University of Alberta

Germline DNA variants as determinants for breast cancer predisposition and prognosis

by

Yadav Sapkota

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Medical Sciences – Laboratory Medicine and Pathology

Fall 2013
Edmonton, Alberta

Dedication

To my parents and family

Abstract

Breast cancer is the most common cancer among women in the developed world. The disease results from the combined effects of genetic, environmental, reproductive and lifestyle risk factors. Germline DNA variations identified thus far by linkage and genome-wide association studies (GWASs) account for less than one-third of variability in breast cancer predisposition, suggesting that more variants exist. Furthermore, despite advancements in breast cancer therapies guided by tumor-based prognostic and predictive factors, approximately 30% of breast cancer patients who receive standard guideline-based therapies experience disease recurrence within ten years post diagnosis. Consequently, there is a clear need of additional markers for disease risk assessment as well as markers of potential prognostic values to better guide treatment modalities. In this thesis, I adopted a comprehensive approach utilizing single nucleotide polymorphisms (SNPs) and germline copy number aberrations (copy number variations (CNVs) and copy neutral-loss of heterozygosities (CN-LOHs)) to identify markers for breast cancer susceptibility and disease prognosis. I used a multi-stage association study design that included cumulative sample sizes of 2,795 invasive breast cancer cases and 4,505 healthy controls of predominantly Caucasian in origin selected from Alberta, Canada. I identified a novel breast cancer susceptibility locus on chromosome 4q31.22 showing a strong statistical significance for overall breast cancer (per allele odds ratio=1.28 and P=1.5 x 10^{-7}), adjusted for body mass index (BMI). I also independently confirmed one literature reported association on chromosome 8q24.21-rs13281615 (BMI adjusted-P<3.1 x 10^{-3})

with breast cancer prognosis. Since epistatic interactions have been hypothesized to capture additional heritability for breast cancer, I extended my studies and identified interactions involving two SNPs and an interaction involving four SNPs. These interactions were from the single-locus effects with weak statistical significance in GWAS and/or candidate-gene studies. Finally, I identified germline CNAs as potential prognostic markers for the predominant luminal A breast cancers (up to 70% of total cases diagnosed), which recur despite the good prognosis. Germline DNA-based markers for disease predisposition and prognosis is an area in its infancy and clearly more work is warranted to substantiate and extend the reported findings to enable eventual translation of research to clinical applications.

Acknowledgements

I would like to thank my supervisor, Dr. Sambasivarao Damaraju, for his support, patience and guidance through my graduate work. I am also thankful to my supervisory committee members: Professors Raymond Lai, Carol E Cass, John R Mackey and Yutaka Yasui for valuable feedback and constructive suggestions throughout my graduate program.

I would like to thank Drs. Sunita Ghosh and Bradley P Coe for useful discussions on various statistical and bioinformatics aspects pertinent to my graduate work.

I also thank Jennifer Dufour, Diana Carandang and Lillian Cook for technical assistance and sample preparation for genotyping experiments. Thanks to Adrian Driga for providing needed information on the clinicopathological characteristics of the cases and controls used in the study.

My sincere thanks to the Canadian Breast Cancer Foundation (CBCF) Tumor Bank and the Tomorrow Project for providing the samples needed for my work and I am extremely indebted to all the participants of the study. Many thanks to the CBCF Tumor Bank and the Tomorrow Project team members for providing technical assistance in sample handling. I would like to acknowledge the Queen Elizabeth II Doctoral Scholarship (University of Alberta) and the CBCF for graduate stipend support.

I would also like to acknowledge Dr. Monika Keelan and Dr. Russell Griener who served as external examiners for my candidacy examination. I am thankful to

Cheryl Titus for helping to accomplish administrative and other necessary tasks during my graduate program.

I would also like to acknowledge members of the Damaraju laboratory Marc St. George, Mohsen Hajiloo, Preethi Krishnan and Ashok Narasimhan for wonderful time and useful discussions.

I am thankful to Dr. Paula J Robson, Dr. Sasha Lupichuk, Dr. Karen Kopciuk, Mr Conrado Franco-Villalobos and Ms Malinee Sridharan for providing valuable edits and comments for the papers that are part of my thesis.

Finally, I thank my family: my parents, sisters and my brother. Their continuous love and support was always there when I needed them. And, of course I am fortunate to have Suyena as my life partner without her love, patience and support; I would never have completed my graduate program. I love you!

Table of Contents

List of Tables

List of Figures

List of Abbreviations

APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1
ATM	Ataxia telangiectasia mutated
BAD	*BCL2*-associated agonist of cell death
BAF	B-allele frequency
BCAC	Breast Cancer Association Consortium
BCR	Breast cancer recurrence
BLM	Bloom syndrome, RecQ helicase-like
BMI	Body mass index
BRCA1	Breast cancer 1, early onset
BRCA2	Breast cancer 2, early onset
BRIP1	*BRCA1*-interacting protein 1
CASP8	Caspase 8
CBCF	Canadian Breast Cancer Foundation
CCNG1	Cyclin G1
CCNO	Cyclin O
CDCC88B	Coiled-coil containing 88B
CDCV	Common disease-common variant
CEU	Central Europeans
CGEMS	National Cancer Institute Cancer Genetics Markers of Susceptibility
CGH	Comparative genomic hybridization
CHEK2	Checkpoint kinase 2

CI	Confidence interval
CLEC18A	C-type lectin fomain family 18, member A
CNA	Copy number aberration
CN-LOH	Copy neutral-loss of heterozygosity
CNV	Copy number variation
COL1A1	Collagen, type I, alpha 1
CQC	Contrast quality control
d.f.	Degree of freedom
DGV	Database of Genomic Variants
DHX15	DEAH (Asp-Glu-Ala-His) box polypeptide 15
EDNRA	Endothein receptor type A
ER	Estrogen receptor
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)
ERCC4	Excision repair cross-complementing rodent repair deficiency, complementation group 4
ERCC5	Excision repair cross-complementing rodent repair deficiency, complementation group 5
ERCC6	Excision repair cross-complementing rodent repair deficiency, complementation group 6
ESRRA	Estrogen-related receptor alpha
FDR	False discovery rate
FGFR2	Fibroblast growth factor receptor 2

FISH	Fluorescence in situ hybridization
GPR137	G protein-coupled receptor 137
GTF2A2	General transcription factor IIA, 12 kDa
GWAS	Genome-wide association study
HER2	Human epidermal growth factor recetor 2
HR	Hazaards ratio
HWE	Hardy-Weinberg Equilibrium
KCNK4	Potassium channel, subfamily K, member 4
KIF3B	Kinesin family member 3B
LD	Linkage disequilibrium
LIG4	Ligase IV, DNA, ATP-dependent
LOD	Logarithm (base 10) of odds
LOH	Loss of heterozygosity
LRRC37B	Leucine rich repeat containing 37B
LSP1	Lymphocyte-specific protein 1
MAF	Minor allele frequency
MAP3K1	Mitogen-activated protein kinase kinase kinase 1, E3 ubiquitin protein ligase
MBD2	Methyl-CpG binding domain protein 2
MBD3	Methyl-CpG binding domain protein 3
MBD5	Methyl-CpG binding domain protein 5
MDM2	Mdm2, p53 E3 ubiquitin protein ligase homolog (mouse)
MGMT	O-6-methylguanine-DNA methyltransferase

MIR3668	MicroRNA 3668
MIR4465	MicroRNA 4465
MLH1	MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
MSH3	MutS homolog 3 (E. coli)
MYH15	Myosin, heavy chain 15
NCBI	National Center for Biotechnology Information
NHS	Nurses' Health Study
OR	Odds ratio
OS	Overall survival
PALB2	Partner and localizer of *BRCA2*
PARP1	Poly (ADP-ribose) polymerase 1
PDPR	Pyruvate dehydrogenase phosphatase regulatory subunit
POLB	Polymerase (DNA directed) beta
PPFIBP2	PTPRF interacting protein, binding protein (liprin beta 2)
PR	Progesterone recepter
PRS	Polygenic risk score
PTEN	Phosphatase and tensin homolog
PXN	Paxillin
QC	Quality control
RAD21	RAD21 homolog (S. pombe)
REMARK	Recommendations for Tumor Marker Prognostic Studies
RFS	Recurrence-free survival

ROP1NL	Rhophilin associated tail protein 1-like
RPAP1	RNA polymerase II associated protein 1
RT-qPCR	Real-time quantitative polymerase chain reaction
SH3GL1P1	SH3-domain GRB2-like 1 pseudogene 1
SLC22A11	Solute carrier family 22 (organic anion/urate transporter)
SLC4A7	Solute carrier family 4, sodium bicarbonate co-transporter, member 7
SNP	Single nucleotide poymorphism
SNP-FASST2	SNP-Fast Adaptive States Segmentation Technique 2
SUZ12	Suppressor of zeste 12 homolog (*Drosophila*)
TNFAIP8	Tumor necrosis factor, alpha-induced protein 8
TNRC9	Trinucleotide repeat-containing 9
TP53	Tumor protein 53
TP63	Tumor protein 63
TPRG1	Tumor protein p63 regulated 1
TPRG1L	Tumor protein p63 regulated 1-like
UPD	Uniparental disomy
VEGFC	Vascular endothelial growth factor C
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1
XRCC3	X-ray repair complementing defective repair in Chinese hamster cells 3
ZNF365	Zinc finger protein 365

ZNF577	Zinc finger protein 577

1. Introduction

1.1 Breast cancer: Epidemiology

Breast cancer is by far the most common cancer among women in the developed world. Worldwide in 2008, over 1.38 million women were diagnosed with breast cancer and more than 458,000 women died from this disease [1]. Likewise, in the European Union, more than 332,000 new breast cancer cases and approximately 89,800 deaths from breast cancer were estimated in 2008. In the United Kingdom in 2010, more than 49,500 women were diagnosed with breast cancer while 11,600 women died from it. Similar statistics were observed in the United States in 2012, with more than 200,000 new cases and 39,000 deaths due to breast cancer [2]. In Canada, approximately 22,700 new breast cancer cases and 5,100 breast cancer related deaths were estimated in 2012 [3].

While the age-standardized incidence rate for breast cancer has increased with the introduction of screening measures, breast cancer survival rates have been improving for the last twenty years. The current five-year survival rate for breast cancer in England is approximately 85% [1]. The five-year survival rate for Canadian women diagnosed with breast cancer during 2004-2006 is estimated as 88% [3]. Similarly, the five-year relative survival rate for breast cancer in the United States is approximately 89% however, caution should be exercised while interpreting these data as they are based on past treatment responses and do not reflect recent advances in breast cancer therapies [2]. Consequently, identification of high-risk populations to initiate preventive and prophylactic measures and interventions are needed to realize the aim of cancer prevention and control.

Current clinical practice for early-stage breast cancer primarily involves excision of localized tumor, followed by adjuvant systemic chemotherapies, endocrine therapies, and/or radiotherapies to eliminate residual micro-metastatic deposits. While systemic chemotherapies and adjuvant endocrine therapies have substantially reduced breast cancer recurrences and deaths, they also have associated life-threatening toxicities [4]. It is therefore of clinical importance to identify patients who benefit the most from these treatments and to spare those who are unlikely to benefit from aggressive therapies. At present, decisions regarding adjuvant therapies for breast cancer patients are predominantly guided by tumor-based prognostic factors such as axillary lymph nodal status, tumor size, tumor histologic grade, lymphatic and vascular invasion, proliferative markers, estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) status [5,6]. However, success from treatments guided by these prognostic factors is limited, in part, due to the molecular heterogeneity of breast cancer. Furthermore, approximately 30% of treated early-stage breast cancer patients show disease recurrence within ten years [7,8], indicating need of improved prognostic and predictive markers with higher sensitivity and specificity.

1.2 Etiology of breast cancer

Breast cancer is a complex multifactorial disease, which results from an interplay of environmental, reproductive, lifestyle and genetic risk factors. It has been estimated that approximately one-third of variations in breast cancer

susceptibility[1] is accounted for by inherited genetic risk factors while environmental and lifestyle risk factors contribute to the remaining two-thirds [9-11]. Age is the strongest risk factor for breast cancer after gender [12]. Breast cancer risk increases with age; the older the woman, the greater her risk of developing breast cancer. Breast cancer risk is also found to be influenced by reproductive factors, such as age at menarche, age at first live birth, parity, breastfeeding and age at menopause. Early age at menarche and late age at menopause have been reported to be associated with increased risk of breast cancer [9,13]. In contrast, childbearing at younger age, higher number of full-term pregnancies and breastfeeding are reported to reduce breast cancer risk [9].

Studies have shown that higher levels of circulating endogenous hormones, such as estrogen and progesterone, increases breast cancer risk by approximately 2-3 times in postmenopausal women [14]. However, the risk caused by these hormones in premenopausal women is not fully understood. Further, women who use oral contraceptives for birth control are at 24% greater risk of breast cancer than those who do not although the excess risk disappears ten years after stopping use [9]. Women receiving hormone replacement therapy were found to be at 66% greater risk of breast cancer than those who were not. Breast cancer risk will be equivalent to non-users five years after stopping use [9]. Studies have shown that family history of breast cancer is a strong risk factor. Women with one first degree relative diagnosed with breast cancer have two-fold higher risk of developing breast cancer than women with no first degree relatives. The risk

[1] Risk of developing a diseased state.

becomes five times higher when two affected first degree relatives are present in the family [9,15]. Over-weight and obese postmenopausal women are at 10-20% increased risk of breast cancer than those with normal body weight [16]. Further, physically active women are at 15-20% reduced risk of breast cancer compared to non-active women [17]. The effect of physical activity on risk of breast cancer is stronger in postmenopausal than in premenopausal women. Epidemiological studies have also shown that women with prior diagnosis of ductal or lobular carcinoma *in situ* (non-invasive forms of breast cancer) are at two-fold higher risk of developing invasive breast cancer [18]. Detailed description of aforementioned and additional environmental, health and lifestyle risk factors for breast cancer is beyond the scope of this thesis however, excellent articles by the Collaborative Group on Hormonal Factors in Breast Cancer [9] and by Lichtenstein *et al.* [11] are suggested for interested readers.

1.3 Genetic basis of breast cancer

While environmental, health and lifestyle risk factors contribute most to the variation in breast cancer susceptibility, multiple twin studies have shown a substantial contribution (approximately 30%) of genetic factors to disease susceptibility [9,11]. Twin studies help to estimate the relative contributions of genetic and environmental factors to diseases or traits and generally consist of pairs of identical (monozygotic) and non-identical (dizygotic) twins. Identical twins are derived from a single zygote and hence share all the genetic material. Non-identical twins are derived from two different zygotes and hence share 50% of the genetic material. Assuming that pairs of twins share a common

environment, identical twins are more likely to develop the same disease than non-identical twins, if the disease has an inherited component. The relative contribution of the inherited component of a disease can then be estimated by statistical modeling, commonly referred to as quantitative genetic analyses [11].

During the past two decades, genetic linkage and association studies have been the predominantly adopted experimental study designs to delineate inherited genetic risk factors for complex diseases or traits, including breast cancer.

1.3.1 Genetic linkage studies

Linkage studies have been successful in identifying predisposition[2] factors for many diseases, including breast cancer [19]. Basic ideas behind linkage studies are (i) closely located genes in chromosomes will co-segregate together when passed to offspring and (ii) if a disease is passed to offspring, together with some known marker genes or loci (often microsatellite markers[3] have been utilized), then gene(s) responsible for the disease are said to be linked with the markers. Such studies require multiple affected families as co-segregation of disease and marker genes needs to be examined in multiple generations. A statistical test, called LOD (logarithm (base 10) of odds) score is used to measure the linkage [20]. The LOD score compares the probabilities of two loci being linked with that of not being linked. The presence of linkage is indicated by a positive LOD score while the absence of linkage is indicated by a negative LOD score.

---

[2] Risk of developing a diseased state.
[3] Repeating sequences of 2-6 base pairs of DNA.

1.3.1.1 High penetrance[4] breast cancer predisposition genes

The first breast cancer predisposition gene to be identified was, *BRCA1*[5], located on chromosome 17q21 region was found in a linkage study in 1990, with a LOD score of 2.35 with a microsatellite marker (D17S74) [21,22]. The linkage was stronger in families with early onset of disease (less than 46 years) with LOD score of 5.98 while linkage vanished in families with late onset of disease, indicating that this gene may not contribute to predisposition to breast cancers that are sporadic in nature. A linkage study conducted in 1994 identified another breast cancer predisposition gene, *BRCA2*[6], located on chromosome 13q12-13 [23]. Both *BRCA1* and *BRCA2* genes play crucial roles in maintaining genomic stability, by their involvment in repair of DNA double strand breaks. Multiple germline mutations in *BRCA1* and *BRCA2* genes have been detected; however, they occur in a small fraction of total breast cancer cases [24]. Studies have shown that most breast cancer-associated germline mutations in *BRCA1* and *BRCA2* genes result in premature truncation of encoded proteins, translational frame shifts and defective splice sites [25,26].

Both *BRCA1* and *BRCA2* are categorized as high penetrant breast cancer genes that confer more than ten-fold increase in disease risk [25-28]. Evidence from epidemiological studies suggests that breast cancer risk by age 70 may

---

[4] Penetrance is defined as the fraction of individuals with a gene or an allele expressing a certain disease or trait. If a gene or an allele is highly penetrant, almost all individuals carrying that gene or allele will express the disease or trait. Penetrance can be equated in terms of relative risk (RR). Breast cancer predisposition genes or alleles with RR>10 are classified as high penetrant, with 2>RR<10 are classified as moderate penetrant and with RR<2 are classified as low penetrant [27,31].

[5] Breast cancer 1, early onset.

[6] Breast cancer 2, early onset.

increase up to 87% in carriers of *BRCA1* mutations and up to 84% in carriers of *BRCA2* mutations carriers [26-28]. Since germline mutations in *BRCA1* and *BRCA2* genes are very rare, these two predisposition genes could only explain 15-20% of genetic risk of breast cancer in the overall population [29-32].

1.3.1.2 Moderate penetrance breast cancer predisposition genes

Continued research efforts to characterize additional breast cancer predisposition genes resulted in identification of multiple genes conferring moderate risk for breast cancer. Germline mutational screening of cancer-related pathway genes identified two cancer predisposition syndromes: Li-Fraumeni and Cowden syndrome [33,34]. Both syndromes were characterized by a variety of different individual germline mutations in their causative tumor suppressor genes, *TP53*[7] [34] and *PTEN*[8] [33], respectively. These syndromes were also found in familial breast cancers[9], conferring increased risks of breast cancer [34,35]. Even though the exact associated risks of breast cancer due to germline mutations in *TP53* and *PTEN* genes are not certain, these are believed to exhibit moderate penetrance for breast cancer predisposition [33,34].

Subsequent germline mutational screening of cancer related candidate genes identified four additional moderate penetrance breast cancer predisposition genes, *CHEK2*[10] [36], *PALB2*[11] [37], *BRIP1*[12] [38] and *ATM*[13] [39]. Germline mutations in

---

[7] Tumor protein 53.
[8] Phosphatase and tensin homolog.
[9] Cases with family history of breast cancer.
[10] Checkpoint kinase 2.
[11] Partner and localizer of *BRCA2*.
[12] *BRCA1*-interacting protein 1.
[13] Ataxia telangiectasia mutated.

these genes are also very rare in the general population and confer moderate risk for breast cancer predisposition. Together, these mutations were suggested to account for an additional 2.3% of genetic risk of breast cancer [37].

1.3.1.3 Other high or moderate penetrance breast cancer predisposition genes

The multiple high and moderate penetrance breast cancer predisposition genes identified thus far are rare in the general population and explain less than 25% of variations in the familial component of disease susceptibility. Further linkage studies did not yield additional *BRCA*-like genes conferring higher penetrance risk for breast cancer predisposition [40]. Search for moderate penetrance genes through germline mutational screenings of cancer-related pathway genes was also not successful. The residual or missing heritability[14] component for breast cancer is explained in terms of the common disease-common variant (CDCV) hypothesis, which states that common diseases are caused by common variants [41,42]. According to the CDCV hypothesis, multiple common low penetrance genes or alleles, either singly or in combination, confer breast cancer risk, conforming to a polygenic model of genetic inheritance. Under the polygenic model, each of the participating genes or alleles, also known as polygenes, has a small additive effect for breast cancer predisposition while linkage among loci and possible influence of environmental factors are ignored [43-46]. The CDCV hypothesis is the basis for several genetic association studies conducted during the last ten years, with an objective of characterizing additional

[14] Phenotypic variations caused by spectrum of genetic variations.

predisposition risk factors for many complex diseases or traits, including breast cancer [47].

1.3.2 Genetic association studies

Genetic association studies are conducted to determine contributions of genetic variants to certain diseases or traits under study. The most commonly used strategies to evaluate genetic contributions to breast cancer in populations are case-control association studies, wherein frequencies of genetic variants in breast cancer cases are compared with those of healthy controls and statistical significance of frequency differences is calculated; controls are free from breast cancer at the time of enrollment in such studies (**Figure 1-1**).

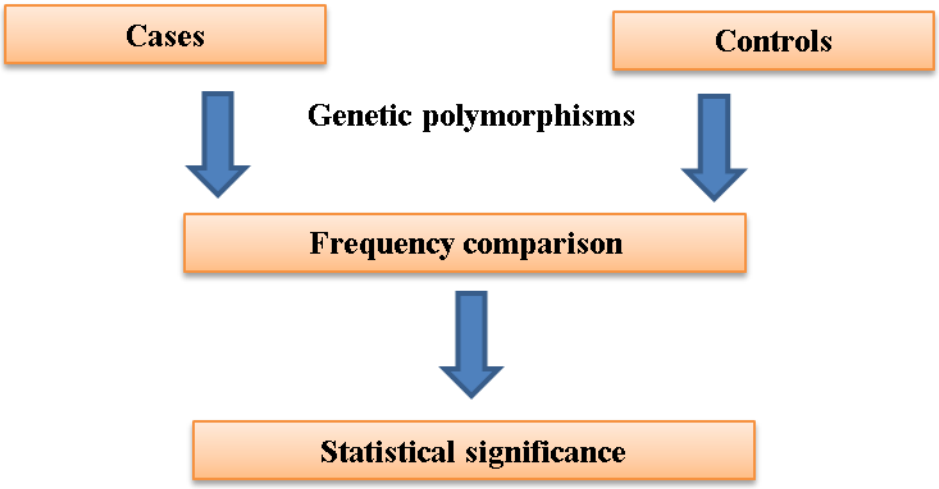


Figure 1-1 A typical genetic case-control association study.

As governed by the CDCV hypothesis, association studies largely rely on common genetic variants[15] with population frequencies >5% as statistically significant frequency differences between cases and controls at this cut-off are

[15] Genetic variants occurring at population frequency of more than 5%.

more easily demonstrable. Single nucleotide base pair changes known as single nucleotide polymorphisms (SNPs) are the predominantly used human genomic variation in association studies [48].

1.3.2.1 Single nucleotide polymorphism (SNPs)

The somatic cells of humans are diploid and contain 22 pairs of homologous chromosomes (autosomes), where each parent contributes one chromosome each and a pair of sex chromosomes (X and Y). These chromosomes are represented by double helix DNA of approximately three billion nucleotide base pairs: adenine (A), thymine (T), cytosine (C) and guanine (G). SNPs are DNA sequence variations at a single nucleotide base between the pairs of homologous chromosomes, occurring with more than 1% population frequency. These are the most common, typically biallelic[16], relatively stable and evolutionary conserved genetic variations in the human genome. On average, SNPs occur at a frequency of one SNP/300 bp in the human genome [49,50].

The allele that is more common in a given population is called the major allele (A), while the allele which is less common is called the minor allele (B). Together, the alleles from paternal and maternal chromosomal loci contribute to three distinct genotypes. When the alleles inherited from both parents are the major allele (AA), the genotype is referred to as wild-type homozygous; when both inherited alleles are minor alleles (BB), the genotype is referred to as variant-type homozygous; and when each parent contribute a major or a minor allele (AB), the genotype is referred to as heterozygous. Allele and genotype

[16]Having two alleles or genetic forms.

frequencies are calculated from the genotype counts of SNPs in a given population. Frequency of the minor allele is referred to as minor allele frequency (MAF) while the frequency of the major allele is referred to as major allele frequency.

SNPs are the most commonly used human genomic markers in genetic association studies to uncover associations of genetic loci in complex diseases or traits. As of 26 June, 2012, the National Center for Biotechnology Information dbSNP Build 137 reported 187,852,828 SNPs in humans [51]. Using SNPs as markers, two primary types of genetic association studies have been used to further characterize residual heritability for breast cancer predisposition: candidate-gene association studies and genome-wide association studies (GWASs).

1.3.2.2 Candidate-gene association studies

Initial genetic association studies aimed to uncover common variants for breast cancer predisposition were largely focused on genes involved in DNA repair processes, apoptosis and cell-cycle regulation, since these had pausible roles in breast cancer [52-55]; such genes are also known as candidate genes. In candidate-gene association studies, SNPs in or close to candidate genes are examined for their roles in breast cancer predisposition, governed by a hypothesis-driven approach. Several candidate gene-association studies for breast cancer have been conducted during the last ten years [56-62]; however, many of these studies were underpowered due to small sample size, variations in the cut-off imposed for minor allele frequencies for candidate SNPs under study, non-

adherence to linkage disequilibrium (LD)[17] among markers, technology platform differences in genotyping[18], differential SNP call rates (missing values) and population stratification. These were further compounded by potential bias in selection of breast cancer cases and controls, resulting in inconsistent and irreproducible findings [43,63]. To date, only one SNP (rs1045485[19]) located in the coding region of *CASP8*[20] identified by a candidate-gene association study has shown promise as a breast cancer predisposition factor [64]. The minor allele of the coding SNP conferred 12% less risk for breast cancer than the major allele (odds ratio (OR)[21]=0.88, 95% confidence interval (CI)=0.84-0.92 and P=1.1 x 10^{-7}). An independent case-control study also reported breast cancer risk reduction due to rs1045485 in *BRCA1* and *BRCA2* carriers [65]. However, the potential causal role of rs1045485 is still not clear.

1.3.2.3 Genome-wide association studies (GWASs)

Advances in genotyping technologies, completion of the Human Genome Project, the International HapMap Project[22] and the 1000 Genomes Projects[23] have led to the paradigm shift of genetic association studies from a limited candidate-gene approach to a genome-wide approach, resulting in more detailed

---

[17] Non-random associations among alleles that inherit together.
[18] An experimental technique to determine the genetic makeup of an individual.
[19] An amino acid substitution of D (Aspartate) to H (Histidine) at residue 302.
[20] Caspase 8.
[21] A measure of strength of association of a categorical independent variable (*i.e.*, SNP genotype) with a dichotomous dependent variable (*i.e.*, case/control status).
[22] The HapMap Project was designed to capture the patterns of common genetic variations in the human genome.
[23] An international research project launched in 2008 with an objective to provide by far the most comprehensive catalogue of human genetic variation, by sequencing whole genomes of approximately 1,000 individuals.

investigation of the CDCV hypothesis. GWAS premise rely on LD among SNPs, which states that many neighboring SNPs are correlated and hence inherited together in a LD block[24] [66]. Although actual sizes of LD blocks are still debated, it has been estimated that LD blocks in the human genome range in size from a few kilobases to more than 100 kilobases [66]. Such correlations (LD) among nearby SNPs enable selection of fewer SNPs (tag SNPs[25]) that essentially capture the information inherent to the block [67]. As such, by just genotyping 500,000 to one million common tag SNPs (with population frequencies >5%), we could effectively capture the information content of more than 80% of all common SNPs in the human genome [68,69]. As of 16 December 2012, 1,120 GWASs have identified >4,500 low penetrance common SNPs associated with over 700 different diseases or traits [47]. Majority of these GWASs have predominantly utilized two types of genotyping platforms: Illumina and Affymetrix SNP arrays. Illumina SNP platforms largely array tag SNPs capturing the genome while Affymetrix SNP platforms represent SNPs across the genome; the SNP arrays capture highly validated SNPs and include representative tag SNPs. Even though there are several differences between Illumina and Affymetrix platforms, ranging from the content of the arrays to their prices, the choice of arrays mainly depends on the application. Illumina arrays are more commonly used in association studies due to their predominant tagSNP capture and large scale consortia efforts and partnerships with Illumina. In contrast, for an unbiased

[24] A chromosomal region where SNPs are highly correlated with each other or are in high LD.
[25] A SNP that is highly correlated with neighboring SNPs and that essentially captures the information of nearby SNPs in a LD block.

detection of copy number variations (CNVs), Affymetrix arrays are popular due to uniform distribution of probes (SNPs and CNVs) across the genome.

1.3.2.4 General experimental design of a GWAS

The majority of GWASs conducted to date have utilized SNPs as genetic markers to identify low penetrance common variants for diseases or traits. GWASs representing other types of human genomic variations, especially CNVs, also exist and reports using these markers in genetic association studies are slowly emerging. Since large-scale GWASs are cost-prohibitive, a typical GWAS follows a multi-stage study design [70-73]. In first or exploratory (hypothesis generating) stage (stage 1), large numbers of SNPs covering the entire human genome are genotyped in a limited number of cases and controls. The number of cases and controls in this stage varies from a few hundred to thousands, depending on investigator chosen premise for statistical power and allelic or genotypic models considered. Frequencies of SNPs in cases and controls are calculated and a statistical test is used to estimate significance of difference in frequencies between cases and controls. Description of standard statistical tests commonly used in GWASs is provided in the next section of this chapter.

If a SNP demonstrates statistically significant frequency difference between cases and controls, it is then considered to be associated with the disease or trait under study. Since adjacent SNPs in a chromosome do not segregate

independently, but rather as haplotype blocks[26] [74], the associated SNP acts as a surrogate marker for the disease-associated chromosomal locus to which it maps. In contrast to candidate-gene association study designs that are limited to SNPs related to a few candidate genes of known or inferred functional significance with diseases or traits, GWAS approaches allow one to investigate the entire human genome and hence are known as hypothesis-free approaches (discovery stage or stage 1) [71,73]. In the second or replication stage (stage 2), the most promising SNPs from stage 1 showing statistically significant associations with diseases or traits are selected and genotyped in a larger but independent set of cases and controls [75]. Most GWASs that use Illumina genotyping platforms consider SNPs showing the strongest associations with diseases or traits (based on significance *P* values) in stage 1 as the most promising SNPs for replication because Illumina platforms array tag SNPs that are surrogate for many nearby SNPs in a LD block. GWASs using Affymetrix as their genotyping platforms in stage 1 tend to consider those SNPs for replication that show not only the strongest associations with diseases or traits under investigation but also are in strong LD (Pearson's correlation coefficient, $r^2 \geq 0.8$) with nearby markers. Such tag SNPs or SNPs in strong LD with nearby SNPs selected for replication capture information from the whole LD block in which they reside in, enabling higher coverage of the genome by simply genotyping fewer SNPs [66,67,74].

[26] Haplotype block is a chromosomal region where there is little evidence for historical recombination; SNPs or alleles in a haplotype block are in high LD. Haplotypes per se are combinations of SNPs or alleles that tend to segregate together from one generation to another.

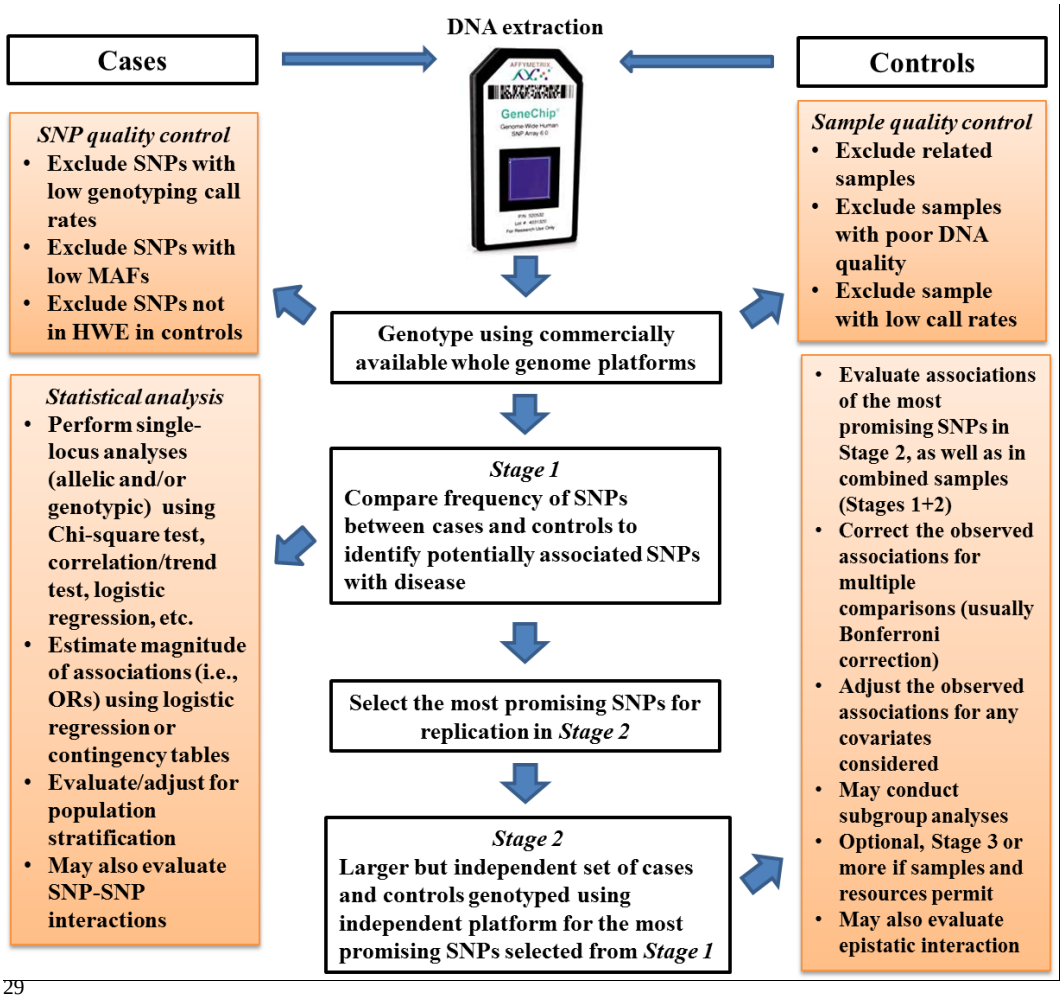
Depending on the availability of samples and genotyping costs, GWASs can be extended to stage 3 and stage 4 wherein SNPs showing statistically significant associations with diseases or traits in stage 1 and/or stage 2 are re-genotyped in a larger but independent set of cases and controls (stages 3 and 4). At the end, cases and controls from all independent stages (1, 2, 3 and 4) are assembled together and associations of SNPs with diseases and traits are examined in a combined sample, using a process commonly referred to as combined analysis [76]. This will help increase statistical power to identify possible associations of low penetrance common variants with diseases or traits under study [71,73]. A typical GWAS with a sample size of approximately 3,000 individuals (1,500 cases and 1,500 controls) will have 80% power to capture associations of SNPs (MAF>10%) with diseases or traits, with odds ratios (effect sizes) ranging from 1.5 to 2.0 [77]. Larger sample sizes (more than 20,000) are required to identify associated risk of SNPs with smaller MAFs (<10%) and odds ratios less than 1.3 and hence collaborative efforts with large international consortia are required to capture such associations.

1.3.2.5 Quality control and statistical considerations in GWASs

Even though GWASs have emerged as powerful tools to identify inherited genetic risk factors for many complex diseases, one needs to consider many aspects and challenges posed by the large amounts of SNP genotype data to ensure accurate interpretation of GWASs findings while minimizing the chances of false positive associations (type 1 error).

A typical workflow for multi-stage association design with quality control measures as well as standard statistical tests is illustrated in **Figure 1-2**. After genotyping, genotype data needs to be cleaned up to remove possible genotype errors [71,73] and other inconsistencies owing to assay reproducibility across genotyping platforms. These include SNPs with low SNP call rates[27], missing genotype data and low MAFs (<5%). Sometimes, samples used for genotyping may be of poor DNA quality leading to low sample call rates[28]; removal of such samples is highly recommended.

[27] Proportion of SNPs reliably genotyped across all samples assayed. A SNP call rate >99% is generally desired.
[28] Proportion of reliably genotyped SNPs per sample over total number of SNPs arrayed.



[29]

Figure 1-2 A workflow for the multi-stage association study with standard quality control measures as well as statistical considerations.

Further, allele and genotype frequencies of SNPs in control subjects need to obey the Hardy-Weinberg Equilibrium (HWE) rule of Mendelian Genetics[30], as deviations from this equilibrium would suggest that either or a combination of natural selection, mutations, non-random mating or migration in or out of the

[29] Bonferroni correction is the most conservative method to address the problem of multiple comparisons. It is calculated by dividing nominal *P* value of 0.05 by the number of statistical tests performed.

[30] Allele and genotype frequencies remain constant from generation to generation.

population under study is occurring [71,73]. Genotype frequencies deviating from HWE proportions may contribute to spurious associations. Another challenge in GWASs is population structure that may also lead to false positive associations. Population structure is the differences in allele or genotype frequencies within a population under study due to ancestry differences. Such population structure may result in false positive associations that are not actually associated with diseases or traits [78,79]. Hence, it is crucial that cases and controls for GWASs are selected from the same ethnic background.

Once the data cleaning step is complete, SNPs or genotype data are analyzed for their potential associations with diseases or traits one SNP or genotype at a time in a given sample size, resulting in significance *P* values and ORs. These approaches are known as single-locus tests, which estimate significance *P* values from contingency tables[31]. Similar statistical tests can also be used for investigating association of haplotypes in a case-control setting, commonly referred to as haplotype association analysis, wherein frequencies of haplotypes in cases and controls are compared to obtain the statistical significance of the frequency difference.

A 2x3 contingency table is created for evaluating association of a SNP assuming either general genotypic[32] or an additive[33] models of inheritance while a 2x2 contingency table is created for examining association of a SNP under

[31] A table in a matrix format that displays frequency distributions of categorical variables.
[32] Under genotypic model, each genotype is treated equally as potential disease associated genotype.
[33] Under additive model, disease risk is one-fold for heterozygotes and two-fold for variant homozygotes.

multiplicative (allelic)[34], dominant[35] and recessive[36] genetic models of inheritance (**Figure 1-3**). Both 2x3 and 2x2 contingency tables can be evaluated by standard statistical software, such as SPSS, SAS, R, Stata, S-plus or even by Microsoft Excel.

(A) Additive model

	AA	AB	BB
Cases	a	b	c
Controls	d	e	f

(B) Multiplicative model

	A	B
Cases	2a+b	b+2c
Controls	2d+e	e+2f

Figure 1-3 Contingency tables for genetic case-control studies, by genetic model of inheritance.

Most commonly used statistical tests to estimate significance *P* values for the association of SNPs with diseases and traits include chi-square tests, correlation/trend tests and logistic regressions for allelic models and Armitage's test for trend for additive models [71,73]. ORs and corresponding 95% CIs can be estimated from contingency tables as well as from unconditional logistic

[34] Under multiplicative or allelic model, disease risk is multiplied *i.e.*, if a minor allele of a SNP confers two-fold disease risk, variant homozygote will confer 2x2=four-fold disease risk.
[35] Under dominant model, one copy of minor allele confers disease risk.
[36] Under recessive model, two copies of minor allele confer disease risk.

regressions implemented in standard statistical software. Whenever the response variable is time-dependent such as time to disease recurrence and death after diagnosis, potential associations of genetic variants with recurrence-free or overall survivals are estimated with the Cox proportional hazards model and represented as hazards ratios (HRs) and corresponding 95% CIs. If we only want to test whether or not there is difference between survival times[37] of different groups (*e.g.*, groups defined by genes or genetic markers), log rank tests are used; however, these tests do not allow other variables (*e.g.*, confounders[38]) to be taken into account, unlike the Cox proportional hazards model.

Individual SNPs analyzed in GWAS for their potential associations with diseases or traits easily exceed few hundred thousands, resulting in a multiple comparison problem. Bonferroni correction has been the predominantly used statistical approach to correct for a multiple comparison problem, in which the nominal *P* (0.05) is divided by the number of SNPs analyzed for their potential associations with diseases or traits (e.g., 800,000), resulting in Bonferroni corrected significance $P=0.05/800{,}000=6.2 \times 10^{-8}$ (often considered as a stringent correction since redundant SNPs, *i.e.*, all those surrogates of tagSNPs from the entire LD block, are also included). However, the Bonferroni corrected *P* can change depending on the number of SNPs analyzed in stage 1 of GWASs [80,81]. SNPs showing statistically significant associations at Bonferroni corrected significance *P* are known to achieve genome-wide significance level and these

[37] The length of time taken to reach a certain end-point, such as disease recurrence and death.
[38] In epidemiology, a confounder is a variable that is associated with both response and independent (explanatory) variables and may result in spurious association of independent variable with response variable if a confounder is not taken accounted for.

SNPs are believed to confer single-locus effects[39]to complex diseases or traits [71,73].

More recently, GWASs for complex diseases have also been extended to stratified analyses wherein frequencies of SNPs are compared between specific subtypes of cases and healthy controls. Such approaches are commonly referred to as subgroup analyses aimed to address disease heterogeneity [82,83]. These subgroup analyses have been commonly adopted to GWASs for breast cancer to identify SNPs associated with subtypes of breast cancer (*e.g.*, estrogen and progesterone receptor status or amplification of human epidermal growth factor receptor 2/Her2 status) based on known clinicopathological characteristics [83,84].

1.3.2.6 GWASs for breast cancer predisposition

To date, more than 40 low penetrance common breast cancer susceptibility SNPs have been identified by multiple GWASs and by independent investigators. A brief summary of breast cancer GWASs is provided in **Table 1-1**. The data presented in the table was retrieved from the National Human Genome Research Institute catalog of published GWASs on 16 December 2012 [47]. The table only shows SNPs conferring single-locus effects for breast cancer at or near conventional genome-wide significance level.

[39] Disease risk conferred by an individual locus.

Table 1-1 A brief summary of breast cancer GWASs conducted during 2007-2012.

Study	Discovery Stage (Cases/controls)	Replication Stage (Cases/controls)	SNPs	Region	MAF	*P* value	ORs	95% CI
Easton et al., 2007	390/364 Familial/Caucasian	26,646/24,889	rs2981582	10q26.13	0.38	2.00E-76	1.26	[1.23-1.30]
			rs3803662	16q12.1	0.25	1.00E-36	1.20	[1.16-1.24]
			rs889312	5q11.2	0.28	7.00E-20	1.13	[1.10-1.16]
			rs13281615	8q24.21	0.4	5.00E-12	1.08	[1.05-1.11]
			rs3817198	11p15.5	0.3	3.00E-09	1.07	[1.04-1.11]
Hunter et al., 2007	1,145/1,142 Sporadic, postmenopausal/ Caucasian	1,176/2,072	rs1219648	10q26.13	0.4	1.00E-10	1.20	[1.07-1.42]
Stacey et al., 2007	1,599/11,546 ER+ve/Icelandic	2,934/5,967	rs3803662	16q12.1	0.27	6.00E-19	1.28	[1.21-1.35]
			rs13387042	2q35	0.5	1.00E-13	1.20	[1.14-1.26]
Stacey et al., 2008	2,277/26,199 ER+ve/Icelandic	6,643/12,922	rs4415084	5p12	0.43	6.40E-10	1.16	[1.10-1.21]
			rs1219648	10q26	0.47	1.30E-17	1.23	[1.17-1.29]
			rs10941679	5p12	0.24	2.90E-11	1.19	[1.13-1.26]
Gold et al., 2008	249/299 Familial/Ashkenazi Jews	1,193/1,166	rs2180341	6q22.33	0.21	3.00E-08	1.41	[1.25-1.59]
Ahmed et al., 2009	390/364 Familial/Caucasian	41,002/43985	rs4973768	3p24	0.46	4.10E-23	1.11	[1.08-1.13]
			rs6504950	17q23	0.27	1.40E-08	0.95	[0.92-0.97]
Thomas et al., 2009	1,145/1,142 Sporadic,postmenopausal/Caucasian	8,625/9,657	rs2981579	10q26.13	0.41	2.00E-10	1.17	[1.07-1.27]
			rs11249433	1p11.2	0.39	7.00E-10	1.16	[1.09-1.24]
			rs3803662	16q12.1	0.27	1.00E-09	1.16	[1.07-1.27]
			rs13387042	2q35	0.51	2.00E-08	1.25	[1.15-1.37]
Zheng et al., 2009	1,505/1,522 Sporadic/Chinese	1,554/1,576	rs2046210	6q25.1	0.37	2.00E-15	1.29	[1.21-1.37]
Antoniou et al., 2010	1,193/1,190 Familial/Caucasian	3,012/2,974	rs8170	19p13.11	0.17	2.00E-09	1.26	[1.17-1.35]
Gaudet et al., 2010	899/804 Familial/Caucasian	1,264/1,222	rs2981575	10q26.13	0.42	1.00E-08	1.28	[1.18-1.39]
Turnbull et al., 2010	3,659/4,897 Familial/Caucasian	12,576/12,223	rs2981579	10q26.13	0.42	4.00E-31	1.43	[1.35-1.53]
			rs3803662	16q12.1	0.26	3.00E-15	1.30	[1.22-1.39]
			rs614367	11q13.3	0.15	3.00E-15	1.15	[1.10-1.20]
			rs10995190	10q21.2	0.85	5.00E-15	1.16	[1.10-1.22]
			rs13387042	2q35	0.49	2.00E-10	1.21	[1.14-1.29]
			rs704010	10q22.3	0.39	4.00E-09	1.07	[1.03-1.11]
			rs889312	5q11.2	0.28	5.00E-09	1.22	[1.14-1.30]
			rs1011970	9p21.3	0.17	3.00E-08	1.09	[1.04-1.14]
Long et al., 2010	2,073/2,084 Sporadic/Chinese	13,395/10,917	rs4784227	16q12.1	0.24	1.00E-28	1.24	[1.20-1.29]
Cai et al., 2011	2,062/2,066 Sporadic/Chinese	15,091/14,877	rs10822013	10q21.2	0.47	6.00E-09	1.12	[1.06-1.18]
Fletcher et al., 2011	2,839/3,507 Familial/Caucasian	7,317/8,124	rs1219648	10q26.13	0.42	1.00E-30	1.31	[1.25-1.37]
			rs1562430	8q24.21	0.6	3.00E-11	1.16	[1.11-1.22]
			rs4415084	5p12	0.42	8.00E-11	1.17	[1.11-1.22]
			rs865686	9q31.2	0.61	2.00E-10	1.12	[1.09-1.18]
			rs13387042	2q35	0.52	2.00E-10	1.16	[1.11-1.22]
			rs3112612	16q12.2	0.43	4.00E-10	1.15	[1.10-1.21]
			rs4973768	3p24.1	0.49	2.00E-08	1.14	[1.09-1.19]

Table 1-1 Continued..

Haiman et al., 2011	2,722/6,415 ER-ve/Caucasian, African	2,222/16,363	rs10069690	5p15.33	0.26	1.00E-10	1.18	[1.13-1.25]
Kim et al., 2012	2,273/2,052 Familial/Korean	4,049/3,845	rs13393577	2q34	0.051	9.00E-14	1.53	[1.37-1.70]
Long et al., 2012	2,918/2,324 Sporadic/Chinese	16,173/18,282	rs9485372	6q25.1	0.55	4.00E-12	1.11	[1.09-1.15]
Siddiq et al., 2012	4,670/28,864 ER-ve/Caucasian, African	946/8,404	rs9383938	6q25.1	NR	2.00E-10	1.28	NR
			rs2284378	20q11.22	0.31	1.00E-08	1.16	[1.10-1.22]
			rs8100241	19p13.11	NA	4.00E-08	1.14	NR
Ghaoussaini et al., 2012	10,052/17,765 Familial/Caucasian	~60,000/~50,000	rs10771399	12p11	0.12	2.70E-35	0.85	[0.83-0.88]
			rs1292011	12q24	0.41	4.30E-19	0.92	[0.91-0.94]
			rs2823093	21q21	0.27	1.10E-12	0.94	[0.92-0.96]
Sapkota et al., 2013*	302/321 Sporadic/Caucasian	2,447/4,149	rs1429142	4q31.22	0.18	1.50E-07	1.28	[1.17-1.41]

MAF, minor allele frequency; OR, odds ratio; CI, confidence interval; NR, not reported; *this study is one of the very few GWASs for sporadic breast cancer predisposition utilizing Caucasian study subjects (Chapter 2 of this thesis)

The first breast cancer GWAS conducted in 2007 by Easton *et al.* interrogated a total of 227,876 SNPs arrayed in a custom-designed Perlegen platform[40] in familial breast cancer cases and healthy controls from the United Kingdom and identified five breast cancer susceptibility loci [84]. Of these, four SNPs are located in gene regions of *FGFR2*[41], *TNRC9*[42], *MAP3K1*[43] and *LSP1*[44] while the fifth SNP is located in an intergenic region on chromosome 8q. Two additional novel breast cancer susceptibility loci on chromosomes 3p24 and 17q23.2 were also identified by further data mining and follow-up of familial breast cancer GWAS comprising Caucasian women from the United Kingdom reported by Easton *et al.* [84,85]. Subsequently, Stacey *et al.* genotyped approximately 300,000 SNPs arrayed in IlluminaHap300 platform for familial breast cancer cases and healthy controls from Iceland and identified two breast

[40] A genotyping platform developed by Perlegen Sciences, Inc.
[41] Fibroblast growth factor receptor 2.
[42] Trinucleotide repeat-containing 9.
[43] Mitogen-activated protein kinase kinase kinase 1, E3 ubiquitin protein ligase.
[44] Lymphocyte-specific protein 1.

cancer susceptibility loci on chromosomes 2q35 and 16q12 [86]. In another breast cancer GWAS, Stacey *et al.* identified novel breast cancer susceptibility locus on chromosome 5p12 using familial breast cancer cases and controls from Iceland [87]. The association was stronger for breast cancer cases with estrogen receptor (ER)-positive than ER-negative tumors. *FGFR2* SNPs recently reported by Easton *et al.* also showed stronger associations for ER-positive than ER-negative breast tumors [84,87].

As part of the National Cancer Institute Cancer Genetics Markers of Susceptibility (CGEMS), Hunter *et al.* conducted a second breast cancer GWAS by genotyping 528,252 SNPs in postmenopausal sporadic breast cancer cases and healthy controls of European ancestry in IlluminaHapMap500 array[45] [88]. The findings from this GWAS confirmed association signals from intron 2 of the *FGFR2* gene reported by Easton *et al.* [84]. Similarly, Thomas *et al.* conducted a follow-up genetic association study of postmenopausal sporadic breast cancer GWAS reported by Hunter *et al.* that included study subjects of European ancestry and identified two additional breast cancer susceptibility loci on chromosomes 1p11.2 and 14q24.1 [89]. The reported loci also showed stronger associations in cases with ER-positive than ER-negative breast tumors. Results from this study also confirmed previous breast cancer susceptibility signals reported by Easton *et al.*, Hunter *et al.* and Stacey *et al.* [84,86-88]. Fletcher *et al.* identified a novel breast cancer susceptibility locus on chromosome 9q31.2 by

[45] Contains approximately 500,000 SNPs.

further data mining and follow-up of the postmenopausal sporadic breast cancer GWAS reported by Hunter *et al.* [88,90].

Turnbull *et al.* conducted a breast cancer GWAS by genotyping familial breast cancer cases and healthy controls from UK on IlluminaInfinium660K array[46] and identified five additional breast cancer susceptibility loci not reported previously [91]. This study also confirmed associations of previously reported breast cancer susceptibility loci reported by Easton *et al.*, Hunter *et al.*, Stacey *et al.*, Thomas *et al.* and Ahmed *et al.* [84-89,92] in their study populations.

Gold *et al.* interrogated 150,080 SNPs arrayed on IlluminaGoldenGate platform in breast cancer cases and controls among non-*BRCA1/2*mutation carriers from a genetically isolated population, Ashkenazi Jews, and identified a novel breast cancer susceptibility locus on chromosome 6q22.33 [93].

Antoniou *et al.* interrogated 620,601 SNPs on IlluminaInfinium 610K array[47] using *BRCA1* mutation carriers of European ancestry with and without breast cancer diagnoses and identified a novel breast cancer susceptibility locus on chromosome 19p13 [94]. The observed association was stronger in cases with triple-negative breast tumors. Gaudet *et al.* genotyped 592,163 SNPs on Affymetrix SNP 6.0 using *BRCA2* mutation carriers of European descent with and without breast cancer diagnoses and identified a breast cancer susceptibility SNP in *FGFR2* intron 2, a locus reported and confirmed by many independent association studies [95].

---

[46] Contains approximately 660,000 SNPs.
[47] Contains approximately 610,000 SNPs.

Haiman *et al.* interrogated 3,154,485 SNPs genotyped on Illumina550-Duo SNP array and imputed[48] in ER-negative breast cancer cases and healthy controls of European ancestry and identified a novel susceptibility locus for ER-negative breast cancer on chromosome 5p15 [96]. A meta-analysis of multiple breast cancer GWASs conducted by Siddiq *et al.* identified two novel breast cancer susceptibility loci on chromosomes 20q11 and 6q14 in women of European ancestry [97]. Of these, rs2284378 on 20q11 showed stronger associations in ER-negative breast cancers as compared to ER-positive and breast cancer cases unselected for ER status.

Zheng *et al.* analyzed 607,728 SNPs genotyped on Affymetrix 500K[49] and SNP 6.0[50] arrays using breast cancer cases and healthy controls of Chinese ancestry and identified a breast cancer susceptibility locus on chromosome 6q25.1 [92]. An additional breast cancer susceptibility SNP on chromosome 6q25.1 for East-Asian women (Chinese, Japanese and Korean) was identified by Long *et al.* through data mining and follow-up of breast cancer GWAS reported by Zheng *et al.* [92,98]. Further data mining and follow-up of breast cancer GWAS by Zheng *et al.* identified a potential causal breast cancer susceptibility SNP in Chinese women at a chromosomal locus reported earlier by Stacey *et al.* (16q12.1) [86,92,99]. Cai *et al.* conducted breast cancer GWAS by genotyping breast cancer cases and healthy controls of Chinese ancestry on Affymetrix SNP 6.0 array and

[48] Imputation is a process by which missing values (*i.e.*, missing genotype calls) are replaced with best possible substitution values.
[49] Arrays approximately 500,000 SNPs for genotyping.
[50] Arrays over 900,000 SNPs and 900,000 copy number probes for both SNP and copy number genotyping.

identified a novel breast cancer susceptibility locus on chromosome 10q21.2, which contained a zinc finger protein encoded by the *ZNF365*[51] gene [100]. Kim *et al.* interrogated 555,525 SNPs genotyped on Affymetrix SNP 6.0 array using breast cancer cases and healthy controls of Korean ancestry and identified a novel breast cancer susceptibility locus on chromosome 2q34, which contained the *ERBB4*[52] gene [101]. This study also successfully reproduced previously GWAS-identified breast cancer susceptibility loci reported by international consortia[53] [84,85,87,88,91].

More recently, Ghoussaini *et al.* combined data from multiple independent breast cancer GWASs and conducted a large-scale replication study that identified three novel breast cancer susceptibility loci on chromosomes 12p11, 12q24 and 21q21 in women with European ancestry [102]. SNPs on 12q24 and 21q21 showed stronger associations for susceptibility to ER-positive than to ER-negative breast cancers whereas a SNP on 12p11 conferred similar risk for both ER-positive and ER-negative breast cancers. This was by far the largest GWAS conducted for breast cancer predisposition until early 2013.

1.3.2.7 Common variants for breast cancer prognosis[54]

Successes from GWASs in identifying low-penetrance common variants for breast cancer predisposition led to investigations examining potential roles of

---

[51] Zinc finger protein 365.
[52] v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian).
[53] The CGEMS, the Breast Cancer Association Consortium (BCAC), the Effectiveness of Additional Reductions in Cholesterol and Homocysteine Collaborative Group, the Nurses' Health Study (NHS) and the National Heart, Lung and Blood Institute Framingham Heart Study.
[54] Natural progression of a disease in absence of treatments.

GWASs-identified breast cancer susceptibility SNPs for breast cancer prognosis. None of the breast cancer susceptibility loci reported by GWASs showed significant associations with breast cancer prognosis, except for a SNP (rs13281615) on chromosome 8q24 reported by Easton *et al.*, which showed statistically significant association with overall survival (P=0.009) in an independent study comprising 13,527 invasive breast cancer cases [103]. In 2010, Azzato *et al.* conducted the first GWAS for breast cancer survival after diagnosis using the follow-up and genotype data of 528,252 SNPs for 1,145 postmenopausal sporadic breast cancer cases from the CGEMS initiative [88,104]. However, the results did not find any SNPs statistically significantly associated with breast cancer prognosis. The authors concluded that a different set of low penetrance common alleles, rather than susceptibility alleles, may be responsible for variations in breast cancer prognosis. In the same year, Azzato *et al.* conducted a second GWAS for breast cancer prognosis by using existing stage 1 GWAS data from Easton *et al.* that consisted of 3,761 invasive breast cancer cases genotyped for 10,621 SNPs on custom-based Perlegen platform [84,105]. The authors reported a SNP (rs4778137) on chromosome 15q13.1 as statistically significantly associated with breast cancer survival for triple-negative breast cancer cases ($P=5.0 \times 10^{-5}$) and the association was successfully replicated in an independent set of 14,096 invasive breast cancer cases. Subsequently, a third GWAS for breast cancer prognosis was conducted by Shu *et al.* in 2012 by interrogating 613,031 SNPs genotyped on Affymetrix SNP 6.0 array for 6,110 invasive breast cancer cases of Chinese ancestry [106]. The results indicated two

SNPs, rs3784099 and rs9934948, located on chromosomes 14 and 16, respectively, as significantly associated with breast cancer survival ($P<5.0 \times 10^{-6}$). Overall, these three studies have shown that inherited germline genetic variations may contribute to the observed variations in breast cancer prognosis and hence larger GWASs in future are warranted to identify additional common variants for breast cancer outcomes.

1.3.2.8 Summary of genetic risk accounted to date and the search for missing heritability

Over the last five years (2007-2012), several GWASs and a candidate-gene association study conducted for breast cancer led to identification of multiple low penetrance common variants conferring single-locus effects for breast cancer risk, lending credence to the CDCV hypothesis and polygenic model of risk for complex diseases [42,43,64]. Except for a breast cancer susceptibility locus at chromosome 2q34 (MAF~5%) reported by Kim *et al.* in the Korean population [101], breast cancer associated SNPs were common in the study population with MAF>10% [64,84-99,102]. However, the effect sizes of these associations were very small, ranging from 1.07 to 1.53 (expressed as ORs), and explain approximately 10% of additional genetic risk for breast cancer predisposition [102]. Taken together, known high and moderate penetrance genes identified through linkage studies and mutational screenings of candidate genes, in addition to recently identified low penetrance common SNPs by genetic association studies, only account for less than 35% of variations in breast cancer predisposition, suggesting that more variants exist.

One of the major challenges to uncover the remainder of breast cancer heritability is the sample size needed for sufficient statistical power since the (yet unidentified) common variants are expected to confer much smaller effect sizes (ORs<1.3). International consortia such as the BCAC, the NHS, the CGEMS and the Breast and Prostate Cancer Cohort Consortia have already made an effort to increase the number of studied individuals to approximately 150,000 by including breast cancer cases and healthy controls from several individual research centers [102]. However, results from such giant consortia were also limited to SNPs with very small effect sizes, indicating that future GWASs are unlikely to identify common variants with very large individual effect sizes (ORs>1.5), regardless of sufficiently large sample sizes. Consequently, there is a clear need to explore other forms of genetic variations contributing to breast cancer predisposition.

Current debates suggest that one of the possible sources for "residual or missing heritability" for breast cancer are contributions of structural variations, such as copy number variations (CNVs), and genetic interactions (gene-gene and gene-environment interactions), in addition to the contributions from strong single-locus effects through continued efforts for sufficiently powered systematic GWASs [107-115]. Such comprehensive approach may identify a larger proportion of breast cancer heritability, leading to possibilities for population level screening and prophylactic interventions in the near future.

1.3.2.9 Copy number variations

One possible source of residual heritability for breast cancer is the contribution of copy number variations (CNVs). CNVs are the most common type

of structural variations in the human genome. These are DNA segments of more than one kilobase in size that vary in their copy numbers due to gains or losses (**Figure 1-4**) [116-118]. Throughout this thesis, I refer to literature and my own results from germline CNVs[55] (and not CNVs from tumor cells/somatic origins) as is my focus with SNPs for their potential value in disease susceptibility or prognosis. As of November 2012, there were 291,801 CNVs reported in the Database of Genomic Variants, Toronto, Canada, a curated catalog of human genomic structural variation and more CNVs may be identified in the coming years. This catalogue is by no means a complete database, but is continually evolving. CNVs are believed to affect expression of many genes, either through gene dosage (gains or losses) or by *cis*-acting regulatory activities [116,119,120]. Studies have shown that germline CNVs may predispose to many complex diseases and SNPs are generally underrepresented in genomic regions harboring CNVs and therefore, GWASs utilizing CNVs are slowly emerging [121-125].

[55]The DNA an individual is born with, which does not change in one's lifetime is referred to as germline (or constitutive) DNA. DNA extracted from blood lymphocytes is considered representative of the germline status. There is ample literature on the CNVs from cancerous cells and is not the focus of this thesis.

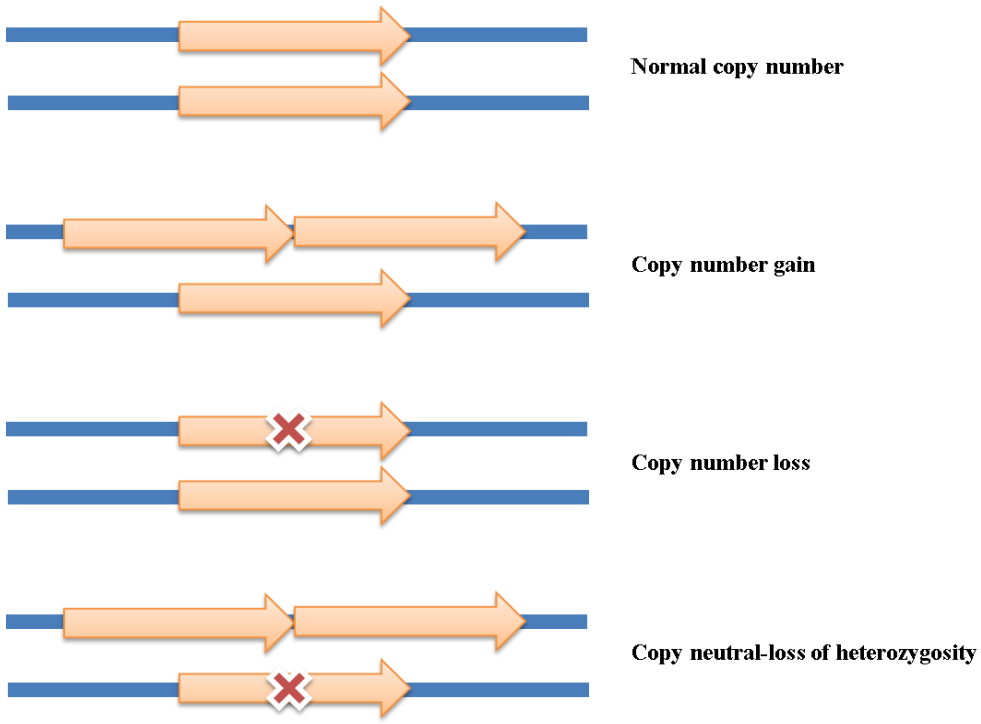


Figure 1-4 Illustration of copy number variations and copy neutral-loss of heterozygosities.

The first large-scale GWAS investigating potential roles of CNVs for genetic susceptibility to complex diseases was conducted by the Welcome Trust Case-Control Consortium in 2010 [126]. The study investigated 3,432 common CNVs for their roles in seven diseases and found no strong evidences for CNVs as better (than SNPs) sources of residual heritability for complex diseases. Even though the study reported lack of associations of common CNVs with complex diseases, perhaps due to limited number of common CNVs considered in the analysis, the study certainly opened possibilities for more systematic and comprehensive analysis (larger coverage) of germline CNVs as predisposition factors for complex diseases.

Another study that evaluated germline CNV profiles between *BRCA1*-associated and sporadic ovarian cancer patients reported substantial differences in copy number gains and losses between these two groups of cancer patients [127]. Germline CNVs were also reported to be associated with susceptibility to familial pancreatic and breast cancers [128].

More recently with the application of high-throughput SNP genotyping arrays, large chromosomal lesions characterized by loss of heterozygosity (LOH) but with diploid copy numbers were observed (as were also in many malignancies), possibly resulting from non-homologous recombination during meiosis, trisomic rescue or mitotic recombination [129-134]. These chromosomal defects, also known as copy-neutral loss of heterozygosities (CN-LOHs) or uniparental disomies (UPDs), are characterized by loss of one allele with simultaneous replacement by an exact copy of another allele, resulting in retention of diploid copy number but loss of polymorphic differences (both alleles are from the same parent) (**Figure 1-4**). CN-LOHs have been reported to be associated with gain of oncogenic alleles and inactivation of tumor suppressors and may be an important mechanism in cancer development [131-134]. With the advent of SNP genotyping platforms that can measure both CNVs and CN-LOHs, it is now possible to investigate potential roles of these large chromosomal defects as genetic determinants for complex diseases using germline DNA. This aspect will be addressed in Chapter 5 of this thesis.

1.3.2.10 Genetic interactions

GWASs representing common SNPs and emerging studies of CNVs or CN-LOHs primarily focus on single-locus effects (also known as main genetic effects). However, risk for complex diseases, including breast cancer, is also attributed to two types of genetic interactions: gene-gene and gene-environment interactions. At present, GWASs inclusive of these genetic interactions are limited because of the need for large sample sizes to achieve the statistical power as well as the need for exposure data (health, lifestyle and reproductive) that are difficult to obtain and, where available, may not have banked DNA in most cohorts.

1.3.2.10.1 Gene-gene interactions

The etiology of complex diseases includes substantial proportion of gene-gene interactions, commonly referred to epistasis. Epistasis is a ubiquitous phenomenon that describes how genes or loci interact to affect phenotypes [114,115]. Such interactions are believed to explain a large proportion of genetic heritability of complex diseases.

Let us assume that SNP A has an effect size of 1.5 and SNP B has an effect size of 1.5. According to the additive model of genetic inheritance, the cumulative genetic effect from SNP A and SNP B would be 1.5+1.5=3. However, if epistasis is present between SNP A and SNP B, the cumulative genetic effects would not be 3 but rather something more or less than 3. In other words, epistasis is a departure from a simple additive model that considers the combined effects of individual single-locus effects.

At present, epistatic interactions involving two loci or SNPs can be evaluated using logistic regressions. However, it has been limited to candidate-gene studies with a small number of SNPs [135]. Testing every combination of pairwise interactions or even extending to multi-SNPs (multi-way) interactions in a GWAS is computationally intensive. More recently, logic regressions have been proposed for testing multi-way interactions among SNPs and have been successfully applied to GWAS of Crohn's disease and a candidate-gene association study of cervical cancer [136,137]. Consequently, future studies that focus on potential epistatic interactions among SNPs, in addition to single-locus effects, in a GWAS or a candidate-gene association study may identify additional heritability for complex diseases, including breast cancer. Chapters 3 and 4 will address this form of genetic interaction for breast cancer predisposition.

1.3.2.10.2 Gene-environment interactions

Complex diseases such as breast cancer also result from combined effects of both genetic and environmental risk factors. Even though these forms of genetic interactions are believed to explain a large proportion of heritability for breast cancer, especially for sporadic breast cancer, investigations of such interactions are limited due to difficulty in obtaining the environmental, lifestyle and reproductive data.

Recently, the Breast and Prostate Cancer Cohort Consortium conducted a comprehensive study to evaluate possible interactions between the common breast cancer susceptibility loci reported earlier by GWASs and the established breast cancer risk factors, such as age at menarche, parity, age at menopause, use of

hormone replacement therapy, body mass index, smoking habit, alcohol consumption, family history and height using data from 8,576 cases and 11,892 controls [138]. The study findings indicated that the common breast cancer susceptibility loci (single-locus associations) do not affect the associations of the examined established risk factors with breast cancer. These findings were also supported by another study conducted by the BCAC that evaluated possible interactions among common breast cancer susceptibility loci and known breast cancer risk factors, using genotype and questionnaire data from 26,349 cases and 32,208 controls from 21 case-control studies [139]. These results indicate that there would be a different set of common SNPs involved in gene-environment interactions contributing to breast cancer predisposition than the susceptibility SNPs. Once sufficient exposure data becomes readily accessible to incorporate in GWAS, future studies may also focus on this form of genetic interaction to address the residual heritability of breast cancer and is currently beyond the scope of this thesis.

1.4 Hypothesis

Common germline DNA variations (SNPs, CNVs and CN-LOHs) are genetic determinants and hence contribute to breast cancer predisposition and disease prognosis, either through single-locus or epistatic effects.

1.5 Specific objectives

Specific objectives of this thesis were as follows:

- To identify common breast cancer susceptibility loci for sporadic breast cancer using a multi-stage association study design (addressed in Chapter 2)
- To evaluate variations in breast cancer susceptibility based on known clinicopathological characteristics, such as hormonal status, menopausal status, family history, tumor stage and grade (addressed in Chapters 2 and 3)
- To identify germline markers (SNPs and structural variations) for breast cancer prognosis (addressed in Chapters 2 and 5)
- To evaluate potential epistatic interactions contributing to breast cancer susceptibility (addressed in Chapters 3 and 4)

1.6 Organization of the thesis

The thesis has been organized into four distinct chapters that will address the specific objectives of this thesis.

Earlier, the Damaraju Laboratory conducted a GWAS for sporadic breast cancer (late onset of disease and absence of family history of breast cancer) by genotyping 906,600 SNPs on Affymetrix SNP 6.0 array in 348 invasive breast cancer cases and 348 healthy controls of predominantly Caucasian origin from Alberta, Canada [140]. After removing potential non-Caucasian subjects during the population stratification step (data clean-up process), the discovery stage included 302 cases and 321 controls for the final analysis (stage 1). After performing the data clean-up step, a total of 782,838 SNPs were considered for single-locus tests, which resulted in 35,859 SNPs at statistical significance

P<0.05. Of these, the 35 most promising SNPs (with the strongest statistical significance *P* values among the analyzed SNPs and in strong LD with nearby markers) were replicated in an independent set of 1,153 breast cancer cases and 1,215 controls (stage 2). Six SNPs survived the replication showing statistical significant associations for breast cancer susceptibility.

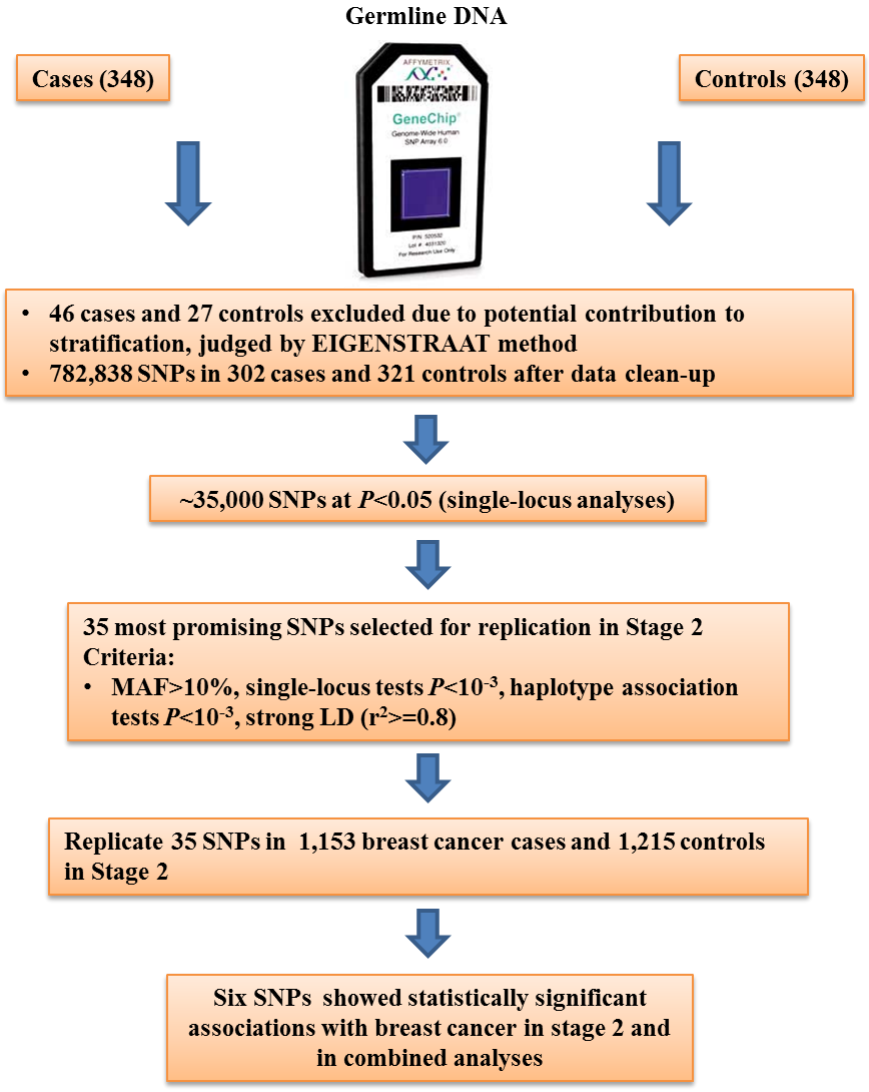


Figure 1-5 An overview of two-stage GWAS for sporadic breast cancer conducted by the Damaraju Laboratory.

An overview of GWAS for sporadic breast cancer conducted in the Damaraju laboratory is illustrated in **Figure 1-5**. The stage 1 data shown in **Figure 1-5** was used in the discovery stage for data mining and further follow-up described in Chapters 2 and 3.

In Chapter 2, I further investigated the observed associations of six putative sporadic breast cancer susceptibility loci using an independent set of 1,294 breast cancer cases and 2,934 controls (stage 3). A combined analysis (stages 1+2+3) was also performed to increase the statistical power. Two SNPs showed statistical significant associations for breast cancer predisposition and of these, a SNP attained commonly adopted genome-wide significance level in combined analysis, adjusted for BMI. Further, I also investigated the robustness of associations of 11 common breast cancer susceptibility SNPs with breast cancer reported by large consortia during the years 2007-2009 in Alberta women. In an attempt to evaluate variations in risk conferred by common breast cancer susceptibility SNPs, I conducted stratified analyses based on luminal A[56] status, menopausal status, family history of breast cancer, tumor stage and grade. Potential associations of common breast cancer susceptibility SNPs with breast cancer outcomes, such as recurrence-free survival and overall survival were also examined. An overview of the study design is provided in **Figure 1-6**.

[56] Breast cancer cases with either ER or PR status positive and HER2 status negative were considered as luminal A cases.

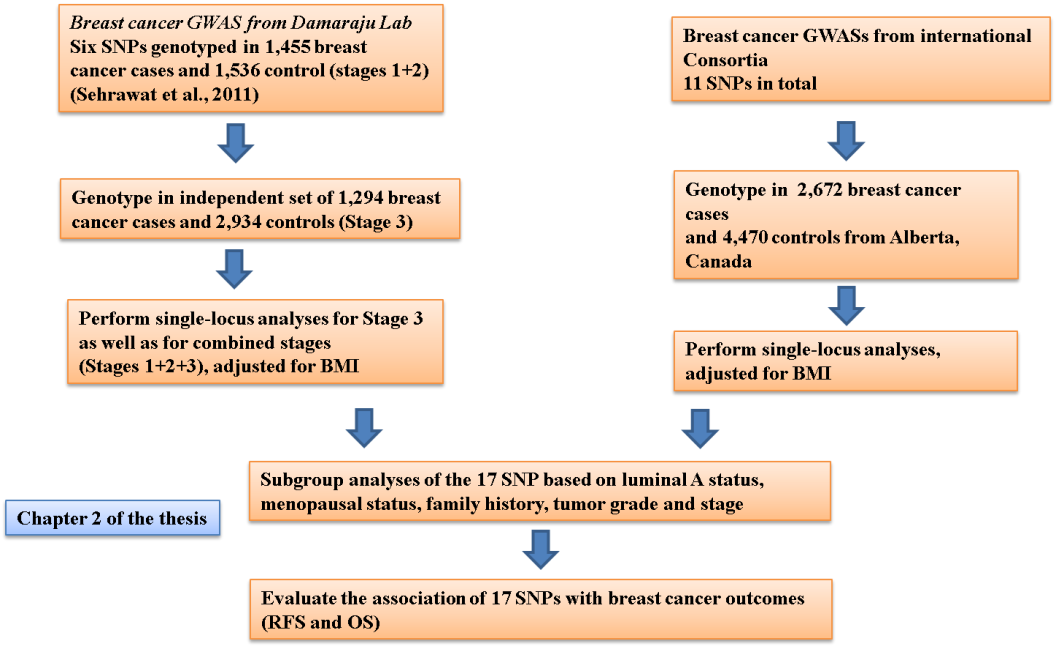


Figure 1-6 An overview of the study design adopted for Chapter 2 of the thesis.

In Chapter 3, I attempted to investigate potential epistatic interactions among common SNPs. I hypothesized that SNPs showing consistently reproducible moderate single-locus effects with weak statistical significance in both discovery and independent replication stages may be useful candidates for evaluating epistatic interactions. This definition for qualifying SNPs for investigations of epistasis interaction analyses were also reported by others [135,141]. Therefore, I first selected 215 SNPs in or close to DNA repair, modification or metabolism pathway related genes for their overall role in breast cancer, from the list of 35,859 SNPs at $P<0.05$ in the single-locus tests of stage 1 GWAS reported by the Damaraju Laboratory [140]. Of these, 22 SNPs (with the strongest statistical significance *P* values among the 215 SNPs and in strong LD with nearby SNPs) were replicated in an independent set of 1,178 breast cancer cases and 1,314

healthy controls from Alberta, Canada. A combined analysis comprising stages 1+2 was also performed to increase statistical power. Six SNPs showed consistent moderate single-locus effects with weak statistical significance for breast cancer across stages 1 and 2 and hence are likely candidates for epistatic interaction analyses. Subgroup analyses based on luminal A status, family history and menopausal status were also performed to examine variations in observed breast cancer susceptibility. An overview of the study design is provided in **Figure 1-7**.

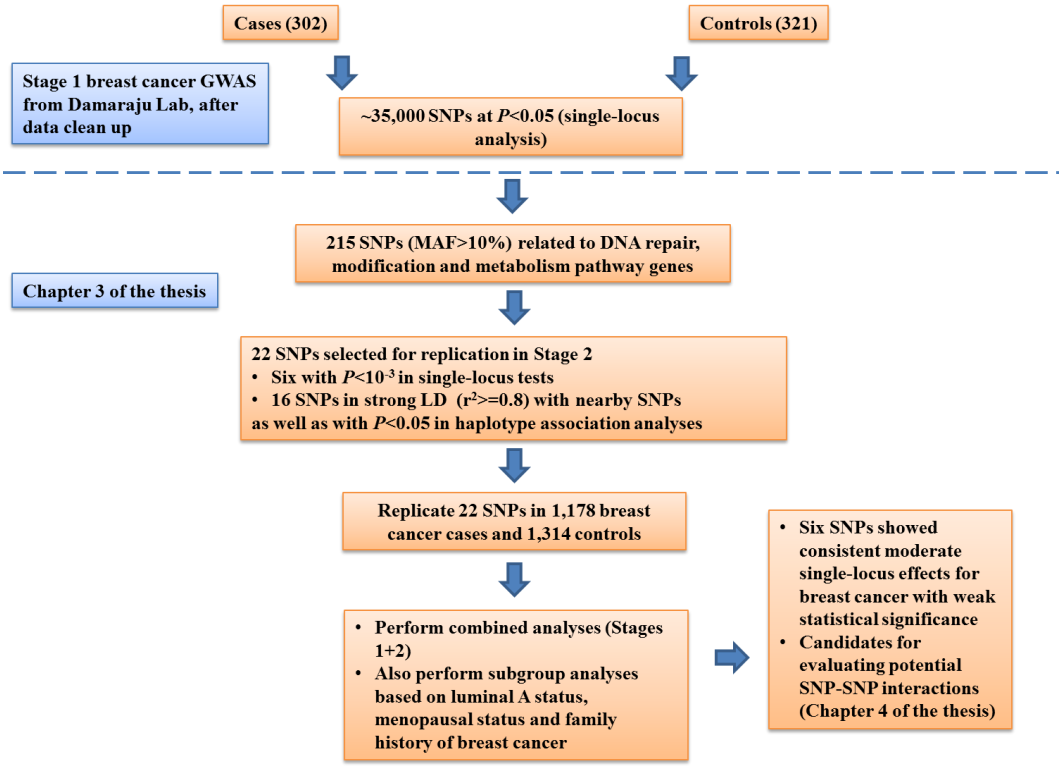


Figure 1-7 An overview of the experimental design used in Chapter 3 of the thesis.

In Chapter 4, I attempted epistatic interaction analyses among the six SNPs identified in Chapter 3, in addition to 11 candidate DNA repair SNPs with prior evidence of their association with breast cancer [45,142-149], using 2,795 breast

cancer cases and 4,505 controls. Logistic regressions were used to assess two-way interactions among the 17 SNPs whereas logic regression was implemented to evaluate interactions involving more than two SNPs. A total of six SNPs were observed in two two-way interactions and a SNP-SNP interaction involving four SNPs were observed. These six SNPs also showed moderate effects with weak statistical significance in single-locus tests. An overview of the study design is provided in **Figure 1-8**.

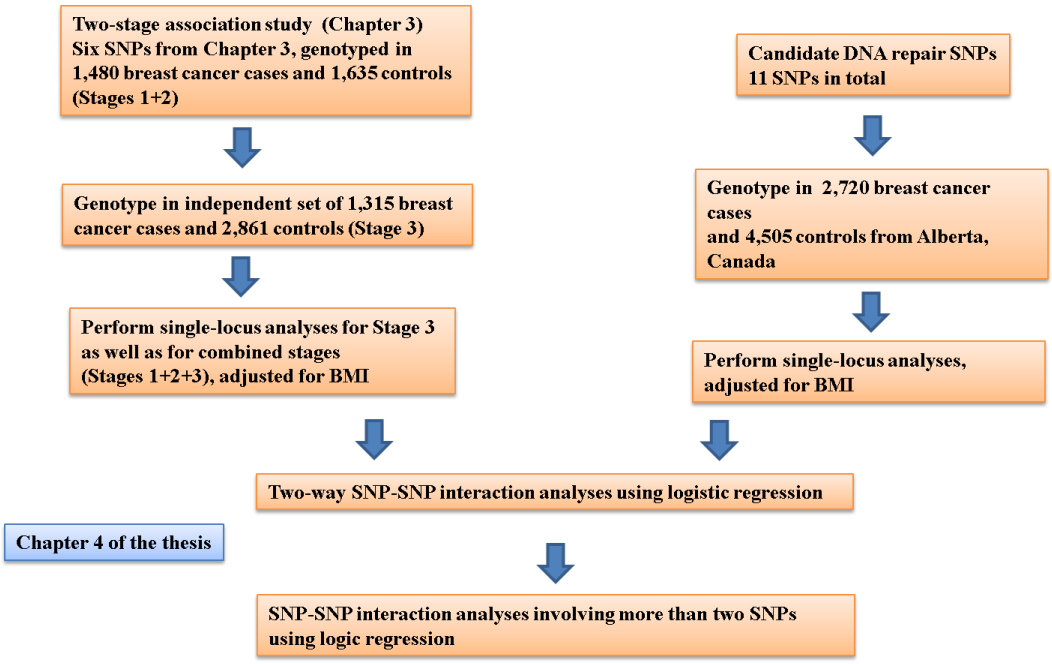


Figure 1-8 An overview of study design used in Chapter 4 of the thesis.

In Chapter 5, I conducted a GWAS utilizing germline CNVs and CN-LOHs (hereafter referred to as copy number aberrations, CNAs) for breast cancer recurrence, using 155 recurred cases and 214 non-recurred cases (a cut-off of at least three years of follow-up for inclusion criteria for a case as non-recurred). I focused on common CNAs (>10% frequency of occurrence) for their potential

associations with breast cancer recurrence. Ten CNAs (two copy number gains and eight CN-LOHs) showed statistical significance difference in recurrence-free survival probabilities in cases with and without the CNA. Of these, I validated three CN-LOHs in a subset of randomly selected cases using real-time polymerase chain reaction. An overview of the study design is provided in **Figure 1-9**.

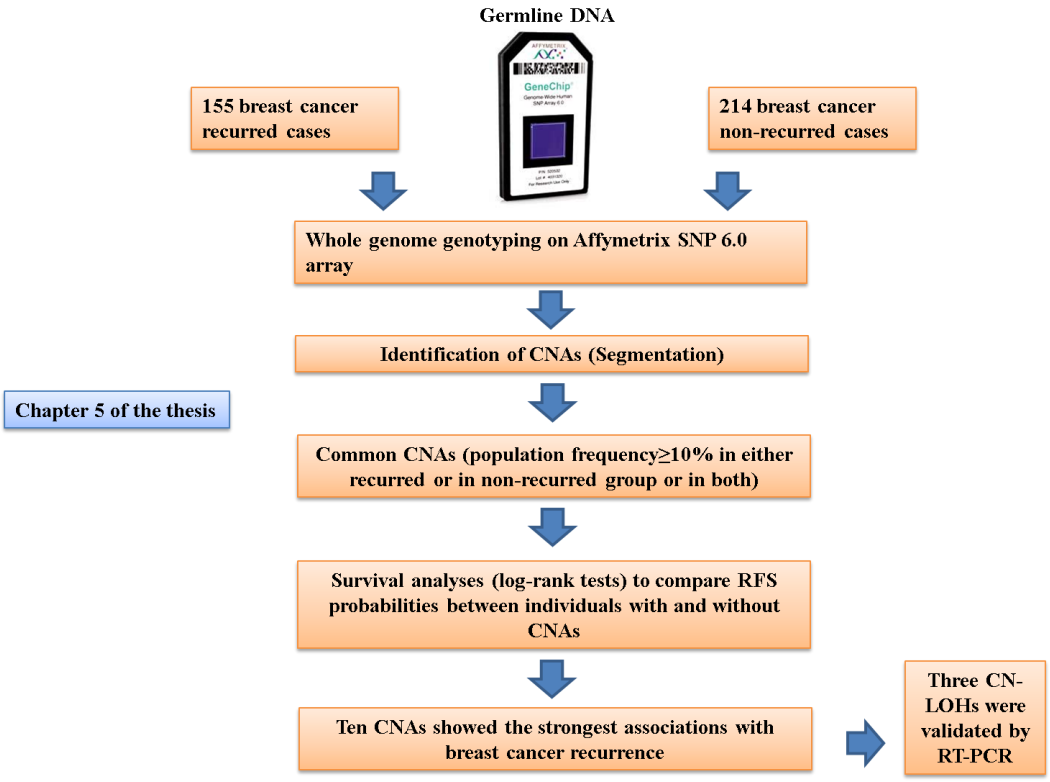


Figure 1-9 Study design for Chapter 5 of the thesis.

In Chapters 6 and 7, I discuss in brief about the overall findings of the thesis, in addition to future work that may be carried out to explore more about the genetic predisposition for breast cancer.

1.7 References

1. Cancer Research UK. (2012) Breast cancer - key facts.

2. Siegel R, Naishadham D, Jemal A. (2012) Cancer statistics, 2012. CA Cancer J Clin 62: 10-29.

3. Canadian Cancer Society's Steering Committee on Cancer Statistics. (2012) Canadian cancer statistics 2012, toronto. ON: Canadian cancer society; 2012.

4. Early Breast Cancer Trialists' Collaborative Group (EBCTCG). (2005) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomised trials. Lancet 365: 1687-1717.

5. Weigel MT, Dowsett M. (2010) Current and emerging biomarkers in breast cancer: Prognosis and prediction. Endocr Relat Cancer 17: R245-62.

6. van der Leij F, Elkhuizen PH, Bartelink H, van de Vijver MJ. (2012) Predictive factors for local recurrence in breast cancer. Semin Radiat Oncol 22: 100-107.

7. Voduc KD, Cheang MC, Tyldesley S, Gelmon K, Nielsen TO, et al. (2010) Breast cancer subtypes and the risk of local and regional relapse. J Clin Oncol 28: 1684-1691.

8. Gonzalez-Angulo AM, Morales-Vasquez F, Hortobagyi GN. (2007) Overview of resistance to systemic therapy in patients with breast cancer. Adv Exp Med Biol 608: 1-22.

9. Collaborative Group on Hormonal Factors in Breast Cancer. (2001) 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 358: 1389-1399.

10. Key TJ, Verkasalo PK, Banks E. (2001) Epidemiology of breast cancer. Lancet Oncology 2: 133-140.
11. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from sweden, denmark, and finland. N Engl J Med 343: 78-85.
12. McPherson K, Steel CM, Dixon JM. (2000) ABC of breast diseases. breast cancer-epidemiology, risk factors, and genetics. BMJ 321: 624-628.
13. Garcia-Closas M, Brinton LA, Lissowska J, Chatterjee N, Peplonska B, et al. (2006) Established breast cancer risk factors by clinically important tumour characteristics. Br J Cancer 95: 123-129.
14. Key T, Appleby P, Barnes I, Reeves G, Endogenous Hormones and Breast Cancer Collaborative Group. (2002) Endogenous sex hormones and breast cancer in postmenopausal women: Reanalysis of nine prospective studies. J Natl Cancer Inst 94: 606-616.
15. Pharoah PDP, Day NE, Duffy S, Easton DF, Ponder BAJ. (1997) Family history and the risk of breast cancer: A systematic review and meta-analysis. International Journal of Cancer 71: 800-809.
16. Reeves GK, Pirie K, Beral V, Green J, Spencer E, et al. (2007) Cancer incidence and mortality in relation to body mass index in the million women study: Cohort study. BMJ 335: 1134.

17. Monninkhof EM, Elias SG, Vlems FA, van der Tweel I, Schuit AJ, et al. (2007) Physical activity and breast cancer - A systematic review. Epidemiology 18: 137-157.

18. Robinson D, Holmberg L, Moller H. (2008) The occurrence of invasive cancers following a diagnosis of breast carcinoma in situ. Br J Cancer 99: 611-615.

19. Easton DF, Bishop DT, Ford D, Crockford GP. (1993) Genetic linkage analysis in familial breast and ovarian cancer: Results from 214 families. the breast cancer linkage consortium. Am J Hum Genet 52: 678-701.

20. MORTON NE. (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7: 277-318.

21. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250: 1684-1689.

22. Solomon E, Ledbetter DH. (1990) Report of the committee on the genetic constitution of chromosome 17. Cytogenet Cell Genet 55: 198-215.

23. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, et al. (1995) Identification of the breast cancer susceptibility gene BRCA2. Nature 378: 789-792.

24. Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, et al. (2007) A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. Am J Hum Genet 81: 873-883.

25. Easton DF, Ford D, Bishop DT. (1995) Breast and ovarian cancer incidence in BRCA1-mutation carriers. breast cancer linkage consortium. Am J Hum Genet 56: 265-271.

26. Ford D, Easton DF, Bishop DT, Narod SA, Goldgar DE. (1994) Risks of cancer in BRCA1-mutation carriers. breast cancer linkage consortium. Lancet 343: 692-695.

27. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, et al. (1998) Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. the breast cancer linkage consortium. Am J Hum Genet 62: 676-689.

28. Thompson D, Easton DF, Breast Cancer Linkage Consortium. (2002) Cancer incidence in BRCA1 mutation carriers. J Natl Cancer Inst 94: 1358-1365.

29. Peto J, Collins N, Barfoot R, Seal S, Warren W, et al. (1999) Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. J Natl Cancer Inst 91: 943-949.

30. Easton DF. (1999) How many more breast cancer predisposition genes are there? Breast Cancer Res 1: 14-17.

31. Antoniou A, Gayther S, Stratton J, Ponder B, Easton D. (2000) Risk models for familial ovarian and breast cancer. Genet Epidemiol 18: 173-190.

32. Antoniou A, Pharoah P, McMullan G, Day N, Stratton M, et al. (2002) A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br J Cancer 86: 76-83.

33. Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, et al. (1997) Germline mutations of the PTEN gene in cowden disease, an inherited breast and thyroid cancer syndrome. Nat Genet 16: 64-67.

34. Malkin D, Li FP, Strong LC, Fraumeni JF,Jr, Nelson CE, et al. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250: 1233-1238.

35. Borresen AL, Andersen TI, Garber J, Barbier-Piraux N, Thorlacius S, et al. (1992) Screening for germ line TP53 mutations in breast cancer patients. Cancer Res 52: 3234-3236.

36. CHEK2 Breast Cancer Case-Control Consortium. (2004) CHEK2*1100delC and susceptibility to breast cancer: A collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. Am J Hum Genet 74: 1175-1182.

37. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, et al. (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet 39: 165-167.

38. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, et al. (2006) Truncating mutations in the fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet 38: 1239-1241.

39. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, et al. (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet 38: 873-875.

40. Smith P, McGuffog L, Easton DF, Mann GJ, Pupo GM, et al. (2006) A genome wide linkage search for breast cancer susceptibility genes. Genes Chromosomes & Cancer 45: 646-655.

41. Chen GK, Jorgenson E, Witte JS. (2007) An empirical evaluation of the common disease-common variant hypothesis. BMC Proc 1 Suppl 1: S5.

42. Hemminki K, Forsti A, Bermejo JL. (2008) The 'common disease-common variant' hypothesis and familial risks. PLoS One 3: e2504.

43. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, et al. (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31: 33-36.

44. Piccolo SR, Camp N, Frey LJ. (2008) Polygenic model for predicting breast cancer risk via genome-wide polymorphisms. AMIA Annu Symp Proc : 1094.

45. Smith TR, Levine EA, Freimanis RI, Akman SA, Allen GO, et al. (2008) Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. Carcinogenesis 29: 2132-2138.

46. Pharoah PDP, Antoniou AC, Easton DF, Ponder BAJ. (2008) Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 358: 2796-2803.

47. Hindorff LA, MacArthur, J. (European Bioinformatics Institute), Morales, J. (European Bioinformatics Institute), Junkins HA, Hall PN, et al. Catalog of published genome-wide association studies. 2012.

48. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933.

49. Bentley D. (2000) The human genome project - an overview. Med Res Rev 20: 189-196.

50. Morgan M, Wallace S. (2004) The international human genome project: An overview. . 23 p.

51. National Center for Biotechnology Information. (June 26, 2012) dbSNP build 137. 2012.

52. Bau DT, Fu YP, Chen ST, Cheng TC, Yu JC, et al. (2004) Breast cancer risk and the DNA double-strand break end-joining capacity of nonhomologous end-joining genes are affected by BRCA1. Cancer Res 64: 5013-5019.

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

54. Lipponen P, Aaltomaa S, Kosma VM, Syrjanen K. (1994) Apoptosis in breast cancer as related to histopathological characteristics and prognosis. Eur J Cancer 30A: 2068-2073.

55. Kastan M, Bartek J. (2004) Cell-cycle checkpoints and cancer. Nature 432: 316-323.

56. Bewick MA, Conlon MS, Lafrenie RM. (2006) Polymorphisms in XRCC1, XRCC3, and CCND1 and survival after treatment for metastatic breast cancer. Journal of Clinical Oncology 24: 5645-5651.

57. Allen-Brady K, Cannon-Albright LA, Neuhausen SL, Camp NJ. (2006) A role for XRCC4 in age at diagnosis and breast cancer risk. Cancer Epidemiology, Biomarkers & Prevention 15: 1306-1310.

58. Haiman CA, Hsu C, de Bakker PI, Frasco M, Sheng X, et al. (2008) Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations. Hum Mol Genet 17: 825-834.

59. Pooley KA, Baynes C, Driver KE, Tyrer J, Azzato EM, et al. (2008) Common single-nucleotide polymorphisms in DNA double-strand break repair genes and breast cancer risk. Cancer Epidemiology, Biomarkers & Prevention 17: 3482-3489.

60. Mangoni M, Bisanzi S, Carozzi F, Sani C, Biti G, et al. (2010) Association between genetic polymorphisms in the XRCC1, XRCC3, XPD, GSTM1, GSTT1, MSH2, MLH1, MSH3, and MGMT genes and radiosensitivity in breast cancer patients. Int J Radiat Oncol Biol Phys 81: 52-58.

61. Sehl ME, Langer LR, Papp JC, Kwan L, Seldon JL, et al. (2009) Associations between single nucleotide polymorphisms in double-stranded DNA repair pathway genes and familial breast cancer. Clinical Cancer Research 15: 2192-2203.

62. Lin WY, Camp NJ, Cannon-Albright LA, Allen-Brady K, Balasubramanian S, et al. (2011) A role for XRCC2 gene polymorphisms in breast cancer risk and survival. J Med Genet 48: 477-484.

63. Dunning A, Healey C, Pharoah P, Teare M, Ponder B, et al. (1999) A systematic review of genetic polymorphisms and breast cancer risk. Cancer Epidemiology Biomarkers & Prevention 8: 843-854.

64. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, et al. (2007) A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet 39: 352-358.

65. Palanca Suela S, Esteban Cardenosa E, Barragan Gonzalez E, de Juan Jimenez I, Chirivella Gonzalez I, et al. (2010) CASP8 D302H polymorphism delays the age of onset of breast cancer in BRCA1 and BRCA2 carriers. Breast Cancer Res Treat 119: 87-93.

66. Reich D, Cargill M, Bolk S, Ireland J, Sabeti P, et al. (2001) Linkage disequilibrium in the human genome. Nature 411: 199-204.

67. Stram DO. (2004) Tag SNP selection for association studies. Genet Epidemiol 27: 365-374.

68. Montpetit A, Nelis M, Laflamme P, Magi R, Ke X, et al. (2006) An evaluation of the performance of tag SNPs derived from HapMap in a caucasian population. Plos Genetics 2: 282-290.

69. Gibbs R, Belmont J, Hardenbol P, Willis T, Yu F, et al. (2003) The international HapMap project. Nature 426: 789-796.

70. Tuma RS. (2009) Genome-wide association studies provoke debate and a new look at strategy. J Natl Cancer Inst 101: 1041-1043.

71. Hardy J, Singleton A. (2009) Genomewide association studies and human disease. N Engl J Med 360: 1759-1768.

72. Cardon LR, Bell JI. (2001) Association study designs for complex diseases. Nature Reviews Genetics 2: 91-99.

73. Hirschhorn JN, Daly MJ. (2005) Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6: 95-108.

74. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-2229.

75. NCI-NHGRI Working Group on Replication in Association,Studies, Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, et al. (2007) Replicating genotype-phenotype associations. Nature 447: 655-660.

76. Skol A, Scott L, Abecasis G, Boehnke M. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38: 209-213.

77. Menashe I, Rosenberg PS, Chen BE. (2008) PGA: Power calculator for case-control genetic association analyses. BMC Genet 9: 36.

78. Edwards TL, Gao X. (2012) Methods for detecting and correcting for population stratification. Curr Protoc Hum Genet Chapter 1: Unit 1.22.1-14.

79. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.

80. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, et al. (2011) Basic statistical analysis in genetic case-control studies. Nat Protoc 6: 121-133.

81. Watanabe RM. (2011) Statistical issues in gene association studies. Methods Mol Biol 700: 17-36.

82. Bernstein L, Lacey JV,Jr. (2011) Receptors, associations, and risk factor differences by breast cancer subtypes: Positive or negative? J Natl Cancer Inst 103: 451-453.

83. Kristensen VN, Borresen-Dale AL. (2008) SNPs associated with molecular subtypes of breast cancer: On the usefulness of stratified genome-wide association studies (GWAS) in the identification of novel susceptibility loci. Molecular Oncology 2: 12-15.

84. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087-U7.

85. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, et al. (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41: 585-590.

86. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, et al. (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39: 865-869.

87. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, et al. (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 40: 703-706.

88. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870-874.

89. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, et al. (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 41: 579-584.

90. Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, et al. (2011) Novel breast cancer susceptibility locus at 9q31.2: Results of a genome-wide association study. J Natl Cancer Inst 103: 425-435.

91. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, et al. (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 42: 504-U47.

92. Zheng W, Long J, Gao YT, Li C, Zheng Y, et al. (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 41: 324-328.

93. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, et al. (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22-33. Proc Natl Acad Sci U S A 105: 4340-4345.

94. Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, et al. (2010) 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 42: 885-892.

95. Gaudet MM, Kirchhoff T, Green T, Vijai J, Korn JM, et al. (2010) Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. Plos Genetics 6: e1001183.

96. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, et al. (2011) A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nat Genet 43: 1210-U61.

97. Siddiq A, Couch FJ, Chen GK, Lindstrom S, Eccles D, et al. (2012) A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. Hum Mol Genet 21: 5373-5384.

98. Long J, Cai Q, Sung H, Shi J, Zhang B, et al. (2012) Genome-wide association study in east asians identifies novel susceptibility loci for breast cancer. Plos Genetics 8: e1002532.

99. Long J, Cai Q, Shu X, Qu S, Li C, et al. (2010) Identification of a functional genetic variant at 16q12.1 for breast cancer risk: Results from the asia breast cancer consortium. Plos Genetics 6: e1001002.

100. Cai Q, Long J, Lu W, Qu S, Wen W, et al. (2011) Genome-wide association study identifies breast cancer risk variant at 10q21.2: Results from the asia breast cancer consortium. Hum Mol Genet 20: 4991-4999.

101. Kim H, Lee J, Sung H, Choi J, Park SK, et al. (2012) A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: Results from the seoul breast cancer study. Breast Cancer Research 14: R56.

102. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, et al. (2012) Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet 44: 312-318.

103. Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, et al. (2008) Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genet 4: e1000054.

104. Azzato EM, Pharoah PDP, Harrington P, Easton DF, Greenberg D, et al. (2010) A genome-wide association study of prognosis in breast cancer. Cancer Epidemiology Biomarkers & Prevention 19: 1140-1143.

105. Azzato EM, Tyrer J, Fasching PA, Beckmann MW, Ekici AB, et al. (2010) Association between a germline OCA2 polymorphism at chromosome 15q13.1 and estrogen receptor-negative breast cancer survival. J Natl Cancer Inst 102: 650-662.

106. Shu XO, Long J, Lu W, Li C, Chen WY, et al. (2012) Novel genetic markers of breast cancer survival identified by a genome-wide association study. Cancer Res 72: 1182-1189.

107. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.

108. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446-450.

109. Gibson G. (2010) Hints of hidden heritability in GWAS. Nat Genet 42: 558-560.

110. [Anonymous]. (2010) On beyond GWAS. Nat Genet 42: 551.

111. Bodmer W, Tomlinson I. (2010) Rare genetic variants and the risk of cancer. Curr Opin Genet Dev 20: 262-267.

112. Thomas D. (2010) Gene--environment-wide association studies: Emerging approaches. Nat Rev Genet 11: 259-272.

113. Zuk O, Hechter E, Sunyaev SR, Lander ES. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109: 1193-1198.

114. Moore JH. (2005) A global view of epistasis. Nat Genet 37: 13-14.

115. Moore J. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered 56: 73-82.

116. Henrichsen CN, Chaignat E, Reymond A. (2009) Copy number variants, diseases and gene expression. Hum Mol Genet 18: R1-8.

117. Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, Geurts van Kessel A. (2010) Germline copy number variation and cancer risk. Curr Opin Genet Dev 20: 282-289.

118. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-454.

119. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848-853.

120. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, et al. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms

within duplicated regions of the human genome. Am J Hum Genet 79: 275-290.

121. McCarroll SA, Altshuler DM. (2007) Copy-number variation and association studies of human disease. Nat Genet 39: S37-42.

122. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40: 1166-1174.

123. Shlien A, Malkin D. (2010) Copy number variations and cancer susceptibility. Curr Opin Oncol 22: 55-63.

124. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nat Genet 37: 727-732.

125. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, et al. (2007) A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80: 91-104.

126. Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, et al. (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464: 713-720.

127. Yoshihara K, Tajima A, Adachi S, Quan J, Sekine M, et al. (2011) Germline copy number variations in BRCA1-associated ovarian cancer patients. Genes Chromosomes Cancer 50: 167-177.

128. Al-Sukhni W, Joe S, Lionel AC, Zwingerman N, Zogopoulos G, et al. (2012) Identification of germline genomic copy number variation in familial pancreatic cancer. Hum Genet 131: 1481-1494.

129. Kryh H, Caren H, Erichsen J, Sjoberg RM, Abrahamsson J, et al. (2011) Comprehensive SNP array study of frequently used neuroblastoma cell lines; copy neutral loss of heterozygosity is common in the cell lines but uncommon in primary tumors. BMC Genomics 12: 443.

130. Lapunzina P, Monk D. (2011) The consequences of uniparental disomy and copy number neutral loss-of-heterozygosity during human development and cancer. Biol Cell 103: 303-317.

131. Melcher R, Hartmann E, Zopf W, Herterich S, Wilke P, et al. (2011) LOH and copy neutral LOH (cnLOH) act as alternative mechanism in sporadic colorectal cancers with chromosomal and microsatellite instability. Carcinogenesis 32: 636-642.

132. Mohamedali A, Gaken J, Twine NA, Ingram W, Westwood N, et al. (2007) Prevalence and prognostic significance of allelic imbalance by single-nucleotide polymorphism analysis in low-risk myelodysplastic syndromes. Blood 110: 3365-3373.

133. O'Keefe C, McDevitt MA, Maciejewski JP. (2010) Copy neutral loss of heterozygosity: A novel chromosomal lesion in myeloid malignancies. Blood 115: 2731-2739.

134. Saeki H, Kitao H, Yoshinaga K, Nakanoko T, Kubo N, et al. (2011) Copy-neutral loss of heterozygosity at the p53 locus in carcinogenesis of

esophageal squamous cell carcinomas associated with p53 mutations. Clin Cancer Res 17: 1731-1740.

135. Onay VU, Briollais L, Knight JA, Shi E, Wang Y, et al. (2006) SNP-SNP interactions in breast cancer susceptibility. BMC Cancer 6: 114.

136. Dinu I, Mahasirimongkol S, Liu Q, Yanai H, Eldin NS, et al. (2012) SNP-SNP interactions discovered by logic regression explain crohn's disease genetics. Plos One 7: e43035.

137. Feng Q, Balasubramanian A, Hawes SE, Toure P, Sow PS, et al. (2005) Detection of hypermethylated genes in women with and without cervical neoplasia. J Natl Cancer Inst 97: 273-282.

138. Campa D, Kaaks R, Le Marchand L, Haiman CA, Travis RC, et al. (2011) Interactions between genetic variants and breast cancer risk factors in the breast and prostate cancer cohort consortium. J Natl Cancer Inst 103: 1252-1263.

139. Milne RL, Gaudet MM, Spurdle AB, Fasching PA, Couch FJ, et al. (2010) Assessing interactions between the associations of common genetic susceptibility variants, reproductive history and body mass index with breast cancer risk in the breast cancer association consortium: A combined case-control study. Breast Cancer Research 12: R110.

140. Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass CE, et al. (2011) Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Hum Genet 130: 529-537.

141. Lo S, Chernoff H, Cong L, Ding Y, Zheng T. (2008) Discovering interactions among BRCA1 and other candidate genes associated with sporadic breast cancer. Proc Natl Acad Sci U S A 105: 12387-12392.

142. Conde J, Silva SN, Azevedo AP, Teixeira V, Pina JE, et al. (2009) Association of common variants in mismatch repair genes and breast cancer susceptibility: A multigene study. BMC Cancer 9: 344-2407-9-344.

143. Boersma BJ, Howe TM, Goodman JE, Yfantis HG, Lee DH, et al. (2006) Association of breast cancer outcome with status of p53 and MDM2 SNP309. J Natl Cancer Inst 98: 911-919.

144. Gochhait S, Bukhari SIA, Bairwa N, Vadhera S, Darvishi K, et al. (2007) Implication of BRCA2-26G > A 5 ' untranslated region polymorphism in susceptibility to sporadic breast cancer and its modulation by p53 codon 72 arg > pro polymorphism. Breast Cancer Research 9: R71.

145. Mitra AK, Singh N, Singh A, Garg VK, Agarwal A, et al. (2008) Association of polymorphisms in base excision repair genes with the risk of breast cancer: A case-control study in north indian women. Oncol Res 17: 127-135.

146. Rajaraman P, Bhatti P, Doody MM, Simon SL, Weinstock RM, et al. (2008) Nucleotide excision repair polymorphisms may modify ionizing radiation-related breast cancer risk in US radiologic technologists. International Journal of Cancer 123: 2713-2716.

147. Economopoulos KP, Sergentanis TN. (2010) XRCC3 Thr241Met polymorphism and breast cancer risk: A meta-analysis. Breast Cancer Res Treat 121: 439-443.

148. Leng S, Bernauer A, Stidley CA, Picchi MA, Sheng X, et al. (2008) Association between common genetic variation in cockayne syndrome A and B genes and nucleotide excision repair capacity among smokers. Cancer Epidemiology Biomarkers & Prevention 17: 2062-2069.

149. Roberts MR, Shields PG, Ambrosone CB, Nie J, Marian C, et al. (2011) Single-nucleotide polymorphisms in DNA repair genes and association with breast cancer risk in the web study. Carcinogenesis 32: 1223-1230.

2. Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm[57]

2.1 Introduction

Breast cancer is the most common cancer in women in the developed world, with 22,700 new cases and 5,100 deaths anticipated in Canada for 2012 [1]. While environmental and lifestyle risk factors contribute to most of the variation in breast cancer risk, twin studies have shown substantial contribution of inherited genetic risk factors to disease susceptibility [2,3]. Linkage and family-based studies have identified high and moderate penetrance mutations in genes such as *BRCA1*[58] [4], *BRCA2*[59] [5], *PTEN*[60] [6], *ATM*[61] [7], *TP53*[62] [8], *BRIP1*[63] [9], *PALB2*[64] [10] and *CHEK2*[65] [11] contributing to hereditary breast cancer; however, these mutations occur rarely in the general population. Further, linkage studies failed to identify additional genes/mutations associated with high or moderate risk of breast cancer. Therefore, it has been hypothesized that most of the genetic risk of breast cancer, for both familial and sporadic cases in the general population, may involve a combination of multiple low penetrance genes/loci, each contributing to an overall genetic risk of breast cancer [12].

---

[57] A version of this chapter has been accepted for publication. *Sapkota et al., 2013. PLoS ONE.* I would like to thank Malinee Sridharan for providing partial genotype data (part of her Masters thesis) of the consortia SNPs included in this chapter.
[58] Breast cancer 1, early onset.
[59] Breast cancer 2, early onset.
[60] Phosphatase and tensin homolog.
[61] Ataxia telangiectasia mutated.
[62] Tumor protein p53.
[63] *BRCA1* interacting protein C-terminal helicase 1.
[64] Partner and localizer of *BRCA2.*
[65] Checkpoint kinase 2.

Over the past five years, several genome-wide association studies (GWASs) have reported breast cancer susceptibility variants (*i.e.*, single nucleotide polymorphisms, SNPs) at multiple loci [13-22]. A large-scale candidate gene study also identified an additional locus (caspase 8 coding SNP, rs1045485) associated with breast cancer risk [23]. The low penetrance common SNPs identified to date explain less than 10% of the genetic risk of breast cancer [22]. Taken together, pathogenic germline mutations and low penetrance variants identified thus far only account for a small fraction of the genetic risk of breast cancer, suggesting that additional variants remain to be identified [24].

Recently, we conducted a two-stage GWAS using sporadic breast cancer cases and healthy controls and identified six SNPs (located at chromosomes 4, 5, 16 and 19) that appeared to be associated with breast cancer susceptibility [21]. In a combined sample size of 1,455 breast cancer cases and 1,536 healthy controls from two independent stages, these SNPs showed modest risk of breast cancer (observed odds ratios (ORs) range: 1.22 – 1.45).

It is an internationally accepted practice to replicate GWAS findings in multiple independent studies with cases and controls of both similar and diverse ethnic backgrounds to assess the robustness and generalizability of the identified associations, respectively. Therefore, in the current study, we further investigated the six putative breast cancer susceptibility SNPs that we have reported previously [21] by conducting an independent replication study (stage 3), using breast cancer cases and controls. The study subjects were predominantly of Caucasian origins, and were drawn from the same geographical region in Canada

as in our previous study. We also evaluated the GWAS variants for breast cancer susceptibility reported by various consortia (including the Breast Cancer Association Consortium [13,18], the Effectiveness of Additional Reductions in Cholesterol and Homocysteine Collaborative Group [13], the Nurses' Health Study [14], the National Cancer Institute Cancer Genetic Markers of Susceptibility Project [14] and the National Heart, Lung, and Blood Institute Framingham Heart Study [15]) in our study population to explore the extent of conformity to previous findings in Caucasian populations, and for the strengths of associations for the sample size utilized in this study. Since obesity is a well-established risk factor for post-menopausal breast cancer [25] and is a heritable trait [26], we also adjusted the identified variant-breast cancer associations for body mass index (BMI[66]) to examine whether the variants are associated with breast cancer risk, through BMI or through different pathways. We assessed variability in disease susceptibility by clinicopathological characteristics such as menopausal status, family history of breast cancer, luminal A status of tumors, tumor grade and tumor stage. Finally, we explored the associations of the six putative susceptibility SNPs identified in our earlier study and the previously published consortia SNPs with breast cancer outcomes.

[66] BMI is an estimation of an individual's body shape and is calculated based on weight and height. BMI=mass(in kg)/(height(in meter))^2.

2.2 Materials and methods

2.2.1 Study participants

All breast cancer cases (n=2,750) used in this study had a confirmed diagnosis of breast cancer in the province of Alberta, Canada, and participated in provincial tumor bank projects in operation since 2001 (the PolyomX Project, 2001-2005 and subsequently merged with the Canadian Breast Cancer Foundation (CBCF) Tumor Bank, 2005 to present; http://www.abtumorbank.com/), Alberta, Canada) [21,27]. The tumor bank accrue tumor tissue and blood samples from patients with confirmed diagnoses of breast and other cancer types, through eight regional hospitals in Edmonton and Calgary in the province of Alberta, Canada and are the comprehensive publicly funded cancer care centres managed by Alberta Health Services (AHS). These centres provide guideline based cancer treatments under the Universal Health Care plan federally legislated in the Canada Health Act. These tumor banks contain well-annotated clinicopathological information for the stored samples. The CBCF Tumor Bank currently holds blood from more than 8,000 individuals from various cancer types, as a source of germline DNA for genotyping. Apparently-healthy (*i.e.*, confirmed not to have had a diagnosis of any cancer) controls (n=4,472) were obtained from the Tomorrow Project (http://in4tomorrow.ca) and were frequency matched to cases based on ten-year age group. The Tomorrow Project is a large prospective cohort study that started in 2000 and successfully recruited approximately 42,000 Albertans (64% women) by 2012 using a combination of random digit dialling (RDD), and random mail-outs, augmented by email campaigns and social media.

Inclusion criteria for initial recruitment to the Tomorrow Project were as follows: (i) aged 35-69 years; (ii) no personal history of cancer, other than non-melanoma skin cancer; (iii) able to complete written questionnaires in English and (iv) currently living in Alberta. Upon enrolment to the Tomorrow Project, participants completed a health and lifestyle questionnaire (including family history of major diseases), and gave written consent to be contacted in the future to provide a blood sample for banking to support research in cancer or chronic diseases, receive invitations to provide updated health and lifestyle information or additional samples in the future, and to linkage with administrative health data to understand patterns of health services utilization and disease occurrence [28]. Absence of prior history of cancer upon study enrolment was confirmed by performing linkage with the Alberta Cancer Registry (http://www.albertahealthservices.ca/poph/hi-poph-surv-cancer-alta-cancer-registry-2009.pdf). As of late 2012, approximately 19,000 Tomorrow Project participants from across Alberta had given a 50 ml non-fasting venous blood sample for banking in multiple aliquots of buffy coat, serum, plasma and red blood cells. Breast cancer cases in this study were of predominantly Caucasian ancestry, and resided in the Edmonton and Calgary regions (sites of tertiary cancer centres in Alberta). The population in these regions accounts for two thirds of the total population of the province of Alberta. Thus, in addition to age matching, the controls were selected from the Tomorrow Project using the same ethnicity and geographic location criteria. Even though socio-economic status (SES) plays a role in health outcomes, differences between SES of cases and

controls used in this study and underlying assumptions need to be validated independently. However, given the universal access to health care as a model adopted in Canada, the influence of SES was therefore considered as minimal, if any. A brief description of demographic characteristics of breast cancer cases and controls is presented in **Table 2-1**. Written informed consent to use banked samples to support research was obtained from all the study participants, and the study was approved by the Alberta Cancer Research Ethics Committee, Alberta, Canada.

Table 2-1 Distribution of age and BMI of breast cancer cases and controls used in the study

Characteristics	Breast cancer cases (n=2750)	Apparently healthy controls (n=4472)
Median age in years at diagnosis/ blood draw [range]	54 [22-92]	54 [35-78]
<40	192	343
40-50	710	1282
50-60	889	1538
60-70	635	1144
70-80	242	162
>80	64	0
Missing	18	3
Median body mass index (kg/m^2) [25th - 75th percentiles]	27.4 [24.1-31.4]	25.5 [22.7-29.3]
<18.5	20	41
18.5-24.99	663	1899
25-39.99	1359	2155
>40	112	148
Missing	596	229

2.2.2 SNPs and samples used

In this replication study (stage 3), we investigated associations of the six putative breast cancer susceptibility SNPs (4q31.22-rs1429142, 5p15.2-

rs1092913, 16q23.2-rs1981867, *ZNF577*[67]-rs10411161, *ZNF577*-rs3848562 and *ZNF577*-rs11878583) [21], that we reported in our previous two-stage GWAS. Stage 3 (total n=4,228) of the study used an independent set of breast cancer cases (n=1,294) and healthy controls (n=2,934). In the combined analyses of all three stages, a cumulative sample size (total n=7,219) was used. We also assessed the strengths of 11 breast cancer susceptibility SNPs that had been reported by consortia until 2009 (*SLC4A7*[68]-rs4973768 [18], 5p12-rs4415084 [16], 5p12-rs10941679 [16], 5q11.2-rs889312 [13], 8q24.21-rs13281615 [13], *FGFR2*[69]-rs2981579 [19], *FGFR2*-rs1219648 [14], *FGFR2*-rs2420946 [14], *FGFR2*-rs2981582 [13], *TNRC9*[70]-rs3803662 [13] and *COL1A1*[71]-rs2075555 [15]). A cumulative sample size of 2,672 breast cancer cases and 4,470 healthy controls were genotyped for these 11 consortia SNPs. Genotype data are available upon request.

2.2.3 SNPs genotyping and quality control

Germline DNA was extracted from peripheral blood samples of both cases and controls using commercially available Qiagen (Mississauga, ON, Canada) DNA isolation kits. All genotyping assays were performed on the Sequenom iPLEX Gold platform (San Diego, CA, USA) using services from the McGill University and Genome Quebec Innovation Center, Montreal, Canada. Within-stage (stage 3 for the six SNPs from our previous GWAS and a single stage for

[67] Zinc finger protein 577.
[68] Solute carrier family 4, sodium bicarbonate co-transporter, member 7.
[69] Fibroblast growth factor receptor 2.
[70] Trinucleotide repeat containing 9.
[71] Collagen, type I, alpha 1.

the 11 consortia SNPs) genotype concordance was assessed with 66 duplicate samples (8 cases and 58 controls). Cross platform (Affymetrix vs. Sequenom *i.e.*, stage 1 vs. stage 3 for the six SNPs) was assessed with 17 duplicate samples (5 cases and 12 controls). Between-stage (stage 2 vs. stage 3 for the six SNPs) genotype concordance was assessed with 632 cases and 452 controls. Duplicate samples used for assessing genotype concordances among various stages were randomly selected. Very stringent criteria of SNP call rate >99% was considered to minimize false positive associations due to missing genotype counts and HWE criteria of $P>10^{-6}$ in control subjects were adopted.

2.2.4 Association analyses and statistical considerations

2.2.4.1 Overall analyses

Allelic associations of SNPs with breast cancer susceptibility were evaluated with correlation/trend tests with one degree of freedom (d.f.). The strengths of allelic and genotypic associations were estimated using unconditional logistic regressions and reported as ORs and 95% confidence intervals (CIs). To increase sample size and hence the statistical power to better capture SNP-breast cancer associations, cases and controls from all independent stages were pooled together and combined analyses were conducted. BMI was included as a covariate in the logistic models to calculate adjusted ORs, 95% CIs and *P* values in Stage 3 and in combined stages.

2.2.4.2 Subgroup analyses

To evaluate variations in SNP-breast cancer associations by clinicopathological characteristics, hence addressing potential heterogeneity in the

observed overall associations, we conducted subgroup analyses (unconditional logistic regressions adjusted for BMI) within the combined breast cancer cases based on menopausal status, luminal A status, family history of breast cancer (captured under the single category representing cases with first, second or third degree relatives), tumor stage and grade. A common set of healthy controls was used to test the SNP-breast cancer associations in these subgroup analyses. Breast tumors that were either estrogen receptor (ER) or progesterone receptor (PR) positive and human epidermal growth factor receptor 2 (HER2) negative were classified as luminal As, and the remainder were classified as non-luminal As. The cases with unknown ER, PR or HER2 status were excluded from the luminal A subgroup analyses. Breast tumors with operable tumor stages (I-IIIA) were classified as one subgroup while tumors with non-operable tumor stages (IIIB, IIIC) were classified as the other subgroup. Heterogeneity in ORs between the subgroups was assessed using multinomial logistic regressions (‘*mlogit’*) and linear combination of estimators (‘*lincom’*) implemented in Stata 12.0 (www.stata.com). Statistical significance of this heterogeneity test was reported as *P* for heterogeneity (P_{het}).

2.2.4.3 Associations of SNPs with breast cancer outcomes

We also evaluated the potential prognostic values of SNPs with breast cancer outcomes, such as recurrence-free survival (RFS) and overall survival (OS), by fitting Cox proportional hazards models available in the “survival” package [29] implemented in R 2.15.1 [30], adjusted for BMI. The associations were reported as hazard ratios (HRs), 95% CIs and adjusted *P* values. Genotypes were recoded

to 0 (wild type homozygotes), 1 (heterozygotes) and 2 (variant homozygotes) before fitting the Cox models.

All statistical tests were two-sided. We assumed an additive model of genetic inheritance to calculate power, as described earlier [21]. As such, our study had adequate power (>80%) to detect associations that were larger than genotypic relative risk of ≥1.2. Whenever multiple SNPs were tested, correction for multiple hypotheses testing was performed by P=0.05/number of tests. We considered all SNPs from our stage 1 GWAS (782,838 SNPs) to calculate genome-wide significance ($P<6.4 \times 10^{-8}$) for the six replicated SNPs. Correlation/trend tests were carried out using SNP and Variation Suite v7.6.11 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) [31]. The observed and adjusted allelic and genotypic ORs and 95% CIs and adjusted P values were estimated using logistic models in PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/) [32]. All the general statistical analyses were conducted using R 2.15.1.

2.3 Results

Genotyping assays of the 17 SNPs considered in this study were successful with a SNP call rate of >99%. Average within-stage genotype concordance was 100% while cross-platform genotype concordance was >99%; between-stage average genotype concordance was also 100%. We reasoned that this negligible percentage (<1%) of discordance was unlikely to influence SNP-breast cancer associations and hence all the genotype data were considered for the downstream association analyses. The genotype distributions from the six SNPs (our previous work) showed conformity with Hardy-Weinberg Equilibrium (HWE) criteria in

control subjects. Similarly, the genotype distributions from the 11 consortia SNPs were also in agreement with HWE. Minor allele frequencies (MAFs) of the six SNPs across all stages and the 11 consortia SNPs were comparable with the published MAFs, reflecting the robustness of the genotyping platform vis-à-vis negligible genotyping errors (**Tables S2-1 and 2-2**) and confidence in the reported associations.

Table S2-1 Associations of the previously identified (consortia SNPs) breast cancer susceptibility loci in the current study.

SNP	Genes/loci	Cytoband	Location (bp)*	Association in published GWAS			
				MA	MAF	Published $OR_{per\text{-}allele}$ [95% CI]	Published *P*
rs4973768	*SLC4A7*	3p24.1	27,416,013	T	0.47	1.11 [1.08-1.13]	4.10E-23
rs4415084	Intergenic	5p12	44,662,515	T	0.43	1.16 [1.10-1.21]	6.40E-10
rs10941679	Intergenic	5p12	44,706,498	G	0.28	1.19 [1.13-1.26]	2.90E-11
rs889312	Intergenic	5q11.2	56,031,884	C	0.28	1.13 [1.10-1.16]	7.00E-20
rs13281615	Intergenic	8q24.21	128,355,618	C	0.40	1.08 [1.05-1.11]	5.00E-12
rs2981579	*FGFR2*	10q26.13	123,337,335	T	0.41	1.17 [1.07-1.27]	1.79E-10
rs1219648	*FGFR2*	10q26.13	123,346,190	G	0.46	1.20 [1.07-1.42]	1.10E-10
rs2420946	*FGFR2*	10q26.13	123,351,324	T	0.38	1.26 [1.18-1.36]	2.00E-10
rs2981582	*FGFR2*	10q26.13	123,352,317	A	0.38	1.26 [1.23-1.30]	2.00E-76
rs3803662	*TNRC9*	16q12.1	52,586,341	T	0.25	1.20 [1.16-1.24]	1.00E-36
rs2075555	*COL1A1*	17q21.33	48,274,291	NA	NA	NA	8.03E-08

*National Center for Biotechnology Information genome build 37
MA, minor allele; MAF, minor allele frequency
**Adjusted for BMI
#*P* values obtained from correlation/trend test with one d.f.

Table S2-1 Continued..

Association in this study (2,672 cases/4,470 controls)						
MA	MAF	Observed $OR_{per-allele}$ [95% CI]	$P_{corr/trend}$ #	Adjusted $OR_{per-allele}$ [95% CI]**	Adjusted P**	Reference
A	0.48	1.14 [1.07-1.22]	1.39E-04	1.14 [1.06-1.23]	5.42E-04	[18]
T	0.42	1.17 [1.09-1.25]	9.41E-06	1.17 [1.09-1.26]	4.29E-05	[16]
C	0.27	1.19 [1.11-1.29]	6.03E-06	1.19 [1.10-1.29]	4.22E-05	[16]
G	0.29	1.22 [1.13-1.31]	2.31E-07	1.22 [1.13-1.32]	1.67E-06	[13]
C	0.42	1.18 [1.10-1.26]	2.65E-06	1.19 [1.11-1.28]	4.10E-06	[13]
A	0.43	1.26 [1.17-1.35]	1.15E-10	1.24 [1.15-1.34]	1.57E-08	[19]
C	0.41	1.26 [1.18-1.35]	5.80E-11	1.26 [1.16-1.35]	4.52E-09	[14]
T	0.41	1.26 [1.17-1.35]	1.54E-10	1.25 [1.16-1.35]	8.62E-09	[14]
T	0.40	1.25 [1.17-1.34]	2.65E-10	1.25 [1.16-1.35]	1.33E-08	[13]
T	0.29	1.26 [1.17-1.36]	7.06E-10	1.25 [1.16-1.36]	4.19E-08	[13]
T	0.14	1.02 [0.92-1.12]	7.63E-01	1.03 [0.92-1.14]	6.37E-01	[15]

2.3.1 Association of previously identified (consortia SNPs) breast cancer susceptibility loci

Except for *COL1A1*-rs2075555, we successfully replicated the association of ten consortia reported breast cancer susceptibility loci in our study population at $P<0.05$ (**Table S2-1**). These SNPs remained statistically significant after correction for multiple hypothesis testing ($P<0.05/11=0.004$). Four *FGFR2* SNPs and *TNRC9*-rs3803662 showed the strongest associations attaining the commonly adopted genome-wide significance level ($P<5.0 \times 10^{-8}$), with similar ORs to the original study findings [13,14,19]. After adjusting for BMI, five SNPs remained statistically significant (adjusted $P<4.2 \times 10^{-8}$) (**Table S2-1**). The adjusted per allele ORs and 95% CIs were also similar to the observed ORs and 95% CI (**Table S2-1),** indicating that these SNP-breast cancer associations are independent of the pathway linking BMI and risk of breast cancer.

2.3.2 Replication of the six putative SNPs in stage 3 analyses

Of the six putative breast cancer susceptibility SNPs that we reported earlier, 4q31.22-rs1429142 showed consistent reproducibility across all three stages. The variant at 5p15.2-rs1092913 also retained statistical significance for increased breast cancer risk in the current independent replication stage 3 study at $P<0.05$ (**Table 2-2**), and remained statistically significant after correction for multiple hypothesis testing ($P<0.05/6=0.008$). The magnitude and direction of per allele ORs and 95% CIs of both SNPs were consistent with our previous findings [21] while slightly elevated ORs and 95% CIs were observed for heterozygotes and variant homozygotes (**Table 2-2**), conforming to the additive model of genetic inheritance. After adjustment for BMI, both 4q31.22-rs1429142 and 5p15.2-rs1092913 remained statistically significant at adjusted $P<0.05$, while both adjusted per allele and genotypic ORs and 95% CIs of 4q31.22-rs1429142 were larger than the observed ORs. The remaining four SNPs did not show statistical significance at $P<0.05$ in this stage 3 study.

Table 2-2 Replication of the six putative breast cancer susceptibility loci in independent stage 3.

SNP	Genes/ loci	Cytoband	Location (bp)*	MA	Stage[#]	MAF	Call rate	Cases/ controls
rs1429142	*EDNRA*/ Intergenic	4q31.22	148,289,389	C	1	0.17	1.00	302/321
					2	0.18	1.00	1153/1215
					3	0.18	1.00	1294/2934
					1+2+3	0.18	1.00	2749/4470
rs1092913	*ROPN1L*/ Intergenic	5p15.2	10,467,702	T	1	0.10	1.00	302/321
					2	0.14	0.99	1153/1215
					3	0.14	1.00	1294/2934
					1+2+3	0.13	0.99	2749/4470
rs1981867	*C16orf61*/ Intergenic	16q23.2	80,923,769	T	1	0.29	1.00	302/321
					2	0.31	1.00	1153/1215
					3	0.32	1.00	1294/2934
					1+2+3	0.31	1.00	2749/4470
rs10411161	*ZNF577*	19q13.41	52,372,976	A	1	0.11	1.00	302/321
					2	0.14	0.99	1153/1215
					3	0.13	1.00	1294/2934
					1+2+3	0.13	1.00	2749/4470
rs3848562	*ZNF577*	19q13.41	52,379,835	C	1	0.11	1.00	302/321
					2	0.13	1.00	1153/1215
					3	0.13	1.00	1294/2934
					1+2+3	0.13	1.00	2749/4470
rs11878583	*ZNF577*	19q13.41	52,388,546	C	1	0.12	1.00	302/321
					2	0.13	1.00	1153/1215
					3	0.13	1.00	1294/2934
					1+2+3	0.13	1.00	2749/4470

*National Center for Biotechnology Information genome build 37

MA, minor allele; MAF, minor allele frequency; ND, not determined

[#]Association analyses (chi-squared tests) of Stages 1 and 2 are reported in Sehrawat et al. 2011. Stage 3 and the combined (1+2+3) stages were conducted in this follow-up study.

**To maintain the consistency in the statistical test, association of the six SNPs with breast cancer in Stages 1 and 2 were further evaluated with correlation/trend test with one d.f. in this study, unlike chi-squared test in the original study (Sehrawat et al. 2011)

***Adjusted for BMI

Table 2-2 Continued..

Observed				Adjusted***			
$OR_{per-allele}$ [95% CI]	$OR_{heterozygote}$ [95% CI]	$OR_{homozygote}$ [95% CI]	$P_{corr/trend}$ **	$OR_{per-allele}$ [95% CI]	$OR_{heterozygote}$ [95% CI]	$OR_{homozygote}$ [95% CI]	*P*
1.55 [1.15-2.08]	1.51 [1.06-2.15]	2.65 [1.06-6.63]	2.82E-03				
1.21 [1.04-1.40]	1.25 [1.05-1.49]	1.29 [0.82-2.04]	1.28E-02		ND		
1.20 [1.07-1.34]	1.23 [1.06-1.42]	1.33 [0.94-1.86]	2.50E-03	1.23 [1.08-1.40]	1.28 [1.09-1.50]	1.34 [0.92-1.95]	1.70E-03
1.22 [1.12-1.33]	1.26 [1.13-1.40]	1.36 [1.06-1.75]	4.71E-06	1.28 [1.17-1.41]	1.32 [1.17-1.48]	1.52 [1.16-2.00]	1.50E-07
1.89 [1.29-2.76]	2.24 [1.46-3.44]	1.23 [0.30-4.97]	7.03E-04				
1.34 [1.14-1.57]	1.27 [1.04-1.55]	2.10 [1.27-3.49]	2.17E-04		ND		
1.15 [1.01-1.30]	1.07 [0.91-1.26]	1.57 [1.10-2.24]	3.14E-02	1.17 [1.02-1.34]	1.07 [0.89-1.28]	1.70 [1.16-2.50]	2.89E-02
1.22 [1.11-1.34]	1.19 [1.06-1.34]	1.62 [1.23-2.14]	2.18E-05	1.21 [1.10-1.34]	1.20 [1.05-1.36]	1.53 [1.13-2.06]	1.96E-04
1.59 [1.23-2.05]	1.50 [1.08-2.08]	2.82 [1.49-5.34]	3.72E-04				
1.14 [1.01-1.29]	1.00 [0.84-1.18]	1.53 [1.15-2.03]	3.17E-02		ND		
0.97 [0.88-1.08]	1.00 [0.87-1.14]	0.92 [0.73-1.16]	6.04E-01	0.99 [0.89-1.11]	1.04 [0.90-1.22]	0.92 [0.71-1.19]	8.46E-01
1.07 [0.99-1.15]	1.02 [0.93-1.13]	1.19 [1.01-1.40]	8.60E-02	1.07 [0.98-1.15]	1.06 [0.95-1.18]	1.15 [0.95-1.37]	1.21E-01
1.79 [1.25-2.57]	1.69 [1.13-2.54]	4.80 [1.01-22.83]	1.08E-03				
1.28 [1.09-1.49]	1.34 [1.09-1.65]	1.49 [0.99-2.24]	6.17E-04		ND		
0.97 [0.84-1.11]	0.98 [0.84-1.15]	0.87 [0.54-1.38]	6.15E-01	1.00 [0.86-1.17]	1.00 [0.83-1.19]	1.03 [0.63-1.70]	9.67E-01
1.13 [1.03-1.24]	1.11 [0.99-1.25]	1.36 [1.02-1.81]	1.06E-02	1.13 [1.02-1.25]	1.11 [0.97-1.26]	1.35 [0.99-1.85]	2.10E-02
1.82 [1.27-2.61]	1.72 [1.15-2.58]	4.82 [1.01-22.93]	8.05E-04				
1.32 [1.11-1.56]	1.34 [1.10-1.64]	1.62 [0.92-2.85]	9.79E-04		ND		
0.96 [0.84-1.10]	0.98 [0.84-1.15]	0.85 [0.54-1.36]	5.91E-01	1.01 [0.87-1.17]	1.00 [0.83-1.19]	1.07 [0.65-1.73]	9.18E-01
1.12 [1.01-1.23]	1.13 [1.01-1.27]	1.19 [0.86-1.65]	2.56E-02	1.13 [1.02-1.26]	1.13 [1.00-1.29]	1.26 [0.89-1.79]	2.60E-02
1.80 [1.25-2.58]	1.65 [1.11-2.45]	8.43 [1.03-69.00]	1.26E-03				
1.25 [1.06-1.47]	1.16 [0.95-1.41]	2.15 [1.21-3.83]	7.60E-03		ND		
0.98 [0.86-1.12]	0.96 [0.82-1.13]	1.01 [0.66-1.56]	7.40E-01	1.00 [0.86-1.16]	0.96 [0.80-1.14]	1.18 [0.74-1.86]	9.98E-01
1.10 [1.00-1.22]	1.07 [0.95-1.20]	1.38 [1.01-1.89]	4.65E-02	1.11 [0.99-1.23]	1.07 [0.94-1.22]	1.38 [0.98-1.94]	6.44E-02

2.3.3 Combined analyses of the six putative SNPs (stages 1+2+3)

In the combined analyses (stages 1+2+3), five of the six SNPs were significantly associated with increased breast cancer risk at $P<0.05$, the exception being 16q23.2-rs1981867 which showed marginal statistical significance ($P=0.06$) (**Table 2-2**). Again, 4q31.22-rs1429142 and 5p15.2-rs1092913 showed the strongest associations after multiple hypotheses correction. The five SNPs retained statistical significance after adjusting for BMI. Interestingly, 4q31.22-rs1429142 achieved near genome-wide significance level with greater per allele and genotypic ORs and 95% CIs (adjusted $P=1.5 \times 10^{-7}$, adjusted per allele OR and 95% CI=1.28 [1.17-1.41], adjusted $OR_{heterozygote}$ and 95% CI=1.32 [1.17-1.48] and adjusted $OR_{homozygote}$ and 95% CI=1.52 [1.16-2.00]), indicating that the 4q31.22-rs1429142-breast cancer association may be linked to the BMI pathway

of breast cancer risk elevation (**Table 2-2**). 5p15.2-rs1092913 also showed a strong association with breast cancer risk (adjusted P=2.0 x 10^{-4}, adjusted per allele OR and 95% CI=1.21 [1.10-1.34], adjusted $OR_{heterozygote}$ and 95%=1.20 [1.05-1.36] and adjusted $OR_{homozygote}$ and 95% CI=1.53 [1.13-2.06]).

2.3.4 Subgroup analyses

The previously reported GWAS variants (consortia SNPs), except *COL1A1*-rs2075555, remained statistically significant in subgroups with both pre and postmenopausal women, luminal A cases, cases with or without family history of breast cancer, low tumor grade and operable tumor stage at adjusted P<0.05 (**Table S2-2**). The adjusted per allele ORs, 95% CIs and P values were also comparable to the overall analyses, with similar magnitudes and directions of risk (**Tables S2-1 and S2-2**).

Table S2-2 Subgroup analysis of the 11 previously GWAS-identified SNPs based on menopausal and luminal A status, family history of breast cancer, tumor grade and stage.

SNP	MA	Controls¶	Premenopausal women			Postmenopausal women			P_{het}
			No. of cases	$OR_{per-allele}$* [95% CI]	Adjusted P*	No. of cases	$OR_{per-allele}$* [95% CI]	Adjusted P*	
rs4973768	A	4470	1033	1.12 [1.01-1.25]	3.76E-02	1553	1.15 [1.05-1.26]	3.76E-03	6.84E-01
rs4415084	T	4470	1033	1.20 [1.08-1.33]	6.98E-04	1553	1.14 [1.04-1.25]	6.59E-03	3.79E-01
rs10941679	C	4470	1033	1.18 [1.05-1.33]	5.23E-03	1553	1.18 [1.07-1.31]	1.37E-03	9.65E-01
rs889312	G	4470	1033	1.25 [1.12-1.40]	1.13E-04	1553	1.21 [1.10-1.34]	1.42E-04	6.26E-01
rs13281615	C	4470	1033	1.19 [1.08-1.33]	8.61E-04	1553	1.19 [1.08-1.30]	2.61E-04	8.69E-01
rs2981579	A	4470	1033	1.25 [1.12-1.39]	5.11E-05	1553	1.22 [1.12-1.34]	1.81E-05	6.13E-01
rs1219648	C	4470	1033	1.26 [1.13-1.40]	2.41E-05	1553	1.23 [1.12-1.35]	9.44E-06	8.00E-01
rs2420946	T	4470	1033	1.26 [1.14-1.40]	1.67E-05	1553	1.22 [1.11-1.34]	2.17E-05	6.37E-01
rs2981582	T	4470	1033	1.25 [1.12-1.39]	5.42E-05	1553	1.23 [1.12-1.35]	1.36E-05	8.65E-01
rs3803662	T	4470	1033	1.32 [1.18-1.47]	1.35E-06	1553	1.21 [1.10-1.34]	1.32E-04	2.06E-01
rs2075555	T	4470	1033	1.09 [0.94-1.26]	2.51E-01	1553	0.99 [0.87-1.13]	9.34E-01	3.12E-01

MA, minor allele; P_{het}, P for heterogeneity

¶Common controls used for conducting each of the subgroup analyses mentioned above

*Adjusted for BMI

Table S2-2 Continued..

Luminal A			Non luminal A			
No. of cases	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	No. of cases	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	P_{het}
1822	1.17 [1.07-1.27]	5.23E-04	751	1.12 [0.99-1.26]	6.32E-02	5.55E-01
1822	1.25 [1.15-1.36]	4.39E-07	751	1.01 [0.89-1.14]	9.14E-01	2.00E-03
1822	1.27 [1.15-1.39]	1.27E-06	751	1.00 [0.87-1.14]	9.67E-01	2.00E-03
1822	1.24 [1.13-1.36]	9.64E-06	751	1.20 [1.05-1.36]	6.65E-03	7.31E-01
1822	1.22 [1.12-1.33]	6.93E-06	751	1.14 [1.01-1.28]	3.58E-02	3.38E-01
1822	1.34 [1.23-1.47]	3.12E-11	751	1.07 [0.95-1.21]	2.79E-01	1.01E-01
1822	1.36 [1.24-1.48]	6.63E-12	751	1.07 [0.95-1.21]	2.50E-01	1.00E-03
1822	1.36 [1.25-1.48]	5.82E-12	751	1.06 [0.93-1.19]	3.86E-01	<0.001
1822	1.35 [1.24-1.47]	1.60E-11	751	1.06 [0.94-1.20]	3.53E-01	1.00E-03
1822	1.24 [1.13-1.36]	8.25E-06	751	1.28 [1.13-1.46]	1.36E-04	5.46E-01
1822	0.99 [0.88-1.12]	8.93E-01	751	1.13 [0.95-1.33]	1.61E-01	1.74E-01

Table S2-2 Continued..

With family history			Without family history			
No. of cases	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	No. of cases	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	P_{het}
1084	1.17 [1.05-1.30]	3.44E-03	1399	1.13 [1.03-1.24]	1.37E-02	5.95E-01
1084	1.11 [1.00-1.23]	5.03E-02	1399	1.19 [1.09-1.31]	2.43E-04	2.75E-01
1084	1.15 [1.03-1.29]	1.70E-02	1399	1.18 [1.06-1.31]	1.76E-03	7.26E-01
1084	1.14 [1.02-1.28]	1.95E-02	1399	1.27 [1.14-1.40]	6.32E-06	1.66E-01
1084	1.18 [1.07-1.31]	1.38E-03	1399	1.20 [1.09-1.31]	2.02E-04	9.03E-01
1084	1.35 [1.22-1.50]	1.63E-08	1399	1.13 [1.03-1.24]	1.30E-02	8.20E-01
1084	1.35 [1.22-1.50]	1.24E-08	1399	1.15 [1.04-1.26]	5.11E-03	1.00E-02
1084	1.35 [1.21-1.50]	1.95E-08	1399	1.14 [1.04-1.26]	5.92E-03	1.10E-02
1084	1.33 [1.20-1.48]	5.97E-08	1399	1.15 [1.04-1.26]	5.28E-03	1.80E-02
1084	1.22 [1.10-1.37]	3.84E-04	1399	1.26 [1.14-1.40]	6.26E-06	6.58E-01
1084	0.95 [0.82-1.10]	5.04E-01	1399	1.07 [0.94-1.22]	3.16E-01	1.98E-01

Table S2-2 Continued..

High tumor grade			Low tumor grade			
No. of cases	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	No. of cases	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	P_{het}
1053	1.04 [0.93-1.15]	5.15E-01	1530	1.24 [1.12-1.36]	1.10E-05	6.00E-03
1053	1.05 [0.95-1.17]	3.33E-01	1530	1.25 [1.14-1.37]	2.65E-06	8.00E-03
1053	1.02 [0.91-1.15]	7.09E-01	1530	1.30 [1.17-1.44]	5.29E-07	1.00E-03
1053	1.15 [1.03-1.29]	1.25E-02	1530	1.26 [1.14-1.40]	5.69E-06	2.14E-01
1053	1.06 [0.95-1.17]	2.99E-01	1530	1.31 [1.19-1.43]	1.91E-08	1.00E-03
1053	1.10 [0.99-1.22]	7.33E-02	1530	1.35 [1.23-1.48]	4.23E-10	2.50E-02
1053	1.10 [0.99-1.23]	6.45E-02	1530	1.37 [1.24-1.50]	7.15E-11	1.00E-03
1053	1.09 [0.99-1.22]	9.32E-02	1530	1.36 [1.24-1.50]	1.13E-10	1.00E-03
1053	1.10 [0.99-1.22]	8.99E-02	1530	1.35 [1.23-1.49]	2.62E-10	1.00E-03
1053	1.28 [1.14-1.42]	1.38E-05	1530	1.23 [1.11-1.36]	6.89E-05	5.33E-01
1053	1.11 [0.97-1.28]	1.36E-01	1530	0.96 [0.84-1.10]	5.42E-01	8.80E-02

Table S2-2 Continued..

Tumor stage I-IIIA			Tumor stage IIIB and IIIC			
No. of cases	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	No. of cases	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	P_{het}
2522	1.14 [1.05-1.23]	1.16E-03	150	1.26 [0.97-1.64]	8.32E-02	4.66E-01
2522	1.17 [1.09-1.27]	4.21E-05	150	1.10 [0.85-1.42]	4.81E-01	6.51E-01
2522	1.20 [1.10-1.30]	3.17E-05	150	1.06 [0.79-1.41]	6.97E-01	4.38E-01
2522	1.21 [1.12-1.32]	5.78E-06	150	1.33 [1.01-1.74]	3.96E-02	4.85E-01
2522	1.20 [1.11-1.30]	2.22E-06	150	1.03 [0.79-1.33]	8.30E-01	2.41E-01
2522	1.26 [1.17-1.37]	3.30E-09	150	0.98 [0.75-1.27]	8.72E-01	5.00E-03
2522	1.27 [1.17-1.37]	2.24E-09	150	1.08 [0.84-1.40]	5.41E-01	2.37E-01
2522	1.26 [1.17-1.36]	4.77E-09	150	1.10 [0.85-1.42]	4.83E-01	2.93E-01
2522	1.26 [1.17-1.36]	6.47E-09	150	1.08 [0.83-1.40]	5.66E-01	2.44E-01
2522	1.26 [1.17-1.37]	2.24E-08	150	1.07 [0.80-1.42]	6.60E-01	2.53E-01
2522	1.04 [0.93-1.15]	5.04E-01	150	0.85 [0.58-1.25]	4.03E-01	3.03E-01

Of these, the four *FGFR2* SNPs retained genome-wide significance level in subgroups with luminal A cases, cases with family history of breast cancer, low tumor grade and operable tumor stage while 8q24.21-rs13281615 and *TNRC9*-rs3803662 showed genome-wide significance level associations only in cases

with low tumor grade and operable tumor stage, respectively. *SLC4A7*-rs4973768, 5q11.2-rs889312, 8q24.21-rs13281615 and *TNRC9*-rs3803662 showed marginal associations in subgroup with non-luminal A cases. Similarly, 5q11.2-rs889312 and *TNRC9*-rs3803662 showed significant associations in cases with high tumor grade. None of the SNPs showed significant associations in cases with non-operable tumor stage, with the possible exception of 5q11.2-rs889312 which showed a marginally statistically significant association (adjusted P=0.04).

The associations of the six GWAS-identified putative SNPs from our populations with breast cancer were consistent across the subgroups, without any substantial modifications in SNP-breast cancer associations observed in overall analyses (**Table 2-2**). 4q31.22-rs1429142 and 5p15.2-rs1092913 remained significantly associated in subgroups with both pre and postmenopausal women, luminal and non-luminal A cases and cases without family history of breast cancer, high and low tumor grades and operable tumor stage at adjusted P<0.05 (**Tables 2-3 and 2-4**). Moreover, 4q31.22-rs1429142 attained genome-wide significance level in subgroups with premenopausal women (adjusted P=6.2 x 10^{-10}), while a strong statistical association was also observed in cases with operable tumor stages (adjusted P=1.6 x 10^{-7}). The *ZNF577* SNPs (rs10411161, rs3848562 and rs11878583) also showed statistically significant associations in subgroups with postmenopausal women, luminal A cases, cases without family history and operable tumor stages (**Tables 2-3 and 2-4**).

Table 2-3 Subgroup analyses of the six putative breast cancer susceptibility SNPs (Table 2-2) based on menopausal and luminal A status and family history of breast cancer.

SNP	MA	Controls¶	Premenopausal women (n=1036)		Postmenopausal women (n=1560)		P_{het}
			$OR_{per\text{-}allele}$* [95% CI]	Adjusted P*	$OR_{per\text{-}allele}$* [95% CI]	Adjusted P*	
rs1429142	C	4470	1.49 [1.31-1.68]	6.22E-10	1.17 [1.04-1.31]	7.79E-03	2.00E-03
rs1092913	T	4470	1.28 [1.12-1.47]	4.53E-04	1.18 [1.04-1.34]	1.01E-02	2.89E-01
rs1981867	T	4470	1.02 [0.91-1.14]	7.19E-01	1.10 [1.00-1.21]	6.02E-02	2.99E-01
rs10411161	A	4470	1.13 [0.98-1.31]	8.81E-02	1.11 [0.98-1.27]	9.74E-02	8.31E-01
rs3848562	C	4470	1.12 [0.96-1.30]	1.52E-01	1.12 [0.98-1.28]	8.83E-02	9.61E-01
rs11878583	C	4470	1.09 [0.94-1.27]	2.50E-01	1.10 [0.97-1.26]	1.44E-01	9.35E-01

MA, minor allele; P_{het}, P for heterogeneity

¶Common controls used for conducting each of the subgroup analyses mentioned above

*Adjusted for BMI

Table 2-3 Continued..

Luminal A (n=1828)		Non luminal A (n=755)		P_{het}
$OR_{per\text{-}allele}$* [95% CI]	Adjusted P*	$OR_{per\text{-}allele}$* [95% CI]	Adjusted P*	
1.26 [1.14-1.41]	1.62E-05	1.33 [1.15-1.54]	1.30E-04	5.99E-01
1.17 [1.04-1.32]	9.56E-03	1.30 [1.12-1.53]	9.14E-04	2.12E-01
1.09 [1.00-1.20]	6.31E-02	1.01 [0.89-1.15]	8.98E-01	2.96E-01
1.17 [1.04-1.31]	1.17E-02	1.02 [0.86-1.21]	7.91E-01	1.87E-01
1.17 [1.04-1.32]	1.06E-02	1.00 [0.84-1.20]	9.72E-01	1.23E-01
1.12 [0.99-1.27]	6.19E-02	1.05 [0.88-1.24]	6.06E-01	4.74E-01

Table 2-3 Continued..

With family history (n=1089)		Without family history (n=1404)		P_{het}
$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	
1.28 [1.13-1.45]	1.53E-04	1.28 [1.14-1.44]	3.07E-05	9.88E-01
1.10 [0.95-1.27]	1.98E-01	1.27 [1.12-1.44]	1.69E-04	8.30E-02
1.10 [0.99-1.23]	8.43E-02	1.06 [0.96-1.18]	2.24E-01	6.03E-01
1.12 [0.97-1.30]	1.16E-01	1.16 [1.02-1.32]	2.49E-02	6.88E-01
1.11 [0.95-1.28]	1.85E-01	1.16 [1.01-1.32]	3.21E-02	6.03E-01
1.10 [0.95-1.27]	2.19E-01	1.13 [0.99-1.29]	6.13E-02	6.89E-01

Table 2-4 Subgroup analyses of the six putative breast cancer susceptibility SNPs (Table 2-2) based on tumor grade and stage.

SNP	MA	Controls¶	High tumor grade (n=1057)		Low tumor grade (n=1536)		P_{het}
			$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	
rs1429142	C	4470	1.34 [1.18-1.52]	5.04E-06	1.24 [1.11-1.39]	2.41E-04	3.47E-01
rs1092913	T	4470	1.29 [1.13-1.48]	2.06E-04	1.12 [0.99-1.28]	7.96E-02	8.50E-02
rs1981867	T	4470	1.11 [0.99-1.23]	7.30E-02	1.05 [0.95-1.16]	3.29E-01	4.51E-01
rs10411161	A	4470	1.14 [0.99-1.32]	6.68E-02	1.12 [0.98-1.27]	9.16E-02	7.75E-01
rs3848562	C	4470	1.14 [0.98-1.32]	7.99E-02	1.11 [0.97-1.27]	1.23E-01	7.32E-01
rs11878583	C	4470	1.12 [0.97-1.30]	1.14E-01	1.08 [0.95-1.24]	2.36E-01	6.46E-01

MA, minor allele; P_{het}, P for heterogeneity

¶Common controls used for conducting each of the subgroup analyses mentioned above

*Adjusted for BMI

Table 2-4 Continued..

Tumor stages I-IIIA (n=2529)		Tumor stage IIIB, IIIC (n=153)		P_{het}
$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	$OR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	
1.29 [1.17-1.42]	1.62E-07	1.25 [0.92-1.70]	1.60E-01	8.28E-01
1.24 [1.12-1.38]	3.91E-05	0.80 [0.53-1.20]	2.83E-01	3.80E-02
1.06 [0.98-1.15]	1.34E-01	1.08 [0.83-1.42]	5.69E-01	9.10E-01
1.12 [1.01-1.25]	3.36E-02	1.23 [0.88-1.72]	2.20E-01	5.73E-01
1.13 [1.01-1.26]	3.60E-02	1.16 [0.81-1.64]	4.18E-01	8.56E-01
1.11 [0.99-1.23]	6.83E-02	1.08 [0.75-1.54]	6.88E-01	8.90E-01

2.3.5 Association of SNPs with breast cancer outcomes

Of the 17 SNPs tested for their associations with breast cancer outcomes, 8q24.21-rs13281615 was significantly associated with reduced risk of both RFS (adjusted P=0.001 and adjusted per allele HR and 95% CI=0.77 [0.65-0.90]) and OS (adjusted P=0.003, adjusted per allele HR and 95% CI=0.76 [0.64-0.91]) (**Table S2-3**). The remaining 16 SNPs did not show statistically significant associations with breast cancer outcomes at adjusted P<0.05.

2.4 Discussion

In this independent replication study in Canadian women involving 2,750 breast cancer cases and 4,472 healthy controls, we successfully reproduced the associations of ten previously GWAS-identified breast cancer susceptibility loci, indicating the robustness of the consortia identified SNPs with breast cancer. In addition, two of the six putative breast cancer susceptibility SNPs (4q31.22-

rs1429142 and 5p15.2-rs1092913) from our previous two-stage GWAS also showed robust associations in an independent set of breast cancer cases and healthy controls (stage 3). After adjusting for BMI, 4q31.22-rs1429142 attained near genome-wide significance level (adjusted P=1.5 x 10^{-7}) (**Table 2**). A major strength of this study is the consideration of BMI, which allowed confirmation that the genetic contributions to breast cancer are independent of one of the major risk factors for breast cancer. An additional strength was our evaluation of the SNP-breast cancer associations as potential prognostic factors for RFS and OS after diagnosis and their relationships with breast cancer clinical and molecular subtypes.

Table S2-3 Association of the 17 SNPs with breast cancer outcomes.

SNP	MA	Relapse-free survival			Overall survival		
		No. of events/All cases	$HR_{per\text{-}allele}$ * [95% CI]	Adjusted P *	No. of events/All cases	$HR_{per\text{-}allele}$ * [95% CI]	Adjusted P *
rs4973768	A	433/1406	0.98 [0.83-1.14]	7.63E-01	366/1410	1.00 [0.84-1.19]	9.86E-01
rs4415084	T	433/1406	0.97 [0.82-1.13]	6.61E-01	366/1410	1.12 [0.93-1.33]	2.34E-01
rs10941679	C	433/1406	1.13 [0.96-1.33]	1.52E-01	366/1410	1.16 [0.96-1.40]	1.15E-01
rs889312	G	433/1406	0.94 [0.79-1.10]	4.31E-01	366/1410	1.01 [0.84-1.22]	9.35E-01
rs13281615	C	433/1406	0.77 [0.65-0.90]	1.00E-03	366/1410	0.76 [0.64-0.91]	3.10E-03
rs2981579	A	433/1406	0.91 [0.78-1.06]	2.36E-01	366/1410	0.93 [0.78-1.10]	3.79E-01
rs1219648	C	433/1406	0.92 [0.79-1.08]	3.03E-01	366/1410	0.91 [0.76-1.08]	2.73E-01
rs2420946	T	433/1406	0.90 [0.76-1.05]	1.70E-01	366/1410	0.87 [0.73-1.04]	1.36E-01
rs2981582	T	433/1406	0.90 [0.77-1.06]	1.96E-01	366/1410	0.88 [0.73-1.04]	1.38E-01
rs3803662	T	433/1406	0.97 [0.82-1.14]	7.20E-01	366/1410	1.03 [0.86-1.24]	7.52E-01
rs2075555	T	433/1406	1.05 [0.86-1.29]	6.10E-01	366/1410	1.10 [0.88-1.38]	3.87E-01
rs1429142	C	439/1416	1.04 [0.86-1.24]	7.02E-01	370/1420	1.05 [0.86-1.29]	6.19E-01
rs1092913	T	439/1416	1.06 [0.87-1.31]	5.56E-01	370/1420	1.11 [0.88-1.39]	3.92E-01
rs1981867	T	439/1416	0.88 [0.75-1.04]	1.34E-01	370/1420	0.93 [0.77-1.12]	4.37E-01
rs10411161	A	439/1416	1.08 [0.89-1.32]	4.24E-01	370/1420	1.07 [0.85-1.33]	5.73E-01
rs3848562	C	439/1416	1.06 [0.86-1.31]	5.72E-01	370/1420	1.10 [0.87-1.39]	4.23E-01
rs11878583	C	439/1416	1.08 [0.87-1.33]	5.00E-01	370/1420	1.12 [0.89-1.41]	3.24E-01

MA, minor allele

*Adjusted for BMI using Cox proportional hazards model

The most notable associations among the ten previously GWAS-identified breast cancer susceptibility loci replicated in this study were with four *FGFR2* SNPs (rs2981579, rs1219648, rs2420946 and rs2981582) and *TNRC9*-rs3803662 (observed $P<7.0 \times 10^{-10}$ and adjusted $P<4.2 \times 10^{-8}$) (**Table S2-1**). The magnitude and direction of the associations were similar to those reported in the original GWASs (observed per allele OR ranges: 1.17 – 1.26) [13-16,18,19], suggesting the robustness of these associations with breast cancer susceptibility. Further, results from the subgroup analyses were consistent with the previous reports [33-35], supporting the hypothesis that *FGFR2* loci (rs1219648, rs2420946 and rs2981582) are associated with increased risk of breast cancer, especially in familial breast cancer cases ($P_{het}<0.02$), and associated with the better prognosis luminal A type or estrogen receptor positive breast cancers ($P_{het}<0.001$) (**Table S2-2**) [33-35].

Of the six putative breast cancer susceptibility SNPs reported in our previous two-stage GWAS, our independent stage 3 analyses successfully replicated the associations of 4q31.22-rs1429142 and 5p15.2-rs1092913 with increased risk of breast cancer. In the combined analyses, five of the six reported associations from our previous GWAS retained statistical significance, the exception being 16q23.2-rs1981867. These five SNPs should be further tested independently in additional cases and controls to assess their role in breast cancer etiology. When adjusted for BMI, we observed near genome-wide significant association for 4q31.22-rs1429142 (adjusted $P=1.7 \times 10^{-7}$) while 5p15.2-rs1092913 remained statistically significant (adjusted $P=1.9 \times 10^{-4}$). For 4q31.22-rs1429142, there was a

substantial increase from the observed ORs (per allele=1.22, $OR_{heterozygote}$=1.26 and $OR_{homozygote}$=1.36) to adjusted ORs (per allele=1.28, $OR_{heterozygote}$=1.32 and $OR_{homozygote}$=1.52). These results indicate that the 4q31.22-rs1429142-breast cancer association may be linked to the BMI pathway of breast cancer risk elevation. This observation is in contrast to the ten GWAS-identified consortia reported SNP-breast cancer associations, and hence requires replication in independent set of breast cancer cases and controls, probably through collaborative efforts involving large international consortia. Both 4q31.22-rs1429142 and 5p15.2-rs1092913 showed statistically significant associations with breast cancer in subgroups with pre and postmenopausal women, cases with luminal and non-luminal A tumors, with and without family history of breast cancer, low and high tumor grade and operable tumor stage at adjusted $P<0.05$ (**Tables 2-3 and 2-4**). However, the association of 4q31.22-rs1429142 was stronger in pre than postmenopausal women (P_{het}=0.002), suggesting that 4q31.22-rs1429142-breast cancer association may vary by menopausal status.

Except for 8q24.21-rs13281615, none of the breast cancer susceptibility SNPs, including 4q31.22-rs1429142, showed significant association with breast cancer outcomes. 8q24.21-rs13281615 was significantly associated with better RFS and OS (adjusted $P<3.1 \times 10^{-3}$) (**Table S2-3**). Similar results for 8q24.21-rs13281615 were also observed in another study involving 13,527 invasive breast cancer cases [33]. To our knowledge, this is the second study to identify the potential prognostic value of 8q24.21-rs13281615 and hence this locus merits further investigation. These results provide further evidence supporting the

hypothesis that the SNPs with prognostic value are yet to be identified using whole genome approaches and that the SNPs associated with breast cancer susceptibility (etiology) are distinct.

4q31.22-rs1429142 is located in a gene desert, with the closest gene endothein receptor type A (*EDNRA*) (**Figure 2-1**) located ~112 kb downstream of the SNP. *EDNRA* gene encoded protein is a cell surface bound receptor involved in several fundamental cellular processes by interacting with endothelins (widely expressed cytokines in various tissues) [36]. SNPs in or near the *EDNRA* gene have been associated with intracranial aneurysm risk [37], hypertension [38] and migraines [39]. This SNP is ~112 kb away from the *EDNRA* gene locus and we therefore queried the SCAN database [40], which uses HapMap human lymphoblastoid cell lines to identify putative expression quantitative trait loci. We found that 4q31.22-rs1429142 is associated with differential expression of five other genes (quantitative transmission disequilibrium test $P<0.0001$, implemented in the SCAN database) involved in at least one type of cancer – *i.e.*, kinesin family member 3B (*KIF3B*) [41], paxillin (*PXN*) [42], general transcription factor IIA, 12 kDa (*GTF2A2*) [43], *PTPRF* interacting protein, binding protein (liprin beta 2) (*PPFIBP2*) [44] and tumor protein p63 regulated 1-like (*TPRG1L*) [45]. However, the allele of 4q31.22-rs1429142 responsible for these is unknown and future fine mapping studies to identify the causal variant and to investigate its allele specific effects are warranted.

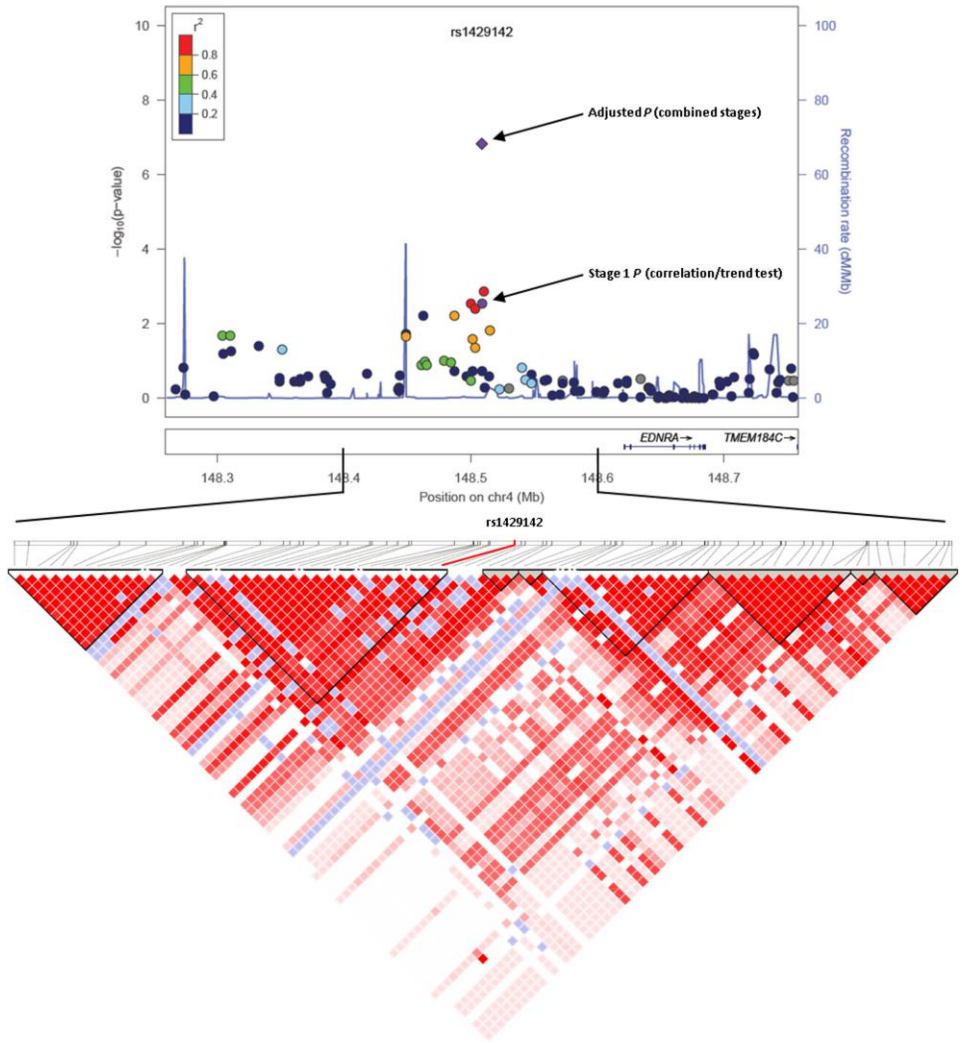


Figure 2-1 Regional association plot (top panel) for 4q31.22-rs1429142 using LocusZoom [48], with the association *P* values (-log10 *P*) on the y-axis and the chromosomal position (hg18) on x-axis. The association of 4q31.22-rs1429142 in stage 1 is shown in purple circle while association in combined stages (1+2+3) after adjusting for BMI is shown in purple diamond. Pair-wise linkage disequilibrium (LD) of 4q31.22-rs1429142 with adjacent SNPs are measured by r^2 values (from HapMap Phase II CEU data) and represented by the color of each circle. Neighbouring Refseq genes are shown

below the plot. LD profiles (bottom panel) among SNPs located within 100 kb up and downstream of the 4q31.22-rs1429142, using HapMap Phase II CEU data are presented.

5p15.2-rs1092913 is also located in a gene desert. The closest gene is rhophilin associated tail protein 1-like (*ROP1NL*) located ~2.5 kb upstream of the polymorphism. *ROP1NL* gene encodes a sperm protein, which is reported to interact with A-kinase anchoring protein. Recently, an independent study (n= 4,325 cases and controls) also showed significant association of 5p15.2-rs1092913 with breast cancer risk in estrogen receptor positive breast cancer of Korean ethnicity, suggesting the potential generalizability of this SNP-breast cancer association in the Korean population [46]. Furthermore, a meta-analysis of two GWASs also found multiple SNPs within the *ROP1NL* locus associated with the phenotype of BMI at 5p15.2, suggesting that this region is important for both breast cancer susceptibility and BMI [47].

In summary, our study not only provided supportive evidence for the robustness of the breast cancer susceptibility SNPs previously identified by consortia, but also identified a new locus at 4q31.22-rs1429142 for contributing to breast cancer susceptibility, lending credence to the continued research efforts in search of common variants for breast cancer.

2.5 References

1. Canadian Cancer Society's Steering Committee on Cancer Statistics. (2012) Canadian cancer statistics 2012, toronto. ON: Canadian cancer society; 2012.
2. Collaborative Group on Hormonal Factors in Breast Cancer. (2001) 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 358: 1389-1399.
3. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from sweden, denmark, and finland. N Engl J Med 343: 78-85.
4. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250: 1684-1689.
5. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, et al. (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science 265: 2088-2090.
6. Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, et al. (1997) Germline mutations of the PTEN gene in cowden disease, an inherited breast and thyroid cancer syndrome. Nat Genet 16: 64-67.
7. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, et al. (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet 38: 873-875.

8. Malkin D, Li FP, Strong LC, Fraumeni JF,Jr, Nelson CE, et al. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250: 1233-1238.

9. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, et al. (2006) Truncating mutations in the fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet 38: 1239-1241.

10. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, et al. (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet 39: 165-167.

11. CHEK2 Breast Cancer Case-Control Consortium. (2004) CHEK2*1100delC and susceptibility to breast cancer: A collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. Am J Hum Genet 74: 1175-1182.

12. Pharoah PDP, Antoniou AC, Easton DF, Ponder BAJ. (2008) Polygenes, risk prediction, and targeted prevention of breast cancer. N Engl J Med 358: 2796-2803.

13. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087-1093.

14. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 39: 870-874.

15. Murabito JM, Rosenberg CL, Finger D, Kreger BE, Levy D, et al. (2007) A genome-wide association study of breast and prostate cancer in the NHLBI's framingham heart study. BMC Med Genet 8 Suppl 1: S6.

16. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, et al. (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 40: 703-706.

17. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, et al. (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci U S A 105: 4340-4345.

18. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, et al. (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41: 585-590.

19. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, et al. (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 41: 579-584.

20. Zheng W, Long J, Gao YT, Li C, Zheng Y, et al. (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 41: 324-328.

21. Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass CE, et al. (2011) Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Hum Genet 130: 529-537.

22. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, et al. (2012) Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet 44: 312-318.

23. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, et al. (2007) A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet 39: 352-358.

24. Thompson D, Easton D. (2004) The genetic epidemiology of breast cancer genes. J Mammary Gland Biol Neoplasia 9: 221-236.

25. McPherson K, Steel CM, Dixon JM. (2000) ABC of breast diseases. breast cancer-epidemiology, risk factors, and genetics. BMJ 321: 624-628.

26. Hjelmborg J, Fagnani C, Silventoinen K, McGue M, Korkeila M, et al. (2008) Genetic influences on growth traits of BMI: A longitudinal study of adult twins. Obesity (Silver Spring) 16: 847-852.

27. Sapkota Y, Robson P, Lai R, Cass CE, Mackey JR, et al. (2012) A two-stage association study identifies methyl-CpG-binding domain protein 2 gene polymorphisms as candidates for breast cancer susceptibility. Eur J Hum Genet 20: 682-689.

28. Bryant H, Robson PJ, Ullman R, Friedenreich C, Dawe U. (2006) Population-based cohort development in Alberta, canada: A feasibility study. Chronic Dis Can 27: 51-59.

29. Therneau T. (2012) A package for survival analysis in S. R package version 2.36-14.

30. R Core Team. (2012) R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

31. Helixtree. SNP & variation suite (version 7.6.11) [software]. bozeman, MT: Golden helix, inc. available from http://www.goldenhelix.com. 2012.

32. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

33. Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, et al. (2008) Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genet 4: e1000054.

34. Campa D, Kaaks R, Le Marchand L, Haiman CA, Travis RC, et al. (2011) Interactions between genetic variants and breast cancer risk factors in the breast and prostate cancer cohort consortium. J Natl Cancer Inst 103: 1252-1263.

35. Garcia-Closas M, Chanock S. (2008) Genetic susceptibility loci for breast cancer by estrogen receptor status. Clin Cancer Res 14: 8000-8009.

36. Wiesmann F, Veeck J, Galm O, Hartmann A, Esteller M, et al. (2009) Frequent loss of endothelin-3 (EDN3) expression due to epigenetic inactivation in human breast cancer. Breast Cancer Res 11: R34.

37. Yasuno K, Bakircioglu M, Low SK, Bilguvar K, Gaal E, et al. (2011) Common variant near the endothelin receptor type A (EDNRA) gene is

associated with intracranial aneurysm risk. Proc Natl Acad Sci U S A 108: 19707-19712.

38. Rahman T, Baker M, Hall D, Avery PJ, Keavney B. (2008) Common genetic variation in the type A endothelin-1 receptor is associated with ambulatory blood pressure: A family study. 22: 282-288.

39. Miao J, Wang F, Fang Y. (2012) Association of 231G>A polymorphism of endothelin type A receptor gene with migraine: A meta-analysis. J Neurol Sci 323: 232-235.

40. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, et al. (2010) SCAN: SNP and copy number annotation. Bioinformatics 26: 259-262.

41. Theriault BL, Pajovic S, Bernardini MQ, Shaw PA, Gallie BL. (2012) Kinesin family member 14: An independent prognostic marker and potential therapeutic target for ovarian cancer. Int J Cancer 130: 1844-1854.

42. Short SM, Yoder BJ, Tarr SM, Prescott NL, Laniauskas S, et al. (2007) The expression of the cytoskeletal focal adhesion protein paxillin in breast cancer correlates with HER2 overexpression and may help predict response to chemotherapy: A retrospective immunohistochemical study. Breast J 13: 130-139.

43. Lee SK, Anzick SL, Choi JE, Bubendorf L, Guan XY, et al. (1999) A nuclear factor, ASC-2, as a cancer-amplified transcriptional coactivator essential for ligand-dependent transactivation by nuclear receptors in vivo. J Biol Chem 274: 34283-34293.

44. Krishnan AV, Shinghal R, Raghavachari N, Brooks JD, Peehl DM, et al. (2004) Analysis of vitamin D-regulated gene expression in LNCaP human prostate cancer cells using cDNA microarrays. Prostate 59: 243-251.

45. Wang X, Zamolyi RQ, Zhang H, Pannain VL, Medeiros F, et al. (2010) Fusion of HMGA1 to the LPP/TPRG1 intergenic region in a lipoma identified by mapping paraffin-embedded tissues. Cancer Genet Cytogenet 196: 64-67.

46. Kim HC, Lee JY, Sung H, Choi JY, Park SK, et al. (2012) A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: Results from the seoul breast cancer study. Breast Cancer Res 14: R56.

47. Wang KS, Liu X, Zheng S, Zeng M, Pan Y, et al. (2012) A novel locus for body mass index on 5p15.2: A meta-analysis of two genome-wide association studies. Gene 500: 80-84.

48. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, et al. (2010) LocusZoom: Regional visualization of genome-wide association scan results. Bioinformatics 26: 2336-2337.

3. A two-stage association study identifies methyl-CpG binding domain protein 2 gene polymorphisms as candidates for breast cancer susceptibility[72]

3.1 Introduction

Breast cancer is a multi-factorial, polygenic disease resulting from the interplay of genetic, environmental and lifestyle risk factors. Linkage studies have revealed that breast cancer tends to cluster in families and disease prevalence is two-fold higher among the first degree relatives of affected individuals [1]. Familial clustering is characterized by early onset of disease often mediated by high-to-moderate penetrance mutations in genes such as those encoding breast cancer (*BRCA1* and *BRCA2*) [2,3], ataxia telangiectasia mutated (*ATM*) [4], cell-cycle checkpoint kinase 2 (*CHEK2*) [5], tumor protein 53 (*TP53*) [6], partner and localizer of *BRCA2* (*PALB2*) [7], *BRCA1* interacting protein C-terminal helicase 1(*BRIP1*) [8] and phosphatase and tensin homolog (*PTEN*) [9]. Nonetheless, these genes in aggregate account for less than 25% of the observed familial genetic risk [10]. A polygenic model has been proposed to explain the remaining genetic risk in non-*BRCA* familial and sporadic breast cancer cases [11]. Single nucleotide polymorphisms (SNPs)-based genome-wide association studies (GWASs have identified low-risk conferring common variants in several complex diseases. For European, Ashkenazi Jewish and Asian population-based GWASs, more than 40 breast cancer susceptibility loci in several genes and intergenic regions have

[72] A version of this chapter has been published. *Sapkota et al., 2012. European Journal of Human Genetics. 20: 682-689.*

already been reported and a subset of these associations have reached genome-wide significance level [12-14]. These variants account for a small proportion of overall genetic risk of breast cancer, leaving open the question of hidden or missing heritability. Current debates suggest that this may be further explained by rare variants, epistasis, epigenetics, gene-environment interactions and copy number variations [15,16].

In a typical GWAS, the frequencies for each SNP (single-locus tests for association) [17] are compared between cases and controls to catalogue polymorphisms potentially associated with the phenotype of interest. The most promising SNPs, sorted based on *p*-value ranking (highest significance) and/or showing significance in haplotype association analysis [18], are selected and replicated in a larger but independent set of cases and controls. In this process, SNPs that are not top-ranked because of their modest *p*-values are ignored and, as a result, potentially informative markers may have been missed. It has been proposed by others [19,20] that even modest associations (*p*-value based), if highly reproducible in independent cohorts, may still be pertinent to the phenotypes under investigation presumably through epistatic interactions (interactions of alleles or genes), a phenomenon strongly implicated in the etiology of breast cancer and the heritable component of genetic risk. Because the majority of the published GWASs concentrate on single locus strategies to identify novel breast cancer susceptibility loci, a candidate gene approach restricted to specific pathway related gene polymorphisms to more effectively mine GWAS data is presented considering moderately associated SNPs. If

reproduced in further independent studies, these may serve as putative candidates for epistatic effects.

Previously reported studies focused on common variants in the genes involved in DNA repair/metabolism pathways and cell cycle regulation, and the markers were selected based on candidate gene approaches [21,22]. In this study, I extend this premise using SNPs in or flanking the DNA repair, modifications and metabolism pathway related genes from the Affymetrix 6.0 array (Stage 1 of GWAS [23]) for independent replication, Stage 2 of the association study design, to identify additional breast cancer susceptibility loci not previously reported.

3.2 Materials and methods

3.2.1 Study population and DNA isolation

I used Stage 1 results of breast cancer GWAS published earlier by the Damaraju Laboratory, described elsewhere [23]. Briefly, sporadic breast cancer cases (n=348), characterized by late onset of disease and controls (n=348) who had no documented history of breast cancer in the first and second degree relatives were selected for Stage 1 of the GWAS [23]. All subjects were predominantly of Caucasian origin. Breast cancer cases (median age=51 years; age range=26-90 years with no. of cases <40 years=35; 40-60 years=241;>60 years=72) were from Alberta, Canada, recruited by the PolyomX Program [24] and the Canadian Breast Cancer Foundation-Tumor Bank (CBCF-TB) [24] during the years 2001-2005 and since 2005-2008, respectively. The two projects, PolyomX Program and CBCF-TB are funded by different granting agencies and nomenclature adopted merely indicates this and in no way reflects bias in

sampling of population. All cases had a histologically confirmed diagnosis of invasive ductal breast carcinoma at the time of enrolment in the study. Gender matched healthy controls (median age=50 years; age range=36-70 years with no. of controls<40 years=50; 40-60 years=226;>60 years=72), also from Alberta, Canada (accessed from the Tomorrow Project [25]), were frequency matched to cases based on age. The proportions of cases and controls for three different age groups (<40 years, 40-60 years and >60 years) were not statistically significant (two-tailed z-test; data not shown). All control subjects enrolled here were free from cancer at the time of recruitment in the study. Potential population confounders were removed leaving cases (n=302) and controls (n=321) for association analysis [23]. Informed consents were obtained from all study participants and the study was approved by Research Ethics Board of Alberta Health Services, Alberta, Canada. Genomic DNA was extracted from the peripheral blood samples of both cases and controls using commercially available Qiagen^{TM} (Mississauga, Ontario, Canada) DNA isolation kits.

3.2.2 SNP selection, genotyping and platform specific genotype concordance

Data filtering and call rate clean-up (Hardy-Weinberg Equilibrium, HWE p>0.001 and SNPs call rate>99%) were carried out as described earlier [23]. Of the 906,600 SNPs genotyped using Affymetrix SNP 6.0, a total of 782,838 SNPs qualified for the downstream analysis. The associations of SNPs with breast cancer were evaluated using correlation/trend tests with 1 degree of freedom (d.f.). Correlation/trend test is similar to chi-square test of independence except

that it is also believed to be a trend test which evaluates correlation of a minor allele with the case status using Pearson's correlation coefficient. The allelic tests with 782,838 SNPs (Stage 1) showed a total of 35,519 SNPs statistically significantly associated with breast cancer at $p<0.05$. Of the 35,519 SNPs, I identified 215 polymorphisms (minor allele frequency, MAF>10%) within or in close proximity to 49 gene regions implicated in pathways or of relevance to DNA repair, modifications and metabolism based on National Center for Biotechnology Information human genome build 37. Six of 215 SNPs were statistically significantly associated with breast cancer at $p<0.001$ (correlation/trend tests with 1 d.f.) and were included for Stage 2 replication study. To reduce the redundancy among the remaining 209 SNPs, I then calculated the pair-wise LD (r^2) among the markers and found that 73 SNPs were strongly correlated ($r^2 \geq 0.8$). Of these 73 short-listed SNPs, 16 were in strong LD ($r^2 \geq 0.8$) with at least one SNP contained within the identified 3,903 haplotype blocks ($p<0.05$) in haplotype association analysis. All haplotypes at a frequency threshold of 1% or more were tested together against the reference haplotype for their associations with breast cancer. The haplotype association analysis *per se* was carried out as described elsewhere [23]. Since my primary objective in this study was to evaluate the moderately associated SNPs from Stage 1 GWAS results, I relaxed the significance threshold in haplotype association analysis to $p<0.05$ as compared to the previous study ($p<0.001$) [23]. Overall, I used allelic tests and haplotype association tests to select SNPs for replication study in an

independent set of 1,178 invasive breast cancer cases and 1,314 healthy individuals serving as controls (Stage 2).

Genotyping assays were performed on Sequenom® iPLEX® Gold platform (services from the McGill University, Genome Quebec Innovation Center, Montreal, Canada).Within- (Sequenom only) and cross-platform (Affymetrix vs. Sequenom) SNP concordances for 22 SNPs were assessed using 205 and 551 duplicate samples, respectively.

3.2.3 Statistical considerations

Allelic associations were evaluated using correlation/trend tests with one d.f. and their corresponding odds ratios (ORs) and 95% confidence intervals (CIs) were estimated using unconditional logistic regression implemented in the SNP & Variation Suite v7.3.1 (Helix Tree Software) [26]. Genotypic associations were also considered for gaining insights in to relative contributions from individual genotypes to breast cancer risk using unconditional logistic regression with two d.f. using the freeware, SNPstats [27] and the results from co-dominant models were summarized in the study. A combined analysis with all samples from Stages 1 and 2 (a total of 1,480 cases and 1,635 controls) was performed to increase the statistical power. The associations for the allelic tests in combined analysis were further examined with 1000-times permutation tests and false discovery rates to identify observations by chance alone (type 1 error) using Helix Tree.

Subgroup analyses were undertaken (correlation/trend tests with 1 d.f.) to identify associations with sub phenotypes within the combined breast cancer cases using a common reference (combined controls) as described previously [28]. The

sub phenotypes examined were family history of breast cancer, menopausal status and luminal A status. Subgroup analyses helps interrogate potential confounding influence of disease heterogeneity on the observed associations and to improve statistical power. Tumors were classified as luminal A based on estrogen and progesterone receptor status (ER^+/PR^+, ER^-/PR^+ and ER^+/PR^-) and human epidermal growth factor receptor 2 status ($HER2^-$) [29]. All the remaining cases were classified as non-luminal A tumors.

The current sample size conferred more than 80% power to detect associations using a co-dominant model for a SNP with 10% MAF, disease prevalence at 1/10 in population for breast cancer, a relative risk of 1.3, type I error of 0.05 and with the LD between markers at r^2 of 0.8 [30].

The LD patterns for regions showing the strongest and consistent associations across Stages 1 and 2 and combined analyses were examined using Haploview v4.2 [31]. For the three methyl-CpG binding domain protein 2 (*MBD2*) SNPs, haplotype frequencies were estimated using SNPstats [27]. The software implements the expectation-maximization algorithm coded into *haplo.stats* package to calculate the estimated relative frequencies for each haplotype [32]. Haplotype association analyses for *MBD2* SNPs were performed with unconditional logistic regression using the default setting of a log-additive model and expressed in terms of ORs and 95% CI (feature available in SNPstats).

3.3 Results

3.3.1 Initial assessment of the data quality

Of the 22 SNPs selected for replication in Stage 2, genotyping for one SNP (rs17519016) was not successful. The cross-platform (Affymetrix vs. Sequenom) SNP call concordance for the remaining 21 SNPs using 551 duplicate samples from Stage 1 was more than 98%. Within-platform (Sequenom) SNP call concordance among the 205 duplicates used in Stage 2 was more than 99.4%. Per sample and per SNP call rates for Stage 2 were >98.3% and >98.4%, respectively, and all 21 SNPs were in HWE proportion at $p>0.001$ in controls (**Table 3-1**).

Cross-platform and within-platform discordances were very low (<2%) and are in agreement with previously reported GWAS studies [12,23]. Further, the MAFs were consistent among the two stages and also comparable to HapMap Central Europeans (CEU) population (data not shown), indicating that the scope of false positive associations due to genotyping errors (systematic or random) was effectively minimized.

Table 3-1 SNPs characteristics used in the study.

SNPs (Cytoband)	Associated gene/ Position[a]	Gene relationship/ distance (bps)	MA	Stage 1[b]			Stage 2[c]			Stages 1+2[d]		
				MAF	HWE $p_{controls}$	Call rate[e]	MAF	HWE $p_{controls}$	Call rate[e]	MAF	HWE $p_{controls}$	Call rate[e]
rs17622933 (4p15.2)	*DHX15* / 24790680	0/Intron	T	0.30	0.06	1.00	0.31	0.52	1.00	0.31	0.89	1.00
rs7200108 (16p13.12)	*ERCC4*/ 13438884	575129/ Upstream	G	0.12	0.61	1.00	0.12	0.94	1.00	0.12	0.66	1.00
rs7317643 (13q33.3)	*LIG4*/ 108536028	323765/ Downstream	A	0.12	0.95	1.00	0.14	0.53	1.00	0.13	0.52	1.00
rs1646807 (18q21.2)	*MBD2*/ 51388197	289773/ Downstream	T	0.17	0.80	0.99	0.19	1.00	1.00	0.18	0.91	1.00
rs4041245 (18q21.2)	*MBD2*/ 51685525	0/Intron	G	0.42	0.68	1.00	0.41	0.01	1.00	0.41	0.04	1.00
rs656923 (18q21.2)	*MBD2*/ 51701796	0/Intron	G	0.19	0.71	1.00	0.20	0.74	1.00	0.20	0.60	1.00
rs7239408 (18q21.2)	*MBD2*/ 51432800	245170/ Downstream	A	0.23	0.61	1.00	0.24	0.51	1.00	0.24	0.43	1.00
rs7614 (18q21.2)	*MBD2*/ 51681244	0/3'UTR	C	0.42	0.27	1.00	0.41	0.03	1.00	0.41	0.17	1.00
rs8094493 (18q21.2)	*MBD2*/ 51700391	0/Intron	G	0.42	0.68	1.00	0.41	0.02	1.00	0.41	0.05	1.00
rs904276 (18q21.2)	*MBD2*/ 51434340	243630/ Downstream	C	0.22	0.47	1.00	0.23	0.39	1.00	0.23	0.29	1.00
rs2044760 (2q23.1)	*MBD5*/ 148989731	0/Intron	T	0.39	0.23	1.00	0.35	0.21	1.00	0.36	0.51	1.00
rs1556459 (10q26.3)	*MGMT*/ 130810711	454766/ Upstream	C	0.16	0.80	1.00	0.15	0.07	1.00	0.15	0.12	1.00
rs3996018 (3q13.13)	*MYH15*/ 108163202	0/Intron	G	0.24	0.43	1.00	0.23	0.28	1.00	0.23	0.18	1.00
rs13250873 (8q24.11)	*RAD21*/ 117806169	52004/ Downstream	G	0.31	0.52	1.00	0.32	0.27	1.00	0.32	0.20	1.00
rs2297381 (15q15.1)	*RPAP1*/ 41827655	0/Intron	G	0.50	0.84	1.00	0.48	0.67	1.00	0.48	0.63	1.00
rs6893184 (5q23.1)	*TNFAIP8*/ 118730867	574/ Downstream	G	0.34	0.63	1.00	0.33	0.0017	1.00	0.33	0.00240	1.00
rs7721752 (5q23.1)	*TNFAIP8*/ 118746110	15817/ Downstream	G	0.33	0.77	1.00	0.30	0.01	1.00	0.31	0.04	1.00
rs6795465 (3q28)	*TP63*/ 189537521	0/Intron	C	0.12	0.42	1.00	0.13	0.27	0.99	0.13	0.17	0.99

Table 3-1 Continued..

SNPs (Cytoband)	Associated gene/ Position[a]	Gene relationship/ distance (bps)	MA	Stage 1[b]			Stage 2[c]			Stages 1+2[d]		
				MAF	HWE $p_{controls}$	Call rate[e]	MAF	HWE $p_{controls}$	Call rate[e]	MAF	HWE $p_{controls}$	Call rate[e]
rs7636114 (3q28)	*TPRG1*/ 189088705	45612/ Downstream	C	0.11	0.80	1.00	0.11	0.14	1.00	0.11	0.21	1.00
rs7700025 (4q34.3)	*VEGFC*/ 177814863	100968/ Upstream	G	0.28	0.50	1.00	0.30	0.79	1.00	0.30	0.91	1.00
rs9992272 (4q34.3)	*VEGFC*/ 177737136	23241/ Upstream	C	0.16	0.71	1.00	0.17	0.99	0.98	0.16	0.93	0.99

[a]From NCBI human genome build GRCh37; *DHX15*, DEAH (Asp-Glu-Ala-His) box polypeptide 15; *ERCC4*, excision repair cross-complementing rodent repair deficiency, complementation group 4; *LIG4*, ligase IV, DNA, ATP-dependent; *MBD2*, methyl-CpG binding domain protein 2; *MBD5*, methyl-CpG binding domain protein 5; *MGMT*, O-6-methylguanine-DNA methyltransferase; *MYH15*, myosin, heavy chain 15; *RAD21*, *RAD21* homolog (*S. pombe*); *RPAP1*, RNA polymerase II associated protein 1; *TNFAIP8*, tumor necrosis factor, alpha-induced protein 8; *TP63*, tumor protein 63; *TPRG1*, tumor protein p63 regulated 1; *VEGFC*, vascular endothelial growth factor C; [b](302 cases and 321 controls); [c](1178 cases and 1314 controls); [d](1480 cases and 1635 controls); MAF, combined minor allele frequency in both cases and controls; ecombined SNP call rate in both cases and controls

3.3.2 Stage 2 analysis

In Stage 2, six SNPs showed suggestive associations with breast cancer (**Table 3-2**).Three SNPs (rs8094493, rs4041245 and rs7614) were from *MBD2* gene regions and were marginally associated with reduced risk for breast cancer (ORs, 0.90, 0.91 and 0.92, respectively) (**Table 3-2**). The other three SNPs rs13250873, rs1556459 and rs2297381 were located in or close proximity of *RAD21* homolog (*S. pombe*) (*RAD21*), O-6-methylguanine-DNA methyltransferase (*MGMT*) and RNA polymerase II associated protein 1 (*RPAP1*) gene regions, respectively, and showed suggestive associations with increased risk for breast cancer.

Table 3-2 Six SNPs with the strongest and consistent associations with breast cancer susceptibility across stages 1, 2 and in combined analysis.

SNPs	Allele or genotype	Stage 1[a]		Stage 2[b]	
		OR, 95% CI	*p* -value[d]	OR, 95% CI	*p* -value[d]
rs8094493	G (minor allele)	0.68 (0.54,0.85)	0.0009	0.90 (0.81,1.01)	0.0773
	GT	0.66 (0.46-0.94)	0.0044	0.80 (0.67-0.95)	0.0410
	GG	0.48 (0.31-0.77)		0.86 (0.68-1.09)	
rs4041245	G (minor allele)	0.68 (0.54,0.85)	0.0009	0.91 (0.81,1.02)	0.0893
	GA	0.66 (0.46-0.94)	0.0044	0.79 (0.67-0.94)	0.0340
	GG	0.48 (0.31-0.77)		0.87 (0.69-1.10)	
rs7614	C (minor allele)	0.67 (0.54,0.84)	0.0006	0.92 (0.82,1.03)	0.1356
	CT	0.70 (0.49-0.99)	0.0038	0.81 (0.68-0.97)	0.0690
	CC	0.47 (0.30-0.74)		0.89 (0.70-1.12)	
rs13250873	G (minor allele)	1.29 (1.01,1.64)	0.0383	1.14 (1.01,1.28)	0.0306
	GA	1.34 (0.96-1.87)	0.1100	1.10 (0.93-1.30)	0.0910
	GG	1.56 (0.91-2.68)		1.33 (1.02-1.73)	
rs1556459	C (minor allele)	1.50 (1.10,2.04)	0.0102	1.13 (0.97,1.32)	0.1151
	CT	1.49 (1.03-2.14)	0.0390	1.05 (0.88-1.26)	0.0740
	CC	2.21 (0.80-6.07)		1.89 (1.07-3.32)	
rs2297381	G (minor allele)	1.27 (1.01,1.58)	0.0368	1.10 (0.98,1.23)	0.0986
	GA	1.28 (0.87-1.89)	0.1100	1.02 (0.85-1.24)	0.1800
	GG	1.60 (1.03-2.50)		1.21 (0.97-1.50)	

Table 3-2 Continued..

SNPs	Allele or genotype	Stages 1+2 (combined analysis)[c]			
		OR, 95% CI	*p*-value[d]	FDR	Permutation *p*-value[e]
rs8094493	G (minor allele)	0.85 (0.77,0.94)	**0.0021**	0.045	0.038
	GT	0.77 (0.66-0.90)	0.0019		
	GG	0.76 (0.61-0.94)			
rs4041245	G (minor allele)	0.86 (0.77,0.95)	**0.0026**	0.027	0.048
	GA	0.76 (0.65-0.89)	0.0018		
	GG	0.77 (0.62-0.95)			
rs7614	C (minor allele)	0.86 (0.78,0.95)	**0.0041**	0.029	0.069
	CT	0.79 (0.67-0.92)	0.0053		
	CC	0.77 (0.63-0.95)			
rs13250873	G (minor allele)	1.17 (1.05,1.30)	**0.0043**	0.023	0.07
	GA	1.14 (0.98-1.33)	0.0190		
	GG	1.37 (1.08-1.74)			
rs1556459	C (minor allele)	1.20 (1.04,1.37)	**0.0103**	0.043	0.161
	CT	1.13 (0.96-1.32)	0.0120		
	CC	1.96 (1.20-3.20)			
rs2297381	G (minor allele)	1.13 (1.02,1.25)	**0.0154**	0.054	0.234
	GA	1.07 (0.90-1.27)	0.0430		
	GG	1.28 (1.05-1.55)			

[a](302 cases and 321 controls); [b](1178 cases and 1314 controls); [c](1480 cases and 1635 controls); OR, odds ratio; CI, confidence interval; [d]*p* -values calculated from correlation/trend test with 1 degree of freedom using multiplicative model and unconditional logistic regression with 2 degrees of freedom using co-dominant genotypic model; FDR, false discovery rate for observed associations in joint analysis using multiplicative model; [e]1000-times permutation p-value for observed associations in combined analysis using multiplicative model.

The association test results for the remaining 15 SNPs are summarized in **Table S3-1.** Fourteen of these showed no statistical significance and one SNP (rs7636114) showed suggestive association trend in Stage 2 (but in opposite direction to the Stage 1 results) and is therefore not considered for further analysis.

Table S3-1 Fifteen SNPs and their association statistics from stages 1, 2 and in combined analysis.

SNPs	Allele or genotype	Stage 1[a]		Stage 2[b]		Stages 1+2 (combined analysis)[c]			
		OR, 95% CI	*p*-value[d]	OR, 95% CI	*p*-value[d]	OR, 95% CI	*p*-value[d]	FDR	Permutation *p*-value[e]
rs1646807	T (minor allele)	0.72 (0.53,0.97)	0.0325	0.99 (0.86,1.14)	0.8956	0.93 (0.82,1.06)	0.2893	0.380	0.996
	TA	0.74 (0.52-1.05)	0.0880	0.98 (0.82-1.16)	0.9700	0.93 (0.79-1.08)	0.5600		
	TT	0.43 (0.15-1.26)		1.02 (0.67-1.56)		0.89 (0.60-1.32)			
rs17622933	T (minor allele)	1.47 (1.16,1.88)	0.0017	1.00 (0.89,1.13)	0.9972	1.08 (0.97,1.20)	0.1630	0.311	0.952
	TA	1.37 (0.99-1.91)	0.0033	1.03 (0.87-1.21)	0.9000	1.09 (0.94-1.26)	0.3700		
	TT	2.76 (1.45-5.23)		0.97 (0.73-1.28)		1.16 (0.90-1.49)			
rs2044760	T (minor allele)	0.75 (0.60,0.94)	0.0142	0.96 (0.85,1.08)	0.4873	0.91 (0.82,1.01)	0.0870	0.203	0.795
	TC	0.64 (0.45-0.90)	0.0230	0.99 (0.83-1.17)	0.7200	0.91 (0.78-1.06)	0.2400		
	TT	0.61 (0.38-1.00)		0.90 (0.70-1.16)		0.84 (0.67-1.05)			
rs3996018	G (minor allele)	1.32 (1.02,1.72)	0.0358	0.91 (0.79,1.03)	0.1415	0.98 (0.87,1.10)	0.7172	0.753	1
	GC	1.43 (1.03-2.00)	0.0840	0.87 (0.73-1.03)	0.2500	0.96 (0.83-1.12)	0.8800		
	GG	1.46 (0.74-2.88)		0.90 (0.65-1.26)		0.99 (0.74-1.34)			
rs656923	G (minor allele)	1.51 (1.13,2.01)	0.0046	1.00 (0.87,1.15)	0.9798	1.08 (0.96,1.22)	0.2113	0.341	0.983
	GA	1.54 (1.09-2.18)	0.0180	0.96 (0.81-1.14)	0.6900	1.05 (0.90-1.22)	0.3900		
	GG	2.08 (0.89-4.85)		1.13 (0.77-1.64)		1.25 (0.89-1.77)			
rs6795465	C (minor allele)	0.57 (0.40,0.81)	0.0016	0.96 (0.81,1.14)	0.6483	0.87 (0.75,1.01)	0.0699	0.183	0.714
	CT	0.61 (0.41-0.91)	0.0055	0.98 (0.81-1.19)	0.8300	0.90 (0.75-1.06)	0.1700		
	CC	0.21 (0.04-0.98)		0.84 (0.47-1.50)		0.67 (0.39-1.14)			
rs6893184	G (minor allele)	1.27 (1.00,1.60)	0.0493	1.08 (0.96,1.21)	0.2145	1.11 (1.00,1.24)	0.0455	0.137	0.546
	GA	1.41 (1.01-1.97)	0.0970	1.05 (0.89-1.24)	0.4700	1.11 (0.96-1.30)	0.1500		
	GG	1.45 (0.86-2.46)		1.17 (0.91-1.49)		1.21 (0.97-1.52)			
rs7200108	G (minor allele)	0.55 (0.39,0.78)	0.0006	1.07 (0.90,1.27)	0.4455	0.93 (0.80,1.09)	0.3862	0.477	0.998
	GA	0.50 (0.34-0.76)	0.0022	1.08 (0.89-1.31)	0.7300	0.93 (0.78-1.11)	0.6900		
	GG	0.51 (0.17-1.55)		1.08 (0.55-2.10)		0.89 (0.50-1.57)			
rs7239408	A (minor allele)	0.66 (0.51,0.86)	0.0024	1.01 (0.89,1.15)	0.8464	0.93 (0.83,1.05)	0.2369	0.332	0.989
	AG	0.68 (0.48-0.94)	0.0076	0.96 (0.81-1.14)	0.6100	0.90 (0.77-1.04)	0.3500		
	AA	0.38 (0.17-0.85)		1.15 (0.81-1.62)		0.95 (0.69-1.30)			
rs7317643	A (minor allele)	1.53 (1.08,2.16)	0.0162	1.06 (0.90,1.24)	0.4929	1.13 (0.98,1.31)	0.1023	0.215	0.838
	AG	1.43 (0.97-2.12)	0.0510	1.06 (0.88-1.28)	0.7900	1.12 (0.95-1.33)	0.2700		
	AA	3.11 (0.81-11.85)		1.09 (0.62-1.90)		1.30 (0.78-2.15)			
rs7636114	C (minor allele)	1.65 (1.16,2.35)	0.0053	0.84 (0.70,1.01)	0.0658	0.97 (0.83,1.14)	0.7249	0.725	1
	CG	1.52 (1.01-2.27)	0.0210	0.79 (0.65-0.97)	0.0690	0.90 (0.76-1.08)	0.1800		
	CC	3.53 (0.94-13.20)		1.16 (0.53-2.55)		1.60 (0.83-3.10)			
rs7700025	G (minor allele)	1.34 (1.05,1.71)	0.0208	0.98 (0.87,1.10)	0.7272	1.04 (0.93,1.16)	0.4903	0.572	1
	GA	1.33 (0.96-1.86)	0.0760	0.95 (0.80-1.12)	0.8000	1.01 (0.87-1.17)	0.7000		
	GG	1.74 (0.97-3.10)		1.00 (0.75-1.32)		1.11 (0.87-1.43)			
rs7721752	G (minor allele)	1.32 (1.04,1.68)	0.0203	1.02 (0.90,1.15)	0.7875	1.07 (0.96,1.20)	0.1897	0.332	0.971
	GA	1.43 (1.03-2.00)	0.0510	1.01 (0.85-1.20)	0.9600	1.09 (0.94-1.26)	0.4300		
	GG	1.66 (0.95-2.90)		1.04 (0.80-1.35)		1.13 (0.89-1.44)			
rs904276	C (minor allele)	0.62 (0.47,0.81)	0.0005	1.02 (0.90,1.17)	0.7229	0.93 (0.83,1.04)	0.2168	0.325	0.983
	CT	0.62 (0.44-0.86)	0.0017	0.96 (0.81-1.13)	0.4400	0.88 (0.76-1.02)	0.2300		
	CC	0.34 (0.15-0.80)		1.21 (0.85-1.73)		0.98 (0.71-1.35)			
rs9992272	C (minor allele)	1.41 (1.03,1.92)	0.0292	0.95 (0.82,1.11)	0.5123	1.03 (0.90,1.17)	0.7145	0.790	1
	CT	1.40 (0.98-2.00)	0.0880	0.96 (0.80-1.15)	0.7900	1.03 (0.88-1.21)	0.9200		
	CC	2.13 (0.70-6.45)		0.87 (0.53-1.43)		1.02 (0.65-1.59)			

[a](302 cases and 321 controls); [b](1178 cases and 1314 controls); [c](1480 cases and 1635 controls); OR, odds ratio; CI, confidence interval; [d]p-values calculated from correlation/trend test with 1 degree of freedom using multiplicative model and unconditional logistic regression with 2 degrees of freedom using co-dominant genotypic model; FDR, false discovery rate for observed associations in joint analysis using multiplicative model; [e]1000-times permutation p-value for observed associations in joint analysis using multiplicative model.

3.3.3 Combined analysis (Stages 1 and 2)

I combined the results for six SNPs from Stages 1 and 2 and conducted a combined analysis and found similar in direction of risk but also stronger association signals for all six variants (**Table 3-2**). The *MBD2* SNPs rs8094493 (OR:0.85, $p<0.0021$), rs4041245 (OR:0.86, $p<0.0026$) and rs7614 (OR:0.86, $p<0.0041$) were significantly associated with reduced risk of breast cancer. The observed false discovery rates (FDR) of 0.045, 0.027 and 0.029, respectively, for the allelic associations in combined analysis provided confidence in the study findings. I also subjected the data to permutation testing (1000-times) and observed permutation *p*-values of 0.038, 0.048 and 0.069, respectively, an indication that the reported findings may not be attributed to associations by chance alone. The heterozygote and variant homozygote genotypes of *MBD2* SNPs from co-dominant models also conferred similar trends of reduced risks of breast cancer (ORs:0.76-0.79).

The remaining polymorphisms analyzed (rs13250873, rs1556459 and rs2297381, **Table 3-2**) also showed significant associations, except the direction of risk for breast cancer (allelic ORs, 1.13-1.20) were in opposite direction to the ones observed for *MBD2* SNPs. The association signals for all three SNPs were characterized by low FDR values (0.023 to 0.054); the 1000-times permutation tests also showed marginal significance for rs13250873. In the co-dominant genotypic models, variant homozygotes (OR $\geq$1.28) showed stronger associations than heterozygotes (OR 1.07 to 1.14) in the combined analysis for rs13250873, rs1556459 and rs2297381.

3.3.4 Subgroup analyses

Due to potential for genetic risk determinants to be associated with specific clinical and molecular subtypes of breast cancer, I reviewed clinicopathological characteristics of the cases in both Stages 1 and 2 and conducted stratified analyses (**Table 3-3**). I evaluated allelic associations for six SNPs with the following subgroups: without and with family history of breast cancer, pre- and post-menopausal status and luminal A and non-luminal A (*i.e.*, good and poor prognostic groups, respectively) breast cancer status of the tumors, using correlation/trend tests with one d.f. I found associations between clinicopathological characteristics and the polymorphisms considered, and the observed ORs were consistent across subgroups (**Table 3-3**).

None of the observed associations were stronger than the single locus effects and hence it is less likely that these clinicopathological characteristics (potential confounders) have significant effects on initial observed associations with unstratified cases (**Table 3-2**).

3.3.5 Pair-wise LD profiling between markers

I examined LD profiles for the six identified variants (**Table 3-2**) using HapMap II CEU genotype data (available from www.hapmap.org). I found that three *MBD2* SNPs (rs8094493, rs4041245 and rs7614) in intron 3, intron 6 and the 3` untranslated region, respectively, were in strong LD with *D*`=1 (**Figure 3-1a**) and these profiles were also observed in the curent study population (**Figure 3-1b**). rs7614 and rs4041245 were located in a LD block spanning ~6 kb region and rs8094493 was located in a LD block of ~9 kb region.

Table 3-3 Subgroup analyses based on family history of breast cancer, menopausal status and luminal A tumors.

SNPs	Premenopausal women[a]		Postmenopausal women[b]		Cases with family history[c]	
	OR, 95% CI	*p* value	OR, 95% CI	*p* value	OR, 95% CI	*p* value
rs8094493	0.84 (0.74,0.96)	0.011	0.86 (0.76,0.97)	0.016	0.86 (0.75,0.99)	0.031
rs4041245	0.85 (0.74,0.97)	0.015	0.86 (0.76,0.97)	0.016	0.87 (0.76,0.99)	0.041
rs7614	0.85 (0.75,0.97)	0.018	0.87 (0.77,0.98)	0.027	0.88 (0.77,1.01)	0.079
rs13250873	1.22 (1.06,1.40)	0.005	1.12 (0.98,1.27)	0.085	1.18 (1.02,1.36)	0.024
rs1556459	1.27 (1.07,1.51)	0.008	1.16 (0.99,1.37)	0.074	1.20 (1.00,1.45)	0.047
rs2297381	1.22 (1.07,1.39)	0.003	1.09 (0.97,1.22)	0.165	1.19 (1.04,1.36)	0.012

[a]623 premenopausal women; [b]829 postmenopausal women: [c]575 cases with family history of breast cancer in their first or second degree relatives; [d]808 cases without family history of breast cancer; [e]761 cases with luminal A tumors and [f]397 cases with non-luminal A tumors; associations of SNPs with each of these subgroups were assessed using common 1635 controls and expressed in terms of odds ratios (OR) and 95% confidence interval (CI); the *p* values were obtained from a correlation/trend test with 1 degree of freedom

Table 3-3 Continued..

SNPs	Cases without family history[d]		Luminal A tumors[e]		Non-luminal A tumors[f]	
	OR, 95% CI	*p* value	OR, 95% CI	*p* value	OR, 95% CI	*p* value
rs8094493	0.85 (0.75,0.96)	0.009	0.88 (0.78,0.99)	0.039	0.80 (0.68,0.94)	0.006
rs4041245	0.85 (0.75,0.96)	0.009	0.88 (0.78,1.00)	0.043	0.80 (0.68,0.94)	0.006
rs7614	0.85 (0.75,0.96)	0.010	0.89 (0.79,1.01)	0.064	0.80 (0.68,0.94)	0.007
rs13250873	1.15 (1.01,1.31)	0.029	1.11 (0.97,1.26)	0.117	1.20 (1.02,1.41)	0.030
rs1556459	1.19 (1.01,1.40)	0.037	1.17 (0.99,1.39)	0.062	1.20 (0.97,1.48)	0.092
rs2297381	1.13 (1.00,1.27)	0.044	1.15 (1.02,1.30)	0.024	1.17 (1.00,1.36)	0.050

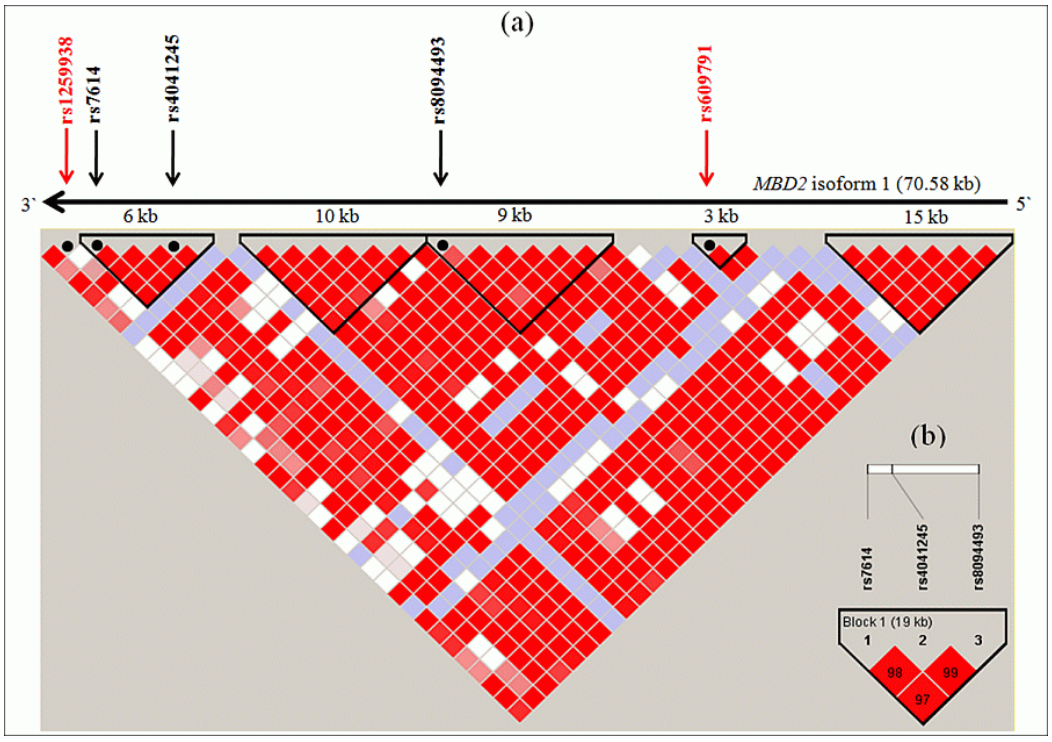


Figure 3-1 Pair wise LD profiles between SNPs from MBD2 gene region. (a) LD profile of whole MBD2 isoform 1 spanning ~70.58 kbps. The gene is in reverse orientation (3`- 5`) on chromosome 18q arm. Five SNPs (three from the current study, shown in black and two from Zhu et al. [35], shown in red) in MBD2 gene regions are shown based on their relative position on HapMap CEU dataset (Phase 1 & 2-full dataset). LD blocks were defined using "confidence interval" method as explained by Gabriel et al. [42]. D` values are given for LD between the markers. The darker the cell, the greater the D` value between the SNPs. (b) LD profile for three MBD2 SNPs from the current study based on the current study population.

3.3.6 Haplotype analysis for MBD2 gene polymorphisms

I reasoned that the highly correlated SNPs from the *MBD2* gene region may form distinct haplotypes that could potentially explain the population diversity. Polymorphisms, rs8094493, rs4041245 and rs7614 formed two major haplotypes, one with common alleles (major allele) and other with variant alleles (minor allele). The common haplotype had a population frequency of 0.58 (0.60 for cases and 0.56 for controls) and the variant haplotype had a population frequency of 0.40 (0.38 for cases and 0.42 for controls). The variant form was significantly associated with the reduced risk of breast cancer (OR:0.86, p<0.003) (**Table 3-4**). The population diversity that could be explained by the two major haplotypes identified in this analysis was 98%.

Table 3-4 Haplotypes for three MBD2 SNPs and their associations with breast cancer risk.

Haplotype			Frequency		Total	OR	95% CI	p value
rs4041245	rs8094493	rs7614	Cases	Controls				
T	A	T	0.601	0.5624	0.5823	1		
C	G	G	0.385	0.4232	0.4053	0.86	0.77-0.95	**0.0029**
Rare	Rare	Rare	*	*	0.0124	1.11	0.70-1.78	0.66

3.4 Discussion

In this study, I identified SNPs associated with breast cancer among genes related to DNA repair, modifications and metabolism. A total of six loci were identified using a two-stage association study design, and these were not previously reported in published GWASs for breast cancer [12-14,23] as putative markers for breast cancer susceptibility. The identified loci were highly

reproducible in an independent study (Stage 2) and the statistical significance of the findings were consistent across study stages, in the combined analysis, and across clinicopathological subtypes of breast cancer. These loci are promising markers and warrant independent validation in Caucasian population or in diverse ethnic cohorts to evaluate the generalizability of my findings.

The six loci identified were from four chromosomes 18, 15, 10 and 8. Both single locus and haplotype association analyses indicated that *MBD2* gene loci (rs8094493, rs4041245 and rs7614) conferred protection against breast cancer. The magnitude and the direction of the association signals in both stages were consistent between allelic and genotypic models (**Table 3-2**). The allelic risk effects were enriched in combined analysis with stronger association *p*-values of $<10^{-3}$. Low FDR values and permutation testing provided further confidence in my findings by ruling out the observations as false positives. Mechanistic relationships to breast carcinogenesis are suggested because *MBD2* is a well characterized gene and the encoded protein binds methylated promoter regions and mediates transcriptional repression of tumor suppressor genes [33]. DNA (cytosine-5)-methyltransferase 1 (*DNMT1*) is reported to interact with the methyl-CpG binding protein complex, *MBD2* and *MBD3* at late S-phase replication foci and as such, these interactions may direct *DNMT1* to hemimethylated sequences following DNA replication and silencing of genes in S-phase [34].

Earlier, Zhu *et al.* reported the associations of two SNPs (rs1259938 and rs609791) in *MBD2* gene regions with the reduced risk of breast cancer in pre-menopausal Caucasian women [35]. I evaluated for possible LD between the

distinct *MBD2* SNPs reported here and those reported by Zhu *et al.* [35] The polymorphisms reported by earlier investigators were not in LD with the SNPs reported here (**Figure 3-1a**). The notable differences between our study and those by Zhu *et al.* [35] are (i) the SNPs rs1259938 and rs609791 in the previous study did not show association with the breast cancer phenotype in unstratified cases although they showed statistical significance when cases were stratified by pre- and post-menopausal status; (ii) I identified distinct *MBD2* gene SNPs and these were all statistically significantly associated with breast cancer as a phenotype even in both unstratified (**Table 3-2**) and stratified cases (**Table 3-3**); (iii) sample sizes were substantially larger in the curent study (total sample size of 1480 cases and 1615 controls) as opposed to 393 cases and 436 controls from the nested case-control study with a Caucasian population reported by Zhu *et al.* [35] In summary, observations with a larger sample size (this study) showed association with breast cancer even without stratification of cases and the haplotypes associated were also distinct. However, it is important to note that the magnitude and direction of risk and the gene identified is similar in both studies. I did not genotype the polymorphisms reported by Zhu *et al.* [35] at this time, and may therefore require independent validation. The SNPs analyzed by Zhu *et al.* [35] were not present in the Affymetrix SNP 6.0 array.

Other genes/loci were identified for breast cancer risk in this study. rs2297381 was located in intron 5 of *RPAP1* and was associated with the risk of breast cancer. *RPAP1* is a poorly understood gene possibly involved in the interaction of RNA polymerase II and its regulators of protein complex formation

[36]. To my knowledge, this is the first report on *RPAP1*gene SNP associated with breast cancer risk. rs13250873 and rs1556459, located ~52 kb downstream of *RAD21* and ~454 kb upstream of *MGMT*, respectively, were significantly associated with the risk of breast cancer across both stages and in combined analysis. Both *RAD21* and *MGMT* are well-studied genes with significant roles in carcinogenesis. The *RAD21* protein is involved in double-strand breaks repair as well as chromatid cohesion during mitosis [37,38]. Intronic polymorphisms in *RAD21*gene have been associated with breast cancer in high-risk population [39]. Similarly, *MGMT* repairs the alkylated guanine due to carcinogenic effects induced by alkylating agents [40]. Coding SNPs of *MGMT* gene are reported to be associated with breast cancer risk [41]. *MGMT* SNP reported in this study is ~454 kb upstream of the *MGMT* gene. Although rs13250873 and rs1556459 were not located in the gene regions, further replication of these findings and fine mapping of these loci is required to determine if the identified polymorphisms exert their action through regulation of the nearby *RAD21* and *MGMT* genes.

None of the associations reached genome-wide significance level in this two-stage association study with the combined sample size of 1,480 cases and 1,635 controls. However, confidence in the reported associations stems from the stringent quality control parameters employed (>98% SNP and sample call rates, HWE p>0.001 and >98% SNP concordance in replicates and good call rate concordance across platforms). Furthermore, the low FDR values and results from permutation testing should favour considering the reported polymorphisms for replication in independent studies. In summary, I identified additional breast

cancer susceptibility loci in Caucasian women by focusing on genes related to DNA repair, modifications and metabolism. The curent study supports the concept of investigating moderate association signals from Stage 1 GWAS using a candidate gene approach restricted to specific pathway related gene polymorphisms. In this study I did not consider all related DNA repair/modifications/metabolism pathway gene polymorphisms or their potential associations with other subtypes of breast cancer (basal, $HER2^{+}$ and luminal B) due to limitations in sample size. Other reported DNA repair/modifications/metabolism gene polymorphisms (that did not reach genome-wide significance) in previously published studies, if replicated in independent cohorts, should also be considered along with the six reported variants here as putative candidates for epistatic models to gain insights to the missing heritability of sporadic breast cancer.

3.5 References

1. Byrne C, Brinton LA, Haile RW, Schairer C. (1991) Heterogeneity of the effect of family history on breast cancer risk. Epidemiology (Cambridge, Mass.) 2: 276-284.

2. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science (New York, N.Y.) 250: 1684-1689.

3. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, et al. (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science (New York, N.Y.) 265: 2088-2090.

4. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, et al. (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature Genetics 38: 873-875.

5. CHEK2 Breast Cancer Case-Control Consortium. (2004) CHEK2*1100delC and susceptibility to breast cancer: A collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. American Journal of Human Genetics 74: 1175-1182.

6. Malkin D, Li FP, Strong LC, FraumeniJF,Jr, Nelson CE, et al. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science (New York, N.Y.) 250: 1233-1238.

7. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, et al. (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nature Genetics 39: 165-167.

8. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, et al. (2006) Truncating mutations in the fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nature Genetics 38: 1239-1241.

9. Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, et al. (1997) Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nature Genetics 16: 64-67.

10. Easton DF. (1999) How many more breast cancer predisposition genes are there? Breast Cancer Research: BCR 1: 14-17.

11. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, et al. (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nature Genetics 31: 33-36.

12. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, et al. (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nature Genetics 41: 585-590.

13. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087-1093.

14. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, et al. (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. Nature Genetics 42: 504-507.

15. Robinson R. (2010) Common disease, multiple rare (and distant) variants. PLoS Biology 8: e1000293.

16. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews.Genetics 11: 446-450.

17. Roeder K, Bacanu SA, Sonpar V, Zhang X, Devlin B. (2005) Analysis of single-locus tests to detect gene/disease associations. Genetic Epidemiology 28: 207-219.

18. Zhang K, Calabrese P, Nordborg M, Sun F. (2002) Haplotype block structure and its applications to association studies: Power and study designs. American Journal of Human Genetics 71: 1386-1394.

19. Lo SH, Chernoff H, Cong L, Ding Y, Zheng T. (2008) Discovering interactions among BRCA1 and other candidate genes associated with sporadic breast cancer. Proceedings of the National Academy of Sciences of the United States of America 105: 12387-12392.

20. Musani SK, Shriner D, Liu N, Feng R, Coffey CS, et al. (2007) Detection of gene x gene interactions in genome-wide association studies of human population data. Human Heredity 63: 67-84.

21. Smith TR, Levine EA, Perrier ND, Miller MS, Freimanis RI, et al. (2003) DNA-repair genetic polymorphisms and breast cancer risk. Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology 12: 1200-1204.

22. Cunningham JM, Vierkant RA, Sellers TA, Phelan C, Rider DN, et al. (2009) Cell cycle genes and ovarian cancer susceptibility: A tagSNP analysis. British Journal of Cancer 101: 1461-1468.

23. Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass CE, et al. (2011) Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human Genetics 130: 529-537.

24. PolyomX 2001 and CBCF-TB 2005. http://www.abtumorbank.com/?about

25. Tomorrow project 2001. http://www.albertahealthservices.ca/tomorrowproject.asp

26. Golden helix, inc.bozeman, MT, USA. HelixTree® software. http://www.goldenhelix.com.

27. Sole X, Guino E, Valls J, Iniesta R, Moreno V. (2006) SNPStats: A web tool for the analysis of association studies. Bioinformatics (Oxford, England) 22: 1928-1929.

28. Mavaddat N, Dunning AM, Ponder BA, Easton DF, Pharoah PD. (2009) Common genetic variation in candidate genes and susceptibility to subtypes of breast cancer. Cancer Epidemiology, Biomarkers &Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology 18: 255-259.

29. Bernstein L, Lacey JV,Jr. (2011) Receptors, associations, and risk factor differences by breast cancer subtypes: Positive or negative? Journal of the National Cancer Institute 103: 451-453.

30. Menashe I, Rosenberg PS, Chen BE. (2008) PGA: Power calculator for case-control genetic association analyses. BMC Genetics 9: 36.

31. Barrett JC, Fry B, Maller J, Daly MJ. (2005) Haploview: Analysis and visualization of LD and haplotype maps. Bioinformatics (Oxford, England) 21: 263-265.

32. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. American Journal of Human Genetics 70: 425-434.

33. Berger J, Bird A. (2005) Role of MBD2 in gene regulation and tumorigenesis. Biochemical Society Transactions 33: 1537-1540.

34. Tatematsu KI, Yamazaki T, Ishikawa F. (2000) MBD2-MBD3 complex binds to hemi-methylated DNA and forms a complex containing DNMT1 at the replication foci in late S phase. Genes to Cells 5: 677-688.

35. Zhu Y, Brown HN, Zhang Y, Holford TR, Zheng T. (2005) Genotypes and haplotypes of the methyl-CpG-binding domain 2 modify breast cancer risk dependent upon menopausal status. Breast Cancer Research : BCR 7: R745-52.

36. Jeronimo C, Langelier MF, Zeghouf M, Cojocaru M, Bergeron D, et al. (2004) RPAP1, a novel human RNA polymerase II-associated protein affinity purified with recombinant wild-type and mutated polymerase subunits. Molecular and Cellular Biology 24: 7043-7058.

37. McKay MJ, Troelstra C, van der Spek P, Kanaar R, Smit B, et al. (1996) Sequence conservation of the rad21 schizosaccharomycespombe DNA

double-strand break repair gene in human and mouse. Genomics 36: 305-315.

38. Sonoda E, Matsusaka T, Morrison C, Vagnarelli P, Hoshi O, et al. (2001) Scc1/Rad21/Mcd1 is required for sister chromatid cohesion and kinetochore function in vertebrate cells. Developmental Cell 1: 759-770.

39. Sehl ME, Langer LR, Papp JC, Kwan L, Seldon JL, et al. (2009) Associations between single nucleotide polymorphisms in double-stranded DNA repair pathway genes and familial breast cancer. Clinical Cancer Research : An Official Journal of the American Association for Cancer Research 15: 2192-2203.

40. Esteller M, Garcia-Foncillas J, Andion E, Goodman SN, Hidalgo OF, et al. (2000) Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. The New England Journal of Medicine 343: 1350-1354.

41. Han J, Tranah GJ, Hankinson SE, Samson LD, Hunter DJ. (2006) Polymorphisms in O6-methylguanine DNA methyltransferase and breast cancer risk. Pharmacogenetics and Genomics 16: 469-474.

42. Gabriel S, Ziaugra L, Tabbaa D. (2009) SNP genotyping using the sequenomMassARRAYiPLEX platform. Current Protocols in Human Genetics / Editorial Board, Jonathan L.Haines ...[Et Al.] Chapter 2: Unit 2.12.

4. Assessing SNP-SNP interactions among DNA repair, modification and metabolism related pathway genes in breast cancer susceptibility[73]

4.1 Introduction

Breast cancer is a multifactorial disease, which results from combined effects of genetic, reproductive, environmental, and lifestyle risk factors. Linkage and twin studies reveal familial clustering of breast cancer, giving an approximately two-fold higher risk for first-degree relatives with family history [1,2]. Although some familial clustering is explained by germline mutations in high or moderate penetrance genes such as *BRCA1*[74][3], *BRCA2*[75][4], *ATM*[76][5], *PTEN*[77][6], *TP53*[78][7], *BRIP1*[79][8], *PALB2*[80][9] and *CHEK2*[81][10], such mutations are rare in the general population [11-13]. Hence, a polygenic model has been proposed to explain the bulk of genetic susceptibility in sporadic and non-*BRCA* breast cancers [13]. Under this model, a combination of multiple low penetrance loci/genes across the genome would contribute to overall genetic risk.

Several genome-wide association studies (GWASs) identified multiple single nucleotide polymorphisms (SNPs) statistically significantly associated with breast cancer susceptibility [12,14-21], supporting the polygenic model. However, these

---

[73] A version of this chapter has been accepted for publication. *Sapkota et al., 2013. PLoS ONE.* I would like to thank Conrado Franco-Villalobos for helping with the logic regression analysis presented in this chapter.
[74] Breast cancer 1, early onset.
[75] Breast cancer 2, early onset.
[76]Ataxia telangiectasia mutated.
[77]Phosphatase and tensin homolog.
[78] Tumor protein 53.
[79]BRCA1 interacting protein C-terminal helicase 1.
[80]Partner and localizer of *BRCA2*.
[81]Checkpoint kinase 2.

low penetrance variants, together with known predisposition genes (*e.g.*, *BRCA1* and *BRCA2*), explain only a small proportion of the total genetic risk of breast cancer [16], suggesting that more variants exist. Identifying additional low penetrance variants is difficult because the effect size is expected to be smaller than the GWAS variants reported thus far, requiring large sample sizes. Collaborative efforts are now underway from international consortia to profile additional low penetrance variants. Current GWAS approaches largely rely on single-locus effects of SNPs with the disease of interest, studied one SNP at a time, while ignoring potential SNP-SNP interactions at two or more loci (*i.e.*, epistatic effects) [22,23]. Epistasis is a ubiquitous phenomenon that describes how genes/loci interact to affect phenotypes. Such interactions are assumed to contribute to breast cancer. In search of the putative genes or SNPs contributing to epistasis, I reasoned that a study design exclusively addressing the value of GWAS or candidate gene SNPs with single-locus effects with weak statistical significance (hereafter referred to as "weak single-locus effects") but acting within a common biological pathway would provide mechanistic support for such a premise, which otherwise might be overlooked in less constrained genetic association studies. There is support for the premise that SNPs with weak single-locus effects are indeed of value to explore for epistatic effects, which in turn may contribute to a substantial proportion of the overall heritable risk [24,25]. While GWAS approaches are still crucial to initially scan genomes to identify variants with appreciable single-locus effects, further analyses capturing the combined effects of two or more SNPs with weak but reproducible single-locus effects in

independent stages/studies may shed light on the unexplained heritability of breast cancer.

In Chapter 3, I conducted a two-stage association study using SNPs selected from GWAS for sporadic breast cancer (recently published [26]). The SNPs selected were located in or close to DNA repair, modification and metabolism pathway related genes and showed weak single-locus effects for breast cancer. In a combined sample size of 1,480 breast cancer cases and 1,635 healthy controls from two independent stages, I observed six SNPs (located on chromosomes 8, 10, 15 and 18) showing weak but consistently reproducible single-locus effects for breast cancer susceptibility (per allele odds ratio (OR) ranged 0.85-0.86 for three protective SNPs and 1.13-1.20 for three risk elevating SNPs). I hypothesized that these variants may be optimal candidates to investigate potential SNP-SNP interactions at two or more loci contributing to breast cancer etiology. Furthermore, I also investigated the single-locus effects of SNPs considered in this study to examine their reproducibility in an independent study population, while adjusting for body mass index (BMI), a known risk factor for breast cancer.

To enable a more comprehensive evaluation of epistatic interactions among SNPs, I also considered additional SNPs from cancer related DNA repair genes, with prior evidence of their weak single-locus effects for breast cancer [27-35]. Genetic variations in DNA repair genes are extensively studied in the context of breast cancer since inter-individual variations in DNA repair capacity are thought to contribute to heritable component of breast cancer [11,36]. Despite large efforts by investigators/consortia, DNA repair genes/loci identified from GWASs that

contribute to breast cancer susceptibility are limited. This further strengthens the premise that DNA repair related SNPs may potentially contribute through the epistatic mechanisms. The bulk of the literature from biochemical characterizations of DNA repair proteins indicate that these gene products are involved in protein-protein and DNA-protein interactions to repair damage to DNA by carcinogens and radiation induced effects. To my knowledge, this is the first study attempting to assess potential SNP-SNP interactions at two or more loci implicated in breast cancer susceptibility, using systematically selected SNPs based on functional criteria from both GWAS and candidate gene approaches.

4.2 Materials and methods

4.2.1 Study participants

Breast cancer cases (n=2,795) used in this study were accessed from the provincial tumor bank located at the Cross Cancer Institute, Edmonton, Alberta, Canada (http://www.abtumorbank.com/), and the description of these has been presented in detail elsewhere [19,26]. This tumor bank contains well-annotated clinicopathological characteristics of the samples stored. The breast cancer cases included in this study had a pathologically confirmed diagnosis of invasive breast cancer predominantly characterized by late onset of disease (*i.e.*, median age and range at diagnosis=54 and 21-92 years, respectively, with >92% of the cases aged 40+ years at the time of diagnosis). The median BMI of breast cancer cases at the time of diagnosis was 27.4 and range 15.6-80.4. Healthy controls (n=4,505) were accessed from the Tomorrow Project (http://in4tomorrow.ca/) [19,26], Edmonton, Alberta, Canada, which aims to capture lifestyle factors and DNA of

approximately 50,000 healthy Albertans enrolled in the prospective cohort study. The median age and range at blood draw were 54 and 34-78 years, respectively, with >92% of the controls aged 40+ years at the time of blood draw. The median BMI of healthy controls at the time of enrollment in the study was 25.5 and range 10.4-60.4. The breast cancer cases and controls were predominantly of Caucasian origin based on their self-declared ethnicity and the overall demographics of the region. All participants provided informed consent and the study was approved by the Alberta Cancer Research Ethics Committee, Alberta, Canada.

4.2.2 SNPs and samples considered

A total of 17 candidate SNPs located in or close to 14 DNA repair, modification and metabolism pathway related genes (*RAD21*, *MGMT*, *RPAP1*, *MBD2*, *PARP1*, *MLH1*, *MSH3*, *ERCC6*, *MDM2*, *BRCA2*, *ERCC5*, *APEX1*, *XRCC3* and *XRCC1*) were considered (**Text S4-1 and Tables S4-1 and S4-2**). Of these, six SNPs (8q24.11-rs13250873, 10q26.3-rs1556459, *RPAP1*[82]-rs2297381, *MBD2*[83]-rs7614, *MBD2*-rs4041245 and *MBD2*-rs8094493) were selected from GWAS and previously replicated in an independent set of breast cancer cases and healthy controls [19,26]. These SNPs were genotyped as part of a stage 3 study in additional breast cancer cases (n=1,315) and healthy controls (n=2,861) and were evaluated for their single-locus effects for breast cancer (**Text S4-1** and **Table S4-1**). Overall, I present my findings from a combined sample size of 2,795 breast cancer cases and 4,663 controls from all three stages to meet

[82]RNA polymerase II associated protein 1.
[83]methyl-CpG binding domain protein 2.

the statistical rigor. The remaining 11 candidate DNA repair SNPs (*PARP1*[84]-rs1136410, *MLH1*[85]-rs1799977, *MSH3*[86]-rs184967, *MSH3*-rs26279, *ERCC6*[87]-rs2228528, *MDM2*[88]-rs769412, *BRCA2*[89]-rs1799943, *ERCC5*[90]-rs17655, *APEX1*[91]-rs1130409, *XRCC3*[92]-rs1799796 and *XRCC1*[93]-rs25487) were selected based on published DNA repair gene polymorphisms and their association with breast cancer susceptibility [27-35], the pilot study by the Damaraju Laboratory screening for more than 100 SNPs from 59 genes showing high minor allele frequency, concordance of genotypes to Hardy-Weinberg Equilibrium (HWE) in controls, statistical significance for the association in overall case-control analysis or promising associations (allelic and/or genotypic) for subtypes of breast cancer addressing the inherent heterogeneity, and high SNP call rates (data not shown). These 11 SNPs were genotyped in 2,720 breast cancer cases and 4,505 controls and were evaluated for their single-locus effects for breast cancer. To evaluate SNP-SNP interactions, I used genotype data of the 17 SNPs represented in a common set of breast cancer cases (n=2,718) and healthy controls (n=4,496). The finite discrepancies between the numbers of samples used for genotyping of the profiled SNPs and those used for SNP-SNP interactions were expected due to multiplexing assays for SNPs and the panels designed for the genotyping

[84]poly (ADP-ribose) polymerase 1.
[85]mutL homolog 1, colon cancer, nonpolyposis type 2 (*E. coli*).
[86]mutS homolog 3 (*E. coli*).
[87]excision repair cross-complementing rodent repair deficiency, complementation group 6.
[88]Mdm2, p53 E3 ubiquitin protein ligase homolog (mouse).
[89] Breast cancer 2, early onset.
[90]excision repair cross-complementing rodent repair deficiency, complementation group 5.
[91]APEX nuclease (multifunctional DNA repair enzyme) 1.
[92] X-ray repair complementing defective repair in Chinese hamster cells 3.
[93] X-ray repair complementing defective repair in Chinese hamster cells 1.

experiments on the Sequenom iPLEX Gold platform. An overview of the study design is presented in **Figure 4-1**.

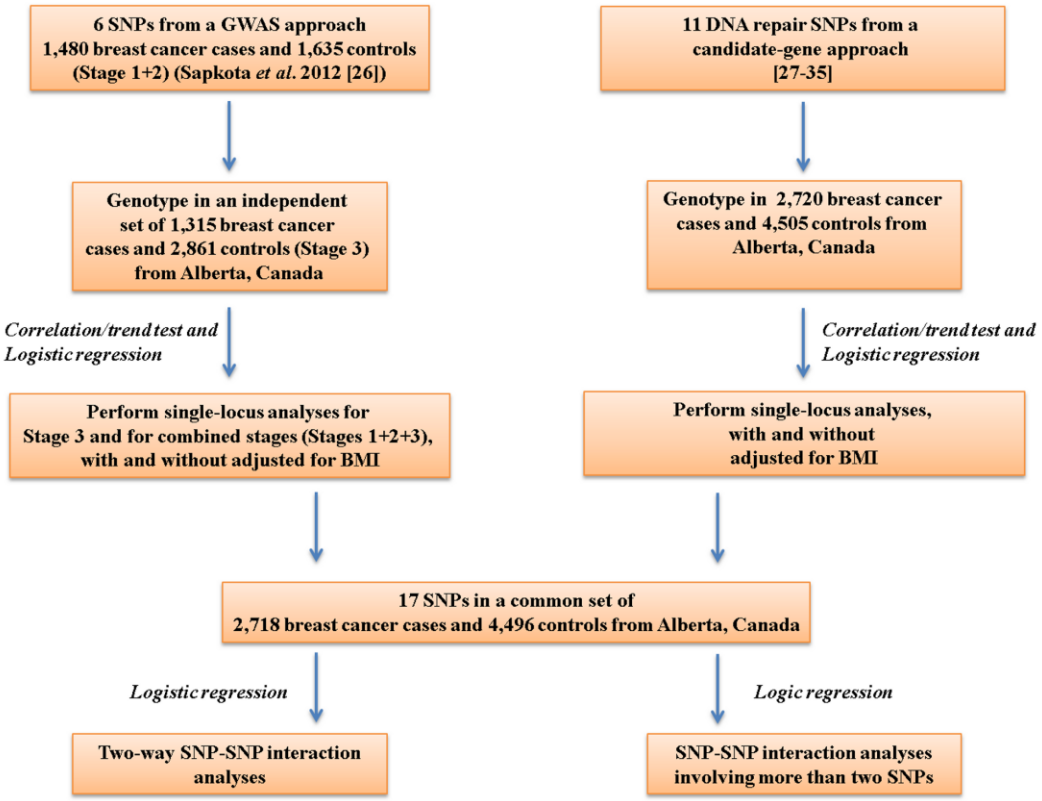


Figure 4-1 An overview of the study design.

4.2.3 Text S4-1 Methodology and pertinent discussion for single-locus association analyses of the 17 SNPs considered in the current study

I initially investigated the single-locus effects of the 17 SNPs considered for potential epistatic effects. Both allelic and genotypic single-locus effects of SNPs for breast cancer were determined. Associations of SNPs with breast cancer susceptibility were evaluated with correlation/trend tests with one degree of freedom (d.f.). The magnitude of allelic and genotypic effects of the six putative breast cancer susceptibility SNPs (previous work from Chapter 3) [26] were estimated using unconditional logistic regression and reported as odds ratios

(ORs) and corresponding 95% confidence intervals (CIs). Cases and controls from all three independent stages were pooled together and combined analysis was conducted. BMI was included as covariate in the logistic regression models: I, therefore, report BMI-adjusted ORs, 95% CI and *P* values of the six susceptibility SNPs and the additional 11 DNA repair SNPs.

Corrections for multiple comparisons were performed by conventional *P*=0.05/number of single-locus tests. Correlation/trend tests were performed using SNP and Variation Suite v7.6.11 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) [45]. The observed and adjusted allelic and genotypic ORs and 95% CI and adjusted *P* values were estimated using logistic models available in PLINK [37]. R.2.15.1 was used for general statistical analyses.

The results from the single-locus tests and combined analyses across three stages of the study are briefly summarized - In stage 3, associations of the six SNPs (previous work Chapter 3) [26] showed consistency in terms of the magnitude and direction of associations in both allelic and genotypic tests but did not show statistical significance at $P<0.05$ (**Table S4-1**). However, in combined analysis, all SNPs demonstrated weak single-locus effects for breast cancer at $P<0.05$, and were independent of BMI. The three *MBD2* SNPs showed more statistically significant associations than the rest and were significant even after correction for multiple comparisons (combined unadjusted $P<3.4 \times 10^{-3}$ and adjusted $P<1.8 \times 10^{-2}$) (**Table S4-1**). Four DNA repair SNPs (*MLH1*-rs1799977, *MDM2*-rs769412, *BRCA2*-rs1799943 and *XRCC1*-rs25487) showed significant associations with breast cancer susceptibility at $P<0.05$ (**Table S4-2**). Of these,

MLH1-rs1799977 and *MDM2*-rs769412 conferred reduced risk of breast cancer with allelic and genotypic ORs ranged from 0.79 to 0.94 while *BRCA2*-rs1799943 and *XRCC1*-rs25487 indicated risk-elevating effects with allelic and genotypic ORs ranged from 1.08 to 1.32. These weak single-locus effects were independent of BMI. *BRCA2*-rs1799943 showed a more statistically significant association and was significant even after correction for multiple comparisons, with per-allele ORs and 95% CI= 1.15 [1.06-1.25], $OR_{heterozygote}$ and 95% CI= 1.15 [1.03-1.28], $OR_{homozygote}$ and 95% CI= 1.32 [1.08-1.61] and P=9.8 x 10^{-4}, adjusted for BMI.

Table S4-1 Associations of the six putative breast cancer susceptibility loci in stage 3 and in combined stages.

SNP	Genes/ loci	Cytoband	Location (bp)*	MA	Stage#	MAF	Call rate	HWE $P_{controls}$	Cases/ controls	$P_{corr/trend}$**
rs13250873	*RAD21*/ Intergenic	8q24.11	117,806,169	G	3	0.33	1.00	1.09E-04	1315/2861	6.49E-01
					combined (1+2+3)	0.33	1.00	8.26E-05	2795/4496	4.67E-02
rs1556459	*MGMT*/ Intergenic	10q26.3	130,810,711	C	3	0.15	1.00	6.16E-01	1315/2861	7.20E-01
					combined (1+2+3)	0.15	1.00	6.83E-01	2795/4496	3.44E-02
rs2297381	*RPAP1*	15q15.1	41,827,655	G	3	0.49	1.00	4.45E-01	1315/2861	3.78E-01
					combined (1+2+3)	0.48	1.00	3.63E-01	2795/4496	3.04E-02
rs7614	*MBD2*	18q21.2	51,681,244	C	3	0.43	1.00	6.57E-01	1315/2861	3.46E-01
					combined (1+2+3)	0.42	1.00	6.76E-01	2795/4496	2.70E-03
rs4041245	*MBD2*	18q21.2	51,685,525	G	3	0.43	1.00	4.82E-01	1315/2861	3.33E-01
					combined (1+2+3)	0.42	1.00	5.51E-01	2795/4496	1.73E-03
rs8094493	*MBD2*	18q21.2	51,700,391	G	3	0.43	1.00	4.37E-01	1315/2861	3.36E-01
					combined (1+2+3)	0.42	1.00	6.33E-01	2795/4496	1.48E-03

*National Center for Biotechnology Information genome build 37

MA, minor allele; MAF, minor allele frequency

***P* values obtained from correlation/trend test with one d.f.

***Adjusted for BMI

#Association analyses of Stages 1 and 2 are reported in Sapkota et al., 2012. Stage 3 and the combined stages were conducted in this study.

Table S4-1 Continued..

Adjusted***			
$OR_{per\text{-}allele}$ [95% CI]	$OR_{heterozygote}$ [95% CI]	$OR_{homozygote}$ [95% CI]	*P*
1.03 [0.93-1.15]	1.08 [0.93-1.26]	1.02 [0.80-1.29]	
1.06 [0.98-1.14]	1.07 [0.96-1.20]	1.10 [0.93-1.30]	1.63E-01
1.04 [0.90-1.20]	1.05 [0.89-1.24]	1.04 [0.64-1.67]	
1.11 [1.01-1.23]	1.09 [0.97-1.23]	1.33 [0.95-1.85]	3.84E-02
1.02 [0.92-1.13]	1.07 [0.89-1.27]	1.05 [0.85-1.29]	
1.07 [0.99-1.15]	1.05 [0.93-1.19]	1.14 [0.99-1.32]	7.99E-02
0.98 [0.88-1.08]	0.97 [0.82-1.14]	0.96 [0.78-1.18]	
0.91 [0.84-0.98]	0.85 [0.75-0.95]	0.84 [0.72-0.98]	8.69E-03
0.98 [0.89-1.09]	0.94 [0.80-1.11]	0.98 [0.79-1.20]	
0.91 [0.85-0.98]	0.85 [0.75-0.95]	0.86 [0.74-1.00]	1.60E-02
0.98 [0.88-1.08]	0.96 [0.82-1.14]	0.96 [0.78-1.18]	
0.90 [0.84-0.97]	0.85 [0.75-0.95]	0.84 [0.72-0.97]	7.66E-03

Table S4-2 Eleven Candidate DNA repair SNPs and their associations with breast cancer susceptibility in 2,720 breast cancer cases and 4,505 healthy controls.

SNP	Genes/ loci	Cytoband	Location (bp)*	MA	MAF	Call rate	HWE $P_{controls}$
rs1136410	*PARP1*	1q42.12	226,555,302	C	0.17	1.00	1.00E+00
rs1799977	*MLH1*	3p22.2	37,053,568	G	0.31	1.00	1.00E+00
rs184967	*MSH3*	5q14.1	80,149,981	A	0.15	1.00	6.95E-01
rs26279	*MSH3*	5q14.1	80,168,937	G	0.28	1.00	2.47E-01
rs2228528	*ERCC6*	10q11.23	50,732,280	A	0.17	1.00	8.80E-01
rs769412	*MDM2*	12q15	69,233,215	G	0.06	1.00	1.00E+00
rs1799943	*BRCA2*	13q13.1	32,890,572	A	0.27	0.99	9.39E-01
rs17655	*ERCC5*	13q33.1	103,528,002	G	0.23	1.00	8.99E-01
rs1130409	*APEX1*	14q11.2	20,925,154	G	0.48	0.99	9.76E-01
rs1799796	*XRCC3*	14q32.33	104,165,927	G	0.33	1.00	9.20E-01
rs25487	*XRCC1*	19q13.31	44,055,726	A	0.35	1.00	4.97E-01

*National Center for Biotechnology Information genome build 37
MA, minor allele; MAF, minor allele frequency
**P values obtained from correlation/trend test with one d.f.
***Adjusted for BMI

Table S4-2 Continued..

$OR_{per\text{-}allele}$ [95% CI]***	$OR_{heterozygote}$ [95% CI]***	$OR_{homozygote}$ [95% CI]***	$P_{corr/trend}$ **	Adjusted P***
1.06 [0.96-1.17]	1.01 [0.90-1.14]	1.33 [0.99-1.78]	1.36E-01	2.34E-01
0.94 [0.87-1.01]	0.89 [0.80-0.99]	0.93 [0.78-1.12]	3.40E-02	1.06E-01
0.97 [0.88-1.08]	0.99 [0.87-1.11]	0.91 [0.64-1.28]	7.38E-01	6.26E-01
1.03 [0.95-1.12]	1.01 [0.90-1.12]	1.10 [0.90-1.34]	4.79E-01	4.90E-01
0.96 [0.87-1.06]	0.95 [0.85-1.07]	0.96 [0.71-1.29]	3.19E-01	4.22E-01
0.86 [0.74-1.01]	0.86 [0.73-1.02]	0.76 [0.32-1.83]	2.51E-02	6.21E-02
1.15 [1.06-1.25]	1.15 [1.03-1.29]	1.31 [1.07-1.61]	3.22E-04	9.32E-04
1.03 [0.94-1.12]	1.01 [0.91-1.13]	1.10 [0.87-1.40]	3.00E-01	5.40E-01
1.01 [0.94-1.09]	1.02 [0.90-1.15]	1.02 [0.88-1.18]	8.84E-01	8.01E-01
1.05 [0.97-1.13]	1.05 [0.94-1.17]	1.10 [0.92-1.31]	1.41E-01	2.59E-01
1.07 [0.99-1.16]	1.12 [1.00-1.25]	1.11 [0.94-1.31]	1.99E-02	7.88E-02

4.2.4 SNP genotyping and quality control

Genotyping assays of the 17 SNPs were designed and performed on the Sequenom iPLEX Gold platform (San Diego, CA, USA) using services from the McGill University and Genome Quebec Innovation Center, Montreal, Canada. Genotype concordance among SNPs was assessed using 66 duplicate samples (8 cases and 58 controls). Thresholds for SNP call rates of >99% and HWE $P>10^{-6}$ in controls were adopted.

4.2.5 Statistical considerations

I evaluated potential interactions among the select 17 candidate SNPs at two loci using logistic regression and multiple loci using logic regression. Logistic regression models with command '*-epistasis*' in PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/) [37] were used to assess two-way interactions and reported as ORs, 95% confidence intervals (CIs) and *P* values associated with the b3 coefficient of the following model:

$$\text{logit}(E[Y])= b0 + b1A + b2B + b3A\times B,\ Y \sim \text{Bernoulli}(E[Y])$$

where b3 captures the two-way interaction between SNP A and SNP B. To correct for multiple comparisons, I calculated the Benjamini-Hochberg False Discovery Rate (FDR) [38].

Logic regression is a method to assess SNP-SNP interaction among multiple loci, and it has been successfully applied to a GWAS SNP data recently [39], in addition to a candidate-gene approach [40]. Logic regression searches for a set of predictors that are Boolean combinations of binary SNP covariates using intersection ("AND") and union ("OR") operations. To explore potential multi-

way SNP-SNP interactions among the 17 SNPs considered in this study, a logic regression model was fitted using the *LogicReg* package [41] available in R 2.15.1 [42]. 157 (2.1%) subjects were excluded due to missing genotype as the *LogicReg* does not allow missing data. Since SNPs can have three possible genotypes (*e.g.*, AA, AB, BB), the 17 SNPs were first recoded into two sets of binary covariates by using both dominant (*e.g.,* AA=1, AB=1, BB=0) and recessive (*e.g.,* AA=0, AB=0, BB=1) and fitted the logic regression of the following form:

$$\text{logit}(E[Y]) = b0 + b1L1 + b2L2 + \ldots + bnLn,\ Y \sim \text{Bernoulli}(E[Y])$$

where L_i is a Boolean combination of the binary SNP covariates such as [(*SNP A=AA OR SNP B=AA) AND SNP C=AB or BB*], also known as a logic tree. A score function (deviance of the model) was then used to evaluate models with the number of trees, *n*, in the range of [2, 5] and the total number of SNPs in the range of [2, 17] using a 10-fold cross validation approach to determine the optimal tree/SNPs size. The statistical significance of a final model was evaluated with the optimal tree/SNPs size using a permutation test with 10,000 permutations of the case control labels. All statistical tests were two-sided.

4.3 Results

Genotyping assays for each of the 17 SNPs were successful with a SNP call rate of >99% and the SNPs also passed HWE ($P>10^{-6}$) in controls (**Tables S4-1 and S4-2**). Average genotype concordance was 100% for the 17 SNPs. Single-locus association tests in independent stages or in combined stage, adjusted for BMI were also profiled. Overall, SNPs considered in this study conferred weak single-locus effects for breast cancer, as I expected. I also analyzed the SNP-

breast cancer associations by removing subjects from cases and controls with extreme ages (<35 yrs. and >80 yrs.) and BMI (<18.5 and >40). The associations did not change materially, suggesting that the small fraction (~6.4% or 468 subjects) of extreme subjects may not have modified the observed overall SNP-breast cancer associations, data not shown.

4.3.1 Two-way SNP-SNP interactions

Logistic models were used to assess all SNP pairs among 17 candidate SNPs. Of these, two SNP pairs (*APEX1*-rs1130409 * *RPAP1*-rs2297381 and *MLH1*-rs1799977 * *MDM2*-rs769412) showed the strongest statistical association with breast cancer ($P<7.3 \times 10^{-3}$), with modest FDR values of 0.30 and 0.49, respectively (**Table 4-1**). Both SNP pairs showed increased risks towards breast cancer with ORs and 95% CIs of 1.16 [1.06-1.28] and 1.33 [1.08-1.64], respectively. The observed risks were similar for cases with luminal A tumors (~70% of the total cases), while the interactions were not statistically significant when analyses were restricted to cases with luminal B, HER2+ and triple negative tumors, data not shown.

Table 4-1 Two-way interactions identified among DNA repair pathway related SNPs.

SNP1 x SNP2	ORs [95% CI]	*P* *	FDR**
APEX1 -rs1130409 * *RPAP1* -rs2297381	1.16 [1.06-1.28]	2.25E-03	0.3
MLH1 -rs1799977 * *MDM2* -rs769412	1.33 [1.08-1.64]	7.31E-03	0.49

*P values obtained from unconditional logistic regression models

**Corrected for multiple comparisons using Benjamini-Hochberg False Discovery Rate method

4.3.2 SNP-SNP interactions involving multiple SNPs using Logic regression

Logic regression including the 17 SNPs identified a logic structure representing a SNP-SNP interaction involving four SNPs and was statistically significant (P=2.4 x 10^{-3}) (**Table 4-2**). The logic structure contained two logic trees, one with three SNPs and another with one SNP. The first logic tree consisted of an intersection of a union of *MBD2*-rs4041245 and *MLH1*-rs1799977 and *MDM2*-rs769412 while the second logic tree contained *BRCA2*-rs1799943. These logic trees formed four logic-based risk groups; a reference group (OR=1.00) and two low risk groups, with ORs 0.79 and 0.90, respectively and a high risk group with OR 1.18. The observed logic structure was tested in subgroups of tumors. It was statistically significant for the subgroup of cases with luminal A tumors (P=3.3 x 10^{-3}), while it was not in other subgroups (luminal B, HER2+ and triple negatives tumors, data not shown).

Table 4-2 Multi-way SNP-SNP interactions identified by logic regression.

		rs4041245 AA	rs1799977 AA	rs769412 AA	rs1799943 AA/AG	Logic-based Risk Groups		
Genotype Frequency	Cases N=2662	952 (35.77%)	1329 (49.92%)	2373 (89.14%)	1306 (49.06%)			
	Controls N=4395	1405 (31.97%)	2071 (47.12%)	3845 (87.49%)	1990 (45.28%)			
Logic 1		(OR) AND				Frequency		Odds Ratio
Logic 2						Cases	Controls	
Logic-based Risk Groups		Logic 1 = No			Logic 2 = No	527	1076	0.79
		Logic 1 = Yes			Logic 2 = No	829	1329	1.00
		Logic 1 = No			Logic 2 = Yes	500	895	0.90
		Logic 1 = Yes			Logic 2 = Yes	806	1095	1.18

4.4 Discussion

In this study of more than seven thousand women, I evaluated the contribution of epistasis to breast cancer susceptibility among 17 SNPs located in

or close proximity to 14 DNA repair, modification and metabolism pathway related genes. I identified two SNP pairs and interactions involving four SNPs among seven candidate SNPs located in seven genes. Except for *APEX1*-rs1130409, the SNPs participating in SNP-SNP interactions also showed weak single-locus effects (both allelic and genotypic) for breast cancer, independent of BMI (**Text S4-1 and Tables S4-1 and S4-2**). Of these, *BRCA2*-rs1799943 showed the strongest single-locus effects. Overall, my findings support the notion that SNPs with reproducible weak single-locus effects are useful candidates for studying their potential epistatic effects contributing to breast cancer susceptibility.

I identified two SNP pairs that demonstrated significant interactive effects on breast cancer risk and carried modest FDR values. Of these, one was *MBD2* SNP I reported earlier (Chapter 3) and the other three were from the candidate DNA repair SNPs considered in this study. The first pair consisted of *APEX1*-rs1130409 and *RPAP1*-rs2297381, with an OR of their interaction as 1.16, which was greater than their individual single-locus effects of 1.00 and 1.07, respectively. Similarly, another pair included *MLH1*-rs1799977 and *MDM2*-rs769412, with an OR of their interaction as 1.35 conferring risk. Interestingly, their individual single-locus effects were in opposite direction with ORs of 0.94 and 0.86, respectively and deserve further independent replication of findings.

Using a logic regression model, I also detected SNP-SNP interactions involving four (*MBD2*-rs4041245, *MLH1*-rs1799977, *MDM2*-rs769412 and *BRCA2*-rs1799943). Interestingly, one of the SNPs I entered in this analysis and

was predicted to participate in epistatic effects from a previous study [26] was also identified to partner with three other SNPs I profiled from the DNA repair genes considered in this study. Except for *MLH1*-rs1799977 and *MDM2*-rs769412, this model captured distinct set of SNPs from the ones profiled in the two-way epistatic interactions, suggesting a possible convergence of multiple DNA repair pathways while conferring breast cancer risk. Future independent studies through large international consortia are warranted to further evaluate the contributions of the observed SNP-SNP interactions to breast cancer predisposition and to understand the underlying important biology of breast cancer. I believe these findings reflect important biology, rather than simply statistical artifacts because of the unprecedented amount of literature indicating DNA-protein and protein interactions involved in DNA repair process.

I further investigated for possible biological insights in to the observed SNP-SNP interactions using a Cytoscape plugin, GENEMANIA [43]. For a given set of genes, GENEMANIA predicts their functional relationships, such as genetic and protein interactions, pathways, co-expressions, co-localization and similar protein domains from mining publicly available knowledgebase (*e.g.*, PubMed, BioGRID, Pathway Commons and Pfam). I observed that the SNP-SNP interactions I identified were also complimented by observed/predicted interactions among the proteins encoded by participating genes. Proteins encoded by *APEX1* and *RPAP1* genes were not in direct cross talk but were mediated by a third protein, cyclin O protein (CCNO). Similarly, protein-protein interactions between proteins encoded by *MLH1* and *MDM2* genes were predicted to be

mediated by cyclin G1 protein (CCNG1). It was noteworthy that CCNG1 was acting as a central molecule interacting with proteins encoded by the four genes involved in the two-way SNP-SNP interactions and the mediating *CCNO* gene. Further, protein-protein interactions facilitated by CCNG1 and GATA zinc finger domain containing 2A (GATAD2A) proteins were also predicted to mediate interactions among proteins encoded by *MDM2*, *MLH1*, *MBD2* and *BRCA2* genes. I was limited in my ability to draw any finer conclusions since the number of genes considered for the study does not represent a comprehensive view of all DNA repair/metabolism genes on the human genome. To-date, the total number of human DNA repair genes annotated is around 130 [44]. The summarized work here merely provides a previously unexplored rationale and may generate hypothesis to test under various experimental designs, both for genetic and biological relevance beyond the provided statistical paradigm. Since the GENEMANIA network analysis is based on experimentally determined functional relationships, it is reasonable to speculate that both the two-way and multi-way SNP-SNP interactions and the known biological relationships among the proteins encoded by corresponding genes suggest possible cross talk and convergence of DNA repair, modification and metabolism pathways contributing to breast cancer etiology; this is consistent with the polygenic nature of complex diseases. The effect sizes from the SNP-SNP interactions were consistent with the predicted polygenic models (small but finite effect sizes from diverse gene/loci) and findings from GWASs to-date (ORs<1.5). Since a majority of breast cancer risk is explained by the intersection of life style factors with genetic

predisposition, future studies may benefit by considering these additional risk factors to comprehensively account for the breast cancer risk in populations. However, caution should be exercised while interpreting the results from the interaction analyses until independent replication by other research groups could as well demonstrate the validity of statistical approaches to this emerging discipline of epistasis as a model to explain the additional missing heritable components of genetic risk.

In summary, I demonstrated both two-way and multi-way SNP-SNP interactions contributing to breast cancer risk, among candidate SNPs related to DNA repair, modification and metabolism pathway genes. The interactions were not previously reported and were mostly among the SNPs with weak but reproducible single-locus effects. My results suggest SNP-SNP interactions among SNPs with weak but reproducible single-locus effects in a typical multi-stage GWAS or candidate-gene studies may identify cross talk among members of multiple cancer-related pathways, and help account for the heritability for complex diseases.

4.5 References

1. Collaborative Group on Hormonal Factors in Breast Cancer. (2001) 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 358: 1389-1399.
2. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from sweden, denmark, and finland. The New England Journal of Medicine 343: 78-85.
3. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science (New York, N.Y.) 250: 1684-1689.
4. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, et al. (1995) Identification of the breast cancer susceptibility gene BRCA2. Nature 378: 789-792.
5. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, et al. (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature Genetics 38: 873-875.
6. Liaw D, Marsh DJ, Li J, Dahia PL, Wang SI, et al. (1997) Germline mutations of the PTEN gene in cowden disease, an inherited breast and thyroid cancer syndrome. Nature Genetics 16: 64-67.
7. Malkin D, Li FP, Strong LC, Fraumeni JF,Jr, Nelson CE, et al. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science (New York, N.Y.) 250: 1233-1238.

8. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, et al. (2006) Truncating mutations in the fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nature Genetics 38: 1239-1241.

9. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, et al. (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nature Genetics 39: 165-167.

10. CHEK2 Breast Cancer Case-Control Consortium. (2004) CHEK2*1100delC and susceptibility to breast cancer: A collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. American Journal of Human Genetics 74: 1175-1182.

11. Shen J, Desai M, Agrawal M, Kennedy DO, Senie RT, et al. (2006) Polymorphisms in nucleotide excision repair genes and DNA repair capacity phenotype in sisters discordant for breast cancer. Cancer Epidemiology, Biomarkers & Prevention 15: 1614-1619.

12. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087-1093.

13. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. (2008) Polygenes, risk prediction, and targeted prevention of breast cancer. The New England Journal of Medicine 358: 2796-2803.

14. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, et al. (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nature Genetics 41: 585-590.

15. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, et al. (2007) A common coding variant in CASP8 is associated with breast cancer risk. Nature Genetics 39: 352-358.

16. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, et al. (2012) Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nature Genetics 44: 312-318.

17. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, et al. (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proceedings of the National Academy of Sciences of the United States of America 105: 4340-4345.

18. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, et al. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature Genetics 39: 870-874.

19. Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass CE, et al. (2011) Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human Genetics 130: 529-537.

20. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, et al. (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nature Genetics 40: 703-706.

21. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, et al. (2009) A multistage genome-wide association study in breast cancer identifies two

new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nature Genetics 41: 579-584.

22. Moore J. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Human Heredity 56: 73-82.

23. Moore JH. (2005) A global view of epistasis. Nature Genetics 37.

24. Lo S, Chernoff H, Cong L, Ding Y, Zheng T. (2008) Discovering interactions among BRCA1 and other candidate genes associated with sporadic breast cancer. Proceedings of the National Academy of Sciences of the United States of America 105.

25. Onay VU, Briollais L, Knight JA, Shi E, Wang Y, et al. (2006) SNP-SNP interactions in breast cancer susceptibility. Bmc Cancer 6: 114.

26. Sapkota Y, Robson P, Lai R, Cass CE, Mackey JR, et al. (2012) A two-stage association study identifies methyl-CpG-binding domain protein 2 gene polymorphisms as candidates for breast cancer susceptibility. European Journal of Human Genetics 20.

27. Smith TR, Levine EA, Freimanis RI, Akman SA, Allen GO, et al. (2008) Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. Carcinogenesis 29.

28. Conde J, Silva SN, Azevedo AP, Teixeira V, Pina JE, et al. (2009) Association of common variants in mismatch repair genes and breast cancer susceptibility: A multigene study. Bmc Cancer 9: 344.

29. Boersma BJ, Howe TM, Goodman JE, Yfantis HG, Lee DH, et al. (2006) Association of breast cancer outcome with status of p53 and MDM2 SNP309. Journal of the National Cancer Institute 98.

30. Gochhait S, Bukhari SIA, Bairwa N, Vadhera S, Darvishi K, et al. (2007) Implication of BRCA2-26G > A 5 ' untranslated region polymorphism in susceptibility to sporadic breast cancer and its modulation by p53 codon 72 arg > pro polymorphism. Breast Cancer Research 9: R71.

31. Rajaraman P, Bhatti P, Doody MM, Simon SL, Weinstock RM, et al. (2008) Nucleotide excision repair polymorphisms may modify ionizing radiation-related breast cancer risk in US radiologic technologists. International Journal of Cancer 123.

32. Mitra AK, Singh N, Singh A, Garg VK, Agarwal A, et al. (2008) Association of polymorphisms in base excision repair genes with the risk of breast cancer: A case-control study in north indian women. Oncology Research 17.

33. Economopoulos KP, Sergentanis TN. (2010) XRCC3 Thr241Met polymorphism and breast cancer risk: A meta-analysis. Breast Cancer Research and Treatment 121.

34. Leng S, Bernauer A, Stidley CA, Picchi MA, Sheng X, et al. (2008) Association between common genetic variation in cockayne syndrome A and B genes and nucleotide excision repair capacity among smokers. Cancer Epidemiology Biomarkers & Prevention 17.

35. Roberts MR, Shields PG, Ambrosone CB, Nie J, Marian C, et al. (2011) Single-nucleotide polymorphisms in DNA repair genes and association with breast cancer risk in the web study. Carcinogenesis 32.

36. Rao NM, Pai SA, Shinde SR, Ghosh SN. (1998) Reduced DNA repair capacity in breast cancer patients and unaffected individuals from breast cancer families. Cancer Genetics and Cytogenetics 102: 65-73.

37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559-575.

38. Benjamini Y, Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289-300.

39. Dinu I, Mahasirimongkol S, Liu Q, Yanai H, Eldin NS, et al. (2012) SNP-SNP interactions discovered by logic regression explain crohn's disease genetics. Plos One 7: e43035.

40. Feng Q, Balasubramanian A, Hawes SE, Toure P, Sow PS, et al. (2005) Detection of hypermethylated genes in women with and without cervical neoplasia. Journal of the National Cancer Institute 97: 273-282.

41. Kooperberg C, Ruczinski I. (2012) LogicReg: Logic regression. R package version 1.5.3. http://CRAN.R-project.org/package=LogicReg.

42. R Core Team. (2012) R: A language and environment for statistical computing. R foundation for statistical computing, vienna, austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

43. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. (2010) The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research 38.

44. Wood RD, Mitchell M, Sgouros J, Lindahl T. (2001) Human DNA repair genes. Science (New York, N.Y.) 291: 1284-1289.

5. Germline DNA Copy Number Aberrations Identified as Potential Prognostic Factors for Breast Cancer Recurrence[94]

5.1 Introduction

Breast cancer is the most common epithelial malignancy among women in the developed world, with more than 200,000 new cases and 39,000 deaths estimated in the United States in 2012 [1]; comparable statistics were also observed in Canada in 2011 [2]. While age-adjusted breast cancer incidence has increased with the introduction of screening measures, there has been a steady decline in breast cancer mortality rates over the last two decades. During the years 1998-2008, cancer related death rates have decreased by more than 1% per year in North American women and breast cancer explains one-third of this total decline [1].

Advances in early diagnosis, increased public awareness and improved adjuvant treatment modalities have contributed to the improvements in prognosis of early-stage breast cancer. Standard guideline-based therapy for non-metastatic breast cancer typically includes surgical excision of localized tumor and involved lymph nodes, followed by adjuvant systemic and radiotherapies to eradicate any residual micro-metastatic deposits. Both systemic chemotherapy and adjuvant

---

[94] A version of this chapter has been published. *Sapkota et al., (2013). PLoS ONE* 8(1):e53580. This paper received wide media coverage both nationally and internationally since this is the first study in literature using germline DNA-based markers to demonstrate germline CN-LOHs as potential prognostic markers for breast cancer.
http://www.globalnews.ca/health/alberta+researchers+discover+dna+marker+that+could+predict+breast+cancer+recurrence/6442790053/story.html
http://www.dailymail.co.uk/health/article-2264672/Simple-blood-test-predicts-womans-breast-cancer-likely-return.html?ito=feeds-newsxml
http://www.medpagetoday.com/HematologyOncology/BreastCancer/36912

endocrine therapy have reduced breast cancer recurrence and death [3]. However, currently used adjuvant therapies have life-threatening and life-altering toxicities, and it therefore is of clinical importance to identify patients who would most benefit from aggressive adjuvant therapies, and to spare those patients unlikely to benefit from aggressive therapy. At present, the determination of those breast cancer patients who are most likely to benefit from adjuvant therapies is primarily guided by tumor-based prognostic factors such as axillary lymph nodal status, tumor size, tumor histologic grade, lymphatic and vascular invasion, proliferative markers, ER/progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) status [4,5]. However, clinicopathological characteristics of tumors remain imperfect prognostic classifiers, in part due to the molecular heterogeneity of breast cancer.

While genomic signatures derived from tumor transcriptome studies such as 21-gene and 70-gene profiles may provide some improvement in prognostic power when added to standard clinicopathologic prognosticators, there are still patients who experience recurrence who are categorized as having an excellent prognosis, and others who remain recurrence free who are categorized as having a very poor prognosis [6,7]. Furthermore, despite incremental improvement in breast cancer therapies, approximately 30% of the treated breast cancer patients (who are non-metastatic at the time of diagnosis) show disease recurrence within ten years [8,9]. Consequently, there remains continued need to identify improved prognostic and predictive markers with higher performance for clinical validation in prospective studies.

Recent studies show that germline DNA variations contribute to disease susceptibility [10-12], prognosis [13-15] and response to therapies [16,17]. The majority of these studies have adopted widely accepted multi-stage association study designs using single nucleotide polymorphisms (SNPs) from candidate genes/pathways or whole genome scans. As a result, thousands of SNPs have been identified that are significantly associated with susceptibility to breast cancer and its subtypes [12,18,19] and some of these are likely to predict overall disease survival [13-15]. In addition to SNPs, germline copy number variations (CNVs) are also found to be an important source of genetic predisposition to many complex phenotypes, including breast cancer [20-24]. CNVs are the most common type of genetic structural variations and by definition show gains or losses of DNA segments comprising more than one kb [21,22]. These DNA variations are believed to exert their affects through gene expression either through gene-dosage or *cis*-acting gene regulatory activities [25,26]. More recently with the application of high-throughput SNP-arrays, large chromosomal lesions characterized by loss of heterozygosity (LOH) but with diploid copy number were observed in many tumor types, possibly resulting from mitotic recombination [27-29]. These unique regions are referred to as copy neutral-loss of heterozygosities (CN-LOHs) or uniparental disomies (UPDs). Interestingly, large CN-LOH regions were also found in germline DNA and these genomic signatures may also be of value as potential markers for susceptibility and prognosis of complex diseases, such as cancers [30-33].

In the present study, I analyzed germline CNVs and CN-LOHs (hereafter referred to as copy number aberrations, CNAs) genotyped with Affymetrix Genome-Wide Human SNP Array 6.0 (Santa Clara, CA, USA) for their role as potential prognostic markers using 369 breast cancer patients from Alberta, Canada, treated with standard guideline-based therapies and followed over extended periods to capture the disease recurrence. I confirmed select CNAs identified from Affymetrix SNP 6.0 array data by independent technology platforms; TaqMan real-time quantitative polymerase chain reaction (RT-qPCR) (Carlsbad, CA, USA) for copy number determination and Sequenom iPLEX Gold Platform (San Diego, CA, USA) for assessing the fraction of heterozygosity in a subset of samples, using services from the McGill University, Genome Quebec Innovation Center, Montreal, Canada.

5.2 Materials and Methods

5.2.1 Patients

Breast cancer cases were accessed from the PolyomX and Canadian Breast Cancer Foundation (CBCF) Tumor Banks, located at the Cross Cancer Institute, Edmonton, Alberta, Canada [10,11]. The subject recruitment criteria and geographic populations of the PolyomX Tumor Bank and its successor, the CBCF Tumor Bank, (accrual during 2001-2005 and 2005-present, respectively) were the same. These tumor banks contain flash frozen tumor specimens, matching buffy coat samples (from over 2,000 subjects, diagnosed between the years 1987 to 2012) and clinicopathological information for breast and other cancers in the province of Alberta (http://www.abtumorbank.com/). In this study, I included 369

Caucasian women (median age=51 years) with a confirmed diagnosis of early-stage non-metastatic breast cancer predominantly characterized by late onset of disease and with the criteria identified below for case selection. Despite standard adjuvant therapy, 155 patients (median follow-up time from diagnosis=6.30 years; range=0.60-21.78 years) experienced recurrence and 214 did not, after a minimum duration of follow up of three years (median follow-up time from diagnosis=8.60 years; range=3.08-13.57 years). Of the 214 cases, follow-up time for (i) 32 (14.95%) was between three to five years, (ii) 40 (18.69%) was between five to seven years, (iii) 105 (49.06%) was between seven to ten years and (iv) 37 (17.29%) was more than ten years.

Of 369 individuals, 286 (77.50%) were ≥45 years old. Following diagnosis, these women received curative-intent primary treatments (surgical resection, chemotherapy with anthracyclines and/or taxanes, trastuzumab, hormonal therapy and radiotherapy) as per standardized provincial breast cancer care. A detailed description of clinicopathological characteristics of breast cancer patients is presented in **Table 5-1**, and the outcome data reflects database updates up to 21st February 2012. Written informed consents were obtained from all study participants and the study was approved by the Research Ethics Board of Alberta Health Services.

Breast cancer patients enrolled in the study were further classified into tumor subtypes based on immunohistochemistry score-based ER, PR and HER2 status of tumors as recorded in pathology reports. Using conventional guidelines commonly used in epidemiological studies [34], tumors were categorized as (i)

luminal A for ER^+ and/or PR^+ and $HER2^-$, (ii) luminal B for ER^+ and/or PR^+ and $HER2^+$, (iii) HER2 type for ER^-, PR^- and $HER2^+$, (iv) triple negative for ER^-, PR^- and $HER2^-$. There were 211 luminal A cases (170 with ER^+ and/or PR^+ and $HER2^-$ and 41 with both ER^+ and PR^+ and unknown HER2 status but characterized by low tumor grade). Among the remaining cases, there were 62 luminal B, 25 HER2 type, and 42 triple negative cases. There were 29 cases with unknown HER2 status and varying combinations of ER and PR (+ or – status) and tumor grades (high or low) that were, therefore, classified as others and excluded from the finer analyses based on stratification of the molecular subtypes of breast cancer. I adhered to the Recommendations for Tumor Marker Prognostic Studies (REMARK) [35] for the results reported, where applicable.

5.2.2 DNA extraction, whole genome genotyping and quality control

DNA was extracted from the buffy coat fractions using commercially available Qiagen™ (Mississauga, Ontario, Canada) DNA isolation kits. Buffy coat fractions collected were stored at -80 ºC until use. Following guidelines provided by manufacturer, whole genome genotyping was conducted using Affymetrix Genome-Wide Human SNP Array 6.0, which consisted of over 1.8 million probes (906,600 SNPs and 946,000 copy number probes) with an overall inter-marker distance of 680 bp. I used Affymetrix recommended contrast quality control (CQC), a measure of performance of genotyping experiments, to assess sample quality. All 369 samples used in this study showed CQC>2.0, a value greater than the default CQC threshold of ≥ 1.7.

5.2.3 Identification of CNAs

I used Nexus Copy Number 6.0 genomics software to process Affymetrix generated signal intensity or CEL files. A reference genome created using 270 HapMap samples was used as a baseline to calculate log2ratios and B-allele frequencies (BAF) in each sample followed by quantile normalization [36]. Probe to probe variance was calculated and reported as quality control (QC) scores to remove extreme outliers due to copy number break-points. I used a default setting for outlier removal, a combined value of 3% at the two extremes, 1.5% at each end. Using these normalized log2ratio and BAF values, CNAs were identified with the SNP-Fast Adaptive States Segmentation Technique 2 (SNP-FASST2) segmentation algorithm in conjunction with quadratic wave correction implemented in the Nexus software. The SNP-FASST2 segmentation algorithm is a Hidden Markov Model-based approach, which uses log2ratio values of ~1.8 million probes to make a CNV call while it considers both log2ratio and BAF values to detect LOHs. Significance threshold for segmentation was set at $P<5 \times 10^{-7}$ with minimum number of ten probes per segment and a maximum probe spacing of 1,000 kb. Single copy gains and losses were defined with log2ratio values of 0.2 and -0.2, respectively while two or more than two copies of gains and losses were defined by log2ratio values of 0.7 and -1.1, respectively. A chromosomal region was called a LOH if ≥95% of the SNP probes in a DNA segment of at least 500 kb exhibited BAF≥0.8 or ≤0.2 -- *i.e.*, ≥95% of the SNP probes in that region are homozygous probes (*e.g.*, AA or BB). Auto gender correction available in Nexus software was applied to call CNAs in X

chromosomes. LOHs with diploid copy number of two were considered as CN-LOHs or UPDs.

5.2.4 Quality control parameters for CNA calling

Pre-processing of CEL files was conducted using the settings described above. Six (three luminal A, two luminal B and one HER2 type tumors) out of 369 samples exhibited very high QC scores (>0.40) and were excluded from final analyses as higher QC scores suggest for elevated noise to signal ratio. Average QC score of remaining 363 samples (152 BCR and 211 non-BCR) was 0.17 (range: 0.08-0.32), acceptable values recommended by the Nexus Copy Number 6.0.

5.2.5 Survival analysis of CNAs and statistical considerations

Of the CNAs identified by the SNP-FASST2 segmentation algorithm, I restricted my analysis to relatively high frequency common CNAs to evaluate their potential role in breast cancer recurrence because common CNVs often harbour cancer-related genes [37]. I excluded LOHs due to copy number losses and more than two copy number gains from the analysis as these were already captured as copy number losses and copy number gains, respectively. I used a frequency cut-off of ≥10% in either group (BCR and non-BCR) or in both to select relatively common CNAs in the curent study population. When overlaps between CNAs selected in BCR and non-BCR groups were noted, I considered the intersecting common CNA regions present in both groups.

Univariate survival analyses showing relationships between select germline CNAs and recurrence-free survival (RFS) were performed using Kaplan-Meier

survival curves. RFS probabilities with and without CNAs in 363 samples were estimated using log-rank tests with one degree of freedom (d.f.). Correction for multiple hypotheses testing was carried out using the Benjamini-Hochberg False Discovery Rate correction method and represented as *Q* value [38]. Association of germline CNAs with BCR was determined with univariate Cox proportional hazards model and reported as hazard ratios (HRs) and corresponding 95% confidence intervals (CIs). Tumor stage and grade were then included as covariates in the Cox proportional hazards model to estimate the adjusted HRs and corresponding 95% CIs.

I also conducted subgroup survival analyses (log-rank tests with one d.f.) to identify additional common CNAs specific to luminal A subtype of breast cancer wherein I compared RFS probabilities with and without CNAs in 208 luminal A samples only. CNAs for association testing were selected using the approach mentioned above (*i.e.*, I focused on relatively common CNAs with ≥10% frequencies in at least one group or in both). Association analyses per se were carried out by fitting Cox proportional hazards models as explained earlier. Subgroup analyses restricted to luminal B, HER2 type and triple negative samples were not attempted due to limited sample size.

All statistical analyses were carried out, either singly or in combination using R 2.14.1 (R Development Core Team, 2011) and SAS software, version 9.3 of the SAS system for Windows. Copyright© 2002-2010 SAS Institute Inc. Cary, NC, USA.

5.2.6 Validation of candidate CNAs using independent genotyping platforms

Potential candidate CNAs were validated in a representative subset of samples. Using services from the McGill University, Genome Quebec Innovation Center, Montreal, Canada, I quantified the copy number of candidate CNAs using pre-designed TaqMan® copy number assays on a RT-qPCR instrument (Applied Biosystems, Foster City, CA, USA). Primers and probes targeted for individual copy number assays were from within the candidate CNA sequence boundaries identified in Nexus. 2 µL per assay of genomic DNA at a final concentration of 20ng/µL was used. All reactions were run in quadruplicates in MicroAmp® optical 96-well plates with barcode sealed with optical adhesive film. Thermal-cycling (7900HT) conditions were: 10 minutes at 95ºC followed by 40 cycles of 15 seconds at 95ºC and 60 seconds at 60ºC. Real-time data was exported to CopyCaller v2.0. RNaseP was used as a reference to calculate the ΔCt values for each sample. Copy numbers were determined using the comparative ΔΔCt cycle threshold method, assuming most frequent sample copy number of two. For CN-LOHs, SNPs (ten per CN-LOH) were also genotyped for same DNAs used in copy number assays using the Sequenom iPLEX Gold Platform to measure percentages of heterozygosity in CN-LOHs. Using HapMap release 24 Central Europeans genotype data, tagSNPs for CN-LOHs were selected with minor allele frequency (MAF) and pair-wise correlation (r^2) cut-offs of 10% and 0.8, respectively, to ensure the large CN-LOH region SNPs selected were non redundant. Whenever number of tagSNPs was less than ten, additional SNPs with ≥10% global MAF (1000 Genomes Project phase 1 population of 629 individuals)

from NCBI dbSNP build 136 were randomly selected ensuring that none of these additional SNPs was tagged by previously selected tagSNPs (see **Table S5-1** for probe selection and relevant assays).

Table S5-1 Details of TaqMan copy number assays and SNPs genotyped for validation.

CN-LOH	Product Type	Assay ID	[a]Context Sequence	Amplicon length (bp)	SNPs genotyped
17q11.2	TaqManÂ® Copy Number Assays, MED	Hs00138078_cn Hs02495547_cn	GCAATTTTGCCCTGTGTTATATGTG CTGCCTCAGGAACCAACTGAAAATT	73 73	rs1870900 rs28649357 rs28704877 rs28708703 rs548957 rs6505274 rs7217438
11p14.1	TaqManÂ® Copy Number Assays, MED	Hs06332650_cn	CCAATGTGAATTATGGAAGGCATCT	78	rs17703479 rs292436 rs294361 rs795237 rs963837 rs1823186 rs555409 rs7105801 rs2467604
11q13.1	TaqManÂ® Copy Number Assays, MED	Hs06324464_cn	TGCTGCTGTGTGCATCTCTGTGTGT	78	rs11231803 rs1529910 rs17299124 rs240698 rs581347 rs604237 rs633218 rs652922 rs2845638
6q24.1	TaqManÂ® Copy Number Assays, MED	Hs06809880_cn	GAGAGAAAAATATCCAATCACCCAT	74	rs9376516 rs9399314 rs57291519 rs11155111 rs12661752 rs13216392 rs17069935 rs2328060

[a]The 25-nucleotide sequence sorrounding the probe

5.3 Results

5.3.1 Patients' clinical characteristics

I identified 369 cases as meeting the criteria for the study of BCR and non-BCR, as described in the methods. I investigated if the clinical characteristics for study subjects (BCR and non-BCR cases) were different and how these might contribute to potential confounding effects. I did not find statistically significant differences for age at diagnosis, menopausal status and family history of breast cancer between BCR and non-BCR while molecular subtypes, tumor overall grade and stage were significantly different between BCR and non-BCR (**Table 5-1**). The identified potential confounders were taken into consideration for the data analysis and interpretations.

Table 5-1 Clinicopathological characteristics of 369 breast cancer cases enrolled in the study.

Characteristics	BCR (n=155)	non-BCR (n=214)	*P* value[a]
Median age at diagnosis (yrs.)	51	51.5	0.90
Follow-up time from diagnosis (days)[b]	2,317 (219-7948)	3,138 (1125-4954)	
Molecular subtypes			**0.01**
Luminal A	82	129	
Luminal B	25	37	
HER2 type	11	14	
Triple negative	28	14	
Other(s)	9	20	
Menopausal status			0.20
Pre	62	80	
Peri	19	17	
Post	73	117	
Unknown	1	0	
Family history of breast cancer			0.37
Yes	59	95	
No	91	115	
Unknown	5	4	
Overall grade			**8.2×10^{-4}**
Low	63	126	
High	89	87	
Unknown	3	1	
Stage			**0.01**
I	21	38	
II	106	159	
III	28	17	

[a] *P* values for Median age at diagnosis (yrs.) was calculated using Mann-Whitney test whereas 2×n Fisher's exact test was used for Molecular subtypes, Menopausal staus, Family history of breast cancer, Overall grade and Stage

[b] Median is presented with range shown in the parentheses

P value<0.05 is indicated in bold

5.3.2 Summary of CNAs identified

SNP-FASST2 algorithm identified 19,591 CNAs (516 copy number gains, 869 copy number losses and 18,206 CN-LOHs) in 363 samples (**Table S5-2**). Of these, 18,561 CNAs (475 copy number gains, 773 copy number losses and 17,313 CN-LOHs) were of ≥one kb (**Figure 5-1**). Majority of copy number gains (n=465), copy number losses (n=746) and CN-LOH (n=15,682) were in chromosomes 1 to 22 while very few events (10 copy number gains, 27 copy number losses and 1,631 CN-LOHs) were observed in X-chromosomes. A total of 7,450CNAs were of >1kb-10kb, 9,523 CNAs were of >10kb-100kb and 1,588 CNAs were very large regions (>100kb-5Mb).

Table S5-2 A total of 19,591 CNVs and CN-LOHs identified in 363 samples.

chrX:99,255,055-99,258,822	3,767	q22.1	CN-LOH	0	0
chrX:99,298,476-99,319,493	21,017	q22.1	CN-LOH	0	0
chrX:99,341,414-99,370,268	28,854	q22.1	CN-LOH	0	0
chrX:99,590,744-99,607,799	17,055	q22.1	CN-LOH	1	0
chrX:99,656,110-99,664,689	8,579	q22.1	CN-LOH	1	0
chrX:99,689,236-99,693,366	4,130	q22.1	CN-LOH	0	0
chrX:99,747,680-99,765,078	17,398	q22.1	CN-LOH	0	0
chrX:99,839,586-99,847,707	8,121	q22.1	CN-LOH	1	0

Note: This is an Excel worksheet with more than 19,500 rows. Alternative on-line source for the complete data can be found at *Sapkota et al., 2013, PLoS ONE, Table S2*.

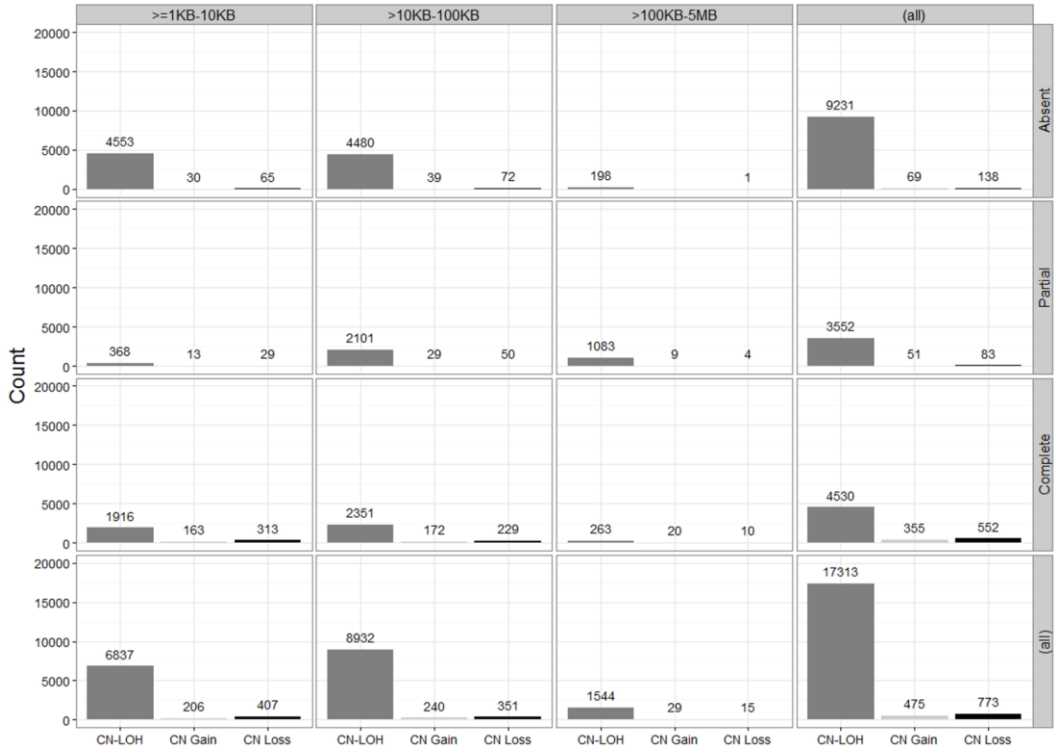


Figure 5-1 Absolute counts of CNAs stratified by overlap with germline CNVs in DGV and their length. Shown in the histograms are total numbers of copy number losses (CN Loss), copy number gains (CN Gain) and copy neutral-loss of heterozygosities (CN-LOHs) identified in 363 samples stratified by their lengths (≥1KB-10KB, >10KB-100KB and >100KB-5MB) and their overlap with known germline CNVs in the Database of Genomic Variants (DGV), Toronto. A 100% overlap is shown as 'Complete', less than 100% but more than 0% is shown as 'Partial' and no overlap is shown as 'Absent'.

I observed three copy number gains (two in chromosome 14 and one in chromosome 2) that were present in all 363 samples. Moreover, 9,123 (approximately 50%) of the CNAs identified in the 363 samples exhibited either

complete (100%) or partial overlap (more than 0% but less than 100%) with known germline CNVs reported in the Database of Genomic Variants (DGV), Toronto (http://projects.tcag.ca/variation/). There were 9,438 (69 copy number gains, 138 copy number losses and 9,231 CN-LOHs) observed in the current study that are absent in the DGV (0% overlap) and hence may be novel chromosomal aberrations that merit independent replication.

5.3.3 CNAs associated with BCR

Of the 18,561 CNAs with more than one kb (152 BCR and 211 non-BCR), I found 9,164 CNAs (145 copy number gains, 241 copy number losses and 8,778 CN-LOHs) with ≥10% frequency either in the BCR or non-BCR groups or in both. When I compared RFS probabilities with and without these CNAs in 363 samples, I found that 585 CNAs (33 copy number gains, 33 copy number losses and 519 CN-LOHs) showed statistically significant differences in RFS probabilities at nominal $P<0.05$ (**Figure 5-2**).

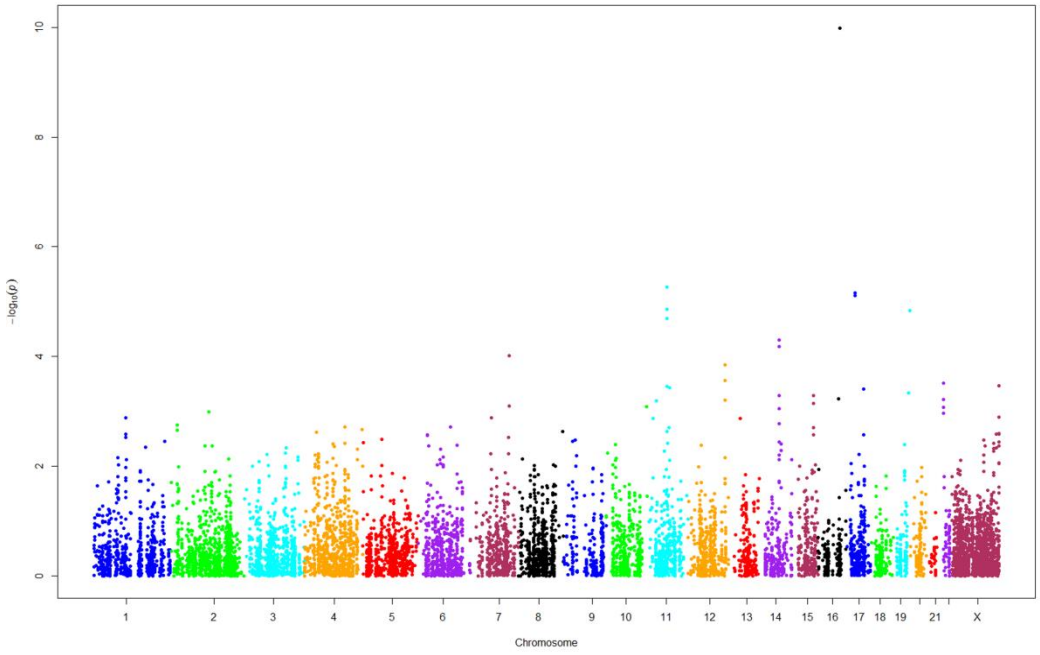


Figure 5-2 Chromosome-wide distributions of 9,164 CNAs tested for association with BCR in unstratified samples. Shown on the x-axis are middle points of chromosomal start and end positions (NCBI Build 37) of 9,161CNAs and on the y-axis are –log10 P values for their association with the phenotype of BCR in unstratified 363 samples. P values were obtained from log-rank tests with one d.f.

Of these, two copy number gains and five CN-LOHs showed the strongest differences in RFS probabilities (P<2.01 x 10^{-5}, Q<0.03) (**Figure 5-2, Table 5-2**) and all seven CNAs (three CNAs at chromosome 11, two at chromosome 17 and one CNA each at chromosomes 16 and 19) were also associated with increased risk of recurrence.

Table 5-2 Chromosomal aberrations statistically significantly associated with BCR in 363 samples.

Chromosomal regions[a]	Cyto band	Length (bps)	Event	Overlap with DGV	Genes/loci	No. of events	*P* value[b]	Q value[c]	$HR_{unadusted}$, 95% CI	$HR_{adjusted}$[d], 95% CI
chr16:70151941-70198049	q22.1	46,109	CN Gain	Complete	*CLEC18A/ PDPR*	21	1.02×10^{-10}	9.35×10^{-7}	4.49 [2.73-7.40]	3.88 [2.35-6.41]
chr17:30556456-30568424	q11.2	11,969	CN-LOH	Absent	Intergenic	66	7.79×10^{-6}	0.02	2.20 [1.54-3.13]	2.09 [1.46-2.99]
chr11:64169509-64201009	q13.1	31,501	CN-LOH	Absent	Intergenic	23	5.46×10^{-6}	0.02	2.98 [1.82-4.88]	2.28 [1.35-3.85]
chr17:30292345-30436095	q11.2	143,751	CN-LOH	Complete	*SUZ12/ LRRC37B/ SH3GL1P1*	68	6.90×10^{-6}	0.02	2.19 [1.54-3.12]	2.06 [1.45-2.94]
chr19:53519960-53538651	q13.41	18,692	CN Gain	Complete	Intergenic	47	1.44×10^{-5}	0.02	2.35 [1.58-3.50]	2.34 [1.56-3.51]
chr11:64228316-64258125	q13.1	29,810	CN-LOH	Absent	Intergenic	36	1.39×10^{-5}	0.02	2.52 [1.64-3.88]	2.08 [1.32-3.26]
chr11:64048319-64143935	q13.1	95,617	CN-LOH	Partial	*BAD/ KCNK4/ GPR137/ ESRRA/ CCDC88B*	25	2.01×10^{-5}	0.03	2.74 [1.69-4.43]	2.15 [1.29-3.58]

[a]chromosomal positions are based on NCBI build 37; DGV, Database of Genomic Variants (Toronto); HR, hazard ratio; CI, confidence interval; [b]*P* value obtained from log-rank test with one d.f; [c]FDR corrected for multiple hypothesis testing; [d]adjusted for tumor stage and grade; *CLEC18A*, C-type lectin domain family 18, member A; *PDPR*, pyruvate dehydrogenase phosphatase regulatory subunit; *SUZ12*, suppressor of zeste 12 homolog (*Drosophila*); *LRRC37B*, leucine rich repeat containing 37B; *SH3GL1P1*, SH3-domain GRB2-like 1 pseudogene 1; *BAD*, BCL2-associated agonist of cell death; *KCNK4*, potassium channel, subfamily K, member 4; *GPR137*, G protein-coupled receptor 137; *ESRRA*, estrogen-related receptor alpha; *CCDC88B*, coiled-coil domain containing 88B

Chromosome 11 CNAs. (i) A CN-LOH of 31,501 bp at chromosome 11q13.1 indicating significant differences in RFS probabilities (P=5.46 x 10^{-6}, Q=0.02) was associated with BCR ($HR_{unadjusted}$, 95% CI=2.98 [1.82-4.88];$HR_{adjusted}$, 95% CI=2.28 [1.35-3.85]). I did not observe any germline CNVs in the DGV overlapping with this CNA. (ii) A CN-LOH of 29,810bp at chromosome 11q13.1 indicating significant differences in RFS probabilities (P=1.39 x 10^{-5}, Q=0.02) was associated with BCR ($HR_{unadjusted}$, 95% CI=2.52 [1.64-3.88];$HR_{adjusted}$, 95% CI=2.08 [1.32-3.26]). There were no overlapping known CNVs reported in the DGV. (iii) Another CN-LOH of 95,617 bp at

chromosome 11q13.1 (exhibiting partial overlap with germline CNVs in DGV) showing significant differences in RFS probabilities (P=2.01 x 10^{-5}, Q=0.03) was associated with BCR ($HR_{unadjusted}$, 95% CI=2.74 [1.69-4.43];$HR_{adjusted}$, 95% CI=2.15 [1.29-3.58]).

Chromosome 17 CNAs. (i) A CN-LOH of 11,969 bp at chromosome 17q11.2 showing significant differences in RFS probabilities (P=7.79 x 10^{-6}, Q=0.02) was associated with BCR ($HR_{unadjusted}$, 95% CI=2.20 [1.54-3.13]; $HR_{adjusted}$, 95% CI=2.09 [1.46-2.99]). I did not find any known germline CNVs in the DGV that overlapped with this CN-LOH, suggesting that it could be a novel CNA. (ii) A CN-LOH of 143,751 bp at chromosome 17q11.2 (exhibiting complete overlap with germline CNVs in DGV) with significant differences in RFS probabilities (P=6.90 x 10^{-6}, Q=0.02) was also associated with BCR ($HR_{unadjusted}$, 95% CI=2.19 [1.54-3.12]; $HR_{adjusted}$, 95% CI=2.06 [1.45-2.94]).

Chromosome 16 CNA. A copy number gain of 46,109 bp at chromosome 16q22.1 (with complete overlap with germline CNVs in DGV) that showed significant differences in RFS probabilities (P=1.02 x 10^{-10}, Q=9.35 x 10-7) was associated with BCR ($HR_{unadjusted}$, 95% CI=4.49 [2.73-7.40]; $HR_{adjusted}$, 95% CI=3.88 [2.35-6.41]).

Chromosome 19 CNA. A copy number gain of 18,692bp at chromosome 19q13.41 (with complete overlap with germline CNVs in DGV) showing significant differences in RFS probabilities (P=1.44 x 10^{-5}, Q=0.02) was associated with BCR ($HR_{unadjusted}$, 95% CI=2.35 [1.58-3.50]; $HR_{adjusted}$, 95% CI=2.34 [1.56-3.51]).

I then compared the RFS probabilities of the above seven CNAs (*i.e.*, with similar start and end positions) in each of the molecular subtypes of breast cancer using log-rank tests with one d.f. to examine for possible overlap of these genomic signatures across molecular subtypes. Differences in RFS probabilities and magnitude and direction of associations (HRs and corresponding 95% CIs) of all seven CNAs with BCR in 208 luminal A samples (80 BCR and 128 non-BCR) were comparable to those observed in entire 363 samples (**Table 5-3**).

Table 5-3 Association of top seven CNAs (Table 5-2) with BCR in 208 luminal A samples.

Chromosomal regions[a]	Event	No. of events	*P* value	$HR_{unadjusted}$, 95% CI	$HR_{adjusted}$[b], 95% CI
chr16:70151941-70198049	CN Gain	10	2.95×10^{-6}	4.92 [2.35-10.33]	4.95 [2.30-10.67]
chr17:30556456-30568424	CN-LOH	39	8.13×10^{-4}	2.23 [1.38-3.60]	2.01 [1.22-3.29]
chr11:64169509-64201009	CN-LOH	12	9.08×10^{-4}	3.07 [1.53-6.17]	1.89 [0.89-4.05]
chr17:30292345-30436095	CN-LOH	39	6.70×10^{-4}	2.26 [1.39-3.65]	1.98 [1.21-3.25]
chr19:53519960-53538651	CN Gain	24	1.83×10^{-7}	3.75 [2.20-6.39]	4.08 [2.29-7.26]
chr11:64228316-64258125	CN-LOH	20	3.25×10^{-7}	3.84 [2.20-6.69]	2.82 [1.54-5.14]
chr11:64048319-64143935	CN-LOH	13	3.89×10^{-4}	3.15 [1.61-6.14]	2.03 [0.98-4.19]

[a]chromosomal positions are based on NCBI build 37; [b]adjusted for tumor stage and grade

However, the differences in RFS probabilities were statistically non-significant in other subtypes (luminal B, HER2 type and triple negative), except for a copy number gain at chromosome 16q22.1 ($P<6.15 \times 10^{-3}$) in luminal B and triple negative subtypes, for a CN-LOH at chromosome 17q11.2 ($P=2.63 \times 10^{-3}$) in the luminal B subtype and for a CN-LOH at chromosome 11q13.1 ($P=5.35 \times 10^{-4}$) in the triple negative subtype (**Table S5-3**).Thus, the seven CNAs reported here appeared to be relatively specific to the luminal A subtype of breast cancer,

as would be expected of the sample composition with luminal A cases as a major subset.

5.3.4 Subgroup analysis restricted to luminal A samples (n=208)

In an attempt to identify additional CNAs, I estimated the differences in RFS probabilities with and without CNAs in the luminal A subtype (80 BCR and 128 non-BCR) of breast cancer. I identified 7,218 CNAs (142 copy number gains, 258 copy number losses and 6,818 CN-LOHs) with ≥10% frequency either in at least one group or in both. Of these, 4,379 CNAs shared commonality with 9,164 CNAs observed in the entire 363 samples while 2,839 CNAs were distinct, owing to the variant start and end positions, chromosomal locations or the indicated frequency threshold of ≥10% in BCR or non-BCR cases or in both. I identified a total of 484 of 7,218 CNAs (27 copy number gains, 32 copy number losses and 425 CN-LOHs) showing statistically significant differences in RFS probabilities with and without CNAs at nominal $P<0.05$ (**Figure 5-3**). Of these, three CN-LOHs showed the strongest statistically significant differences in RFS probabilities in the luminal A subtype of breast cancer (**Table 5-4**), vis-à-vis from the 2,839 distinct CNAs in this sub group.

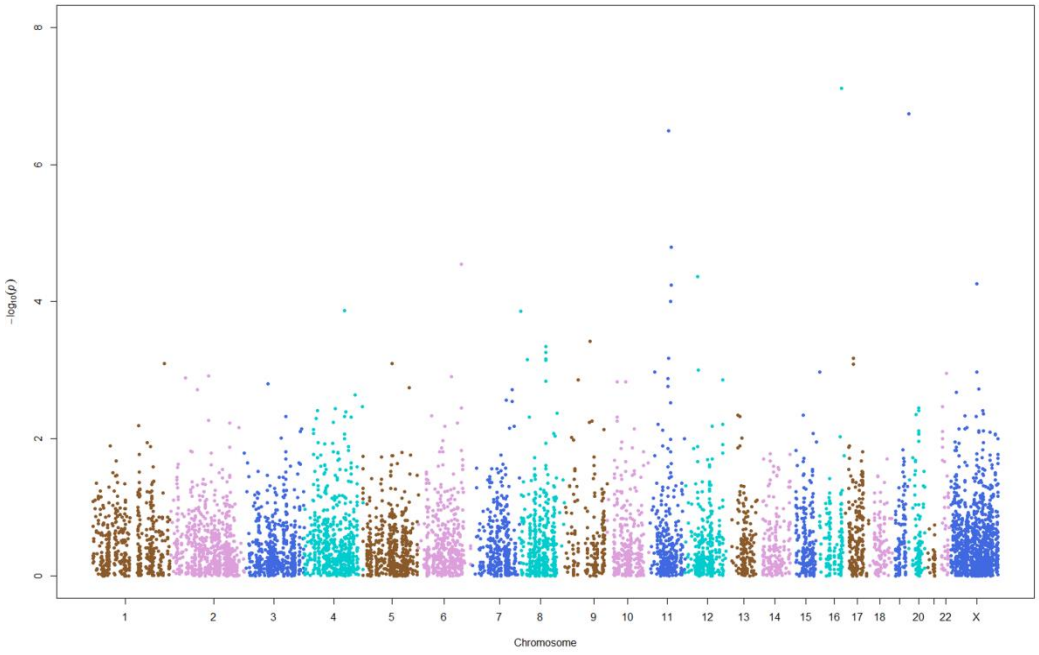


Figure 5-3 Chromosome-wide distribution of 7,218 CNAs tested for association with BCR in 208 luminal A cases. Shown on the x-axis are middle points of chromosomal start and end positions (NCBI Build 37) of 7,218 CNAs and on the y-axis are –log10 P values for their association with the phenotype of BCR in 208 luminal A cases. P values were obtained from log-rank tests with one d.f.

The three distinct CNAs identified in this sub group showed the following characteristics. (i) A CN-LOH of 57,590 bp at chromosome 11q13.1 showing significant differences in RFS probabilities (P=3.25 x 10^{-7}, Q=7.82 x 10^{-4}) was associated with increased risk of BCR ($HR_{unadjusted}$, 95% CI=3.84 [2.20-6.69]; $HR_{adjusted}$, 95% CI=2.82 [1.54-5.14]). (ii) A CN-LOH of length 9,850 bp at chromosome 11q13.4 indicating significant differences in RFS probabilities (P=1.62 x 10^{-5}, Q=0.03) was associated with increased risk of BCR ($HR_{unadjusted}$,

95% CI=3.58 [1.93-6.66]; $HR_{adjusted}$, 95% CI=2.60 [1.27-5.31]). (iii) And lastly, a CN-LOH of length 91,670 bp at chromosome 6q24.1 showing significant differences in RFS probabilities (P=2.86 x 10^{-5}, Q=0.04)was associated with increased risk of BCR ($HR_{unadjusted}$, 95% CI=2.54 [1.62-3.98]; $HR_{adjusted}$, 95% CI=2.38 [1.50-3.76]). I did not find any known germline CNVs in the DGV that overlapped with these three CN-LOHs.

Table 5-4 Additional CNAs statistically significantly associated with BCR in 208 luminal A samples.

Chromosomal regions[a]	Cyto band	Length (bp)	Event	Overlap with DGV	Genes/ loci	No. of events	*P* value[b]	Q value[c]	$HR_{unadjusted}$, 95% CI	$HR_{adjusted}$[d], 95% CI
chr11:64228316-64285905	q13.1	57,590	CN-LOH	Absent	Intergenic	20	3.25 x 10^{-7}	7.82 x 10^{-4}	3.84 [2.20-6.69]	2.82 [1.54-5.14]
chr11:72198387-72208236	q13.4	9,850	CN-LOH	Absent	Intergenic	16	1.62 x 10^{-5}	0.03	3.58 [1.93-6.66]	2.60 [1.27-5.31]
chr6:140631638-140723307	q24.1	91,670	CN-LOH	Absent	Intergenic	52	2.86 x 10^{-5}	0.04	2.54 [1.62-3.98]	2.38 [1.50-3.76]

[a]chromosomal positions are based on NCBI build 37; DGV, Database of Genomic Variants (Toronto); HR, hazard ratio; CI, confidence interval; [b]*P* value obtained from log-rank test with one d.f.; [c]FDR corrected for multiple hypothesis testing; [d]adjusted for tumor stage and grade

I did not find statistically significant differences in RFS probabilities with and without above three CN-LOHs (11q13.1, 11q13.4 and 6q24.1) in luminal B, HER2 type or triple negative subtypes suggesting that these CN-LOHs were relatively specific to the luminal A subtype of breast cancer (**Table S5-3**). Overall, chromosome 11 appears to harbour multiple CN-LOHs (identified 5 CNAs in total, **Tables 5-2 to 5-4** and **Table S5-3**) and these showed increased risk of BCR in the luminal A subtype of breast cancer.

Moreover, adjustment in HRs and 95% CI for tumor grade and stage in the analyses presented thus far revealed minimal or no evidence of potential

confounding effects. Hence, these clinicopathological characteristics are less likely to significantly modify the observed associations of identified CNAs with BCR at the indicated sample size (**Tables 5-1 to 5-4**).

Table S5-3 Relationship of top ten CNAs (Tables 5-2 and 5-4) with BCR in luminal B, HER2 type and triple negative subtypes of breast cancer samples.

Chromosomal regions[a]	Event	Luminal B (n=60)		HER2 type (n=24)		Triple negative (n=42)	
		No. of events	*P* value[b]	No. of events	*P* value[b]	No. of events	*P* value[b]
chr16:70151941-70198049	CN Gain	2	2.89×10^{-3}	2	0.13	5	6.15×10^{-3}
chr17:30556456-30568424	CN-LOH	9	0.01	4	0.55	10	0.08
chr11:64169509-64201009	CN-LOH	5	0.13	1	0.54	4	$5.35\ 10^{-4}$
chr17:30292345-30436095	CN-LOH	10	2.63×10^{-3}	4	0.55	10	0.08
chr19:53519960-53538651	CN Gain	7	0.96	3	0.41	10	0.55
chr11:64228316-64258125	CN-LOH	6	0.79	1	0.54	6	0.11
chr11:64048319-64143935	CN-LOH	5	0.13	1	0.54	5	0.16
chr11:64228316-64285905	CN-LOH	6	0.79	1	0.54	5	0.22
chr11:72198387-72208236	CN-LOH	5	0.05	1	0.69	2	0.11
chr6:140631638-140723307	CN-LOH	9	0.07	3	0.50	5	0.44

[a]chromosomal positions are based on NCBI build 37; [b]*P* values obtained from log-rank tests with one degree of freedom

5.3.5 RT-qPCR validation of select CNAs in representative samples

Of the ten CNAs (eight CN-LOHs and two copy number gains) showing statistically significant association with the phenotype of BCR, I chose to validate three relatively longer CN-LOHs (143,751 bp at 17q11.2, 57,590 bp at 11q13.1 and 91,670 bp at 6q24.1) in a subset of 363 samples (a combination of randomly selected samples harbouring the CN-LOHs plus approximately equal proportion of samples without these CN-LOHs as evaluated by the Nexus Copy Number 6.0) by RT-qPCR and Sequenom genotyping. There is a growing consensus that CN-LOHs are important in the genomes profiled using germline DNA in the recent

literature [27-29,32] and this formed the basis for validating the most predominant chromosomal aberrations in this study. These newly emerging chromosomal aberrations (CN-LOHs), if confirmed, may be included in future investigations alongside the copy loss or gain aberrations for a comprehensive catalogue of CNAs relevant for complex/polygenic phenotypes. Remaining two CN-LOHs were shorter in size and were not considered for validation. First, the copy number status of these three CN-LOHs was quantified using copy number assays (Hs00138078_cn and Hs02495547_cn for CN-LOH at 17q11.2 in 38 samples (interrogated using two assays in this region, largest of the CN-LOH identified in this study), Hs06324464_cn for CN-LOH at 11q13.1 in 33 samples and Hs06809880_cn for CN-LOH at 6q24.1 in 36 samples). Concordance between copy number calls made from Nexus read-out and RT-qPCR was 100% showing copy number of two (**Table S5-4**). Second, 24 of 30 SNPs initially selected for three CN-LOHs in the same DNA samples used for copy number assays were successfully genotyped to measure percentage of heterozygosity in each CN-LOH; assays of six SNPs were not successful. Percentages of heterozygosity were calculated using seven SNPs for CN-LOH at 17q11.2, nine SNPs for CN-LOH at 11q13.1 and eight SNPs for CN-LOH at 6q24.1. Pearson correlations coefficients between heterozygote frequencies measured from Affymetrix SNP 6.0 array data and from Sequenom iPLEX Gold Platform for CN-LOH at 17q11.2, CN-LOH at 11q13.1 and CN-LOH at 6q24.1 were 0.97, 0.98 and 0.99, respectively (**Table S5-4**).

Table S5-4 Validation of three CN-LOHs in subset of 208 samples using RT-qPCR and Sequenom genotyping.

	CN-LOH of 143,751 bp at 17q11.2										
Sample_ID	Affymetrix 6_logratio	Affymetrix 6_CN	Affymetrix 6_hetfreq	Affymetrix 6_CN call	qPCR_assay1	qPCR_assay 2	qPCR_average_CN	Sequenom_hetfreq	Sequenom+qPCR_CN call	Status
C223	-0.07	1.91	0.00	CN-LOH	1.97	1.95	1.96	0.00	CN-LOH	BCR
C24	-0.02	1.98	10.00	CN-LOH	2.05	1.99	2.02	0.00	CN-LOH	BCR
C249	-0.01	1.99	10.00	CN-LOH	1.83	2.03	1.93	0.00	CN-LOH	non-BCR
C34	0.03	2.05	0.00	CN-LOH	1.87	1.88	1.88	0.00	CN-LOH	non-BCR
C369	-0.01	1.99	0.00	CN-LOH	1.95	1.96	1.96	0.00	CN-LOH	non-BCR
C398	-0.04	1.95	20.00	CN-LOH	1.96	1.85	1.91	0.00	CN-LOH	BCR
C440	-0.08	1.90	10.00	CN-LOH	1.86	2.05	1.96	0.00	CN-LOH	non-BCR
C534	-0.14	1.82	0.00	CN-LOH	1.90	1.96	1.93	0.00	CN-LOH	non-BCR
C81	0.03	2.05	0.00	CN-LOH	2.20	2.31	2.26	0.00	CN-LOH	BCR
MT1180	0.03	2.04	10.00	CN-LOH	1.85	1.90	1.88	0.00	CN-LOH	BCR
MT209	-0.03	1.95	10.00	CN-LOH	1.77	1.94	1.86	0.00	CN-LOH	BCR
MT371	-0.02	1.98	0.00	CN-LOH	1.93	2.07	2.00	0.00	CN-LOH	BCR
C18	0.00	2.01	0.00	CN-LOH	2.21	2.01	2.11	0.14	CN-LOH	BCR
C49	-0.04	1.95	10.00	CN-LOH	2.17	2.19	2.18	0.14	CN-LOH	non-BCR
C56	-0.03	1.95	0.00	CN-LOH	2.00	2.10	2.05	0.14	CN-LOH	BCR
C702	-0.02	1.97	0.00	CN-LOH	2.08	2.08	2.08	0.14	CN-LOH	BCR
GT13	-0.04	1.94	0.00	CN-LOH	2.21	2.28	2.25	0.14	CN-LOH	non-BCR
MT1213	-0.05	1.93	10.00	CN-LOH	2.17	2.10	2.14	0.14	CN-LOH	BCR
MT441	0.01	2.02	0.00	CN-LOH	2.00	2.12	2.06	0.14	CN-LOH	BCR
MT452	0.05	2.07	10.00	CN-LOH	2.09	2.05	2.07	0.14	CN-LOH	BCR
MT813	-0.09	1.88	10.00	CN-LOH	2.45	2.18	2.32	0.14	CN-LOH	BCR
C135	-0.03	1.96	60.00	Normal	1.81	1.93	1.87	0.71	Normal	non-BCR
C209	0.00	2.01	60.00	Normal	1.86	1.94	1.90	0.71	Normal	BCR
C218	-0.04	1.95	70.00	Normal	2.08	2.02	2.05	0.71	Normal	BCR
C22	-0.08	1.90	70.00	Normal	2.07	1.82	1.95	0.71	Normal	non-BCR
C371	0.11	2.16	70.00	Normal	1.77	1.80	1.79	0.71	Normal	BCR
C388	-0.08	1.89	60.00	Normal	1.80	1.88	1.84	0.71	Normal	non-BCR
C448	0.01	2.01	70.00	Normal	2.10	1.98	2.04	0.71	Normal	BCR
C494	-0.07	1.90	60.00	Normal	1.78	1.81	1.80	0.71	Normal	BCR
C498	-0.10	1.87	70.00	Normal	1.88	2.01	1.95	0.71	Normal	non-BCR
C571	-0.02	1.97	70.00	Normal	1.99	2.01	2.00	0.71	Normal	non-BCR
MT103	-0.01	1.99	70.00	Normal	1.91	1.84	1.88	0.71	Normal	non-BCR
MT121	0.00	2.00	70.00	Normal	1.90	1.94	1.92	0.71	Normal	non-BCR
MT196	-0.06	1.92	70.00	Normal	1.74	1.98	1.86	0.71	Normal	non-BCR
MT480	-0.03	1.97	70.00	Normal	1.99	2.01	2.00	0.71	Normal	BCR
MT606	0.01	2.02	60.00	Normal	1.86	2.05	1.96	0.71	Normal	non-BCR
MT79	0.05	2.07	60.00	Normal	2.00	2.04	2.02	0.71	Normal	BCR
MT848	-0.01	1.99	70.00	Normal	2.12	2.19	2.16	0.71	Normal	non-BCR

Table S5-4 Continued..

					CN-LOH of 57,590 bp at 11q13.1					
C81	0.06	2.08	0.00	CN-LOH	2.21	NA	NA	0.00	CN-LOH	BCR
MT452	0.05	2.07	0.00	CN-LOH	2.04	NA	NA	0.00	CN-LOH	BCR
MT478	-0.01	1.98	0.00	CN-LOH	2.00	NA	NA	0.00	CN-LOH	BCR
MT813	-0.04	1.94	0.00	CN-LOH	2.30	NA	NA	0.00	CN-LOH	BCR
MT867	-0.02	1.98	0.00	CN-LOH	2.19	NA	NA	0.00	CN-LOH	BCR
C108	-0.11	1.85	0.00	CN-LOH	2.13	NA	NA	11.11	CN-LOH	non-BCR
C144	-0.08	1.89	0.00	CN-LOH	1.94	NA	NA	11.11	CN-LOH	BCR
C209	0.02	2.03	0.00	CN-LOH	2.00	NA	NA	11.11	CN-LOH	BCR
C60	-0.10	1.86	8.33	CN-LOH	1.97	NA	NA	11.11	CN-LOH	BCR
MT689	-0.06	1.92	0.00	CN-LOH	2.20	NA	NA	11.11	CN-LOH	BCR
GT13	0.02	2.02	50.00	Normal	2.24	NA	NA	55.56	Normal	non-BCR
GT2	-0.10	1.87	50.00	Normal	2.01	NA	NA	55.56	Normal	non-BCR
GT7	0.03	2.04	50.00	Normal	2.09	NA	NA	55.56	Normal	non-BCR
GT74	0.05	2.07	50.00	Normal	2.14	NA	NA	55.56	Normal	non-BCR
MT103	0.04	2.06	50.00	Normal	1.84	NA	NA	55.56	Normal	non-BCR
MT121	0.12	2.17	50.00	Normal	1.89	NA	NA	55.56	Normal	non-BCR
MT209	0.03	2.04	50.00	Normal	1.94	NA	NA	55.56	Normal	BCR
MT237	0.03	2.04	50.00	Normal	1.74	NA	NA	55.56	Normal	BCR
MT275	0.03	2.04	50.00	Normal	1.78	NA	NA	55.56	Normal	BCR
MT289	0.12	2.18	50.00	Normal	1.95	NA	NA	55.56	Normal	BCR
MT290	0.14	2.21	50.00	Normal	1.86	NA	NA	55.56	Normal	non-BCR
MT355	0.08	2.11	50.00	Normal	1.86	NA	NA	55.56	Normal	non-BCR
MT371	-0.08	1.89	50.00	Normal	2.02	NA	NA	55.56	Normal	BCR
MT379	0.01	2.02	50.00	Normal	1.93	NA	NA	55.56	Normal	non-BCR
MT49	-0.05	1.93	50.00	Normal	1.82	NA	NA	55.56	Normal	non-BCR
MT517	0.00	2.00	50.00	Normal	1.99	NA	NA	55.56	Normal	BCR
MT606	0.07	2.10	50.00	Normal	2.01	NA	NA	55.56	Normal	non-BCR
MT66	-0.02	1.97	50.00	Normal	2.16	NA	NA	55.56	Normal	BCR
MT94	-0.13	1.83	50.00	Normal	1.90	NA	NA	55.56	Normal	BCR
BM01-15	0.06	2.08	50.00	Normal	2.11	NA	NA	66.67	Normal	non-BCR
C109	0.14	2.20	50.00	Normal	2.09	NA	NA	66.67	Normal	non-BCR
C11	0.08	2.11	50.00	Normal	1.95	NA	NA	66.67	Normal	BCR
C134	0.16	2.23	50.00	Normal	1.91	NA	NA	66.67	Normal	non-BCR

Table S5-4 Continued..

				CN-LOH of 91,670 bp at 6q24.1						
C152	-0.03	1.95	0.00	CN-LOH	2.06	NA	NA	0	CN-LOH	BCR
C203	-0.09	1.87	0.00	CN-LOH	1.95	NA	NA	0	CN-LOH	BCR
C223	0.02	2.02	0.00	CN-LOH	2.02	NA	NA	0	CN-LOH	BCR
C24	-0.02	1.97	0.00	CN-LOH	1.89	NA	NA	0	CN-LOH	BCR
C281	-0.01	1.98	0.00	CN-LOH	1.87	NA	NA	0	CN-LOH	BCR
C284	0.04	2.05	0.00	CN-LOH	1.95	NA	NA	0	CN-LOH	non-BCR
C285	0.03	2.04	0.00	CN-LOH	1.91	NA	NA	0	CN-LOH	non-BCR
C293	-0.01	1.98	0.00	CN-LOH	1.83	NA	NA	0	CN-LOH	non-BCR
C341	-0.02	1.98	0.00	CN-LOH	1.78	NA	NA	0	CN-LOH	non-BCR
C423	-0.01	1.99	0.00	CN-LOH	2.01	NA	NA	0	CN-LOH	BCR
C482	-0.01	1.98	0.00	CN-LOH	2.04	NA	NA	0	CN-LOH	non-BCR
C515	-0.09	1.88	0.00	CN-LOH	1.96	NA	NA	0	CN-LOH	non-BCR
C524	-0.01	1.98	0.00	CN-LOH	2.07	NA	NA	0	CN-LOH	BCR
C702	-0.01	1.98	0.00	CN-LOH	1.97	NA	NA	0	CN-LOH	BCR
C81	-0.06	1.91	0.00	CN-LOH	2.13	NA	NA	0	CN-LOH	BCR
MT101	0.01	2.02	0.00	CN-LOH	1.89	NA	NA	0	CN-LOH	BCR
MT155	-0.04	1.95	0.00	CN-LOH	1.79	NA	NA	0	CN-LOH	non-BCR
MT196	-0.02	1.97	0.00	CN-LOH	1.85	NA	NA	0	CN-LOH	non-BCR
MT289	-0.03	1.96	0.00	CN-LOH	2.08	NA	NA	0	CN-LOH	BCR
MT370	-0.04	1.95	0.00	CN-LOH	2.2	NA	NA	0	CN-LOH	non-BCR
MT38	0.05	2.07	0.00	CN-LOH	1.88	NA	NA	0	CN-LOH	BCR
MT380	0.03	2.04	0.00	CN-LOH	1.88	NA	NA	0	CN-LOH	BCR
MT441	-0.05	1.93	0.00	CN-LOH	2.06	NA	NA	0	CN-LOH	BCR
MT517	-0.05	1.93	0.00	CN-LOH	1.92	NA	NA	0	CN-LOH	BCR
MT79	0.03	2.05	0.00	CN-LOH	2.05	NA	NA	0	CN-LOH	BCR
MT87	0.04	2.06	0.00	CN-LOH	2.22	NA	NA	0	CN-LOH	BCR
C250	-0.09	1.88	0.00	CN-LOH	1.96	NA	NA	12.5	CN-LOH	non-BCR
C339	0.06	2.09	0.00	CN-LOH	1.82	NA	NA	12.5	CN-LOH	BCR
PK70	0.02	2.03	0.00	CN-LOH	2.09	NA	NA	12.5	CN-LOH	non-BCR
BM01-15	0.01	2.01	60.00	Normal	2.09	NA	NA	62.5	Normal	non-BCR
BM01-28	0.01	2.01	60.00	Normal	2.01	NA	NA	62.5	Normal	non-BCR
BM01-4	0.08	2.12	60.00	Normal	2.45	NA	NA	62.5	Normal	non-BCR
C11	-0.01	1.98	60.00	Normal	2	NA	NA	62.5	Normal	BCR
C134	-0.01	1.99	60.00	Normal	2.15	NA	NA	62.5	Normal	non-BCR
C138	-0.05	1.93	60.00	Normal	2.11	NA	NA	62.5	Normal	non-BCR
C139	0.01	2.01	50.00	Normal	1.86	NA	NA	62.5	Normal	non-BCR

5.4 Discussion and Conclusion

In this study, I identified ten germline CNAs as potential prognostic factors for disease recurrence in the early-stage non-metastatic breast cancer. These germline signatures were particularly relevant to the luminal A subtype as large number of breast cancer cases with luminal A tumors experience disease recurrence despite their good prognosis based on tumor characteristics. Using a sample size of 363 breast cancer patients who received standard guideline-based therapy upon diagnosis, I demonstrated statistically significant associations of ten CNAs (two copy number gains and eight CN-LOHs) with the phenotype of BCR in both univariate and multivariate analyses (adjusted for tumor stage and grade). Three CN-LOHs (17q11.2, 11q13.1 and 6q24.1) were validated in a subset of 363 samples using RT-qPCR and Sequenom iPLEX Gold Platform technologies. Adjustment for tumor stage and grade did not influence the direction or effect size reported in terms of the HRs and 95% CI, suggesting that these clinicopathological characteristics did not influence the observed association results. As such, these germline CNAs may offer significant prognostic value for breast cancer, independent of tumor clinicopathological characteristics considered in this study.

While many studies have evaluated potential role of copy number gains, copy number losses and classical LOHs, only a few have investigated the impact of CN-LOH in complex diseases such as cancer [27-29]. This may be due to inadequate karyotyping technology as conventional cytogenetics (array-CGH) and fluorescence in situ hybridization (FISH) cannot detect these small unique

chromosomal aberrations. However, with the availability of high-resolution SNP-arrays containing both copy number and SNP probes, it is now possible to identify previously hidden CN-LOHs. Mitotic recombination between pairs of homologous chromosomes is believed to be the underlying mechanism generating CN-LOHs [32,33]. Studies have shown that CN-LOHs tend to localize within fragile sites, previously known regions of frequent genomic instability [32,33]. Potential clinical utility of CN-LOHs is recently being appreciated, as CN-LOHs are associated with duplication of oncogenic alleles with simultaneous loss of normal functional alleles.

I have validated three CN-LOHs in an independent genotyping platform and with the following generalized features:

(i) A CN-LOH at 17q11.2 showed significant associations with BCR in unstratified 363 samples while comparable log-rank *P* values and HRs (increased risk) were also observed in the molecularly stratified 208 luminal A samples. The CN-LOH also demonstrated entire overlap with multiple germline CNVs in DGV, including both copy number gains and losses. This CN-LOH harboured three known genes such as suppressor of zeste12 homolog (*Drosophila*) (*SUZ12*), leucine rich repeat containing 37B (*LRRC37B*) and SH3-domain GRB2-like 1 pseudogene 1 (*SH3GL1P1*). *SUZ12* is a zinc finger gene often found at the breakpoints of recurrent chromosomal translocation in endometrial stromal sarcoma [39]. It has also been shown to act as a transcriptional repressor of Homeo box protein Hox-A9 gene in primary breast cancers through DNA hypermethylation and recruitment of DNA methyltransferases [40]. Protein

encoded by *LRRC37B* gene is not well-characterized yet. However, a recent study has reported that the *LRRC37B* locus may harbour non-allelic homologous recombination hotspot, a major mechanism involved in chromosomal rearrangements [41]. *SH3GL1P1*is a pseudogene with no known function. Chromosome 17q11.2 region is also known to harbour CNVs as this loci is a hot spot for segmental duplications [42].

The CN-LOH was more specific to the luminal A subtype of breast cancer as log-rank *P* values and HRs were comparable in luminal A samples only but were statistically insignificant in other sub-phenotypes such as HER2 type and triple negative, except in the luminal B subtype (albeit at the limited sample size for other subtypes of breast cancer). Recently, distinct CNA profiles were reported for molecular subtypes of breast cancer [43] and the findings from the curent study not only support such a premise but also extend these observations to the disease outcomes. On the other hand, intrinsic molecular similarities between the luminal A and luminal B subtypes of breast cancer, especially in terms of ER and PR status, may be attributed to similar log-rank *P* values in both groups;

(ii) The two remaining CNAs at 11q13.1 and 6q24.1 were detected in subgroup analyses restricted to luminal A cases (BCR=80, non-BCR=128) showing significant differences in RFS probabilities and conferred risk to BCR. Both CN-LOHs at 11q13.1 and 6q24.1 are novel and did not harbour any known genes; however, these may still influence the phenotype through cis-acting regulatory activities. CN-LOH at 11q13.1 did not contain any known genes but solute carrier family 22 (organic anion/urate transporter) (*SLC22A11*) was located

~37.19 kb down-stream of the CNA. The integral membrane protein encoded by the *SLC22A11* gene acts as an organic anion transporter, which mainly involves in transfer of estrone 3 sulfate through plasma membrane [44]. CN-LOH at 6q24.1 also did not contain any known genes; however, microRNA 3668 (*MIR3668*) and microRNA 4465 (*MIR4465*) were found ~105.18 kb upstream and ~281.64 kb down-stream of this CNA. Both *MIR3668* and *MIR4465* encode microRNAs, short non-coding RNA molecules involved in post-transcriptional modifications of eukaryotic organisms.

Even though I did not perform independent survival analysis with the non luminal A molecular sub-phenotypes of breast cancer (luminal B, HER2 type and triple negative) owing to limited sample size, my results provide a rationale for conducting such analyses to identify germline CNAs specific to these molecular subgroups. Analyses based on finer classification of molecular subtypes of breast cancer encompassing ki67 marker in addition to the cell surface receptor (ER, PR and HER2 status) based classifications described here and the newly described molecular subtypes in breast cancer [43,45], may help identify more informative germline CNAs that potentially explain larger proportion of heterogeneity in breast cancer prognosis. Clinical utility of the identified germline CNAs showing strong prognostic value will be favorable if these markers are reproduced in larger but independent studies.

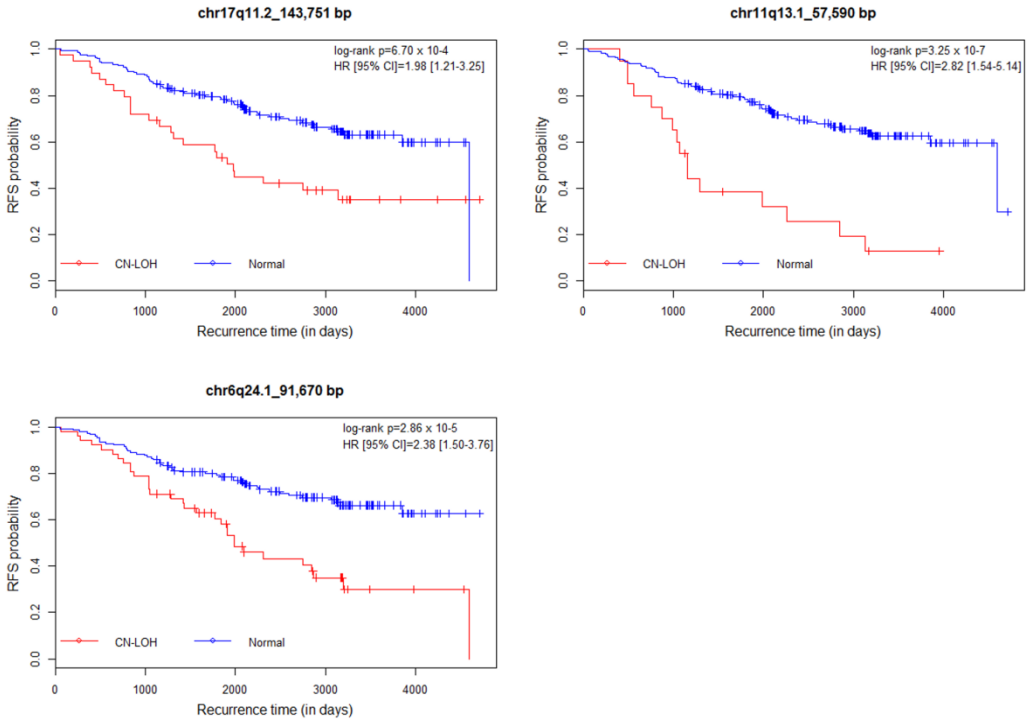


Figure 5-4 Relationships between RFS and three CN-LOHs validated by RT-qPCR and Sequenom genotyping. Using the data from 208 luminal A cases, Kaplan-Meir survival plots were generated to evaluate the predictive power of three CN-LOHs validated in independent platform for RFS. The x-axes in all three plots show recurrence time in days and the y-axes show RFS probabilities with and without CN-LOHs. Differences in RFS probabilities were assessed by log-rank tests with one d.f. HRs and 95% CIs were estimated by Cox proportional hazards model adjusted for tumor stage and grade.

In summary, I found multiple germline CNAs at chromosomes 6, 11 and 17 (results confirmed from independent genotyping platforms, **Figure 5-4** and **Table S5-4**) with potential prognostic value, independent of tumor grade and

tumor stage for early-stage non-metastatic luminal A subtype of breast cancer. Despite the large collection of recurrent cases from a single source (derived from Alberta) with extensive follow-up and outcomes data, the sample size needed for independent replication of these findings therefore warrant large international collaborations. Further investigations in to the biochemical and molecular basis for the prognostic significance of the genomic signatures may aid in the development of targeted therapeutics and molecularly driven strategies to reduce the risk of BCR.

5.5 References

1. Siegel R, Naishadham D, Jemal A. (2012) Cancer statistics, 2012. CA Cancer J Clin 62: 10-29.
2. Canadian Cancer Society's Steering Committee on Cancer Statistics.Canadian Cancer Statistics 2011. Toronto, ON: Canadian Cancer Society; 2011.
3. Early Breast Cancer Trialists' Collaborative Group (EBCTCG).(2005) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomised trials. Lancet 365: 1687-1717.
4. Weigel MT, Dowsett M. (2010) Current and emerging biomarkers in breast cancer: Prognosis and prediction.EndocrRelat Cancer 17: R245-62.
5. van der Leij F, Elkhuizen PH, Bartelink H, van de VijverMJ. (2012) Predictive factors for local recurrence in breast cancer. SeminRadiatOncol 22: 100-107.
6. Paik S, Shak S, Tang G, Kim C, Baker J, et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817-2826.
7. van de VijverMJ, He YD, van't Veer LJ, Dai H, Hart AA, et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347: 1999-2009.
8. Gonzalez-AnguloAM, Morales-Vasquez F, Hortobagyi GN. (2007) Overview of resistance to systemic therapy in patients with breast cancer.AdvExp Med Biol 608: 1-22.

9. VoducKD, Cheang MC, Tyldesley S, Gelmon K, Nielsen TO, et al. (2010) Breast cancer subtypes and the risk of local and regional relapse. J Clin Oncol 28: 1684-1691.

10. Sapkota Y, Robson P, Lai R, Cass CE, Mackey JR, et al. (2012) A two-stage association study identifies methyl-CpG-binding domain protein 2 gene polymorphisms as candidates for breast cancer susceptibility. Eur J Hum Genet 20: 682-689.

11. Sehrawat B, Sridharan M, Ghosh S, Robson P, Cass CE, et al. (2011) Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Hum Genet 130: 529-537.

12. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, et al. (2012) Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet 44: 312-318.

13. Shu XO, Long J, Lu W, Li C, Chen WY, et al. (2012) Novel genetic markers of breast cancer survival identified by a genome-wide association study. Cancer Res 72: 1182-1189.

14. Martin DN, Boersma BJ, Howe TM, Goodman JE, Mechanic LE, et al. (2006) Association of MTHFR gene polymorphisms with breast cancer survival. BMC Cancer 6: 257.

15. Lin WY, Camp NJ, Cannon-Albright LA, Allen-Brady K, Balasubramanian S, et al. (2011) A role for XRCC2 gene polymorphisms in breast cancer risk and survival. J Med Genet 48: 477-484.

16. Bray J, Sludden J, Griffin MJ, Cole M, Verrill M, et al. (2010) Influence of pharmacogenetics on response and toxicity in breast cancer patients treated with doxorubicin and cyclophosphamide. Br J Cancer 102: 1003-1009.

17. Wang L, Ellsworth KA, Moon I, Pelleymounter LL, Eckloff BW, et al. (2010) Functional genetic polymorphisms in the aromatase gene CYP19 vary the response of breast cancer patients to neoadjuvant therapy with aromatase inhibitors. Cancer Res 70: 319-328.

18. Easton DF, Pooley KA, Dunning AM, PharoahPD, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087-1093.

19. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, et al. (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 40: 703-706.

20. Krepischi AC, Achatz MI, Santos EM, Costa SS, Lisboa BC, et al. (2012) Germline DNA copy number variation in familial and early-onset breast cancer. Breast Cancer Res 14: R24.

21. Kuiper RP, LigtenbergMJ, Hoogerbrugge N, Geurts van Kessel A. (2010) Germline copy number variation and cancer risk. CurrOpin Genet Dev 20: 282-289.

22. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-454.

23. Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, et al. (2010) Genome-wide association study of CNVs in

16,000 cases of eight common diseases and 3,000 shared controls. Nature 464: 713-720.

24. Yoshihara K, Tajima A, Adachi S, Quan J, Sekine M, et al. (2011) Germline copy number variations in BRCA1-associated ovarian cancer patients. Genes Chromosomes Cancer 50: 167-177.

25. HenrichsenCN, Chaignat E, Reymond A. (2009) Copy number variants, diseases and gene expression. Hum Mol Genet 18: R1-8.

26. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848-853.

27. Melcher R, Hartmann E, Zopf W, Herterich S, Wilke P, et al. (2011) LOH and copy neutral LOH (cnLOH) act as alternative mechanism in sporadic colorectal cancers with chromosomal and microsatellite instability. Carcinogenesis 32: 636-642.

28. Saeki H, Kitao H, Yoshinaga K, Nakanoko T, Kubo N, et al. (2011) Copy-neutral loss of heterozygosity at the p53 locus in carcinogenesis of esophageal squamous cell carcinomas associated with p53 mutations. Clin Cancer Res 17: 1731-1740.

29. Kryh H, Caren H, Erichsen J, SjobergRM, Abrahamsson J, et al. (2011) Comprehensive SNP array study of frequently used neuroblastoma cell lines; copy neutral loss of heterozygosity is common in the cell lines but uncommon in primary tumors. BMC Genomics 12: 443.

30. Makishima H, Maciejewski JP.(2011) Pathogenesis and consequences of uniparental disomy in cancer.Clin Cancer Res 17: 3913-3923.

31. Lapunzina P, Monk D. (2011) The consequences of uniparental disomy and copy number neutral loss-of-heterozygosity during human development and cancer. Biol Cell 103: 303-317.

32. O'Keefe C, McDevitt MA, Maciejewski JP. (2010) Copy neutral loss of heterozygosity: A novel chromosomal lesion in myeloid malignancies. Blood 115: 2731-2739.

33. Mohamedali A, Gaken J, Twine NA, Ingram W, Westwood N, et al. (2007) Prevalence and prognostic significance of allelic imbalance by single-nucleotide polymorphism analysis in low-risk myelodysplastic syndromes. Blood 110: 3365-3373.

34. Bernstein L, LaceyJV,Jr. (2011) Receptors, associations, and risk factor differences by breast cancer subtypes: Positive or negative? J Natl Cancer Inst 103: 451-453.

35. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, et al. (2005) REporting recommendations for tumour MARKer prognostic studies (REMARK). Br J Cancer 93: 387-391.

36. Bolstad BM, Irizarry RA, Astrand M, Speed TP. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.

37. Shlien A, Malkin D. (2010) Copy number variations and cancer susceptibility. CurrOpinOncol 22: 55-63.

38. Benjamini Y, Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289-300.

39. Koontz JI, Soreng AL, Nucci M, Kuo FC, Pauwels P, et al. (2001) Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors. ProcNatlAcadSci U S A 98: 6348-6353.

40. Reynolds PA, Sigaroudinia M, Zardo G, Wilson MB, Benton GM, et al. (2006) Tumor suppressor p16INK4A regulates polycomb-mediated DNA hypermethylation in human mammary epithelial cells. J BiolChem 281: 24790-24802.

41. Zickler AM, Hampp S, Messiaen L, Bengesser K, Mussotter T, et al. (2012) Characterization of the nonallelic homologous recombination hotspot PRS3 associated with type-3 NF1 deletions. Hum Mutat 33: 372-383.

42. Nakajima T, Kaur G, Mehra N, Kimura A. (2008) HIV-1/AIDS susceptibility and copy number variation in CCL3L1, a gene encoding a natural ligand for HIV-1 co-receptor CCR5. Cytogenet Genome Res 123: 156-160.

43. Guedj M, Marisa L, de Reynies A, Orsetti B, Schiappa R, et al. (2012) A refined molecular taxonomy of breast cancer. Oncogene 31: 1196-1206.

44.Grube M, Reuther S, Meyer ZuSchwabedissen H, Kock K, Draber K, et al. (2007) Organic anion transporting polypeptide 2B1 and breast cancer resistance protein interact in the transepithelial transport of steroid sulfates in human placenta. Drug Metabolism & Disposition 35: 30-35.

45. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486: 395-399.

6. Discussion and conclusions

In this thesis, I evaluated potential contributions of germline common variants (SNPs, CNVs and CN-LOHs) for breast cancer predisposition and disease prognosis using well-established fundamentals of population genetics and GWAS paradigm. Analyses of these variants for their single-locus as well as epistatic effects have led to the following three primary conclusions:

1. A SNP (rs1429142) at chromosome 4q31.22 is a disease predisposition risk factor for sporadic breast cancer. The SNP-breast cancer association was stronger in premenopausal than postmenopausal women. To my knowledge, such a report for sporadic breast cancer in Caucasian population is first of its kind in Canada and third in the world.
2. SNP pairs (*APEX1*[95]-rs1130409 and *RPAP1*[96]-rs2297381; *MLH1*[97]-rs1799977 and *MDM2*[98]-rs769412) and SNP-SNP interactions involving four SNPs (*MLH1*-rs1799977, *MDM2*-rs769412, *BRCA2*[99]-rs1799943 and *MBD2*[100]-rs4041245) confer risk to breast cancer.
3. Multiple germline CNAs showed promising associations as potential prognostic factors for breast cancer.

These results have stressed the potential utility and continued search for germline common variants as breast cancer predisposition and prognostic factors, in addition to currently used tumor-based prognostic and predictive markers.

---

[95] APEX nuclease (multifunctional DNA repair enzyme) 1.
[96] RNA polymerase II associated protein 1.
[97] mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli).
[98] Mdm2, p53 E3 ubiquitin protein ligase homolog (mouse).
[99] breast cancer 2, early onset.
[100] methyl-CpG binding domain protein 2.

Moreover, small effect sizes observed for SNPs conferring breast cancer susceptibility, (ORs<1.5) either through single-locus effects or by epistatic interactions, further support the notion of polygenic nature of breast cancer.

Using a multi-stage association study design (combined sample size=7,219 cases and controls), I identified two breast cancer susceptibility loci at 4q31.22 and 5p15.2. When adjusted for BMI, SNP at 4q31.22 showed a strong statistical significane for overall breast cancer (P=1.5 x 10^{-7}) while achieved genome-wide level of statistical significance (P=6.2 x 10^{-10}) in premenopausal women. The SNP at 5p15.2 showed strong statistical significance for breast cancer (P=2.0 x 10^{-4}). These SNPs showed stronger associations for premenopausal than postmenopausal women and in cases with operable (I-IIIA) than non-operable tumor stages (IIIB, IIIC). In addition, I also successfully replicated common breast cancer susceptibility SNPs reported by international consortia between the years 2007 and 2009, suggesting the robustness of these associations. These results indicate that more common SNPs may be identified through systematic GWASs that may explain additional residual heritability of breast cancer.

In addition to single-locus effects of common SNPs, I also evaluated epistatic interactions among SNPs selected from a breast cancer GWAS and candidate gene-association studies as a potential source of inherited genetic contribution for breast cancer. I focused on SNPs in or close to cancer related pathway (DNA repair, metabolism and modification) genes as an exploratory attempt that showed reasonably consistent moderate single-locus effects with weak statistical significance in both discovery and replication stages in GWAS and/or candidate-

gene association studies. Epistatic interaction analyses led to the identification of two two-way SNP-SNP interactions (*APEX1*-rs1130409 and *RPAP1*-rs2297381; *MLH1*-rs1799977 and *MDM2*-rs769412) as well as an interaction involving four SNPs (*MLH1*-rs1799977, *MDM2*-rs769412, *BRCA2*[101]-rs1799943 and *MBD2*[102]-rs4041245) that conferred risk for breast cancer. Known biological interactions involving DNA-protein and protein-protein interactions in DNA repair process, in addition to epistatic interactions observed among multiple SNPs, indicate possible cross talk and convergence of multiple DNA repair pathways. These results also indicate that potential epistatic interaction analyses, in addition to the single-locus tests of association primarily adopted in GWASs and candidate-gene association studies, may explain larger proportion of heritability for breast cancer.

Identification of large number breast cancer predisposition factors (usually common SNPs), either with single-locus or with epistatic effects, could be of use for breast cancer risk assessment. In 2011, Wacholder *et al.* for the first time attempted to assess the disease risk conferred by common breast cancer susceptibility variants [1]. The authors demonstrated that inclusion of ten common breast cancer susceptibility variants identified through GWASs into the widely used Gail model moderately improved the performance of risk models for breast cancer from 58.0% to 61.8%, as measured by the area under the curve. While this scant improvement in risk assessment may not be sufficient for inclusion of common variants to identify women who might benefit from prophylactic intervention, it is possible that many more variants remain to be

[101] breast cancer 2, early onset.
[102] methyl-CpG binding domain protein 2.

identified that could eventually improve clinical risk assessment for breast cancer. More recently, Sawyer *et al.* evaluated associated familial breast cancer risk conferred by 22 common breast cancer susceptibility variants identified through multiple GWASs, using a polygenic risk score (PRS), calculated as the sum of the log odds ratio for each allele [2]. Using PRS for risk assessment, the 22 common variants could explain 18.5% of genetic risk for breast cancer while predictive power of PRS in non-*BRCA1/2* familial breast cancer cases was 65.4%, as measured by area under the curve. These results also indicated that PRS was significantly higher among individuals with familial breast cancer than in healthy controls (P=1.0 x 10^{-16}). Moreover, the PRS was significantly higher among familial cases without *BRCA1/2* mutations than cases with mutation carriers (P=2.3 x 10^{-6}). Women who tested negative for *BRCA1/2* mutations but higher PRS were more likely to have early-onset of disease before 30 years of age (OR=3.37) and higher chance of second breast cancer (OR=1.96) as compared to women with low PRS.

Presently, the current model of genetic testing for familial breast cancer only identifies *BRCA1/2* mutations in approximately one in five women. The test is uninformative for familial cases that test negative for *BRCA1/2* mutations. However, after addition of common variants (*i.e.*, PRS) into the current model of genetic testing, it may be now possible to subdivide non-*BRCA1/2* familial breast cancer cases into high, intermediate and low risk groups, as described by these investigators. Similar model of genetic testing, also taking into account genetic interactions (gene-gene and gene-environment interactions), can be attempted for

risk assessment of sporadic breast cancer when sufficient common variants will be identified through more GWASs and candidate-gene association studies aided by large international consortia.

In this work, I also identified common germline variants with potential prognostic values for breast cancer. SNP rs13281615 on chromosome 8q24.21 [3,4] was for the first time shown to be of prognostic value with breast cancer outcomes (RFS and OS) by independent investigators and I confirmed these findings. In GWAS literature, corroborative evidence of this kind is highly recommended in diverse ethnic groups. Even though the prognostic SNP replicated here in this study is also from Caucasian subjects, its replication in other ethnic groups is awaited. While the SNP profiled here is located in non-gene region, this locus warrants further investigation for mechanistic insights. In addition, I also identified multiple CNAs (two copy number gains and eight CN-LOHs) showing statistically significant differences in RFS probabilities in breast cancer cases with and without these CNAs. Of these, three CN-LOHs were validated by an independent platform, RT-qPCR in a proof of concept study. These findings indicate the importance of germline common variants as potential prognostic markers for breast cancer. More research is needed to identify additional germline variants of potential prognostic and predictive values. The most promising markers that show consistent statistically significant associations with breast cancer prognosis can be further evaluated in prospective clinical trials. If successful, these germline markers, in addition to currently utilized tumor-based prognostic and predictive factors, may help us realize practical value of

breast cancer prevention and control through applications of genetically stratified populations to benefit from emerging genomics medicine.

6.1 References

1. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, et al. (2010) Performance of common genetic variants in breast-cancer risk models. The New England Journal of Medicine 362: 986-993.

2. Sawyer S, Mitchell G, McKinley J, Chenevix-Trench G, Beesley J, et al. (2012) A role for common genomic variants in the assessment of familialbreast cancer. Journal of Clinical Oncology 30: 4330-4336.

3. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447: 1087-U7.

4. Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, et al. (2008) Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genetics 4: e1000054.

7. Future directions

1. This thesis work has explored several aspects of the genetic predisposition to breast cancer and disease prognosis and showed excellent replication in independent stages of the study. These results upon further validation would help generalize the findings (Chapters 2 and 5). Other aspects of the work that warrant further investigations are summarized below -
    a. The identified association of rs1429142 on chromosome 4q31.22 may be further investigated to identify causal variant(s) through either fine mapping or targeted deep sequencing within approximately 250 kb up and downstream of the locus.
    b. The association of this locus was observed to be influenced by BMI. Hence, the connection among BMI, rs1429142 and breast cancer requires further investigation.
2. I identified the statistical significant single-locus effect of rs1092913 on chromosome 5p15.2 in breast cancer (Chapter 2). The independent replication in a larger but distinct set of breast cancer cases and controls may be conducted to further evaluate its association with breast cancer.
3. I also provided a framework to evaluate risk conferred by potential epistatic interactions among SNPs showing moderate single-locus effects with weak statistical significance in GWAS and/or candidate-gene association studies for breast cancer (Chapters 3 and 4). Our analyses identified two two-way interactions and an interaction involving four SNPs conferring risk for breast cancer. These interactions warrant

replication in an independent set of cases and controls. Further, this approach can be extended to larger datasets as well as to other phenotypes.

4. In an exploratory analysis, I also identified germline CNAs with potential prognostic value for breast cancer. These markers showed stronger associations for luminal A type breast cancer than for others. Further investigation of these markers in independent breast cancer cases is warranted. A similar approach could also be adopted to identify germline CNAs with potential prognostic value for other types of breast cancer such as luminal B, HER2 and triple negative.
5. Finally, the comprehensive approach used in this thesis, which included SNPs, their potential interactions, and CNAs may be applied to other complex diseases or traits in order to identify germline DNA variations for susceptibility and disease prognosis.