Incidence: North American Central Cancer Registeries, 2009. Mortality: National Center for Health Statistics.

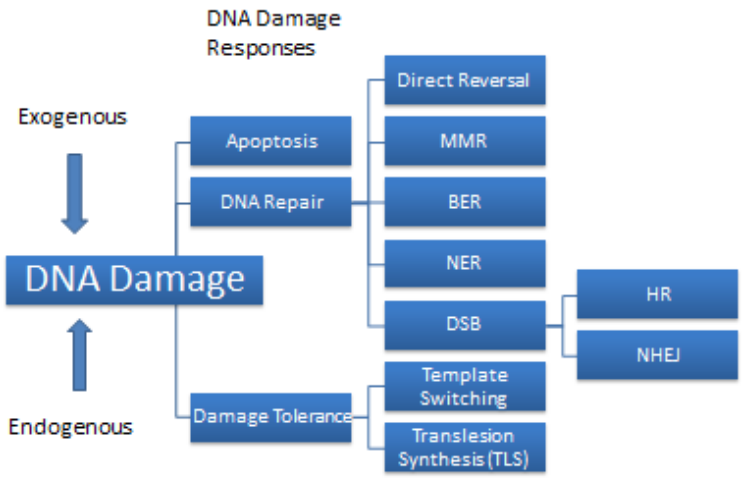


Figure 3. DNA Damage Responses

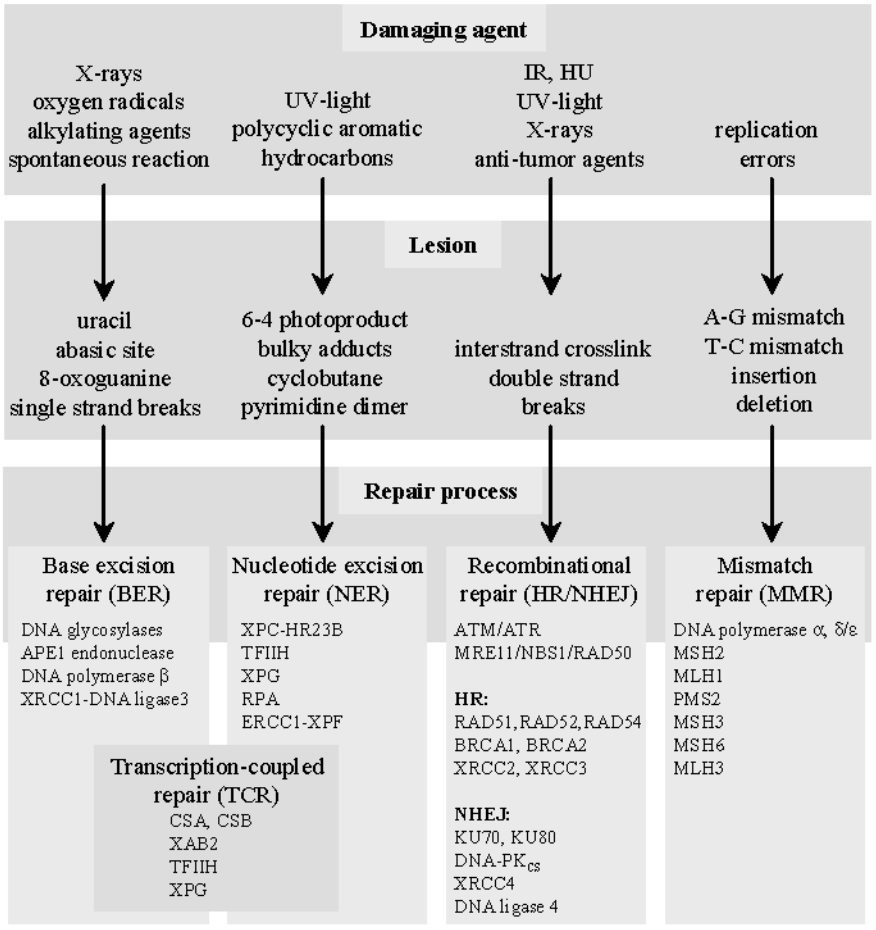


Figure 4. Sources of DNA Damage and associated lesion and repair pathway genes
(Adapted from Wood 2005)

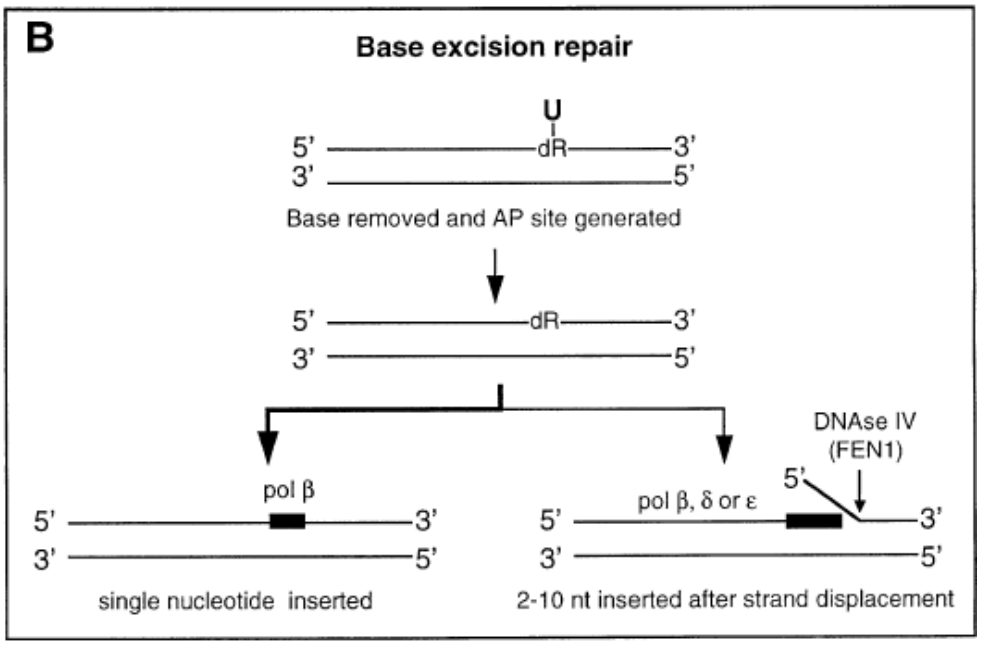


Figure 5. Short-patch vs. long-patch BER

CHAPTER 2. METHODS

2.1 Specific Aims

The American Cancer Society estimates 226,870 new cases of invasive breast cancer and 64,640 new cases of carcinoma of the breast *in situ,* which will represent 29% of all female cancer cases in the United States in 2012 (3). Previous studies have identified both non-genetic and genetic risk factors for breast cancer. Among the most well-known genetic factors are mutations in *BRCA1,* a DNA repair susceptibility marker. It has been hypothesized that deficient DNA repair due to mutations in *BRCA1* may contribute to increased breast cancer risk. However, *BRCA1* mutations are rare and together with *BRCA2* only account for 5-10% of all breast cancer and 15-20% of familial cancers, leaving a large proportion of breast cancer without a known genetic component. Consequently, other genes including DNA repair genes in multiple DNA repair pathways have been investigated for their association with breast cancer incidence. DNA repair is one of several DNA damage response mechanisms that have evolved to respond to ubiquitous DNA damage and base excision repair, is one such repair pathway. Other non-repair DNA tolerance mechanisms such as TLS (translesion synthesis) may also play a role in breast cancer. The impact of genetic variation in BER and TLS pathways will be evaluated using genotyped data from the Carolina Breast Cancer Study (CBCS), a large population-based case-control study. To assess the individual and combined effects of DNA repair and DNA tolerance, adjusted unconditional logical regression models will be used to estimate odds ratios and 95% confidence intervals.

The specific aims of this study are as follows:

Specific Aim 1: To estimate the association between breast cancer risk and genetic variation in base excision repair genes (BER).

A) To assess the race-specific effects of SNPs in BER genes on breast cancer risk, odds ratios (ORs) and 95% confidence intervals (CI) will be estimated using unconditional logistic regression, adjusting for ancestry informative markers (AIMs) and offset term.

B) To assess the subtype-specific effects of SNPs in BER genes on breast cancer risk, odds ratios (ORs) and 95% confidence intervals (CI) will be estimated for SNPs using unconditional logistic regression, comparing cases of each subtype (combined luminal, HER2+/ER-, and basal-like) to all controls.

C) To assess the combined pathway effects of SNPs within the base excision repair pathway on breast cancer risk (SNP-set Kernel Association Test (SKAT) will be used to estimate global p values for 2 SNPs sets (White and African American).

Specific Aim 2: To estimate the association between breast cancer risk and genetic variation in DNA bypass polymerase genes.

A) To assess the race-specific effects of SNPs in bypass polymerase genes on breast cancer risk, odds ratios (ORs) and 95% confidence intervals (CI) will be estimated using unconditional logistic regression, adjusting for ancestry informative markers (AIMs) and offset term.

B) To assess the subtype-specific effects of SNPs in XX bypass polymerase genes on breast cancer risk, odds ratios (ORs) and 95% confidence intervals (CI) will be estimated for SNPs using unconditional logistic regression, comparing cases of each subtype (combined luminal, HER2+/ER-, and basal-like) to all controls.

C) To assess the combined pathway effects of SNPs within the DNA bypass polymerase genes on breast cancer risk (SNP-set Kernel Association Test (SKAT) will be used to estimate global p values for 2 SNPs sets (White and African-American).

2.2 Study population: Carolina Breast Cancer Study (CBCS)

To accomplish these specific research aims, we will utilize genotype data from extant DNA extracted from blood samples from Phase I (1993-1996) and Phase II (1996-2001) of the Carolina Breast Cancer Study (CBCS). Study design and methods have been described extensively in (274, 275). CBCS is a large population-based case-control study that incorporates both randomized recruitment to oversample understudied populations such as younger and African-American women as well as rapid case ascertainment system which allows access to state reported data in a time efficient manner In addition, as a part of the study, biologic samples were collected which allowed for the DNA extraction and genotyping of various putative breast cancer genes, including DNA repair and bypass polymerase genes. CBCS also collected tumor tissue samples from participants which allowed for tumor subtyping via immunohistochemistry (IHC) as a surrogate for gene expression.

2.2.1 Case ascertainment

Case eligibility was determined using the following criteria: female, between the ages 20 and 74 years at the time of diagnosis, living within the 24 county study area in North Carolina, primary diagnosis of an invasive breast cancer between May 1, 1993 and September 30, 1995 (Phase I enrollment) or primary diagnosis of an invasive or *in situ* breast cancer between May 1, 1996 and September 30, 2001 (Phase II enrollment).

Eligible cases were identified from the Rapid Case Ascertainment program within the North Carolina Central Cancer Registry (NCCCR). By law, all breast cancer cases in North Carolina (invasive and *in situ*) are reportable to the North Carolina Center Cancer Registry

(NCCCR). Hospitals are required to send timely reports to the registry for all newly diagnosed cases, while physicians are required to report cancer cases that are not diagnosed in the hospital (276). With some cases of breast cancer that are rapidly fatal, timeliness of reporting can be critical. To ensure that cases were reported in a timely manner, CBCS collaborated with CCR to develop and implement a rapid case ascertainment system (RCA) (277). The CCR closely coordinated with hospital registries and were given an incentive to forward pathology reports to CCR as soon as they were received. Therefore, CBCS received expedited reports from CCR usually within a month of the diagnosis. Cases were invited to participate in the study based on the county of residence during their time of diagnosis, which included 24 central and eastern North Carolina (Figure 6). In addition, participants were required to live in the same county as they did at the time of their diagnosis.

2.2.2 Control ascertainment

Controls for the study were also female residents of North Carolina residing in one of the 24 study counties. Controls ages 20-64 at study entry were selected from the North Carolina Department of Motor Vehicle (DMV), while controls ages 65-74 were selected from the U.S. Health Care Financing Administration (Medicare) records. Controls represented the pool of women ages 20-74 living in the 24-county study region without a previous diagnosis of invasive or *in situ* breast cancer at the time of selection into the study. Controls were also matched to cases based on 5-year age categories and self-reported race (African-American and White).

2.2.3 Randomized recruitment

As an alternative to frequency matching, "randomized recruitment" or probability matching individually randomizes subjects to be recruited or not based on available screening

variables and disease status (278). In CBCS, these screening variables were the participant's race and age abstracted from pathology reports for cases and DMV and Medicare records for controls. This information was used to ensure that the sampling probabilities were approximately equal across race and age categories. These sampling probabilities were different for invasive cases in Phases 1 and 2. Table 5 shows the sampling probabilities for invasive cancers in both phases of the study, stratified by age and race. In phase II of the study, 100% of the *in situ* cases were sampled. To increase power, African-American cases and all cases younger than 50 years old were oversampled. Controls were probabilistic matched to cases by race and five year age group. To account for this "biased" sampling design, race and age was adjusted for in all logistic regression models using an offset term. The offset term is defined as the natural log of the ratio of the sampling probability for a case in the specific age-race strata to the sampling probability for a control in the same age-race strata (i.e. a non-black case aged 30-34 will have same offset term as non-black control aged 30-34, despite different sampling probabilities). Therefore, each CBCS participant will have their own offset term based on their race and age category.

2.2.4 Subject recruitment and enrollment

After receiving identifying information from NCCCR about a potential study participant, the participant's treating physician was sent a letter requesting permission to contact their patient. If physician permission was obtained, cases were sent a study brochure and a letter inviting them to participate. Physician consent was not obtained for 7% of eligible cases. If physician consent was obtained, a CBCS recruiter contacted the potential study case via telephone to assess interest and study eligibility.

Contact rates was defined as the percentage of women who were identified as potential study participants with whom contact was achieved (279). While contact information was readily

available for cases, contact information (i.e. telephone numbers) were not provided from DMV or HCFA records and a variety of strategies were employed to contact eligible controls (280). Contact rates were 98% for cases (3,292 out of 3,360) and 83% for controls (3,706 out of 4,465).

Cooperation rates were defined as the number of completed interviews divided by the number of women who were successfully contacted and eligible. Cooperation rates differed by case status (79% for cases and 71% for controls). In addition, the age/race specific cooperation rate ranged from 72% for older African-American women to 84% for young white women. If the study recruitment specialist confirmed the woman met all eligibility criteria and verbally agreed to participate in the study, an at-home interview with a trained study nurse was scheduled.

The overall response rate was defined as the number of completed interviews divided by the number of potentially eligible women selected for the study. Overall response rates for both phases of the study were 77% for cases and 57% for controls. Total enrollment included 1,803 invasive cases and 1,564 matched controls, and 508 *in situ* cases and 458 matched controls. Among cases, older African-Americans had the lowest overall response rate (70.8%) while younger Whites had the highest overall response rate (82.7%). Among controls, younger African-Americans had the lowest overall response rates (47.8%) while older Whites had the highest (77.9%). African-American *in situ* cases and controls were also were less likely to be selected into the study.

2.2.5 Baseline study interview

Prior to beginning the interview, a written signed informed consent was obtained. The subjects were required to initial a special section describing potential genetic research on their samples. The consent also assured participants that their blood samples would be used only for research purposes and safeguards were in place to ensure their confidentiality. Any questions or

concerns were addressed by the study nurse and participants were assured that their participation was voluntary. In addition, participants were given medical record release forms and HIPAA forms to sign allowing CBCS to obtain pathology reports and formalin-fixed paraffin-embedded (FFPE) tumor blocks. The tumor tissue blocks were used to both confirmed diagnosis by a pathologist as well as to conduct IHC subtyping. FFPE tumor blocks were obtained and successfully sectioned for 80% of cases and immunochemistry was completed for 62% of cases (281).

The baseline interview consisted of a nurse-administered questionnaire consisting of known and suspected breast cancer risk factors such as family history, personal medical history, occupational history, and exposure to known reproductive and lifestyle factors. In addition, participants completed a self-administered quality of life survey, and height, weight, waist circumstance, and hip circumstance were measured by the study nurse.

At the end of the interview, the nurse interviewer collected a 30 mL blood sample. Whites were more likely to provide blood samples than African-Americans. There were no significant differences in other risk factors for those who provide a biological sample and those who did not (281, 282). Women who refused the blood sample were given the option of providing a buccal cell sample using mouthwash or having their blood drawn at their physician's office to be sent into the study. If the biologic sample was collected at the interview site, the study nurse transported the sample back to the laboratory at UNC to be processed. DNA was extracted from peripheral blood lymphocytes using an automated ABI-DNA extractor in the UNC Tissue Procurement Facility and stored at $-80°C$. Blood samples were collected and DNA successfully extracted from peripheral blood lymphocytes on 89% of cases and 90% of controls.

2.3 Exposure assessment

2.3.1 CBCS SNP selection

Single nucleotide polymorphisms or SNPs are the most common type of genetic variation. It is estimated that there are 100,000-300-000 nonsynonymous coding SNPs in humans, representing 1% of all SNPs. The other 99% of SNPs include intronic (63%), untranslated regions (11%), synonymous (1%), locus regions (24%), splice site (<1%) and uncoding variants (<1%). A SNP occurs when a single nucleotide at a particular DNA location differs between members of a population and occurs at a frequency greater than 1% in the general population (283). In this study, a candidate gene approach was utilized to select SNPs in the BER and TLS pathways. The candidate gene approach focuses on associations between genetic variation within pre-specified genes of interest (i.e. DNA repair pathway genes) and phenotype (i.e. breast cancer). This is in contrast to genome-wide association studies (GWAS) which scan the entire genome for common genetic variation. Candidate genes are most often selected for study based on *a priori* knowledge of the gene's biological functional impact on the trait or disease in question. The candidate gene approach is hypothesis-based, relying on prior functional SNP data from laboratory studies or computer simulations (284, 285). Identifying potential functional SNPs may help to define a biological mechanism through which genotype is causally associated with breast cancer (286). A functional SNP is defined as a polymorphism in a codon that leads to an amino acid change that alters gene product and function and case-control status (287, 288).

Two publicly available SNP databases, the Single Nucleotide Polymorphism Database (dbSNP) and SNP500Cancer, were used to select 1,536 SNPs for CBCS. Preference was given to non-synonymous and promoter SNPs in genes that have been implicated in one or more cancers

(289). SNPs in the CBCS were selected based on the following criteria: gene previously identified in the base excision DNA repair or bypass polymerase pathway; experimental evidence (*In vitro/in vivo/in silico* studies) demonstrating functional effect; non-synonymous missense coding variants, upstream regulatory regions, splice variants, or 5'UTR variants, and at least 5% minor allele frequency in African-American or White populations.

2.3.2 Genotyping analysis

A total of 1,536 SNPs which passed all four Illumina reviews were selected in each pathway to be genotyped. There were 284 SNPs chosen in DNA repair pathways, including 59 SNPs in BER and 30 SNPs in bypass polymerase genes (Table 6, Table 7). SNPs were genotyped from biological samples collected at the time of baseline interview. High-throughput genotyping of selected SNPs was conducted at the Mammalian Genotyping Core Facility at the University of North Carolina at Chapel Hill using the Illumina high-multiplex GoldenGate Genotyping Assay with Sentrix Array matrix. This process has been documented in detail previously by (290). Briefly, the GoldenGate Assay queries genomic DNA with three oligonucleotide probes for each locus and creates DNA fragments that can be amplified by standard PCR methods using universal primers (290).The oligo mix contains two allele-specific and one locus-specific probe. The 3′ ends of the two alternative allele specific probes are complementary to two universal primers, U1 and U2, with the 5′ end complementary to the 3′ end of the locus. Each probe sequence terminates at the SNP that is to be assayed with an allele specific base. The third probe, the locus specific probe, is complementary to the genomic DNA. DNA polymerase is added to close the gap between the allele specific and the locus specific probes and the paired fragments are ligated together. The probe fragments are then separated from the genomic DNA and PCR results in a single strand hybridized to the BeadArray (290,

291). The genotype of an individual at a SNP is thus determined by comparing the relative hybridization intensities of the two probling sequences (Teo 2012). Large-scale genotyping depends on automated strategies (i.e. genotype calling) to translate the hybridization intensities for the two alleles at each SNP into a categorical genotype call.

2.3.3 Genotyping quality control

CBCS investigators utilized several quality control measures to measure and improve overall data accuracy. Upon arrival, all samples were labeled with a unique bar code with the BSP identifier, type of contained material, volume, concentration, date of creation, and locked in a secure storage facility. Case and control samples were randomly distributed on the panel. Systematic bias can arise if cases and controls are genotyped separately, since different error rates or genotyping success rates may lead to falsely different allele frequencies. In addition, blind duplicate samples, and negative and positive lab controls were used. A 4% (169 out of 3,857) random subsample of genotypes was repeated for each locus to test concordance with the original sample. Replicates that did not show greater than 99.5% concordance were excluded (291). Six subjects (3 cases and 3 controls) were excluded due to issues in non-blind DNA samples.

In addition, there were several potential sources of pre-genotyping error such as poor assay design and post-genotyping error such as low call rates, low signal intensity, indistinguishable genotype clusters. Departures from Hardy-Weinberg equilibrium may also indicate genotyping error. To assess these errors, individual call rates were examined, and there was careful inspection of assay intensity data and genotype clustering images. 103 samples with a call rate <95% were excluded.

Out of 1,536 SNPs, 163 (11%) SNPs were excluded due to genotyping error. Out of

2,311 cases and 2,022 controls enrolled in CBCS, 2,045 (88%) cases and 1,818 controls (90%) submitted a DNA biological sample. After all exclusions, a total of 1,972(85%) enrolled cases and 1,776(88%) enrolled controls were successfully genotyped (Figure 8).

2.4 Outcome Assessment

Diagnosis of invasive or *in situ* breast cancer from the pathology reports were confirmed via medical records. Centralized review of pathology was conducted for all cases using original or recut H&E sections (292). Details on the quality control procedures for tumor blocks are outlined in Dressler 1999 (293). A total of 1,845 cases had tumor tissue available.

2.4.1 Ascertainment of intrinsic subtype markers

Formalin-fixed paraffin-embedded (FFPE) tumor tissue samples were available for 80% of invasive cases and sent to the UNC Immunohistochemistry Core Laboratory to be sectioned and subtyped (22, 59, 72). Since gene expression analysis using DNA microarray technology was not possible on FFPE samples at the time, immunohistochemistry markers were used as a surrogate method to subtype the tumors (58).

A total of 1,424 (77% of available tumor blocks) were successfully subtyped and classified into one of five "intrinsic" subtype groups: luminal A (ER+ and/or PR+, *HER2*-), luminal B (ER+ and/or PR+, *HER2*+), *HER2*+/ER- (ER-, PR-, *HER2*+), and basal-like (ER-, PR-, *HER2*-, HER1+ and/or *CK 5/6*+), with those negative for all 5 markers considered 'unclassified' (59)(Figure 8).

Estrogen and progesterone status was abstracted from medical records for 80% of cases. For the remaining 20% of cases, ER and PR IHC assays were conducted using stored tumor tissue. Tumors with more than 5% of cells showing nuclei-specific staining were considered

receptor positive (53). A 10% random sample of ER+ and ER- tumors reported in medical reports were tested in the lab to evaluate concordance between the two data sources. There was a kappa statistic of 0.62 between the medical records and the lab data based on a previous CBCS report (281). *HER2* status was detected using the CB11 monoclonal antibody. A case was considered *HER2 positive* if at least 10% of observed cells showed signs of staining. This method had high concordance (81%) with PCR-based measures of *HER2* gene expression (293). *EGFR* (*HER1*) and *CK 5/6* assays were defined as being positively expressed if the tumor displayed any signs of staining (58). Table 8 shows the distribution of subtypes by race.

For the current study, we classified tumors as either luminal (ER+ and/or PR+; n=788), basal-like (ER-, PR-, HER2-, CK 5/6+ and/or EGFR+; n=199) or HER2+/ER- (n=94). We excluded 'unclassified' tumors from further analysis due to their uncertain status. The major distinction between the two luminal subtypes are their proliferation signatures, measured by the expression of *CCNB1*, *MKI67*, and *MYBL2* (49). HER2 expression only identifies about 30% of luminal B tumors. In the current study, we did not have information about these proliferation markers and therefore combined Luminal A and B tumors into a single 'luminal' category (48, 49) (Figure 11). Additionally, most other studies do not have subtype data available and only have estrogen receptor status data. Therefore, we conducted an additional exploratory analysis using estrogen receptor (ER) status to evaluate comparability to "intrinsic" subtype results. We found that ER positive effects were concordant with luminal subtype results; while ER negative (ER-) effects correlated with those of basal-like and HER2+/ER- subtypes (Table 15). There were no differences between CBCS cases with and without subtyping data in terms of age, menopausal status, or family history.

2.4.2 IHC for *in situ* cases

Phase 2 of the CBCS also included women diagnosed with ductal carcinoma *in situ* breast cancer (DCIS), lobular carcinoma *in situ* (LCIS), and mixed DCIS and LCIS (72). Tumor tissue was collected from 79% of *in situ* cases and sent to the UNC Immunohistochemistry Core Laboratory for subtyping (22). IHC subtyping procedures were slightly modified for *in situ* tumors due to availability of tumor samples. ER positive tumors were defined as having an Allred score >2 with nuclear staining. PR status was not determined independently due to the high correlation between ER and PR positivity (294). However a recent study suggested that IHC of PR could add prognostic value and identified cases with better outcomes (295). *HER2* positive tumors were defined as having more than 10% of cells stained greater than 3 using DAB chromogen or greater than 2 or 3 using SG chromogen. As for *CK 5/6* and *EGFR* in invasive cases, tissue with staining of 1+ or greater was defined as positive for expression (72). DCIS was the most common subgroup of *in situ* breast cancer and was defined in the CBCS by microinvasion of less than or equal to 2mm (296).

Of the 2,311 cases enrolled into the study, 1,220 (53%) cases had both complete subtype and genotyped data (Figure 8). The subtype distribution of those with genotyped data is very similar to the subtype distribution of all CBCS participants (22).

2.5 Covariate Assessment

2.5.1 Traditional Confounding

Confounders are "factors (exposures, interventions, treatments) that explain or produce all or part of the difference between the measure of association and the measure of effect that would be obtained with a counterfactual ideal"(288). Confounding has also been described as a mixing of two or more effects. The bias caused by traditional confounding in conventional

epidemiological studies typically does not apply to genetic epidemiological studies. Potential non-genetic confounders (i.e. reproductive factors) can be associated with the outcome (i.e. breast cancer) but they are unlikely to be related to the genotype (Figure 9), especially with the genes under investigation in this project. If they are an intermediate variable between genotype and breast cancer, adjusting for these covariates could induce bias.

2.5.2 Confounding by ancestry (population stratification)

While traditional confounders may not be relevant in genetic association studies, there is the possibility for confounding by race/ancestry or population stratification. Population stratification or may occur if one or more subpopulations have a higher prevalence of an allele and a higher risk of disease (283). According to Barnholtz-Sloan, two criteria must be fulfilled in order for population stratification to exist. First, the frequency of the marker gene of interest must vary significantly by race/ethnicity and second, the background disease prevalence must also vary significantly by race/ethnicity (297) (Figure 11). Therefore, population stratification refers to differences in allele frequencies between cases and controls due to systematic differences in ancestry rather than association of genes with disease (298).

Population stratification may be a possible source of bias among admixed groups such as African-Americans and Latinos (299). Several studies of African populations have indicated that levels and patterns of LD in these populations differ from those in non-African populations due to admixture with other African and non-African populations. LD block size tends to be shorter in individuals of African ancestry and longer in Caucasians due to genetic drift and recombination (298, 300). Barnholtz-Sloan also argues that "classifying individuals into classes that represent heterogeneous racial/ethnic groups may also misclassify a person's actual ancestral background and limit assessment of variation within racial/ethnic groups that is relevant for

understanding disease risk or outcome (297). Therefore estimation of individual ancestry should better capture the variation in ancestry within a subpopulation and account for residual confounding. These arguments validated the necessity of controlling for admixture in race-stratified analysis in the current study.

There have been several analytic approaches proposed to control for population stratification in genetic association studies. The two most common methods for estimating individual ancestry are using maximum likelihood estimates (MLE) or a structural association approach (301, 302). Structured association can use markers pre-selected to differ between ancestral populations (AIMs) or random genetic markers (297). STRUCTURE, a structured association program, uses Bayesian Markov Chain Monte Carlo method to estimate allele frequencies in subpopulations and individual ancestry proportions (302).

In addition, the genomic control method, proposed by Devlin and Roeder, calculates a variance inflation factor for a set of random, unlinked SNPs across the genome, and adjusts all SNP association tests by the inflation factor (303). Genomic control makes the assumption that the variance inflation is constant across all loci being tested. If this assumption is violated, overadjustment or underadjustment of variance may occur for different loci, which may result in reduced power to detect risk alleles. Marchini argues hat using too few markers for genomic control could lead to false positives, while using too many markers could lead to decreased power (304).

Finally, a principal components method has been touted as a method to assess population stratification. This method uses genotype data to estimate axes of variation that can be interpreted as describing continuous ancestral heterogeneity within a group of individuals (305). These axes of variation are defined as the top eigenvectors of a covariance matrix between

individuals in the study population that was formed using genotype information from random markers or AIMs. One of the main advantages of the principal components method is that it is more efficient in determining population structure using a large number of markers (i.e for GWAS). However the principal component methods can have a higher rate of type I error when the number of markers is low and may not be appropriate for a study such as CBCS (305).

In this proposed study, we will use the MLE methods proposed by Barnholtz-Sloan (297). Estimates of genetic ancestry will be derived from genotyped AIMs, which are unlinked markers found throughout the genome that show large allele frequency differences between ancestral populations (297).In CBCS, a set of 144 ancestry informative markers (AIMs) were selected to maximize the differences or δ in allele frequencies between African-Americans and White participants in the YRI and CEU HapMap data respectively (306) (Table 9). Proportions of ancestry for each ancestral population should sum to 1 since there are only two ancestral populations used for this study (301) Initially, a set of 200 AIM SNPs were selected from the panel of 5,400 AIMs identified at UC Davis Rowe Program representing four genetically diverse populations: two West African populations, European Americans, and African Americans. 158 (79%) passed the initial Illumina review and a subset of those SNPs (144 or 91%) were successfully genotyped. This set of SNPs provides nearly uniform coverage of the genome. Among African-American CBCS participants, the median proportion of European ancestry was 0.19 (average= 0.22), with most women in the 0 to 0.50 range. In Whites, most women had between 80% and 100% European ancestry, with a median proportion of 0.94 (average = 0.93) (307).

2.6 Statistical Analysis

2.6.1 Assessment of Hardy-Weinberg Equilibrium

The Hardy-Weinberg Equilibrium (HWE) states that "under certain assumptions, the genotype and allele frequencies in a large, randomly mating population remain stable over generations and that there is a fixed relationship between allele and genotype frequencies" (283).

$$\chi^2 = \sum_{i=1}^{3} \frac{(O_i - E_i)}{E_i}, \text{ with 1 degree of freedom (df)}$$

Deviations from HWE will be measured through a Pearson's chi-square test with the null hypothesis assuming that alleles are chosen randomly and that the observed genotype proportions match the expected genotype proportions (p^2, 2pq, q^2). However this goodness of fit test is sensitive to small sample size or rare allele frequencies and an exact test will be performed in these scenarios (308).

There are several reasons that a SNP may deviate from HWE among controls, including genotyping error, chance, failure of HWE assumptions (i.e. random mating), population stratification, and even a true genetic disease association. Examining deviations from HWE among cases is not performed since this may reflect a true mutation and cannot be distinguished from genotyping error. Barring chance and assuming the other conditions under HWE to be minimal, any SNP with a p-value less than 0.05 will be considered in violation of HWE due to genotyping errors and excluded from further analyses.

2.6.2 Genetic Model Specification

There are several genetic model choices available to test whether a specific SNP is associated to breast cancer (309). If there is no prior information about mode of inheritance of

the variant, an additive or general (co-dominant) genetic model can be used to estimate having one or two copies of the variant. The additive model makes an additional assumption that the relationship between the log ORs is linear. For the additive model, genotypes are coded as an ordinal variable ('0' for no risk alleles, '1' for a single copy of risk allele, and '2' for both copies of the minor allele). This model generates two ORs, one comparing homozygote variant to homozygous wildtype (referent) and one comparing the heterozygote to the homozygous wildtype (referent). The additive model uses a 2 df test, while the dominant model uses a 1 df test. The dominant model assumes that one copy of the variant allele increases risk. Since many of our selected SNPs had non-polymorphic or rare homozygous variants, we used a dominant model for all SNPs. We combined the homozygous variant and heterozygous genotypes and compared them to the homozygous wildtype genotype to obtain a single effect estimate and corresponding 95% confidence intervals.

2.6.3 Race-specific effects

We explored whether the effects of BER and TLS SNPs vary by race. Race was coded as a dichotomous variable, 0 for White and 1 for African-American, based on participant's self-reported race. Less than 2% of participants self-identified as another race and will be excluded from the analysis.

Unconditional binary logistic regression will be used to estimate odds ratios (ORs) and 95% confidence intervals (CIs) to capture race-specific effects of base excision genes (Specific Aim 1B) and DNA bypass polymerases (Specific Aim 2B) with breast cancer, adjusted for age, proportion African ancestry, and offset term using SAS version 9.3 (SAS Institute, Cary, NC) . The following binary model was used:

$$\text{Logit}\,[D=1|X=x] = \alpha + \beta_1 X_1 + \beta_2\,\text{age} + \beta_3\,\text{ancestry} + \text{offset}$$

where D represents the outcome of interest (invasive or *in situ* breast cancer) coded dichotomously as 0 for control and 1 for case, α represents the model intercept, β_1 is equal to the log OR for the effect of each additional copy of the variant alleles and x_1 is equal to the number of copies of the variant allele. The offset term is designed to adjust for selection bias induced by randomized recrutiment sampling method. Each CBCS participant will have a value for the variable CBCSOFF (offset term).

2.6.4 Subtype-specific effects

Specific Aims 1B and 2B will assess potential heterogeneity of SNP effects across strata of breast cancer subtype. "Intrinsic" subtypes have been classified into 5 different categories: luminal A, luminal B, HER2+/ER-, basal-like, and unclassifed. However, different markers continue to emerge in defining subtypes and there are no universally accepted classifications of breast cancer subtype across studies. The major distinction between the two luminal subtypes are their proliferation signatures, measured by the expression of *CCNB1*, *MKI67*, and *MYBL2* (49). *HER2* expression only identifies about 30% of luminal B tumors. In the current study, we did not have information about these proliferation markers and therefore will combine luminal A and B tumors into a single 'luminal' category (48, 49). Furthermore, Leong et al. describe methodological issues in using IHC to detect *HER2* expression (310). A recent study comparing concordance of PAM50 with IHC showed that ER positivity by IHC was strongly associated with luminal (A and B) subtypes (92%) (311). Therefore, we will conduct a case-control analysis estimating three ORs: luminal cases compared to controls, HER2+/ER- cases compared to controls, and basal-like cases compared to controls.

2.6.4 Correction for multiple testing

Conducting multiple tests in genetic association studies may increase the likelihood of obtaining false positives. A false positive occurs when a test statistic suggests that the null hypothesis should be rejected even though it is true. We considered two different methods that control for the type 1 error rate. The Bonferroni method controls the family-wise error rate or the probability of at least one false positive. This method is computationally simple, dividing the p-value cutoff (usually alpha= 0.05) by the number of tests conducted. While the Bonferroni correction method may be overly conservative for studies with thousands of multiple comparisons such as GWAS, it has been shown to be robust for up to a few hundred tests and easy to calculate (312). Furthermore, it is important to control the false negative rate. The false discovery rate or FDR has been touted a less conservative alternative to the Bonferroni method (313). In FDR, p-values of each SNP are ranked, and all SNPs except the largest are corrected by multiplying by the total number of SNPs being tests divided by the p-value's rank. Therefore, the FDR is the proportion of the rejected null hypotheses which were incorrectly rejected, or a type II error. FDR can be estimated using PROC MULTITEST in SAS (314). Ideally, it is important to balance both types of error. In this study, we will use FDR to adjust for multiple comparisons.

However, both the Bonferroni and FDR make the independence assumption which may be violated if SNPs are found to be correlated (in high LD). LD refers to the non-random association between two alleles at two loci on a chromosome in a natural breeding population (283).Two SNPs are in LD if they are inherited together more often than expected by chance (285). There are several methods that measure linkage disequilbirum between SNPs including D' and r^2 (209, 315). The r^2 statistic or correlation coefficient squared is a measure of how well the

identity of one allele at a polymorphic locus predicts the identity of the allele at another polymorphic locus. An r^2=1.0 indicates that the examined loci are in "perfect LD" (209). The measure r^2 is complementary to D'. r^2 is equal to D^2 divided by the product of the allele frequencies at the two loci. The absolute value of D' is determined by dividing D by its maximum possible value, given the allele frequencies at the two loci. If D'=1, then SNPs are in complete LD. Values of D'<1 indicate that the complete LD has been disrupted. D' values <1 can be biased in small sample, therefore only D' values close to one provide a useful information. R^2 has a more inituitive interpretation and will be used to evaluate potential LD in this study.

2.6.5 Combined within-pathway effects

When the number of susceptibility loci is small, the logistic regression model is an appropriate method for evaluating SNP-SNP interactions; however when there are multiple loci and interactions, the classical modeling approach may lack power due to high dimensionality of the data. There are several statistical methods available to reduce the dimensionality of the data and detect higher-order statistical interactions such as Monte Carlo methods, hierarchical modeling, machine learning, MDR, classification trees, and recursive partitioning (316-326) A common approach is to test the individual significance of each SNP, using the most significant p value as the p value for all the SNPs, then adjusting for multiple comparisons (327). However this test will have low power if the individual SNP are not in high LD with the causal variant. Omnibus tests for multiple SNPs or haplotypes allow for simultaneous analysis of all SNPs, but are based on a large number of degrees of freedom. To reduce the number of degrees of freedom, several approaches have used U-statstics, which summarizes the genomic similarity (genotype) to phenotype similarity (disease status) (328-331) . Kernel regression methods are closely related

to U-statistics in that they convert genomic information for a pair of individuals to a kernel score representing either similarity or dissimilarity, creating a positive semidefinite matrix when applied to all pairs of the individuals (332).

In this proposed study, we will use a logistic kernel machine test (LKMT) as proposed by Wu to evaluate the combined effects of SNP in two biologically driven pathways (BER and TLS) using the software package SKAT(SNP-set Kernel Association Test) package in R. This pathway-based method combines machine learning and kernel regression models. First, a set of "different but correlated SNPs are grouped based on prior biological knowledge" to create a SNP-set. The formation of SNP-sets harnesses the LD between SNPs to increase power (328). This prior biological knowledge could be based on several potential correlations between SNPs, including physical proximity to a gene, evolutionarily conserved regions, and SNPs within a haplotype block (333). For the purposes of this study, we will group our SNPs based on established DNA repair pathways (BER and TLS). This will allow us to assess the combined effects of a panel of predetermined SNPs that interact in the same pathway.

The second step evaluates the association between each SNP and breast cancer using logistic kernel-machine-based multi-locus test. This test combines the logistic kernel-machine testing approach of Liu (334) with the kernel framework suggested by Kwee (335). The LKMT uses a semi-definite kernel function to represent the influence of all SNPs in the SNP set. The choice of kernel changes the underlying basis for the nonparametric function defining the relationship between case-control status and the SNPs in the SNP-set. Choosing an appropriate kernel will increase power of the study. There are several choices for kernel type including linear, Gaussian, Identical-by-state (IBS), and weighted IBS (330, 333). In this study, we will use

a linear kernel since we are assuming a log linear model. The probability of being a case depends on the SNPs only through the function h(Z) therefore H0=h(z)=0. If the focus is hypothesis testing, the null hypothesis is h(Z)=0. Using a variance component score test, we will get an estimate for testing the global null hypothesis equals zero (333). A significant score test indicates that combined effects exist between SNPs. The estimate derived from this pathway-based test is a global p value representing the combined effect of individual SNPs in the BER and DNA bypass polymerases (TLS) pathways. Therefore, it may not be possible to estimate the interactions between pairs or sets of individual SNPs or distinguish which SNP is actually driving the association, if a significant association is found. Also, it is not possible to capture multi-SNP or epistatic effects among SNPs in separate SNP sets.

This method has a number of advantages over other multi-SNP methods (330, 331, 336). There is no need for a parametric model *a priori* which allows for estimation of joint and nonlinear effects. While Schaid's method makes the assumption that all variants have the same direction of effect, this method allows for flexibility in the functional relationship between the SNPs in a SNP set and the outcome. (330). Additionally, similar to hierarchical modeling, kernel-based machine learning logistic regression reduces the number of hypothesis being tested, which lowers the significance threshold and increases power. This is especially relevant for SNPs which have moderate or low individual effects. In summary, the advantages of this method are the reduced numbers of hypothesis being tested, improved power when SNP have modest effects, and model flexibility to account for non-linear effects. In addition, one of the most important features of this model is that it can simultaneously account for covariates, which is a limitation of most other similar methods (333).

2.7 Power calculations

QUANTO Version 1.2.4 was used for power calculations (337). The study had a fixed sample size of 3,748 genotyped cases and controls with a control-case ratio of 0.90 (1,972 cases and 1,776 controls). For Whites, the overall sample size was 2,346 (1,229 cases and 1,117 controls) with a control-case ratio of 0.91. For African-Americans, the overall sample size was 1,400 (742 cases and 658 controls) with a control-case ratio of 0.89. For luminal subtype calculations, the overall sample size was 2,571 (795 luminal cases vs. 1,776 controls) with a control-case ratio of 2.23. For basal-like subtype calculations, the overall sample size was 1,976 (200 basal-like vs. 1,776 controls) with a control-case ratio of 8.88. For HER2+/ER- subtype calculations, the overall sample size was 1,870 (94 HER2+/ER- cases vs. 1,776 controls) with a control-case ratio of 18.9.We anticipated that the estimates for HER2+/ER- and basal-like subtype tumors will be less precise than those for luminal subtype tumors due to sample size limitations. Based on the previous literature of genetic association studies of DNA repair and breast cancer, we estimated effects ranging from OR=1.0 to OR=2.0. Tests for statistical significance were two-sided with an alpha level of 0.05. Given our sample sizes, assuming 80% power, we can detect minimum ORs of 1.2-1.4 depending on the MAF in Whites, 1.3-1.6 in African-Americans, and 1.5-1.9 for luminal subtype (Figures 12-16).

2.8 Limitations

2.8.1 Exposure (genotype) misclassification

There is the potential for exposure misclassification due to genotyping errors in the laboratory. However, several measures were in place to minimize genotyping errors. In the overall study, blind duplicates of 169 samples were assayed to measure the reproducibility and

no SNPs were excluded. Assay intensity data and genotype cluster images were reviewed for all SNPs. Out of 2,039 cases and 1,818 controls, 103 subjects (64 cases and 39 controls) and 204 samples had low call rates (<95%) for SNPs and were therefore excluded.

In the current study, a total of 8 SNPs were excluded due to low signal intensity or indistinguishable genotype clusters. In addition, tests of Hardy-Weinberg equilibrium were conducted. Four SNPs in the BER pathway failed HWE and were excluded from subsequent analyses. Overall, there were 3,748 or 97% of enrolled participants (1,972 cases and 1,776 controls) with successfully genotyped data. A comparison of participants with and without genotyped data did not show significant differences (data not shown).

2.8.2 Outcome (phenotype) misclassification

CBCS had detailed subtype data on tumors from a majority of cases (62%) allowing a unique investigation of the genetics of specific breast cancer subtypes. However, cases with subtype data were more likely to be African American and to have a later stage at diagnosis, which may bias estimates for SNPs related to race or disease aggressiveness (22). However, there were no significant differences for age, menopausal status, or family history between CBCS cases with and without subtyping data.

In phase 2 of the CBCS, *in situ* cancers were enrolled in the study. There has been some debate to whether *in situ* cases should be included along with the invasive cases. Millikan argues that identifying risk factors in *in situ* tumors that occur during an earlier or intermediate stage of cancer progression may be informative for developing new preventive and treatment measures (74). Furthermore, studies have shown that *in situ* and invasive tumors share similar risk factor profiles and clinical features with effects in the same direction, albeit with varying magnitude of effects (70, 72, 338, 339).

Due to a lack of fresh frozen tissue samples, gene expression microarray analysis in CBCS phase 1 and 2 was not feasible and IHC was used as a proxy method in Phase 1 and 2 to subtype tumors. Several studies have evaluated the concordance between IHC and gene expression. While studies showed good correlation for ER and *HER2* IHC markers with gene expression, there has been some debate on the lack of sensitivity for staining CK 5/6 in identifying basal-like cancers using IHC methods (58, 296).

The definitions for "intrinsic" molecular subtypes are constantly evolving. We have yet to develop a standardized definition for the molecular subtypes. Therefore, comparability between study results may be compromised. In many studies, triple negative is used as a proxy for basal-like tumors. While approximately 70% of triple-negative breast cancers express basal markers, the remaining 30% are grouped together as unclassified (66). Cheang identified differences in prognostic values between basal-type and triple-negative cancers, with basal-like phenotype as a better prognostic predictor than triple-negative phenotype (340).Therefore, It is important to distinguish between triple-negative and basal-like subtypes (341). ER/PR status was successfully abstracted from medical records for 80% of cases. For the remaining 20% of cases, this data was obtained using IHC methods if tissue was available (281).

2.8.3 Covariate misclassification

CBCS participants self-reported their race during the baseline interview. Participants who self-reported race as other than White or African-American were excluded due to small sample size (2%). Barnholtz-Sloan 2005 reported that adjusting for individual European ancestry provided a better fit to the data compared with adjusting for self-reported race only (301).Therefore, all models in the study will be adjusted by AIMs to control for residual confounding by race (i.e. population stratification).

2.8.4 Selection bias

The parent study made an intentional effort to oversample African-Americans into the study. Younger African-American women have been traditionally understudied in breast cancer research therefore special efforts were made to include study counties with high proportions of African American women living in rural areas. There are also statistical advantages to randomized recruitment. Potential sampling bias from randomized recruitment was adjusted using an offset term in the analysis (278). Despite randomized design and intensive recruitment efforts, African-American women were less likely to be enrolled in the study compared to Whites. Among cases, older African-Americans had the lowest overall response rate (70.8%) while younger Whites had the highest overall response rate (82.7%). Among controls, younger African-Americans had the lowest overall response rates (47.8%) while older Whites had the highest (77.9%). African-American *in situ* cases and controls were also were less likely to be selected into the study.

The final data set included 1,809 white women (55%) compared to 1,505 African-American women (45%). This could potentially have implications for power in detecting effects in African-Americans. Comparing MAFs in CBCS controls stratified by race to MAFs in public databases (i.e. HapMap) would be one method to assess potential selection bias.

2.8.5 Missing data

Out of total of 4,333 enrolled participants, 2,045 (88%) of cases and 1,818 (90%) controls provided a DNA sample at interview leaving 272 cases and 204 controls either had insufficient DNA for genotyping or did not have a sample for genotyping. There were differences between by race and case status for those who were genotyped successfully and those who were not. African-Americans and cases were less likely to have genotyped data (281).

However, there were no differences between genotyped and non-genotyped participants for age, race, and family history. In addition, 47% of enrolled cases did not have IHC data available for subtyping, but there were no differences between those with and without subtype information (307).

There were a few novel bypass polymerase SNPs that were identified after CBCS completed its genotyping phase and as a result are not included in the analysis. Two recently identified bypass polymerases (*POLK* and *POLN*) were not genotyped in CBCS. Of significance, several SNPs in *POLK* (rs3213801, rs5744533, rs3756558) have been found to be significantly associated with pre-menopausal breast risk ($p<0.05$) (202) and interact with other Y family members (243, 255).

In addition, although we had measured data on environmental/lifestyle risk factors, no gene x environment interactions were tested. This was after consideration of low power to detect interaction effects in the current study.

2.9 Strengths of the study

The innovative study design of integrating population-based epidemiology with genetic and molecular data is one of the main strengths of the study (275). The molecular subtyping of tumors has revealed new insights into the heterogeneity of breast cancer (22, 59, 72, 307). In addition, the use of population-based controls representing women from the same geographic region strengthens the external validity of the study. The 24 study counties were selected to represent a larger African-American and rural population (281). In addition, CBCS took the initiative to collaborate with hospital registrars in the state to design the Rapid Case Ascertainment System, a system designed to minimize the delay in contacting cancer patients who may be otherwise been lost to study participation due to death or relocation.

Another benefit of randomized recruitment was the oversampling of younger and African-American women to allow better representation of these two understudied subgroups (277). Probability matching increases the relative sample size for younger African-Americans. There are also advantages over traditional matching techniques such as frequency matching including the simultaneous recruitment of cases and controls in a more time efficient manner. Another advantage is the ability to estimate effects associated with matching factors with better precision than under random sampling (278).

CBCS included a racially diverse study population that allowed for results to be stratified by race. Several CBCS studies have been able to report significant race-specific effects, which is important due to the differences in genetic architecture among those with African descent (342, 343). . As mentioned above, LD block size tends to be shorter in admixed populations such as African-Americans due to genetic drift and recombination.

Despite finding no significant main effects, a CBCS report found a *XRCC1* SNP to be associated with African-American race.(92). Another CBCS report found several SNPs in NER genes to be associated with an increased risk in African Americans (344). Many earlier studies lacked the power to obtain precise estimates for African-Americans. Similarly, this proposed study will have the power to report race-specific estimates.

This is one of the first studies to look at the effects of bypass polymerases on breast cancer risk. To date, only two other reports from the NHS II cohort have evaluated bypass polymerase SNPs in breast cancer (202, 213) .

2.10 Public health significance

The results of this proposed study may enhance our understanding of the complex biological processes involved in the development of breast carcinogenesis. The role of bypass polymerases in breast cancer has not been fully elucidated and this proposed study will be one of the first to further examine this association. Ultimately these results may inform future research on DNA bypass polymerases as potential targeted preventative strategies and therapies for breast cancer, especially for women with basal-like tumors that are resistant to traditional chemotherapy (345, 346). Furthermore, the CBCS dataset allows for the evaluation of both race-specific and subtype-specific associations. Since younger African-Americans carry a disproportionate burden of basal-like disease, the results derived from this study will be directly generalizable to this high-risk subgroup. In the future, the hope is that cancer therapies can be selected based on genomic profiles that identify tumor subtypes and other biological markers (347).

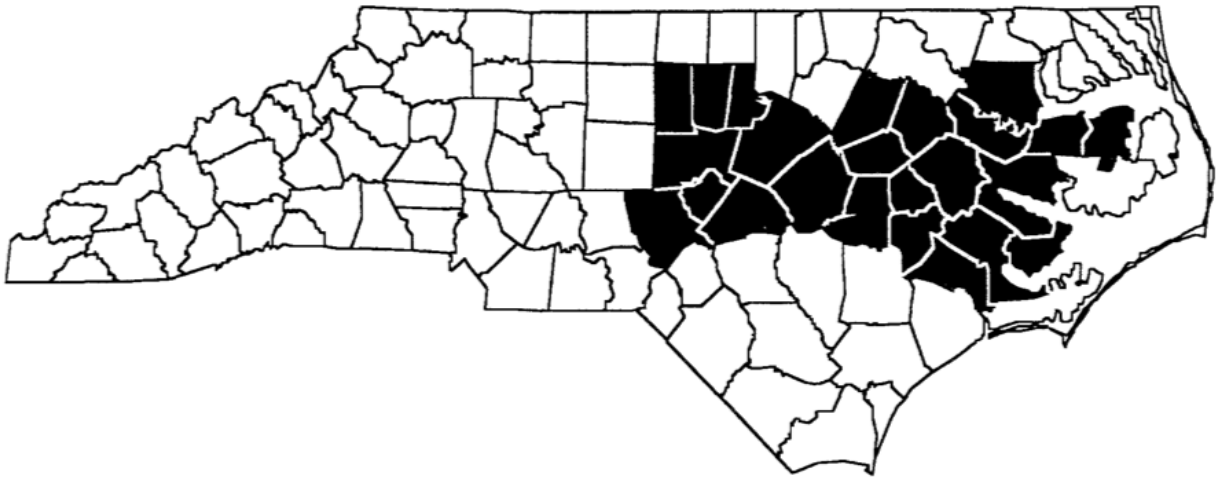


Figure 6. Carolina Breast Cancer Study Area (Phase 1 and 2)

Table 5. CBCS Sampling Probabilities

Phase I invasive cases	**Age <50**	**Age ≥ 50**
African American	100%	75%
White	67%	20%
Phase II invasive cases		
African American	100%	100%
White	50%	20%

Table 6. Base Excision Repair SNPs

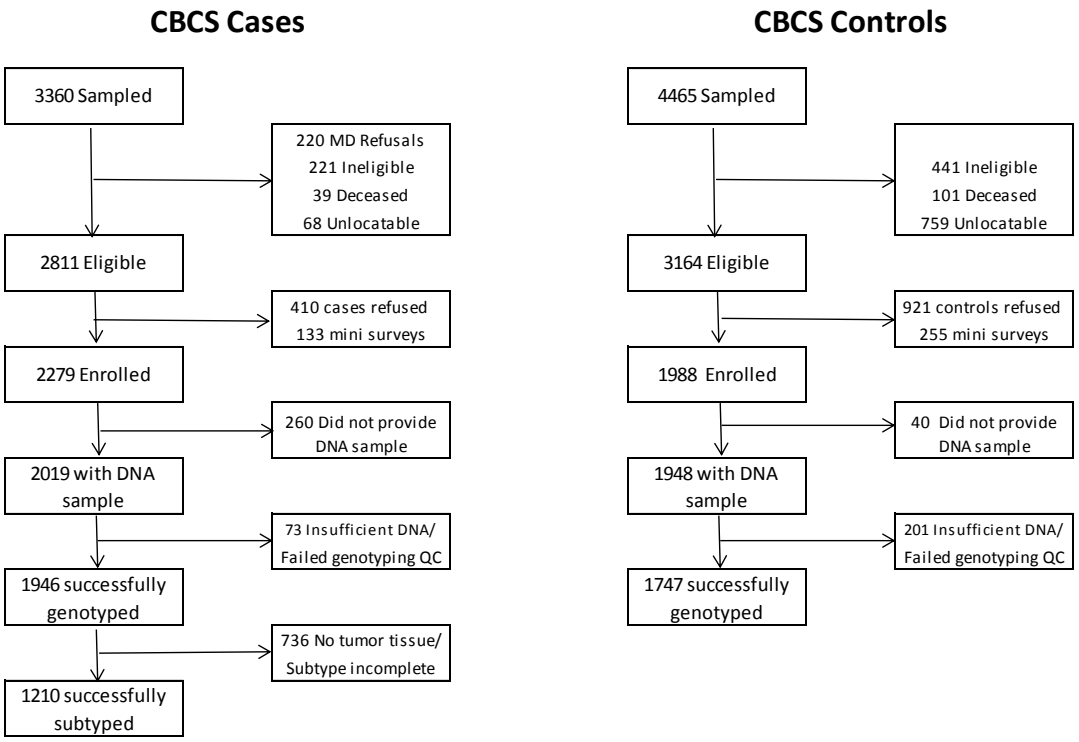
Gene	Type of variant	Original SNP set selected	Failed Pre-genotyping	Failed Post-genotyping	Successfully Genotyped	Allele Not Polymorphic	Failed HWE	Included in Final Analyis
XRCC1	Arg194Trp	rs1799782			rs1799782			rs1799782
	Arg280His	rs25489			rs25489			rs25489
	Arg399Gln	rs25487			rs25487			rs25487
	N576T	rs2307177	rs2307177					
	V72A	rs25496			rs25496	rs25496 (W)		rs25496 (AA)
	3'UTR	rs2682558			rs2682558		rs2682558	
	T304A	rs25490	rs25490					
	5'UTR	rs3213245	rs3213245					
APE1	Asp148Glu	rs3136820			rs3136820			rs3136820
	5'UTR	rs1760944		rs1760944				
	Q51H	rs1048945			rs1048945	rs1048945 (AA)		rs1048945 (W)
OGG1	Ser326Cys	rs1052133			rs1052133			rs1052133
	A85S	rs17050550	rs17050550					
	R229Q	rs1805373			rs1805373	rs1805373 (W)		rs1805373 (AA)
MUTYH	Gln324His	rs3219489			rs3219489			rs3219489
	R507Q	rs3219497			rs3219497	rs3219497 (W)		rs3219497 (AA)
	V8M	rs3219484			rs3219484			rs3219484
MBD4	splice	rs140696			rs140696			rs140696
	E346K	rs140693	rs140693					
	S342P	rs2307289			rs2307289	rs2307289 (W)		rs2307289 (AA)
	A/T/ S 273	rs10342	rs10342					
MPG	5'UTR	rs710079	rs710079					
		rs3176380	rs3176380					
		rs2234890	rs2234890					
		rs710080			rs 710080		rs710080	
NTHL1	D239Y	rs3087468			rs3087468	rs3087468		
TDG	G199S	rs4135113			rs4135113		rs4135113	
	V367L	rs2888805	rs2888805					
	5'UTR	rs4135038		rs4135038				
UNG	3'UTR	rs1018784	rs1018784					
	3'UTR	rs3219275			rs3219275	rs3219275 (W)		rs3219275 (AA)

Table 6. continued

POLB	Splice	rs2307155	rs2307155					
	P 242 R	rs3136797			rs3136797	rs3136797 (AA)		rs3136797 (W)
LIG3	R 867 H	rs3136025			rs3136025	rs3136025 (W)		rs3136025 (AA)
	5'UTR	rs12945428	rs12945428					
	3'UTR	rs4796030			rs4796030			rs4796030
NEIL1	D 252 N	rs5745926			rs5745926	rs5745926 (W)		rs5745926 (AA)
NEIL2	R 103 W	rs8191612	rs8191612					
	R 103 Q	rs8191613			rs8191613			rs8191613
	R 257 L	rs8191664			rs8191664	rs8191664 (AA)		rs8191664 (W)
	3'UTR	rs1534862			rs1534862			rs1534862
SMUG1	3'UTR	rs3136391			rs3136391	rs3136391 (W)		rs3136391 (AA)
	5'UTR	rs3087404			rs3087404			rs3087404
POLE2	L 458 V	rs34574266		rs34574266				
PCNA	intron	rs25406			rs25406			rs25406
	intron	rs17352			rs17352			rs17352
	splice	rs17349			rs17349			rs17349
	3'UTR	rs3626	rs3626					
RFC1	splice	rs17288820			rs17288820	rs17288820 (W)		rs17288820 (AA)
	I 598 V	rs2066791			rs2066791	rs2066791 (W)		rs2066791 (AA)
	5'UTR	rs17287851			rs17287851	rs17287851 (W)		rs17287851 (AA)
FEN1	5'UTR	rs412334			rs412334			rs412334
	3'UTR	rs4246215	rs4246215					
PARP1	K 123 R	rs1805407		rs1805407				
	V 762 A	rs1136410			rs1136410			rs1136410
	A 188 T	rs1805409			rs1805409	rs1805409		
	5'UTR	rs907187	rs907187					
PARP3	S 92 N	rs34224216			rs34224216	rs34224216		
	Q 270 R	rs323870			rs323870		rs323870	

Table 7. Bypass polymerase SNPs

Gene	Type of variant	Original SNP set selected	Failed Pre-genotyping	Failed Post-genotyping	Successfully Genotyped	Allele Not Polymorphic	Failed HWE	Included in Final Analyis
POLH	T 329 I	rs 35675573			rs 35675573	rs 35675573		
	M 647 L	rs 6941583		rs 6941583				
	M 595 V	rs 9333555			rs 9333555			rs 9333555
	3'UTR	rs 6899628			rs 6899628			rs 6899628
	upstream	rs 9333500		rs 9333500				
POLI	H 449 R	rs 3730823			rs 3730823	rs 3730823		
	F 507 S	rs 3218786			rs 3218786			rs 3218786
	A 706 T	rs 8305			rs 8305			rs 8305
POLL	R 438 W	rs 3730477			rs 3730477			rs 3730477
	splice	rs 3730475			rs 3730475			rs 3730475
	T 221 P	rs 3730463			rs 3730463			rs 3730463
POLM	V 246 F	rs 28382653			rs 28382653	rs 28382653		
	G 220 A	rs 28382644			rs 28382644			rs 28382644
	E 107 D	rs 28382635			rs 28382635	rs 28382635		
POLQ	A 581 V	rs 487848			rs 487848			rs 487848
	H 1201 R	rs 3218651			rs 3218651			rs 3218651
	A 2304 V	rs 532411			rs 532411			rs 532411
	Q 2513 R	rs 1381057			rs 1381057			rs 1381057
	L 2538 V	rs 3218634			rs 3218634			rs 3218634
	R 1953 Q	rs 3218637			rs 3218637	rs 3218637		
	T 982 R	rs 3218649			rs 3218649			rs 3218649
	R 66 I	rs 702017			rs 702017	rs 702017		
REV1L	V 138 M	rs 3087403			rs 3087403			rs 3087403
	F 257 S	rs 3087386			rs 3087386			rs 3087386
	N 373 S	rs 3087399			rs 3087399			rs 3087399
REV3L	Y 1078 C	rs 458017			rs 458017			rs 458017
	S 1142 L	rs 3218600	rs 3218600					
	T 11461 I	rs 462779		rs 462779				
	P 1713 S	rs 17539651			rs 17539651			rs 17539651
	V 2986 I	rs 3204953		rs 3204953				



Contact rate: 3292/3360=98%

Cooperation rate: 2412/3042=79.3%

Overall response rate: 2412/3110=77.6%

Contact rate: 3706/4465=83.0%

Cooperation rate: 2243/3164=70.9%

Overall response rate: 2243/3923=57.2%

Figure 7. Enrolled cases with genotyped data

Table 8. Subtype distribution by race

Tumor Subtype	**White N (%)**	**African-American N (%)**
Luminal (n= 788)	601(70.9)	332 (57.0)
Basal-like (n=199)	103(12.2)	122 (21.0)
HER2+/ER- (n=94)	68 (8.1)	48 (8.3)
Unclassified (n=129)	71 (8.4)	79 (13.6)
Total N=1210	843	581

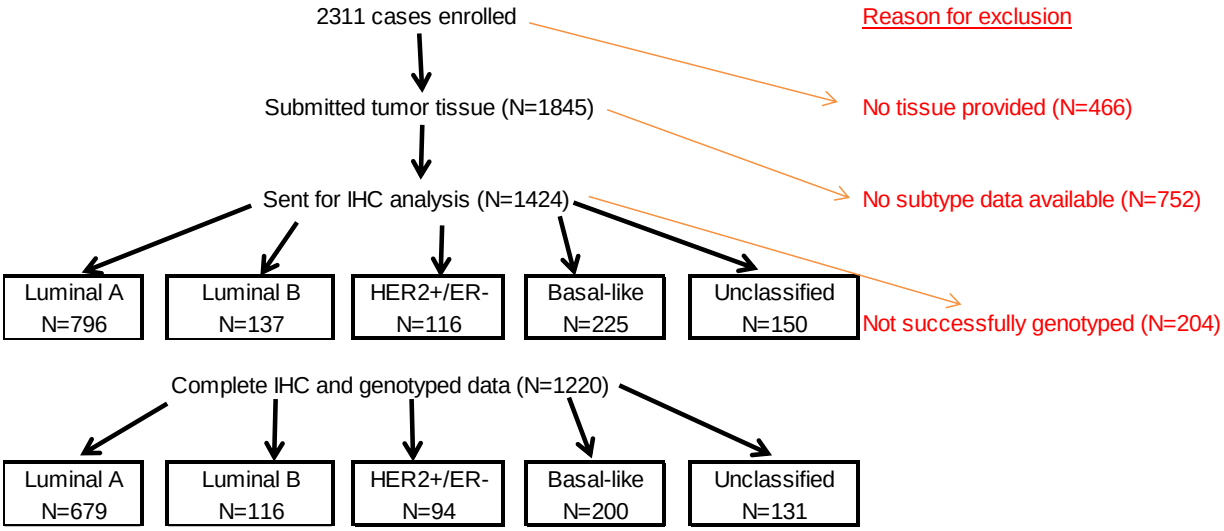
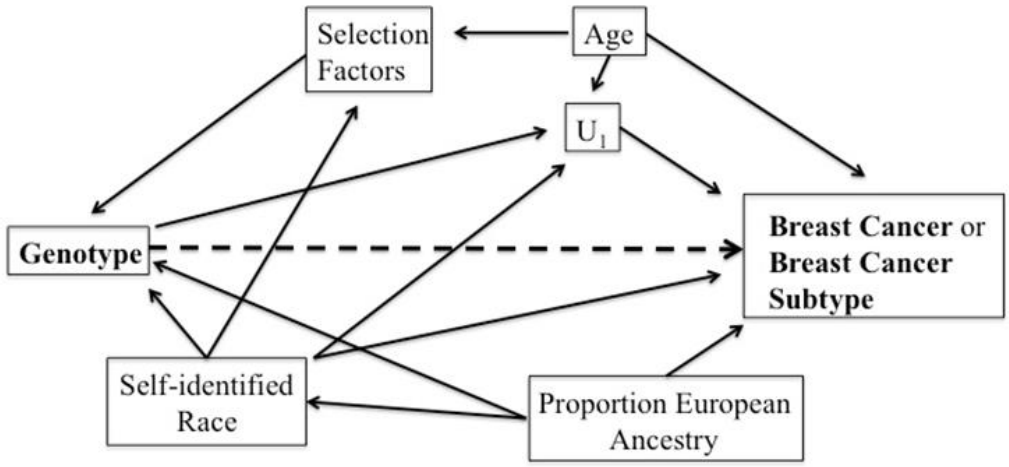


Figure 8. Enrolled cases with complete IHC and genotyped data



From O'Brien 2013

Figure 9. Directed Acyclic Graph (DAG)

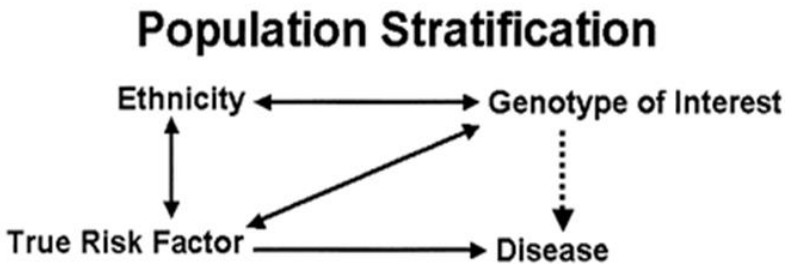


Figure 10. Confounding by ancestry
From Walchoder 2000

Table 9. Set of 144 Ancestry Informative Markers (AIMs)

rs12094678	rs11264110	rs10908312*	rs7161*	rs6666101	rs7512316	rs4659762	rs12129648	rs798443	rs12612040	rs1508061	rs7575147*
rs3755446	rs10195705	rs1257010	rs4149436	rs17049450	rs17261772	rs1117382	rs1372115	rs12692701	rs1982235	rs7424137	rs12997060
rs10202705	rs3791896	rs11901793	rs155409*	rs1303629	rs13318432	rs2660769	rs1462309	rs6414248	rs1256197	rs13080353*	rs6765491
rs9849733	rs833282	rs4859147	rs6820509	rs2687427	rs9306906	rs4619931	rs12640848	rs7689609	rs10028057*	rs6535244	rs385194
rs1372894	rs316598	rs13169284	rs16891982	rs10056388	rs13173738	rs10041728	rs33957	rs1917028	rs1380014	rs13178470	rs6556352
rs857440	rs2451563	rs10806263	rs6937164	rs4896780*	rs10952147	rs7810554	rs7788641	rs17520733	rs10254729	rs10255169	rs344454
rs4602918	rs4143633	rs1870571	rs12676654	rs13261248	rs9297712	rs7021690	rs10124991	rs1415723	rs3861709	rs10962612*	rs1885167*
rs2777804	rs1412521	rs870272	rs2488465	rs1335826	rs9416972	rs1733731	rs2184033	rs4529792	rs503677	rs9416026	rs11000419
rs1911999	rs1125217*	rs7107482	rs11607932	rs7111814	rs11223503	rs2416791	rs1490728	rs10842753	rs7134682	rs328744	rs3759171
rs2596793	rs645510	rs9525462	rs9543532	rs4885162	rs9530646	rs6491743	rs1477921	rs222674	rs2246695	rs710052	rs12900552
rs1470608	rs12900262	rs4489979	rs7086	rs4923940	rs12594483	rs567357	rs735480	rs1426654*	rs17269594	rs6494466	rs9806307
rs4506877	rs4350528	rs9923864	rs7187359	rs12926237	rs11150219	rs7189172	rs1862819	rs4792105	rs12945601	rs1043809	rs2593595
rs4793237	rs228768	rs11652805	rs4789070	rs897351	rs8113143	rs1991818	rs1011643	rs2426515	rs6023376	rs4811651*	rs2075902
rs4823460											

*SNPs which failed genotyping (i.e. weak signal intensity or in distinguishable genotype clusters)

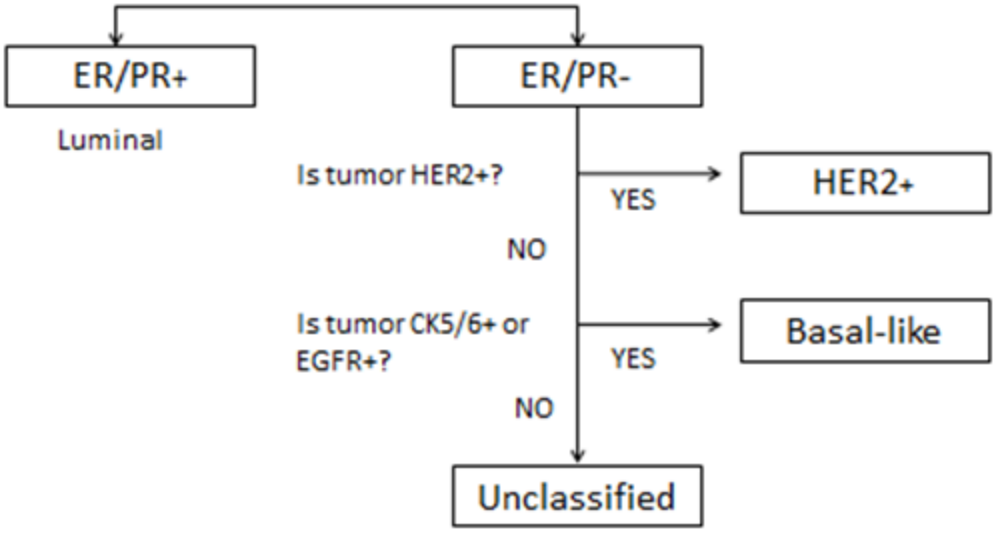


Figure 11. Classification schema for tumor subtypes

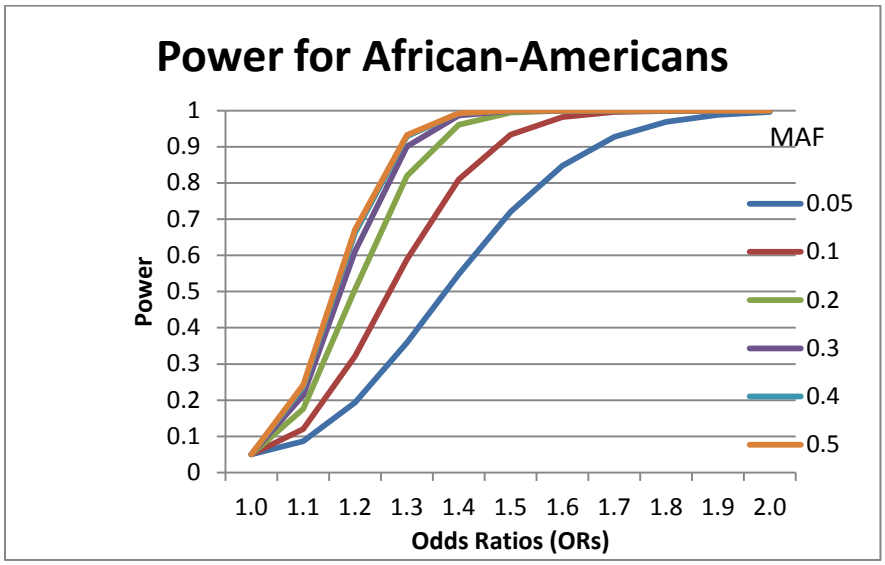


Figure 12. Power curves for African Americans

Assumptions: Additive genetic model: two-sided α=0.05; Control to Case Ratio: 0.89 P(Breast Ca at baseline)=0.01

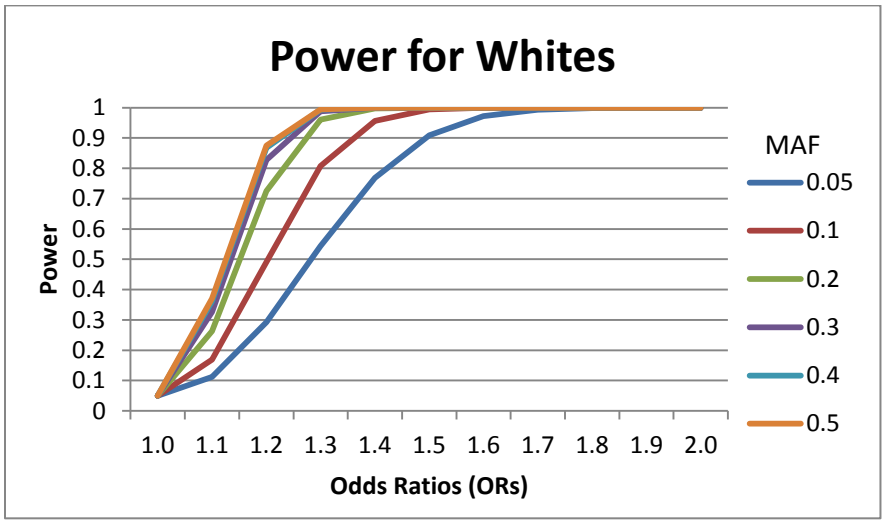


Figure 13. Power curves for Whites

Assumptions: Additive genetic model, two-sided α=0.05; Control to Case Ratio: 0.91; P(Breast Ca at baseline)=0.01

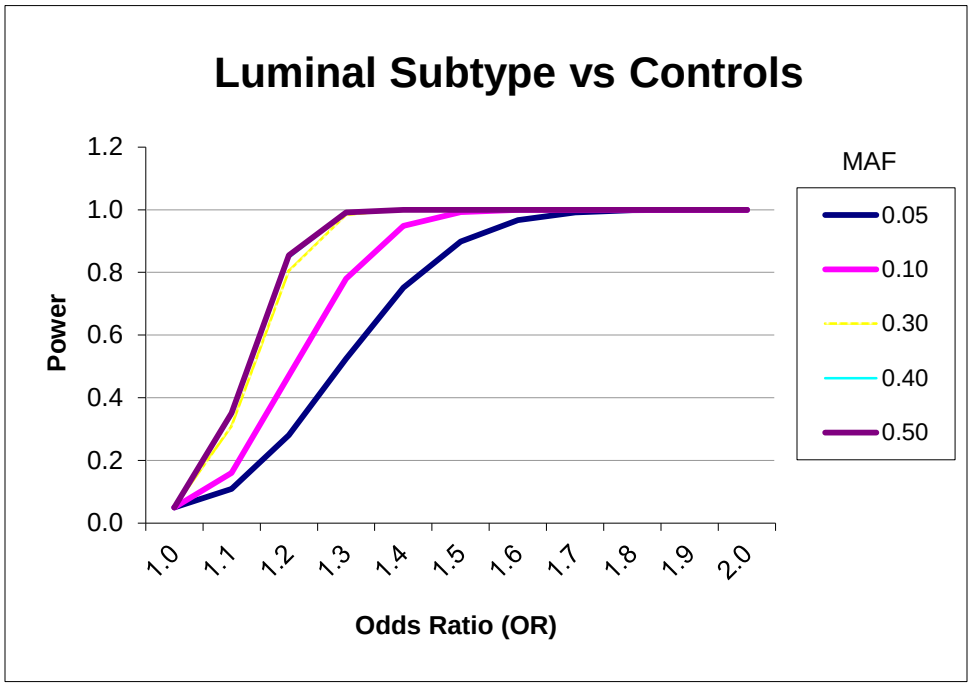


Figure 14. Power curves for luminal vs. controls

Assumptions: Additive genetic model, two-sided α=0.05; Control to Case Ratio: 2.2
P(Breast Ca at baseline)=0.01

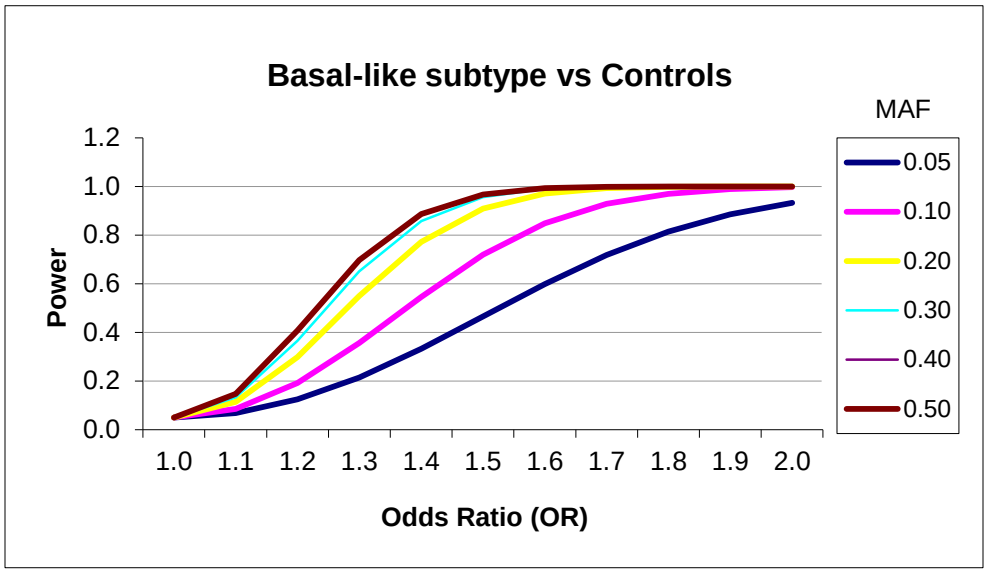


Figure 15. Power curves for Basal-like vs controls

Assumptions: Additive genetic model, two-sided α=0.05; Control to Case Ratio: 8.8
P(Breast Ca at baseline)=0.01

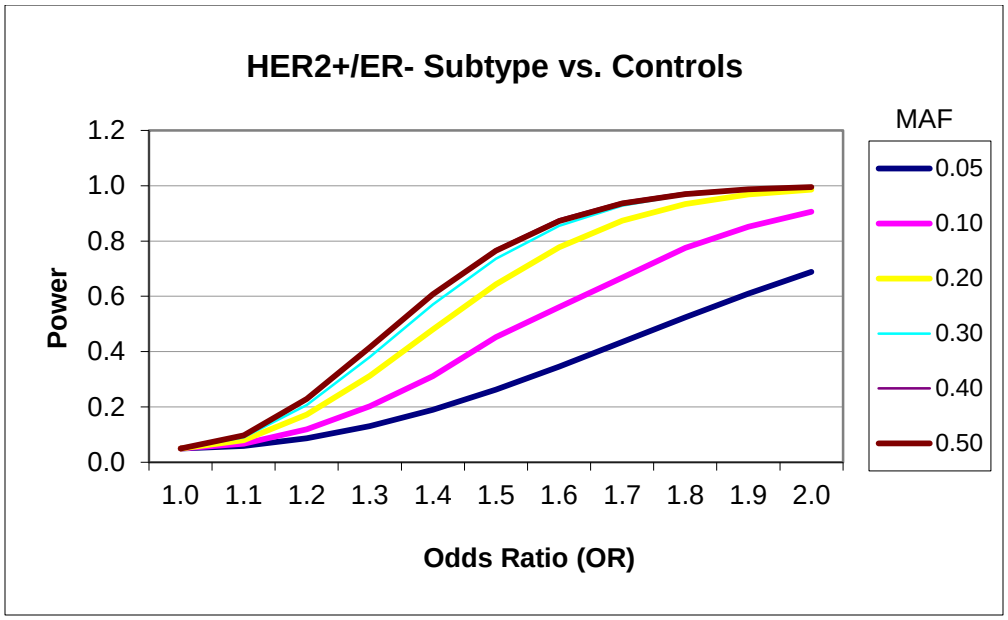


Figure 16. Power curves for HER2+/ER- vs controls

Assumptions: Additive genetic model, two-sided α=0.05; Control to Case Ratio:18.8
P(Breast Ca at baseline)=0.01

CHAPTER 3. SINGLE NUCLEOTIDE POLYMORPHISMS IN BASE EXCISION REPAIR PATHWAY GENES AND ASSOCIATION WITH BREAST CANCER AND BREAST CANCER SUBTYPES AMONG AFRICAN AMERICANS AND WHITES

3.1 Introduction

The role of DNA repair in the initiation and progression of cancer has been the subject of much investigation, both experimental and epidemiologic. Evidence has supported the role of deficient DNA repair as biologically relevant for breast tumorigenesis, including rare and highly penetrant mutations in *BRCA1,* a tumor suppressor gene that plays an essential role in the promotion and regulation of DNA repair (348). However, *BRCA1* mutations and rare variants of other genes appear to only account for 15-20% of suspected genetic predisposition to breast cancer, leaving the majority of genetic risk of breast cancer unexplained (349).

DNA repair pathways, including base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR) and double-strand break repair (DSB), homologous recombination (HR) and non-homologous end-joining (NHEJ) have been investigated in experimental systems and epidemiologic studies. The BER pathway is primarily responsible for the repair of DNA damage induced by X-rays, oxygen radicals, and alkylating agents. BER is specialized to repair non-bulky DNA base lesions such as base adducts and abasic sites (105). No consistent associations between common genetic variation in BER genes and breast cancer risk were observed in previous genetic association studies conducted to date including several meta-analyses (137-139, 141-143, 145, 146, 162, 177-180, 182-184, 186, 187, 189, 190, 194, 206-208, 212, 350-355). It is possible that conflicting results among study findings may be explained by

heterogeneity of association according to tumor subtype and limited coverage of the BER pathway. There may also be different associations by race.

We conducted a candidate pathway analysis of BER gene variants using data from the Carolina Breast Cancer Study (CBCS). CBCS, a large population-based case-control study with a racially diverse study population (40% African American and other non-Whites) and data on tumor subtype, offered an important resource to evaluate both subtype and race specific effects. Previous CBCS reports examined functional SNPs in XRCC1 (rs1799782, rs25487, and rs25489) (92, 208); in this report we had substantial coverage of the candidate SNPs in BER, including 15 BER genes and 31 associated SNPs.

3.2 Materials and Methods

3.2.1 Study population

The Carolina Breast Cancer Study (CBCS) is a population-based case-control study of breast cancer conducted in 24 counties of central and eastern North Carolina and has been described previously (275, 356). Briefly, rapid case ascertainment was implemented to identify cases from the North Carolina Central Cancer Registry (NCCCR) (277). Eligible cases included women ages 20-74, living in the study procurement area during the period of study enrollment, diagnosed with a primary invasive breast cancer between 1993 and 1996 (Phase 1) and 1996 and 2001 (Phase 2). In Phase 2 of the study, *in situ* cases of breast cancer were also eligible. Eligible controls were identified using Department of Motor Vehicles (DMV) records for women under age 65 and Health Care Financing Administration lists for women ages 65 and older. Controls were frequency matched to cases based on race and age using randomized recruitment to oversample African American and younger women (278). This study was approved by the Institutional Review Board of the University of North Carolina at Chapel Hill.

3.2.2 Baseline Study Visit

Study subjects who met eligibility criteria and provided written informed consent were scheduled for an in-home visit that included an interview and specimen collection by a trained study nurse. In addition, during the in-home visit breast cancer cases were asked to provide permission to obtain medical records and tumor tissue. The nurse-administered interview collected information about demographics and known and suspected breast cancer risk factors such as family history, personal medical history, occupational history, and reproductive factors. At the end of the interview, the nurse interviewer collected a 30 mL blood sample. Blood samples were collected from 88% of cases and 90% of controls. Whites were more likely to provide blood samples than African Americans (88% vs. 83%), but there were no significant differences in other risk factors for those who provided a blood sample and those who did not (281, 282). A total of 2,311 cases (894 African American and 1,417 Whites) and 2,022 controls (788 African Americans and 1,234 Whites) were successfully enrolled in Phase 1 and 2 of the study. This included 862 cases and 790 from Phase 1. The overall response rates for cases and controls were 78% and 57% respectively. Other study response rates have been reported previously (281).

3.2.3 SNP selection and genotyping

We searched SNP500 (http://snp500cancer.nci.nih.gov) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP) databases and selected 58 SNPs in 19 BER genes based on *in vitro* or *in silico* functional effect in BER or previously published studies in the breast cancer literature. These SNPs included non-synonymous missense, regulatory (5'UTR and 3' UTR), and intronic variants (including splice SNPs) with a minor allele frequency (MAF) of at least 5% in African Americans or Whites (Table 11).

DNA was extracted from peripheral blood lymphocytes by standard methods using an automated ABI-DNA extractor (Nuclei Acid Purification System, Applied Biosystems, Foster City, CA, USA) (356). High-throughput genotyping of selected SNPs was conducted as part of a larger set of 1536 SNPs by the UNC Mammalian Genotyping Core using Illumina GoldenGate assay (Illumina, Inc., San Diego, CA) (290). Assay intensity data and genotype cluster images for all SNPs were reviewed individually. Overall, 1,373 of 1536 (89%) SNPs passed quality control. Out of the 41 genotyped BER SNPs, we excluded 4 SNPs for which genotyping resulted in poor signal intensity or genotyping clustering, as well as, loci that were non-polymorphic overall (rs1805409 and rs34224216) or in either race (11 SNPs in Whites, 3 SNP in African Americans). Among the remaining SNPs, 4 SNPs (rs2682558, rs710080, rs4135113, and rs323870) failed Hardy Weinberg Equilibrium (HWE) ($p<0.05$) and were excluded from further analysis (Table 12). Our final analysis included genotyped data for 31 SNPs in the base excision pathway in 1972 of 2311 (85%) cases and 1776 of 2022 (88%) controls. In addition 144 ancestry informative markers (AIMs) were also genotyped to estimate African and European ancestry (281).

3.2.4 IHC analysis and subtype ascertainment

Immunohistochemical (IHC) markers were used as a surrogate for gene expression based subtyping (58). IHC staining and scoring procedures have been explained previously in detail (22, 53, 58, 59). Briefly, tumor tissue blocks were used to confirm diagnosis by a pathologist and to conduct IHC subtyping. Formalin-fixed paraffin-embedded (FFPE) tumor tissue was available 80% of cases and immunohistochemistry was completed for 62% of cases. ER/PR status was abstracted from medical records for 80% of cases while IHC was used for the remaining 20% of cases. The concordance between these two methods was 81% (307). A total of 1424 (77% of

available tumor blocks) were successfully subtyped and classified into one of five "intrinsic" subtype groups: luminal A (ER+ and/or PR+, *HER2*-, luminal B (ER+ and/or PR+, *HER2*+), *HER2*+/ER- (ER-, PR-, *HER2*+), and basal-like (ER-, PR-, *HER2*-, HER1+ and/or *CK 5/6*+), with those negative for all 5 markers considered 'unclassified' (59).

For the current study, we classified tumors as either luminal (ER+ and/or PR+; n=788), basal-like (ER-, PR-, HER2-, CK 5/6+ and/or EGFR+; n=199) or HER2+/ER- (n=94). We excluded 'unclassified' tumors from further analysis due to their uncertain status. The major distinction between the two luminal subtypes are their proliferation signatures, measured by the expression of *CCNB1*, *MKI67*, and *MYBL2* (49). HER2 expression only identifies about 30% of luminal B tumors. In the current study, we did not have information about these proliferation markers and therefore combined Luminal A and B tumors into a single 'luminal' category (48, 49). Additionally, most other studies do not have subtype data available and only have estrogen receptor status data. Therefore, we conducted an additional exploratory analysis using estrogen receptor (ER) status to evaluate comparability to "intrinsic" subtype results. We found that ER positive effects were concordant with luminal subtype results; while ER negative (ER-) effects correlated with those of basal-like and HER2+/ER- subtypes (Table 15). There were no differences between CBCS cases with and without subtyping data in terms of age, menopausal status, or family history.

3.2.5 Statistical analysis

We calculated allele and genotype frequencies stratified by case status and self-reported race (African American or White). We assessed departure from HWE for each locus by comparing expected versus observed genotype frequencies among race-specific (White and African American) controls using exact $\chi 2$ tests ($p<0.05$). We calculated pairwise linkage

disequilibrium (LD) r^2 using SAS Genetics (version9.1.3) (SAS Institute, Cary, NC) stratified by race (Table 17).

We used unconditional logistic regression models to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for race-stratifed effects of base excision repair SNPs on breast cancer, based on the additive model. We coded genotype as an ordinal variable (0, 1, or 2 for the number of minor alleles carried by the individual). If the minor allele frequency (MAF) differed by race, the more common allele in Whites was used as the referent group for both populations. We excluded non-polymorphic SNPs or SNPs with a minor allele frequency of less than 0.05 in either race. Less than 2% of participants self-identified as another race and were not included in the final analysis. We adjusted for proportion of African ancestry, as measured with a set of 144 ancestry informative markers (AIMs) (297, 306). Final models were adjusted for age at diagnosis, proportion of African ancestry and offset term for the sampling design (278).

3.2.6 Subtype analyses

We coded breast cancer subtype as a categorical variable with four levels (control, luminal, HER2+/ER-, and basal-like). We used unconditional polytomous regression models to estimate ORs and 95% CI for each subtype compared to controls.

3.2.7 Correction for multiple testing

We used FDR correction for multiple testing, following the method of Benjamini and Hochberg (313). The false discovery rate is defined as "the expected proportion of errors among the rejected hypotheses" (313). Corrections were based on the number of SNPs tested and were performed separately for African American and Whites in the race-stratified analysis and separately for luminal, HER2+/ER- and basal-like categories in the subtype analysis. Observed

p-values from the additive model were used to determine q-values. The q-value is defined as the minimum FDR that can be attained when calling a SNP significant (i.e., expected proportion of false positives) (314). Q-values were computed using the software package R. Statistical significance was set at $q<0.10$.

3.2.8 Pathway-based analysis

We used SKAT (SNP-set Kernel Association Test) to evaluate the combined effects of the genotyped SNPs in the BER pathway (333). A SNP-set refers to a set of related SNPs that are grouped based on prior biological knowledge. In the case of the current study, SNP groups were defined based on the base excision repair pathway (333). The formation of SNP-sets harnesses the potential correlation between SNPs to increase power (328). We chose a linear kernel since we assumed a log linear model. Kernel regression methods convert genomic information for a pair of individuals to a kernel score representing either similarity or dissimilarity. When applied to all pairs of the individuals, this information formed a positive semi-definite matrix (332). We tested the global hypothesis for SNPs in the pathway separately for White and African American participants (333).

3.3 Results

Characteristics of the study population with genotyping data are described in Table 10. The distributions of age, proportion of African ancestry, and menopausal status were similar between cases and controls. African American cases were more likely to be diagnosed at a later stage and were more likely to have tumors that were ER negative. African Americans were more likely to be classified as having basal-like tumors compared to Whites (22% vs. 11%).

3.3.1 Genotype associations by race

The race-stratified odds ratios for BER SNPs are summarized in Table 13. Across both race groups, six SNPs from 4 BER genes (*OGG1*, *NEIL2*, *PCNA*, and *UNG*) were associated with an increased or decreased breast cancer risk under the additive model ($p<0.05$). Among Whites, the results revealed that *OGG1* rs1052133 and *NEIL2* rs1534862 were significantly associated with an increased risk in breast cancer (rs1052133 CG/CC vs. GG, OR= 1.17, 95% CI: 1.01-1.36, P = 0.036; rs1534862 CT/TT vs. CC, OR=1.24; 95% CI: 1.07-1.44, P=0.004). Two SNPs in *PCNA* were inversely associated with breast cancer (rs17349 CT/TT vs. CC, OR=0.79; 95% CI: 0.64- 0.96, P=0.019; rs17352 AC/CC vs. AA, OR=0.76; 95% CI: 0.63-0.93, P=0.007), respectively. Among African Americans, we found another *NEIL2* SNP to be inversely associated with risk of breast cancer (rs8191613 AG/AA vs. GG, OR=0.72; 95% CI: 0.52-0.98, P=0.038). *UNG* rs3219275 was also associated with a significant increased risk of breast cancer (rs3219275 AT/AA vs. TT, OR=1.44; 95% CI: 1.01-2.06, P=0.044). After adjustment for multiple testing, only 2 SNPs (*NEIL2* 1534862 and *PCNA* 17352) remained significant (q=0.10).

3.3.2 Genotype associations by subtype

In the tumor subtype analysis, the *NEIL2* SNP (rs1534862) was positively associated with luminal and HER2+/ER- breast cancer (rs1534862 CT/TT vs. CC; OR=1.27; 95% CI: 1.06-1.52; P=0.009 and OR=1.68; 95% CI: 1.09-2.57; P=0.018), respectively (Table 14). We also found a significant inverse association between *FEN1* SNP (rs412334) and basal-like breast cancer (rs412334 AG/AA vs. GG; OR=0.56; 95% CI: 0.35-0.88; P=0.011). The ER+ SNPs

correlated with luminal SNPs while ER- SNPs correlated with HER2+/ER- and basal-like SNPs (Table 15). However, after FDR adjustment for multiple testing, none of these SNPs remained significant (q=0.10).

3.3.3 Pathway-based analysis

We assessed the global p-value for two different SNP-sets (African American and White) using the SNP-set Kernel Association Test (SKAT), adjusted for AIMs, and offset term. We did not find any significant associations. A Kernel machine test of no linear effects yielded a global p-value of 0.84 and 0.16 for African Americans, and Whites, respectively (Table 16).

3.4 Discussion

We found evidence for both race- and subtype -specific associations between BER variants and breast cancer risk. Some associations represent new findings. In Whites, 2 SNPs were associated with an increased risk (*OGG1* rs1052133 and *NEIL2* rs1534862) and 2 SNPs in high LD (r^2=0.95) in *PCNA* (rs17349 and rs17352), had an inverse association (Table 13). In African Americans, we found a *NEIL2* SNP (rs8191613) to be associated with a reduced risk and *UNG* rs3219725 to be associated with an increased risk. Two previous CBCS studies, based on the first study phase (1993-1996) had evaluated the association between *XRCC1* SNPs (rs1799782, rs25487, and rs25489) and breast cancer risk (92, 208). Duell et al. found *XRCC1* rs1799782 to be significantly associated with risk among African Americans (92); however we were unable to replicate this finding in the current analysis, underscoring the contribution of small study size to unstable genetic associations. The current study, that includes participants recruited from 1993 to 2001, has increased power, essentially doubling the sample size from Phase 1 only.

Also contributing to previous discordance in BER-pathway genetic associations, distinct tumor subtypes show heterogeneity in their associations with BER SNPs. This subtype- specific analysis showed *NEIL2* rs1534862 to be associated with luminal and *HER2*+/ER- subtype. *FEN1* rs41334 was associated with basal-like subtype. Previous studies have indicated that other risk factor profiles (both genetic and environmental) may differ by tumor subtype and race and our suggestive associations require further investigation (22, 357)

We identified SNPs in several DNA glycosylases *(OGG1, UNG*, and *NEIL2*) as being associated with breast cancer. To date, there are 12 known DNA human glycosylases that play an important role in the initial recognition and repair of a DNA lesion (106). DNA glycosylates initiate repair by releasing the modified/damaged base out of the double helix and cleaving the N-glycosidic bond of the damaged base (105). The human 8-oxoguanine DNA glycosylase (*hOGG1*) gene located on chromosome 3p26 in exon 7 encodes the bifunctional glycosylase that is primarily responsible for the accurate excision of 7,8-dihydro-8-oxoguanine (8-oxoG) (358). 8-OxoG, a product of oxidative stress, can cause a G-T transversion during DNA replication if it not removed. OGG1 variants have been shown to be highly mutagenic in mice and *in vitro* studies (219, 359). Additionally, *OGG1* rs1052133 has been one of the most studied variants in breast cancer genetic association studies. Several initial functional studies showed the *OGG1* variant to be associated with reduced DNA repair activity. In one such study, Vodicka et al. found that the capacity to repair oxidative DNA damage was significantly decreased in individuals homozygous for the variant (GG) genotype compared to other genotypes (360). Subsequently, functional and epidemiological studies that evaluated the role of *OGG1* rs1052133 with breast cancer risk in White and Asian populations showed inconsistent main results (137, 139, 141, 143, 145, 361). With the exception of a Thai case-control study that reported a

subgroup effect with postmenopausal breast cancer (OR=2.05; 95% CI: 1.14-3.67) (138), all other studies yielded effect estimates close to or at the null. In the current study, we also found CC genotype to be associated with slightly increased breast cancer risk in Whites, however we did not have sufficient sample size to evaluate the *OGG1* SNP by premenopausal status.

While various *UNG* variants have been shown to be associated with colorectal cancer, glioblastoma, B cell lymphoma, and esophageal squamous cell carcinoma, (106, 120) previous studies have not identified associations between variants in *UNG* and breast cancer risk. We are the first to report a significant increased risk of breast cancer among African Americans. *UNG* is a monofunctional glycosylase that is involved in removing uracil from DNA as a result of spontaneous deamination (362). This spontaneous deamination reaction occurs during hydrolysis of cytosine into uracil. If the uracil is not removed before DNA replication, deamination of cytosine can result in a GC to AT transition mutation, which may potentially lead to carcinogenesis. Our result for rs3219725, which is located in the 3'UTR of the gene, may indicate a novel regulatory SNP associated with breast cancer risk in African Americans; however this finding requires replication in a larger group of African Americans.

NEIL2 is a part of a newly discovered family of monofunctional DNA glycosylases (106). Laboratory studies have shown that *NEIL2* plays an important role in the repair of oxidized bases such as pymidines and cytosines (109, 148). Specifically, this *NEIL* protein cleaves the DNA backbone to generate a single-strand break at the site of the removed base with both 3'- and 5'-phosphates (114). *NEIL2* was shown to interact with *POLB* and *LIG3* in the short-patch pathway (109, 110, 150). Variants in *NEIL2* have been previously associated with increased risk in colorectal, head and neck and lung cancers. One report from the Cancer Genetic Markers of Susceptibility (CGEMS) Project noted a pair of SNPs in *NEIL2* (rs8191649 and

rs8191642) to be significantly associated with premenopausal breast cancer ($p<0.02$) (202). In the current study, we found a non-synonymous missense mutation in *NEIL2* to be associated with a decreased risk of breast cancer (rs8191613 AG/AA vs. GG; OR=0.72, 95% CI: 0.52-0.98) in African Americans. Further, we found *NEIL2* rs1534862, located in the 3'UTR of the gene, to be associated with an increased risk of breast cancer in Whites and two subtypes (Luminal and HER2+/ER). Therefore, these *NEIL2* SNPs represent a novel finding in association with breast cancer risk by race, and subtype.

FEN1 and *PCNA* are both genes involved in the ligation step of BER, specifically the long-patch repair pathway (112). While most ligation occurs via the short-patch, long-patch repair is activated when polymerase beta lyase activity is unavailable. Specifically, *PCNA* elongates the 3'-OH into the repair gap and *FENI1* acts as an endonuclease to remove the 5' flap. We found *PCNA* rs17349 and 17352 to be associated with a significant decreased risk of breast cancer in Whites. These SNPs were also in almost complete LD ($r^2=0.95$). Several animal models have associated *PCNA* mutations with cancer and genomic instability (203). A 2000 study that sequenced 9 coding variants in *PCNA* (all different that the ones selected herein) showed no associations with melanoma, breast cancer or lung cancer in a small group of 60 individuals compared with healthy controls (204). Furthermore, we found *FEN1* rs412334 to be associated with both basal-like and ER negative- breast cancer. A recent study showed overexpression and hypomethylation of *FEN1* in breast cancer cell lines (197). To our knowledge, there are no previous epidemiological studies that evaluated *FEN1* SNPs and breast cancer risk.

These findings should be considered in light of strengths and limitations of our study. Compared to other genetic association studies of breast cancer, CBCS has a sufficient sample of

African Americans. In addition, CBCS has detailed subtype data on tumors from a large population-based sample of women allowing a unique investigation of the genetics of specific breast cancer subtypes. Stratification of the dataset by subtype does reduce power for some race-specific and subtype comparisons, especially for HER2+/ER- and basal-like tumors. Phase 3 of the CBCS which is underway, uses a case-only design to add 3,000 newly diagnosed breast cancer cases with equal numbers for the four race/age subgroups (750 each). Future research in Phase 3 will have improved power to clarify these subtype associations.

We had genotype and subtype data for a large proportion of CBCS participants. However, 47% were missing IHC-based subtype data due primarily to unavailable tumor blocks. A comparison of subtyped and non-subtyped CBCS cases showed that the subtyped cases were not different from the CBCS as a whole with respect to age and menopausal status. However, cases with subtype data were more likely to be African American and to have a later stage at diagnosis, which may bias estimates for SNPs related to race or disease aggressiveness (22). In addition, we cannot rule out the role of false positives. With the exception of two SNPs, the other associated SNPs did not remain significant after adjustment for multiple comparisons. It is also noteworthy that definitions for luminal breast cancer have evolved since original CBCS IHC subtyping methods were published (58). As a result, the CBCS -defined Luminal A and Luminal B cases do not capitalize on current subtyping differences that suggest use of PR positivity or Ki-67 to distinguish the two. Therefore, it is likely that there is heterogeneity within the group of luminal breast cancers identified herein. Nonetheless, our subtyping methods here have the advantage of excluding tumors that were negative for all markers tested. Only triple negatives that were also positive for a basal-like marker are included among basal-like cancers, reducing outcome misclassification potential in this important subgroup. To our knowledge, this

represents one of the largest collections of African American breast cancer cases with tumor subtype classification.

Although we did not find any significant combined effects of SNPs in the BER pathway using SKAT, to our knowledge, this is one of the few studies to have used this recent kernel-based machine learning method to assess pathway effects in cancer (175, 213). We recognize that our pathway analysis was limited by the density of SNP coverage across BER pathway genes. Thus SKAT may be better applied to studies that utilize tag-SNP approaches in candidate pathways or GWAS.

In summary, this study adds important new information for the role of BER in breast cancer etiology by using tumor tissue to evaluate subtype-specific effects and considers carefully selected regulatory and coding SNP-sets in a biologically established DNA repair pathway using innovative statistical approaches. After controlling for multiple comparisons, we found two SNPs significantly associated with breast cancer in Whites. We identified other suggestive associations for breast cancer in SNPs not previously evaluated for their relationship with breast cancer incidence. Larger studies such as the CBCS Phase 3 with improved power for race- and subtype-specific analyses and collaborative consortia will help gain further insight into the role of genetic variation in the base excision repair pathway and the risk of breast cancer.

Table 10. Characteristics of CBCS participants with genotyping data

Characteristic	Cases				Controls			
	African American		White		African American		White	
	N	%	N	%	N	%	N	%
Total (N)	742	100.0	1,204	100.0	658	100.0	1,089	100.0
Average age at selection	52		52		52		53	
Average proportion of African ancestry	0.78		0.064		0.774		0.066	
Menopausal status								
Pre-menopausal	324	41.4	539	30.5	290	59.1	456	62.7
Post-menopausal	418	58.6	665	69.5	368	40.9	633	37.3
Stage								
0[a]	88	11.9	349	29.0				
1	216	29.1	393	32.6				
2	299	40.3	328	27.2				
3	76	10.2	68	5.7				
4	27	3.6	15	1.3				
Missing	36	4.9	51	4.2				
Subtype								
Luminal	269	56.3	519	75.6				
HER2+/ER-	38	7.6	56	5.8				
Basal-like	108	22.0	91	10.7				
Estrogen Receptor (ER) Status								
Positive	235	49.6	482	68.9				
Negative	251	50.4	242	31.1				

[a] *carcinoma in situ*

Table 11. List of successfully genotyped BER SNP in HWE

Gene	rs#	Type of variant	Amino Acid Change	Allele Change	SIFT score[a]
XRCC1	rs1799782	missense	R194W	C/T	0
	rs25489	missense	R280H	A/G	0.05
	rs25487	missense	Q399R	A/G	0.05
	rs25496	missense	V72A	C/T	0.07
APE1	rs1130409	missense	D148E	G/T	0.09
	rs1048945	missense	Q51H	C/G	0
OGG1	rs1052133	missense	S326C	C/G	0.05
	rs1805373	missense	R229Q	A/G	0.05
MUTYH	rs3219489	missense	Q324H	C/G	0.4
	rs3219497	missense	R507Q	A/G	0.22
	rs3219484	missense	V8M	A/G	0.02
MBD4	rs140696	splice	N/A	C/T	N/A
	rs2307289	missense	S342P	C/G	0.29
NTHL1	rs3087468	missense	D239Y	G/T	0
UNG	rs3219275	3'UTR	N/A	A/T	N/A
POLB	rs3136797	missense	P242R	C/G	0
LIG3	rs3136025	missense	R867H	A/G	0.18
	rs4796030	3'UTR	N/A	A/C	N/A
NEIL1	rs5745926	missense	D252N	A/G	0.27
NEIL2	rs8191613	missense	R103Q	A/G	0.61
	rs8191664	missense	R257L	G/T	0.28
	rs1534862	3'UTR	N/A	C/T	N/A
SMUG1	rs3136391	3'UTR	N/A	C/T	N/A
	rs3087404	5'UTR	N/A	A/G	N/A
PCNA	rs25406	intron	N/A	C/T	N/A
	rs17349	splice	N/A	C/T	N/A
RFC1	rs17288820	splice	N/A	A/G	N/A
	rs2066791	missense	I598V	A/G	0.08
	rs17287851	5'UTR	N/A	C/T	N/A
FEN1	rs412334	5'UTR	N/A	A/G	N/A
PARP1	rs1136410	missense	V762A	C/T	0.16

[a]Ranges from 0 to 1. The amino acid substitution is predicted damaging if the score is <= 0.05, and tolerated if the score is > 0.05.

Table 12. Minor Allele Frequencies (MAFs) of BER variants stratified by race and case status

		Whites				African Americans			
Gene	SNP[a]	Minor Allele	Cases	Controls	p HWE[b]	Minor Allele	Cases	Controls	p HWE[b]
XRCC1	rs1799782	T	0.063	0.068	0.366	T	0.065	0.067	0.971
XRCC1	rs25489	A	0.049	0.044	0.929	A	0.036	0.027	0.457
XRCC1	rs25487	A	0.365	0.352	0.343	A	0.161	0.146	0.371
XRCC1	rs25496	C	0.000	0.001	N/A	C	0.055	0.062	0.300
XRCC1	rs2682558	A	0.210	0.201	0.000	A	0.179	0.181	0.000
APE1	rs3136820	G	0.475	0.481	0.456	G	0.373	0.393	0.708
APE1	rs1048945	C	0.040	0.040	0.810	C	0.010	0.005	N/A
OGG1	rs1052133	C	0.223	0.239	0.629	C	0.167	0.159	0.636
OGG1	rs1805373	A	0.001	0.001	N/A	A	0.080	0.084	0.094
MUTYH	rs3219489	C	0.244	0.248	0.324	C	0.257	0.262	0.700
MUTYH	rs3219497	A	0.000	0.000	N/A	A	0.037	0.027	0.421
MUTYH	rs3219484	A	0.069	0.069	0.562	A	0.013	0.013	0.737
MBD4	rs140696	T	0.098	0.092	0.645	T	0.201	0.210	0.997
MBD4	rs2307289	C	0.002	0.000	N/A	C	0.129	0.110	0.429
MPG	rs710080	G	0.014	0.017	0.000	A	0.351	0.354	0.444
TDG	rs4135113	A	0.020	0.022	0.457	A	0.158	0.149	0.000
UNG	rs3219275	A	0.000	0.001	N/A	A	0.055	0.043	0.439
POLB	rs3136797	G	0.020	0.020	0.496	G	0.004	0.003	N/A
LIG3	rs3136025	A	0.002	0.001	N/A	A	0.090	0.087	0.146
LIG3	rs4796030	A	0.433	0.455	0.498	A	0.133	0.145	0.022
NEIL1	rs5745926	A	0.000	0.001	N/A	A	0.018	0.014	0.722
NEIL2	rs8191613	A	0.018	0.017	0.579	A	0.053	0.072	0.336
NEIL2	rs8191664	T	0.018	0.014	0.634	T	0.004	0.003	N/A
NEIL2	rs1534862	T	0.238	0.208	0.020	T	0.327	0.321	0.136
SMUG1	rs3136391	C	0.000	0.000	N/A	C	0.049	0.039	0.290
SMUG1	rs3087404	A	0.435	0.436	0.853	G	0.324	0.332	0.532

PCNA	rs25406	T	0.408	0.413	0.600	T	0.466	0.452	0.008
PCNA	rs17352	C	0.101	0.123	0.505	A	0.435	0.447	0.920
PCNA	rs17349	T	0.098	0.116	0.098	T	0.280	0.285	0.101
RFC1	rs17288820	G	0.000	0.000	N/A	G	0.019	0.021	0.189
RFC1	rs2066791	G	0.000	0.000	N/A	G	0.018	0.017	0.663
RFC1	rs17287851	T	0.000	0.000	N/A	T	0.083	0.081	0.481
FEN1	rs412334	A	0.162	0.166	0.808	A	0.032	0.037	0.237
PARP1	rs1136410	C	0.161	0.144	0.688	C	0.047	0.062	0.314
PARP1	rs1805409	A	0.000	0.000	N/A	A	0.007	0.011	N/A
PARP3	rs34224216	A	0.000	0.000	N/A	A	0.009	0.009	N/A
PARP3	rs323870	G	0.003	0.002	0.000	G	0.168	0.156	0.237

[a]SNP, single nucleotide polymorphism

[b]P HWE, Hardy-Weinberg equilibrium, $p<0.05$

N/A indicates non-polymorphic loci or MAF<0.05

Table 13. Association of BER variants with breast cancer stratified by race

			Whites							African Americans						
			Cases		Controls		OR (95% CI)[b]	p value[c]	q value[d]	Cases		Controls		OR (95% CI)[b]	p value[c]	q value[d]
Gene	SNP	Genotype[a]	N	%	N	%				N	%	N	%			
XRCC1	rs1799782	CC	1057	0.879	947	0.870	1.00			647	0.872	573	0.871	1.00		
		CT	140	0.116	135	0.124	0.93 (0.71, 1.21)	0.8691		94	0.127	82	0.125	0.93(0.67, 1.30)	0.4699	
		TT	6	0.005	7	0.006	0.78 (0.24, 2.48)	0.7165		1	0.001	3	0.005	0.37 (0.04, 3.55)	0.4034	
		CT +TT	146	0.121	142	0.130	0.92 (0.72, 1.18)	0.5133	0.8723	95	0.128	85	0.129	0.90 (0.66, 1.23)	0.5146	0.8078
	rs25489	GG	1082	0.902	991	0.913	1.00			689	0.930	623	0.947	1.00		
		AG	118	0.098	92	0.085	1.25 (0.93, 1.69)	0.9620		51	0.069	34	0.052	1.37 (0.87, 2.17)	0.2239	
		AA	0	0.000	2	0.002	NA	NA		1	0.001	1	0.002	1.11 (0.07, 18.11)	0.9704	
		AG + AA	118	0.098	94	0.087	1.19 (0.89, 1.60)	0.2507	0.7123	52	0.070	35	0.053	1.34 (0.87, 2.07)	0.1895	0.7727
	rs25487	GG	475	0.403	459	0.427	1.00			519	0.706	473	0.726	1.00		
		AG	548	0.465	476	0.443	1.06 (0.88, 1.29)	0.7979		195	0.265	168	0.258	1.12 (0.87, 1.44)	0.4914	
		AA	156	0.132	140	0.130	1.08 (0.82, 1.42)	0.7356		21	0.029	11	0.017	1.71 (0.79. 3.69)	0.2201	
		AG + AA	704	0.597	616	0.573	1.05 (0.92, 1.19)	0.4940	0.8723	216	0.294	179	0.275	1.17 (0.94, 1.46)	0.1533	0.7727
	rs25496	TT	1203	0.999	1087	0.998	1.00			666	0.898	577	0.877	1.00		
		CT	1	0.001	2	0.0018	NA	NA		71	0.096	80	0.122	0.80 (0.56, 1.13)	0.0791	
		CC	0	0.000	0	0.000	NA	NA		5	0.007	1	0.002	4.87 (0.55, 42.89)	0.1270	
		CT +TT	1	0.0008	2	0	NA	NA		76	0.102	81	0.123	0.91 (0.66, 1.26)	0.5571	0.8028
APE1	rs3136820	GG	321	0.268	287	0.264	1.00			297	0.402	244	0.371	1.00		
		GT	617	0.515	556	0.511	1.09 (0.87, 1.36)	0.6621		332	0.449	309	0.470	1.00 (0.73, 1.38)	0.5316	
		TT	261	0.218	246	0.226	1.08 (0.84, 1.39)	0.7177		110	0.149	104	0.158	1.17 (0.84, 1.62)	0.2069	
		GT + TT	878	0.7323	802	0.7365	1.04 (0.92, 1.18)	0.5578	0.8723	442	0.598	413	0.629	1.10 (0.94 1.28)	0.2398	0.7727
	rs1048945	GG	1107	0.921	1004	0.923	1.00			726	0.980	651	0.989	1.00		
		CG	94	0.078	82	0.075	1.06 (0.77, 1.47)	0.6820		15	0.020	7	0.011	NA	NA	
		CC	1	0.001	2	0.002	0.67 (0.06, 7.41)	0.7259		0	0.000	0	0.000	NA	NA	
		CG + CC	95	0.079	84	0.0772	1.04 (0.76, 1.43)	0.7991	0.9316	15	0.02	7	0.011	NA	NA	
OGG1	rs1052133	GG	721	0.602	628	0.577	1.00			516	0.696	463	0.705	1.00		
		CG	419	0.350	401	0.369	1.19 (0.79, 1.81)	0.9329		203	0.274	179	0.273	0.72 (0.36, 1.44)	0.5032	
		CC	57	0.048	59	0.054	1.39 (0.93, 2.08)	0.0464		22	0.030	15	0.023	0.67 (0.34, 1.33)	0.2340	
		CG + CC	476	0.3976	460	0.4228	1.17 (1.01, 1.36)	0.0359	0.1795	225	0.304	194	0.295	0.90 (0.73, 1.11)	0.3217	0.7774
	rs1805373	GG	1202	0.998	1086	0.997	1.00			628	0.846	555	0.844	1.00		
		AG	2	0.002	3	0.003	NA	NA		109	0.147	95	0.144	0.94 (0.69. 1.28)	0.3592	
		AA	0	0.000	0	0.000	NA	NA		5	0.007	8	0.012	0.47 (0.14, 1.55)	0.2395	
		AG + AA	2	0.0017	3	0.0028	NA	NA		114	0.154	103	0.157	0.88 (0.66, 1.16)	0.3561	0.7944

			Whites							African Americans						
			Cases		Controls		OR (95% CI)[b]	p value[c]	q value[d]	Cases		Controls		OR (95% CI)[b]	p value[c]	q value[d]
Gene	SNP	Genotype[a]	N	%	N	%				N	%	N	%			
MUTYH	rs3219489	GG	689	0.577	608	0.559	1.00			417	0.563	355	0.542	1.00		
		CG	429	0.359	418	0.385	0.89 (0.74, 1.07)	0.1166		267	0.360	257	0.392	0.93 (0.74, 1.17)	0.2291	
		CC	77	0.064	61	0.056	1.16 (0.80, 1.69)	0.2670		57	0.077	43	0.066	1.22 (0.79. 1.87)	0.2785	
		CG +CC	506	0.4234	479	0.4406	0.98 (0.85, 1.13)	0.7842	0.9316	324	0.437	300	0.458	1.02 (0.86, 1.21)	0.8199	0.8729
	rs3219497	GG	1203	0.999	1088	0.999	1.00			687	0.926	624	0.948	1.00		
		AG	1	0.001	1	0.001	NA	NA		55	0.074	33	0.050	1.58 (1.00, 2.49)	0.9696	
		AA	0	0.000	0	0.000	NA	NA		0	0.000	1	0.002	NA		
		AG + AA	1	0.0008	1	0.0009	NA	NA		55	0.074	34	0.052	1.47 (0.95, 2.30)	0.0866	0.6612
	rs3219484	GG	1042	0.865	942	0.865	1.00			723	0.974	641	0.974	1.00		
		AG	158	0.131	143	0.131	1.05 (0.81, 1.35)	0.6112		18	0.024	17	0.026	1.04 (0.51, 2.11)	0.9764	
		AA	4	0.003	4	0.004	0.72 (0.15, 3.35)	0.6538		1	0.001	0	0.000	NA		
		AG + AA	162	0.1345	147	0.135	1.02 (0.80, 1.31)	0.8503	0.9316	19	0.026	17	0.026	1.07 (0.54, 2.14)	0.8428	0.8729
MBD4	rs140696	CC	976	0.811	896	0.823	1.00			475	0.640	410	0.624	1.00		
		CT	219	0.182	185	0.170	1.15 (0.91, 1.44)	0.1978		235	0.317	218	0.332	0.92 (0.73, 1.17)	0.7963	
		TT	9	0.008	8	0.007	0.64 (0.24, 1.76)	0.3215		32	0.043	29	0.044	0.78 (0.45, 1.35)	0.4546	
		CT +TT	228	0.1894	193	0.1772	1.09 (0.88, 1.34)	0.4525	0.8723	267	0.36	247	0.376	0.91 (0.75 1.10)	0.3143	0.7774
	rs2307289	TT	1200	0.997	1088	0.999	1.00			564	0.760	519	0.789	1.00		
		CT	3	0.003	1	0.001	NA	NA		165	0.222	133	0.202	1.13 (0.87, 1.48)	0.6462	
		CC	1	0.001	0	0.000	NA	NA		13	0.018	6	0.009	1.67 (0.80, 4.66)	0.3910	
		CT +TT	4	0.0033	1	0.0009	NA	NA		178	0.24	139	0.211	1.16 (0.91, 1.48)	0.2223	0.7727
UNG	rs3219275	TT	1204	1.000	1087	0.998	1.00			662	0.892	604	0.918	1.00		
		AT	0	0.000	2	0.002	NA	NA		78	0.105	52	0.079	1.54 (1.06, 2.24)	0.2644	
		AA	0	0.000	0	0.000	NA	NA		2	0.003	2	0.003	0.70 (0.09, 5.28)	0.5807	
		AT+AA	0	0	2	0.0018	NA	NA		80	0.108	54	0.082	1.44 (1.01, 2.06)	0.0446	0.6467
POLB	rs3136797	CC	1156	0.960	1045	0.960	1.00			736	0.992	654	0.994	1.00		
		CG	48	0.040	44	0.040	0.97 (0.62, 1.51)	0.8850		6	0.008	4	0.006	NA	NA	
		GG	0	0.000	0	0.000	0.00	NA		0	0.000	0	0.000	NA	NA	
		CG + GG	48	0.0399	44	0.0404	0.97 (0.62, 1.51)	0.8850	0.9316	6	0.008	4	0.006	NA	NA	
LIG3	rs3136025	GG	1200	0.997	1087	0.998	1.00			615	0.829	551	0.837	1.00		
		AG	4	0.003	2	0.002	NA	NA		121	0.163	99	0.151	1.09 (0.81, 1.48)	0.0540	
		AA	0	0.000	0	0.000	NA	NA		6	0.008	8	0.012	0.33 (0.10, 1.09)	0.0587	
		AG + AA	4	0.0033	2	0.0018	NA	NA		127	0.171	107	0.163	0.97 (0.74, 1.27)	0.8067	0.8729

			Whites							African Americans						
			Cases		Controls		OR (95% CI)[b]	p value[c]	q value[d]	Cases		Controls		OR (95% CI)[b]	p value[c]	q value[d]
Gene	SNP	Genotype[a]	N	%	N	%				N	%	N	%			
	rs4796030	CC	379	0.315	329	0.302	1.00			565	0.762	488	0.743	1.00		
		AC	608	0.505	529	0.486	1.02 (0.83, 1.24)	0.0958		157	0.212	148	0.225	0.90 (0.69, 1.18)	0.7877	
		AA	217	0.180	231	0.212	0.76 (0.59, 0.98)	0.0142		20	0.027	21	0.032	0.91 (0.48, 1.73)	0.8894	
		AC + AA	825	0.6852	760	0.6979	0.89 (0.79, 1.01)	0.0645	0.2580	177	0.239	169	0.257	0.92 (0.74, 1.15)	0.4662	0.8078
NEIL1	rs5745926	GG	1203	0.999	1087	0.998	1.00			716	0.965	640	0.973	1.00		
		AG	1	0.001	2	0.002	NA	NA		26	0.035	18	0.027	1.28 (0.68, 2.42)	0.4417	
		AA	0	0.000	0	0.000	NA	NA		0	0.000	0	0.000	NA	NA	
		AG + AA	1	0.0008	2	0.0018	NA	NA		26	0.035	18	0.027	1.28 (0.68, 2,42)	0.4417	0.8078
NEIL2	rs8191613	GG	1158	0.963	1053	0.967	1.00			663	0.897	568	0.865	1.00		
		AG	44	0.037	36	0.033	0.99 (0.62, 1.60)	0.9757		73	0.099	84	0.128	0.72 (0.51, 1.02)	0.9547	
		AA	0	0.000	0	0.000	NA	NA		3	0.004	5	0.008	0.49 (0.11, 2.19)	0.4769	
		AG + AA	44	0.0366	36	0.0331	0.99 (0.62, 1.60)	0.9757	0.9757	76	0.103	89	0.136	0.72 (0.52, 0.98)	0.0386	0.6467
	rs8191664	GG	1160	0.964	1058	0.972	1.00			736	0.992	654	0.994	1.00		
		GT	44	0.037	31	0.029	1.16 (0.70, 1.91)	0.5670		6	0.008	4	0.006	1.48 (0.41, 5.38)	0.5525	
		TT	0	0.000	0	0.000	NA	NA		0	0.000	0	0.000	NA	NA	
		GT + TT	44	0.0365	31	0.0285	1.16 (0.70, 1.91)	0.5670	0.8723	6	0.008	4	0.006	1.48 (0.41, 5.38)	0.5525	0.8078
	rs1534862	CC	691	0.574	695	0.638	1.00			341	0.460	312	0.474	1.00		
		CT	452	0.376	334	0.307	1.44 (1.20, 1.74)	0.0118		317	0.427	270	0.410	1.07 (0.85, 1.34)	0.5825	
		TT	60	0.050	60	0.055	1.10 (0.74, 1.63)	0.6581		84	0.113	76	0.116	0.99 (0.69, 1.42)	0.8035	
		CT +TT	512	0.4256	394	0.3618	1.24 (1.07, 1.44)	0.0038	0.0750	401	0.54	346	0.526	1.02 (0.87, 1.19)	0.8417	0.8729
SMUG1	rs3136391	TT	1204	1.000	1088	1.000	1.00			669	0.904	609	0.926	1.00		
		CT	0	0.000	0	0.000	NA	NA		70	0.095	47	0.071	1.39 (0.94, 2.08)	0.2817	
		CC	0	0.000	0	0.000	NA	NA		1	0.001	2	0.003	0.48 (0.04, 5.47)	0.4689	
		CT +TT	0	0	0	0	NA	NA		71	0.096	49	0.074	1.30 (0.89, 1.90)	0.1727	0.7727
	rs3087404	GG	385	0.320	345	0.317	1.00			79	0.107	69	0.105	1.00		
		AG	590	0.490	538	0.495	0.97 (0.80, 1.19)	0.5883		323	0.435	299	0.454	0.90 (0.72, 1.14)	0.3843	
		AA	229	0.190	205	0.188	1.05 (0.81, 1.35)	0.6143		340	0.458	290	0.441	1.02 (0.70, 1.48)	0.7083	
		AG + AA	819	0.6802	743	0.6829	1.02 (0.90, 1.15)	0.8056	0.9316	663	0.894	589	0.895	0.97 (0.82, 1.15)	0.7299	0.8729
PCNA	rs25406	CC	419	0.348	368	0.341	1.00			213	0.289	213	0.326	1.00		
		CT	586	0.487	532	0.493	0.91 (0.75, 1.11)	0.3895		362	0.491	290	0.444	1.27 (0.99, 1.64)	0.0664	
		TT	198	0.165	180	0.167	0.97 (0.75, 1.26)	0.8812		163	0.221	150	0.230	1.07 (0.79, 1.44)	0.6836	
		CT +TT	784	0.6517	712	0.6593	0.97 (0.86, 1.10)	0.6515	0.9307	525	0.711	440	0.674	1.05 (0.90, 1.22)	0.5375	0.8078

			Whites							African Americans						
			Cases		Controls		OR (95% CI)[b]	p value[c]	q value[d]	Cases		Controls		OR (95% CI)[b]	p value[c]	q value[d]
Gene	SNP	Genotype[a]	N	%	N	%				N	%	N	%			
	rs17352	AA	978	0.812	836	0.768	1.00			141	0.190	132	0.201	1.00		
		AC	209	0.174	239	0.220	0.70 (0.56, 0.87)	0.0655		363	0.489	324	0.492	0.97 (0.76, 1.24)	0.9831	
		CC	17	0.014	14	0.013	1.11 (0.50, 2.44)	0.4845		238	0.321	202	0.307	0.94 (0.68, 1.29)	0.7201	
		AC + CC	226	0.1877	253	0.2324	0.76 (0.63, 0.93)	0.0075	0.0750	601	0.81	526	0.799	0.97 (0.83, 1.13)	0.6804	0.8729
	rs17349	CC	982	0.816	846	0.777	1.00			382	0.515	345	0.524	1.00		
		CT	207	0.172	234	0.215	0.71 (0.57, 0.89)	0.0960		304	0.410	251	0.382	1.13 (0.89, 1.42)	0.0706	
		TT	15	0.013	9	0.008	1.46 (0.61, 3.52)	0.2184		56	0.076	62	0.094	0.77 (0.51, 1.16)	0.1178	
		CT + TT	222	0.1844	243	0.2232	0.79 (0.64, 0.96)	0.0198	0.1320	360	0.485	313	0.476	0.98 (0.82, 1.16)	0.7707	0.8729
RFC1	rs17288820	AA	1204	1.000	1089	1.000	1.00			715	0.964	631	0.959	1.00		
		AG	0	0.000	0	0.000	NA	NA		26	0.035	26	0.040	1.24 (0.06, 26.51)	0.9910	
		GG	0	0.000	0	0.000	NA	NA		1	0.001	1	0.002	1.58 (0.08, 31.94)	0.6572	
		AG + GG	0	0	0	0	NA	NA		27	0.036	27	0.041	1.27 (0.74, 2.18)	0.3922	0.8078
	rs2066791	AA	1203	0.999	1089	1.000	1.00			715	0.964	636	0.967	1.00		
		AG	1	0.001	0	0.000	NA	NA		27	0.036	22	0.033	1.10 (0.61, 1.98)	0.7536	
		GG	0	0.000	0	0.000	NA	NA		0	0.000	0	0.000	NA	NA	
		AG +GG	1	0.0008	0	0	NA	NA		27	0.036	22	0.033	1.10 (0.61, 1.98)	0.7536	0.8729
	rs17287851	CC	1204	1.000	1089	1.000	1.00			623	0.840	554	0.842	1.00		
		CT	0	0.000	0	0.000	NA	NA		115	0.155	101	0.154	1.00 (0.74, 1.35)	0.8643	
		TT	0	0.000	0	0.000	NA	NA		4	0.005	3	0.005	0.86 (0.18, 4.14)	0.8499	
		CT + TT	0	0.000	0	0.000	NA	NA		119	0.160	104	0.158	0.99 (0.74, 1.31)	0.9297	0.9297
FEN1	rs412334	GG	850	0.706	755	0.694	1.00			695	0.937	611	0.929	1.00		
		AG	318	0.264	304	0.279	0.86 (0.71, 1.05)	0.2725		47	0.063	45	0.068	0.85 (0.54, 1.33)	0.9777	
		AA	36	0.030	29	0.027	1.06 (0.62, 1.81)	0.6239		0	0.000	2	0.003	NA	NA	
		AG + AA	354	0.294	333	0.306	0.92 (0.78, 1.08)	0.3076	0.7690	47	0.063	47	0.071	0.79 (0.51, 1.04)	0.2768	0.7774
PARP1	rs1136410	TT	854	0.709	796	0.731	1.00			673	0.907	578	0.878	1.00		
		CT	313	0.260	272	0.250	1.03 (0.84, 1.26)	0.1252		68	0.092	79	0.120	0.73 (0.51, 1.04)	0.6317	
		CC	37	0.031	21	0.019	1.78 (1.01, 3.17)	0.0509		1	0.001	1	0.002	1.06 (0.07, 17.33)	0.8764	
		CT + CC	350	0.291	293	0.269	1.12 (0.95. 1.33)	0.1850	0.6167	69	0.093	80	0.122	0.74 (0.52, 1.05)	0.0912	0.6612

a Minor allele frequency for Whites used as reference for both races

b Odds ratio and 95% confidence interval, adjusted for age, self-identified race, African ancestry, and offset term

c P-value is unadjusted for multiple comparisons

Table 14. Association of BER variants with breast cancer stratified by subtype

			Luminal			HER2+/ER-			Basal-like		
Gene	**SNP**	**Genotype**	**OR (95% CI)[a]**	**p value[b]**	**q value[c]**	**OR (95% CI)[a]**	**p value[b]**	**q value[c]**	**OR (95% CI)[a]**	**p value[b]**	**q value[c]**
XRCC1	rs1799782	CC	Referent			Referent			Referent		
		CT	0.94 (0.72, 1.23)	0.7634		0.89 (0.46, 1.70)	0.9710		1.12 (0.73, 1.72)	0.9701	
		TT	1.08 (0.33, 3.58)	0.8566		NA	0.9705		NA	0.9705	
		CT +TT	0.95 (0.73, 1.23)	0.6865	0.8966	0.86 (0.45, 1.64)	0.6391	0.9054	1.08 (0.70, 1.66)	0.7295	0.9787
	rs25489	GG	Referent			Referent			Referent		
		AG	1.33 (0.97, 1.82)	0.9482		1.39 (0.68, 2.86)	0.9826		1.55 (0.92, 2.61)	0.9672	
		AA	NA	0.9502		NA	0.9834		NA	0.9692	
		AG + AA	1.30 (0.95, 1.87)	0.1033	0.4474	1.40 (0.68, 2.87)	0.3998	0.8496	1.54 (0.92, 2.59)	0.1206	0.5453
	rs25487	GG	Referent			Referent			Referent		
		AG	1.04 (0.85, 1.26)	0.3977		1.07 (0.68, 1.69)	0.2245		0.99 (0.71, 1.39)	0.8529	
		AA	1.29 (0.94, 1.77)	0.1289		0.54 (0.19. 1.54)	0.2087		0.91 (0.47, 1.75)	0.7747	
		AG + AA	1.08 (0.90, 1.30)	0.4252	0.8234	0.98 (0.63, 1.53)	0.9487	0.9717	0.97 (0.70, 1.35)	0.8983	0.9787
APE1	rs3136820	GG	Referent			Referent			Referent		
		GT	1.08 (0.86, 1.37)	0.7465		1.13 (0.64, 2.01)	0.6510		0.84 (0.56, 1.26)	0.2221	
		TT	1.11 (0.86, 1.43)	0.5420		1.05 (0.56, 1.97)	0.9535		1.04 (0.68, 1.60)	0.4708	
		GT + TT	1.09 (0.87, 1.37)	0.4359	0.8234	1.10 (0.64, 1.90)	0.7340	0.9187	0.91 (0.62, 1.34)	0.6413	0.9787
OGG1	rs1052133	GG	Referent			Referent			Referent		
		CG	1.05 (0.66, 1.68)	0.9129		1.88 (0.44, 8.10)	0.5228		0.96 (0.42, 2.23)	0.7277	
		CC	1.14 (0.72, 1.80)	0.4228		2.09 (0.50, 8.81)	0.2862		1.10 (0.49, 2.49)	0.6194	
		CG + CC	1.11 (0.71, 1.74)	0.6554	0.8966	2.01 (0.48, 8.37)	0.3359	0.8362	1.04 (0.47, 2.34)	0.9014	0.9787
MUTYH	rs3219489	GG	Referent			Referent			Referent		
		CG	0.81 (0.67, 0.98)	0.0485		0.98 (0.63, 1.53)	0.8219		0.94 (0.69, 1.30)	0.8433	
		CC	1.05 (0.73, 1.52)	0.3900		1.11 (0.46, 2.67)	0.8299		0.99 (0.52, 1.87)	0.9918	
		CG +CC	0.84 (0.70, 1.01)	0.0587	0.4474	0.99 (0.65, 1.52)	0.9717	0.9717	0.95 (0.70, 1.29)	0.7276	0.9787
	rs3219484	GG	Referent			Referent			Referent		
		AG	1.20 (0.89, 1.62)	0.7771		1.68 (0.87, 3.23)	0.9660		1.32 (0.77, 2.28)	0.9700	
		AA	1.81 (0.39, 8.43)	0.5209		NA	0.9685		NA	0.9711	
		AG + AA	1.22 (0.90, 1.63)	0.1957	0.5545	1.64 (0.85, 3.16)	0.1446	0.8362	1.31 (0.76, 2.24)	0.3636	0.9287
MBD4	rs140696	CC	Referent			Referent			Referent		
		CT	0.90 (0.72, 1.12)	0.6702		1.04 (0.63, 1.71)	0.3745		1.23 (0,87, 1.74)	0.9393	
		TT	0.94 (0.52, 1.73)	0.9878		0.40 (0.05, 3.03)	0.3666		1.46 (0.65, 3.29(	0.5052	
		CT +TT	0.90 (0.73, 1.12)	0.3497	0.8234	0.97 (0.59, 1.58)	0.9366	0.9717	1.23 (0.88, 1.71)	0.1811	0.6157

Gene	SNP	Genotype	Luminal			HER2+/ER-			Basal-like		
			OR (95% CI)[a]	p value[b]	q value[c]	OR (95% CI)[a]	p value[b]	q value[c]	OR (95% CI)[a]	p value[b]	q value[c]
LIG3	rs4796030	CC	Referent			Referent			Referent		
		AC	1.04 (0.85, 1.28)	0.1820		0.79 (.47, 1.31)	0.2511		0.78(0.54, 1.11)	0.6950	
		AA	0.67 (0.50, 0.91)	0.0032		1.08 (0.57, 206)	0.5112		0.70 (0.41, 1.20)	0.3737	
		AC + AA	0.95 (0.78, 1.16)	0.6023	0.8966	0.86 (0.54, 1.38)	0.5287	0.8798	0.77 (0.55, 1.08)	0.1101	0.5453
NEIL2	rs8191613	GG	Referent			Referent			Referent		
		AG	0.77 (0.53, 1.12)	0.9508		1.29 (0.62, 2.68)	0.9827		0.81 (0.44, 1.46)	0.9669	
		AA	NA	0.9489		NA	0.9833		NA	0.9659	
		AG + AA	0.74 (0.51, 1.08)	0.1192	0.4474	1.22 (0.59, 2.55)	0.5693	0.8798	0.76 (0.42, 1.37)	0.3824	0.9287
	rs1534862	CC	Referent			Referent			Referent		
		CT	1.28 (1.06, 1.54)	0.2130		1.82 (1.17, 2.82)	0.0327		0.98 (0.70, 1.36)	0.8100	
		TT	1.23 (0.88, 1.73)	0.6041		1.02 (0.42, 2.49)	0.5352		1.05 (0.61, 1.79)	0.8552	
		CT +TT	1.27 (1.06, 1.52)	0.0092	0.1564	1.68 (1.09, 2.57)	0.0178	0.3026	0.99 (0.72, 1.34)	0.9278	0.9787
SMUG1	rs3087404	GG	Referent			Referent			Referent		
		AG	0.99 (0.79, 1.24)	0.5452		0.58 (0.34, 1.01)	0.0040		1.14 (0.75, 1.72)	0.6663	
		AA	1.09 (0.84, 1.41)	0.3990		1.23 (0.70, 2.16)	0.3084		1.10 (0.70, 1.75)	0.7888	
		AG + AA	1.02 (0.82, 1.26)	0.8678	0.9220	0.78 (0.48, 1.29)	0.3354	0.8362	1.12 (0.75, 1.68)	0.5657	0.9787
PCNA	rs25406	CC	Referent			Referent			Referent		
		CT	1.05 (0.86, 1.28)	0.5564		1.33 (0.81, 2.17)	0.3441		1.31 (0.92, 1.86)	0.2615	
		TT	0.99 (0.77, 1.28)	0.7617		1.19 (0.64, 2.23)	0.9297		1.24 (0.80, 1.92)	0.7128	
		CT +TT	1.03 (0.86, 1.25)	0.7384	0.8966	1.29 (0.81, 2.07)	0.2819	0.8362	130 (0.93, 1.81)	0.1283	0.5453
	rs17352	AA	Referent			Referent			Referent		
		AC	0.91 (0.73, 1.14)	0.2418		0.95 (0.56, 1.61)	0.8211		0.94 (0.63, 1.40)	0.1677	
		CC	1.07 (0.76, 1.50)	0.4694		0.80 (0.36, 1.80)	0.5928		1.41 (0.85 2.32)	0.8620	
		AC + CC	0.94 (0.75, 1.17)	0.5497	0.8966	0.91 (0.54, 1.55)	0.7566	0.9187	1.01 (0.69, 1.48)	0.8848	0.9787
	rs17349	CC	Referent			Referent			Referent		
		CT	0.98 (0.80, 1.21	0.9440		0.80 (0.48, 1.31)	0.9546		0.98 (0.69, 1.38)	0.4042	
		TT	0.95 (.59, 1.53)	0.8499		0.66 (0.20, 2.21)	0.6219		1.35 (0.72, 2.55)	0.3179	
		CT + TT	0.98 (0.80, 1.20)	0.8416	0.9220	0.77 (0.47, 1.25)	0.3120	0.8362	1.00 (0.72, 1.39)	0.8744	0.9787
FEN1	rs412334	GG	Referent			Referent			Referent		
		AG	0.86 (0.69, 1.09)	0.6016		0.79 (0.44, 1.41)	0.2689		0.51 (0.31, 0.83)	0.0617	
		AA	0.59 (0.28, 1.27)	0.2487		1.53 (0.44, 5.39)	0.3948		1.00 (0.33, 3.07)	0.5513	
		AG + AA	0.84 (0.67, 1.05)	0.1316	0.4474	0.86 (0.50, 1.48)	0.5586	0.8798	0.56 (0.35, 0.88)	0.0111	0.1887
PARP1	rs1136410	TT	Referent			Referent			Referent		
		CT	0.96 (0.77, 1.20)	0.1182		1.22 (0.73, 2.03)	0.6763		0.93 (0.62, 1.39)	0.1269	
		CC	1.72 (0.87, 3.42)	0.1079		2.15 (0.48, 9.59)	0.3805		2.41 (0.79, 7. 34)	0.1079	
		CT + CC	1.00 (0.81, 1.24)	0.9979	0.9979	1.26 (0.77, 2.08)	0.3443	0.8362	0.99 (0.67, 1.46)	0.9787	0.9787

a Odds ratio and 95% confidence interval, adjuste

b P-value is unadjusted for multiple comparisons

Table 15. Assocation of BER variants with breast cancer stratified by estrogen receptor (ER) status.

Gene	SNP	Genotype	ER positive	p value	q value	ER negative	p value	q value
XRCC1	rs1799782	CC	Referent			Referent		
		CT	0.97 (0.73, 1.27)			0.92 (0.67, 1.26)		
		TT	0.92 (0.25, 3.43)			0.91 (0.19, 4.29)		
		CT +TT	0.97 (0.74, 1.27)	0.7948	0.9651	0.92 (0.67, 1.25)	0.5805	0.8998
	rs25489	GG	Referent			Referent		
		AG	1.39 (1.01, 1.92)			1.57 (1.09, 2.27)		
		AA	NA			1.41 (0.14, 14.45)		
		AG + AA	1.36 (0.99, 1.87)	0.0581	0.2910	1.57 (1.09, 2.25)	0.0144	0.2448
	rs25487	GG	Referent			Referent		
		AG	1.04 (0.85, 1.28)			1.03 (0.82, 1.30)		
		AA	1.32 (0.96, 1.83)			0.94 (0.61, 1.44)		
		AG + AA	1.09 (0.90, 1.32)	0.3639	0.7733	1.02 (0.81, 1.27)	0.8929	0.9670
APE1	rs3136820	GG	Referent			Referent		
		GT	1.07 (0.84, 1.36)			1.04 (0.78, 1.38)		
		TT	1.07 (0.82, 1.39)			1.17 (0.86, 1.59)		
		GT + TT	1.06 (0.85, 1.34)	0.5804	0.8677	1.09 (0.83, 1.43)	0.5484	0.8998
OGG1	rs1052133	GG	Referent			Referent		
		CG	1.08 (0.67, 1.75)			0.87 (0.50, 1.52)		
		CC	1.15 (0.72, 1.84)			1.06 (0.62, 1.82)		
		CG + CC	1.13 (0.71, 1.79)	0.6125	0.8677	.99 (0.58, 1.69)	0.9764	0.9764
MUTYH	rs3219489	GG	Referent			Referent		
		CG	0.81 (0.67, 0.99)			1.00 (0.80, 1.24)		
		CC	1.11 (0.76, 1.61)			1.10 (0.71, 1.70)		
		CG +CC	0.85 (0.71, 1.02)	0.0856	0.291	1.01 (0.82, 1.25)	0.9101	0.967
	rs3219484	GG	Referent			Referent		
		AG	1.29 (0.95, 1.74)			1.10 (0.75, 1.63)		
		AA	1.95 (0.42, 9.02)			NA		
		AG + AA	1.31 (0.97, 1.76)	0.0783	0.2910	1.07 (0.72, 1.58)	0.7410	0.8998
MBD4	rs140696	CC	Referent			Referent		
		CT	0.91 (0.73, 1.14)			1.07 (0.84, 1.37)		
		TT	0.96 (0.52, 1.77)			0.81 (0.41, 1.59)		
		CT +TT	0.92 (0.74, 1.14)	0.4234	0.7998	1.04 (0.82, 1.32)	0.7215	0.8998

Gene	SNP	Genotype	ER positive	p value	q value	ER negative	p value	q value
LIG3	rs4796030	CC	Referent			Referent		
		AC	1.02 (0.83, 1.26)			0.91 (0.71, 1.16)		
		AA	0.66 (0.48, 0.91)			0.85 (0.60, 1.22)		
		AC + AA	0.93 (0.76, 1.14)	0.4937	0.8393	0.90 (0.71, 1.13)	0.3540	0.7523
NEIL2	rs8191613	GG	Referent			Referent		
		AG	0.73 (0.49, 1.09)			0.82 (0.54, 1.23)		
		AA	NA			NA		
		AG + AA	0.71 (0.47, 1.05)	0.0841	0.2910	0.78 (0.52, 1.18)	0.2342	0.6885
	rs1534862	CC	Referent			Referent		
		CT	1.25 (1.03, 1.52)			1.24 (0.99, 1.55)		
		TT	1.23 (0.87, 1.74)			1.00 (0.67, 1.49)		
		CT +TT	1.25 (1.04, 1.50)	0.0185	0.291	1.19 (0.97, 1.47)	0.1026	0.6885
SMUG1	rs3087404	GG	Referent			Referent		
		AG	0.94 (0.77, 1.15)			0.90 (0.71, 1.13)		
		AA	0.93, (0.70, 1.22)			1.06 (0.78, 1.45)		
		AG + AA	1.04 (0.84, 1.30)	0.7198	0.9413	0.87 (0.67, 1.13)	0.2994	0.5814
PCNA	rs25406	CC	Referent			Referent		
		CT	1.01 (0.82, 1.24)			1.24 (0.97, 1.57)		
		TT	0.94 (0.72, 1.23)			1.23 (0.91, 1.67)		
		CT +TT	0.99 (0.82, 1.20)	0.9313	0.9871	1.24 (0.98, 1.55)	0.6996	0.7271
	rs17352	AA	Referent			Referent		
		AC	0.97 (0.77, 1.22)			0.85 (0.65, 1.19)		
		CC	1.09 (0.76, 1.56)			1.07 (0.74, 1.55)		
		AC + CC	0.86 (0.70, 1.06)	0.1590	0.3908	0.76 (0.60, 0.97)	0.0243	0.8998
	rs17349	CC	Referent			Referent		
		CT	1.02 (0.82, 1.26)			0.85 (0.67, 1.09)		
		TT	1.00 (0.61, 1.63)			0.97 (0.59, 1.58)		
		CT + TT	1.02 (0.83, 1.25)	0.8763	0.9871	0.87 (0.69, 1.09)	0.2303	0.6885
FEN1	rs412334	GG	Referent			Referent		
		AG	0.88 (0.69, 1.11)			0.69 (0.51, 0.93)		
		AA	0.53 (0.23, 1.20)			1.17 (0.55, 2.51)		
		AG + AA	0.85 (0.67, 1.07)	0.1609	0.3908	0.73 (0.54, 0.97)	0.0315	0.2678
PARP1	rs1136410	TT	Referent			Referent		
		CT	0.96 (0.76, 1.21)			1.00 (0.76, 1.31)		
		CC	1.74 (0.87, 3.49)			2.29 (1.04, 5.05)		
		CT + CC	1.00 (0.81, 1.25)	0.9871	0.9871	1.06 (.82, 1.38)	0.6621	0.8998

a Odds ratio and 95% confidence interval, adjusted for age, proportion European ancestry, and offset term

b P-value is unadjusted for multiple comparisons

c FDR adjusted

Table 16. SKAT analysis for base excision repair SNPs

Group	Number of SNPs	SNP-set	Global p-value
African Americans	29	rs1799782 rs25489 rs25487 rs25496 rs3136820 rs1052133 rs1805373 rs3219489 rs3219497 rs3219484 rs140696 rs2307289 rs3219275 rs3136025 rs4796030 rs5745926 rs8191613 rs1534862 rs3136391 rs3087404 rs25406 rs17352 rs17349 rs17288820 rs2066791 rs17287851 rs412334 rs1136410 rs323870	0.8638
Whites	20	rs1799782 rs25489 rs25487 rs3136820 rs1048945 rs1052133 rs3219489 rs3219484 rs140696 rs4135113 rs3136797 rs4796030 rs8191613 rs8191664 rs1534862 rs3087404 rs34857719 rs25406 rs17352 rs17349 rs412334 rs1136410 rs323870	0.1601

Table 17. Linkage Disequilibrium by race

Whites

XRCC1	rs1799782	rs25489	rs25487
rs1799782	1.00		
rs25489	0.00	1.00	
rs25487	0.04	0.02	1.00
rs25496	0.00	0.01	0.00

APE1	rs3136820	rs1048945
rs3136820	1.00	
rs1048945	0.16	1.00

OGG1	rs1052133	rs1805373
rs1052133	1.00	
rs1805373	0.04	1.00

MUTYH	rs3219489	rs3219497	rs3219484
rs3219489	1.00		
rs3219497	0.12	1.00	
rs3219484	0.02	0.01	1.00

MBD4	rs140696	rs2307289
rs140696	1.00	
rs2307289	0.11	1.00

LIG3	rs3136025	rs4796030
rs3136025	1.00	
rs4796030	0.01	1.00

NEIL2	rs8191613	rs8191664	rs1534862
rs8191613	1.00		
rs8191664	0.91	1.00	
rs1534862	0.06	0.24	1.00

PCNA	rs25406	rs17352	rs17349
rs25406	1.00		
rs17352	0.26	1.00	
rs17349	0.29	0.95	1.00

African Americans

XRCC1	rs1799782	rs25489	rs25487
rs1799782	1.00		
rs25489	0.00	1.00	
rs25487	0.01	0.07	1.00
rs25496	0.00	0.13	0.01

APE1	rs3136820	rs1048945
rs3136820	1.00	
rs1048945	0.00	1.00

OGG1	rs1052133	rs1805373
rs1052133	1.00	
rs1805373	0.02	1.00

MUTYH	rs3219489	rs3219497	rs3219484
rs3219489	1.00		
rs3219497	0.11	1.00	
rs3219484	0.00	0.00	1.00

MBD4	rs140696	rs2307289
rs140696	1.00	
rs2307289	0.53	1.00

LIG3	rs3136025	rs4796030
rs3136025	1.00	
rs4796030	0.16	1.00

NEIL2	rs8191613	rs8191664	rs1534862
rs8191613	1.00		
rs8191664	0.05	1.00	
rs1534862	0.03	0.01	1.00

PCNA	rs25406	rs17352	rs17349
rs25406	1.00		
rs17352	0.00	1.00	
rs17349	0.34	0.31	1.00

[a] r^2= correlation coefficient squared

CHAPTER 4. SINGLE NUCLEOTIDE POLYMORPHISMS IN DNA BYPASS POLYMERASE GENES AND ASSOCIATION WITH BREAST CANCER AND BREAST CANCER SUBTYPES AMONG AFRICAN AMERICANS AND WHITES

4.1 Introduction

The integrity of DNA is constantly threatened by DNA damage from both endogenous and exogenous sources. DNA may be damaged as much as a million times per cell per day (103). Unrepaired DNA damage can result in genomic instability, leading to point mutations, deletions and insertions, as well as chromosomal alterations. These defects increase the probability of a hit to an oncogene or tumor suppressor and may ultimately lead to carcinogenesis. To maintain genomic integrity, there is an intricate system of damage response mechanisms (363). Researchers have identified at least 15 different DNA polymerases in humans which are essential for DNA replication, DNA repair and the tolerance of DNA damage (222).

DNA replicative polymerases which carry out the bulk of DNA synthesis have evolved to be very precise and efficient (224). Despite this high fidelity, a replication error may generate a one-sided double-strand break (DSB) or degrade to a full DSB if it not repaired prior to initiation of DNA replication (225, 226). In order to resume DNA replication at a stalled replication fork, two damage tolerance mechanisms have been proposed; template switching in homologous recombination (HR) and translesion synthesis (TLS) (227). During template switching, synthesis on the undamaged template strands can continue to a limited extent (119, 222, 227). In contrast, translesion synthesis is conducted by specialized DNA polymerases that do not directly repair the damage, but rather bypass the damage to prevent replication fork stalling. Unlike replicative polymerases, bypass polymerases lack 3' to 5' exonuclease

(proofreading) activity and are able to resume replication without an undamaged template. However, this also contributes to their low fidelity and potential mis-incorporation of nucleotides (234).

Previous research has shown that mutations in bypass polymerases may be associated with the risk of cancer. *POLH* (pol eta) was shown to be highly efficient in the bypass of UV lesions, such as cyclobutane pyrimidine dimer (CPD). Germline mutations in *POLH* were identified in patients with xeroderma pigmentosum (XP), an autosomal recessive genetic disorder of DNA repair in which individuals are unable to repair damage caused by UV light and thus are at high risk of developing skin cancer. Typically mutations in the nucleotide excision repair (NER) genes results in XP, however this was the first evidence that bypass polymerases may be involved in human cancer (249). While genetic variation in other DNA repair pathway genes have been studied extensively in association with breast cancer, the focus on DNA bypass polymerases is relatively recent. Several studies have evaluated bypass polymerases in association with breast cancer risk (166, 202, 213, 254, 261, 262). Two reports from a comprehensive analysis of DNA repair genes in nested case-control study within the NHS (Nurses' Health Study) II cohort evaluated SNPs from 5 bypass polymerase genes (*POLB*, *POLD1*, *POLE, POLL, POLK*) (202, 213). Han et al. reported that 44 SNPs, including 3 SNPs in *POLK* (rs3213801, rs5744533, and rs3756558), were significantly associated with premenopausal breast cancer risk (239 cases, 477 controls) ($p<0.05$) (202). However, in the study of postmenopausal women (1,145 cases, 1,142 controls), there were no associations with any of the studied bypass polymerase SNPs (213). In an *in vivo* study of breast cancer cells, Yang et al. reported elevated *POLI* expression when exposed to UV radiation (254). In a gene sequencing study, Wang et al. identified several mutations in *POLB*, including an 87-bp deletion

in the catalytic domain of the gene (166). In two other reports, *POLQ* overexpression in tumors was associated with poor prognosis of breast cancer (261, 273).

We evaluated the association between DNA bypass polymerases variants and breast cancer risk in the Carolina Breast Cancer Study, a large population-based case-control study with a racially diverse study population and tumor subtype data. This analysis offered a unique opportunity to evaluate both breast cancer subtype and race specific effects of 7 bypass polymerase genes.

4.2 Materials and Methods

4.2.1 Study population

The Carolina Breast Cancer Study (CBCS) is a population-based case-control study of breast cancer conducted in 24 counties of central and eastern North Carolina and has been described previously (275, 356). Briefly, rapid case ascertainment was implemented to identify eligible cases from the North Carolina Central Cancer Registry (NCCCR) (277). Eligible cases included women ages 20-74, living in the selected North Carolina counties during their primary breast cancer diagnosis. There were 2 phases of enrollment: Phase 1 (1993-1996) enrolled only invasive cancers, while Phase 2 (1996-2001) also included women with *in situ* cancer. Eligible controls were identified using Department of Motor Vehicles (DMV) records for women under age 65 and Health Care Financing Administration lists for women ages 65 and older. Controls were frequency matched to cases based on race and age using randomized recruitment to oversample African American and younger women, a subgroup often underrepresented in research studies of breast cancer (278). This study was approved by the Institutional Review Board of the University of North Carolina at Chapel Hill.

4.2.2 Baseline Study Visit

During an in-home visit, a written signed informed consent was obtained from cases and controls Release forms for medical records and tumor tissue were obtained from cases. The in-home interview consisted of a nurse-administered questionnaire asking about demographic factors and known and suspected breast cancer risk factors. A 30mL blood sample was collected at the end of the nurse visit. Blood samples were collected from 88% of cases and 90% of controls. Whites were more likely to provide blood samples than African Americans (88% vs. 83%), but there were no significant differences in other risk factors for those who provided a blood sample and those who did not (281, 282). A total of 2,311 cases (894 African American and 1,417 Whites) and 2,022 controls (788 African Americans and 1,234 Whites) were successfully enrolled in the study. The overall response rates for cases and controls were 78% and 57% respectively (281).

4.2.3 SNP selection

We searched SNP500 (http://snp500cancer.nci.nih.gov) and dbSNP (http://www.ncbi.nlm.nih.gov/SNP) databases and selected 30 SNPs in bypass polymerase genes to be genotyped based on reported *in vitro* or *in silico* functional effect and the DNA repair literature (364). These SNPs included non-synonymous missense, regulatory (5'UTR and 3' UTR), and intronic variants (including splice SNPs) with a minor allele frequency (MAF) of at least 5% in African Americans or Whites (Table 19).

4.2.4 Genotyping methods and quality control

DNA was extracted from peripheral blood lymphocytes by standard methods using an automated ABI-DNA extractor (Nuclei Acid Purification System, Applied Biosystems, Foster City, CA, USA) (356). High-throughput genotyping of selected SNPs was conducted at the

Mammalian Genotyping Core Facility at the University of North Carolina at Chapel Hill. An Illumina high-multiplex GoldenGate Genotyping with 1536 SNP Sentrix Array matrix included 30 SNPs in 7 bypass polymerases genes (POLH, POLI, POLM, POLQ, REV1L, and REV3L) (Illumina Inc, San Diego, CA) (290). Assay intensity data and genotype cluster images for all SNPs were reviewed individually. Overall, 1,373 of 1536 (89%) SNPs passed quality control. Out of the 30 TLS SNPs selected, we excluded 4 SNPs (rs6941583, rs9333500, rs462779 and rs3204953) for which genotyping resulted in poor signal intensity or genotyping clustering, 1 SNP that failed due to poor assay design (rs3218600), and 3 non-polymorphic SNPs (rs3730823, rs28382644, and rs28382635). All SNPs were in HWE (P>0.05) (Table 20). Our final analysis included genotyped data for 22 SNPs in bypass polymerase genes for 1,972 cases and 1,776 controls. In addition 144 ancestry informative markers (AIMs) were genotyped to estimate African and European ancestry (281).

4.2.5 IHC analysis and subtype ascertainment

Immunohistochemical (IHC) markers were used as a surrogate for gene expression-based subtyping (58). IHC staining and scoring procedures have been explained previously in detail (22, 53, 58, 59). Briefly, tumor tissue blocks were used to confirm diagnosis by a pathologist and to conduct IHC subtyping. Formalin-fixed paraffin-embedded (FFPE) tumor tissue was available 80% of cases and immunohistochemistry was completed for 62% of cases. ER/PR status was abstracted from medical records for 80% of cases while IHC was used for the remaining 20% of cases (for whom clinical status was not available). In cases that had both medical records and IHC data available, the concordance of ER/PR status was 81% (307). A total of 1424 (77% of available tumor blocks) were successfully subtyped and classified tumors as either luminal (ER+ and/or PR+; n=788), basal-like (ER-, PR-, HER2-, CK 5/6+ and/or EGFR+; n=199) or

HER2+/ER- (n=94). We excluded 'unclassified' tumors from further analysis. The major distinction between the two luminal subtypes are their proliferation signatures, measured by the expression of *CCNB1*, *MKI67*, and *MYBL2* (49). HER2 expression only identifies about 30% of luminal B tumors. In the current study, we did not have information about these proliferation markers and therefore combined Luminal A and B tumors into a single 'luminal' category (48, 49). Additionally, most other studies do not have subtype data available and only have estrogen receptor status data. Therefore, we conducted an additional exploratory analysis using estrogen receptor (ER) status to evaluate comparability to "intrinsic" subtype results. We found that ER positive effects were concordant with luminal subtype results (Table 23). There were no differences between CBCS cases with and without subtyping data in terms of age, menopausal status, or family history.

4.2.6 Statistical analysis

We calculated allele and genotype frequencies stratified by case status and self-reported race (African American or White). We assessed departure from HWE for each locus by comparing expected versus observed genotype frequencies among race-specific (White and African American) controls using exact tests ($p<0.05$). We calculated pairwise linkage disequilibrium (LD) r^2 stratified by race using SAS Genetics (version9.1.3) (SAS Institute, Cary, NC).

We used unconditional logistic regression models to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for race-stratiifed effects of bypass polymerase SNPs on breast cancer based on the additive model. Less than 2% of participants self-identified as a race other than Caucasian or African American and were not included in the final analysis. We coded each genotype as an ordinal variable (0, 1, or 2 for the number of minor alleles carried by the

individual). If the minor allele frequency (MAF) differed by race, the more common allele in Whites was used as the referent group for both populations. We also adjusted for proportion of African ancestry, as measured with a set of 144 ancestry informative markers (AIMs) (297, 306). We excluded non-polymorphic SNPs or SNPs with a minor allele frequency of less than 0.05 in either race, leaving 20 SNPs in Whites and 29 SNPs in African Americans available for analysis. Because of the high number of rare homozygote variants (18 out of 22 SNPs), homozygotes for the variant allele were combined with heterozygotes and effect estimates were reported based on the dominant model. Final models were adjusted for age at diagnosis, proportion of African ancestry and offset term for the sampling design (278).

4.2.7 Subtype analyses

We coded breast cancer subtype as a categorical variable with four levels (control, luminal, HER2+/ER-, and basal-like). We used unconditional polytomous regression to estimate ORs and 95% CI for each subtype compared to controls.

4.2.8 Correction for multiple testing

We used FDR correction for multiple testing, following the method of Benjamini and Hochberg (313). The false discovery rate is defined as "the expected proportion of errors among the rejected hypotheses" (313). Corrections were based on the number of SNPs tested and were performed separately for African American and Whites in the race-stratified analysis and separately for luminal, HER2+/ER- and basal-like categories in the subtype analysis. Observed p-values from the additive model were used to determine q-values. The q-value is defined as the minimum FDR that can be attained when calling a SNP significant (i.e., expected proportion of false positives) (314). Q-values were computed using the software package R. Statistical significance was set at $q<0.10$.

4.2.9 Pathway-based analysis

We used SKAT (SNP-set Kernel Association Test) to evaluate the combined effects of the genotyped SNPs in bypass polymerases (333). A SNP-set refers to a set of related SNPs that are grouped based on prior biological knowledge. Per our study aim, we used a SNP-set that contained bypass polymerases. We chose a linear kernel since we assumed a log linear model. Kernel regression methods convert genomic information for a pair of individuals to a kernel score representing either similarity or dissimilarity. The formation of SNP-sets harnesses the potential correlation between SNPs to increase power (328). When applied to all pairs of the individuals, this information formed a positive semi-definite matrix (332). We tested the global null hypothesis (none are related to breast cancer) for SNPs in the pathway separately for White and African American participants (333).

4.3 Results

Characteristics of the CBCS population with genotyping data are described in Table 18. The distributions of age, proportion of African ancestry, and menopausal status were similar between cases and controls. African American cases were more likely to be diagnosed at a later stage and were more likely to have tumors that were ER negative. African Americans were twice as likely to be classified as having basal-like tumors compared to Whites.

4.3.1.Genotype associations by race

Genotype distributions in race-stratified controls were all in HWE ($p < 0.05$) (Table 20). The race-stratified adjusted odds ratios for BER SNPs are summarized in Table 21. Most SNPs did not show a meaningfully increased or decreased odds ratio. However, for both race groups, 3 SNPs in *POLQ* were associated with an increased odds ratio under the dominant model ($p<0.05$).

POLQ rs487848 (AG+AA vs GG) showed a statistically significant (uncorrected) positive association with breast cancer risk in Whites (OR=1.31; 95% CI= 1.08, 1.68) and African Americans (OR=1.22; 95% CI=1.00, 1.49). *POLQ* SNP rs532411 (CT+TT vs. CC) was also significantly associated with increased breast cancer among both races (OR=1.31; 95% CI= 1.02, 1.66) and (OR=1.22; 95% CI=1.00, 1.48), respectively. Finally, *POLQ* SNP rs3218634 (CG+CC vs. GG) showed an increased risk in breast cancer in Whites (OR=1.29; 95% CI: 1.02, 1.65) and in African Americans (OR=1.20; 95% CI=0.98, 1.47). After adjustment for multiple testing, none of the SNPs remained significant at the 0.10 FDR level.

4.3.2 Genotype associations by subtype

In subtype-specific analyses, the 3 *POLQ* SNPs were significantly associated with luminal breast cancer ($p<0.05$ without FDR correction): rs487848 AG+AA vs. GG (OR=1.34, 95% CI: 1.02-1.67); rs532411 CT+TT vs. CC, (OR=1.33, 95% CI: 1.06-1.65); rs3218634 CG+CC vs. GG, (OR=1.26, 95% CI: 1.01-1.57). Additionally, another *POLQ* SNP rs1381057 (CT+TT vs CC) was significantly associated with HER+/ER- breast cancer (OR=1.44; 95% CI= 1.06, 1.93) (Table 22). The same set of *POLQ SNPs* was significantly associated with ER+ breast cancer (Table 23). However, after FDR adjustment for multiple testing, none of these SNPs remained significant (q=0.10).

4.3.3 Pathway-based analysis

We assessed the global p-value for two different SNP-sets (SNPs successfully genotyped in African Americans and SNPs successfully genotyped in Whites) using the SNP-set Kernel Association Test (SKAT), adjusted for AIMs, and offset term. We did not find any significant associations for SNP-sets. A Kernel machine test of no linear effects yielded a global p-value of 0.40 and 0.54 for African Americans and Whites, respectively (Table 24).

4.4 Discussion

Given the relatively low fidelity and high mutational potential of bypass polymerases, it was initially hypothesized that SNPs in DNA bypass polymerases may be linked to increased cancer risk. We did not find a consistent pattern of association with breast cancer risk overall or within a given subtype for most SNPs we evaluated. Subsequently, specialized bypass polymerases were shown to bypass lesions in an error-free manner (233, 246, 365). Therefore, functional redundancy in this pathway may partially explain the lack of associations between bypass polymerases and breast cancer. Indeed, lesion specificity and functional redundancy are both evolutionary tactics which may ensure that genomic integrity is maintained.(366).

Despite the weak results for most bypass polymerases, we did observe evidence for both race- and subtype -specific associations between three *POLQ* variants and an increased breast cancer risk. To our knowledge, other studies have not investigated these associations. Interestingly, all of the SNPs showing an association appeared to predict increased risk of luminal breast cancer. Although not statistically significant using the FDR these findings are suggestive and warrant replication in other studies. Within each race, these three *POLQ* SNPs were in linkage disequilibrium with each other making it difficult to identify which, if any SNPs were most likely to have functional effects. *POLQ* rs3218634 had a SIFT score of 0.01 indicative of being a damaging functional SNP (Table 19), possibly implicating the SNP as the most likely causal variant, however functional studies and fine mapping of the region is needed.

The *POLQ* gene, located at chromosome 3q, is a member of the A Family that encodes the protein polymerase theta. *POLQ* has also been implicated as playing a role in other DNA repair mechanisms such as base excision repair (BER) and crosslink repair (367, 368). *POLQ* is able to efficiently bypass oxidative DNA lesions such as abasic (AP) sites and thymine glycol *in*

vitro (246, 369-371). Another lab study showed that *POLQ* successfully extends from mismatches and bases opposite (6-4) photoproducts (246). On the other hand, POLQ-deficient mutants exhibited hypersensitivity to oxidative base damage induced by H2O2 (263). The results of the current study, together with previous experimental evidence, suggest POLQ may play an important role in breast cancer risk.

Recently, a pair of studies have linked *POLQ* overexpression in tumors to breast cancer progression and poorer prognoses (253, 261, 262). Lemee et al. examined gene expression profiles of tumors from two cohorts of European women with untreated primary breast cancer. Patients' tumor cells that overexpressed *POLQ* had a 4.3-fold increased risk of death compared those with normal expression (261). Higgins et al. also found elevated levels of *POLQ* expression in breast cancer cells, which was linked to poor prognosis in early breast cancer patients (262). While these findings emphasize the role of *POLQ* after disease onset, genes that influence progression also have been shown to influence early disease/etiology and therefore these findings also suggest that *POLQ* merits further investigation.

These findings should be considered in light of strengths and limitations of our study. Compared to other genetic association studies of breast cancer, CBCS has a larger proportion of African Americans (over 40%). In addition, CBCS has detailed subtype data on tumors from a large population-based sample of women allowing a unique investigation of the genetics of specific breast cancer subtypes as well as the ability to extend study results to the population as a whole. Stratification by subtype does reduce power for some race-specific and subtype comparisons, especially for HER2+/ER- and basal-like tumors. Future research that includes large numbers of breast cancer cases with less common subtypes and focuses on oversampling

African American cases should have improved power to more precisely estimate subtype associations, especially among African American women.

We had genotype and subtype data for a large proportion of CBCS participants. Tumor tissue was available for 1,845 of 2,311 cases (80%) and subtyping using IHC was completed for 1,424 of 2,311 cases (62%) (307). A comparison of subtyped and non-subtyped CBCS cases showed that the subtyped cases were not significantly different from the CBCS as a whole with respect to age and menopausal status. However, cases with subtype data were more likely to be African American and to have a later stage at diagnosis, which may bias estimates for SNPs related to race or disease aggressiveness (22).

It is also noteworthy that definitions for luminal breast cancer have evolved since original CBCS IHC subtyping methods were published (58). As a result, we were unable to divide luminal breast cancers into finer categories (Luminal A vs. Luminal B); current methods require use of PR positivity (on a quantitative scale) or Ki-67 to distinguish the two (295, 372). Therefore, there is heterogeneity within the group of luminal breast cancers defined here. Nonetheless, our subtyping methods here have the advantage of excluding tumors that were negative for all markers tested. Only triple negatives that were also positive for a basal-like marker are included among basal-like cancers, reducing outcome misclassification potential in this important subgroup.

Although we did not find any significant combined effects of SNPs in the TLS pathway using SKAT, use of kernel-based machine learning to assess pathway effects in breast cancer is an important advance in studying gene-gene interactions (175, 213). While our pathway analysis was limited by the density of SNP coverage across TLS pathway genes, it is important to understand gene-gene interactions in breast cancer pathways. Future application of SKAT to

similar data should consider tag-SNP approaches, which may better capture variation in candidate pathways.

4.5 Conclusions

In summary, this study adds important new information on the role of bypass polymerases in breast cancer etiology by using tumor tissue to evaluate subtype-specific effects and considers carefully selected regulatory and coding SNP-sets in a biologically established DNA repair pathway. We identified three novel SNPs in the *POLQ* gene, not previously evaluated in an epidemiologic study. With the exception of *POLQ*, we did not find any other bypass polymerase variants to be significantly associated with breast cancer risk. Larger studies such as the CBCS Phase 3 with improved power for race- and subtype-specific analyses and collaborative consortia will help gain further insight into the role of genetic variation in the DNA bypass polymerases and the risk of breast cancer.

Table 18. Characteristics of CBCS participants with genotyped data

Characteristic	Cases				Controls			
	African American		White		African American		White	
	N	%	N	%	N	%	N	%
Total (N)	742	100	1,204	100	658	100	1,089	100
Average age at selection	52		52		52		53	
Average proportion of African ancestry	0.78		0.064		0.774		0.066	
Menopausal status								
Pre-menopausal	324	41.36	539	30.53	290	59.09	456	62.67
Post-menopausal	418	58.64	665	69.47	368	40.91	633	37.33
Stage								
CIS*	88	11.90	349	29				
1	216	29.10	393	32.6				
2	299	40.30	328	27.2				
3	76	10.20	68	5.7				
4	27	3.60	15	1.3				
Missing	36	4.90	51	4.2				
Subtype								
Luminal	269	56.31	519	75.63				
HER2+/ER-	38	7.59	56	5.82				
Basal-like	108	21.98	91	10.68				
Estrogen Receptor (ER) Status								
Positive	235	49.57	482	68.91				
Negative	251	50.43	242	31.09				

**carcinoma in situ*

Table 19. List of successfully genotyped TLS variants

Gene	rs#	Type of variant	Amino Acid Change	Allele Change	SIFT score[a]
POLH	rs35675573	missense	T329I	C/T	0.01
	rs9333555	missense	M595V	A/G	0.13
	rs6899628	3'UTR	N/A	C/T	N/A
POLI	rs3730823	missense	H449R	A/G	0.52
	rs3218786	missense	F507S	C/T	0
	rs8305	missense	A706T	A/G	0.86
POLL	rs3730477	missense	R438W	C/T	0
	rs3730475	splice	N/A	C/T	N/A
	rs3730463	missense	T221P	A/C	0
POLM	rs28382653	missense	V246F	G/T	0
POLQ	rs487848	missense	A581V	A/G	0.48
	rs3218651	missense	H1201R	A/G	0.17
	rs532411	missense	A2304V	C/G	0.05
	rs1381057	missense	Q2513R	C/T	0.26
	rs3218634	missense	L2538V	C/G	0.01
	rs3218637	missense	R1953Q	A/G	0.58
	rs3218649	missense	T982R	C/G	0.61
	rs702017	missense	R66I	G/T	0.21
REV1L	rs 3087403	missense	V138M	A/G	0.09
	rs 3087386	missense	F257S	C/T	0.38
	rs3087399	missense	N373S	A/G	0.78
REV3L	rs458017	missense	Y1078C	C/T	0.22
	rs17539651	missense	P1713S	C/T	0.61

[a]Ranges from 0 to 1. The amino acid substitution is predicted damaging if the score is <=0.05, and tolerated if the score is > 0.05

Table 20. Minor Alleles Frequencies of bypass polymerase SNPs stratified by race

		Whites				African American			
Gene	SNP	Minor Allele	Cases	Controls	p HWE	Minor Allele	Cases	Controls	p HWE
POLH	rs35675573	T	0.000	0.000	N/A	T	0.012	0.014	0.707
	rs9333555	G	0.026	0.026	0.735	G	0.006	0.004	N/A
	rs6899628	T	0.039	0.039	0.278	T	0.328	0.342	0.085
POLI	rs3730823	G	0.000	0.000	N/A	G	0.005	0.007	N/A
	rs3218786	C	0.035	0.026	0.375	C	0.007	0.006	N/A
	rs8305	G	0.301	0.293	0.914	G	0.060	0.067	0.971
POLL	rs3730477	T	0.211	0.224	0.656	T	0.056	0.073	0.773
	rs3730475	C	0.284	0.288	0.601	C	0.169	0.186	0.135
	rs3730463	A	0.072	0.065	0.199	C	0.103	0.106	0.580
POLM	rs28382653	T	0.000	0.001	N/A	T	0.051	0.051	0.569
	rs28382635	T	0.000	0.000	N/A	T	0.000	0.000	N/A
	rs28382644	C	0.012	0.008	N/A	C	0.000	0.002	N/A
POLQ	rs487848	A	0.078	0.062	0.899	A	0.205	0.181	0.354
	rs3218651	G	0.167	0.163	0.073	G	0.118	0.109	0.725
	rs532411	T	0.078	0.063	0.874	T	0.204	0.181	0.354
	rs1381057	T	0.309	0.298	0.696	T	0.367	0.350	0.765
	rs3218634	C	0.078	0.063	0.849	C	0.201	0.179	0.103
	rs3218637	A	0.001	0.000	N/A	A	0.049	0.055	0.446
	rs3218649	C	0.361	0.352	0.926	C	0.454	0.441	0.447
	rs702017	G	0.000	0.000	N/A	G	0.044	0.044	0.504
REV1L	rs3087403	A	0.281	0.291	0.464	A	0.291	0.277	0.476
	rs3087386	T	0.443	0.451	0.886	T	0.287	0.290	0.550
	rs3087399	G	0.127	0.119	0.488	G	0.257	0.257	0.487
REV3L	rs458017	C	0.065	0.070	0.208	C	0.040	0.049	0.743
	rs17539651	T	0.001	0.002	N/A	T	0.113	0.119	0.226

Table 21. Assocations between bypass polymerase variants with breast cancer stratified by race

			White					African Americans				
Gene	SNP	Genotype	Cases (N)	Controls (N)	OR (95% CI)	p	q value	Cases (N)	Controls (N)	OR (95% CI)	p	q value
POLH	rs9333555	AA	1142	1034	Referent			733	653	Referent		
		AG+GG	62	55	0.94 (0.64, 1.38)	0.7398	0.8088	9	5	1.92 (0.62,6.01)	0.2601	0.6134
	rs6899628	CC	1117	1007	Referent			336	275	Referent		
		CT+ TT	87	82	0.92(0.65,1.29)	0.2778	0.716	405	383	0.96 (0.81,1.13)	0.5979	0.7972
POLI	rs3218786	TT	1120	1032	Referent			732	650	Referent		
		CT+CC	84	57	1.41 (0.98, 2.03)	0.065	0.2763	10	8	1.04 (0.39,2.76)	0.9358	0.9943
POLL	rs8305	AA	599	543	Referent			657	573	Referent		
		AG+GG	605	546	1.06 (0.93, 1.21)	0.3791	0.7161	85	85	0.84 (0.61, 1.16)	0.2876	0.6134
	rs3730477	CC	739	659	Referent			662	565	Referent		
		CT+TT	464	430	0.93 (0.80, 1.08)	0.3286	0.7161	80	93	0.79 (0.57,1.08)	0.1438	0.6134
	rs3730475	TT	613	548	Referent			511	430	Referent		
		CT+CC	591	541	0.98 (0.85, 1.12)	0.7612	0.8088	231	228	0.91 (0.74,1.12)	0.3761	0.6134
POLM	rs3730463	AA	1041	950	Referent			599	525	Referent		
		AC+CC	163	139	1.12 (0.88, 1.43)	0.3416	0.7161	143	133	1.00 (0.78, 1.28)	0.9802	0.9943
	rs28382653	GG	1204	1086	Referent			670	592	Referent		
		GT + TT	0	3	N/A	N/A		72	66	1.03 (0.73, 1.46)	0.8481	0.9943
POLQ	rs487848	GG	1024	957	Referent			469	438	Referent		
		AG+AA	180	132	1.31 (1.03, 1.68)	0.0279	0.2131	273	220	1.22 (1.00, 1.49)	0.0487	0.482
	rs3218651	AA	840	771	Referent			577	521	Referent		
		AG+GG	364	318	0.96 (0.82, 1.13)	0.6463	0.8088	165	137	1.13 (0.88,1.43)	0.3411	0.6134
	rs532411	CC	1024	956	Referent			469	438	Referent		
		CT+ TT	180	133	1.31 (1.02, 1.66)	0.0318	0.2131	272	220	1.22 (1.00,1.48)	0.0484	0.482
	rs1381057	CC	575	534	Referent			298	276	Referent		
		CT+ TT	629	555	1.02 (0.90, 1.17)	0.7355	0.8088	444	382	1.10 (0.94,1.29)	0.2464	0.6134
	rs3218634	GG	1024	955	Referent			30	15	Referent		
		CG+ CC	180	134	1.29 (1.02, 1.65)	0.0376	0.2131	712	643	1.20 (0.98,1.47)	0.0723	0.482
	rs3218637	GG	1202	1089	Referent			672	586	Referent		
		AG+AA	2	0	N/A	N/A		69	72	0.88 (0.62, 1.24)	0.4559	0.6513
	rs3218649	GG	497	457	Referent			221	201	Referent		
		CG+CC	704	632	1.01 (0.89, 1.15)	0.8453	0.8453	521	457	1.08 (0.92,1.26)	0.3358	0.6134

REV1L	rs3087403	GG	623	543	Referent			377	348	Referent		
		AG+AA	580	546	0.95 (0.83, 1.09)	0.4451	0.7424	365	310	1.09 (0.92,1.29)	0.3406	0.6134
	rs3087386	CC	385	329	Referent			383	329	Referent		
		CT+TT	819	760	0.96 (0.85, 1.09)	0.5317	0.7532	359	329	0.98 (0.83,1.17)	0.8511	0.9943
	rs3087399	AA	915	843	Referent			411	360	Referent		
		AG+GG	306	259	1.07 (0.89, 1.29)	0.4804	0.7424	331	298	1.00 (0.84, 1.19)	0.9943	0.9943
REV3L	rs458017	TT	1054	945	Referent			684	595	Referent		
		CT+CC	150	144	0.87 (0.68, 1.11)	0.2642	0.7161	58	63	0.83 (0.58, 1.20)	0.3222	0.6134
	rs17539651	CC	1201	1085	Referent			582	508	Referent		
		CT+TT	3	4	N/A	N/A		160	150	0.90 (0.70,1.15)	0.3987	0.6134

N/A indicates non-polymorphic

Table 22. Association of bypass polymerase variants with breast cancer stratified by subtype

			Luminal			HER2+/ER-			Basal-like		
Gene	SNP	Genotype	OR (95% CI)	p	q value	OR (95% CI)	p	q value	OR (95% CI)	p	q value
POLH	rs6899628	CC	Referent			Referent			Referent		
		CT+ TT	0.94 (0.74, 1.21)	0.6402	0.873	1.00 (0.57, 1.77)	0.9991	0.9991	0.80 (0.55, 1.17)	0.255	0.425
POLL	rs8305	AA	Referent			Referent			Referent		
		AG+GG	0.99 (0.81, 1.20)	0.8863	0.8863	1.22 (0.76, 1.97)	0.4115	0.8763	0.91 (0.63, 1.30)	0.5897	0.6719
	rs3730477	CC	Referent			Referent			Referent		
		CT+TT	0.98 (0.80, 1.20)	0.8344	0.8863	0.72 (0.43, 1.21)	0.2144	0.6432	0.79 (0.54, 1.15)	0.2168	0.425
	rs3730475	TT	Referent			Referent			Referent		
		CT+CC	0.97 (0.81, 1.17)	0.7731	0.8863	0.68 (0.43, 1.06)	0.867	0.9389	0.93 (0.68, 1.26)	0.6271	0.6719
POLM	rs3730463	AA	Referent			Referent			Referent		
		AC+CC	.92 (0.71, 1.18)	0.4994	0.873	0.86 (0.47, 1.58)	0.6358	0.9166	1.34 (0.92, 1.96)	0.1236	0.425
POLQ	rs487848	GG	Referent			Referent			Referent		
		AG+AA	1.34 (1.02, 1.67)	0.0096	0.144	1.21 (0.72, 2.02)	0.4753	0.8763	1.23 (0.86, 1.76)	0.2492	0.425
	rs3218651	AA	Referent			Referent			Referent		
		AG+GG	1.16 (0.65, 2.07)	0.6258	0.873	1.18 (0.28, 5.01)	0.8215	0.9289	1.10 (0.38, 3.15)	0.2492	0.425
	rs532411	CC	Referent			Referent			Referent		
		CT+ TT	1.33 (1.06, 1.65)	0.118	0.3918	1.20 (0.72, 2.02)	0.479	0.8763	1.23 (0.86, 1.76)	0.2487	0.425
	rs1381057	CC	Referent			Referent			Referent		
		CT+ TT	1.11 (0.97, 1.27)	0.1306	0.3918	1.44 (1.06, 1.93)	0.0193	0.2895	1.10 (0.88, 1.38)	0.389	0.5305
	rs3218634	GG	Referent			Referent			Referent		
		CG+ CC	1.26 (1.01, 1.57)	0.0376	0.282	1.18 (0.71, 1.98)	0.5258	0.8763	1.24 (0.87, 1.77)	0.2331	0.425
	rs3218649	GG	Referent			Referent			Referent		
		CG+CC	1.08 (0.95, 1.23)	0.2505	0.6263	1.27 (0.94, 1.71)	0.1235	0.6308	1.04 (0.84, 1.30)	0.7263	0.7263
REV1L	rs3087403	GG	Referent			Referent			Referent		
		AG+AA	1.01 (0.88, 1.16)	0.8611	0.8863	0.78 (0.55, 1.11)	0.1682	0.6308	1.22 (0.97, 1.53)	0.0931	0.425
	rs3087386	CC	Referent			Referent			Referent		
		CT+TT	0.95 (0.83, 1.08)	0.4133	0.873	1.25 (0.93, 1.69)	0.1404	0.6308	0.86 (0.68, 1.07)	0.1771	0.425
	rs3087399	AA	Referent			Referent			Referent		
		AG+GG	1.05 (0.87, 1.28)	0.6069	0.873	0.91 (0.57, 1.46)	0.7042	0.9166	0.84 (0.60, 1.17)	0.3085	0.4626
REV3L	rs458017	TT	Referent			Referent			Referent		
		CT+CC	0.76 (0.57, 1.02)	0.0651	0.3255	1.11 (0.60, 2.05)	0.7333	0.9166	0.84 (0.52, 1.38)	0.4933	0.6166

Table 23. Association of bypass polymerase variant with breast cancer stratified by ER status

			ER+			ER-		
Gene	SNP	Genotype	OR (95% CI)	p	q value	OR (95% CI)	p	q value
POLH	rs6899628	CC	Referent			Referent		
		CT+ TT	0.99 (0.80, 1.21)	0.8827	0.9225	1.03 (0.83, 1.27)	0.7925	0.8491
POLL	rs8305	AA	Referent			Referent		
		AG+GG	1.08 (0.92, 1.27)	0.3581	0.6714	1.05 (0.86, 1.28)	0.6496	0.847
	rs3730477	CC	Referent			Referent		
		CT+TT	0.96 (0.80, 1.14)	0.6103	0.8322	0.85 (0.68, 1.06)	0.1413	0.6038
	rs3730475	TT	Referent			Referent		
		CT+CC	0.98 (0.84, 1.14)	0.7862	0.9225	0.95 (0.79, 1.14)	0.566	0.847
POLM	rs3730463	AA	Referent			Referent		
		AC+CC	0.99 (0.78, 1.26)	0.9225	0.9225	1.12 (0.87, 1.46)	0.3804	0.8151
POLQ	rs487848	GG	Referent			Referent		
		AG+AA	1.30 (1.06, 1.60)	0.0113	0.1028	1.18 (0.94, 1.48)	0.1526	0.6038
	rs3218651	AA	Referent			Referent		
		AG+GG	0.87 (0.73, 1.04)	0.1307	0.2801	1.05 (0.85, 1.31)	0.6393	0.847
	rs532411	CC	Referent			Referent		
		CT+ TT	1.29 (1.05, 1.59)	0.0137	0.1028	1.18 (0.94, 1.48)	0.1537	0.6038
	rs1381057	CC	Referent			Referent		
		CT+ TT	1.14 (0.99, 1.31)	0.0638	0.2392	1.04 (0.89, 1.22)	0.5935	0.847
	rs3218634	GG	Referent			Referent		
		CG+ CC	1.26 (1.02, 1.44)	0.03	0.15	1.18 (0.94, 1.48)	0.161	0.6038
	rs3218649	GG	Referent			Referent		
		CG+CC	1.11 (0.98, 1.27)	0.1131	0.2801	1.02 (0.88, 1.19)	0.7717	0.8491
REV1L	rs3087403	GG	Referent			Referent		
		AG+AA	0.99 (0.86, 1.14)	0.8769	0.9225	1.08 (0.92, 1.27)	0.3422	0.8151
	rs3087386	CC	Referent			Referent		
		CT+TT	0.96 (0.84, 1.10)	0.57	0.8322	0.97 (0.83, 1.13)	0.6776	0.847
	rs3087399	AA	Referent			Referent		
		AG+GG	1.05 (0.88, 1.25)	0.582	0.8322	0.98 (0.81, 1.19)	0.8539	0.8539
REV3L	rs458017	TT	Referent			Referent		
		CT+CC	0.80 (0.61, 1.060	0.1184	0.2801	0.86 (0.63, 1.19)	0.3663	0.8151

Table 24. SKAT analysis of bypass polymerase SNP sets

Group	Number of SNPs	SNP-set	Global p-value
African American	20	rs35675573 rs6899628 rs8305 rs3730477 rs3730475 rs3730463 rs2838653 rs487848 rs3218651 rs532411 rs1381057 rs3218634 rs3218637 rs3218649 rs702017 rs3087403 rs3087386 rs3087399 rs458017 rs17539651	0.4027
White	16	rs933555 rs6899628 rs3218786 rs8305 rs3730477 rs3730475 rs3730463 rs487848 rs3218651 rs532411 rs1381057 rs3218634 rs3218649 rs3087403 rs3087386 rs3087399 rs458017	0.5453

Table 25. Linkage disequilibrium by race

African American

POLI	rs3218786	rs8305
rs3218786	1.00	
rs8305	0.00	1.00

POLL	rs3730477	rs3730475	rs3730463
rs3730477	1.00		
rs3730475	0.99	1.00	
rs3730463	0.00	0.54	1.00

REV1L	rs3087403	rs3087386	rs3087399
rs3087403	1.00		
rs3087386	0.16	1.00	
rs3087399	0.14	0.14	1.00

REV3L	rs458017	rs17539651
rs458017	1.00	
rs17539651	0.00	1.00

White

POLI	rs3730823	rs3218786	rs8305
rs3730823	1.00		
rs3218786	0.00	1.00	
rs8305	0.00	0.02	1.00

POLL	rs3730477	rs3730475	rs3730463
rs3730477	1.00		
rs3730475	0.69	1.00	
rs3730463	0.02	0.18	1.00

REV1L	rs3087403	rs3087386	rs3087399
rs3087403	1.00		
rs3087386	0.32	1.00	
rs3087399	0.06	0.11	1.00

REV3L	rs458017	rs17539651
rs458017	1.00	
rs17539651	0.00	1.00

Table 25 continued.

White

POLQ	rs487848	rs3218651	rs532411	rs1381057	rs3218634	rs3218637	rs3218649	rs702017
rs487848	1.00							
rs3218651	0.03	1.00						
rs532411	1.00	0.03	1.00					
rs1381057	0.43	0.07	0.43	1.00				
rs3218634	0.94	0.03	0.94	0.42	1.00			
rs3218637	0.01	0.00	0.01	0.03	0.01	1.00		
rs3218649	0.30	0.10	0.30	0.68	0.28	0.07	1.00	
rs702017	0.01	0.01	0.01	0.08	0.11	0.00	0.06	1.00

African American

POLQ	rs487848	rs3218651	rs532411	rs1381057	rs3218634	rs3218637	rs3218649	rs702017
rs487848	1.00							
rs3218651	0.01	1.00						
rs532411	1.00	0.02	1.00					
rs1381057	0.17	0.86	0.17	1.00				
rs3218634	0.99	0.02	1.00	0.17	1.00			
rs3218637	0.00	0.00	0.00	0.00	0.00	1.00		
rs3218649	0.14	0.11	0.14	0.79	0.14	0.00	1.00	
rs702017	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

CHAPTER 5. DISCUSSION

Since the discovery of mutations in *BRCA1*, a prominent DNA repair gene, in the early 1990's, researchers have been investigating other DNA repair genes in relation to breast cancer risk. Results from many of these DNA repair gene variant studies have been inconclusive; however previous CBCS studies have indicated that risk factor profiles may differ by tumor subtype and race (22, 59, 373).The purpose of this dissertation was to examine whether there are DNA repair gene subgroup SNP effects on breast cancer risk by race or breast cancer subtype. We further examined combined SNP effects using a biological pathway-based approach separately in each pathway.

We used genotype data from the Carolina Breast Cancer Study, a population-based study of White and African American women in North Carolina the majority of whom had tumor samples available for subtyping. We used unconditional logistic regression to estimate the odds ratios (ORs) and 95% confidence intervals (CIs) for the association between 31 SNPs in 15 genes in the base excision repair pathway and 22 SNPs in DNA bypass polymerase genes. We categorized race as White or African-American and classified tumors using immunochemistry as luminal (ER+ and/or PR+; n=788), basal-like (ER-, PR-, HER2-, CK 5/6+ and/or EGFR+; n=199) or HER2+/ER- (n=94).

5.1 Summary of Results

We found evidence for both race- and subtype -specific associations between BER and bypass polymerase variants and breast cancer risk. In the BER pathway, two SNPs were associated with an increased risk (*OGG1* rs1052133 and *NEIL2* rs1534862) and two *PCNA* SNPs

(rs17349 and rs17352) in high LD (r^2=0.95) were associated with an inverse association in Whites. Among African Americans, we found a *NEIL2* SNP (rs8191613) to be associated with a 28% decreased risk of breast cancer and *UNG* rs3219725 with a 44% increased risk. In the tumor subtype analysis, the *NEIL2* SNP (rs1534862) was positively associated with a moderately increased risk of luminal and HER2+/ER- breast cancer. We also found an inverse association between *FEN1* SNP (rs412334) and basal-like breast cancer.

The majority of our findings for the BER pathway were in a set of DNA glycosylase genes, *OGG1, UNG,* and *NEIL2*. These are involved in the initial recognition and response to DNA lesions. Therefore, it could be biologically plausible that a mutation in any one of these genes may lead to dysfunction in DNA replication or BER repair.

OGG1 is a glycosylase that is primarily responsible for the accurate excision of 7,8-dihydro-8-oxoguanine (8-oxoG), a product of oxidative stress, which can cause a G-T transversion during DNA replication if it not removed. *OGG1* variants have been shown to be highly mutagenic in mice and *in vitro* studies and associated with reduced DNA repair activity (219, 359). However, the evidence from epidemiologic literature has not been as consistent. The majority of studies that evaluated the role of *OGG1* rs1052133 with breast cancer risk in White and Asian populations showed inconsistent or generally null results (137, 139, 141, 143, 145, 361). In the current study, the positive association that we observed in Whites did not remain statistically significant after adjustment for multiple comparisons.

We are the first to report a statistically significant increased risk of breast cancer among African Americans for *UNG* SNP rs3219725. Mutational analyses have identified *UNG* missense variants in colorectal cancer, glioblastoma, B cell lymphoma, and esophageal squamous cell carcinoma, however there is no evidence for breast cancer. Our result for rs3219725, which is

located in the 3'UTR of the gene, represents a regulatory SNP not previously reported for breast cancer risk in African Americans. However this finding requires replication in a larger group of African Americans since it did not remain significant after adjusting for multiple comparisons.

NEIL2 is a part of a newly discovered family of monofunctional DNA glycosylases (106). Laboratory studies have shown that *NEIL2* plays an important role in the repair of oxidized bases such as pyrimidines and cytosines (109, 148). Variants in *NEIL2* have been previously associated with increased risk in colorectal, head and neck and lung cancers. One report from the Cancer Genetic Markers of Susceptibility (CGEMS) Project noted a pair of SNPs in *NEIL2* (rs8191649 and rs8191642) to be significantly associated with premenopausal breast cancer ($p<0.02$) (202).

Our analysis showed a different *NEIL2* SNP, rs1534862, was associated with increased risk of breast cancer in Whites, and also in luminal and *HER2*+/ER- subtypes. Another *NEIL2* SNP (rs8191613) was associated with a decreased risk in African Americans.

After controlling for multiple comparisons using the FDR, two BER SNPs (*PCNA* rs17352 and *NEIL2* rs1534862) remained statistically significantly associated with breast cancer. The T allele of *NEIL2* rs1534862, located in the 3'UTR of the gene, was associated with an increased risk of breast cancer in Whites and two subtypes (Luminal and HER2+/ER-). The C allele of *PCNA* rs17352, located in an intronic region, was associated with a decreased risk of breast cancer in Whites. These findings, especially for *NEIL2* are new and require further replication in larger studies.

Among bypass polymerase genes, we found evidence for both race- and subtype -specific associations between three *POLQ* variants and an increased breast cancer risk. To our knowledge, this is the first study to report these associations. Additionally, all of these SNPs

were associated with an increased risk of luminal breast cancer. Among each race, all three *POLQ* SNPs were in high LD. *POLQ* rs3218634, a missense SNP, had a SIFT score of 0.01 indicative of being a damaging functional SNP, implicating that it may be the causal variant. Although not statistically significant after adjusting for multiple comparisons, these findings are suggestive and warrant investigation in future studies.

POLQ, located at chromosome 3q, is a member of the A Family of DNA polymerases that encodes the protein polymerase theta. Recently, a pair of *in vitro* studies has linked *POLQ* overexpression in tumors to breast cancer progression and poorer prognoses (253, 261, 262). *POLQ*-deficient mutants exhibited hypersensitivity to oxidative base damage induced by hydrogen peroxide (263). The results of the current study corroborate the experimental evidence of the potential mutagenicity of *POLQ* variants.

With the exception of *POLQ*, we did not find any other bypass polymerase variants to be significantly associated with breast cancer risk. There is also evidence for functional redundancy within the oxidative DNA damage repair system. Both BER and bypass polymerases deal with DNA damage caused by oxidative stress. Functional redundancy within and between pathway genes may in part explain the lack of SNPs associations with breast cancer. Several studies have suggested that DNA bypass polymerases are involved in BER and *vice versa*. *POLQ* is one example, purportedly implicated in base excision repair and crosslink repair (367, 368). *POLQ* is able to efficiently bypass oxidative DNA lesions such as abasic (AP) sites and thymine glycol *in vitro* (246, 369-371). Another lab study showed that *POLQ* successfully extends from mismatches and bases opposite (6-4) photoproducts (246). *NEIL2* was shown to interact with *POLB* and *LIG3* in the short-patch pathway of BER (109, 110, 150). In addition, posttranslational modification of PCNA by ubiquitin may play a role in determining which DNA

response mechanism to activate. Studies showed that the mono-ubiquitylation of PCNA may allow for translesion synthesis by damage-tolerant DNA polymerases, while poly-ubiquitylation may initiate error-free pathway involving template switching in homologous recombination (HR) (228-231). Therefore, our data may reflect the fact that there is an intricate system of functionally redundant DNA damage response mechanisms in place to protect our cells from genomic instability and prevent carcinogenesis.

5.2 Strengths and Limitations

5.2.1 Study design

The Carolina Breast Cancer study is one of the first studies of its kind to integrate molecular biology and genetics with population-based epidemiology. One of the main strengths of this study was the large proportion of African Americans enrolled in the CBCS. In an unpublished review of 10 breast cancer studies, only CBCS and two other studies had an adequate proportion of African American cases to evaluate race-specific effects. Randomized recruitment, a novel sampling method, was used to oversample younger and African American women to improve power to detect associations in these often understudied subgroups of women (278). As an alternative to frequency matching, "randomized recruitment" or probability matching individually randomizes subjects to be recruited or not based on available screening variables and disease status (278). In CBCS, these screening variables, race and age, were abstracted from pathology reports for cases and DMV and Medicare records for controls. The final dataset included 1,809 White women (55%) and 1,505 African American women (45%).

In addition to reporting results stratified by self-reported race, we also used AIMs to estimate African and European ancestry to control for any residual confounding (i.e. population

stratification). Rapid case ascertainment improved access to data from North Carolina's Cancer Registry in a more time-efficient manner that allowed for more complete case ascertainment.

5.2.2 Genotyping methods

CBCS researchers had several quality controls measures in place to minimize potential genotyping errors. Blind duplicates were genotyped to verify the reproducibility of genotype calls. Any genotype with a call rate <95% was excluded. In addition, tests of Hardy-Weinberg equilibrium were conducted. Four SNPs in the BER pathway failed HWE and were excluded from subsequent analyses. Assay intensity data and genotype cluster images were reviewed for all SNPs. Across both pathways, a total of 8 SNPs were excluded due to low signal intensity or indistinguishable genotype clusters. Overall, there were 3,748 or 97% of enrolled participants (1,972 cases and 1,776 controls) with successfully genotyped data.

5.2.3 Tumor Subtyping

CBCS had detailed subtype data on tumors from a majority of cases (62%) allowing a unique investigation of the genetics of specific breast cancer subtypes. However, cases with subtype data were more likely to be African American and to have a later stage at diagnosis, which may bias estimates for SNPs related to race or disease aggressiveness (22). However, there were no significant differences for age, menopausal status, or family history between CBCS cases with and without subtyping data.

Definitions for luminal breast cancer have evolved since original CBCS IHC subtyping methods were published (58). As a result, we defined tumor subtypes differently than previous studies. The major distinction between the two luminal subtypes are their proliferation signatures, measured by the expression of *CCNB1*, *MKI67*, and *MYBL2* (49). *HER2* expression only identifies about 30% of luminal B tumors. In the current study, we did not have

information about these proliferation markers and therefore combined Luminal A and B tumors into a single 'luminal' category (48, 49). We also excluded 'unclassified' tumors from further analysis due to their heterogeneity.

Our final subtype analysis was based on three subtypes (luminal, HER2+/ER- and basal-like). Our subtyping methods have the advantage of excluding tumors that were negative for all markers tested. Only triple negatives that were also positive for a basal-like marker are included among basal-like cancers, reducing outcome misclassification potential in this important subgroup. Since many other studies do not have detailed subtype data but do have ER status, we conducted an exploratory analysis using estrogen receptor (ER) status to evaluate comparability to "intrinsic" subtype results and found that ER positive effects were for the most part concordant with luminal subtype results and ER negative effects with HER2+/ER- or basal-like subtype results.

5.2.4 SKAT analysis

Although we did not find any statistically significant combined effects of SNPs in either pathway using SKAT, to our knowledge, this is one of the few studies to have used this recent kernel-based machine learning method to assess pathway effects in cancer (175, 213). We chose this pathway-based method to harness correlation between biologically related genes. However, we recognize that our association analyses including the pathway analysis was limited by the density of SNP coverage across our two pathways and perhaps our choice of kernel. Thus SKAT may be better applied to GWAS studies with greater SNP coverage.

5.2.5 Power issues

While Phase 2 of the CBCS improved power by recruiting more invasive breast cancer along with *in situ* cases, with the exception of two SNPs in the BER pathway, other associated

SNPs did not remain significant after adjustment for multiple comparisons using the false discovery rate (FDR). Therefore, we cannot rule out the role of chance for our observed associations. Traditionally, genetic association studies have been criticized for lack of replicable results. In our BER literature review, we noticed that many studies were drastically underpowered and as a result had imprecise results (Chapter 1, Table 3). In the current study, even with over 1,900 cases and 1,700 controls, we may have been underpowered to detect small SNP effects and also had reduced power to detect subtype associations, especially for HER2+/ER- and basal-like tumors. Therefore, it is possible that our significant subtype findings for *NEIL2* rs1534862 with HER2+/ER- subtype and *FEN1* rs412334 with basal-like subtype are due to chance and need to be replicated in a larger group of women with detailed subtype classification.

5.3 Public health significance

While advances in screening and treatment have improved outcomes in breast cancer, breast cancer is still the most common (non-skin cancer) and the second most deadly cancer in U.S. women. In particular, premenopausal African American women have a disproportionate increased risk of mortality due to breast cancer compared to other subgroups of women. Therefore, this subgroup was targeted for enrollment into the study. As a result, CBCS represents one of the most comprehensive datasets of African American breast cancer cases with tumor subtype information.

CBCS was one of the first studies to report that breast cancer risk and prognostic factors may vary by both race and tumor subtype. Millikan et al. reported that risk factors for basal-like subtype included increased parity, younger age at first full term pregnancy, lack of breastfeeding, high waist-to-hip ratio, young age at menarche, and higher BMI (22). Of note, many of these risk

factors were contrary to risk factors established in luminal tumors. Furthermore, Carey et al. reported that premenopausal African Americans had a higher proportion of basal-like breast cancer (39%) than luminal cancers (36%). African American women were also two times as likely to be diagnosed with basal-like cancer compared to their White counterparts (22% vs. 11%). Additionally, Carey et al. showed that compared with luminal A tumors, basal-like tumors had poorer prognostic factors such as higher mitotic index, higher grade, and lower survival (59). These findings for basal-like breast cancer may partially explain the survival disparity in this subgroup of younger, African American women.

Since younger African-Americans carry a disproportionate burden of basal-like disease, this high risk group should also be targeted for early screening mammography and increased surveillance by their clinicians and patients themselves. Our knowledge of genetic variation in DNA repair or bypass polymerases may also inform research on potential targeted therapies. Targeted treatments that exploit DNA repair could greatly benefit women whose tumors are not responsive to traditional chemotherapy. To date, several *PARP1* and *POLB* inhibitors are in development and have been investigated as adjuvant therapies for cancer (345, 346).

5.4 Future research

We identified several potentially significant SNPs in both our race-stratified and subtype-specific analyses. Further work is needed to replicate these findings in other study populations with an adequate proportion of African Americans and with complete subtype information. The AMBER consortium, a large collaborative study of African American women with breast cancer subtype information fulfills both of these criteria (374). A similar large collaborative study of White women with subtype information is needed to replicate SNP findings specific to Whites.

In the past decade, advances in technology and statistical methods have greatly accelerated the field of genetic and molecular epidemiology. Our knowledge of DNA repair genes and pathway has also expanded since the early 2000s. Several new DNA repair and bypass polymerases genes have been discovered since the initially genotyping of the CBCS which could enhance the gene coverage in future studies (119). Our knowledge and technology in defining breast cancer heterogeneity of tumors is also evolving rapidly. Future studies should take advantage of more comprehensive marker panels (i.e. PAM 50) used for gene expression patterns of tumor subtypes (311, 375, 376).

We also suggest a using a more comprehensive set of SNPs for future SKAT analyses. In addition we would further explore if other kernels are more appropriate fit for our genetic model. We would also expand our pathway-based analysis to explore combined SNP effects between DNA pathways (i.e since the literature has suggested interactions between genes in multiple DNA repair pathways. We could also use different pathway-based statistical methods such as other machine learning methods and hierarchical modeling to evaluate multiple SNP-SNP interactions.

Finally, several environmental factors are known to be associated with DNA damage such as X-ray radiation, UV radiation, and folate deficiency, as well as yet to be discovered environmental sources of damage. Further research is needed to investigate the effect on environment interaction on the relationship between common variation in DNA damage genes and breast cancer susceptibility.

5.5 Conclusion

Since its inception two decades ago, the Carolina Breast Cancer Study has been at the forefront of many breast cancer research discoveries and innovative methods such as randomized

recruitment and rapid case ascertainment. Furthermore, the study has strived to stay relevant and has adapted to the ever-evolving changes in technology and methods. The results from this current study have added to the repository of over 100 CBCS publications as well as the breast cancer literature. We believe this research has contributed to our understanding of the relationship between genetic variation in DNA damage response genes and breast cancer risk. In the future, we hope that results from the CBCS will continue to add to our understanding of breast cancer as it has in the past.

REFERENCES

1. Narod SA, Foulkes WD. BRCA1 and BRCA2: 1994 and beyond. Nat Rev Cancer. 2004 Sep;4(9):665-76.

2. American Cancer Society. Breast cancer facts & figures 2011-2012. Atlanta: American Cancer Society, Inc.; 2012.

3. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. CA Cancer J Clin. 2012 Jan-Feb;62(1):10-29.

4. Howlader N, Noone AM, Krapcho M, Neyman N, Aminou R, Altekruse SF, Kosary CL, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). *SEER cancer statistics review, 1975-2009 (vintage 2009 populations)*, National Cancer Institute. Bethesda, MD, http://Seer.cancer.gov/csr/1975_2009_pops09/, based on November 2011 SEER data submission, posted to the SEER web site, 2012. 2012.

5. Krieger N, Chen JT, Waterman PD. Decline in US breast cancer rates after the women's health initiative: Socioeconomic and racial/ethnic differentials. Am J Public Health. 2010 Apr 1;100 Suppl 1:S132-9.

6. Beral V, Bull D, Doll R, Peto R, Reeves G, Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and abortion: Collaborative reanalysis of data from 53 epidemiological studies, including 83?000 women with breast cancer from 16 countries. Lancet. 2004 Mar 27;363(9414):1007-16.

7. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: Collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and women without the disease. Lancet. 2002;360:187-95.

8. Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. Lancet Oncol. 2012 Nov;13(11):1141-51.

9. Hamajima N, Hirose K, Tajima K, Rohan T, Calle EE, Heath CW,Jr, et al. Alcohol, tobacco and breast cancer--collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. Br J Cancer. 2002 Nov 18;87(11):1234-45.

10. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: Collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet. 2001 Oct 27;358(9291):1389-99.

11. Ahn J, Schatzkin A, Lacey JV,Jr, Albanes D, Ballard-Barbash R, Adams KF, et al. Adiposity, adult weight change, and postmenopausal breast cancer risk. Arch Intern Med. 2007 Oct 22;167(19):2091-102.

12. Morimoto LM, White E, Chen Z, Chlebowski RT, Hays J, Kuller L, et al. Obesity, body size, and risk of postmenopausal breast cancer: The Women's Health Initiative (United States). Cancer Causes Control. 2002 Oct;13(8):741-51.

13. Shin A, Matthews CE, Shu XO, Gao YT, Lu W, Gu K, et al. Joint effects of body size, energy intake, and physical activity on breast cancer risk. Breast Cancer Res Treat. 2009 Jan;113(1):153-61.

14. Slattery ML, Edwards S, Murtaugh MA, Sweeney C, Herrick J, Byers T, et al. Physical activity and breast cancer risk among women in the southwestern united states. Ann Epidemiol. 2007 May;17(5):342-53.

15. Adami HO, Lund E, Bergstrom R, Meirik O. Cigarette smoking, alcohol consumption and risk of breast cancer in young women. Br J Cancer. 1988 Dec;58(6):832-7.

16. van den Brandt PA, Spiegelman D, Yaun SS, Adami HO, Beeson L, Folsom AR, et al. Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. Am J Epidemiol. 2000 Sep 15;152(6):514-27.

17. Marcus PM, Newman B, Millikan RC, Moorman PG, Baird DD, Qaqish B. The associations of adolescent cigarette smoking, alcoholic beverage consumption, environmental tobacco smoke, and ionizing radiation with subsequent breast cancer risk (united states). Cancer Causes Control. 2000 Mar;11(3):271-8.

18. Lahmann PH, Hoffmann K, Allen N, van Gils CH, Khaw KT, Tehard B, et al. Body size and breast cancer risk: Findings from the european prospective investigation into cancer and nutrition (EPIC). Int J Cancer. 2004 Sep 20;111(5):762-71.

19. Hiatt RA, Klatsky A, Armstrong MA. Alcohol and breast cancer. Prev Med. 1988 Nov;17(6):683-5.

20. Newman LA. Breast cancer in African-American women. Oncologist. 2005 Jan;10(1):1-14.

21. Mayberry RM, Stoddard-Wright C. Breast cancer risk factors among black women and white women: Similarities and differences. Am J Epidemiol. 1992 Dec 15;136(12):1445-56.

22. Millikan RC, Newman B, Tse CK, Moorman PG, Conway K, Dressler LG, et al. Epidemiology of basal-like breast cancer. Breast Cancer Res Treat. 2008 May;109(1):123-39.

23. Colditz G, Willett WC, Hunter DJ, Stampfer MJ, Manson JE, Hennekens CH, et al. Family history, age, and risk of breast cancer. prospective data from the nurses' health study. JAMA. 1993;270(3):338-43.

24. Welsh ML, Buist DS, Aiello Bowles EJ, Anderson ML, Elmore JG, Li CI. Population-based estimates of the relation between breast cancer risk, tumor subtype, and family history. Breast Cancer Res Treat. 2009 Apr;114(3):549-58.

25. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995 Dec 21-28;378(6559):789-92.

26. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. Am J Hum Genet. 2003 May;72(5):1117-30.

27. Chen F, Chen GK, Stram DO, Millikan RC, Ambrosone CB, John EM, et al. A genome-wide association study of breast cancer in women of african ancestry. Hum Genet. 2013 Jan;132(1):39-48.

28. Turnbull C, Rahman N. Genetic predisposition to breast cancer: Past, present, and future. Annu Rev Genomics Hum Genet. 2008;9:321-45.

29. Bogdanova N, Feshchenko S, Cybulski C, Dork T. CHEK2 mutation and hereditary breast cancer. J Clin Oncol. 2007 Jul 1;25(19):e26.

30. Hellebrand H, Sutter C, Honisch E, Gross E, Wappenschmidt B, Schem C, et al. Germline mutations in the PALB2 gene are population specific and occur with low frequencies in familial breast cancer. Hum Mutat. 2011 Jun;32(6):E2176-88.

31. Wong MW, Nordfors C, Mossman D, Pecenpetelovska G, Avery-Kiejda KA, Talseth-Palmer B, et al. BRIP1, PALB2, and RAD51C mutation analysis reveals their relative importance as genetic susceptibility factors for breast cancer. Breast Cancer Res Treat. 2011 Jun;127(3):853-9.

32. Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, et al. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nat Genet. 2010 Oct;42(10):885-92.

33. Couch FJ, Gaudet MM, Antoniou AC, Ramus SJ, Kuchenbaecker KB, Soucy P, et al. Common variants at the 19p13.1 and ZNF365 loci are associated with ER subtypes of breast cancer and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers. Cancer Epidemiol Biomarkers Prev. 2012 Apr;21(4):645-57.

34. Antoniou AC, Kuchenbaecker KB, Soucy P, Beesley J, Chen X, McGuffog L, et al. Common variants at 12p11, 12q24, 9p21, 9q31.2 and in ZNF365 are associated with breast cancer risk for BRCA1 and/or BRCA2 mutation carriers. Breast Cancer Res. 2012 Feb 20;14(1):R33.

35. Rebbeck TR, Mitra N, Domchek SM, Wan F, Friebel TM, Tran TV, et al. Modification of BRCA1-associated breast and ovarian cancer risk by BRCA1-interacting genes. Cancer Res. 2011 Sep 1;71(17):5792-805.

36. Osorio A, Milne RL, Alonso R, Pita G, Peterlongo P, Teule A, et al. Evaluation of the XRCC1 gene as a phenotypic modifier in BRCA1/2 mutation carriers. results from the consortium of investigators of modifiers of BRCA1/BRCA2. Br J Cancer. 2011 Apr 12;104(8):1356-61.

37. Figueroa JD, Garcia-Closas M, Humphreys M, Platte R, Hopper JL, Southey MC, et al. Associations of common variants at 1p11.2 and 14q24.1 (RAD51L1) with breast cancer risk and heterogeneity by tumor subtype: Findings from the breast cancer association consortium. Hum Mol Genet. 2011 Dec 1;20(23):4693-706.

38. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet. 2009 May;41(5):579-84.

39. Antoniou AC, Sinilnikova OM, McGuffog L, Healey S, Nevanlinna H, Heikkinen T, et al. Common variants in LSP1, 2q35 and 8q24 and breast cancer risk for BRCA1 and BRCA2 mutation carriers. Hum Mol Genet. 2009 Nov 15;18(22):4442-56.

40. Wang X, Pankratz VS, Fredericksen Z, Tarrell R, Karaus M, McGuffog L, et al. Common variants associated with breast cancer in genome-wide association studies are modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers. Hum Mol Genet. 2010 Jul 15;19(14):2886-97.

41. U.S. Department of Energy Genome Programs. Human Genome Project http://Genomics.energy.gov. . 2013.

42. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet. 2009 May;41(5):585-90.

43. Stevens KN, Fredericksen Z, Vachon CM, Wang X, Margolin S, Lindblom A, et al. 19p13.1 is a triple-negative-specific breast cancer susceptibility locus. Cancer Res. 2012 Apr 1;72(7):1795-803.

44. Gaudet MM, Milne RL, Cox A, Camp NJ, Goode EL, Humphreys MK, et al. Five polymorphisms and breast cancer risk: Results from the breast cancer association consortium. Cancer Epidemiol Biomarkers Prev. 2009 May;18(5):1610-6.

45. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet. 2012 Jan 22;44(3):312-8.

46. Kirchhoff T, Gaudet MM, Antoniou AC, McGuffog L, Humphreys MK, Dunning AM, et al. Breast cancer risk and 6q22.33: Combined results from Breast Cancer Association Consortium and consortium of investigators on modifiers of BRCA1/2. PLoS One. 2012;7(6):e35706.

47. Lambrechts D, Truong T, Justenhoven C, Humphreys MK, Wang J, Hopper JL, et al. 11q13 is a susceptibility locus for hormone receptor positive breast cancer. Hum Mutat. 2012 Jul;33(7):1123-32.

48. Newman B, Mu H, Butler LM, Millikan RC, Moorman PG, King MC. Frequency of breast cancer attributable to BRCA1 in a population-based series of American women. JAMA. 1998 Mar 25;279(12):915-21.

49. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet. 2007 Jul;39(7):870-4.

50. Zheng W, Cai Q, Signorello LB, Long J, Hargreaves MK, Deming SL, et al. Evaluation of 11 breast cancer susceptibility loci in African-American women. Cancer Epidemiol Biomarkers Prev. 2009 Oct;18(10):2761-4.

51. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet. 2007 Jul;39(7):865-9.

52. Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol. 2007 Nov 20;25(33):5287-312.

53. Huang WY, Newman B, Millikan RC, Schell MJ, Hulka BS, Moorman PG. Hormone-related factors and risk of breast cancer in relation to estrogen receptor and progesterone receptor status. Am J Epidemiol. 2000 Apr 1;151(7):703-14.

54. Colditz GA, Rosner BA, Chen WY, Holmes MD, Hankinson SE. Risk factors for breast cancer according to estrogen and progesterone receptor status. J Natl Cancer Inst. 2004 Feb 4;96(3):218-28.

55. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci U S A. 1999 Aug 3;96(16):9212-7.

56. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747-52.

57. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001 Sep 11;98(19):10869-74.

58. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. Clin Cancer Res. 2004 Aug 15;10(16):5367-74.

59. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. JAMA. 2006 Jun 7;295(21):2492-502.

60. Phipps AI, Chlebowski RT, Prentice R, McTiernan A, Stefanick ML, Wactawski-Wende J, et al. Body size, physical activity, and risk of triple-negative and estrogen receptor-positive breast cancer. Cancer Epidemiol Biomarkers Prev. 2011 Mar;20(3):454-63.

61. Yang XR, Sherman ME, Rimm DL, Lissowska J, Brinton LA, Peplonska B, et al. Differences in risk factors for breast cancer molecular subtypes in a population-based study. Cancer Epidemiol Biomarkers Prev. 2007 Mar;16(3):439-43.

62. Trivers KF, Lund MJ, Porter PL, Liff JM, Flagg EW, Coates RJ, et al. The epidemiology of triple-negative breast cancer, including race. Cancer Causes Control. 2009 Sep;20(7):1071-82.

63. Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, et al. Associations of breast cancer risk factors with tumor subtypes: A pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst. 2011 Feb 2;103(3):250-63.

64. Gaudet MM, Press MF, Haile RW, Lynch CF, Glaser SL, Schildkraut J, et al. Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger. Breast Cancer Res Treat. 2011 Nov;130(2):587-97.

65. Stark A, Schultz D, Kapke A, Nadkarni P, Burke M, Linden M, et al. Obesity and risk of the less commonly diagnosed subtypes of breast cancer. Eur J Surg Oncol. 2009 Sep;35(9):928-35.

66. Kwan ML, Kushi LH, Weltzien E, Maring B, Kutner SE, Fulton RS, et al. Epidemiology of breast cancer subtypes in two prospective cohort studies of breast cancer survivors. Breast Cancer Res. 2009;11(3):R31.

67. Shinde SS, Forman MR, Kuerer HM, Yan K, Peintinger F, Hunt KK, et al. Higher parity and shorter breastfeeding duration: Association with triple-negative phenotype of breast cancer. Cancer. 2010 Nov 1;116(21):4933-43.

68. Tamimi RM, Colditz GA, Hazra A, Baer HJ, Hankinson SE, Rosner B, et al. Traditional breast cancer risk factors in relation to molecular subtypes of breast cancer. Breast Cancer Res Treat. 2012 Jan;131(1):159-67.

69. Phipps AI, Buist DS, Malone KE, Barlow WE, Porter PL, Kerlikowske K, et al. Family history of breast cancer in first-degree relatives and triple-negative breast cancer risk. Breast Cancer Res Treat. 2011 Apr;126(3):671-8.

70. Phillips LS, Millikan RC, Schroeder JC, Barnholtz-Sloan JS, Levine BJ. Reproductive and hormonal risk factors for ductal carcinoma in situ of the breast. Cancer Epidemiol Biomarkers Prev. 2009 May;18(5):1507-14.

71. Bauer KR, Brown M, Cress RD, Parise CA, Caggiano V. Descriptive analysis of estrogen receptor (ER)-negative, progesterone receptor (PR)-negative, and HER2-negative invasive breast cancer, the so-called triple-negative phenotype: A population-based study from the California Cancer Registry. Cancer. 2007 May 1;109(9):1721-8.

72. Livasy CA, Perou CM, Karaca G, Cowan DW, Maia D, Jackson S, et al. Identification of a basal-like subtype of breast ductal carcinoma in situ. Hum Pathol. 2007 Feb;38(2):197-204.

73. Stark A, Kapke A, Schultz D, Brown R, Linden M, Raju U. Advanced stages and poorly differentiated grade are associated with an increased risk of HER2/neu positive breast carcinoma only in white women: Findings from a prospective cohort study of African-American and White-American women. Breast Cancer Res Treat. 2008 Feb;107(3):405-14.

74. Huo D, Ikpatt F, Khramtsov A, Dangou JM, Nanda R, Dignam J, et al. Population differences in breast cancer: Survey in indigenous African women reveals over-representation of triple-negative breast cancer. J Clin Oncol. 2009 Sep 20;27(27):4515-21.

75. Stark A, Kleer CG, Martin I, Awuah B, Nsiah-Asare A, Takyi V, et al. African ancestry and higher prevalence of triple-negative breast cancer: Findings from an international study. Cancer. 2010 Nov 1;116(21):4926-32.

76. Guler G, Himmetoglu C, Jimenez RE, Geyer SM, Wang WP, Costinean S, et al. Aberrant expression of DNA damage response proteins is associated with breast cancer subtype and clinical features. Breast Cancer Res Treat. 2011 Sep;129(2):421-32.

77. Turner NC, Reis-Filho JS, Russell AM, Springall RJ, Ryder K, Steele D, et al. BRCA1 dysfunction in sporadic basal-like breast cancer. Oncogene. 2007 Mar 29;26(14):2126-32.

78. Turner NC, Reis-Filho JS. Basal-like breast cancer and the BRCA1 phenotype. Oncogene. 2006 Sep 25;25(43):5846-53.

79. Young SR, Pilarski RT, Donenberg T, Shapiro C, Hammond LS, Miller J, et al. The prevalence of BRCA1 mutations among young women with triple-negative breast cancer. BMC Cancer. 2009 Mar 19;9:86.

80. Evans DG, Howell A, Ward D, Lalloo F, Jones JL, Eccles DM. Prevalence of BRCA1 and BRCA2 mutations in triple negative breast cancer. J Med Genet. 2011 Aug;48(8):520-2.

81. Foulkes WD, Stefansson IM, Chappuis PO, Begin LR, Goffin JR, Wong N, et al. Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer. J Natl Cancer Inst. 2003 Oct 1;95(19):1482-5.

82. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A. 2003 Jul 8;100(14):8418-23.

83. Bergamaschi A, Kim YH, Wang P, Sorlie T, Hernandez-Boussard T, Lonning PE, et al. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer. 2006 Nov;45(11):1033-40.

84. Weigman VJ, Chao HH, Shabalin AA, He X, Parker JS, Nordgard SH, et al. Basal-like breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. Breast Cancer Res Treat. 2012 Jun;133(3):865-80.

85. Wang Y, Localio R, Rebbeck TR. Evaluating bias due to population stratification in case-control association studies of admixed populations. Genet Epidemiol. 2004 Jul;27(1):14-20.

86. Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010 Sep 28;107(39):16910-5.

87. Broeks A, Schmidt MK, Sherman ME, Couch FJ, Hopper JL, Dite GS, et al. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: Findings from the Breast Cancer Association Consortium. Hum Mol Genet. 2011 Aug 15;20(16):3289-303.

88. Garcia-Closas M, Chanock S. Genetic susceptibility loci for breast cancer by estrogen receptor status. Clin Cancer Res. 2008 Dec 15;14(24):8000-9.

89. Warren H, Dudbridge F, Fletcher O, Orr N, Johnson N, Hopper JL, et al. 9q31.2-rs865686 as a susceptibility locus for estrogen receptor-positive breast cancer: Evidence from the Breast Cancer Association Consortium. Cancer Epidemiol Biomarkers Prev. 2012 Aug 2.

90. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nat Genet. 2011 Oct 30;43(12):1210-4.

91. Zheng Y, Ogundiran TO, Adebamowo C, Nathanson KL, Domchek SM, Rebbeck TR, et al. Lack of association between common single nucleotide polymorphisms in the TERT-CLPTM1L locus and breast cancer in women of African ancestry. Breast Cancer Res Treat. 2012 Feb;132(1):341-5.

92. Duell EJ, Millikan RC, Pittman GS, Winkel S, Lunn RM, Tse CK, et al. Polymorphisms in the DNA repair gene XRCC1 and breast cancer. Cancer Epidemiol Biomarkers Prev. 2001 Mar;10(3):217-22.

93. Domagala P, Wokolorczyk D, Cybulski C, Huzarski T, Lubinski J, Domagala W. Different CHEK2 germline mutations are associated with distinct immunophenotypic molecular subtypes of breast cancer. Breast Cancer Res Treat. 2012 Apr;132(3):937-45.

94. Mohrenweiser HW, Wilson DM,3rd, Jones IM. 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 May 15;526(1-2):93-125.

95. Parshad R, Price FM, Bohr VA, Cowans KH, Zujewski JA, Sanford KK. Deficient DNA repair capacity, a predisposing factor in breast cancer. Br J Cancer. 1996 Jul;74(1):1-5.

96. Mohrenweiser HW, Jones IM. Variation in DNA repair is a factor in cancer susceptibility: A paradigm for the promises and perils of individual and population risk estimation? Mutat Res. 1998 May 25;400(1-2):15-24.

97. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet. 2006 Aug;38(8):873-5.

98. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet. 2007 Feb;39(2):165-7.

99. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, et al. Truncating mutations in the fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet. 2006 Nov;38(11):1239-41.

100. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet. 2002 May;31(1):33-6.

101. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med. 2008 Jun 26;358(26):2796-803.

102. Xi T, Jones IM, Mohrenweiser HW. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics. 2004 Jun;83(6):970-9.

103. Lodish H, Berk A, Matsudaira P. Molecular biology of the cell. 5th ed. New York, NY: WH Freeman; 2004.

104. Wood RD, Mitchell M, Sgouros J, Lindahl T. Human DNA repair genes. Science. 2001 Feb 16;291(5507):1284-9.

105. Kim YJ, Wilson DM,3rd. Overview of base excision repair biochemistry. Curr Mol Pharmacol. 2012 Jan;5(1):3-13.

106. Broderick P, Bagratuni T, Vijayakrishnan J, Lubbe S, Chandler I, Houlston RS. Evaluation of NTHL1, NEIL1, NEIL2, MPG, TDG, UNG and SMUG1 genes in familial colorectal cancer predisposition. BMC Cancer. 2006 Oct 9;6:243.

107. Hazra TK, Kow YW, Hatahet Z, Imhoff B, Boldogh I, Mokkapati SK, et al. Identification and characterization of a novel human DNA glycosylase for repair of cytosine-derived lesions. J Biol Chem. 2002 Aug 23;277(34):30417-20.

108. Dou H, Mitra S, Hazra TK. Repair of oxidized bases in DNA bubble structures by human DNA glycosylases NEIL1 and NEIL2. J Biol Chem. 2003 Dec 12;278(50):49679-84.

109. Das A, Wiederhold L, Leppard JB, Kedar P, Prasad R, Wang H, et al. NEIL2-initiated, APE-independent repair of oxidized bases in DNA: Evidence for a repair complex in human cells. DNA Repair (Amst). 2006 Dec 9;5(12):1439-48.

110. Wiederhold L, Leppard JB, Kedar P, Karimi-Busheri F, Rasouli-Nia A, Weinfeld M, et al. AP endonuclease-independent DNA base excision repair in human cells. Mol Cell. 2004 Jul 23;15(2):209-20.

111. Liu Y, Prasad R, Beard WA, Kedar PS, Hou EW, Shock DD, et al. Coordination of steps in single-nucleotide base excision repair mediated by apurinic/apyrimidinic endonuclease 1 and DNA polymerase beta. J Biol Chem. 2007 May 4;282(18):13532-41.

112. Fortini P, Dogliotti E. Base damage and single-strand break repair: Mechanisms and functional significance of short- and long-patch repair subpathways. DNA Repair (Amst). 2007 Apr 1;6(4):398-409.

113. Robertson AB, Klungland A, Rognes T, Leiros I. DNA repair in mammalian cells: Base excision repair: The long and short of it. Cell Mol Life Sci. 2009 Mar;66(6):981-93.

114. Wallace SS, Murphy DL, Sweasy JB. Base excision repair and cancer. Cancer Lett. 2012 Dec 31;327(1-2):73-89.

115. Boehler C, Gauthier LR, Mortusewicz O, Biard DS, Saliou JM, Bresson A, et al. Poly(ADP-ribose) polymerase 3 (PARP3), a newcomer in cellular response to DNA damage and mitotic progression. Proc Natl Acad Sci U S A. 2011 Feb 15;108(7):2783-8.

116. Fotedar R, Mossi R, Fitzgerald P, Rousselle T, Maga G, Brickner H, et al. A conserved domain of the large subunit of replication factor C binds PCNA and acts like a dominant negative inhibitor of DNA replication in mammalian cells. EMBO J. 1996 Aug 15;15(16):4423-33.

117. Maga G, Mossi R, Fischer R, Berchtold MW, Hubscher U. Phosphorylation of the PCNA binding domain of the large subunit of replication factor C by Ca2+/calmodulin-dependent protein kinase II inhibits DNA synthesis. Biochemistry. 1997 May 6;36(18):5300-10.

118. Lindahl T. Suppression of spontaneous mutagenesis in human cells by DNA base excision-repair. Mutat Res. 2000 Apr;462(2-3):129-35.

119. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. Mutat Res. 2005 Sep 4;577(1-2):275-83.

120. Moon YW, Park WS, Vortmeyer AO, Weil RJ, Lee YS, Winters TA, et al. Mutation of the uracil DNA glycosylase gene detected in glioblastoma. Mutat Res. 1998 Nov 3;421(2):191-6.

121. Nilsen H, Stamp G, Andersen S, Hrivnak G, Krokan HE, Lindahl T, et al. Gene-targeted mice lacking the ung uracil-DNA glycosylase develop B-cell lymphomas. Oncogene. 2003 Aug 21;22(35):5381-6.

122. Marian C, Tao M, Mason JB, Goerlitz DS, Nie J, Chanson A, et al. Single nucleotide polymorphisms in uracil-processing genes, intake of one-carbon nutrients and breast cancer risk. Eur J Clin Nutr. 2011 Jun;65(6):683-9.

123. Chanson A, Parnell LD, Ciappio ED, Liu Z, Crott JW, Tucker KL, et al. Polymorphisms in uracil-processing genes, but not one-carbon nutrients, are associated with altered DNA uracil concentrations in an urban puerto rican population. Am J Clin Nutr. 2009 Jun;89(6):1927-36.

124. Yamada T, Koyama T, Ohwada S, Tago K, Sakamoto I, Yoshimura S, et al. Frameshift mutations in the MBD4/MED1 gene in primary gastric cancer with high-frequency microsatellite instability. Cancer Lett. 2002 Jul 8;181(1):115-20.

125. Song JH, Maeng EJ, Cao Z, Kim SY, Nam SW, Lee JY, et al. The Glu346Lys polymorphism and frameshift mutations of the methyl-CpG binding domain 4 gene in gastrointestinal cancer. Neoplasma. 2009;56(4):343-7.

126. Hao B, Wang H, Zhou K, Li Y, Chen X, Zhou G, et al. Identification of genetic variants in base excision repair pathway and their associations with risk of esophageal squamous cell carcinoma. Cancer Res. 2004 Jun 15;64(12):4378-84.

127. Miao R, Gu H, Liu H, Hu Z, Jin G, Wang H, et al. Tagging single nucleotide polymorphisms in MBD4 are associated with risk of lung cancer in a chinese population. Lung Cancer. 2008 Dec;62(3):281-6.

128. Sohn TJ, Kim NK, An HJ, Ko JJ, Hahn TR, Oh D, et al. Gene amplification and expression of the DNA repair enzyme, N-methylpurine-DNA glycosylase (MPG) in HPV-infected cervical neoplasias. Anticancer Res. 2001 Jul-Aug;21(4A):2405-11.

129. Kim NK, An HJ, Kim HJ, Sohn TJ, Roy R, Oh D, et al. Altered expression of the DNA repair protein, N-methylpurine-DNA glycosylase (MPG) in human gonads. Anticancer Res. 2002 Mar-Apr;22(2A):793-8.

130. Kim NK, Ahn JY, Song J, Kim JK, Han JH, An HJ, et al. Expression of the DNA repair enzyme, N-methylpurine-DNA glycosylase (MPG) in astrocytic tumors. Anticancer Res. 2003 Mar-Apr;23(2B):1417-23.

131. Dallosso AR, Dolwani S, Jones N, Jones S, Colley J, Maynard J, et al. Inherited predisposition to colorectal adenomas caused by multiple rare alleles of MUTYH but not OGG1, NUDT1, NTH1 or NEIL 1, 2 or 3. Gut. 2008 Sep;57(9):1252-5.

132. Out AA, Wasielewski M, Huijts PE, van Minderhout IJ, Houwing-Duistermaat JJ, Tops CM, et al. MUTYH gene variants and breast cancer in a Dutch case-control study. Breast Cancer Res Treat. 2012 Jul;134(1):219-27.

133. Wasielewski M, Out AA, Vermeulen J, Nielsen M, van den Ouweland A, Tops CM, et al. Increased MUTYH mutation frequency among Dutch families with breast cancer and colorectal cancer. Breast Cancer Res Treat. 2010 Dec;124(3):635-41.

134. Zhu M, Chen X, Zhang H, Xiao N, Zhu C, He Q, et al. AluYb8 insertion in the MUTYH gene and risk of early-onset breast and gastric cancers in the Chinese population. Asian Pac J Cancer Prev. 2011;12(6):1451-5.

135. Krzesniak M, Butkiewicz D, Samojedny A, Chorazy M, Rusin M. Polymorphisms in TDG and MGMT genes - epidemiological and functional study in lung cancer patients from Poland. Ann Hum Genet. 2004 Jul;68(Pt 4):300-12.

136. Tani M, Shinmura K, Kohno T, Shiroishi T, Wakana S, Kim SR, et al. Genomic structure and chromosomal localization of the mouse Ogg1 gene that is involved in the repair of 8-hydroxyguanine in DNA damage. Mamm Genome. 1998 Jan;9(1):32-7.

137. Roberts MR, Shields PG, Ambrosone CB, Nie J, Marian C, Krishnan SS, et al. Single-nucleotide polymorphisms in DNA repair genes and association with breast cancer risk in the WEB study. Carcinogenesis. 2011 Aug;32(8):1223-30.

138. Sangrajrang S, Schmezer P, Burkholder I, Waas P, Boffetta P, Brennan P, et al. Polymorphisms in three base excision repair genes and breast cancer risk in Thai women. Breast Cancer Res Treat. 2008 Sep;111(2):279-88.

139. Sterpone S, Mastellone V, Padua L, Novelli F, Patrono C, Cornetta T, et al. Single-nucleotide polymorphisms in BER and HRR genes, XRCC1 haplotypes and breast cancer risk in Caucasian women. J Cancer Res Clin Oncol. 2010 Apr;136(4):631-6.

140. Rossner P,Jr, Terry MB, Gammon MD, Zhang FF, Teitelbaum SL, Eng SM, et al. OGG1 polymorphisms and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2006 Apr;15(4):811-5.

141. Vogel U, Nexo BA, Olsen A, Thomsen B, Jacobsen NR, Wallin H, et al. No association between OGG1 Ser326Cys polymorphism and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2003 Feb;12(2):170-1.

142. Choi J, Kim DY, Hyun JW, Yoon SH, Choi EM, Hahm KB, et al. Measurement of oxidative damage at individual gene levels by quantitative PCR using 8-hydroxyguanine glycosylase (OGG1). Mutat Res. 2003 Feb-Mar;523-524:225-35.

143. Zhang Y, Newcomb PA, Egan KM, Titus-Ernstoff L, Chanock S, Welch R, et al. Genetic polymorphisms in base-excision repair pathway genes and risk of breast cancer. Cancer Epidemiol Biomarkers Prev. 2006 Feb;15(2):353-8.

144. Goode EL, Ulrich CM, Potter JD. Polymorphisms in DNA repair genes and associations with cancer risk. Cancer Epidemiol Biomarkers Prev. 2002 Dec;11(12):1513-30.

145. Yuan W, Xu L, Feng Y, Yang Y, Chen W, Wang J, et al. The hOGG1 Ser326Cys polymorphism and breast cancer risk: A meta-analysis. Breast Cancer Res Treat. 2010 Aug;122(3):835-42.

146. Weiss JM, Goode EL, Ladiges WC, Ulrich CM. Polymorphic variation in hOGG1 and risk of cancer: A review of the functional and epidemiologic literature. Mol Carcinog. 2005 Mar;42(3):127-41.

147. Chen X, Wang J, Guo W, Liu X, Sun C, Cai Z, et al. Two functional variations in 5'-UTR of hoGG1 gene associated with the risk of breast cancer in chinese. Breast Cancer Res Treat. 2011 Jun;127(3):795-803.

148. Dou H, Theriot CA, Das A, Hegde ML, Matsumoto Y, Boldogh I, et al. Interaction of the human DNA glycosylase NEIL1 with proliferating cell nuclear antigen. the potential for replication-associated repair of oxidized bases in mammalian genomes. J Biol Chem. 2008 Feb 8;283(6):3130-40.

149. Maiti AK, Boldogh I, Spratt H, Mitra S, Hazra TK. Mutator phenotype of mammalian cells due to deficiency of NEIL1 DNA glycosylase, an oxidized base-specific repair enzyme. DNA Repair (Amst). 2008 Aug 2;7(8):1213-20.

150. Conlon KA, Miller H, Rosenquist TA, Zharkov DO, Berrios M. The murine DNA glycosylase NEIL2 (mNEIL2) and human DNA polymerase beta bind microtubules in situ and in vitro. DNA Repair (Amst). 2005 Apr 4;4(4):419-31.

151. Dey S, Maiti AK, Hegde ML, Hegde PM, Boldogh I, Sarkar PS, et al. Increased risk of lung cancer associated with a functionally impaired polymorphic variant of the human DNA glycosylase NEIL2. DNA Repair (Amst). 2012 Jun 1;11(6):570-8.

152. Zhai X, Zhao H, Liu Z, Wang LE, El-Naggar AK, Sturgis EM, et al. Functional variants of the NEIL1 and NEIL2 genes and risk and progression of squamous cell carcinoma of the oral cavity and oropharynx. Clin Cancer Res. 2008 Jul 1;14(13):4345-52.

153. Kinslow CJ, El-Zein RA, Hill CE, Wickliffe JK, Abdel-Rahman SZ. Single nucleotide polymorphisms 5' upstream the coding region of the NEIL2 gene influence gene transcription levels and alter levels of genetic damage. Genes Chromosomes Cancer. 2008 Nov;47(11):923-32.

154. Haiman CA, Hsu C, de Bakker PI, Frasco M, Sheng X, Van Den Berg D, et al. Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations. Hum Mol Genet. 2008 Mar 15;17(6):825-34.

155. Hadi MZ, Coleman MA, Fidelis K, Mohrenweiser HW, Wilson DM,3rd. Functional characterization of Ape1 variants identified in the human population. Nucleic Acids Res. 2000 Oct 15;28(20):3871-9.

156. Agachan B, Kucukhuseyin O, Aksoy P, Turna A, Yaylim I, Gormus U, et al. Apurinic/apyrimidinic endonuclease (APE1) gene polymorphisms and lung cancer risk in relation to tobacco smoking. Anticancer Res. 2009 Jun;29(6):2417-20.

157. Kuasne H, Rodrigues IS, Losi-Guembarovski R, Reis MB, Fuganti PE, Gregorio EP, et al. Base excision repair genes XRCC1 and APEX1 and the risk for prostate cancer. Mol Biol Rep. 2011 Mar;38(3):1585-91.

158. Popanda O, Schattenberg T, Phong CT, Butkiewicz D, Risch A, Edler L, et al. Specific combinations of DNA repair gene variants and increased risk for non-small cell lung cancer. Carcinogenesis. 2004 Dec;25(12):2433-41.

159. Canbay E, Agachan B, Gulluoglu M, Isbir T, Balik E, Yamaner S, et al. Possible associations of APE1 polymorphism with susceptibility and HOGG1 polymorphism with prognosis in gastric cancer. Anticancer Res. 2010 Apr;30(4):1359-64.

160. Zawahir Z, Dayam R, Deng J, Pereira C, Neamati N. Pharmacophore guided discovery of small-molecule human apurinic/apyrimidinic endonuclease 1 inhibitors. J Med Chem. 2009 Jan 8;52(1):20-32.

161. Smith TR, Levine EA, Freimanis RI, Akman SA, Allen GO, Hoang KN, et al. Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. Carcinogenesis. 2008 Nov;29(11):2132-8.

162. Wu B, Liu HL, Zhang S, Dong XR, Wu G. Lack of an association between two BER gene polymorphisms and breast cancer risk: A meta-analysis. PLoS One. 2012;7(12):e50857.

163. Poletto M, Di Loreto C, Marasco D, Poletto E, Puglisi F, Damante G, et al. Acetylation on critical lysine residues of apurinic/apyrimidinic endonuclease 1 (APE1) in triple negative breast cancers. Biochem Biophys Res Commun. 2012 Jul 20;424(1):34-9.

164. Lang T, Dalal S, Chikova A, DiMaio D, Sweasy JB. The E295K DNA polymerase beta gastric cancer-associated variant interferes with base excision repair and induces cellular transformation. Mol Cell Biol. 2007 Aug;27(15):5587-96.

165. Yamtich J, Nemec AA, Keh A, Sweasy JB. A germline polymorphism of DNA polymerase beta induces genomic instability and cellular transformation. PLoS Genet. 2012 Nov;8(11):e1003052.

166. Wang L, Banerjee S. Mutations in DNA-polymerase-beta occur in breast, prostate and colorectal tumors. Int J Oncol. 1995 Feb;6(2):459-63.

167. Barakat KH, Gajewski MM, Tuszynski JA. DNA polymerase beta (pol beta) inhibitors: A comprehensive overview. Drug Discov Today. 2012 Aug;17(15-16):913-20.

168. Gieseking S, Bergen K, Di Pasquale F, Diederichs K, Welte W, Marx A. Human DNA polymerase beta mutations allowing efficient abasic site bypass. J Biol Chem. 2011 Feb 4;286(5):4011-20.

169. Makridakis NM, Reichardt JK. Translesion DNA polymerases and cancer. Front Genet. 2012;3:174.

170. Starcevic D, Dalal S, Sweasy JB. Is there a link between DNA polymerase beta and cancer? Cell Cycle. 2004 Aug;3(8):998-1001.

171. Dalal S, Hile S, Eckert KA, Sun KW, Starcevic D, Sweasy JB. Prostate-cancer-associated I260M variant of DNA polymerase beta is a sequence-specific mutator. Biochemistry. 2005 Dec 6;44(48):15664-73.

172. Donigan KA, Hile SE, Eckert KA, Sweasy JB. The human gastric cancer-associated DNA polymerase beta variant D160N is a mutator that induces cellular transformation. DNA Repair (Amst). 2012 Apr 1;11(4):381-90.

173. Sweasy JB. Fidelity mechanisms of DNA polymerase beta. Prog Nucleic Acid Res Mol Biol. 2003;73:137-69.

174. Donigan KA, Sun KW, Nemec AA, Murphy DL, Cong X, Northrup V, et al. Human POLB gene is mutated in high percentage of colorectal tumors. J Biol Chem. 2012 Jul 6;287(28):23830-9.

175. Kazma R, Babron MC, Gaborieau V, Genin E, Brennan P, Hung RJ, et al. Lung cancer and DNA repair genes: Multilevel association analysis from the international lung cancer consortium. Carcinogenesis. 2012 May;33(5):1059-64.

176. Nemec AA, Donigan KA, Murphy DL, Jaeger J, Sweasy JB. Colon cancer-associated DNA polymerase beta variant induces genomic instability and cellular transformation. J Biol Chem. 2012 Jul 6;287(28):23840-9.

177. Deligezer U, Dalay N. Association of the XRCC1 gene polymorphisms with cancer risk in turkish breast cancer patients. Exp Mol Med. 2004 Dec 31;36(6):572-5.

178. Figueiredo JC, Knight JA, Briollais L, Andrulis IL, Ozcelik H. Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario site of the breast cancer family registry. Cancer Epidemiol Biomarkers Prev. 2004 Apr;13(4):583-91.

179. Moullan N, Cox DG, Angele S, Romestaing P, Gerard JP, Hall J. Polymorphisms in the DNA repair gene XRCC1, breast cancer risk, and response to radiotherapy. Cancer Epidemiol Biomarkers Prev. 2003 Nov;12(11 Pt 1):1168-74.

180. Smith TR, Levine EA, Perrier ND, Miller MS, Freimanis RI, Lohman K, et al. DNA-repair genetic polymorphisms and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2003 Nov;12(11 Pt 1):1200-4.

181. Smith TR, Miller MS, Lohman K, Lange EM, Case LD, Mohrenweiser HW, et al. Polymorphisms of XRCC1 and XRCC3 genes and susceptibility to breast cancer. Cancer Lett. 2003 Feb 20;190(2):183-90.

182. Thyagarajan B, Anderson KE, Folsom AR, Jacobs DR,Jr, Lynch CF, Bargaje A, et al. No association between XRCC1 and XRCC3 gene polymorphisms and breast cancer risk: Iowa women's health study. Cancer Detect Prev. 2006;30(4):313-21.

183. Shen J, Gammon MD, Terry MB, Wang L, Wang Q, Zhang F, et al. Polymorphisms in XRCC1 modify the association between polycyclic aromatic hydrocarbon-DNA adducts, cigarette smoking, dietary antioxidants, and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2005 Feb;14(2):336-42.

184. Costa S, Pinto D, Pereira D, Rodrigues H, Cameselle-Teijeiro J, Medeiros R, et al. DNA repair polymorphisms might contribute differentially on familial and sporadic breast cancer susceptibility: A study on a portuguese population. Breast Cancer Res Treat. 2007 Jun;103(2):209-17.

185. Ali MF, Meza JL, Rogan EG, Chakravarti D. Prevalence of BER gene polymorphisms in sporadic breast cancer. Oncol Rep. 2008 Apr;19(4):1033-8.

186. Hussien YM, Gharib AF, Awad HA, Karam RA, Elsawy WH. Impact of DNA repair genes polymorphism (XPD and XRCC1) on the risk of breast cancer in Egyptian female patients. Mol Biol Rep. 2012 Feb;39(2):1895-901.

187. Mitra AK, Singh N, Singh A, Garg VK, Agarwal A, Sharma M, et al. Association of polymorphisms in base excision repair genes with the risk of breast cancer: A case-control study in North Indian women. Oncol Res. 2008;17(3):127-35.

188. Silva SN, Moita R, Azevedo AP, Gouveia R, Manita I, Pina JE, et al. Menopausal age and XRCC1 gene polymorphisms: Role in breast cancer risk. Cancer Detect Prev. 2007;31(4):303-9.

189. Chacko P, Rajan B, Joseph T, Mathew BS, Pillai MR. Polymorphisms in DNA repair gene XRCC1 and increased genetic susceptibility to breast cancer. Breast Cancer Res Treat. 2005 Jan;89(1):15-21.

190. Silva SN, Tomar M, Paulo C, Gomes BC, Azevedo AP, Teixeira V, et al. Breast cancer risk and common single nucleotide polymorphisms in homologous recombination DNA repair pathway genes XRCC2, XRCC3, NBS1 and RAD51. Cancer Epidemiol. 2010 Feb;34(1):85-92.

191. Li H, Ha TC, Tai BC. XRCC1 gene polymorphisms and breast cancer risk in different populations: A meta-analysis. Breast. 2009 Jun;18(3):183-91.

192. Saadat M, Ansari-Lari M. Polymorphism of XRCC1 (at codon 399) and susceptibility to breast cancer, a meta-analysis of the literatures. Breast Cancer Res Treat. 2009 May;115(1):137-44.

193. Huang Y, Li L, Yu L. XRCC1 Arg399Gln, Arg194Trp and Arg280His polymorphisms in breast cancer risk: A meta-analysis. Mutagenesis. 2009 Jul;24(4):331-9.

194. Loizidou MA, Michael T, Neuhausen SL, Newbold RF, Marcou Y, Kakouri E, et al. Genetic polymorphisms in the DNA repair genes XRCC1, XRCC2 and XRCC3 and risk of breast cancer in Cyprus. Breast Cancer Res Treat. 2008 Dec;112(3):575-9.

195. Hu Z, Ma H, Chen F, Wei Q, Shen H. XRCC1 polymorphisms and cancer risk: A meta-analysis of 38 case-control studies. Cancer Epidemiol Biomarkers Prev. 2005 Jul;14(7):1810-8.

196. Puebla-Osorio N, Lacey DB, Alt FW, Zhu C. Early embryonic lethality due to targeted inactivation of DNA ligase III. Mol Cell Biol. 2006 May;26(10):3935-41.

197. Singh P, Yang M, Dai H, Yu D, Huang Q, Tan W, et al. Overexpression and hypomethylation of flap endonuclease 1 gene in breast and other cancers. Mol Cancer Res. 2008 Nov;6(11):1710-7.

198. Allinson SL, Dianova II, Dianov GL. Poly(ADP-ribose) polymerase in base excision repair: Always engaged, but not essential for DNA damage processing. Acta Biochim Pol. 2003;50(1):169-79.

199. Dantzer F, de La Rubia G, Menissier-De Murcia J, Hostomsky Z, de Murcia G, Schreiber V. Base excision repair is impaired in mammalian cells lacking poly(ADP-ribose) polymerase-1. Biochemistry. 2000 Jun 27;39(25):7559-69.

200. Bieche I, de Murcia G, Lidereau R. Poly(ADP-ribose) polymerase gene expression status and genomic instability in human breast cancer. Clin Cancer Res. 1996 Jul;2(7):1163-7.

201. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, et al. Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. N Engl J Med. 2009 Jul 9;361(2):123-34.

202. Han J, Haiman C, Niu T, Guo Q, Cox DG, Willett WC, et al. Genetic variation in DNA repair pathway genes and premenopausal breast cancer risk. Breast Cancer Res Treat. 2009 Jun;115(3):613-22.

203. Malkas LH, Herbert BS, Abdel-Aziz W, Dobrolecki LE, Liu Y, Agarwal B, et al. A cancer-associated PCNA expressed in breast cancer has implications as a potential biomarker. Proc Natl Acad Sci U S A. 2006 Dec 19;103(51):19472-7.

204. Ma X, Jin Q, Forsti A, Hemminki K, Kumar R. Single nucleotide polymorphism analyses of the human proliferating cell nuclear antigen (pCNA) and flap endonuclease (FEN1) genes. Int J Cancer. 2000 Dec 15;88(6):938-42.

205. Overmeer RM, Gourdin AM, Giglia-Mari A, Kool H, Houtsmuller AB, Siegal G, et al. Replication factor C recruits DNA polymerase delta to sites of nucleotide excision repair but is not required for PCNA recruitment. Mol Cell Biol. 2010 Oct;30(20):4828-39.

206. Dufloth RM, Costa S, Schmitt F, Zeferino LC. DNA repair gene polymorphisms and susceptibility to familial breast cancer in a group of patients from Campinas, Brazil. Genet Mol Res. 2005 Dec 30;4(4):771-82.

207. Metsola K, Kataja V, Sillanpaa P, Siivola P, Heikinheimo L, Eskelinen M, et al. XRCC1 and XPD genetic polymorphisms, smoking and breast cancer risk in a Finnish case-control study. Breast Cancer Res. 2005;7(6):R987-97.

208. Pachkowski BF, Winkel S, Kubota Y, Swenberg JA, Millikan RC, Nakamura J. XRCC1 genotype and breast cancer: Functional studies and epidemiologic data show interactions between XRCC1 codon 280 his and smoking. Cancer Res. 2006 Mar 1;66(5):2860-8.

209. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet. 2004 Jan;74(1):106-20.

210. Harlid S, Ivarsson MI, Butt S, Grzybowska E, Eyfjord JE, Lenner P, et al. Combined effect of low-penetrant SNPs on breast cancer risk. Br J Cancer. 2012 Jan 17;106(2):389-96.

211. Cheng TC, Chen ST, Huang CS, Fu YP, Yu JC, Cheng CW, et al. Breast cancer risk associated with genotype polymorphism of the catechol estrogen-metabolizing genes: A multigenic study on cancer susceptibility. Int J Cancer. 2005 Jan 20;113(3):345-53.

212. Hung RJ, Hall J, Brennan P, Boffetta P. Genetic polymorphisms in the base excision repair pathway and cancer risk: A HuGE review. Am J Epidemiol. 2005 Nov 15;162(10):925-42.

213. Monsees GM, Kraft P, Chanock SJ, Hunter DJ, Han J. Comprehensive screen of genetic variation in DNA repair pathway genes and postmenopausal breast cancer risk. Breast Cancer Res Treat. 2011 Jan;125(1):207-14.

214. Tebbs RS, Flannery ML, Meneses JJ, Hartmann A, Tucker JD, Thompson LH, et al. Requirement for the XRCC1 DNA base excision repair gene during early mouse development. Dev Biol. 1999 Apr 15;208(2):513-29.

215. Cabelof DC, Guo Z, Raffoul JJ, Sobol RW, Wilson SH, Richardson A, et al. Base excision repair deficiency caused by polymerase beta haploinsufficiency: Accelerated DNA damage and increased mutational response to carcinogens. Cancer Res. 2003 Sep 15;63(18):5799-807.

216. Sobol RW, Horton JK, Kuhn R, Gu H, Singhal RK, Prasad R, et al. Requirement of mammalian DNA polymerase-beta in base-excision repair. Nature. 1996 Jan 11;379(6561):183-6.

217. Larsen E, Gran C, Saether BE, Seeberg E, Klungland A. Proliferation failure and gamma radiation sensitivity of Fen1 null mutant mice at the blastocyst stage. Mol Cell Biol. 2003 Aug;23(15):5346-53.

218. Chan MK, Ocampo-Hafalla MT, Vartanian V, Jaruga P, Kirkali G, Koenig KL, et al. Targeted deletion of the genes encoding NTH1 and NEIL1 DNA N-glycosylases reveals the existence of novel carcinogenic oxidative damage to DNA. DNA Repair (Amst). 2009 Jul 4;8(7):786-94.

219. Xie Y, Yang H, Cunanan C, Okamoto K, Shibata D, Pan J, et al. Deficiencies in mouse myh and Ogg1 result in tumor predisposition and G to T mutations in codon 12 of the K-ras oncogene in lung tumors. Cancer Res. 2004 May 1;64(9):3096-102.

220. Cooper G. The cell: A molecular approach. 2nd ed. Sunderland, MA: Sinauer Associates; 2000.

221. Alba M. Replicative DNA polymerases. Genome Biol. 2001;2(1):REVIEWS3002.

222. Lange SS, Takata K, Wood RD. DNA polymerases and cancer. Nat Rev Cancer. 2011 Feb;11(2):96-110.

223. Chung DW, Zhang JA, Tan CK, Davie EW, So AG, Downey KM. Primary structure of the catalytic subunit of human DNA polymerase delta and chromosomal location of the gene. Proc Natl Acad Sci U S A. 1991 Dec 15;88(24):11197-201.

224. McCulloch SD, Kunkel TA. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Res. 2008 Jan;18(1):148-61.

225. Cleaver JE. Mechanisms by which human cells bypass damaged bases during DNA replication after ultraviolet irradiation. ScientificWorldJournal. 2002 May 14;2:1296-305.

226. Li X, Heyer WD. Homologous recombination in DNA repair and DNA damage tolerance. Cell Res. 2008 Jan;18(1):99-113.

227. Budzowska M, Kanaar R. Mechanisms of dealing with DNA damage-induced replication problems. Cell Biochem Biophys. 2009;53(1):17-31.

228. Jung YS, Hakem A, Hakem R, Chen X. Pirh2 E3 ubiquitin ligase monoubiquitinates DNA polymerase eta to suppress translesion DNA synthesis. Mol Cell Biol. 2011 Oct;31(19):3997-4006.

229. Guo C, Fischhaber PL, Luk-Paszyc MJ, Masuda Y, Zhou J, Kamiya K, et al. Mouse Rev1 protein interacts with multiple DNA polymerases involved in translesion DNA synthesis. EMBO J. 2003 Dec 15;22(24):6621-30.

230. Zlatanou A, Despras E, Braz-Petta T, Boubakour-Azzouz I, Pouvelle C, Stewart GS, et al. The hMsh2-hMsh6 complex acts in concert with monoubiquitinated PCNA and pol eta in response to oxidative DNA damage in human cells. Mol Cell. 2011 Aug 19;43(4):649-62.

231. Ulrich HD. Timing and spacing of ubiquitin-dependent DNA damage bypass. FEBS Lett. 2011 Sep 16;585(18):2861-7.

232. Zhang Y, Yuan F, Wu X, Wang M, Rechkoblit O, Taylor JS, et al. Error-free and error-prone lesion bypass by human DNA polymerase kappa in vitro. Nucleic Acids Res. 2000 Nov 1;28(21):4138-46.

233. Johnson RE, Haracska L, Prakash S, Prakash L. Role of DNA polymerase zeta in the bypass of a (6-4) TT photoproduct. Mol Cell Biol. 2001 May;21(10):3558-63.

234. Friedberg EC, Gerlach VL. Novel DNA polymerases offer clues to the molecular basis of mutagenesis. Cell. 1999 Aug 20;98(4):413-6.

235. Nelson JR, Lawrence CW, Hinkle DC. Deoxycytidyl transferase activity of yeast REV1 protein. Nature. 1996 Aug 22;382(6593):729-31.

236. Nelson JR, Lawrence CW, Hinkle DC. Thymine-thymine dimer bypass by yeast DNA polymerase zeta. Science. 1996 Jun 14;272(5268):1646-9.

237. Cordonnier AM, Fuchs RP. Replication of damaged DNA: Molecular defect in xeroderma pigmentosum variant cells. Mutat Res. 1999 Oct 22;435(2):111-9.

238. Lehmann AR. Replication of UV-damaged DNA: New insights into links between DNA polymerases, mutagenesis and human disease. Gene. 2000 Jul 25;253(1):1-12.

239. Beard WA, Prasad R, Wilson SH. Activities and mechanism of DNA polymerase beta. Methods Enzymol. 2006;408:91-107.

240. Goodman MF. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. Annu Rev Biochem. 2002;71:17-50.

241. Pages V, Fuchs RP. How DNA lesions are turned into mutations within cells? Oncogene. 2002 Dec 16;21(58):8957-66.

242. Marx A, Summerer D. Molecular insights into error-prone DNA replication and error-free lesion bypass. Chembiochem. 2002 May 3;3(5):405-7.

243. Livneh Z, Ziv O, Shachar S. Multiple two-polymerase mechanisms in mammalian translesion DNA synthesis. Cell Cycle. 2010 Feb 15;9(4):729-35.

244. Zhu F, Zhang M. DNA polymerase zeta: New insight into eukaryotic mutagenesis and mammalian embryonic development. World J Gastroenterol. 2003 Jun;9(6):1165-9.

245. Yuan F, Zhang Y, Rajpal DK, Wu X, Guo D, Wang M, et al. Specificity of DNA lesion bypass by the yeast DNA polymerase eta. J Biol Chem. 2000 Mar 17;275(11):8233-9.

246. Seki M, Wood RD. DNA polymerase theta (POLQ) can extend from mismatches and from bases opposite a (6-4) photoproduct. DNA Repair (Amst). 2008 Jan 1;7(1):119-27.

247. Wittschieben JP, Patil V, Glushets V, Robinson LJ, Kusewitt DF, Wood RD. Loss of DNA polymerase zeta enhances spontaneous tumorigenesis. Cancer Res. 2010 Apr 1;70(7):2770-8.

248. Masutani C, Araki M, Yamada A, Kusumoto R, Nogimori T, Maekawa T, et al. Xeroderma pigmentosum variant (XP-V) correcting protein from HeLa cells has a thymine dimer bypass DNA polymerase activity. EMBO J. 1999 Jun 15;18(12):3491-501.

249. Johnson RE, Kondratick CM, Prakash S, Prakash L. hRAD30 mutations in the variant form of xeroderma pigmentosum. Science. 1999 Jul 9;285(5425):263-5.

250. Albertella MR, Green CM, Lehmann AR, O'Connor MJ. A role for polymerase eta in the cellular tolerance to cisplatin-induced damage. Cancer Res. 2005 Nov 1;65(21):9799-806.

251. McGregor WG, Wei D, Maher VM, McCormick JJ. Abnormal, error-prone bypass of photoproducts by xeroderma pigmentosum variant cell extracts results in extreme strand bias for the kinds of mutations induced by UV light. Mol Cell Biol. 1999 Jan;19(1):147-54.

252. Glick E, White LM, Elliott NA, Berg D, Kiviat NB, Loeb LA. Mutations in DNA polymerase eta are not detected in squamous cell carcinoma of the skin. Int J Cancer. 2006 Nov 1;119(9):2225-7.

253. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006 Oct 13;314(5797):268-74.

254. Yang J, Chen Z, Liu Y, Hickey RJ, Malkas LH. Altered DNA polymerase iota expression in breast cancer cells leads to a reduction in DNA replication fidelity and a higher rate of mutagenesis. Cancer Res. 2004 Aug 15;64(16):5597-607.

255. Tissier A, Kannouche P, Reck MP, Lehmann AR, Fuchs RP, Cordonnier A. Co-localization in replication foci and interaction of human Y-family members, DNA polymerase pol eta and REVl protein. DNA Repair (Amst). 2004 Nov 2;3(11):1503-14.

256. Hashimoto K, Cho Y, Yang IY, Akagi J, Ohashi E, Tateishi S, et al. The vital role of polymerase zeta and REV1 in mutagenic, but not correct, DNA synthesis across benzo[a]pyrene-dG and recruitment of polymerase zeta by REV1 to replication-stalled site. J Biol Chem. 2012 Mar 16;287(12):9613-22.

257. Zhang Y, Wu X, Rechkoblit O, Geacintov NE, Taylor JS, Wang Z. Response of human REV1 to different DNA damage: Preferential dCMP insertion opposite the lesion. Nucleic Acids Res. 2002 Apr 1;30(7):1630-8.

258. Lawrence CW. Cellular roles of DNA polymerase zeta and Rev1 protein. DNA Repair (Amst). 2002 Jun 21;1(6):425-35.

259. Sakiyama T, Kohno T, Mimaki S, Ohta T, Yanagitani N, Sobue T, et al. Association of amino acid substitution polymorphisms in DNA repair genes TP53, POLI, REV1 and LIG4 with lung cancer risk. Int J Cancer. 2005 May 1;114(5):730-7.

260. Clark DR, Zacharias W, Panaitescu L, McGregor WG. Ribozyme-mediated REV1 inhibition reduces the frequency of UV-induced mutations in the human HPRT gene. Nucleic Acids Res. 2003 Sep 1;31(17):4981-8.

261. Lemee F, Bergoglio V, Fernandez-Vidal A, Machado-Silva A, Pillaire MJ, Bieth A, et al. DNA polymerase theta up-regulation is associated with poor survival in breast cancer, perturbs DNA replication, and promotes genetic instability. Proc Natl Acad Sci U S A. 2010 Jul 27;107(30):13390-5.

262. Higgins GS, Harris AL, Prevo R, Helleday T, McKenna WG, Buffa FM. Overexpression of POLQ confers a poor prognosis in early breast cancer patients. Oncotarget. 2010 Jul;1(3):175-84.

263. Yoshimura M, Kohzaki M, Nakamura J, Asagoshi K, Sonoda E, Hou E, et al. Vertebrate POLQ and POLbeta cooperate in base excision repair of oxidative DNA damage. Mol Cell. 2006 Oct 6;24(1):115-25.

264. Hendel A, Ziv O, Gueranger Q, Geacintov N, Livneh Z. Reduced efficiency and increased mutagenicity of translesion DNA synthesis across a TT cyclobutane pyrimidine dimer, but not a

TT 6-4 photoproduct, in human cells lacking DNA polymerase eta. DNA Repair (Amst). 2008 Oct 1;7(10):1636-46.

265. Kunz BA, Straffon AF, Vonarx EJ. DNA damage-induced mutation: Tolerance via translesion synthesis. Mutat Res. 2000 Jun 30;451(1-2):169-85.

266. Stone JE, Lujan SA, Kunkel TA, Kunkel TA. DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in saccharomyces cerevisiae. Environ Mol Mutagen. 2012 Sep 11.

267. Capp JP, Boudsocq F, Besnard AG, Lopez BS, Cazaux C, Hoffmann JS, et al. Involvement of DNA polymerase mu in the repair of a specific subset of DNA double-strand breaks in mammalian cells. Nucleic Acids Res. 2007;35(11):3551-60.

268. Lieber MR. The polymerases for V(D)J recombination. Immunity. 2006 Jul;25(1):7-9.

269. Lee JW, Blanco L, Zhou T, Garcia-Diaz M, Bebenek K, Kunkel TA, et al. Implication of DNA polymerase lambda in alignment-based gap filling for nonhomologous DNA end joining in human nuclear extracts. J Biol Chem. 2004 Jan 2;279(1):805-11.

270. Garcia-Diaz M, Dominguez O, Lopez-Fernandez LA, de Lera LT, Saniger ML, Ruiz JF, et al. DNA polymerase lambda (pol lambda), a novel eukaryotic DNA polymerase with a potential role in meiosis. J Mol Biol. 2000 Aug 25;301(4):851-67.

271. Aoufouchi S, Flatter E, Dahan A, Faili A, Bertocci B, Storck S, et al. Two novel human and mouse DNA polymerases of the polX family. Nucleic Acids Res. 2000 Sep 15;28(18):3684-93.

272. Terrados G, Capp JP, Canitrot Y, Garcia-Diaz M, Bebenek K, Kirchhoff T, et al. Characterization of a natural mutator variant of human DNA polymerase lambda which promotes chromosomal instability by compromising NHEJ. PLoS One. 2009 Oct 6;4(10):e7290.

273. Higgins MJ, Baselga J. Breast cancer in 2010: Novel targets and therapies for a personalized approach. Nat Rev Clin Oncol. 2011 Feb;8(2):65-6.

274. Millikan RC, Hummer AJ, Wolff MS, Hishida A, Begg CB. HER2 codon 655 polymorphism and breast cancer: Results from kin-cohort and case-control analyses. Breast Cancer Res Treat. 2005 Feb;89(3):309-12.

275. Newman B, Moorman PG, Millikan R, Qaqish BF, Geradts J, Aldrich TE, et al. The carolina breast cancer study: Integrating population-based epidemiology and molecular biology. Breast Cancer Res Treat. 1995 Jul;35(1):51-60.

276. North Carolina Central Cancer Registry. 2011 cancer collection and reporting manual. July 2011.

277. Aldrich TE, Vann D, Moorman PG, Newman B. Rapid reporting of cancer incidence in a population-based study of breast cancer: One constructive use of a central cancer registry. Breast Cancer Res Treat. 1995 Jul;35(1):61-4.

278. Weinberg CR, Sandler DP. Randomized recruitment in case-control studies. Am J Epidemiol. 1991 Aug 15;134(4):421-32.

279. Slattery ML, Edwards SL, Caan BJ, Kerber RA, Potter JD. Response rates among control subjects in case-control studies. Ann Epidemiol. 1995 May;5(3):245-9.

280. Moorman PG, Newman B, Millikan RC, Tse CK, Sandler DP. Participation rates in a case-control study: The impact of age, race, and race of interviewer. Ann Epidemiol. 1999 Apr;9(3):188-95.

281. Nyante SJ. Single nucleotide polymorphisms and the etiology of basal-like and luminal A breast cancer: A pathway-based approach [dissertation]. Chapel Hill, NC: University of North Carolina at Chapel Hill; 2009.

282. Millikan RC, Player JS, Decotret AR, Tse CK, Keku T. Polymorphisms in DNA repair genes, medical exposure to ionizing radiation, and breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2005 Oct;14(10):2326-34.

283. Ziegler A, Konig I. A statistical approach to genetic epidemiology. Germany: Wiley- VCH; 2006.

284. Zhu M, Zhao S. Candidate gene identification approach: Progress and challenges. Int J Biol Sci. 2007 Oct 25;3(7):420-7.

285. Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. Nat Protoc. 2007;2(10):2492-501.

286. Little J, Bradley L, Bray MS, Clyne M, Dorman J, Ellsworth DL, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. Am J Epidemiol. 2002 Aug 15;156(4):300-10.

287. Cordell HJ, Clayton DG. Genetic association studies. Lancet. 2005 Sep 24-30;366(9491):1121-31.

288. Rothman, K. J., Greenland, S., & Lash, T. L. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams, and Wilkins.; 2008.

289. Packer BR, Yeager M, Staats B, Welch R, Crenshaw A, Kiley M, et al. SNP500Cancer: A public resource for sequence validation and assay development for genetic variation in candidate genes. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D528-32.

290. Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, et al. High-throughput SNP genotyping on universal bead arrays. Mutat Res. 2005 Jun 3;573(1-2):70-82.

291. Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Wickham Garcia E, et al. Illumina universal bead arrays. Methods Enzymol. 2006;410:57-73.

292. Furberg H, Millikan R, Dressler L, Newman B, Geradts J. Tumor characteristics in African American and White women. Breast Cancer Res Treat. 2001 Jul;68(1):33-43.

293. Dressler LG, Geradts J, Burroughs M, Cowan D, Millikan RC, Newman B. Policy guidelines for the utilization of formalin-fixed, paraffin-embedded tissue sections: The UNC SPORE experience. University of North Carolina Specialized Program of Research Excellence. Breast Cancer Res Treat. 1999 Nov;58(1):31-9.

294. Ma H, Wang Y, Sullivan-Halley J, Weiss L, Burkman RT, Simon MS, et al. Breast cancer receptor status: Do results from a centralized pathology laboratory agree with SEER registry reports? Cancer Epidemiol Biomarkers Prev. 2009 Aug;18(8):2214-20.

295. Prat A, Cheang MC, Martin M, Parker JS, Carrasco E, Caballero R, et al. Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal a breast cancer. J Clin Oncol. 2013 Jan 10;31(2):203-9.

296. Livasy CA, Karaca G, Nanda R, Tretiakova MS, Olopade OI, Moore DT, et al. Phenotypic evaluation of the basal-like subtype of invasive breast carcinoma. Mod Pathol. 2006 Feb;19(2):264-71.

297. Barnholtz-Sloan JS, McEvoy B, Shriver MD, Rebbeck TR. Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. Cancer Epidemiol Biomarkers Prev. 2008 Mar;17(3):471-7.

298. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004 Apr;36(4):388-93.

299. Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, et al. Population stratification confounds genetic association studies among Latinos. Hum Genet. 2006 Jan;118(5):652-64.

300. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet. 1999 Jul;65(1):220-8.

301. Barnholtz-Sloan JS, Chakraborty R, Sellers TA, Schwartz AG. Examining population stratification via individual ancestry estimates versus self-reported race. Cancer Epidemiol Biomarkers Prev. 2005 Jun;14(6):1545-51.

302. Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor Popul Biol. 2001 Nov;60(3):227-37.

303. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999 Dec;55(4):997-1004.

304. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet. 2005 Apr;37(4):413-7.

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

306. Pfaff CL, Barnholtz-Sloan J, Wagner JK, Long JC. Information on ancestry from genetic markers. Genet Epidemiol. 2004 May;26(4):305-15.

307. Nyante SJ, Gammon MD, Kaufman JS, Bensen JT, Lin DY, Barnholtz-Sloan JS, et al. Common genetic variation in adiponectin, leptin, and leptin receptor and association with breast cancer subtypes. Breast Cancer Res Treat. 2011 Sep;129(2):593-606.

308. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005 May;76(5):887-93.

309. Lewis CM. Genetic association studies: Design, analysis and interpretation. Brief Bioinform. 2002 Jun;3(2):146-53.

310. Leong TY, Leong AS. Controversies in the assessment of HER-2: More questions than answers. Adv Anat Pathol. 2006 Sep;13(5):263-9.

311. Bastien RR, Rodriguez-Lescure A, Ebbert MT, Prat A, Munarriz B, Rowe L, et al. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. BMC Med Genomics. 2012 Oct 4;5:44,8794-5-44.

312. Rice TK, Schork NJ, Rao DC. Methods for handling multiple testing. Adv Genet. 2008;60:293-308.

313. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. JSTOR. 1995;57(1):289.

314. Storey JD, Tibshirani R. Statistical methods for identifying differentially expressed genes in DNA microarrays. Methods Mol Biol. 2003;224:149-57.

315. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002 Jun 21;296(5576):2225-9.

316. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009 Jun;10(6):392-404.

317. McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions: A review. Appl Bioinformatics. 2006;5(2):77-88.

318. Brassat D, Motsinger AA, Caillier SJ, Erlich HA, Walker K, Steiner LL, et al. Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in african americans. Genes Immun. 2006 Jun;7(4):310-5.

319. Hahn LW, Ritchie MD, Moore JH. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics. 2003 Feb 12;19(3):376-82.

320. Motsinger AA, Ritchie MD. Multifactor dimensionality reduction: An analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. Hum Genomics. 2006 Mar;2(5):318-28.

321. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001 Jul;69(1):138-47.

322. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol. 2003 Feb;24(2):150-7.

323. Zhang H, Bonney G. Use of classification trees for association studies. Genet Epidemiol. 2000 Dec;19(4):323-32.

324. Breiman L. Random forests. Mach Learn. 2001;45:5.

325. Chen GK, Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. Am J Hum Genet. 2007 Aug;81(2):397-404.

326. Hung RJ, Brennan P, Malaveille C, Porru S, Donato F, Boffetta P, et al. Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer. Cancer Epidemiol Biomarkers Prev. 2004 Jun;13(6):1013-21.

327. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004 Apr;74(4):765-9.

328. Li H. U-statistics in genetic association studies. Hum Genet. 2012 May 20.

329. Stram DO, Leigh Pearce C, Bretsky P, Freedman M, Hirschhorn JN, Altshuler D, et al. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum Hered. 2003;55(4):179-90.

330. Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet. 2005 May;76(5):780-93.

331. Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet. 2006 Nov;79(5):792-806.

332. Schaid DJ. Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. Hum Hered. 2010 Jul 3;70(2):109-31.

333. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010 Jun 11;86(6):929-42.

334. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics. 2008 Jun 24;9:292,2105-9-292.

335. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet. 2008 Feb;82(2):386-97.

336. Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. Genet Epidemiol. 2010 Apr;34(3):213-21.

337. Gauderman W, Morrison S. QUANTO documentation (technical report no. 157) Los Angeles, CA: Department of preventive medicine, University of Southern California. 2001.

338. Kerlikowske K, Barclay J, Grady D, Sickles EA, Ernster V. Comparison of risk factors for ductal carcinoma in situ and invasive breast cancer. J Natl Cancer Inst. 1997 Jan 1;89(1):76-82.

339. Lambe M, Hsieh CC, Tsaih SW, Ekbom A, Trichopoulos D, Adami HO. Parity, age at first birth and the risk of carcinoma in situ of the breast. Int J Cancer. 1998 Jul 29;77(3):330-2.

340. Cheang MC, Voduc D, Bajdik C, Leung S, McKinney S, Chia SK, et al. Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype. Clin Cancer Res. 2008 Mar 1;14(5):1368-76.

341. Carey L, Winer E, Viale G, Cameron D, Gianni L. Triple-negative breast cancer: Disease entity or title of convenience? Nat Rev Clin Oncol. 2010 Dec;7(12):683-92.

342. Greene CS, Penrod NM, Williams SM, Moore JH. Failure to replicate a genetic association may provide important clues about genetic architecture. PLoS One. 2009 Jun 2;4(6):e5639.

343. Chen F, Chen GK, Millikan RC, John EM, Ambrosone CB, Bernstein L, et al. Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. Hum Mol Genet. 2011 Nov 15;20(22):4491-503.

344. Mechanic LE, Millikan RC, Player J, de Cotret AR, Winkel S, Worley K, 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 Jul;27(7):1377-85.

345. Barakat K, Gajewski M, Tuszynski J. DNA repair inhibitors: Our last disposal to improve cancer therapy. Curr Top Med Chem. 2012 Jun 7.

346. Bryant HE, Schultz N, Thomas HD, Parker KM, Flower D, Lopez E, et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature. 2005 Apr 14;434(7035):913-7.

347. Mullan PB, Millikan RC. Molecular subtyping of breast cancer: Opportunities for new therapeutic approaches. Cell Mol Life Sci. 2007 Dec;64(24):3219-32.

348. Wu J, Lu LY, Yu X. The role of BRCA1 in DNA damage response. Protein Cell. 2010 Feb;1(2):117-23.

349. Rebbeck TR. Inherited genetic predisposition in breast cancer. A population-based perspective. Cancer. 1999 Dec 1;86(11 Suppl):2493-501.

350. Brewster AM, Jorgensen TJ, Ruczinski I, Huang HY, Hoffman S, Thuita L, et al. Polymorphisms of the DNA repair genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): Relationship to breast cancer risk and familial predisposition to breast cancer. Breast Cancer Res Treat. 2006 Jan;95(1):73-80.

351. Forsti A, Angelini S, Festa F, Sanyal S, Zhang Z, Grzybowska E, et al. Single nucleotide polymorphisms in breast cancer. Oncol Rep. 2004 Apr;11(4):917-22.

352. Zhai X, Liu J, Hu Z, Wang S, Qing J, Wang X, et al. Polymorphisms of ADPRT Val762Ala and XRCC1 Arg399Glu and risk of breast cancer in Chinese women: A case control analysis. Oncol Rep. 2006 Jan;15(1):247-52.

353. Shu XO, Cai Q, Gao YT, Wen W, Jin F, Zheng W. A population-based case-control study of the Arg399Gln polymorphism in DNA repair gene XRCC1 and risk of breast cancer. Cancer Epidemiol Biomarkers Prev. 2003 Dec;12(12):1462-7.

354. Kim SU, Park SK, Yoo KY, Yoon KS, Choi JY, Seo JS, et al. XRCC1 genetic polymorphism and breast cancer risk. Pharmacogenetics. 2002 Jun;12(4):335-8.

355. Patel AV, Calle EE, Pavluck AL, Feigelson HS, Thun MJ, Rodriguez C. A prospective study of XRCC1 (X-ray cross-complementing group 1) polymorphisms and breast cancer risk. Breast Cancer Res. 2005;7(6):R1168-73.

356. Millikan R, Eaton A, Worley K, Biscocho L, Hodgson E, Huang WY, et al. HER2 codon 655 polymorphism and risk of breast cancer in African Americans and Whites. Breast Cancer Res Treat. 2003 Jun;79(3):355-64.

357. Ma H, Wang Y, Sullivan-Halley J, Weiss L, Marchbanks PA, Spirtas R, et al. Use of four biomarkers to evaluate the risk of breast cancer subtypes in the Women's Contraceptive and Reproductive Experiences study. Cancer Res. 2010 Jan 15;70(2):575-87.

358. Shinmura K, Yokota J. The OGG1 gene encodes a repair enzyme for oxidatively damaged DNA and is involved in human carcinogenesis. Antioxid Redox Signal. 2001 Aug;3(4):597-609.

359. Boiteux S, Radicella JP. The human OGG1 gene: Structure, functions, and its implication in the process of carcinogenesis. Arch Biochem Biophys. 2000 May 1;377(1):1-8.

360. Vodicka P, Stetina R, Polakova V, Tulupova E, Naccarati A, Vodickova L, et al. Association of DNA repair polymorphisms with DNA repair functional outcomes in healthy human subjects. Carcinogenesis. 2007 Mar;28(3):657-64.

361. Gu D, Wang M, Zhang Z, Chen J. Lack of association between the hOGG1 Ser326Cys polymorphism and breast cancer risk: Evidence from 11 case-control studies. Breast Cancer Res Treat. 2010 Jul;122(2):527-31.

362. Visnes T, Akbari M, Hagen L, Slupphaug G, Krokan HE. The rate of base excision repair of uracil is controlled by the initiating glycosylase. DNA Repair (Amst). 2008 Nov 1;7(11):1869-81.

363. Hoeijmakers JH. Genome maintenance mechanisms for preventing cancer. Nature. 2001 May 17;411(6835):366-74.

364. Mohrenweiser HW, Xi T, Vazquez-Matias J, Jones IM. Identification of 127 amino acid substitution variants in screening 37 DNA repair genes in humans. Cancer Epidemiol Biomarkers Prev. 2002 Oct;11(10 Pt 1):1054-64.

365. Choi JY, Lim S, Kim EJ, Jo A, Guengerich FP. Translesion synthesis across abasic lesions by human B-family and Y-family DNA polymerases alpha, delta, eta, iota, kappa, and REV1. J Mol Biol. 2010 Nov 19;404(1):34-44.

366. Wang Z. DNA damage-induced mutagenesis : A novel target for cancer prevention. Mol Interv. 2001 Dec;1(5):269-81.

367. Shcherbakova PV, Bebenek K, Kunkel TA. Functions of eukaryotic DNA polymerases. Sci Aging Knowledge Environ. 2003 Feb 26;2003(8):RE3.

368. Muzzini DM, Plevani P, Boulton SJ, Cassata G, Marini F. Caenorhabditis elegans POLQ-1 and HEL-308 function in two distinct DNA interstrand cross-link repair pathways. DNA Repair (Amst). 2008 Jun 1;7(6):941-50.

369. Hogg M, Seki M, Wood RD, Doublie S, Wallace SS. Lesion bypass activity of DNA polymerase theta (POLQ) is an intrinsic property of the pol domain and depends on unique sequence inserts. J Mol Biol. 2011 Jan 21;405(3):642-52.

370. Yoon JH, Bhatia G, Prakash S, Prakash L. Error-free replicative bypass of thymine glycol by the combined action of DNA polymerases kappa and zeta in human cells. Proc Natl Acad Sci U S A. 2010 Aug 10;107(32):14116-21.

371. Prasad R, Longley MJ, Sharief FS, Hou EW, Copeland WC, Wilson SH. Human DNA polymerase theta possesses 5'-dRP lyase activity and functions in single-nucleotide base excision repair in vitro. Nucleic Acids Res. 2009 Apr;37(6):1868-77.

372. Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. J Natl Cancer Inst. 2009 May 20;101(10):736-50.

373. Hall IJ, Moorman PG, Millikan RC, Newman B. Comparative analysis of breast cancer risk factors among African-American women and White women. Am J Epidemiol. 2005 Jan 1;161(1):40-51.

374. Palmer JR, Ambrosone CB, Olshan AF. A collaborative study of the etiology of breast cancer subtypes in African American women: The AMBER consortium. Cancer Causes Control. 2014 Mar;25(3):309-19.

375. Chia SK, Bramwell VH, Tu D, Shepherd LE, Jiang S, Vickery T, et al. A 50-gene intrinsic subtype classifier for prognosis and prediction of benefit from adjuvant tamoxifen. Clin Cancer Res. 2012 Aug 15;18(16):4465-72.

376. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct 4;490(7418):61-70.