University of Alberta

Association Mapping of Genes Using Whole Genome Polymorphism Arrays: Identification of Markers of Breast Cancer Susceptibility in Alberta Women

by

Malinee Chakravarthy Sridharan

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

Master of Science

in

Medical Sciences – Laboratory Medicine and Pathology

Edmonton, Alberta
Fall 2010

Examining Committee Page

Dr. Sambasivarao Damaraju, Laboratory Medicine and Pathology

Dr. Carol Cass, Oncology

Dr. John R. Mackey, Oncology

Dr. Yutaka Yasui, Public Health

Dedicated to

my husband

for his love, support and faith

Abstract

Breast cancer is a heterogeneous, polygenic disease and is influenced by genetic, environmental and life-style factors. Many single nucleotide polymorphisms (SNPs) associated with breast cancer risk have been identified in genome-wide association studies (GWASs) by several research groups for different populations. However, the variants identified so far contribute to a small proportion of disease risk. The objectives of the work described in this thesis were (i) to seek relevance/replicability of reported risk alleles from SNP scans to our study population; and (ii) to perform an independent GWAS for identification of additional/novel polymorphisms in the Albertan population. We approached these two end points by using cases and controls recruited in Alberta (total sample size, *n*=3064) in a two-stage association study (discovery study followed by replication study). We reproduced 14 of the 28 variants reported by others and also identified seven novel variants associated with breast cancer risk in our study population.

Acknowledgements

I thank my supervisor Prof. Sambasivarao Damaraju, for his excellent guidance. The discussions I had with him during my work over the last two and a half years shaped my research and brought the necessary depth and breadth to my thesis. Right from the initial phases, Prof. Damaraju was very encouraging in my attempts to gain mastery over the experimentation techniques, was keenly interested in the experimental outcomes and also gave his input on the progress of the experiments and ideas on improving the work regularly. Later, when I was pursuing graduate courses, he encouraged me to give my fullest attention to my courses and participated in discussions with me regarding specific research problems which arose as a part of my course work. In addition, he laid a lot of emphasis on learning the state-of-the-art in our field through an in-depth literature review, weekly lab meetings and journal club discussions at which we critically analysed a number of research papers related to our field of research. He also encouraged me to attend national and international research conferences where I have been able to present my work, attend presentations and engage in technical discussions with experts in my field. He was always available for discussions regarding my work and through his regular appraisal I have been able to effectively utilize my time. Overall, my graduate study with Dr. Damaraju has been very rewarding.

I am thankful to Dr. Carol Cass and Dr. John Mackey for serving as my supervisory committee members and giving positive and constructive input on the progress of my research work.

I am grateful to Dr. Badan Sehrawat for initiation into the area of 'whole genome' research and to my laboratory colleague Mr. Yadav Sapkota for stimulating discussions during lab meetings. I thank Dr. Paula Robson for providing control samples for association studies. I also thank Drs. Sunita Ghosh and Russ Greiner for discussions related to the statistical aspects of my research work. I thank the research staff, Ms. Jennifer Dufour and Ms. Diana Carandang, for technical help and Ms. Lillian Cook, Ms. Kathryn Calder, Mr. Adrian Driga and the rest of the tumour bank team for assistance and access with the clinical database. I would also like to thank Dr. Monika Keelan for serving as my examining committee chair. I would like to thank Ms. Cheryl Titus for guidance on administrative affairs, departmental procedures and regulations; and last but not the least Drs. Jonathan Martin and Jason Acker for providing all the needed help during the course of my research program.

I would like to thank Prof. Damaraju and the Department of Laboratory Medicine and Pathology in the University of Alberta for giving me an opportunity to pursue the Master of Science program. I would also like to thank the Alberta Cancer Foundation and Alberta Health Services for studentship support through operating grants to my supervisor.

I have been fortunate to pursue academic research at one of the best universities in the world. The high quality of biological and statistical courses helped me to start

my research with confidence. I thank Profs. Gary Eitzen, Andrew Simmonds, Giseon Heo, Carlos Fernandez-Patron, Charles Holmes, David Brindley, Luis Schang and Shairaz Baksh for their high technical standards and dedication to teaching. I thank my colleagues in the Laboratory Medicine and Pathology department for providing a stimulating academic environment.

I reserve special thanks to my friends, Akila, Aishwarya, Sirish and Anju for the wonderful moments we shared in Edmonton.

I am indebted to my husband, mother, father, mother-in-law, father-in-law, sister-in-law & family and sister & family for their love and support without which this work would not have been possible and for their faith and trust in me.

Table of contents

List of Tables

List of Figures

Abbreviations

ACB	Alberta Cancer Board
ACRI	Alberta Cancer Research Institute
ATM	Ataxia telangiectasia mutated
BRIP1	BRCA1 interacting protein 1
BRCA1	breast cancer susceptibility gene 1
BRCA2	breast cancer susceptibility gene 2
bp	base pairs
CBCF	Canadian Breast Cancer Foundation
CDCV	Common disease common variant
CEU	Utah residents with northern and western European ancestry from the CEPH collection
CGEMS	Cancer Genetic Markers of Susceptibility
CHEK2	checkpoint kinase 2
CNV	copy number variation
COL1A1	collagen, type 1, alpha 1
DNA	deoxyribonucleic acid
EDNRA	endothelin receptor type A
ER	oestrogen receptor
FGF	fibroblast growth factor
FDR	false discovery rate
FGFR2	fibroblast growth factor receptor 2

FHS	Framingham Heart Study
GWAS	genome-wide association study
HDAC	histone deacetylase
Her2	human epidermal growth factor receptor
JPT	Japanese in Tokyo
kb	kilobases
KRAB	Kruppel-associated box
HWE	Hardy–Weinberg equilibrium
LD	linkage disequilibrium
MAF	minor allele frequency
MALDI-ToF	matrix assisted laser desorption ionization – time of flight
MAPK	mitogen activated protein kinase
MAP3K1	mitogen activated protein kinase kinase kinase 1
NCBI	National Center for Biotechnology Information
Oct-1	octamer 1
OR	odds ratio
orf	open reading frame
PALB2	partner and localizer of *BRCA2*
PCR	polymerase chain reaction
PHKA	phosphorylase kinase, alpha 1
PI3K	phosphoinositide 3 kinase
PR	progesterone receptor
PTEN	phosphatase and tensin homolog

Q–Q	quantile–quantile
Ras	rat sarcoma
Runx2	runt-related transcription factor 2
SLC4A7	solute carrier family 4, sodium bicarbonate cotransporter, member 7
SNP	Single nucleotide polymorphisms
TNRC9	trinucleotide repeat containing 9
UTR	untranslated region
YRI	Yoruba in Ibadan
ZBRK1	zinc finger and BRCA1-interacting protein with a KRAB domain 1
ZNF577	zinc finger protein 577
ZNF365	zinc finger protein 365

1 Introduction

Breast cancer is a complex disease strongly influenced by genetic, environmental and life-style factors. There is substantial inter-individual variation in terms of age of onset and expression patterns of the disease phenotype. An extensive search for genetic and molecular factors underlying breast cancer yielded new insights and created new opportunities particularly in the post-genomic era. A major breakthrough in the genetics of breast cancer happened with the identification of high penetrance mutations in the breast cancer 1 and 2 (*BRCA1* and *BRCA2*) DNA repair genes increasing the risk of breast and ovarian cancers (1, 2). Affected individuals are mostly characterized by an early onset of the disease with multiple affected cases within the families. Subsequent research efforts identified mutations in certain other DNA repair genes such as ataxia telangiectasia mutated (*ATM*), partner and localizer of *BRCA2* (*PALB2*), phosphatase and tensin homolog (*PTEN*) and *p53* which contribute to a small proportion of hereditary breast cancers (3-6). These familial cancer genes were mostly identified from analyses of pedigrees of high-risk families. In aggregate the known familial cancer genes contribute to approximately 20% of the disease risk (7, 8). Intensive research efforts to identify *BRCA*-like genes have not been successful (9, 10).

The next question is to address the remaining missing information on the heritability of breast cancer. With the availability of the entire genome sequence from the Human Genome Project, it is now possible to study the influence of genetic variations across the genome that potentially contribute to the missing heritability and overall disease risks. Several genome-wide association studies (GWASs) using single nucleotide polymorphisms (SNPs) as markers have identified multiple potential novel susceptibility loci associated with breast cancer risk (8, 11-19). Candidate-gene studies have also proven successful in identifying low-penetrance variants associated with disease susceptibility (20, 21). Other potential novel targets are being identified by screening structural variations, mainly copy number aberrations (amplifications and/or deletions) in cases and controls (22). Gene expression and microRNA expression patterns are also being explored to identify potential markers associated with breast cancer predisposition (23, 24). Pathway-based approaches (*e.g.*, using genes in DNA repair pathways or other signal transduction pathways) are also being explored for a comprehensive understanding of the genetic architecture of breast cancer (25).

Several clinical characteristics of breast cancer, including receptor status, tumour grade, stage, and invasion status, have been considered as variables in the search for breast cancer predisposition as these clinicopathological markers also serve as prognostic or predictive markers (11, 16, 17, 26, 27). The well-established prognostic and predictive factors, mainly oestrogen and progesterone receptor status, are tumour-based markers and as such there is an express need to

identify markers at the level of germ-line DNA (constitutive DNA) to enable screening of populations. SNPs are considered for association studies since these are germ-line DNA markers suitable for the association studies in disease susceptibility as well as potential prognostic markers. Progress in DNA sequencing methods will further enhance the understanding of genetic alterations through mutational screens. The work summarized in this thesis specifically focused on screening the genome for SNPs in a GWAS that showed associations with breast cancer phenotype in Alberta women.

1.1 Genome-wide association study

Whole genome association studies (also referred to as GWASs) are used to compare the frequency of affected cases with that of unaffected controls for single base nucleotide alterations, commonly referred to as SNPs (**Figure 1.1**). These markers, which are evolutionarily conserved, high frequency alleles with a mean inter-SNP distance of 300 base pairs (bp) across the genome (28), offer clues about the gene region or loci associated with a particular condition when statistically significant differences are observed between cases and controls. Large coverage of the genome is accomplished by the use of commercially available high-density oligonucleotide microarrays which typically interrogate up to a million SNPs. A comprehensive statistical analysis with appropriate quality control metrics is needed to identify the few statistically significant genetic

variants/markers that may be associated with the disease susceptibility and such tools are now well evolved.

GWAS can be divided into four major steps: (i) careful selection of cases and controls from a population, or populations with similar ethnic background and from the same geographical region; (ii) isolation of DNA, genotyping and initial quality control measures to enrich the dataset; (iii) application of appropriate statistical tests to identify differences in allele frequencies between cases and controls; and (iv) finally, replication of the GWAS findings with independent cases and controls (29). The term replication defines studies wherein initial findings are reproduced in independent cohorts from the same geographical region. The term validation in the context of SNP association studies usually refers to reproducibility of findings from population-based cohorts distinct and not restricted to the same geographical region (nor ethnicity) relative to those of the initial findings (30).

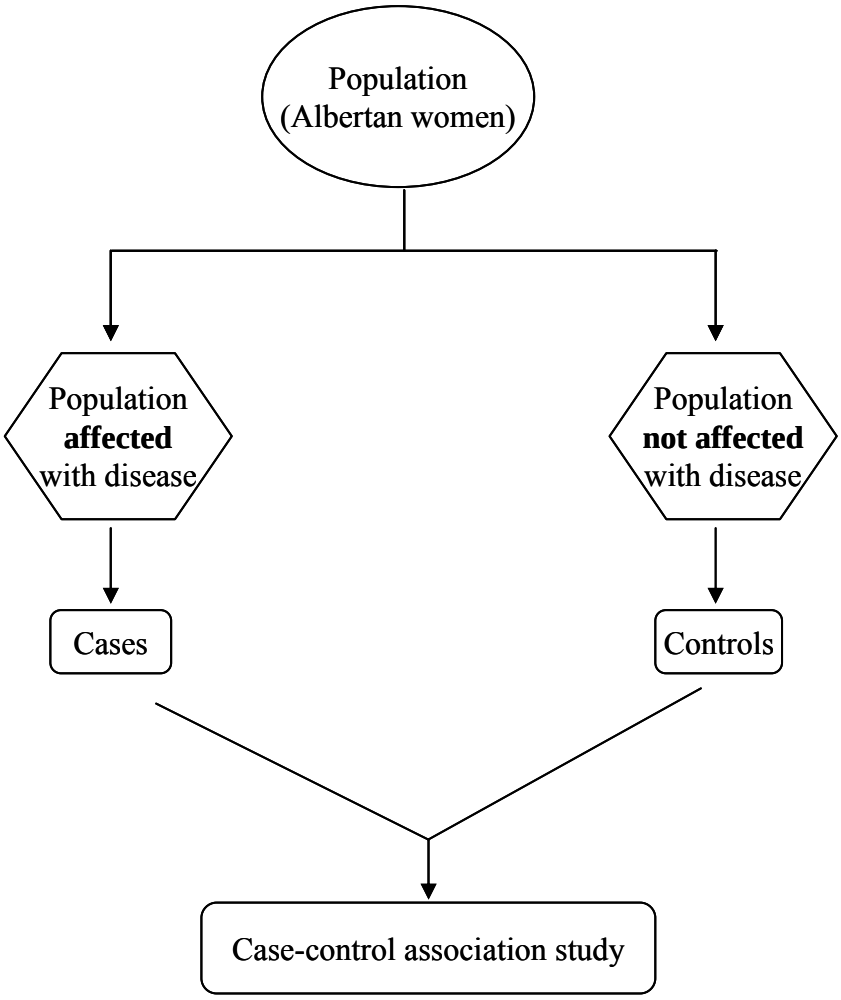


Figure 1.1: Flow chart depicting a case–control association study.

GWAS thus is an exploratory data generating method meant to narrow down putative candidate loci associated with disease/phenotype. Confirmatory evidence linking the association of a subset of loci is accomplished in the subsequent replication/and or validation cohorts. Genome-wide approaches are also free of bias for selection of markers as these are derived from all genomic regions of defined intervals (1 SNP/300 bp). There has been considerable progress in GWASs since 2005 and several loci have been identified to be associated with increased susceptibility to diseases, including breast cancer (12, 14), prostate

cancer (31, 32) and Alzheimer's disease (33). An updated version of all published GWASs is available in *A Catalog of published genome-wide association studies* (34) with emphasis on several (>100) complex diseases/phenotypes.

1.1.1 Assumptions of GWAS

(i) The key assumption of GWAS is that common, complex diseases are caused by common genetic variants. This is known as the 'common disease common variant' (CDCV) hypothesis. These genetic variants, which are common polymorphisms that are evolutionarily conserved, occur at a relatively high frequency in a population but exhibit relatively low penetrance conferring modest risk, accounting for only a small percentage of the disease risk (35). It is believed that highly deleterious variants undergo negative selection pressure and are eliminated during the course of evolution (36).

(ii) The second assumption in GWAS is that a SNP showing strong association may not be a causal variant but more often serves as a surrogate marker for the causal variant. As such, the surrogate marker, when in strong linkage disequilibrium (LD) with nearby SNPs, may help identify the causal variant. The genomes of humans and several other species are inherited in large blocks (haplotype blocks) during evolution and a subset of SNPs within a block could represent nearby SNPs (termed tagSNP). If tagSNPs are genotyped, one could potentially reduce the number of markers for genotyping across the genome. Genotyping fewer SNPs, which best represent the genome, improve statistical efficiency by reducing the number of association tests, in turn reducing the

number of false-positive associations at 5% level (28, 37). Several studies in GWAS have been reported using tagSNPs (11, 14, 16-18).

1.2 Need for genome-wide association studies

Mapping genes associated with disease is a commonly used method to elucidate the genetic basis of the phenotype of interest. Several approaches have been used to decipher the underlying molecular mechanisms for disease susceptibility. Traditionally, linkage analysis was used to investigate and identify the transmission of disease-causing genes from parents to offspring or even extended family members. Linkage analysis is performed with the assumption that the genetic loci or alleles are jointly inherited with the disease genes. Linkage studies have been successful in mapping monogenic disorders (*e.g.*, cystic fibrosis) exhibiting Mendelian patterns of inheritance. In general, these rare single-gene mutations are highly penetrant and deleterious, and in most instances are eliminated from the population due to negative selection pressures (36). Linkage studies have been successful in identifying genes such as *BRCA1* and *BRCA2* predisposing to certain complex diseases such as breast cancer (1, 2, 38). These genes are highly penetrant, characterized by early onset of breast cancer, and in most instances first and second degree relatives are affected. Since linkage studies predominantly address the mutations that segregate in families, the coverage of the finding is confined to a small subset of the affected population

(21). A severe limitation in linkage analysis is the compromised ability in identifying polygenic diseases involving common, multiple low-penetrance variants contributing to the disease phenotype (36).

Candidate gene association studies are an alternative to linkage analysis to address complex diseases. Similar to GWAS, a candidate gene association study can be used to study the genetic basis of complex diseases by determining the statistical correlation between the genetic variant and disease of interest. Selection of candidate genes for interrogation are based on *a priori* knowledge about their role in disease pathogenesis (39). For example, the selected genes may play a critical role in major cancer-related pathways such as DNA repair, apoptosis and signal transduction, thereby increasing the chance to identify the causal gene. This procedure increases the chances of introducing bias in selection strategies since one may not identify all putative candidate genes in the disease pathway/aetiology.

A GWAS is the best option when prior knowledge about the physical location and role of the causal gene/loci is unknown. It enables us to address the common variants scattered across the entire genome involved in common, complex diseases (36). Linkage studies focus on specific recombination patterns for a particular gene within a family and there will be limited recombination events (frequency) within the selected region (40). On the contrary, a GWAS accounts for meiotic recombination events across the entire genome at a

population level (41). Due to these advantages, a GWAS can provide a comprehensive understanding of the genetic aetiology of the disease.

1.3 Single nucleotide polymorphisms

As discussed earlier, a SNP is a type of genetic variation wherein a single nucleotide change on the DNA sequence serves as a bi-allelic marker (*e.g.*, C>T change) as opposed to short tandem repeats, insertions or deletions which are multi-allelic markers (**Figure 1.2**). Single nucleotide changes are classified as SNPs only when they occur at >1% allele frequency in a population. A SNP allele that is common in one population may be rarer in other populations. SNPs are stable and heritable. SNPs arise due to ancestral mutations and two unrelated individuals having a similar polymorphism pattern and allele frequencies are considered to belong to the "common evolutionary heritage" (42). It is believed that the frequency of a novel SNP originating in a particular generation is fairly low, approximately 10^{-8} per site per generation, which would give rise to approximately 30 new polymorphisms per haploid gamete (28, 43).

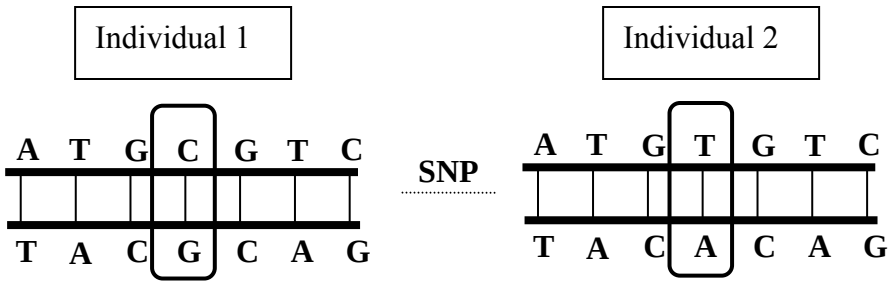


Figure 1.2: Diagrammatic representation of a SNP on a DNA double strand. Two individuals differ in a single nucleotide in a DNA segment.

Types and location of SNPs: SNPs can be grouped as: *synonymous* (no amino acid change) and *non-synonymous* (amino acid change). The majority of SNPs are likely to occur outside of exons, the gene-encoding regions (44, 45). Synonymous SNPs that occur within coding regions would not give rise to structural modifications of the encoded proteins, resulting in functional changes. However, because synonymous SNPs can potentially contribute to codon bias (may result in changes in relative protein abundance in the cells) and may also serve as surrogate markers for nearby causal alleles, they are useful in association studies. On the other hand, non-synonymous SNPs have a higher probability of affecting the structure or catalytic/binding activity of the proteins encoded by the genes carrying the SNPs. Once again, there are exceptions such as SNPs that result in conservative vs. non-conservative amino acid changes (***e.g.***, glycine to alanine vs. alanine to threonine). A vast majority of SNPs are found in **introns**, and although they do not alter encoded proteins, can serve as important markers. If present at a sufficient density on the chromosome(s), such SNPs will facilitate

fine mapping of the region of interest (12, 46). Alternatively, SNPs in introns can expose cryptic promoters. SNPs in **exons** can cause alterations (non-synonymous SNPs) in protein structure and function, leading to development of disease or can alter metabolism of a drug and hence response to therapy (47). SNPs at **intron/exon boundaries** can potentially alter the splicing events leading to alternative transcripts, unmask cryptic promoters and regulatory elements. SNPs in regulatory regions, such as **promoters, enhancers or non-coding regions at the 3' or 5'** ends of genes can affect binding of transcription factors altering translational efficiency and the relative abundance of encoded proteins. SNPs in **3' untranslated regions** (UTR) can affect the transcription/translation of mRNA to protein, affect mRNA stability and poly-adenylation signals (44, 48).

The entire SNP archive is available at a database known as (db)SNP, which is accessible at the National Center for Biotechnology Information (49, 50). The information included for each SNP is the following: (i) flanking sequence around the SNP; (ii) frequency of the SNP in a population; and (iii) experimental methods used to assay the SNP. Each SNP submitted is assigned a reference SNP accession ID (rs number) and the ones in the same physical location are given the same rs number.

In the post-genomic era, SNPs have been the markers of choice because of their relative abundance in the genome in comparison with other genetic variations. There is approximately 1 SNP for every 300 bp. Therefore, the 3.2

billion bp human genome harbours approximately 10 million SNPs (28). Several SNPs may serve as surrogates for a SNP in LD (co-segregation of certain SNPs due to LD), investigators are faced with this redundant information in genotyping projects. Assuming a ten-fold redundancy due to surrogate SNPs, screening a million tagSNPs could potentially eliminate this redundancy. The stability, heritability, random distribution across the entire genome, and lower chance of mutational events within a generation justify SNPs as an appropriate choice for a study of the genetic basis of disease over other genetic markers (*e.g.*, microsatellite, insertion/deletion or amplification polymorphisms).

Using SNPs as biomarkers, a gene association can be related to a phenotype of interest either by direct or indirect methods (**Figure 1.3**). In the direct method, detection of the causative SNP shows association and high statistical significance and confers measurable, though modest, risk. Typical examples are drug metabolism gene SNPs in the promoter regions affecting protein abundance or function (47). In the indirect method, the causative SNP is generally in LD with the marker locus. Therefore, the interrogated SNP acts as surrogate marker (through tagSNPs) and helps identify the putative genomic location of the causal SNP (36).

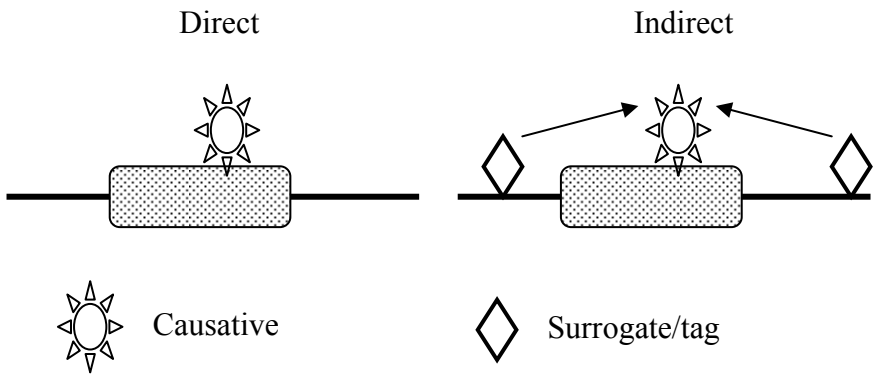


Figure 1.3: Direct and indirect methods to identify causative alleles (adapted from Hirschhorn and Daly (36)).

Mapping the location of causal or surrogate SNPs for a disease may enable screening of population susceptibility to disease. This information in turn may be useful to implement prophylactic measures if causality is determined. The polygenic nature of complex diseases addressed through SNP research could open new avenues for rationalizing therapies or identifying markers predictive of treatment outcomes.

1.4 Factors contributing to the success of GWAS

GWAS is a fairly new approach to identify low-penetrance variants associated with disease pathogenesis in a polygenic disease predisposition model. There are three main factors contributing to the success of GWAS.

(i) **Completion of the Human Genome Project:** The first advancement that enabled association studies was the completion of the Human Genome Project. The project commenced in 1990, the first working draft was submitted in 2000 and the completed version was published in 2003. Canada was the seventh country to participate in the sequencing project in 1992 (51). The key objectives of the Human Genome Project were to understand the genetic make-up of humans by identifying the sequential arrangement of base pairs along the length of DNA, identifying the physical location and putative functional properties of the genes, depositing the information in publicly accessible databases and developing software tools to mine the information (52). The sequenced genome is made up of approximately 3.2 billion base pairs with approximately 22,000 genes. The long-term vision of this project included identification of causal genes contributing to susceptibility to various diseases that would allow the design of drugs that would specifically target susceptibility genes, which, in turn, would enhance patient care. Deciphering the human genome sequence has made it possible to create a reference sequence against which all new sequence data will be interrogated/searched or aligned. Extensive research efforts are underway to understand the genetic basis of common, complex diseases at a molecular level.

(ii) **Human genome variations:** The second advancement was the identification of genetic variants across the genome. Ready and easy accessibility to the human genome sequence has provided the means to scan the genome to identify single nucleotide changes, particularly SNPs, in human populations

(53). From the human genome sequencing effort, it has become evident that there is remarkably high DNA sequence homology between unrelated individuals but the small proportion of variation contributes to the uniqueness of each individual such as susceptibility to diseases, varying response to drugs and treatment outcomes. In 2001, The SNP Consortium and International Human Genome Sequencing Consortium jointly published a map of 1.42 million SNPs spread across the entire genome with an average density of one SNP every 1.9 kilobases (kb) (54). Subsequently, The International HapMap Project was initiated in October 2002 to identify common haplotypes in different populations and also to identify tagSNPs. This project was a joint effort of researchers from Canada, United States, Nigeria, China and Japan. Initially, this consortium assessed 270 samples from four different populations: 30 trios from US Utah population (CEU) with Northern and Western ancestry (originally, samples were collected in 1980 by the Centre d'Etude du Polymorphisme Humain [CEPH]); 30 trios (mother, father and child) from the Yoruba (YRI) in Ibadan, Nigeria; 45 unrelated Japanese (JPT) in Tokyo, Japan; and 45 unrelated Han Chinese (CHB) in Beijing, China. Since Japanese and Han Chinese allele frequencies are nearly the same, some of the analyses considered them as a single population.

Phase I of the HapMap project aimed at genotyping approximately 1 million SNPs with a genetic distance of 1 SNP every 5 kb and a minor allele frequency (MAF) of $\geq$0.05. A description of Phase I HapMap project was published in 2005 (55). Phase II of the project attempted to genotype an

additional 4.4 million SNPs in the 270 samples but successfully genotyped only ~3.1 million SNPs and published in 2007 (56). The rest of the SNPs could not be genotyped, were monomorphic SNPs (SNPs with single form or allele, i.e., no heterozygous individuals in the population) or failed to pass the quality control metrics (43). The information and dataset are freely accessible and readily retrievable from the HapMap website (57). The patterns of DNA genetic variation identified through the HapMap project have been successfully used to identify putative loci for several common, complex diseases using GWAS.

Apart from SNPs that are being used as molecular genetic markers in mapping the loci/genes associated with complex diseases and for pharmacogenomic studies, other genetic variants that contribute to structural changes in the genome such as copy number variation (CNV) and microsatellites are also explored.

(iii) **Advancement in technology:** The third advancement that enabled the use of SNPs as genetic markers is the improvement in genotyping technology with a corresponding decrease in the cost of genotyping. Currently, there are two main competing commercial organisations offering whole genome arrays for SNP genotyping: Affymetrix and Illumina. They offer DNA microarrays that interrogate approximately 1 million SNPs and 1 million CNVs. Affymetrix′s Genome-wide Human SNP Array 6.0 features 906,600 SNPs and 946,000 CNV probes and Illumina's High Density Human 1M-Duo chip interrogates approximately 1.2 million polymorphic loci per sample. Despite the similarity

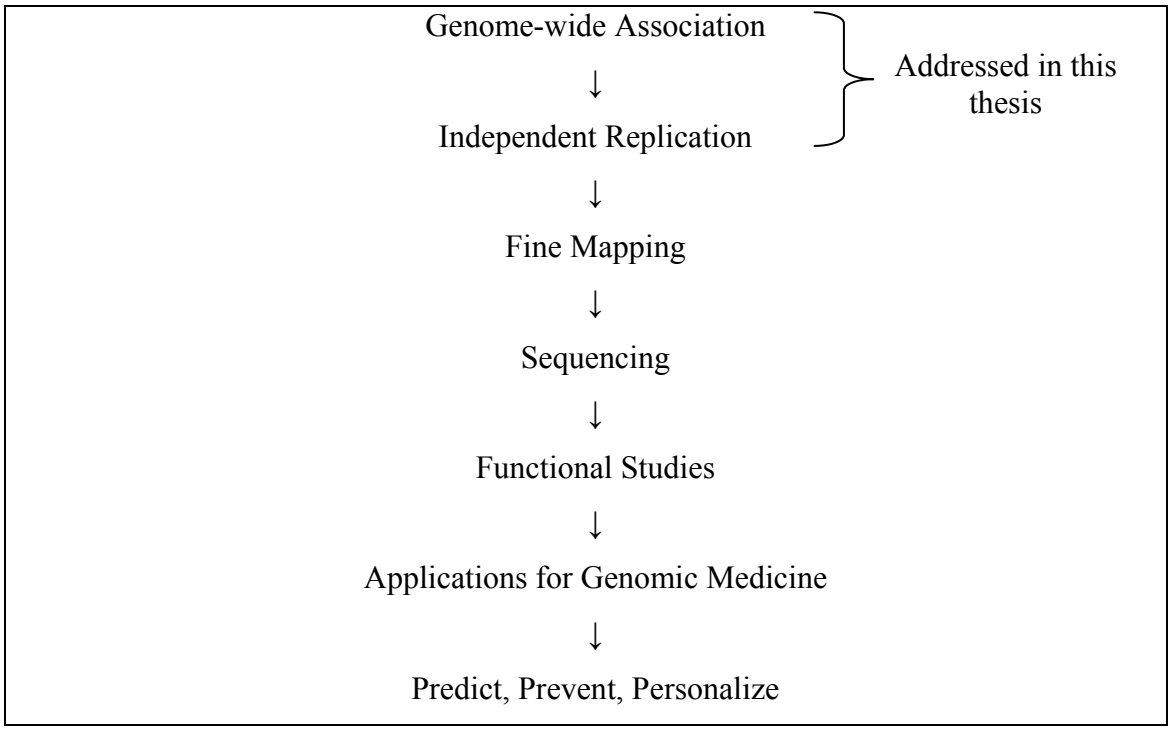
of approximately 1 million SNPs and CNVs each on both chips, there are substantial differences between the two platforms. Illumina arrays use 50-mer oligonucleotides (up to several replicates) whereas Affymetrix arrays use 25-mer oligonucleotides and each SNP is represented by four to six probes/replicate (totally, 6.8 million probes on the array). SNP selection strategies used by the two platforms are different. Illumina's probes predominantly include the tagSNPs (somewhat biased but eliminate redundancy of markers) identified by the International HapMap Consortium whereas half of the SNP probes included on the Affymetrix array 6.0 are tagSNPs, and the others are randomly distributed across the genome (determined by the presence of restriction enzyme cutting sites, *Nsp*I and *Sty*I) and include SNPs in recombination hotspots and the newly annotated SNPs from the (db)SNP database (58). Both technologies have advantages and disadvantages and the main challenge is in data mining and the statistical power (sample size and associated costs) needed to detect associations taking into account the genome level correction for multiple-hypothesis testing (59).

1.5 Road map to GWAS

The scheme shown in **Table 1.1** depicts the proposed GWAS workflow. One of my research objectives was to reproduce initial findings (Stage I, 348 cases and 348 controls) from a genome-wide scan conducted in our laboratory for breast cancer susceptibility (see **Appendix A**); I selected statistically significant

markers from Stage I and replicated them using an independent technology platform (Sequenom). I also utilized case–control cohorts distinct from the Stage I study from Alberta. In all, 1153 cases and 1215 controls were used for replication (Stage II, will be reviewed in detail in Chapter 4). The scheme in **Table 1.1** shows the general sequence of gene/loci discovery in the context of GWAS even though all of these end points are not relevant to my thesis objectives. However, the full scheme depicts the overall study premise in Dr. Damaraju's laboratory with the end points of biomarker discovery for cancer and pharmacogenomics. Fine mapping and sequencing are essential to identify the causal allele(s) that will ultimately facilitate building predictive models for disease risk and prognostication.

Table 1.1: Translation from whole genome association to genomic medicine



1.6 Breast cancer

The post-genomic era has opened up novel opportunities to approach a comprehensive understanding of the genetic basis of cancer and other complex human diseases. Cancer development is accompanied by multiple genetic changes which include single nucleotide changes, small or large DNA amplifications or deletions contributing to structural changes in DNA. Breast cancer is a heterogeneous disease and is influenced by genetic, environmental and life-style factors. This interplay has complicated the identification of genes contributing to increased susceptibility to breast cancer.

While the aetiology of breast cancer is variable and heterogeneous, the current understanding of this disease can be broadly divided into familial and sporadic breast cancers.

1.6.1 Familial breast cancer

One of the strongest risk factors is the inherited genetic (familial) component of breast cancer. It involves the inheritance of disease susceptibility genes within the family more often than is expected by chance. Several linkage studies have successfully mapped high-to-moderate penetrance genes associated with the disease (1, 2). Mutations or structural aberrations within such genes resulting in truncation of the encoded polypeptide or affecting protein structure and function result in a higher incidence of familial breast cancers. Individuals with a prior family history of breast cancer are more likely to develop the disease than individuals with no family history. Based on our current understanding, familial breast cancer genes can be grouped as those with high and moderate penetrance mutations.

1.6.1.1 High penetrance mutations

A major breakthrough in understanding the genetic basis of breast cancer occurred with the identification of a locus on chromosome 17q21, a region of a gene responsible for causing the disease in families (1). Subsequently, genetic

linkage analysis with multiple case breast cancer families led to the characterizing and cloning of the now highly celebrated gene, *BRCA1* (60). *BRCA1* accounted for only a small proportion of hereditary breast cancer susceptibility, leading to the conclusion that other, as yet unidentified, susceptibility genes must also exist. This optimism was soon rewarded by the discovery and characterization of a second breast cancer susceptibility locus *BRCA2* on chromosome 13q12–q13 (2).

Germ-line mutations in the *BRCA1* and *BRCA2* tumour-suppressor genes lead to early onset of breast cancer and often first and second degree relatives are affected. These genes also increase susceptibility to ovarian, prostate, and pancreatic cancers as well as male breast cancer. The *BRCA* genes are highly expressed during DNA replication. *BRCA1* and *BRCA2* interact with other DNA repair genes (*e.g.*, RAD51 homolog (RecA homolog, *E. coli*) (*S. cerevisiae*), *RAD51*) to mend the damage caused to DNA due to external (ionizing radiation) and/or internal (reactive oxygen species) agents/factors (61). Mutations or deletions in these genes can result in impaired DNA repair activity and compromised repair capacity of the cell leading to more mutations and genomic instability. The accumulation of damaged DNA products may lead to tumourigenesis. In aggregate, *BRCA1* and *BRCA2* account for approximately 15–20% of the familial breast cancer risk in several populations including European, Asian, African and Ashkenazi Jewish populations (62-65).

1.6.1.2 Moderate penetrance mutations

Some potential moderate penetrance genes have been identified. *ATM* gene, localized on chromosome 11q22–23, is known to confer the phenotype, ataxia telangiectasia (a neurodegenerative disorder) and plays an important role in the regulation of cell division and DNA repair. In 1999, it was shown that mis-sense variants in the *ATM* gene that do not cause ataxia telangiectasia increase susceptibility to breast cancer (66). Subsequent studies using multiple case breast cancer families with non-carriers of *BRCA* mutations indicated the involvement of *ATM* gene in breast cancer susceptibility with a modest risk effect on the disease pathogenesis (6).

CHEK2 (checkpoint kinase) is a protein kinase the gene for which is localized on chromosome 22q. It is known to interact with the *BRCA1*, *p53* and *ATM* genes to bring about cell cycle inhibition and function in DNA repair processes in response to DNA damage. CHEK2*100delC, a protein-truncating mutation in *CHEK2* that abrogates control of cell cycle leading to uncontrolled cell division, was found to segregate with breast cancer (67). *CHEK2* does not confer elevated risk of breast cancer in carriers of BRCA gene mutations. It was also shown that both ATM and CHEK2*100delC mutations do not occur together (6).

PALB2 localized on chromosome 16p12.2, is known to interact with *BRCA2*. The protein encoded by this gene is implicated in nuclear localization and stability and also assists in *BRCA2* functions such as homologous recombination and DNA repair processes. Mutation in these genes can directly or indirectly affect DNA repair processes and increases female breast cancer risk by approximately two fold (5).

Predisposing mutations in *BRIP1* (*BRCA1* interacting protein 1), *PTEN*, *p53* and other DNA repair pathway genes have also been implicated in breast cancer susceptibility (3, 4, 68). They show high-to-moderate penetrance and may influence the age at onset of the disease. A number of familial cancer genes with high-to-moderate penetrance mutations have been identified thus far, but they account for only a small proportion of disease risk. Residual genetic variance could possibly be resolved by addressing the alleles that are of low penetrance and confer finite risk.

1.6.2 Sporadic cancer

With the emerging consensus that breast cancer is polygenic in that several genes contribute to disease susceptibility with each polymorphism/allele conferring a finite risk (69), the search for additional variants has intensified. Several research studies have been undertaken to identify markers associated with breast cancer susceptibility. The pioneers in GWAS on breast cancer published

their findings in June 2007 (12). Subsequently, several independent research groups also identified additional loci showing highly statistically significant associations with disease susceptibility (8, 11, 13-19). The commonality in all these studies, except that of Murabito et al. (15), is the multi-stage study design. A multi-stage approach involves at least a two-stage study design, and the number of stages varies depending on the availability of resources and sample size. There are two main reasons to follow a multi-stage study design: (i) screening for approximately 1 million SNP for association may result in nearly 50,000 SNPs possibly showing false-positive association at 5% significance level (p <0.05), thus warranting further replication studies to identify true-positive findings; (ii) although there has been a sharp decline in the cost of genotyping over the years, it is still considered expensive to genotype a million SNPs across all stages. In the first stage, a comprehensive set of SNPs representative of the entire genome are genotyped in a fraction of the samples (hypothesis generation step); and subsequent stages ideally should involve replication of all markers from Stage I. However, due to cost constraints a subset of SNPs are often selected and genotyped for replication in Stage II in a similar or larger sample size (for additional details refer to Chapter 2, Section 2.2) (7). While there is similarity in the study design, there are differences in the number of SNPs genotyped, genotyping platforms, quality control metrics and statistical measures applied for data analysis. Given below is a brief summary of the findings of the nine major research articles in their quest for potential SNPs associated with breast cancer pathogenesis.

GWAS 1: Easton et al. (12) were the first to report five putative SNPs to be associated with breast cancer susceptibility using a three-stage study design. The samples included in the study were women affected with breast cancer and healthy controls from the United Kingdom, with a European ancestry. For Stage I, 266,722 SNPs were genotyped in 390 breast cancer cases all with a family history of the disease and 364 healthy controls using the custom-made microarrays offered by Perlegen Sciences. They selected 12,711 SNPs that were statistically significant from Stage I and genotyped them in 3990 cases and 3916 controls for Stage II. For the third stage, 30 SNPs were selected and genotyped using 5' nuclease assay offered by Taqman in 22,848 cases and 22,578 controls from 22 different cohorts. A case–control association analysis revealed five SNPs to be highly significant across the three stages with *p*-values $<10^{-7}$. The five SNPs identified were rs2981582 (fibroblast growth factor receptor 2, *FGFR2*), rs3803662 (trinucleotide repeat containing 9, *TNRC9/LOC643714*), rs889312 (mitogen activated protein kinase kinase kinase 1, *MAP3K1*), rs13281615 (8q), and rs3817198 (lymphocyte-specific protein 1, *LSP1*) (**Table 1.2**). Efforts were undertaken to identify the causal genetic variant in the *FGFR2* gene and *TNRC9/LOC643714* locus by fine mapping.

Fine mapping of FGFR2: Polymorphism rs2981582, localized in the intron 2 of the *FGFR2* gene on chromosome 10q, showed the strongest statistical significance in this study. rs2981582 lies within a 25-kb LD block and the entire block could be captured with six other tagging SNPs. None of the six SNPs

showed statistical evidence of association with the disease phenotype. Haplotype analysis of the seven SNPs (including rs2981582) indicated that multiple haplotypes harbouring minor alleles of rs2981582 conferred elevated risk for the disease, indicating its possible role as a causal variant or a surrogate marker. Resequencing of the SNPs in the region that was strongly correlated with rs2981582 uncovered another SNP rs7895676 that showed strong association with disease susceptibility. The strong correlation of several SNPs with rs2981582 confounded whether the effect of this particular SNP is causal or is closely correlated with the causal variant(s).

Fine mapping of TNRC9/LOC643714 locus: rs3803662 is a synonymous SNP of gene region *LOC643714* present 8 kb upstream of *TNRC9*. This SNP showed the strongest association with breast cancer. Other tagging SNPs spanning the coding region of *TNRC9* did not show any association with the disease. The strong correlation of SNPs rs17271951, rs1362548, rs3095604 and rs478422 with rs3803662 makes it difficult to exactly determine the causal variant(s).

GWAS 2: Hunter et al. (14) performed a two-stage GWAS with post-menopausal women of European ancestry. For Stage I, 528,173 SNPs were genotyped in 1145 breast cancer cases and 1142 controls using the Illumina HumanHap500 assay. For Stage II, six highly significant SNPs, two from the *FGFR2* gene region and four from other loci, selected from the preliminary analysis and two other SNPs that best define the *FGFR2* risk haplotype were

genotyped in an independent set of 1776 cases and 2072 controls from three different studies. Polymorphism rs1219648 retained high statistical significance in both stages and in joint analysis across all four studies. rs1219648 positioned in intron 2 of the *FGFR2* gene is known to be in strong LD with the previously identified SNP rs2981582 (12) with an r^2 (pairwise comparison between markers) close to 1.0. This SNP is also known to be in strong LD with other neighbouring SNPs in intron 2 of *FGFR2* with r^2 values of 1.0 with rs2420946, 0.97 with rs2981579 and 0.96 with rs11200014 in the HapMap CEU samples. These results indicated that SNPs in strong LD identified in independent studies/populations were detected in GWAS, confirming the primary findings and lending credibility to the approach.

The two independent studies described above have given an impetus to understand the functional aspects of the *FGFR2* gene and to identify the causative SNP(s). The *FGFR2* gene is already known to be over-expressed in breast cancer and is localized on chromosome 10q (70). It plays a critical role in mammary gland development and tumourigenesis in mice (71). Functional analysis of a haplotype of eight strongly linked SNPs revealed that the minor and major alleles of rs2981578 and rs7895676, respectively, tightly bind the Oct-1/Runx2 and C-EBPβ transcription factors leading to over-expression of the *FGFR2* gene (72). Runx2 forms a complex with the ubiquitous transcription factor Oct-1 that is known to play an important role in mammary gland-specific expression (73, 74). These functional studies have not completely explained the biological relevance

to the disease by pinpointing the causal variant but definitely are a pointer for future research.

GWAS 3: Stacey et al. (16) carried out a six-stage GWAS with cases and controls collected from Iceland, Sweden, Spain, Holland and the US multi-ethnic cohort. For Stage I, approximately 300,000 SNPs were genotyped in 1600 Icelandic breast cancer cases and 11,563 controls using the IlluminaHap300 platform. SNPs were selected for replication based on ranking the signals by *p*-values and the SNPs that represented the 10 best loci were chosen. There were five replication sets, which included a combined 4554 breast cancer cases and 17,577 controls. Two SNPs consistently retained statistical significance in all five replication sets: rs13387042 (2q35) and rs3803662 (16q12). This study revealed that individuals with minor allele for the polymorphisms shown associated with breast cancer are at greater risk in oestrogen-receptor positive breast cancers. There is no gene annotation available for the sequences flanking rs13387042. A previous GWAS also showed that rs3803662 is associated with the disease risk and is located in *LOC643714* gene region, 8 kb upstream of *TNRC9* (12). Loss of heterozygosity in 16q is a common event in breast cancer, leading to the speculation that the region might harbour tumour suppressor genes (75, 76). Also, the over-expression of *TNRC9* has been implicated in metastasis of breast cancer cells to bones (77). The above-mentioned evidence warrants exploration of functional and biochemical aspects of the gene.

GWAS 4: Stacey et al. (17) conducted yet another GWAS with samples collected from Iceland, Sweden, Spain, Holland, the US multi-ethnic cohort, Nigeria and Cancer Genetic Markers of Susceptibility (CGEMS) study, yielding a total of 6145 cases and 33,016 controls, of which 5028 cases and 32,090 controls were of European ancestry. Two SNPs on chromosome 5p12, rs4415084 and rs10941679, showed strong association and conferred increased risk for oestrogen receptor positive breast cancers than for oestrogen receptor negative breast cancers. They also examined the *FGFR2* locus, which showed a high statistical significance for rs1219648, which was in agreement with the previous findings (12, 14). rs1219648 conferred greater risk for oestrogen receptor positive tumours and no risk for oestrogen receptor negative tumours.

GWAS 5: Gold et al. (13) conducted a GWAS with Ashkenazi Jewish women, a genetically distinct population of Eastern European descent, affected with breast cancer and normal, healthy controls and replicated select markers in two additional stages. For Stage I (GWAS) of the study, 150,080 SNPs were genotyped in 249 familial cases and 299 controls using the Affymetrix 500K SNP array. For Stage II, 343 highly significant SNPs were selected from Stage I, along with 4 SNPs from the *FGFR2* region, and genotyped in 950 cases and 979 controls using the Illumina GoldenGate assay. For Stage III, a subset of SNPs that showed greater association was genotyped in 243 cases and 187 controls using the Affymetrix 500K SNP array. This study identified seven SNPs that showed association with breast cancer in three stages and in joint analysis (combining

genotype data of all stages and performing association analysis for the seven SNPs): rs6569479 (enoyl CoA hydratase domain containing 1, *ECHDC1*; RING finger protein 146, *RNF146*), rs7776136 (*ECHDC1*, *RNF146*), rs2180341 (*ECHDC1*, *RNF146*), rs6569480 (*ECHDC1*, *RNF146*), rs1078806 (*FGFR2*), rs3012642 (phosphorylase kinase, alpha 1, *PHKA1*; histone deacetylase 8, *HDAC8*; not significant after joint analysis), and rs7203563 (ataxin-2-binding protein 1, *A2BP1*). They observed a strong and consistent association across all stages with the *RNF146*/*ECHDC1* region at 6q22. The protein encoded by *ECHDC1* plays a critical role in mitochondrial fatty acid oxidation and *RNF146* encodes an ubiquitin protein ligase. These genes are involved in pathways in breast cancer pathogenesis. Although the *FGFR2* locus rs1078806 retained marginal significance in Ashkenazi Jews, it was found in strong LD with the *FGFR2* SNPs reported by Easton et al. (12) and Hunter et al. (14) (rs2981582, rs1219648, rs2420946 and rs2981579).

GWAS 6: Most of the studies presented thus far have reported association analyses from samples of women with European ancestry. Genetic architecture can vary greatly with different ancestry groups. Zheng et al. (19) conducted a three-stage GWAS among Chinese women using samples obtained from the Shanghai Breast Cancer Study and Shanghai Breast Cancer Survival Study. This is one of the large-scale studies to report association analysis results for women with non-European ancestry. For Stage I, 906,602 SNPs were genotyped in 1505 cases and 1522 controls. This study is the first to interrogate approximately 1

million SNPs for Stage I using Affymetrix Genome-wide Human SNP array 6.0. For Stage II, 29 SNPs selected from preliminary analysis were genotyped in an independent sample set with 1554 cases and 1576 controls using Sequenom Mass-ARRAY iPlex technology. For fast-track replication, markers were selected which had (i) MAF of ≥10%; (ii) distinct genotype clusters; (iii) not been previously reported in other studies; (iv) $p \leq 0.01$ for all SNPs; and (v) best SNP with lowest *p*-trend and r^2 values of ≥0.8. For Stage III, four SNPs showing promising associations were genotyped in yet another independent sample set of 3472 cases and 900 controls. They identified a putative SNP associated with breast cancer risk, rs2046210 at chromosome 6q25.1. This SNP was also evaluated for its association with breast cancer among 1591 cases and 1466 controls of European ancestry. Consistent with the findings for the Chinese women, the minor allele showed an increased risk of breast cancer and the association was stronger in post-menopausal than in pre-menopausal women. rs2046210 lies upstream of the *ESR1* gene, which encodes oestrogen receptor α, which is known to play a critical role in hormone binding, signal transduction, DNA binding, and activation of transcription. This study also validated previously identified polymorphisms from Easton et al. (12) and Gold et al. (13) in their Stage I samples and showed rs1219648, rs2981582 and rs3803662 to be associated with breast cancer risk in Chinese population.

GWAS 7: Thomas et al. (18) carried out a three-stage GWAS using the samples obtained from the CGEMS project. For Stage I, 528,173 SNPs were

genotyped in 1145 breast cancer cases and 1142 controls using the Illumina HumanHap500 assay. For Stage II, 24,909 highly significant SNPs from Stage I were selected and genotyped in 4547 cases and 4434 controls using a custom-made Illumina chip. Additional SNPs were selected and genotyped to monitor population stratification and candidate genes/loci identified in previous studies, including the *FGFR2* region polymorphisms. For Stage III, 21 SNPs were genotyped in 4078 cases and 5223 controls using the Taqman assay. They confirmed strong association signals for six loci (2q35, 5p12, 5q11.2, 8q24, 10q26, and 16q12.1) previously reported to be associated with predisposition to breast cancer (12, 14, 16, 17). Two novel susceptibility loci that reached high statistical significance were identified at chromosomes 1p11.2 (rs11249433) and 14q24.1 (rs999737). rs11249433 is located in the pericentromeric region of the chromosome at which there is minimum possibility of recombination events. Hence, the SNP is expected to be representative of a large LD block. Distal to the SNP is the NOTCH2 promoter known to play a crucial role in cellular signalling. The second SNP rs999737 is located in the *RAD51L1* gene (RAD51-like 1 (*S. cerevisiae*)), which has an important role in double-strand DNA break repair and homologous recombination. RAD51 is also known to interact directly with *BRCA2* gene and indirectly with the *BRCA1* gene to bring about the above-mentioned functions (61).

GWAS 8: The study by Ahmed et al. (11) was an extension of the previous work by Easton et al. (12) with the objective of identifying additional

susceptibility loci associated with risk of breast cancer. They selected 925 statistically significant SNPs from the first two stages of the prior study. For Stage III of this study, these SNPs were genotyped in independent set of samples with 3878 cases and 3928 controls using a custom Illumina iSelect array. Joint analysis of the current data and previous GWAS data yielded three SNPs, rs4973768, rs4132417 and rs6504950, to be significant at $p < 10^{-5}$. For Stage IV, three SNPs were evaluated in 33,134 cases and 36,141 controls obtained from different studies. Samples from 27 study cohorts were genotyped using Taqman and five study cohorts using Sequenom Mass-ARRAY iPlex technology. Two novel susceptibility loci that reached high statistical significance were identified at chromosomes 3p (rs4973768) and 17q (rs6504950). The genomic region flanking rs4973768 contains two genes, *NEK10* and *SLC4A7*. *NEK10* encodes a kinase involved in cell cycle control and *SLC4A7* encodes a tyrosine kinase substrate known to regulate cellular pH and the activity of the gene is down-regulated in breast cancer cells (78). rs6504950 lies in intron 1 of STXBP4 (syntaxin binding protein 4) known to play a role in glucose transport and GLUT4 vesicle translocation (79).

GWAS 9: Murabito et al. (15) reported results for two cancers, breast cancer and prostate cancer, in 1335 participants from 330 families. The samples were obtained from the Framingham Heart Study (FHS) which offers the advantage of being a family-based association study. FHS cohort is a longitudinal study with extensive information collected on the participating subjects for health

and life-style risk factors and history of diseases including cancer in the families. The samples were genotyped using the Affymetrix 100K SNP array. SNPs selected for the association test were from the candidate genes previously implicated in breast cancer susceptibility. Five novel breast cancer susceptibility loci were identified from different genomic locations – rs2075555, rs6556756, rs1154865, rs1978503, and rs1926657.

Table 1.2: Summary of GWAS-identified variants

S. No	Reference	Population	SNPs identified	Associated genes	Sample size (Cases/Controls)		
					Stage 1	**Stage 2**	**Stage 3**
1)	Easton et al., 2007	European	rs2981582	*FGFR2*	390/ 364	3990/ 3916	21,860/ 22,578
			rs3803662	*TNRC9*/*LOC643714*			
			rs889312	*MAP3K1*			
			rs13281615	8q			
			rs3817198	*LSP1*			
2)	Stacey et al., 2008	European	rs4415084	5p12	6145/33,016 (all stages, >3 stages)		
			rs10941679				
3)	Stacey et al., 2007	European	rs13387042	2q35	4554/17,577 (all stages, >3 stages)		
		European	rs3803662	*TNRC9*/*LOC643714*			
4)	Zheng et al., 2009	Chinese, European	rs2046210	6q25.1	1505/1522	1554/1576	3472/ 900
5)	Thomas et al., 2008	European	rs11249433	1p11.2	1145/1142	4547/4434	4078/ 5223
			rs999737	*RAD51L1*			
			rs7716600	*MRPS30*			
			rs2067980				
6)	Gold et al., 2008	Ashkenazi Jews	rs6569479	*ECHDC1*; *RNF146*	249/ 299	950/ 979	243/187
			rs7776136				
			rs2180341				
			rs6569480				
			rs1078806	*FGFR2*			
			rs3012642	*PHKA1*; *HDAC8*			
			rs7203563	*ALG1*			
7)	Hunter et al., 2007	European	rs1219648	*FGFR2*	1145/1142	1776/2072	
			rs2420946				
			rs11200014				
			rs2981579				
			rs17157903	*RELN*			
			rs7696175	*TLR1* *TLR6*			
8)	Ahmed et al., 2009	European, Korean, Taiwan	rs4973768	*SLC4A7*	390/ 364	3990/3916	3878/ 3928*
			rs6504950	*STXBP4*			
9)	Murabito et al., 2007		rs2075555	*COL1A1*	1335 participants**		
			rs6556756				
			rs1154865	–			
			rs1978503	*FLJ45743*			
			rs1926657	*ABCC4*			

*This study also included Stage IV with 33,134 cases and 36,141 controls.

**A family-based association study using participants from 330 families.

1.7 Organisation of the thesis

The primary objective of the thesis was to investigate the relevance of polymorphisms in breast cancer susceptibility recently described from GWAS by validating in a study population in a case–control setting.

The secondary objective of the thesis was to reproduce initial findings (Stage I, Affymetrix data) from the genome-wide scan for breast cancer susceptibility using an independent technology platform (Sequenom) for a select set of informative, statistically significant markers in additional independent cohorts (Stage II). Stage I of the GWAS generated the hypothesis that the polymorphisms found associated confer breast cancer susceptibility and that breast cancer is a polygenic disease. Biological and functional relevance of the putative loci was beyond the scope of this thesis.

Currently, global efforts at GWAS are limited to select laboratories (total of 11 groups including the study from Dr. Damaraju's laboratory). Taking into account the current research status in GWAS, we are the only group within Canada to perform a whole genome study addressing the breast cancer phenotype using an association study design.

1.7.1 Methods

The second chapter deals with the experimental study design, description of case and control samples selected for the association study to generate hypothesis, generation of genotype data and data filtering criteria. We also describe genotyping platform, number of SNPs genotyped, software used for analysis, subjecting data to different quality control measures, statistics used for association analysis and the rationale behind selecting markers for replication studies.

1.7.2 Other chapters

Chapter 3 summarizes the results of our investigation of the relevance of SNPs from GWAS reported between 2007 and 2009 to our study population in a case–control setting. An allelic association analysis and a subgroup analysis based on the receptor status were performed using chi-square test for 28 SNPs to determine the allelic frequency differences between breast cancer cases and normal, healthy controls. We confirmed that associations with breast cancer risk were similar to those reported in the literature (Caucasian population).

Chapter 4 mainly focuses on the results obtained from the replication phase (Stage II). We also performed a joint analysis by pooling all the samples from Stages I and II and conducted an association analysis using chi-square test.

The hypothesis for Stage II was generated from the association analysis performed in Stage I, which is summarized in **Appendix A**. In the Appendix, we also describe the process of selection of quality control metrics for genotype filtering by applying varied filters including detection of population stratification in our study population by comparing with the HapMap samples (used as a reference). An allelic association analysis using chi-square test was performed to determine the total number of SNPs that show statistically significant association to the disease pathogenesis. Finally, a subset of markers was chosen for replication in Stage II in an independent sample set.

Chapter 5 briefly summarizes the conclusions based on the results reported in the thesis. This chapter also addresses the future work that is possible by making use of the readily available data generated in this study period.

2 Materials, methods and statistics

2.1 Experimental design

Many factors were taken into consideration while designing the experiments described in the ensuing sections. Particular emphasis was paid to different experimental designs discussed in the literature for association studies and whole genome scans using polymorphisms, including their strengths and limitations. Availability of resources, both DNA samples and their associated clinical characteristics from subjects, as well as technology platform strengths and analytical methods currently available were also of paramount importance.

2.2 Multi-stage association study design

A multi-stage approach is the widely accepted experimental design for a GWAS. In general, identification of putative loci is carried out in two or more stages. The work described in this thesis was divided into two stages. In *Stage I*, the *Discovery phase*, a large set of SNPs scattered across the entire genome selected without bias were genotyped in a limited number of breast cancer cases and controls (**Figure 2.1**). An association analysis was performed to identify a small proportion of SNPs showing statistically significant association with disease risk at a nominal *p*-value threshold of <0.05 (36). Stage I was the hypothesis-

generating step in which select markers that reached nominal statistical significance (p <0.05) were prioritised for replication based on stringent selection criteria.

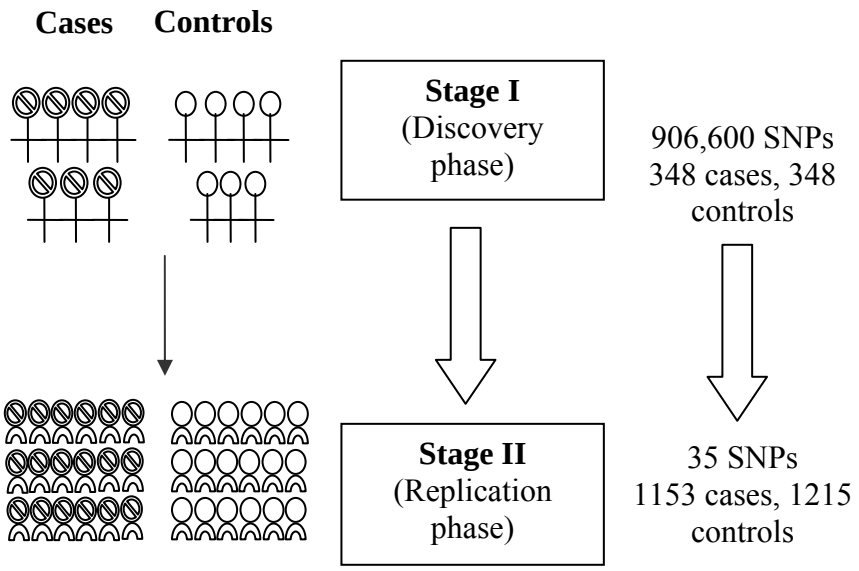


Figure 2.1: Multi-stage study design for GWAS. Flow chart on the far right indicates the number of SNPs interrogated in each stage along with the sample size for cases and controls.

In *Stage II*, the *Replication phase*, SNPs selected from the preliminary analysis using independent set of cases and controls were investigated further (**Figure 2.1**). Replication of the initial findings is an important component of GWAS because a significant fraction of the SNPs that reached statistical significance could potentially be false-positive signals. Therefore, it is considered essential to replicate the findings in multiple stages to eliminate the possibility of association of markers by chance (7, 80, 81). In spite of a sharp decline in the cost of whole genome genotyping over the past few years, replicating the findings in a large cohort for a comprehensive set of SNPs is still not economical. Hence, in the work described in this thesis a subset of statistically significant SNPs selected from the initial screen were re-evaluated in a larger set of independent cases and controls recruited from the same geographical region. Due to the modest effect of the polymorphisms on disease susceptibility, a large sample size is crucial in the replication phase to find true association signals (36, 80).

2.3 Study population

Case and control samples were obtained from women in Alberta, Canada. Informed consent was obtained from each participant included in the study and this research study was approved by the Institutional Ethics Board.

2.3.1 Cases

Germ-line DNA from women with breast cancer were obtained from the PolyomX study cohort and associated clinical characteristics from the database housed within Canadian Breast Cancer Foundation (CBCF) tumour bank located at the Cross Cancer Institute, Edmonton, Alberta. The PolyomX program, which was started in 2001 with funding from Alberta Health and Wellness and the Alberta Cancer Foundation, was a multidisciplinary collaborative research initiative in cancer genomics with the aim of identifying and characterizing biomarkers of value in diagnostication, prognostication and prediction. CBCF tumour bank was launched in 2005 as a provincial project to collect, annotate and bank tumour specimens along with buffy coat and serum. The CBCF tumour bank operates from two sites within the province of Alberta, one in Edmonton and the other in Calgary. Detailed clinicopathological characteristics from PolyomX cohort were also entered for each patient in an in-house built relational database called DORA (Database for Online Retrieval and Analysis) housed within the CBCF tumour bank

Cases had histologically confirmed invasive breast cancer, predominantly ductal carcinoma, non-metastatic at presentation, and with a median age at diagnosis of 53 years; central laboratory testing of receptor status was performed in a single institution (Cross Cancer Institute, Edmonton, Alberta).

Buffy coat (white blood cell enriched fraction of blood) or total blood samples were obtained from each participating subject, from which the DNA was isolated (tumour DNA was not used in the studies reported here). DNA from lymphocytes is also referred to as constitutive DNA or germ-line DNA in literature (82). Germinal cells however, typically are haploid genomes, and for association studies, diploid genomes (lymphocyte DNA as a surrogate for germinal cells) are used routinely to represent maternal and paternal chromosomes. Genomic DNA isolation from the buffy coat or blood samples was done using commercially available Qiagen™ (Mississauga, Ontario, Canada) DNA isolation kits for both cases and controls (QiAamp DNA blood mini kit [250 reactions], catalog no: 51106). Briefly, the protocol recommended by Qiagen is described below. Samples typically comprised 200 μL of whole blood or buffy coat stored since the time of collection at -80°C. Frozen samples were first warmed to room temperature and the cells were treated with proteinase K to remove/digest extraneous protein first (a result of fractionation of blood). To obtain RNA-free genomic DNA, the cells were treated with RNase A. Then, Buffer AL (lysis buffer) was added to the cells, mixed into a homogeneous solution, and incubated at 56°C for 10 min to bring about efficient lysis of the cells and to allow proteinase K to act on cellular proteins. The lysate was then centrifuged to bring down the liquid present inside of the lid. Then the lysed cells were treated with 96–100% ethanol to precipitate the DNA from other cell components and centrifuged to remove drops from inside of the lid. The mixture was then transferred to the QIAamp spin column provided along with the kit. The

column was placed in a 2 ml collection tube and the mixture was centrifuged at 6000× *g* (8000 rpm) for 1 min. The filtrate in the collection tube was discarded. The resulting precipitate was first treated with Buffer AW1 (wash buffer), centrifuged at 6000× *g* (8000 rpm) for 1 min and the filtrate in the collection tube was discarded. Then, the precipitate was treated with Buffer AW2 (wash buffer), centrifuged at 20,000× g (14,000 rpm) for 3 min and the filtrate in the collection tube was discarded. Treatment with wash buffer removes any residual unlysed cells or contaminants. Finally, elution of DNA was done using Buffer AE (elution buffer) or distilled water, incubated at room temperature for 1 min and then centrifuged at 6000× g (8000 rpm) for 1 min. Simultaneous processing of multiple samples was possible and yielded pure DNA amenable for downstream processing (83). The concentration of the isolated DNA was determined at A_{260}/A_{280} nm using NanoDrop 2000 spectrophotometer (Wilmington, Delaware, USA) and stored at -20°C.

A total of 348 cases in Stage I were genotyped using the Affymetrix Human Genome-wide SNP array 6.0. A total of 1153 cases in Stage II were genotyped using Sequenom Mass-ARRAY iPlex technology.

2.3.2 Controls

Control subjects were matched for age based on frequency and matched for gender with that of cases. They were healthy women from the Edmonton and

Calgary regions of Alberta who were free of cancer at the time of recruitment for this study (during the years 2003–2009). These subjects were obtained through collaboration with the Tomorrow Project (84). The Tomorrow Project is a longitudinal study with molecular epidemiological end points in which extensive health and life-style data is being gathered from study participants to investigate genetic and environmental variables on cancer susceptibility.

The immediate objective of the Alberta Tomorrow Project is to recruit 50,000 healthy individuals between the ages of 35 to 69. Currently, nearly 30,000 participants have joined the study since 2001 and new participants continue to be recruited (85). Informed consent was obtained from each participant for inclusion in the study.

A total of 348 control samples were included in Stage I with a full record of self-declared ethnicity information (predominantly of Caucasian origin) and with no documented prior history of breast cancer in first and second degree relatives. Stage I samples were genotyped using the Affymetrix SNP 6.0 array. A total of 1215 controls were accessed in Stage II from the Tomorrow Project and were genotyped using Sequenom Mass-ARRAY iPlex technology. Procedures for sample collection and DNA isolation were the same as described above for cases.

Our Stage I cohort used cases and controls with no family history of breast cancer in the first and second degree relatives. This was in contrast with the

GWAS from literature wherein the cases selected were stratified for positive family history of breast cancer (first and second generations) or menopausal status among the Caucasian populations (12, 14). Since earlier studies addressed the selection of cases based on prior family history, our interest is to study the sporadic cases of breast cancer.

2.4 Genotyping

2.4.1 General protocol for SNP genotyping using Affymetrix

A genome-wide scan of cases and controls (*n*=348, each) was performed using the Affymetrix Genome-wide Human SNP Array 6.0 featuring 906,600 SNPs with each probe represented four to six times in an array. Each of the samples was subjected to a standardised experimental protocol designed by Affymetrix. The protocol consisted of the following critical steps.

(i) *Restriction digestion*: Total genomic DNA (250 ng/5 µl) was digested with *Nsp*I and *Sty*I restriction enzymes. These restriction enzymes recognise specific nucleotide sequences on the DNA and make double-stranded cuts (**Table 2.1**). Restriction digestion generates DNA fragments with single-stranded extensions on the strand that facilitate ligation of complementary sequences. These are also known as cohesive or sticky ends (**Table 2.1**).

(ii) *Ligation*: The process of joining the complementary sequences of DNA strands is termed as ligation. *Nsp*I and *Sty*I restriction enzymes digests were annealed to the *Nsp*I and *Sty*I adaptors, respectively, that specifically recognise the 4 bp overhangs (cohesive ends), with DNA ligase. Regardless of the size of the restriction enzyme digests, all fragments are substrates for adaptor ligation.

(iii) *Polymerase chain reaction* (*PCR*): A generic primer (supplied by the manufacturer) that recognises the adaptor sequences of the ligated DNA was used to amplify the DNA fragments in triplicates or quadruplicates to ensure adequate yield of the target DNA for subsequent analysis. PCR conditions were adjusted to selectively amplify fragments ranging in size from 200 to 1100 bp. The PCR conditions adopted were the same as suggested by the manufacturer (Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0 User Guide).

(iv) *Purification*: Amplified DNA from each restriction enzyme digest was pooled from replicates and purified using magnetic beads (Agencourt AMPure, Beverly, MA, USA). Magnetic beads help in recovering large amplicons, greater than 100 bp and unincorporated dNTPs, primers, primer dimers, salts and other contaminants can be efficiently separated in this process. Purified PCR products were stored at 4°C until further analysis.

(v) *Fragmentation and labelling*: Purified DNA was fragmented with DNAseI enzyme to approximately 50 bp. Subsequently, a biotin-labelled reagent was used for end-labelling fragmented PCR amplicons, with terminal

deoxynucleotidyl transferase. End labelling reactions were carried out at 4°C as described by the manufacturer.

(vi) *Hybridisation and Scanning*: Hybridisation is a process in which the target (end-labelled DNA product) interacts with the probe sequences embedded on the array. The recommended condition to carry out hybridisation was at 49°C for approximately 16–18 hours in an oven with 60 revolutions per minute. Hybridisation mix contains denaturing agents to generate single stranded DNA to bind to the single stranded probe sequences on the array. Mixing minimises background binding and results in homogeneous hybridisation. The hybridised arrays were washed with buffers rigorously to remove non-specific hybridisation (target molecules bound to a wrong probe) (86). The array was then stained with streptavidin–phycoerythrin conjugate that binds with high affinity to the biotin-labelled target molecules. The washing and staining procedures were carried out using the Affymetrix fluidics station 450 (Santa Clara, California, USA). Finally, the array was scanned using the GeneChip® Scanner 3000 7G (Affymetrix, Santa Clara, California, USA). This scanner has the ability to scan smaller features ranging in size from 2.5 μm to 0.51 μm. The raw optical images called the *.dat* files are generated. The pixel values were used to calculate the signal intensities for each feature (every SNP) and these were automatically stored as *.cel* files (58).

Table 2.1: Recognition sequences of *Nsp*I and *Sty*I restriction enzymes before and after digestion

*Nsp*I recognition sequence	*Sty*I recognition sequence
Before cut 5′ * * *RCATGY* * * * * *YGTACR* * * 5′	Before cut 5′ * * *CCWWGG* * * * * *GGWWCC* * * 5′
After cut (cohesive ends) 5′ * * *RCATG Y* * * * * *Y GTACR* * * 5′	After cut (cohesive ends) 5′ * * *C CWWGG* * * * * *GGWWC C* * * 5′

A: Adenine; T: Thymine; C: Cytosine; G: Guanine; R: A/G; Y: C/T; W: A/T.

A detailed protocol and the sources of the reagents used in genotyping are available in Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0 User Guide (Affymetrix, Santa Clara, California).

2.5 Data acquisition – Genotype calling

Genotype calling is the process of assigning one of three possible genotypes at corresponding alleles to a specific SNP. All the SNPs interrogated are bi-allelic. For any SNP with A and B alleles, there are three possible genotypes: homozygous (AA, BB) or heterozygous (AB). In this procedure, genotype information for each SNP is generated from the raw intensity files (*.cel* files) available after scanning each array. The signal intensities for each SNP are

measured to categorise a SNP to a particular allele based on the signal strength. If the intensity of one is high and the other low, then it is homozygous and if intensities of both alleles are equally high, then it is heterozygous (87).

The Birdseed v2 algorithm was developed by researchers at the BROAD Institute of Harvard and MIT specifically for the Affymetrix genome-wide human SNP 6.0 array. Birdseed v2 is a multi-chip algorithm used to assign genotype calls. In addition to genotype calling, the algorithm also performs data normalisation to eliminate any probe-specific effects to increase precision. The algorithm performs efficiently with sample sizes of 50 or more (88). The library files available at the Affymetrix website are used as a reference to make the genotype calls (58).

A total of 348 breast cancer cases and 348 healthy controls were genotyped using the Affymetrix genome-wide human SNP 6.0 array. Genotype calling was done in eight batches with 96 randomly-chosen samples in a batch. Intensity files were used to generate the genotype information (*.chp* files), which were subjected to downstream analysis. The genotype data in one of the 96 sample sets contained randomly selected 72 replicates (47 cases and 25 controls) and also representative samples across all batches to assess genotype call concordance within and across batches. The mean genotype concordance rate for the samples achieved in this analysis was very high (99.9%).

2.6 Statistical considerations

2.6.1 Data quality assessment

Any experiment is subject to random and systematic variations which can be attributed to several sources, including heterogeneity of the phenotype under investigation and experimental conditions. High-throughput technologies, such as DNA microarrays, are no exception. Although a whole genome study design includes data normalisation and replication of experiments to minimise variability, it is possible to further minimise these errors by applying several quality control measures. Therefore, different quality control measures at both the SNP and sample levels are used to ensure good quality of genotype data to perform the association analysis and these methods have evolved recently, and been successfully implemented in all GWASs (89). The data quality control measures are presented in the sequence in which they were carried out for our analysis.

2.6.1.1 SNP quality control measures

Genotype filtering is a step-wise selection procedure that minimises errors (false-positive results), which helps increase the overall power of a study. This was the first quality control measure applied (i) to check the accuracy of the genotype calls for each SNP; and (ii) to detect and remove poor-quality SNPs

from further data analysis. Several SNP quality control measures were applied to the dataset and it was a prerequisite to meet them all to be included in the downstream analysis.

2.6.1.1.1 Hardy–Weinberg equilibrium

Hardy–Weinberg equilibrium (HWE) was independently proposed by H.G. Hardy and W. Weinberg in 1908. In a perfect world, genotype and allele frequencies are expected to remain constant from generation to generation in a randomly mating population (90). Testing for departure from HWE may point to problems in the genotyping procedure; errors can occur at any stage of the sample processing, such as sample handling and problems with hybridisation, leading to incorrect or biased (e.g., excessive homozygosity or heterozygosity) genotype calls for a particular SNP. But assigning faulty genotypes may not be an exclusive reason for deviations from HWE. The other possible factors include:

(i) Small-sized population is largely influenced by random variation in the distribution of alleles to successive generations, a process known as genetic drift, which results in reduced genetic variation in small populations (90). For example, the population size of northern elephant seals drastically dropped after intensive human hunting in the late 19th century. Though the number rebounded after awareness about extinction, the population exhibited limited

genetic diversity (as opposed to southern elephant seals) due to the reduction in the size of the population (91).

(ii) Non-random mating or inbreeding consists of mating of two genetically related members and leads to limited genetic diversity. There are two major drawbacks of inbreeding: (i) increased chance of transmission of recessive deleterious genes (arising due to repeated mutations within a population) for generations resulting in manifestation of the altered phenotype (90); and (ii) violation of the principle of independence, i.e. the occurrence of one event should not be predictable by the occurrence of the other. It implies that there should be no cryptic relatedness, i.e., members of cases or controls should not be related, existing between the subjects included in the analysis.

(iii) Certain alleles or genotypes become more common in population over successive generations. For example, African-Americans are known to have higher susceptibility to hypertension due to impaired excretion of salts leading to expansion of water volume in the blood vessels, resulting in elevated blood pressure (92).

We assessed for deviations from HWE using the chi-square test, with 1 degree of freedom. It was expected that SNPs tested should be in HWE and those that showed deviation at a stringent *p*-value of <0.001 (user-defined) were excluded from the analysis. In other words, there was a 1 in 1000 chance for the observed frequencies to deviate from the expected frequencies.

2.6.1.1.2 Missing genotype calls

Inaccurate genotype calling, due to technical problems such as defective hybridisation and faulty array, can introduce bias and affect the quality of the data. Missing genotype calls can be directly related to incompleteness of the data. Failure to assign genotypes for a particular SNP or subsets of SNPs in a sample results in suboptimal call rates. Therefore, it is ideal to exclude such SNPs from further analysis to improve the overall result. Many reported GWASs to-date have adopted different cut-off points ranging from 80% to 99.7% (8, 11-14, 16-19). We applied a cut-off of ≥99% genotype call rate and SNPs below the cut-off were excluded from further analysis.

2.6.1.1.3 Minor allele frequency

Minor allele frequency (MAF) of a SNP is the frequency of the less frequent allele in a population. As stated earlier, single nucleotide changes are classified as SNPs when the allele frequency is >1%. MAF falls in the range of 1% to <50%. Alleles occurring at a frequency ≥50% are termed major alleles. Due to the low-penetrance nature of the polymorphisms, detection of association of rare variants (very low frequency SNPs) with disease risk is difficult. Adhering to the assumption that common diseases are the result of common variants (CDCV hypothesis), detecting association of rare variants with disease requires large effect sizes, which can be accomplished by a very large sample size (93). It is

advisable to eliminate those SNPs occurring at a lesser frequency to enrich the dataset and thereby reduce the number of association tests to be performed. Therefore, SNPs with MAF <10% were not included in the replication phase of our study. Most reported studies apply a filter criterion ranging from 5% to 10% (11, 12, 18, 19).

2.6.1.1.4 Signal intensity plots

A signal intensity plot, otherwise known as a cluster plot, is a two-dimensional plot (A versus B) giving a graphical display of precise resolution of SNP signals into three distinct genotype clusters (AA, AB and BB) using the intensity values. Low intensity values directly affect the genotype calls of a marker leading to suboptimal separation of the three genotype clusters. Possible reasons for skewed genotype are low SNP intensities, homologous SNP flanking sequences in different parts of the genome and SNPs positioned in the regions of structural aberrations (87, 89).

It would have been labour intensive to screen the intensity plots for 906,600 SNPs and to scrutinize them according to distinct separation of the three genotypes. Therefore, the cluster plots were analysed for only those SNPs short listed for the replication study and the SNPs that showed three distinct genotype clusters were retained in the analysis.

2.6.1.2 Sample quality control

Quality control of samples is as important as SNP quality control. Samples can be of poor quality due to several reasons: poor DNA quality, technical problems with hybridisation, not enough DNA for hybridisation or faulty array. These discrepancies may lead to confounding results. It is essential to remove bad-quality samples to increase the overall accuracy of the results (89).

2.6.1.2.1 Individual chip call rate

Affymetrix GeneChip Operating Software automatically generates an overall quality control call rate for every sample/chip after scanning. The overall call rate is defined as the number of SNPs receiving genotype call (AA, AB, or BB) divided by the total number of SNPs (94). Affymetrix recommends a chip call rate threshold of >86%. In our study, an average chip call rate of 97.6% was observed for the 696 samples.

In addition, quality for each sample was determined by contrast quality control as recommended for the SNP Array 6.0 by the manufacturers. Contrast quality control is a cluster based algorithm that is a good predictor of sample genotyping performance. It measures the distinct clustering of the three genotypes (based on genotype calls of AA, AB or BB) by using a subset of probes (58). Default average contrast quality control for a sample to be included in further

analysis is set at ≥1.7. Most of our samples used in this study had a contrast quality control of >2.0.

2.6.1.2.2 Detection and removal of outliers and correction for population stratification

Detection and removal of outliers and correction for population stratification are important to minimise false-positive findings. The ancestry differences within the study population can result in spurious associations due to differing genotype and disease prevalence patterns. For example, **Figure 2.2** depicts the differences in the genotypes between two populations at a particular SNP locus. A mixture of a few individuals from population 2 into the study population (population 1) resulting in differences in allele and genotype frequencies within the population of interest will impact the association analysis results. The figure shows that 'AA' genotype occurs at a higher frequency and 'BB' genotype occurs at a lower frequency in population 1 than in population 2 in cases; but, the genotype frequency of the controls between the two populations is nearly the same. The effect of stratification is largely dependent on the total number of admixture samples and the loci at which subpopulations differ (95).

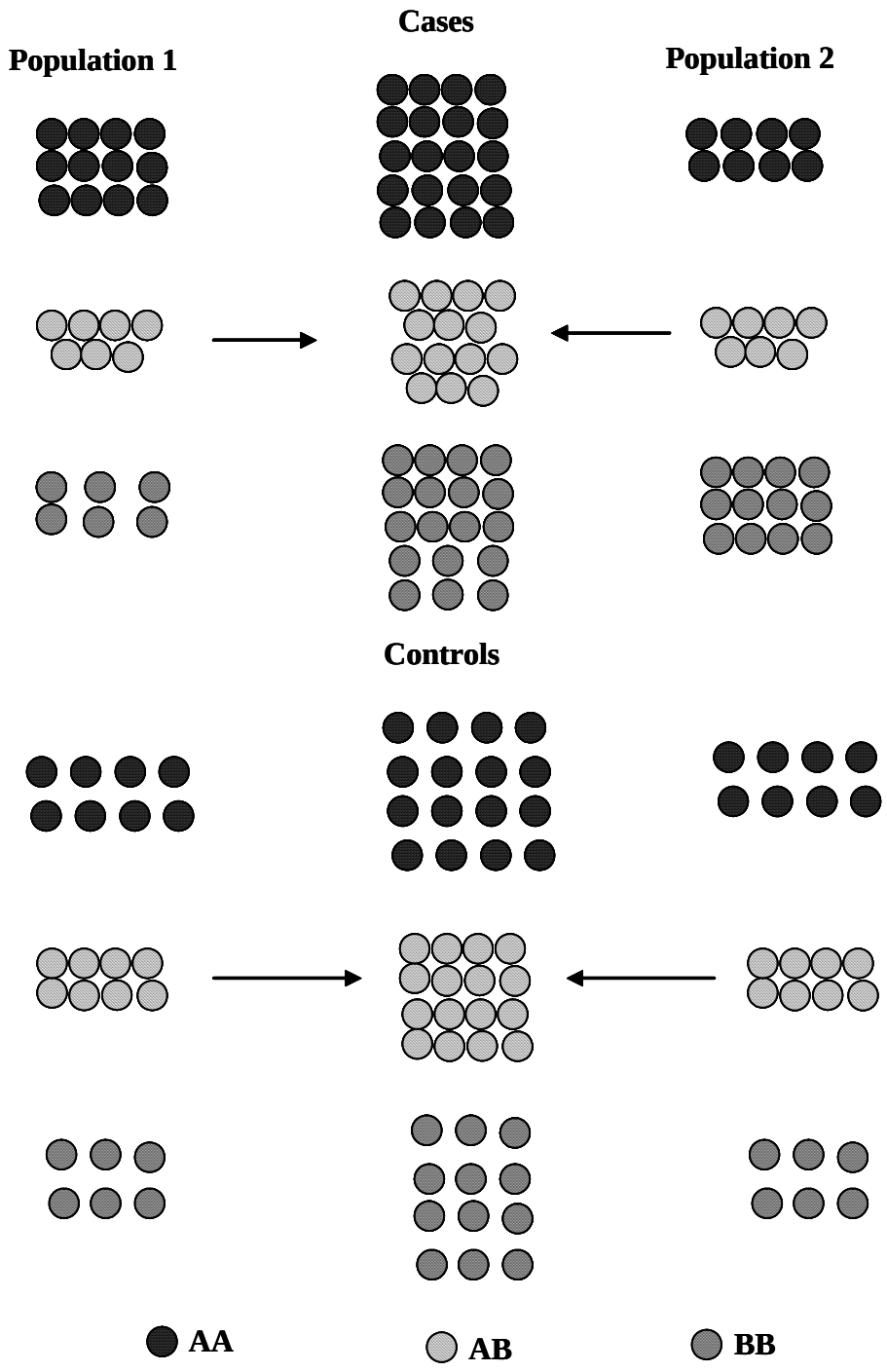


Figure 2.2: Changes in genotype distribution of the study population for a particular SNP upon deriving the samples from two different populations (adapted from Marchini et al. (95)).

Technique used for correcting population stratification: In our study, the ethnicity of the participating individual was based on the self-completed ethnicity questionnaires at the time of recruitment. However, self-declared ethnicity could be biased in families with multiethnic roots. Therefore, it is important to comprehensively assess the population substructure (eliminate those samples that could probably be an admixture) to increase the homogeneity of the dataset at least at the level of selection of markers from Stage I.

EIGENSTRAT is a technique developed by team of researchers at the BROAD Institute that uses principal components analysis to detect and correct for population stratification (96). HelixTree software uses similar methodology as in EIGENSTRAT with further enhancements.

We applied principal components analysis methodology to detect the ancestry differences within our study population (cases and controls). The approach was to project the data on to axes, known as principal components, which capture the maximum variability in the data and end up reducing the dimensionality of the data. We chose to adjust for ancestry along two principal components. EIGENSTRAT emphasises mainly on capturing the maximum variability (effects of stratification) within the defined axes of variation and is not sensitive to the number of principal components considered in the analysis (96). The distribution of variation in two dimensions enabled us to visually inspect and to assess the similarities and differences between samples by comparing with the

reference samples. Outliers were identified and removed from the analyses to enrich the dataset. Outliers were defined as individuals whose ancestry was at least three standard deviations from the mean on one of the two principal components after performing five iterations. We identified 73 outliers which fell beyond the limit of three standard deviations (as opposed to the six standard deviations, which is less stringent, applied in Price et al. (96)) leaving 302 breast cancer cases and 321 controls for further analysis.

Comparison of our study cohort with HapMap samples: We also assessed the genetic homogeneity of our study population by comparing with the HapMap Phase I dataset which included three different populations: CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; YRI, Yoruba in Ibadan; and CHB, Han Chinese in Beijing, China. Each population included a representative set of 90 samples (30 trios: mother, father and child) (28). For a meaningful comparison, we obtained genotype information of all 270 HapMap samples genotyped using Affymetrix SNP 6.0 array. Genotype calling of these samples was done at our facility using the Birdseed v2 algorithm. We inspected the clustering patterns of our study cohort (before and after outlier removal) and HapMap samples by superimposing the two datasets.

2.6.2 Association analysis

2.6.2.1 Single-locus association analysis

After applying quality control measures, we obtained an enriched dataset with 302 breast cancer cases and 321 healthy controls with 782,838 SNPs for single-locus association analysis. The objective of the analysis was to compare the allele frequencies of the affected individuals with those of healthy, unaffected individuals. The differences in frequencies of SNPs implied that there may be associations existing between the markers and the disease outcome. Genotype data were grouped under independent, categorical variable and case–control labels were used as dependent variable. Allelic association between the cases and controls was done using a 2 × 2 contingency table according to the format shown in **Table 2.2**. We performed a chi-square test with 1 degree of freedom to test the null hypothesis of no association between alleles and disease outcome. For every SNP, a 2 × 2 contingency table was constructed by counting the number of times a particular allele occurred for a SNP individually in case and control samples.

Table 2.2: 2 × 2 Contingency table with allele counts for cases and controls

Alleles	Cases	Controls
A	n_A	$\boldsymbol{n_A}$
B	n_B	$\boldsymbol{n_B}$

After association analysis, we selected all the SNPs significant at $p < 0.05$. Due to possible false-positive associations, select markers were replicated in independent cohorts. However, selection of markers for replication was itself not straight forward. One could simply rank the significance level and select SNPs; this method does not account for redundant markers due to LD. We adopted the method described by Zheng et al. (19) which takes into account the LD patterns of the genome. LD patterns were taken into consideration mainly for selection of markers for the Stage II replication study.

2.6.2.2 Haplotype association analysis

Single-locus association analysis provides information on individual allele, independent of its association with neighbouring markers. Genome-level studies using SNP markers have a lot of redundant information due to LD patterns. Therefore, we performed a multiple-marker analysis to potentially overcome redundancy of markers. Recent studies have shown that the entire genome can be parsed into block-like structure called '*haplotype blocks*', discrete segments with low recombination frequencies containing strongly associated

SNPs (37, 97). These are regions of high LD with limited haplotype diversity separated by regions of very low LD known as 'recombination hotspots'. The correlation between markers in these very low LD regions can be addressed by typing many polymorphisms. Representative SNPs from the region with very high LD or a haplotype known as tagSNPs, act as proxy for the neighbouring SNPs in the haplotype and knowledge of the alleles of the tagSNPs can predict the allelic architecture of the adjacent SNPs. The Affymetrix SNP 6.0 array not only has a substantial number of tagSNPs but also SNPs from other regions such as mitochondrial SNPs and SNPs in recombination hotspots

We computed haplotype blocks for the entire genome based on correlation between markers across a chromosome. The correlation between the alleles of SNPs in a block is measured by the r^2 statistic, a measure for statistical correlation between two SNPs. r^2 values range between 0 (no disequilibrium) and 1 (high disequilibrium). Defining a block was done using the Expectation-Maximisation algorithm, a feature available in the software. Default parameters were used for block detection – i.e., maximum length of 160 kb and maximum of 30 markers per block. A haplotype association analysis (pair-wise comparison) was performed in a case–control setting to improve statistical power due to the lesser number of hypotheses tested. A chi-square test was used to perform association analysis. SNPs with chi-square p-value <0.001, $r^2 \geq 0.8$ (SNPs in high LD), with more than two SNPs per block were chosen.

2.6.2.3 Multiple-hypothesis testing

In single-hypothesis testing, when the significance level (*p*-value) is set at 0.05, there is a 5% probability of a SNP showing association to the disease susceptibility by chance. This may result in falsely rejecting the null hypothesis of no association. In our SNP microarray data 906,600 SNPs were interrogated, resulting in a large number of statistical tests. It is commonly referred to as multiple-hypothesis testing, defined as when more than one hypothesis is tested at a time. There is a high chance that false-positive associations will outnumber the true positives and can have serious consequences while evaluating the results (98). A correction method for *p*-values widely accepted by the GWAS research community is the *Bonferroni correction*, a method developed by the mathematician Carlo Emilio Bonferroni (99, 100). This method applies a stringent *p*-value to the entire dataset by adjusting the significance level. The adjusted significance level can be represented as $\alpha^* = \alpha/n$, where α is the unadjusted significance level (0.05) and n is the total number of SNPs included in the association analysis (101). Applying this measure to our dataset,

Unadjusted significance level (*p*-value) = 0.05

Number of SNPs included in single-locus association analysis = 782,838

Overall expected adjusted *p*-value = (0.05/782,838) = 6.4×10^{-8}

This simple calculation indicated that the SNPs that are potentially associated with disease susceptibility are more likely to reach a Bonferroni corrected significance value of $<10^{-8}$. A Bonferroni *p*-value of 10^{-8} corresponds to a *p*-value of 0.05 used in single hypothesis testing. The Bonferroni *p*-value becomes stringent with the increase in the number of markers and does not take into account the redundant markers due to LD. A Bonferroni correction is therefore overly stringent and may lead to discarding true-positive SNPs showing association below this threshold.

2.6.2.4 False-discovery rate

Although the Bonferroni method for multiple comparisons is commonly used for SNP microarray data, it is highly conservative because it increases the specificity by significantly reducing the number of false-positives but on the flip side it compromises on sensitivity by increasing the number of false-negatives (102). A less conservative, alternative to the Bonferroni method is the method of false-discovery rate (FDR). The false-discovery rate (FDR) was also estimated to quantify the proportion of "discoveries" (i.e., rejections of the null hypotheses of 'no association') that were false (103, 104).

Based on the above-mentioned analysis strategies, select SNPs were considered for replication in an independent study using 1153 breast cancer cases and 1215 controls. Selection of markers is discussed in **Appendix A**.

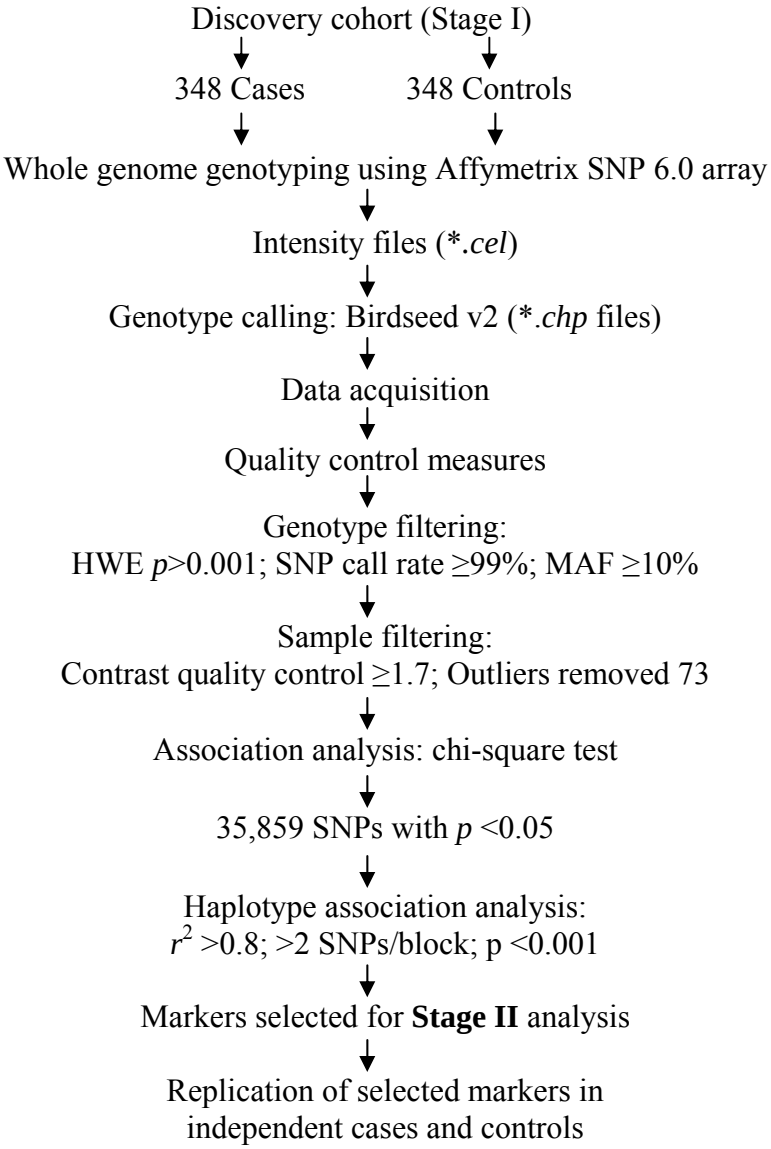


Figure 2.3: Flow chart summarizing the entire protocol of Stage I association analysis

2.7 Replication

There were two vital components in our study: firstly, validation of previously reported markers from the literature to seek relevance to our study population in Alberta; and secondly, replication of markers selected from Stage I from Albertan population in additional independent cases and controls. Both the GWAS-identified variants (literature reports) and the novel markers identified from the preliminary analysis (Stage I) were genotyped using Sequenom® Mass-ARRAY iPlex technology. The Sequenom Mass-ARRAY iPlex technology is based on locus-specific PCR reactions. The PCR product is used in the extension reaction to produce products of different sizes for each allele of a SNP and the size of the product is determined using Sequenom MALDI-ToF mass spectrometry, which in turn is converted into genotype data (105). Genotyping services were provided by Genome Quebec Innovation Centre (Montreal, Quebec, Canada).

2.7.1 Validation of GWAS-identified variants

SNP selection and genotyping: We conducted a systematic PubMed literature review using the search terms 'whole genome association study', 'breast cancer susceptibility loci', and 'novel SNPs', along with manual review of bibliographies of published articles about breast cancer, to identify reported SNPs/loci associated with breast cancer risk. Thirty-three SNPs that were reported

as statistically significant (p <0.05) were selected from nine research articles published between May 2007 and May 2009 (11-19). Of the 33 SNPs from this search, 28 were genotyped in this study using Sequenom Mass-ARRAY iPlex technology (105). One SNP was monomorphic in our study population (rs1078806 (13)) and four other SNPs did not meet the criteria for multiplexing on the Sequenom platform (rs2046210 (19), rs6556756 and rs1154865 (15), rs7776136 (13)) and were excluded. Since the SNPs validated were extensively studied by others, we did not adopt a multi-stage study design. However, the total number of cases and controls used to interrogate these literature findings were the same as our multi-stage association study to identify novel/additional SNPs in the Albertan population. This approach also served as an internal control wherein the sample size and power if sufficient for the validation of previously published associations would also give confidence in the interpretations for the genome-wide scan in Albertan population.

2.7.2 Replication of novel markers

For the replication phase (Stage II), 1153 breast cancer cases and 1215 controls were genotyped using the Sequenom Mass-ARRAY iPlex technology. Thirty-five SNPs selected from Stage I analysis were genotyped. Consistent statistically significant results in independent stages and in joint analysis for novel SNPs were considered as positive associations.

2.7.3 Genotype calling quality control

Genotyping on an independent technology platform provides a technical validation across genotyping platforms. It ensures the quality of the genotype data. There are two possible ways to validate the genotype quality: (i) within platform concordance by repeating the same set of samples in duplicates; and (ii) across-platform concordance to assess the consistency in genotype calling of the same set of samples between the two genotyping platforms (Affymetrix versus Sequenom). As quality control measures for genotype calling between platforms, the Stage I samples were re-genotyped on the Sequenom platform to evaluate the genotype concordance between the platforms, prior to replication of select markers in an independent cohort (Stage II). It enabled us to compare the genotypes of the select SNPs between Affymetrix and Sequenom platforms.

The Affymetrix genotyping protocol is summarized in this chapter since half of the samples were genotyped by me but the results generated from Affymetrix data are presented in the Appendix section due to the participation of other members in the laboratory. Chapters 3 and 4 will address the results obtained from GWAS-identified variants and novel SNPs, respectively, along with discussion.

3 Association analysis of candidate SNPs selected from literature – a validation study[a]

Many SNPs associated with breast cancer risk have been identified in GWAS by several research groups and for different populations. Subsets of these SNPs have been successfully replicated within the initially identified population or have been validated in independent studies from geographically diverse regions. However, no single study has attempted to validate all previously GWAS-identified variants, and none has been reported from an ethnically defined Canadian population.

Since 2007, a series of publications about GWASs on breast cancer have shown that several loci are potentially associated with disease risk (11-19). Disease heterogeneity was also addressed by determining the associations based on the clinical subphenotypes, such as tumour grade, receptor status, family history and stage (11, 16, 17, 26, 27). Thus, it is evident from the literature findings that many genes/loci play critical roles in disease pathogenesis. However, the outcomes from multiple GWASs (original and validation studies from independent laboratories) can vary due to application of different quality control metrics (SNP call rates, genotypic models, cut-offs imposed for deviations from HWE and disease heterogeneity). Therefore, it is important to validate the

[a] A version of this chapter has been submitted for publication.

reported findings in several independent studies and in different populations, albeit of similar or diverse genetic background, to determine the association of the defined variant with the disease risk.

Determining correlation of the candidate SNPs in our study population provides confirmation of the association of the previously reported markers with breast cancer susceptibility. The SNPs selected for validation were from the nine research articles reviewed in the Introduction (see Section 1.6.2). A total of 1439 breast cancer case subjects and 1596 healthy control subjects were genotyped for the selected SNPs. We reviewed the medical records of affected individuals for their oestrogen (ER), progesterone (PR) and human epidermal growth factor (Her2) receptor status and assessed association of the SNPs with breast cancer based on the receptor status (**Figure 3.1**).

Twenty-eight of the 33 SNPs reported to be significantly associated with breast cancer were successfully genotyped in our population on the Sequenom platform. Of the 28 SNPs that were successfully genotyped, only one SNP deviated from HWE (rs3012642, *HDAC8* gene SNP) in our population (13). We included this SNP also for analysis since we intend to compare our results with those reported in literature. Association analysis was carried out using the commercial software, HelixTree. A chi-square analysis was performed to determine the association of the SNPs. As a quality control measure for within platform genotype calling, 132 replicate samples (67 cases and 65 controls) were

randomly distributed in each of the 96-well plate assays. The mean genotype concordance rate of the replicates was 98.6%.

Of the 28 SNPs subjected to allelic association analysis, 14 SNPs from nine genes were identified to be statistically significant (p <0.05) in the case–control breast cancer association study in our population. All the SNPs (except rs3012642, with a MAF of 0.03) had an MAF >0.10. The risk (minor) allele frequencies of SNPs in our study population were comparable to those reported in the GWASs (11-18).

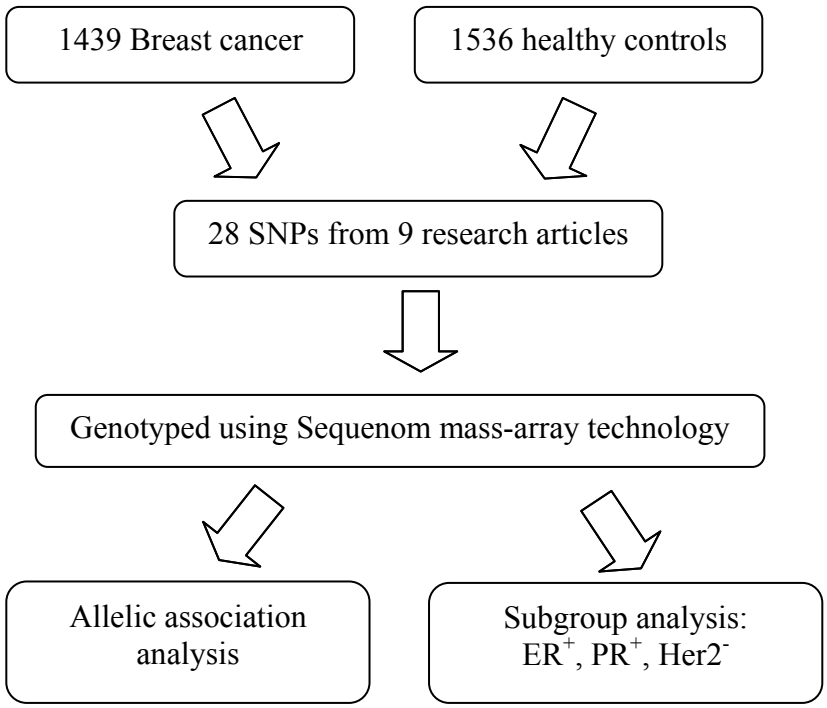


Figure 3.1: Flow chart depicting the overview of analysis performed using candidate SNPs.

3.1 Chromosome 10 region polymorphisms

Fibroblast growth factor receptor 2 (*FGFR2*) gene, located in chromosome 10q26, has been well-studied and is known to function upstream of the *Ras*/*MAPK* and *PI3K*/*Akt* cellular signalling pathways (106). It has been reported to be over-expressed in cultured breast cancer cells, resulting in constitutive downstream signalling to maintain the transformed cell in an activated state (70). The multiple alleles identified in this region have been studied in diverse ethnic cohorts (19, 107). In confirmation with previous reports (12, 14), we observed statistically significant associations for all the five SNPs from *FGFR2* in our study cohort: rs1219648 (*p*-value 7.07×10^{-6}, FDR 1.98×10^{-4}), rs2981579 (*p*-value 2.66×10^{-5}, FDR 2.48×10^{-4}), rs2420946 (*p*-value 2.94×10^{-5}, FDR 2.06×10^{-4}), rs2981582 (*p*-value 5.32×10^{-5}, FDR 2.98×10^{-4}), and rs11200014 (*p*-value 8.52×10^{-5}, FDR 3.96×10^{-4}). The ORs for all risk alleles in *FGFR2* were in the range of 1.23–1.27 (**Table 3.1**). The association of *FGFR2* alleles remained significant even in the subgroup analysis (**Table 3.2**). The direction of risk (OR >1.0) was retained for all *FGFR2* risk alleles in all independent analyses based on ER^+, PR^+ and $Her2^-$ receptor status.

3.2 Chromosome 5 region polymorphisms

A number of GWASs have reported SNPs within chromosome 5 as significant in conferring breast cancer risk (12, 16-18). Two SNPs, rs4415084 and rs10941679, from the Icelandic population and other populations of European origin, were initially reported to be associated with breast cancer risk (17). These two SNPs, along with four other SNPs from the same gene region, were also evaluated in African-American women (108). Thomas et al. (18) evaluated two novel SNPs, rs7716600 and rs2067980, from the same region in a GWAS. Easton et al. (12) showed a mitogen activated protein kinase 1 (*MAP3K1*) SNP rs889312, located in the q arm of chromosome 5, to be associated with breast cancer risk, an observation that was validated in post-menopausal European-American and African-American women (109). In all, a total of nine SNPs were reported in independent studies from chromosome 5 to be significantly associated with breast cancer risk.

In our study, we found statistical significance (**Table 3.1**) for two of the SNPs from the study by Stacey et al. (17): rs4415084 (p-value 1.22×10^{-3}, FDR 3.42×10^{-3}) and rs10941679 (p-value 5.25×10^{-3}, FDR 1.23×10^{-2}), and only marginal significance for one SNP (rs2067980) reported by Thomas et al. (18) (p-value 0.05, FDR 0.10). Polymorphism rs889312 from the study of Easton et al. (12) showed high statistical significance in our study population (p-value 1.76×10^{-4}, FDR 6.16×10^{-4}) in the overall association analysis. The OR (95% CI) of this

risk allele was 1.24 (1.11–1.38). The SNP, rs7716600, from chromosome 5 was not significant in the overall association analysis (**Table 3.3**).

Of the four SNPs that were significant in our study, rs4415084 and rs889312 showed strong association with breast cancer risk in the subgroup analysis. rs4415084 showed association with ER^+ (*p*-value 3.52×10^{-4}, OR 1.23), PR^+ (*p*-value 1.23×10^{-3}, OR 1.22) and $Her2^-$ (*p*-value 4.27×10^{-3}, OR 1.18); rs889312 showed association with ER^+ (*p*-value 8.28×10^{-4}, OR 1.23), PR^+ (*p*-value 4.59×10^{-3}, OR 1.20) and $Her2^-$ (*p*-value 2.04×10^{-3}, OR 1.22) in the subgroup analysis (**Table 3.2**), whereas rs10941679 showed marginal significance for ER^+ (*p*-value 1.04×10^{-2}, OR 1.18), PR^+ (*p*-value 3.15×10^{-2}, OR 1.16) and $Her2^-$ (*p*-value 3.94×10^{-2}, OR 1.14). While rs7716600 did not show significance in the overall association analysis, it did show marginal significance for ER^+ (*p*-value 3.97×10^{-2}, OR 1.15), PR^+ (*p*-value 3.90×10^{-2}, OR 1.16) but not for $Her2^-$ (**Table 3.2**). Also rs2067980 was marginally significant in the overall association analysis but was not statistically significant in any of the subgroups.

3.3 Other significantly associated breast cancer risk alleles

(i) *Chromosome 16 region polymorphism*: Two GWASs independently identified rs3803662 to be associated with breast cancer risk (12, 16). This

polymorphism is located at chromosome 16q12, 8 kb upstream of the *TNRC9* and the hypothetical gene *LOC643714*. It has been shown that the *TNRC9* gene contains a high mobility group box motif, suggesting its role as a transcription factor (12). The expression of the gene has been implicated in the metastasis of breast cancer to bone (77). Based on *p*-values, rs3803662 was the second most statistically significant SNP in our study population (*p*-value 1.13×10^{-5}, FDR 1.58×10^{-4}) with an OR (95% CI) of 1.29 (1.15–1.44) conferring disease risk (**Table 3.1**). The SNP showed strong association with ER^+ (*p*-value 1.17×10^{-3}, OR 1.23), PR^+ (*p*-value 5.93×10^{-3}, OR 1.20) and $Her2^-$ (*p*-value 1.45×10^{-3}, OR 1.23) (**Table 3.2**).

(ii) *Chromosome 8 region polymorphism*: Polymorphism rs13281615 lies in the 8q24 chromosomal region with no flanking annotated gene regions. This 8q24 region has been previously implicated in prostate cancer (31, 32). Easton et al. (12) were the first to show the association of rs13281615 with breast cancer susceptibility. In our study cohort, this variant was statistically significant (*p*-value 3.52×10^{-4}, FDR 1.09×10^{-3}). The OR (95% CI) of 1.21 (1.09–1.34) of the minor allele indicated disease risk (**Table 3.1**). Variant from rs13281615 also showed strong association with ER^+ (*p*-value 7.20×10^{-4}, OR 1.22), PR^+ (*p*-value 1.02×10^{-3}, OR 1.22), and $Her2^-$ (*p*-value 1.22×10^{-3}, OR 1.21) (**Table 3.2**).

(iii) *Chromosome 3 region polymorphism*: A polymorphism rs4973768, located in chromosome 3p24, was identified in the flanking region of the gene

encoding solute carrier family 4, sodium bicarbonate cotransporter, member 7 (*SLC4A7*), showing an association with breast cancer risk (11). *SLC4A7* is known to be a transporter of bicarbonate (HCO_3^-), which plays a key role in the regulation of pH in the body. Chen et al. (78) suggested that *SLC4A7* is a tyrosine kinase substrate with decreased expression in cultured breast cancer cells. We observed that rs4973768 was significantly associated (*p*-value 0.005, FDR 0.012) in the overall association analysis with an OR (95% CI) of 1.16 (1.05–1.29) and the minor allele conferring risk (**Table 3.1**). This polymorphism also showed association with ER^+ (*p*-value 0.008, OR 1.17), PR^+ (*p*-value 0.009, OR 1.18) and $Her2^-$ (*p*-value 0.013, OR 1.16) (**Table 3.2**).

(iv) *Chromosome 17 region polymorphism*: rs2075555 located within 60 kb of the collagen, type 1, alpha 1 (*COL1A1*) gene on the q arm of chromosome 17 was shown to be associated with breast cancer risk in a family-based association study (15). In our study cohort, we observed a marginal statistical significance of rs2075555 (*p*-value 0.023, FDR 0.05) in the overall association analysis. An OR (95% CI) of 1.19 (1.02–1.37) was indicative of risk effect conferred by minor allele (**Table 3.1**). This SNP did not show any association in the subgroup analysis.

(v) *Chromosome-X region polymorphism*: The histone deacetylases (*HDAC*) genes are known to play a crucial role in cell cycle control, cell differentiation and histone assembly. *HDAC1* and *HDAC3* are also known to be

over-expressed in tumour cells (110). A direct association of *HDAC8* from chromosome-X and breast cancer has not yet been established. Gold et al. (13) showed rs3012642 to be marginally associated with breast cancer across independent stages but not in their combined analysis. In our study cohort, rs3012642 was statistically significant (*p*-value 1.39×10^{-4}, FDR 5.54×10^{-4}). An OR of 1.71 (95% CI of 1.29–2.25) for the risk allele was observed in the overall association analysis (**Table 3.1**) but not in the subgroup analysis.

Table 3.1: Polymorphisms associated with breast cancer susceptibility in the women from Alberta

dbSNP ID	Associated genes	Chr	χ^2 p	FDR	OR (95% CI)	MA	MAF	Ref
rs1219648	*FGFR2*	10q26	7.071E-06	1.980E-04	1.27 (1.14-1.40)	C	0.42	(14)
rs3803662	*TNRC9*	16q	1.129E-05	1.580E-04	1.29 (1.15-1.44)	T	0.29	(12)
rs2981579	*FGFR2*	10q26	2.661E-05	2.483E-04	1.25 (1.12-1.38)	A	0.43	(14)
rs2420946	*FGFR2*	10q26	2.937E-05	2.056E-04	1.25 (1.12-1.38)	T	0.42	(14)
rs2981582	*FGFR2*	10q26	5.323E-05	2.981E-04	1.24 (1.12-1.37)	T	0.41	(12)
rs11200014	*FGFR2*	10q26	8.517E-05	3.975E-04	1.23 (1.11-1.36)	T	0.42	(14)
rs3012642	*PHKA/HDAC8*	Xq13.1	1.386E-04	5.544E-04	1.71 (1.29-2.25)	C	0.04	(13)
rs889312	*MAP3K1*	5q	1.760E-04	6.159E-04	1.24 (1.11-1.38)	G	0.30	(12)
rs13281615	8q	8q	3.517E-04	1.094E-03	1.21 (1.09-1.34)	C	0.43	(12)
rs4415084	5p12	5p12	1.221E-03	3.420E-03	1.19 (1.07-1.31)	T	0.43	(17)
rs4973768	*SLC4A7*	3p24	5.073E-03	1.291E-02	1.16 (1.05-1.29)	A	0.49	(11)
rs10941679	5p12	5p12	5.249E-03	1.225E-02	1.18 (1.05-1.32)	C	0.27	(17)
rs2075555	*COL1A1*	17	2.324E-02	5.005E-02	1.19 (1.02-1.37)	T	0.14	(15)
rs2067980	*MRPS30*	5p12	4.976E-02	9.952E-02	1.15 (1.00-1.32)	C	0.16	(18)

FGFR2, fibroblast growth factor receptor 2; *TNRC9*, trinucleotide repeat containing 9; *PHKA1*, phosphorylase kinase, alpha 1; *HDAC8*, histone deacetylase 8; *MAP3K1*, mitogen activated protein kinase kinase kinase 1; *SLC4A7*, solute carrier family 4, sodium bicarbonate cotransporter, member 7; *COL1A1*, collagen, type 1, alpha 1; *MRPS30*, mitochondrial ribosomal protein S30; Chr, chromosome; χ^2, chi-square; FDR, false-discovery rate; OR, odds ratio; CI, confidence interval; MA: minor allele; MAF, minor allele frequency.

Table 3.2: Subgroup analysis of polymorphisms based on the hormone receptor status in the women from Alberta

dbSNP ID	Genes	ER^+ vs Controls		PR^+ vs Controls		$Her2^-$ vs Controls		Ref
		χ^2 p	OR (95% CI)	χ^2 p	OR (95% CI)	χ^2 p	OR (95% CI)	
rs1219648	*FGFR2*	1.08E-05	1.29 (1.15,1.45)	1.65E-05	1.30 (1.15,1.46)	3.28E-05	1.28 (1.14,1.43)	(14)
rs2420946	*FGFR2*	2.22E-05	1.28 (1.14,1.43)	3.99E-05	1.28 (1.14,1.44)	1.18E-04	1.25 (1.12,1.41)	(14)
rs2981579	*FGFR2*	4.74E-05	1.26 (1.13,1.42)	2.93E-05	1.29 (1.14,1.45)	9.81E-05	1.26 (1.12,1.41)	(14)
rs2981582	*FGFR2*	6.13E-05	1.26 (1.13,1.41)	9.90E-05	1.27 (1.12,1.43)	3.60E-04	1.23 (1.10,1.39)	(12)
rs11200014	*FGFR2*	1.09E-04	1.25 (1.12,1.40)	7.24E-05	1.27 (1.13,1.43)	3.69E-04	1.23 (1.10,1.38)	(14)
rs4415084	5p12	3.52E-04	1.23 (1.10,1.38)	1.23E-03	1.22 (1.08,1.37)	4.27E-03	1.18 (1.05,1.33)	(17)
rs13281615	8q	7.20E-04	1.22 (1.09,1.37)	1.02E-03	1.22 (1.08,1.38)	1.22E-03	1.21 (1.08,1.36)	(12)
rs889312	*MAP3K1*	8.28E-04	1.23 (1.09,1.39)	4.59E-03	1.20 (1.06,1.37)	2.04E-03	1.22 (1.07,1.38)	(12)
rs3803662	*TNRC9*	1.17E-03	1.23 (1.08,1.39)	5.93E-03	1.20 (1.05,1.37)	1.45E-03	1.23 (1.08,1.39)	(12)
rs4973768	*SLC4A7*	8.08E-03	1.17 (1.04,1.31)	8.62E-03	1.18 (1.04,1.33)	1.26E-02	1.16 (1.03,1.30)	(11)
rs10941679	5p12	1.03E-02	1.18 (1.04,1.34)	3.15E-02	1.16 (1.01,1.32)	3.94E-02	1.14 (1.01,1.30)	(17)
rs3012642	*PHKA1*	2.58E-02	1.42 (1.04,1.95)	2.19E-02	1.46 (1.05,2.01)	1.73E-04	1.77 (1.31,2.39)	(13)
rs7716600	*MRPS30*	3.97E-02	1.15 (1.01,1.32)	3.90E-02	1.16 (1.01,1.34)	1.89E-01	1.10 (0.96,1.26)	(18)

FGFR2, fibroblast growth factor receptor 2; *TNRC9*, trinucleotide repeat containing 9; *PHKA1*, phosphorylase kinase, alpha 1; *HDAC8*, histone deacetylase 8; *MAP3K1*, mitogen activated protein kinase kinase kinase 1; *SLC4A7*, solute carrier family 4, sodium bicarbonate cotransporter, member 7; *MRPS30*, mitochondrial ribosomal protein S30; Chr, chromosome; ER, estrogen receptor; PR, progesterone receptor; Her2, human epidermal growth factor receptor 2; χ^2, chi-square; FDR, false-discovery rate; OR, odds ratio; CI, confidence interval.

Table 3.3: SNPs not significant from among the selected polymorphisms (28 SNPs) in the women from Alberta

dbSNP ID	Associated genes	Chr	χ^2 *p*	FDR	OR (95% CI)	MA	MAF	Ref
rs7716600	*MRPS30*	5p12	8.14E-02	1.52E-01	1.12 (0.99-1.27)	T	0.23	(18)
rs7696175	*TLR1*/*TLR6*	4p	1.14E-01	2.00E-01	0.92 (0.83-1.02)	T	0.45	(14)
rs11249433	1p11.2	1p11.2	1.50E-01	2.47E-01	1.08 (0.97-1.20)	G	0.41	(18)
rs13387042	2q35	2q35	3.23E-01	5.02E-01	0.95 (0.86-1.05)	C	0.48	(16)
rs999737	*RAD51L1*	14q24.1	3.25E-01	4.79E-01	0.94 (0.83-1.06)	T	0.22	(18)
rs7203563	*A2BP1*	16p	4.23E-01	5.92E-01	1.07 (0.91-1.27)	C	0.10	(13)
rs17157903	*RELN*	7q	5.21E-01	6.95E-01	1.05 (0.90-1.22)	T	0.13	(14)
rs6504950	*STXBP4*	17q23	6.10E-01	7.76E-01	0.97 (0.86-1.09)	A	0.27	(11)
rs1978503	*FLJ45743*	18	6.71E-01	8.17E-01	1.03 (0.90-1.17)	G	0.18	(15)
rs6569479	*ECHDC1*/*RNF146*	6q22.33	6.75E-01	7.88E-01	1.03 (0.91-1.15)	A	0.25	(13)
rs3817198	*LSP1*	11p	6.85E-01	7.68E-01	1.02 (0.92-1.14)	C	0.32	(12)
rs2180341	*ECHDC1*/*RNF146*	6q22.33	6.99E-01	7.53E-01	1.02 (0.91-1.15)	C	0.25	(13)
rs6569480	*ECHDC1*/*RNF146*	6q22.33	7.17E-01	7.44E-01	1.02 (0.91-1.15)	T	0.25	(13)
rs1926657	*ABCC4*	13	8.51E-01	8.51E-01	1.01 (0.88-1.16)	A	0.17	(15)

MRPS30, mitochondrial ribosomal protein S30; *TLR1*/*TLR6*, toll-like receptor 1/6; *RAD51L1*, RAD51-like 1 (*S. cerevisiae*); *A2BP1*, ataxin-2-binding protein 1; *RELN*, reelin; *STXBP4*, syntaxin binding protein 4; FLJ45743 (hypothetical protein); *ECHDC1*/*RNF146*, enoyl Coenzyme A hydratase domain containing 1/ring finger protein 146; *LSP1*, lymphocyte-specific protein 1; *ABCC4*, ATP-binding cassette sub-family C member 4; Chr, chromosome; χ^2, chi-square; FDR, false discovery rate; OR, odds ratio; CI, confidence interval; MA, minor allele; MAF, minor allele frequency.

3.4 Discussion

Inter-individual variations in germ-line DNA contributes to several phenotypes in health and disease. In general, allele frequency differences are more prominent in geographically distinct populations. Therefore, it is important to investigate the relevance of reported breast cancer susceptibility variants within specific populations. Few studies have considered validation of initial GWAS findings from the literature as an integral part of their study (18, 19) and few other studies have exclusively validated a subset of GWAS SNPs from a specific gene region (107-109, 111). However, in our study we selected the most promising SNPs from the published literature for an independent, geographically confined confirmation. Our study cohort included 1439 breast cancer cases and 1536 controls from women in Alberta, Canada, and is the first study to report the polymorphisms associated with breast cancer in the Canadian population. Although 1028 breast cancer subjects and 329 control subjects from the Ontario Familial Breast Cancer Registry were part of a large-scale study (11), the population subset was not individually analysed for their association with the polymorphisms.

A recent GWAS has validated 13 SNPs previously reported to be associated with breast cancer, and for each, explored the association with that published SNP or one of its neighbouring SNPs known to be in strong correlation with the published SNPs (8). In their study, all 13 SNPs genotyped showed

significant association with breast cancer risk, with SNPs from the *FGFR2* gene region showing the strongest associations. Our study included these SNPs (except rs10931936, from *CASP8* gene (20)) and also included 16 other SNPs from GWASs for validation (13-15, 17, 18), totalling 28 SNPs in our study.

The *FGFR2* gene is over-expressed in a small subset of breast cancers (70) and is known to play a critical role in mammary gland development and tumourigenesis in mice (71). In addition, association of *FGFR2* polymorphisms with breast cancer risk has been documented and tumours with the receptor profiles that were mainly ER^+ and PR^+ were overrepresented (11, 16, 17, 26, 112), while $Her2^-$ tumours were underrepresented (109). Two independent studies have shown polymorphisms in intron 2 of the *FGFR2* gene region to be associated with breast cancer (12, 14). Fine mapping showed that eight SNPs are in strong LD spanning a 7.5 kb region (12). Functional analysis revealed that the minor allele of rs2981578 and major allele of rs7895676 tightly bind the Oct-1/Runx2 and C/EBPβ transcription factors, respectively, leading to over-expression of *FGFR2* gene (72). Runx2 forms a complex with the ubiquitous transcription factor Oct-1 that is known to play an important role in mammary gland-specific expression (73, 74). Udler et al. (113) performed a joint analysis of *FGFR2* polymorphism data from African-American, European and Asian populations to search for the most commonly conserved polymorphisms as an approach to identify the causative allele(s); up-regulation of expression of *FGFR2* in breast cancer was correlated with chromatin structure (DNAseI hypersensitive sites), and it was

reasoned that the presence of SNP rs2981578 within this accessible region of chromatin was a plausible mechanism to explain breast carcinogenesis. While unequivocal evidence for *FGFR2* as a causative locus in breast cancer is awaited, *FGFR2* has been validated in several independent studies, over diverse genetic backgrounds and differing ethnicities. Most studies show a strong association in the overall and/or subgroup analyses and support an association in Caucasians (8, 18, 109), Sephardi Jews (107), Ashkenazi Jews (13, 107), and the Chinese population (19), although in Arab Israeli (107) and African-American (109) women *FGFR2* SNPs were not significantly associated with breast cancer. SNPs in intron 2 of the *FGFR2* gene region have consistently showed statistical significance in Caucasian populations and our results supported these findings.

Initial findings showed that rs4415084 (*p*-value 1.8×10^{-11}) and rs10941679 (*p*-value 2.5×10^{-12}) from chromosome 5 were strongly associated with ER^+ breast cancer cases of European ancestry (17). Validation of SNPs in a subsequent whole genome study with women from European ancestry also yielded similar results in the overall analysis (rs4415084: *p*-value 4.53×10^{-5}; rs10941679: *p*-value 5.50×10^{-3}) (18). We confirmed this association in our Caucasian population and observed associations in the same direction for the overall, ER^+, PR^+ and $Her2^-$ subgroup analyses. Others report that both SNPs in African-American women confer little risk, in that rs4415084 was marginally associated with breast cancer in the overall analysis (*p*-value 0.06) and ER^+ receptor status (*p*-value 0.03) and rs10941679 was not significant in the overall and subgroup analyses (108). In aggregate, it

appears that this association is more pronounced for women from European ancestry than for women from African ancestry. Another SNP from chromosome 5, rs2067980, was only marginally significant in our study cohort and did not show any association with receptor status.

MAP3K1 polymorphism rs889312, located on chromosome 5q, showed high statistical significance in the original association study (*p*-value 7×10^{-20}) in women from European ancestry (12). The *MAP3K1* gene product plays a crucial role in cellular signalling by responding to fibroblast growth factor 2 (FGF2) and activating the MAPK/Erk pathway (109). Validation of the SNP in women from European ancestry confirmed the initial findings with a *p*-value 4.6×10^{-9} (8). Independent validation of this SNP showed an association based on receptor status in African-American women, but failed to show an association in European-American (Caucasian ancestry) women (109). In our Alberta cohort (predominantly Caucasian), we observed rs889312 to be associated with breast cancer susceptibility in both overall and subgroup association analyses.

The *TNRC9* locus SNP rs3803662, located on chromosome 16q12, was shown to be highly significant in two independent studies (12, 16). Fine mapping of the tagSNPs representing the variants in entire *TNRC9* and *LOC643714* genes consistently showed a strong association of rs3803662 with breast cancer risk (12). Validation of the SNP in women from European ancestry also showed high statistical significance in two GWASs: *p*-value 1.11×10^{-9} (18) and *p*-value $3.2\times10^{-}$

[15] (8). On the contrary, validation of this polymorphism in two different studies in Chinese population yielded marginal significance (*p*-value 0.012) in one study (19) and did not show any association in another study (114). Consistent with the original findings in the Caucasian population, we observed statistical significance in the Alberta population for this polymorphic variant in the overall association analysis and also when stratified by receptor status (subgroup analysis).

Ahmed et al. (11) showed that polymorphism rs4973768, in the proximity of the *SLC4A7* gene, was highly significant (*p*-value 4.1×10^{-23}). This gene plays a critical role in the regulation of cellular pH balance and its down-regulation in breast cancer cells was correlated with tumour progression (78). Validation of the SNP in a recent GWAS with women of European ancestry showed a strong association (*p*-value 5.8×10^{-7}) with disease risk (8) and we also confirmed the association of this SNP in our population (*p*-value 5.07×10^{-3}).

The chromosome-X SNP, rs3012642 was originally shown to be only marginally statistically significant in multi-stage study design (and not in the joint analysis) in Ashkenazi Jews (13). We were interested in evaluating the association of this SNP in our study population, which was predominantly of Caucasian origin. Although it was found to be highly significant, this SNP deviated from HWE both in cases (HWE *p*-value 3.67×10^{-72}) and in controls (HWE *p*-value 2.86×10^{-39}), suggesting that it is undergoing shifts in population frequency and/or may harbour copy number alterations. The observed HWE deviation was more

pronounced in cases than in controls. This polymorphism also exhibited low MAF (0.03 both in our study cohort and Ashkenazi Jews). This region warrants further investigation using a larger cohort; interrogation of copy number aberrations in this region may offer additional insights.

Easton et al. (12) previously showed that polymorphism rs13281615 from chromosome 8 (undefined gene) was highly associated (p-value 5×10^{-12}) with breast cancer risk. Consistent with the initial findings, validation of this SNP in women of European ancestry in two other studies also showed high statistical significance in breast cancer susceptibility (p-value: 3×10^{-5} (115) and 2.2×10^{-5} (8)). We observed a similar association (p-value of 3.5×10^{-4}, **Table 3.1**) in our study cohort.

A family-based association study identified rs207555 to be associated (p-value 8.3×10^{-8}) with breast cancer susceptibility (15). The reason for considering the SNP for validation was its inclusion in the GWAS karyogram offered by HapMap website (57). It was shown that high stromal collagen in the mammary tissue of mice increases the risk of breast cancer (116, 117). This is the first study to validate the SNP in the context of breast cancer, as it showed a marginally significant association in the overall analysis.

In summary, we verified the significant associations with breast cancer risk for 14 of the 28 SNPs reported in the original studies; the risk estimates were

of similar magnitude and were concordant in direction. The strongest associations were exhibited by SNPs from intron 2 of the *FGFR2* gene region, which was concordant with the original reports and other validation studies. We could not validate the remaining SNPs from published association studies in our Alberta population, presumably due to limitations in the sample size, disease heterogeneity or some, as yet, unidentified fine population-specific differences. Due to the importance of understanding the multifactorial contributions to breast cancer risk, the remaining SNPs warrant further study by independent groups and in other populations.

4 Association analysis of select markers from whole genome scan – a replication study[b]

We performed a two-stage association study on cohorts from Alberta, Canada, to identify potential novel loci associated with breast cancer susceptibility. Whole genome association analysis from Stage I enabled us to identify a subset of markers with nominal *p*-values (<0.05) that showed association with the disease trait (**Appendix A**). As this subset of markers may also include several false associations, we attempted to replicate these primary findings by testing them on an independent cohort with higher sample sizes than in Stage I. We selected the markers for replication from Stage I in a systematic manner proposed by Zheng et al. (19). We selected 35 SNPs for replication in Stage II with a completely independent series of 1153 cases and 1215 controls. Association analysis and tests of significance were carried out independently for Stage II and in combined samples for Stages I and II. The joint analysis consisted of a total of 1455 breast cancer cases and 1536 controls obtained by combining the samples from two stages of the study.

Genotyping was done using the Sequenom Mass-ARRAY iPlex platform for Stage II samples. As a quality control measure, the Stage I samples were re-genotyped on the Sequenom platform to evaluate the genotype concordance

[b] A version of this chapter has been submitted for publication.

between the platforms (Affymetrix versus Sequenom), prior to replication of select markers in an independent cohort (Stage II). The SNPs that were statistically significant (p <0.001) in Stage I on the Affymetrix platform and those that were selected for replication were re-genotyped in 647 samples (326 cases and 321 controls) on the Sequenom Mass-ARRAY iPlex platform. We observed high mean genotype concordance rates of 95% between these two genotyping platforms. To assess the genotype concordance within the Sequenom platform, 132 replicate samples (67 cases and 65 controls) were randomly distributed in each of the 96-well plate assays. The mean genotype concordance rate of the replicates was again high at 98.6%.

4.1 Replication of markers (Stage II)

In Stage II, we genotyped 35 SNPs using Sequenom Mass-ARRAY iPlex technology in independent case and control subjects (1153 cases and 1215 controls). We identified 10 of the 35 SNPs from Stage I as significant (p <0.05) in the Stage II samples (**Tables 4.1** and **4.2**). We also performed a joint analysis which is considered the best way to confer power and confidence in the results, as well as to address the possible sampling bias and inherent heterogeneity of breast cancer as a phenotype (118). The joint analysis consisted of a total of 1455 breast cancer cases and 1536 controls obtained by combining the samples from the two stages of the study. Of the 10 SNPs that were significant in the replication (Stage II) sample set, seven SNPs retained statistical significance in the joint analysis

(**Table 4.1**). Polymorphisms rs9644134 and rs7119677 in the intronic regions of orf80 and orf141 from chromosomes 8 and 11, respectively, showed significance in the individual stages but not in the joint analysis. Three SNPs from chromosomes 8, 9 and 11 (rs6997395, rs6478296 and rs9630178) showed significance in Stage I and not in Stage II; however, these three SNPs retained overall significance in the joint analysis (**Table 4.2**). **Table 4.1** also lists the OR, 95% confidence interval (CI), false discovery rate (FDR), SNP call rate and MAF obtained from joint analysis.

The association of rs3935234 present on chromosome 20p11.21 was the strongest (*p*-value 1.81×10^{-12}, FDR 6.33×10^{-11}) of the markers in this study, and also was the only polymorphism with genome-wide significance (Bonferroni *p*-value $<6.33\times10^{-11}$). An OR of 0.69 (95% CI of 0.62–0.77) of the minor allele G indicated reduced risk. The other polymorphisms with high significance (10^{-4} to 10^{-6}) in joint analysis and conferring risk for breast cancer are in chromosomes 4, 5, 16 and 19. Of these, rs1092913 is located on chromosome 5p15.2 (*p*-value 1.89×10^{-6}, FDR 3.30×10^{-5}, OR (95% CI) 1.45 (1.24–1.69)), with the ropporin-1 like (*ROPN1L*) gene present 2.5 kb downstream of the polymorphism; the three SNPs present on chromosome 19q13.33, *ZNF577* (zinc finger protein 577) gene are (i) rs10411161 (*p*-value 7.09×10^{-6}, FDR 8.27×10^{-5}) located in the 3′ untranslated region (UTR; 2.8 kb downstream of the stop codon, Goldenpath-hg 18/db SNP build 130); (ii) rs3848562 (*p*-value 9.23×10^{-6}, FDR 8.08×10^{-5}); and (iii) rs11878583 (*p*-value 1.35×10^{-4}, FDR 9.45×10^{-4}) located in the introns 6 and

2, respectively. We observed ORs (95% CI) of 1.42 (1.22–1.65), 1.42 (1.22–1.66), 1.35 (1.16–1.57), respectively, for the three *ZNF577* SNPs. The sixth SNP, rs1429142 is located on chromosome 4q31.23 (*p*-value 3.60×10^{-4}, FDR 2.10×10^{-3}), with *EDNRA* (endothelin receptor type A) gene present approximately 112.5 kb downstream of the polymorphism. An OR of 1.27 (95% CI of 1.11–1.45) was noted for the minor allele C. Lastly, rs1981867 located on chromosome 16q23.2 showed satisfactory statistical significance in Stage I (*p*-value 3.7×10^{-4}) and in joint analysis (*p*-value 4.32×10^{-4}, FDR 2.16×10^{-3}) but showed only marginal significance in Stage II (*p*-value 0.03). An OR of 1.22 (95% CI of 1.09–1.36) for the minor allele A was noted.

Table 4.1: Seven novel loci showing consistent association with breast cancer in both stages of the study

dbSNP rs#	Chr	AG	RL	Stage2 χ^2 *p*	Joint analysis				
					χ^2 *p*	FDR	OR (95% CI)	MA	MAF
rs3935234	20p11.21	*C20orf56*	93 kb DS	7.80E-11	1.81E-12	6.33E-11	0.69 (0.62,0.77)	G	0.43
rs1092913	5p15.2	*ROPN1L*	2.5 kb DS	2.17E-04	1.89E-06	3.30E-05	1.45 (1.24,1.69)	T	0.13
rs10411161	19q13.33	*ZNF577*	3' UTR	6.16E-04	7.09E-06	8.27E-05	1.42 (1.22,1.65)	T	0.13
rs3848562	19q13.33	*ZNF577*	Intron	9.78E-04	9.23E-06	8.08E-05	1.42 (1.22,1.66)	C	0.12
rs11878583	19q13.33	*ZNF577*	Intron	7.59E-03	1.35E-04	9.45E-04	1.35 (1.16,1.57)	C	0.13
rs1429142	4q31.23	*EDNRA*	112 kb US	1.28E-02	3.59E-04	2.10E-03	1.27 (1.11,1.45)	C	0.18
rs1981867	16q23.2	*C16orf61*	85.9 kb DS	3.17E-02	4.32E-04	2.16E-03	1.22 (1.09,1.36)	A	0.31

Chr, chromosome; AG, associated genes; *C20orf56*, chromosome 20 open reading frame 56; *ROPN1L*, ropporin-1 like; *ZNF577*, zinc finger 577; *EDNRA*, endothelin receptor A; *C16orf61*, chromosome 16 open reading frame 61; RL, relative location; kb, kilobases; UTR, untranslated region; DS, downstream; US, upstream; χ^2, chi-square; FDR, false-discovery rate; OR, odds ratio; CI, confidence interval; MA, minor allele; MAF, minor allele frequency.

Table 4.2: Polymorphisms not significant in Stage II or in joint analysis

dbSNP rs#	Chr	AG	Stage2 $\chi^2 p$	Joint Analysis				
				χ^2 p	FDR	OR (95% CI)	MA	MAF
rs6478296	9q33.1	*ASTN2*	1.91E-01	2.74E-03	1.20E-02	0.85 (0.77,0.95)	C	0.41
rs6997395	8q22.1	*PTDSS1/SDC2*	1.61E-01	5.46E-03	2.12E-02	0.86 (0.77,0.96)	C	0.30
rs9630178	11p12	*LRRC4C/RAG2*	1.68E-01	5.47E-03	1.91E-02	1.25 (1.07,1.47)	G	0.12
rs10506269	12q13.11	*AMIGO2/SLC38A4*	3.44E-01	1.50E-02	4.76E-02	0.81 (0.69,0.96)	G	0.10
rs1059307	6q14.3	*SNHG5*	2.99E-01	1.76E-02	5.15E-02	0.88 (0.80,0.98)	G	0.47
rs7908500	10q26.13	*OAT/CHST15*	7.00E-01	3.20E-02	8.62E-02	1.12 (1.01,1.25)	C	0.38
rs2080976	5q34	*ODZ2*	5.43E-01	3.70E-02	9.25E-02	1.12 (1.01,1.25)	A	0.35
rs7099921	10p13	*OPTN/CCDC3*	6.12E-01	3.84E-02	8.96E-02	1.15 (1.01,1.31)	G	0.22
rs2546513	12q15	*NUP107*	6.08E-01	4.34E-02	9.50E-02	0.89 (0.80,1.00)	G	0.29
rs10794182	10q26.13	*OAT/CHST15*	9.09E-01	6.02E-02	1.24E-01	1.11 (1.00,1.23)	G	0.38
rs11138489	9q21.31	*TLE1/TLE4*	9.63E-01	7.08E-02	1.38E-01	0.87 (0.75,1.01)	A	0.13
rs8075722	17p13.3	*OR3A2/OR1D5*	9.43E-01	8.16E-02	1.50E-01	1.17 (0.98,1.38)	A	0.10
rs11195949	10q25.2	*ACSL5*	9.68E-01	1.22E-01	2.14E-01	0.92 (0.83,1.02)	T	0.50
rs11257153	10p14	*USP6NL*	9.23E-01	1.34E-01	2.23E-01	0.89 (0.77,1.04)	A	0.15

Table 4.2 continued…

dbSNP rs#	Chr	AG	Stage2 χ^2 p	Joint Analysis				
				χ^2 p	FDR	OR (95% CI)	MA	MAF
rs6493076	15q15.2	*UBR1*	9.79E-01	1.47E-01	2.34E-01	0.89 (0.75,1.04)	A	0.11
rs6561682	13q14.3	*LECT1/SUGT1*	6.03E-01	2.04E-01	3.11E-01	1.07 (0.96,1.20)	G	0.31
rs1857434	9p21.3	*MLLT3/SLC24A2*	8.88E-01	2.05E-01	2.99E-01	1.09 (0.95,1.25)	G	0.18
rs1911864	5p14.3	*GUSBL2/CDH18*	6.90E-01	2.09E-01	2.93E-01	1.07 (0.96,1.18)	T	0.43
rs13299280	9q21.31	*TLE1/TLE4*	6.60E-01	2.75E-01	3.70E-01	0.92 (0.80,1.07)	G	0.15
rs6852237	4q35.1	*DCTD/ODZ3*	5.85E-01	3.27E-01	4.24E-01	1.05 (0.95,1.17)	G	0.43
rs9644134	8p21.1	*C8orf80*	6.97E-03	3.64E-01	4.55E-01	1.05 (0.95,1.16)	C	0.41
rs7119677	11p13	*C11orf41*	7.33E-03	4.00E-01	4.83E-01	1.05 (0.94,1.18)	T	0.26
rs6991277	8q22.1	*SDC2/PTDSS1*	3.99E-02	6.40E-01	7.47E-01	1.04 (0.89,1.20)	C	0.13
rs8095374	18q21.1	*C18orf25*	1.07E-01	7.85E-01	8.86E-01	0.99 (0.89,1.09)	A	0.47
rs12433708	14q23.2	*PPP2R5E*	1.79E-01	8.55E-01	9.35E-01	0.99 (0.85,1.14)	A	0.15
rs268840	14q23.1	*SLC35F4/C14orf105*	9.17E-02	9.12E-01	9.67E-01	0.99 (0.89,1.11)	A	0.36
rs7818355	8q22.1	*SDC2/PTDSS1*	8.64E-02	9.49E-01	9.77E-01	1.01 (0.86,1.17)	G	0.12
rs1451991	8q21.13	*LOC728643/SNX16*	1.25E-01	9.95E-01	9.95E-01	1.00 (0.88,1.14)	G	0.19

Chr, chromosome; AG, associated genes; *ASTN2*, astrotactin 2; *PTDSS1*/*SDC2*, phosphatidylserine synthase 1/Syndecan-2; *LRRC4C*/*RAG2*, leucine rich repeat containing 4C/recombination activating gene 2; *AMIGO2*/*SLC38A4*, adhesion molecule with Ig-like domain 2/solute carrier family 38, member 4; *SNHG5*, small nucleolar RNA host gene 5; *OAT*/*CHST15*, ornithine aminotransferase/carbohydrate (N-acetylgalactosamine 4-sulfate 6-O) sulfotransferase 15; *ODZ2*, odd Oz/ten-m homolog 2 (*Drosophila*); *OPTN*/*CCDC3*, optineurin/coiled-coil domain containing 3; *NUP107*, nucleoporin 107kDa; *TLE1*/*TLE4*, transducin-like enhancer of split 1/4 (E(sp1) homolog, *Drosophila*) *OR3A2*/*OR1D5*, olfactory receptor, family 3, subfamily A, member 2/olfactory receptor, family 1, subfamily D, member 5; *ACSL5*, acyl-CoA

synthetase long-chain family member 5; *USP6NL*, USP6 N-terminal like; *UBR1*, ubiquitin protein ligase E3 component n-recognin 1; *LECT1*/*SUGT1*, leukocyte cell derived chemotaxin 1/SGT1, suppressor of G2 allele of SKP1 (*S. cerevisiae*); *MLLT3*/*SLC24A2*, myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, *Drosophila*)/solute carrier family 24 (sodium/potassium/calcium exchanger), member 2; *GUSBL2*/*CDH18*, glucuronidase, beta-like 2/cadherin 18, type 2; *DCTD*/*ODZ3*, dCMP deaminase/odd Oz/ten-m homolog 3 (*Drosophila*); *C8orf80*, chromosome 8 open reading frame 80; *C11orf41*, chromosome 11 open reading frame 41; *C18orf25*, chromosome 18 open reading frame 25; *PPP2R5E*, protein phosphatase 2, regulatory subunit B', epsilon isoform; *SLC35F4*/*C14orf105*, solute carrier family 35, member F4/chromosome 14 open reading frame 105; *LOC728643*/*SNX16*, sorting nexin 16; χ^2, chi-square; FDR, false-discovery rate; OR, odds ratio; CI, confidence interval; MA, minor allele; MAF, minor allele frequency.

4.2 Discussion

Increasingly, assessing genetic risk in complex traits requires identification of multiple loci conferring risk and/or mining of the data in an integrated manner to identify potential gene–gene interactions that together may explain a higher proportion of risk than the single locus analysis (119, 120). This approach calls for identification of potential novel risk-associated SNPs, and their subsequent validation to improve the accuracy of genetic risk assessment models. We performed a whole genome analysis to identify novel markers (single locus analysis) associated with breast cancer, and confirmed several new loci in a larger, independent replication set of cases and controls.

The conduct of our study was appropriate in that the sample size used in our Stage I GWAS (348 cases and 348 controls) closely matched those used in Stage I of earlier studies, *e.g.*, 390 familial breast cancer cases and 364 controls used by Easton et al. (12) and 249 Ashkenazi Jew familial cases and 299 controls used by Gold et al. (13). Furthermore, this was only the second study, after Zheng et al. (19), to use high density Affymetrix SNP arrays (906,600 SNPs/array), which provides a vast physical coverage of the genome in an unbiased manner to identify markers associated with disease susceptibility. A precedent exists for using high-density SNP arrays to identify breast cancer susceptibility loci, and to subsequently replicate those identified variants in large cohorts (8, 11-19). The SNPs identified, so far, in GWASs including our study were largely surrogate

markers; fine mapping and independent validation studies are underway to identify causal variants.

In this study we have identified polymorphisms in open reading frames associated with breast cancer: (i) rs3935234 in the vicinity of the open reading frame from chromosome 20 (*C20orf56*). *C20orf56* found 93.2 kb downstream of rs3935234 was not previously implicated in breast cancer, but, Korkola et al. (121) showed that the expression of the gene has a strong predictive power for yolk sac non-seminomatous male germ-cell tumours. (ii) SNP rs1981867 in the open reading frame on chromosome 16 (*C16orf61*) showed significant association with breast cancer. *C16orf61* gene has been shown to be associated with multi-drug resistance (122). It has also been established that loss of heterozygosity in the q arm of chromosome 16 is a common genetic event in breast cancer resulting in copy number changes and sometimes the loss of the arm (76, 123, 124). This led to the speculation that the region may harbour tumour-suppressor genes (125). Easton et al. (12) and Stacey et al. (16) previously reported that rs3803662 positioned on chromosome 16q showed association with breast cancer, in two independent GWAS. The identified SNP, rs1981867 in this study from chromosome 16 further emphasise the importance on this region in breast cancer. While these results and interpretations require larger scale studies and independent confirmation, repeated and independent observations by several research groups suggesting breast cancer risk related to these open reading frames

underscores the importance of this region. Functional characterisation of these variants is warranted.

Zinc finger proteins are commonly involved in transcriptional regulation of genes. Tan et al. (126) have shown that C-terminal transcriptional repression domain of zinc finger protein ZBRK1 interacts with *BRCA1* tumour suppressor gene to repress transcription. Previous linkage studies have shown that mutations in *BRCA1* gene are a common event in early-onset, multiple-case breast cancer families (1). The C-terminal extension of *ZNF577* shares sequence homology with ZBRK1 (127). It remains to be determined whether *ZNF577* identified in our study also plays a role in transcriptional repression by binding to the *BRCA1* protein. We found three SNPs from *ZNF577* gene region to be associated with disease susceptibility: rs10411161 found 2.8 kb downstream of the gene in the 3′ UTR and the other two markers rs3848562 and rs11878583 are present in the introns 2 and 6, respectively. *ZNF577* has also been shown to be differentially expressed in tissues and cells (128). Similarly, a recent GWAS has shown a polymorphism rs10995190 located within the intron 4 of zinc finger protein 365 (*ZNF365*) to be associated with breast cancer susceptibility (8). Further studies are required to understand the functional role of *ZNF577*.

In a steady-state condition of the cell, endothelin receptor type A (*EDNRA*) coupled with endothelin-1 plays an important role in tissue differentiation and development, cell proliferation and hormone production (129).

Interestingly, constitutive co-expression of endothelin-1 growth factor and *EDNRA* often results in ovarian carcinoma (130), and also contributes to bone metastases in different primary tumours (131). Previous studies have shown that polymorphisms in *EDNRA* gene are associated with pulse pressure in myocardial infarction (132), idiopathic dilated cardiomyopathy (133) and hypertension (134). The *EDNRA* gene is located 112.5 kb upstream of the polymorphism rs1981867 which might act as a surrogate marker to identify the causal variant associated with breast cancer susceptibility.

The *ROPN1L* gene on chromosome 5p15.2 is present 2.5 kb downstream of the polymorphism rs1092913. There is no previous evidence on the association of the gene to breast cancer susceptibility. The *ROPN1L* gene encodes for a sperm protein known to interact with A-kinase anchoring protein (135). It is evident from previous GWAS that the p arm of chromosome 5 harbours several polymorphisms implicated in breast cancer susceptibility (17, 18). Also, Lowe et al. (136) showed that *ROPN1L* gene is highly expressed in pancreatic cancer when compared to the normal pancreatic tissues and other tumours in their dataset. Again, our results motivate further investigation in this gene.

A future study will explore association of the reported polymorphisms to sub-phenotypes of breast cancer, i.e., receptor status. A preliminary analysis of the novel SNPs reported in this communication also showed association to ER^+ breast cancers, an observation consistent with the previously characterized

polymorphisms from GWAS. Replication of our findings in Stage III in Albertan population and independent validation of these findings in cohorts elsewhere will help explore the biological relevance of these findings.

5 Conclusions and future work

The post-genomic era has awakened the need to understand the genetic underpinnings of complex traits. A major role of geneticists is to associate phenotypes with genotypes. Familial clustering of breast cancer was evident with the identification of risk due to mutations of high penetrance in *BRCA1* (1) and *BRCA2* (2) tumour suppressor genes. Subsequently, certain genes of moderate penetrance such as *ATM* (6), *CHEK2* (67), and *PALB2* (5) were shown to indicate predisposition to breast cancer susceptibility. However, these genes account only for a small proportion of the genetic risk. Intensive research efforts to identify *BRCA*-like genes to explain the familial risk have not been successful (9, 10). Unlike single-gene disorders which follow a Mendelian pattern of inheritance, complex disorders are often caused by multiple interacting disease genes. This led Pharoah et al. (69) to propose the polygenic nature of disease susceptibility to explain the remaining risk in non-familial or sporadic breast cancer cases. This model enables identification of several common genetic variants that each individually confers only modest risk for the disease. The unexplained fraction of the disease heritability could possibly be explained by the low-penetrance variants. In addition, several confounding factors such as disease heterogeneity, environment, life-style, and low penetrance nature of the allele make the identification of disease causative genes more challenging.

Recent years have seen a rapid surge in the number of loci identified to be associated with breast cancer susceptibility. Some of these have been successfully replicated in the original study populations and/or validated in several other studies in different populations providing confirmation of the initial findings. Some of the reasons for rising success in the replication of the findings are the accuracy of genotype calling by various genotyping platforms and the highly evolved quality control measures applied to datasets to filter out polymorphisms that showed strong association with the disease in initial studies.

Over the past three years, there have been ten major whole genome studies identifying risk alleles associated with breast cancer risk (8, 11-19). Most of these studies were conducted in women of European ancestry with the exception of two, which investigated risk alleles in Chinese (19) and Ashkenazi Jewish (13) populations. The women of European ancestry included in the earlier studies were mostly affected and unaffected individuals from Europe and USA. There is no known study with the exclusive emphasis on gaining insight into the genetic architecture of a complex disease such as breast cancer in Canadian population. Although a majority of the Canadian population have a European ancestry, it is reasonable to assume that the exposure to different environmental and life-style factors has an impact on disease predisposition. It is of interest to explore the allelic architecture of an ethnically defined Canadian population (Province of Alberta). The work summarized in this thesis is an initial step to reach the ultimate goal of obtaining the allelic architecture of susceptibility to breast cancer.

This thesis has presented the results of screening the most abundant genetic polymorphism (SNPs) in the genome in the context of breast cancer phenotype. The study reported here represents the 11th GWAS in the literature and 9th one from the Caucasian study population, and the first independent GWAS in Canada as well as in Alberta. There were two main objectives of the research described in this thesis:

5.1 Validation of candidate polymorphisms

Many SNPs have been identified by GWAS to be associated with breast cancer risk and subsets of them have been replicated and/or validated in different populations. One of the objectives of this thesis was to seek relevance of the reported risk alleles from whole genome scans to our study population in a case–control setting.

We genotyped 28 SNPs previously identified in the literature from GWAS in 1439 breast cancer cases and 1536 from women predominantly of Caucasian origin. An overall case–control association analysis showed significant associations for 14 of the 28 SNPs with breast cancer risk in our study population. The direction of odds ratio (>1) of all significant SNPs indicated that the minor allele of the SNP conferred risk to disease susceptibility. The strongest associations were found with SNPs from intron 2 of the *FGFR2* gene region, which helps confirm original findings and other validation studies. Due to the

potentially different genetic underpinnings of breast cancer subtypes, we conducted an independent subgroup analysis based on the case receptor status – ER^+, PR^+ and $Her2^-$. We confirmed that a majority of SNPs/loci – *FGFR2*, 5p12, 8q, *TNRC9*, *SLC4A7*, *MAP3K1* and *HDAC8* – interrogated showed strong association with breast cancer phenotypes exhibiting ER^+, PR^+ and $Her2^-$ receptor status.

A recent study evaluated the effect size ("$ES=-2\beta^2f(1-f)$, where the coefficient β measures the regression effect of the locus per copy of the variant allele, and f denotes the MAF" as defined by Park et al. (119)) and the power of five breast cancer susceptibility variants based on OR and MAF of the initial study and determined the expected number of loci yet to be identified for a given power and effect size. The effect sizes of the five SNPs (rs3817198, rs13281615, rs3803662, rs2981582 and rs889312) were in the range of 0.002–0.025 and power of the study in the range of 0.010–0.930. Four out of the five SNPs reported in the literature showed association with breast cancer in our study population. The SNPs reported as having high effect size and statistical power were the ones that we could readily replicate. The SNPs that showed replication in only one stage have to be evaluated at a higher sample size in future as these may potentially represent the low effect size as defined by Park et al. (119) or owing to disease heterogeneity or even some yet to be identified fine population-specific differences.

Having characterized our cohort for reproducibility of findings from the reported GWASs, we designed an independent GWAS using the same set of cases and controls from Alberta to detect additional (novel) associations that may have been missed by others.

5.2 Replication study and joint analysis

Considering several large sample size studies have identified susceptibility alleles from different gene regions, the effect size of each of the variants is small. A few GWASs may not be sufficient to identify most of the risk alleles. Therefore, it is of continuing importance to explore the full spectrum of breast cancer susceptibility loci by conducting GWASs in women of European ancestry and ethnically diverse populations. Our second objective was to replicate subset of markers identified from our preliminary analysis, i.e., ones not previously reported in the literature (11-19), in independent cohorts.

We performed a two-stage association study in a cohort of 3064 women from Alberta to identify novel loci associated with breast cancer susceptibility. In Stage I of our association study, we genotyped 348 breast cancer cases and 348 control subjects using an Affymetrix SNP 6.0 array featuring 906,600 SNPs. The dataset was subjected to stringent SNP and sample quality control measures and the SNPs that adhered to the requirements were included in the subsequent analysis. Overall allelic and haplotype association analyses were carried out in the

enriched dataset. Based on stringent selection criteria, 35 markers were selected for replication in the subsequent stage. In Stage II, we attempted to replicate the 35 significant markers in an independent study of 1153 cases and 1215 controls. Genotyping was carried out using the Sequenom iPlex technology. An allelic association analysis – individually for Stage II and joint analysis (combining all samples) – was performed to determine associations with the disease. We identified seven loci from five different gene regions (chromosomes 4, 5, 16, 19 and 20) that showed statistically significant differences between cases and controls in both Stage I and Stage II testing, and also in joint analysis. Although these loci have been independently replicated within this study, they warrant further evaluation in larger studies both within Alberta and in other populations.

5.3 Recommendations for future work

This study identified a number of possible directions for future work. Some of these are as follows:

(i) Replication of the markers (Stage III) in yet another independent set of cases and controls would provide confirmatory evidence of the current findings. Repeat testing of association in a larger sample size and conducting a joint analysis will enhance the power of the study and accuracy of the results. This would give a pointer to select the candidate polymorphisms for fine mapping studies. Performing a subgroup analysis based on subphenotypes of breast

cancer will provide additional insights into the molecular basis of the disease. Since receptor status is both a prognostic and predictive factors and holds direct relevance to specific treatments, stratifying cases based on different combinations of receptor status such as (ER^+, PR^+, $Her2^-$), (ER^+, PR^+, $Her2^+$), (ER^-, PR^-, $Her2^+$), and (ER^-, PR^-, $Her2^-$) will help identify novel markers missed in overall association analysis.

(ii) Apart from genotyping common variants, it is essential to genotype rare variants which are not generally interrogated in the available whole genome arrays. Identifying and acquiring information on both common and rare variants will help fill in the missing gaps in genetic heritability information of the disease. It will enable us to choose candidate polymorphisms (strongly associated with breast cancer) with high precision for functional studies.

(iii) It may be valuable to consider gene expression signatures in the candidate polymorphisms selection process. This strategy is largely justified when a given SNP or haplotype is present in the promoter and/or 3' UTR of the gene. These regions play a critical role in the initiation of transcription and stability of mRNA, respectively. This approach would give a confirmation of near causativeness of the gene or the polymorphisms around the gene. Functional validation after this confirmation would be more meaningful.

(iv) It will also be beneficial to conduct a whole genome haplotype association analysis instead of single locus association analysis because it will provide an idea about the evolutionary pattern at population level and also reduce the number of hypotheses tested, thereby increasing the power of the study.

(v) It would also be interesting to study the polymorphisms in tumour DNA, which may provide useful clues on the expressed functional markers. A comparison of genotypes for a particular set of SNPs between the lymphocyte DNA and tumour DNA from the same individual will help display differential DNA level signatures indicative of tumour cell derived changes (allelic imbalance or loss of heterozygosity).

As demonstrated in this thesis, there is potential to identify novel variants associated with disease susceptibility through GWASs. It is of continuing importance to conduct whole genome and validation studies in diverse populations to exhaustively identify susceptibility loci associated with breast cancer risk. It will then be meaningful to perform an integrated analysis to identify potential gene–gene interactions that together may explain a higher proportion of disease risk.

6 Bibliography

1. Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990;250(4988):1684-9.
2. Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature* 1995;378(6559):789-92.
3. Li J, Yen C, Liaw D, et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* 1997;275(5308):1943-7.
4. Malkin D, Li FP, Strong LC, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 1990;250(4985):1233-8.
5. Rahman N, Seal S, Thompson D, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* 2007;39(2):165-7.
6. Renwick A, Thompson D, Seal S, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* 2006;38(8):873-5.
7. Garcia-Closas M, Chanock S. Genetic susceptibility loci for breast cancer by estrogen receptor status. *Clin Cancer Res* 2008;14(24):8000-9.
8. Turnbull C, Ahmed S, Morrison J, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* 2010;42(6):504-7.

9. Pharoah PD, Antoniou AC, Easton DF, et al. Polygenes, risk prediction, and targeted prevention of breast cancer. *N Engl J Med* 2008;358(26):2796-803.

10. Smith P, McGuffog L, Easton DF, et al. A genome wide linkage search for breast cancer susceptibility genes. *Genes Chromosomes Cancer* 2006;45(7):646-55.

11. Ahmed S, Thomas G, Ghoussaini M, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet* 2009;41(5):585-90.

12. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 2007;447(7148):1087-93.

13. Gold B, Kirchhoff T, Stefanov S, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci U S A* 2008;105(11):4340-5.

14. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* 2007;39(7):870-4.

15. Murabito JM, Rosenberg CL, Finger D, et al. A genome-wide association study of breast and prostate cancer in the NHLBI's Framingham Heart Study. *BMC Med Genet* 2007;8 Suppl 1:S6.

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

17. Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* 2008;40(6):703-6.

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

19. Zheng W, Long J, Gao YT, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* 2009;41(3):324-8.

20. Cox A, Dunning AM, Garcia-Closas M, et al. A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* 2007;39(3):352-8.

21. Kammerer S, Roth RB, Reneland R, et al. Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus. *Cancer Res* 2004;64(24):8906-10.

22. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med* 2009;1(6):62.

23. Iorio MV, Ferracin M, Liu CG, et al. MicroRNA gene expression deregulation in human breast cancer. *Cancer Res* 2005;65(16):7065-70.

24. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. *N Engl J Med* 2009;360(8):790-800.

25. Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am J Hum Genet* 2007;81(6).

26. Huijts PE, Vreeswijk MP, Kroeze-Jansema KH, et al. Clinical correlates of low-risk variants in FGFR2, TNRC9, MAP3K1, LSP1 and 8q24 in a Dutch cohort of incident breast cancer cases. *Breast Cancer Res* 2007;9(6):R78.

27. Mavaddat N, Dunning AM, Ponder BA, et al. Common genetic variation in candidate genes and susceptibility to subtypes of breast cancer. *Cancer Epidemiol Biomarkers Prev* 2009;18(1):255-9.

28. International HapMap Consortium. The International HapMap Project. *Nature* 2003;426(6968):789-96.

29. Pearson TA, Manolio TA. How to interpret a genome-wide association study. *JAMA* 2008;299(11):1335-44.

30. Igl BW, Konig IR, Ziegler A. What do we mean by 'replication' and 'validation' in genome-wide association studies? *Hum Hered* 2009;67(1):66-8.

31. Gudmundsson J, Sulem P, Manolescu A, et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet* 2007;39(5):631-7.

32. Yeager M, Orr N, Hayes RB, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet* 2007;39(5):645-9.

33. Carrasquillo MM, Zou F, Pankratz VS, et al. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nat Genet* 2009;41(2):192-8.

34. Hindorff LA JH, Hall PN, Mehta JP, Manolio TA. A Catalog of Published Genome-Wide Association Studies. (www.genome.gov/gwastudies). (Accessed 16 June 2009).

35. Schork NJ, Murray SS, Frazer KA, et al. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 2009;19(3):212-9.

36. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;6(2):95-108.

37. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* 2003;4(8):587-97.

38. Futreal PA, Liu Q, Shattuck-Eidens D, et al. BRCA1 mutations in primary breast and ovarian carcinomas. *Science* 1994;266(5182):120-2.

39. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002;3(5):391-7.

40. Boerwinkle E, Hixson JE, Hanis CL. Peeking under the peaks: following up genome-wide linkage analyses. *Circulation* 2000;102(16):1877-8.

41. Keith T. Human genome-wide association studies. *Genet Eng Biotechnol News*, 2007:1.

42. Stoneking M. Single nucleotide polymorphisms. From the evolutionary past. *Nature* 2001;409(6822):821-2.

43. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118(5):1590-605.

44. Kim S, Misra A. SNP genotyping: technologies and biomedical applications. *Annu Rev Biomed Eng* 2007;9:289-320.

45. Li WH, Sadler LA. Low nucleotide diversity in man. *Genetics* 1991;129(2):513-23.

46. Liu YL, Fann CS, Liu CM, et al. A single nucleotide polymorphism fine mapping study of chromosome 1q42.1 reveals the vulnerability genes for schizophrenia, GNPAT and DISC1: Association with impairment of sustained attention. *Biol Psychiatry* 2006;60(6):554-62.

47. Flaherty DK. Single nucleotide polymorphisms, drug metabolism and untoward health effects. *J Med Biol Sci* 2007;1(2).

48. Hesketh J. 3'-Untranslated regions are important in mRNA localization and translation: lessons from selenium and metallothionein. *Biochem Soc Trans* 2004;32(Pt 6):990-3.

49. NCBI. (http://www.ncbi.nlm.nih.gov/)

50. Smigielski EM, Sirotkin K, Ward M, et al. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 2000;28(1):352-5.

51. Norris S. The human genome project and beyond: Canada's role. *Parliamentary Information and Research Service*, 2005.

52. HGP. Human Genome Project Information. (http://www.ornl.gov/). (Accessed 3 March 2010).

53. Bentley DR. The Human Genome Project--an overview. *Med Res Rev* 2000;20(3):189-96.

54. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409(6822):928-33.

55. International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005;437(7063):1299-320.

56. International HapMap Consortium, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449(7164):851-61.

57. HapMap. (http://hapmap.ncbi.nlm.nih.gov/). (Accessed 19 Oct 2009).

58. Affymetrix. (www.affymetrix.com). (Accessed April 2010).

59. Perkel J. SNP genotyping: six technologies that keyed a revolution. *Nat Meth* 2008;5(5):447-53.

60. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994;266(5182):66-71.

61. Venkitaraman AR. Functions of BRCA1 and BRCA2 in the biological response to DNA damage. *J Cell Sci* 2001;114(Pt 20):3591-8.

62. Gao Q, Adebamowo CA, Fackenthal J, et al. Protein truncating BRCA1 and BRCA2 mutations in African women with pre-menopausal breast cancer. *Hum Genet* 2000;107(2):192-4.

63. Shiri-Sverdlov R, Oefner P, Green L, et al. Mutational analyses of BRCA1 and BRCA2 in Ashkenazi and non-Ashkenazi Jewish women with familial breast and ovarian cancer. *Hum Mutat* 2000;16(6):491-501.

64. Thirthagiri E, Lee SY, Kang P, et al. Evaluation of BRCA1 and BRCA2 mutations and risk-prediction models in a typical Asian country (Malaysia) with a relatively low incidence of breast cancer. *Breast Cancer Res* 2008;10(4):R59.

65. Balmain A, Gray J, Ponder B. The genetics and genomics of cancer. *Nat Genet* 2003;33 Suppl:238-44.

66. Gatti RA, Tward A, Concannon P. Cancer risk in ATM heterozygotes: a model of phenotypic and mechanistic differences between missense and truncating mutations. *Mol Genet Metab* 1999;68(4):419-23.

67. Meijers-Heijboer H, van den Ouweland A, Klijn J, et al. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* 2002;31(1):55-9.

68. Seal S, Thompson D, Renwick A, et al. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet* 2006;38(11):1239-41.

69. Pharoah PD, Antoniou A, Bobrow M, et al. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* 2002;31(1):33-6.

70. Moffa AB, Ethier SP. Differential signal transduction of alternatively spliced FGFR2 variants expressed in human mammary epithelial cells. *J Cell Physiol* 2007;210(3):720-31.

71. Dickson C, Spencer-Dene B, Dillon C, et al. Tyrosine kinase signalling in breast cancer: fibroblast growth factors and their receptors. *Breast Cancer Res* 2000;2(3):191-6.

72. Meyer KB, Maia AT, O'Reilly M, et al. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biol* 2008;6(5):e108.

73. Inman CK, Li N, Shore P. Oct-1 counteracts autoinhibition of Runx2 DNA binding to form a novel Runx2/Oct-1 complex on the promoter of the mammary gland-specific gene beta-casein. *Mol Cell Biol* 2005;25(8):3182-93.

74. Inman CK, Shore P. The osteoblast transcription factor Runx2 is expressed in mammary epithelial cells and mediates osteopontin expression. *J Biol Chem* 2003;278(49):48684-9.

75. Hulten MA, Hill SM, Rodgers CS. Chromosomes 1 and 16 in sporadic breast cancer. *Genes Chromosomes Cancer* 1993;8(3):204.

76. Cleton-Jansen AM, Callen DF, Seshadri R, et al. Loss of heterozygosity mapping at chromosome arm 16q in 712 breast tumors reveals factors that influence delineation of candidate regions. *Cancer Res* 2001;61(3):1171-7.

77. Smid M, Wang Y, Klijn JG, et al. Genes associated with breast cancer metastatic to bone. *J Clin Oncol* 2006;24(15):2261-7.

78. Chen Y, Choong LY, Lin Q, et al. Differential expression of novel tyrosine kinase substrates during breast cancer development. *Mol Cell Proteomics* 2007;6(12):2072-87.

79. Min J, Okada S, Kanzaki M, et al. Synip: a novel insulin-regulated syntaxin 4-binding protein mediating GLUT4 translocation in adipocytes. *Mol Cell* 1999;3(6):751-60.

80. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008;9(5):356-69.

81. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. *Nat Rev Genet* 2009;10(5):318-29.

82. Innocenti F ed. Genomics and Pharmacogenomics in Anticancer Drug Development and Clinical Response. Totowa, NJ: Humana Press, 2009.

83. Qiagen. QIAamp DNA Mini and Blood Mini handbook. (www.qiagen.com).

84. The Tomorrow Project. (http://www.cancerboard.ab.ca/tomorrow/).

85. Alberta Health Services. (http://www.albertahealthservices.ca).

86. Dufva M ed. DNA Microarrays for Biomedical Research: Methods and Protocols. Humana Press, 2009.

87. Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. *Biom J* 2008;50(1):8-28.

88. BROAD Institute. Birdsuite: Birdseed. (Accessed 27 July 2010).

89. Pham T ed. Computational Biology: Issues and Applications in Oncology. Springer, 2010.

90. Ryckman K, Williams SM. Calculation and use of the Hardy-Weinberg model in association studies. *Curr Protoc Hum Genet* 2008;Chapter 1:Unit 1 18.

91. Weber DS, Stewart BS, Garza JC, et al. An empirical genetic assessment of the severity of the northern elephant seal population bottleneck. *Curr Biol* 2000;10(20):1287-90.

92. Cooper RS, Rotimi CN, Ward R. The puzzle of hypertension in African-Americans. *Sci Am* 1999;280(2):56-63.

93. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461(7265):747-53.

94. Huentelman MJ, Craig DW, Shieh AD, et al. SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. *BMC Genomics* 2005;6:149.

95. Marchini J, Cardon LR, Phillips MS, et al. The effects of human population structure on large genetic association studies. *Nat Genet* 2004;36(5):512-7.

96. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38(8):904-9.

97. Cardon LR, Abecasis GR. Using haplotype blocks to map human complex trait loci. *Trends Genet* 2003;19(3):135-40.

98. Shaffer JP. Multiple Hypothesis Testing. *Ann Rev Psychol* 1995;46(1):561-84.

99. Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*. Rome, Italy, 1935:13-60.

100. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 1936;8:3-62.

101. Daniel PB, Werner D, Martin G. A Practical Approach to Microarray Data Analysis. Springer Publishing Company, Incorporated; 2009.

102. Korenberg MJ ed. Microarray Data Analysis: Methods and Applications. Totowa, NJ: Humana Press, 2007.

103. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B (Methodological)* 1995;57(1):289-300.

104. Storey J. A direct approach to false discovery rates. *J R Stat Soc B (Statistical Methodology)* 2002;64(3):479-98.

105. Gabriel S, Ziaugra L, Tabbaa D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr Protoc Hum Genet* 2009;Chapter 2:Unit 2 12.

106. Eswarakumar VP, Lax I, Schlessinger J. Cellular signaling by fibroblast growth factor receptors. *Cytokine Growth Factor Rev* 2005;16(2):139-49.

107. Raskin L, Pinchev M, Arad C, et al. FGFR2 is a breast cancer susceptibility gene in Jewish and Arab Israeli populations. *Cancer Epidemiol Biomarkers Prev* 2008;17(5):1060-5.

108. Ruiz-Narvaez EA, Rosenberg L, Rotimi CN, et al. Genetic variants on chromosome 5p12 are associated with risk of breast cancer in African

American women: the Black Women's Health Study. *Breast Cancer Res Treat* 2010; 123(2):525-30.

109. Rebbeck TR, DeMichele A, Tran TV, et al. Hormone-dependent effects of FGFR2 and MAP3K1 in breast cancer susceptibility in a population-based sample of post-menopausal African-American and European-American women. *Carcinogenesis* 2009;30(2):269-74.

110. Krusche CA, Wulfing P, Kersting C, et al. Histone deacetylase-1 and -3 protein expression in human breast cancer: a tissue microarray analysis. *Breast Cancer Res Treat* 2005;90(1):15-23.

111. Zhang J, Qiu LX, Wang ZH, et al. Current evidence on the relationship between three polymorphisms in the FGFR2 gene and breast cancer risk: a meta-analysis. *Breast Cancer Res Treat* (Epub ahead of print).

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

113. Udler MS, Meyer KB, Pooley KA, et al. FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum Mol Genet* 2009;18(9):1692-703.

114. Li L, Zhou X, Huang Z, et al. TNRC9/LOC643714 polymorphisms are not associated with breast cancer risk in Chinese women. *Eur J Cancer Prev* 2009;18(4):285-90.

115. Fletcher O, Johnson N, Gibson L, et al. Association of genetic variants at 8q24 with breast cancer risk. *Cancer Epidemiol Biomarkers Prev* 2008;17(3):702-5.

116. Provenzano PP, Eliceiri KW, Campbell JM, et al. Collagen reorganization at the tumor-stromal interface facilitates local invasion. *BMC Med* 2006;4(1):38.

117. Provenzano PP, Inman DR, Eliceiri KW, et al. Collagen density promotes mammary tumor initiation and progression. *BMC Med* 2008;6:11.

118. Skol AD, Scott LJ, Abecasis GR, et al. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38(2):209-13.

119. Park JH, Wacholder S, Gail MH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet* 2010;42(7):570-5.

120. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 2010;42(7):565-9.

121. Korkola JE, Houldsworth J, Dobrzynski D, et al. Gene expression-based classification of nonseminomatous male germ cell tumors. *Oncogene* 2005;24(32):5101-7.

122. Campone M, Campion L, Roche H, et al. Prediction of metastatic relapse in node-positive breast cancer: establishment of a clinicogenomic model

after FEC100 adjuvant regimen. *Breast Cancer Res Treat* 2008;109(3):491-501.

123. Cleton-Jansen AM, Moerland EW, Kuipers-Dijkshoorn NJ, et al. At least two different regions are involved in allelic imbalance on chromosome arm 16q in breast cancer. *Genes Chromosomes Cancer* 1994;9(2):101-7.

124. Cleton-Jansen AM, Buerger H, Haar N, et al. Different mechanisms of chromosome 16 loss of heterozygosity in well- versus poorly differentiated ductal breast cancer. *Genes Chromosomes Cancer* 2004;41(2):109-16.

125. Rakha EA, Green AR, Powe DG, et al. Chromosome 16 tumor-suppressor genes in breast cancer. *Genes Chromosomes Cancer* 2006;45(6):527-35.

126. Tan W, Zheng L, Lee WH, et al. Functional dissection of transcription factor ZBRK1 reveals zinc fingers with dual roles in DNA-binding and BRCA1-dependent transcriptional repression. *J Biol Chem* 2004;279(8):6576-87.

127. Tan W, Kim S, Boyer TG. Tetrameric oligomerization mediates transcriptional repression by the BRCA1-dependent Kruppel-associated box-zinc finger protein ZBRK1. *J Biol Chem* 2004;279(53):55153-60.

128. Chen R, Morgan AA, Dudley J, et al. FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome Biol* 2008;9(12):R170.

129. Nelson J, Bagnato A, Battistini B, et al. The endothelin axis: emerging role in cancer. *Nat Rev Cancer* 2003;3(2):110-6.

130. Salani D, Di Castro V, Nicotra MR, et al. Role of endothelin-1 in neovascularization of ovarian carcinoma. *Am J Pathol* 2000;157(5):1537-47.

131. Medinger M, Adler CP, Schmidt-Gersbach C, et al. Angiogenesis and the ET-1/ETA receptor system: immunohistochemical expression analysis in bone metastases from patients with different primary tumors. *Angiogenesis* 2003;6(3):225-31.

132. Nicaud V, Poirier O, Behague I, et al. Polymorphisms of the endothelin-A and -B receptor genes in relation to blood pressure and myocardial infarction: the Etude Cas-Temoins sur l'Infarctus du Myocarde (ECTIM) Study. *Am J Hypertens* 1999;12(3):304-10.

133. Charron P, Tesson F, Poirier O, et al. Identification of a genetic risk factor for idiopathic dilated cardiomyopathy. Involvement of a polymorphism in the endothelin receptor type A gene. CARDIGENE group. *Eur Heart J* 1999;20(21):1587-91.

134. Benjafield AV, Katyk K, Morris BJ. Association of EDNRA, but not WNK4 or FKBP1B, polymorphisms with essential hypertension. *Clin Genet* 2003;64(5):433-8.

135. Gene Cards. (http://www.genecards.org/). (Accessed 2010 5 May).

136. Lowe AW, Olsen M, Hao Y, et al. Gene expression patterns in pancreatic tumors, cells and tissues. *PLoS One* 2007;2(3):e323.

7 Appendix A: Preliminary GWAS: Stage I

Genome-wide analysis using high density arrays are very expensive and often a preliminary work in the laboratory typically contributes to the generation of hypotheses. This work referred to as *Stage I* should be followed by subsequent stages (multi-stage wherever possible) to confirm or extend the initial findings. The initial GWAS in Dr. Damaraju's laboratory was conducted with a sample size of 348 cases and 348 controls. Since I joined the laboratory in January 2008, I have also contributed to the generation of the dataset using the Affymetrix platform. However, this work has not been included in the main part of my thesis since it would be distracting to the main body of the thesis and the set study objectives; in addition, other members of Dr. Damaraju's laboratory also contributed to this effort. In Stage I of the study, participation of more than one person is not uncommon since the data complexity is huge and requires extensive bioinformatics support as well as technical support. Following the Stage I analysis using the Affymetrix platform, I selected polymorphisms from Affymetrix data (those showing statistical significance) and replicated using Sequenom Mass-ARRAY iPlex technology (using Stage I cohort and additional samples) to assess the cross-platform concordance and this has set the stage for the rest of my thesis work.

I preferred to summarize all this background work as part of an appendix to provide a window into association study design, selection of markers for subsequent replication and, finally, exhaustive attention that we paid to quality control aspects of generating this data. I was involved with all these steps of data generation and analysis. I am reporting the results presented in the main body of the thesis chapters entirely from my efforts and the data generated is a comprehensive summary of work (design, execution, analysis and interpretations) from the total case and control cohorts summarized originating from a single genotyping platform (Sequenom).

Data analysis of Stage I samples can be arbitrarily divided into two levels. The first-level analysis is the process of data review to enrich the dataset by applying SNP and sample quality control filters. The second-level analysis includes the allelic and haplotype association analyses using chi-square test to determine markers that are statistically significantly associated with breast cancer susceptibility. Results generated from the quality control filters and association analyses are presented in this section (**Figure 7.1**).

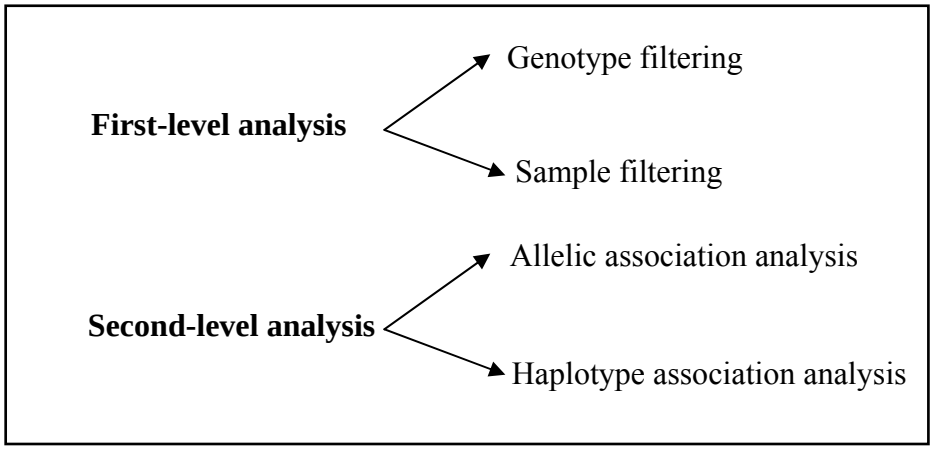


Figure 7.1: Diagram showing the two levels of analysis in Stage I association study.

7.1 Quality control filters

7.1.1 Genotype filtering

Genotype filtering is done at two levels: SNPs deviating from HWE and SNP call rate clean-up to address the missing genotype calls as described in the methods section.

Departure from HWE: When we applied a quality control filter of p <0.001 (i.e. any SNP that falls below the user-defined threshold), we identified significant deviations for 30,636 SNPs (3.38% of 906,600 SNPs) and these were excluded from further analysis.

SNP call rate clean-up: We restricted the analysis to those SNPs that had ≥99% genotype calls. A total of 93,126 SNPs (10.3% of 906,600 SNPs) did not provide complete genotype results with a stringent SNP call rate cut-off and these were excluded from further analysis.

Several GWASs of breast cancer have applied different quality control metrics for HWE (ranging from 10^{-10} to 0.02) and genotype completeness (ranging from 80% to 99.7%). We applied varying filtering parameters to determine the metric that best suits our dataset, also taking into account the number of markers retained for the final analysis (**Table 7.1**). Eliminating a large set of SNPs would mean losing some informative markers and the expenses incurred in genotyping them. After closely assessing the different cut-off, we chose to apply a *p*-value <0.001 for deviations from HWE and ≥99% for genotype completeness for our dataset.

Table 7.1: Varying cut-off values applied for HWE and missing genotype calls to determine the number of SNPs to be retained for downstream analysis

HWE	SNP call rate	SNPs retained	SNPs excluded
0.001	99.99%	506,836	399,764
0.01	99.99%	497,912	408,688
0.05	99.99%	475,096	431,504
0.001	**99%**	**782,838**	**123,762**
0.01	99%	768,818	137,782
0.05	99%	733,075	173,525
0.001	98%	831,106	75,494
0.01	98%	815,991	90,609
0.05	98%	777,827	128,773
0.001	95%	863,086	43,514
0.01	95%	846,906	59,694
0.05	95%	806,745	99,855

7.1.2 Evaluation of population stratification (Sample filtering)

Evaluation of population stratification was carried out to assess the genetic homogeneity of cases and controls. As discussed earlier, we used the principal components analysis-based method – EIGENSTRAT – developed by Price et al. (96), embedded in the HelixTree software to conduct the analysis. SNP data for cases and controls from our study were compared with the multi-ethnic cohort

from HapMap SNP (matched Affymetrix SNP 6.0) dataset. Firstly, principal components analysis plot was generated for the three different populations – European, African, and Asian – of the HapMap samples to demonstrate the distinct clustering of the three isolated populations in a two-dimensional plot (**Figure 7.2**). Due to the tight clustering, using HapMap samples as the reference will enable us to visually assess the confounding stratification in our study population.

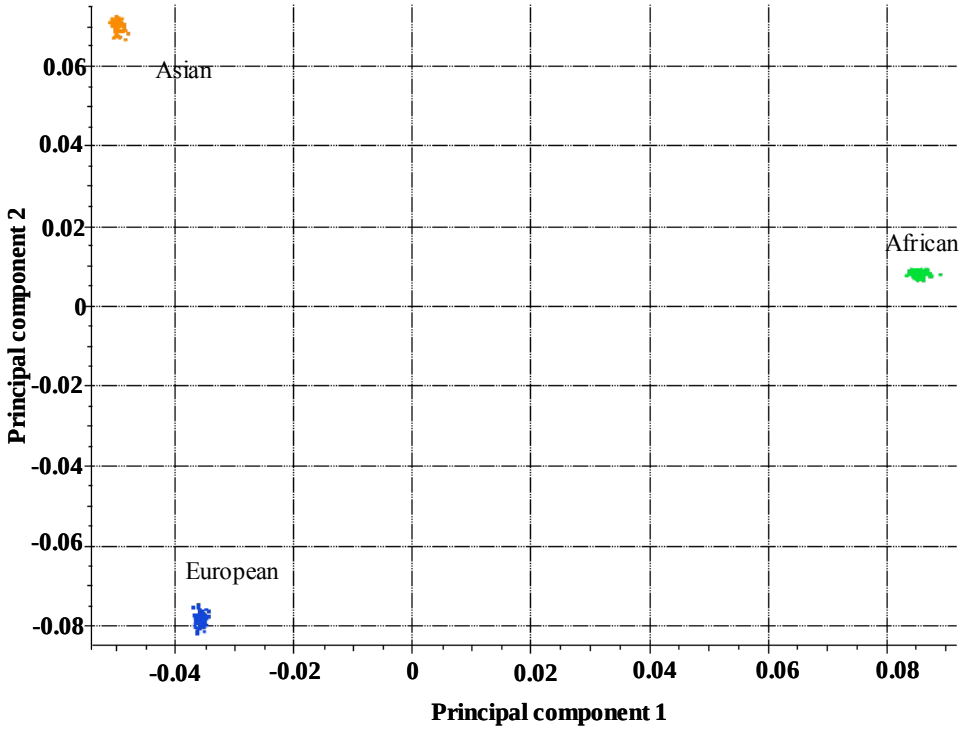


Figure 7.2: Distinct genotype clusters of three isolated populations of HapMap samples: European (blue), Asian (orange) and African (green). Principal components analysis plot is generated using two principal components. On the *x*-axis is the principal component 1 and on the *y*-axis is the principal component 2.

Secondly, we superimposed our entire study population (348 cases and 348 controls) onto HapMap samples without removing outliers to visualise the clustering of the samples. We found that most of our samples clustered around the European population as shown in **Figure 7.3**. Considering the demographics of Canada, the trail shown in the figure demonstrates admixture of African and Asian populations in our study population. The observed substructure can lead to false-positive associations in disease pathogenesis.

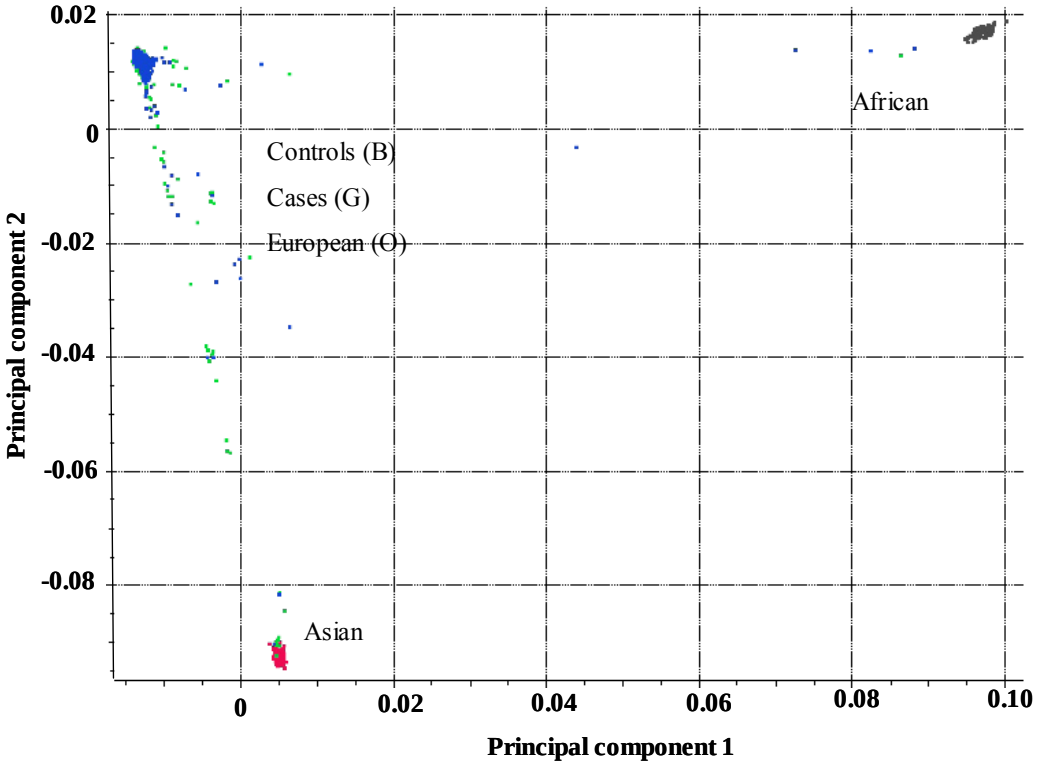


Figure 7.3: Genotype clusters after super-imposing our study population onto the HapMap samples without removing outliers. The cluster showing cases (green) and controls (blue) mostly overlap with the cluster from Central European population (orange) of the HapMap samples and also show some admixture of African (black) and Asian (pink) populations in our study population.

Thirdly, to address the stratification issue in our study population, we identified the samples that are not genetically homogeneous (away from the major cluster of Caucasian/Central European origin/descent) using EIGENSTRAT method. A total of 73 outliers were detected that were ≥3 standard deviations away from the mean on one of the two principal components. The detected outliers were removed from our dataset leaving with 302 cases and 321 controls for further scrutiny. Our samples were again super-imposed onto the HapMap samples to verify the genetic homogeneity of our study population. Case and control samples of this study showed significant overlap with Central European population cluster of HapMap samples (**Figure 7.4**). This indicates that our predominantly Caucasian population has (i) high genetic similarity with the European population as compared with Asian or Yoruba Indians (African); and (ii) the cases and controls from the Alberta region showed near genetic homogeneity, i.e., both appear to be of European ancestry.

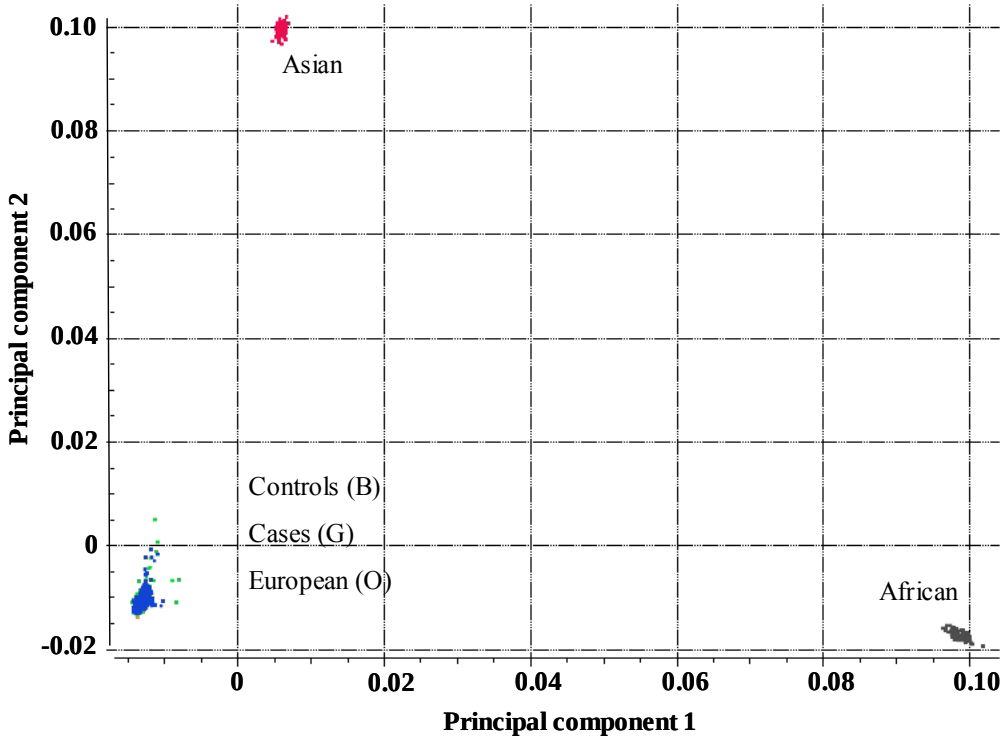


Figure 7.4: Genotype clusters after super-imposing our study population onto the HapMap samples after removing outliers. The cluster showing cases (green) and controls (blue) fully overlap with the cluster from Central European population (orange) of the HapMap samples in the figure.

7.2 GWAS in 348 cases and 348 controls (Stage I)

Allelic association analysis with 782,838 SNPs showed statistically significant (p <0.05) differences between the cases and the controls at multiple genomic locations (35,859 SNPs) scattered across all chromosomes. The best way to summarize the p-values of thousands of markers is by using a Manhattan plot. **Figure 7.5** is a visual representation of all statistically significant markers (plotted as $-\log_{10}$ chi-square values) grouped according to the chromosomes 1–22 and X

in which the SNPs are localized. From the figure it is evident that bulk of markers fall under *p*-value range of 0.05–0.001 with only few hundred markers below 0.001. The distribution of significant SNPs across all chromosomes confirms the polygenic nature of the disease.

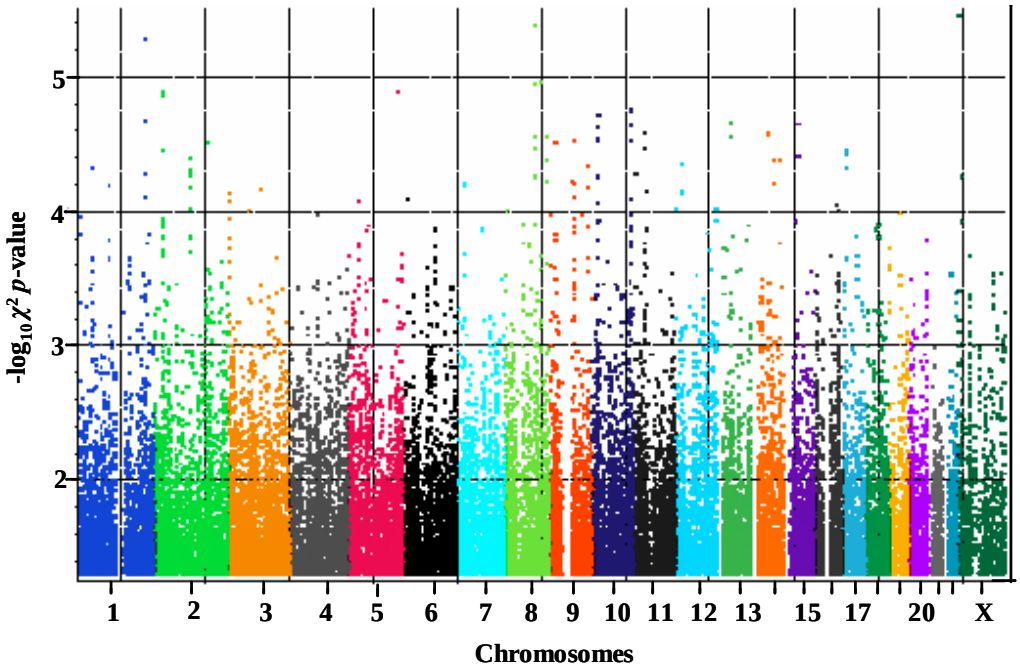


Figure 7.5: Scatter plot (Manhattan plot) for Stage I association study showing 35,589 markers (p <0.05) distributed across chromosomes. This graph is plotted against allelic chi-square p-values on $-\log_{10}$ scale to indicate polygenic nature of breast cancer susceptibility.

Q–Q plot is a graphical display comparing the observed distribution versus the expected distribution. In an ideal condition, observed statistic should conform to the expected statistic. **Figure 7.6** indicates that most of the SNPs lie along the

expected line (line of best fit) conforming to the null hypothesis of no association. There is no explicit deviation of the observed values from the expected values except a fraction of SNPs show deviation.

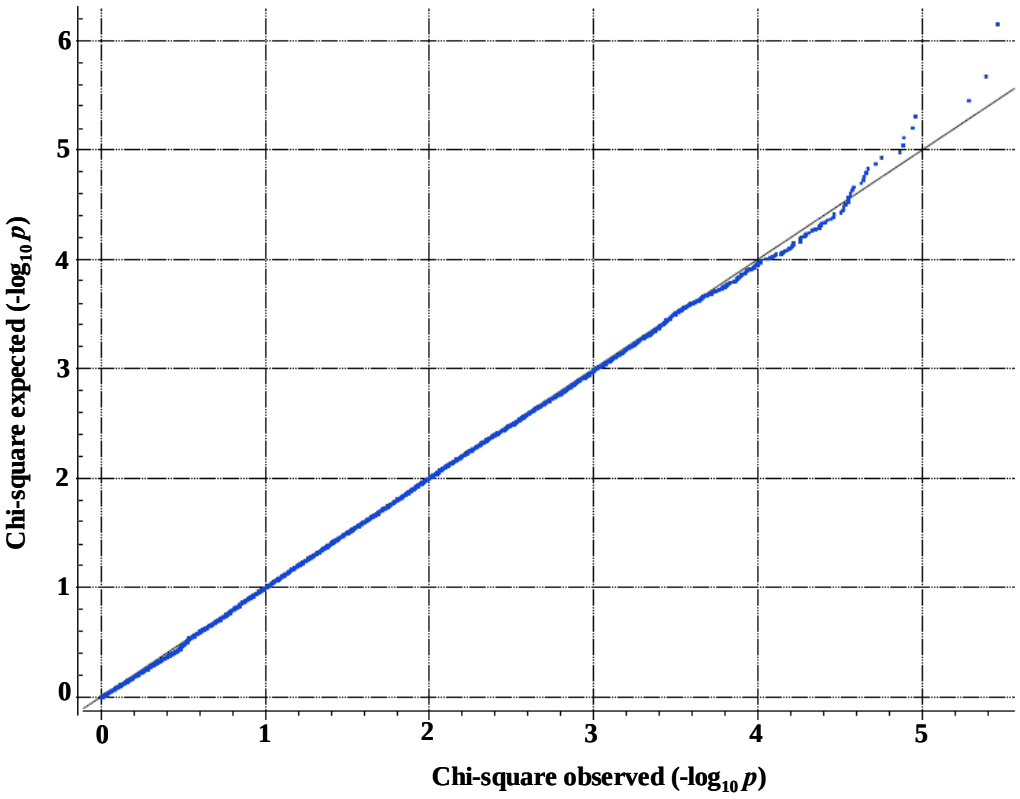


Figure 7.6: Quantile–Quantile plot displaying the conformity of observed versus expected statistic for select SNPs. Most of the SNPs lie along the line of best fit (solid central diagonal line in black) with only a subset of SNPs deviating from the expected distribution. On the *x*-axis is the –log10 *p*-value of chi-square observed distribution and on the *y*-axis is the –log10 *p*-value of chi-square expected distribution.

7.3 Selection of markers for replication

Selection of SNPs for replication is a crucial step after the genotype data is rigorously subjected to quality control measures before and after association analysis to reduce the inconsistencies and minimise false-positive association signals. At 0.05 cut-off, most or all of the observations (35,859 SNPs) could be by chance alone (most Stage I studies) necessitates replication of markers in independent cohorts. However, selection of markers for replication is itself not straightforward. Therefore, the selection of markers for replication from Stage I was carried out in a systematic manner proposed by Zheng et al. (19): (i) $p < 0.001$ for selected markers from allelic association analysis; (ii) MAF $\geq 10\%$; (iii) distinct genotype clusters or signal intensity plots; (iv) not previously reported in literature; and (v) $r^2 \geq 0.8$ (measure of LD). In addition, marker selection also was based on the presence of more than two SNPs in a single haplotype block and $p < 0.001$ for haplotype association analysis. Initially, a total of 109 SNPs that adhered to the above-mentioned criteria were selected but that included several markers from the same haplotype blocks that may possibly be redundant information. Owing to the costs associated with genotyping, we confined the replication steps for select SNPs following the Stage I whole genome scan by selecting a representative SNP per block with the lowest chi-square *p*-value which resulted in a selection of 35 SNPs.

Table 7.2: Selected markers from Stage I association analysis

dbSNP rs#	Associated gene	Relative location	χ^2 P	OR (95% CI)	HWE *p*
rs11138489	*TLE1 /TLE4*	1485 kb DS/371.7 kb DS	6.20E-05	0.48 (0.34,0.69)	0.573
rs6478296	*ASTN2*	Intron	6.68E-05	0.63 (0.50,0.79)	0.234
rs7908500	*OAT/CHST15*	105.8 kb DS/173.7 kb US	8.18E-05	1.59 (1.26,1.99)	0.356
rs10794182	*OAT/CHST15*	105.5 kb DS/174 kb US	1.04E-04	1.57 (1.25,1.98)	0.402
rs6561682	*LECT1 /SUGT1*	11 kb DS/3.8 kb DS	1.53E-04	1.59 (1.25,2.03)	0.996
rs8095374	*C18orf25*	Intron	1.57E-04	0.65 (0.52,0.81)	0.018
rs7099921	*OPTN /CCDC3*	16.1 kb US/82.3 kb US	3.53E-04	1.66 (1.26,2.20)	0.325
rs1981867	*C16orf61*	85.9 kb DS	3.70E-04	1.56 (1.22,2.00)	0.300
rs1911864	*GUSBL2 /CDH18*	1493 kb US/126.9 kb US	3.92E-04	1.51 (1.20,1.89)	0.354
rs268840	*SLC35F4 /C14orf105*	30.5 kb DS/39.5 kb US	4.59E-04	0.66 (0.52,0.83)	0.223
rs9630178	*LRRC4C /RAG2*	432.4 kb DS/3083.4 kb US	4.86E-04	1.91 (1.32,2.77)	0.114
rs10506269	*AMIGO2 /SLC38A4*	173 kb DS/76.6 kb US	5.05E-04	0.51 (0.35,0.75)	0.086
rs8075722	*OR3A2 /OR1D5*	5.5 kb DS/5.8 kb DS	5.26E-04	1.90 (1.32,2.74)	0.434
rs11195949	*ACSL5*	Intron	5.41E-04	0.67 (0.54,0.84)	0.374
rs2546513	*NUP107*	Intron	6.24E-04	0.65 (0.50,0.83)	0.346
rs13299280	*TLE1 /TLE4*	1508 kb DS/348.8 kb DS	6.39E-04	0.56 (0.40,0.78)	0.920
rs6493076	*UBR1*	Intron	6.47E-04	0.51 (0.34,0.75)	0.656

Table 7.2 continued…

dbSNP rs#	Associated gene	Relative location	χ^2 P	OR (95% CI)	HWE*p*
rs2080976	*ODZ2*	Intron	6.58E-04	1.51 (1.19,1.91)	0.455
rs1092913	*ROPN1L*	2.5 kb DS	7.00E-04	1.91 (1.31,2.80)	0.061
rs7119677	*C11orf41*	Intron	7.20E-04	0.64 (0.50,0.83)	0.731
rs3848562	*ZNF577*	Intron	8.01E-04	1.85 (1.28,2.65)	0.890
rs3935234	*C20orf56*	93.2 kb DS	8.64E-04	0.62 (0.47,0.82)	0.171
rs11257153	*USP6NL*	Intron	9.57E-04	0.60 (0.44,0.81)	0.668
rs6997395	*PTDSS1/SDC2*	19.9 kb DS/139.2 kb US	9.98E-04	0.67 (0.53,0.85)	0.832
rs10411161	*ZNF577*	3' UTR	1.08E-03	1.82 (1.27,2.62)	0.890
rs7818355	*SDC2 /PTDSS1*	146.9 kb US/12.1 kb DS	1.21E-03	0.56 (0.40,0.80)	0.887
rs9644134	*C8orf80*	Intron	1.21E-03	0.69 (0.55,0.86)	0.380
rs11878583	*ZNF577*	Intron	1.25E-03	1.78 (1.25,2.55)	0.312
rs6852237	*DCTD /ODZ3*	2.8 kb DS/84 kb DS	1.38E-03	1.44 (1.15,1.80)	0.736
rs1857434	*MLLT3 /SLC24A2*	364 kb DS/193.7 kb US	1.49E-03	1.64 (1.21,2.23)	0.377
rs1059307	*SNHG5*	Exon	1.54E-03	0.70 (0.56,0.87)	0.928
rs1451991	*LOC728643 /SNX16*	341 kb DS/108 kb US	2.19E-03	0.63 (0.47,0.85)	0.545
rs12433708	*PPP2R5E*	Intron	2.42E-03	0.61 (0.44,0.84)	0.533
rs6991277	*SDC2 /PTDSS1*	105.7 kb US/53.3 kb DS	2.63E-03	0.60 (0.43,0.84)	0.472
rs1429142	*EDNRA*	112.5 kb US	2.80E-03	1.57 (1.17,2.11)	0.853

DS, downstream; US, upstream; χ^2, chi-square; OR, odds ratio; CI, confidence interval; HWE, Hardy–Weinberg equilibrium; *C20orf56*, chromosome 20 open reading frame 56; *ROPN1L*, ropporin-1 like; *ZNF577*, zinc finger 577; *EDNRA*, endothelin receptor A; *C16orf61*, chromosome 16 open reading frame 61; *ASTN2*, astrotactin 2; *PTDSS1*/*SDC2*, phosphatidylserine synthase 1/Syndecan-2; *LRRC4C*/*RAG2*, leucine rich repeat containing 4C/recombination activating gene 2; *AMIGO2*/*SLC38A4*, adhesion molecule with Ig-like domain 2/solute carrier family 38, member 4; *SNHG5*, small nucleolar RNA host gene 5;

OAT/*CHST15*, ornithine aminotransferase/carbohydrate (N-acetylgalactosamine 4-sulfate 6-O) sulfotransferase 15; *ODZ2*, odd Oz/ten-m homolog 2 (*Drosophila*); *OPTN*/*CCDC3*, optineurin/coiled-coil domain containing 3; *NUP107*, nucleoporin 107kDa; *TLE1*/*TLE4*, transducin-like enhancer of split 1/4 (E(sp1) homolog, *Drosophila*) *OR3A2*/*OR1D5*, olfactory receptor, family 3, subfamily A, member 2/olfactory receptor, family 1, subfamily D, member 5; *ACSL5*, acyl-CoA synthetase long-chain family member 5; *USP6NL*, USP6 N-terminal like; *UBR1*, ubiquitin protein ligase E3 component n-recognin 1; *LECT1*/*SUGT1*, leukocyte cell derived chemotaxin 1/SGT1, suppressor of G2 allele of SKP1 (*S. cerevisiae*); *MLLT3*/*SLC24A2*, myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, *Drosophila*)/solute carrier family 24 (sodium/potassium/calcium exchanger), member 2; *GUSBL2*/*CDH18*, glucuronidase, beta-like 2/cadherin 18, type 2; *DCTD*/*ODZ3*, dCMP deaminase/odd Oz/ten-m homolog 3 (*Drosophila*); *C8orf80*, chromosome 8 open reading frame 80; *C11orf41*, chromosome 11 open reading frame 41; *C18orf25*, chromosome 18 open reading frame 25; *PPP2R5E*, protein phosphatase 2, regulatory subunit B', epsilon isoform; *SLC35F4*/*C14orf105*, solute carrier family 35, member F4/chromosome 14 open reading frame 105; *LOC728643*/*SNX16*, sorting nexin 16

The selected SNPs were genotyped in an independent study with 1153 breast cancer cases and 1215 controls from Alberta. Genotyping was carried out in Sequenom Mass-ARRAY iPlex technology. Results of Stage II and joint analysis are presented in Chapter 4.