**RARE CODING VARIANTS in GWAS identified loci with BREAST CANCER RISK**

By

Mi-Ryung Han

Dissertation

Submitted to the Faculty of the Graduate School of Vanderbilt University

in partial fulfillment of the requirements  for the degree of

DOCTOR OF PHILOSOPHY

in

Epidemiology

May, 2016

Nashville, Tennessee

Approved:

Professor Jirong Long

Professor Wei Zheng

Professor Todd L. Edwards

Professor Bingshan Li

To my parents, Hyung-Suk Han and Suk-Hee Yu, for being a great example, giving me

unconditional support and love, and unwavering faith to pursue my dreams

To my sister Shin-Young Han for believing in me and supporting me through this journey.

and

To God, for everything

# ACKNOWLEDGEMENTS

I would like to thank my committee chair and mentor, Dr. Jirong Long, and the members of my committee, Dr. Wei Zheng, Dr. Todd L. Edwards, and Dr. Bingshan Li for their invaluable guidance, and thoughtful comments and suggestions throughout this process. I am especially grateful to Dr. Long and Dr. Zheng for their support of my research at Vanderbilt University and their professional guidance and teaching about both scientific research and life in general. I also want to thank Dr. Edwards and Dr. Li for sharing their knowledge, providing guidance and helping me to learn in-depth scientific methods. I also thank the members of our lab, especially Dr. Guo, Dr. Wen and Jing He who taught me a great deal about scientific research and provided thoughtful feedback throughout this work.

I would like to thank study participants, staff and funding agencies involved in current project, without whom this dissertation would not have been possible. I also want to acknowledge my father, Hyung-Suk Han, and my mother, Suk-Hee Yu for always supporting and believing in me. Without their love and encouragement, none of my accomplishments would be possible. Thank you both for giving me strength to pursue my dreams. My sister, Shin-Young Han deserves my wholehearted thanks as well.

Thank you God for the light I have been able to see and for being there for me all the time.

**TABLE OF CONTENTS**

**Page**

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AA | African Americans |
| AKAP12 | A Kinase (PRKA) Anchor Protein 12 |
| AKR1C2 | Aldo-Keto Reductase Family 1, Member C2 |
| ANKRD35 | Ankyrin Repeat Domain 35 |
| ANNOVAR | Annotate Variation |
| ANO1 | Anoctamin 1, Calcium Activated Chloride Channel |
| BPIFA2 | BPI Fold Containing Family A, Member 2 |
| BPIFB6 | BPI Fold Containing Family B, Member 6 |
| BRCA2 | Breast Cancer 2, Early Onset |
| CCDC38 | Coiled-Coil Domain Containing 38 |
| CENPW | Centromere Protein W |
| CMC | Combined Multivariate and Collapsing |
| CPA1 | Carboxypeptidase A1 |
| DCLRE1A | DNA Cross-Link Repair 1A |
| DFFA | DNA Fragmentation Factor, 45kDa, Alpha Polypeptide |
| EA | European Americans |
| ELK3 | ETS-Domain Protein (SRF Accessory Protein 2) |
| eQTL | Expression Quantitative Trait Loci |
| FGF10 | Fibroblast Growth Factor 10 |
| FKBP8 | FK506 Binding Protein 8, 38kDa |
| GPR98 | G Protein-Coupled Receptor 98 |
| gTDT | group-wise Transmission/Disequilibrium Tests |
| GTEx | Genotype-Tissue Expression |
| GWAS | Genome-Wide Association Studies |
| IBD | Identity-By-Descent |
| ITGA10 | Integrin, Alpha 1 |
| LAPTM4A | Lysosomal Protein Transmembrane 4 Alpha |
| LD | Linkage Disequilibrium |

| | |
|---|---|
| LOC100294362 | RNA Gene |
| LOF | Loss-of-function variants |
| LOFTEE | Loss Of Function Transcript Effect Estimator |
| MAF | Minor Allele Frequency |
| MB | Madsen-Browning test |
| METABRIC | Molecular Taxonomy of Breast Cancer International Consortium |
| MPP4 | Membrane Protein, Palmitoylated 4 |
| mRNA | messenger Ribonucleic Acid |
| MTMR11 | Myotubularin Related Protein 11 |
| MUS81 | MUS81 Structure-Specific Endonuclease Subunit |
| NBHS | Nashville Breast Health Study |
| OR2J2 | Olfactory Receptor, Family 2, Subfamily J, Member 2 |
| PCs | Principal Components |
| PLBD1 | Phospholipase B Domain Containing 1 |
| PLEKHS1 | Pleckstrin Homology Domain Containing, Family S Member 1 |
| PSG5 | Pregnancy Specific Beta-1-Glycoprotein 5 |
| PSG6 | Pregnancy Specific Beta-1-Glycoprotein 6 |
| QC | Quality Control |
| RefSeq | NCBI Reference Sequence Database |
| SBCGS | Shanghai Breast Cancer Genetics Study |
| SBCS | Shanghai Breast Cancer Study |
| SBCSS | Shanghai Breast Cancer Survival Study |
| SCCS | Southern Community Cohort Study |
| SECS | Shanghai Endometrial Cancer Study |
| SHAPEIT | Segmented HAPlotype Estimation and Imputation Tool |
| SKAT | Sequence Kernel Association Test |
| SLC25A42 | Solute Carrier Family 25, Member 42 |
| SLC25A45 | Solute Carrier Family 25, Member 45 |
| SLC6A18 | Solute Carrier Family 6 (Neutral Amino Acid Transporter), Member 18 |
| SNPs | Single Nucleotide Polymorphisms |
| SWHS | Shanghai Women's Health Study |

SYT8          Synaptotagmin VIII

TCF7L2        Transcription Factor 7-Like 2

TCGA          The Cancer Genome Atlas

THEMIS        Thymocyte Selection Associated

TRPS1         Trichorhinophalangeal Syndrome I

UBR7          Ubiquitin Protein Ligase E3 Component N-Recognin 7

VANGARD       Vanderbilt Technologies for Advanced Genomics Analysis and Research Design

VANTAGE       Vanderbilt Technologies for Advanced Genomics

VEP           Ensembl Variant Effect Predictor

VT            Variable Threshold

ZFYVE26       Zinc Finger, FYVE Domain Containing 26

# CHAPTER I

## INTRODUCTION AND SPECIFIC AIMS

Breast cancer is the most common invasive cancer in females worldwide and in East Asian countries (1). To date, common genetic variants in ~ 109 loci have been identified for breast cancer risk via genome-wide association studies (GWAS), which primarily focus on evaluating common single nucleotide polymorphisms (SNPs) (2–12). Collectively, these common variants only explain approximately 16% of the heritability of breast cancer, so it is suspected that rare/low-frequency variants in these loci may also contribute to breast cancer risk (13). Currently, many studies are investigating low-frequency (MAF 0.01-0.05) and rare (MAF < 0.01) variants. Studies showed that genetic variants with lower allele frequency are more likely to be functional than common variants (14). So far, several genes have been shown to harbor rare coding variants associated with breast cancer risk such as *BRCA2*, *EDEM1*, *EFEMP2*, *FBXO18*, *ERBB2*, *CHEK2*, *ATM*, *BRIP1*, *PALB2*, *RAD51C*, *RAD51D*, and *PPM1D* genes (15–23).

In this dissertation, we describe approaches for genetic analyses of breast cancer risk associated with rare coding variants using whole-exome chip data. Exome-based genotyping has the capacity to discover rare/low-frequency variants in exon regions associated with complex diseases in a large population (24, 25). Exome-based genotyping arrays, such as the Illumina HumanExome BeadChip and the Affymetrix Axiom exome array, are cost effective and have recently been used as alternative platforms to whole-exome sequencing (26, 27). We focused on 109 loci identified from previous GWAS in order to investigate rare nonsense/missense variants and their corresponding genes in different ethnic groups, including Chinese, European

Americans (EA), and African Americans (AA). Only functional variants within protein-coding regions (e.g., missense, nonsense, and loss-of-function variants) are included since they result in alteration in the encoded amino acid and they can help pinpoint causative genes. We did not include synonymous variants since they do not result in a change of amino acid in the protein.

The list of 109 GWAS identified SNPs is provided in Appendix 1. A total of 9,004 cases and 11,996 controls from three ethnic groups were examined. Included in this study were 5,766 cases and 5,703 controls of Chinese women from the Shanghai Breast Cancer Genetics Study (28), 1,509 cases and 1,456 controls of EA women and 500 cases and 272 controls of AA women from the Nashville Breast Health Study (NBHS) (29, 30), 534 cases and 781 controls of AA women from the Southern Community Cohort Study (SCCS) (31), and 695 cases and 3,784 controls of EA women from the Vanderbilt electronic medical record-linked DNA repository, BioVU (32).

The following aims are developed to carefully investigate the relationship between rare coding variants and breast cancer risk.

**Specific Aim 1:** To identify candidate genes in the previously reported GWAS loci for breast cancer. Most breast cancer-associated GWAS loci are located in noncoding regions and are therefore thought to be regulatory in nature. The mechanistic basis for the association between breast cancer and most of the common variants discovered in GWAS is still largely unknown. Rare variant in GWAS identified genes may in part explain this limitation and contribute significantly to breast cancer risk. Candidate genes were identified through expression quantitative trait loci (eQTL) analyses: expression level of which genes were affected by the GWAS identified SNPs. Data generated from breast tissues in three major sources were used for eQTL analyses: the Cancer Genome Atlas (TCGA) (33), the Genotype-Tissue Expression

(GTEx) (34), and Molecular Taxonomy of Breast Cancer International Consortium

(METABRIC) (35). *Cis*-based eQTL analyses for all genes 1Mb flanking the GWAS index SNPs

were performed. Genes were selected with expression level associated with breast cancer risk-

associated SNPs.

Hypothesis: Most common variants found in GWAS are located in non-coding region, thus

impeding the direct interpretation of their functional effects. They may be involved in regulation

of gene expression, and rare functional variants in the coding region of these genes may change

gene structure and function. Therefore, we hypothesize that these rare variants may contribute to

breast cancer.

**Specific Aim 2:** To investigate rare variants in the eQTL genes identified in Aim 1 in association

with breast cancer risk.

**Sub-aim 1:** Functional prediction of rare coding variants. We investigated rare variants using

ANNOVAR (Annotate Variation) in order to annotate nonsynonymous variants that result in a

change of amino acid in the protein. Loss-of-function variants (LOF) were predicted using

LOFTEE (Loss Of Function Transcript Effect Estimator) in order to categorize stop_gain, splice

site disrupting, and frameshift variants.

Hypothesis of Sub-aim 1: We hypothesize that rare coding variants in the eQTL genes will alter

translation or protein function that impact breast cancer with potentially deleterious outcome.

**Sub-aim 2:** Investigating associations of rare variants with breast cancer. Whole-exome chip

data from Chinese (5,766 cases and 5,703 controls), EA (2,204 cases and 5,240 controls), and

AA populations (1,034 cases and 1,053 controls) were used for association analysis of rare

variants. Associations of rare variants with breast cancer risk were evaluated using single-variant and gene-based aggregation tests. We used two sets of variants (nonsynonymous and LOF variants) prioritized from Sub-aim 1, in order to conduct gene-based aggregation tests. Single-variant tests including score test, firth test, and fisher's exact test, and gene-based tests including burden and non-burden tests were conducted. All of these tests assume additive models. In order to detect rare variants that confer significant risk in a recessive manner, we performed compound heterozygous (CH) analysis. Both single-variant and gene-based meta-analyses within Chinese, EA, and AA were conducted to establish association between rare variants and breast cancer risk. Using these meta-analyses, we could have more power to detect true associations between rare variants and breast cancer risk. After investigating associations, we predicted function of identified variants using SIFT algorithm (Sorting Intolerant From Tolerant), PolyPhen-2 (Polymorphism Phenotyping v2), and PROVEAN (Protein Variation Effect Analyzer).

Hypothesis of Sub-aim 2: We hypothesize that a significant proportion of the inherited susceptibility to breast cancer disease may be due to the summation of the effects of rare variants of a variety of different genes, each conferring a moderate but detectable increase in relative risk. We expect to find several rare coding variants associated with breast cancer risk in Chinese, EA, and AA populations.

# CHAPTER II

# BACKGROUND

## A.    Current Status of Genetic Research on Breast Cancer

Breast cancer is the most common malignancy among women in the United States and many other countries around the world (36). It is a complex disease in which genetic factors play an important role (10, 37). In the 1990s, the two major susceptibility genes for breast cancer, *BRCA1* (38) and *BRCA2* (39), were identified through family-based linkage studies. Due to the limitation of linkage studies which aimed at identifying rare and high-risk disease-associated mutations based on multiple individuals in a family, a large number of candidate gene studies were conducted over the following decade. Candidate gene approaches have focused on selecting genes based on their known biological function and aimed at identifying moderate and low penetrance alleles believed to be responsible for the remaining familial risk. Several DNA repair genes including *ATM* (40), *CHEK2* (41), *BRIP1* (42) and *PALB2* (43) and an apoptosis gene, *CASP8* (44, 45), have been implicated in susceptibility to breast cancer. However, the majority of reported SNP associations in candidate genes could not be replicated.

Since 2005, GWAS have made an important contribution to find many novel variants for human diseases that were not found by the candidate gene approach. GWAS are designed to detect associations through linkage disequilibrium (LD) between genotyped (or imputed) common SNP markers and unknown causal variants. Approximately 109 common genetic susceptibility loci for breast cancer risk have been found, including those identified in our own study among Asian women (4, 6, 24, 33).

Extensive genetic studies have identified high-penetrance genes (*BRCA1*, *BRCA2*, *PTEN* and *TP53*), moderate-penetrance genes (*CHEK2, ATM, BRIP1, PALB2, RAD51C, STK11, CDH1, RAD50,* and *NBN*), and more than 109 low-penetrance loci that contribute to the risk of breast cancer over the past 20 years (4–6, 13, 28, 37, 46–51). It has been shown that pathogenic mutations in the *BRCA1* and *BRCA2* genes are associated with a 10- to 20-fold increased risk of breast cancer which corresponds to a cumulative risk of breast cancer by age 70 years of 55%-65% for *BRCA1* mutation carriers and 45-47% for *BRCA2* mutation carriers (52, 53). Recently, it has been reported that female *PTEN* mutation carriers have an 85% lifetime risk of developing breast cancer with 50% penetrance by 50 years of age (54). These findings were subsequently confirmed by two other studies (55, 56). Mutations in the *TP53* gene are associated with at least a 10-fold increased risk of breast cancer and account for 2-7% of early-onset breast cancer (57, 58). It is estimated that the cumulative risk of breast cancer by 70 years old is approximately 14% for women who carry *CHEK2 1100delC*, and a subsequent meta-analysis based on 29,154 cases and 37,064 controls from 25 case-control studies reported a significant association between *CHEK2 1100delC* heterozygotes and breast cancer risk with OR (95% CI) of 2.75 (2.25-3.36) (59, 60). Similarly, the approximate risk of breast cancer is 15% for those who carry *ATM* mutations (61). It is estimated that the eight confirmed high and moderate-penetrance genes (*BRCA1, BRCA2, PTEN, TP53, CHEK2, ATM, BRIP1*, and *PALP2*), explain approximately 20% of the familial risk of breast cancer (46).

Despite the recent success of GWAS, the majority of the genetic component of many complex traits remains unexplained. In addition, although the statistical evidence for an association between SNP and breast cancer risk is overwhelming, the biologically relevant variants and the mechanism by which they lead to increased risk are unknown and require further

genetic and functional characterization. As rare variants have been comparatively less well-studied than common variants, attention has shifted to exome-chip, exome or genome sequencing approaches to identifying additional risk factors. We used exome-chip data since it is cost-effective and feasible for large studies to identify rare genetic variants in thousands of individuals.

## B.    Breast Cancer Susceptibility: The Role of Rare Variants

GWAS are designed to evaluate common genetic variants, typically with a $MAF > 0.05$, therefore examining only a portion of the genomic landscape of complex traits. GWAS identified more than 100 common genetic susceptibility loci associated with breast cancer so far; however, these loci collectively explain approximately 16% of the heritability of breast cancer (13). It is reasonable to assume that most common and highly penetrant susceptibility genes have already been discovered for breast cancer. Currently, many studies are investigating rare ($MAF < 0.01$) variants which have been more challenging to assess.

## B1.    Missing Heritability

More than 20 years ago, the identification of the two high-penetrance genes in breast cancer, *BRCA1* and *BRCA2*, launched a sustained effort to uncover new genes explaining the "missing heritability" in the disease. The best known high or moderate penetrance genes include *BRCA1*, *BRCA2*, *TP53*, *PTEN*, *STK11*, *PALB2*, and *ATM*, and these genes globally account for around 35% of the familial breast cancer cases (62). Many explanations, such as rare variants, epistatic interactions, gene-environment interactions, structural variants, heritable epigenetic factors, parent-of-origin effects, or inflated heritability estimates have been proposed to illustrate the "missing heritability" that the GWAS loci and high-penetrance genes could not explain (63–

66). The major debates over the nature of the genetic contribution to individual susceptibility to common complex diseases are common disease common variant (CDCV) and common disease rare variant (CDRV) hypotheses. The CDCV hypothesis argues that genetic variations with appreciable frequencies in the population at large, but relatively low penetrance (or the probability that a carrier of the relevant variants will express the disease), are the major contributors to genetic susceptibility to common diseases (67). CDRV argues that multiple rare DNA sequence variants, each with relatively high penetrance, could account for the genetic variance in disease susceptibility (67).

Many investigators have tried the alternative CDRV hypothesis. Pritchard argued that the notion that multiple, very recent rare variations contributing to disease arising in the last two centuries is more consistent with human population pathobiology than the notion that older, common variations are contributing to disease (68). This is because rare variants are often evolved from more recent mutations and subjected to less natural selection. Leal pointed out that rare variants, although individually rare, are collectively frequent, and even though their effect sizes are greater than those observed for common variants, they are not large enough to produce familial aggregation (66). In this light, reports on the frequency of human alleles and their likely 'functional' or phenotypic effects suggest that rare coding variants are enriched for functional importance (14). We are in the era to investigate rare variants that might play an important role in explaining the "missing heritability" of complex traits including breast cancer.

**B2.      Rare Variants Associated with Breast Cancer and Other Diseases**

It has been increasingly recognized that the "missing heritability" for breast cancer could be partially explained by low-frequency (MAF 0.01-0.05) and rare (MAF < 0.01) variants. There is strong evidence that rare genetic variation is important in breast cancer predisposition (69). In

the 1990s, genome-wide linkage analysis and positional cloning led to the identification of the DNA repair genes *BRCA1* and *BRCA2*, and rare mutations of those genes in noncoding region confer substantial risks to breast cancer (69). More recently, through case-control resequencing studies of candidate genes, several rare coding variants have been shown to be associated with breast cancer risk such as *ERBB2*, *CHEK2*, *ATM*, *BRIP1*, *PALB2*, *RAD51C*, *RAD51D*, and *PPM1D* genes (16–23). Rare protein truncating variants (PTV) mutations in the p53 inducible protein phosphatase gene *PPM1D* are associated with predisposition to breast cancer (18).

In addition, recently, a known moderate susceptibility indel variant (*CHEK2 1100delC*) and a catalogue of 11 rare variants in other genes (*FANCM*, *WNT8A*, *MAPKAP1*, *TNFSF8*, *PTPRF*, *UBA3*, *AXIN1*, *TIMP3*, *SLBP*, *CNTROB*, and *S1PR3*), presenting signs of association with breast cancer, were identified through whole-exome sequencing (62).

Zhang *et al*. recently investigated rare missense/nonsense variants with MAF ≤ 0.05 located in flanking 1Mb of each of the index SNP in 67 GWAS loci from the Shanghai Breast Cancer Study including 3,472 cases and 3,595 controls (15). Notably, 5 rare variants in different genes (*BRCA2*, *EDEM1*, *EFEMP2*, and *FBXO18*) were associated with breast cancer risk at P-value < 0.01 (15). Compared to Zhang's study, the current study included an increased number of Chinese (5,766 cases and 5,703 controls) and investigated other ethnic groups, EA (2,204 cases and 5,240 controls) and AA (1,034 cases and 1,053 controls) as well. We performed more comprehensive functional and eQTL analyses to prioritize candidate genes in the 1Mb regions flanking the breast cancer 109 GWAS loci using three major databases, and we assessed rare recessive variants in addition to additive models. With increased number of populations and improved statistical methods, we had more power to detect rare variants associated with breast cancer risk compared with Zhang's study.

Recently, multiple papers reported that low frequency or rare variants in GWAS loci have been identified for other diseases through target sequencing or fine-mapping (70–72). Beaudoin *et al*. have used a targeted sequencing approach in 200 ulcerative colitis cases and 150 healthy controls, all of French Canadian descent, to study 55 genes in regions associated with ulcerative colitis (70). They found significant association with rare non-synonymous variants in both *IL23R* and *CARD9*, previously identified from sequencing of Crohn's disease loci, as well as a novel association in *RNF186* (70). Fine mapping of GWAS loci associated with low-density lipoprotein cholesterol also discovered several low frequency or rare variants (71). In addition, Johansen *et al*. reported that an accumulation of rare variants is present in GWAS identified genes, and that these contribute to the heritability of complex traits among individuals at the extreme of a lipid phenotype (72). These studies support our hypothesis that rare coding variants in GWAS loci may contribute to breast cancer risk.

## B3.  eQTL analysis

GWAS have identified thousands of variants that are associated with complex traits and diseases. However, because most variants are noncoding and located in intronic or intergenic regions, it is difficult to identify causal genes. Polymorphisms associated with messenger RNA (mRNA) levels are typically referred to as eQTLs (50). eQTLs have provided key insights into genes and pathways as well as the genetic architecture of gene expression (73). Several eQTL-mapping studies have shown that disease-predisposing variants often affect the gene expression levels of nearby genes (*cis*-eQTLs) (74–76). *Cis*-acting regulation is due to DNA variation that directly influences the transcription process in an allele-specific manner. Alternatively, *trans*-acting regulation affects the gene expression by modifying the activity (or abundance) of the factors that regulate the gene (77). Regarding rare variants studies, Cheng *et al*. discovered rare

variants associated with autism spectrum disorders in the GWAS candidate gene (*SEMA5A*) using *cis*-eQTL mapping (78). Recently, eQTL analyses of 15 previously reported breast cancer risk loci resulted in the discovery of three variants (at 2q35 (*IGFBP5*), 5q11 (*C5orf35*), and 16q12 (*TOX3*)) that are significantly associated with transcript levels (73).

The eQTL approach is valuable when causal variants exert remote regulatory effects on genes whose coding regions lie outside the region of association, and this approach has potential to find candidate genes and their functional variants. To investigate rare variants in the eQTL genes might be particularly informative since the associated rare variants for complex diseases will be more facile to evaluate for functional impact. Therefore, in current study, we have examined genes within breast cancer-associated GWAS loci using eQTL mapping. Using our approach, we would be able to find rare variants that may contribute to breast cancer.

# CHAPTER III

# RESEARCH GAP

To the best of our knowledge, this is the first study to systemically examine the associations between rare variants in potential functional genes in breast cancer risk loci. Few studies have identified rare recessive variants associated with breast cancer risk. He *et al*. identified a recessive missense variant of *XRCC4* in non-*BRCA1/2* breast cancer patients in the Chinese population (68). Kuligina *et al*. recently found rare recessive homozygous variant in *GEN1* that has been associated with bilateral breast cancer (79). We investigated rare recessive variants in GWAS loci that have been found to predispose people to breast cancer. Unlike other studies, this study provides a strong basis for the rare recessive variant in GWAS loci through comprehensive analysis strategy that might reveal the important mechanism and biology underlying breast cancer.

As far as we know, this is the largest study to date investigating rare coding variants for association with breast cancer risk in the East and Southeast Asian populations. Our research group has recruited a large number of subjects and collected adequate biological samples and clinical data for genetic epidemiology study in the Eastern and South-eastern Asian population. In addition, data from European American and African American populations from the NBHS, SCCS, and BioVU were also available. These data were used to compare the results from Asians to investigate the generalizability. Therefore, with a large number of subjects from these three ethnic groups (a total of 9,004 cases and 11,996 controls), we would be able to comprehensively evaluate rare genetic variants for breast cancer risk.

This study capitalizes on the most recent resources of three major datasets for breast tissue (TCGA, GTEx, and METABRIC) in order to idnetify genes that may cause the disease. We are not aware of any study that evaluates eQTLs using all three databases. We have more power to conduct eQTL analyses to identify genes associated with breast cancer risk in GWAS identified loci by integrating three databases (4, 6, 10, 11). In addition, currently identified GWAS loci included 12 novel genetic variants discovered in our own GWAS from Chinese population (6, 28, 49–51, 80–82).

CHAPTER IV


METHODS


**A. Methods for Specific Aim 1: To identify potential functional genes underlying the associations in breast cancer GWAS loci**

Hypothesis: The mechanistic basis for the association between breast cancer and most of the common variants discovered in GWAS is still largely unknown. Common variants found in GWAS studies can regulate gene expression level. Functional variants in the coding region of these genes may change gene expression level, structure and function. Therefore, we hypothesize that these variants may contribute to breast cancer.

GWAS have identified novel and known loci associated with breast cancer risk. Although GWAS continue to reveal new associations, each newly associated variant has a smaller effect size and contributes only marginally to the cumulative variation of complex diseases. This suggests that GWAS of population-based subjects may be reaching the limits of their ability to reveal genetic variation underlying complex traits. Then, a question has arisen whether additional forms of genetic variation, such as rare variants with large individual effects, could contribute to the heritability of complex traits such as breast cancer. Due to many successful results from GWAS, we are able to use resources of GWAS identified loci associated with breast cancer. It remains possible that rare variants in GWAS identified genes may contribute significantly to breast cancer risk (70–72, 83, 84). In this study, our exposure is defined as SNPs which are investigated in this study, and outcome of interest is breast cancer cases.

**A1. GWAS loci for breast cancer**

We systemically investigated all GWAS loci associated with breast cancer risk using publicly available databases. A Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwastudies/) and MEDLINE were used to find GWAS loci related to breast cancer risk. From the search results, we extracted information including SNP ID, GWAS identified loci, risk allele, effect size (odds ratio), and P-value. In order to check the direction of the association, risk alleles were carefully selected. We included SNPs identified from GWAS among Asian populations and other ethnic groups including European and African. For each population, SNPs were excluded if they were in strong LD ($r^2 > 0.8$) using HaploReg V3 (http://www.broadinstitute.org/mammals/haploreg/haploreg_v3.php). Finally, we included all GWAS loci associated with breast cancer risk including two loci recently identified from our group (Han *et al*. submitted manuscript, 2016).

**A2. eQTL analysis**

**A2.1. Data sources**

There are three major sources that provide data generated in breast tissues; The Cancer Genome Atlas (TCGA), The Genotype-Tissue Expression (GTEx), and The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC).

The Cancer Genome Atlas (TCGA), is a comprehensive database which focuses on genomic alterations in diverse cell types at different sites in the body that give rise to hundreds of different forms of cancer (33). We downloaded RNA-Seq V2 data (level 3), DNA methylation data and somatic copy number alterations (CNA) data using the CGDS-R package from the cBioPortal (http://www.cbioportal.org/public-portal/), which provided a basic set of functions for

15

extracting data from the Cancer Genomic Data Server (CGDS). We also downloaded level 2

SNP data genotyped using Affymetrix SNP 6.0 array from TCGA data portal (The Cancer

Genome Atlas, http://cancergenome.nih.gov/). Genotype data from the flanking 1Mb region for

the 109 GWAS loci were extracted. We analyzed a total of 709 breast tumor tissues (653

European population and 56 Asian population), including matched CNV, genotype, and

expression data.

In addition to TCGA, we conducted eQTL analyses using the GTEx database

(http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi) and data from the METABRIC project

(35). The GTEx project of the NIH Common Fund aims to establish a resource database and

associated tissue bank in which to study the relationship between genetic variation and gene

expression and other molecular phenotypes in multiple reference tissues (34). The GTEx Portal

has been updated to data release V6 in October, 2015. We accessed *cis*-eQTL results of 183

breast normal tissues from the most recent GTEx database which were calculated from linear

regression analysis using Matrix eQTL (85). The Molecular Taxonomy of Breast Cancer

International Consortium (METABRIC) is a Canada-UK project that aims to classify breast

tumors into further subcategories, based on molecular signatures that will help determine the

optimal course of treatment (35). We extracted matched genotypes and gene expression levels in

a total of 1,981 breast cancer tumor tissues from the METABRIC project. Gene expression

profiling was generated on the Illumina HT12 arrays and downloaded from the Synapse

(syn1757063, https://www.synapse.org/). A total of 49,576 transcripts are included in gene

expression profiling and have been normalized as described previously (35). Genotype data

using the Affymetrix SNP 6.0 array was downloaded from EBI (EGAD00010000164,

https://www.ebi.ac.uk/). We used R package CRLMM

16

(http://bioconductor.org/packages/crlmm/) to generate genotype calls from the original image array-based data for METABRIC (86). Only probes of high quality with intensity more than 3,000 at a 95% calling rate were included. A total of 1,981 tumor tissue samples with matched gene expression and genotype data were included.

## A2.2. Imputation

Genotype data from the flanking 1Mb region for the 109 GWAS loci were imputed for the TCGA and METABRIC data. Imputation was performed using SHAPEIT to derive phased genotypes and Minimac2 to perform imputation on the phased data (87, 88). Minimac2 is a low memory, computationally efficient implementation of the MaCH algorithm for genotype imputation. The 1000 Genome Project phase 3 was used as the reference data for imputation (http://www.1000genomes.org/). A total of 2,504 subjects and 84.7 million SNPs are included in the 1000 Genome Project phase 3. SNPs with high imputation quality (RSQR > 0.3) and MAF > 0.05 within the 1Mb regions flanking the 109 GWAS loci were included in the analysis.

## A2.3. Data analysis

The eQTL analysis was performed in TCGA tumor tissues as previously described (75, 89). Briefly, the RSEM (RNA-Seq by Expectation-Maximization) value of each gene was log2 transformed and those genes with a median expression level of 0 across tissues were removed. We then performed principal component correction on gene expression data to remove potential batch effects. To make the data better conform to the linear model for the eQTL analysis, we further transformed the gene expression level to fit a quantile of $N(0,1)$ distributions based on the rank of the expression values to their respective quantiles.

A full linear regression analysis was then used to detect eQTLs while adjusting for methylation, CNV, and ethnicity. For a given gene *i* and a SNP locus *j*, three factors of transcript abundance (T) were considered; the germline genotypes as the genetic determinants (G), the somatic copy number alterations (Sc), the CpG methylation levels (M), and ethnicity (E):

$$T_i = G_i + Sc_i + M_i + E_i + \varepsilon_i$$

Using this model, we evaluated the association between genotypes and genes located within the 1Mb regions flanking the 109 GWAS loci to identify *cis*-eQTLs.

In the METABRIC dataset, eQTL analysis was performed using Matrix eQTL to evaluate the association between genotypes and gene expression levels using linear regression model (85). We were not able to adjust methylation and CNV since data were not available for METABRIC. The eQTL results of GTEx were also calculated using Matrix eQTL, and available on GTEx Portal. For all datasets (TCGA, METABRIC and GTEx), a significance threshold P-value of < 0.05 was used to determine candidate *cis*-eQTLs.

**B. Methods for Specific Aim 2: To investigate rare variants associated with breast cancer risk**

**B1. Sub-Aim 1: Functional prediction of rare coding variants**

Hypothesis: We hypothesize that rare coding variants in the eQTL genes will alter translation or protein function that impact breast cancer with potentially deleterious outcome.

**B1.1.  Rare Variants**

Most genetic variation is considered neutral but single base changes in and around a gene can affect its expression or the function of its protein products (90, 91). Among the sequence variants currently known to be directly linked with human Mendelian disease, 57% are due to

nonsynonymous mutations. An additional 23% of disease variants are due to small insertions and deletions (indels) in genes. Because nonsynonymous SNPs can affect protein function, they are believed to have the largest impact on human health compared with SNPs in other regions of the genome. Therefore, it is important to distinguish those nonsynonymous SNPs that affect protein function from those that are functionally neutral.

Nonsynonymous mutations are further classified into missense and nonsense mutations. For protein-coding regions, there are three classes of mutations: silent, missense, and disruptive (defined as nonsense, splice site, and frameshift mutations). In genetics, silent mutations are DNA mutations that do not significantly alter the phenotype of the organism in which they occur. A missense mutation is a point mutation in which a single nucleotide change results in a codon that codes for a different amino acid. A nonsense mutation is a point mutation in a sequence of DNA that results in a premature stop codon (stop_gain), and ultimately resulting in the production of a truncated protein. A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript. A splice site mutation is a genetic mutation that inserts, deletes or changes nucleotides in the specific site at which splicing of an intron takes place during the processing of precursor messenger RNA into mature messenger RNA. A frameshift mutation is a genetic mutation caused by a disruption of the translational reading frame because the number of nucleotides inserted or deleted is not a multiple of three.

According to the common disease-rare variant hypothesis, low-frequency variants with strong effects at a locus can contribute to disease (68). In this study, we focused on studying missense and nonsense variants (nonsynonymous) as well as splice site and frameshift variants for rare variant analysis.

## B1.2. Assessing protein function prediction

We investigated rare variants using ANNOVAR (Annotate Variation) in order to annotate nonsynonymous variants that result in a change of amino acid in the protein. ANNOVAR is a program for functional annotation of genetic variants from high-throughput sequencing data such as RefSeq (NCBI Reference Sequence Database) (92). We included all missense and nonsense variants located flanking 1Mb of the indexed SNP of 109 GWAS loci, and ANNOVAR program was used to annotate all SNPs (92). Then, LOF variants were annotated using LOFTEE (https://github.com/konradjk/loftee) since every human carries at least a hundred loss-of-function variants predicted to severely disrupt the function of protein-coding genes. LOFTEE has recently been developed using pipeline inspired by MacArthur *et al* (93). It removes variants within short distance (less than 15bp) intronic regions, non-canonical (e.g., intron does not start with GT and end with AG) splice regions, LOF variants in the last 5% of the transcript, and variants where the LOF allele is the ancestral allele for that position.

Current methods for predicting the LOF variants are insensitive to many important classes of LOF variant such as splice-disrupting variants outside canonical splice sites. In addition, LOFTEE has been systematically validated their results using large-scale functional data sets to assess their accuracy in the detection of LOF variants. Therefore, each of their decisions on assessing LOF variants help to improve our confidence to predict variants. Using LOFTEE which can be run through the Ensembl Variant Effect Predictor (VEP) plugin, we were able to categorize stop_gain, splice site disrupting, and frameshift variants.

**B2. Sub-Aim 2: Investigating associations of rare-variants with breast cancer**

Hypothesis: In Aim 1, we hypothesized that common variants found in GWAS studies regulate gene expression, and rare coding variants of these genes may contribute to breast cancer. In order to investigate whether these rare coding variants contribute to breast cancer, we hypothesize that a significant proportion of the inherited susceptibility to breast cancer may be due to the summation of the effects of rare variants of a variety of different genes, each conferring a moderate but detectable increase in relative risk.

**B2.1 Study populations**

*Shanghai Breast Cancer Genetics Study (SBCGS)*

The Chinese participants were drawn from Shanghai Breast Cancer Genetics Study (SBCGS), which consists of the Shanghai Breast Cancer Study (SBCS), Shanghai Breast Cancer Survival Study (SBCSS), Shanghai Endometrial Cancer Study (SECS, contributed control data only), and the Shanghai Women's Health Study (SWHS), four large population-based studies in urban Shanghai. The SBCS is a two-phase (SBCS-I and SBCS-II) population-based case-control study that recruited incident patients with breast cancer and controls in urban Shanghai, the largest commercial center in China (49). In the initial phase (SBCS-I), subjects were recruited between August 1996 and March 1998. Two senior pathologists reviewed and confirmed cancer diagnoses for all patients. Controls were randomly selected from the general population using the Shanghai Resident Registry, a population registry containing demographic information for all residents of urban Shanghai. The inclusion criteria for controls were identical to those for cases with the exception of a breast cancer diagnosis. Our study used a structured questionnaire to elicit detailed information on demographic factors, and known/suspected risk factors for breast

cancer. All participants were measured for their current weight, height, and circumference of the waist and hips. All interviews were tape-recorded and reviewed by the field supervisor and quality control staff to monitor the quality of interview data. For both cases and controls, blood samples (10 ml from each woman) were obtained who completed the in-person interview. Using cotton swabs, a sample of exfoliated buccal cells was obtained from virtually all study participants who did not provide a blood sample. The second round of subject recruitment (SBCS-II) occurred between April 2002 and February 2005 using a protocol similar to the one used in the initial phase. Similar to the SBCS-I subjects, the majority of newly-recruited cases and controls provided a blood sample or an exfoliated buccal cell sample to the study. Our study used modified mouthwash method from initially reported by Lum *et al.* and provided, on average, approximately 34 µg of DNA per sample (94). Eligibility criteria for study participation were identical for SBCS-I and SBCS-II except age. The age ranged from 25 to 65 years for SBCS-I, and from 25 to 70 years in SBCS-II.

The SBCSS included newly diagnosed breast cancer cases ascertained via the population-based Shanghai Cancer Registry between April 2002 and December 2006 (49). In-person interviews were conducted to collect information on known breast cancer risk factors as well as anthropometrics using a protocol and questionnaire similar to that used in the SBCS. Patient medical charts were also reviewed to obtain detailed information on disease related characteristics and cancer treatment. Using the modified mouthwash method, buccal cell samples were collected from 96% of study participants.

The SECS is a population-based, case-control study of endometrial cancer conducted between January 1997 and December 2003 using a protocol similar to the SBCS; only community controls from the SECS were included in the present study (49). Except a few

questionnaires related specifically to breast or endometrial cancer risk, the questionnaires used in the SECS and the SBCS were virtually identical. Eligible cases were identified through the population-based Shanghai Cancer Registry and controls were randomly selected from the general population of Shanghai using the Shanghai Resident Registry and were age frequency matched to cases. Women with a history of cancer or hysterectomy were not eligible. Trained interviewers conducted in-person interviews to collect detailed information on demographic factors, menstrual and reproductive history, hormone use, prior disease history, physical activity, tobacco and alcohol use, weight, and family history of cancer. Anthropometrics measurements were taken.

The SWHS is a population-based cohort study that were recruited from urban Shanghai between 1997 and 2000 (95). The cohort has been followed by a combination of record linkage and active follow-ups (49). All these SBCGS studies are conducted among Chinese women in Shanghai, a genetically homogenous population, using virtually identical protocols in data and sample collection. Genomic DNA for all included participants was extracted using commercial DNA purification kits. All participants provided written informed consent prior to interview, and institutional review boards of all institutes in both China and the United States approved the study. Included in this study are 610 cases and 697 controls of SBCS-I, 1,651 cases and 1,539 controls of SBCS-II, 2,919 cases of SBCSS, 855 controls of SECS, and 586 cases and 2,612 controls of SWHS. A total of 11,469 (5,766 cases and 5,703 controls) participants from SBCGS were included in this study. Descriptive characteristics for study participants are presented in Table 1.

*Nashville Breast Health Study (NBHS)*

The Nashville Breast Health Study (NBHS) is a population based case-control study conducted between February 1, 2001 and December 31, 2008, in Nashville, Tennessee. Through a rapid case ascertainment system, we identified newly-diagnosed breast cancer cases through the Tennessee State Cancer Registry and five major hospitals in the city that provide medical care for breast cancer patients. Eligible cases were women who were newly diagnosed with primary breast cancer (invasive ductal or ductal carcinoma in situ) between the ages of 25 and 75 years old. They had no prior history of cancer other than non-melanoma skin cancer. The majority of participants (92%) were residents of the Nashville eight-county metropolitan area. Eligibility criteria for study participation included a resident telephone, English speaking, and capable of providing informed consent. Control subjects had virtually identical criteria to cases with the exception that they had no prior breast cancer diagnosis. Controls were identified mostly via random-digit dialing of households in the same geographic area as cases. Controls were frequency matched to cases on 5-year age group, race, and county of residence. Information on demographic factors, as well as known and suspected risk factors for breast cancer, was ascertained through a structured questionnaire administered via telephone interview. Two methods were used to collect buccal cell samples: Oragene saliva collection kits (DNA GenoteK, Ottawa, Canada) and mouthwash samples. This study was approved from the institutional review boards of Vanderbilt University Medical Center and of the individual collaborating institutions. All participants provided informed consent prior to enrollment in this study. A total of 2,965 (1,509 cases and 1,456 controls) European and 772 (500 cases and 272 controls) African American ancestry participants from NBHS were included in this study. Descriptive characteristics for European women are presented in Table 1.

*Southern Community Cohort Study (SCCS)*

The Southern Community Cohort Study (SCCS) is a prospective cohort study initiated in 2002 investigating racial disparities in the risk of cancer and other various chronic diseases (31). SCCS includes approximately 86,000 participants with two-thirds African American recruited in 12 southern states. Participants completed a comprehensive, in-person, baseline interview or completed a study questionnaire asking various aspects of health conditions, behavioral factors, personal and family medical history, and other lifestyle factors. Once these participants return a completed questionnaire and signed consent form, they were asked to self-collect a buccal cell using the swishing method and mail it back to the lab at Vanderbilt. In the SCCS, 534 breast cancer cases of African American women were included in this study who were diagnosed with breast cancer. In addition, 534 controls of AA women were selected randomly from those who were cancer-free and frequency-matched to cases in a 1:1 ratio on age at enrollment (±1 year), recruitment method, and sample type (blood/buccal cell). Additional AA controls (n = 247) were selected from cancer-free SCCS participants and frequency-matched to NBHS cases by age (±1 y), family income, and education in order to increase the statistical power. A total of 1,034 cases and 1,053 controls of African American participants from the SCCS and NBHS were included in the current study.

*BioVU (the Vanderbilt DNA Databank)*

The Vanderbilt's biorepository, BioVU, is composed of electronic medical records scrubbed of personal identifiers, linked to coded DNA samples. BioVU accrues DNA samples extracted from blood remaining from routine clinical testing after the samples have been retained for three days and are scheduled to be discarded. A full description of BioVU including its design, collection methods, and ethical considerations have been published elsewhere (32). Biological samples from BioVU are linked through an anonymous research unique identifiers to the Synthetic Derivative, a de-identified version of Vanderbilt's electronic medical record. Using the Synthetic Derivative, candidates were identified using informatics methods, cancer registry data and ICD-9 code. In the BioVU Breast Cancer Study, 695 cases and 3,784 controls of European women were included in this study.

**Table 1. Participants included in current study**

| Study (N = sample size) | Cases | Controls | Ethnicity | Age (year, mean ± sd) |
|---|---|---|---|---|
| SBCGS (N = 11,469) | 5,766 | 5,703 | Chinese | 53.08 ± 9.46 |
| NBHS (N = 2,965) | 1,509 | 1,456 | European American | 52.73 ± 9.19 |
| NBHS/SCCS (N = 2,087) | 1,034 | 1,053 | African American | 54.59 ± 9.79 |
| BioVU (N = 4,479) | 695 | 3,784 | European American | 57.55 ± 20.40 |
| **Total (N = 21,000)** | **9,004** | **11,996** | | |

## B2.2. Genotyping Method

All Chinese women from SBCGS were genotyped using the Asian Exomechip, an expanded Illumina HumanExome-12v1_A Beadchip. In order to improve the coverage for the low frequency variants in Asian population, we added additional customer content variants onto the Illumina HumanExome-12v1_A Beadchip. The original Exome array includes 247,870 markers focused on protein-altering variants selected from sequencing data in >12,000 subjects, mostly from European ancestry populations. In the Asian Exomechip, the additional variants were primarily selected from exome sequencing in 581 Chinese women from SBCS, exome sequencing in 496 Singapore Chinese, and Asian data in the 1000 Genomes Project. We added nonsynonymous, splicing and stop-altering variants observed two or more times in any of these datasets or once in any two of the three datasets.

All EA and AA women from NBHS and SCCS data were genotyped using the Illumina HumanExome-12v1_A Beadchip, which includes 247,870 markers focused on protein-altering variants selected from sequencing data in >12,000 subjects. Details about SNP content and selection strategies were described at http://genome.sph.umich.edu/wiki/Exome_Chip_Design. In brief, the Illumina HumanExome BeadChip is enriched for rare and low frequency coding variations previously identified from the sequenced exomes of approximately 12,000 individuals of diverse populations for variations seen in more than two individuals and in more than two sequencing projects. Nonsynonymous variants had to be observed three or more times in at least two studies, and splicing and stop-altering variants had to be observed two or more times in at least two studies.

All samples included in SBCGS, NBHS, and SCCS were genotyped at the Genome Quebec Innovation Centre (Montreal, Quebec, Canada) following Illumina's protocol. On each

96-well plate, blind duplicate samples and two HapMap samples were included as quality control (QC). Genotype calling was carried out using Illumina's GenTrain version 2.0 clustering algorithm in GenomeStudio version 2011.1. We used study samples to determine cluster boundaries. After clustering, ~80,000 variants were manually reviewed and clusters were edited for 27,506 variants.

For BioVU data, all EA women were genotyped using the Illumina HumanExome Beadchip. Genotyping was performed at the Vanderbilt Technologies for Advanced Genomics (VANTAGE) Core, and genomic data were processed by the Vanderbilt Technologies for Advanced Genomics Analysis and Research Design (VANGARD) Core. GenomeStudio's GenTrain and GenCall were used for clustering and genotype calling. The details have been previously described (https://victr.vanderbilt.edu/pub/biovu).

**B2.3. Quality control (QC) for genotype data**

After calling genotypes using GenomeStudio version 2011.1, further QC procedures are conducted using plink (http://pngu.mgh.harvard.edu/~purcell/plink/). Concordance rates are evaluated for HapMap samples genotyped in our study and sequenced by the 1,000 Genomes Project (http://www.1000genomes.org/). Pair-wise proportion of identity-by-descent (IBD) is estimated to identify potentially genetically identical, unexpected duplicated samples or close relatives.

The samples were excluded in the following criteria. First, samples with genotype call rates of < 98% were removed from analysis. The genotype call rate is defined as the fraction of called SNPs per sample over the total number of SNPs in the dataset. Second, samples were excluded if consistence rates between the HapMap samples with 1000 Genomes data is < 99%. Additionally, samples with wrong sex, heterozygosity outlier, ethnic outliers, or consistence rates

among duplicated samples < 99% were removed. Due to the assumption of independent sampling in our dataset, individuals who have familial relations with each other need to be removed. Thus, samples with close relationship were excluded by IBD analyses.

The SNPs were excluded in the following criteria. First, SNPs with MAF of 0, or SNP call rates of < 98%, or genotyping concordance rate < 98% in QC samples were removed. Additionally SNPs that violate the Hardy Weinberg Equilibrium at a P-value threshold of < $1\times10^{-5}$, or redundant SNPs were excluded from the dataset. We also excluded cautious SNPs discovered by the exome-chip design group (http://genome.sph.umich.edu/wiki/Exome_Chip_Design#Cautious_Sites).

## B2.4. Treatment of Confounding

Population stratification refers to differences in allele frequencies between cases and controls due to systematic differences in ancestry rather than association of genes with disease. In GWAS, population stratification is a major confounding factor for case-control association studies and can result in false positive associations since the association found could be due to the underlying structure of the population and not a disease associated locus (96, 97). Therefore, population stratification can be confounders depending on which data people used in their GWAS.

When analyzing rare variants, it is especially important to adequately control for population substructure since rare variants tend to have occurred more recently and therefore have greater population diversity than common variants. Figure 1 shows that the rarer a genetic variant is within a population, the less likely it is to be found in all ethnic groups (98). If a GWAS identified genetic marker is linked to a mixture of common and rare causal alleles, some

of the rare ones are likely to differ in frequency in different populations, or even be completely

absent in some populations (98).

**Figure 1. Rare Alleles More Likely Population-Specific (One hundred people were sampled from each population).**
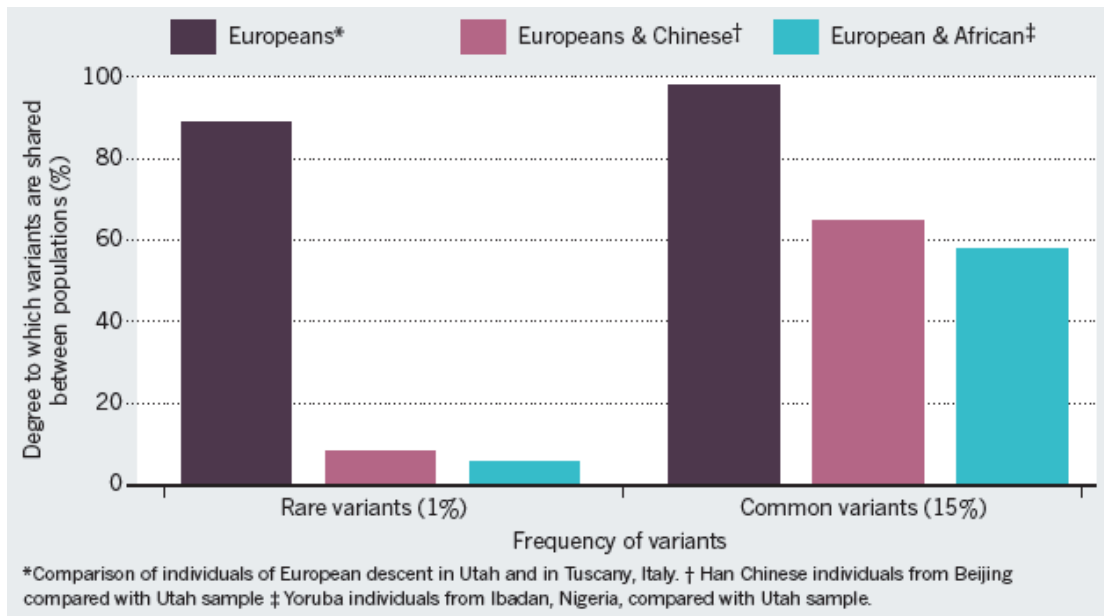


*Figure reprinted from Bustamante et al. Genomics for the world. Nature 2011;475(7355):163-5 (98)*

In GWAS, principal-component analysis (PCA) and linear mixed models are commonly

used to adjust for population stratification (99). In this study, principal components analyses

(PCA) were conducted using EIGENSTRAT

(http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm) to identify population outliers with

the 1,000 Genomes Project data as reference.

## B2.5.  Data analysis

### B2.5.1. Single-variant tests

The association between each genetic variant and a disease trait is typically evaluated by logistic regression for binary traits. The standard approach in GWAS to testing for association between genetic variants and complex traits is a single-variant test under an additive genetic model. Single-variant tests can also identify association with rare, low-frequency variants if sample sizes are large enough. However, single-variant test of rare variants has very low power for detecting association than common variants with identical effect sizes, due to extremely low frequency (usually $< 0.01$) (100).

Single-variant tests are still a useful tool for rare-variant analysis if the sample sizes and the effects are large enough. It should be considered that single-variant-based P-value estimates based on standard regression methods might not be accurate if the number of subjects with the variant is small since the requisite multiple test corrections are poorly understood. This issue will require more methodological development. Due to the large sample size of SBCGS Chinese population, we first conducted single-variant analysis adjusted for the five first principal components (PCs). Studies have shown that firth test is best for joint analysis in both balanced and unbalanced studies, and the score test is best for meta-analysis in balanced studies only (101, 102). Firth logistic regression introduces a more effective score function by adding a term that counteracts the first-order term from the asymptotic expansion of the bias of the maximum likelihood estimation (102, 103). We used score and firth test for logistic regression implemented in Rvtests (http://genome.sph.umich.edu/wiki/Rvtests). First five principal components were adjusted for both score and firth tests. In addition to logistic regression, Fisher's exact test is another commonly used case-control test for rare variants since it guarantees type I error control

for small sample sizes or low variant frequency. It is a conservative test and thus has its power

diminished to some extent. However, due to the very low allele frequency (usually $< 0.01$) for

rare variant, the expected number in any cell of the contingency table might be small. Therefore,

we conducted Fisher's exact test as a secondary test to confirm the association signal from

logistic regression. Further conditional analyses were conducted by adjusting the index SNP in

each locus (SNPs within 1Mb flanking regions of the index SNP) in order to recognize

independent association signals.

## B2.5.2. Gene-Based Aggregation Tests

Rare variants are more abundant than common variants in the human genome, and

controlling for multiple testing problems becomes a severe issue for any single-variant-based

analysis. To address these questions, researchers have recently developed statistical methods

specifically configured for rare-variant association analysis to increase power. These methods

evaluate cumulative effects of multiple variants in a biologically relevant region, such as a gene,

instead of testing the effects of single variants which is commonly done in GWAS. Power will be

increased when multiple variants in the group are associated with a given disease or trait.

Numerous methods have been developed to aggregate information across several variant

sites within a gene to enrich association signals and to reduce the penalty of multiple testing. The

methods we are using are all regression-based methods. The simplest approach is the burden test,

which creates a burden score for each subject by taking a weighted linear combination of the

mutation counts within a gene or indicating whether there is any mutation within a gene (104,

105). The summary genetic score, $s_i$ (weighted sum test) is then:

$$s_i = \sum_{j=1}^{m} w_j g_{ij}$$

where $w_j$ is a threshold indicator or weight for variant $j$, $i$ is subject, and $g_{ij}$ is allele counts for

$m$ variants of interest. For the Combined Multivariate and Collapsing (CMC) method, the genetic

variables contain the genotypes of common variants and the burden scores of rare variants (105).

CMC method provides the ability to easily adjust for covariates. The Madsen-Browning burden

test calculates $S_i$ based on allele frequency in control group (104). These simple methods are

powerful when a large proportion of variants are causal and effects are in the same direction

(106). However, they might lose power in the presence of both trait-increasing and trait-

decreasing variants or a small fraction of causal variants (106). A second approach is the variable

threshold (VT) test, which performs a burden test for variants with MAFs below a certain

threshold and minimizes the P-value over the observed MAF thresholds (107, 108). This

adaptive burden test is generally more robust than the original burden methods since they use

fixed weights or thresholds (106). However, adaptive burden tests are often computationally

intensive due to the required permutation to estimate P-values. A third approach is the variance-

component test, which is designed to detect variants with opposite effects within a gene using a

variance-component test within a random-effects model (109–111). The sequence kernel

association test (SKAT) is one of the variance-component tests in which regions can be defined

by genes (in candidate-gene or whole-exome studies) or moving windows across the genome (in

whole-genome studies) (111). For each region, SKAT analytically calculates a P-value for

association while adjusting for covariates (111). Because SKAT evaluates significance via a

score test, which operates under the null hypothesis, the type I error is protected irrespective of

the kernel and the weights used (111). The SKAT method is powerful in the presence of both

trait-increasing and trait-decreasing variants, but less powerful when most variants are causal and

effects are in the same direction (106). Also, all three approaches are based on varying

assumptions about the underlying genetic model, and power for each test depends on the true disease model. For these reasons, we applied all three approaches in this study. We conducted gene-based tests including Madsen-Browning burden test, CMC, variable threshold burden test, and SKAT for the Chinese, EA, and AA populations. All methods are implemented in Rvtests (http://genome.sph.umich.edu/wiki/Rvtests). Rare variants with MAF ≤ 0.01 or MAF ≤ 0.05 within each gene were aggregated for nonsynonymous and LOF variants separately based on the functional prediction results from Sub-Aim 1. Because covariates can be incorporated in CMC and SKAT tests, we adjusted for the five first principal components (PCs) in order to control for potential confounders. In addition, we carried out conditional analyses to recognize independent association signals by adjusting the index SNP in each locus (SNPs within 1Mb flanking regions of the index SNP).

### B2.5.3. Rare-variant Meta-Analysis

Meta-analysis of GWAS has led to the discoveries of common genetic variants for many complex human diseases by providing an effective way to combine data from multiple studies (112–114). Rare-variant meta-analysis can be performed efficiently with simple study-specific summary statistics for the construction of rare-variant test statistics across large numbers of samples. Meta-analysis is especially important in rare-variant association studies since detecting rare-variant associations requires large sample sizes. The traditional meta-analysis method is to combine P-values across studies by using Fisher's or Stouffer's Z score methods (112, 115). However, this approach has been known that it is less powerful than joint analysis of individual-level data and fixed-effects meta-analysis (112). Fixed-effects meta-analysis can use individual-level data to achieve power essentially identical to that of joint analysis (116, 117).

In the current study, meta-analyses of single-variant results from both score and firth tests were conducted using the fixed-effect inverse variance method to combine the $\beta$ estimates and standard errors from each dataset (SBCGS, NBHS, SCCS and BioVU). This method is implemented in METAL software (118). This approach weights the effect size estimates, or $\beta$-coefficients using the inverse of the corresponding standard errors, and it also requires effect size estimates and their standard errors to be in consistent units across studies (118). For meta-analyses of single-variant results from Fisher's exact test, a P-value based combined method proposed by Michael *et al.*(119) was used. Similar to the approach used by METAL software, P-value based combined method used the z-statistic which summarizes the direction of effect relative to the reference allele, and then an overall z-statistic and p-value are calculated from a weighted sum of the individual statistics. Weights are proportional to the square-root of the number of samples in each study. For a study with unequal numbers of cases and controls, we used the effective sample size, where $N_{eff} = 4/(1/N_{cases}+1/N_{ctrls})$ (118).

For meta-analysis of gene-based results, first, summary statistics and covariance matrices of score statistics for each study were generated by RAREMETALWORKER (120). Compared to the traditional meta-analysis, combined score statistics improve computational efficiency (given that only a null model shared between markers needs to be fit) and provide numerical stability (since one does not need to estimate regression coefficients and their standard errors, which is difficult for rare variants) (106). Then, RAREMETAL was used to conduct gene-based meta-analysis using Madsen-Browning burden test, CMC, variable threshold burden test, and SKAT (120). The main idea for RAREMETAL is that gene-level test statistics can be reconstructed from single variant score statistics and their covariance matrix, and that, when LD

relationships between variants are known, the distribution of gene-level statistics can be derived to evaluate significance (120).

Using these meta-analyses, we have more power to detect true associations between rare variants and breast cancer risk by combining the Chinese, EA, and AA populations.

### B2.5.4. Compound Heterozygous analysis

Common variants in GWAS loci were detected by the additive approaches since there are many homozygotes observed for these common variants, and therefore strong signal exists even under the additive model (121). However, the power of the additive model to detect recessive alleles could reduce dramatically at lower frequencies since the numbers of homozygotes observed are far fewer. Therefore, it is very important to detect rare recessive variants that confer significant risk in a recessive manner. Several studies have reported the importance of rare recessive variants associated with complex diseases including autism and schizophrenia (122, 123).

The compound heterozygous (CH) is a recessive model in which the two haplotypes have to carry at least one rare allele each (124). For Autism Spectrum Disorders (ASD), Lim *et al*. found and confirmed the inheritance of two previously unreported compound heterozygous nonsense mutations in *USH2A* gene from both parents (125). Recently, Chen *et al*. developed a general framework for group-wise TDT (gTDT) which is haplotype-based and models the transmission of rare variant carrying haplotypes (124). Their study focused on Transmission/Disequilibrium Tests (TDT) based on family designs, and evaluated the power of gTDT using CH analysis (124). Although CH is commonly observed in Mendelian diseases, it may also play a role in complex disease (122, 126).

In the current study, we conducted CH analysis in order to investigate recessive models within each gene that are associated with breast cancer risk (124). To our knowledge, there is no study to conduct CH analysis in breast cancer using whole-exome chip data. First, genotypes had been phased using SHAPEIT v2 (segmented haplotype estimation and imputation tool) in order to identify CH (127). After phasing, we designed coding schemes to see which sample is CH for a particular gene: i) we identified the haplotypes for each gene and each sample, ii) if both haplotypes carry rare variants, then it is selected as a CH (coded as 1), and non-CH (coded as 0) otherwise. After identifying CHs, we constructed a 2x2 table for each gene by counting number of samples with CH and without CH in cases and controls. Fisher's exact test has been used to test statistically significant associations between CH and case/control status for each gene. All analyses have been conducted using R statistical language (http://www.r-project.org/) and Perl programming language (http://www.perl.org/). For meta-analyses of CH results from each study, a P-value based combined method proposed by Michael *et al.*(119) was used.

## B2.5.5. Functional prediction of identified variants

For the identified variants from association results of Sub-Aim 2, we predicted potential damaging effects using three different algorithms; SIFT algorithm (Sorting Intolerant From Tolerant), PolyPhen-2 (Polymorphism Phenotyping v2), and PROVEAN (Protein Variation Effect Analyzer). All three algorithms use alignment-based score in order to predict the damaging effects of variants.

The SIFT algorithm was developed to predict whether an amino acid substitution affects protein function (128). SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST (128). If the SIFT score is equal to or below a predefined threshold (0.05), the variant is

predicted to have a "deleterious" effect, and if the score is above the threshold, the variant is predicted to have a "tolerated" effect. PolyPhen-2 predictions are calculated for all resulting amino acid residue substitutions in human UniProtKB proteins with the maximum coding sequences (CDS) sequence overlap and identity (129). If the PolyPhen-2 score is greater or equal to 0.957, the variant is predicted to have a "probably damaging" effect, if the score is in between 0.453 and 0.956, then the variant is predicted to have a "possibly damaging" effect, otherwise it is predicted as benign ($\leq 0.452$). Choi *et al*. have developed a novel prediction algorithm with expanded functions, PROVEAN, which supports functional predictions for SNPs as well as insertions, deletions, and replacements of amino acids at the protein level (130, 131). If the PROVEAN score is equal to or below a predefined threshold (e.g. -2.5), the protein variant is predicted to have a "deleterious" effect. If the PROVEAN score is above the threshold, the variant is predicted to have a "neutral" effect.

Those algorithms provided us computational predictions of whether identified variants are likely to be damaging (whether missense alleles are null or neutral).

# CHAPTER V

# FINDINGS FOR SPECIFIC AIM 1: IDENTIFY POTENTIAL FUNCTIONAL GENES IN THE PREVIOUSLY REPORTED GWAS LOCI ASSOCIATED WITH BREAST CANCER RISK.

## A.    Results

We selected 109 GWAS loci associated with breast cancer risk using publicly available databases. There were 3,851 genes in 1 Mb flanking regions of 109 GWAS loci. For eQTL analysis using TCGA data, we adjusted CNV and DNA methylation since tumors acquire frequent genetic and epigenetic alterations, which can affect gene expression (89). Although CNV and DNA methylation are related to somatic mutation and gene expression, there is still lack of correlation between CNV and DNA methylation, and genetic variants. Heyn *et al*. showed that one-third of the DNA methylation differences were not associated with any genetic variation (132). Wagner *et al*. recently reported significant correlation between gene expression and DNA methylation in developmentally significant regions having little or no discernible involvement of DNA sequence variation (133). By definition of confounding in epidemiology area, confounding variable has to be causally associated with the outcome (gene expression) and non-causally or causally associated with the exposure (genetic variants). By definition, CNV and DNA methylation are not strong confounders, but they still have potential minimal confounding effect.

We checked the relationship between adjusted and unadjusted TCGA data for CNV and DNA methylation. In total, 27,831 SNPs were overlapped between adjusted and unadjusted TCGA data after using certain criteria (eQTL P-value < 0.05, RSQR > 0.8, MAF > 0.05). We

found potential confounding effect of CNV and DNA methylation between adjusted and unadjusted TCGA data (Figure 2). Plot showed the -log10 P-values (y-axis and x-axis) for each SNP.

**Figure 2. Plot showing relationship between adjusted and unadjusted TCGA data for CNV and DNA methylation (eQTL P-value < 0.05, RSQR > 0.8, MAF > 0.05)**
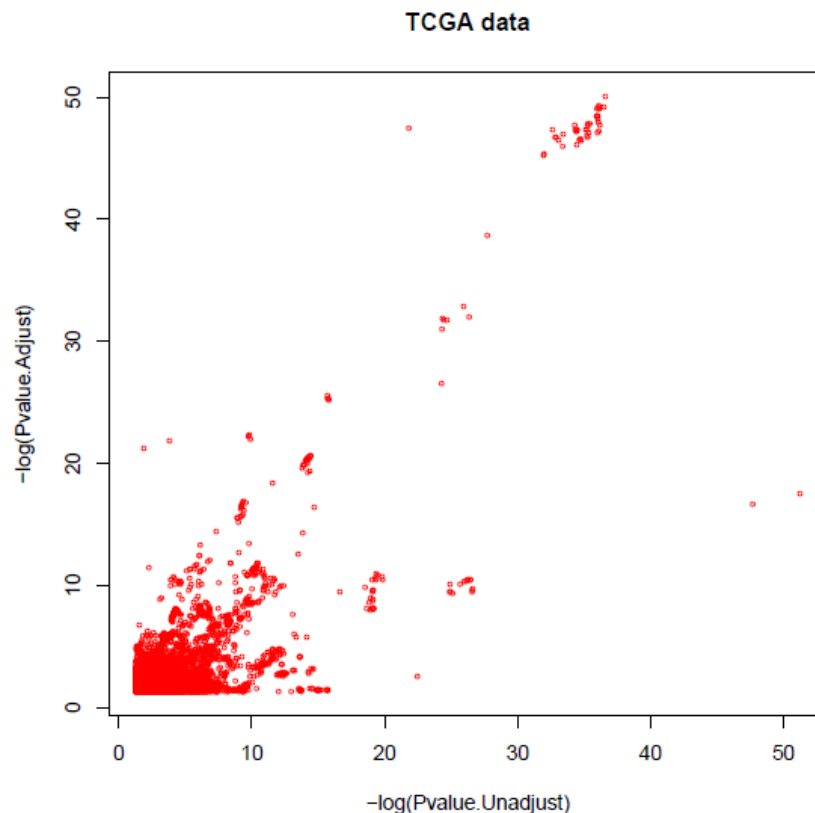


For eQTL analysis using METABRIC data, we were not able to adjust CNV and DNA methylation because they were not provided. We might have a reduced power to conduct eQTL analysis using METABRIC since CNV and DNA methylation have potential minimal confounding effect, but we would not have an inflated type I error. Finally, we compared eQTL results using all three datasets (TCGA, METABRIC, GTEx) with and without adjusting CNV and DNA methylation for TCGA data (Figure 3 and Figure 4 respectively). Total 1,050 genes were overlapped among all three datasets after adjusting CNV and DNA methylation for TCGA

data (1,198 genes from TCGA, 1,235 genes from METABRIC, and 1,640 genes from GTEx) (Figure 3). Total 1,046 genes were overlapped among all three datasets without adjusting CNV and DNA methylation for TCGA data (1,253 genes from TCGA, 1,235 genes from METABRIC, and 1,640 genes from GTEx) (Figure 4). We also checked the number of genes identified from eQTL analyses with different thresholds (Table 2).

**Figure 3. Venn diagrams showing number of breast cancer candidate genes from TCGA, METABRIC, and GTEx[a]**
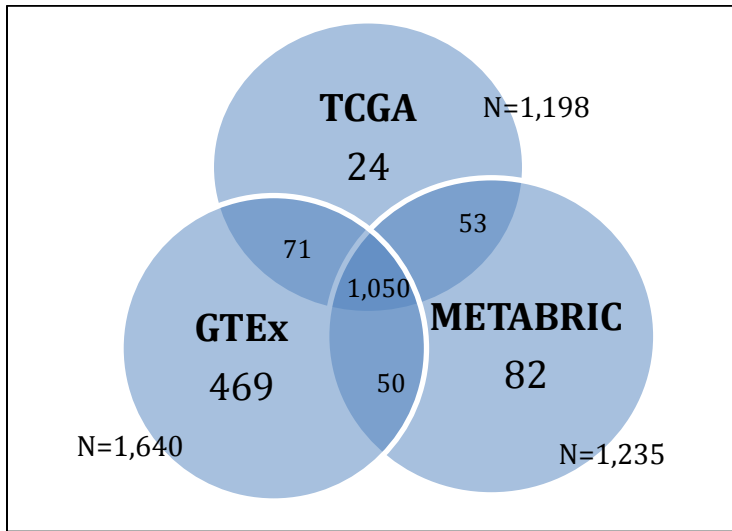


[a] TCGA breast cancer tumor tissue was adjusted for CNV and DNA methylation.

**Figure 4. Venn diagrams showing number of breast cancer candidate genes from TCGA, METABRIC, and GTEx[a]**
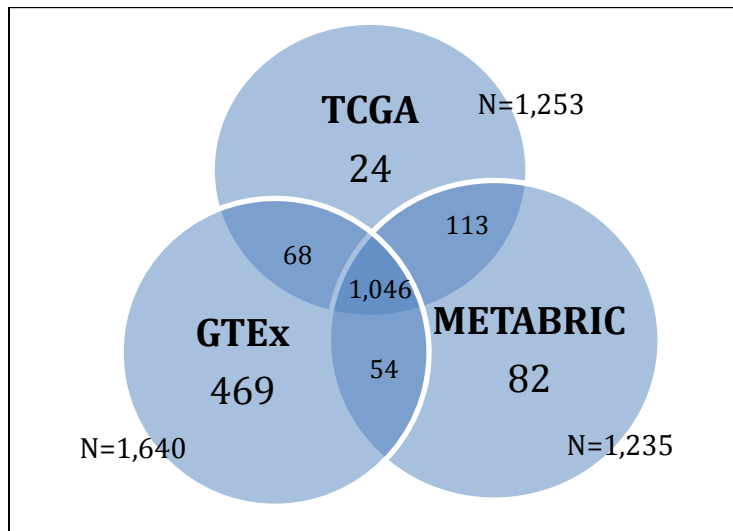


[a] TCGA breast cancer tumor tissue was not adjusted for CNV and DNA methylation.

**Table 2. Number of genes identified from eQTL analysis using three datasets.**

| | | Number of Genes | |
|---|---|---|---|
| | | A | B |
| P < 0.01 | TCGA | 41 | 886 |
| | METABRIC | 54 | 544 |
| | GTEx | 72 | 913 |
| | Union of 3 datasets | 133 | 1,343 |
| P < 0.05 | TCGA | 101 | 1,198 |
| | METABRIC | 133 | 1,235 |
| | GTEx | 150 | 1,640 |
| | Union of 3 datasets | 329 | **1,799** |
| FDR < 0.05 | TCGA | 14 | 539 |
| | METABRIC | 28 | 813 |
| | GTEx | 34 | 731 |
| | Union of 3 datasets | 60 | 860 |

A: eQTL results using 109 GWAS index SNPs (adjusting CNV and DNA methylation for TCGA)

B: eQTL results using flanking 1Mb of 109 GWAS index SNPs (adjusting CNV and DNA methylation for TCGA)

## B.     Discussion

We performed *cis*-based eQTL analyses for all genes 1Mb flanking the 109 GWAS loci which have been identified from breast cancer GWAS. Since we have CNV and DNA methylation data for TCGA, but not for METABRIC, we compared the differences between adjusted and unadjusted TCGA data for CNV and DNA methylation. Results showed that there is potential minimal confounding effect of CNV and DNA methylation on gene expression.

From all three datasets (TCGA, METABRIC, GTEx), a total of 1,799 genes (union of three datasets) 1Mb flanking the 109 GWAS loci at a P-value of 0.05 were selected with expression level associated with breast cancer risk-associated SNPs after adjusting CNV and DNA methylation for TCGA. With the 109 GWAS index SNPs, a total of 329 genes (union of three datasets) at a P-value of 0.05 were selected from eQTL analysis. We did not consider association directions for genes because underlying biological mechanisms have not been characterized for genetic variants that are involved in gene pathways. Even though a gene is known to be associated with increased risk of breast cancer, we are not sure whether genetic variants in this gene are associated with increased risk of breast cancer or decreased risk of breast cancer.

Human gene regulation is often mediated by distal enhancer elements. Also, including more genes 1Mb flanking the 109 GWAS loci could give us strong signal due to the many multiple comparisons. Therefore, we decided to use 1,799 genes 1Mb flanking the 109 GWAS loci (P-value < 0.05) for further analysis in the following aims. We are not aware of any study that evaluates eQTLs using all three major databases. We have used the most updated datasets for all three databases, and no study has yet reported eQTL results using all 109 GWAS loci that have been found to be associated with breast cancer risk so far. The mechanistic basis for the

association between breast cancer and most of the common variants discovered in GWAS is still

largely unknown. Common variants found in GWAS studies can affect gene expression level,

and missense/nonsense variants in these genes may change expression level, structure and

function. Therefore, our comprehensive eQTL analysis would help to find these rare variants

which may contribute to breast cancer.

# CHAPTER VI

# FINDINGS FOR SPECIFIC AIM 2: INVESTIGATE RARE VARIANTS ASSOCIATED WITH BREAST CANCER RISK.

## A. Sub-Aim 1: Functional prediction of rare coding variants neighboring common GWAS loci

### A1.    Results

We predicted nonsynonymous variants located flanking 1Mb of the indexed SNP of 109 GWAS loci using ANNOVAR. As shown in Table 3, we predicted 7,161 variants from SBCGS, 7,233 variants from NBHS, 8,192 variants from NBHS/SCCS, and 7,681 variants from BioVU. Then, we compared total number of LOF variants predicted from ANNOVAR and LOFTEE since there are no classically defined LOF variants. ANNOVAR includes frameshift, stop_gain, and stop_loss as LOF category, and LOFTEE includes frameshift, stop_gain, splice_donor_variant, and splice_acceptor_variant as LOF category. In total, 130 (160) (ANNOVAR (LOFTEE)) variants from SBCGS, 121 (152) variants from NBHS, 128 (159) variants from NBHS/SCCS, and 143 (176) variants from BioVU were predicted as LOF variants. There were no frameshift variants predicted by either ANNOVAR or LOFTEE. Most of the LOF variants were predicted from stop_gain variants category which are sequence variants whereby at least one base of a codon is changed, resulting in a premature stop codon or leading to a shortened transcript.

**Table 3. Summary of annotation from ANNOVAR and LOFTEE (Number of Nonsynonymous, Synonymous, LOF variants) in 1Mb flanking the 109 GWAS loci[a]**

| | Per study | | | |
|---|---|---|---|---|
| | SBCGS (Asian) | NBHS (EA) | NBHS/SCCS (AA) | BioVU (EA) |
| Nonsynonymous (ANNOVAR) | 7,161 | 7,233 | 8,192 | 7,681 |
| Synonymous (ANNOVAR) | 376 | 334 | 404 | 357 |
| LOF from ANNOVAR | 130 | 121 | 128 | 143 |
| frameshift | 0 | 0 | 0 | 0 |
| stop_gain | 129 | 114 | 119 | 135 |
| stop_loss | 1 | 7 | 9 | 8 |
| | | | | |
| LOF from LOFTEE | 160 | 152 | 159 | 176 |
| frameshift_variant | 0 | 0 | 0 | 0 |
| stop_gain | 121 | 104 | 110 | 126 |
| splice_donor_variant | 20 | 24 | 30 | 25 |
| splice_acceptor_variant | 19 | 24 | 19 | 25 |

[a]Shown are the total number of Nonsynonymous, Synonymous, LOF variants observed.

We checked the number of candidate LOF variants in 1Mb flanking the 109 GWAS loci per individual in each study using LOFTEE. The average number of candidate LOF variants was 1.5 for Asian from SBCGS, 1.9 for EA from NBHS, 2.9 for AA from NBHS/SCCS, and 1.6 for EA from BioVU (Table 4). On average, AA woman has the largest number of candidate LOF variants. The mean number of candidate LOF variants was similar in both Asian and EA women.

**Table 4. Number of LOF variants per individual in 1Mb flanking the 109 GWAS loci**

|  | Average per individual | | | |
|---|---|---|---|---|
|  | **Mean** | **SD** | **Min** | **Max** |
| SBCGS (Asian) | 1.5 | 1.0 | 0.0 | 7.0 |
| NBHS (EA) | 1.9 | 1.2 | 0.0 | 8.0 |
| NBHS/SCCS (AA) | 2.9 | 1.6 | 0.0 | 10.0 |
| BioVU (EA) | 1.6 | 1.1 | 0.0 | 7.0 |

## A2.    Discussion

The most recognized deleterious variants are those that disrupt a protein-coding gene either by leading to loss of function or by altering an amino acid. Among the analysis steps for rare variant study, functional prediction (of being deleterious) plays an important role in filtering or prioritizing nonsynonymous and LOF variants for further analysis. Thus, we prioritized those variants separately using ANNOVAR and LOFTEE for further rare variant association analysis.

Every human carries at least a hundred LOF variants predicted to severely disrupt the function of protein-coding genes. Discovering LOF variants in the human population remains a significant challenge since these variants can be annotated inaccurately. In order to overcome this issue, LOFTEE has been introduced for predicting many important classes of LOF variant, such as splice-disrupting variants outside canonical splice sites which are not captured by other methods such as ANNOVAR. LOFTEE removes variants within short distance (less than 15bp) intronic regions, and non-canonical splice region (i.e. intron does not start with GT and end with

AG). They have been systematically validated their software using large-scale functional data sets to assess their accuracy in the detection of LOF variants. They did not include stop_loss and start_gain variants since they found that those variants may or may not truncate or remove any sequence. They had also considered adding start_loss variant since this could ablate transcript, but they found that these variants are not too deleterious based on allele frequency and functional data information (they will be publishing these data soon). Therefore, using LOFTEE, we were able to predict more LOF variants than ANNOVAR.

As shown in Table 3, we found that most candidate LOF variants are predicted from stop_gain variants category from both ANNOVAR and LOFTEE. Stop_gain variants are prevalent, having an estimated number of 100 to 200 occurrences per human genome (134, 135). As discussed previously, different LOF prediction algorithms (ANNOVAR and LOFTEE) use different information to prioritize LOF variants. We used LOF variants predicted from LOFTEE since they have improved variant annotations. Our functional prediction approaches provided us meaningful candidate nonsynonymous and LOF variants which we used for further analysis.

**B. Sub-Aim 2: Associations between rare-variants and breast cancer risk in Chinese, European American, and African American populations.**

**B1.     Results**

*Single-variant analysis results*

Total 15,033 SNPs from SBCGS, 7,729 SNPs from NBHS (EA population), 8,686 SNPs from NBHS/SCCS (AA population), and 8,160 SNPs from BioVU were included in analyses. All analyses were adjusted for index SNPs. We found several rare variants with MAF < 0.01 that were associated with breast cancer risk at P-value < 0.01 (unless otherwise stated, "P-value" refers to the P-value obtained from logistic regression (score test)); 7 SNPs from the Asian population (Table 5), 12 SNPs from the EA population (NBHS) (Table 6), 7 SNPs from the AA population (Table 7), and 38 SNPs from the EA population (BioVU) (Table 8). For BioVU, 6 missense variants were associated with breast cancer risk at P-value < 0.001 (MAF < 0.01) among 38 rare variants. For Asian, 6 missense variants (chr4:107181660, rs190673256, rs114365673, rs201444816, chr11:68530105, and chr22:29656346) were associated with breast cancer risk at P-value < 0.01 (Table 5). Among them, 4 SNPs were predicted to be "damaging" based on three functional prediction algorithms (chr4:107181660 (Ile->Asn), rs114365673 (Arg->His), rs201444816 (Asp->Glu), and chr11:68530105 (Pro->Leu)).

For EA and AA populations (NBHS), all rare variants (MAF < 0.01) were identified as missense variants at a P-value of < 0.01. Ten of the 12 missense variants, and 5 of the 7 missense variants were predicted to be "damaging" in EA (NBHS) and AA (NBHS) populations, respectively. For BioVU, a total of 36 missense variants and 2 stop_gain variants were associated with breast cancer risk at P-value < 0.01 (MAF < 0.01) (Table 8). Among 36 missense variants, 26 missense variants were predicted to be "damaging" from at least one of the three functional

49

prediction algorithms. We also compared P-values obtained from logistic regression analyses of score test and firth test with fisher's exact test since it controls type I error for low variant frequency. We found consistent results from three test statistics in our datasets. Results from conditional analysis adjusted for index SNPs were consistent with the results without adjustment for index SNPs.

**Table 5. Associations of breast cancer with SNPs with MAF<0.01 and P-value<0.01 among Asian population (SBCGS)[a]**

| Gene | SNP (Alleles)[b] | Chr:Pos[c] | Annotation | Amino acid change (Polyphen-2 score/SIFT score/PROVEAN score) | No. of Samples (case/control)[d] | OR (95%CI) [e] | P-value[f] | P-value[g] | P-value[h] |
|---|---|---|---|---|---|---|---|---|---|
| *ATP2B4* | chr1:203671173 (C/T) | chr1:203671173 | Synonymous | | 21/6 | 3.46 (1.62-7.36) | 0.004 | 0.006 | 0.009 |
| *TBCK* | chr4:107181660 (T/A) | chr4:107181660 | Missense | Ile->Asn (0.547/0.001/-5.07) | 4/16 | 0.25 (0.10-0.59) | 0.007 | 0.007 | 0.015 |
| *SYNE1* | rs190673256 (T/C) | chr6:152560708 | Missense | Arg->Gln (0.002/0.76/0.41) | 28/8 | 2.57 (1.38-4.79) | 0.007 | 0.003 | 0.015 |
| *CPA1* | rs114365673 (A/G) | chr7:130023254 | Missense | Arg->His (0.915/0.077/-3.93) | 5/17 | 0.29 (0.12-0.66) | 0.009 | 0.010 | 0.017 |
| *DMRTA1* | rs201444816 (G/C) | chr9:22447094 | Missense | Asp->Glu (0.888/0/-0.54) | 35/14 | 2.35 (1.34-4.13) | 0.006 | 0.004 | 0.008 |
| *CPT1A* | chr11:68530105 (A/G) | chr11:68530105 | Missense | Pro->Leu (0.996/0.035/-5.24) | 13/2 | 6.4 (2.32-17.64) | 0.005 | 0.007 | 0.015 |
| *RHBDD3* | chr22:29656346 (G/A) | chr22:29656346 | Missense | Trp->Arg (0.003/0.842/-0.13) | 11/29 | 0.37 (0.20-0.69) | 0.004 | 0.0041 | 0.007 |

[a] All SNPs with MAF < 0.01. All SNPs with P-value < 0.01 in at least two of the test statistics.

[b] Effect allele/reference allele.

[c] Chromosome position (bp) based on NCBI Human Genome Build 37.

[d] Number of samples carrying heterozygous variant.

[e] OR (95% CI) was adjusted for first five principal components.

[f] P-value obtained from logistic regression analysis (score test)..

[g] P-value obtained from fisher's exact test.

[h] P-value obtained from firth logistic regression analysis.

**Table 6. Associations of breast cancer with SNPs with MAF<0.01 and P-value<0.01 among European American population (NBHS)[a]**

| Gene | SNP (Alleles)[b] | Chr:Pos[c] | Annotation | Amino acid change (Polyphen-2 score/SIFT score/PROVEAN score) | No. of Samples (case/control)[d] | OR (95%CI) [e] | P-value[f] | P-value[g] | P-value[h] |
|---|---|---|---|---|---|---|---|---|---|
| *DFFA* | rs138842024 (G/A) | chr1:10529326 | Missense | Ile->Thr (1/0.121/-2.32) | 30/12 | 2.42 (1.32-4.46) | 0.008 | 0.008 | 0.012 |
| *CLCA2* | rs55736627 (T/C) | chr1:86894231 | Missense | Thr->Ile (0.003/0.250/-0.57) | 38/15 | 2.49 (1.45-4.30) | 0.002 | 0.002 | 0.004 |
| *MTMR11* | rs145659444 (T/C) | chr1:149902342 | Missense | Arg->His (0.999/0.003/-1.33) | 9/24 | 0.36 (0.18-0.72) | 0.007 | 0.008 | 0.012 |
| *CA14* | rs140320147 (T/G) | chr1:150230578 | Missense | Gly->Cys (0.998/0.002/-3.08) | 2/11 | 0.17 (0.06-0.50) | 0.009 | 0.011 | 0.025 |
| *IGFN1* | rs143014998 (A/G) | chr1:201186501 | Missense | Gly->Ser (1/0.001/-5.42) | 19/5 | 3.64 (1.63-8.14) | 0.006 | 0.007 | 0.013 |
| *LPCAT1* | rs144081179 (C/G) | chr5:1463932 | Missense | Ala->Gly (0.049/0.03/-3.07) | 1/10 | 0.09 (0.03-0.31) | 0.005 | 0.005 | 0.022 |
| *ZSCAN12* | rs2232432 (G/A) | chr6:28359073 | Missense | Cys->Arg (1/0/-10.79) | 20/6 | 3.22 (1.49-7.00) | 0.008 | 0.009 | 0.015 |
| *BRF2* | rs138763430 (T/C) | chr8:37707277 | Missense | Asp->Asn (0.026/0.029/-0.17) | 3/14 | 0.21 (0.08-0.54) | 0.007 | 0.007 | 0.017 |
| *ELK3* | rs118124881 (T/C) | chr12:96641121 | Missense | Thr->Met (0.988/0.012/-1.38) | 4/19 | 0.21 (0.09-0.47) | 0.002 | 0.001 | 0.005 |
| *PVR* | rs35959395 (C/G) | chr19:45153113 | Missense | Val->Leu (0.986/0.509/0.80) | 2/12 | 0.15 (0.05-0.44) | 0.005 | 0.006 | 0.016 |
| *BPIFA3* | rs142257117 (A/G) | chr20:31813013 | Missense | Asp->Asn (0/0.121/-2.32) | 1/10 | 0.10 (0.03-0.32) | 0.006 | 0.005 | 0.023 |
| *ASCC2* | rs1894473 (A/G) | chr22:30221201 | Missense | Arg->Cys (0.013/0.02/-4.47) | 5/19 | 0.25 (0.11-0.55) | 0.003 | 0.003 | 0.007 |

[a] All SNPs with MAF < 0.01. All SNPs with P-value < 0.01 in at least two of the test statistics.

[b] Effect allele/reference allele.

[c] Chromosome position (bp) based on NCBI Human Genome Build 37.

[d] Number of samples carrying heterozygous variant.

[e] OR (95% CI) was adjusted for first five principal components.

[f] P-value obtained from logistic regression analysis (score test).

[g] P-value obtained from fisher's exact test.

[h] P-value obtained from firth logistic regression analysis.

**Table 7. Associations of breast cancer with SNPs with MAF<0.01 and P-value<0.01 among African American population (NBHS/SCCS)[a]**

| Gene | SNP (Alleles)[b] | Chr:Pos[c] | Annotation | Amino acid change (Polyphen-2 score/SIFT score/PROVEAN score) | No. of Samples (case/control)[d] | OR (95%CI) [e] | P-value[f] | P-value[g] | P-value[h] |
|---|---|---|---|---|---|---|---|---|---|
| *TMIE, PRSS50* | rs146386127 (T/C) | chr3:46754539 | Missense | Arg->Gln (0.836/0.397/-0.88) | 11/1 | 11.54 (3.70-35.97) | 0.003 | 0.003 | 0.015 |
| *GINM1* | rs1137086 (G/A) | chr6:149903597 | Missense | Lys->Glu (0.228/0.4/-0.24) | 24/8 | 3.06 (1.51-6.21) | 0.005 | 0.004 | 0.008 |
| *INTS9* | rs141707027 (T/C) | chr8:28695158 | Missense | Gly->Ser (0.999/0.004/-5.15) | 3/14 | 0.22 (0.08-0.56) | 0.008 | 0.012 | 0.020 |
| *SFTPD* | rs150968324 (A/G) | chr10:81706274 | Missense | Pro->Ser (0.992/0.013/-3.82) | 9/1 | 9.43 (2.72-32.73) | 0.009 | 0.011 | 0.029 |
| *BLM* | rs142551229 (G/A) | chr15:91310209 | Missense | Lys->Glu (0.153/0.075/-3.13) | 7/22 | 0.32 (0.15-0.67) | 0.006 | 0.008 | 0.012 |
| *NSRP1* | rs117582579 (A/G) | chr17:28512405 | Missense | Asp->Asn (0.007/0.051/-0.98) | 2/13 | 0.15 (0.05-0.42) | 0.004 | 0.007 | 0.014 |
| *TBC1D16* | rs143618029 (A/G) | chr17:77984172 | Missense | Thr->Met (0.798/0.024/-1.86) | 1/10 | 0.10 (0.03-0.33) | 0.007 | 0.012 | 0.026 |

[a] All SNPs with MAF < 0.01. All SNPs with P-value < 0.01 in at least two of the test statistics.

[b] Effect allele/reference allele.

[c] Chromosome position (bp) based on NCBI Human Genome Build 37.

[d] Number of samples carrying heterozygous variant.

[e] OR (95% CI) was adjusted for first five principal components.

[f] P-value obtained from logistic regression analysis (score test).

[g] P-value obtained from fisher's exact test.

[h] P-value obtained from firth logistic regression analysis.

**Table 8. Associations of breast cancer with SNPs with MAF<0.01 and P-value<0.01 among European American population (BioVU)[a]**

| Gene | SNP (Alleles)[b] | Chr:Pos[c] | Annotation | Amino acid change (Polyphen-2 score/SIFT score/PROVEAN score) | No. of Samples (case/control)[d] | OR (95%CI) [e] | P-value[f] | P-value[g] | P-value[h] |
|---|---|---|---|---|---|---|---|---|---|
| *CASZ1* | rs143629495 (C/T) | chr1:10725188 | Missense | Gly->Arg (0.019/0.048/-1.13) | 3/1 | 16.57 (1.10-248.92) | 0.001 | 0.013 | 0.022 |
| *RSBN1* | rs41283514 (C/T) | chr1:114340502 | Missense | Arg->His (0.999/0.009/-1.74) | 8/11 | 4.05 (1.16-14.14) | 0.001 | 0.005 | 0.004 |
| *ADAM30* | rs147294252 (G/A) | chr1:120438577 | Missense | Thr->Ile (0.989/0.024/-4.84) | 12/25 | 2.60 (1.07-6.34) | 0.005 | 0.010 | 0.007 |
| *ITGA10* | rs35515885 (G/A) | chr1:145536012 | Missense | Ala->Thr (0.999/0.074/-1.48) | 10/17 | 3.01 (1.08-8.37) | 0.004 | 0.005 | 0.007 |
| *ANKRD35* | rs139709279 (C/T) | chr1:145561330 | Stop_gain | | 3/1 | 15.93 (1.10-230.40) | 0.001 | 0.013 | 0.024 |
| *LMOD1* | rs202184893 (G/T) | chr1:201868510 | Missense | Pro->His (1/0/-5.02) | 4/3 | 7.25 (0.94-55.86) | 0.002 | 0.014 | 0.013 |
| *NFASC* | rs139099286 (G/A) | chr1:204978777 | Missense | Val->Ile (0.944/0.048/-0.66) | 4/4 | 5.43 (0.80-36.70) | 0.007 | 0.024 | 0.021 |
| *CASP10* | rs143882052 (C/T) | chr2:202060670 | Missense | Pro->Leu (0.049/0.017/-3.38) | 3/2 | 8.73 (0.74-103.24) | 0.004 | 0.029 | 0.025 |
| *SMARCAL1* | rs190386780 (A/C) | chr2:217347476 | Missense | Lys->Gln (0.607/0.123/-0.86) | 4/3 | 6.61 (0.92-47.67) | 0.005 | 0.014 | 0.018 |
| *LTF* | rs61739313 (C/T) | chr3:46487937 | Missense | Val->Met (0.995/0.034/-0.78) | 2/64 | 0.16 (0.08-0.31) | 0.004 | 0.002 | 0.004 |
| *CEP44* | rs146429616 (G/A) | chr4:175224861 | Missense | Arg->His (0.76/0.002/-4.62) | 6/8 | 4.13 (0.97-17.64) | 0.005 | 0.013 | 0.010 |
| *SLC12A7* | rs141825245 (G/A) | chr5:1081844 | Missense | Thr->Met (0.008/0.137/-1.54) | 3/2 | 9.13 (0.74-112.44) | 0.003 | 0.029 | 0.022 |
| *SLC6A18* | rs200802505 (C/T) | chr5:1239577 | Stop_gain | | 3/2 | 8.69 (0.74-102.04) | 0.005 | 0.029 | 0.025 |
| *GPR98* | rs200541858 (A/G) | chr5:89923041 | Missense | Asp->Gly (1/0.006/-3.68) | 3/2 | 8.96 (0.74-108.85) | 0.004 | 0.029 | 0.023 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| GPR98 | rs200816323 (G/T) | chr5:90059182 | Missense | Val->Phe (0.934/0.015/-1.77) | 6/6 | 5.80 (1.18-28.53) | 0.0006 | 0.005 | 0.003 |
| GPR98 | rs201890097 (A/G) | chr5:90119357 | Missense | Thr->Ala (0.037/0.339/-1.72) | 5/3 | 9.04 (1.34-60.84) | 0.0003 | 0.003 | 0.004 |
| THEMIS | rs139859697 (A/T) | chr6:128134314 | Missense | Leu->His (1/0.001/-4.49) | 4/2 | 10.72 (1.20-95.95) | 0.0007 | 0.007 | 0.009 |
| AKAP12 | rs200662204 (G/A) | chr6:151672950 | Missense | Glu->Lys (0.778/0.064/-2.28) | 4/3 | 7.59 (0.96-60.23) | 0.002 | 0.014 | 0.012 |
| SYNE1 | rs138745849 (C/T) | chr6:152639250 | Missense | Ala->Val (0.003/0.506/-1.64) | 2/1 | 11.97 (0.48-300.81) | 0.010 | 0.065 | 0.045 |
| RAB11FIP1 | rs146427711 (G/A) | chr8:37732819 | Missense | Val->Met (0.01/0.017/-1.36) | 2/1 | 12.81 (0.47-346.90) | 0.007 | 0.065 | 0.054 |
| ANO1 | rs201870990 (G/A) | chr11:70007354 | Missense | Gln->Pro (0.993/0.046/-0.53) | 16/35 | 2.58 (1.20-5.55) | 0.001 | 0.005 | 0.002 |
| PRDM10 | rs141740226 (T/G) | chr11:129780436 | Missense | Phe->Ser (0.001/0/-6.75) | 8/14 | 3.04 (0.97-9.55) | 0.009 | 0.014 | 0.014 |
| ART4 | rs150640567 (A/G) | chr12:14994047 | Missense | Glu->Gly (1/0.062/-3.11) | 4/4 | 5.64 (0.82-39.09) | 0.006 | 0.024 | 0.018 |
| BRCA2 | rs56403624 (A/G) | chr13:32907000 | Missense | Ile->Thr (0.993/0.002/-1.89) | 2/1 | 12.15 (0.48-309.60) | 0.009 | 0.065 | 0.020 |
| ZFYVE26 | rs139163400 (A/G) | chr14:68229462 | Missense | Ile->Thr (1/0.002/-1.89) | 10/14 | 3.93 (1.30-11.91) | 0.0004 | 0.002 | 0.001 |
| IQGAP1 | rs147346534 (G/A) | chr15:90997745 | Missense | Val->Met (0.497/0.002/-2) | 2/1 | 12.62 (0.47-335.46) | 0.008 | 0.065 | 0.056 |
| BLM | rs149754073 (C/A) | chr15:91312417 | Missense | Leu->Ile (0.982/0.012/-1.80) | 3/1 | 17.94 (1.10-291.89) | 0.0006 | 0.013 | 0.019 |
| DYNLRB2 | rs149421698 (G/A) | chr16:80577179 | Missense | Val->Met (0.139/0.001/-2.78) | 2/1 | 12.78 (0.47-345.95) | 0.007 | 0.065 | 0.054 |
| EFCAB5 | rs185083328 (C/T) | chr17:28270598 | Missense | Pro->Ser (0.003/0/-0.21) | 5/2 | 13.97 (1.78-109.48) | 0.00004 | 0.001 | 0.003 |
| RNF213 | rs141301945 (C/T) | chr17:78355462 | Missense | Thr->Ile (0.994/0.005/-3.68) | 5/6 | 4.52 (0.89-23.05) | 0.007 | 0.018 | 0.015 |
| NWD1 | rs142852841 (C/G) | chr19:16918562 | Missense | Ala->Gly (0.958/0.611/-0.29) | 4/3 | 7.21 (0.94-55.31) | 0.003 | 0.014 | 0.013 |
| ARRDC2 | rs201831893 (G/A) | chr19:18121104 | Missense | Val->Met (1/0.44/-0.48) | 6/8 | 4.32 (0.99-18.85) | 0.003 | 0.013 | 0.007 |
| FKBP8 | rs113307565 (G/C) | chr19:18649227 | Missense | Pro->Ala (0.157/0.133/-4.54) | 7/9 | 4.42 (1.12-17.38) | 0.001 | 0.007 | 0.004 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *SLC25A42* | rs144256360 (C/T) | chr19:19218779 | Missense | Pro->Ser (0.999/0.009/-5.04) | 3/2 | 9.33 (0.75-115.87) | 0.003 | 0.029 | 0.022 |
| *ZNF235* | rs141976678 (A/G) | chr19:44793278 | Missense | Se->Pro (0.001/0.142/-1.47) | 6/6 | 5.52 (1.16-26.40) | 0.001 | 0.005 | 0.004 |
| *NRIP1* | rs61755059 (G/A) | chr21:16338139 | Missense | Ala->Val (0.031/0.027/-2.31) | 4/4 | 5.77 (0.82-40.54) | 0.005 | 0.024 | 0.017 |
| *NRIP1* | rs2228507 (T/A) | chr21:16339570 | Missense | Val->Phe (0/1/2.02) | 3/2 | 8.78 (0.74-103.99) | 0.004 | 0.029 | 0.025 |
| *USP25* | rs142929561 (G/C) | chr21:17236674 | Missense | Val->Leu (0.214/0.006/-2.04) | 7/8 | 4.70 (1.17-18.83) | 0.001 | 0.004 | 0.004 |

[a] All SNPs with MAF < 0.01. All SNPs with P-value < 0.01 in at least two of the test statistics.

[b] Effect allele/reference allele.

[c] Chromosome position (bp) based on NCBI Human Genome Build 37.

[d] Number of samples carrying heterozygous variant.

[e] OR (95% CI) was adjusted for first five principal components.

[f] P-value obtained from logistic regression analysis (score test).

[g] P-value obtained from fisher's exact test.

[h] P-value obtained from firth logistic regression analysis.

*Single-variant Meta-Analysis results*

For rare variants with MAF < 0.01 in the combined SBCGS, NBHS, SCCS and BioVU datasets, we found 7 missense variants that were associated with breast cancer risk at P-value < 0.01 (Table 9). Out of 7, 5 SNPs were not available in Asian (SBCGS) due to allele with zero MAF. A total of 3 missense variants were predicted to be "damaging" from at least one of the three functional prediction algorithms (rs145659444 (Arg->His) in the *MTMR11* gene, rs201870990 (Val->Met) in the *ANO1* gene, and rs139163400 (Ile->Thr) in the *ZFYVE26* gene). All of those 3 missense variants showed same association directions across all studies included in meta-analysis.

**Table 9. Meta-analysis result: Associations of breast cancer with SNPs with MAF<0.01 and meta P-value<0.01[a]**

| Gene | SNP (Alleles) Annotation[b] | Chr:Position[c] | Amino acid change (Polyphen-2 score/SIFT score/PROVEAN score) | Study | No. of Samples (case/control)[d] | OR (95%CI) [e] | P-value[f] | P-value[g] | P-value[h] |
|------|------|------|------|------|------|------|------|------|------|
| *DFFA* | rs138842024 (A/G) | chr1:10529326 | Ile->Thr | Asian | 3/0 | - | - | - | - |
| | Missense | | (0.999/0.121/-2.32) | European | 30/12 | 2.42 (1.32-4.46) | 0.008 | 0.008 | 0.012 |
| | | | | African American | 1/4 | 0.25 (0.04-1.47) | 0.186 | 0.375 | 0.294 |
| | | | | BioVU_European | 9/23 | 2.00 (0.78-5.12) | 0.075 | 0.080 | 0.076 |
| | | | | Meta-analysis | | 2.05 (1.24-3.38) | 0.037 | 0.023 | 0.033 |
| *ITGA10* | rs35515885 (G/A) | chr1:145536012 | Ala->Pro | European | 16/6 | 2.63 (1.14-6.25) | 0.036 | 0.052 | 0.052 |
| | Missense | | (0.999/0.074/-1.48) | African American | 12/13 | 1.07 (0.48-2.35) | 0.871 | 1.000 | 0.878 |
| | | | | BioVU_European | 10/17 | 3.01 (1.08-8.37) | 0.004 | 0.005 | 0.007 |
| | | | | Meta-analysis | | 2.04 (1.06-3.92) | 0.002 | 0.003 | 0.004 |
| *MTMR11* | rs145659444 (C/T) | chr1:149902342 | Arg->His | European | 9/24 | 0.36 (0.18-0.72) | 0.007 | 0.008 | 0.012 |
| | Missense | | (0.999/0.002/-1.33) | African American | 1/2 | 0.51 (0.05-5.00) | 0.574 | 1.000 | 0.679 |
| | | | | BioVU_European | 4/43 | 0.51 (0.23-1.15) | 0.196 | 0.226 | 0.248 |
| | | | | Meta-analysis | | 0.41 (0.21-0.77) | 0.008 | 0.021 | 0.017 |
| *LAPTM4A* | rs145912850 (C/A) | chr2:20234122 | Val->Leu | Asian | 41/23 | 1.82 (1.11-2.94) | 0.020 | 0.032 | 0.024 |
| | Missense | | (0.001/0.704/-0.14) | European | 12/10 | 1.16 (0.50-2.70) | 0.732 | 0.832 | 0.747 |
| | | | | African American | 3/3 | 1.03 (0.21-5.15) | 0.973 | 1.000 | 0.974 |
| | | | | BioVU_European | 11/30 | 2.08 (0.88-5.00) | 0.033 | 0.052 | 0.038 |
| | | | | Meta-analysis | | 1.71 (1.18-2.47) | 0.005 | 0.010 | 0.006 |
| *AKAP12* | rs142810400 (T/C) | chr6:151670656 | Val->Ala | European | 34/21 | 1.60 (0.94-2.73) | 0.091 | 0.105 | 0.119 |
| | Missense | | (0.013/0.092/-1.54) | African American | 5/1 | 4.92 (0.99-24.55) | 0.108 | 0.121 | 0.186 |
| | | | | BioVU_European | 17/63 | 1.50 (0.81-2.76) | 0.143 | 0.160 | 0.141 |
| | | | | Meta-analysis | | 1.60 (1.08-2.36) | 0.007 | 0.010 | 0.012 |
| *ANO1* | rs201870990 (G/A) | chr11:70007354 | Val->Met | European | 29/20 | 1.45 (0.82-2.56) | 0.210 | 0.253 | 0.225 |
| | Missense | | (0.972/0.017/-1.36) | African American | 5/2 | 2.50 (0.57-11.11) | 0.252 | 0.284 | 0.330 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | BioVU_European | 16/35 | 2.58 (1.2-5.55) | 0.001 | 0.005 | 0.002 |
| | | | | Meta-analysis | | 1.97 (1.20-3.24) | 0.001 | 0.002 | 0.001 |
| *ZFYVE26* | rs139163400 (A/G) | chr14:68229462 | Ile->Thr | European | 12/5 | 2.32 (0.89-6.03) | 0.105 | 0.143 | 0.136 |
| | Missense | | (1/0.002/-1.89) | African American | 2/1 | 1.96 (0.20-18.95) | 0.577 | 0.621 | 0.682 |
| | | | | BioVU_European | 10/14 | 3.93 (1.30-11.91) | 0.0004 | 0.002 | 0.001 |
| | | | | Meta-analysis | | 3.20 (1.68-6.09) | 0.0003 | 0.001 | 0.001 |

[a] All SNPs with MAF < 0.01. All SNPs with P-value < 0.01 in at least two of the test statistics. Analysis has been adjusted for index SNPs.

[b] Effect allele/reference allele.

[c] Chromosome position (bp) based on NCBI Human Genome Build 37.

[d] Number of samples carrying heterozygous variant.

[e] OR (95% CI) was adjusted for first five principal components.

[f] P-value in each study obtained from logistic regression analysis (score test). Meta-analysis p-value derived from a weighted z statistic-based meta-analysis.

[g] P-value in each study obtained from Fisher's exact test. Meta-analysis p-value derived from a P-value based combined method.

[h] P-value obtained from firth logistic regression analysis. Meta-analysis p-value derived from a weighted z statistic-based meta-analysis.

If the number of samples carrying heterozygous variant are zero, then they are indicated as '-'.

*Gene-Based Aggregation analysis results*

We conducted gene-based analysis on LOF and nonsynonymous variants separately based on the functional prediction in Sub-Aim 1. Rare variants with MAF ≤ 0.01 or MAF ≤ 0.005 within each gene were aggregated. At MAF ≤ 0.01, total 152 genes were tested for LOF, and 1,140 genes were tested for nonsynonymous variants. At MAF ≤ 0.005, total 147 genes were tested for LOF, and 1,133 genes were tested for nonsynonymous variants. Analysis has been adjusted for index SNPs. Results are shown in Table 10-1 ~ Table 17-2 (P-values were obtained from Madsen-Browning test (MB), Combined Multivariate and Collapsing (CMC), variable threshold (VT), and sequence kernel association test (SKAT)). Since MB test uses permutations, it is possible that permutation statistics are more extreme than the observed statistic, and thus the permuted P-value can be zero. For CMC and SKAT tests, we adjusted for the five first principal components (PCs) for all datasets (SBCGS, NBHS, SCCS, and BioVU).

For LOF variants, when collapsing variants with MAF ≤ 0.01 within each gene, *SYT8* gene was found to be associated with breast cancer risk at P-value < 0.01 in EA population (NBHS) (Table 12-1); *PSG5* gene was found to be associated with breast cancer risk at P-value < 0.01 in AA population (Table 14); and *SLC6A18* gene was found to be associated with breast cancer risk at P-value < 0.01 in EA population (BioVU) (Table 16-1). Similar results have been found when collapsing variants with MAF ≤ 0.005 within each gene except AA population (Table 10-2, Table 12-2 and Table 16-2). No genes were selected when collapsing variants with MAF ≤ 0.005 for AA population.

For nonsynonymous variants, when collapsing variants with MAF ≤ 0.01 within each gene, *ELK3* and *TRPS1* genes were found to be associated with breast cancer risk at P-value < 0.001 in EA population (NBHS) (Table 13-1); *CCDC38* genes was found to be associated with

breast cancer risk at P-value < 0.001 in AA population (Table 15-1); and *FKBP8*, *THEMIS*, and

*MUS81* genes were found to be associated with breast cancer risk at P-value < 0.001 in EA

population (BioVU) (Table 17-1). Similar results have been found when collapsing variants with

MAF ≤ 0.005 within each gene except AA population (Table 13-2, Table 15-2, and Table 17-2).

No genes were selected at a P-value < 0.001 in Asian population (Table 11-1 and Table 11-2).

Results from conditional analysis adjusted for index SNPs were consistent with the results

without adjustment for index SNPs.

**Table 10-1. LOF Variants: Gene-based analysis result among Asian population (MAF≤0.01)[a]**

| Gene | MAF≤0.01 | | | | |
|------|----------|-------|-------|-------|--------|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| SLC25A21 | 3 | 0.037 | 0.010 | 0.041 | 0.065 |
| IGF2 | 2 | 0.045 | 0.154 | 0.046 | 0.029 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 10-2. LOF Variants: Gene-based analysis result among Asian population (MAF≤0.005)[a]**

| Gene | MAF≤0.005 | | | | |
|------|-----------|-------|-------|-------|--------|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| IGF2 | 2 | 0.045 | 0.145 | 0.043 | 0.036 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 11-1. Nonsynonymous Variants: Gene-based analysis result among Asian population (MAF≤0.01)[a]**

| Gene | MAF≤0.01 | | | | |
|------|----------|-------|-------|-------|--------|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| TCF7L2 | 5 | 0.003 | 0.016 | 0.003 | 0.009 |
| UBR7 | 3 | 0.004 | 0.003 | 0.006 | 0.014 |
| CPA1 | 8 | 0.008 | 0.052 | 0.027 | 0.039 |
| IL12B | 5 | 0.009 | 0.023 | 0.004 | 0.008 |
| RBM4 | 5 | 0.027 | 0.003 | 0.015 | 0.175 |
| EXOC2 | 12 | 0.028 | 0.003 | 0.009 | 0.235 |
| RBM14 | 5 | 0.027 | 0.004 | 0.015 | 0.167 |
| TNS1 | 29 | 0.013 | 0.004 | 0.023 | 0.131 |
| ZNF225 | 7 | 0.048 | 0.005 | 0.018 | 0.377 |
| NBEAL2 | 33 | 0.048 | 0.006 | 0.058 | 0.613 |
| LGR6 | 13 | 0.013 | 0.006 | 0.029 | 0.043 |
| PLEK2 | 2 | 0.251 | 0 | 0.239 | 0.003 |
| RHBDD3 | 2 | 0.048 | 0.383 | 0.086 | 0.003 |
| FGF19 | 3 | 0.481 | 0.641 | 0.441 | 0.008 |
| DMRTA1 | 3 | 0.223 | 0.210 | 0.397 | 0.009 |
| MPP4 | 6 | 0.019 | 0.014 | 0.033 | 0.036 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 11-2. Nonsynonymous Variants: Gene-based analysis result among Asian population (MAF≤0.005)[a]**

| Gene | MAF ≤0.005 | | | | |
|------|-----------|------|------|------|------|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| TCF7L2 | 5 | 0.003 | 0.014 | 0.004 | 0.008 |
| UBR7 | 3 | 0.004 | 0.003 | 0.006 | 0.013 |
| CPA1 | 8 | 0.008 | 0.051 | 0.027 | 0.037 |
| IL12B | 5 | 0.009 | 0.023 | 0.005 | 0.009 |
| TNS1 | 27 | 0.010 | 0.009 | 0.062 | 0.067 |
| MUS81 | 4 | 0.010 | 0.008 | 0.020 | 0.055 |
| RBM4 | 5 | 0.027 | 0.003 | 0.015 | 0.173 |
| EXOC2 | 12 | 0.028 | 0.004 | 0.009 | 0.223 |
| RBM14 | 5 | 0.027 | 0.004 | 0.015 | 0.173 |
| GPAM | 5 | 0.023 | 0.006 | 0.026 | 0.223 |
| ZNF225 | 6 | 0.109 | 0.006 | 0.014 | 0.424 |
| NBEAL2 | 32 | 0.055 | 0.007 | 0.056 | 0.582 |
| COMMD3 | 2 | 0.036 | 0.008 | 0.060 | 0.102 |
| PLEK2 | 2 | 0.251 | 0 | 0.252 | 0.003 |
| RHBDD3 | 2 | 0.048 | 0.364 | 0.089 | 0.004 |
| DMRTA1 | 3 | 0.223 | 0.195 | 0.399 | 0.007 |
| FGF19 | 3 | 0.481 | 0.612 | 0.456 | 0.008 |
| MPP4 | 6 | 0.019 | 0.014 | 0.033 | 0.036 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 12-1. LOF Variants: Gene-based analysis result among European American population (NBHS) (MAF≤0.01)[a]**

| Gene | MAF ≤0.01 | | | | |
|------|-----------|------|------|------|------|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| SYT8 | 2 | 0.006 | 0.008 | 0.003 | 0.004 |
| SCTR | 2 | 0.050 | 0.090 | 0.089 | 0.197 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 12-2. LOF Variants: Gene-based analysis result among European American population (NBHS) (MAF≤0.005)[a]**

| Gene | MAF≤0.005 | | | | |
|:---:|:---:|:---:|:---:|:---:|:---:|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| SYT8 | 2 | 0.006 | 0.009 | 0.004 | 0.004 |
| SCTR | 2 | 0.050 | 0.093 | 0.092 | 0.187 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 13-1. Nonsynonymous Variants: Gene-based analysis result among European American population (NBHS) (MAF≤0.01)[a]**

| Gene | MAF≤0.01 | | | | |
|:---:|:---:|:---:|:---:|:---:|:---:|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| ELK3 | 4 | 0.0006 | 0.012 | 0.0002 | 0.001 |
| TMCC3 | 8 | 0.003 | 0.002 | 0.008 | 0.110 |
| RNF135 | 3 | 0.004 | 0.010 | 0.004 | 0.002 |
| CLCA2 | 9 | 0.004 | 0.021 | 0.012 | 0.003 |
| SGOL2 | 15 | 0.004 | 0.079 | 0.005 | 0.010 |
| CA14 | 2 | 0.005 | 0.004 | 0.004 | 0.004 |
| EXD2 | 4 | 0.005 | 0.003 | 0.005 | 0.061 |
| PKP1 | 10 | 0.006 | 0.010 | 0.016 | 0.140 |
| ZC3HC1 | 6 | 0.007 | 0.048 | 0.015 | 0.011 |
| BPIFB4 | 10 | 0.008 | 0.023 | 0.030 | 0.111 |
| CBFA2T2 | 3 | 0.008 | 0.003 | 0.007 | 0.022 |
| SLC25A42 | 2 | 0.021 | 0.005 | 0.025 | 0.013 |
| UNC13A | 6 | 0.017 | 0.006 | 0.026 | 0.046 |
| SCAP | 10 | 0.045 | 0.010 | 0.004 | 0.181 |
| TRPS1 | 6 | 0.047 | 0.237 | 0.045 | 0.0006 |
| KIAA0408 | 4 | 0.015 | 0.141 | 0.011 | 0.002 |
| CEP44 | 8 | 0.895 | 0.230 | 0.086 | 0.003 |
| LPCAT1 | 5 | 0.015 | 0.282 | 0.013 | 0.003 |
| CHID1 | 2 | 0.012 | 0.049 | 0.011 | 0.004 |
| RP11-15A1.2 | 3 | 0.539 | 0.762 | 0.551 | 0.004 |
| BRF2 | 4 | 0.279 | 0.822 | 0.479 | 0.004 |
| ZNF45 | 3 | 0.539 | 0.758 | 0.555 | 0.005 |
| MTMR11 | 9 | 0.089 | 0.972 | 0.261 | 0.006 |
| FES | 6 | 0.188 | 0.604 | 0.639 | 0.007 |
| UTP23 | 3 | 0.163 | 0.369 | 0.548 | 0.010 |
| PLEKHS1 | 5 | 0.013 | 0.026 | 0.019 | 0.024 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 13-2. Nonsynonymous Variants: Gene-based analysis result among European American population (NBHS) (MAF≤0.005)[a]**

| Gene | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
|------|-----------------|-------|------|------|--------|
| | | MAF ≤0.005 | | | |
| *ELK3* | 4 | 0.0006 | 0.009 | 0.0003 | 0.001 |
| *TMCC3* | 7 | 0.004 | 0.004 | 0.008 | 0.112 |
| *SGOL2* | 15 | 0.004 | 0.079 | 0.004 | 0.013 |
| *CA14* | 2 | 0.005 | 0.006 | 0.004 | 0.005 |
| *EXD2* | 4 | 0.005 | 0.004 | 0.005 | 0.057 |
| *ZC3HC1* | 6 | 0.007 | 0.051 | 0.012 | 0.013 |
| *BPIFB4* | 10 | 0.008 | 0.021 | 0.032 | 0.120 |
| *ITGA10* | 20 | 0.008 | 0.016 | 0.023 | 0.050 |
| *SOGA3* | 5 | 0.008 | 0.086 | 0.006 | 0.002 |
| *PKP1* | 9 | 0.009 | 0.022 | 0.023 | 0.108 |
| *SLC25A42* | 2 | 0.021 | 0.005 | 0.028 | 0.017 |
| *DMBT1* | 14 | 0.059 | 0.008 | 0.017 | 0.634 |
| *SCAP* | 10 | 0.045 | 0.010 | 0.002 | 0.178 |
| *CHID1* | 2 | 0.012 | 0.046 | 0.009 | 0.005 |
| *TRPS1* | 6 | 0.047 | 0.243 | 0.052 | 0.0009 |
| *ASCC2* | 3 | 0.023 | 0.260 | 0.027 | 0.002 |
| *TSSC4* | 5 | 0.516 | 0.488 | 0.378 | 0.002 |
| *LPCAT1* | 5 | 0.015 | 0.278 | 0.013 | 0.003 |
| *KIAA0408* | 4 | 0.015 | 0.145 | 0.014 | 0.003 |
| *CEP44* | 8 | 0.895 | 0.232 | 0.086 | 0.003 |
| *RP11-15A1.2* | 3 | 0.539 | 0.752 | 0.560 | 0.003 |
| *BRF2* | 4 | 0.279 | 0.812 | 0.473 | 0.004 |
| *ZNF45* | 3 | 0.539 | 0.738 | 0.567 | 0.004 |
| *FES* | 6 | 0.188 | 0.584 | 0.641 | 0.007 |
| *UTP23* | 3 | 0.163 | 0.368 | 0.535 | 0.008 |
| *IGFN1* | 24 | 0.078 | 0.296 | 0.135 | 0.009 |
| *PLEKHS1* | 5 | 0.013 | 0.028 | 0.019 | 0.024 |
| *CCRL2* | 4 | 0.017 | 0.017 | 0.059 | 0.053 |
| *ZC3H11A* | 5 | 0.021 | 0.019 | 0.027 | 0.060 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 14. LOF Variants: Gene-based analysis result among African American population (MAF≤0.01)[a]**

| Gene | MAF ≤0.01 | | | | |
|------|-----------|------|------|------|--------|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| PSG5 | 2 | 0.006 | 0.004 | 0.011 | 0.018 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 15-1. Nonsynonymous Variants: Gene-based analysis result among African American population (MAF≤0.01)[a]**

| Gene | MAF ≤0.01 | | | | |
|------|-----------|------|------|------|--------|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| CCDC38 | 11 | 0.0003 | 0.003 | 0.002 | 0.010 |
| ATAD5 | 8 | 0.001 | 0.008 | 0.002 | 0.030 |
| SNRPF | 9 | 0.001 | 0.005 | 0.006 | 0.017 |
| CCDC170 | 11 | 0.003 | 0.006 | 0.009 | 0.134 |
| ISYNA1 | 3 | 0.004 | 0.013 | 0.014 | 0.019 |
| ZNF404 | 3 | 0.005 | 0.003 | 0.006 | 0.031 |
| SIRT5 | 6 | 0.006 | 0.005 | 0.009 | 0.117 |
| CTD-2349P21.11 | 6 | 0.007 | 0.040 | 0.015 | 0.038 |
| FBXL7 | 3 | 0.007 | 0.004 | 0.004 | 0.015 |
| CDCA7 | 2 | 0.008 | 0.007 | 0.010 | 0.015 |
| GINM1 | 2 | 0.009 | 0.020 | 0.004 | 0.006 |
| SLC29A2 | 4 | 0.010 | 0.074 | 0.016 | 0.015 |
| RXFP2 | 9 | 0.061 | 0.006 | 0.006 | 0.492 |
| PRSS50 | 4 | 0.108 | 0.078 | 0.308 | 0.003 |
| MUC6 | 27 | 0.339 | 0.079 | 0.402 | 0.005 |
| MAGI3 | 8 | 0.014 | 0.031 | 0.046 | 0.026 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 15-2. Nonsynonymous Variants: Gene-based analysis result among African American population (MAF≤0.005)[a]**

| Gene | MAF ≤ 0.005 | | | | |
| --- | --- | --- | --- | --- | --- |
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| ATAD5 | 8 | 0.001 | 0.007 | 0.002 | 0.027 |
| RP3-508I15.14 | 3 | 0.002 | 0.009 | 0.009 | 0.017 |
| RIN1 | 8 | 0.003 | 0.040 | 0.007 | 0.036 |
| RXFP2 | 8 | 0.003 | 0.002 | 0.004 | 0.226 |
| ZNF404 | 3 | 0.005 | 0.004 | 0.005 | 0.033 |
| FYCO1 | 18 | 0.007 | 0.105 | 0.029 | 0.460 |
| CTD-2349P21.11 | 6 | 0.007 | 0.040 | 0.017 | 0.034 |
| FBXL7 | 3 | 0.007 | 0.004 | 0.004 | 0.015 |
| SIRT5 | 5 | 0.008 | 0.009 | 0.008 | 0.084 |
| CDCA7 | 2 | 0.008 | 0.009 | 0.013 | 0.016 |
| AP1B1 | 5 | 0.013 | 0.003 | 0.007 | 0.082 |
| MATN3 | 8 | 0.012 | 0.009 | 0.027 | 0.042 |
| PRSS50 | 4 | 0.108 | 0.077 | 0.295 | 0.003 |
| MBL1P | 3 | 0.224 | 0.086 | 0.437 | 0.004 |
| SFTPD | 3 | 0.224 | 0.091 | 0.439 | 0.004 |
| SIPA1 | 7 | 0.057 | 0.013 | 0.136 | 0.006 |
| TBC1D16 | 7 | 0.668 | 0.873 | 0.883 | 0.009 |
| CCDC170 | 10 | 0.011 | 0.014 | 0.024 | 0.155 |
| ZC3H11A | 5 | 0.071 | 0.056 | 0.149 | 0.021 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 16-1. LOF Variants: Gene-based analysis result among European American population (BioVU) (MAF≤0.01)[a]**

| Gene | MAF≤0.01 | | | | |
|---|---|---|---|---|---|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| *SLC6A18* | 2 | 0.002 | 0.014 | 0.029 | 0.006 |
| *ANKRD35* | 3 | 0.029 | 0.057 | 0.181 | 0.011 |
| *PSG9* | 2 | 0.063 | 0.063 | 0.034 | 0.064 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 16-2. LOF Variants: Gene-based analysis result among European American population (BioVU) (MAF≤0.005)[a]**

| Gene | MAF≤0.005 | | | | |
|---|---|---|---|---|---|
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| *SLC6A18* | 2 | 0.002 | 0.014 | 0.029 | 0.006 |
| *ANKRD35* | 3 | 0.029 | 0.057 | 0.181 | 0.011 |
| *PSG9* | 2 | 0.063 | 0.058 | 0.035 | 0.065 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 17-1. Nonsynonymous Variants: Gene-based analysis result among European American population (BioVU) (MAF≤0.01)[a]**

| Gene | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
|------|-----------------|-------|------|------|--------|
| | | | MAF ≤ 0.01 | | |
| FKBP8 | 2 | 0.0002 | 0.0004 | 0.0007 | 0.002 |
| THEMIS | 6 | 0.0006 | 0.004 | 0.003 | 0.004 |
| ANO1 | 6 | 0.001 | 0.006 | 0.019 | 0.002 |
| RSBN1 | 2 | 0.002 | 0.013 | 0.005 | 0.002 |
| GDF15 | 2 | 0.002 | 0.006 | 0.008 | 0.006 |
| MUS81 | 2 | 0.002 | 0.009 | 0.006 | 0.0009 |
| SCYL1 | 8 | 0.004 | 0.023 | 0.026 | 0.046 |
| SPDYC | 3 | 0.005 | 0.010 | 0.018 | 0.014 |
| GPR98 | 84 | 0.007 | 0.011 | 0.007 | 0.030 |
| SLC6A4 | 2 | 0.009 | 0.019 | 0.024 | 0.031 |
| CDC42BPG | 20 | 0.009 | 0.057 | 0.056 | 0.099 |
| MGP | 3 | 0.009 | 0.022 | 0.018 | 0.023 |
| PIGU | 5 | 0.039 | 0.002 | 0.003 | 0.184 |
| GPAM | 3 | 0.045 | 0.004 | 0.002 | 0.053 |
| ATP5O | 3 | 0.051 | 0.004 | 0.014 | 0.146 |
| ANP32E | 2 | 0.055 | 0.005 | 0.014 | 0.127 |
| RP11-867G23.12 | 4 | 0.010 | 0.007 | 0.020 | 0.050 |
| XRCC4 | 7 | 0.377 | 0.110 | 0.001 | 0.135 |
| ADAM29 | 4 | 0.429 | 0.026 | 0.002 | 0.043 |
| GAPT | 2 | 0.249 | 0.023 | 0.005 | 0.106 |
| SEMA4B | 13 | 0.291 | 0.030 | 0.005 | 0.490 |
| NRIP1 | 17 | 0.237 | 0.011 | 0.007 | 0.168 |
| MAN2A2 | 16 | 0.013 | 0.011 | 0.009 | 0.025 |
| LTF | 13 | 0.105 | 0.498 | 0.459 | 0.004 |
| MUC2 | 41 | 0.335 | 0.916 | 0.289 | 0.008 |
| SLC25A42 | 2 | 0.029 | 0.090 | 0.053 | 0.009 |
| ANKRD55 | 10 | 0.022 | 0.452 | 0.102 | 0.010 |
| AC098820.3 | 2 | 0.011 | 0.024 | 0.013 | 0.011 |
| RP11-307B6.3 | 3 | 0.026 | 0.081 | 0.181 | 0.011 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

**Table 17-2. Nonsynonymous Variants: Gene-based analysis result among European American population (BioVU) (MAF≤0.005)[a]**

| Gene | MAF ≤ 0.005 | | | | |
| :---: | :---: | :---: | :---: | :---: | :---: |
| | No. of variants | P_CMC | P_MB | P_VT | P_SKAT |
| *FKBP8* | 2 | 0.0002 | 0.0002 | 0.0009 | 0.001 |
| *THEMIS* | 6 | 0.0006 | 0.005 | 0.004 | 0.004 |
| *ATP5O* | 2 | 0.002 | 0.022 | 0.011 | 0.021 |
| *RSBN1* | 2 | 0.002 | 0.012 | 0.005 | 0.002 |
| *GDF15* | 2 | 0.002 | 0.005 | 0.008 | 0.008 |
| *MUS81* | 2 | 0.002 | 0.012 | 0.004 | 0.0006 |
| *SPDYC* | 3 | 0.005 | 0.010 | 0.019 | 0.016 |
| *EFCAB5* | 12 | 0.007 | 0.012 | 0.040 | 0.031 |
| *GPR98* | 77 | 0.008 | 0.016 | 0.021 | 0.006 |
| *SLC6A4* | 2 | 0.009 | 0.016 | 0.025 | 0.031 |
| *ANP32E* | 2 | 0.055 | 0.003 | 0.014 | 0.135 |
| *PIGU* | 4 | 0.014 | 0.005 | 0.003 | 0.053 |
| *GPAM* | 3 | 0.045 | 0.005 | 0.002 | 0.054 |
| *RP11-867G23.12* | 4 | 0.010 | 0.007 | 0.021 | 0.047 |
| *MAN2A2* | 16 | 0.013 | 0.010 | 0.007 | 0.027 |
| *USP25* | 7 | 0.031 | 0.010 | 0.032 | 0.037 |
| *XRCC4* | 6 | 0.948 | 0.037 | 0.002 | 0.064 |
| *ADAM29* | 4 | 0.429 | 0.027 | 0.002 | 0.045 |
| *NRIP1* | 16 | 0.136 | 0.015 | 0.005 | 0.051 |
| *SEMA4B* | 13 | 0.291 | 0.031 | 0.007 | 0.473 |
| *SLC25A42* | 2 | 0.029 | 0.096 | 0.052 | 0.007 |

[a] Gene-based analysis P-value < 0.05 in at least one of the test statistics.

*Gene-Based Meta-Analysis results*

We also conducted gene-based meta-analysis on LOF and nonsynonymous variants separately at two MAF thresholds (MAF $\leq$ 0.01 or MAF $\leq$ 0.005) in the combined SBCGS, NBHS, SCCS and BioVU datasets. Results are shown in Table 18-1 ~ Table 19-2 (P-values were obtained from MB, CMC, VT, and SKAT). For CMC and SKAT tests, we adjusted for the five first PCs for all datasets (SBCGS, NBHS, SCCS, BioVU).

For LOF variants, collapsing variants with MAF $\leq$ 0.01 within each gene suggested two genes (*PSG5* and *ANKRD35*) associated with breast cancer at P < 0.01 (Table 18-1). At the 19q13.2, the *PSG5* (consisting of 2 variants with MAF $\leq$ 0.01) was significantly associated with breast cancer risk (P-values = $7.0\times10^{-3}$ and $6.0\times10^{-3}$) from CMC and MB tests. At the 1q21.1, the *ANKRD35* (consisting of 5 variants with MAF $\leq$ 0.01) was significantly associated with breast cancer risk (P-value = $5.0\times10^{-3}$) from SKAT test. This gene was also significantly associated with breast cancer risk from VT test (P-value = $7.0\times10^{-3}$) based on 3 variants (MAF $\leq$ 0.01). These associations did not change when collapsing variants with MAF $\leq$ 0.005 within each gene (Table 18-2).

For nonsynonymous variants, collapsing variants with MAF $\leq$ 0.01 within each gene suggested two genes (*FKBP8* and *ANO1*) associated with breast cancer at P < 0.001 (Table 19-1). At the 19p12, the *FKBP8* (consisting of 2 variants with MAF $\leq$ 0.01) was significantly associated with breast cancer risk (P-values = $3.0\times10^{-4}$, $3.0\times10^{-4}$, and $5.0\times10^{-4}$) from CMC, MB, and VT tests. At the 11q13.3, the *ANO1* (consisting of 14 variants with MAF $\leq$ 0.01) was significantly associated with breast cancer risk (P-values = $4.0\times10^{-4}$, $9.0\times10^{-4}$, and $6.0\times10^{-4}$) from CMC, MB, and SKAT tests. In addition to these two genes, the *PLEKHS1* was significantly

associated with breast cancer risk (P-value = $3.0\times10^{-4}$ from CMC test) at the 10q25.3 when

collapsing variants with MAF ≤ 0.005 (consisting of 9 variants) (Table 19-2).

**Table 18-1. LOF Variants: Gene-based Meta-analysis result from all four datasets (MAF≤0.01)[a]**

| Gene | MAF≤0.01 | | | | | |
|------|-----------------|-------|------|--------|-----------------|------|
| | No. of variants | P_CMC | P_MB | P_SKAT | No. of variants | P_VT |
| PSG5 | 2 | 0.007 | 0.006 | 0.031 | 2 | 0.013 |
| OR2J2 | 2 | 0.012 | 0.011 | 0.037 | 2 | 0.022 |
| SLC6A18 | 3 | 0.033 | 0.033 | 0.045 | 3 | 0.076 |
| CCR5 | 2 | 0.053 | 0.038 | 0.108 | 2 | 0.100 |
| ALS2CR11 | 2 | 0.056 | 0.049 | 0.085 | 2 | 0.107 |
| ANKRD35 | 5 | 0.237 | 0.190 | 0.005 | 3 | 0.007 |

[a] Gene-based Meta-analysis P-value < 0.05 in at least one of the test statistics.

**Table 18-2. LOF Variants: Gene-based Meta-analysis result from all four datasets (MAF≤0.005)[a]**

| Gene | MAF≤0.005 | | | | | |
|------|-----------------|-------|------|--------|-----------------|------|
| | No. of variants | P_CMC | P_MB | P_SKAT | No. of variants | P_VT |
| PSG5 | 2 | 0.007 | 0.006 | 0.031 | 2 | 0.013 |
| OR2J2 | 2 | 0.012 | 0.011 | 0.037 | 2 | 0.022 |
| SLC6A18 | 3 | 0.033 | 0.033 | 0.045 | 3 | 0.076 |
| CCR5 | 2 | 0.053 | 0.038 | 0.108 | 2 | 0.100 |
| ALS2CR11 | 2 | 0.056 | 0.049 | 0.085 | 2 | 0.107 |
| ANKRD35 | 5 | 0.237 | 0.190 | 0.005 | 3 | 0.008 |

[a] Gene-based Meta-analysis P-value < 0.05 in at least one of the test statistics.

**Table 19-1. Nonsynonymous Variants: Gene-based Meta-analysis result from all four datasets (MAF≤0.01)[a]**

| Gene | MAF ≤ 0.01 | | | | | |
|---|---|---|---|---|---|---|
| | No. of variants | P_CMC | P_MB | P_SKAT | No. of variants | P_VT |
| FKBP8 | 2 | 0.0003 | 0.0003 | 0.0011 | 2 | 0.0005 |
| ANO1 | 14 | 0.0004 | 0.0009 | 0.0006 | 14 | 0.0038 |
| PLEK2 | 5 | 0.0011 | 0.1526 | 0.0010 | 5 | 0.0053 |
| SCYL1 | 16 | 0.0019 | 0.0113 | 0.1423 | 16 | 0.0165 |
| CPA1 | 6 | 0.0028 | 0.0045 | 0.0448 | 6 | 0.0127 |
| UBR7 | 3 | 0.0032 | 0.0108 | 0.0133 | 3 | 0.0087 |
| LAPTM4A | 2 | 0.0033 | 0.0143 | 0.0045 | 2 | 0.0066 |
| NKD2 | 17 | 0.0069 | 0.0940 | 0.0431 | 17 | 0.0598 |
| AKR1C2 | 4 | 0.0079 | 0.0101 | 0.0073 | 4 | 0.0263 |
| ALS2CL | 11 | 0.0097 | 0.8760 | 0.0053 | 11 | 0.0699 |
| SUN2 | 27 | 0.0099 | 0.0284 | 0.1976 | 24 | 0.0300 |
| CARD14 | 11 | 0.0168 | 0.0028 | 0.1393 | 7 | 0.0704 |
| ZSCAN12 | 4 | 0.0101 | 0.0035 | 0.0608 | 4 | 0.0368 |
| LIPI | 7 | 0.0387 | 0.0084 | 0.1822 | 3 | 0.0669 |
| PLEKHS1 | 11 | 0.0127 | 0.0095 | 0.0137 | 9 | 0.0024 |
| CAPN1 | 3 | 0.4175 | 0.0119 | 0.2207 | 2 | 0.0020 |
| GLT25D1 | 16 | 0.2271 | 0.0243 | 0.6331 | 11 | 0.0038 |
| NT5C1B/NT5C1B-RDH14 | 18 | 0.6124 | 0.1473 | 0.2859 | 5 | 0.0045 |
| SLC25A42 | 4 | 0.0323 | 0.0192 | 0.0069 | 3 | 0.0051 |
| PIGU | 9 | 0.3347 | 0.0309 | 0.5441 | 4 | 0.0051 |
| ZFYVE26 | 13 | 0.0274 | 0.0278 | 0.0560 | 11 | 0.0083 |
| CCDC38 | 3 | 0.4579 | 0.1128 | 0.0441 | 2 | 0.0089 |
| SIPA1 | 16 | 0.6173 | 0.0341 | 0.5455 | 11 | 0.0098 |
| FTO | 9 | 0.0847 | 0.0135 | 0.6913 | 5 | 0.0099 |
| PLEKHH1 | 27 | 0.0614 | 0.4895 | 0.0020 | 27 | 0.4121 |
| RSBN1 | 4 | 0.0121 | 0.1122 | 0.0034 | 4 | 0.0429 |
| BRCA2 | 89 | 0.0459 | 0.2049 | 0.0087 | 22 | 0.3392 |

[a] Gene-based Meta-analysis P-value < 0.01 in at least one of the test statistics.

**Table 19-2. Nonsynonymous Variants: Gene-based Meta-analysis result from all four datasets (MAF≤0.005)**[a]

| Gene | MAF ≤0.005 | | | | | |
|---|---|---|---|---|---|---|
| | No. of variants | P_CMC | P_MB | P_SKAT | No. of variants | P_VT |
| FKBP8 | 2 | 0.0003 | 0.0003 | 0.0011 | 2 | 0.0005 |
| PLEKHS1 | 9 | 0.0003 | 0.0058 | 0.0025 | 9 | 0.0019 |
| ANO1 | 14 | 0.0004 | 0.0009 | 0.0006 | 14 | 0.0038 |
| SCYL1 | 16 | 0.0019 | 0.0113 | 0.1423 | 16 | 0.0166 |
| CPA1 | 6 | 0.0028 | 0.0045 | 0.0448 | 6 | 0.0132 |
| UBR7 | 3 | 0.0032 | 0.0108 | 0.0133 | 3 | 0.0089 |
| LAPTM4A | 2 | 0.0033 | 0.0143 | 0.0045 | 2 | 0.0066 |
| RUNX1 | 2 | 0.0045 | 0.0046 | 0.0180 | 2 | 0.0083 |
| ZFYVE26 | 12 | 0.0059 | 0.0272 | 0.0021 | 11 | 0.0074 |
| NKD2 | 17 | 0.0069 | 0.0940 | 0.0431 | 17 | 0.0596 |
| AKR1C2 | 4 | 0.0079 | 0.0101 | 0.0073 | 4 | 0.0268 |
| SUN2 | 27 | 0.0099 | 0.0284 | 0.1976 | 24 | 0.0291 |
| ZSCAN12 | 4 | 0.0101 | 0.0035 | 0.0608 | 4 | 0.0366 |
| LIPI | 7 | 0.0387 | 0.0084 | 0.1822 | 3 | 0.0670 |
| BARX2 | 4 | 0.0140 | 0.0095 | 0.1028 | 4 | 0.0466 |
| CARD14 | 10 | 0.0763 | 0.0095 | 0.4001 | 7 | 0.0630 |
| CAPN1 | 3 | 0.4175 | 0.0119 | 0.2207 | 2 | 0.0020 |
| GLT25D1 | 16 | 0.2271 | 0.0243 | 0.6331 | 11 | 0.0038 |
| NT5C1B\|NT5C1B-RDH14 | 18 | 0.6124 | 0.1473 | 0.2859 | 5 | 0.0044 |
| PIGU | 9 | 0.3347 | 0.0309 | 0.5441 | 4 | 0.0049 |
| SLC25A42 | 4 | 0.0323 | 0.0192 | 0.0069 | 3 | 0.0057 |
| CCDC38 | 3 | 0.4579 | 0.1128 | 0.0441 | 2 | 0.0084 |
| SIPA1 | 14 | 0.3726 | 0.0210 | 0.2487 | 11 | 0.0085 |
| FTO | 9 | 0.0847 | 0.0135 | 0.6913 | 5 | 0.0098 |
| RSBN1 | 4 | 0.0121 | 0.1122 | 0.0034 | 4 | 0.0425 |
| ANKRD55 | 15 | 0.0598 | 0.1657 | 0.0096 | 15 | 0.3578 |

[a] Gene-based Meta-analysis P-value < 0.01 in at least one of the test statistics.

*Compound Heterozygous (CH) analysis results*

We performed compound heterozygous analysis on LOF and nonsynonymous variants separately at two MAF thresholds (MAF ≤ 0.01 or MAF ≤ 0.005). Results are shown in Table 20 ~ Table 22. For LOF variant, no significant results have been found in any of the four datasets (SBCGS, NBHS, SCCS and BioVU). For nonsynonymous variants, when collapsing variants with MAF ≤ 0.01 within each gene, *LOC100294362* gene from Asian population, *AKAP12* and *GPR98* genes from AA population, and *BRCA2* and *GPR98* genes from EA population (BioVU) were found to be associated with breast cancer risk at P-value < 0.05 (Table 20, Table 21, and Table 22). At the 17q25.3, the *LOC100294362* (consisting of 28 variants with MAF ≤ 0.01) was significantly associated with breast cancer risk (P-value = $3.9 \times 10^{-2}$) with 8 CH in breast cancer cases and 1 CH in control from Asian population at P-value < 0.05. At the 6q25.1, the *AKAP12* (consisting of 31 variants with MAF ≤ 0.01) was significantly associated with breast cancer risk (P-value = $2.0 \times 10^{-2}$) with 8 CH in breast cancer cases and 1 CH in control from AA population at P-value < 0.05. At the 5q14.3, the *GPR98* was significantly associated with breast cancer risk (at P-value < 0.05) in both AA (6 CH in cases/17 CH in controls) and EA (BioVU) (14 CH in cases/39 CH in controls) populations with P-values of $3.4 \times 10^{-2}$ and $3.5 \times 10^{-2}$, respectively. One of the most well-known breast cancer related genes, *BRCA2* gene at the 13q12.3, showed significant association with breast cancer risk (P-value = $1.1 \times 10^{-2}$) in EA population (BioVU) (5 CH in cases/5 CH in controls) at P-value < 0.05. No significant results have been found when collapsing variants with MAF ≤ 0.005 within each gene, and from combined analysis of all four datasets (SBCGS, NBHS, SCCS and BioVU).

**Table 20. Nonsynonymous Variants: CH-analysis result among Asian population[a]**

| Gene | MAF ≤ 0.01 | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| | No. of variants | No. of CH (case/control) | OR[b] | Lower 95% CI | Upper 95% CI | P-value[b] |
| *LOC100294362* | 28 | 8/1 | 0.126 | 0.003 | 0.942 | 0.039 |

[a] Genes with CH-analysis P-value < 0.05.

[b] Obtained from Fisher's exact test.

**Table 21. Nonsynonymous Variants: CH-analysis result among African American population[a]**

| Gene | MAF ≤ 0.01 | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| | No. of variants | No. of CH (case/control) | OR[b] | Lower 95% CI | Upper 95% CI | P-value[b] |
| *AKAP12* | 31 | 8/1 | 0.122 | 0.003 | 0.915 | 0.020 |
| *GPR98* | 70 | 6/17 | 2.818 | 1.055 | 8.765 | 0.034 |

[a] Genes with CH-analysis P-value < 0.05.

[b] Obtained from Fisher's exact test.

**Table 22. Nonsynonymous Variants: CH-analysis result among European American population (BioVU)[a]**

| Gene | MAF ≤ 0.01 | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| | No. of variants | No. of CH (case/control) | OR[b] | Lower 95% CI | Upper 95% CI | P-value[b] |
| *BRCA2* | 42 | 5/5 | 0.183 | 0.042 | 0.796 | 0.011 |
| *GPR98* | 84 | 14/39 | 0.507 | 0.267 | 1.016 | 0.035 |

[a] Genes with CH-analysis P-value < 0.05.

[b] Obtained from Fisher's exact test.

## B2.    Discussion

We investigated all LOF and nonsynonymous variants located in flanking 1Mb of all index SNPs in 109 GWAS loci using two MAF thresholds (MAF ≤ 0.01 or MAF ≤ 0.005). From single-variant analysis, we identified several novel missense variants predicted to be damaging that were associated with breast cancer risk at P-value < 0.01: 4 from Asian; 10 from EA (NBHS), 5 from AA, and 26 from EA (BioVU). When combined all datasets together, we identified total 3 novel missense variants that were predicted to be damaging; rs145659444 (Arg->His) in the *MTMR11* gene, rs201870990 (Val->Met) in the *ANO1* gene, and rs139163400 (Ile->Thr) in the *ZFYVE26* gene.

Interestingly, 3 genes (*LAPTM4A*, *ANO1*, and *ZFYVE26*) were identified from both gene-based meta-analysis and single-variant meta-analysis (MAF < 0.01 and meta P-value < 0.01).

At the 2p24.1, the *LAPTM4A* (consisting of 2 variants with MAF ≤ 0.01) was significantly associated with breast cancer risk from CMC, VT, and SKAT tests (P-values = $3.3 \times 10^{-3}$, $6.6 \times 10^{-3}$, and $4.5 \times 10^{-3}$, respectively). Li *et al*. reported *LAPTM4B* (lysosomal protein transmembrane 4 beta) as a novel cancer-associated gene including breast cancer (136). They revealed a new role of *LAPTM4B-35* in promoting multidrug resistance of cancer cells (136). It has been known that the putative protein of *LAPTM4B* is highly conserved, with 46% homologous at the amino-acid level to a *LAPTM4A* gene (human lysosome-associated transmembrane-4 protein) (137). Also, *LAPTM4A* gene is thought to function as a transporter protein that transfers nucleosides (and/or nucleoside metabolites) between the cytosol and intracellular organelles (138). Therefore, although the underlying biological mechanism is still not known, it might be possible that *LAPTM4B-35* functions as multidrug transporter through the

help of *LAPTM4A* (136, 138). Here we provide evidence that rare variants in the *LAPTM4A* gene may also contribute to the risk of breast cancer.

At the 11q13.3, the *ANO1* (consisting of 14 variants with MAF ≤ 0.01) was significantly associated with breast cancer risk from CMC, MB, and SKAT tests (P-values = $4.0 \times 10^{-4}$, $9.0 \times 10^{-4}$, $6.0 \times 10^{-4}$, respectively). Britschgi *et al*. have found that the *ANO1* gene is amplified in breast cancer, and amplification of the *ANO1* gene is associated with poor prognosis of breast cancer patients (139). Subsequently, Wu *et al*. revealed that *ANO1* overexpression was associated with good prognosis in PR-positive, or HER2-negative patients following tamoxifen treatment (140). The gene-based results in our study provide further evidence that the *ANO1* gene is significantly associated with breast cancer risk.

At the 14q24.1, the *ZFYVE26* (consisting of 11 variants with MAF ≤ 0.01) was significantly associated with breast cancer risk from VT test (P-values = $8.3 \times 10^{-3}$). Recently, Sagona *et al*. found that both *Beclin 1* (Autophagy Related) and *ZFYVE26* (Zinc Finger, FYVE Domain Containing 26) were down-regulated in advanced breast cancers (141). Their findings suggest a novel potential tumor suppressor mechanism for *Beclin 1* through a positive feedback loop for recruitment of *ZFYVE26* and *Beclin 1* to the intercellular bridge during cytokinesis (141). Here, we provide evidence that rare variants in the *ZFYVE26* gene may contribute to the risk of breast cancer.

Our CH analysis identified *LOC100294362* at the 17q25.3 from Asian population, *AKAP12* (6q25.1) and *GPR98* (5q14.3) from AA population, and *BRCA2* (13q12.3) and *GPR98* (5q14.3) from EA population (BioVU) (MAF ≤ 0.01 and P-value < 0.05). The less significant P-values for CH models are due to the rarity of compound heterozygotes of rare variants. Among our findings, *GPR98* gene was identified from both AA (6 CH in cases/17 CH in controls) and

EA (BioVU) (14 CH in cases/39 CH in controls) populations. In *GPR98* gene, two missense

variants, rs200541858 (Asp->Gly) and rs200816323 (Val->Phe), were significantly associated

with breast cancer risk (P-values = $4.0 \times 10^{-3}$ and $6.0 \times 10^{-4}$) in EA (BioVU) population. Both

variants were predicted to be "damaging". *GPR98* mutations are known to cause familial febrile

seizures and one form of Usher syndrome, which is the most common genetic cause of combined

blindness and deafness (142). Hilgert *et al*. found a large *GPR98* deletion of 136,017 bp

segregates with *USH2C* in an Iranian family (143). The function of *GPR98* is still unknown.

Also, missense variant rs142810400 in *AKAP12* was significantly associated with breast cancer

risk from single-marker meta-analysis (meta P-value = $7.0 \times 10^{-3}$). In humans, *AKAP12* maps to

6q25.1, a deletion hotspot in advanced breast cancer, implicating a role for the loss of *AKAP12*

in cancer progression (144). The CH results in current study provide further evidence that the

*GPR98* and *AKAP12* genes are significantly associated with breast cancer risk.

Importantly, *SLC25A42* gene at the 19p13.1, one of the top genes from eQTL result using

TCGA (P-value = 0.04), was significantly associated with breast cancer risk from SKAT and VT

tests (P-values = $6.9 \times 10^{-3}$ and $5.7 \times 10^{-3}$, respectively). The *SLC25A42* gene is located 603 kb

downstream of GWAS SNP rs4808801. We found that risk allele of SNP rs4808801 was

associated with lower gene expression (P-value = 0.04) which indicates increased risk of breast

cancer. Total 3 out of 4 rare variants (MAF $\leq$ 0.005) in this gene were predicted as "damaging".

Therefore, although the function of *SLC25A42* remains unexplored, this gene would be important

to further investigate in larger populations.

Our findings from CH analysis can be used as valuable genetic information since CH is

hard to be found in a large, randomly mating population due to its' recessive manner. This is the

first study to specifically examine the associations between rare recessive variants and breast

cancer risk using CH analysis. Most CH studies had been conducted in Mendelian-disease genes in family-based sequencing studies. We used large dataset in order to investigate rare coding variants and breast cancer risk through systematic analyses of *cis*-eQTLs, functional predictions, and comprehensive association tests. Therefore, our study provides additional insights into the genetics and biology of breast cancer.

# CHAPTER VII

# SYNOPSIS AND FUTURE DIRECTIONS

## A.      Conclusions

As the first study to examine rare coding variants associated with breast cancer risk using CH analysis, three major databases for *cis*-eQTLs, and the largest breast cancer datasets for Asian population, our results identified multiple rare coding variants associated with breast cancer in GWAS identified loci. We found 3 novel missense variants that were predicted to be "damaging" in the combined data; rs145659444 (Arg->His) in the *MTMR11* gene, rs201870990 (Val->Met) in the *ANO1* gene, and rs139163400 (Ile->Thr) in the *ZFYVE26* gene. Especially, three genes at 2p24.1 (*LAPTM4A*), 11q13.3 (*ANO1*), and 14q24.1 (*ZFYVE26*) from single-variant meta-analysis were consistently found to be significantly associated with breast cancer risk in gene-based meta-analysis. Importantly, we found that *SLC25A42* gene, one of the top genes from eQTL result, was significantly associated with breast cancer risk from gene-based meta-analysis with evidence of GWAS SNP rs4808801 and 3 rare variants (predicted as damaging). Based on previous genetic studies, we provided evidence that rare variants in these genes may also contribute to the risk of breast cancer (136, 140, 141).

Results from CH analysis indicated an important role of *GPR98* and *AKAP12* genes in breast cancer, and missense variants included in these genes were significantly associated with breast cancer risk from our single-variant analyses. For CH, the most significant gene was *BRCA2* in EA (BioVU) population with a P-value of 0.011. We expected to observe less significant P-values for CH models since there are a few compound heterozygotes for rare

variants (105, 145). Due to the study design limitation (no family-based sequencing studies) and the rarity of compound heterozygotes of rare variants, our CH findings from the large population-based studies are more valuable in breast cancer genetics. For example, in the case of a family where some members express a rare phenotype trait, we could easily examine the family's genomic pedigree for rare alleles being inherited in a recessive manner in accordance with the trait. And this is not the case for a large, randomly mating population-based studies. As we might expect, two copies of the same rare allele are unlikely to be inherited together in a large, randomly mating population. Due to the rarity of CH findings, especially in a large population-based study, our CH results provide new genetic clues for breast cancer risk inherited in a recessive manner, and further experimental validation are warranted.

In conclusion, results from our study provide additional insights into the genetics and biology of breast cancer. Further studies are required to explain the underlying biological mechanisms of our findings.

## B.    Considerations

Studies previously evaluating rare coding variants associated with breast cancer have been limited to their sample sizes. We used large sample size to examine rare coding variants, and our functional prediction approaches provided us meaningful candidate nonsynonymous and LOF variants which we used for further analysis. For gene-based meta-analysis, RAREMETAL might not be an optimal method for binary model, but it is close to optimal. There would be little power loss. The developers of this method compared single variants test using binary and quantitative models, and they found that the results were almost similar. Therefore, our gene-based meta-analysis results are reliable based on gene-level test statistics implemented in

RAREMETAL which are reconstructed from single variant score statistics and their covariance matrix (120).

Our *cis*-eQTL analysis to select genes in 109 GWAS loci associated with breast cancer risk used three major databases with the most updated genotype and gene expression information. We are not aware of any study that evaluates eQTLs using all three major databases. We also examined the potential differences between adjusted and unadjusted CNV and DNA methylation for TCGA in eQTL studies. We found potential minimal confounding effect of CNV and DNA methylation on gene expression. Thus, we might have a reduced power to conduct eQTL analysis using METABRIC since they do not provide CNV and DNA methylation information. Although we would not have an inflated type I error since those are not strong confounders, further consideration need to be taken in eQTL studies.

## C.   Future directions

In this study, we found several novel missense variants and significant genes associated with breast cancer risk from comprehensive association analyses. If a rare variant is predicted to have a functional effect according to several functional prediction algorithms, further biological validation is required to prove any suspected functional effect. Specifically, LOF variants are expected to be found at lower frequencies in the genome due to evolutionary pressure which results in an enrichment for false positives among such variants (93, 135). Therefore, proper biological validation of these variants is especially important.

Although classic dominant inheritance model is still useful for rare variant evaluation, recessive patterns of compound heterozygotes of rare variants can also expose the function altering effects of rare variants. For validation of our CH findings and explanation of the functional effect of rare variants, family-based sequencing studies will serve as valuable

resources. Therefore, further studies including experimental validations are necessary to explain our findings in the genetics of breast cancer.

**Appendix 1. Previously identified GWAS loci associated with Breast Cancer risk.**

| SNP | Chr | Position (hg19) | Alleles[a] | Reported Gene | OR (95% CI) | P-value[b] | Ethnicity | Study (reference) |
|---|---|---|---|---|---|---|---|---|
| rs616488 | 1 | 10566215 | G/A | PEX14 | 0.94 (0.92-0.96) | $2.0 \times 10^{-10}$ | European | Michailidou et al. (11) |
| rs12118297 | 1 | 87779217 | T/G | LMO4 | 0.91 (0.88-0.94) | $4.5 \times 10^{-8}$ | East Asian | Han et al. (Under review) |
| rs11552449 | 1 | 114448389 | T/C | PTPN22-BCL2L15-AP4B1-DCLRE1B-HIPK1 | 1.07 (1.04-1.09) | $1.8 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs11249433 | 1 | 121280613 | G/A | None | 1.09 (1.07-1.11) | $2.0 \times 10^{-26}$ | European | Michailidou et al. (11) |
| rs12405132 | 1 | 145644984 | T/C | LOC10028814,NBPF10,RNF115 | 0.95 (0.93-0.97) | $7.9 \times 10^{-9}$ | European | Michailidou et al. (13) |
| rs12048493 | 1 | 149927034 | C/A | None | 1.07 (1.05-1.10) | $1.1 \times 10^{-9}$ | European | Michailidou et al. (13) |
| rs6678914 | 1 | 202187176 | ?/? | LGR6,UBE2T,PTPN7 | 1.10 (1.06-1.13) | $1.4 \times 10^{-8}$ | European | Garcia-Closas et al. (7) |
| rs4951011 | 1 | 203766331 | G/A | ZC3H11A | 1.09 (1.06-1.12) | $8.8 \times 10^{-9}$ | East Asian | Cai et al. (6) |
| rs4245739 | 1 | 204518842 | ?/? | LRRN2,PIK3C2B,MDM4 | 1.14 (1.10-1.18) | $2.1 \times 10^{-12}$ | European | Garcia-Closas et al. (7) |
| rs72755295 | 1 | 242034263 | G/A | EXO1 | 1.15 (1.09-1.22) | $1.8 \times 10^{-8}$ | European | Michailidou et al. (13) |
| rs12710696 | 2 | 19320803 | ?/? | None | 1.10 (1.06-1.13) | $4.6 \times 10^{-8}$ | European | Garcia-Closas et al. (7) |
| rs4849887 | 2 | 121245122 | T/C | None | 0.91 (0.88-0.94) | $3.7 \times 10^{-11}$ | European | Michailidou et al. (11) |
| rs2016394 | 2 | 172972971 | A/G | METAP1D-DLX1-DLX2 | 0.95 (0.93-0.97) | $1.2 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs1550623 | 2 | 174212894 | G/A | CDCA7 | 0.94 (0.92-0.97) | $3.0 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs10931936 | 2 | 202143928 | C/T | CASP8 | 0.88 (0.82-0.94) | $1.5 \times 10^{-4}$ | European | Turnbull et al. (8) |
| rs1045485 | 2 | 202149589 | C/G | CASP8 | 0.88 (0.84-0.92) | $5.7 \times 10^{-7}$ | European | Cox et al. (45) |
| rs13387042 | 2 | 217905832 | G/A | None | 0.88 (0.86-0.90) | $2.2 \times 10^{-57}$ | European | Michailidou et al. (11) |
| rs16857609 | 2 | 218296508 | T/C | DIRC3 | 1.08 (1.06-1.10) | $1.1 \times 10^{-15}$ | European | Michailidou et al. (11) |
| rs6762644 | 3 | 4742276 | G/A | ITPR1-EGOT | 1.07 (1.04-1.09) | $2.2 \times 10^{-12}$ | European | Michailidou et al. (11) |
| rs4973768 | 3 | 27416013 | T/C | SLC4A7 | 1.10 (1.08-1.12) | $2.3 \times 10^{-30}$ | European | Michailidou et al. (11) |
| rs12493607 | 3 | 30682939 | C/G | TGFBR2 | 1.06 (1.03-1.08) | $2.3 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs6796502 | 3 | 46866866 | A/G | None | 0.92 (0.89-0.95) | $1.8 \times 10^{-8}$ | European | Michailidou et al. (13) |
| rs9790517 | 4 | 106084778 | T/C | TET2 | 1.05 (1.03-1.08) | $4.2 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs6828523 | 4 | 175846426 | A/C | ADAM29 | 0.90 (0.87-0.92) | $3.5 \times 10^{-16}$ | European | Michailidou et al. (11) |
| rs10069690 | 5 | 1279790 | T/C | TERT | 1.06 (1.04-1.09) | $7.2 \times 10^{-9}$ | European | Michailidou et al. (11) |
| rs13162653 | 5 | 16187528 | T/G | None | 0.95 (0.93-0.97) | $1.1 \times 10^{-10}$ | European | Michailidou et al. (13) |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs2012709 | 5 | 32567732 | T/C | None | 1.05 (1.03-1.08) | $6.4 \times 10^{-9}$ | European | Michailidou et al. (13) |
| rs10941679 | 5 | 44706498 | G/A | None | 1.13 (1.10-1.15) | $1.7 \times 10^{-37}$ | European | Michailidou et al. (11) |
| rs9790879 | 5 | 44899885 | C/T | None | 1.10 (1.03-1.17) | $3.2 \times 10^{-3}$ | European | Turnbull et al. (8) |
| rs889312 | 5 | 56031884 | C/A | MAP3K1 | 1.12 (1.10-1.15) | $2.7 \times 10^{-36}$ | European | Michailidou et al. (11) |
| rs10472076 | 5 | 58184061 | C/T | RAB3C | 1.05 (1.03-1.07) | $2.9 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs1353747 | 5 | 58337481 | G/T | PDE4D | 0.92 (0.89-0.95) | $2.5 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs7707921 | 5 | 81538046 | T/A | ATG10 | 0.93 (0.91-0.95) | $5.0 \times 10^{-11}$ | European | Michailidou et al. (13) |
| rs10474352 | 5 | 90732225 | C/T | ARRDC3 | 1.09 (1.06-1.12) | $1.7 \times 10^{-9}$ | East Asian | Cai et al. (6) |
| rs1432679 | 5 | 158244083 | C/T | EBF1 | 1.07 (1.05-1.09) | $2.0 \times 10^{-14}$ | European | Michailidou et al. (11) |
| rs11242675 | 6 | 1318878 | C/T | FOXQ1 | 0.94 (0.92-0.96) | $7.1 \times 10^{-9}$ | European | Michailidou et al. (11) |
| rs204247 | 6 | 13722523 | G/A | RANBP9 | 1.05 (1.03-1.07) | $8.3 \times 10^{-9}$ | European | Michailidou et al. (11) |
| rs9257408 | 6 | 28926220 | C/G | None | 1.05 (1.03-1.08) | $4.8 \times 10^{-8}$ | European | Michailidou et al. (13) |
| rs17529111 | 6 | 82128386 | G/A | None | 1.06 (1.04-1.09) | $3.2 \times 10^{-7}$ | Caucasian | Purrington et al. (146) |
| rs17530068 | 6 | 82193109 | G/A | None | 1.05 (1.03-1.08) | $8.2 \times 10^{-9}$ | European | Michailidou et al. (11) |
| rs2180341 | 6 | 127600630 | G/A | ECHDC1 | 1.41 (1.25-1.59) | $2.9 \times 10^{-8}$ | Ashkenazi Jews | Gold et al. (9) |
| rs9485372 | 6 | 149608874 | A/G | TAB2 | 0.90 (0.87-0.92) | $3.8 \times 10^{-12}$ | East Asian | Long et al. (50) |
| rs3757318 | 6 | 151914113 | A/G | ESR1 | 1.16 (1.12-1.21) | $2.2 \times 10^{-21}$ | European | Michailidou et al. (11) |
| rs2046210 | 6 | 151948366 | A/G | ESR1 | 1.08 (1.06-1.10) | $2.0 \times 10^{-19}$ | European | Michailidou et al. (11) |
| rs4593472 | 7 | 130667121 | T/C | FLJ43663 | 0.95 (0.94-0.97) | $1.8 \times 10^{-9}$ | European | Michailidou et al. (13) |
| rs720475 | 7 | 144074929 | A/G | ARHGEF5-NOBOX | 0.94 (0.92-0.96) | $7.0 \times 10^{-11}$ | European | Michailidou et al. (11) |
| rs9693444 | 8 | 29509616 | A/C | None | 1.07 (1.05-1.09) | $9.2 \times 10^{-14}$ | European | Michailidou et al. (11) |
| rs13365225 | 8 | 36858483 | G/A | None | 0.95 (0.93-0.98) | $1.1 \times 10^{-8}$ | European | Michailidou et al. (13) |
| rs6472903 | 8 | 76230301 | G/T | None | 0.91 (0.89-0.93) | $1.7 \times 10^{-17}$ | European | Michailidou et al. (11) |
| rs2943559 | 8 | 76417937 | G/A | HNF4G | 1.13 (1.09-1.17) | $5.7 \times 10^{-15}$ | European | Michailidou et al. (11) |
| rs13267382 | 8 | 117209548 | A/G | LINC00536 | 1.05 (1.03-1.07) | $1.7 \times 10^{-8}$ | European | Michailidou et al. (13) |
| rs13281615 | 8 | 128355618 | G/A | None | 1.09 (1.07-1.12) | $9.6 \times 10^{-28}$ | European | Michailidou et al. (11) |
| rs1562430 | 8 | 128387852 | T/C | None | 1.17 (1.10-1.25) | $5.8 \times 10^{-7}$ | European | Turnbull et al. (8) |
| rs11780156 | 8 | 129194641 | T/C | MIR1208 | 1.07 (1.04-1.10) | $3.4 \times 10^{-11}$ | European | Michailidou et al. (11) |
| rs1011970 | 9 | 22062134 | T/G | CDKN2A/B | 1.06 (1.03-1.08) | $5.5 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs10759243 | 9 | 110306115 | A/C | None | 1.06 (1.03-1.08) | $1.2 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs865686 | 9 | 110888478 | G/T | None | 0.89 (0.88-0.91) | $9.5 \times 10^{-35}$ | European | Michailidou et al. (11) |
| rs2380205 | 10 | 5886734 | T/C | ANKRD16,FBXO18 | 0.94 (0.91-0.98) | $4.6 \times 10^{-7}$ | European | Turnbull et al. (8) |
| rs7072776 | 10 | 22032942 | A/G | MLLT10-DNAJC1 | 1.07 (1.05-1.09) | $4.3 \times 10^{-14}$ | European | Michailidou et al. (11) |

| SNP | Chr | Position | Allele | Gene | OR (95% CI) | P-value | Population | Reference |
|---|---|---|---|---|---|---|---|---|
| rs11814448 | 10 | 22315843 | C/A | DNAJC1 | 1.26 (1.18-1.35) | $9.3 \times 10^{-16}$ | European | Michailidou et al. (11) |
| rs10822013 | 10 | 64251977 | T/C | ZNF365 | 1.12 (1.06-1.18) | $5.9 \times 10^{-9}$ | East Asian | Cai et al. (28) |
| rs10995190 | 10 | 64278682 | A/G | ZNF365 | 0.86 (0.84-0.88) | $1.3 \times 10^{-36}$ | European | Michailidou et al. (11) |
| rs704010 | 10 | 80841148 | T/C | ZMIZ1 | 1.08 (1.06-1.10) | $7.4 \times 10^{-22}$ | European | Michailidou et al. (11) |
| rs7904519 | 10 | 114773927 | G/A | TCF7L2 | 1.06 (1.04-1.08) | $3.1 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs11199914 | 10 | 123093901 | T/C | None | 0.95 (0.93-0.97) | $1.9 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs2981579 | 10 | 123337335 | A/G | FGFR2 | 1.27 (1.24-1.29) | $1.9 \times 10^{-170}$ | European | Michailidou et al. (11) |
| rs1219648 | 10 | 123346190 | G/A | FGFR2 | 1.20 (1.07-1.42) | $1.1 \times 10^{-10}$ | European | Hunter et al. (147) |
| rs2981582 | 10 | 123352317 | A/G | FGFR2 | 1.26(1.23-1.30) | $2.0 \times 10^{-76}$ | European | Easton et al. (148) |
| rs3817198 | 11 | 1909006 | C/T | LSP1 | 1.07 (1.05-1.09) | $1.5 \times 10^{-11}$ | European | Michailidou et al. (11) |
| rs909116 | 11 | 1941946 | T/C | LSP1 | 1.17 (1.10-1.24) | $7.3 \times 10^{-7}$ | European | Turnbull et al. (8) |
| rs12575663 | 11 | 65574535 | A/G | OVOL1 | 0.95 (0.93-0.96) | $8.6 \times 10^{-12}$ | Asian and European | Shi et al. (82) |
| rs3903072 | 11 | 65583066 | T/G | DKFZp761E198-OVOL1-SNX32-CFL1-MUS81 | 0.95 (0.93-0.96) | $8.6 \times 10^{-12}$ | European | Michailidou et al. (11) |
| rs614367 | 11 | 69328764 | T/C | None | 1.21 (1.18-1.24) | $2.2 \times 10^{-63}$ | European | Michailidou et al. (11) |
| rs11820646 | 11 | 129461171 | T/C | None | 0.95 (0.93-0.97) | $1.1 \times 10^{-9}$ | European | Michailidou et al. (11) |
| rs7107217 | 11 | 129473690 | C/A | TMEM45B,BARX2 | 1.08(1.05-1.11) | $4.6 \times 10^{-7}$ | East Asian | Long et al. (50) |
| rs12422552 | 12 | 14413931 | C/G | None | 1.05 (1.03-1.07) | $3.7 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs10771399 | 12 | 28155080 | G/A | PTHLH | 0.86 (0.83-0.88) | $8.1 \times 10^{-31}$ | European | Michailidou et al. (11) |
| rs17356907 | 12 | 96027759 | G/A | NTN4 | 0.91 (0.89-0.93) | $1.8 \times 10^{-22}$ | European | Michailidou et al. (11) |
| rs1292011 | 12 | 115836522 | G/A | None | 0.92 (0.90-0.94) | $8.9 \times 10^{-22}$ | European | Michailidou et al. (11) |
| rs11571833 | 13 | 32972626 | T/A | BRCA2-N4BP2L1-N4BP2L2 | 1.26 (1.14-1.39) | $4.9 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs2236007 | 14 | 37132769 | A/G | PAX9-SLC25A21 | 0.93 (0.91-0.95) | $1.7 \times 10^{-13}$ | European | Michailidou et al. (11) |
| rs2588809 | 14 | 68660428 | T/C | RAD51L1 | 1.08 (1.05-1.11) | $1.4 \times 10^{-10}$ | European | Michailidou et al. (11) |
| rs999737 | 14 | 69034682 | T/C | RAD51L1 | 0.92 (0.90-0.94) | $2.5 \times 10^{-19}$ | European | Michailidou et al. (11) |
| rs8009944 | 14 | 69039588 | A/C | RAD51L1 | 0.88 (0.82-0.95) | $4.0 \times 10^{-4}$ | European | Turnbull et al. (8) |
| rs941764 | 14 | 91841069 | G/A | CCDC88C | 1.06 (1.04-1.09) | $3.7 \times 10^{-10}$ | European | Michailidou et al. (11) |
| rs11627032 | 14 | 93104072 | C/T | RIN3 | 0.94 (0.92-0.96) | $4.5 \times 10^{-9}$ | European | Michailidou et al. (13) |
| rs2290203 | 15 | 91512067 | G/A | PRC1 | 1.08 (1.05-1.11) | $4.3 \times 10^{-8}$ | East Asian | Cai et al. (6) |
| rs12443621 | 16 | 52548037 | G/A | TNRC9/LOC643714 | 1.11(1.08-1.14) | $2.0 \times 10^{-19}$ | European | Easton et al. (148) |
| rs3803662 | 16 | 52586341 | A/G | TOX3 | 1.24 (1.21-1.27) | $2.1 \times 10^{-114}$ | European | Michailidou et al. (11) |

| rs4784227 | 16 | 52599188 | T/C | TOX3 | 1.24 (1.20-1.29) | $1.3 \times 10^{-28}$ | Combined (Asian and European American) | Long et al. (80) |
|---|---|---|---|---|---|---|---|---|
| rs17817449 | 16 | 53813367 | G/T | MIR1972-2-FTO | 0.93 (0.91-0.95) | $6.4 \times 10^{-14}$ | European | Michailidou et al. (11) |
| rs11075995 | 16 | 53855291 | ?/? | FTO,KIAA1752 | 1.11 (1.07-1.15) | $4.0 \times 10^{-8}$ | European | Garcia-Closas et al. (7) |
| rs13329835 | 16 | 80650805 | G/A | CDYL2 | 1.08 (1.05-1.10) | $2.1 \times 10^{-16}$ | European | Michailidou et al. (11) |
| chr17:29230520:D | 17 | 29230520 | G/GGT | ATAD5 | 0.93 (0.91-0.96) | $3.3 \times 10^{-8}$ | European | Michailidou et al. (13) |
| rs6504950 | 17 | 53056471 | A/G | COX11 | 0.94 (0.92-0.96) | $2.3 \times 10^{-13}$ | European | Michailidou et al. (11) |
| rs745570 | 17 | 77781725 | G/A | None | 0.95 (0.93-0.97) | $1.4 \times 10^{-9}$ | European | Michailidou et al. (13) |
| rs527616 | 18 | 24337424 | C/G | None | 0.95 (0.93-0.97) | $1.6 \times 10^{-10}$ | European | Michailidou et al. (11) |
| rs1436904 | 18 | 24570667 | G/T | CHST9 | 0.96 (0.94-0.98) | $3.2 \times 10^{-8}$ | European | Michailidou et al. (11) |
| rs6507583 | 18 | 42399590 | G/A | SETBP1 | 0.91 (0.88-0.95) | $3.2 \times 10^{-8}$ | European | Michailidou et al. (13) |
| rs8170 | 19 | 17389704 | A/G | ANKLE1,C19orf62,ABHD8 | 1.26 (1.17-1.35) | $2.3 \times 10^{-9}$ | European | Antoniou et al. (149) |
| rs2363956 | 19 | 17394124 | A/C | None | 0.82 (0.77-0.88) | $2.3 \times 10^{-8}$ | Caucasian | Purrington et al. (146) |
| rs4808801 | 19 | 18571141 | G/A | SSBP4-ISYNA1-ELL | 0.93 (0.91-0.95) | $4.6 \times 10^{-15}$ | European | Michailidou et al. (11) |
| rs3760982 | 19 | 44286513 | A/G | C19orf61-KCNN4-LYPD5-ZNF283 | 1.06 (1.04-1.08) | $2.1 \times 10^{-10}$ | European | Michailidou et al. (11) |
| rs2284378 | 20 | 32588095 | T/C | RALY,EIF2S2,ASIP | 1.16 (1.10-1.22) | $1.1 \times 10^{-8}$ | European | Siddiq et al. (3) |
| rs2823093 | 21 | 16520832 | A/G | NRIP1 | 0.92 (0.90-0.94) | $6.8 \times 10^{-16}$ | European | Michailidou et al. (11) |
| rs16992204 | 21 | 36111201 | C/T | LINC00160 | 1.13 (1.07-1.18) | $4.6 \times 10^{-8}$ | East Asian | Han et al. (Under review) |
| rs132390 | 22 | 29621477 | C/T | EMID1-RHBDD3-EWSR1 | 1.12 (1.07-1.18) | $3.1 \times 10^{-9}$ | European | Michailidou et al. (11) |
| rs12628403 | 22 | 39358037 | A/C | APOBEC3 | 1.18 (1.12-1.25) | $2.9 \times 10^{-9}$ | East Asian | Long et al. (51) |
| rs6001930 | 22 | 40876234 | C/T | MKL1 | 1.12 (1.09-1.16) | $8.8 \times 10^{-19}$ | European | Michailidou et al. (11) |

[a] Effect/reference alleles.

[b] Obtained from the meta-analyses for each study.

# REFERENCES

1. Forouzanfar,M.H., Foreman,K.J., Delossantos,A.M., Lozano,R., Lopez,A.D., Murray,C.J.L. and Naghavi,M. (2011) Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *The Lancet*, **378**, 1461–1484.

2. Sueta,A., Ito,H., Kawase,T., Hirose,K., Hosono,S., Yatabe,Y., Tajima,K., Tanaka,H., Iwata,H., Iwase,H., *et al.* (2012) A genetic risk predictor for breast cancer using a combination of low-penetrance polymorphisms in a Japanese population. *Breast Cancer Res. Treat.*, **132**, 711–721.

3. Siddiq,A., Couch,F.J., Chen,G.K., Lindstrom,S., Eccles,D., Millikan,R.C., Michailidou,K., Stram,D.O., Beckmann,L., Rhie,S.K., *et al.* (2012) A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum. Mol. Genet.*, **21**, 5373–5384.

4. Zheng,W., Zhang,B., Cai,Q., Sung,H., Michailidou,K., Shi,J., Choi,J.-Y., Long,J., Dennis,J., Humphreys,M.K., *et al.* (2013) Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. *Hum. Mol. Genet.*, **22**, 2539–2550.

5. Long,J., Shu,X.-O., Cai,Q., Gao,Y.-T., Zheng,Y., Li,G., Li,C., Gu,K., Wen,W., Xiang,Y.-B., *et al.* (2010) Evaluation of breast cancer susceptibility loci in Chinese women. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.*, **19**, 2357–2365.

6. Cai,Q., Zhang,B., Sung,H., Low,S.-K., Kweon,S.-S., Lu,W., Shi,J., Long,J., Wen,W., Choi,J.-Y., *et al.* (2014) Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat. Genet.*, **46**, 886–890.

7. Garcia-Closas,M., Couch,F.J., Lindstrom,S., Michailidou,K., Schmidt,M.K., Brook,M.N., Orr,N., Rhie,S.K., Riboli,E., Feigelson,H.S., *et al.* (2013) Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat. Genet.*, **45**, 392–398, 398e1–2.

8. Turnbull,C., Ahmed,S., Morrison,J., Pernet,D., Renwick,A., Maranian,M., Seal,S., Ghoussaini,M., Hines,S., Healey,C.S., *et al.* (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.*, **42**, 504–507.

9. Gold,B., Kirchhoff,T., Stefanov,S., Lautenberger,J., Viale,A., Garber,J., Friedman,E., Narod,S., Olshen,A.B., Gregersen,P., *et al.* (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 4340–4345.

10. Ghoussaini,M., Pharoah,P.D.P. and Easton,D.F. (2013) Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning? *Am. J. Pathol.*, **183**, 1038–1051.

11. Michailidou,K., Hall,P., Gonzalez-Neira,A., Ghoussaini,M., Dennis,J., Milne,R.L., Schmidt,M.K., Chang-Claude,J., Bojesen,S.E., Bolla,M.K., *et al.* (2013) Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.*, **45**, 353–361.

12. Fletcher,O., Johnson,N., Orr,N., Hosking,F.J., Gibson,L.J., Walker,K., Zelenika,D., Gut,I., Heath,S., Palles,C., *et al.* (2011) Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study. *J. Natl. Cancer Inst.*, **103**, 425–435.

13. Michailidou,K., Beesley,J., Lindstrom,S., Canisius,S., Dennis,J., Lush,M.J., Maranian,M.J., Bolla,M.K., Wang,Q., Shah,M., *et al.* (2015) Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.*, **47**, 373–380.

14. Gorlov,I.P., Gorlova,O.Y., Sunyaev,S.R., Spitz,M.R. and Amos,C.I. (2008) Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. *Am. J. Hum. Genet.*, **82**, 100–112.

15. Zhang,Y., Long,J., Lu,W., Shu,X.O., Cai,Q., Zheng,Y., Li,C., Li,B., Gao,Y.T. and Zheng,W. (2014) Rare coding variants and breast cancer risk: Evaluation of susceptibility loci identified in genome-wide association studies. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.*, **23**, 622–628.

16. Frank,B., Hemminki,K., Wirtenberger,M., Bermejo,J.L., Bugert,P., Klaes,R.., Schmutzler,R.K., Wappenschmidt,B., Bartram,C.R. and Burwinkel,B. (2005) The rare ERBB2 variant Ile654Val is associated with an increased familial breast cancer risk. *Carcinogenesis*, **26**, 643–647.

17. McInerney,N.M., Miller,N., Rowan,A., Colleran,G., Barclay,E., Curran,C., Kerin,M.J., Tomlinson,I.P. and Sawyer,E. (2010) Evaluation of variants in the CHEK2, BRIP1 and PALB2 genes in an Irish breast cancer cohort. *Breast Cancer Res. Treat.*, **121**, 203–210.

18. Ruark,E., Snape,K., Humburg,P., Loveday,C., Bajrami,I., Brough,R., Rodrigues,D.N., Renwick,A., Seal,S., Ramsay,E., *et al.* (2013) Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature*, **493**, 406–410.

19. Lu,W., Wang,X., Lin,H., Lindor,N.M. and Couch,F.J. (2012) Mutation screening of RAD51C in high-risk breast and ovarian cancer families. *Fam. Cancer*, **11**, 381–385.

20. Tavtigian,S.V., Oefner,P.J., Babikyan,D., Hartmann,A., Healey,S., Le Calvez-Kelm,F., Lesueur,F., Byrnes,G.B., Chuang,S.C., Forey,N., *et al.* (2009) Rare, Evolutionarily

Unlikely Missense Substitutions in ATM Confer Increased Risk of Breast Cancer. *Am. J. Hum. Genet.*, **85**, 427–446.

21. Le Calvez-Kelm,F., Lesueur,F., Damiola,F., Vall+�e,M., Voegele,C., Babikyan,D., Durand,G., Forey,N., McKay-Chopin,S., Robinot,N., *et al.* (2011) Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study. *Breast Cancer Res. BCR*, **13**, R6–R6.

22. Dowty,J., Lose,F., Jenkins,M., Chang,J.H., Chen,X., Beesley,J., Dite,G., Southey,M., Byrnes,G., Tesoriero,A., *et al.* (2008) The RAD51D E233G variant and breast cancer risk: population-based and clinic-based family studies of Australian women. *Breast Cancer Res. Treat.*, **112**, 35–39.

23. Goldgar,D.E., Healey,S., Dowty,J.G., Da Silva,L., Chen,X., Spurdle,A.B., Terry,M.B., Daly,M.J., Buys,S.M., Southey,M.C., *et al.* (2011) Rare variants in the ATM gene and risk of breast cancer. *Breast Cancer Res. BCR*, **13**, R73–R73.

24. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.

25. Stitziel,N.O., Binkowski,T.A., Tseng,Y.Y., Kasif,S. and Liang,J. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.*, **32**, D520–D522.

26. Korn,J.M., Kuruvilla,F.G., McCarroll,S.A., Wysoker,A., Nemesh,J., Cawley,S., Hubbell,E., Veitch,J., Collins,P.J., Darvishi,K., *et al.* (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, **40**, 1253–1260.

27. Ritchie,M.E., Liu,R., Carvalho,B.S. and Irizarry,R.A. (2011) Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. *BMC Bioinformatics*, **12**, 68.

28. Cai,Q., Long,J., Lu,W., Qu,S., Wen,W., Kang,D., Lee,J.-Y., Chen,K., Shen,H., Shen,C.-Y., *et al.* (2011) Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. *Hum. Mol. Genet.*, **20**, 4991–4999.

29. Han,M.-R., Deming-Halverson,S., Cai,Q., Wen,W., Shrubsole,M.J., Shu,X.-O., Zheng,W. and Long,J. (2014) Evaluating 17 breast cancer susceptibility loci in the Nashville breast health study. *Breast Cancer*, **22**, 544–551.

30. Cui,Y., Deming-Halverson,S.L., Shrubsole,M.J., Beeghly-Fadiel,A., Fair,A.M., Sanderson,M., Shu,X.-O., Kelley,M.C. and Zheng,W. (2014) Associations of Hormone-Related Factors With Breast Cancer Risk According to Hormone Receptor Status Among White and African American Women. *Clin. Breast Cancer*, **14**, 417–425.

31. Signorello,L.B., Hargreaves,M.K., Steinwandel,M.D., Zheng,W., Cai,Q., Schlundt,D.G., Buchowski,M.S., Arnold,C.W., McLaughlin,J.K. and Blot,W.J. (2005) Southern community cohort study: establishing a cohort to investigate health disparities. *J. Natl. Med. Assoc.*, **97**, 972–979.

32. Roden,D., Pulley,J., Basford,M., Bernard,G., Clayton,E., Balser,J. and Masys,D. (2008) Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin. Pharmacol. Ther.*, **84**, 362–369.

33. The Cancer Genome Atlas Research Network, Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

34. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N., *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

35. Curtis,C., Shah,S.P., Chin,S.-F., Turashvili,G., Rueda,O.M., Dunning,M.J., Speed,D., Lynch,A.G., Samarajiwa,S., Yuan,Y., *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

36. Siegel,R., Naishadham,D. and Jemal,A. (2012) Cancer statistics, 2012. *CA. Cancer J. Clin.*, **62**, 10–29.

37. Zhang,B., Beeghly-Fadiel,A., Long,J. and Zheng,W. (2011) Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Lancet Oncol.*, **12**, 477–488.

38. Miki,Y., Swensen,J., Shattuck-Eidens,D., Futreal,P.A., Harshman,K., Tavtigian,S., Liu,Q., Cochran,C., Bennett,L.M., Ding,W., *et al.* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, **266**, 66–71.

39. Wooster,R., Bignell,G., Lancaster,J., Swift,S., Seal,S., Mangion,J., Collins,N., Gregory,S., Gumbs,C., Micklem,G., *et al.* (1995) Identification of the breast cancer susceptibility gene BRCA2. *Nature*, **378**, 789–792.

40. Renwick,A., Thompson,D., Seal,S., Kelly,P., Chagtai,T., Ahmed,M., North,B., Jayatilake,H., Barfoot,R., Spanova,K., *et al.* (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.*, **38**, 873–875.

41. Meijers-Heijboer,H., Ouweland,A. van den, Klijn,J., Wasielewski,M., Snoo,A. de, Oldenburg,R., Hollestelle,A., Houben,M., Crepin,E., Veghel-Plandsoen,M. van, *et al.* (2002) Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.*, **31**, 55–59.

42. Seal,S., Thompson,D., Renwick,A., Elliott,A., Kelly,P., Barfoot,R., Chagtai,T., Jayatilake,H., Ahmed,M., Spanova,K., *et al.* (2006) Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.*, **38**, 1239–1241.

43. Rahman,N., Seal,S., Thompson,D., Kelly,P., Renwick,A., Elliott,A., Reid,S., Spanova,K., Barfoot,R., Chagtai,T., *et al.* (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.*, **39**, 165–167.

44. Lin,W.-Y., Camp,N.J., Ghoussaini,M., Beesley,J., Michailidou,K., Hopper,J.L., Apicella,C., Southey,M.C., Stone,J., Schmidt,M.K., *et al.* (2015) Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. *Hum. Mol. Genet.*, **24**, 285–298.

45. Cox,A., Dunning,A.M., Garcia-Closas,M., Balasubramanian,S., Reed,M.W.R., Pooley,K.A., Scollen,S., Baynes,C., Ponder,B.A.J., Chanock,S., *et al.* (2007) A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.*, **39**, 352–358.

46. Stratton,M.R. and Rahman,N. (2008) The emerging landscape of breast cancer susceptibility. *Nat. Genet.*, **40**, 17–22.

47. Meindl,A., Hellebrand,H., Wiek,C., Erven,V., Wappenschmidt,B., Niederacher,D., Freund,M., Lichtner,P., Hartmann,L., Schaal,H., *et al.* (2010) Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. *Nat. Genet.*, **42**, 410–414.

48. Walsh,T. and King,M.C. (2007) Ten Genes for Inherited Breast Cancer. *Cancer Cell*, **11**, 103–105.

49. Zheng,W., Long,J., Gao,Y.-T., Li,C., Zheng,Y., Xiang,Y.-B., Wen,W., Levy,S., Deming,S.L., Haines,J.L., *et al.* (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.*, **41**, 324–328.

50. Long,J., Cai,Q., Sung,H., Shi,J., Zhang,B., Choi,J.-Y., Wen,W., Delahanty,R.J., Lu,W., Gao,Y.-T., *et al.* (2012) Genome-Wide Association Study in East Asians Identifies Novel Susceptibility Loci for Breast Cancer. *PLoS Genet.*, **8**, e1002532.

51. Long,J., Delahanty,R.J., Li,G., Gao,Y.-T., Lu,W., Cai,Q., Xiang,Y.-B., Li,C., Ji,B.-T., Zheng,Y., *et al.* (2013) A Common Deletion in the APOBEC3 Genes and Breast Cancer Risk. *J. Natl. Cancer Inst.*, **105**, 573–579.

52. Antoniou,A., Pharoah,P.D.P., Narod,S., Risch,H.A., Eyfjord,J.E., Hopper,J.L., Loman,N., Olsson,H., Johannsson,O., Borg,+�, *et al.* (2003) Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. *Am. J. Hum. Genet.*, **72**, 1117–1130.

53. Chen,S. and Parmigiani,G. (2007) Meta-Analysis of BRCA1 and BRCA2 Penetrance. *J. Clin. Oncol.*, **25**, 1329–1333.

54. Tan,M.H., Mester,J.L., Ngeow,J., Rybicki,L.A., Orloff,M.S. and Eng,C. (2012) Lifetime Cancer Risks in Individuals with Germline PTEN Mutations. *Clin. Cancer Res.*, **18**, 400–407.

55. Bubien,V., Bonnet,F., Brouste,V., Hoppe,S., Barouk-Simonet,E., David,A., Edery,P., Bottani,A., Layet,V., Caron,O., *et al.* (2013) High cumulative risks of cancer in patients with PTEN hamartoma tumour syndrome. *J. Med. Genet.*, **50**, 255–263.

56. Nieuwenhuis,M.H., Kets,C.M., Murphy-Ryan,M., Yntema,H.G., Evans,D.G., Colas,C., Møller,P., Hes,F.J., Hodgson,S.V., Olderode-Berends,M.J.W., *et al.* (2013) Cancer risk and genotype–phenotype correlations in PTEN hamartoma tumor syndrome. *Fam. Cancer*, **13**, 57–63.

57. Lalloo,F., Varley,J., Moran,A., Ellis,D., O'Dair,L., Pharoah,P., Antoniou,A., Hartley,R., Shenton,A., Seal,S., *et al.* (2006) BRCA1, BRCA2 and TP53 mutations in very early-onset breast cancer with associated risks to relatives. *Eur. J. Cancer*, **42**, 1143–1150.

58. Walsh,T., Casadei,S. and Coats,K. (2006) Spectrum of mutations in brca1, brca2, chek2, and tp53 in families at high risk of breast cancer. *JAMA*, **295**, 1379–1388.

59. Consortium,C.B.C.C.-C. (2004) CHEK2*1100delC and Susceptibility to Breast Cancer: A Collaborative Analysis Involving 10,860 Breast Cancer Cases and 9,065 Controls from 10 Studies. *Am. J. Hum. Genet.*, **74**, 1175–1182.

60. Yang,Y., Zhang,F., Wang,Y. and Liu,S.-C. (2012) CHEK2 1100delC variant and breast cancer risk in Caucasians: a meta-analysis based on 25 studies with 29,154 cases and 37,064 controls. *Asian Pac. J. Cancer Prev.*, **13**, 3501–3505.

61. Ahmed,M. and Rahman,N. (2006) ATM and breast cancer susceptibility. *Oncogene*, **25**, 5906–5911.

62. Gracia-Aznarez,F.J., Fernandez,V., Pita,G., Peterlongo,P., Dominguez,O., de la Hoya,M., Duran,M., Osorio,A., Moreno,L., Gonzalez-Neira,A., *et al.* (2013) Whole Exome Sequencing Suggests Much of Non-BRCA1/BRCA2 Familial Breast Cancer Is Due to Moderate and Low Penetrance Susceptibility Alleles. *PLoS ONE*, **8**, e55681.

63. Zaitlen,N., Kraft,P., Patterson,N., Pasaniuc,B., Bhatia,G., Pollack,S. and Price,A.L. (2013) Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits. *PLoS Genet*, **9**, e1003520.

64. Zuk,O., Hechter,E., Sunyaev,S.R. and Lander,E.S. (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.*, **109**, 1193–1198.

65. Gibson,G. (2012) Rare and Common Variants: Twenty arguments. *Nat. Rev. Genet.*, **13**, 135–145.

66. Eichler,E.E., Flint,J., Gibson,G., Kong,A., Leal,S.M., Moore,J.H. and Nadeau,J.H. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

67. Schork,N.J., Murray,S.S., Frazer,K.A. and Topol,E.J. (2009) Common vs. Rare Allele Hypotheses for Complex Diseases. *Curr. Opin. Genet. Dev.*, **19**, 212–219.

68. Pritchard,J.K. (2001) Are Rare Variants Responsible for Susceptibility to Complex Diseases? *Am. J. Hum. Genet.*, **69**, 124–137.

69. Turnbull,C. and Rahman,N. (2008) Genetic Predisposition to Breast Cancer: Past, Present, and Future. *Annu. Rev. Genomics Hum. Genet.*, **9**, 321–345.

70. Beaudoin,M., Goyette,P., Boucher,G., Lo,K.S., Rivas,M.A., Stevens,C., Alikashani,A., Ladouceur,M., Ellinghaus,D., Törkvist,L., *et al.* (2013) Deep Resequencing of GWAS Loci Identifies Rare Variants in CARD9, IL23R and RNF186 That Are Associated with Ulcerative Colitis. *PLoS Genet*, **9**, e1003723.

71. Sanna,S., Li,B., Mulas,A., Sidore,C., Kang,H.M., Jackson,A.U., Piras,M.G., Usala,G., Maninchedda,G., Sassu,A., *et al.* (2011) Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLoS Genet*, **7**, e1002198.

72. Johansen,C.T., Wang,J., Lanktree,M.B., Cao,H., McIntyre,A.D., Ban,M.R., Martins,R.A., Kennedy,B.A., Hassell,R.G., Visser,M.E., *et al.* (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nat. Genet.*, **42**, 684–687.

73. Li,Q., Seo,J.-H., Stranger,B., McKenna,A., Pe'er,I., LaFramboise,T., Brown,M., Tyekucheva,S. and Freedman,M.L. (2013) A novel eQTL-based analysis reveals the biology of breast cancer risk loci. *Cell*, **152**, 633–641.

74. Fehrmann,R.S.N., Jansen,R.C., Veldink,J.H., Westra,H.-J., Arends,D., Bonder,M.J., Fu,J., Deelen,P., Groen,H.J.M., Smolonska,A., *et al.* (2011) Trans-eQTLs Reveal That Independent Genetic Variants Associated with a Complex Phenotype Converge on Intermediate Genes, with a Major Role for the HLA. *PLoS Genet*, **7**, e1002197.

75. Pickrell,J.K., Marioni,J.C., Pai,A.A., Degner,J.F., Engelhardt,B.E., Nkadori,E., Veyrieras,J.-B., Stephens,M., Gilad,Y. and Pritchard,J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–772.

76. Nicolae,D.L., Gamazon,E., Zhang,W., Duan,S., Dolan,M.E. and Cox,N.J. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLoS Genet*, **6**, e1000888.

77. Wittkopp,P.J., Haerum,B.K. and Clark,A.G. (2004) Evolutionary changes in cis and trans gene regulation. *Nature*, **430**, 85–88.

78. Cheng,Y., Quinn,J.F. and Weiss,L.A. (2013) An eQTL mapping approach reveals that rare variants in the SEMA5A regulatory network impact autism risk. *Hum. Mol. Genet.*, **22**, 2960–2972.

79. Kuligina,E.S., Sokolenko,A.P., Mitiushkina,N.V., Abysheva,S.N., Preobrazhenskaya,E.V., Gorodnova,T.V., Yanus,G.A., Togo,A.V., Cherdyntseva,N.V., Bekhtereva,S.A., *et al.* (2012) Value of bilateral breast cancer for identification of rare recessive at-risk alleles: evidence for the role of homozygous GEN1 c.2515_2519delAAGTT mutation. *Fam. Cancer*, **12**, 129–132.

80. Long,J., Cai,Q., Shu,X.-O., Qu,S., Li,C., Zheng,Y., Gu,K., Wang,W., Xiang,Y.-B., Cheng,J., *et al.* (2010) Identification of a Functional Genetic Variant at 16q12.1 for Breast Cancer Risk: Results from the Asia Breast Cancer Consortium. *PLoS Genet.*, **6**, e1001002.

81. Ma,X., Beeghly-Fadiel,A., Lu,W., Shi,J., Xiang,Y.-B., Cai,Q., Shen,H., Shen,C.-Y., Ren,Z., Matsuo,K., *et al.* (2012) Pathway Analyses Identify TGFBR2 as Potential Breast Cancer Susceptibility Gene: Results from a Consortium Study among Asians. *Cancer Epidemiol. Biomarkers Prev.*, **21**, 1176–1184.

82. Shi,J., Sung,H., Zhang,B., Lu,W., Choi,J.-Y., Xiang,Y.-B., Kim,M.K., Iwasaki,M., Long,J., Ji,B.-T., *et al.* (2013) New Breast Cancer Risk Variant Discovered at 10q25 in East Asian Women. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.*, **22**, 1297–1303.

83. Cruchaga,C., Karch,C.M., Jin,S.C., Benitez,B.A., Cai,Y., Guerreiro,R., Harari,O., Norton,J., Budde,J., Bertelsen,S., *et al.* (2014) Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer/'s disease. *Nature*, **505**, 550–554.

84. Morris,A.P. (2014) Fine Mapping of Type 2 Diabetes Susceptibility Loci. *Curr. Diab. Rep.*, **14**.

85. Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

86. Scharpf,R.B., Irizarry,R.A., Ritchie,M.E., Carvalho,B. and Ruczinski,I. (2011) Using the R Package crlmm for Genotyping and Copy Number Estimation. *J. Stat. Softw.*, **40**, 1–32.

87. Howie,B., Fuchsberger,C., Stephens,M., Marchini,J. and Abecasis,G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.*, **44**, 955–959.

88. Fuchsberger,C., Abecasis,G.R. and Hinds,D.A. (2015) minimac2: faster genotype imputation. *Bioinformatics*, **31**, 782–784.

89. Li,Q., Seo,J.-H., Stranger,B., McKenna,A., Pe'er,I., Laframboise,T., Brown,M., Tyekucheva,S. and Freedman,M.L. (2013) Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, **152**, 633–641.

90. Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

91. Collins,F.S., Guyer,M.S. and Chakravarti,A. (1997) Variations on a Theme: Cataloging Human DNA Sequence Variation. *Science*, **278**, 1580–1581.

92. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164–e164.

93. MacArthur,D.G., Balasubramanian,S., Frankish,A., Huang,N., Morris,J., Walter,K., Jostins,L., Habegger,L., Pickrell,J.K., Montgomery,S.B., *et al.* (2012) A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, **335**, 823–828.

94. Lum,A. and Le Marchand,L. (1998) A simple mouthwash method for obtaining genomic DNA in molecular epidemiological studies. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.*, **7**, 719–724.

95. Zheng,W., Chow,W.-H., Yang,G., Jin,F., Rothman,N., Blair,A., Li,H.-L., Wen,W., Ji,B.-T., Li,Q., *et al.* (2005) The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am. J. Epidemiol.*, **162**, 1123–1131.

96. Devlin,B. and Roeder,K. (1999) Genomic Control for Association Studies. *Biometrics*, **55**, 997–1004.

97. Lee,S., Wright,F.A. and Zou, and F. (2011) Control of population stratification by correlation-selected principal components. *Biometrics*, **67**, 967–974.

98. Bustamante,C.D., De La Vega,F.M. and Burchard,E.G. (2011) Genomics for the world. *Nature*, **475**, 163–165.

99. Price,A.L., Zaitlen,N.A., Reich,D. and Patterson,N. (2010) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 459–463.

100. Pasaniuc,B., Zaitlen,N., Lettre,G., Chen,G.K., Tandon,A., Kao,W.H.L., Ruczinski,I., Fornage,M., Siscovick,D.S., Zhu,X., *et al.* (2011) Enhanced Statistical Tests for GWAS

in Admixed Populations: Assessment using African Americans from CARe and a Breast Cancer Consortium. *PLoS Genet*, **7**, e1001371.

101. Ma,C., Blackwell,T., Boehnke,M. and Scott,L.J. (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.*, **37**, 539–550.

102. Wang,X. (2014) Firth logistic regression for rare variant association tests. *Front. Genet.*, **5**.

103. Firth,D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

104. Madsen,B.E. and Browning,S.R. (2009) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet*, **5**, e1000384.

105. Li,B. and Leal,S.M. (2008) Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am. J. Hum. Genet.*, **83**, 311–321.

106. Lee,S., Abecasis,G.R., Boehnke,M. and Lin,X. (2014) Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.*, **95**, 5–23.

107. Price,A.L., Kryukov,G.V., de Bakker,P.I.W., Purcell,S.M., Staples,J., Wei,L.-J. and Sunyaev,S.R. (2010) Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.*, **86**, 832–838.

108. Lin,D.-Y. and Tang,Z.-Z. (2011) A General Framework for Detecting Disease Associations with Rare Variants in Sequencing Studies. *Am. J. Hum. Genet.*, **89**, 354–367.

109. Tzeng,J.-Y. and Zhang,D. (2007) Haplotype-Based Association Analysis via Variance-Components Score Test. *Am. J. Hum. Genet.*, **81**, 927–938.

110. Neale,B.M., Rivas,M.A., Voight,B.F., Altshuler,D., Devlin,B., Orho-Melander,M., Kathiresan,S., Purcell,S.M., Roeder,K. and Daly,M.J. (2011) Testing for an Unusual Distribution of Rare Variants. *PLoS Genet*, **7**, e1001322.

111. Wu,M.C., Lee,S., Cai,T., Li,Y., Boehnke,M. and Lin,X. (2011) Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am. J. Hum. Genet.*, **89**, 82–93.

112. Liu,L., Sabo,A., Neale,B.M., Nagaswamy,U., Stevens,C., Lim,E., Bodea,C.A., Muzny,D., Reid,J.G., Banks,E., *et al.* (2013) Analysis of Rare, Exonic Variation amongst Subjects with Autism Spectrum Disorders and Population Controls. *PLoS Genet*, **9**, e1003443.

113. Evangelou,E. and Ioannidis,J.P.A. (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.*, **14**, 379–389.

114. Lin,D.Y. and Zeng,D. (2010) Meta-Analysis of Genome-Wide Association Studies: No Efficiency Gain in Using Individual Participant Data. *Genet. Epidemiol.*, **34**.

115. Rödel,E. (1971) Fisher, R. A.: Statistical Methods for Research Workers (Oliver & Boyd, Edinburgh). *Biom. Z.*, **13**, 429–430.

116. Liu,D.J., Peloso,G.M., Zhan,X., Holmen,O.L., Zawistowski,M., Feng,S., Nikpay,M., Auer,P.L., Goel,A., Zhang,H., *et al.* (2014) Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.*, **46**, 200–204.

117. Lee,S., Teslovich,T.M., Boehnke,M. and Lin,X. (2013) General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.*, **93**, 42–53.

118. Willer,C.J., Li,Y. and Abecasis,G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.

119. Borenstein,M., Hedges,L.V., Higgins,J.P.T. and Rothstein,H.R. (2009) Introduction to Meta-Analysis. John Wiley & Sons.

120. Feng,S., Liu,D., Zhan,X., Wing,M.K. and Abecasis,G.R. (2014) RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics*, **30**, 2828–2829.

121. Vukcevic,D., Hechter,E., Spencer,C. and Donnelly,P. (2011) Disease model distortion in association studies. *Genet. Epidemiol.*, **35**, 278–290.

122. Chahrour,M.H., Yu,T.W., Lim,E.T., Ataman,B., Coulter,M.E., Hill,R.S., Stevens,C.R., Schubert,C.R., Greenberg,M.E., Gabriel,S.B., *et al.* (2012) Whole-Exome Sequencing and Homozygosity Analysis Implicate Depolarization-Regulated Neuronal Genes in Autism. *PLoS Genet*, **8**, e1002635.

123. Keller,M.C., Simonson,M.A., Ripke,S., Neale,B.M., Gejman,P.V., Howrigan,D.P., Lee,S.H., Lencz,T., Levinson,D.F., Sullivan,P.F., *et al.* (2012) Runs of Homozygosity Implicate Autozygosity as a Schizophrenia Risk Factor. *PLoS Genet*, **8**, e1002656.

124. Chen,R., Wei,Q., Zhan,X., Zhong,X., Sutcliffe,J., Cox,N., Cook,E.H., Li,C., Chen,W. and Li,B. (2015) A haplotype-based framework for group-wise transmission/disequilibrium tests for rare variant association analysis. *Bioinformatics*, 10.1093/bioinformatics/btu860.

125. Lim,E.T., Raychaudhuri,S., Sanders,S.J., Stevens,C., Sabo,A., MacArthur,D.G., Neale,B.M., Kirby,A., Ruderfer,D.M., Fromer,M., *et al.* (2013) Rare Complete Knockouts in Humans: Population Distribution and Significant Role in Autism Spectrum Disorders. *Neuron*, **77**, 235–242.

126. Mao,H., Yang,W., Lee,P.P.W., Ho,M.H.-K., Yang,J., Zeng,S., Chong,C.-Y., Lee,T.-L., Tu,W. and Lau,Y.-L. (2012) Exome sequencing identifies novel compound heterozygous mutations of IL-10 receptor 1 in neonatal-onset Crohn's disease. *Genes Immun.*, **13**, 437–442.

127. Delaneau,O., Zagury,J.-F. and Marchini,J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*, **10**, 5–6.

128. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.

129. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat Meth*, **7**, 248–249.

130. Choi,Y., Sims,G.E., Murphy,S., Miller,J.R. and Chan,A.P. (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, **7**, e46688.

131. Choi,Y. (2012) A Fast Computation of Pairwise Sequence Alignment Scores Between a Protein and a Set of Single-locus Variants of Another Protein. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '12. ACM, New York, NY, USA, pp. 414–417.

132. Heyn,H., Moran,S., Hernando-Herraez,I., Sayols,S., Gomez,A., Sandoval,J., Monk,D., Hata,K., Marques-Bonet,T., Wang,L., *et al.* (2013) DNA methylation contributes to natural human variation. *Genome Res.*, 10.1101/gr.154187.112.

133. Wagner,J.R., Busche,S., Ge,B., Kwan,T., Pastinen,T. and Blanchette,M. (2014) The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.*, **15**, R37.

134. Consortium,T. 1000 G.P. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

135. MacArthur,D.G. and Tyler-Smith,C. (2010) Loss-of-function variants in the genomes of healthy humans. *Hum. Mol. Genet.*, **19**, R125–R130.

136. Li,L., Wei,X.H., Pan,Y.P., Li,H.C., Yang,H., He,Q.H., Pang,Y., Shan,Y., Xiong,F.X., Shao,G.Z., *et al.* (2010) LAPTM4B: A novel cancer-associated gene motivates multidrug resistance through efflux and activating PI3K/AKT signaling. *Oncogene*, **29**, 5785–5795.

137. Shao,G.-Z., Zhou,R.-L., Zhang,Q.-Y., Zhang,Y., Liu,J.-J., Rui,J.-A., Wei,X. and Ye,D.-X. (2003) Molecular cloning and characterization of LAPTM4B, a novel gene upregulated in hepatocellular carcinoma. *Oncogene*, **22**, 5060–5069.

138. Hogue,D.L., Ellison,M.J., Young,J.D. and Cass,C.E. (1996) Identification of a Novel Membrane Transporter Associated with Intracellular Membranes by Phenotypic Complementation in the Yeast Saccharomyces cerevisiae. *J. Biol. Chem.*, **271**, 9801–9808.

139. Britschgi,A., Bill,A., Brinkhaus,H., Rothwell,C., Clay,I., Duss,S., Rebhan,M., Raman,P., Guy,C.T., Wetzel,K., *et al.* (2013) Calcium-activated chloride channel ANO1 promotes breast cancer progression by activating EGFR and CAMK signaling. *Proc. Natl. Acad. Sci.*, **110**, E1026–E1034.

140. Wu,H., Guan,S., Sun,M., Yu,Z., Zhao,L., He,M., Zhao,H., Yao,W., Wang,E., Jin,F., *et al.* (2015) Ano1/TMEM16A Overexpression Is Associated with Good Prognosis in PR-Positive or HER2-Negative Breast Cancer Patients following Tamoxifen Treatment. *PLoS ONE*, **10**, e0126128.

141. Sagona,A.P., Nezis,I.P., Bache,K.G., Haglund,K., Bakken,A.C., Skotheim,R.I. and Stenmark,H. (2011) A Tumor-Associated Mutation of FYVE-CENT Prevents Its Interaction with Beclin 1 and Interferes with Cytokinesis. *PLoS ONE*, **6**, e17086.

142. McMillan,D.R. and White,P.C. (2010) Studies on the very large G protein-coupled receptor: from initial discovery to determining its role in sensorineural deafness in higher animals. *Adv. Exp. Med. Biol.*, **706**, 76–86.

143. Hilgert,N., Kahrizi,K., Dieltjens,N., Bazazzadegan,N., Najmabadi,H., Smith,R.J.H. and Camp,G.V. (2009) A large deletion in GPR98 causes type IIC Usher syndrome in male and female members of an Iranian family. *J. Med. Genet.*, **46**, 272–276.

144. Gelman,I.H. (2002) The role of SSeCKS/gravin/AKAP12 scaffolding proteins in the spaciotemporal control of signaling pathways in oncogenesis and development. *Front. Biosci. J. Virtual Libr.*, **7**, d1782–1797.

145. Li,B., Liu,D.J. and Leal,S.M. (2001) Identifying Rare Variants Associated with Complex Traits via Sequencing. In *Current Protocols in Human Genetics*. John Wiley & Sons, Inc.

146. Purrington,K.S., Slager,S., Eccles,D., Yannoukakos,D., Fasching,P.A., Miron,P., Carpenter,J., Chang-Claude,J., Martin,N.G., Montgomery,G.W., *et al.* (2013) Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple negative breast cancer. *Carcinogenesis*, 10.1093/carcin/bgt404.

147. Hunter,D.J., Kraft,P., Jacobs,K.B., Cox,D.G., Yeager,M., Hankinson,S.E., Wacholder,S., Wang,Z., Welch,R., Hutchinson,A., *et al.* (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.

148. Easton,D.F., Pooley,K.A., Dunning,A.M., Pharoah,P.D.P., Thompson,D., Ballinger,D.G., Struewing,J.P., Morrison,J., Field,H., Luben,R., *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.

149. Antoniou,A.C., Wang,X., Fredericksen,Z.S., McGuffog,L., Tarrell,R., Sinilnikova,O.M., Healey,S., Morrison,J., Kartsonaki,C., Lesnick,T., *et al.* (2010) A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat. Genet.*, **42**, 885–892.