Identification and Characterization of a Novel Premenopausal Breast Cancer Locus and Insights into Copy Number Variations for Disease Predisposition and Prognosis

by

Mahalakshmi Kumaran

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Laboratory Medicine and Pathology
University of Alberta

Abstract

Breast cancer is a complex multifactorial disease with the interplay of genetic, environmental and lifestyle factors contributing to the disease risk. Studies based on twins estimated that ~30% of the risk is due to genetic factors. High and moderate penetrant mutations along with low penetrance variants accounted for a proportion of the total heritable risk. Remaining heritability is yet to be accounted for.

My thesis is based on genome-wide analysis of both SNPs and Copy Number Variations (CNVs) as genetic determinants of breast cancer risk.

(i) Characterization of the SNP rs1429142 conferring premenopausal breast cancer risk

I focused on SNP (rs1429142 on chromosome locus 4q31.22) associated with premenopausal breast cancer risk, first of its kind in literature reported by the Damaraju laboratory (Stages 1-3). In the current study additional cases were genotyped (Stage 4). In the analysis of the combined samples (Stage1-4; 4331 cases/4271 controls) the index SNP showed genome-wide significance (OR 1.25, p-value 4.35x10-8). Analysis of rs1429142 showed elevated risk in premenopausal women (n=1503 cases/4271 controls; odds ratio (OR) 1.40, p-value 5.81x10-10). Postmenopausal Caucasian women (n=2700 cases/4271 controls) showed modest risk (OR 1.17; p-value 7.81x10-04) and this finding was confirmed in the postmenopausal cohort from Cancer Genetic Markers of Susceptibility study (CGEMS, USA). SNP rs1429142 showed an association among premenopausal women with African ancestry (OR minor allele 0.82; p-value-1.45x10^{-02}).

Since the index SNP, rs1429142, was in an intergenic region[a], fine-scale mapping of the locus 4q31.22 revealed 135 SNPs to be associated with premenopausal risk. Conditional regression analysis did not reveal any additional peaks of association. Likelihood ratio analysis excluded five variants that were less likely causal compared to the strongly associated SNP. I further refined the putative loci (130 SNPs) by linkage disequilibrium (LD) block mapping and compared patterns for Caucasian and African populations (HapMap data).

I examined active enhancer functions based on chromatin state (histone marks, DNase hypersensitive sites) in human breast cell lines (HMEC, vHEMC) and breast myoepithelial primary cells using data from publicly available resources. I found evidence for the binding of the transcription factors (C-FOS, STAT1/3, POL2/3) at SNP sites in the human breast cell line MCF10A-Er-Src. Three SNPs (rs1366691, rs1429139, rs7667633) were identified as potentially causal and appeared to be part of the predicted Topologically Associated Domain (TAD), helping to explain short-range interactions and enhancer-promoter cross-talk.

(ii) CNV association studies: I studied CNVs, which are larger in size (>50 bp and up to 1Mb) relative to the single base changes of SNPs. CNVs harbor both coding and non-coding genes and may exert gene-dosage effects or regulatory functions. Whole genome CNVs were captured in 422 cases and 348 controls using the Human Affymetrix SNP 6 array platform (discovery dataset). Whole genome copy number estimation was

[a] Intergenic regions are also referred as ‘gene desert regions’ in the thesis

performed and the CNVs with frequencies > 10% and overlapping protein-coding genes were considered further. Association analysis revealed a total of 200 contiguous CNV regions (CNVRs) or CNVs associated with breast cancer risk (q-value < 0.05).

I investigated if any of the breast cancer associated CNVs show prognostic relevance since SNP GWAS attempts to identify prognostic markers were thus far unsuccessful. Among the 200 associated CNVs/CNVRs, 21 CNVRs (overlapping with 22 genes) showed association with Overall survival (OS) and Recurrence Free Survival (RFS). CNVs were interrogated for gene dosage effects by correlating copy number status with breast tumor tissue gene expression. Also, I interrogated the role of germline CNVs harboring small-noncoding RNAs in conferring breast cancer risk. Further, I investigated the breast tissue specific expression of CNV-embedded small-noncoding RNAs (CNV-sncRNAs) to understand the post-transcriptional gene regulatory mechanisms and how they might contribute to breast cancer. I used 495 samples (Affymetrix 6 array data) available in the TCGA as my validation set and identified 1812 breast cancer associated CNVs harboring miRNAs (n=38), piRNAs (n=9865), snoRNAs (n=71) and tRNAs (n=12) genes. A subset of CNV-sncRNAs expressed in breast tissue (tumor and normal) in TCGA dataset, also showed correlation with germline copy numbers.

In summary, I have fine-mapped premenopausal breast cancer locus and identified potential causal variants which are predicted to have enhancer functions Germline CNVs also are useful markers for breast cancer susceptibility and prognosis.

Preface

This thesis is an original study conducted by Ms. Mahalalakshmi Kumaran. The research work conducted as part of this thesis was approved by local Institutional Research Ethics Committee - Health Research Ethics Board of Alberta-Cancer Committee under Protocols # 26180 and #26126.

Work from chapters 3 and 4 have been published in peer-reviewed journals. Individual contributions from all authors are listed below.

Contents of chapter 3 of this thesis has been published as Mahalakshmi Kumaran, Carol E Cass, Kathryn Graham, John R Mackey, Roland Hubaux, Wan Lam, Yutaka Yasui and Sambasivarao Damaraju "***Germline copy number variations are associated with breast cancer risk and prognosis***," Scientific Reports, volume 7, Article number: 14621. 2017 October 2017. doi:10.1038/s41598-017-14799-7. I performed the experiments, contributed to the study design, statistical and bioinformatics analysis and interpretations. Dr. Sambasivarao Damaraju conceived the study. Dr. Carol E. Cass and Dr. Yutaka Yasui provided insightful suggestions for the study design and interpretations. Dr. Kathryn Graham, Dr. John R Mackey provided access to breast gene expression dataset. Myself and Dr. Sambasivarao Damaraju prepared the manuscript; all contributing authors reviewed the manuscript and provided edits and suggestions.

Contents of chapter 4 of this thesis has been published as Mahalakshmi Kumaran, Preethi Krishnan, Carol E. Cass, Roland Hubaux, Wan Lam, Yutaka Yasui and Sambasivarao Damaraju **"*Breast cancer associated germline structural variants harboring small noncoding RNAs impact post-transcriptional gene regulation,*"** Scientific Reports, volume 8, Article number: 7529. 2018 April 27. doi:10.1038/s41598-018-25801-1. I performed the experiments, contributed to the study design, statistical and bioinformatics analysis and interpretations. Dr. Sambasivarao Damaraju conceived the study. Drs. Carol E. Cass and Yutaka Yasui provided insightful suggestions for the study design and interpretations. Myself and Dr. Sambasivarao Damaraju prepared the manuscript; all contributing authors reviewed the manuscript and provided edits and suggestions.

Dedicated to

To my ever-loving parents and family

Acknowledgements

I take this opportunity to sincerely thank my supervisor Prof. Sambasivarao Damaraju for giving me an opportunity to pursue my doctoral research in the exciting and upcoming field of Genomics. He has identified my strengths and gave me challenging research questions that have tremendously improved my technical and intellectual skills and taught me to aim high. He was very patient and greatly supported my learning curve during my program. He was always available and provided continuous insights into my research. He gave me the opportunity and environment to grow as a confident researcher.

I sincerely thank Prof. Carol E. Cass for being a very supportive committee member. She has always been there to give me valuable inputs throughout my program. Over the years, I have improved my writing skills from her meticulous and constructive feedback on my manuscripts.

I sincerely thank Prof. Yutaka Yasui for being a very supportive committee member. He has provided his expertise and critical feedback on my work. He has provided valuable insights and suggestions in handling challenging statistical problems. It was a great rewarding experience working with him.

I would like to thank Dr. Sunita Ghosh for being friendly and approachable. She has provided valuable feedbacks and critical help with my statistical problems. She has been very supportive and encouraged me throughout.

I thank Dr. John Mackey for his expertise in breast cancer oncology aspects, for providing the interesting perspectives to my research problem, and Dr. Anil A. Joy for clinical expertise as well as valuable suggestions to my manuscripts.

I would like to thank Dr. Russell Greiner for thoughtful discussions for projects and for suggesting high-performance computing platform for my work.

I am thankful to Dr. Monika Keelan for being a very supportive graduate coordinator and chairing my exam. She has motivated me in various instances and have been very resourceful.

I would like to thank Dr. Paul Stothard for serving on my examining committee for my candidacy as well as for my PhD defense. His intriguing questions and suggestions in my candidacy have helped in many ways.

I thank the external examiner Prof. Wei Zheng for being part of my thesis defense examining committee.

I would like to thank Ms. Cheryl Titus, Mr. Ronald Marinas and Ms. Femka Williams for being very friendly and helpful with various administrative work and importantly for being great colleagues.

I would like to thank Jennifer Dufour for her technical assistance, being a supportive and friendly colleague.

I would like to thank Dr. Ashok Narasimhan and Dr. Preethi Krishnan for a being wonderful lab colleague and an extended family in Edmonton. I would cherish the time working with them and time spent together.

I thank Ms. Bhumi Bhatt for her support and as a cheerful lab colleague.

I would like to thank Alberta Cancer Research Biobank and Alberta Tomorrow Project for housing the precious biospecimens and for providing the specimen access to carry out my research, without which this study could not have been possible. I would like to immensely thank all the study participants. I would also like to acknowledge Canadian Breast Cancer Foundation - Prairies/NWT Chapter for providing the operating funds to Prof. Damaraju. I thank the scholarship supports from 75th Anniversary and Medical Science Graduate Program Scholarships from Faculty of Medicine & Dentistry

I would like to thank various online public repositories (dbGAP, TCGA) for providing access to valuable datasets (CGEMS, African Diaspora) that helped me for various critical analyses during my research program. I would like to acknowledge the computational resources made available by West Grid, Compute Canada and very helpful support teams.

I would like to thank my beloved friends Sonia Chech Mathew, Aishwarya Iyer and Fathima Khajamoinudeen for being very supportive and motivating during challenging times.

I am forever indebted to my family for their love and support. My mother believed in my dreams, supported and motivated me to stay positive throughout. My father had been a strong support and provided the means to achieve my dream. My brother Ramesh and sister Rajalakshmi have supported me throughout and made my journey to Canada possible. My beloved brother Karthik Kumaran motivated me to pursue the research career and doctoral degree. My cousins Venkateshwaran, Vigneshwaran, Vidya for being supportive. Special thanks to my father-in-law and mother-in-law for being supportive. I would like to thank my childhood friend Ms. Divya and my school teacher Mrs. Anandhi.

Finally, my heartfelt thanks for my beloved husband and a cheerful friend Dr. Sai Rajajayavel, who has been absolute support for me. His love and motivation helped me to navigate challenging times. He was always proud of my success and has been there for me when I needed the most, without him this journey would not be happy and memorable.

Table of contents

List of Tables

List of Figures

Abbreviations

ABHD8	*Abhydrolase Domain Containing 8*
ANKLE1	*Ankyrin Repeat and LEM Domain Containing 1*
ANKRD20A1	*Ankyrin Repeat Domain 20 Family Member A1*
ANKRD20A3	*Ankyrin Repeat Domain 20 Family Member A3*
ANKS1B	*Ankyrin Repeat and Sterile Alpha Motif Domain Containing 1B*
APOBEC3A_B	*APOBEC3A And APOBEC3B Deletion Hybrid*
ARHGAP10	*Rho Gtpase Activating Protein 10*
ATF7IP	*Activating Transcription Factor 7 Interacting Protein*
ATM	*ATM Serine/Threonine Kinase*
BAGE	*B Melanoma Antigen*
BCAC	Breast Cancer Association Consortium
BMI	Body Mass Index
BRCA1	*BRCA1, DNA Repair Associated*
BRCA2	*BRCA2, DNA Repair Associated*
BRIP1	*BRCA1 Interacting Protein C-Terminal Helicase 1*
BTNL3	*Butyrophilin Like 3*
CACNA1C	*Calcium Voltage-Gated Channel Subunit Alpha1 C*
CASP8	*Caspase 8*
CDCV	Common Disease Common Variant
CDH1	*Cadherin 1*
CDK5	*Cyclin Dependent Kinase 5*
CGEMS	*Cancer Genetic Markers of Susceptibility*
CHEK2	*Checkpoint Kinase 2*
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-seq	Chromatin Immunoprecipitation Combined with DNA Sequencing
CNV	Copy Number Variation
CN-LOH	Copy Neutral Loss of Heterozygosity
CTCF	Ccctc-Binding Factor

DLD	*Dihydrolipoamide Dehydrogenase*
DNAJC1	*Dnaj Heat Shock Protein Family Member C1*
DOCK3	*Dedicator of Cytokinesis 3*
DSBs	*Double strand breaks*
ECHDC1	*Ethylmalonyl-Coa Decarboxylase 1*
EDNRA	*Endothelin Receptor Type A*
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression Quantitative Trait Loci
ER	Estrogen Receptor
ERBB4	*Erb-B2 Receptor Tyrosine Kinase 4*
ESR1	*Estrogen Receptor 1*
ETS2	*ETS Proto-Oncogene 2, Transcription Factor*
EYA1	*EYA Transcriptional Coactivator and Phosphatase 1*
FAIRE-seq	Formaldehyde-Assisted Isolation af Regulatory Elements Combined with DNA Sequencing
FAM27B	*Family with Sequence Similarity 27 Member B*
FAM27E3	*Family with Sequence Similarity 27 Member E3*
FAM66E	*Family with Sequence Similarity 66 Member E*
FGFR2	Fibroblast Growth Factor Receptor 2
FLT3	*Fms Related Tyrosine Kinase 3*
FOS	*Fos Proto-Oncogene, AP-1 Transcription Factor Subunit*
FoTES	*Fork stall and template switching*
GAB1	*GRB2 Associated Binding Protein 1*
GSTM1	*Glutathione S-Transferase Mu 1*
GSTM2	*Glutathione S-Transferase Mu 2*
GSTT1	*Glutathione S-Transferase Theta 1*
GUSBP3	*Glucuronidase, Beta Pseudogene 3*
GUSBP9	*Glucuronidase, Beta Pseudogene 9*
HDAC2	*Histone Deacetylase 2*
HER2	Human Epidermal Growth Factor Receptor 2

HLA-DRB5	*Major Histocompatibility Complex, Class II, DR Beta 5*
HLA-DRB6	*Major Histocompatibility Complex, Class II, DR Beta 6*
HMEC	Human Mammary Epithelial Cells
HOXA4	*Homeobox A4*
iCHAV	Independent Set of Correlated Highly Trait-Associated Variants
JAK1	*Janus Kinase 1*
L1	Long interspersed elements-1
LCE3C	*Late Cornified Envelope 3C*
LCR	Low copy repeats
LD	Linkage Disequilibrium
LGALS9B	*Galectin 9B*
LPA	*Lipoprotein*
MAF	Minor allele frequency
MAP3K1	*Mitogen-Activated Protein Kinase Kinase Kinase 1*
MGLL	*Monoglyceride Lipase*
miRNA	MicroRNA
MLIP	*Muscular LMNA Interacting Protein*
MRPS30	*Mitochondrial Ribosomal Protein S30*
MUC20	*Mucin 20, Cell Surface Associated*
N4BP2L1	*NEDD4 Binding Protein 2 Like 1*
N4BP2L2	*NEDD4 Binding Protein 2 Like 2*
NAIP	*NLR Family Apoptosis Inhibitory Protein*
NAHR	*-Nonallelic homologous recombination*
NF-AT	*Nuclear Factor of Activated T-Cells*
NGF	*Nerve Growth Factor*
NGS	Next Generation Sequencing
NHEJ	Nonhomologous end-joining
NME7	NME/NM23 Family Member 7
NSUN5P1	*NOP2/Sun RNA Methyltransferase Family Member 5 Pseudogene 1*
OR2G6	*Olfactory Receptor Family 2 Subfamily G Member 6*

OR2T11	*Olfactory Receptor Family 2 Subfamily T Member 11*
OR4C6	*Olfactory Receptor Family 4 Subfamily C Member 6*
OR4F16	*Olfactory Receptor Family 4 Subfamily F Member 16*
OR4F29	*Olfactory Receptor Family 4 Subfamily F Member 29*
OR4F3	*Olfactory Receptor Family 4 Subfamily F Member 3*
OR4P4	*Olfactory Receptor Family 4 Subfamily P Member 4*
OR4S2	*Olfactory Receptor Family 4 Subfamily S Member 2*
p27	*Cyclin-Dependent Kinase Inhibitor 1B (P27)*
PALB2	*Partner and Localizer of BRCA2*
PAX4	*Paired Box 4*
PCDH9	*Protocadherin 9*
PDGFRA	*Platelet Derived Growth Factor Receptor Alpha*
piRNA	Piwi Interacting RNA
PML	*Promyelocytic Leukemia*
POLE	*DNA Polymerase Epsilon, Catalytic Subunit*
POLR2A	*Rna Polymerase Ii Subunit A*
POU2F2	*POU Class 2 Homeobox 2*
POU3F2	*POU Class 3 Homeobox 2*
PPIAL4A	*Peptidylprolyl Isomerase A Like 4A*
PPIAL4C	*Peptidylprolyl Isomerase A Like 4C*
PR	Progesteron Receptor
PRKACB	*Protein Kinase Camp-Activated Catalytic Subunit Beta*
PRMT10	*Protein Arginine Methyltransferase 10*
PTEN	*Phosphatase and Tensin Homolog*
PTHLH	*Parathyroid Hormone Like Hormone*
PWM	Position Weighted Matrix
RAB11FIP3	*RAB11 Family Interacting Protein 3*
RAB40B	*RAB40B, Member RAS Oncogene Family*
RAD51B	*RAD51 Paralog B*
RAD51L1	*Rad51 Paralog B*

RAN	*RAN, Member RAS Oncogene Family*
RB1	*RB Transcriptional Corepressor 1*
RBL1	*RB Transcriptional Corepressor Like 1*
RNF146	*Ring Finger Protein 146*
ROPN1L	*Rhophilin Associated Tail Protein 1 Like*
RUNX1T1	*RUNX1 Translocation Partner 1*
SERF1B	*Small EDRK-Rich Factor 1B*
SGCZ	*Sarcoglycan Zeta*
SIAH2	*Siah E3 Ubiquitin Protein Ligase 2*
SLC45A1	*Solute Carrier Family 45 Member 1*
SMA5	*Glucuronidase Beta Pseudogene*
SMN1	*Survival of Motor Neuron 1, Telomeric*
SMN2	*Survival of Motor Neuron 2, Centromeric*
SNORD	*C/D Box Snornas*
snoRNAs	*Small Nucleolar Rnas*
SNPs	Single Nucleotide Polymorphims
SORBS2	*Sorbin And SH3 Domain Containing 2*
SPDEF	*SAM Pointed Domain Containing ETS Transcription Factor*
STAT3	*Signal Transducer and Activator of Transcription 3*
STK11	*Serine/Threonine Kinase 11*
STK11/LKB1	*Serine/Threonine Kinase 11*
TAD	Topologically Associated Domain
TCGA	The Cancer Genome Atlas
TEKT5	*Tektin 5*
TERT	*Telomerase Reverse Transcriptase*
TF	*Transcription Factor*
TMEM18C	*Transmembrane Protein 18*
TNIP3	*TNFAIP3 Interacting Protein 3*
TNRC9	*Tox High Mobility Group Box Family Member 3*
TOX3	*TOX High Mobility Group Box Family Member 3*

TP53	*Tumor Protein P53*
tRNA	*Transfer RNA*
UGT2B15	*UDP Glucuronosyltransferase Family 2 Member B15*
UGT2B17	*UDP Glucuronosyltransferase Family 2 Member B17*
USP17L8	*Ubiquitin Specific Peptidase 17-Like Family Member 8*
ZFP14	*ZFP14 Zinc Finger Protein*
ZNF577	*Zinc Finger Protein 577*
ZNF658	*Zinc Finger Protein 658*

1 Introduction and Review of Literature

1.1. Breast cancer epidemiology

Breast cancer is the second most commonly diagnosed cancer in the world and the most prevalent cancer among women. Nearly 1.7 million breast cancer cases were diagnosed globally in 2012, representing 25% of all cancers diagnosed[1]. The incidence rate varies across different countries; however, breast cancer remains the leading cancer diagnosis in women in both developed countries as well as in developing/under developed countries[1]. The differences in incidence rates across the countries can be ascribed to the better awareness, screening programs and access to health care in the developed world.

Breast cancer is a disease with one of highest mortality rates and ranks fifth among overall cancer related deaths1. Mortality rates are higher in the developing or under developed countries, relative to the developed world due to poorer access to health care1. Early diagnosis and treatments specific to subtypes and availability of treatment modalities (surgery, radiation and chemotherapies) contribute to better outcomes2.

According to the 2017 Canadian Cancer Society statistics[3] breast cancer is the third most commonly diagnosed cancer in Canada. However, it is the leading cancer diagnosis representing 25.5% of all cancer diagnoses among women. One in 8 Canadian women is expected to develop breast cancer during their lifetime. The age distribution of breast cancer incidence in Canada shows that, of women diagnosed with breast cancer, 17% are < 50 years (predominantly pre-menopausal), nearly 51% are between 50-69 years of age (predominantly post-menopausal) and 32% are above the age of 70 years. Traditionally incidence rates were reported based on age at diagnosis and not based on menopausal status. Considering the age cut-offs, the above statistics do not fully explain the individual incidence rates for perimenopausal- and premenopausal women with breast cancer, since the average age at menopause is ~52 in Europe and North America[4].

Also, breast cancer continues to be the second leading cause of cancer related death (13%) among women in Canada. However, the Age-Standardized Mortality Rates (ASMR) have declined since 1988 from 41.7 to 23.2 deaths per 100,000 in 2017[3]. This steady decline is due to more effective screening and better therapies. Similar trends of decline in ASMR have been noted in other developed countries such as the United States, the United Kingdom and Australia[3].

1.1.1. Risk factors

Breast cancer is a complex multifactorial disease. There is strong interplay of genetic, lifestyle and environmental factors in conferring disease risk[5]. There are two major types of risk factors: (i) non-modifiable risk factors such as genetic factors, race or ethnicity, family history, age, age at menarche, age at menopause., and (ii) modifiable risk factors such as body mass index (BMI), and lifestyle factors (including smoking, alcohol consumption, physical activity, breast feeding, oral contraceptive use, hormone replacement therapy). A combination of the above factors influences the overall risk[6].

1.2. Genetic risk factors for breast cancer susceptibility

Epidemiological studies have identified health, lifestyle and environmental factors as the major contributors to risk of breast cancer. However, strong familial clustering[b] of breast cancer cases point to a predominant genetic contribution irrespective of the shared environmental factors. In support of this premise, a study based on identical (monozygotic) and non-identical (dizygotic) twins was conducted under the assumption that identical twins share the genetic and common environmental, while non-identical twins share only the environmental, components. These findings were based on 47,788 pairs of twins from Sweden, Denmark and Finland and contributed to the current understanding on the role of health, lifestyle and environmental factors as the major contributors to risk of breast cancer. It is estimated that up to 30% of the risk associated with breast cancer is from heritable factors[5]. Therefore, to understand the genetic architecture of breast cancer, several approaches, including linkage analysis and genetic association study designs, have been adopted to address breast cancer heritability in populations.

1.2.1. Genetic linkage analysis

The initial searches for genetic risk factors based on families with multiple individuals affected with breast cancer using linkage analysis[7] were successful in identification of high and moderate penetrance[c] variants (explained in detail in ensuing text). Linkage

[b] Familial clustering is the ratio of the risk of breast cancer for a relative of an affected individual compared to the general population

[c] Penetrance measures the proportion of individuals in a population who carry a specific allele and express the related trait.

analysis is a powerful tool to identify disease gene(s) since genes that physically reside in nearby locations on a chromosome are likely to co-segregate during meiosis, an indication that they are linked. If a disease gene is in linkage with known marker genes in the locus, the affected individual is likely to pass the disease gene to the offspring who inherit the marker. Based on the patterns of segregation, disease loci can be mapped. However, this approach requires large numbers of families with multiple affected individuals. The linkage between two loci can be estimated using a statistical approach of comparing the probability of two loci being linked versus not being linked. The estimated score is called the logarithm (base 10) of odds (LOD) score. Positive and negative LOD scores indicate the presence and absence of linkages,[8] respectively, and explain a proportion of the genetic risk associated with breast cancer.

(i) High penetrance mutations

The strong familial clustering seen among breast cancer cases was explained in part by single alleles conferring high risk. These high-risk variants are extremely rare but confer high penetrance. Family based linkage studies based on high-risk breast cancer cases have led to the discovery of disease loci which helped identify tumor suppressor genes, i.e., **BR**east **CA**ncer genes (*BRCA1*, in the year 1993[7,9,10] and BRCA2, in the year 1994)[11], with the odds ratios ranging from ~10 to 20. These findings also led to the discovery that women harboring mutations in *BRCA1* and/or *BRCA2* are predisposed to ovarian cancer. While the role of BRCA genes is acknowledged in conferring familial risk, these genes explained only a small portion of heritable component: 52% of the breast cancer patients with multiple affected family members carried *BRCA1* mutations, 32% carried *BRCA2* mutations and patients with breast and ovarian cancers carried either *BRCA1* (84%) or *BRCA2* (14%) gene mutations.

The *BRCA1* gene is located on chr17q21 with 24 exons (including two non-translating exons) and encodes a protein of 1863 amino acids. *BRCA1*, now a known tumor suppressor gene, plays a role in cell cycle and DNA damage repair. *BRCA2* is located on chr13q12 with 27 exons (including one non-translating exon) and encodes a protein of 3418 amino acids. BRCA2 binds with BRCA1 in response to DNA damage and aids in repair. The functional mutations are often small deletions or insertions, of which 85% are frameshift or nonsense mutations leading to translation of truncated proteins[12]. The frequency of these mutations[d] is extremely rare, and the frequency and mutational sites vary by population. For instance, in the Ashkenazi Jewish population, the mutational hot spots are at 185delAG at frequency of 1.09%[13] and 5382insC at frequency of 0.13%[13] in *BRCA1*[14] and at 6174delT at frequency of 1.52% in *BRCA2*, whereas in high risk Swedish families, *BRCA1* mutations are often at 3171ins5. The lifetime risk of breast cancer among carriers of these mutations varies from 60-80%[9,16].

For the ease of discussion, I refer to familial breast cancers as those affected individuals with a family history of breast cancer but without any known gene mutations or specific patterns of inheritance[6]. On the other hand, hereditary breast cancers are those in which familial clustering has been ascribed to gene mutations, often high penetrance (*e.g.*, BRCA1/2) with clear patterns of inheritance[18]. Both hereditary and familial forms of breast cancers tend to occur with early age of onset. The emphasis in this thesis is on

[d] Mutation is a small change in the DNA sequence and frequency <1% in the population

breast cancers with late age at onset (>45 years) and with no family history, cases which are often mentioned in the literature as sporadic breast cancers[6,19]. Sporadic breast cancers comprise 80% or more of all breast cancers diagnosed. There is paucity of literature in terms of the genetic basis for sporadic breast cancers, which has been addressed recently by adopting a population-based case-control design and identifying common low penetrant variants[19,20] (see ensuing text for more in depth discussion).

The pathology of hereditary breast cancers among *BRCA1*/2 carriers is different compared to that of non-*BRCA1/2* familial breast cancers or sporadic breast cancers. Breast cancers among *BRCA1* carriers are often "basal-like" tumors[21] with high grades, high mitotic rates, and receptors including estrogen (ER), progesterone (PR) and HER2 are negative[22,23]. Expression of basal markers including basal keratins[24], P-cadherin and epidermal growth factor receptor and over-expression of cell-cycle proteins including cyclins A, B1 and E, and S-phase kinase-associated protein 2 are frequent[25]. On the other hand, breast cancers in *BRCA2* carriers are rarely basal-like tumors[22], but are of high grade and are ER/PR positive [26,27]. Also, higher expression of cell cycle proteins such as cyclin D1 and p27 is noted. Overall, non-*BRCA* related tumors are less aggressive with low grade and mitotic counts compared to *BRCA1/2* positive tumors [25].

There are other high penetrance mutations associated with breast cancer in genes including *TP53, PTEN, STK11/LKB1* and*CDH1*. These mutations are rare in populations and confer about two to ten-fold increased risk for breast cancer (Table 1.1).

TP53 is a tumor suppressor gene in which mutations confer Li-Fraumeni syndrome in children and adults. About 5% of *TP53* mutation carriers diagnosed with breast cancer before age of 30 [28]. Compared to the general population, the mutation carriers have an 18 to 60-fold increased risk for early age of onset breast cancer <45 years old [29-32]. The lifetime cancer risk for individuals with a mutation in *TP53* is more than 90% and breast cancer is the most frequent cancer.

PTEN is a tumor suppressor gene in which mutations confer Cowden syndrome. The disease is characterized by multiple hamartomas (normally benign tumors in tissue of origin), but with high risk of both benign and malignant tumors in thyroid, breast and endometrium. The lifetime risk for developing breast cancer among *PTEN* carriers is about 50%.

STK11/LKB1 is a tumor suppressor gene with a role in apoptosis and the cell cycle. Mutational carriers are at risk for developing Peutz-Jeghers Syndrome, characterized by mucocutaneous pigmentation and hamartomatous polyps[33] and there is also an increased risk for developing cancers of breast, lung, ovary, cervix, testis, pancreas, and/or the gastrointestinal tract including esophagus, stomach, small bowel and colon[34]. The lifetime risk for developing any of these cancers among *STK11/LKB1* mutation carriers is about 85%[35].

CDH belongs to the E-cadherin family of genes and is a calcium dependent cell-cell adhesion molecule expressed in epithelial cell junctions. Often mutational carriers are at risk of developing diffuse gastric carcinoma and have increased risk of developing lobular breast and colon cancers. About 40-54% of the women carriers develop breast cancer during their lifetime [36].

(ii) Moderate penetrance variants

Family based linkage studies failed to identify additional highly or moderately penetrant variants and that has led to alternate approaches to address the heritability of breast cancers. Candidate genes chosen for investigation were based on their cellular functions in studies that were conducted among familial breast cancer cases. This approach has successfully identified moderately penetrant variants (confers about two-fold increased risk) in genes such as *CHEK2*[37], *ATM*[38], *PALB2*[39], and *BRIP1 (BACH1)*[40]. The proteins encoded by these genes play a role in DNA repair by interacting with *BRCA* pathways.

CHEK2 encodes for the protein kinase that regulates the G2 phase of the cell cycle in response to DNA damage. CHEK2 gets phosphorylated to become the active form, which stabilizes TP53 and interacts with BRCA1. *CHEK2*1100delC* is the most commonly seen mutation (up to 1-2%) in the general population and up to 5% among individuals with familial or hereditary breast cancers. It confers a two-fold increase among female carriers and a ten-fold increase among male carriers for breast cancer risk[37]. There are additional rare mutations in *CHEK2,* identified in the Ashkenazi Jewish population, that are suggestive of a founder effect[41]. However, there is no additional risk among co-carriers of mutations in *BRCA* and *CHEK2*, suggestive of an overlapping effect in the DNA repair pathways[37].

ATM encodes a protein kinase that was shown to play a role in repair of double stranded breaks in DNA, and in regulation of *BRCA1* and *CHEK2*. Impaired regulation of DNA repair pathways increases the risk of developing cancers. Biallelic mutations in *ATM* causes the autosomal recessive disease, ataxia telangiectasia, and such homozygous variants confer susceptibility to breast cancer with a relative risk 2.3-fold higher than that of women in the general population.

BRIP1 protein interacts with the C-Terminus (BRCT) domain of BRCA1. Mutations in *BRIP1* are rare (< 1%) among breast cancer cases, and the majority lead to formation of truncated proteins. Biallelic mutations in *BRIP1* are associated with Fanconi anemia. It is estimated that there is a two-fold higher relative risk for early onset breast cancer among the mutational carriers with strong family history.

PALB2 encodes a protein that interacts with BRCA2.The relative risk for breast cancer among women < 50 years with PALP2 mutations is three-fold higher[39,42]. Biallelic mutations cause Fanconi anemia type N and a higher incidence of childhood cancers. Also, a higher incidence of male breast cancer is associated with *PALB2* mutations[43], however it only accounts for a minority of familial breast cancer cases.

Table 1.1 High and moderately penetrant variants associated with breast cancer

Locus	Genes	RAF	Relative risk	Familial relative risk	Breast cancer incidence
High penetrance variants					
17q21	*BRCA1*	0.0006	5-45	10%	82% lifetime risk
13q12.3	*BRCA2*	0.001	9-21	12%	
17p13.1	*TP53*	rare	2-10	ND	25% by age 74
10q23.3	*PTEN*	rare	2-10	ND	85% lifetime risk
19p13.3	*STK11*	rare	2-10	ND	32% by age 60
16q22.1	*CDH1*	rare	2-10	ND	39% lifetime risk of lobular breast cancer
Moderate penetrance variants					
11q22.3	*ATM*	0.003	2-3	5%	ND
22q12.1	*CHEK2*	0.004	2-3		
17q22-q24	*BRIP1*	0.001	2-3		
16p12.1	*PALB2*	rare	2-4		

ND, not determined; RAF, risk allele frequency

Relative risk is the ratio of the probability of event (breast cancer) occurring in an exposed group (mutation carriers) to the probability of event (breast cancer) occurring in an unexposed group (non-mutation carriers

Familial relative risk is the relative risk of breast cancer incidence within the families of breast cancer affected individuals

1.2.2. Common Disease-Common Variant hypothesis:

Family-based linkage studies and identification of loci associated with breast cancer together with subsequent sequencing studies have led to discovery of high penetrance variants (*e.g.*, *BRCA1/2* mutations). However, these attempts explained only a proportion of the heritability associated with familial breast cancer. Most of the unexplained risk among familial cases was thought to be explained by a polygenic model of inheritance, in which multiple low penetrance variants (>5% frequency) contribute to the phenotype[44]. However, because the majority of breast cancer cases are sporadic (*i.e.*, no family history of breast cancer), linkage studies are not feasible. Their sporadic nature implies that breast cancers, and other commonly occurring sporadic diseases, have a different genetic architecture. This premise has led to the hypothesis of Common Disease-Common Variants (CDCV)[45-48], which states that common genetic variations (frequency more than 5%) in a population contribute to a small but finite risk to explain genetic susceptibility. The Human Genome Project Consortium[49] efforts led to the current understanding that up to 99.9% of all human populations share a similar DNA sequence, and yet small genetic variations of 0.1% could still account for large phenotypic variations, lending credence to the CDCV hypothesis.

1.2.3. Genetic association studies

In family-based linkage studies genetic loci are mapped based on their segregation with phenotypes within pedigrees. In contrast, genetic association aims to detect variants associated with phenotypes based on family or population-based study designs. The two commonly adopted association study approaches are (i) candidate gene associations, and (ii) genome-wide association study (GWAS) designs. In both approaches, the frequencies of genetic variants are compared between cases and controls using a statistical test. In family-based linkage studies, microsatellite markers are more commonly used, however these have limitations, *i.e.*, microsatellites are fewer in number (~4000), meaning less dense, and therefore the resolution of mapping of loci is lower. Microsatellites are also unstable because they are mutable. Currently high resolution genetic mapping is feasible using polymorphisms (see below) whose densities in the genome are several orders of magnitude higher than microsatellite markers. The three main classes of DNA variations include single-base-pair variants or *Single Nucleotide* Polymorphisms[e] *(SNPs)*, insertions/deletions[f] and structural variants[g] (including copy number variations, or CNVs). Due to their high densities across the genome and their stability, being evolutionarily conserved across populations, SNPs and CNVs are the preferred genetic markers for association studies.

(i) Candidate gene association studies

Candidate gene association studies aim to identify common variants (those polymorphisms with allele frequencies >5%) present within the select candidate gene(s) or flanking regions (5' and 3' untranslated regions of the gene in question) that may affect functions (*e.g.,* translational efficiency, splicing, gene regulation) and thereby

[e] SNPs are single base-pair changes in the DNA sequence that occur at frequencies of more than 1% in the general population.

[f] Insertion or deletion of a single stretch of DNA sequence, from two to hundreds of base-pairs in length.

[g] Structural changes in the DNA sequence, including copy number variations (CNVs) and chromosomal rearrangements.

confer a phenotype. Candidate gene studies are hypothesis driven and focus on genes with known cellular functions (*e.g.*, DNA repair, apoptosis, cell cycle), investigating the role of common variants in conferring breast cancer risk. Although several candidate gene association studies have been conducted to identify variants associated with breast cancer[50-56], only one, a study of *CASP8*, has successfully identified a SNP rs1045485[48] in the coding region that confers risk for breast cancer. This finding was replicated by several independent studies[57]. Despite several decades of effort, candidate gene association studies have been largely unsuccessful in identifying additional breast cancer risk variants[58]. Inherent limitations of candidate SNP association studies include inadequate study design power (small sample size), selection of SNPs, genotyping errors, sampling bias and population stratification leading to failures in the replication of the findings. Therefore, technological and methodological developments were needed to design well powered studies, including access to large numbers of cases and controls and a sound statistical framework.

(ii) Genome-wide association studies (GWASs)

GWASs offer a systematic and unbiased approach for genome wide screening for common variants. GWASs are hypothesis-free wherein the entire genome is screened for variants associated with the phenotype being investigated, followed by multiple replication stages. GWASs embrace the CDCV hypothesis to identify multiple common variants (albeit, with low effect sizes[h]). Such studies utilize large sample sizes, providing

[h] Effect size is the quantitative measure of the magnitude of risk or a phenomenon. Odds Ratio and Relative risk are measure of effect size. In this thesis effect sizes and odds ratios are used interchangeably.

statistical power, and reliable high throughput genotyping platforms. Further, the association statistics are adjusted for multiple marker hypothesis testing by various methods[59] (Bonferroni correction, Benjamini and Hochberg false discovery rate correction[60]) to limit false positive associations. Human Genome, HapMap and 1000 Genomes projects helped catalogue variants with their allele frequencies, with estimates of linkage disequilibrium (LD) in diverse populations. Patterns of inheritance are guided by the process of meiotic recombination events. Large chunks (referred to as LD blocks) of the genome are stably inherited from parents by off spring, and these patterns are highly specific for individual ancestries. Typically, LD blocks vary in size from 1 kilobase (kb) to 100 kb. SNPs present within LD blocks are highly correlated, and high throughput platforms have evolved to reduce the redundancy of genotyping based on the LD patterns. The technique of selecting fewer SNPs representing an entire LD block is termed tagging, and the resulting SNPs are called tagSNPs. This approach is cost-efficient, making genome wide coverage of markers feasible for mapping of associated genes/loci.

1.3. Success stories of breast cancer GWASs

Over the last decade, GWAS approaches were widely utilized to identify the common genetic variants associated with breast cancer. To date GWASs have identified about 172 risk variants with effect sizes (odds ratio, OR) 1.04 to 1.53 and explain ~18% of the total heritable risk associated with breast cancer[61]. Genotyping platforms have evolved over the last decade, and consortia efforts are now more predominant in the study designs to enable large sample sizes, and statistical power to detect variants even with modest effect size. Tools for detecting population stratification, data analytics and strategies to identify causal variants have also contributed to the overall success of GWASs. The timeline of GWASs along with the approaches adopted can be divided into three groups:

(i) Early era (2007-2013)

The first breast cancer GWASs were published in 2007 by Easton et al.[62] and Hunter etal.[19] and subsequently additional studies were published reporting novel findings and replication of previously reported variants. During this early era, the studies typically utilized whole genome genotyping platforms from a small number of samples (also called the discovery stage, ~300-500 each of cases and controls), and highly statistically significant SNPs from the discovery stage were further replicated in larger sample sizes (independent replication stages). I compiled the following data from the catalogue of GWAS variants[63] for the putative breast cancer susceptibility loci. This catalogue adopted a cut-off p-value $<10^{-5}$. A total of 17 breast cancer GWASs were attempted during the early phase on Caucasian populations, including one from the Damaraju laboratory[20]. Of the reported novel loci, 21 had effect sizes >1.20, 36 had effect sizes between 1.1-1.9, and 42 had effect sizes between 1.04-1.1. A total of 13 GWASs were published in non-Caucasian populations, including three studies each in Chinese [64-66] and Japanese[67-69], two each in Ashkenazi Jewish[70] [71] and African[72] and one in Korean[73] populations. In total, 33 novel loci were identified from diverse populations, of which 22 SNPs had effect sizes > 1.2, and the remaining 11 SNPs had effect sizes of 1.08-1.9. Table 1.2 below summarizes the studies on breast cancer associated SNPs with effect sizes >1.2 for both Caucasian and other ancestries.

Table 1.2 GWASs reported for breast cancer risk between 2007 to 2013

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
1	Hunter DJ et al. (2007)[19] A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.	1,145 European ancestry cases, 1,142 European ancestry controls	874 European ancestry cases, 1,478 European ancestry controls, 302 cases, 594 controls	10q26.13	FGFR2	rs1219648-G	0.4	1.00E-10	1.2 [1.07-1.42]
2	Stacey et al. (2007)[74] Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.	1,599 European ancestry cases, 11,546 European ancestry controls	2,954 European ancestry cases, 5,967 European ancestry controls, Up to 561 Japanese ancestry cases, Up to 565 Japanese ancestry control, Up to 422 African American cases, Up to	2q35	intergenic	rs13387042-A	0.5	1.00E-13	1.2 [1.14-1.26]
				16q12.1	TNRC9	rs3803662-T	0.27	6.00E-19	1.28 [1.21-1.35]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
			448 African American controls, Up to 418 Hispanic cases, Up to 422 Hispanic controls, Up to 148 cases, Up to 293 controls						
3	Easton DF et al. (2007)[62] Genome-wide association study identifies novel breast cancer susceptibility loci.	390 European ancestry cases, 364 European ancestry controls	4,364 East Asian ancestry cases, 24,174 European ancestry controls, 3,564 East Asian ancestry controls, 24,391 European ancestry controls	10q26.13	FGFR2	rs2981582-A	0.38	2.00E-76	1.26 [1.23-1.30]
				16q12.1	TNRC9, LOC643714	rs3803662-T	0.25	1.00E-36	1.2 [1.16-1.24]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
4	Thomas G et al. (2009)[75] A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1).	1,145 European ancestry cases, 1,142 European ancestry controls	8,625 European ancestry cases, 9,657 European ancestry controls	5q11.2	MAP3K1	rs16886165-G	0.15	5.00E-07	1.23 [1.12-1.35] (Het)
				2q35	intergenic	rs13387042-A	0.51	2.00E-08	1.25 [1.15-1.37] (Het)
5	Turnbull C et al. (2010)[76] Genome-wide association study identifies five new breast cancer susceptibility loci.	3,659 European ancestry cases, 4,897 European ancestry controls	12,576 European ancestry cases, 12,223 European ancestry controls	10q26.13	FGFR2	rs2981579-A	0.42	4.00E-31	1.43 [1.35-1.53]
				16q12.1	TOX3	rs3803662-A	0.26	3.00E-15	1.3 [1.22-1.39]
				5q11.2	MAP3K1	rs889312-C	0.28	5.00E-09	1.22 [1.14-1.30]
				2q35	intergenic	rs13387042-A	0.49	2.00E-10	1.21 [1.14-1.29]
				6q25.1	ESR1, C6orf97	rs3757318-A	0.07	3.00E-06	1.3 [1.17-1.46]
6	Antoniou AC et al. (2010)[77] A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the	1,193 European ancestry cases, 1,190 European ancestry controls	2,974 European ancestry cases, 3,012 European ancestry controls	19p13.11	ANKLE, C19orf6, ABHD8	rs8170-A	0.17	2.00E-09	1.26 [1.17-1.35]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
	general population.								
7	Fletcher O et al. (2011)[78] Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study.	2,839 European ancestry cases, 3,507 European ancestry controls	9,041 European ancestry cases, 8,980 European ancestry controls	10q26.13	FGFR2	rs1219648-?	0.42	1.00E-30	1.31 [1.25-1.37]
8	Li J et al. (2010)[79] A combined analysis of genome-wide association studies in breast cancer.	2,702 European ancestry female cases, 5,726 European ancestry controls	Up to 7,386 cases, 7,576 controls	10q26.13	FGFR2	rs1219648-G	0.42	2.00E-13	1.32 [1.22-1.42]
				16q12.1	TOX3	rs3803662-A	0.3	4.00E-07	1.22 [1.13-1.32]
				5p12	MRPS30	rs7716600-A	0.23	7.00E-07	1.24 [1.14-1.34]
9	Sehrawat B et al. (2011)[20] Potential novel candidate polymorphisms identified in genome-wide	302 European ancestry female cases, 321	1,153 European ancestry female cases, 1,215 European ancestry	5p15.2	ROPN1L	rs1092913	0.13	2.00E-06	1.45 [1.24-1.69]
				19q13.41	ZNF577	rs10411161	0.13	7.00E-07	1.42 [1.22-1.65]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
	association study for breast cancer susceptibility.	European ancestry female controls	female controls						
10	Siddiq A et al. (2012)[80] A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11.	3,666 European ancestry cases, 28,864 European ancestry controls, 1,004 African American cases, 2,744 African American controls	562 European ancestry cases, 6,410 European ancestry controls, 84 Japanese ancestry cases, 830 Japanese ancestry controls, 300 Latino cases, 1,164 Latino controls	6q25.1	-	rs9383938	-	2.00E-10	1.28
11	Orr N et al. (2012)[81] Genome-wide association study identifies a common variant in RAD51B associated with	823 European ancestry cases, 2,795 European ancestry	438 European ancestry cases, 474 European ancestry controls	1p31.1	PRKACB	rs903263	-	1.00E-06	1.27 [1.10-1.34]
				14q24.1	RAD51B	rs1314913	-	3.00E-13	1.57 [1.39-1.77]
				16q12.1	LOC64374, TOX3	rs3803662	-	4.00E-15	1.5 [1.35-1.66]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
	male breast cancer risk.	controls							
12	Garcia-Closas M et al. (2013)[82] Genome-wide association studies identify four ER negative-specific breast cancer risk loci.	4,193 European ancestry cases, 35,194 European ancestry controls	6,514 European ancestry cases, 41,455 European ancestry controls	12p11.22	PTHLH	rs10771399	0.89	2.00E-12	1.2 [1.15-1.27]
				13q13.1	BRCA2, N4BP2L1	rs11571833	0.5	6.00E-07	1.52 [1.31-1.77]
13	Michailidou K (2013)[83] Large-scale genotyping identifies 41 new loci associated with breast cancer risk.	10,052 European ancestry cases, 12,575 European ancestry controls	45,290 European ancestry cases, 41,880 European ancestry controls	13q13.1	BRCA2, N4BP2L, N4BP2L2	rs11571833	0.008	5.00E-08	1.26 [1.14-1.39]
				10q26.13	FGFR2	rs2981579	0.4	2.00E-170	1.27 [1.24-1.29]
				11q13.3	intergenic	rs614367	0.15	2.00E-63	1.21 [1.18-1.24]
				16q12.1	TOX3	rs3803662	0.26	2.00E-114	1.24 [1.21-1.27]
				10p12.31	DNAJC1	rs11814448	0.02	9.00E-16	1.26 [1.18-1.35]
14	Purrington KS (2013)[84] Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for	1,529 European ancestry cases, 3,399 European ancestry controls	2,148 European ancestry cases, 1,309 European ancestry controls	5p15.33	TERT	rs10069690	-	1.00E-07	1.24 [1.14-1.34]
				6q25.1	ESR1	rs3757318	-	9.00E-07	1.33 [1.17-1.51]
				19p13.11	intergenic	rs2363956	-	2.00E-08	1.22 [1.14-1.3]
				12p11.22	PTHLH	rs10771399		2.00E-08	1.39 [1.25-1.56]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
	triple-negative breast cancer.								
15	Gold B et al. (2008)[70] Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33.	249 Ashkenazi Jewish non-BRCA1/2 carriers cases, 299 Ashkenazi Jewish non-BRCA1/2 carriers controls	1,193 Ashkenazi Jewish non-BRCA1/2 carriers cases, 1,166 Ashkenazi Jewish non-BRCA1/2 carriers controls	6q22.33	ECHDC, RNF146	rs2180341	0.21	3.00E-08	1.41 [1.25-1.59]
16	Zheng et al. (2009)[64] Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1.	1,505 Chinese ancestry cases, 1,522 Chinese ancestry controls	5,026 Chinese ancestry cases, 2,476 Chinese ancestry controls, 1,591 European ancestry cases, 1,466 European ancestry	6q25.1	ESR1, C6orf97	rs2046210	0.37	2.00E-15	1.29 [1.21-1.37]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
			controls						
17	Long et al. (2010)[65] Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium.	2,073 Chinese ancestry cases, 2,084 Chinese ancestry controls	15,159 East Asian ancestry cases, 12,993 East Asian ancestry controls, 2,797 European ancestry cases, 2,662 European ancestry controls	16q12.1	TOX3	rs4784227	0.24	1.00E-28	1.24 [1.20-1.29]
18	Shu Xo et al. (2012)[85] Novel genetic markers of breast cancer survival identified by a genome-wide association study.	1,950 Chinese ancestry cases	4,160 Chinese ancestry cases	14q24.1	RAD51L1	rs3784099	-	1.00E-07	1.49 [1.28-1.72]
				14q24.1	RAD51L1	rs3784099	-	3.00E-07	1.43 [1.25-1.64]
				16q22.3	intergenic	rs9934948	-	6.00E-06	1.29 [1.16-1.44]
19	Kim HC et al. (2012)[86] A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34:	2,273 Korean ancestry cases, 2,052 Korean ancestry controls	4,049 Korean ancestry cases, 3,845 Korean ancestry controls	2q34	ERBB4	rs13393577	0.05	9.00E-14	1.53 [1.37-1.70]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
	results from the Seoul Breast Cancer Study.								
20	Elgazzar S et al. (2012)[68] A genome-wide association study identifies a genetic variant in the SIAH2 locus associated with hormonal receptor-positive breast cancer in Japanese.	1,086 Japanese ancestry cases, 1,816 Japanese ancestry controls	1,653 Japanese ancestry cases, 2,797 Japanese ancestry controls	3q25.1	SIAH2	rs6788895	0.65	9.00E-08	1.22 [1.13-1.31]
				10q26.13	FGFR2	rs3750817	0.49	8.00E-08	1.22
21	Rinella ES et al. (2013)[71] Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable BRCA1/2 mutation.	477 Ashkenazi Jewish cases, 524 Ashkenazi Jewish controls	203 Ashkenazi Jewish cases, 263 Ashkenazi Jewish controls	10q26.13	FGFR2	rs1078806	0.39	2.00E-06	1.43
				6p22.3	intergenic	rs16882214	0.81	2.00E-06	1.43
				15q24.3	intergenic	rs12906542	0.93	7.00E-07	2
22	Song C et al. (2013)[87] A genome-wide scan for breast	3,016 African American cases,	NA	10q22.3		rs12355688	0.22	6.00E-06	1.24 [1.13-1.36]
			NA	1p36.23	SLC45A1	rs2305016; rs7535752;	-	5.00E-06	1.23 [1.12-1.35]

#	Study	Initial Sample Size	Replication Sample Size	Region	Gene	Risk SNP-allele	RAF	P-value	OR [95% CI]
	cancer risk haplotypes among African American women.	2,745 African American controls				rs9628987; rs12711517; rs2289731			
			NA	4q27	TNIP3	rs17435444; rs13116936	0.64	3.00E-07	1.23 [1.13-1.33]
			NA	10p15.1		rs4414128; rs2386661; rs17141741	-	5.00E-06	1.27 [1.14-1.39]
			NA	14q24.1		rs765899; rs757369; rs10132579; rs2842347; rs737387; rs2842346	-	2.00E-06	1.67 [1.35-2.08]
23	Low SK (2013)[69] Genome-wide association study of breast cancer in the Japanese population.	2,642 Japanese ancestry cases, 2,099 Japanese ancestry controls	2,885 Japanese ancestry cases, 3,395 Japanese ancestry controls	10q26.13	FGFR2	rs2981578	0.51	1.00E-12	1.23 [1.15-1.29]
				16q12.2	TOX3, LOC643714	rs12922061	0.24	4.00E-10	1.23 [1.15-1.31]
				16q12.1	TOX3, LOC643714	rs3803662	0.52	3.00E-11	1.21 [1.15-1.28]
				12p13.1	ATF7IP	rs17221259	0.20	7.00E-06	1.25 [1.14-1.38]

The above table represents the GWASs published between 2007 to 2013 that reported one or more variants with OR ≥1.2 in both Caucasian and non-Caucasian population. RAF -Risk Allele Frequency

(ii) Collaborative Oncologic Gene-environment Study (COGS) Era (2012- 2015)

GWASs reported 27 common variants in the early era and accounted for ~9% of estimated breast cancer risk. To identify additional genomic variants, a collaborative effort led by the COGS consortium (http://www.cogseu.org/) focused on identification of gene-environment interactions contributing to the risk of breast, prostate and ovarian cancers. A custom panel of genotyping array was designed with ~200,000 SNPs and used an Illumina genotyping platform (called as iCOG array). In 2013, an association study using the iCOG array was reported by Michailidou et al.[88]. The study utilized 45,290 breast cancer cases and 41,880 controls of European ancestry (from 41 studies from the Breast Cancer Association Consortium (BCAC)). The study identified 41 new loci and replicated 27 previously identified breast cancer loci of various effect sizes. Of the 41 newly identified loci, 13 SNPs showed specific association with ER positive and one with ER negative breast cancers. Independent studies also utilized the iCOGS array with 4,193 ER negative cases and 35,194 controls from 40 BCAC studies to identify four SNPs associated with ER negative breast cancer[82]. Together, all reported GWASs and large-scale replication studies have identified 79 variants accounting for about 14% of heritability associated with familial breast cancer. In 2015, a meta-analysis[89] based on 11 previously published GWASs (15,748 breast cancer cases and 18,084 controls) and 41 BCAC studies (46,785 cases and 42,892 controls) using genotypes based on the iCOG array were performed and replicated 71 previously reported loci. Furthermore, imputation and excluding variants within 500 kb of the previously identified SNPs led to the identification of 15 additional new SNPs[89] associated with breast cancer. In summary, the total number of identified loci in these iCOGS attempts were 94 SNPs with a total estimated familial breast cancer heritability of 16%. Overall, the iCOGS array catalyzed utilization of samples from consortia and independent studies to effectively mine for additional risk variants that were otherwise missed due to inadequate sample size. It is also expected that higher sample sizes and mining the same genotype data sources will likely identify variants of lower effect size. In line with these expectations, 94 variants showed effect size of <1.2. There are potentially other variants from the above studies waiting to be discovered to account for the overall heritability of breast cancer beyond the 16%.

(iii) OncoArray era (2015-present)

The OncoArray Network[90], a collaborative effort to uncover the genetic architecture of breast, ovarian, prostrate, colorectal and lung cancers, used the iCOG array, a custom high-density array from Illumina, also known as the OncoArray BeadChip (~570,000 SNPs). The iCOG array includes ~260,000 tagSNPs, providing extensive coverage of common variants across the genome, and GWASs identified SNPs for each of the cancer types and SNPs from fine-mapping studies of previously identified loci. The OncoArray study reported in 2017 utilized 61,282 breast cancer cases and 45,494 controls of European ancestry which are part of the previously published reports from 68 studies, including BCAC and Discovery, Biology and Risk of Inherited Variants in Breast Cancer Consortium (DRIVE). The OncoArray study used the iCOG array for genotyping followed by subsequent imputation resulting in a total of 11.8 million SNPs (MAF > 0.5% and imputation quality score > 0.3). A meta-analysis combining the results from the above study and other previous studies based on iCOG arrays with 11 previously reported

GWASs were conducted. Together a total of 122,977 breast cancer cases and 105,974 controls of European ancestry and 14,068 breast cancer cases and 13,104 controls of East Asian ancestry were utilized. The meta-analysis reported the association of 65 new breast cancer risk loci with genome wide significance among European ancestry[61] of which 19 of the 65 SNPs were associated with ER-positive and two with ER negative breast cancers. A majority of the variants identified thus far are associated with ER positive breast cancer (Figure 1.1). Therefore, another study, which stratified the above cases based on ER status, identified ten additional variants associated with ER negative breast cancer[91]. This summarizes the massive data mining attempts by international consortia to identify all potential variants associated with breast cancer. However, the estimates from all 172 common SNPs/loci identified contribute to a heritable risk of ~18%, suggesting additional variants are yet to be discovered.

1.3.1. SNPs associated with pathological subtypes of breast cancer and BRCA

Following identification of loci associated with breast cancer risk, several subsequent studies investigated the association of these risk loci with histopathology of breast tumors including triple negative breast cancer[92-94] and the risk conferred by common variants among the *BRCA1*[77,95] and *BRCA2* mutation carriers[96]. However, detailed discussion into these topics is beyond the scope of the thesis and I have included the pertinent references for interested readers.

Figure 1.1 GWAS-identified variants associated with breast cancer based on estrogen receptor status

GWAS-identified breast cancer associated variants with respect to (a) ER positive and (b) ER negative breast cancers. The Y-axis indicates the effect size (odds ratio, OR) and the X- axis indicates the effect allele frequency (EAF). The figure is from Lilyquist et al (2018)[97]. The arrow indicates OR 1.2, to draw attention to the small number of SNPs with this effect size relative to all variants identified so far.

1.3.2. Post-GWAS era in breast cancer

In the post-GWAS era in breast cancer, GWAS designs for mapping disease associations were based on using SNPs across the genome (equidistant and dense representation of markers), rather than using SNPs with putative functional consequences as in candidate gene studies. The very premise of GWAS is based on LD patterns, and GWAS-identified SNPs are likely proxies for causal variants. Strategies to identify the causal variants underlying disease associations were sought through fine-mapping approaches. Interrogation of the catalogue of variants identified through GWASs of various phenotypes indicated that a large proportion of the SNPs (~88%) were in the intergenic (gene desert) or intronic regions[98]. The scenario is no different for breast cancer in that the challenge is to find putative biological functions for GWAS-identified variants. To date, a limited number of studies have performed fine-mapping of hits from GWASs, and the approaches and strategies used in fine-mapping are discussed elsewhere[99]. The overview of the steps post GWAS to gain functional insights of the loci so identified is depicted[100] in Figure 1.2.

Figure 1.2 Roadmap from GWAS to elucidation of functional relevance of disease associated loci

This figure illustrates the roadmap from association to functional characterization of a GWAS identified variant (a) Outline of GWAS study design identifying common variants associated with disease, (b) the linkage disequilibrium pattern of the associated region, (c-e) functional annotation indicating the state of the chromatin and binding of potential transcription factors in the associated loci and (f) different functional assays for validating the SNPs in predicted function. The figure is from Harismendy et al (2009)[100].

1.4. Fine-mapping approaches

1.4.1. Dense genotyping and imputation

The GWAS approach reveals associations of genomic loci with phenotypes. Since it greatly relies on tagSNPs, the GWAS-associated tagSNPs may not necessarily have direct functional consequences but may be in LD with potential causal SNPs[101]. Pair-wise correlations of SNPs in a region or "block," indicated by r^2 (in the range 0-1), with a value of 0 indicating no LD and 1 being in perfect LD, signifies that all SNPs in the block are correlated to varying degree. The size of the LD blocks varies in different ethnic groups, for instance LD blocks are larger in European populations (used interchangeably in this thesis as Caucasian populations) compared to African or Asian populations in which a LD block may have been broken down due to extensive meiotic recombination[102]. Therefore, fewer tagSNPs for each LD block are sufficient to provide coverage for populations of European ancestry, compared to populations of other ancestries[103]. Because GWAS-associated SNPs are often not directly linked to function, a successful fine-mapping approach is needed to identify the functional variants underlying the GWAS-associated signal. The first step in the fine-mapping approach is to capture all the variants that are correlated with the GWAS-associated tagSNPs.

(i) Targeted sequencing

The initial approaches utilized targeted sequencing of the GWAS-associated locus in a limited number of subjects, ensuring identification of all variants that could have been associated[104]. However, sequencing small numbers of samples detects common variants whereas sequencing large numbers of samples is required to detect associations with rare variants in the loci, making targeted sequencing a technically challenging and expensive approach.

(ii) High density Arrays

The 1000 Genomes project, which has comprehensively sequenced the DNA of 1092 subjects of different ethnic groups, sufficiently captured and catalogued the variants with minor allele frequencies >1%[105]. The collaborative effort of the consortia (Wellcome Trust Case Control Consortium (WTCCC), Genetic Investigation of Anthropometric Traits, and BCAC) put forth their common interest in developing high density genotyping chips such as Immunochip[106], Metabochip[107] and iCOGs array[88] enabling fine-scale mapping of GWAS loci based on an array design for affordable genotyping in larger cohorts. The consortia's efforts are thus to genotype large numbers of samples, with increased power to detect association signals. Since the array designs are based on selected SNPs, the coverage and density of SNPs on the genome are biased towards previously identified loci for fine-mapping[108]. These array techniques helped identify a limited number of causal variants although the array design limits the numbers of SNPs selected in a region, unlike imputation-based approaches. However, custom arrays also helped to reconfirm the originally reported associations (index SNPs), and the array based fine-mapping yielded additional variants conferring breast cancer risk which also showed genome wide significance, but the effect sizes were low[109]. Even though large sample sizes are available through the consortia, the cost associated with genotyping and the inability of the genotyping arrays to capture all SNPs in a given LD block has led to different strategies for fine-mapping of disease associated loci.

(iii) Imputation

Imputation is a statistical technique which is used to estimate the genotype probabilities of ungenotyped SNPs for a subject. Imputation relies on the concept of LD and the high correlation between SNP genotypes. As the array-based platforms mostly use tagSNPs and not all SNPs in the LD block are captured, imputation is a way of finding the missing genotypes. The imputation algorithms utilize a reference genome panel (1000 Genomes Panel) to predict the missing genotypes. Imputation has advantages over other methods (described above) for fine-mapping of associated disease loci and in identifying putative causal variants. Imputation is now widely used in fine-mapping studies. The two commonly used imputation algorithms are IMPUTE2[110] and MACH[111]. In the work presented in this thesis, the IMPUTE2 algorithm was used to predict the missing genotypes. The different steps in imputation analysis are discussed below for IMPUTE2, an algorithm that has optimal performance when used in combination with the 1000 Genomes Panel as the reference dataset[112].

Steps in imputation:

a. Pre-processing

This step involves quality control for the data at the sample and genotype levels. The sample level quality controls include call rate filtering, heterozygosity, and relatedness between genotyped individuals. The genotype level quality controls include call rates, Hardy-Weinberg Equilibrium and exclude SNPs with low minor allele frequencies. I used National Center for Biotechnology Information (NCBI) genome build 37 (hg19) for all genomic annotations in this thesis. I aligned the genotype data to the same strand convention as the reference panel. Often the SNP probes in the genotyping array are optimized either for the positive or negative strand. However, the reference genome is always aligned with the positive strand and GTOOL can align the study genotypes from the negative strand to the positive stand. The genotype output file from GTOOL is saved for each individual chromosome in “GEN file” format and a “Sample file” which has sample identifiers.

b. Pre-phasing

This step reduces the computational burden by calling the haplotypes prior to imputation. IMPUTE2 and SHAPEIT algorithms estimate the phased haplotypes as input genotypes.

c. Imputation

Imputation is the process of filling the missing genotypes in the input phased haplotypes based on the reference panel haplotypes. The imputation algorithm handles small chunks of data at a time (5 Mb), however it is not necessary to physically split the chromosome, instead the IMPUTE2[110] algorithm can take arguments defining the 5-Mb chunks. The output from the imputation is provided as a probability of individual genotypes along the physical length of the chromosome. The quality control metric for imputation is indicated as a concordance table that captures the estimates of concordance between the genotyped and imputed SNPs using one-fold cross validation. Detailed usage of the IMPUTE2 algorithm is described in https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#home and https://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook#Imputation. Since whole genome imputation is computationally intense, the analysis should be performed using a high-performance computing cluster (*e.g.*, Compute Canada Server, https://www.computecanada.ca/).

1.4.2. Fine-mapping based on LD patterns

As discussed earlier, GWASs and fine-mapping greatly rely on LD patterns. However, LD patterns vary across ancestral populations. This approach is also referred to in the literature as cross-ethnic mapping. Most GWASs have been performed in the European population wherein larger LD block patterns are common, and several SNPs in the fine-mapped regions correlate with GWAS-identified SNPs, making it challenging to identify putative causal variants. To address this problem, cross ethnic mapping has been adopted in the literature [113] wherein associations are tested in different ethnic groups, usually African or Asian populations. Often the original GWAS-identified SNP may show potential associations in these diverse populations, although the size of LD blocks may vary (and smaller LD blocks are more informative) in these populations relative to those of European ancestry. Thus, the index GWAS SNP may now be confined to a smaller LD block or putative causal SNPs may be in a different LD block and with fewer correlated SNPs (of equal or higher statistical significance in the association tests). This approach narrows the region in which the putative causal allele resides[103]. Even though this approach is logical and appears simple, the underlying assumptions are that the GWAS-identified SNP also shows statistically significant associations in other populations and that finer LD block patterns need to be discernable across diverse populations for the given genomic locus of interest.

1.4.3. Conditional regression to identify independent peaks of association

Fine-mapping across a locus of interest (> 100kb long) may yield several independent peaks of association flanking a GWAS-identified SNP with several correlated SNPs within each peak. Such sets of correlated SNPs are also termed "independent Correlated Highly trait-Associated Variants" (iCHAVs)[99,114]. The initial step in identification of casual variants and exclusion of non-causal variants is to determine the number of iCHAV peaks, which can typically be done using forward conditional logistic regression. Conditional regression, extension of the logistic regression method, in this context, the regression analysis is subjected to conditioning based on top associated SNP to identify independently associated SNPs. In the fine-mapping of the 11q13[114] breast cancer loci identified multiple iCHAVs by adopting conditional regression analysis and revealed stronger signals compared to the original GWAS SNP.

1.4.4. Likelihood ratio analysis to identify potential causal variants among the associated SNPs

Likelihood ratio test is a statistical test to exclude non-causal variants present within each independent association peak. Comparing the risk of each variant with that of the strongly associated potential causal variants within the iCHAV allows exclusion of variants with likelihood ratios >100[109]. This method is informative provided the sample size is adequate, and the above statistical methods have reduced the number of highly correlated SNPs. Nonetheless, it is still possible to end up with several potential causal variants that may require independent data pruning strategies[104,114,115]. An additional benefit of fine-mapping is that SNPs from multiple iCHAVs (each with a finite risk) may explain larger proportions of heritability than estimated from the original GWAS[116]. However, each of the iCHAV SNPs may regulate target genes by independent mechanisms[88,114]. While statistical approaches eliminate the less likely causal SNPs, the challenges are in elucidating biological functions for each causal variant. For determining the functions of non-coding variants, there are an array of computational approaches, databases and online resources discussed in detail in the following section.

1.4.5. Functional annotation

Most GWAS-identified variants are in the non-coding regions of the genome. There are several steps in elucidating potential regulatory functions for such SNPs. Transcription of a gene is a complex process that depends on interactions between proteins and DNA. The transcriptional machinery involves binding of RNA polymerase II (RNA Pol II) and transcription factors (TFs) at gene promoters. The active state of transcription depends on histone modifications, vis-à-vis, chromatin accessibility. Regulatory signals can act over long distances influencing interactions of promoter and enhancer elements via the three-dimensional conformation of DNA. To elucidate potential regulatory functions, several lines of experimental data need to be integrated. ENCODE[117] and the National Institutes of Health (NIH)-Roadmap Epigenomics Projects[118] have generated data that is available in the public domain and are of immense help for understanding the functional roles of regulatory variants. These databases provide experimental evidence for open chromatin structure, histone modifications, TF binding, and high throughput sequencing and genotyping data from diverse cell types of both normal and cancer cell lines. The information from these databases can be directly accessed or interrogated using online bioinformatic tools such as RegulomeDB[119] and HaploReg[120]. Table 1.3 summarizes the different datasets, their descriptions and the online resources.

(i) Open chromatin

The open chromatin state in DNA is due to depletion of nucleosomes, which may indicate sites of active gene transcription. Openness of a chromatin state is assayed using DNase-Seq and FAIRE-seq. DNase-Seq targets DNase hypersensitivity sites which are open and not bound by nucleosomes, indicating open chromatin states at the loci of interest. FAIRE-seq[121] (Formaldehyde-Assisted Isolation of Regulatory Elements) uses a different approach, wherein DNA is cross linked with bound nucleosomes using formaldehyde, fragmented and extracted using phenol-chloroform. The nucleosome depleted DNA is separated from the DNA with bound protein during the phase separation. The nucleosome depleted DNA is later sequenced[i]. Both methods are complimentary and offer insights into the open chromatin states.

(ii) DNA-protein interactions

[i] I refer to the use of Next Generation Sequencing (NGS) technologies in the context of DNA sequencing throughout this thesis, unless specified otherwise

Binding of different types of proteins to DNA sequences may lead to gene expression or regulatory functions, depending on the nature of the protein and its sequence specificity. For instance, the binding of TFs to DNA can be computationally predicted using Position Weighted Matrices (PWM). However, the experimental evidence of protein binding to DNA is assayed using ChIP-seq[122] and DNase foot printing[123]. In ChIP-seq, DNA is cross-linked with bound protein using formaldehyde and fragmented, after which specific antibodies attached to magnetic beads are used to pull down the bound protein of interest. The enriched DNA bound protein is de-crosslinked, and the DNA is sequenced. This assay specifically detects DNA sequence motifs for binding to the protein of interest. DNase foot printing can also detect binding of proteins to DNA, using enzymatic cleavage, wherein the DNA with bound protein is often protected from the enzymatic reaction compared to free DNA. The bound and unbound DNA fragments can be distinguished from each other because they migrate during gel electrophoresis at different mobilities.

(iii) DNA methylation

Methylation of cytosine residues in CpG islands indicates gene silencing or repression of gene expression[124,125]. DNA methylation patterns determine if a gene is off or on. Methylation patterns can be captured using a number of high throughput techniques such as methylation array[126] and bisulphite sequencing[127].

(iv) RNA expression

The level of transcriptional activity can be measured based on quantification of transcribed RNAs. There are different types of RNAs - protein coding RNAs, non-coding

RNAs (small and long non-coding RNAs) and alternatively spliced RNAs (isoforms). The individual species of RNAs can be profiled and using gene expression microarray platforms as well as RNA-seq experiment (NGS platform)[99]. NGS offers an absolute quantification of expression of transcripts, whereas microarray-based technologies offer relative quantification of transcripts.

(v) Histone Modifications

Histone proteins together with nucleosomes bound to DNA form the fundamental blocks of eukaryotic chromatin. Modifications of residues in the tail domains of histone proteins play an important role in epigenetic regulatory activities[128]. There are different modifications including methylation and acetylation at different lysine residues. The combinations of these histone modifications (histone code) can determine the state of chromatin as either active or inactive. The histone modifications are conserved across the cell types and are tissue specific. Methylation patterns (mono, di or tri methylations) or acetylation on histones are specific in promoters or enhancers. For instance, H3K4me1, H3K4me2 or H3K4me3 indicate active promoters or enhancers; H3K27me3 indicates inactive promoters; H3K79me2 indicates transcription transition; H3K27ac indicates active regulatory regions; H3K9ac indicates promoters; H3K9me1indicates active chromatin; H3K9me3indicates repressed chromatin[128]. Histone modifications can be assayed using ChIP-seq method utilizing specific antibodies.

(vi) Chromatin interactions

The interactions facilitated by DNA looping brings together regulatory motifs (such as enhancers and promoters) to impart gene regulation. These mechanisms are complex and tissue-specific. With comprehensive genomics approaches, and the data deposited in the public domain, delineation of complex gene regulatory mechanisms is now feasible. The higher dimensional interactions of DNA can be captured using techniques such as Chromosome Conformation Capture (3C)[129], Circular Chromosome Conformation Capture (4C)[130,131], Carbon-Copy Chromosome Conformation Capture (5C)[132], Combined 3C-ChIP-Cloning (6C); Hi-C (High Throughout Sequencing and an extension of the technique of 3C)[133], Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)[134]. All the above techniques are derived from 3C, which typically captures the three-dimensional conformation of DNA using DNA crosslinking, ligating cross-linked ends, de-crosslinking and sequencing. A schematic representation of these methods is illustrated in Figure 1.3.

Figure 1.3 Illustration of special organization of chromatin within a cell by chromatin conformation technologies

Figure is from Li et al (2014)[135]. The basic methodology of conformation capture assays depicted here (top panel) involves crosslinking of the chromatin using formaldehyde to freeze the interacting genomic loci, followed by digestion with restriction enzymes and random ligation that favors ligation of the ends that are crosslinked fragments compared to non-interacting fragments. The bottom panel explaining different methods including 3C,4C,5C, Hi-C, ChiP-loop and ChiA-PET. Finally, interacting loci are quantified using PCR with known primers in 3C. **4C** captures the interaction between one locus versus all other genomic loci. It involves digestion with restriction enzyme (every 4 base pairs) and ligation to form self-circularized DNA fragments and followed by inverse PCR. Microarray or sequencing of the PCR amplification can capture about million interactions. **5C** method captures interactions between all restriction digested fragments. Universal Primers were ligated to the fragments and amplified. The amplified fragments were captured using microarray or sequencing. Capturing the genome-wide complex interactions by the 5C method were limited by the number of primers needed. **Hi-C** method used high throughput sequencing to detect the fragments of obtained by restriction digestion and detected by adopting pair-end sequencing to captures the interacting fragments. **ChiP-loop** combines 3C with ChiP-seq to detect interacting loci mediated by protein of interest. **ChIA-PET** is combination of Hi-C and ChiP-seq, to detect all interacting loci mediated by specific protein.

(vii) Expression Quantitative Trait Loci (eQTL)

A subset of genomic variants is capable of conferring phenotypes (termed Quantitative Trait Loci) and those variants regulating tissue specific gene expression are termed expression Quantitative Trait Loci (eQTL). The heritable nature of the germline variants and their correlations with genotypes are useful to explain a proportion of genetic variance in gene expression phenotypes[99]. Several studies have shown that the SNPs in the GWAS-identified loci are eQTLs regulating putative target genes. There are several statistical methods available to identify genotype-gene expression correlations. However, eQTL mapping studies are informative only if the genotype data and specific tissue level gene expression data are available from the same individuals. There are online resources[136,137] wherein such matched data sets (in normal and cancerous tissues) are used and a summary of the eQTLs is available for interrogation. Such databases require input of SNP identifiers.

(viii) Allelic specific expression:

In allelic specific expression, the effects of the alleles (major and minor) are investigated for their influence on gene expression in contrast with eQTL correlations wherein genotypic influence on gene expression is investigated. If the variant of interest is in a regulatory region, allelic specific analysis will reveal if binding of TFs to the allele can influence the gene expression[114]. This helps to understand the effect of the risk allele on gene expression compared to the referent or wild type allele[99,105]. If the SNP is in a coding exon, its allelic expression reflects the preferential transcript expressed in the cell. However, for allele specific analysis, the genotype and gene expression profiles from heterozygote individuals or relevant cell lines are needed. Overall allelic expression is also influenced by histone modifications and the open chromatin state of the DNA.

Table 1.3 Bioinformatics tools and resources for functional annotation of regulatory variants

Feature	Experimental Approach	Bioinformatic Tools and Online Resources
Open chromatin	DNase-seq, FAIRE sequencing	ENCODE, NIH Roadmap, Epigenomics Project, RegulomeDB, HaploReg, FunciSNP
TF-binding prediction	Position Weight Matrices	TRANSFAC, JASPAR, MAPPER2
DNA-protein interaction	ChIP-seq, DNase foot-printing	ENCODE, NRCistrome, RegulomeDB, HaploReg
DNA methylation	methylation array, bisulphite sequencing	ENCODE, NIH Roadmap, Epigenomics Project, MethDB, EpiGraph
RNA expression	RNA-seq, RNA-PET, CAGE	ENCODE, Gene Expression Omnibus, Galaxy
Histone modifications	ChIP-seq	ENCODE, NIH Roadmap Epigenomics Project, NRCistrome, RegulomeDB, HaploReg, ChromHMM, GWAS3D, Segway, ChroMoS
Chromatin interactions	3C, 4C, 5C, 6C, Hi-C, ChIA-PET	GWAS3D, Hi-C Project, ChIA-PET Browser

1.5. Copy Number Variants

Germline CNVs are a class of structural variants of DNA, involving loss or gain of segments of size >50bp[138]. The base pair coverage by all genomic CNVs is an order of magnitude higher than the cumulative genomic coverage by all SNPs[139-144]. CNVs are also polymorphic and those with population frequencies of >5% are termed common CNVs (similar definitions are ascribed to SNPs). CNVs are relatively stable, heritable and contribute to genetic predisposition of diseases and traits. CNVs have not been studied much for genetic heritability in breast cancer although studies on rare CNVs and predisposition to breast cancer have been reported[145-148]. However, the common CNVs and breast cancer risk are currently a subject of intensive investigations in the Damaraju laboratory.

CNVs are complex and copy status can be anywhere from a total deletion (single copy or both copies) to multicopy amplification of the same region. As such, CNVs may confer gene dosage effects and therefore a higher phenotypic variance can be explained at a population level. Phenotypic effects may vary, *i.e.*, those that confer survival advantage to species (adaptive traits), or cause diseases or embryonic lethality. Such deleterious CNVs may be selectively eliminated during evolution[149,150]. For instance, CNVs affecting the gene encoding alpha-amylase contribute to the adaptation to starch consumption[151]. CNVs have also been linked to a number of disease conditions such as autism[152,153], schizophrenia[154], Crohn's disease[141,155], rheumatoid arthritis[141], type1 diabetes[141], obesity[156] and developmental disorders[142,157-159]. Germline CNVs have also been investigated for their role in susceptibility to familial breast cancer[145-148,160,161] and cancers of prostate[162-164], ovary[161,165-167], pancreas[168-170], colon and rectum[147,171-175], endometrium[176], lung[177-179] and melanoma[180,181].

1.5.1. Mechanism of CNV formation

The genomic rearrangements implicated from recombination-based mechanisms such as nonallelic homologous recombination (NAHR), nonhomologous end-joining[182] (NHEJ) and retrotransposition[183-186] result in the formation of CNVs. Recently replication-based mechanisms, fork stalling and template switching (FoSTeS) mechanisms [187,188] were also proposed to contribute to the formation of CNVs (Figure 1.4). The CNV formation and the role of DNA recombination pathways are complex. The following models were proposed as a basis to understand the CNV origins.

(i) Nonallelic homologous recombination (NAHR):

NAHR occurs during meiosis and mitosis, involving alignment and crossover of two non-allelic or paralogous DNA sequences at the region of sequence repeats sharing high similarity[190]. However, if repeats are on the same chromosome, and the same orientation, a duplication or deletion event can occur, wherein inverted repeats mediate inversion of the genomic interval flanked by the repeats. If the repeats are on different chromosomes, they may lead to chromosomal translocation. Substrates for NAHR are the low copy repeats (LCR) or segmental duplication of size more than 10kb with > 95% sequence similarity[190,191]. NAHR rates on the genome are determined by genetic and environmental factors. Thus, NAHR contributes to genomic rearrangements and the resulting phenotypic variations in populations. NAHR during meiosis results in unequal crossing over leading to genomic rearrangements. CNVs originating from NAHR may be benign or contribute to inherited genomic disorders[144,182,192]. Another class of CNVs to which NAHR contributes are called *de novo* CNVs which may once again be benign or disease causing. Autism spectrum[193,194], neurodevelopmental diseases and schizophrenia[195-197] are representative genetic disorders with *de novo* CNVs contributing to the disease etiology.

(ii) Nonhomologous end-joining (NHEJ):

Nonhomologous end-joining (Figure 1.4) is a mechanism utilized by human cells to repair double strand breaks (DSBs) in DNA caused by ionizing radiation or reactive oxygen species[198-200]. NHEJ is distinct from NAHR in that NHEJ does not require substrates with extended homologies and in the process can lose or add several nucleotides at the joined end.

(iii) Fork stalling and template switching (FoSTeS):

Lee et al.[187] proposed the mechanism of fork stalling and template switching (FoSTeS) as one possible mechanism for genomic rearrangements. According to this model, the DNA replication fork stalls, and the lagging strand uncouples from the original template and switches to another replication fork, restarting DNA synthesis with a new fork. This happens via small homology between the switched arm and the original fork [187]. The new template formed may not be adjacent to the original replication fork at the primary sequence but may be in proximity in three-dimensional space. Depending on the fork progression and location downstream or upstream of the original fork, template switching may result in deletion or duplication.

(iv) L1 Retrotransposition:

Long interspersed elements-1 (L1) cover up to 17% of human genomic DNA and are known to contribute to CNVs [183,201]. L1 elements are known as active transposons in human genomes. Nearly 15% of the structural variants that are detected are due to retro transposition events[184].

Figure 1.4 Mechanism of copy number variation

The figure illustrates the mechanism of copy number variants described above (a) Nonallelic homologous recombination (NAHR) - regions of recombination at repeats such as low copy repeats regions, *Alu* element or L1- element. (b) Nonhomologous end-joining (NHEJ)- double strand break repair mechanism via recombination (c) Fork stalling and template switching (FoSTeS)- single FoSTeS (x1) and multiple FoSTeS (x2) causes simple and complex rearrangements respectively. (d) L1 retrotransposition. TS, target site and TSD, Target Site duplication. The figure is from Zhang F et al.[189] (2009). Thick colored bars indicate different genomic fragments and different colors (orange and red in NHEJ/L1 transposition or orange/red/green in FoSTeS×2) indicate that there is no homology between the two fragments. Bars represented in shades of blue (NAHR) indicate extensive homology with each other. The triangles (filled or empty) symbolize short sequences sharing microhomologies.

1.5.2. Function of CNVs in gene regulation

The DNA sequence coverage for CNVs is ~10% of the genome. CNVs harbor coding regions and non-coding regulatory regions and may confer profound phenotypic effects relative to effects caused by SNPs[202-204]. CNVs have a multitude of effects based on their genomic location, including gene dosage effects and *cis*-regulatory functions[164]. Since the distribution of CNVs across the genome is disproportionate with a higher proportion in non-coding than coding regions, their functional impact on phenotype is not clear. However, CNVs that overlap protein coding genes offer insights into disease phenotypes and associated biology[142]. Nearly 80% of cancer genes harbor CNVs[205] and support the premise that CNVs in genes contribute to phenotypic variance.

A study based on the HapMap dataset, which includes data from 270 human lymphoblastoid cell lines, assessed the impact of CNVs on gene expression. It has been estimated that ~20% of measurable genetic impact on gene expression is due to CNVs[206]. CNVs can modify gene expression by gene dosage through either amplifications or deletions. Figure 1.5 may be consulted for potential mechanisms of CNVs influencing gene expression or regulation. CNVs can disrupt gene structure, including gene fusion events that lead to formation of novel transcripts[207]. CNVs can also influence regulation of genes from long distances through *cis*- or *trans*- mechanisms, and not necessarily by gene dosage effects[207-211]. Gene dosage effects can occur if the gene overlaps a structural variant due to inversion or translocation[207]. There are also other mechanisms by which the regulatory molecules such as the microRNAs and other small non-coding RNAs harbored within the CNV regions can potentially play a role in gene regulation.

Figure 1.5 Potential mechanisms of how CNVs influence phenotypes

The above figure describes the possible mechanism by which structural variants can influence gene expression and contribute to phenotypes. The figure is used from Feuk et al (2006)[207]. The green bars in the figure is shown in pairs (homologous chromosomes) to indicate the diploid status of human genome. (a) Genes that are encompassed by structural variants are affected by dosage sensitivity. Deletion or duplication of dosage sensitive gene will result in the phenotype. Deletion is depicted in the figure. Deletion of dosage insensitive gene may result in phenotype by activation of the recessive mutant allele on the homologous chromosome. (b) Genes that overlap structural variants can be disrupted directly by inversion (upper panel), deletion or translocation (lower panel) which leads to the reduced expression of dosage-sensitive genes. (c) Genes that flank a structural variant can also result in dosage sensitivity, upper panel depicts the deletion of the regulatory element can alter the gene expression or may unmask of a functional polymorphism. (d) Genes that are involved in complex disorders, where a combination of variations can produce phenotype.

CNVs are known to play a role in several disease phenotypes. They have been exhaustively investigated for their role in neurodevelopmental disorders, however their role in cancer predisposition is slowly evolving. Understanding of the role of germline CNVs in breast cancer is in its early stages, with the majority of studies reporting rare CNVs associated in familial breast cancer. Studies describing CNVs as genetic determinants of sporadic breast cancer are limited. Long et al.[212] considered a candidate CNV for detailed analysis. The study used a case-control approach in subjects of Chinese ancestry (Stage 1: 2623 breast cancer patients and 1946 control subjects and Stage 2: 4254 breast cancer patients and 4387 control subjects) and reported the association of a common deletion in *APOBEC* genes with breast cancer. The study reported that the effect size associated with one-copy deletion is 1.31 (95% CI = 1.21 to 1.42) and two-copy deletion is 1.76 (95% CI = 1.57 to 1.97). Later the association was replicated in European[213] and Iranian populations[214]. I have described the association of a number of candidate common CNVs associated with sporadic breast cancer in Chapter 3[215] and

Chapter 4[216]. The Damaraju laboratory is the first to report CNV GWAS for sporadic breast cancer in Caucasian populations.

1.6. Genetic risk factors for predisposition to breast cancer prognosis

Even though breast cancer prognosis is often determined by histopathological features of the tumor, there are a subset of patients who experience poor outcomes irrespective of the predicted good prognosis. Current tumor-based markers for prognosis are useful in guiding treatments but markers with higher specificity would be more useful in addressing inter-individual variations in breast cancer prognosis. Several gene expression profiling studies from tumors have identified potential prognostic mRNA-based[217,218] and miRNA-based[219] markers. However, germline DNA markers for prognosis are unexplored. I reasoned that germline prognostic markers may complement the existing tumor-based markers to yield prognostic models of higher specificity and accuracy. According to the gene predisposition model for prognosis[220], it is believed that the genetic burden of the host can play a role in the expression of metastatic phenotypes of tumors. There are attempts in the literature to find SNPs and CNVs of prognostic value (discussed below). GWAS-identified SNPs showing association with breast cancer susceptibility were not prognostic[221,222]. Also, independent SNP based GWASs for prognosis in breast cancer were not informative[3,221-224]. However, the Damaraju laboratory previously described that germline Copy Neutral Loss of Heterozygosity (CN-LOH, a class of CNVs) are associated with recurrence free survival in breast cancer[225]. These initial findings prompted me to undertake an in-depth investigation of the role of common CNVs as prognostic markers. I address these in Chapter 3 of this thesis[215]. CNVs are informative compared to SNPs, since CNVs and their embedded genes may confer higher levels of penetrance (relative to SNPs) owing to loss or gain of functions. Germline CNVs have been identified as prognostic markers for several cancer types including prostate cancer[226], ovarian cancer[166] and colorectal cancer[227]. CNVs provide mechanistic insights and allow deciphering the biological roles of the affected genes. Understanding the genes and/or pathways affected may offer therapeutics developments. In my current efforts, I focused on the prognostic relevance of the CNVs which showed association with breast cancer. This work therefore should lay the foundation that CNVs play a role in both breast cancer susceptibility and prognosis. An independent CNV-GWAS study for breast cancer prognosis was beyond the scope of the study.

1.7. Gaps in the literature

The rationale to conduct the current study was to uncover genetic variants associated with breast cancer. GWASs have identified several variants to be associated with breast cancer susceptibility[61]. Yet the variants reported so far showed increased risk among predominantly postmenopausal cases (both familial and sporadic cases)[19,62]. However, premenopausal women also develop breast cancer in sporadic cases (age at onset >45 and without any family history). Previous studies from the Damaraju laboratory identified a novel locus on chr4q31.22 to be associated with premenopausal breast cancer risk[20,222]. In my study, I have reconfirmed these findings and fine-mapped the locus to identify putative causal variants. While SNP-based GWASs could not fully account for breast cancer heritability, there is a need to identify other genetic variants which can potentially account for the missing heritability. I have investigated the role of CNVs in breast cancer susceptibility. As mentioned above, SNPs showing association with breast cancer susceptibility were not prognostic[221,222], and independent SNP-based GWASs did not reveal variants associated with breast cancer prognosis[221-224,226,227]. Therefore, I further explored the contribution of breast cancer associated CNVs in prognosis.

1.8. Hypothesis

Common germline polymorphisms (SNPs and CNVs) are heritable determinants for breast cancer susceptibility and prognosis.

1.9. Objective

The specific objectives of the research described in this thesis were as follows:

1. To replicate and validate the association at the chr4q31.22 locus with premenopausal breast cancer risk in Caucasian and non-Caucasian populations **(Chapter 2)**.

2. To fine-map the chr4q31.22 locus to identify putative causal variants **(Chapter 2)**.

3. To identify the germline copy number variants harboring coding genes and show association with breast cancer susceptibility and prognosis **(Chapter 3)**.

4. To identify the germline copy number variants harboring small non-coding RNA genes and their role in conferring breast cancer risk (**Chapter 4**).

1.10. Organization of the thesis

The thesis has been organized into six chapters, each addressing a specific objective as described below. Introduction pertinent to individual study objectives are provided in the corresponding chapters. At the outset, the following historical account offers the premise for the findings summarized in this thesis.

Earlier studies published from the Damaraju laboratory reported six putative variants [20] associated with sporadic breast cancer. Stage 1 of the study consisted of 348 breast cancer cases and 348 controls of Caucasian ancestry and utilized whole genome genotyping platform Affymetrix Human SNP 6.0 array (~906,600 SNPs). Following population stratification analysis, 302 cases and 321 controls that clustered with HapMap Caucasian subjects were retained for association analysis. Genotype level data filtering resulted in a total of 782,838 SNPs that were amenable for single-locus association tests. Association analysis revealed 35,859 SNPs associated with breast cancer at statistical significance $P<0.05$.

Of the associated SNPs, 35 were selected as described for Stage 2 replication[20]. Stage 2 consisted of 1,153 breast cancer cases and 1,215 controls. Six of the 35 SNPs (rs1429142 on chr4q31.22, rs1092913 on chr5p15.2, rs10411161, rs3848562, rs11878583 on chr19q13.33, rs1981867 on chr16q23.2) were replicated. Subsequently, an independent Stage 3 replication study [222] consisting of 1,294 breast cancer cases and 2,934 controls of Caucasian ancestry from Alberta, Canada, replicated the association of the two SNPs rs1429142 on chr4q31.22 and rs1092913 on chr5p15.2. In the combined analysis of Stages 1-3, rs1429142 on chr4q31.22 showed association with overall risk for breast cancer association reaching near genome level significance (adjusted for BMI, $P = 1.5 \times 10^{-7}$). In the stratified analysis of Stages 1-3 cases and controls, SNP rs1429142 showed elevated risk with premenopausal breast cancer risk compared to post-menopausal breast cancer and showed genome wide significance (adjusted for BMI, $P = 10^{-10}$). Stratified analysis based on luminal A status, menopausal status, family history of breast cancer, tumor stage and grade did not reveal any elevated risk associated with the SNPs. Characterization of the second variant (rs1092913) described by the Damaraju laboratory, and which showed replication in three stages, warrants further investigations.

In chapter 2, I address objectives 1 and 2. I reconfirmed the association for rs1429142 for elevated breast cancer risk among premenopausal women using an independent set of 1502 breast cancer cases (Stage 4). In Stage 4 of the study, the association of SNP rs1429142 with overall risk for breast cancer was also replicated. In the combined Stages 1-4, the association of SNP rs1429142 reached genome-wide significance. I also investigated the association of SNP rs1429142 in independent external datasets of Caucasian (CGEMs study: 1144 cases and 1143 controls, all postmenopausal), and non-Caucasian populations (African Diaspora: 1607 cases and 2041 controls).

Objective 2 consisted of fine-mapping of the chr4q31.22 locus previously shown to be associated with premenopausal breast cancer risk. The fine-mapping approaches adopted in the study were aimed at identification of potential causal variants.

In chapter 3, I describe (objective 3) the identification of common CNVs overlapping with protein coding genes and their association with breast cancer susceptibility. I also describe the contribution of a subset of breast cancer associated CNVs in conferring genetic predisposition for prognosis (Overall Survival and Recurrence Free Survival).

In chapter 4, I describe (objective 4) the identification of CNVs harboring small-non-coding RNA genes (microRNA, piwi-interacting RNA, small nucleolar RNA and transfer RNA) and their association with breast cancer susceptibility. I investigated expression of these small non-coding RNAs in breast tissue and their role in post transcriptional gene regulation. Profiling of small non-coding RNAs in breast tissues and their role in prognosis was addressed earlier by the Damaraju laboratory[219,228-230]. The role of CNV embedded small non-coding RNAs in prognosis is not addressed here due to the lower frequencies of these CNVs. I used common CNVs harboring protein coding genes (>10% frequency). Therefore, larger sample size is needed to capture the low frequency variants, and adequate number of events (survival or recurrence events) to address prognostic relevance.

Further, the overall discussion (**chapter 5**), conclusions and future directions (**chapter 6**) are described, followed by appendices and bibliography.

1.11. References

1 Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136, E359-386, doi:10.1002/ijc.29210 (2015).

2 Ghoncheh, M., Pournamdar, Z. & Salehiniya, H. Incidence and Mortality and Epidemiology of Breast Cancer in the World. Asian Pac J Cancer Prev 17, 43-46 (2016).

3 Canadian Cancer Society. Breast Cancer Statistics. (2017).

4 Palacios, S., Henderson, V. W., Siseles, N., Tan, D. & Villaseca, P. Age of menopause and impact of climacteric symptoms by geographical region. Climacteric 13, 419-428, doi:10.3109/13697137.2010.507886 (2010).

5 Lichtenstein, P. et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343, 78-85, doi:10.1056/NEJM200007133430201 (2000).

6 McPherson, K., Steel, C. M. & Dixon, J. M. ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. BMJ 321, 624-628 (2000).

7 Hall, J. M. et al. Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250, 1684-1689 (1990).

8 Morton, N. E. Sequential tests for the detection of linkage. Am J Hum Genet 7, 277-318 (1955).

9 Easton, D. F., Bishop, D. T., Ford, D. & Crockford, G. P. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. Am J Hum Genet 52, 678-701 (1993).

10 Miki, Y. et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266, 66-71 (1994).

11 Wooster, R. et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science 265, 2088-2090 (1994).

12 Struewing, J. P. et al. The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals. Nat Genet 11, 198-200, doi:10.1038/ng1095-198 (1995).

13 Roa, B. B., Boyd, A. A., Volcik, K. & Richards, C. S. Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. Nat Genet 14, 185-187, doi:10.1038/ng1096-185 (1996).

14 Abeliovich, D. et al. The founder mutations 185delAG and 5382insC in BRCA1 and 6174delT in BRCA2 appear in 60% of ovarian cancer and 30% of early-onset breast cancer patients among Ashkenazi women. Am J Hum Genet 60, 505-514 (1997).

15 Syrjakoski, K. et al. Population-based study of BRCA1 and BRCA2 mutations in 1035 unselected Finnish breast cancer patients. J Natl Cancer Inst 92, 1529-1531 (2000).

16 Struewing, J. P., Tarone, R. E., Brody, L. C., Li, F. P. & Boice, J. D., Jr. BRCA1 mutations in young women with breast cancer. Lancet 347, 1493 (1996).

17 Zelada-Hedman, M. et al. A screening for BRCA1 mutations in breast and breast-ovarian cancer families from the Stockholm region. Cancer Res 57, 2474-2477 (1997).

18 Hill, A. D., Doyle, J. M., McDermott, E. W. & O'Higgins, N. J. Hereditary breast cancer. Br J Surg 84, 1334-1339 (1997).

19 Hunter, D. J. et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics 39, 870-874, doi:ng2075 [pii] (2007).

20 Sehrawat, B. et al. Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human genetics 130, 529-537, doi:10.1007/s00439-011-0973-1 [doi] (2011).

21 Perou, C. M. et al. Molecular portraits of human breast tumours. Nature 406, 747-752, doi:10.1038/35021093 (2000).

22 Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 or BRCA2 mutations and sporadic cases. Breast Cancer Linkage Consortium. Lancet 349, 1505-1510 (1997).

23 Lakhani, S. R. et al. Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations. J Natl Cancer Inst 90, 1138-1145 (1998).

24 Foulkes, W. D. et al. Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer. J Natl Cancer Inst 95, 1482-1485 (2003).

25 Honrado, E., Benitez, J. & Palacios, J. The molecular pathology of hereditary breast cancer: genetic testing and therapeutic implications. Mod Pathol 18, 1305-1320, doi:10.1038/modpathol.3800453 (2005).

26 Lakhani, S. R. et al. The pathology of familial breast cancer: predictive value of immunohistochemical markers estrogen receptor, progesterone receptor, HER-2, and p53 in patients with mutations in BRCA1 and BRCA2. J Clin Oncol 20, 2310-2318, doi:10.1200/JCO.2002.09.023 (2002).

27 Robson, M. E. et al. A combined analysis of outcome following breast cancer: differences in survival based on BRCA1/BRCA2 mutation status and administration of adjuvant treatment. Breast Cancer Res 6, R8-R17, doi:10.1186/bcr658 (2004).

28 Gonzalez, K. D. et al. Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. J Clin Oncol 27, 1250-1256, doi:10.1200/JCO.2008.16.6959 (2009).

29 Birch, J. M. et al. Linkage studies in a Li-Fraumeni family with increased expression of p53 protein but no germline mutation in p53. Br J Cancer 70, 1176-1181 (1994).

30 Garber, J. E. & Offit, K. Hereditary cancer predisposition syndromes. J Clin Oncol 23, 276-292, doi:10.1200/JCO.2005.10.042 (2005).

31 Olivier, M. et al. Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype. Cancer Res 63, 6643-6650 (2003).

32 Walsh, T. et al. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA 295, 1379-1388, doi:10.1001/jama.295.12.1379 (2006).

33 Tomlinson, I. P. & Houlston, R. S. Peutz-Jeghers syndrome. J Med Genet 34, 1007-1011 (1997).

34 van Lier, M. G. et al. High cancer risk in Peutz-Jeghers syndrome: a systematic review and surveillance recommendations. Am J Gastroenterol 105, 1258-1264; author reply 1265, doi:10.1038/ajg.2009.725 (2010).

35 Hearle, N. et al. Frequency and spectrum of cancers in the Peutz-Jeghers syndrome. Clin Cancer Res 12, 3209-3215, doi:10.1158/1078-0432.CCR-06-0083 (2006).

36 Pharoah, P. D., Guilford, P., Caldas, C. & International Gastric Cancer Linkage, C. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. Gastroenterology 121, 1348-1353 (2001).

37 Meijers-Heijboer, H. et al. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet 31, 55-59, doi:10.1038/ng879 (2002).

38 Renwick, A. et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet 38, 873-875, doi:10.1038/ng1837 (2006).

39 Rahman, N. et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet 39, 165-167, doi:10.1038/ng1959 (2007).

40 Seal, S. et al. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet 38, 1239-1241, doi:10.1038/ng1902 (2006).

41 Shaag, A. et al. Functional and genomic approaches reveal an ancient CHEK2 allele associated with breast cancer in the Ashkenazi Jewish population. Hum Mol Genet 14, 555-563, doi:10.1093/hmg/ddi052 (2005).

42 Teo, Z. L. et al. Prevalence of PALB2 mutations in Australasian multiple-case breast cancer families. Breast Cancer Res 15, R17, doi:10.1186/bcr3392 (2013).

43 Adank, M. A., van Mil, S. E., Gille, J. J., Waisfisz, Q. & Meijers-Heijboer, H. PALB2 analysis in BRCA2-like families. Breast Cancer Res Treat 127, 357-362, doi:10.1007/s10549-010-1001-1 (2011).

44 Antoniou, A. C. et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br J Cancer 86, 76-83, doi:10.1038/sj.bjc.6600008 (2002).

45 Lander, E. S. The new genomics: global views of biology. Science 274, 536-539 (1996).

46 Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. Trends Genet 17, 502-510 (2001).

47 Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11, 2417-2423 (2002).

48 Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl, 228-237, doi:10.1038/ng1090 (2003).

49 Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860-921, doi:10.1038/35057062 (2001).

50 Bewick, M. A., Conlon, M. S. & Lafrenie, R. M. Polymorphisms in XRCC1, XRCC3, and CCND1 and survival after treatment for metastatic breast cancer. J Clin Oncol 24, 5645-5651, doi:10.1200/JCO.2006.05.9923 (2006).

51 Allen-Brady, K., Cannon-Albright, L. A., Neuhausen, S. L. & Camp, N. J. A role for XRCC4 in age at diagnosis and breast cancer risk. Cancer Epidemiol Biomarkers Prev 15, 1306-1310, doi:10.1158/1055-9965.EPI-05-0959 (2006).

52 Haiman, C. A. et al. Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations. Hum Mol Genet 17, 825-834, doi:10.1093/hmg/ddm354 (2008).

53 Pooley, K. A. et al. Common single-nucleotide polymorphisms in DNA double-strand break repair genes and breast cancer risk. Cancer Epidemiol Biomarkers Prev 17, 3482-3489, doi:10.1158/1055-9965.EPI-08-0594 (2008).

54 Mangoni, M. et al. Association between genetic polymorphisms in the XRCC1, XRCC3, XPD, GSTM1, GSTT1, MSH2, MLH1, MSH3, and MGMT genes and radiosensitivity in breast cancer patients. Int J Radiat Oncol Biol Phys 81, 52-58, doi:10.1016/j.ijrobp.2010.04.023 (2011).

55 Sehl, M. E. et al. Associations between single nucleotide polymorphisms in double-stranded DNA repair pathway genes and familial breast cancer. Clin Cancer Res 15, 2192-2203, doi:10.1158/1078-0432.CCR-08-1417 (2009).

56 Lin, W. Y. et al. A role for XRCC2 gene polymorphisms in breast cancer risk and survival. J Med Genet 48, 477-484, doi:10.1136/jmedgenet-2011-100018 (2011).

57 Couch, F. J. et al. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. Nat Commun 7, 11375, doi:10.1038/ncomms11375 (2016).

58 Breast Cancer Association, C. Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. J Natl Cancer Inst 98, 1382-1396, doi:10.1093/jnci/djj374 (2006).

59 van den Oord, E. J. Controlling false discoveries in genetic studies. Am J Med Genet B Neuropsychiatr Genet 147B, 637-644, doi:10.1002/ajmg.b.30650 (2008).

60 Hochberg, Y. & Benjamini, Y. More powerful procedures for multiple significance testing. Stat Med 9, 811-818 (1990).

61 Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92-94, doi:10.1038/nature24284 (2017).

62 Easton, D. F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. (2007).

63 MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45, D896-D901, doi:10.1093/nar/gkw1133 (2017).

64 Zheng, W. et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nature genetics 41, 324-328, doi:10.1038/ng.318 [doi] (2009).

65 Long, J. et al. Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. PLoS genetics 6, e1001002, doi:10.1371/journal.pgen.1001002 [doi] (2010).

66 Long, J. et al. Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. PLoS genetics 8, e1002532, doi:10.1371/journal.pgen.1002532 [doi] (2012).

67 Kiyotani, K. et al. A genome-wide association study identifies locus at 10q22 associated with clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients in Japanese. Human molecular genetics 21, 1665-1672, doi:10.1093/hmg/ddr597 [doi] (2012).

68 Elgazzar, S. et al. A genome-wide association study identifies a genetic variant in the SIAH2 locus associated with hormonal receptor-positive breast cancer in Japanese. Journal of human genetics 57, 766-771, doi:10.1038/jhg.2012.108 [doi] (2012).

69 Low, S. K. et al. Genome-wide association study of breast cancer in the Japanese population. PloS one 8, e76463, doi:10.1371/journal.pone.0076463 [doi] (2013).

70 Gold, B. et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proceedings of the National Academy of Sciences of the United States of America 105, 4340-4345, doi:10.1073/pnas.0800441105 [doi] (2008).

71 Rinella, E. S. et al. Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable BRCA1/2 mutation. Human genetics 132, 523-536, doi:10.1007/s00439-013-1269-4 [doi] (2013).

72 Chen, F. et al. A genome-wide association study of breast cancer in women of African ancestry. Human genetics 132, 39-48, doi:10.1007/s00439-012-1214-y [doi] (2013).

73 Kim, H. C. et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast cancer research : BCR 14, R56, doi:bcr3158 [pii] (2012).

74 Stacey, S. N. et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics 39, 865-869, doi:ng2064 [pii] (2007).

75 Thomas, G. et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nature Genetics 41, 579, doi:10.1038/ng.353

https://www.nature.com/articles/ng.353#supplementary-information (2009).

76 Turnbull, C. et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 42, 504-507, doi:10.1038/ng.586 (2010).

77 Antoniou, A. C. et al. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nat Genet 42, 885-892, doi:10.1038/ng.669 (2010).

78 Fletcher, O. et al. Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study. JNCI: Journal of the National Cancer Institute 103, 425-435, doi:10.1093/jnci/djq563 (2011).

79 Li, J. et al. A combined analysis of genome-wide association studies in breast cancer. Breast Cancer Res Treat 126, 717-727, doi:10.1007/s10549-010-1172-9 (2011).

80 Siddiq, A. et al. 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, doi:10.1093/hmg/dds381 (2012).

81 Orr, N. et al. Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk. Nat Genet 44, 1182-1184, doi:10.1038/ng.2417 (2012).

82 Garcia-Closas, M. et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet 45, 392-398, 398e391-392, doi:10.1038/ng.2561 (2013).

83 Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics 45, 353-361, 361e351-352, doi:10.1038/ng.2563 [doi] (2013).

84 Purrington, K. S. et al. Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. Carcinogenesis 35, 1012-1019, doi:10.1093/carcin/bgt404 (2014).

85 Shu, X. O. et al. Novel genetic markers of breast cancer survival identified by a genome-wide association study. Cancer Res 72, 1182-1189, doi:10.1158/0008-5472.CAN-11-2561 (2012).

86 Kim, H.-c. et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Research 14, R56, doi:10.1186/bcr3158 (2012).

87 Song, C. et al. A Genome-Wide Scan for Breast Cancer Risk Haplotypes among African American Women. PLOS ONE 8, e57298, doi:10.1371/journal.pone.0057298 (2013).

88 Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 45, doi:10.1038/ng.2563 (2013).

89 Michailidou, K. et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 47, doi:10.1038/ng.3242 (2015).

90 Amos, C. I. et al. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev 26, 126-135, doi:10.1158/1055-9965.EPI-16-0106 (2017).

91 Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet 49, 1767-1778, doi:10.1038/ng.3785 (2017).

92 Haiman, C. A. et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nature genetics 43, 1210-1214, doi:10.1038/ng.985 [doi] (2011).

93 Stevens, K. N. et al. 19p13.1 is a triple-negative-specific breast cancer susceptibility locus. Cancer Res 72, 1795-1803, doi:10.1158/0008-5472.CAN-11-3364 (2012).

94 Stevens, K. N. et al. Common breast cancer susceptibility loci are associated with triple-negative breast cancer. Cancer Res 71, 6240-6249, doi:10.1158/0008-5472.CAN-11-1266 (2011).

95 Couch, F. J. et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS genetics 9, e1003212, doi:10.1371/journal.pgen.1003212 [doi] (2013).

96 Mulligan, A. M. et al. Common breast cancer susceptibility alleles are associated with tumour subtypes in BRCA1 and BRCA2 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2. Breast Cancer Res 13, R110, doi:10.1186/bcr3052 (2011).

97 Lilyquist, J., Ruddy, K. J., Vachon, C. M. & Couch, F. J. Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. Cancer Epidemiol Biomarkers Prev 27, 380-394, doi:10.1158/1055-9965.EPI-17-1144 (2018).

98 Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42, D1001-1006, doi:10.1093/nar/gkt1229 (2014).

99 Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet 93, 779-797, doi:10.1016/j.ajhg.2013.10.012 (2013).

100 Harismendy, O. & Frazer, K. A. Elucidating the role of 8q24 in colorectal cancer. Nat Genet 41, 868-869, doi:10.1038/ng0809-868 (2009).

101 Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6, 95-108, doi:10.1038/nrg1521 (2005).

102 Reich, D. E. et al. Linkage disequilibrium in the human genome. Nature 411, 199-204, doi:10.1038/35075590 (2001).

103 Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22, 139-144, doi:10.1038/9642 (1999).

104 Udler, M. S. et al. Fine scale mapping of the breast cancer 16q12 locus. Human molecular genetics 19, 2507-2515, doi:10.1093/hmg/ddq122 (2010).

105 Genomes Project, C. et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65, doi:10.1038/nature11632 (2012).

106 Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 43, 1193-1201, doi:10.1038/ng.998 (2011).

107 Voight, B. F. et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 8, e1002793, doi:10.1371/journal.pgen.1002793 (2012).

108 Dunning, A. M. et al. The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet 67, 1544-1554, doi:10.1086/316906 (2000).

109 Udler, M. S., Tyrer, J. & Easton, D. F. Evaluating the Power to Discriminate Between Highly Correlated SNPs in Genetic Association Studies. Genet Epidemiol 34, 463-468, doi:10.1002/gepi.20504 (2010).

110 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

111 Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34, 816-834, doi:10.1002/gepi.20533 (2010).

112 Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. G3 (Bethesda) 1, 457-470, doi:10.1534/g3.111.001198 (2011).

113 Sexton, T., Bantignies, F. & Cavalli, G. Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. Semin Cell Dev Biol 20, 849-855, doi:10.1016/j.semcdb.2009.06.004 (2009).

114 French, JulietÂ D. et al. Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. American Journal of Human Genetics 92, 489-503, doi:10.1016/j.ajhg.2013.01.002 (2013).

115 Bojesen, S. E. A. et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nature genetics, 371 (2013).

116 Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747-753, doi:10.1038/nature08494 (2009).

117 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74, doi:10.1038/nature11247 (2012).

118 Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28, 1045-1048, doi:10.1038/nbt1010-1045 (2010).

119 Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research 22, 1790-1797 (2012).

120 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Research 40, D930-D934, doi:10.1093/nar/gkr917 (2012).

121 Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17, 877-885, doi:10.1101/gr.5533506 (2007).

122 Landt, S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22, 1813-1831, doi:10.1101/gr.136184.111 (2012).

123 Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489, 83-90, doi:10.1038/nature11212 (2012).

124 Thurman, R. E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75-82, doi:10.1038/nature11232 (2012).

125 Ball, M. P. et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol 27, 361-368, doi:10.1038/nbt.1533 (2009).

126 Bibikova, M. et al. Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics 1, 177-200, doi:10.2217/epi.09.14 (2009).

127 Frommer, M. et al. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89, 1827-1831 (1992).

128 Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. Nature 403, 41-45, doi:10.1038/47412 (2000).

129 Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. Nat Methods 3, 17-21, doi:10.1038/nmeth823 (2006).

130 Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38, 1348-1354, doi:10.1038/ng1896 (2006).

131 Zhao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet 38, 1341-1347, doi:10.1038/ng1891 (2006).

132 Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc 2, 988-1002, doi:10.1038/nprot.2007.116 (2007).

133 Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289-293, doi:10.1126/science.1181369 (2009).

134 Fullwood, M. J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462, 58-64, doi:10.1038/nature08497 (2009).

135 Li, G. et al. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics 15 Suppl 12, S11, doi:10.1186/1471-2164-15-S12-S11 (2014).

136 Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580-585, doi:10.1038/ng.2653 (2013).

137 Gong, J. et al. PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res 46, D971-D976, doi:10.1093/nar/gkx861 (2018).

138 Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. Nat Rev Genet 16, 172-183, doi:10.1038/nrg3871 (2015).

139 Levy, S. et al. The diploid genome sequence of an individual human. PLoS Biol 5, e254, doi:10.1371/journal.pbio.0050254 (2007).

140 Conrad, D. F. & Hurles, M. E. The population genetics of structural variation. Nat Genet 39, S30-36, doi:10.1038/ng2042 (2007).

141 Craddock, N. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, doi:10.1038/nature08979 (2010).

142 Lee, C. & Scherer, S. W. The clinical context of copy number variation in the human genome. Expert Rev Mol Med 12, e8, doi:10.1017/S1462399410001390 (2010).

143 Pang, A. W., Macdonald, J. R., Yuen, R. K., Hayes, V. M. & Scherer, S. W. Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. G3 (Bethesda) 4, 63-65, doi:10.1534/g3.113.008797 (2014).

144 Lupski, J. R. Genomic rearrangements and sporadic disease. Nat Genet 39, S43-47, doi:10.1038/ng2084 (2007).

145 Krepischi, A. C. et al. Germline DNA copy number variation in familial and early-onset breast cancer. Breast Cancer Res 14, R24, doi:10.1186/bcr3109 (2012).

146 Pylkas, K. et al. Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. PLoS Genet 8, e1002734, doi:10.1371/journal.pgen.1002734 (2012).

147 Villacis, R. A. et al. ROBO1 deletion as a novel germline alteration in breast and colorectal cancer patients. Tumour Biol 37, 3145-3153, doi:10.1007/s13277-015-4145-0 (2016).

148 Walker, L. C. et al. Increased genomic burden of germline copy number variants is associated with early onset breast cancer: Australian breast cancer family registry. Breast Cancer Res 19, 30, doi:10.1186/s13058-017-0825-6 (2017).

149 Beckmann, J. S., Estivill, X. & Antonarakis, S. E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8, 639-646, doi:10.1038/nrg2149 (2007).

150 Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. Trends Genet 24, 238-245, doi:10.1016/j.tig.2008.03.001 (2008).

151 Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. Nature genetics 39, doi:10.1038/ng2123 (2007).

152 Pinto, D. et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet 94, 677-694, doi:10.1016/j.ajhg.2014.03.018 (2014).

153 Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466, 368-372, doi:10.1038/nature09146 (2010).

154 Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell 148, 1223-1241, doi:10.1016/j.cell.2012.02.039 (2012).

155 Cantsilieris, S. & White, S. J. Correlating multiallelic copy number polymorphisms with disease susceptibility. Hum Mutat 34, 1-13, doi:10.1002/humu.22172 (2013).

156 Jacquemont, S. et al. Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478, 97-102, doi:10.1038/nature10406 (2011).

157 Firth, H. V. et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84, doi:10.1016/j.ajhg.2009.03.010 (2009).

158 Riggs, E. R. et al. Towards an evidence-based process for the clinical interpretation of copy number variation. Clin Genet 81, 403-412, doi:10.1111/j.1399-0004.2011.01818.x (2012).

159 de Vries, B. B. et al. Diagnostic genome profiling in mental retardation. Am J Hum Genet 77, 606-616, doi:10.1086/491719 (2005).

160 Masson, A. L. et al. Expanding the genetic basis of copy number variation in familial breast cancer. Hered Cancer Clin Pract 12, 15, doi:10.1186/1897-4287-12-15 (2014).

161 Kuusisto, K. M. et al. copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. PLoS ONE [Electronic Resource] 8, e71802, doi:http://dx.doi.org/10.1371/journal.pone.0071802 (2013).

162 Laitinen, V. H. et al. Germline copy number variation analysis in Finnish families with hereditary prostate cancer. Prostate 76, 316-324, doi:10.1002/pros.23123 (2016).

163 Ledet, E. M. et al. Characterization of germline copy number variation in high-risk African American families with prostate cancer. Prostate 73, 614-623, doi:10.1002/pros.22602 (2013).

164 Demichelis, F. et al. Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc Natl Acad Sci U S A 109, 6686-6691, doi:10.1073/pnas.1117405109 (2012).

165 Pedersen, B. S., Konstantinopoulos, P. A., Spillman, M. A. & De, S. Copy neutral loss of heterozygosity is more frequent in older ovarian cancer patients. Genes Chromosomes Cancer 52, 794-801, doi:10.1002/gcc.22075 (2013).

166 Fridley, B. L. et al. Germline copy number variation and ovarian cancer survival. Front Genet 3, 142, doi:10.3389/fgene.2012.00142 (2012).

167 Yoshihara, K. et al. Germline copy number variations in BRCA1-associated ovarian cancer patients. Genes Chromosomes Cancer 50, 167-177, doi:10.1002/gcc.20841 (2011).

168 Fanale, D. et al. Germline copy number variation in the YTHDC2 gene: does it have a role in finding a novel potential molecular target involved in pancreatic adenocarcinoma susceptibility? Expert Opin Ther Targets 18, 841-850, doi:10.1517/14728222.2014.920324 (2014).

169 Fanale, D. et al. Analysis of germline gene copy number variants of patients with sporadic pancreatic adenocarcinoma reveals specific variations. Oncology 85, 306-311, doi:10.1159/000354737 (2013).

170 Al-Sukhni, W. et al. Identification of germline genomic copy number variation in familial pancreatic cancer. Hum Genet 131, 1481-1494, doi:10.1007/s00439-012-1183-1 (2012).

171 Brea-Fernandez, A. J. et al. Candidate predisposing germline copy number variants in early onset colorectal cancer patients. Clin Transl Oncol, doi:10.1007/s12094-016-1576-z (2016).

172 Weren, R. D. et al. Germline deletions in the tumour suppressor gene FOCAD are associated with polyposis and colorectal cancer development. J Pathol 236, 155-164, doi:10.1002/path.4520 (2015).

173 Yang, R. et al. Genome-wide analysis associates familial colorectal cancer with increases in copy number variations and a rare structural variation at 12p12.3. Carcinogenesis 35, 315-323, doi:10.1093/carcin/bgt344 (2014).

174 Masson, A. L. et al. Copy number variation in hereditary non-polyposis colorectal cancer. Genes (Basel) 4, 536-555, doi:10.3390/genes4040536 (2013).

175 Venkatachalam, R. et al. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. Int J Cancer 129, 1635-1642, doi:10.1002/ijc.25821 (2011).

176 Moir-Meyer, G. L. et al. Rare germline copy number deletions of likely functional importance are implicated in endometrial cancer predisposition. Hum Genet 134, 269-278, doi:10.1007/s00439-014-1507-4 (2015).

177 Liu, B. et al. A Functional Copy-Number Variation in MAPKAPK2 Predicts Risk and Prognosis of Lung Cancer. The American Journal of Human Genetics 91, 384-390, doi:http://dx.doi.org/10.1016/j.ajhg.2012.07.003 (2012).

178 Iwakawa, R. et al. Contribution of germline mutations to PARK2 gene inactivation in lung adenocarcinoma. Genes Chromosomes Cancer 51, 462-472, doi:10.1002/gcc.21933 (2012).

179 Butler, M. W. et al. Glutathione S-transferase copy number variation alters lung gene expression. Eur Respir J 38, 15-28, doi:10.1183/09031936.00029210 (2011).

180 Shi, J. et al. Rare Germline Copy Number Variations and Disease Susceptibility in Familial Melanoma. J Invest Dermatol 136, 2436-2443, doi:10.1016/j.jid.2016.07.023 (2016).

181 Fidalgo, F. et al. Role of rare germline copy number variation in melanoma-prone patients. Future Oncol 12, 1345-1357, doi:10.2217/fon.16.22 (2016).

182 Lupski, J. R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet 1, e49, doi:10.1371/journal.pgen.0010049 (2005).

183 Kazazian, H. H., Jr. & Moran, J. V. The impact of L1 retrotransposons on the human genome. Nat Genet 19, 19-24, doi:10.1038/ng0598-19 (1998).

184 Kidd, J. M. et al. Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56-64, doi:10.1038/nature06862 (2008).

185 Korbel, J. O. et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420-426, doi:10.1126/science.1149504 (2007).

186 Xing, J. et al. Mobile elements create structural variation: analysis of a complete human genome. Genome Res 19, 1516-1526, doi:10.1101/gr.091827.109 (2009).

187 Lee, J. A., Carvalho, C. M. & Lupski, J. R. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131, 1235-1247, doi:10.1016/j.cell.2007.11.037 (2007).

188 Perry, G. H. et al. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82, 685-695, doi:10.1016/j.ajhg.2007.12.010 (2008).

189 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy Number Variation in Human Health, Disease, and Evolution. Annual review of genomics and human genetics 10, 451-481, doi:10.1146/annurev.genom.9.081307.164217 (2009).

190 Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. Trends Genet 18, 74-82 (2002).

191 Shaw, C. J. & Lupski, J. R. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet 13 Spec No 1, R57-64, doi:10.1093/hmg/ddh073 (2004).

192 Turner, D. J. et al. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat Genet 40, 90-95, doi:10.1038/ng.2007.40 (2008).

193 Sebat, J. et al. Strong association of de novo copy number mutations with autism. Science 316, doi:10.1126/science.1138659 (2007).

194 Weiss, L. A. et al. Association between microdeletion and microduplication at 16p11.2 and autism. The New England Journal of Medicine 358, doi:10.1056/NEJMoa075974 (2008).

195 Moreno-De-Luca, D. et al. Deletion 17q12 Is a Recurrent Copy Number Variant that Confers High Risk of Autism and Schizophrenia. American Journal of Human Genetics 87, 618-630, doi:10.1016/j.ajhg.2010.10.004 (2010).

196 Warnica, W. et al. Copy Number Variable MicroRNAs in Schizophrenia and Their Neurodevelopmental Gene Targets. Biological Psychiatry 77, 158-166, doi:http://dx.doi.org/10.1016/j.biopsych.2014.05.011 (2015).

197 Xu, B. et al. Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet 40, 880-885, doi:10.1038/ng.162 (2008).

198 Lieber, M. R., Lu, H., Gu, J. & Schwarz, K. Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. Cell Res 18, 125-133, doi:10.1038/cr.2007.108 (2008).

199 Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. Nat Rev Mol Cell Biol 4, 712-720, doi:10.1038/nrm1202 (2003).

200 Schwarz, K., Ma, Y., Pannicke, U. & Lieber, M. R. Human severe combined immune deficiency and DNA repair. Bioessays 25, 1061-1070, doi:10.1002/bies.10344 (2003).

201 Goodier, J. L. & Kazazian, H. H., Jr. Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135, 23-35, doi:10.1016/j.cell.2008.09.022 (2008).

202 Sebat, J. et al. Large-scale copy number polymorphism in the human genome. Science 305, 525-528, doi:10.1126/science.1098918 (2004).

203 Iafrate, A. J. et al. Detection of large-scale variation in the human genome. Nature genetics 36, doi:10.1038/ng1416 (2004).

204 Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. Nature 464, 704-712, doi:10.1038/nature08516 (2010).

205 Pang, A. W. et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11, R52, doi:10.1186/gb-2010-11-5-r52 (2010).

206 Stranger, B. E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315, doi:10.1126/science.1136678 (2007).

207 Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. Nat Rev Genet 7, 85-97, doi:10.1038/nrg1767 (2006).

208 Merla, G. et al. Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. Am J Hum Genet 79, 332-341, doi:10.1086/506371 (2006).

209 Gamazon, E. R., Nicolae, D. L. & Cox, N. J. A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. PLoS Genet 7, e1001292, doi:10.1371/journal.pgen.1001292 (2011).

210 Kleinjan, D. A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76, 8-32, doi:10.1086/426833 (2005).

211 Kleinjan, D. J. & van Heyningen, V. Position effect in human genetic disease. Hum Mol Genet 7, 1611-1618 (1998).

212 Long, J. et al. A common deletion in the APOBEC3 genes and breast cancer risk. J Natl Cancer Inst 105, 573-579, doi:10.1093/jnci/djt018 (2013).

213 Xuan, D. et al. APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis 34, 2240-2243, doi:10.1093/carcin/bgt185 (2013).

214 Rezaei, M., Hashemi, M., Hashemi, S. M., Mashhadi, M. A. & Taheri, M. APOBEC3 Deletion is Associated with Breast Cancer Risk in a Sample of Southeast Iranian Population. International Journal of Molecular and Cellular Medicine 4, 103-108 (2015).

215 Kumaran, M. et al. Germline copy number variations are associated with breast cancer risk and prognosis. Scientific Reports 7, 14621, doi:10.1038/s41598-017-14799-7 (2017).

216 Kumaran, M. et al. Breast cancer associated germline structural variants harboring small noncoding RNAs impact post-transcriptional gene regulation. Sci Rep 8, 7529, doi:10.1038/s41598-018-25801-1 (2018).

217 Dabbs, D. J. et al. High false-negative rate of HER2 quantitative reverse transcription polymerase chain reaction of the Oncotype DX test: an independent quality assurance study. J Clin Oncol 29, 4279-4285, doi:10.1200/JCO.2011.34.7963 (2011).

218 van 't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 415, 530-536, doi:10.1038/415530a (2002).

219 Krishnan, P. et al. Next generation sequencing profiling identifies miR-574-3p and miR-660-5p as potential novel prognostic markers for breast cancer. BMC Genomics 16, 735, doi:10.1186/s12864-015-1899-0 (2015).

220 Ribelles, N., Santonja, A., Pajares, B., Llacer, C. & Alba, E. The seed and soil hypothesis revisited: current state of knowledge of inherited genes on prognosis in breast cancer. Cancer Treat Rev 40, 293-299, doi:10.1016/j.ctrv.2013.09.010 (2014).

221 Azzato, E. M. et al. A genome-wide association study of prognosis in breast cancer. Cancer Epidemiol Biomarkers Prev 19, 1140-1143, doi:10.1158/1055-9965.EPI-10-0085 (2010).

222 Sapkota, Y. et al. Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm. PLoS One 8, e62550, doi:10.1371/journal.pone.0062550 (2013).

223 Rafiq, S. et al. A genome wide meta-analysis study for identification of common variation associated with breast cancer prognosis. PloS one 9, e101488, doi:10.1371/journal.pone.0101488 [doi] (2014).

224 Azzato, E. M. et al. Association Between a Germline OCA2 Polymorphism at Chromosome 15q13.1 and Estrogen Receptor–Negative Breast Cancer Survival. Journal of the National Cancer Institute 102, 650-662, doi:10.1093/jnci/djq057 (2010).

225 Sapkota, Y. et al. Germline DNA copy number aberrations identified as potential prognostic factors for breast cancer recurrence. PLoS One 8, e53850, doi:10.1371/journal.pone.0053850 (2013).

226 Jin, G. et al. Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer. Carcinogenesis 32, 1057-1062, doi:10.1093/carcin/bgr082 (2011).

227 Andersen, C. L. et al. Frequent genomic loss at chr16p13.2 is associated with poor prognosis in colorectal cancer. Int J Cancer 129, 1848-1858, doi:10.1002/ijc.25841 (2011).

228 Krishnan, P. et al. Genome-wide profiling of transfer RNAs and their role as novel prognostic markers for breast cancer. 6, 32843, doi:10.1038/srep32843

https://www.nature.com/articles/srep32843#supplementary-information (2016).

229 Krishnan, P. et al. Profiling of Small Nucleolar RNAs by Next Generation Sequencing: Potential New Players for Breast Cancer Prognosis. PLoS One 11, e0162622, doi:10.1371/journal.pone.0162622 (2016).

230 Krishnan, P. et al. Piwi-interacting RNAs and PIWI genes as novel prognostic markers for breast cancer. Oncotarget 7, 37944 (2016).

2 Fine-mapping of a novel premenopausal breast cancer susceptibility locus at Chr4q31.22 in Caucasian women and validation in African women[j]

2.1. Introduction

Breast cancer is the most commonly diagnosed cancer among women worldwide[1,2]. Genome Wide Association Study (GWAS) approaches have identified to-date a total of 172 common low penetrance variants associated with breast cancer risk[3]. SNPs identified by GWAS approaches using high-density genotyping arrays are usually tagSNPs. GWAS-identified SNPs are often in linkage disequilibrium (LD) with putative causal variant(s) contributing to the phenotype[4]. Therefore, it is necessary to comprehensively investigate GWAS-identified loci by fine-scale mapping to identify putative causal variants and their functional significance[5]. While fine-mapping approaches are well described in the literature, it is challenging to elucidate functional relevance of GWAS SNPs, which are predominantly from gene deserts potentially conferring gene regulation. Thus far only 14 breast cancer associated GWAS variants have been fine-mapped and characterized for putative biological roles[6-19].

[j] The GWAS of Breast Cancer in the African Diaspora was conducted by the University of Chicago and supported by the National Cancer Institute (R01 CA142996-02). This manuscript was not prepared in collaboration with investigators of the GWAS of Breast Cancer in the African Diaspora and does not necessarily reflect the opinions or views of University of Chicago, or NCI. The dataset was accessed from dbGaP and study Accession: phs000383.v1.p1. I would like to thank Jennifer Dufour and Preethi Krishnan for providing technical assistance.

A previous study from the Damaraju laboratory reported six putative variants[20] from a GWAS in a Caucasian population (Alberta, Canada), of which four were from different chromosomes showing association with sporadic (>40 years of age at onset and no family history) breast cancer risk. One SNP, rs1429142 on Chr4q31.22, showed consistent associations in two independent replication studies for the overall risk (Stages 1-3, $P=1.5x10^{-7}$ adjusted for BMI; OR 1.28). The GWAS discovery stage (Affymetrix SNP 6.0 array) had 348 cases/348 controls; Stages 2 and 3 replication cohorts had 1,153 cases/1,215 controls[20] and 1,294 cases/2,934 controls, respectively[21]. Analysis based on menopausal status (Stages 1-3) revealed that SNP rs1429142 had an elevated risk for breast cancer among premenopausal women[21] (BMI adjusted p-value of $6.22x10^{-10}$ and $OR_{\text{per-allele}}$ of 1.49) compared to postmenopausal women (BMI adjusted p-value of $7.79x10^{-03}$ and $OR_{\text{per-allele}}$ of 1.17) with a p-value of heterogeneity $< 10^{-03}$.

In the current study, we (i) accessed an additional 1502 breast cancer cases (Stage 4) from Alberta, Canada, and reanalyzed the SNP rs1429142 for overall breast cancer risk in a stratified analysis based on menopausal status; (ii) extended the study to validate findings in women of African ancestry, and (iii) conducted a fine-scale mapping of the Chr4q31.22 locus. The goal was to identify the potential causal variants and their putative functions.

2.2. Methods

I performed all the experiments and analysis, unless otherwise indicated in the text.

2.2.1. Study population

Written informed consent was obtained from all study participants, and the study protocol was approved by the Health Research Ethics Board of Alberta (HREBA)-Cancer Committee. Samples from Alberta, Canada (Internal dataset, Stages 1-4)

The study includes breast cancer cases and apparently healthy control samples recruited from the province of Alberta, Canada. The description of described cases for the Stages 1-3 (age matched 2,750 breast cancer cases and 4472 controls) is available elsewhere[20,21]. The cases were accessed from the Alberta Cancer Research Biobank (http://www.acrb.ca/about-us/), which enrolled patients into the bank between 2001–2005. The study inclusion criteria for cases were: (i) invasive breast cancer, and (ii) non-metastatic at the time of diagnosis. The cases in Stages 1and 2 had no documented family history of breast cancer. For Stage 4 of the study, we accessed independent breast cancer cases (n=1722) diagnosed between 2002 till 2015 from the Alberta Research Tumor Bank and the study inclusion criteria were the same as in the previously described[20,21]. Cases recruited were independent of family history for Stages 3 and 4 to facilitate comparisons of the variants identified by GWAS in a stratified analysis based on family history.

Controls were accessed from the Tomorrow Project, a longitudinal cohort study that is described elsewhere (www. https://myatp.ca/)[20,21]. Inclusion criteria for controls included no personal history of cancer at the time of enrolment, resident of Alberta, Canada, age between 35-69 Y. The controls were progressively followed for incidence of cancer. The control samples from individuals who had developed cancer (n= 201) since the time of enrollment in the study were excluded from the current analysis, bringing the total number of controls to 4271. All case and control subjects were of Caucasian origin.

The biobanks provided buffy coat samples for both cases and controls to isolate germline DNA, and pertinent demographical and patient clinical characteristics (Appendix Table A.1).

2.2.2. Patient demographics

Total sample size (n=9235) for the current study included 4964 (cases) and 4271 (controls). Among the cases, 33% and 67% were pre- and post-menopausal cases (self-declared at the time of diagnosis), respectively. Luminal cancers were predominant (77%) and this frequency was maintained when cases were stratified by menopausal status. Up to 94% of the total breast cancer cases in this study were >40 Y of age. The cases and controls showed similar frequencies for age and BMI distribution (Appendix Table A.1 and Figure A.1).

External datasets

(i) CGEMS: The Cancer Genetic Markers of Susceptibility (CGEMS) case-control study for breast cancer was based on postmenopausal women of European ancestry and is a subset of the longitudinal cohort from the Nurses' Health Study (NHS)[22,23]. The study includes invasive breast cancer cases (n=1,145) and controls (n=1,142) and we analysed rs1429142 (C>T polymorphism) in this cohort, wherein the whole genome data was generated on Illumina HumanHap550 and genotypes of 528,173 SNPs were available in the open access database. Genotype and phenotype information was accessed from dbGaP under study Accession: phs000147.v1.p1. Briefly, the study is based on Caucasian populations, cases had confirmed diagnosis of invasive breast cancer, and controls were matched for age and menopausal status.

(ii) African Diaspora: Dataset for breast cancer GWASs was accessed from dbGaP (Study Accession: phs000383.v1.p1) to analyze rs1429142 (T>C polymorphism). In this population T is the minor allele, whereas C is the minor allele in Caucasian populations. The study includes women of African ancestry (n= 3766) living in Nigeria, Barbados and the United States of America. Genotyping was performed using Illumina HumanOmni2.5-Quad platform. Following data filtering as described below, we retained 2091 controls and 1641 breast cancer cases for association analysis.

2.2.3. DNA extraction and genotyping

Genomic DNA was extracted from buffy coat samples using a commercially available Qiagen Tm kit (Mississauga, Ontario, Canada). Genotyping was performed using Sequenom iPLEX Gold platform (San Diego, CA, USA) and utilized the services provided by McGill University and Genome Quebec Innovation Center, Montreal, Canada.

2.2.4. SNP selection and genotyping

Stage 1 of the study had whole genome genotype data available in Human Affymetrix SNP 6.0 array (906,600 SNPs) for 348 cases and 348 controls. Principal component analysis was used to identify outliers (n=72) and the remaining 624 samples clustered with HapMap population of Caucasian ancestry[20]. I applied a call rate filter (>99%) and assessed for deviations from Hardy-Weinberg equilibrium (cut-off of p<0.001 on controls). I also performed identity by decent analysis[24] based on the genotypes to identify cryptic relatedness (with pairwise correlation $r^2 > 0.25$). Chromosome (Chr) 4 with 40,146 SNPs with genotype calls on Affymetrix array was used for imputation. I used GTOOL for flipping the strand for the SNPs genotyped from the minus strand in Affymetrix to the same strand convention as the reference panel. Followed by strand flipping Chr4 is phased using SHAPEIT algorithm[25] prior to imputation. For imputation we used the best guess method, implemented within IMPUTE2 algorithm[26] and the 1000 Genomes panel based on diverse populations was used as the reference for imputation.

I imputed 952,002 SNPs with imputation info score > 0.7. SNPs imputed were filtered for genotype call rate > 95% and minor allele frequency > 1%. I selected 2019 SNPs in the 1 MB region flanking the index SNP rs1429142 and tagSNP were selected from the locus. Of the 2019 SNPs, 209 are genotypes from the Affymetrix platform and the rest are imputed SNPs. Instead of genotyping all the 2019 SNPs across all samples as cost effective strategy, we selected SNPs that will give coverage across the 1MB region and that enabled second round of imputation in all Stages 1-4 samples. I used Tagger, a SNP selection tool implemented within Haploview ver4.2 and selected 63 tagSNPs. Multiplex assay system on Sequenom iPLEX Gold platform was validated for 56 SNPs (including SNP rs1429142). I genotyped all cases and controls from Stages 1-4, and 4331 case and 4271 controls passed genotyping. (Supplementary Table3). The 56 SNPs (spanning Chr4:147,802,550-148,781,409, Hg19 build) are in LD (r2 > 0.2) with rs1429142. SNP call rates for 56 SNPs were > 92%. I also estimated the imputation and genotyping concordance for these 56 tagSNPs in the Stage 1 samples. All the SNPs had a correlation (r^2) of > 0.80, of which 44 SNPs had r^2 of > 0.90. I included several technical replicates for each SNP, and genotype concordance was 100%. I estimated the concordance between genotyping batches (previous genotype calls for Stage 1-3 samples) which also showed 100% concordance.

I re-imputed data from 56 SNPs and from pre-menopausal cases (n=1503) and controls (n=4271), as the focus of this investigation was on assessing breast cancer risk and replicating previous findings. I imputed 1715 SNPs using one-phase imputation approach with imputation info score value > 0.7. After applying genotyping quality filter, 587 SNPs were retained with 85% genotype call rate and minor allele frequency ≥5% for fine-mapping association analysis.

2.2.5. Statistical analysis

I used correlation/trend test for allelic correlation tests with one degree of freedom (d.f) for unadjusted analysis in the association study between cases and controls. Unconditional logistic regression was used to estimate the odds ratio (OR) with 95% confidence interval (adjusted for BMI). Subgroup analysis were carried out based on menopausal status, disease stage (I, II versus III), grade (high versus low), molecular subtype (luminal A versus non-luminal). P-heterogeneity was estimated between the subgroups. All association analyses were performed using Golden Helix SNP & Variation suite and Plink v1.07 [27]. Conditional logistic regression analysis was conducted with adjustments for the highly associated variants (rs13134510, rs1366691, rs1429139 and rs12501429) using binary logistic regression analysis in PLINK. Likelihood ratio analyses were carried out using IBM SPSS Statistics (IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp) to identify the potential casual variants. The top associated SNP rs13134510 was used as a reference, to test fine-mapped SNPs with 4 degrees of freedom. I excluded SNPs with p-value > 0.01.

2.2.6. *In silico* predictions for functional relevance of the fine-mapped SNPs

To elucidate the functional relevance, we annotated a total of 130 breast cancer risk variants (p-value < 0.05). The annotation used data from ENCODE (Encyclopedia of DNA Elements)[28], Roadmap Epigenomics consortium[29] available through Regulome DB ver1.1[30], HaploReg v4.1[31] and Washington University Epigenome Browser (https://epigenomegateway.wustl.edu/). I scored all 130 variants using RegulomeDB, variants with scores of 1- 4 were considered and these variants were annotated for histone marks such as H3K4me1, H3K4Me3 indicative for enhancer and promoter activity respectively. I used the histone marks data generated in normal breast epithelial cell lines such as Mammary Epithelial Primary Cells (HMEC), Breast variant Human Mammary Epithelial Cells (vHMEC) and Breast Myoepithelial Primary Cells. I also utilized datasets for DNase Hypersensitivity sites informative about the open chromatin state in the breast epithelial cell lines. For transcription factor (TF) binding, we used the ChIP-seq datasets generated for the breast cancer cell lines MCF10A-ER-Src, HMEC and MCF7. Polymorphisms potentially affecting the TF binding motifs were predicted using position weighted matrix (PWM) for each variant, when applicable. I accessed the encode Hi-C datasets for HMEC and ChIA-PET data for POL2A and CTCF in the MCF-7 cell line. TAD domain predictions based on the Hi-C data was predicted using the 3D genome browser[32] (http://promoter.bx.psu.edu/hi-c/view.php). Interaction arcs based on the ChIA-

PET data was generated based on Washington University Epigenome Browser. I also captured the expression of nearby genes (~2MB spanning the SNP rs1429142) based on the RNA-Seq for the HMEC cell line.

2.2.7. Expression quantitative trait loci (eQTL) analysis

eQTL data for normal breast tissues and heart left ventricle were used for the interpretation of the results based on GTEx database (GTEx portal was accessed on 07/04/2018, GTEx analysis V7 (dbGaP Accession phs000424.v7.p2)). eQTL based on lymphoblastoid cell lines were inferred from ENCODE project.

2.3. Results

2.3.1. Association of SNP rs1429142 at Chr4q31.22 with overall and premenopausal breast cancer risk in Caucasian women

I replicated the association of the previously identified SNP rs1429142 (C/T) with breast cancer risk among Caucasian women. The SNP is located at Chr4:148289398 (GRCh37/hg19), with minor allele 'C' (frequency, MAF ~18%) among the Caucasian population. The association p-value (adjusted for BMI) for overall breast cancer risk (Stage 4) was 1.20x10^{-04} with ORs of 1.23 [1.11-1.37] (Table 2.1). In the combined analysis for overall breast cancer risk (Stages 1-4; total n= 4331 cases/4271 controls), SNP rs1429142 showed genome level significance with adjusted p-value 4.35x10^{-08} and OR of 1.25 [1.15-1.35]. The genome wide significance threshold was calculated based on testing 782,838 SNPs for association in Stage I study (0.05/782,838=6.4x10^{-8}).

In a subgroup analysis (samples from Stages 1-4) based on menopausal status, the association of rs1429142 with premenopausal breast cancer risk in women of Caucasian ancestry reached genome level significance with adjusted p-value of $5.81\text{x}10^{-10}$ and OR of 1.40 [1.26-1.56]. However, the association among postmenopausal women of Caucasian ancestry was moderate even upon adjusting for BMI (OR of 1.17 [1.07-1.28], p-value of $7.81\text{x}10^{-04}$) (Table 2.1). The p-value for the test of heterogeneity comparing the ORs between premenopausal and post-menopausal women was statistically significant at $1.84\text{x}10^{-02}$ (Table 2.2), consistent with the earlier findings[21].

Data in Appendix Table A.1 summarize the patient demographic data for the study samples, Stages 1-4. SNP rs1429142 was initially shown to be associated with sporadic breast cancer (Stages 1 and 2). In subsequent replication studies (Stages 3 and 4), we recruited cases irrespective of family history. Association analysis of SNP rs1429142 based on family history in all stages 1-4, showed a trend of elevated risk and stronger association of SNP with sporadic breast cancer (n= 1886 cases /4271 controls, p-value $5.09\text{x}10^{-8}$ OR 1.31) compared to cases with family history (n=1640 cases/4271 controls, p-value 1.86 $\text{x}10^{-4}$ OR 1.21) (Table 2.2). Even though, the p-value of heterogeneity (p-het 0.37) between these strata is not significant, the trend of association validates the study premise. Other subgroup analysis based on clinicopathological features such as molecular subtype (luminal versus non-luminal), tumor grade (high versus low), and stage (<III versus ≥III) were also considered. None of these associations showed trends of elevated risk between the strata and the p-value for heterogeneity was not significant (Table 2.2).

I analyzed the association of SNP rs1429142 in the Cancer Genetic Markers of Susceptibility dataset (CGEMs; 1144 cases/1143 controls) consisting of all postmenopausal women as study subjects. SNP rs1429142 showed modest risk, OR 1.05; p-value - 6.8x10^{-01} (Table 2.1).

2.3.2. Association of SNP rs1429142 with premenopausal breast cancer risk in women of African ancestry

The association of SNP rs1429142 was tested using datasets from the African Diaspora study. SNP rs1429142 has a T/C polymorphism in the African population with a minor allele (T) frequency of 25%. Since the C allele is a risk allele in Caucasian population, the data represented for the association study findings are in reference to the C allele. I initially tested rs1429142 in 1607 cases/2041 controls for overall risk of breast cancer and the SNP did not show statistically significant association (p-value 6.08x10^{-01}). The C allele showed trends for risk (1.08 [0.92-1.14]). Interestingly, in the stratified analysis, SNP rs1429142 was associated with breast cancer risk among premenopausal women and the C allele showed risk (p-value 1.45x10^{-02}; OR of 1.2 [1.03-1.40]). Risk for postmenopausal women was not statistically significant (8.56x10^{-01}, OR of 1.01 [0.87-1.17]).

Therefore, based on this study findings, I report a novel premenopausal risk variant with a moderately high effect size for breast cancer in the Caucasian population (OR 1.40). This variant was validated in premenopausal African women (Table 2.1). These findings warrant further fine-scale mapping of the locus to identify potential causal variant(s) and their putative roles in conferring breast cancer susceptibility.

Table 2.1 Replication and validation of SNP rs1429142 at Chr4q31.22 and association with premenopausal breast cancer risk

	Sample size, n	**Status**	**Risk Allele /Allele frequency**	**P-value**	**Allelic OR [95% CI]**
Replication (Caucasian population)					
Caucasian, Stages 1-3[a] (Canada)*	2829 cases/4271 controls	Overall	C/0.18	6.17E-07	1.26 [1.15-1.38]
Caucasian, Stage 4[a] (Canada)	1502 cases/4271 controls	Overall	C/0.18	1.20E-04	1.23 [1.11-1.37]
Caucasian, Stages 1-3[a] (Canada)	**4331 cases /4271 controls**	**Overall**	**C/0.18**	**4.35E-08**	**1.25 [1.15-1.35]**
	1503 cases /4271 controls	**Premenopausal**	**C/0.17**	**5.81E-10**	**1.40 [1.26-1.56]**
	2700 cases /4271 controls	Postmenopausal	C/0.18	7.81E-04	1.17 [1.07-1.28]
Caucasian (CGEMs study)	1144 cases /1143 controls	Postmenopausal	C/0.17	6.80E-01	1.05[0.89-1.22]
Validation (Diverse population)					
African Diaspora	1607 cases /2041 controls	Overall	C/0.75	6.08E-01	1.03 [0.92-1.14]
	645 cases /2041 controls	**Premenopausal**	**C/0.75**	**1.45E-02**	**1.21 [1.04-1.40]**
	663 cases /2041 controls	Postmenopausal	T/0.75	8.56E-01	1.01 [0.88-1.17]

*Indicates the association analysis adjusted for Body Mass Index (BMI) available for cases and controls in Canadian populations. BMI information was not available or missing for several samples for other cohorts. *Data for Stages 1-3 was based on reanalyzed samples from a previous study[21] and SNP rs1429142 was independently genotyped, taking into account the longitudinal follow-up on cases and controls as described in methods. Replication of the association with respect to menopausal status in the Caucasian population is indicated using internal dataset (Stage 4 and Stages 1-4 combined analysis) and CGEMS cohorts. Validation study utilized African population. For SNP rs1429142, the minor allele is C in the Caucasian whereas it is T in the African population (T/C). Note that the frequencies of the minor alleles across the populations are different. The results are presented with respect to the risk allele ‘C’.

Table 2.2 Association of SNP rs1429142 at chr4q31.22 with breast cancer risk

	Sample size (cases/ controls)	Adjusted analysis (allelic)		P_{het}
		P-value	OR [95% CI]	
Family history				
Yes	1640/4271	1.86E-04	1.21 [1.10-1.35]	3.69E-01
No	1886/4271	5.09E-08	1.31 [1.19-1.44]	
Subtype				
Luminal A cases	2421/4271	7.16E-07	1.22 [1.11-1.34]	6.60E-01
Non luminal cases	1058/4271	7.48E-04	1.30 [1.12-1.51]	
Grade				
High	1582/4271	1.89E-06	1.28 [1.16-1.42]	4.99E-01
Low	2074/4271	4.18E-05	1.22 [1.11-1.34]	
Stage				
Stage <III	3472/4271	3.03E-07	1.24 [1.14-1.34]	1
Stage >III	1013/4271	2.48E-04	1.45 [1.19-1.78]	

All analysis was adjusted for BMI. This table represents the association of SNP rs1429142 across different subgroups and the P_{het} is indicated.

2.3.3. Identification of potential causal variants by fine-scale mapping of Chr4q31.22

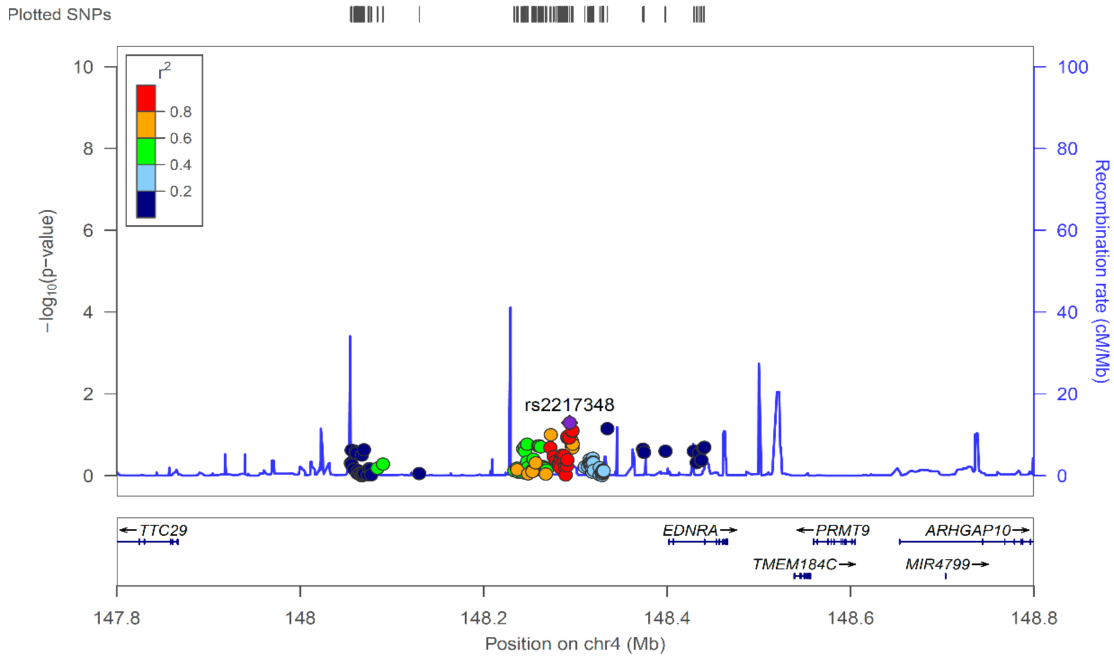
Fine-scale mapping of SNP rs1429142 was performed to identify putative causal variants. I fine-mapped a ~1 MB region, 147802550 to 148781409 (GRCh37/hg19) flanking the SNP, rs1429142 located at Chr4:148289389. Whole genome imputation of Chr4 was performed for the Stage 1 samples for which the data from the Affymetrix Human SNP 6 array was available. IMPUTE2 algorithm was used for imputation and the 1000 Genomes data (multiethnic populations) as a reference panel as recommended elsewhere[33].

Following imputation, imputed genotype data with an imputation info score >0.7, call rate > 95% and MAF >1% were retained. I selected 2019 SNPs within 1 MB region for further analysis. A total of 63 Tag SNPs (see methods for SNP selection strategy) were selected using the HAPLOVIEW algorithm. Selected SNPs were genotyped in all samples (cases and controls) from Stages 1-4 of which 56 SNPs were amenable for multiplex genotyping and passed the internal quality control criteria (Appendix Table A.3). Based on the 56 genotyped tagSNPs, I re-imputed (one phased imputation method) for all premenopausal cases and all controls. A total of 1715 SNPs with an imputation info score value > 0.7 were obtained and 587 SNPs were retained based on > 85% genotype call rate and MAF $\geq$5%.

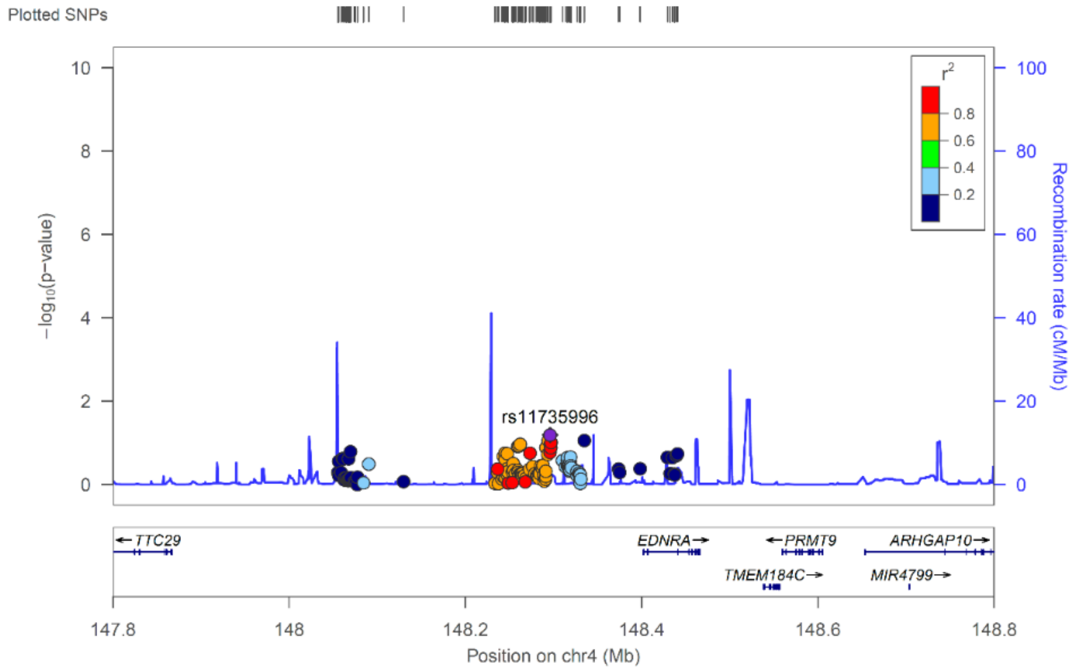
Association testing of 587 fine-mapped SNPs in the premenopausal cases and controls identified 135 SNPs with p-values of < 0.05 and 49 SNPs at < 10^{-8} (Figure 2.1a and Appendix Table A.2, p-values unadjusted and adjusted for BMI). There were four SNPs (rs13134510, rs1366691, rs1429139 and rs12501429) showing highest association with p-values of < 10^{-11}. All these four fine-mapped SNPs are in LD with the originally identified SNP rs1429142. SNP rs13134510 showed highest statistical significance (unadjusted p-value $1.11x10^{-12}$). Conditional regression analysis based on these four SNPs did not reveal any additional independent signals (Figure 2.2: a-d and Appendix Table A.4).

Figure 2.1 Association of the fine-mapped SNPs with premenopausal breast cancer risk and their functional annotation

This figure represents the association of the fine-mapped SNPs with premenopausal breast cancer risk and the functional relevance of the SNP is indicated in cell line data. (a.) The locus zoom plot indicates the association p-value (log scale) on the Y-axis and genomic location on the x-axis. The 587 fine-mapped SNPs are represented as squares (imputed) and circles (genotyped), and the LD (r2) between the SNPs were indicated according to the color scale. The GWAS SNP rs1429142 is indicated. (b) The functional relevance of the fine-mapped SNPs was indicated using human breast cell lines (HMEC, HMF and MCF-7). The DNase hypersensitive sites (HMEC, HMF), histone marks (HMEC and MCF-7) and chromatin states (Encode cell lines) were inferred from corresponding cell lines. The SNPs with RegulomeDb score (1-4) are indicated.



Conditional regression on the top SNP rs13134510



Conditional regression on the top SNP rs1366691

Figure 2.2 Conditional regression analysis

The data in this figure (a-d) represents the conditional regression plots generated based on the top four associated SNPs. Each plot represents the analysis adjusted for (a) rs13134510, (b) rs1366691, (c) rs1429139 and (d) rs12501429. The plot represents the association of the fine-mapped SNPs after conditioning. The Y-axis represents the p-value in logarithmic scale and genomic co-ordinates on the X - axis. Conditional regression analysis did not reveal any additional independent signal.

Multiple methods, tools and annotation algorithms were used assess the functional relevance of the associated and fine-mapped SNPs and described below.

(i) ***Log likelihood ratio analysis***- This was carried out as an independent pruning method which revealed five SNPs with a p-value of >0.05. These five SNPs were excluded and the remaining 130 SNPs (including the top four SNPs showing highest association) were identified as potentially causal variants showing a statistical significance at $p < 0.01$ (Appendix Table A.5).

(ii) ***LD*** mapping- Given the expected small LD block patterns in African populations and the statistical significance observed among premenopausal women. The fine-mapped region (130 SNPs) was refined based on the LD block patterns using the HapMap dataset. I noted that the Caucasian population had fewer but larger LD blocks consisting of the fine-mapped SNPs and the GWAS SNP rs1429142 (Figure. 8a). As expected, we observed multiple smaller LD blocks in African populations in the fine-mapped region in contrast to Caucasian populations. The fine-mapped variants (130 SNPs) were scattered across multiple LD blocks in African populations. In the African population, ten of the highly significant fine-mapped SNPs (p-values < 10^{-10}) (rs1366691, rs1429139, rs12501429, rs1583003, rs2163012, rs2163011, rs12498595, rs13120678, rs1366679, rs13134510) were clustered in a single LD block and the remaining SNPs, including the GWAS index SNP rs1429142, were scattered over multiple LD blocks (Figure. 8b). This contrasts with the Caucasian population wherein the index SNP along with the ten highly associated SNPs were found in a single LD block.

Figure 2.3 Linkage disequilibrium plot for the fine-mapped locus chr4q22.31 in Caucasian and African population

The LD plot for the (a) Caucasian and (b) African populations for the fine-mapped SNPs generated based on HapMap populations. The fine-mapped SNPs indicated as **(▼) are** highly associated (p-value < 10^{-08}) with premenopausal breast cancer risk and GWAS SNP rs1429142 is indicated as (▼). The GWAS SNP is in a different but nearby LD block to the fine-mapped region in both populations. The LD plots were generated using the tool (https://snpinfo.niehs.nih.gov/snpinfo/snptag.html).

(iii)*Putative regulatory functions for the causal variants*- I have annotated all 130 variants for functional relevance. I used RegulomeDB-ver1.1 (Appendix Table A.6 and A.7) and HaploReg-v4.1 (Appendix Table A.8) for functional annotations. I identified 19 SNPs (Appendix Table A.7) with Regulome scores between 1 to 4 (1 being the most informative); these are derived from composite scores from the inferred regulatory functional states such as DNase hypersensitivity sites, transcription factor binding, chromatin state, histone marks and changes in binding motifs of bound proteins. Among the 19 SNPs with putative regulatory functions, five SNPs (with p-values) were predicted to have enhancer roles inferred from chromatin marks (post translational modification of histone protein): rs1366691 ($1.91\text{x}10^{-12}$), rs1429139 ($6.64\ \text{x}10^{-12}$), rs7667633 ($5.05\text{x}10^{-08}$), rs6836670 ($1.41\text{x}10^{-07}$) and rs17023196 ($1.01\text{x}10^{-04}$). The combination of the chromatin marks was used to predict enhancer functions using the method chromHMM (multivariate hidden Markov model). The chromatin state at the locus of interest harbored the histone marks: H3K4me1, H3K27ac, and H3K9ac, captured by ChIP-seq assay in normal breast cell lines: Mammary Epithelial Primary Cells (HMEC) and Breast variant Human Mammary Epithelial Cells (vHMEC) (Appendix Table A.8). There was evidence of DNase hypersensitivity peaks near these SNPs captured in HMEC, vHMEC and Breast Myoepithelial Primary Cells (Appendix Table A.8).

Among the 19 SNPs that were annotated for putative regulatory functions, SNPs rs1568136, rs6821368 and rs6822565 were present within the intron of the *EDNRA* gene. The histone marks at these loci indicated weak transcriptional activity in HMEC, vHMEC and Breast Myoepithelial Primary Cells. Additionally, SNP rs1568136 affected binding of transcription factors such as EN1 and SNP rs6821368 affected binding of NF-AT, SOX, HDAC2, HOXA4, PAX-4, POU2F2, POU3F2, and SIN3AK-20 (Appendix Table A.8) judged from the Position Weighted Matrix (PWM) scores.

(iv)*Binding of transcription factors at the SNP sites-* The dataset from the ENCODE project offered further insights into binding of transcription factors (TFs) at three SNPs, rs1366691, rs7667633 and rs7668383. Evidence for binding of three TFs (FOS, STAT3 and POL2A) at these sites was obtained from the MCF10A-Er-Src cell line (derived from parental MCF-10A cells which are negative for estrogen receptor expression). However, MCF10A-Er-Src contains a variant of the Src kinase oncoprotein that is fused to the ligand binding domain of the estrogen receptor and is induced by adding Tamoxifen (TAM) (Appendix Table A.7). Src expression leads to transformation of cells as evidenced by visible morphological changes between 24-36 hours. ENCODE project has also captured binding of TFs to target sites in TAM treated and untreated cells at 4-hr,12-hr and 36-hr time intervals. Based on ChIP-sequencing, FOS binding was noted to be high at rs1366691, rs7667633 and rs7668383 loci in the TAM-treated group relative to the untreated group when analysed at different time intervals in the MCF10A-Er-Src cell line (Figure 2.4).

In summary, the evidence presented from the various methods described above indicated that a select number of SNPs (*i and ii*) among the fine-mapped region appeared to be active enhancer domains judged from the collective experimental evidence (*iii* and *iv*) from various cell lines (epigenetic marks and transcription factor binding). Three SNPs, rs1366691, rs1429139 (p-value $<10^{-10}$) and rs7667633 (p-value 10^{-08}) were identified to be the likely causal SNPs. I based the conclusions on the combined evidence from strengths of association and functionality as enhancers (inferred from chromatin state and binding of transcription factors). These loci may exhibit complex long or short-range DNA interactions, and such interactions between the enhancer(s) and promoters may contribute to the overall regulatory effects.

Figure 2.4 Transcriptional activity at the fine-mapped locus

The figure represents transcriptional activity at the fine-mapped locus. The binding of the transcription factors (left top corner) were determined using ChIP-Seq data capturing the binding of -fos, STAT1/3 and Pol2/3 were described in breast cell lines (MCF10A-Er-Src, HMEC) and Encode cell lines. Similarly, transcriptional activity (left bottom panel) estimated from the

RNA-seq data generated in HMEC cell line. The binding of the transcription factors (right-side top) such as EN1, SOX and NF-AT may potentially be affected by polymorphism in the intron of the EDNRA gene estimated from position weighted matrix. The source of the data is shown in the column (ChIP-seq for c-FOS, POL2, STAT3 based on MCF10A-Er-Src were generated from Harvard, for the encode cell lines: c-FOS captured in HUVEC from University of Southern California; STAT1 captured in GM12878 from Stanford University; C-FOS and Pol3 captured in GM12878 from Yale University. Figure was generated based on the output from the browser http://epigenomegateway.wustl.edu/browser/

2.3.4. Gene regulation by short range DNA interactions

The fine-mapped region was interrogated for possible short-range interactions based on the Hi-C data available for the HMEC cell line. The fine-mapped regions harbored multiple interactions with the neighboring region and were predicted to be present within Topologically Associated Domains (TADs) (Figure 2.5a). TADs consist of regions of DNA that preferentially interact with each other. The interactions are predominantly seen within the TAD boundaries and are less likely to interact outside of the TAD[34]. Since TADs are derived by complex DNA looping and interactions, they play a role in gene regulation, wherein promoters interact with local enhancer elements. CCCTC-binding factor (CTCF) and Cohesin (a multi-subunit protein complex) are the common DNA binding proteins known to be enriched in TAD regions. DNA looping is mediated by the binding of CTCF proteins and that brings about the physical contact of the DNA domains. I analysed the data from the Chromatin Interaction Analysis by Paired End Tag (ChIA-PET) data generated from MCF-7 enriched for CTCF and POL2 (Figure 2.5b). I observed multiple interactions between fine-mapped SNPs and upstream promoter elements of nearby genes including *EDNRA, PRMT10, ARHGAP10 and TMEM18C* (potential eQTLs, Appendix Table A.9). Further experiments are needed to gain mechanistic insights on the regulation of the target genes and interactions with the identified potential causal variants.

Figure 2.5 TADs and short-range interactions captured by Hi-C and ChIA-PET data

This figure represents Topological Associated Domain (TAD) and short range interactions in the fine-mapped region (chr4: 147000000-149000000) estimated from Hi-

C and ChIA-PET dataset in breast cell lines. (a.) The TADs were predicted based on the Hi-C in HMEC cell line, the heat map presents the frequency of the interaction, and the intensity of the heatmap varies according to the interaction frequency (http://promoter.bx.psu.edu/hi-c/view.php). (b) The short-range interactions indicated by arcs; estimated in MCF-7 cell lines and ChIA-PET enriched for POL2 and CTCF proteins.

2.4. Discussion

I report three potential causal variants (rs1366691, rs1429139 and rs7667633) from fine-mapping and annotation analysis which are strongly associated with premenopausal breast cancer risk. The effect size for the three novel variants are in line with the originally described index SNP rs1429142 (OR 1.4, Table 2.1 and Appendix Table A.2). Analysis of the GWAS literature identified fewer variants with effect sizes in the range 1.25-1.4, largely from familial breast cancers and breast cancers in postmenopausal women. The index SNP rs1429142, which was originally described to be associated with sporadic breast cancer, also showed association among cases with family history, albeit at lower risk than the sporadic cases in the combined analysis of Stages 1-4 , confirming the original findings[21]. In stratified analyses based on disease stage, grade and ER status, the SNP rs1429142 did not show differences in risk between the groups (Table 2.2). FGFR2 variants and others from previous GWAS literature which were known to confer risk in familial cases were also reproduced in a this study samples (Stage 1-3), i.e., the effect size was higher in familial cases than in sporadic cases[21].

A previous study from the Damaraju laboratory reported a SNP from GWASs (Stages 1-3), rs1429142 at Chr4q31.22, to be a novel locus associated with breast cancer risk and that the risk was higher for premenopausal women[21]. In this study, I further replicated the association of SNP rs1429142 with breast cancer risk using an independent set of breast cancer cases (Stage 4). In the combined analysis of all Stages (1-4, n = 4331 cases and 4271 controls) for overall risk, SNP rs1429142 reached genome level significance at p-value $4.35x10^{-08}$ with an OR of 1.25 [1.15-1.35]. In the stratified analysis based on menopausal status, SNP rs1429142 showed strong association with premenopausal breast cancer, p-value $5.81x10^{-10}$ with an OR 1.40 [1.26-1.56] (Table 2.1). The overall breast cancer risk conferred by SNP rs1429142 was not affected by luminal status, tumor grade or stage (Table 2.2). In an independent analysis, the SNP rs1429142 did not show elevated risk to estrogen receptor (ER) status (ER positive vs. ER negative cases, Table 2.2). The majority of the GWAS identified SNPs in earlier studies were shown to confer risk in women with ER positive disease[35,36] and in postmenopausal cases[22].

I stratified cases based on menopausal status to identify risk with an emphasis on identifying risk variants for breast cancer in women with age of onset of disease >40Y, which has hitherto not been addressed in the breast cancer GWAS literature. Limited GWASs addressed sporadic breast cancer without emphasis to menopausal status [22,37,38], or those that addressed focused predominantly on postmenopausal women with familial component. I have validated the study premise by analyzing the postmenopausal cohort from CGEMS and showed that SNP rs1429142 was not associated with breast cancer risk, lending credence to the observations on premenopausal women. In a previous study[21], the association of the literature reported GWAS SNPs were replicated, and these SNPs showed stronger association with familial breast cancer risk in the study population (Alberta, Canada) stratified based on family history. I replicated these findings in that the effect size was higher in familial cases compared to sporadic breast cancer cases (age of onset of disease >40 and no family history of breast cancer). However, the literature-reported SNPs did not show elevated risk when cases were stratified based on menopausal status[21]. These findings, taken together, demonstrate that the variant rs1429142 described in this study is novel and confers breast cancer risk in premenopausal women.

Among African populations, an allele reversal was noted wherein C is the major allele and T is the minor allele with 75% and 25% frequencies, respectively. In the overall association, SNP rs1429142 was not associated with breast cancer, however in the subgroup analysis its association was significant among premenopausal breast cancer risk (p-value < 0.05). The C allele remained the risk allele across different populations irrespective of its association with breast cancer risk (Table 2.1), an observation that aligns with the higher prevalence of premenopausal breast cancer among women of African ancestry[39-42].

In the fine-scale mapping of the associated region at the Chr4q31.22 locus, we identified 587 SNPs within the 1Mb region flanking SNP rs1429142. Of the 587 SNPs, 135 were associated with premenopausal breast cancer risk. Conditional regression analysis did not reveal any independently associated signals. Likelihood analysis retained 130 as putatively causal SNPs with p-values < 0.01. The fine-mapped region and the SNPs showing association with premenopausal breast cancer risk were present within fewer but large LD blocks in the Caucasian population, whereas there were multiple but smaller LD blocks for the same region in the African population. These findings agree with the higher level of recombination events and resultant decay of LD in African populations (Figure 2.3). Consistent with current knowledge of LD in diverse populations,

Functional scoring revealed five SNPs (rs1366691, rs1429139, rs7667633, rs6836670 and rs17023196) at highest predicted levels of functionality (i.e., as enhancers). The DNase hypersensitivity peaks revealed an open chromatin state at these loci. In addition, the histone methylation pattern, H3K4me1 and acetylation of H3K9ac and H3K27ac suggested potential enhancer roles based on HMEC, vHMEC and breast myoepithelial primary cell lines. To decipher transcription factors binding at these loci, we utilized the ChIP-Seq data from ENCODE for the MCF10-src cell line. The characteristic feature of MCF10-Src cells is that upon transformation by Tamoxifen induction, the cells exhibit increased motility, invasion, formation of foci, formation of single cell colonies, mammospheres and confer formation of tumor in mouse xenografts[43,44]. Based on the ENCODE data, transcription factors including FOS, STAT3 and POL2RA were bound to SNPs rs136691, r7667633 and rs7668383 from among the fine-mapped loci. These results suggested active enhancer regions at the putative causal loci which potentially regulate the expression of downstream target genes flanking the index SNP. For instance, the nearest target gene identified was EDNRA, located 2 kb gene downstream of putative causal SNP rs1366691.

STAT3 protein is a well characterized transcription factor implicated in many cancer types[45-47]. STAT3 expression alone was sufficient to initiate tumorigenesis, and its over expression brings about transformation of both human fibroblast [48] and MCF10 derived (MCF10-ER-Src) [49]cell lines. Induction of Src expression transforms the cells, conferring the phenotypic changes characteristics of cancers [43,44]. The process of transformation involves epigenetic switch and inflammatory pathway gene expressions. STAT3 exclusively binds to open chromatin regions and regulates expression of NFKB1 which in turn regulates expression of IL6, a cascade of events that is part of the well characterized feed-back loop involving these transcription factors and inflammatory mediators[50]. Often STAT3 and FOS proteins coregulate the transcription of genes. In this study, STAT3 and FOS bound to the sequences at SNP sites, rs1366691 and rs7667633 in the MCF10-ER-Src cell line during the process of transformation.

Since the fine-mapped variants were predicted to have an enhancer function, they are likely to influence promoters of the nearby genes by DNA looping. Based on the DNA interaction profiles generated in HMEC cells, we confirmed that the fine-mapped loci have multiple local interactions and were present within TAD domains. TAD domains, which were recently described[34], consist of regions of DNA that are likely to interact with each other within the TAD boundaries. These are complex mechanisms of gene regulation and TAD domains are conserved across the tissues and species[34,51].

Several SNPs from the fine-mapped region appeared to be eQTLs (in different tissues other than breast) regulating nearby genes ENDRA, ARHGAP10 present within ~800kb (Appendix Table A.9). ENDRA is well known for its role in vasoconstriction and in arterial diseases. However these genes are also often noted to be dysregulated in cancer; EDNRA bound by endothelin-1 triggers a cascade of signaling pathways leading to proliferation[52,53], angiogenesis[54], invasion/ tumor progression[55,56] and inhibition of cell death[57,58], when activated by Hypoxia induced factor1-Alpha. Overexpression of EDNRA has been observed in several cancer types[53,56,57] and is an independent predictor of prognosis[59]. Similarly, ARHGAP10 belongs to the family of Rho GTPase-activating proteins that are known to play a role in cell cytoskeleton organization, cellular migration and adhesion, regulation of transcription[60]. ARHGAP10 was associated with invasive breast cancer prognosis[61], pediatric leukemia[62], and ovarian[63] and lung cancers[64]. ARHGAP10 is often downregulated in tumors and may play a role as a tumor suppressor[63,64]. The eQTL role for the fine-mapped variants in breast tissues warrants further work and is recognized as a potential limitation for generalizability of the findings.

The fine-mapped variants in this study are common polymorphisms (MAF 18%). A higher sample size might have enabled the identification of low frequency putative causal variants within the susceptibility locus to gain additional biological insights[5,18]. Due to the challenges in the functional characterization of the fine-mapped loci, only a limited number of breast cancer studies successfully identified the target genes (FGFR2[11], CCND1[10],MAP3K1[13], TERT[9], IGFBP5[12], TET2[14], STXBP4[16]) with role in breast cancer etiology.

In summary, we have identified three potential causal variants (rs1366691, rs1429139, rs7667633) strongly associated with premenopausal breast cancer risk and the variants appear to have enhancer functions, likely regulating the nearby target genes. Further experimental evidence is needed to elucidate the mechanism by which these genes may increase the risk for breast cancer among premenopausal women. The novel locus associated with premenopausal breast cancer in this study and a fine-mapping analysis of the locus revealed binding of transcription factors known to play a role in inflammatory pathways, also a common etiological basis of many cancers.

2.5. References

1 Canadian Cancer Society. Breast Cancer Statistics. (2017).

2 Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. CA Cancer J Clin 64, doi:10.3322/caac.21208 (2014).

3 Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42, D1001-1006, doi:10.1093/nar/gkt1229 (2014).

4 McCarthy, M. I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9, 356-369, doi:10.1038/nrg2344 (2008).

5 Manolio, T. A. et al. Finding the missing heritability of complex diseases. Nature 461, 747-753, doi:10.1038/nature08494 (2009).

6 Udler, M. S. et al. FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. Human molecular genetics 18, 1692-1703, doi:10.1093/hmg/ddp078 (2009).

7 Udler, M. S. et al. Fine scale mapping of the breast cancer 16q12 locus. Human molecular genetics 19, 2507-2515, doi:10.1093/hmg/ddq122 (2010).

8 Chen, F. *et al.* Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. *Human molecular genetics* **20**, 4491-4503, doi:10.1093/hmg/ddr367. (PMID: 21852243) (2011).

9 Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, doi:10.1038/ng.2563 (2013).

10 French, JulietÂ D. *et al.* Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. *American Journal of Human Genetics* **92**, 489-503, doi:10.1016/j.ajhg.2013.01.002 (2013).

11 Meyer, KerstinÂ B. *et al.* Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1. *American Journal of Human Genetics* **93**, 1046-1060, doi:10.1016/j.ajhg.2013.10.026 (2013).

12 Ghoussaini, M. *et al.* Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat Commun* **4** (2014).

13 Glubb, Dylan M. *et al.* Fine-Scale Mapping of the 5q11.2 Breast Cancer Locus Reveals at Least Three Independent Risk Variants Regulating MAP3K1. *The American Journal of Human Genetics* **96**, 5-20, doi:10.1016/j.ajhg.2014.11.009 (2015).

14 Guo, X. et al. Fine-scale mapping of the 4q24 locus identifies two independent loci associated with breast cancer risk. Cancer Epidemiol Biomarkers Prev 24, 1680-1691, doi:10.1158/1055-9965.EPI-15-0363 (2015).

15 Orr, N. et al. Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Human molecular genetics 24, 2966-2984, doi:10.1093/hmg/ddv035. (PMID: 25652398) (2015).

16 Darabi, H. et al. Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). Scientific Reports 6, 32512, doi:10.1038/srep32512 (2016).

17 Horne, H. N. et al. Fine-Mapping of the 1p11.2 Breast Cancer Susceptibility Locus. PLOS ONE 11, e0160316, doi:10.1371/journal.pone.0160316 (2016).

18 Shi, J. et al. Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. Int J Cancer 139, 1303-1317, doi:10.1002/ijc.30150 (2016).

19 Zeng, C. et al. Identification of independent association signals and putative functional variants for breast cancer risk through fine-scale mapping of the 12p11 locus. Breast Cancer Research 18, 64, doi:10.1186/s13058-016-0718-0 (2016).

20 Sehrawat, B. et al. Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human genetics 130, 529-537, doi:10.1007/s00439-011-0973-1 [doi] (2011).

21 Sapkota, Y. et al. Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm. PLoS One 8, e62550, doi:10.1371/journal.pone.0062550 (2013).

22 Hunter, D. J. et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics 39, 870-874, doi:ng2075 [pii] (2007).

23 Haiman, C. A. et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nature genetics 43, 1210-1214, doi:10.1038/ng.985 [doi] (2011).

24 Pang, A. W. et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11, R52, doi:10.1186/gb-2010-11-5-r52 (2010).

25 Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. Nat Methods 9, 179-181, doi:10.1038/nmeth.1785 (2011).

26 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

27 Purcell, S. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81, 559-575, doi:10.1086/519795.

28 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74, doi:10.1038/nature11247 (2012).

29 Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317-330, doi:10.1038/nature14248 (2015).

30 Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research 22, 1790-1797 (2012).

31 Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Research 40, D930-D934, doi:10.1093/nar/gkr917 (2012).

32 Wang, Y. et al. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. bioRxiv (2017).

33 Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. G3 (Bethesda) 1, 457-470, doi:10.1534/g3.111.001198 (2011).

34 Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376-380, doi:10.1038/nature11082 (2012).

35 Lilyquist, J., Ruddy, K. J., Vachon, C. M. & Couch, F. J. Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. Cancer Epidemiol Biomarkers Prev 27, 380-394, doi:10.1158/1055-9965.EPI-17-1144 (2018).

36 Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92-94, doi:10.1038/nature24284 (2017).

37 Stacey, S. N. et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics 39, 865-869, doi:ng2064 [pii] (2007).

38 Zheng, W. et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nature genetics 41, 324-328, doi:10.1038/ng.318 [doi] (2009).

39 Bharat, A., Aft, R. L., Gao, F. & Margenthaler, J. A. Patient and tumor characteristics associated with increased mortality in young women (< or =40 years) with breast cancer. J Surg Oncol 100, 248-251, doi:10.1002/jso.21268 (2009).

40 Jedy-Agba, E. et al. Cancer incidence in Nigeria: a report from population-based cancer registries. Cancer Epidemiol 36, e271-278, doi:10.1016/j.canep.2012.04.007 (2012).

41 Sighoko, D. et al. Population-based breast (female) and cervix cancer rates in the Gambia: evidence of ethnicity-related variations. Int J Cancer 127, 2248-2256, doi:10.1002/ijc.25244 (2010).

42 Sighoko, D. et al. Breast cancer in pre-menopausal women in West Africa: analysis of temporal trends and evaluation of risk factors associated with reproductive life. Breast 22, 828-835, doi:10.1016/j.breast.2013.02.011 (2013).

43 Aziz, N., Cherwinski, H. & McMahon, M. Complementation of defective colony-stimulating factor 1 receptor signaling and mitogenesis by Raf and v-Src. Mol Cell Biol 19, 1101-1115 (1999).

44 Soule, H. D. et al. Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. Cancer Res 50, 6075-6086 (1990).

45 Bowman, T., Garcia, R., Turkson, J. & Jove, R. STATs in oncogenesis. Oncogene 19, 2474-2488, doi:10.1038/sj.onc.1203527 (2000).

46 Frank, D. A. STAT3 as a central mediator of neoplastic cellular transformation. Cancer Letters 251, 199-210, doi:10.1016/j.canlet.2006.10.017.

47 Yu, H., Pardoll, D. & Jove, R. STATs in cancer inflammation and immunity: a leading role for STAT3. Nat Rev Cancer 9, 798-809, doi:10.1038/nrc2734 (2009).

48 Bromberg, J. F. et al. <em>Stat3</em> as an Oncogene. Cell 98, 295-303, doi:10.1016/S0092-8674(00)81959-5.

49 Dechow, T. N. et al. Requirement of matrix metalloproteinase-9 for the transformation of human mammary epithelial cells by Stat3-C. Proc Natl Acad Sci U S A 101, 10602-10607, doi:10.1073/pnas.0404100101 (2004).

50 Fleming, J. D. et al. STAT3 acts through pre-existing nucleosome-depleted regions bound by FOS during an epigenetic switch linking inflammation to cancer. Epigenetics & Chromatin 8, 7, doi:10.1186/1756-8935-8-7 (2015).

51 Vietri Rudan, M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep 10, 1297-1309, doi:10.1016/j.celrep.2015.02.004 (2015).

52 Grant, K. et al. Mechanisms of endothelin 1-stimulated proliferation in colorectal cancer cell lines. Br J Surg 94, 106-112, doi:10.1002/bjs.5536 (2007).

53 Zhang, W. M., Zhou, J. & Ye, Q. J. Endothelin-1 enhances proliferation of lung cancer cells by increasing intracellular free Ca2+. Life Sci 82, 764-771, doi:10.1016/j.lfs.2008.01.008 (2008).

54 Wulfing, P. et al. Endothelin-1-, endothelin-A-, and endothelin-B-receptor expression is correlated with vascular endothelial growth factor expression and angiogenesis in breast cancer. Clin Cancer Res 10, 2393-2400 (2004).

55 Rosano, L. et al. Beta-arrestin links endothelin A receptor to beta-catenin signaling to induce ovarian cancer cell invasion and metastasis. Proc Natl Acad Sci U S A 106, 2806-2811, doi:10.1073/pnas.0807158106 (2009).

56 Wilson, J. L., Burchell, J. & Grimshaw, M. J. Endothelins induce CCR7 expression by breast tumor cells via endothelin receptor A and hypoxia-inducible factor-1. Cancer Res 66, 11802-11807, doi:10.1158/0008-5472.CAN-06-1222 (2006).

57 Del Bufalo, D. et al. Endothelin-1 protects ovarian carcinoma cells against paclitaxel-induced apoptosis: requirement for Akt activation. Mol Pharmacol 61, 524-532 (2002).

58 Nelson, J. B., Udan, M. S., Guruli, G. & Pflug, B. R. Endothelin-1 inhibits apoptosis in prostate cancer. Neoplasia 7, 631-637 (2005).

59 Wulfing, P. et al. Expression of endothelin-1, endothelin-A, and endothelin-B receptor in human breast cancer and correlation with long-term follow-up. Clin Cancer Res 9, 4125-4131 (2003).

60 Jaffe, A. B. & Hall, A. Rho GTPases: biochemistry and biology. Annu Rev Cell Dev Biol 21, 247-269, doi:10.1146/annurev.cellbio.21.020604.150721 (2005).

61 Azzato, E. M. et al. A genome-wide association study of prognosis in breast cancer. Cancer Epidemiol Biomarkers Prev 19, 1140-1143, doi:10.1158/1055-9965.EPI-10-0085 (2010).

62 Wong, N. C. et al. Stability of gene expression and epigenetic profiles highlights the utility of patient-derived paediatric acute lymphoblastic leukaemia xenografts for

investigating molecular mechanisms of drug resistance. BMC Genomics 15, 416, doi:10.1186/1471-2164-15-416 (2014).

63 Luo, N. et al. ARHGAP10, downregulated in ovarian cancer, suppresses tumorigenicity of ovarian cancer cells. Cell Death Dis 7, e2157, doi:10.1038/cddis.2015.401 (2016).

64 Teng, J. P. et al. The roles of ARHGAP10 in the proliferation, migration and invasion of lung cancer cells. Oncol Lett 14, 4613-4618, doi:10.3892/ol.2017.6729 (2017).

3 Germline copy number variations are associated with breast cancer risk and prognosis[k]

3.1. Introduction

Breast Cancer is one of the commonly diagnosed cancers among women worldwide1, in Canada, breast cancer accounts for about 25% of all diagnosed cancers, and 15% of all cancer deaths2. Based on twin studies, estimated heritable genetic factors contribute to about 30% for breast cancer risk, the remaining risk being due to environmental and lifestyle factors3. Family based linkage and genome sequencing studies have identified high and moderate penetrant mutations in genes such as BRCA 1 or BRCA 24,5 PTEN6, PALB27, ATM8, TP539, and CHECK210 that contribute to the genetic risk of breast cancers. Subsequently, large scale population based Genome Wide Association Studies (GWAS) were successful in identifying several low penetrant common genetic variants (Single Nucleotide Polymorphisms, SNPs) associated with breast cancer risk. Among these, a limited number of GWAS SNPs (7 SNPs) showed effect sizes (odds ratio or ORs) between 1.25 – 1.5 and the remaining SNPs showed effect sizes <1.2511,12. SNP based GWAS served as a valuable tool in uncovering novel genes or loci associated with breast cancer aetiology. Low, moderate and high penetrant SNPs and mutations together explain up to 50% of the genetic risk associated with breast cancer11,12, and the remaining variants to explain the "missing heritability" are yet to be discovered. Copy

[k] A version of this chapter has been published. Kumaran et al., 2017. Scientific Reports. © 2017 Kumaran et al. This article published in Scientific Reports is licensed under a Creative Commons Attribution 4.0 International License. The author owns the copyright for the article. I would like to thank Dr. Sunita Ghosh and Dr. Preethi Krishnan for helpful suggestions. Ms. Jennifer Dufour is acknowledged for technical support.

Number Variations (CNVs) in the germline DNA are currently being investigated to explain missing heritable risk for breast cancer[13].

Germline CNVs are a class of structural variations and are defined as loss or gain of genomic DNA in size range of 50bp to 1Mb[14]. Germline CNVs are studied as genetic determinants for susceptibility for familial breast cancer[15-20] and also cancers of prostate[21-23], ovary[18,24-26], pancreas[27-29], colon, rectum[16,30-34], endometrium[35], lung[36-38] and melanoma[39,40].

The DNA sequence coverage for CNVs is ~10% of the genome. CNVs harbour coding regions and non-coding regulatory regions and may confer profound phenotypic effects relative to effects caused by SNPs[41-43]. CNVs have a multitude of effects based on their genomic location including gene dosage effects and *cis*-regulatory functions[23]. Since the distribution of CNVs across the genome is disproportionate with a higher proportion in non-coding than coding regions, their functional impact on phenotype is not clear. However, CNVs that overlap protein coding genes offer insights into disease phenotypes and associated biology[44]. Nearly 80% of cancer genes harbour CNVs[45] and support the above premise.

The majority of the CNVs that have been identified to-date for breast cancer are rare (frequency < 1%) and may potentially confer high penetrance (odds ratios >3.0) in familial breast cancer[18,20]. Associations of low penetrant common CNVs identified using GWAS have been shown in prostate[21,22] and pancreatic[29] cancers. CNV-GWAS has met with considerable success in several complex disease phenotypes[46] but is lagging in breast cancer with a limited number of studies adopting this approach. Long et al. in 2013 was the first to report a common CNV (deletion) in a coding gene using GWAS, wherein *APOBEC3* loci were shown to be associated with breast cancer risk in a Chinese population[47]. This deletion polymorphism was also validated in a Caucasian population[48]. These results support the goal of searching for common germline CNVs associated with sporadic breast cancer to address missing heritability in populations. This is in contrast to earlier claims that common CNVs were not associated with breast cancer[49].

Tumor based markers for prognosis are useful in guiding treatments but markers with higher specificity are needed to account for inter-individual variations in breast cancer prognosis. DNA level aberrations (CNVs) from tumor (somatic) genomes were shown to be prognostic. However, such studies do not distinguish origins from germline CNVs or de novo copy number aberrations in somatic cells due to genomic instability. Current emphasis is to assess the role of germline copy number variations for their prognostic value. SNPs showing association with breast cancer susceptibility were not prognostic[50,51]. Because independent SNP based GWAS for prognosis in breast cancer were not informative[2,50-53], I focused on identifying germline CNVs associated with breast cancer susceptibility and prognosis.

Since germline structural variations and their coverage on the genome is higher than SNPs, I reasoned that CNVs are suitable candidates to explore for their associations with prognosis. Germline CNVs have been identified as prognostic markers for several cancer types including prostate cancer[54], ovarian cancer[25] and colorectal cancer[55]. Our group showed that germline Copy Neutral Loss of Heterozygosity (CN-LOH), a class of CNVs, are associated with recurrence free survival in breast cancer[56].

Our aim was to conduct GWAS to identify common germline CNVs associated with breast cancer risk and assess if subsets of the risk associated CNVs are also associated with prognosis. Earlier studies on CNV association in familial breast cancer were restricted to identifying disease risk variants but not prognosis[18-20]. Specifically, I conducted CNV-GWAS, firstly focusing on identifying common CNVs overlapping with protein coding genes for association with breast cancer risk, secondly investigating the prognostic significance of the risk associated CNVs and thirdly correlating breast cancer risk associated CNVs with breast tumor tissue specific gene expression. have identified several common CNVs associated with breast cancer and determined that subsets of these CNVs are associated with both disease risk and prognosis. These findings highlight the importance of pursuing common germline CNVs to address the knowledge gap in the literature.

3.2. Methods

I performed all the experiments and analysis, unless otherwise indicated in the text

3.2.1. Study ethics approval

The study was approved by the local Health Research Ethics Board of Alberta (HREBA) - Cancer Committee. Written informed consents were obtained from all study participants. All experiments performed using specimens from study samples were carried out under approved guidelines and regulation.

3.2.2. Study population

Women with confirmed diagnosis of invasive breast cancer (cases, n=422) were recruited from Alberta, Canada between 1987 to 2006[51,56], and were described earlier. Briefly, the cases were non-metastatic at the time of diagnosis. Median age at diagnosis was 52 years, and 90% of cases were diagnosed at age >40 years (late age at onset); these are referred to as sporadic cases. Germline DNA and the clinical pathological information was accessed from the provincial tumor bank, the Alberta Cancer Research Biobank (formerly Canadian Breast Cancer Foundation (CBCF) Tumor Bank), located at the Cross-Cancer Institute, Edmonton, Alberta, Canada (http://www.acrb.ca/about-us/). At the time of study completion, the median follow-up time was 8.96 years and the number of events of breast cancer recurrence and death were n=171 and n=150, respectively. The controls (n=348) were healthy women (median age 50 years) with no personal or family history of cancer at the time of recruitment. The controls were accessed from a prospective cohort study called the Tomorrow Project ((http://in4tomorrow.ca) from Alberta, Canada. Comprehensive information about study participants (cases and controls) and methods to extract germline DNA from buffy coats are described elsewhere[56,57].

3.2.3. Genotyping and quality control

DNA extracted from buffy coat samples were genotyped using Affymetrix Genome-Wide Human SNP 6.0 array following manufacture's protocol[56]. Affymetrix SNP 6 array has independent probes for SNPs (~ 906,600 probes) and CNVs (~ 946,000 probes). Genotyping quality control was assessed using Birdseed V2 algorithm in Affymetrix genotyping console. Sample Contrast Quality Control (CQC) ≥1.7 indicates acceptable genotyping quality. All our study samples had a CQC value more than 2.

3.2.4. Population stratification

Principle Component Analysis (PCA) using EIGENSTRAT algorithm implemented in Golden Helix SNP and Variation suite v8.5.0 uses SNP genotypes generated on study samples (n=762) to infer the population stratification. Genotype data from 270 HapMap samples were used as a reference to infer the genetic ancestry of the study samples, and these were described previously[56,58]. After removing the outlier samples, I had 366 cases and 320 controls classified as European ancestry, and these were used for copy number analysis.

I also carried out Identity by Descent (IBD) analysis based on SNP probes using Golden Helix SNP and Variation suite v8.5.0. These analyses did not reveal any cryptic relatedness in samples with pair-wise correlation cut off < 0.25.

3.2.5. Copy number detection and gene annotation

Study design is described in Figure 3.1. Copy number analysis was performed using Partek® Genomics Suite™ 6.6 (PGS). Affymetrix array generated CEL files were used as input files for the program. GC wave correction was applied using default functions. I created a reference baseline (all sample normalization) using all the study samples to assign a diploid status and to infer the relative copy number estimates in individual cases and controls. A genomic segmentation algorithm implemented in the software was used to call the genomic segments with the following default criteria: genomic markers >10;

P-value threshold = 0.001; Signal/Noise (S/N) ratio = 0.3. The copy number status was assigned for each inferred segment relative to the normalized intensity (*i.e.*, 1.7-2.3 was considered as diploid); intensity values of >2.3 and <1.7 were called copy gains and losses, respectively. The CNVs were annotated using RefSeq genes using human genome build Hg19 (GRCh 37). The CNVs occurring at a frequency of >10% (termed common CNVs) of the study samples and mapping (or overlapping) to the protein coding gene regions were considered for downstream analysis. I excluded the regions that mapped to small and long non-coding RNA genes and pseudogenes. Multiple CNVs with contiguous genomic break points and similar copy status in a genomic region were merged into a single Copy Number Variation Region (CNVR).

3.2.6. Mapping to publicly available CNV databases

The identified CNVs were mapped to the Database for Genomic Variants[59] (DGV, to ascertain CNVs calls). The structural variant data currently available through 1000 Genomes Project phase 3 has information about 60,000 structural variations captured at the population level. The project utilized low coverage whole genome sequencing and exome sequencing and microarray technologies. These germline datasets were utilized to compare the break points estimated for CNVs in our study and for potential overlap with coding genes[60].

3.2.7. Statistical Analysis

(i) Power calculations:

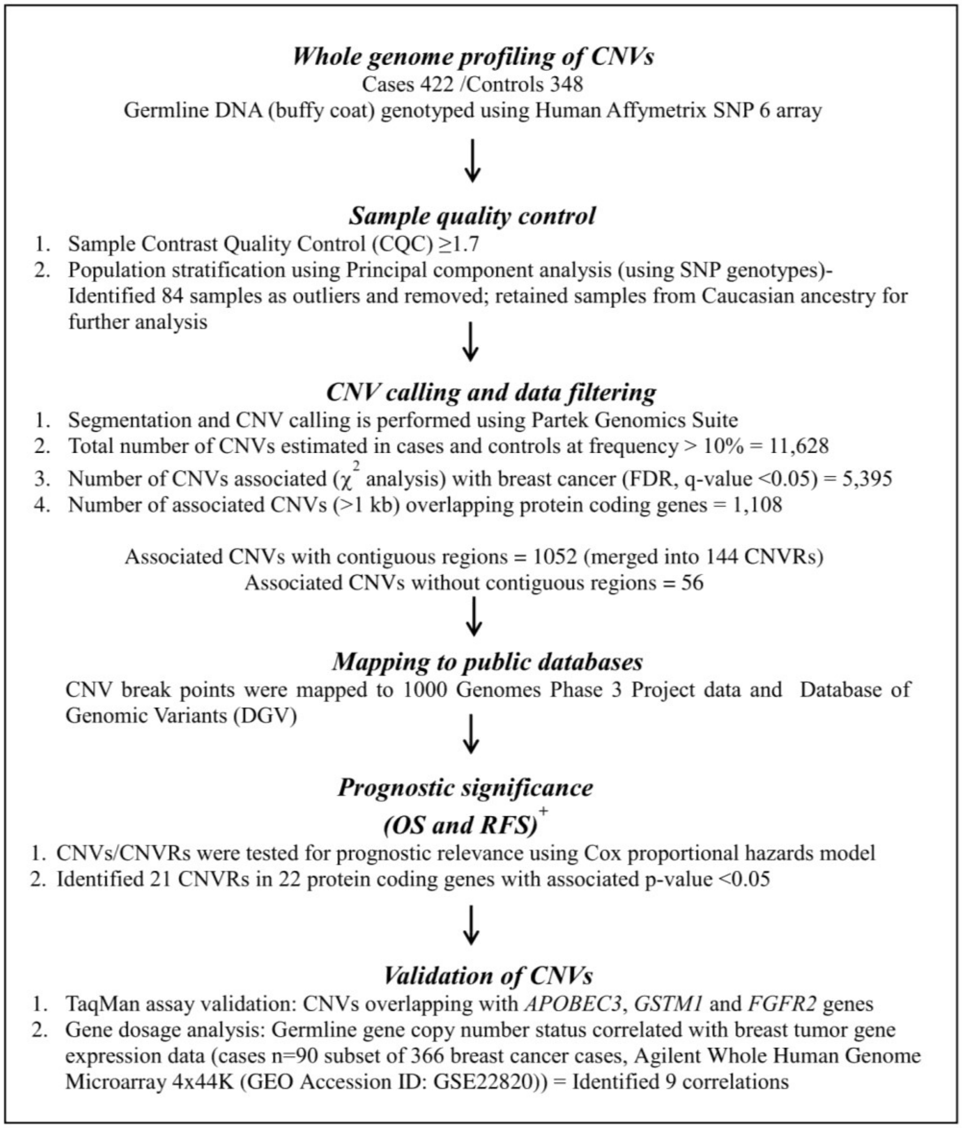
Power to detect CNVs associated with breast cancer susceptibility was calculated with "gap" package[61,62] using R program[63] I estimate that the study design and the sample size used will confer 94% power to detect associations for breast cancer risk. The following assumptions were made to compute power with a sample size of n=770: an additive model for genetic inheritance, the lifetime risk for breast cancer is 11% (1 in 9 among Caucasians) and at a genotype relative risk of 2 and a risk allele frequency of 10%.

(ii) Association analysis:

The association frequencies of the CNVs (diploid, gain and loss) between sample categories (cases, controls) were compared using chi-square (2X3) test implemented in Partek® Genomics Suite™ 6.6. A multiple hypothesis testing was accounted for using a false discovery rate method (reported as q-value). CNVs were considered significant if q-value were < 0.05.

(iii) Survival analysis and Cox-proportional hazards model:

CNVRs significantly associated with breast cancer risk by chi-square test were assessed for their prognostic significance of overall survival (OS) and recurrence free survival (RFS) using Cox-proportional hazards model, estimating Hazards Ratios (HRs) by the copy number status (diploid vs. gain/loss). Differences in survival probabilities among cases by the copy status (diploid vs gain/loss) were described using Kaplan-Meier survival curves. Survival analysis and Cox proportional hazards model were performed using “KMsurv” and “survival”[64,65] packages, respectively, implemented in R[63]. Since only breast cancer associated CNVs with overlap to coding genes (n=200 CNVs/CNVRs) and corrected for false discovery (q-value <0.05) were considered for Cox analysis, I did not apply additional multiple hypothesis corrections.



Whole genome profiling of CNVs

Cases 422 /Controls 348

Germline DNA (buffy coat) genotyped using Human Affymetrix SNP 6 array

Sample quality control

1. Sample Contrast Quality Control (CQC) ≥1.7
2. Population stratification using Principal component analysis (using SNP genotypes)- Identified 84 samples as outliers and removed; retained samples from Caucasian ancestry for further analysis

CNV calling and data filtering

1. Segmentation and CNV calling is performed using Partek Genomics Suite
2. Total number of CNVs estimated in cases and controls at frequency > 10% = 11,628
3. Number of CNVs associated (χ^2 analysis) with breast cancer (FDR, q-value <0.05) = 5,395
4. Number of associated CNVs (>1 kb) overlapping protein coding genes = 1,108

Associated CNVs with contiguous regions = 1052 (merged into 144 CNVRs)

Associated CNVs without contiguous regions = 56

Mapping to public databases

CNV break points were mapped to 1000 Genomes Phase 3 Project data and Database of Genomic Variants (DGV)

Prognostic significance

(OS and RFS)[+]

1. CNVs/CNVRs were tested for prognostic relevance using Cox proportional hazards model
2. Identified 21 CNVRs in 22 protein coding genes with associated p-value <0.05

Validation of CNVs

1. TaqMan assay validation: CNVs overlapping with *APOBEC3*, *GSTM1* and *FGFR2* genes
2. Gene dosage analysis: Germline gene copy number status correlated with breast tumor gene expression data (cases n=90 subset of 366 breast cancer cases, Agilent Whole Human Genome Microarray 4x44K (GEO Accession ID: GSE22820)) = Identified 9 correlations

Figure 3.1 Study Design

3.2.8. TaqMan copy number assays for validation of CNVs

CNVs were validated using TaqMan copy number assays from Applied Biosystems. Copy caller software supplied from Applied Biosystems was used for the data analysis. Representative CNVs were selected from three genes. I used predesigned assays for APOBEC3B (Hs04504055_cn), GSTM1 (Hs00273142_cn) and a custom assay for FGFR2 gene (assay location, chr10:123346308). Selection of genes for validation was based on the frequency of CNVs in our study cohort, availability of DNA in the corresponding samples with the inferred copy status for each sample from the copy number analysis. APOBE3B[47] and GSTM1 loci[66] were previously characterized to show copy number deletions. I used RNAase P as an internal control and followed the manufacturer-supplied protocols. I used two genomic DNA specimens from the Coriell DNA panel as positive controls. NA18635, which is of Chinese ancestry and diploid for all three genes tested, was used for data normalization. NA05299 belongs to European ancestry and has deletion in FGFR2 region.

3.2.9. Gene expression (mRNA) analysis in breast tumor tissues

mRNA dataset (Gene expression dataset) generated on breast tumor samples using Agilent Whole Human Genome Microarray 4x44K (GEO Accession ID: GSE22820) was available in-house with patient clinical characteristics (n=90). The 90 breast cancer cases were a subset of 366 (PCA stratified) cases with copy number profiles. Raw intensity files were quantile normalized, and log2 transformed using Partek Genomics Suite v6.6. The linear correlation was estimated between the germline copy number status and gene expression using PGS algorithms. In the correlation analysis, I considered only those gene expression probes whose location is within the breakpoints of the CNVs interrogated.

The objectives were to characterize the gene dosage effects and the relative expression of CNV-genes in breast tissues: (i) The dosage sensitive genes were determined by Pearson's correlation analysis (using PGS) between copy number and gene expression, and correlation value r>0.20. For the significantly correlated CNVs, dot plots of breast tumor gene expression versus germline copy number status were plotted. (ii) The prognostic significance of the genes overlapping in the germline CNV-genes from RFS and OS were also examined for breast tumor tissue specific gene expression. Fifteen of the 16 genes overlapping in the CNVR associated with OS were expressed. For ten genes in CNVR associated with RFS, eight genes were expressed in the mRNA dataset. Considering these genes as continuous variables, Univariate Cox proportional hazards regression was performed using SPSS v21.

3.3. Results

3.3.1. CNV-GWAS: Identification of breast cancer associated CNVs in coding regions

I identified 11628 CNVs in autosomes in an analysis that was restricted to common variants at frequency >10% in the study samples (see Figure 3.1 for study design). CNV frequencies compared between cases and controls (2x3 chi-square test) resulted in identification of 5395 CNVs which were statistically significantly associated with breast cancer at q-values <0.05. I only considered CNVs with size more than 1 kb for further analysis to increase confidence in CNV segments estimated by the algorithm. Although I identified CNVs in both protein coding and non-coding genes, those overlapping protein-coding genes have higher potential to contribute to phenotypic variation[44] and therefore focused on identification of CNVs overlapping with protein coding genes. CNVs were annotated for protein coding genes using RefSeq (GRCh37/ Human genome, Hg19 build) gene annotations. Of the 5395 CNVs that were significantly associated ($q<0.05$) with breast cancer, 1108 CNVs were mapped to 258 protein coding genes. I merged multiple contiguous CNVs from the set of 1108 into a single Copy Number Variable Region (CNVR) and interrogation of the overlapping genes for association with breast cancer yielded 200 altogether (144 CNVRs and 56 CNVs). The size ranges of the CNVRs and CNVs were 1.1 – 237 kb and 1.1 – 9Mb, respectively. The list of all associated CNVs/CNVRs is given (provide as electronic **Supplementary dataset 1** https://doi.org/10.1038/s41598-017-14799-7) and the list of the top CNVRs/CNVs (with q-values $<10^{-5}$) is given in Table 3.1.

(i) Mapping of CNVs to publicly available structural variation databases

Different genomic segmentation algorithms have their strengths and limitations58; the CNV break points called by different algorithms may or may not overlap and some algorithms tend to overcall CNVs58. Therefore, it was important to ascertain that the called CNVs were reliable by independent methods, and CNVs were mapped to the DGV and 1000 Genomes Project phase 3 data to assess concordances for the CNVs identified in this study. Ninety percent of CNVs associated with breast cancer mapped to the DGV, and while this is a common approach, this database has limitations. DGV curation is ongoing; its datasets are generated on diverse microarray platforms and by diverse CNV calling algorithms58. I therefore, considered a second method using higher resolution structural variation data available in the public domain from the 1000 Genomes Project (Phase 3). I mapped 76% of the 200 CNVRs/CNVs to the 1000 Genomes Project data and most of these (94%) also had hits in DGV, giving confidence in the CNV calling methods utilized in this study.

Table 3.1 Top associated germ line CNV/CNVR signature associated with breast cancer risk

CNV region	Cytoband	Size (bp)	Total CNV /CNVR Frequency in cohort	Average Frequency of CNV cases (gain/loss)	Average Frequency of CNV Controls (gain/loss)	q-value	Overlapping gene	Mapping
Chr5-69784291-70254895	5q13.2	470605	44	31 (13/18)	59 (3/56)	$1.46x10^{-21}$	SMN2, ERF1A, GUSBP9, SERF1B, SMN1, SMA5, GUSBP3	1000g, DGV
Chr5-70254905-70328368	5q13.2	73469	31	26 (11/15)	37 (7/30)	$3x10^{-02}$ to $1.76x10^{-13}$	NAIP	1000g.DGV
Chr21-40184963-40190820	21q22.2	2792	15	7 (3/4)	24 (0/24)	$1.58x10^{-10}$ to 4.3 $x10^{-12}$	ETS2	-
Chr9-40784158-40800446	9p13.1	60428	19	12 (5/7)	28 (3/25)	$1.09x10^{-11}$ to 5.23^{-12}	ZNF658	DGV
Chr8-7827144-7831849	8p23.1	4707	24	15 (7/8)	33 (4/29)	$1.02x10^{-09}$ to $1.65x10^{-09}$	FAM66E, USP17L8	DGV
Chr9-67899911-68067313	9q13	167404	18	8 (2/6)	29 (4/25)	$1.86x10^{-08}$ to $1.52x10^{-09}$	ANKRD20A1, ANKRD20A3	DGV
Chr1-248683401-248687808	1q44	4409	29	23 (8/15)	35 (1/34)	$2.38x10^{-08}$ to $6.47x10^{-09}$	OR2G6	DGV
Chr11-55418110-55421252	11q11	3143	85	94 (49/45)	76 (32/44)	$1.21x10^{-08}$	OR4S2	1000g, DGV
Chr8-93005629-93015066	8q21.3	9444	11	5 (2/3)	18 (0/18)	$7.69x10^{-08}$ to $5.94x10^{-09}$	RUNX1T1	-
Chr6-34516636-34517772	6p21.31	1143	11	17 (13/4)	6 (0/6)	$1.34x10^{-07}$ to $1.02x10^{-08}$	SPDEF	DGV
Chr11-55403771-55407672	11q11	3902	85	93 (49/44)	77 (33/44)	$4.18x10^{-08}$	OR4P4	1000g, DGV
Chr1-149548719-149563724	1q21.2	15005	30	26 (10/16)	35 (2/33)	$6.61x10^{-08}$	PPIAL4A, PPIAL4C	1000g, DGV
Chr10-123346484-123348045	10q26.13	1569	11	7 (3/4)	15 (0/15)	$6.04x10^{-07}$ to $1.05x10^{-07}$	FGFR2	-
Chr16-10788745-10790882	16p13.13	2137	10	7 (4/3)	14 (0/14)	$4.24x10^{-07}$	TEKT5	1000g, DGV

CNV region	Cytoband	Size (bp)	Total CNV /CNVR Frequency in cohort	Average Frequency of CNV cases (gain/loss)	Average Frequency of CNV Controls (gain/loss)	q-value	Overlapping gene	Mapping
Chr1-356492-380356	1p36.33	23865	21	16 (8/8)	28 (4/24)	5.62x10^{-07}	OR4F16, OR4F29, OR4F3	1000g, DGV
Chr9-67789400-67808579	9q13	19180	19	10 (2/8)	28 (3/25)	7.98x10^{-07}	FAM27B	1000g, DGV
Chr4-144288613-144293270	4q31.21	4667	18	11 (5/6)	26 (2/24)	1.5x10^{-05} to 2.4x10^{-11}	GAB1	DGV
Chr4-69505724-69536970	4q13.2	31250	32	29 (12/17)	35 (5/30)	1.29x10^{-03} to 1.10x10^{-06}	UGT2B15	1000g, DGV
Chr11-55430518-55436423	11q11	5907	81	87 (46/41)	73 (30/43)	1.68x10^{-05} to 2.79x10^{-08}	OR4C6	DGV
Chr9-67753281-67808579	9q13	55300	19	11 (2/9)	28 (3/25)	1.46x10^{-06} to 7.87x10^{-07}	FAM27E3,	1000g, DGV
Chr13-67509369-67513167	13q21.32	3811	11	7 (3/4)	14 (1/14)	1.24x10^{-03} to 2.07x10^{-06}	PCDH9	DGV
Chr7-75044860-75062133	7q11.23	17277	12	7 (3/4)	17 (0/17)	2.09x10^{-06} to 1.76x10^{-07}	NSUN5P1, POM121C	DGV
Chr17-20346165-20366887	17p11.2	20725	11	7 (3/4)	15 (0/15)	2.08x10^{-06} to 6.78x10^{-07}	LGALS9B	1000g, DGV
Chr4-55106768-55120708	4q12	13940	17	15 (6/9)	19 (0/19)	5.21x10^{-03} to 6.14x10^{-08}	PDGFRA	-
Chr13-48968806-48977635	13q14.2	8835	11	7 (3/4)	17 (0/17)	1.53x10^{-06} to 6.19x10^{-07}	RB1	1000g
Chr3-127422064-127423993	3q21.3	1931	10	6 (2/4)	15 (0/15)	6.29x10^{-06} to 4.01x10^{-06}	MGLL	1000g, DGV
Chr5-180425664-180437832	5q35.3	12170	19	19 (9/10)	18 (1/17)	4.71x10^{-05} to 2.62x10^{-05}	BTNL3	1000g, DGV
Chr1-152572873-152574332	1q21.3	2728	75	83 (40/43)	67 (24/43)	4.71x10^{-05} to 2.64x10^{-05}	LCE3C	1000g, DGV
Chr22-39363651-39371629	22q13.1	1119	19	21 (3/18)	17 (3/14)	3.65x10^{-02} to 2.73x10^{-02}	APOBEC3A_B	1000g, DGV

3.3.2. CNVRs associated with breast cancer prognosis

Since SNPs associated with breast cancer risk are poor prognosticators[52] , I investigated if the CNVs associated with breast cancer risk would have prognostic significance. I tested the 200 CNVRs/CNVs that showed association with breast cancer risk for prognostic significance using the Cox proportional hazards model. I compared the hazard function among the cases with diploid gene copy versus copy gain or loss. The identified prognostic CNVRs for Overall Survival (OS) and Recurrence Free Survival (RFS) are summarized in Tables 3.2 and 3.3. I identified 21 CNVRs overlapping 22 genes that showed associations with both breast cancer risk and prognosis.

Table 3.2 CNVRs associated with breast cancer risk and OS

CNVR region	**Gene name**	**CNVR Size (kb)**	**Copy number status**	**P-value**	**Hazards Ratio [95% CI]**
chr19:36846012-36847567*	*ZFP14*	1.55	gain	4.78×10^{-3}	2.38 [1.3-4.36]
chr1:65393459-65410228*	*JAK1*	16.77	gain	$1.07\ x10^{-2}$	3.24 [1.31-8.01]
chr1:110225034-110226615	*GSTM2*	1.58	gain	$1.30\ x10^{-2}$	1.81 [1.13-2.89]
chr17:80646036-80647251	*RAB40B*	1.21	gain	$1.60 x10^{-2}$	2.57 [1.19-5.52]
chr6:32487136-32497161	*HLA-DRB5, HLA-DRB6*	10.02	gain	$2.25x10^{-2}$	0.59 [0.38-0.93]
chr8:72213838-72215337	*EYA1*	1.49	gain	$3.09x10^{-2}$	1.59 [1.04-2.43]
chr6:161032642-161068568*	*LPA*	35.92	gain	$3.13x10^{-2}$	0.37 [0.15-0.91]
chr3:50951343-50960775	*DOCK3*	9.43	gain	$3.18x10^{-2}$	2.20 [1.07-4.52]
chr12:99796328-99797863	*ANKS1B*	1.53	gain	$3.35x10^{-2}$	1.94 [1.05-3.57]
chr12:2254285-2256046	*CACNA1C*	1.76	gain	$3.49x10^{-2}$	0.48 [0.24-0.95]
chr4:55111660-55120708*	*PDGFRA*	9.05	loss	$6.58x10^{-3}$	0.35 [0.16-0.74]

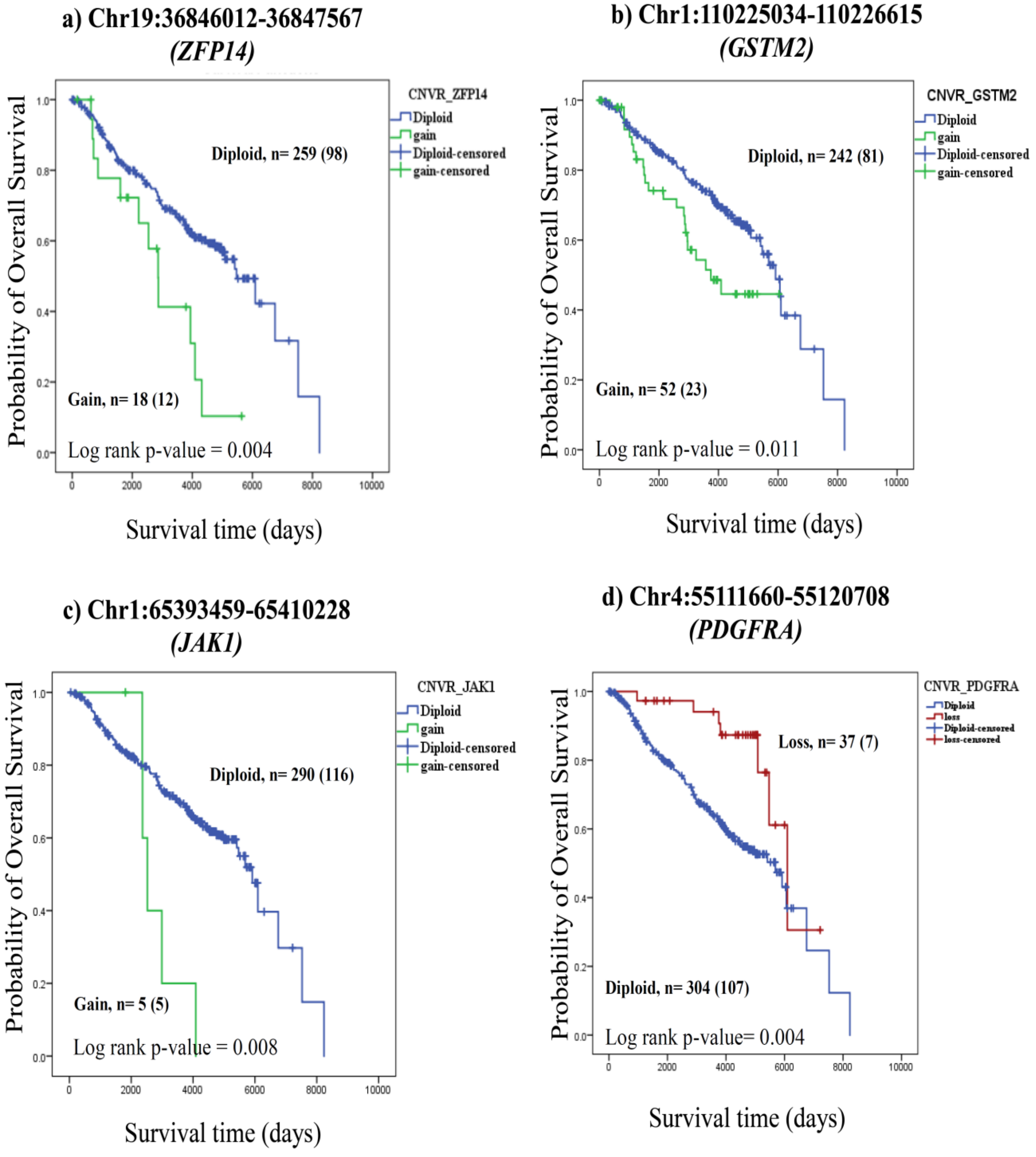
chr16:515664-536683	*RAB11FIP3*	21.02	loss	1.66x10^{-2}	0.43 [0.22-0.86]
chr21:11053457-11069332	*BAGE*	15.87	loss	2.01x10^{-2}	0.40 [0.19-0.87]
chr8:14284477-14288732	*SGCZ*	4.25	loss	2.41x10^{-2}	0.27 [0.08-0.84]
chr7:75044860-75054268	*POM121c*	9.41	loss	4.77x10^{-2}	0.20 [0.06-0.98]

This table describes the list of CNVRs associated with both risk and overall survival identified using Cox proportional hazard model. Only the associated copy number status (either loss or gain) compared with diploid is indicated in the table. The CNVR region marked with “*” indicate common CNVRs between OS and RFS. Abbreviation: CI – Confidence Interval.

(i) Germline CNVRs and OS in Breast cancer

I identified 15 CNVRs (with 16 overlapping genes) associated with breast cancer risk and OS (Table 3.2). Among these, 11 CNVRs overlapped with 12 genes (*GSTM2, RAB40B, HLA_DRB5, HLA_DRB6, EYA1, DOCK3, ANKS1B, CACNA1C, RAB11FIP3, BAGE, SGCZ, POM121c*) and were specifically associated with breast cancer risk and OS. The remaining four CNVRs overlapped with genes *ZFP14, JAK1, LPA, PDGFRA* and were also associated with RFS in breast cancer. The P-values for the identified 15 CNVRs were in the range of 4.77x10^{-2} to 4.78x10-^{3}. Both gains and losses contributed to prognostic significance. Copy gains showed both risk elevating and protective effects whereas copy losses showed only protective effects. The Kaplan-Meier (KM) survival plot for the top associated CNVR with OS is shown in Figure 3.2. Copy number gains in the genes *ZFP14, GSTM2* and *JAK1* were shown to be associated with poor OS in the univariate Cox analysis. P-values and HRs estimated for these genes were as follows: *ZFP14* (P-value =4.78x10^{-3} and HR 2.38), *GSTM2* (P=1.30x10^{-2} and HR 1.81) and *JAK1*

(P-value =1.07x10^{-2} and HR 3.24). KM plots describing the survival differences and estimated log rank p-values are shown in Figure 3.2 (a-c). The estimated survival differences (log rank p-values) for cases with copy gains compared to cases with diploid copies of the genes *ZFP14, GSTM2*, and *JAK1* were 0.004, 0.11 and 0.008 respectively. Copy number loss of *PDGFRA* was associated with OS (P-value 6.58x10^{-3} and HR 0.35) and cases with copy loss had better survival outcomes compared with cases with diploid copies, the log rank p-value estimated for the difference in survival value was 4x10^{-3}.



a) Chr19:36846012-36847567 *(ZFP14)*

b) Chr1:110225034-110226615 *(GSTM2)*

c) Chr1:65393459-65410228 *(JAK1)*

d) Chr4:55111660-55120708 *(PDGFRA)*

Figure 3.2 KM plots for CNVRs associated with overall Survival

KM plots were constructed based on the copy number status of each gene to determine the difference in overall survival (OS) between cases with genes harbouring copy number variation (gain/loss) versus diploid status. Blue indicates Diploid copy number; Green indicates Copy number gain; Red indicates Copy number loss. “ + ” indicates the censored events. The number of cases, n, in the analysis is indicated and the number of events in the study for each survival curve is indicated in parenthesis. Log rank p-value for significance between the curves is indicated at the bottom of each panel within the figure.

(ii) Germline CNVRs and RFS in Breast cancer

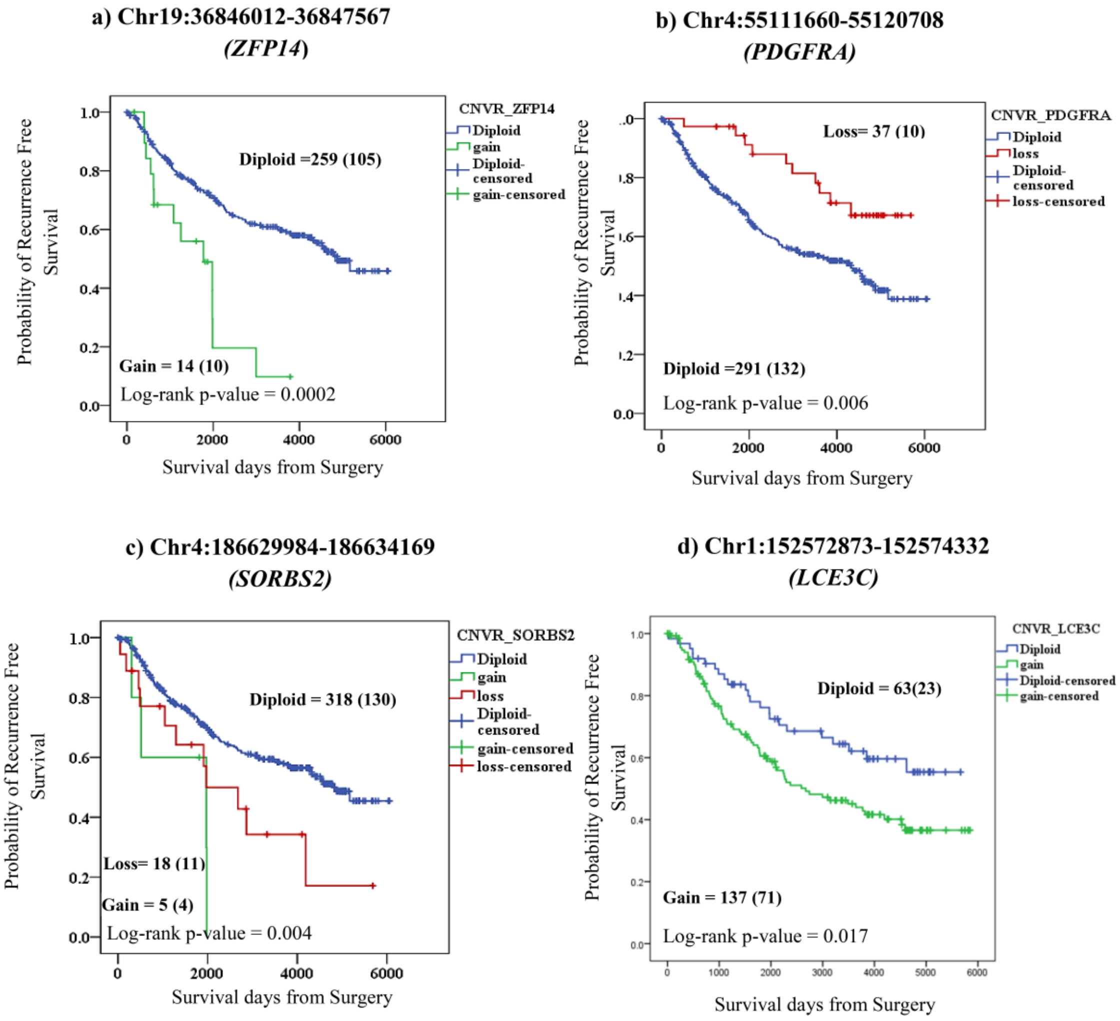
I identified a total of ten CNVRs associated with breast cancer risk and RFS (Table 3.3). Among the ten CNVRs, six CNVRs overlapped with the genes (*SORBS2, LCE3C, MLIP, OR2T11, MUC20, LGALS*) that were specifically associated with RFS; and four CNVRs (*ZFP14, JAK1, LPA, PDGFRA)* were also associated with OS. The associated CNVRs had P-values in the range of 3.65x10^{-2} to 3.82x10^{-4}. Both copy gains and losses were associated with elevated risk or protective effects. The KM plots for the top associated CNVRs with RFS are illustrated in Figure 3.3. I observed that copy gains in *ZFP14* and *LEC3C* were associated with poor RFS with P-values 3.82x10^{-4} and 1.94x10^{-2} and HRs 2.89 and 1.75, respectively. The log rank p-value estimated from KM plots (Figure 3.3a, 3.3d) for the genes ZFP14 and *LEC3C* were 2.0x10^{-4} and 1.7x10^{-2}, respectively. In *PDGRA* gene copy loss associated with RFS and cases with copy loss had better survival outcomes compared with diploid copy status (RFS, P-value 7.92 x 10^{-3} and HR 0.42). The log rank p-value estimated was 6x10^{-3} based on KM plot (Figure 3.3b). A similar trend was observed for OS as well. Another interesting CNVR was in the SORBS2 gene in which both copy gain and loss were associated with poor RFS. For copy gain, the P-value was 1.35 x 10^{-2} and HR was 3.54; for copy loss, the P-value was 3.65 x 10^{-2}, and the HR was 1.93. The log rank p-value for the difference in the copy gain/loss versus diploid copy status was 4x10^{-3} (Figure 3.3c).

I observed that copy number deletion in *APOBEC3A_B* was not associated with either RFS and OS in breast cancer, which agrees with published findings[67].

Table 3.3 CNVRs associated with breast cancer risk and RFS

CNVR region	Gene name	CNVR Size (kb)	CNV type	Cox P-value	Hazards Ratio [95% CI]
chr19:36846012-36847567*	*ZFP14*	1.55	Gain	3.82x10^{-4}	2.89 [1.61-5.19]
chr4:186629984-186634169	*SORBS2*$^{+}$	4.18	Gain	1.35x10^{-2}	3.54 [1.3-9.64]
chr1:152572873-152574332	*LCE3C*	1.46	Gain	1.94x10^{-2}	1.75 [1.1-2.81]
chr1:248787969-248794876	*OR2T11*	6.91	Gain	2.64x10^{-2}	2.09 [1.09-4]
chr3:195456468-195461506	*MUC20*	5.04	Gain	3.46x10^{-2}	0.62 [0.39-0.97]
chr1:65393459-65410228*	*JAK1*	16.77	Gain	3.47x10^{-2}	2.6 [1.07-6.47]
chr6:161032642-161068568*	*LPA*	35.92	Gain	5.08x10^{-3}	0.31 [0.13-0.70]
chr17:20346165-20366887	*LGALS9B*	20.72	Gain	3.52x10^{-2}	2.27 [1.06-4.87]
chr4:55111660-55120708*	*PDGFRA*	9.05	Loss	7.92x10^{-3}	0.42 [0.22-0.8]
chr6:53931117-53933601	*MLIP*	2.48	Loss	2.53x10^{-2}	0.62 [0.4-0.94]
chr4:186629984-186634169	*SORBS2*$^{+}$	4.18	Loss	3.65x10^{-2}	1.93 [1.04-3.58]

This table represents the list of CNVRs associated with both risk and RFS identified using Cox proportional hazard model. Only the associated copy number status (either loss or gain) compared with diploid is indicated in the table. The CNVR region marked with "*" indicate common CNVRs between OS and RFS "+" Indicates that gene that has both gain and loss associated with recurrence free survival when compared to diploid. Abbreviation: CI – Confidence Interval.



a) Chr19:36846012-36847567 *(ZFP14)*

b) Chr4:55111660-55120708 *(PDGFRA)*

c) Chr4:186629984-186634169 *(SORBS2)*

d) Chr1:152572873-152574332 *(LCE3C)*

Figure 3.3 KM plots for CNVRs associated with RFS

KM plots were constructed based on the copy number status of each gene to determine the difference in recurrence free survival (RFS) between cases with genes harbouring copy number variation (gain/loss) versus diploid status. Blue indicates Diploid copy number; Green indicates Copy number gain; Red indicates Copy number loss. “ + ” indicates the censored events. Number of cases, n in the analysis is indicated and the number of events in the study for each survival curve is indicated in parenthesis. Log rank p-value for significance between the curves is indicated at the bottom of each panel within the figure.

3.3.3. C: Validation of associated CNVs

(i) Cross platform validation of CNVs using the TaqMan Assay

Breast cancer associated CNVs overlapping with the genes *APOBEC3B, GSTM1 and FGFR2* were validated using the TaqMan assay. For APOBEC3B, 13 samples were tested (Figure 3.4a): one sample (healthy control) had two copy deletions, ten samples had one copy deletion (4 healthy controls and 6 breast cancer cases) and two samples (breast cancer cases) had diploid copy numbers. For *GSTM1*, I identified 16 samples (7 controls, 9 cases) with two copy deletions and 11 samples (3 controls and 8 cases) with one copy deletion (Figure 3.4b). Both *APOBEC3* and *GSTM1* quantifications by the TaqMan assays showed excellent agreement with the predicted copy status from PGS (this study) and the 1000 genomes data.

CNVs identified in *FGFR2* predominantly showed copy deletions as inferred by PGS; the same CNVs, when mapped to the 1000 genomes data, showed diploid status. I tested 29 samples (19 controls and 10 cases) by the TaqMan assay to verify copy status; all samples showed diploid status. To ensure the quality of the assay design, I used the Coriell DNA sample (NA05299) that had one copy deletion in *FGFR2* as a positive control for *FGFR2* deletion thereby demonstrating that the technical aspects of the TaqMan assay did not contribute to disagreement in the copy deletions noted (data not shown). A targeted re-sequencing of this region is needed to confirm these findings.

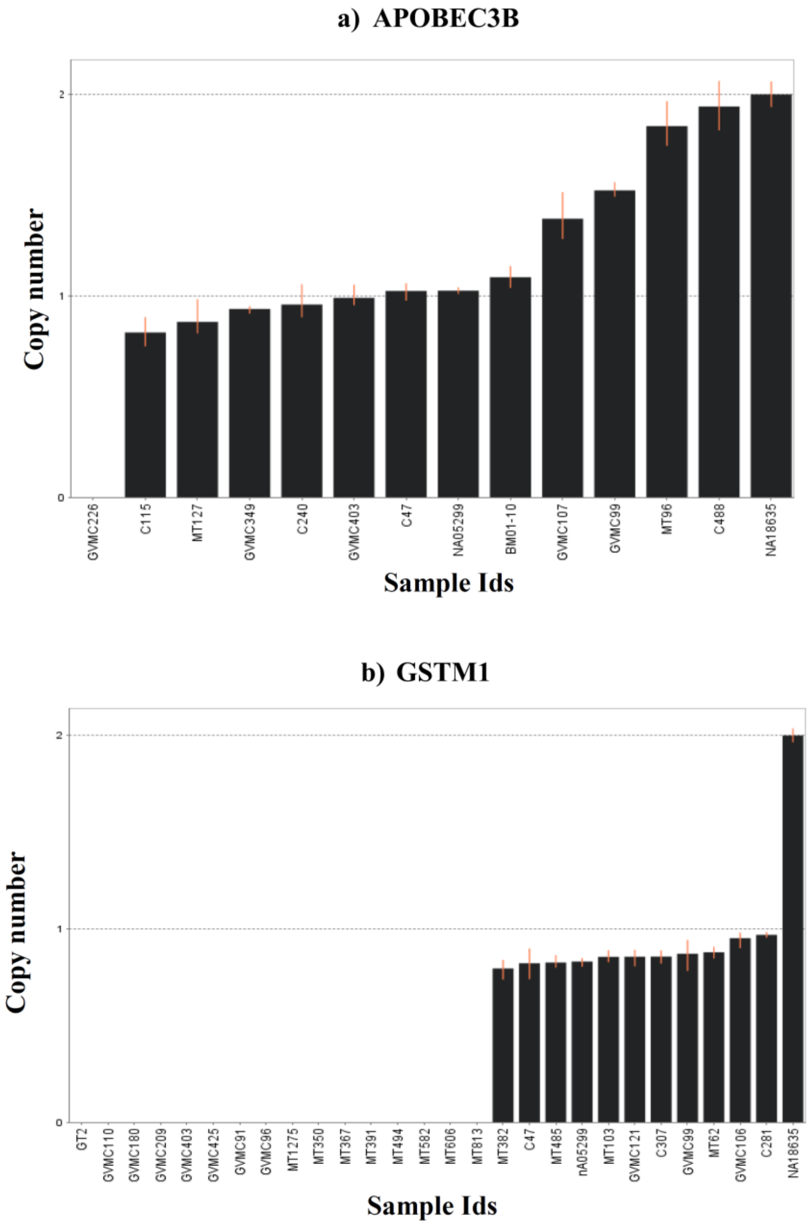


Figure 3.4 Copy number status estimated study samples using TaqMan Assay

Copy number status of genes *APOBEC3B* (**a**) and *GSTM1* (**b**) are represented for each sample. The Human *RNAase P* was used as internal normalization and the Coriell sample NA18635, which is diploid for both genes, were also used in copy number estimation.

(ii) Detailed characteristics of the validated CNVs: (a) *APOBEC3A_B* loci: A deletion of *APOBEC3A_B* was previously reported to be associated with breast cancer risk in Chinese[47], European[48] and Iranian[68] populations. In this study, I have also identified CNVs showing a deletion in the APOBEC3B gene and associated with breast cancer risk (Table 3.1). I validated the deletion in our cohort using the TaqMan assay as an independent genotyping platform. A single copy deletion of *APOBEC3A_B* was observed at frequencies of 14% among controls and 18% of cases (Caucasian ancestry), which is comparable with results of previous reports[48]. This is the second such study based on a Caucasian population to independently validate a common CNV and its association with breast cancer.

(b) *GSTM1*: Although the role of germline CNVs in the *GSTM* family of genes, which are involved in xenobiotic detoxification and drug metabolism pathways is well documented in other cancer types[69], their role in breast cancer is not clear. I identified CNVs (both gains and losses) in GSTM1 and *GSTM2* and their frequencies in the total cohort were 78% and 27% in the Caucasian population, respectively (provided as electronic Supplementary dataset 1 https://doi.org/10.1038/s41598-017-14799-7). The relative frequencies of deletions in GSTM1 (Cases, 40%; Controls, 31%) and *GSTM2* (Cases, 15%; Controls, 8%). CNVs were higher among the cases compared to the controls. The CNVs identified in *GSTM* loci were also observed in 1000 Genomes Project data as a copy variable region.

(c) Correlation of germline CNV copy status of protein coding genes with gene expression in breast tumors: One of the mechanisms by which germline CNVs may bring about phenotypic effects is gene dosage, and in this context "functionality" refers to underlying gene expression changes in breast tumor tissues rather than specific changes in cellular morphology or proliferation rates. To identify gene dosage effects due to germline CNVs, I looked for correlations between gene expression profiles derived from breast tumor biopsy samples (n=90) and the germline CNV data available from the same cases. I expected only a subset of genes to be expressed in a tissue specific manner and our observations support this premise. The expression of nine genes correlated with corresponding germline CNVs with correlation coefficients in the range 0.2 to 0.39 (Appendix Table A.10). Seven of the nine genes also were statistically significant at $p<0.05$ and two showed trends of association ($p<0.1$). The association of gene expression as a function of the germline copy number status is illustrated in Figure 3.5. Mean expression levels among cases with copy number deletions were consistently less than among cases with diploid copy number or amplification. The correlated genes are well known to harbour germline copy number variations[70-72], and the association of CNVs in these genes with breast cancer risk and the altered expression of these genes in breast tumor tissues is noteworthy.

In addition to the linear correlation of gene expression with CNVs, I also tested if the genes overlapping in the prognostic CNVs (n=22) were also associated with RFS and OS. Eighteen of the 22 genes overlapping in the CNVRs also showed expression in breast tumor tissues. Of these, expression of five genes *(GSTM2, SGCZ, HLA_DRB5, ZFP14, LCE3C*) showed association with prognosis (Appendix Table A.11).

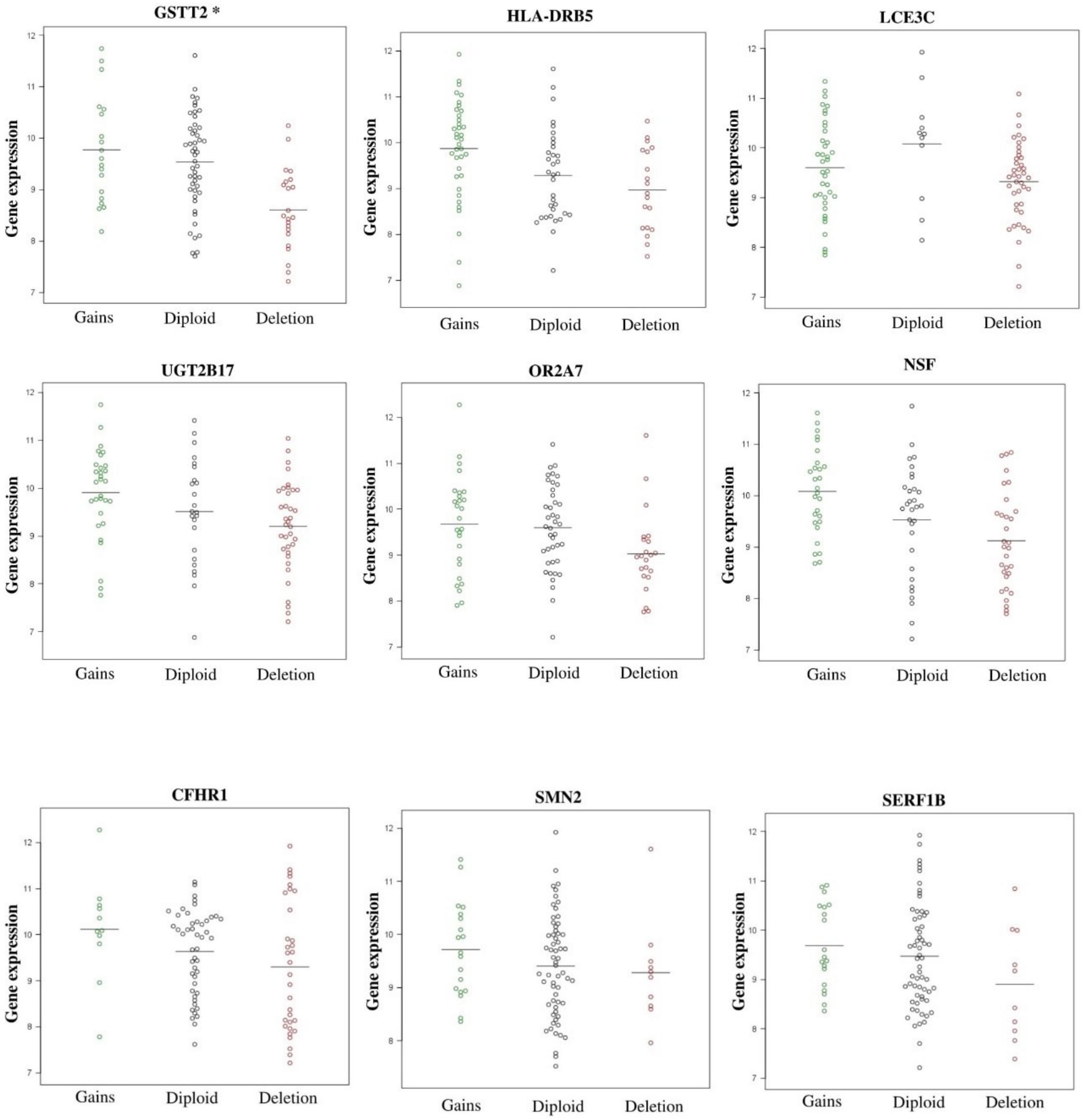


Figure 3.5 Association of germline copy number status and gene expression in breast tumor tissue

Germline copy number status of individual genes was plotted against gene expression in breast tumors from matched samples. The colours indicated in green, grey and red represent gain, diploid and deletion, respectively.

3.4. Discussion

In this study, I sought to identify germline CNVs that predispose to both breast cancer susceptibility and prognosis. Using 686 samples for copy number analysis, I identified 200 CNVs/CNVRs (frequencies >10%) that overlapped with protein coding genes at q-values <0.05. I compared the identified CNVs/CNVRs break points to the structural variation data available from the 1000 Genomes Project to ascertain CNV calls, an approach that was unique to our study. Another novel aspect was the assessment of prognostic relevance of breast cancer susceptibility CNVs. I demonstrated that some CNVs were only associated with disease risk whereas some were associated with both disease risk and prognosis. These findings are in contrast to SNP based association studies in which susceptibility SNPs from GWAS did not show prognostic relevance, with one exception, the SNP rs13281615[73] on chromosome 8q24.21 locus which myself and others showed as associated with both OS and RFS in breast cancer[51]. Further, independent SNP based GWAS was not successful in identifying variants associated with breast cancer prognosis[52]. CNVs cover 10% of the genome based on nucleotide coverage and our study rationale assumed that CNVs overlapping with coding genes (deletions or gains) influence phenotypes.

Of relevance was the replication in my study of the *APOBEC3A_B* gene deletion (Chr22-39363651-39364770), which was originally reported in Chinese populations as a breast cancer susceptibility CNV in sporadic cases[47]. Subsequently the same was replicated in European[48] and Iranian populations[68]. There were both gains and losses at this locus in this study; frequencies of gains were the same in both cases and controls (at 3%) whereas the above published studies reported only copy loss. The copy number deletion is the risk allele and the frequencies were 18% and 14%, respectively, in cases and controls (this study). These were in agreement with reported studies[74] in Caucasian populations (Table 3.1). *APOBEC3B* gene was not shown to be associated with prognosis (OS)[67], which I confirmed in this study.

I have identified a CNV (Chr1:110230244-110233070) showing association with breast cancer and harbouring the *GSTM1* gene. Earlier candidate gene studies identified SNPs in *GSTM1* to be associated with breast cancer risk[75]. I report a common CNV approximately 3kb in size in a locus encompassing *GSTM1* associated with breast cancer risk. The 1000 genome annotation indicates that a CNV in this genomic locus spans about 20kb in size and encompasses the entire gene. The CNV encompassing *GSTM1* showed both gains and losses at high frequencies in cases and controls (provided as electronic **Supplementary dataset 1** https://doi.org/10.1038/s41598-017-14799-7). The frequencies were approximately the same for gains in cases and controls (43% vs. 42%). However, deletion frequencies differed between cases and controls (40% vs. 31%), with cases showing higher frequencies. Although a germline CNV overlapping *GSTM1* was shown to be associated with prognosis in prostate and bladder cancers[69], this CNV was not associated with prognosis in this study. SNP based studies in the *GSTM1* gene associated with breast cancer risk but not with prognosis[76,77]. I validated both *APOBEC3 and GSTM1* CNV deletions using the TaqMan assays. Interestingly, the representative genes *(APOBEC3B and GSTM1)* validated by the TaqMan assays were also identified as copy variable genes by the 1000 genomes project.

The characteristics and putative biological roles for representative genes associated with breast cancer susceptibility and/or prognosis are summarized here:

(i) *PDGFRA*, Platelet-Derived Growth Factor Receptor Alpha is a tyrosine kinase receptor that is overexpressed in malignancies including the breast. I observed a CNV in *PDGFRA* is not only associated with BC risk and but a copy loss in this gene is conferring protective effect for RFS and OS. A higher frequency of copy gain was seen in cases (~6%) compared to 0% frequency among controls. However, frequency of deletion observed in controls was higher (19%) compared to cases (9%). Overexpression of *PDGFRA* is also known to play a role in tumorigenesis and its amplification or genetic alteration is believed to activate the *PDGFRA* mediated signaling pathway[78].

LPA (Lysophosphatidic acid), a lipid biomolecule that functions as a growth factor mediating cell proliferation, migration and progression, processes that are central to tumorigenesis[79,80]. Both CNV and gene expression profiles of LPA are associated with both susceptibility and prognosis. Copy number gain was associated with protective effect for OS and RFS.

A germline CNV in *ZFP14* (Zinc Finger protein) was associated with risk and prognosis in our analysis. CNV in *ZFP14* is associated with prostate cancer[23], in which a deletion is protective for prostate cancer risk. I observed copy gains among the cases and there were associated with poor prognosis. Somatic copy number aberration is also observed in *ZFP14* gene in breast tumors[81,82].

The CNV association studies in breast cancer reported thus far have focused on cases that are BRCA positive or with family history with or without BRCA mutations[18] and with limited sample sizes (n=30-60). These studies identified rare CNVs (frequency <1% in total cohort). Recently a CNV-GWAS study was conducted using cases with early onset of breast cancer (age <40 Years; 200 cases and 293 controls) and genotyping was performed using Illumina Human610-Quad BeadChip[15] and CNV calls were inferred based on SNP probe intensities. Our study utilized cases that were diagnosed with invasive breast cancer with late age at onset of the disease (>40 Years; 422 cases and 348 controls) and focused on common CNVs. I used Affymetrix SNP 6 arrays and CNV calls were based both on SNP and CNV probes. Because SNP density is lower in CNV dense regions, our study benefitted from using the Affymetrix arrays. Most existing studies on CNV associations with breast cancer have relied on SNP probes, and CNV calling algorithms are also diverse. Hence potential overlap of the genes identified in our study with those previously described are likely to be highly restrictive. Our use of both CNV and SNP probes to infer copy status may have contributed to higher numbers of CNVs associated with breast cancer. As with any GWAS study, Stage-1 study identifies several variants associated with the phenotype, and our data conforms with the GWAS literature. However, I addressed multiple hypothesis testing by implementing q-value (<0.05) thresholds. In addition, I also mapped the associated CNVs with breast cancer to 1000 Genomes Project database and confirmed that a majority of CNVs identified were indeed common CNVs. I have replicated CNVs (n=5) from the familial breast cancer study, including CNVs in genes *ANKS1B*[19] , *OR4C11, OR4P4, UGT2B17, OR4C6, OR4S2*[15]. Even though previous studies have ascribed these CNV overlapping genes to early onset of breast cancer, independent replication of these findings in late age at onset of breast cancer (this study) suggests that some CNVs may be common and emphasizes the more general role these genes play in the aetiology of breast cancer.

The breast cancer risk associated CNVs (Table 3.1) that mapped to 1000 genomes *(NME7, RB1, UGT2B15, BTNL3, RBL1, LGALS9B, MGLL, GSTM1, and PML*) were also captured in a recent breast tumor tissue (somatic) profiling study, confirming that the identified genes are primarily in copy number variable regions[82].

I tested the 200 CNVRs overlapping protein coding genes for their associations with breast cancer RFS and OS using the Cox proportional hazard model. The cases in our study have well annotated clinical data and long years of follow up, and compared the survival benefit of cases based on the germline copy number status (gain or loss) against diploid copy for a given CNVR. I identified CNVRs to be associated with RFS and/or OS among the cases. Genes within the four CNVRs (*i.e.*, *ZFP14, JAK1, LPA, PDGFRA*) were associated with both RFS and OS; these genes are also known to harbour somatic copy number aberrations in breast tumors[81-83].

It is critical to demonstrate the functionality of genes overlapping with CNVs. I therefore examined their dosage sensitivities and identified nine genes whose expression is breast tissue specific. The dot plots (Figure 3.5) clearly indicate the differences in expression levels between deletion versus diploid genes. The well-known germline CNV harbouring genes, *GSTT1, UGT2B17*, are involved in detoxification, steroid and drug metabolism pathways. and their dosage sensitivities are well studied[76,84,85]. These genes are also associated with breast cancer risk and demonstrating dosage sensitivity at the tissue level will contribute to an understanding of the mechanistic basis for disease aetiology. Even though GST family of genes showed associations at the CNV level, their correlation with gene expression was not significant due to the unequal distribution of samples across different copy number states and the limited sample size of 90. A larger sample size with gene expression and germline CNV profiles will allow us to detect correlations between CNVs and gene expression.

3.5. Conclusion

In this study I restricted the analysis to CNVs overlapping with protein coding regions, the preferred approach in most CNV based association studies reported in the literature[44,47]. Although intergenic CNVs in non-coding regions also merits attention, access to matched data sets (germline CNVs and gene expression data) is needed and these are to be addressed in future studies. Such data mining approaches have shown promising leads in disease settings other than breast cancer[86,87]. In this study, the identified CNVs associated with breast cancer phenotypes, vis-à-vis, heritable determinants for disease susceptibility and prognosis and predict that our results also apply to CNVs that harbour non-coding RNA genes.

3.6. Availability of data and material

All data generated or analysed during this study are included in the published article and its supplementary information files. The dataset is provided as electronic Supplementary dataset 1 https://doi.org/10.1038/s41598-017-14799-7.

3.7. References

1 Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136**, E359-386, doi:10.1002/ijc.29210 (2015).

2 Canadian Cancer Society. Breast Cancer Statistics. (2017).

3 Locatelli, I., Lichtenstein, P. & Yashin, A. I. The heritability of breast cancer: a Bayesian correlated frailty model applied to Swedish twins data. Twin Res 7, 182-191, doi:10.1375/136905204323016168 (2004).

4 Wooster, R. et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 378, 789-792, doi:10.1038/378789a0 (1995).

5 Miki, Y. et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266, 66-71 (1994).

6 Liaw, D. et al. Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nat Genet 16, 64-67, doi:10.1038/ng0597-64 (1997).

7 Rahman, N. et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet 39, 165-167, doi:http://www.nature.com/ng/journal/v39/n2/suppinfo/ng1959_S1.html (2007).

8 Renwick, A. et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet 38, 873-875, doi:http://www.nature.com/ng/journal/v38/n8/suppinfo/ng1837_S1.html (2006).

9 Malkin, D. et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 250, 1233 (1990).

10 Meijers-Heijboer, H. et al. Low-penetrance susceptibility to breast cancer due to CHEK2[ast]1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet 31, 55-59 (2002).

11 Fachal, L. & Dunning, A. M. From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. Curr Opin Genet Dev 30, 32-41, doi:10.1016/j.gde.2015.01.004 (2015).

12 Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics 45, 353-361, 361e351-352, doi:10.1038/ng.2563 [doi] (2013).

13 Kuiper, R. P., Ligtenberg, M. J., Hoogerbrugge, N. & Geurts van Kessel, A. Germline copy number variation and cancer risk. Curr Opin Genet Dev 20, 282-289, doi:10.1016/j.gde.2010.03.005 (2010).

14 Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. Nat Rev Genet 16, 172-183, doi:10.1038/nrg3871 (2015).

15 Walker, L. C. et al. Increased genomic burden of germline copy number variants is associated with early onset breast cancer: Australian breast cancer family registry. Breast Cancer Res 19, 30, doi:10.1186/s13058-017-0825-6 (2017).

16 Villacis, R. A. et al. ROBO1 deletion as a novel germline alteration in breast and colorectal cancer patients. Tumour Biol 37, 3145-3153, doi:10.1007/s13277-015-4145-0 (2016).

17 Masson, A. L. et al. Expanding the genetic basis of copy number variation in familial breast cancer. Hered Cancer Clin Pract 12, 15, doi:10.1186/1897-4287-12-15 (2014).

18 Kuusisto, K. M. et al. copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. PLoS ONE [Electronic Resource] 8, e71802, doi:http://dx.doi.org/10.1371/journal.pone.0071802 (2013).

19 Pylkas, K. et al. Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. PLoS Genet 8, e1002734, doi:10.1371/journal.pgen.1002734 (2012).

20 Krepischi, A. C. et al. Germline DNA copy number variation in familial and early-onset breast cancer. Breast Cancer Res 14, R24, doi:10.1186/bcr3109 (2012).

21 Laitinen, V. H. et al. Germline copy number variation analysis in Finnish families with hereditary prostate cancer. Prostate 76, 316-324, doi:10.1002/pros.23123 (2016).

22 Ledet, E. M. et al. Characterization of germline copy number variation in high-risk African American families with prostate cancer. Prostate 73, 614-623, doi:10.1002/pros.22602 (2013).

23 Demichelis, F. et al. Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc Natl Acad Sci U S A 109, 6686-6691, doi:10.1073/pnas.1117405109 (2012).

24 Pedersen, B. S., Konstantinopoulos, P. A., Spillman, M. A. & De, S. Copy neutral loss of heterozygosity is more frequent in older ovarian cancer patients. Genes Chromosomes Cancer 52, 794-801, doi:10.1002/gcc.22075 (2013).

25 Fridley, B. L. et al. Germline copy number variation and ovarian cancer survival. Front Genet 3, 142, doi:10.3389/fgene.2012.00142 (2012).

26 Yoshihara, K. et al. Germline copy number variations in BRCA1-associated ovarian cancer patients. Genes Chromosomes Cancer 50, 167-177, doi:10.1002/gcc.20841 (2011).

27 Fanale, D. et al. Germline copy number variation in the YTHDC2 gene: does it have a role in finding a novel potential molecular target involved in pancreatic adenocarcinoma susceptibility? Expert Opin Ther Targets 18, 841-850, doi:10.1517/14728222.2014.920324 (2014).

28 Fanale, D. et al. Analysis of germline gene copy number variants of patients with sporadic pancreatic adenocarcinoma reveals specific variations. Oncology 85, 306-311, doi:10.1159/000354737 (2013).

29 Al-Sukhni, W. et al. Identification of germline genomic copy number variation in familial pancreatic cancer. Hum Genet 131, 1481-1494, doi:10.1007/s00439-012-1183-1 (2012).

30 Brea-Fernandez, A. J. et al. Candidate predisposing germline copy number variants in early onset colorectal cancer patients. Clin Transl Oncol, doi:10.1007/s12094-016-1576-z (2016).

31 Weren, R. D. et al. Germline deletions in the tumour suppressor gene FOCAD are associated with polyposis and colorectal cancer development. J Pathol 236, 155-164, doi:10.1002/path.4520 (2015).

32 Yang, R. et al. Genome-wide analysis associates familial colorectal cancer with increases in copy number variations and a rare structural variation at 12p12.3. Carcinogenesis 35, 315-323, doi:10.1093/carcin/bgt344 (2014).

33 Masson, A. L. et al. Copy number variation in hereditary non-polyposis colorectal cancer. Genes (Basel) 4, 536-555, doi:10.3390/genes4040536 (2013).

34 Venkatachalam, R. et al. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. Int J Cancer 129, 1635-1642, doi:10.1002/ijc.25821 (2011).

35 Moir-Meyer, G. L. et al. Rare germline copy number deletions of likely functional importance are implicated in endometrial cancer predisposition. Hum Genet 134, 269-278, doi:10.1007/s00439-014-1507-4 (2015).

36 Liu, B. et al. A Functional Copy-Number Variation in MAPKAPK2 Predicts Risk and Prognosis of Lung Cancer. The American Journal of Human Genetics 91, 384-390, doi:http://dx.doi.org/10.1016/j.ajhg.2012.07.003 (2012).

37 Iwakawa, R. et al. Contribution of germline mutations to PARK2 gene inactivation in lung adenocarcinoma. Genes Chromosomes Cancer 51, 462-472, doi:10.1002/gcc.21933 (2012).

38 Butler, M. W. et al. Glutathione S-transferase copy number variation alters lung gene expression. Eur Respir J 38, 15-28, doi:10.1183/09031936.00029210 (2011).

39 Shi, J. et al. Rare Germline Copy Number Variations and Disease Susceptibility in Familial Melanoma. J Invest Dermatol 136, 2436-2443, doi:10.1016/j.jid.2016.07.023 (2016).

40 Fidalgo, F. et al. Role of rare germline copy number variation in melanoma-prone patients. Future Oncol 12, 1345-1357, doi:10.2217/fon.16.22 (2016).

41 Sebat, J. et al. Large-scale copy number polymorphism in the human genome. Science 305, 525-528, doi:10.1126/science.1098918 (2004).

42 Iafrate, A. J. et al. Detection of large-scale variation in the human genome. Nat Genet 36, 949-951, doi:10.1038/ng1416 (2004).

43 Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. Nature 464, 704-712, doi:10.1038/nature08516 (2010).

44 Lee, C. & Scherer, S. W. The clinical context of copy number variation in the human genome. Expert Rev Mol Med 12, e8, doi:10.1017/S1462399410001390 (2010).

45 Pang, A. W. et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11, R52, doi:10.1186/gb-2010-11-5-r52 (2010).

46 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy Number Variation in Human Health, Disease, and Evolution. Annual review of genomics and human genetics 10, 451-481, doi:10.1146/annurev.genom.9.081307.164217 (2009).

47 Long, J. et al. A common deletion in the APOBEC3 genes and breast cancer risk. J Natl Cancer Inst 105, 573-579, doi:10.1093/jnci/djt018 (2013).

48 Xuan, D. et al. APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis 34, 2240-2243, doi:10.1093/carcin/bgt185 (2013).

49 Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713-720, doi:http://www.nature.com/nature/journal/v464/n7289/suppinfo/nature08979_S1.html (2010).

50 Azzato, E. M. et al. A genome-wide association study of prognosis in breast cancer. Cancer Epidemiol Biomarkers Prev 19, 1140-1143, doi:10.1158/1055-9965.EPI-10-0085 (2010).

51 Sapkota, Y. et al. Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm. PLoS One 8, e62550, doi:10.1371/journal.pone.0062550 (2013).

52 Rafiq, S. et al. A genome wide meta-analysis study for identification of common variation associated with breast cancer prognosis. PloS one 9, e101488, doi:10.1371/journal.pone.0101488 [doi] (2014).

53 Azzato, E. M. et al. Association Between a Germline OCA2 Polymorphism at Chromosome 15q13.1 and Estrogen Receptor–Negative Breast Cancer Survival. Journal of the National Cancer Institute 102, 650-662, doi:10.1093/jnci/djq057 (2010).

54 Jin, G. et al. Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer. Carcinogenesis 32, 1057-1062, doi:10.1093/carcin/bgr082 (2011).

55 Andersen, C. L. et al. Frequent genomic loss at chr16p13.2 is associated with poor prognosis in colorectal cancer. Int J Cancer 129, 1848-1858, doi:10.1002/ijc.25841 (2011).

56 Sapkota, Y. et al. Germline DNA copy number aberrations identified as potential prognostic factors for breast cancer recurrence. PLoS One 8, e53850, doi:10.1371/journal.pone.0053850 (2013).

57 Sehrawat, B. et al. Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human genetics 130, 529-537, doi:10.1007/s00439-011-0973-1 [doi] (2011).

58 Sapkota, Y., Narasimhan, A., Kumaran, M., Sehrawat, B. S. & Damaraju, S. A Genome-Wide Association Study to Identify Potential Germline Copy Number Variants

for Sporadic Breast Cancer Susceptibility. Cytogenet Genome Res 149, 156-164, doi:10.1159/000448558 (2016).

59 MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42, D986-992, doi:10.1093/nar/gkt958 (2014).

60 Genomes Project, C. et al. A global reference for human genetic variation. Nature 526, 68-74, doi:10.1038/nature15393 (2015).

61 Zhao, J. H. gap: Genetic Analysis Package. 2007 23, 18, doi:10.18637/jss.v023.i08 (2007).

62 H, Z. J. gap: Genetic Analysis Package. R package version 1.1-17. (2017).

63 Team, R. C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

. (2014).

64 Grambsch, T. M. T. a. P. M. Modeling Survival Data: Extending the Cox Model. Springer (2000).

65 Therneau, T. M. A Package for Survival Analysis in S . version 2.38. (2015).

66 Rose-Zerilli, M. J., Barton, S. J., Henderson, A. J., Shaheen, S. O. & Holloway, J. W. Copy-number variation genotyping of GSTT1 and GSTM1 gene deletions by real-time PCR. Clin Chem 55, 1680-1685, doi:10.1373/clinchem.2008.120105 (2009).

67 Liu, J. et al. The 29.5 kb APOBEC3B Deletion Polymorphism Is Not Associated with Clinical Outcome of Breast Cancer. PLoS One 11, e0161731, doi:10.1371/journal.pone.0161731 (2016).

68 Rezaei, M., Hashemi, M., Hashemi, S. M., Mashhadi, M. A. & Taheri, M. APOBEC3 Deletion is Associated with Breast Cancer Risk in a Sample of Southeast Iranian Population. International Journal of Molecular and Cellular Medicine 4, 103-108 (2015).

69 Norskov, M. S. et al. Copy number variation in glutathione-S-transferase T1 and M1 predicts incidence and 5-year survival from prostate and bladder cancer, and incidence of corpus uteri cancer in the general population. Pharmacogenomics J 11, 292-299, doi:10.1038/tpj.2010.38 (2011).

70 Yang, T. L. et al. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *Am J Hum Genet* **83**, 663-674, doi:10.1016/j.ajhg.2008.10.006 (2008).

71 Armengol, L. et al. Identification of Copy Number Variants Defining Genomic Differences among Major Human Groups. PLOS ONE 4, e7230, doi:10.1371/journal.pone.0007230 (2009).

72 de Cid, R. et al. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet 41, 211-215, doi:http://www.nature.com/ng/journal/v41/n2/suppinfo/ng.313_S1.html (2009).

73 Easton, D. F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. (2007).

74 Xuan, D. *et al.* APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis 34, doi:10.1093/carcin/bgt185 (2013).

75 Charrier, J., Maugard, C. M., Mevel, B. L. & Bignon, Y. J. Allelotype influence at glutathione S-transferase M1 locus on breast cancer susceptibility. Br J Cancer 79, 346-353, doi:10.1038/sj.bjc.6690055 (1999).

76 Syamala, V. S. et al. Influence of germline polymorphisms of GSTT1, GSTM1, and GSTP1 in familial versus sporadic breast cancer susceptibility and survival. Fam Cancer 7, 213-220, doi:10.1007/s10689-007-9177-1 (2008).

77 Yu, K.-D. et al. Genetic variants in GSTM3 gene within GSTM4-GSTM2-GSTM1-GSTM5-GSTM3 cluster influence breast cancer susceptibility depending on GSTM1. Breast Cancer Research and Treatment 121, 485-496, doi:10.1007/s10549-009-0585-9 (2010).

78 Carvalho, I., Milanezi, F., Martins, A., Reis, R. M. & Schmitt, F. Overexpression of platelet-derived growth factor receptor α in breast cancer is associated with tumour progression. Breast Cancer Research 7, R788, doi:10.1186/bcr1304 (2005).

79 Mills, G. B. & Moolenaar, W. H. The emerging role of lysophosphatidic acid in cancer. Nat Rev Cancer 3, 582-591, doi:10.1038/nrc1143 (2003).

80 van Corven, E. J., Groenink, A., Jalink, K., Eichholtz, T. & Moolenaar, W. H. Lysophosphatidate-induced cell proliferation: identification and dissection of signaling pathways mediated by G proteins. Cell 59, 45-54 (1989).

81 Geyer, F. C. et al. Genomic profiling of mitochondrion-rich breast carcinoma: chromosomal changes may be relevant for mitochondria accumulation and tumour biology. Breast Cancer Research and Treatment 132, 15-28, doi:10.1007/s10549-011-1504-4 (2012).

82 Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346-352, doi:10.1038/nature10983 (2012).

83 Kan, Z. et al. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466, 869-873, doi:10.1038/nature09208 (2010).

84 Liu, W. et al. Genetic factors affecting gene transcription and catalytic activity of UDP-glucuronosyltransferases in human liver. Hum Mol Genet 23, 5558-5569, doi:10.1093/hmg/ddu268 (2014).

85 Yu, K. D. et al. A functional polymorphism in the promoter region of GSTM1 implies a complex role for GSTM1 in breast cancer. FASEB J 23, 2274-2287, doi:10.1096/fj.08-124073 (2009).

86 Persengiev, S., Kondova, I. & Bontrop, R. Insights on the functional interactions between miRNAs and copy number variations in the aging brain. Frontiers in Molecular Neuroscience 6, 32 (2013).

87 Marcinkowska, M., Szymanski, M., Krzyzosiak, W. J. & Kozlowski, P. Copy number variation of microRNA genes in the human genome. BMC Geno*mics* **12**, 183, doi:10.1186/1471-2164-12-183 (2011).

4 Breast cancer associated germline structural variants harboring small noncoding RNAs impact post-transcriptional gene regulation[1]

4.1. Introduction

Globally, breast cancer (BC) is one of the most common cancers diagnosed among women[1]. It is estimated from twin studies that genetic factors contribute up to 30% of the risk for breast cancer[2]. To date, high, moderate and low penetrance single nucleotide variants associated with breast cancer explained only 50% of the heritable risk and much of the remaining genetic susceptibility (so-called missing heritability) remains unexplored[3,4]. However, majority of these variants are present in the intronic or intergenic regions and therefore precludes delineation of their role in breast cancer pathogenesis. Therefore, there is a need to explore the significance of other forms of genetic variants for their role in breast heritability.

Copy Number Variations (CNVs), a class of structural variations of DNA (> 50 bp in size), which includes amplification or deletion of genomic segments. CNVs can influence phenotype in a variety of ways: through gene dosage (correlation of copy status and ensuing tissue specific gene expression changes), partial deletions in genic regions

[1] A version of this chapter has been published. Kumaran et al., 2018. Scientific Reports. © 2018 Kumaran et al. This article published in Scientific Reports is licensed under a Creative Commons Attribution 4.0 International License. The author owns the copyright for the article. Ms. Jennifer Dufour is acknowledged for technical support. The results published or shown here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

leading to fusion genes, or complete deletions of genes, and lastly, changes that lead to more complex levels of *cis* or *trans* regulatory functions[5,6].

Recently, genetic susceptibility has been explained in part by common germline CNVs (>5% in frequency) and rare germline CNVs (1-5% in frequency) for sporadic and familial breast cancers, respectively[6,7]. A common germline CNV deletion affecting *APOBEC3* loci resulted in a fusion protein, *APOBEC3A_B*, which was reported to confer breast cancer susceptibility in diverse populations[6,8,9]. Recently, I demonstrated that germline CNVs overlapping with protein coding genes are associated with breast cancer risk and prognosis. Also the associated CNVs showed gene dosage effects*, i.e.,* germline copy status (gain, loss or diploid status) and showed correlation with breast tissue gene expression[7]. Even though previous studies have suggested that a significant proportion of CNVs reside in the intergenic regions which harbor non-coding genes, there were no direct studies to address their relevance to breast cancer. I reasoned that studies of germline CNVs harboring small non-coding RNAs (hereafter referred to as CNV-sncRNAs) such as microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs) and transfer RNAs (tRNAs) and their relative levels of expression in breast tissues potentially offers biological insights into the role of CNV-sncRNAs in breast cancer risk.

The sncRNAs are less than 200 nucleotides in size and include different classes of RNAs – miRNAs, piRNAs, snoRNAs and tRNAs. While miRNAs and piRNAs are known post-transcriptional regulators of gene expression, snoRNAs and tRNAs are also currently being investigated as potential regulators of gene expression. Although the canonical roles of snoRNAs and tRNAs include RNA modification/splicing and translation, respectively, novel functions of these RNAs are emerging. The nucleotide sequences within these RNAs show sequence homology with mature miRNAs and piRNAs. snoRNAs and tRNAs may undergo nucleolytic processing to unmask cryptic miRNAs and piRNAs. Dysregulation of all four classes of sncRNAs has been observed in various cancer types, including breast cancer, and its clinical significance has been addressed in some detail (miRNAs and piRNAs)[10,11] or is emerging (snoRNAs and tRNAs)[12,13].

Germline single nucleotide polymorphisms (SNPs) present in pre-miRNA regions are known to affect their biogenesis and target binding efficiencies of miRNAs, thereby influencing disease predisposition[14-16]. Germline CNVs may also affect disease predisposition by independent mechanisms. For instance, a copy number deletion of a miRNA cluster present on chr22q11.2 locus is a classic example of a germline CNV as a genetic determinant of schizophrenia[17-19]. Additionally, germline CNVs and their embedded miRNAs (CNV-miRNAs) were shown to be associated with autism[20], roles in brain aging and neurodegeneration[21] and congenital heart disease[22]. Prior studies have predicted that the target genes conferring the phenotypes are likely regulated by CNV-miRNAs[19]. However, there is no direct experimental evidence to support this premise.

I hypothesized that germline CNVs are associated with the phenotype of breast cancer, and that CNV-sncRNAs are indeed expressed in breast tissues, show gene dosage effects and mediate the regulation of downstream target genes. I show evidence in support of this hypothesis and offer insights on the role of disease associated CNVs. Firstly, I identified germline breast cancer associated CNVs using a genome wide association study (GWAS) design (Fig. 1) and identified embedded sncRNA gene regions. Secondly, I showed that sncRNAs originating in CNVs are indeed expressed in breast tissues and show correlation with germline copy status. Thirdly, I identified the target mRNAs regulated by CNV-miRNAs. I therefore infer that cancer associated CNVs harboring sncRNAs contribute to the pathogenesis of breast cancer.

4.2. Methods

I performed all the experiments and analysis, unless otherwise indicated in the text

4.2.1. Study ethics approval

The study was approved by the local Health Research Ethics Board of Alberta (HREBA) - Cancer Committee. Written informed consents were obtained from all study participants. All experiments performed using specimens from study samples were carried out under approved guidelines and regulation.

4.2.2. Study subjects and whole genome platforms

A schematic of the overall study design is summarized (Figure 4.1) and details of the protocols followed are summarized below. The flowchart depicts the overall study design, summary of the datasets, and experimental platforms used at each stage of the analysis. Detailed protocols and data analysis methods are discussed in the methods section.

4.2.3. Discovery dataset

The study included women from Alberta, Canada with confirmed diagnosis of invasive breast cancer (cases, n=422)[7,23]. The cases were non-metastatic at the time of diagnosis. Biological specimens and clinical-pathological information were accessed from the

Alberta Cancer Research Biobank, located at the Cross-Cancer Institute, Edmonton, Alberta, Canada[24].The controls (n=348) included in this study were age matched healthy women (no personal or family history of cancer at the time of recruitment). The controls were accessed from a prospective cohort study called the Tomorrow Project[25] based in Alberta, Canada. Affymetrix Human SNP 6.0 array data and information about the study participants and the specimens can be found elsewhere[23,26] and in the ensuing text.

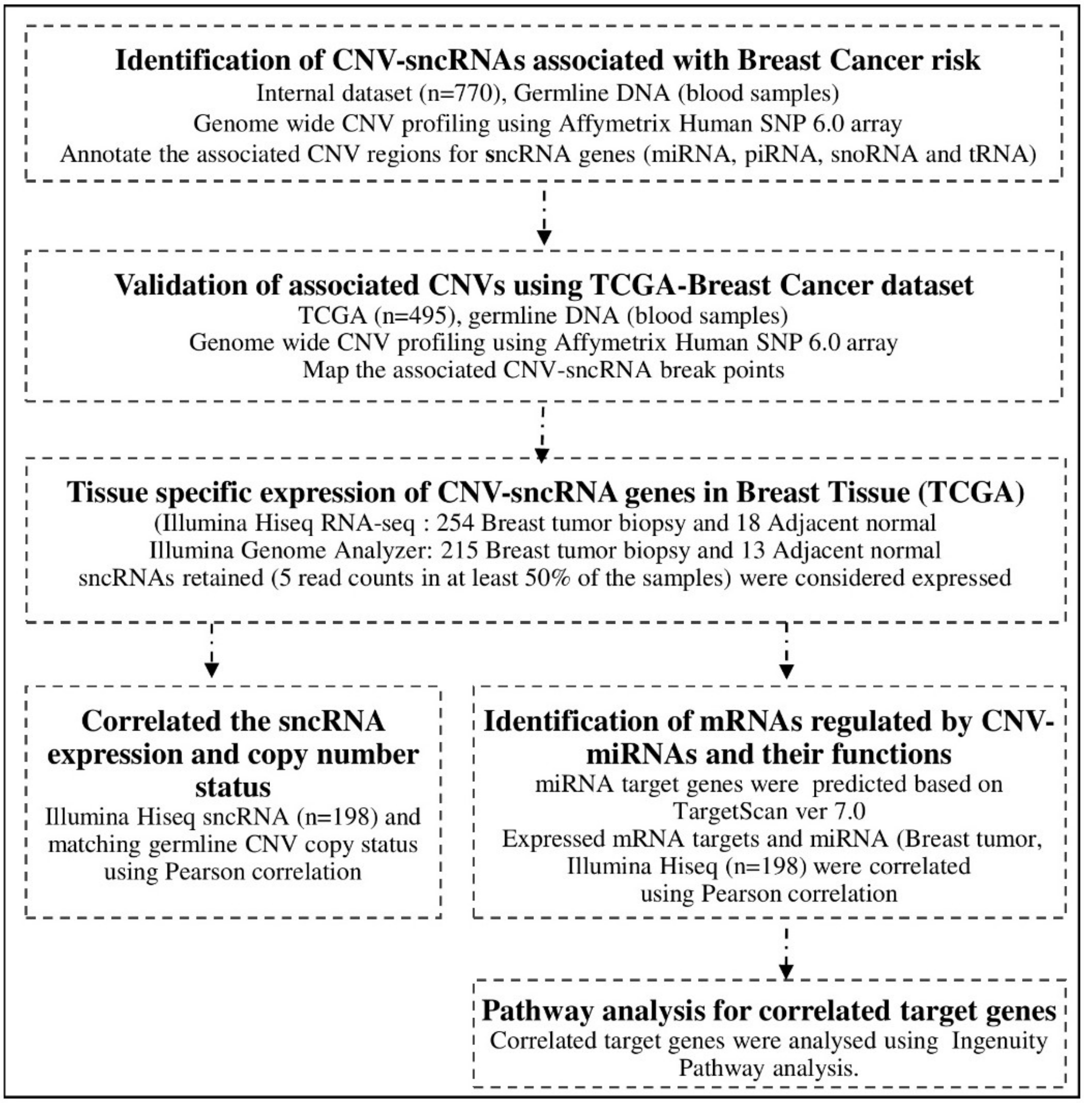

Identification of CNV-sncRNAs associated with Breast Cancer risk
Internal dataset (n=770), Germline DNA (blood samples)
Genome wide CNV profiling using Affymetrix Human SNP 6.0 array
Annotate the associated CNV regions for sncRNA genes (miRNA, piRNA, snoRNA and tRNA)
Validation of associated CNVs using TCGA-Breast Cancer dataset
TCGA (n=495), germline DNA (blood samples)
Genome wide CNV profiling using Affymetrix Human SNP 6.0 array
Map the associated CNV-sncRNA break points
Tissue specific expression of CNV-sncRNA genes in Breast Tissue (TCGA)
(Illumina Hiseq RNA-seq : 254 Breast tumor biopsy and 18 Adjacent normal
Illumina Genome Analyzer: 215 Breast tumor biopsy and 13 Adjacent normal
sncRNAs retained (5 read counts in at least 50% of the samples) were considered expressed
Correlated the sncRNA expression and copy number status
Illumina Hiseq sncRNA (n=198) and matching germline CNV copy status using Pearson correlation
Identification of mRNAs regulated by CNV-miRNAs and their functions
miRNA target genes were predicted based on TargetScan ver 7.0
Expressed mRNA targets and miRNA (Breast tumor, Illumina Hiseq (n=198) were correlated using Pearson correlation
Pathway analysis for correlated target genes
Correlated target genes were analysed using Ingenuity Pathway analysis.

Figure 4.1 Study design

The flowchart depicts the overall study design, summary of the datasets, and experimental platforms used at each stage of the analysis. Detailed protocols and data analysis methods are discussed in the methods section.

4.2.4. Validation dataset (The Cancer Genome Atlas Project, TCGA)

I have accessed the dataset from TCGA study with cases diagnosed with invasive breast cancer. This study meets the publication guidelines provided by TCGA (http://cancergenome.nih.gov/publications/publicationguidelines). I accessed level 1 and level 3 TCGA datasets for Whole Genome Copy number profiles, small RNA sequencing data and mRNA sequencing datasets, respectively. The datasets were available for 1088 Invasive breast cancer cases. I selected 516 cases based on the study inclusion criteria: i) no history of other malignancy, ii) no metastasis at the time of diagnosis and iii) diagnosis of invasive ductal or lobular carcinoma.

4.2.5. Germline CNV dataset from TCGA: Affymetrix Human SNP array 6.0 platform

I utilized Affymetrix generated (.CEL files) data from germline DNA. Based on the SNP genotype calls for the 516 cases, I performed population stratification analysis using Principal Component Analysis (PCA) as described in the ensuing text. I identified 495 cases with Caucasian ancestry which were used for the down-stream analysis.

4.2.6. Breast tissue transcriptome data set from TCGA for small non-coding RNAs: Next Generation Sequencing platform

I accessed datasets for small RNA sequencing files (level 1 data; .bam files) matching to 495 cases of Caucasian ancestry. Of these, sequencing data were available for 469 breast tumor tissues. However, for a subset of cases data were available on both tumor and adjacent normal tissues specimens. Sequencing data from Illumina HiSeq and Genome Analyzer (GA) platforms from TCGA were accessed (254 breast tumor samples and 18 adjacent normal samples from HiSeq and 215 breast tumor samples and 13 adjacent normal samples from GA).

4.2.7. Breast tissue transcriptome data set from TCGA for mRNAs: Next Generation Sequencing platform

I accessed mRNA sequencing data from breast tumors generated on Illumina HiSeq platform. Level 3 data (Reads Per Kilobase Million, RPKM normalized) was used for all analysis. mRNA sequencing data was available for 198 cases and these were matched with the data available for sncRNAs on the same HiSeq platform. This enabled the identification of post-transcriptionally regulated target mRNAs by CNV-miRNAs.

4.2.8. DNA extraction

DNA was extracted from peripheral blood samples of cases and controls (discovery dataset, n=770). DNA isolation was carried out by using commercially available QiagenTM (Mississauga, Ontario, Canada) DNA isolation kits, as described earlier [23,26].

4.2.9. Genotyping and quality control

DNAs extracted from study samples was genotyped using Affymetrix Human SNP array 6.0 following manufacturer's protocol and are described elsewhere [26]. Affymetrix SNP array 6.0 has an independent set of probes for SNPs and CNVs. Genotyping quality control was assessed using Birdseed V2 algorithm in Affymetrix genotyping console. Sample Contrast Quality Control (CQC) ≥ 1.7 indicates acceptable genotyping quality. All study samples (both discovery and validation data) had a CQC values > 2.

4.2.10. Population stratification

Principle component analysis was performed using EIGENSTRAT algorithm implemented in Golden Helix SNP and Variation suite v8.5.0. Genotype data from 270 HapMap samples were used as reference to infer genetic ancestry of the study samples. Variance was accounted for by the top two principal components and a threshold of three standard deviations was set to determine the outliers.

Of the 770 samples in the discovery dataset, 686 samples co-clustered with the European ancestry samples from the HapMap data, and 84 samples were identified as outliers. Of the 516 TCGA samples, 495 samples were identified as belonging to the European ancestry and 21 samples were removed as outliers. Identity by descent (IBD) analysis did not reveal any cryptic relatedness among the study subjects as judged from the pair-wise correlation cut off < 0.25 in both datasets.

4.2.11. Copy number estimation and association analysis

Copy Number Analysis was performed using Partek® Genomics Suite™ 6.6 (PGS) and the default parameters as described below. Affymetrix. CEL files served as the source files. The CNV analysis was performed for 686 samples (320 controls and 366 cases) and all sample normalization was used to create a reference baseline to infer the relative copy number estimate. Genomic segmentation algorithm implemented in the software was used to call the genomic segments based on the following default criteria: genomic markers >10; segmentation p-value threshold = 0.001; Signal/Noise (S/N) ratio = 0.3. The copy number status for each inferred segment was assigned based on the normalized intensity as diploid copy number = 1.7-2.3, copy gain >2.3 and copy loss <1.7. CNV association analysis was performed using 2X3 Chi-square association test estimates the difference in frequency of a CNV (gain/loss/diploid) between the cases and controls. Data was corrected for multiple hypothesis testing using Benjamin-Hochberg false discovery rate method and CNVs with q-value < 0.05 were considered significant.

CNV estimation for the 495-breast cancer TCGA samples (validation set) was performed similar to the discovery dataset, except for the normalization. I used HapMap 270 samples as a reference for a diploid status (controls) to infer copy status in TCGA samples (cases). Associated CNV regions and break-points from the discovery data set were mapped to the CNV profiles and break-points in TCGA samples.

4.2.12. Gene annotation for the CNV regions

Breast cancer associated CNV regions were annotated for sncRNAs from the following sources: mature miRNAs using miRBase ver20 [27], snoRNAs using Ensembl [28], piRNAs using piRNAdb [29] and tRNAs [30] using UCSC genome browser. Protein coding and lncRNA genes were annotated using UCSC.

4.2.13. Expression analysis of sncRNAs

Partek® Genomics Suite was used for the analysis of sncRNAs and .bam files as a source of sequence data. TCGA samples (both breast tumor and adjacent normal tissues) sequenced using Illumina HiSeq platform and Genome Analyzer were analyzed separately using PGS. sncRNA annotation was based on the database sources described above. For sncRNA expression analysis, a cut-off at least 5 read counts in 50% of the samples was considered for further analysis. I restricted integrative analysis of CNV status, sncRNAs and mRNAs to HiSeq data because read depths may vary between HiSeq and GA platforms.

4.2.14. Correlation of the breast tissue expression of sncRNAs with germline copy number estimates

It was important to ascertain if there was a correlation between CNV copy status and expression of CNV embedded genes (e.g., encoding sncRNAs) in breast tumor tissues to assess the role of the latter in disease risk. I used Pearson Correlation analysis (p-value <0.1) to demonstrate the relationship between copy status and sncRNA expression. I used 198 samples with germline CNV data and compared with sncRNA expression in matched breast tumor tissues from the TCGA cohort. sncRNA read counts (5 counts in at least 50% of the samples as a cut-off) were RPKM normalized and log-transformed to compare with the germline copy status as a categorical variable. Copy number status for each inferred segment was assigned based on the normalized intensity as diploid copy number (i.e., 1.7-2.3), with copy gain > 2.3 and copy loss < 1.7, as described above. Even though sncRNAs may originate from multiple genomic locations, I considered only expression of RNAs present within the breast cancer associated CNV regions.

4.2.15. Target predictions for miRNAs embedded within CNVs, tissue level mRNA-miRNA expressions and correlations with copy status

Target mRNAs for the 10 miRNAs were predicted *in silico* using TargetScan version 7.1. I accessed level 3 data for mRNA (HiSeq) from the TCGA cohort which is RPKM normalized and log-transformed. All of the predicted targets were expressed in the HiSeq mRNA data (albeit at varying expression levels). I performed RPKM normalization and log transformation of the miRNA expression data from HiSeq. The samples (n=198) were initially classified into two groups based on their copy number status; Diploid and copy gains. Correlated mRNA-miRNAs were identified using Pearson Correlation coefficients and a negative correlation with $r \leq - 0.2$ and p-value <0.05 was considered as indicative of regulated genes.

4.2.16. Ingenuity Pathway Analysis (IPA)

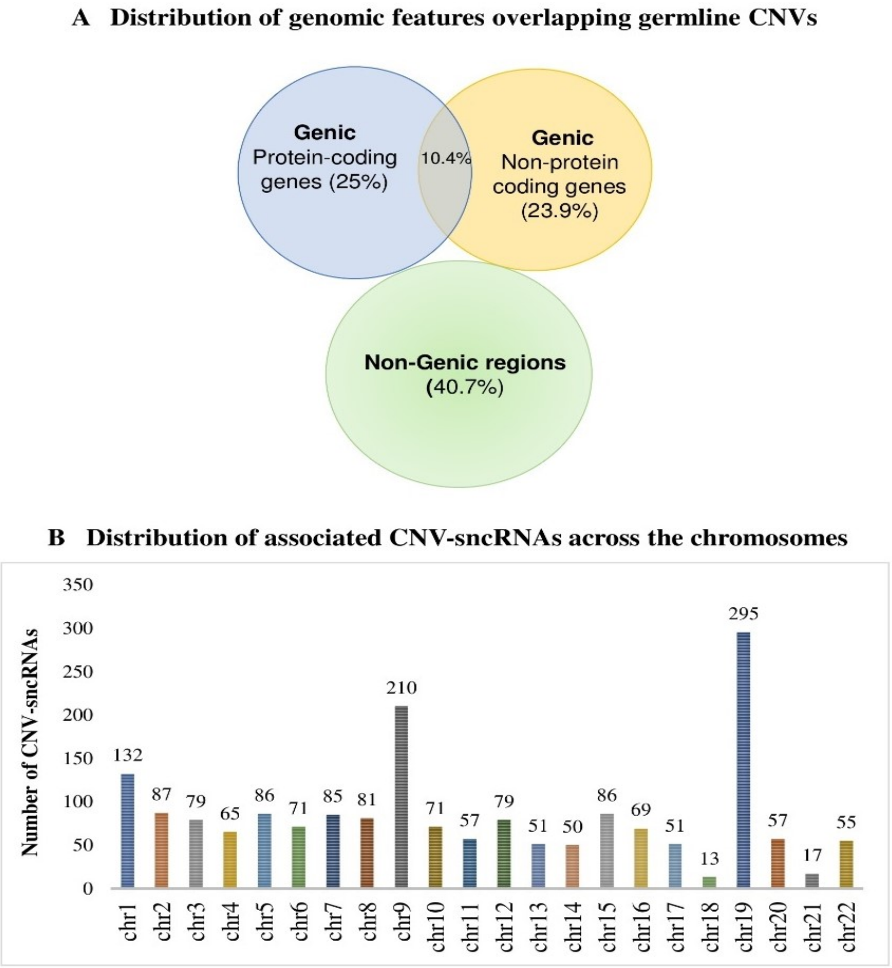
Data were analyzed using the IPA platform (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis) to identify potentially affected pathways. Coding genes targeted by miRNAs were used as an input to assess the pathways involved. Separate analysis was conducted for the genes identified in the stratified groups based on copy status. Enrichment p-value <0.05 was considered significant.

4.3. Results

4.3.1. Identification of germline CNVs encompassing sncRNA genes and their association with breast cancer risk

I conducted a GWAS (discovery dataset) using 366 cases/320 controls and germline CNVs as polymorphic markers. I identified 7496 CNVs that were associated with breast cancer risk (q-value < 0.05) [7]. Of these, 59.3% of the CNVs mapped to genic regions including protein coding genes, non-coding RNA genes and pseudogenes and the remaining 40.7% mapped to the non-genic regions. Among, the CNVs mapping to the genic regions, 25.0% (n=1876) mapped to protein coding genes and another 23.9% CNVs (n=1789) mapped to non-coding RNA genes, including genes for long non-coding RNAs, sncRNAs and to pseudogenes. I observed that 10.4% of the breast cancer associated CNVs (n=776) mapped to both protein coding and non-coding genes because introns of the protein coding genes also serve as a source of non-coding RNAs **(**Figure 4.2a). I have earlier described CNVs with embedded protein coding genes and their relevance to breast cancer [7]. Of the total 2565 CNVs (1789 non-coding RNA genes plus 776 non-coding RNA genes originating from protein coding introns), I considered 1812 CNVs harboring four classes of sncRNA genes (miRNAs, piRNAs, snoRNAs and tRNAs) for further analysis as these are known to play a role in post-transcriptional gene regulatory mechanisms.

The distribution of sncRNA genes within the 1812 breast cancer associated CNVs included miRNA (n=38) and tRNA genes (n=15), embedded within 26 and 10 CNVs, respectively. Each of the miRNA and tRNA genes that mapped within CNVs were non-redundant, in that none originated from multiple chromosomal locations. In contrast, piRNAs and snoRNAs showed redundancy, in that the same piRNA or snoRNA genes were found within multiple CNV loci across chromosomes. For instance, 9865 redundant piRNA genes were mapped to 1760 CNVs regions, of which 1292 piRNAs were unique. Seventy-one (or 66 non-redundant) snoRNAs were mapped to 52 CNV regions. (provided as electronic Supplementary Table S1 at https://doi.org/10.1038/s41598-018-25801-1). Individual frequencies of CNVs in cases and controls as well as the copy gain or copy loss frequencies are also summarized to facilitate comparisons. The average size of the associated CNVs was about 25kb (range 50bp to 9Mbp). The number of sncRNA genes present within a CNV varied from 2 and 240, depending on the size of the CNV. About 36 CNVs harbored more than one class of sncRNAs, and piRNAs genes were predominant (provided as electronic Supplementary Table S1 at https://doi.org/10.1038/s41598-018-25801-1). Chromosomes 19, 9 and 1 showed the highest number of breast cancer associated CNVs, (295, 210 and 132, respectively), harboring sncRNAs (Figure 4.2b), relative to other chromosomes. In summary, I have not only identified CNVs associated with breast cancer risk across the genome, but also the embedded CNV-sncRNAs. I identified CNVs that overlapped with SNORD-115 and SNORD-116 clusters (chr15: 25296245-25326762) and were found to be associated with breast cancer (provided as electronic Supplementary Table S1 at https://doi.org/10.1038/s41598-018-25801-1). Deletion of these clusters were initially described in patients with Prader-Willi Syndrome (PWS)[31]. In our study, the SNORD locus showed both copy-gain (5 -14%) and copy-loss (3-8%) in the cases but not in controls.



A Distribution of genomic features overlapping germline CNVs

B Distribution of associated CNV-sncRNAs across the chromosomes

Figure 4.2 Genome wide distribution of germline CNVs

In the Figure 4.2a, the distribution of genomic features overlapping germline CNVs are shown. Figure shows a Venn diagram of the genome wide distribution of germline CNVs associated (q<0.05) with breast cancer. Represented genic regions were: protein coding (25%) and non-protein coding genes including pseudogenes and small and long non-coding RNAs (23.9%). An overlap of these regions (10.4%) capture non-coding RNAs originating from the intronic regions of the coding genes. 40.7% of CNVs do not show embedded genes (genome build hg19), hence labelled as non-genic regions. In Figure 4.2b, Distribution of associated CNV-sncRNAs across the chromosomes are shown. This figure illustrates the distribution of breast cancer associated CNVs (q<0.05) harboring small non-coding RNA genes (miRNA, piRNA, tRNA and snoRNAs) for all chromosomes

.

4.3.2. Validation of CNV breakpoints in TCGA dataset

GWAS (n=686) allowed us to identify CNVs (with embedded sncRNAs) that are associated with breast cancer risk. I used the TCGA cohort as a validation dataset to address the following: Firstly, to validate the CNVs from the discovery stage GWAS and to assess the replicability of copy number estimates between the datasets called by the same algorithm. Secondly, to examine breast tissue specific expression of sncRNAs embedded within CNVs. Thirdly, to identify regulatory potential of miRNAs (subset of all sncRNAs identified) using mRNA expression dataset from the same breast tumors from which sncRNAs were profiled.

I successfully mapped the 1812 CNVs (with embedded sncRNAs) from the discovery dataset to the TCGA dataset, thus validating the copy number estimates called by the algorithm (provided as electronic Supplementary Table S2 at https://doi.org/10.1038/s41598-018-25801-1). For comparisons of CNV break points in the discovery and TCGA data sets, I defined 100% overlap as those CNVs that had break points exactly matching or embedded within CNVs identified from either of the datasets. CNVs may have an influence on the level of expression of sncRNAs, and regulation of their downstream target mRNAs by diverse mechanisms. There is evidence to suggest that CNVs overlapping miRNA genes are more likely to exhibit phenotypic effects[32], and I now extend this premise for other sncRNAs. Subsequent data analysis was based on TCGA cohorts for breast tissue expression analysis of sncRNAs and mRNAs from the matched samples**.**

4.3.3. Breast tissue specific expression of CNV-sncRNAs in TCGA dataset

Detailed analysis of sncRNAs identified in breast tumors and adjacent normal tissues using HiSeq (n=254) and Genome Analyzer, (GA) (n=215) platforms are summarized in Appendix Table A.12. Breast tissue specific expression of sncRNAs (miRNAs, piRNAs, snoRNAs and tRNAs) were analyzed. I compared the total number of sncRNAs expressed with the total number of sncRNAs originating from within the CNV regions. The total number of sncRNAs expressed were comparable between normal and tumor tissues. Similarly, I have also compared the total number of CNV-sncRNAs showing expression in normal and tumor tissues. (Figure 4.3). Overall, I have identified 38 CNV-sncRNAs (14 miRNAs, 1 piRNA, 11 snoRNAs and12 tRNAs) expressed in both breast tumors and adjacent normal tissues. While CNV embedded snoRNAs, tRNAs and piRNAs were expressed similarly in both tumor or adjacent normal tissues, a subset of miRNAs detected were present either in tumor or normal tissues. Five of the miRNAs (hsa-miR-154-3p, hsa-miR-4999-5p, hsa-miR-382-3p, hsa-miR-487a-5p, hsa-miR-539-5p) were expressed only in adjacent normal tissues, at the cut-off criteria of 5 read counts in 50% of the samples. Using a similar cut-off criterion, one miRNA (hsa-miR-4746-5p) was expressed only in tumor tissues (Appendix Table A.13). A higher number of piRNA genes mapped to the breast cancer associated CNVs. However, CNV-piRNA, hsa-piR-20636 was the only one expressed in breast tumor tissue. In case of the snoRNA, I noted the C/D box SNORD 116 from the PWS loci showed expression in both breast tumors and adjacent normal tissues.

Breast cancer associated CNV regions showing overlap between discovery and validation datasets and harboring the embedded sncRNAs (n=38) are summarized (Table 4.1). It is interesting to note that 27% of CNVs (showing expression of embedded sncRNAs) were also reported as copy variable regions in the 1000 Genomes Phase 3 Project. A majority of the CNV frequencies were higher in cases relative to controls, thereby explaining the limited overlap with the 1000 Genomes data which is generated from the control populations.

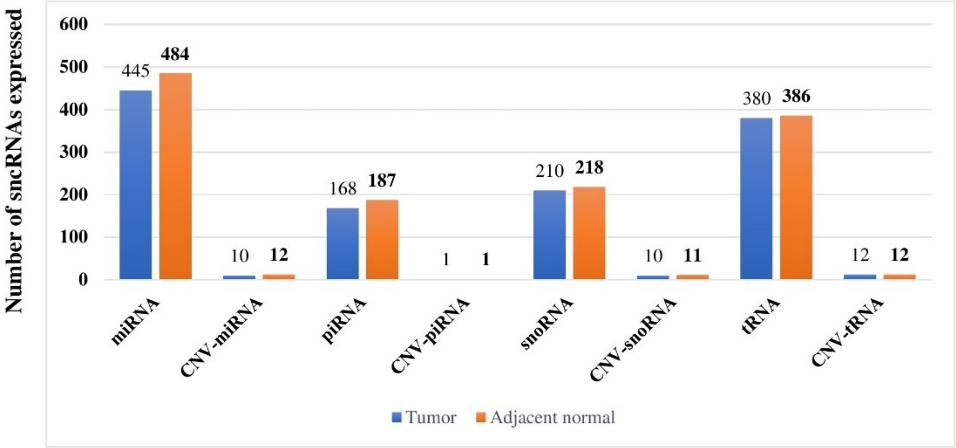


Figure 4.3 Expression profiles of small non-coding RNAs in breast tumor and adjacent normal tissues (HiSeq)

This figure illustrates the expression profiles from the four classes of sncRNAs between tumor and adjacent normal tissues. Individual bar graphs capture the expressed total sncRNAs and CNV-sncRNAs. Data presented is from TCGA Illumina Hiseq (n=254 cases and 18 adjacent normal).

Table 4.1 Germline CNVs in discovery cohort showing association with breast cancer risk and expression of embedded small RNAs in breast tumor tissues from TCGA

Discovery Dataset							TCGA Dataset	
CNV region	Cytoband	length (bps)	p-value	q-value	CNV frequency gain/loss (%)		CNV region	Small RNAs expressed in breast tumors
					Cases	Controls		
*chr14:101513466-101514318	14q32.31	853	7.71E-05	9.21E-04	5/1	0/0	chr14:101513466-101517099	hsa-miR-539-5p (+), hsa-miR-889-3p (+)
*chr14:101515194-101519779	14q32.31	4586	4.84E-05	6.52E-04	5/1	0/0	chr14:101513466-101517099; chr14:101517099-101527707	hsa-miR-655-3p (+), hsa-miR-487a-5p
*chr14:101519779-101525402	14q32.31	5624	5.53E-05	7.27E-04	5/1	0/0	chr14:101517099-101527707	hsa-miR-134-3p (+), hsa-miR-134-5p (+), hsa-miR-323b-3p (+), hsa-miR-382-5p (+), hsa-miR-485-3p (+), hsa-miR-382-3p
*chr14:101525779-101527707	14q32.31	1929	8.94E-04	5.41E-03	4/1	0/0	chr14:101517099-101527707	hsa-miR-154-3p (+), hsa-miR-154-5p (+),
chr19:4437681-4494605	19p13.3	56925	3.09E-04	2.53E-03	3/2	0/0	chr19:4424993-4664433	hsa-miR-4746-5p (+)
chr1:149676729-149684202	1q21.2	7474	9.33E-06	1.77E-04	2/5	0/16	chr1:149676729-149684202	hsa-piR-20636
chr15:25296245-25297449	15q11.2	1205	4.32E-04	3.26E-03	5/1	0/0	chr15:25296245-25297449	snoRNA_SNORD116-1-201 (+)
chr15:25297449-25300158	15q11.2	2710	5.92E-07	1.92E-05	8/1	0/0	chr15:25298903-25300158	snoRNA_SNORD116-2-201 (+)
*chr15:25300158-25306451	15q11.2	6294	2.26E-07	8.49E-06	9/1	0/0	chr15:25300158-25304384; chr15:25305396-25308383	snoRNA_SNORD116-3-201 (+)

Discovery Dataset							TCGA Dataset	
CNV region	Cytoband	length (bps)	p-value	q-value	CNV frequency gain/loss (%)		CNV region	Small RNAs expressed in breast tumors
					Cases	Controls		
chr15:25307985-25310508	15q11.2	2524	6.12E-08	2.82E-06	9/1	0/0	chr15:25305396-25308383; chr15:25308383-25310928	snoRNA_SNORD116-6-201 (+)
chr15:25310508-25316405	15q11.2	5898	9.95E-08	4.25E-06	9/1	0/0	chr15:25310928-25318258	snoRNA_SNORD116-8-201 (+)
chr15:25316405-25318258	15q11.2	1854	2.62E-07	9.64E-06	8/1	0/0	chr15:25310928-25318258	snoRNA_SNORD116-9-201 (+)
chr15:25318258-25324279	15q11.2	6022	9.95E-08	4.25E-06	8/2	0/0	chr15:25318258-25325686	snoRNA_SNORD116-9-201 (+) ,
chr15:25324512-25325686	15q11.2	1175	2.87E-06	6.76E-05	6/2	0/0	chr15:25318258-25325686	snoRNA_SNORD116-14-201 (+)
chr15:25325686-25326762	15q11.2	1077	4.61E-06	9.87E-05	6/1	0/0	chr15:25325686-25326762	snoRNA_SNORD116-15-201 (+)
chr16:2011427-2016398	16p13.3	4972	6.98E-04	4.58E-03	3/2	0/1	chr16:2011427-2016398	snoRNA_SNORA10-201 (-), snoRNA_SNORA64-201 (-)
chr19:3975155-3984201	19p13.3	9047	3.09E-04	2.53E-03	3/2	0/0	chr19:3768181-4110048	snoRNA_SNORD37-201 (-)
chr1:148580449-148606453	1q21.2	26005	7.50E-09	4.65E-07	7/14	10/32	chr1:148580449-148632305	chr1.trna108-AsnGTT (-)
chr1:148705208-148768557	1q21.2	63350	7.26E-04	4.72E-03	4/11	4/22	chr1:148662374-148789654	chr1.trna107-AsnGTT (-)
chr1:149598086-149617469	1q21.2	19384	4.48E-10	4.08E-08	9/12	2/29	chr1:149598086-149631220	chr1.trna30-AsnGTT (+),
chr1:149661965-149670179	1q21.2	8215	3.70E-06	8.35E-05	4/8	1/19	chr1:149652461-149676729	chr1.trna94-GluTTC (-)

Discovery Dataset							TCGA Dataset	
CNV region	Cytoband	length (bps)	p-value	q-value	CNV frequency gain/loss (%)		CNV region	Small RNAs expressed in breast tumors
					Cases	Controls		
chr1:149670179-149676729	1q21.2	6551	3.60E-06	8.17E-05	2/6	0/17	chr1:149652461-149676729	chr1.trna92-PheGAA (-)
chr1:149676729-149684202	1q21.2	7474	9.33E-06	1.77E-04	2/5	0/16	chr1:149676729-149684202	chr1.trna90-ValCAC (-), chr1.trna91-GlyCCC (-)
chr6:26286287-26287456	6p22.2	1170	2.38E-04	2.13E-03	3/4	0/1	chr6:26274458-26287456	chr6.trna2-MetCAT (+)
*chr19:1381502-1407359	19p13.3	25858	1.23E-04	1.29E-03	4/2	0/0	chr19:1342160-1547869	chr19.trna1-AsnGTT (+), chr19.trna14-PheGAA (-)
*chr19:4658652-4771070	19p13.3	112419	3.09E-04	2.53E-03	3/2	0/0	chr19:4714925-4751218	chr19.trna13-ValCAC (-), chr19.trna2-GlyTCC (+)

The above table represents the selected CNV regions associated with breast cancer that also included one of the four classes of sncRNAs. The statistics represented in this table are based on the discovery dataset (cases/control =686) and includes the CNV region mapped in validation dataset (TCGA). These sncRNAs were expressed in the breast tissue (either breast tumor or adjacent normal tissues or both) in the TCGA dataset. The rows marked with * symbol indicates the CNVs that are also seen as copy number variable regions in 1000 genomes Phase 3 project.

4.3.4. Correlation of expressed CNV-sncRNAs to copy status

CNVs are known to confer gene dosage effects among protein coding genes[7,33], and whether or not CNV-sncRNAs also show gene dosage effects was investigated. Correlation of the expression of the CNV-sncRNAs with corresponding copy status was addressed using Pearson Correlation analysis. Overall, 15 sncRNAs (one piRNA, eight tRNAs, six snoRNAs) showed correlation (Appendix Table A.14 and Appendix Figure A.2); of these 13 correlated at p-value <0.05 and two correlated at p-value < 0.1. One piRNA and five tRNAs showed positive correlation whereas three tRNAs and six snoRNAs showed negative correlations. The positively correlated sncRNA genes showed r=14% to 21% and p-values 10^{-2} to 10^{-3}. Negatively correlated snoRNAs showed r= -13% to -45% and p-values 10^{-2} to 10^{-11}. Expression and regulation of sncRNAs are thus complex; while a positive correlation with copy status indicates potential gene dosage effects, a negative correlation may potentially indicate gene disruption or epigenetic regulation. This kind of negative correlations were also noted by others[34] and there is no clear consensus mechanisms identified to explain these correlations. I observed that negatively correlated tRNAs originated from intergenic regions, whereas negatively correlated snoRNAs originated from intronic regions. I did not observe any significant correlations between copy status and miRNA expression. This could be due to the diverse mechanisms regulating miRNA expression. I could not distinguish if the CNV-miRNA itself is regulated by upstream elements within the CNV region or a combination of all the above.

4.3.5. Gene targets for CNV-miRNAs and pathway analysis

I reasoned that a germline copy status for CNV-miRNA may show pronounced effects on downstream mRNA targets. To demonstrate such effects, I stratified breast cancer cases (mRNA expressions from n=198 breast tumors from HiSeq Platform) based on germline status. Therefore, a correlation between miRNA and mRNA expressions may reveal higher number of targets that are regulated as a function of CNV copy status, as an indirect measure of miRNA copies. For instance, I examined CNV embedded hsa-miR-4746-5p in 198 breast cancer cases; 52 cases exhibited copy gains and 146 were diploid. Gene targets for the CNV-hsa-miR-4746-5p were predicted using TargetScan and these predicted targets were identified in the mRNA expression data sets (HiSeq platform). A correlation analysis revealed 25 common target genes for both diploid and copy gain cases; an additional 29 targets were identified for copy-gain cases (Appendix Table A.15). The miRNA-mRNA correlation (r) values were from -0.20 to-0.34; and from -0.27 to -0.42, for the diploid and copy gain cases respectively. The targets regulated by hsa-miR-4746-5p among the copy gain cases were enriched for key signaling molecules (growth hormone*, FLT3, NGF, PTEN,* G-protein coupled receptor) and glutamine biosynthesis pathways. The identified targets in this study have been well addressed in literature for their association with cancer [35-37].

Except for the CNV region overlapping with hsa-miRNA-4746-5p, copy status for other nine CNV-miRNAs showed predominantly a diploid status, and therefore the correlation between miRNA and mRNA expressions were restricted to cases (n=195) with diploid status (Appendix Table A.15). Ingenuity Pathway Analysis of the identified target genes regulated by hsa-miR-655, hsa-miR-134-3p, hsa-miR-4746 showed significant enrichment of several pathways (Appendix Table A.16). hsa-miR-655-3p and hsa-miR-134-3p had a common target gene, *DLD* (dihydrolipoamide dehydrogenase*)* which plays an important role in cellular biosynthesis and degradation of amino acid pathways. In addition, miRNA-134-3p targeted *CDK5* (Cyclin Dependent kinase 5)[38,39], *POLE* (DNA polymerase epsilon, catalytic subunit)[40] and *RAN* (member RAS oncogene family)[41] with potential role in cell cycle.

4.4. Discussion

GWAS approaches have identified several SNPs of low penetrance that contributed to the genetic risk of breast cancer [26,42,43]. However, the putative causal variants have not been identified for a majority of GWAS-identified loci and thus limit our understanding of the role of these variants in disease etiology. CNVs are complex genomic variants which may show an overlap with protein coding and non-coding regions. Therefore, characterizing CNVs associated with breast cancer may offer potential mechanistic insights. CNVs can influence gene expression in several ways, including gene dosage effects and *cis*/*trans* regulation. In this study, I have addressed the role of germline CNVs with embedded sncRNAs in breast cancer. Although CNV embedded sncRNAs may play a role in disease pathogenesis, a direct demonstration of expression of sncRNA genes from CNV-sncRNAs was lacking [5]. This is the first study to identify associated CNVs containing four different classes of sncRNAs including miRNAs. I identified 1812 CNVs mapping small RNA genes (38 miRNAs, 9865 piRNAs, 15 tRNAs and 71 snoRNAs) significantly associated with breast cancer risk using a case-control approach. I gained insights into the associated CNV loci by quantifying the expression of the embedded sncRNA genes in both breast tumors and adjacent normal tissues.

The sncRNAs play key roles in post-transcriptional gene regulation events, and variations in expression of sncRNAs may potentially affect their downstream targets. I identified a subset of CNV-sncRNAs that were expressed in both breast tumor and adjacent normal tissues. Since gene expressions are tissue specific, I expect only a small subset of sncRNAs to be expressed in breast tissues despite several sncRNA genes were annotated to the CNV regions. Recent studies on neurodevelopmental disorders have also identified CNVs were shown to be enriched with miRNA genes[17-21]. Several mechanisms have been proposed to explain the impact on the miRNAs based on the extent of CNV overlap with miRNA genes *e.g.,* dosage effects attributed to loss of expression depending on the extent of overlap[32]. Other key findings of the study were as follows.

(i) Among the breast cancer associated CNVs (Table 4.1), four CNVs at 14q32.31 locus with embedded miRNA genes were confirmed as copy variable regions in the 1000 Genomes Phase 3 project. These CNV-miRNAs showed tissue specific expression in this study. Literature evidence suggests that regulated targets are influenced by levels of miRNA expression which in turn are regulated by feedback mechanisms [44]. Extending this premise, I reasoned that CNV-miRNA gene can potentially modulate expression levels and therefore affect downstream targets. However, I did not observe direct correlation of copy status and expression of the embedded-miRNAs. Instead, I observed that cases with germline copy gain regions with hsa-miR-4746-5p regulated more target genes than cases with diploid copy status for the same miRNA. Pathway analysis of the regulated genes indicated their involvement in cell cycle, receptor mediated signaling, proliferation and/or apoptosis.

(ii) piRNAs are known to play a role in maintaining genomic stability by repression of transposons through gene silencing mechanisms[45] and are well studied in gonadal cells [46]. However, the role of piRNAs in somatic tissues and in cancer context are beginning to emerge. I showed piRNAs were differentially expressed between breast tumor and normal tissues and that piRNAs and their biogenesis pathway molecules (PIWI proteins) are prognostic[47]. miRNAs bind to the 3'-untranslated regions (UTR) of protein-coding genes and piRNAs also share similar mechanisms to mediate translational arrest or mRNA degradation[10]. In the Autism genetic database (AGD)[48] which catalogs autism related CNV signatures, a higher proportion of CNVs harbored piRNA genes compared to other classes of small non-coding RNA genes. A similar trend was seen in this study wherein CNVs harbored several piRNAs compared to other sncRNAs, which cannot be fully attributed to multiple copies of piRNA genes. Instead, their tendency to be enriched in CNV regions may have evolutionary significance since earlier studies have noted that there are selective constraints on the origins of piRNA[49] clusters in African populations. This is corroborated by the observed rates of insertion of transposable elements in African populations [17]. Although I mapped several piRNA genes to the breast cancer associated CNVs, only one (hsa-piR-20636) was expressed in both the breast tissues and showed trends of dosage effects. The functional significance of hsa-piR-20636 in the context of breast cancer warrants further studies.

(iii) I identified breast cancer associated CNVs (q-value $<10^{-3}$) overlapping with SNORD-115 and 116 clusters (15q11.2). Theses CNV were present only among breast cancer cases and showed a higher frequency of copy gain than copy loss. A previous study reported a CNV overlapping with the above loci at 15q11.2-13, spanning many protein and non-protein coding genes including the SNORD-115 and 116 clusters, which have been implicated in PWS[31]. In another study, wherein copy number gain in loci (chr15:24738239-24749581) upstream of the SNORD-116 cluster but in PWS loci was associated with obesity[50]. These findings suggest that copy gain or loss at these loci may confer diverse phenotypes including breast cancer. Genotyping platforms and CNV calling algorithms may contribute to the variation in the detected CNV breakpoints, therefore fine scale analysis is needed to confirm the exact breakpoints to delineate the mechanisms by which germline CNVs exerts pleotropic effects. I observed expression of eight snoRNAs from the SNORD116 cluster, and the expression of SNORD37, SNORA10 and SNORA 64 in both tumor and adjacent normal breast tissues. There are no known target RNAs regulated by SNORD116 in humans. However, SNORD 37 (target: 28S rRNA A3697) guides methylation, snoRNA 10 (target RNA: 18S rRNA U210 and 28S rRNA U4491) and SNORA 64 (target RNA: 28S rRNA U4975) directs pseudouridylation of the corresponding target rRNAs[51]. This supports the premise, that CNV embedded snoRNAs may play a role in regulation and maturation of the rRNA targets, although more direct experimental evidence is needed. Understanding the biological functions of these RNAs in the context of breast cancer susceptibility or tumorigenesis is needed.

(iv) tRNAs play a critical role in protein translation and previous studies have shown that expression of tRNAs and tRNA derived fragments were dysregulated in breast tumors[13]. Although the 1000 Genomes Phase 3 project has catalogued CNVs overlapping tRNA genes in the human genome, the role of germline CNVs with embedded tRNA genes was not studied in a disease context. Studies with model organisms demonstrated that copy number variation of tRNA genes alter the relative abundance of tRNAs, thereby altering codon usage [18,31,52,53] and potentially stalling translation leading to formation of misfolded proteins [54,55]. The current study is the first to report the association of CNV-tRNAs with breast cancer and demonstrated their expression in breast tissues. Even though I correlated tRNA expression in breast tissues with germline copy status, our study limitation is in the direct extrapolation of findings to the tRNA abundance and their effects on translational mechanisms. While the current study focused on sncRNA, long non-coding RNAs are also known to regulate genes at the post-transcriptional level and their effects warrant independent investigations.

4.5. Conclusion

In summary, I identified and validated germline CNVs associated with breast cancer. The break points identified in the discovery cohort were independently confirmed using the TCGA dataset. I was able to use the TCGA datasets since the discovery data set and the TCGA datasets were profiled for CNVs with the Affymetrix Human SNP 6.0 array platform. I acknowledge the potential limitation in the absolute calls of copy status due to differences in the control populations used as a reference. However, the unique aspect of the study was the integrative analysis of CNV calls, sncRNA and mRNA expressions in matched TCGA subjects. I showed that germline CNVs can potentially influence tissue level gene expression through their embedded sncRNA genes. Our findings provide a compelling rationale that germline CNVs have functional consequences, possibly mediated through gene dosage mechanisms.

4.6. Availability of data and material

All data generated or analysed during this study are included in this published article and its supplementary information files. The dataset is provided as electronic Supplementary Table S1 and S2 at https://doi.org/10.1038/s41598-017-14799-7

4.7. References

1 Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 136, E359-386, doi:10.1002/ijc.29210 (2015).

2 Locatelli, I., Lichtenstein, P. & Yashin, A. I. The heritability of breast cancer: a Bayesian correlated frailty model applied to Swedish twins data. Twin Res 7, 182-191, doi:10.1375/136905204323016168 (2004).

3 Michailidou, K. et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 47, doi:10.1038/ng.3242 (2015).

4 Fachal, L. & Dunning, A. M. From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. Curr Opin Genet Dev 30, 32-41, doi:10.1016/j.gde.2015.01.004 (2015).

5 Conrad, D. F. et al. Origins and functional impact of copy number variation in the human genome. Nature 464, 704-712, doi:10.1038/nature08516 (2010).

6 Long, J. et al. A common deletion in the APOBEC3 genes and breast cancer risk. J Natl Cancer Inst 105, 573-579, doi:10.1093/jnci/djt018 (2013).

7 Kumaran, M. et al. Germline copy number variations are associated with breast cancer risk and prognosis. Scientific Reports 7, 14621, doi:10.1038/s41598-017-14799-7 (2017).

8 Rezaei, M., Hashemi, M., Hashemi, S. M., Mashhadi, M. A. & Taheri, M. APOBEC3 Deletion is Associated with Breast Cancer Risk in a Sample of Southeast Iranian Population. International Journal of Molecular and Cellular Medicine 4, 103-108 (2015).

9 Xuan, D. et al. APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis 34, 2240-2243, doi:10.1093/carcin/bgt185 (2013).

10 Krishnan, P. et al. Piwi-interacting RNAs and PIWI genes as novel prognostic markers for breast cancer. *Oncotarget* **7**, 37944 (2016).

11 Krishnan, P. et al. Next generation sequencing profiling identifies miR-574-3p and miR-660-5p as potential novel prognostic markers for breast cancer. BMC Genomics 16, 735, doi:10.1186/s12864-015-1899-0 (2015).

12 Krishnan, P. et al. Profiling of Small Nucleolar RNAs by Next Generation Sequencing: Potential New Players for Breast Cancer Prognosis. PLoS One 11, e0162622, doi:10.1371/journal.pone.0162622 (2016).

13 Krishnan, P. et al. Genome-wide profiling of transfer RNAs and their role as novel prognostic markers for breast cancer. 6, 32843, doi:10.1038/srep32843

https://www.nature.com/articles/srep32843#supplementary-information (2016).

14 Duan, R., Pak, C. & Jin, P. Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. Hum Mol Genet 16, 1124-1131, doi:10.1093/hmg/ddm062 (2007).

15 Saunders, M. A., Liang, H. & Li, W.-H. Human polymorphism at microRNAs and microRNA target sites. Proceedings of the National Academy of Sciences of the United States of America 104, 3300-3305, doi:10.1073/pnas.0611347104 (2007).

16 Sun, G. et al. SNPs in human miRNA genes affect biogenesis and function. RNA 15, 1640-1651, doi:10.1261/rna.1560209 (2009).

17 Beveridge, N. J. & Cairns, M. J. MicroRNA dysregulation in schizophrenia. Neurobiol Dis 46, 263-271, doi:10.1016/j.nbd.2011.12.029 (2012).

18 Brzustowicz, L. & Bassett, A. miRNA-mediated risk for schizophrenia in 22q11.2 deletion syndrome. Frontiers in Genetics 3, doi:10.3389/fgene.2012.00291 (2012).

19 Warnica, W. et al. Copy Number Variable MicroRNAs in Schizophrenia and Their Neurodevelopmental Gene Targets. Biological Psychiatry 77, 158-166, doi:http://dx.doi.org/10.1016/j.biopsych.2014.05.011 (2015).

20 Matuszek, G. & Talebizadeh, Z. Autism genetic database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with

known noncoding RNAs and fragile sites. *BMC Medical Genetics* **10**, 102, doi:10.1186/1471-2350-10-102 (2009).

21 Persengiev, S., Kondova, I. & Bontrop, R. Insights on the functional interactions between miRNAs and copy number variations in the aging brain. Frontiers in Molecular Neuroscience 6, 32 (2013).

22 Xing, H. J. et al. Identification of microRNAs present in congenital heart disease associated copy number variants. Eur Rev Med Pharmacol Sci 17, 2114-2120 (2013).

23 Sapkota, Y. et al. Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm. PLoS One 8, e62550, doi:10.1371/journal.pone.0062550 (2013).

24 Alberta Cancer Research biobank, <http://www.acrb.ca/about-us/> (2001).

25 Shi, J. et al. Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. Int J Cancer 139, 1303-1317, doi:10.1002/ijc.30150 (2016).

26 Sehrawat, B. et al. Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human genetics 130, 529-537, doi:10.1007/s00439-011-0973-1 [doi] (2011).

27 Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42, doi:10.1093/nar/gkt1181 (2014).

28 Kersey, P. J. et al. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res 44, D574-580, doi:10.1093/nar/gkv1209 (2016).

29 (2016).

30 Karolchik, D. et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32, D493-496, doi:10.1093/nar/gkh103 (2004).

31 Sahoo, T. et al. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet 40, 719-721, doi:10.1038/ng.158 (2008).

32 Marcinkowska, M., Szymanski, M., Krzyzosiak, W. J. & Kozlowski, P. Copy number variation of microRNA genes in the human genome. BMC Genomics 12, 183, doi:10.1186/1471-2164-12-183 (2011).

33 Rose-Zerilli, M. J., Barton, S. J., Henderson, A. J., Shaheen, S. O. & Holloway, J. W. Copy-number variation genotyping of GSTT1 and GSTM1 gene deletions by real-time PCR. Clin Chem 55, 1680-1685, doi:10.1373/clinchem.2008.120105 (2009).

34 Stranger, B. E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315, doi:10.1126/science.1136678 (2007).

35 Dolle, L. et al. Nerve growth factor receptors and signaling in breast cancer. Curr Cancer Drug Targets 4, 463-470 (2004).

36 Li, J. et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science 275, 1943-1947 (1997).

37 Spiegelberg, B. D. & Hamm, H. E. Roles of G-protein-coupled receptor signaling in cancer biology and gene transcription. Curr Opin Genet Dev 17, 40-44, doi:10.1016/j.gde.2006.12.002 (2007).

38 Xiong, Y., Zhang, H. & Beach, D. D type cyclins associate with multiple protein kinases and the DNA replication and repair factor PCNA. Cell 71, 505-514 (1992).

39 Zhang, H., Xiong, Y. & Beach, D. Proliferating cell nuclear antigen and p21 are components of multiple cell cycle kinase complexes. Mol Biol Cell 4, 897-906 (1993).

40 Fuss, J. & Linn, S. Human DNA Polymerase ε Colocalizes with Proliferating Cell Nuclear Antigen and DNA Replication Late, but Not Early, in S Phase. Journal of Biological Chemistry 277, 8658-8666, doi:10.1074/jbc.M110615200 (2002).

41 Clarke, P. R. & Zhang, C. Spatial and temporal coordination of mitosis by Ran GTPase. Nat Rev Mol Cell Biol 9, 464-477, doi:10.1038/nrm2410 (2008).

42 Easton, D. F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. (2007).

43 Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics 45, 353-361, 361e351-352, doi:10.1038/ng.2563 [doi] (2013).

44 Shu, J. et al. Dose-dependent differential mRNA target selection and regulation by let-7a-7f and miR-17-92 cluster microRNAs. RNA Biol 9, 1275-1287, doi:10.4161/rna.21998 (2012).

45 Ross, R. J., Weiner, M. M. & Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. Nature 505, 353-359, doi:10.1038/nature12987 (2014).

46 Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol 12, 246-258, doi:10.1038/nrm3089 (2011).

47 Krishnan, P. et al. Piwi-interacting RNAs and PIWI genes as novel prognostic markers for breast cancer. Oncotarget 7, 37944-37956, doi:10.18632/oncotarget.9272 (2016).

48 Matuszek, G. & Talebizadeh, Z. Autism Genetic Database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. BMC Med Genet 10, 102, doi:10.1186/1471-2350-10-102 (2009).

49 Gould, D. W., Lukic, S. & Chen, K. C. Selective constraint on copy number variation in human piwi-interacting RNA Loci. PLoS One 7, e46611, doi:10.1371/journal.pone.0046611 (2012).

50 Chen, Y. et al. Copy Number Variations at the Prader–Willi Syndrome Region on Chromosome 15 and associations with Obesity in Whites. Obesity (Silver Spring, Md.) 19, 1229-1234, doi:10.1038/oby.2010.323 (2011).

51 Lestrade, L. & Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Research* **34**, D158-D162, doi:10.1093/nar/gkj002 (2006).

52 Iben, J. R. et al. Comparative whole genome sequencing reveals phenotypic tRNA gene duplication in spontaneous Schizosaccharomyces pombe La mutants. Nucleic Acids Res 39, 4728-4742, doi:10.1093/nar/gkr066 (2011).

53 Iben, J. R. & Maraia, R. J. tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon:codon wobble pair in a eukaryote. RNA 18, 1358-1372, doi:10.1261/rna.032151.111 (2012).

54 Zhou, M. et al. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature 495, 111-115, doi:10.1038/nature11833 (2013).

55 Iben, J. R. & Maraia, R. J. tRNA gene copy number variation in humans. Gene 536, 376-384, doi:10.1016/j.gene.2013.11.049 (2014).

5 Discussion

In this thesis, I investigated the genetic architecture of breast cancer. Breast cancer is a complex, multifactorial and polygenic disease. I investigated the role of common polymorphisms (SNPs and CNVs) and their contributions to the heritability in breast cancer. Several independent GWASs have collectively reported 172 variants to be associated with breast cancer accounting for about 18% of heritability[1]. Previous studies from the Damaraju laboratory were among the GWASs reported for breast cancer, wherein a multistage GWAS study design was implemented, discovery (Stage 1) and replication (Stages 2-3)[2,3] which led to the identification of the SNP rs1429142 (in chr4q31.22) associated with sporadic breast cancer risk as well as a trend of elevated risk for premenopausal breast cancer[3]. In my study, (chapter 2) I utilized an independent replication cohort (Stage 4) also based on a Caucasian population from Alberta, Canada to reproduce the findings. I replicated the association of the SNP rs1429142 with breast cancer risk (Stages 1-4 combined cases and controls from previous Stages 1-3) which is now significant after genome wide correction (OR 1.25, 4.35×10^{-8}). Further, in the combined analysis of all premenopausal cases and controls from Stages 1-4, the SNP showed genome-wide significance at P-value < 10^{-10} (OR 1.4) Also consistent with previous studies, I replicated the marginal association of the SNP with post-menopausal breast cancer risk. I also tested for associations based on luminal vs non-luminal, high vs low tumor grade, and ER positive vs ER negative and noted that the difference in risk (OR) between these subgroups were not statistically significant (P-heterogeneity>0.05). I also used external datasets for replication and validation of the association: (i) CGEMS dataset includes postmenopausal cases and controls of Caucasian ancestry. I showed that rs1429142 is not significant among postmenopausal breast cancer and at the sample sizes indicated (total n=2287). However, at a larger sample size (n=6971), this SNP showed statistically significant associations but was not genome wide significant. The effect size was modest in both CGEMs and the lab datasets (OR, 1.03 to 1.17) for post-menopausal women, in agreement with my hypothesis and consistent with the previous reports. (ii) I used a dataset of African ancestry (African Diaspora study from dbGap) to validate the association of the index SNP. Once again, rs1429142 was associated with premenopausal breast cancer risk and not with postmenopausal breast cancer risk. In summary, I was able to confirm the association of the SNP rs1429142 with breast cancer risk among Caucasian women and specific risk associated with premenopausal breast cancer among Caucasian and African populations.

I fine-mapped the chr4q31.22 locus to identify putative causal variants and sought functional relevance to breast cancer. I used several fine-mapping approaches (imputation, genotyping of the imputed SNPs, functional annotation for regulatory variants). In the fine-mapped locus, I identified 135 SNPs associated with premenopausal breast cancer risk. Based on data filtering and annotation techniques (as discussed in chapter 2), I identified SNPs (rs1366691, rs1429139, rs7667633, rs6836670 and rs17023196) at highest predicted level of functionality as enhancers. In support of this interpretation, I identified DNase I hypersensitivity peaks (indicated open chromatin state), histone methylation (H3K4me1) and acetylation (H3K9ac and H3K27ac) patterns in breast cell lines. Also, ChIP-Seq data based on MCF10-src cell line revealed the binding of FOS, STAT3 and POL2RA of the transcription factors at SNP locus rs136691, r7667633, rs7668383. The binding of transcription factors at the SNP locus was shown during the process of transformation in the of MCF10-Src cell line (exhibits increased motility, invasion, formation of foci, single cell colonies and mammospheres[4,5]) and suggests that transcription factors binding to these regions impart the cellular phenotypes. STAT3 is well known for its role as a transcriptional regulator in many cancer types, and during the process of transformation, STAT3 acts as an epigenetic switch regulating the inflammatory pathways including NFKB1 and IL6 cascade[6].

The fine-mapped variants were predicted to have enhancer functions, and they are likely to interact with the promoters of nearby gene(s) and regulate them. Interaction of enhancers and promoters are facilitated by DNA looping. The SNP locus is present within a topologically associated domain (TAD), wherein the interactions are likely to be short range and within the domain boundaries. The data from high throughput DNA conformation assays using the HMEC cell line revealed multiple short-range interactions at the SNP locus supporting the premise of TADs.

I also investigated for potential eQTLs between the fine-mapped SNPs and the neighboring genes within 1 Mb distance. I identified eQTLs and the evidence presented supported the regulation of *ENDRA* and *ARHGAP10* in heart left ventricle and lymphoblastoid tissues. Functional roles of EDNRA[7-14] and ARHGAP10[15-19] were previously described in cancer.

In summary, I fine-mapped and identified rs1366691, rs1429139, rs7667633 potential causal variants associated with premenopausal breast cancer risk. However, further experimental evidence is needed in model systems to delineate the mechanisms by which these variants regulate the targets and confer breast cancer risk.

In chapters 3 and 4, I investigated the role of germline CNVs and their contribution to breast cancer risk. The function of the CNVs vary according to the genomic locations (genic and gene desert/intergenic region) and genes they harbor (protein coding gene regions, non-coding RNA genes). In my thesis, I explored the functional consequences of the CNVs overlapping with the protein coding genes (in Chapter 3) and small-non-coding RNA genes (in chapter 4) which are key players in post transcriptional gene regulation.

CNVs overlapping protein coding genes may offer insights to the target genes and their role in breast cancer susceptibility. I utilized a case-control approach (as described in chapter 3) and identified 200 common CNVs/contiguous CNV Regions or CNVRs (>10%) overlapping protein coding genes associated with breast cancer risk[20]. Long et al. identified a common deletion polymorphism in APOBEC3 loci associated with breast cancer risk in Chinese ancestry[21]. These findings were further validated in different populations[22,23]. I replicated the association of deletion of APOBEC3 genes with breast cancer risk in Caucasian population (Alberta, Canada). I also validated the deletion of APOBEC3 genes and GSTM1 using the TaqMan assay. The majority of the CNVs identified in my study are also catalogued as common CNVs in the 1000 Genomes phase 3 project, serving as a confirmatory analysis for common CNVs. I showed CNVs associated with breast cancer risk that overlap with protein coding genes resulting in gene dosage effects. I identified nine genes whose expression correlates with germline copy status. I replicated the previously reported association of the CNVs (*ANKS1B19, OR4C11, OR4P4, UGT2B17, OR4C6, OR4S215*) from a familial breast cancer study[20]. Germline CNVs and their embedded genes are expressed in breast tissues, thus offering functional insights. CNVs as susceptibility determinants could serve the dual purpose of identifying high risk individuals, and the embedded genes and the pathways regulated can serve as potential therapeutic targets.

I investigated the prognostic potential of the breast cancer associated genes. Of the 200 CNVs/CNVRs associated with breast cancer risk, 21 CNVRs were associated with breast cancer prognosis (OS and RFS). Four CNVRs showed overlap with the genes *ZFP14, JAK1, LPA, PDGFRA* and were associated with both RFS and OS. Six CNVs overlapping the genes (*SORBS2, LCE3C, MLIP, OR2T11, MUC20, LGALS*) were specifically associated with RFS. 11 CNVRs overlapped with 12 genes (*GSTM2, RAB40B, HLA_DRB5, HLA_DRB6, EYA1, DOCK3, ANKS1B, CACNA1C, RAB11FIP3, BAGE, SGCZ, POM121c*) were specifically associated with OS[20]. This is the first study in the literature to describe the prognostic relevance for breast cancer risk associated CNVs. Given that CNVs have the potential to confer risk for both susceptibility and prognosis, therapeutics development based on these markers may help in breast cancer prevention as well as in treatments for better outcomes.

In chapter 4, I investigated the effects of the CNVs on embedded small-non-coding RNAs and their role at the post transcriptional level of gene regulatory mechanisms. Distribution of the CNVs across the genome is disproportionate and most CNVs are harbored in the non-coding genome. However, the functional significance of such CNVs in the disease context is not clear. Therefore, in my study I identified CNVs associated with breast cancer (at p-value <0.05) using the case-control approach (as described in chapter 4). Of the associated CNVs, 1812 had embedded small non-coding RNAs (38 miRNA, 9865 piRNA, 71 snoRNA and 15 tRNA) genes[24]. I also utilized an external dataset (TCGA) and validated the CNV breakpoints. Next, I interrogated the expression of the CNV embedded small-RNA (CNV-sncRNAs) genes in breast tissue. Even though several sncRNAs were harbored at the CNV regions, only a subset of the snc-RNAs showed expression in breast tissues[24]. Since sncRNAs are key regulators in post transcriptional gene regulatory events, any variation in the expression of sncRNA due to CNVs may affect downstream target genes. Similar studies have identified CNV overlap with miRNA genes enriched in neurodevelopmental disorders[25-29] using *in silico* predictions.

I demonstrated for the first time the expression of CNV embedded protein coding and small RNA genes in breast tissues, hence their functional relevance. Gene dosage effects were more pronounced for protein coding genes. I noted the copy gain region with embedded hsa-miR-4746-5p regulated more target genes (compared to diploid copy status) and these genes regulated cell cycle, receptor mediated signaling, proliferation and/or apoptosis. Similarly, I identified several piRNAs to be embedded within the associated CNVs [30] but only one piRNA (hsa-piR-20636) was expressed in breast tissue, and showed gene-dosage effects. The expression of a number of piRNAs in the breast tissue, but the expression of CNV embedded piRNAs are limited. The functional significance in the context of breast cancer needs further investigations.

I also identified snoRNAs harbored in the CNV region, the key findings include the CNV overlapping the SNORD-115 and 116 clusters (15q11.2). CNV in the same cluster is also implicated in Prader-Willi Syndrome[31] and obesity[32]. The expression analysis indicated eight snoRNAs from the SNORD116 cluster, and the expression of SNORD37, SNORA10 and SNORA 64 in the breast tissues. snoRNAs guide in the methylation and pseudouridylation of the corresponding target rRNAs[33] and play a role in regulation/maturation of the rRNAs. However, the functional consequences of these rRNAs in breast cancer is yet to be determined.

The tRNAs have a unique role in the modulation of protein translation. Based on animal models, studies have described that the relative abundance and variation of the expression of tRNA can directly affect the codon usage[26,31,34,35] and potentially stalling translation leading to formation of misfolded proteins[36,37]. The current study is the first to describe the role of CNV of tRNAs in the context of breast cancer, and I also described correlation between copy status and tRNA expression.

5.1. Study limitations and strategies to overcome

Potential limitations of the described studies are indicated below.

The lack of access to GWAS data sets in literature and in the open access databases limited the stratified analysis based on menopausal status in Caucasian women based on external datasets. However, this limitation does not hamper the generalizability of my findings. I was able to access the African diaspora which helped to confirm the major findings in this study, *i.e.*, premenopausal risk conferred is by rs1429142.

(i) The sample size of this study (overall cases and controls ~9000), which is moderate compared to consortium-based studies (~40,000 cases and controls each)[38] is a limitation. However, with minor allele frequency (at ~18%) and OR at ~1.25 (overall breast cancer risk) and OR at ~1.40 (premenopausal breast cancer risk), the estimated power is ~0.99 under additive or multiplicative models of risk, and a population disease prevalence of 1/8[39]. Therefore, higher sample size would have potentially added strength to the association (p-value) but would not have influenced the estimated risk (OR) for premenopausal or overall breast cancer risk.

(ii) Higher sample size is needed to identify putative causal variants at low minor allele frequencies. As such the current study may have underestimated the number of causal/regulatory variants. This limitation can be overcome by consortia led studies wherein sample sizes upwards of 100,000 each of cases and controls are used[1]. Collaborations with breast cancer consortia are needed to address this gap.

(iii) Experimental evidences are needed in model systems to delineate the mechanisms by which these variants regulate the targets and confer breast cancer risk.

(iv) I identified several CNVs associated with breast cancer risk and prognosis. (a) The study lacked a replication Stage (as in traditional multistage GWAS). Publicly available data sets were limited, and where available, only data from cases could be utilized (TCGA). Since, no matching germline controls were available within the TCGA, I could not attempt an independent case control analysis for CNVs associated with breast cancer susceptibility. To overcome this limitation, I ascertained the CNV calls by comparing with the 1000 Genome Project data from controls. (b) I validated representative CNVs on TaqMan assays (APOBE3C, GSTM1, known breast cancer susceptibility alleles, hence my study met the needed power to detect associations for common CNVs >10%). (c) Further, the Damaraju laboratory has data on matched samples (gene expression data and CNV data generated on array platforms for the same individual breast cancer cases). Mapping for the embedded protein coding genes within the CNVs and showing gene-dosage effects is a unique strength of my study. (d) The break points observed for cases were from the Partek algorithm implemented in this study, and these were compared with the break points from the TCGA data. Both TCGA and the data was generated on the same Affymetrix array 6 platforms, and CNVs were called on both datasets using Partek bioinformatics platform to maintain consistency. Again, the embedded small non-coding RNA gene expressions (TCGA), and gene-dosage effects were assessed, lending confidence to the study findings.

5.2. References

1 Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92-94, doi:10.1038/nature24284 (2017).

2 Sehrawat, B. et al. Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human genetics 130, 529-537, doi:10.1007/s00439-011-0973-1 [doi] (2011).

3 Sapkota, Y. et al. Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm. PLoS One 8, e62550, doi:10.1371/journal.pone.0062550 (2013).

4 Aziz, N., Cherwinski, H. & McMahon, M. Complementation of defective colony-stimulating factor 1 receptor signaling and mitogenesis by Raf and v-Src. Mol Cell Biol 19, 1101-1115 (1999).

5 Soule, H. D. et al. Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. Cancer Res 50, 6075-6086 (1990).

6 Fleming, J. D. et al. STAT3 acts through pre-existing nucleosome-depleted regions bound by FOS during an epigenetic switch linking inflammation to cancer. Epigenetics & Chromatin 8, 7, doi:10.1186/1756-8935-8-7 (2015).

7 Grant, K. et al. Mechanisms of endothelin 1-stimulated proliferation in colorectal cancer cell lines. Br J Surg 94, 106-112, doi:10.1002/bjs.5536 (2007).

8 Zhang, W. M., Zhou, J. & Ye, Q. J. Endothelin-1 enhances proliferation of lung cancer cells by increasing intracellular free Ca2+. Life Sci 82, 764-771, doi:10.1016/j.lfs.2008.01.008 (2008).

9 Wulfing, P. et al. Endothelin-1-, endothelin-A-, and endothelin-B-receptor expression is correlated with vascular endothelial growth factor expression and angiogenesis in breast cancer. Clin Cancer Res 10, 2393-2400 (2004).

10 Rosano, L. et al. Beta-arrestin links endothelin A receptor to beta-catenin signaling to induce ovarian cancer cell invasion and metastasis. Proc Natl Acad Sci U S A 106, 2806-2811, doi:10.1073/pnas.0807158106 (2009).

11 Wilson, J. L., Burchell, J. & Grimshaw, M. J. Endothelins induce CCR7 expression by breast tumor cells via endothelin receptor A and hypoxia-inducible factor-1. Cancer Res 66, 11802-11807, doi:10.1158/0008-5472.CAN-06-1222 (2006).

12 Del Bufalo, D. et al. Endothelin-1 protects ovarian carcinoma cells against paclitaxel-induced apoptosis: requirement for Akt activation. Mol Pharmacol 61, 524-532 (2002).

13 Nelson, J. B., Udan, M. S., Guruli, G. & Pflug, B. R. Endothelin-1 inhibits apoptosis in prostate cancer. Neoplasia 7, 631-637 (2005).

14 Wulfing, P. et al. Expression of endothelin-1, endothelin-A, and endothelin-B receptor in human breast cancer and correlation with long-term follow-up. Clin Cancer Res 9, 4125-4131 (2003).

15 Jaffe, A. B. & Hall, A. Rho GTPases: biochemistry and biology. Annu Rev Cell Dev Biol 21, 247-269, doi:10.1146/annurev.cellbio.21.020604.150721 (2005).

16 Azzato, E. M. et al. A genome-wide association study of prognosis in breast cancer. Cancer Epidemiol Biomarkers Prev 19, 1140-1143, doi:10.1158/1055-9965.EPI-10-0085 (2010).

17 Wong, N. C. et al. Stability of gene expression and epigenetic profiles highlights the utility of patient-derived paediatric acute lymphoblastic leukaemia xenografts for investigating molecular mechanisms of drug resistance. BMC Genomics 15, 416, doi:10.1186/1471-2164-15-416 (2014).

18 Luo, N. et al. ARHGAP10, downregulated in ovarian cancer, suppresses tumorigenicity of ovarian cancer cells. Cell Death Dis 7, e2157, doi:10.1038/cddis.2015.401 (2016).

19 Teng, J. P. et al. The roles of ARHGAP10 in the proliferation, migration and invasion of lung cancer cells. Oncol Lett 14, 4613-4618, doi:10.3892/ol.2017.6729 (2017).

20 Kumaran, M. et al. Germline copy number variations are associated with breast cancer risk and prognosis. Scientific Reports 7, 14621, doi:10.1038/s41598-017-14799-7 (2017).

21 Long, J. et al. A common deletion in the APOBEC3 genes and breast cancer risk. J Natl Cancer Inst 105, 573-579, doi:10.1093/jnci/djt018 (2013).

22 Rezaei, M., Hashemi, M., Hashemi, S. M., Mashhadi, M. A. & Taheri, M. APOBEC3 Deletion is Associated with Breast Cancer Risk in a Sample of Southeast Iranian Population. International Journal of Molecular and Cellular Medicine 4, 103-108 (2015).

23 Xuan, D. et al. APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis 34, 2240-2243, doi:10.1093/carcin/bgt185 (2013).

24 Kumaran, M. et al. Breast cancer associated germline structural variants harboring small noncoding RNAs impact post-transcriptional gene regulation. Sci Rep 8, 7529, doi:10.1038/s41598-018-25801-1 (2018).

25 Beveridge, N. J. & Cairns, M. J. MicroRNA dysregulation in schizophrenia. Neurobiol Dis 46, 263-271, doi:10.1016/j.nbd.2011.12.029 (2012).

26 Brzustowicz, L. & Bassett, A. miRNA-mediated risk for schizophrenia in 22q11.2 deletion syndrome. Frontiers in Genetics 3, doi:10.3389/fgene.2012.00291 (2012).

27 Matuszek, G. & Talebizadeh, Z. Autism genetic database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. BMC Medical Genetics 10, 102, doi:10.1186/1471-2350-10-102 (2009).

28 Persengiev, S., Kondova, I. & Bontrop, R. Insights on the functional interactions between miRNAs and copy number variations in the aging brain. Frontiers in Molecular Neuroscience 6, 32 (2013).

29 Warnica, W. et al. Copy Number Variable MicroRNAs in Schizophrenia and Their Neurodevelopmental Gene Targets. Biological Psychiatry 77, 158-166, doi:http://dx.doi.org/10.1016/j.biopsych.2014.05.011 (2015).

30 Gould, D. W., Lukic, S. & Chen, K. C. Selective constraint on copy number variation in human piwi-interacting RNA Loci. PLoS One 7, e46611, doi:10.1371/journal.pone.0046611 (2012).

31 Sahoo, T. et al. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet 40, 719-721, doi:10.1038/ng.158 (2008).

32 Chen, Y. et al. Copy Number Variations at the Prader–Willi Syndrome Region on Chromosome 15 and associations with Obesity in Whites. Obesity (Silver Spring, Md.) 19, 1229-1234, doi:10.1038/oby.2010.323 (2011).

33 Lestrade, L. & Weber, M. J. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Research 34, D158-D162, doi:10.1093/nar/gkj002 (2006).

34 Iben, J. R. et al. Comparative whole genome sequencing reveals phenotypic tRNA gene duplication in spontaneous Schizosaccharomyces pombe La mutants. Nucleic Acids Res 39, 4728-4742, doi:10.1093/nar/gkr066 (2011).

35 Iben, J. R. & Maraia, R. J. tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon:codon wobble pair in a eukaryote. RNA 18, 1358-1372, doi:10.1261/rna.032151.111 (2012).

36 Zhou, M. et al. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature 495, 111-115, doi:10.1038/nature11833 (2013).

37 Iben, J. R. & Maraia, R. J. tRNA gene copy number variation in humans. Gene 536, 376-384, doi:10.1016/j.gene.2013.11.049 (2014).

38 Michailidou, K. et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 47, doi:10.1038/ng.3242 (2015).

39 H, Z. J. gap: Genetic Analysis Package. R package version 1.1-17. (2017).

6 Future directions and Conclusions

6.1. Future directions

In my thesis, I identified genetic variants associated with breast cancer susceptibility and prognosis. Previous GWASs from the Damaraju laboratory reported a novel SNP rs1429142[1,2] associated with breast cancer. This is an addition to the known high, moderate and low penetrant variants reported thus far. I also identified CNVs as potential breast cancer susceptibility determinants, an emerging theme in breast cancer literature in accounting for "missing heritability".

In chapter 2, I replicated and validated the previously identified GWAS locus. I further fine-mapped the locus and identified putative causal SNP variants. I also described potential functions based on available online annotation resources. Fine-mapped variants function as potential enhancer regions, likely interact with the promoters of the target genes by DNA looping. Further investigations are needed to elucidate the mechanisms by which the causal variants regulate the target genes and confer the breast cancer risk. Future investigations should include demonstrating binding of the transcription factors (STAT, FOS) to the SNP sites and electrophoretic mobility shift assays to identify allele-specific binding of these factors[3-5]. Currently, the datasets available through ENCODE are based on MCF-10, HMEC, vHMEC or breast myoepithelial cell lines. However, choosing the appropriate cell line or model system closely depicting premenopausal breast cancer would be advantageous. Binding of the TFs could be assayed at different conditions, competitive binding[3,4,6] with other transcription factors could be tested. Future investigations should also confirm the physical interactions between the enhancer and promoter based on high throughput DNA looping experiments in different cell lines. This will help identify novel target genes that are regulated by the interaction of the enhancer and promoter, which in turn may provide new insights into biological pathways in conferring the breast cancer risk among the premenopausal women.

In chapters 3 and 4, I described several CNVs to be associated with breast cancer. It would be valuable to replicate at least a subset of the CNVs in large sample sizes similar to GWAS stages, a study design adopted by Long et al[7]in identifying a CNV in APOBE3C locus as a breast cancer susceptibility determinant. To enable large scale replication of candidate CNVs described in this thesis, the currently available CNV genotyping platforms are not adequate or cost-effective. High throughput and multiplex platforms are needed to advance these studies to the level of SNP studies.

CNVs have the potential to be associated with risk as well as prognosis. Compared to SNPs, CNVs are amenable for interpretation of the putative functions, including embedded genes and gene-dosage effects. There is the potential to adopt germline CNVs as therapeutic targets and genetic biomarkers. Utility of CNV based biomarkers for screening and diagnosis of several inherited genetic conditions or developmental disorders have demonstrated the feasibility of such approaches.

6.2. Conclusions

Overall, I investigated the genetic variants that play a role in genetic architecture of breast cancer. SNP based GWAS approaches, as well as fine-mapping of GWAS variants, were widely adopted to identify novel variants associated with breast cancer. I fine-mapped the locus associated with premenopausal breast cancer risk based on bioinformatics and statistical approaches. I report several variants in the locus that are highly correlated. I adopted different strategies (statistical and functional annotations) to narrow down the set of putative causal variants. I inferred the functional significance of these variants based on a number of experimental datasets (*e.g.*, ENCODE[8], Roadmap epigenomics project[9]). I identified potential target genes that are regulated by these variants. My study has laid the foundations for future studies to identify mechanistic insights on how the target genes are regulated and their effects on the phenotype. Despite exhaustive searches based on SNP GWAS approaches, there are variants yet to be discovered to account for the missing heritability. I showed that CNVs are candidates to explore and to identify the missing heritability. I comprehensively investigated the role of germline CNVs in conferring breast cancer risk and prognosis by adopting CNV GWAS study design (described in detail in chapters 3 and 4). I investigated the effects of the CNVs overlapping with the protein-coding genes and the small-non-coding RNA genes. A correlation between the copy status and gene expression is demonstrated to explore the possible biological significances. I identified several candidate CNVs overlapping with both protein-coding and small-non-coding RNA genes for future replication studies and the potential to explain a proportion of the missing heritability.

6.3. References

1 Sapkota, Y. et al. Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm. PLoS One 8, e62550, doi:10.1371/journal.pone.0062550 (2013).

2 Sehrawat, B. et al. Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human genetics 130, 529-537, doi:10.1007/s00439-011-0973-1 [doi] (2011).

3 French, JulietÂ D. et al. Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. American Journal of Human Genetics 92, 489-503, doi:10.1016/j.ajhg.2013.01.002 (2013).

4 Meyer, KerstinÂ B. et al. Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1. American Journal of Human Genetics 93, 1046-1060, doi:10.1016/j.ajhg.2013.10.026 (2013).

5 Udler, M. S. et al. Fine scale mapping of the breast cancer 16q12 locus. Human molecular genetics 19, 2507-2515, doi:10.1093/hmg/ddq122 (2010).

6 Glubb, Dylan M. et al. Fine-Scale Mapping of the 5q11.2 Breast Cancer Locus Reveals at Least Three Independent Risk Variants Regulating MAP3K1. The American Journal of Human Genetics 96, 5-20, doi:10.1016/j.ajhg.2014.11.009 (2015).

7 Long, J. et al. A common deletion in the APOBEC3 genes and breast cancer risk. J Natl Cancer Inst 105, 573-579, doi:10.1093/jnci/djt018 (2013).

8 Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Research 22, 1790-1797 (2012).

9 Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 28, 1045-1048, doi:10.1038/nbt1010-1045 (2010).

Bibiliography

Abeliovich, D., Kaduri, L., Lerer, I., Weinberg, N., Amir, G., Sagi, M., . . . Peretz, T. (1997). The founder mutations 185delAG and 5382insC in BRCA1 and 6174delT in BRCA2 appear in 60% of ovarian cancer and 30% of early-onset breast cancer patients among Ashkenazi women. Am J Hum Genet, 60(3), 505-514.

Adank, M. A., van Mil, S. E., Gille, J. J., Waisfisz, Q., & Meijers-Heijboer, H. (2011). PALB2 analysis in BRCA2-like families. Breast Cancer Res Treat, 127(2), 357-362. doi:10.1007/s10549-010-1001-1

Al-Sukhni, W., Joe, S., Lionel, A. C., Zwingerman, N., Zogopoulos, G., Marshall, C. R., . Gallinger, S. (2012). Identification of germline genomic copy number variation in familial pancreatic cancer. Hum Genet, 131(9), 1481-1494. doi:10.1007/s00439-012-1183-1

Alberta Cancer Research biobank. (2001). Retrieved from http://www.acrb.ca/about-us/

Allen-Brady, K., Cannon-Albright, L. A., Neuhausen, S. L., & Camp, N. J. (2006). A role for XRCC4 in age at diagnosis and breast cancer risk. Cancer Epidemiol Biomarkers Prev, 15(7), 1306-1310. doi:10.1158/1055-9965.EPI-05-0959

Amos, C. I., Dennis, J., Wang, Z., Byun, J., Schumacher, F. R., Gayther, S. A., . . . Easton, D. F. (2017). The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev, 26(1), 126-135. doi:10.1158/1055-9965.EPI-16-0106

Andersen, C. L., Lamy, P., Thorsen, K., Kjeldsen, E., Wikman, F., Villesen, P., . Orntoft,

T. F. (2011). Frequent genomic loss at chr16p13.2 is associated with poor prognosis in colorectal cancer. Int J Cancer, 129(8), 1848-1858. doi:10.1002/ijc.25841

Antoniou, A. C., Pharoah, P. D., McMullan, G., Day, N. E., Stratton, M. R., Peto, J., . Easton, D. F. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br J Cancer, 86(1), 76-83. doi:10.1038/sj.bjc.6600008

Antoniou, A. C., Wang, X., Fredericksen, Z. S., McGuffog, L., Tarrell, R., Sinilnikova, O. M., Couch, F. J. (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(10), 885-892. doi:10.1038/ng.669

Armengol, L., Villatoro, S., González, J. R., Pantano, L., García-Aragonés, M., Rabionet, R., .Estivill, X. (2009). Identification of Copy Number Variants Defining Genomic Differences among Major Human Groups. PloS one, 4(9), e7230. doi:10.1371/journal.pone.0007230

Aziz, N., Cherwinski, H., & McMahon, M. (1999). Complementation of defective colony-stimulating factor 1 receptor signaling and mitogenesis by Raf and v-Src. Mol Cell Biol, 19(2), 1101-1115.

Azzato, E. M., Pharoah, P. D., Harrington, P., Easton, D. F., Greenberg, D., Caporaso, N. E., Kraft, P. (2010). A genome-wide association study of prognosis in breast cancer. Cancer Epidemiol Biomarkers Prev, 19(4), 1140-1143. doi:10.1158/1055-9965.EPI-10-0085

Azzato, E. M., Tyrer, J., Fasching, P. A., Beckmann, M. W., Ekici, A. B., Schulz-

Wendtland, R., . . . Pharoah, P. D. P. (2010). Association Between a Germline OCA2 Polymorphism at Chromosome 15q13.1 and Estrogen Receptor–Negative Breast Cancer Survival. Journal of the National Cancer Institute, 102(9), 650-662. doi:10.1093/jnci/djq057

Ball, M. P., Li, J. B., Gao, Y., Lee, J. H., LeProust, E. M., Park, I. H., . . . Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol, 27(4), 361-368. doi:10.1038/nbt.1533

Beckmann, J. S., Estivill, X., & Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet, 8(8), 639-646. doi:10.1038/nrg2149

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol, 28(10), 1045-1048. doi:10.1038/nbt1010-1045

Beveridge, N. J., & Cairns, M. J. (2012). MicroRNA dysregulation in schizophrenia. Neurobiol Dis, 46(2), 263-271. doi:10.1016/j.nbd.2011.12.029

Bewick, M. A., Conlon, M. S., & Lafrenie, R. M. (2006). Polymorphisms in XRCC1, XRCC3, and CCND1 and survival after treatment for metastatic breast cancer. J Clin Oncol, 24(36), 5645-5651. doi:10.1200/JCO.2006.05.9923

Bharat, A., Aft, R. L., Gao, F., & Margenthaler, J. A. (2009). Patient and tumor characteristics associated with increased mortality in young women (< or =40 years) with breast cancer. J Surg Oncol, 100(3), 248-251. doi:10.1002/jso.21268

Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., & Gunderson, K. L. (2009). Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics, 1(1), 177-200. doi:10.2217/epi.09.14

Birch, J. M., Heighway, J., Teare, M. D., Kelsey, A. M., Hartley, A. L., Tricker, K. J., . . . Santibanez-Koref, M. F. (1994). Linkage studies in a Li-Fraumeni family with increased expression of p53 protein but no germline mutation in p53. Br J Cancer, 70(6), 1176-1181.

Bojesen, S. E. A., Pooley, K. A. A., Johnatty, S. E. A., Beesley, J. A., Michailidou, K. A., Tyrer, J. P. A., . . . Umeå universitet, M. f. I. f. s. O. (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nature genetics, 371.

Botstein, D., & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet, 33 Suppl, 228-237. doi:10.1038/ng1090

Bowman, T., Garcia, R., Turkson, J., & Jove, R. (2000). STATs in oncogenesis. Oncogene, 19(21), 2474-2488. doi:10.1038/sj.onc.1203527

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., . . . Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. Genome research, 22(9), 1790-1797.

Brea-Fernandez, A. J., Fernandez-Rozadilla, C., Alvarez-Barona, M., Azuara, D., Ginesta, M. M., Clofent, J., Ruiz-Ponte, C. (2016). Candidate predisposing germline copy number variants in early onset colorectal cancer patients. Clin Transl Oncol. doi:10.1007/s12094-016-1576-z

Breast Cancer Association, C. (2006). Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. J Natl Cancer Inst, 98(19), 1382-1396. doi:10.1093/jnci/djj374

Bromberg, J. F., Wrzeszczynska, M. H., Devgan, G., Zhao, Y., Pestell, R. G., Albanese, C., & Darnell, J. E., Jr. <em>Stat3</em> as an Oncogene. Cell, 98(3), 295-303. doi:10.1016/S0092-8674(00)81959-5

Brzustowicz, L., & Bassett, A. (2012). miRNA-mediated risk for schizophrenia in 22q11.2 deletion syndrome. Frontiers in Genetics, 3(291). doi:10.3389/fgene.2012.00291

Butler, M. W., Hackett, N. R., Salit, J., Strulovici-Barel, Y., Omberg, L., Mezey, J., & Crystal, R. G. (2011). Glutathione S-transferase copy number variation alters lung gene expression. Eur Respir J, 38(1), 15-28. doi:10.1183/09031936.00029210

Canadian Cancer Society. (2017). Breast Cancer Statistics.

Cantsilieris, S., & White, S. J. (2013). Correlating multiallelic copy number polymorphisms with disease susceptibility. Hum Mutat, 34(1), 1-13. doi:10.1002/humu.22172

Carvalho, I., Milanezi, F., Martins, A., Reis, R. M., & Schmitt, F. (2005). Overexpression of platelet-derived growth factor receptor α in breast cancer is associated with tumour progression. Breast Cancer Research, 7(5), R788. doi:10.1186/bcr1304

Charrier, J., Maugard, C. M., Mevel, B. L., & Bignon, Y. J. (1999). Allelotype influence at glutathione S-transferase M1 locus on breast cancer susceptibility. Br J Cancer, 79(2), 346-353. doi:10.1038/sj.bjc.6690055

Chen, F., Chen, G. K., Millikan, R. C., John, E. M., Ambrosone, C. B., Bernstein, L., . . . Haiman, C. A. (2011). Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. Human molecular genetics, 20(22), 4491-4503. doi:10.1093/hmg/ddr367. (PMID: 21852243)

Chen, F., Chen, G. K., Stram, D. O., Millikan, R. C., Ambrosone, C. B., John, E. M., . . . Haiman, C. A. (2013). A genome-wide association study of breast cancer in women of African ancestry. Human genetics, 132(1), 39-48. doi:10.1007/s00439-012-1214-y [doi]

Chen, Y., Liu, Y.-J., Pei, Y.-F., Yang, T.-L., Deng, F.-Y., Liu, X.-G., . . . Deng, H.-W. (2011). Copy Number Variations at the Prader–Willi Syndrome Region on Chromosome 15 and associations with Obesity in Whites. Obesity (Silver Spring, Md.), 19(6), 1229-1234. doi:10.1038/oby.2010.323

Clarke, P. R., & Zhang, C. (2008). Spatial and temporal coordination of mitosis by Ran GTPase. Nat Rev Mol Cell Biol, 9(6), 464-477. doi:10.1038/nrm2410

Conrad, D. F., & Hurles, M. E. (2007). The population genetics of structural variation. Nat Genet, 39(7 Suppl), S30-36. doi:10.1038/ng2042

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., . . . Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. Nature, 464(7289), 704-712. doi:10.1038/nature08516

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57-74. doi:10.1038/nature11247

Consortium, G. T. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet, 45(6), 580-585. doi:10.1038/ng.2653

Couch, F. J., Kuchenbaecker, K. B., Michailidou, K., Mendoza-Fandino, G. A., Nord, S., Lilyquist, J., Antoniou, A. C. (2016). Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. Nat Commun, 7, 11375. doi:10.1038/ncomms11375

Couch, F. J., Wang, X., McGuffog, L., Lee, A., Olswold, C., Kuchenbaecker, K. B., . . . Cimba. (2013). Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS genetics, 9(3), e1003212. doi:10.1371/journal.pgen.1003212 [doi]

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., . . . Giannoulatou, E. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature, 464. doi:10.1038/nature08979

Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., . . . Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403), 346-352. doi:10.1038/nature10983

Dabbs, D. J., Klein, M. E., Mohsin, S. K., Tubbs, R. R., Shuai, Y., & Bhargava, R. (2011). High false-negative rate of HER2 quantitative reverse transcription polymerase chain reaction of the Oncotype DX test: an independent quality assurance study. J Clin

Oncol, 29(32), 4279-4285. doi:10.1200/JCO.2011.34.7963

Darabi, H., Beesley, J., Droit, A., Kar, S., Nord, S., Moradi Marjaneh, M., Dunning, A. M. (2016). Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). Scientific Reports, 6, 32512. doi:10.1038/srep32512

de Cid, R., Riveira-Munoz, E., Zeeuwen, P. L. J. M., Robarge, J., Liao, W., Dannhauser, E. N., . . . Estivill, X. (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet, 41(2), 211-215. doi:http://www.nature.com/ng/journal/v41/n2/suppinfo/ng.313_S1.html

de Vries, B. B., Pfundt, R., Leisink, M., Koolen, D. A., Vissers, L. E., Janssen, I. M., Veltman, J. A. (2005). Diagnostic genome profiling in mental retardation. Am J Hum Genet, 77(4), 606-616. doi:10.1086/491719

Dechow, T. N., Pedranzini, L., Leitch, A., Leslie, K., Gerald, W. L., Linkov, I., & Bromberg, J. F. (2004). Requirement of matrix metalloproteinase-9 for the transformation of human mammary epithelial cells by Stat3-C. Proc Natl Acad Sci U S A, 101(29), 10602-10607. doi:10.1073/pnas.0404100101

Dekker, J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. Nat Methods, 3(1), 17-21. doi:10.1038/nmeth823

Del Bufalo, D., Di Castro, V., Biroccio, A., Varmi, M., Salani, D., Rosano, L., Bagnato, A. (2002). Endothelin-1 protects ovarian carcinoma cells against paclitaxel-induced apoptosis: requirement for Akt activation. Mol Pharmacol, 61(3), 524-532.

Delaneau, O., Marchini, J., & Zagury, J. F. (2011). A linear complexity phasing method for thousands of genomes. Nat Methods, 9(2), 179-181. doi:10.1038/nmeth.1785

Demichelis, F., Setlur, S. R., Banerjee, S., Chakravarty, D., Chen, J. Y., Chen, C. X., . . . Rubin, M. A. (2012). Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc Natl Acad Sci U S A, 109(17), 6686-6691. doi:10.1073/pnas.1117405109

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485(7398), 376-380. doi:10.1038/nature11082

Dolle, L., Adriaenssens, E., El Yazidi-Belkoura, I., Le Bourhis, X., Nurcombe, V., & Hondermarck, H. (2004). Nerve growth factor receptors and signaling in breast cancer. Curr Cancer Drug Targets, 4(6), 463-470.

Dostie, J., & Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc, 2(4), 988-1002. doi:10.1038/nprot.2007.116

Duan, R., Pak, C., & Jin, P. (2007). Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. Hum Mol Genet, 16(9), 1124-1131. doi:10.1093/hmg/ddm062

Dunning, A. M., Durocher, F., Healey, C. S., Teare, M. D., McBride, S. E., Carlomagno, F., Ponder, B. A. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet, 67(6), 1544-1554. doi:10.1086/316906

Easton, D. F., Bishop, D. T., Ford, D., & Crockford, G. P. (1993). Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. Am J Hum Genet, 52(4), 678-701.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Ponder, B. A. J. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci.

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet, 93(5), 779-797. doi:10.1016/j.ajhg.2013.10.012

Elgazzar, S., Zembutsu, H., Takahashi, A., Kubo, M., Aki, F., Hirata, K., Nakamura, Y. (2012). A genome-wide association study identifies a genetic variant in the SIAH2 locus associated with hormonal receptor-positive breast cancer in Japanese. Journal of human genetics, 57(12), 766-771. doi:10.1038/jhg.2012.108 [doi]

Fachal, L., & Dunning, A. M. (2015). From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. Curr Opin Genet Dev, 30, 32-41. doi:10.1016/j.gde.2015.01.004

Fanale, D., Iovanna, J. L., Calvo, E. L., Berthezene, P., Belleau, P., Dagorn, J. C., Russo, A. (2013). Analysis of germline gene copy number variants of patients with sporadic pancreatic adenocarcinoma reveals specific variations. Oncology, 85(5), 306-311. doi:10.1159/000354737

Fanale, D., Iovanna, J. L., Calvo, E. L., Berthezene, P., Belleau, P., Dagorn, J. C., . . .

Russo, A. (2014). Germline copy number variation in the YTHDC2 gene: does it have a role in finding a novel potential molecular target involved in pancreatic adenocarcinoma susceptibility? Expert Opin Ther Targets, 18(8), 841-850. doi:10.1517/14728222.2014.920324

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., . . . Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer, 136(5), E359-386. doi:10.1002/ijc.29210

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nat Rev Genet, 7(2), 85-97. doi:10.1038/nrg1767

Fidalgo, F., Rodrigues, T. C., Silva, A. G., Facure, L., de Sa, B. C., Duprat, J. P., . Krepischi, A. C. (2016). Role of rare germline copy number variation in melanoma-prone patients. Future Oncol, 12(11), 1345-1357. doi:10.2217/fon.16.22

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet, 84. doi:10.1016/j.ajhg.2009.03.010

Fleming, J. D., Giresi, P. G., Lindahl-Allen, M., Krall, E. B., Lieb, J. D., & Struhl, K. (2015). STAT3 acts through pre-existing nucleosome-depleted regions bound by FOS during an epigenetic switch linking inflammation to cancer. Epigenetics & Chromatin, 8(1), 7. doi:10.1186/1756-8935-8-7

Foulkes, W. D., Stefansson, I. M., Chappuis, P. O., Begin, L. R., Goffin, J. R., Wong, N., Akslen, L. A. (2003). Germline BRCA1 mutations and a basal epithelial phenotype in breast cancer. J Natl Cancer Inst, 95(19), 1482-1485.

Frank, D. A. STAT3 as a central mediator of neoplastic cellular transformation. Cancer Letters, 251(2), 199-210. doi:10.1016/j.canlet.2006.10.017

French, JulietÂ D., Ghoussaini, M., Edwards, StaceyÂ L., Meyer, KerstinÂ B., Michailidou, K., Ahmed, S., Dunning, AlisonÂ M. (2013). Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers. American Journal of Human Genetics, 92(4), 489-503. doi:10.1016/j.ajhg.2013.01.002

Fridley, B. L., Chalise, P., Tsai, Y. Y., Sun, Z., Vierkant, R. A., Larson, M. C., . . . Goode, E. L. (2012). Germline copy number variation and ovarian cancer survival. Front Genet, 3, 142. doi:10.3389/fgene.2012.00142

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., . . . Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A, 89(5), 1827-1831.

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462(7269), 58-64. doi:10.1038/nature08497

Fuss, J., & Linn, S. (2002). Human DNA Polymerase ε Colocalizes with Proliferating Cell Nuclear Antigen and DNA Replication Late, but Not Early, in S Phase. Journal of Biological Chemistry, 277(10), 8658-8666. doi:10.1074/jbc.M110615200

Gamazon, E. R., Nicolae, D. L., & Cox, N. J. (2011). A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. PLoS Genet, 7(2), e1001292. doi:10.1371/journal.pgen.1001292

Garber, J. E., & Offit, K. (2005). Hereditary cancer predisposition syndromes. J Clin Oncol, 23(2), 276-292. doi:10.1200/JCO.2005.10.042

Garcia-Closas, M., Couch, F. J., Lindstrom, S., Michailidou, K., Schmidt, M. K., Brook, M. N., . . . Kraft, P. (2013). Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet, 45(4), 392-398, 398e391-392. doi:10.1038/ng.2561

Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. (2010). Nature, 464(7289), 713-720. doi:http://www.nature.com/nature/journal/v464/n7289/suppinfo/nature08979_S1.html

Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature, 491(7422), 56-65. doi:10.1038/nature11632

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Abecasis, G. R. (2015). A global reference for human genetic variation. Nature, 526(7571), 68-74. doi:10.1038/nature15393

Geyer, F. C., de Biase, D., Lambros, M. B. K., Ragazzi, M., Lopez-Garcia, M. A., Natrajan, R., Tallini, G. (2012). Genomic profiling of mitochondrion-rich breast carcinoma: chromosomal changes may be relevant for mitochondria accumulation and tumour biology. Breast cancer research and treatment, 132(1), 15-28. doi:10.1007/s10549-011-1504-4

Ghoncheh, M., Pournamdar, Z., & Salehiniya, H. (2016). Incidence and Mortality and Epidemiology of Breast Cancer in the World. Asian Pac J Cancer Prev, 17(S3), 43-46.

Ghoussaini, M., Edwards, S. L., Michailidou, K., Nord, S., Cowper-SalÂ·lari, R., Desai, K., . Australian Ovarian Cancer, M. G. (2014). Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. Nat Commun, 4.

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., & Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res, 17(6), 877-885. doi:10.1101/gr.5533506

Glubb, Dylan M., Maranian, Mel J., Michailidou, K., Pooley, Karen A., Meyer, Kerstin B., Kar, S., French, Juliet D. (2015). Fine-Scale Mapping of the 5q11.2 Breast Cancer Locus Reveals at Least Three Independent Risk Variants Regulating MAP3K1. The American Journal of Human Genetics, 96(1), 5-20. doi:10.1016/j.ajhg.2014.11.009

Gold, B., Kirchhoff, T., Stefanov, S., Lautenberger, J., Viale, A., Garber, J., . . . Offit, K. (2008). Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proceedings of the National Academy of Sciences of the United States of America, 105(11), 4340-4345. doi:10.1073/pnas.0800441105 [doi]

Gong, J., Mei, S., Liu, C., Xiang, Y., Ye, Y., Zhang, Z., . . . Han, L. (2018). PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res, 46(D1), D971-D976. doi:10.1093/nar/gkx861

Gonzalez, K. D., Noltner, K. A., Buzin, C. H., Gu, D., Wen-Fong, C. Y., Nguyen, V. Q., Weitzel, J. N. (2009). Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. J Clin Oncol, 27(8), 1250-1256. doi:10.1200/JCO.2008.16.6959

Goodier, J. L., & Kazazian, H. H., Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell, 135(1), 23-35. doi:10.1016/j.cell.2008.09.022

Gould, D. W., Lukic, S., & Chen, K. C. (2012). Selective constraint on copy number variation in human piwi-interacting RNA Loci. PloS one, 7(10), e46611. doi:10.1371/journal.pone.0046611

Grambsch, T. M. T. a. P. M. (2000). Modeling Survival Data: Extending the Cox Model. Springer.

Grant, K., Knowles, J., Dawas, K., Burnstock, G., Taylor, I., & Loizidou, M. (2007). Mechanisms of endothelin 1-stimulated proliferation in colorectal cancer cell lines. Br J Surg, 94(1), 106-112. doi:10.1002/bjs.5536

Guo, X., Long, J., Zeng, C., Michailidou, K., Ghoussaini, M., Bolla, M. K., . Zheng, W. (2015). Fine-scale mapping of the 4q24 locus identifies two independent loci associated with breast cancer risk. Cancer Epidemiol Biomarkers Prev, 24(11), 1680-1691. doi:10.1158/1055-9965.EPI-15-0363

H, Z. J. (2017). gap: Genetic Analysis Package. R package version 1.1-17.

Haiman, C. A., Chen, G. K., Vachon, C. M., Canzian, F., Dunning, A., Millikan, R. C., Couch, F. J. (2011). A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nature genetics, 43(12), 1210-1214. doi:10.1038/ng.985 [doi]

Haiman, C. A., Hsu, C., de Bakker, P. I., Frasco, M., Sheng, X., Van Den Berg, D., Henderson, B. E. (2008). Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations. Hum Mol Genet, 17(6), 825-834. doi:10.1093/hmg/ddm354

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., & King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science, 250(4988), 1684-1689.

Harismendy, O., & Frazer, K. A. (2009). Elucidating the role of 8q24 in colorectal cancer. Nat Genet, 41(8), 868-869. doi:10.1038/ng0809-868

Hearle, N., Schumacher, V., Menko, F. H., Olschwang, S., Boardman, L. A., Gille, J. J., Houlston, R. S. (2006). Frequency and spectrum of cancers in the Peutz-Jeghers syndrome. Clin Cancer Res, 12(10), 3209-3215. doi:10.1158/1078-0432.CCR-06-0083

Hill, A. D., Doyle, J. M., McDermott, E. W., & O'Higgins, N. J. (1997). Hereditary breast cancer. Br J Surg, 84(10), 1334-1339.

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. Nat Rev Genet, 6(2), 95-108. doi:10.1038/nrg1521

Honrado, E., Benitez, J., & Palacios, J. (2005). The molecular pathology of hereditary breast cancer: genetic testing and therapeutic implications. Mod Pathol, 18(10), 1305-1320. doi:10.1038/modpathol.3800453

Horne, H. N., Chung, C. C., Zhang, H., Yu, K., Prokunina-Olsson, L., Michailidou, K., Figueroa, J. D. (2016). Fine-Mapping of the 1p11.2 Breast Cancer Susceptibility Locus. PloS one, 11(8), e0160316. doi:10.1371/journal.pone.0160316

Howie, B., Marchini, J., & Stephens, M. (2011). Genotype imputation with thousands of genomes. G3 (Bethesda), 1(6), 457-470. doi:10.1534/g3.111.001198

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet, 5(6), e1000529. doi:10.1371/journal.pgen.1000529

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Chanock, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nature genetics, 39(7), 870-874. doi:ng2075 [pii]

Hurles, M. E., Dermitzakis, E. T., & Tyler-Smith, C. (2008). The functional impact of structural variation in humans. Trends Genet, 24(5), 238-245. doi:10.1016/j.tig.2008.03.001

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Lee, C. (2004a). Detection of large-scale variation in the human genome. Nature genetics, 36. doi:10.1038/ng1416

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Lee, C. (2004b). Detection of large-scale variation in the human genome. Nat Genet, 36(9), 949-951. doi:10.1038/ng1416

Iben, J. R., Epstein, J. A., Bayfield, M. A., Bruinsma, M. W., Hasson, S., Bacikova, D., Maraia, R. J. (2011). Comparative whole genome sequencing reveals phenotypic tRNA gene duplication in spontaneous Schizosaccharomyces pombe La mutants. Nucleic Acids Res, 39(11), 4728-4742. doi:10.1093/nar/gkr066

Iben, J. R., & Maraia, R. J. (2012). tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon:codon wobble pair in a eukaryote. RNA, 18(7), 1358-1372. doi:10.1261/rna.032151.111

Iben, J. R., & Maraia, R. J. (2014). tRNA gene copy number variation in humans. Gene, 536(2), 376-384. doi:10.1016/j.gene.2013.11.049

Iwakawa, R., Okayama, H., Kohno, T., Sato-Otsubo, A., Ogawa, S., & Yokota, J. (2012). Contribution of germline mutations to PARK2 gene inactivation in lung adenocarcinoma. Genes Chromosomes Cancer, 51(5), 462-472. doi:10.1002/gcc.21933

Jacquemont, S., Reymond, A., Zufferey, F., Harewood, L., Walters, R. G., Kutalik, Z., Froguel, P. (2011). Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature, 478(7367), 97-102. doi:10.1038/nature10406

Jaffe, A. B., & Hall, A. (2005). Rho GTPases: biochemistry and biology. Annu Rev Cell Dev Biol, 21, 247-269. doi:10.1146/annurev.cellbio.21.020604.150721

Jedy-Agba, E., Curado, M. P., Ogunbiyi, O., Oga, E., Fabowale, T., Igbinoba, F., Adebamowo, C. A. (2012). Cancer incidence in Nigeria: a report from population-based cancer registries. Cancer Epidemiol, 36(5), e271-278. doi:10.1016/j.canep.2012.04.007

Jin, G., Sun, J., Liu, W., Zhang, Z., Chu, L. W., Kim, S.-T., . . . Xu, J. (2011). Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer. Carcinogenesis, 32(7), 1057-1062. doi:10.1093/carcin/bgr082

Kan, Z., Jaiswal, B. S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H. M., Seshagiri, S. (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. Nature, 466(7308), 869-873. doi:10.1038/nature09208

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res, 32(Database issue), D493-496. doi:10.1093/nar/gkh103

Kazazian, H. H., Jr., & Moran, J. V. (1998). The impact of L1 retrotransposons on the human genome. Nat Genet, 19(1), 19-24. doi:10.1038/ng0598-19

Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Staines, D. M. (2016). Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res, 44(D1), D574-580. doi:10.1093/nar/gkv1209

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Eichler, E. E. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature, 453(7191), 56-64. doi:10.1038/nature06862

Kim, H. C., Lee, J. Y., Sung, H., Choi, J. Y., Park, S. K., Lee, K. M., Kang, D. (2012). A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34:

results from the Seoul Breast Cancer Study. Breast cancer research : BCR, 14(2), R56. doi:bcr3158 [pii]

Kiyotani, K., Mushiroda, T., Tsunoda, T., Morizono, T., Hosono, N., Kubo, M., Zembutsu, H. (2012). A genome-wide association study identifies locus at 10q22 associated with clinical outcomes of adjuvant tamoxifen therapy for breast cancer patients in Japanese. Human molecular genetics, 21(7), 1665-1672. doi:10.1093/hmg/ddr597 [doi]

Kleinjan, D. A., & van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet, 76(1), 8-32. doi:10.1086/426833

Kleinjan, D. J., & van Heyningen, V. (1998). Position effect in human genetic disease. Hum Mol Genet, 7(10), 1611-1618.

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Snyder, M. (2007). Paired-end mapping reveals extensive structural variation in the human genome. Science, 318(5849), 420-426. doi:10.1126/science.1149504

Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res, 42. doi:10.1093/nar/gkt1181

Krepischi, A. C., Achatz, M. I., Santos, E. M., Costa, S. S., Lisboa, B. C., Brentani, H., . . . Rosenberg, C. (2012). Germline DNA copy number variation in familial and early-onset breast cancer. Breast Cancer Res, 14(1), R24. doi:10.1186/bcr3109

Krishnan, P., Ghosh, S., Graham, K., Mackey, J. R., Kovalchuk, O., & Damaraju, S. (2016). Piwi-interacting RNAs and PIWI genes as novel prognostic markers for breast cancer. Oncotarget, 7(25), 37944.

Krishnan, P., Ghosh, S., Graham, K., Mackey, J. R., Kovalchuk, O., & Damaraju, S. (2016). Piwi-interacting RNAs and PIWI genes as novel prognostic markers for breast cancer. Oncotarget, 7(25), 37944-37956. doi:10.18632/oncotarget.9272

Krishnan, P., Ghosh, S., Wang, B., Heyns, M., Graham, K., Mackey, J. R., Damaraju, S. (2016). Profiling of Small Nucleolar RNAs by Next Generation Sequencing: Potential New Players for Breast Cancer Prognosis. PloS one, 11(9), e0162622. doi:10.1371/journal.pone.0162622

Krishnan, P., Ghosh, S., Wang, B., Heyns, M., Li, D., Mackey, J. R., Damaraju, S. (2016). Genome-wide profiling of transfer RNAs and their role as novel prognostic markers for breast cancer. 6, 32843. doi:10.1038/srep32843

https://www.nature.com/articles/srep32843#supplementary-information

Krishnan, P., Ghosh, S., Wang, B., Li, D., Narasimhan, A., Berendt, R., Damaraju, S. (2015). Next generation sequencing profiling identifies miR-574-3p and miR-660-5p as potential novel prognostic markers for breast cancer. BMC Genomics, 16(1), 735. doi:10.1186/s12864-015-1899-0

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet, 22(2), 139-144. doi:10.1038/9642

Kuiper, R. P., Ligtenberg, M. J., Hoogerbrugge, N., & Geurts van Kessel, A. (2010). Germline copy number variation and cancer risk. Curr Opin Genet Dev, 20(3), 282-289. doi:10.1016/j.gde.2010.03.005

Kumaran, M., Cass, C. E., Graham, K., Mackey, J. R., Hubaux, R., Lam, W., . . . Damaraju, S. (2017). Germline copy number variations are associated with breast cancer risk and prognosis. Scientific Reports, 7(1), 14621. doi:10.1038/s41598-017-14799-7

Kumaran, M., Krishnan, P., Cass, C. E., Hubaux, R., Lam, W., Yasui, Y., & Damaraju, S. (2018). Breast cancer associated germline structural variants harboring small noncoding RNAs impact post-transcriptional gene regulation. Sci Rep, 8(1), 7529. doi:10.1038/s41598-018-25801-1

Kuusisto, K. M., Akinrinade, O., Vihinen, M., Kankuri-Tammilehto, M., Laasanen, S. L., & Schleutker, J. (2013). copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. PLoS ONE [Electronic Resource], 8(8), e71802. doi:http://dx.doi.org/10.1371/journal.pone.0071802

Laitinen, V. H., Akinrinade, O., Rantapero, T., Tammela, T. L., Wahlfors, T., & Schleutker, J. (2016). Germline copy number variation analysis in Finnish families with hereditary prostate cancer. Prostate, 76(3), 316-324. doi:10.1002/pros.23123

Lakhani, S. R., Jacquemier, J., Sloane, J. P., Gusterson, B. A., Anderson, T. J., van de Vijver, M. J., . . . Easton, D. F. (1998). Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations. J Natl Cancer Inst, 90(15), 1138-1145.

Lakhani, S. R., Van De Vijver, M. J., Jacquemier, J., Anderson, T. J., Osin, P. P., McGuffog, L., & Easton, D. F. (2002). The pathology of familial breast cancer: predictive value of immunohistochemical markers estrogen receptor, progesterone receptor, HER-2, and p53 in patients with mutations in BRCA1 and BRCA2. J Clin Oncol, 20(9), 2310-2318. doi:10.1200/JCO.2002.09.023

Lander, E. S. (1996). The new genomics: global views of biology. Science, 274(5287), 536-539.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. Nature, 409(6822), 860-921. doi:10.1038/35057062

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res, 22(9), 1813-1831. doi:10.1101/gr.136184.111

Ledet, E. M., Hu, X., Sartor, O., Rayford, W., Li, M., & Mandal, D. (2013). Characterization of germline copy number variation in high-risk African American families with prostate cancer. Prostate, 73(6), 614-623. doi:10.1002/pros.22602

Lee, C., & Scherer, S. W. (2010). The clinical context of copy number variation in the human genome. Expert Rev Mol Med, 12, e8. doi:10.1017/S1462399410001390

Lee, J. A., Carvalho, C. M., & Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell, 131(7), 1235-1247. doi:10.1016/j.cell.2007.11.037

Lestrade, L., & Weber, M. J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic acids research, 34(suppl_1), D158-D162. doi:10.1093/nar/gkj002

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., . . . Venter, J. C. (2007). The diploid genome sequence of an individual human. PLoS Biol, 5(10), e254. doi:10.1371/journal.pbio.0050254

Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E. V., . . . Ruan, Y. (2014). Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics, 15 Suppl 12, S11. doi:10.1186/1471-2164-15-S12-S11

Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S. I., . . . Parsons, R. (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science, 275(5308), 1943-1947.

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol, 34(8), 816-834. doi:10.1002/gepi.20533

Liaw, D., Marsh, D. J., Li, J., Dahia, P. L., Wang, S. I., Zheng, Z., . . . Parsons, R. (1997). Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. Nat Genet, 16(1), 64-67. doi:10.1038/ng0597-64

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer--

analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med, 343(2), 78-85. doi:10.1056/NEJM200007133430201

Lieber, M. R., Lu, H., Gu, J., & Schwarz, K. (2008). Flexibility in the order of action and in the enzymology of the nuclease, polymerases, and ligase of vertebrate non-homologous DNA end joining: relevance to cancer, aging, and the immune system. Cell Res, 18(1), 125-133. doi:10.1038/cr.2007.108

Lieber, M. R., Ma, Y., Pannicke, U., & Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. Nat Rev Mol Cell Biol, 4(9), 712-720. doi:10.1038/nrm1202

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326(5950), 289-293. doi:10.1126/science.1181369

Lilyquist, J., Ruddy, K. J., Vachon, C. M., & Couch, F. J. (2018). Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. Cancer Epidemiol Biomarkers Prev, 27(4), 380-394. doi:10.1158/1055-9965.EPI-17-1144

Lin, W. Y., Camp, N. J., Cannon-Albright, L. A., Allen-Brady, K., Balasubramanian, S., Reed, M. W., . . . Cox, A. (2011). A role for XRCC2 gene polymorphisms in breast cancer risk and survival. J Med Genet, 48(7), 477-484. doi:10.1136/jmedgenet-2011-100018

Liu, B., Yang, L., Huang, B., Cheng, M., Wang, H., Li, Y., . . . Lu, J. (2012). A Functional Copy-Number Variation in MAPKAPK2 Predicts Risk and Prognosis of Lung Cancer. The American Journal of Human Genetics, 91(2), 384-390. doi:http://dx.doi.org/10.1016/j.ajhg.2012.07.003

Liu, J., Sieuwerts, A. M., Look, M. P., van der Vlugt-Daane, M., Meijer-van Gelder, M. E., Foekens, J. A., Martens, J. W. (2016). The 29.5 kb APOBEC3B Deletion Polymorphism Is Not Associated with Clinical Outcome of Breast Cancer. PloS one, 11(8), e0161731. doi:10.1371/journal.pone.0161731

Liu, W., Ramirez, J., Gamazon, E. R., Mirkov, S., Chen, P., Wu, K., Ratain, M. J. (2014). Genetic factors affecting gene transcription and catalytic activity of UDP-glucuronosyltransferases in human liver. Hum Mol Genet, 23(20), 5558-5569. doi:10.1093/hmg/ddu268

Locatelli, I., Lichtenstein, P., & Yashin, A. I. (2004). The heritability of breast cancer: a Bayesian correlated frailty model applied to Swedish twins data. Twin Res, 7(2), 182-191. doi:10.1375/136905204323016168

Long, J., Cai, Q., Shu, X. O., Qu, S., Li, C., Zheng, Y., Zheng, W. (2010). Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. PLoS genetics, 6(6), e1001002. doi:10.1371/journal.pgen.1001002 [doi]

Long, J., Cai, Q., Sung, H., Shi, J., Zhang, B., Choi, J. Y., Zheng, W. (2012). Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. PLoS genetics, 8(2), e1002532. doi:10.1371/journal.pgen.1002532 [doi]

Long, J., Delahanty, R. J., Li, G., Gao, Y. T., Lu, W., Cai, Q., Zheng, W. (2013). A common deletion in the APOBEC3 genes and breast cancer risk. J Natl Cancer Inst, 105(8), 573-579. doi:10.1093/jnci/djt018

Low, S. K., Takahashi, A., Ashikawa, K., Inazawa, J., Miki, Y., Kubo, M., Katagiri, T. (2013). Genome-wide association study of breast cancer in the Japanese population. PloS one, 8(10), e76463. doi:10.1371/journal.pone.0076463 [doi]

Luo, N., Guo, J., Chen, L., Yang, W., Qu, X., & Cheng, Z. (2016). ARHGAP10, downregulated in ovarian cancer, suppresses tumorigenicity of ovarian cancer cells. Cell Death Dis, 7, e2157. doi:10.1038/cddis.2015.401

Lupski, J. R. (2007). Genomic rearrangements and sporadic disease. Nat Genet, 39(7 Suppl), S43-47. doi:10.1038/ng2084

Lupski, J. R., & Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet, 1(6), e49. doi:10.1371/journal.pgen.0010049

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res, 45(D1), D896-D901. doi:10.1093/nar/gkw1133

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res, 42(Database issue), D986-992. doi:10.1093/nar/gkt958

Malhotra, D., & Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. Cell, 148(6), 1223-1241. doi:10.1016/j.cell.2012.02.039

Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Nelson, C. E., Kim, D. H., et, a. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science, 250(4985), 1233.

Mangoni, M., Bisanzi, S., Carozzi, F., Sani, C., Biti, G., Livi, L., Gorini, G. (2011). Association between genetic polymorphisms in the XRCC1, XRCC3, XPD, GSTM1, GSTT1, MSH2, MLH1, MSH3, and MGMT genes and radiosensitivity in breast cancer patients. Int J Radiat Oncol Biol Phys, 81(1), 52-58. doi:10.1016/j.ijrobp.2010.04.023

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. Nature, 461(7265), 747-753. doi:10.1038/nature08494

Marcinkowska, M., Szymanski, M., Krzyzosiak, W. J., & Kozlowski, P. (2011). Copy number variation of microRNA genes in the human genome. BMC Genomics, 12, 183. doi:10.1186/1471-2164-12-183

Marcinkowska, M., Szymanski, M., Krzyzosiak, W. J., & Kozlowski, P. (2011). Copy number variation of microRNA genes in the human genome. BMC Genomics, 12(1), 183. doi:10.1186/1471-2164-12-183

Masson, A. L., Talseth-Palmer, B. A., Evans, T. J., Grice, D. M., Duesing, K., Hannan, G. N., & Scott, R. J. (2013). Copy number variation in hereditary non-polyposis colorectal cancer. Genes (Basel), 4(4), 536-555. doi:10.3390/genes4040536

Masson, A. L., Talseth-Palmer, B. A., Evans, T. J., Grice, D. M., Hannan, G. N., & Scott, R. J. (2014). Expanding the genetic basis of copy number variation in familial breast cancer. Hered Cancer Clin Pract, 12(1), 15. doi:10.1186/1897-4287-12-15

Matuszek, G., & Talebizadeh, Z. (2009). Autism genetic database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. BMC medical genetics, 10(1), 102. doi:10.1186/1471-2350-10-102

Matuszek, G., & Talebizadeh, Z. (2009). Autism Genetic Database (AGD): a comprehensive database including autism susceptibility gene-CNVs integrated with known noncoding RNAs and fragile sites. BMC Med Genet, 10, 102. doi:10.1186/1471-2350-10-102

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet, 9(5), 356-369. doi:10.1038/nrg2344

McPherson, K., Steel, C. M., & Dixon, J. M. (2000). ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. BMJ, 321(7261), 624-628.

Meijers-Heijboer, H., Van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., . . . Van Veghel-Plandsoen, M. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2[ast]1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet, 31(1), 55-59.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., . . . Consortium, C. H.-B. C. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet, 31(1), 55-59. doi:10.1038/ng879

Merla, G., Howald, C., Henrichsen, C. N., Lyle, R., Wyss, C., Zabot, M. T., . . . Reymond, A. (2006). Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. Am J Hum Genet, 79(2), 332-341. doi:10.1086/506371

Meyer, KerstinÂ B., Oâ€™Reilly, M., Michailidou, K., Carlebur, S., Edwards, StaceyÂ L., French, JulietÂ D., . . . Easton, DouglasÂ F. (2013). Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1. American Journal of Human Genetics, 93(6), 1046-1060. doi:10.1016/j.ajhg.2013.10.026

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Shah, M. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet, 47. doi:10.1038/ng.3242

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Bolla, M. K. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet, 45. doi:10.1038/ng.2563

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Easton, D. F. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics, 45(4), 353-361, 361e351-352. doi:10.1038/ng.2563 [doi]

Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Easton, D. F. (2017). Association analysis identifies 65 new breast cancer risk loci. Nature, 551(7678), 92-94. doi:10.1038/nature24284

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science, 266(5182), 66-71.

Mills, G. B., & Moolenaar, W. H. (2003). The emerging role of lysophosphatidic acid in cancer. Nat Rev Cancer, 3(8), 582-591. doi:10.1038/nrc1143

Milne, R. L., Kuchenbaecker, K. B., Michailidou, K., Beesley, J., Kar, S., Lindstrom, S., Simard, J. (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet, 49(12), 1767-1778. doi:10.1038/ng.3785

Moir-Meyer, G. L., Pearson, J. F., Lose, F., Australian National Endometrial Cancer Study, G., Scott, R. J., McEvoy, M., . . . Walker, L. C. (2015). Rare germline copy number deletions of likely functional importance are implicated in endometrial cancer predisposition. Hum Genet, 134(3), 269-278. doi:10.1007/s00439-014-1507-4

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature, 464, 773. doi:10.1038/nature08903

https://www.nature.com/articles/nature08903#supplementary-information

Moreno-De-Luca, D., Consortium, S., Mulle, J. G., Simons Simplex Collection Genetics, C., Kaminsky, E. B., Sanders, S. J., . . . Ledbetter, D. H. (2010). Deletion 17q12 Is a

Recurrent Copy Number Variant that Confers High Risk of Autism and Schizophrenia. American Journal of Human Genetics, 87(5), 618-630. doi:10.1016/j.ajhg.2010.10.004

Morton, N. E. (1955). Sequential tests for the detection of linkage. Am J Hum Genet, 7(3), 277-318.

Mulligan, A. M., Couch, F. J., Barrowdale, D., Domchek, S. M., Eccles, D., Nevanlinna, H., Cimba. (2011). Common breast cancer susceptibility alleles are associated with tumour subtypes in BRCA1 and BRCA2 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2. Breast Cancer Res, 13(6), R110. doi:10.1186/bcr3052

Nelson, J. B., Udan, M. S., Guruli, G., & Pflug, B. R. (2005). Endothelin-1 inhibits apoptosis in prostate cancer. Neoplasia, 7(7), 631-637.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., . . . Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature, 489(7414), 83-90. doi:10.1038/nature11212

Norskov, M. S., Frikke-Schmidt, R., Bojesen, S. E., Nordestgaard, B. G., Loft, S., & Tybjaerg-Hansen, A. (2011). Copy number variation in glutathione-S-transferase T1 and M1 predicts incidence and 5-year survival from prostate and bladder cancer, and incidence of corpus uteri cancer in the general population. Pharmacogenomics J, 11(4), 292-299. doi:10.1038/tpj.2010.38

Olivier, M., Goldgar, D. E., Sodha, N., Ohgaki, H., Kleihues, P., Hainaut, P., & Eeles, R. A. (2003). Li-Fraumeni and related syndromes: correlation between tumor type, family structure, and TP53 genotype. Cancer Res, 63(20), 6643-6650.

Orr, N., Dudbridge, F., Dryden, N., Maguire, S., Novo, D., Perrakis, E., Peto, J. (2015). Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Human molecular genetics, 24(10), 2966-2984. doi:10.1093/hmg/ddv035. (PMID: 25652398)

Palacios, S., Henderson, V. W., Siseles, N., Tan, D., & Villaseca, P. (2010). Age of menopause and impact of climacteric symptoms by geographical region. Climacteric, 13(5), 419-428. doi:10.3109/13697137.2010.507886

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Scherer, S. W. (2010). Towards a comprehensive structural variation map of an individual human genome. Genome Biol, 11(5), R52. doi:10.1186/gb-2010-11-5-r52

Pang, A. W., Macdonald, J. R., Yuen, R. K., Hayes, V. M., & Scherer, S. W. (2014). Performance of high-throughput sequencing for the discovery of genetic variation across the complete size spectrum. G3 (Bethesda), 4(1), 63-65. doi:10.1534/g3.113.008797

Pathology of familial breast cancer: differences between breast cancers in carriers of BRCA1 or BRCA2 mutations and sporadic cases. Breast Cancer Linkage Consortium. (1997). Lancet, 349(9064), 1505-1510.

Pedersen, B. S., Konstantinopoulos, P. A., Spillman, M. A., & De, S. (2013). Copy neutral loss of heterozygosity is more frequent in older ovarian cancer patients. Genes Chromosomes Cancer, 52(9), 794-801. doi:10.1002/gcc.22075

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Botstein, D. (2000). Molecular portraits of human breast tumours. Nature, 406(6797), 747-752. doi:10.1038/35021093

Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., . . . Lee, C. (2008). The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet, 82(3), 685-695. doi:10.1016/j.ajhg.2007.12.010

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Misra, R. (2007). Diet and the evolution of human amylase gene copy number variation. Nature genetics, 39. doi:10.1038/ng2123

Persengiev, S., Kondova, I., & Bontrop, R. (2013). Insights on the functional interactions between miRNAs and copy number variations in the aging brain. Frontiers in Molecular Neuroscience, 6, 32.

Pharoah, P. D., Guilford, P., Caldas, C., & International Gastric Cancer Linkage, C. (2001). Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. Gastroenterology, 121(6), 1348-1353.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., . . . Scherer, S. W. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet, 94(5), 677-694. doi:10.1016/j.ajhg.2014.03.018

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., . . . Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature, 466(7304), 368-372. doi:10.1038/nature09146

piRNAdb https://www.bioinfo.mochsl.org.br/~rpiuco/pirna/. (2016). Retrieved May 2017 https://www.bioinfo.mochsl.org.br/~rpiuco/pirna/

Pooley, K. A., Baynes, C., Driver, K. E., Tyrer, J., Azzato, E. M., Pharoah, P. D., Dunning, A. M. (2008). Common single-nucleotide polymorphisms in DNA double-strand break repair genes and breast cancer risk. Cancer Epidemiol Biomarkers Prev, 17(12), 3482-3489. doi:10.1158/1055-9965.EPI-08-0594

Pritchard, J. K., & Cox, N. J. (2002). The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet, 11(20), 2417-2423.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., . . . Sham, P. C. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics, 81(3), 559-575. doi:10.1086/519795

Pylkas, K., Vuorela, M., Otsukka, M., Kallioniemi, A., Jukkola-Vuorinen, A., & Winqvist, R. (2012). Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. PLoS Genet, 8(6), e1002734. doi:10.1371/journal.pgen.1002734

Rafiq, S., Khan, S., Tapper, W., Collins, A., Upstill-Goddard, R., Gerty, S., Eccles, D. (2014). A genome wide meta-analysis study for identification of common variation associated with breast cancer prognosis. PloS one, 9(12), e101488. doi:10.1371/journal.pone.0101488 [doi]

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Stratton, M. R. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet, 39(2), 165-167. doi:10.1038/ng1959

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Stratton, M. R. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet, 39(2), 165-167. doi:http://www.nature.com/ng/journal/v39/n2/suppinfo/ng1959_S1.html

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lander, E. S. (2001). Linkage disequilibrium in the human genome. Nature, 411(6834), 199-204. doi:10.1038/35075590

Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. Trends Genet, 17(9), 502-510.

Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., Rahman, N. (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet, 38(8), 873-875. doi:10.1038/ng1837

Renwick, A., Thompson, D., Seal, S., Kelly, P., Chagtai, T., Ahmed, M., Rahman, N. (2006). ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet, 38(8), 873-875. doi:http://www.nature.com/ng/journal/v38/n8/suppinfo/ng1837_S1.html

Rezaei, M., Hashemi, M., Hashemi, S. M., Mashhadi, M. A., & Taheri, M. (2015). APOBEC3 Deletion is Associated with Breast Cancer Risk in a Sample of Southeast Iranian Population. International Journal of Molecular and Cellular Medicine, 4(2), 103-108.

Ribelles, N., Santonja, A., Pajares, B., Llacer, C., & Alba, E. (2014). The seed and soil hypothesis revisited: current state of knowledge of inherited genes on prognosis in breast cancer. Cancer Treat Rev, 40(2), 293-299. doi:10.1016/j.ctrv.2013.09.010

Riggs, E. R., Church, D. M., Hanson, K., Horner, V. L., Kaminsky, E. B., Kuhn, R. M., Martin, C. L. (2012). Towards an evidence-based process for the clinical interpretation of copy number variation. Clin Genet, 81(5), 403-412. doi:10.1111/j.1399-0004.2011.01818.x

Rinella, E. S., Shao, Y., Yackowski, L., Pramanik, S., Oratz, R., Schnabel, F., Ostrer, H. (2013). Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable BRCA1/2 mutation. Human genetics, 132(5), 523-536. doi:10.1007/s00439-013-1269-4 [doi]

Roa, B. B., Boyd, A. A., Volcik, K., & Richards, C. S. (1996). Ashkenazi Jewish population frequencies for common mutations in BRCA1 and BRCA2. Nat Genet, 14(2), 185-187. doi:10.1038/ng1096-185

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. Nature, 518(7539), 317-330. doi:10.1038/nature14248

Robson, M. E., Chappuis, P. O., Satagopan, J., Wong, N., Boyd, J., Goffin, J. R., Foulkes, W. D. (2004). A combined analysis of outcome following breast cancer: differences in survival based on BRCA1/BRCA2 mutation status and administration of adjuvant treatment. Breast Cancer Res, 6(1), R8-R17. doi:10.1186/bcr658

Rosano, L., Cianfrocca, R., Masi, S., Spinella, F., Di Castro, V., Biroccio, A., Bagnato, A. (2009). Beta-arrestin links endothelin A receptor to beta-catenin signaling to induce ovarian cancer cell invasion and metastasis. Proc Natl Acad Sci U S A, 106(8), 2806-2811. doi:10.1073/pnas.0807158106

Rose-Zerilli, M. J., Barton, S. J., Henderson, A. J., Shaheen, S. O., & Holloway, J. W. (2009). Copy-number variation genotyping of GSTT1 and GSTM1 gene deletions by real-time PCR. Clin Chem, 55(9), 1680-1685. doi:10.1373/clinchem.2008.120105

Ross, R. J., Weiner, M. M., & Lin, H. (2014). PIWI proteins and PIWI-interacting RNAs in the soma. Nature, 505(7483), 353-359. doi:10.1038/nature12987

Sahoo, T., del Gaudio, D., German, J. R., Shinawi, M., Peters, S. U., Person, R. E., Beaudet, A. L. (2008). Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet, 40(6), 719-721. doi:10.1038/ng.158

Sapkota, Y., Ghosh, S., Lai, R., Coe, B. P., Cass, C. E., Yasui, Y., . . . Damaraju, S. (2013). Germline DNA copy number aberrations identified as potential prognostic factors for breast cancer recurrence. PloS one, 8(1), e53850. doi:10.1371/journal.pone.0053850

Sapkota, Y., Narasimhan, A., Kumaran, M., Sehrawat, B. S., & Damaraju, S. (2016). A Genome-Wide Association Study to Identify Potential Germline Copy Number Variants for Sporadic Breast Cancer Susceptibility. Cytogenet Genome Res, 149(3), 156-164. doi:10.1159/000448558

Sapkota, Y., Yasui, Y., Lai, R., Sridharan, M., Robson, P. J., Cass, C. E., Damaraju, S. (2013). Identification of a breast cancer susceptibility locus at 4q31.22 using a genome-wide association study paradigm. PloS one, 8(5), e62550. doi:10.1371/journal.pone.0062550

Saunders, M. A., Liang, H., & Li, W.-H. (2007). Human polymorphism at microRNAs and microRNA target sites. Proceedings of the National Academy of Sciences of the United States of America, 104(9), 3300-3305. doi:10.1073/pnas.0611347104

Schwarz, K., Ma, Y., Pannicke, U., & Lieber, M. R. (2003). Human severe combined immune deficiency and DNA repair. Bioessays, 25(11), 1061-1070. doi:10.1002/bies.10344

Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Rahman, N. (2006). Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet, 38(11), 1239-1241. doi:10.1038/ng1902

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Wigler, M. (2007). Strong association of de novo copy number mutations with autism. Science, 316. doi:10.1126/science.1138659

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. Science, 305(5683), 525-528. doi:10.1126/science.1098918

Sehl, M. E., Langer, L. R., Papp, J. C., Kwan, L., Seldon, J. L., Arellano, G., Ganz, P. A. (2009). Associations between single nucleotide polymorphisms in double-stranded DNA repair pathway genes and familial breast cancer. Clin Cancer Res, 15(6), 2192-2203. doi:10.1158/1078-0432.CCR-08-1417

Sehrawat, B., Sridharan, M., Ghosh, S., Robson, P., Cass, C. E., Mackey, J. R., Damaraju, S. (2011). Potential novel candidate polymorphisms identified in genome-wide association study for breast cancer susceptibility. Human genetics, 130(4), 529-537. doi:10.1007/s00439-011-0973-1 [doi]

Sexton, T., Bantignies, F., & Cavalli, G. (2009). Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. Semin Cell Dev Biol, 20(7), 849-855. doi:10.1016/j.semcdb.2009.06.004

Shaag, A., Walsh, T., Renbaum, P., Kirchhoff, T., Nafa, K., Shiovitz, S., King, M. C. (2005). Functional and genomic approaches reveal an ancient CHEK2 allele associated with breast cancer in the Ashkenazi Jewish population. Hum Mol Genet, 14(4), 555-563. doi:10.1093/hmg/ddi052

Shaw, C. J., & Lupski, J. R. (2004). Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet, 13 Spec No 1, R57-64. doi:10.1093/hmg/ddh073

Shi, J., Zhang, Y., Zheng, W., Michailidou, K., Ghoussaini, M., Bolla, M. K., Long, J. (2016). Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. Int J Cancer, 139(6), 1303-1317. doi:10.1002/ijc.30150

Shi, J., Zhou, W., Zhu, B., Hyland, P. L., Bennett, H., Xiao, Y., Yang, X. R. (2016). Rare Germline Copy Number Variations and Disease Susceptibility in Familial Melanoma. J Invest Dermatol, 136(12), 2436-2443. doi:10.1016/j.jid.2016.07.023

Shu, J., Xia, Z., Li, L., Liang, E. T., Slipek, N., Shen, D., Steer, C. J. (2012). Dose-dependent differential mRNA target selection and regulation by let-7a-7f and miR-17-92 cluster microRNAs. RNA Biol, 9(10), 1275-1287. doi:10.4161/rna.21998

Siegel, R., Ma, J., Zou, Z., & Jemal, A. (2014). Cancer statistics, 2014. CA Cancer J Clin, 64. doi:10.3322/caac.21208

Sighoko, D., Bah, E., Haukka, J., McCormack, V. A., Aka, E. P., Bourgeois, D., Hainaut, P. (2010). Population-based breast (female) and cervix cancer rates in the Gambia: evidence of ethnicity-related variations. Int J Cancer, 127(10), 2248-2256. doi:10.1002/ijc.25244

Sighoko, D., Kamate, B., Traore, C., Malle, B., Coulibaly, B., Karidiatou, A., Hainaut, P. (2013). Breast cancer in pre-menopausal women in West Africa: analysis of temporal trends and evaluation of risk factors associated with reproductive life. Breast, 22(5), 828-835. doi:10.1016/j.breast.2013.02.011

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., . . . de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet, 38(11), 1348-1354. doi:10.1038/ng1896

Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol, 12(4), 246-258. doi:10.1038/nrm3089

SNP & Variation Suite ™ (Version 8.x) [Software]. Bozeman, MT: Golden Helix, Inc. Available from http://www.goldenhelix.com.

Soule, H. D., Maloney, T. M., Wolman, S. R., Peterson, W. D., Jr., Brenz, R., McGrath, C. M., Brooks, S. C. (1990). Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. Cancer Res, 50(18), 6075-6086.

Spiegelberg, B. D., & Hamm, H. E. (2007). Roles of G-protein-coupled receptor signaling in cancer biology and gene transcription. Curr Opin Genet Dev, 17(1), 40-44. doi:10.1016/j.gde.2006.12.002

Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., Stefansson, K. (2007). Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics, 39(7), 865-869. doi:ng2064 [pii]

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. Trends Genet, 18(2), 74-82.

Stevens, K. N., Fredericksen, Z., Vachon, C. M., Wang, X., Margolin, S., Lindblom, A., Couch, F. J. (2012). 19p13.1 is a triple-negative-specific breast cancer susceptibility locus. Cancer Res, 72(7), 1795-1803. doi:10.1158/0008-5472.CAN-11-3364

Stevens, K. N., Vachon, C. M., Lee, A. M., Slager, S., Lesnick, T., Olswold, C., . . . Couch, F. J. (2011). Common breast cancer susceptibility loci are associated with triple-negative breast cancer. Cancer Res, 71(19), 6240-6249. doi:10.1158/0008-5472.CAN-11-1266

Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. Nature, 403(6765), 41-45. doi:10.1038/47412

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Lee, C. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science, 315. doi:10.1126/science.1136678

Struewing, J. P., Abeliovich, D., Peretz, T., Avishai, N., Kaback, M. M., Collins, F. S., & Brody, L. C. (1995). The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals. Nat Genet, 11(2), 198-200. doi:10.1038/ng1095-198

Struewing, J. P., Tarone, R. E., Brody, L. C., Li, F. P., & Boice, J. D., Jr. (1996). BRCA1 mutations in young women with breast cancer. Lancet, 347(9013), 1493.

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D. A., Rossi, J. J. (2009). SNPs in human miRNA genes affect biogenesis and function. RNA, 15(9), 1640-1651. doi:10.1261/rna.1560209

Syamala, V. S., Sreeja, L., Syamala, V., Raveendran, P. B., Balakrishnan, R., Kuttan, R., & Ankathil, R. (2008). Influence of germline polymorphisms of GSTT1, GSTM1, and GSTP1 in familial versus sporadic breast cancer susceptibility and survival. Fam Cancer, 7(3), 213-220. doi:10.1007/s10689-007-9177-1

Syrjakoski, K., Vahteristo, P., Eerola, H., Tamminen, A., Kivinummi, K., Sarantaus, L., Nevanlinna, H. (2000). Population-based study of BRCA1 and BRCA2 mutations in 1035 unselected Finnish breast cancer patients. J Natl Cancer Inst, 92(18), 1529-1531.

Team, R. C. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Teng, J. P., Yang, Z. Y., Zhu, Y. M., Ni, D., Zhu, Z. J., & Li, X. Q. (2017). The roles of ARHGAP10 in the proliferation, migration and invasion of lung cancer cells. Oncol Lett, 14(4), 4613-4618. doi:10.3892/ol.2017.6729

Teo, Z. L., Park, D. J., Provenzano, E., Chatfield, C. A., Odefrey, F. A., Nguyen-Dumont, T., Southey, M. C. (2013). Prevalence of PALB2 mutations in Australasian multiple-case breast cancer families. Breast Cancer Res, 15(1), R17. doi:10.1186/bcr3392

Therneau, T. M. (2015). A Package for Survival Analysis in S . version 2.38.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. Nature, 489(7414), 75-82. doi:10.1038/nature11232

Tomlinson, I. P., & Houlston, R. S. (1997). Peutz-Jeghers syndrome. J Med Genet, 34(12), 1007-1011.

Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A., van Heel, D. A. (2011). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet, 43(12), 1193-1201. doi:10.1038/ng.998

Turner, D. J., Miretti, M., Rajan, D., Fiegler, H., Carter, N. P., Blayney, M. L., Hurles, M. E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat Genet, 40(1), 90-95. doi:10.1038/ng.2007.40

Udler, M. S., Ahmed, S., Healey, C. S., Meyer, K., Struewing, J., Maranian, M., Dunning, A. M. (2010). Fine scale mapping of the breast cancer 16q12 locus. Human molecular genetics, 19(12), 2507-2515. doi:10.1093/hmg/ddq122

Udler, M. S., Meyer, K. B., Pooley, K. A., Karlins, E., Struewing, J. P., Zhang, J., Easton, D. F. (2009). FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. Human molecular genetics, 18(9), 1692-1703. doi:10.1093/hmg/ddp078

Udler, M. S., Tyrer, J., & Easton, D. F. (2010). Evaluating the Power to Discriminate Between Highly Correlated SNPs in Genetic Association Studies. Genetic epidemiology, 34(5), 463-468. doi:10.1002/gepi.20504

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871), 530-536. doi:10.1038/415530a van Corven, E. J., Groenink, A., Jalink, K., Eichholtz, T., & Moolenaar, W. H. (1989). Lysophosphatidate-induced cell proliferation: identification and dissection of signaling pathways mediated by G proteins. Cell, 59(1), 45-54.

van Lier, M. G., Wagner, A., Mathus-Vliegen, E. M., Kuipers, E. J., Steyerberg, E. W., & van Leerdam, M. E. (2010). High cancer risk in Peutz-Jeghers syndrome: a systematic review and surveillance recommendations. Am J Gastroenterol, 105(6), 1258-1264; author reply 1265. doi:10.1038/ajg.2009.725

Venkatachalam, R., Verwiel, E. T., Kamping, E. J., Hoenselaar, E., Gorgens, H., Schackert, H. K., . . . Kuiper, R. P. (2011). Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. Int J Cancer, 129(7), 1635-1642. doi:10.1002/ijc.25821

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A., & Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep, 10(8), 1297-1309. doi:10.1016/j.celrep.2015.02.004

Villacis, R. A., Abreu, F. B., Miranda, P. M., Domingues, M. A., Carraro, D. M., Santos, E. M., . . . Rogatto, S. R. (2016). ROBO1 deletion as a novel germline alteration in breast and colorectal cancer patients. Tumour Biol, 37(3), 3145-3153. doi:10.1007/s13277-015-4145-0

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., . . . Boehnke, M. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet, 8(8), e1002793. doi:10.1371/journal.pgen.1002793

Walker, L. C., Pearson, J. F., Wiggins, G. A., Giles, G. G., Hopper, J. L., & Southey, M. C. (2017). Increased genomic burden of germline copy number variants is associated with early onset breast cancer: Australian breast cancer family registry. Breast Cancer Res, 19(1), 30. doi:10.1186/s13058-017-0825-6

Walsh, T., Casadei, S., Coats, K. H., Swisher, E., Stray, S. M., Higgins, J., King, M. C. (2006). Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. JAMA, 295(12), 1379-1388. doi:10.1001/jama.295.12.1379

Wang, Y., Zhang, B., Zhang, L., An, L., Xu, J., Li, D., Yue, F. (2017). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. bioRxiv.

Ward, L. D., & Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic acids research, 40(D1), D930-D934. doi:10.1093/nar/gkr917

Warnica, W., Merico, D., Costain, G., Alfred, S. E., Wei, J., Marshall, C. R., Bassett, A. S. (2015). Copy Number Variable MicroRNAs in Schizophrenia and Their Neurodevelopmental Gene Targets. Biological Psychiatry, 77(2), 158-166. doi:http://dx.doi.org/10.1016/j.biopsych.2014.05.011

Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R., Daly, M. J. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. The New England Journal of Medicine, 358. doi:10.1056/NEJMoa075974

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res, 42(Database issue), D1001-1006. doi:10.1093/nar/gkt1229

Weren, R. D., Venkatachalam, R., Cazier, J. B., Farin, H. F., Kets, C. M., de Voer, R. M., Kuiper, R. P. (2015). Germline deletions in the tumour suppressor gene FOCAD are associated with polyposis and colorectal cancer development. J Pathol, 236(2), 155-164. doi:10.1002/path.4520

Wilson, J. L., Burchell, J., & Grimshaw, M. J. (2006). Endothelins induce CCR7 expression by breast tumor cells via endothelin receptor A and hypoxia-inducible factor-1. Cancer Res, 66(24), 11802-11807. doi:10.1158/0008-5472.CAN-06-1222

Wong, N. C., Bhadri, V. A., Maksimovic, J., Parkinson-Bates, M., Ng, J., Craig, J. M., . . . Lock, R. B. (2014). Stability of gene expression and epigenetic profiles highlights the utility of patient-derived paediatric acute lymphoblastic leukaemia xenografts for investigating molecular mechanisms of drug resistance. BMC Genomics, 15, 416. doi:10.1186/1471-2164-15-416

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature, 378(6559), 789-792. doi:10.1038/378789a0

Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., et al. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science, 265(5181), 2088-2090.

Wulfing, P., Diallo, R., Kersting, C., Wulfing, C., Poremba, C., Rody, A., Kiesel, L. (2003). Expression of endothelin-1, endothelin-A, and endothelin-B receptor in human breast cancer and correlation with long-term follow-up. Clin Cancer Res, 9(11), 4125-4131.

Wulfing, P., Kersting, C., Tio, J., Fischer, R. J., Wulfing, C., Poremba, C., Kiesel, L. (2004). Endothelin-1-, endothelin-A-, and endothelin-B-receptor expression is correlated with vascular endothelial growth factor expression and angiogenesis in breast cancer. Clin Cancer Res, 10(7), 2393-2400.

Xing, H. J., Li, Y. J., Ma, Q. M., Wang, A. M., Wang, J. L., Sun, M., Wang, L. (2013). Identification of microRNAs present in congenital heart disease associated copy number variants. Eur Rev Med Pharmacol Sci, 17(15), 2114-2120.

Xing, J., Zhang, Y., Han, K., Salem, A. H., Sen, S. K., Huff, C. D., Jorde, L. B. (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res, 19(9), 1516-1526. doi:10.1101/gr.091827.109

Xiong, Y., Zhang, H., & Beach, D. (1992). D type cyclins associate with multiple protein kinases and the DNA replication and repair factor PCNA. Cell, 71(3), 505-514.

Xu, B., Roos, J. L., Levy, S., van Rensburg, E. J., Gogos, J. A., & Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet, 40(7), 880-885. doi:10.1038/ng.162

Xuan, D., Li, G., Cai, Q., Deming-Halverson, S., Shrubsole, M. J., Shu, X. O., Long, J. (2013a). APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis, 34. doi:10.1093/carcin/bgt185

Xuan, D., Li, G., Cai, Q., Deming-Halverson, S., Shrubsole, M. J., Shu, X. O., Long, J. (2013b). APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis, 34(10), 2240-2243. doi:10.1093/carcin/bgt185

Yang, R., Chen, B., Pfütze, K., Buch, S., Steinke, V., Holinski-Feder, E., Burwinkel, B. (2014). Genome-wide analysis associates familial colorectal cancer with increases in copy number variations and a rare structural variation at 12p12.3. Carcinogenesis, 35(2), 315-323. doi:10.1093/carcin/bgt344

Yang, T. L., Chen, X. D., Guo, Y., Lei, S. F., Wang, J. T., Zhou, Q., Deng, H. W. (2008). Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. Am J Hum Genet, 83(6), 663-674. doi:10.1016/j.ajhg.2008.10.006

Yoshihara, K., Tajima, A., Adachi, S., Quan, J., Sekine, M., Kase, H., Tanaka, K. (2011). Germline copy number variations in BRCA1-associated ovarian cancer patients. Genes Chromosomes Cancer, 50(3), 167-177. doi:10.1002/gcc.20841

Yu, H., Pardoll, D., & Jove, R. (2009). STATs in cancer inflammation and immunity: a leading role for STAT3. Nat Rev Cancer, 9(11), 798-809. doi:10.1038/nrc2734

Yu, K.-D., Fan, L., Di, G.-H., Yuan, W.-T., Zheng, Y., Huang, W., Shao, Z.-M. (2010). Genetic variants in GSTM3 gene within GSTM4-GSTM2-GSTM1-GSTM5-GSTM3 cluster influence breast cancer susceptibility depending on GSTM1. Breast cancer research and treatment, 121(2), 485-496. doi:10.1007/s10549-009-0585-9

Yu, K. D., Di, G. H., Fan, L., Wu, J., Hu, Z., Shen, Z. Z., Shao, Z. M. (2009). A functional polymorphism in the promoter region of GSTM1 implies a complex role for GSTM1 in breast cancer. FASEB J, 23(7), 2274-2287. doi:10.1096/fj.08-124073

Zarrei, M., MacDonald, J. R., Merico, D., & Scherer, S. W. (2015). A copy number variation map of the human genome. Nat Rev Genet, 16(3), 172-183. doi:10.1038/nrg3871

Zelada-Hedman, M., Wasteson Arver, B., Claro, A., Chen, J., Werelius, B., Kok, H., Lindblom, A. (1997). A screening for BRCA1 mutations in breast and breast-ovarian cancer families from the Stockholm region. Cancer Res, 57(12), 2474-2477.

Zeng, C., Guo, X., Long, J., Kuchenbaecker, K. B., Droit, A., Michailidou, K., Zheng, W. (2016). Identification of independent association signals and putative functional variants for breast cancer risk through fine-scale mapping of the 12p11 locus. Breast Cancer Research, 18(1), 64. doi:10.1186/s13058-016-0718-0

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. Annual review of genomics and human genetics, 10, 451-481. doi:10.1146/annurev.genom.9.081307.164217

Zhang, H., Xiong, Y., & Beach, D. (1993). Proliferating cell nuclear antigen and p21 are components of multiple cell cycle kinase complexes. Mol Biol Cell, 4(9), 897-906.

Zhang, W. M., Zhou, J., & Ye, Q. J. (2008). Endothelin-1 enhances proliferation of lung cancer cells by increasing intracellular free Ca2+. Life Sci, 82(13-14), 764-771. doi:10.1016/j.lfs.2008.01.008

Zhao, J. H. (2007). gap: Genetic Analysis Package. 2007, 23(8), 18. doi:10.18637/jss.v023.i08

Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., . . . Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet, 38(11), 1341-1347. doi:10.1038/ng1891

Zheng, W., Long, J., Gao, Y. T., Li, C., Zheng, Y., Xiang, Y. B., Shu, X. O. (2009). Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nature genetics, 41(3), 324-328. doi:10.1038/ng.318 [doi]

Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J. M., Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature, 495(7439), 111-115. doi:10.1038/nature11833

Appendix

a. Distribution of Age in cases and controls (Internal dataset Stages 1-4)

Age group	Cases			All Controls (n)
	Pre-menopausal (n)	Post-menopausal (n)	All (n)	
Median	46 [21-70]	62 [35-93]	57 [21-93]	53 [34-78]
<40	284	7	295	340
40-50	876	105	1007	1382
50-60	353	952	1355	1561
60-70	15	1055	1095	1090
70-80		467	486	144
>80		154	160	

b. Distribution of the Body Mass Index between cases and controls (Internal dataset Stages 1 - 4)

Sample status	Median [25th -75th percentile]
All cases	27.55 [24.22-31.92]
Premenopausal cases	25.98 [23.24-30.40]
Postmenopausal cases	28.19[25.05-32.45]
Controls	25.40 [22.73-29.23]

Figure A.1 Distribution of Age and Body Mass Index in the study population

Table A.1 Patient Demographics for the internal dataset (Stages 1-4)

	Premenopausal cases (total n=1670)	**Postmenopausal cases (n=3163)**	**All cases combined n=4964***
Subtype			
Luminal A	1006 (60%)	2146 (68%)	3229 (65%)
Luminal B	269 (16%)	329 (10%)	610 (12%)
HER 2+	75 (4%)	119 (4%)	203 (4%)

Triple Negative	192 (11%)	325 (10%)	525 (11%)
Unknown	128 (8%)	244 (8%)	397 (8%)
Stage			
0-111A	1588 (95%)	2991 (95%)	4693 (95%)
IIIB	82 (5%)	174 (6%)	271 (5%)
Grade			
Low	652 (39%)	1293 (41%)	1993 (40%)
High	609 (36%)	772 (24%)	1409 (28%)
Unknown	409 (24%)	1098 (35%)	1562 (31%)
Family history			
Yes	729 (44%)	1208 (38%)	1974 (40%)
No	826 (49%)	1796 (57%)	2654 (53%)
Unknown	115 (7%)	159 (5%)	336 (7%)

Table A.2 Association of fine-mapped SNPs with a premenopausal breast cancer risk

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
rs13134510	C	0.23	1.11E-12	1.43 [1.3-1.58]	1.32 [1.15-1.5]	2.17 [1.72-2.74]	7.78E-12	1.43 [1.29-1.58]	1.34 [1.17-1.54]	2.22 [1.74-2.84]	Genotyped
rs1366691	C	0.21	1.91E-12	1.43 [1.29-1.58]	1.27 [1.11-1.45]	2.1 [1.71-2.71]	2.96E-11	1.41 [1.27-1.56]	1.3 [1.13-1.49]	2.22 [1.74-2.82]	Imputed
rs1429139	T	0.22	6.64E-12	1.42 [1.29-1.57]	1.25 [1.09-1.43]	2.16 [1.72-2.72]	3.33E-11	1.41 [1.27-1.56]	1.29 [1.12-1.49]	2.24 [1.76-2.85]	Imputed
rs12501429	T	0.21	1.19E-11	1.42 [1.28-1.57]	1.27 [1.11-1.45]	2.1 [1.67-2.65]	1.65E-10	1.39 [1.26-1.54]	1.28 [1.11-1.48]	2.18 [1.71-2.78]	Imputed
rs1583003	A	0.22	1.30E-11	1.39 [1.27-1.54]	1.31 [1.16-1.49]	2.07 [1.63-2.62]	1.51E-11	1.41 [1.28-1.56]	1.34 [1.17-1.53]	2.18 [1.7-2.79]	Genotyped
rs2163012	G	0.22	2.45E-11	1.4 [1.27-1.55]	1.25 [1.09-1.43]	2.05 [1.64-2.57]	1.56E-10	1.39 [1.25-1.53]	1.28 [1.11-1.48]	2.12 [1.68-2.69]	Imputed
rs10519886	T	0.24	2.93E-11	1.38 [1.25-1.52]	1.26 [1.11-1.43]	2.07 [1.66-2.59]	1.07E-10	1.38 [1.25-1.53]	1.29 [1.13-1.47]	2.12 [1.67-2.67]	Genotyped
rs2163011	A	0.23	4.13E-11	1.39 [1.26-1.53]	1.27 [1.12-1.45]	1.99 [1.59-2.49]	2.40E-10	1.38 [1.25-1.52]	1.31 [1.14-1.5]	2.04 [1.61-2.57]	Imputed
rs12498595	C	0.23	6.85E-11	1.38 [1.26-	1.26 [1.11-	1.99 [1.59-	4.13E-10	1.37 [1.24-1.52]	1.3 [1.13-1.49]	2.03 [1.61-	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
				1.53]	1.44]	2.48]				2.57]	
rs13120678	G	0.22	1.16E-10	1.39 [1.26-1.54]	1.24 [1.08-1.42]	2.04 [1.63-2.57]	1.00 E-09	1.37 [1.24-1.52]	1.27 [1.1-1.46]	2.1 [1.65-2.67]	Imputed
rs12511935	T	0.21	2.41E-10	1.39 [1.25-1.53]	1.29 [1.13-1.47]	1.97 [1.55-2.51]	1.15 E-09	1.38 [1.24-1.53]	1.31 [1.14-1.5]	2.06 [1.6-2.65]	Imputed
rs12500103	G	0.21	2.85E-10	1.39 [1.25-1.54]	1.29 [1.13-1.47]	1.96 [1.54-2.5]	1.43 E-09	1.38 [1.24-1.53]	1.31 [1.14-1.51]	2.05 [1.59-2.64]	Imputed
rs1366679	G	0.21	4.57E-10	1.38 [1.25-1.53]	1.29 [1.13-1.47]	1.94 [1.53-2.48]	2.45 E-09	1.37 [1.24-1.52]	1.3 [1.13-1.5]	2.03 [1.57-2.61]	Imputed
rs11735996	T	0.21	5.14E-10	1.37 [1.24-1.52]	1.28 [1.13-1.46]	1.94 [1.53-2.46]	1.83 E-09	1.37 [1.24-1.51]	1.3 [1.13-1.5]	2.02 [1.57-2.6]	Imputed
rs28645698	C	0.18	1.57E-09	1.37 [1.24-1.52]	1.36 [1.2-1.54]	1.85 [1.4-2.45]	1.22 E-09	1.39 [1.25-1.55]	1.4 [1.22-1.6]	1.93 [1.44-2.6]	Genotyped
rs1429133	C	0.21	3.33E-09	1.34 [1.22-1.48]	1.26 [1.11-1.43]	1.94 [1.52-2.48]	2.74 E-09	1.36 [1.23-1.5]	1.29 [1.13-1.47]	2.03 [1.58-2.62]	Genotyped
rs6810798	A	0.18	3.58E-09	1.36 [1.23-1.51]	1.36 [1.2-1.55]	1.8 [1.36-2.39]	2.83 E-09	1.38 [1.24-1.54]	1.4 [1.22-1.6]	1.87 [1.39-2.52]	Genotyped
rs28720373	T	0.18	4.73E-09	1.36 [1.23-1.51]	1.37 [1.2-1.56]	1.77 [1.32-2.36]	2.03 E-09	1.39 [1.25-1.55]	1.4 [1.23-1.61]	1.89 [1.4-2.55]	Genotyped
rs1429142	C	0.1	4.99E-	1.35	1.33	1.89	5.81	1.40 [1.26-	1.40	1.96	Genotyped

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
		8	09	[1.22-1.50]	[1.17-1.52]	[1.43-2.51]	E-10	1.55]	[1.22-1.60]	[1.46-2.63]	
rs1429134	T	0.19	7.74E-09	1.36 [1.22-1.51]	1.22 [1.07-1.4]	1.99 [1.55-2.54]	1.18 E-08	1.36 [1.22-1.51]	1.25 [1.08-1.44]	2.1 [1.62-2.72]	Imputed
rs1346600	A	0.20	1.24E-08	1.35 [1.22-1.5]	1.22 [1.07-1.4]	1.94 [1.52-2.48]	2.69 E-08	1.34 [1.21-1.49]	1.25 [1.08-1.44]	2.03 [1.57-2.62]	Imputed
rs1864248	C	0.19	1.25E-08	1.35 [1.22-1.5]	1.22 [1.07-1.4]	1.94 [1.52-2.48]	2.49 E-08	1.34 [1.21-1.49]	1.25 [1.09-1.44]	2.03 [1.57-2.62]	Imputed
rs2562873	T	0.19	1.29E-08	1.35 [1.22-1.5]	1.22 [1.06-1.39]	1.97 [1.54-2.52]	1.94 E-08	1.35 [1.22-1.5]	1.25 [1.08-1.44]	2.08 [1.61-2.69]	Imputed
rs1429112	G	0.19	1.34E-08	1.35 [1.22-1.5]	1.21 [1.06-1.39]	1.98 [1.54-2.54]	2.58 E-08	1.35 [1.21-1.5]	1.24 [1.07-1.43]	2.09 [1.61-2.71]	Imputed
rs2562871	T	0.19	1.39E-08	1.35 [1.22-1.5]	1.22 [1.06-1.39]	1.97 [1.53-2.52]	2.07 E-08	1.35 [1.22-1.5]	1.24 [1.08-1.44]	2.08 [1.6-2.69]	Imputed
rs2435095	A	0.19	1.53E-08	1.35 [1.22-1.5]	1.21 [1.06-1.39]	1.97 [1.53-2.52]	2.21 E-08	1.35 [1.22-1.5]	1.24 [1.08-1.43]	2.08 [1.6-2.69]	Imputed
rs28623525	C	0.18	1.60E-08	1.36 [1.22-1.51]	1.32 [1.16-1.51]	1.81 [1.37-2.41]	9.02 E-09	1.38 [1.23-1.53]	1.37 [1.19-1.57]	1.92 [1.43-2.58]	Imputed
rs2562875	T	0.19	1.62E-08	1.35 [1.22-1.5]	1.21 [1.06-1.39]	1.97 [1.54-2.52]	2.66 E-08	1.35 [1.21-1.5]	1.24 [1.07-1.43]	2.08 [1.61-2.69]	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
rs1346598	T	0.19	1.68E-08	1.36 [1.22-1.51]	1.2 [1.04-1.38]	2.01 [1.57-2.58]	3.43 E-08	1.35 [1.21-1.5]	1.23 [1.06-1.42]	2.1 [1.62-2.73]	Imputed
rs2562877	T	0.19	2.04E-08	1.35 [1.22-1.5]	1.21 [1.05-1.38]	1.97 [1.53-2.52]	3.87 E-08	1.34 [1.21-1.49]	1.23 [1.06-1.42]	2.08 [1.6-2.69]	Imputed
rs2562878	G	0.19	2.04E-08	1.35 [1.22-1.5]	1.21 [1.05-1.38]	1.97 [1.53-2.52]	3.87 E-08	1.34 [1.21-1.49]	1.23 [1.06-1.42]	2.08 [1.6-2.69]	Imputed
rs11737107	G	0.19	2.06E-08	1.35 [1.21-1.5]	1.21 [1.06-1.39]	1.95 [1.52-2.5]	3.48 E-08	1.34 [1.21-1.49]	1.24 [1.07-1.43]	2.06 [1.59-2.67]	Imputed
rs2059904	G	0.18	2.08E-08	1.35 [1.21-1.49]	1.31 [1.15-1.49]	1.85 [1.39-2.45]	1.81 E-08	1.36 [1.22-1.52]	1.34 [1.17-1.54]	1.93 [1.43-2.59]	Genotyped
rs2562870	C	0.19	2.18E-08	1.35 [1.21-1.5]	1.21 [1.06-1.39]	1.96 [1.53-2.51]	3.13 E-08	1.35 [1.21-1.49]	1.24 [1.07-1.43]	2.07 [1.59-2.68]	Imputed
rs1864247	C	0.20	2.30E-08	1.35 [1.21-1.5]	1.21 [1.05-1.39]	1.94 [1.52-2.48]	5.38 E-08	1.34 [1.2-1.49]	1.24 [1.07-1.43]	2.02 [1.56-2.61]	Imputed
rs2435094	C	0.20	2.38E-08	1.35 [1.21-1.5]	1.2 [1.05-1.38]	1.96 [1.53-2.51]	5.82 E-08	1.34 [1.2-1.49]	1.23 [1.06-1.42]	2.05 [1.58-2.65]	Imputed
rs934146	C	0.19	2.57E-08	1.35 [1.22-1.51]	1.19 [1.04-1.37]	2.01 [1.56-2.58]	4.91 E-08	1.35 [1.21-1.5]	1.21 [1.05-1.4]	2.14 [1.64-2.78]	Imputed
rs1560226	C	0.19	2.67E-08	1.35 [1.21-1.5]	1.19 [1.04-	2 [1.56-2.57]	4.63 E-08	1.34 [1.21-1.49]	1.21 [1.05-1.4]	2.12 [1.63-	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
					1.36]					2.74]	
rs2714900	T	0.20	2.74E-08	1.35 [1.21-1.5]	1.21 [1.05-1.38]	1.94 [1.52-2.48]	6.73E-08	1.34 [1.2-1.48]	1.23 [1.07-1.42]	2.02 [1.56-2.61]	Imputed
rs9654228	T	0.17	3.43E-08	1.35 [1.22-1.51]	1.3 [1.14-1.49]	1.84 [1.38-2.44]	3.27E-08	1.36 [1.22-1.52]	1.34 [1.17-1.55]	1.92 [1.43-2.59]	Imputed
rs2562879	G	0.19	3.52E-08	1.35 [1.21-1.5]	1.2 [1.04-1.37]	1.97 [1.54-2.52]	6.87E-08	1.34 [1.2-1.49]	1.22 [1.05-1.41]	2.08 [1.6-2.69]	Imputed
rs2714905	A	0.19	3.79E-08	1.34 [1.21-1.49]	1.21 [1.06-1.39]	1.93 [1.51-2.48]	6.75E-08	1.34 [1.2-1.49]	1.23 [1.07-1.42]	2.04 [1.57-2.65]	Imputed
rs2562880	C	0.19	3.82E-08	1.35 [1.21-1.5]	1.2 [1.04-1.38]	1.96 [1.52-2.51]	5.31E-08	1.34 [1.21-1.49]	1.22 [1.06-1.42]	2.08 [1.61-2.7]	Imputed
rs1429141	C	0.17	3.96E-08	1.35 [1.21-1.5]	1.3 [1.14-1.49]	1.83 [1.38-2.43]	3.32E-08	1.36 [1.22-1.52]	1.35 [1.17-1.55]	1.91 [1.42-2.57]	Imputed
rs2562876	G	0.19	4.07E-08	1.35 [1.21-1.5]	1.19 [1.04-1.37]	1.96 [1.53-2.51]	7.10E-08	1.34 [1.2-1.49]	1.22 [1.05-1.41]	2.07 [1.6-2.68]	Imputed
rs2714901	T	0.19	4.21E-08	1.34 [1.21-1.49]	1.19 [1.04-1.37]	1.97 [1.53-2.52]	6.70E-08	1.34 [1.2-1.49]	1.22 [1.05-1.41]	2.07 [1.6-2.69]	Imputed
rs2562869	G	0.20	4.23E-08	1.34 [1.21-1.49]	1.18 [1.02-1.35]	1.99 [1.56-2.55]	7.07E-08	1.34 [1.2-1.49]	1.21 [1.05-1.4]	2.07 [1.6-2.68]	Imputed
rs6812819	A	0.1	4.24E-	1.35	1.3 [1.13-	1.83	3.27	1.36 [1.22-	1.34	1.92	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
		7	08	[1.21-1.51]	1.49]	[1.38-2.44]	E-08	1.52]	[1.17-1.55]	[1.43-2.58]	
rs1975060	T	0.19	4.96E-08	1.34 [1.21-1.5]	1.19 [1.04-1.37]	1.99 [1.54-2.56]	6.93E-08	1.34 [1.21-1.49]	1.21 [1.05-1.4]	2.11 [1.62-2.75]	Imputed
rs7667633	C	0.15	5.05E-08	1.38 [1.23-1.54]	1.32 [1.15-1.52]	1.9 [1.39-2.59]	1.79E-08	1.4 [1.24-1.57]	1.36 [1.18-1.58]	2.07 [1.5-2.87]	Imputed
rs2562874	C	0.20	5.31E-08	1.34 [1.21-1.49]	1.18 [1.03-1.35]	1.98 [1.55-2.53]	1.12E-07	1.33 [1.2-1.48]	1.21 [1.04-1.4]	2.06 [1.59-2.66]	Imputed
rs1816280	A	0.19	5.54E-08	1.34 [1.21-1.49]	1.19 [1.04-1.36]	1.97 [1.53-2.53]	1.08E-07	1.33 [1.2-1.48]	1.21 [1.05-1.4]	2.08 [1.6-2.7]	Imputed
rs2562882	C	0.19	5.99E-08	1.34 [1.21-1.49]	1.19 [1.03-1.36]	1.97 [1.53-2.53]	1.02E-07	1.33 [1.2-1.48]	1.21 [1.05-1.4]	2.08 [1.6-2.7]	Imputed
rs2303839	A	0.19	7.49E-08	1.34 [1.2-1.49]	1.19 [1.03-1.36]	1.96 [1.53-2.53]	1.46E-07	1.33 [1.2-1.48]	1.21 [1.05-1.4]	2.05 [1.58-2.67]	Imputed
rs17023141	A	0.16	1.09E-07	1.36 [1.21-1.52]	1.31 [1.14-1.51]	1.8 [1.33-2.44]	4.81E-08	1.38 [1.23-1.54]	1.36 [1.17-1.57]	1.96 [1.43-2.69]	Imputed
rs1594082	G	0.18	1.14E-07	1.32 [1.19-1.47]	1.34 [1.17-1.52]	1.65 [1.23-2.21]	3.21E-08	1.36 [1.22-1.51]	1.38 [1.21-1.58]	1.75 [1.29-2.37]	Genotyped
rs13147231	G	0.26	1.16E-07	1.29 [1.18-1.42]	1.23 [1.08-1.4]	1.7 [1.37-2.11]	9.12E-07	1.28 [1.16-1.41]	1.25 [1.09-1.43]	1.68 [1.34-2.11]	Imputed

Marker	MA	MAF	Corr/Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
rs6836525	T	0.14	1.41E-07	1.37 [1.22-1.54]	1.34 [1.16-1.54]	1.8 [1.3-2.5]	1.02 E-07	1.38 [1.23-1.56]	1.38 [1.18-1.6]	1.94 [1.38-2.72]	Imputed
rs6836670	C	0.18	1.41E-07	1.32 [1.19-1.47]	1.29 [1.13-1.47]	1.79 [1.34-2.38]	2.69 E-08	1.36 [1.22-1.51]	1.35 [1.18-1.55]	1.88 [1.4-2.54]	Genotyped
rs4593108	G	0.15	1.43E-07	1.37 [1.22-1.54]	1.32 [1.14-1.52]	1.87 [1.36-2.58]	7.64 E-08	1.39 [1.23-1.56]	1.36 [1.17-1.58]	2.04 [1.46-2.86]	Imputed
rs11728738	C	0.26	1.44E-07	1.29 [1.17-1.42]	1.23 [1.08-1.4]	1.69 [1.36-2.1]	1.10 E-06	1.28 [1.16-1.41]	1.25 [1.09-1.43]	1.68 [1.33-2.11]	Imputed
rs6836562	T	0.14	1.91E-07	1.37 [1.22-1.54]	1.33 [1.15-1.54]	1.8 [1.3-2.5]	1.42 E-07	1.38 [1.22-1.56]	1.37 [1.18-1.59]	1.94 [1.38-2.72]	Imputed
rs1429137	T	0.15	2.06E-07	1.36 [1.21-1.53]	1.3 [1.13-1.5]	1.9 [1.38-2.62]	9.96 E-08	1.38 [1.23-1.56]	1.34 [1.15-1.55]	2.08 [1.49-2.9]	Imputed
rs6812432	G	0.26	2.63E-07	1.28 [1.17-1.41]	1.22 [1.07-1.39]	1.69 [1.36-2.1]	1.56 E-06	1.27 [1.15-1.4]	1.24 [1.08-1.42]	1.68 [1.33-2.11]	Imputed
rs11100960	A	0.26	2.79E-07	1.28 [1.17-1.41]	1.21 [1.07-1.38]	1.69 [1.36-2.09]	2.39 E-06	1.27 [1.15-1.4]	1.23 [1.07-1.41]	1.67 [1.33-2.1]	Imputed
rs2357778	G	0.26	2.99E-07	1.28 [1.17-1.41]	1.21 [1.06-1.38]	1.69 [1.36-2.09]	2.31 E-06	1.27 [1.15-1.4]	1.23 [1.07-1.41]	1.67 [1.33-2.1]	Imputed
rs2357779	T	0.26	2.99E-07	1.28 [1.17-	1.21 [1.06-	1.69 [1.36-	2.31 E-06	1.27 [1.15-1.4]	1.23 [1.07-	1.67 [1.33-	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
				1.41]	1.38]	2.09]			1.41]	2.1]	
rs11726718	C	0.26	5.93E-07	1.27 [1.16-1.4]	1.21 [1.06-1.37]	1.67 [1.35-2.08]	3.12 E-06	1.26 [1.15-1.39]	1.22 [1.07-1.4]	1.66 [1.32-2.1]	Imputed
rs7671190	C	0.26	6.09E-07	1.27 [1.16-1.4]	1.21 [1.06-1.37]	1.67 [1.34-2.07]	3.06 E-06	1.26 [1.15-1.39]	1.22 [1.07-1.4]	1.66 [1.32-2.09]	Imputed
rs1429106	C	0.26	6.39E-07	1.27 [1.16-1.4]	1.21 [1.06-1.37]	1.67 [1.34-2.07]	3.21 E-06	1.26 [1.14-1.39]	1.22 [1.07-1.4]	1.66 [1.32-2.09]	Imputed
rs1429105	C	0.26	6.96E-07	1.27 [1.16-1.4]	1.2 [1.06-1.37]	1.67 [1.34-2.07]	3.49 E-06	1.26 [1.14-1.39]	1.22 [1.07-1.4]	1.66 [1.32-2.09]	Imputed
rs1346594	T	0.26	8.34E-07	1.26 [1.15-1.39]	1.22 [1.07-1.38]	1.63 [1.32-2.03]	3.51 E-06	1.26 [1.14-1.39]	1.23 [1.08-1.4]	1.64 [1.3-2.06]	Imputed
rs13105529	C	0.26	9.59E-07	1.27 [1.15-1.39]	1.2 [1.06-1.37]	1.66 [1.33-2.06]	5.47 E-06	1.26 [1.14-1.39]	1.22 [1.06-1.39]	1.64 [1.31-2.07]	Imputed
rs1346595	G	0.26	1.01E-06	1.26 [1.15-1.39]	1.23 [1.08-1.39]	1.62 [1.3-2.02]	3.31 E-06	1.26 [1.14-1.39]	1.24 [1.08-1.41]	1.63 [1.29-2.06]	Genotyped
rs28612496	A	0.17	1.08E-06	1.32 [1.18-1.47]	1.3 [1.14-1.5]	1.62 [1.21-2.18]	1.01 E-06	1.32 [1.18-1.48]	1.34 [1.16-1.55]	1.7 [1.24-2.31]	Imputed
rs28406843	T	0.17	1.78E-06	1.31 [1.17-1.46]	1.3 [1.13-1.48]	1.61 [1.2-2.16]	1.18 E-06	1.32 [1.18-1.47]	1.34 [1.16-1.54]	1.68 [1.23-2.29]	Imputed
rs1429100	A	0.1	1.84E-	1.32	1.3 [1.13-	1.65	1.32	1.33 [1.19-	1.34	1.72	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
		6	06	[1.18-1.48]	1.49]	[1.21-2.24]	E-06	1.49]	[1.16-1.56]	[1.25-2.38]	
rs1579452	A	0.16	1.97E-06	1.31 [1.17-1.47]	1.31 [1.14-1.51]	1.59 [1.17-2.15]	1.82E-06	1.32 [1.18-1.48]	1.35 [1.17-1.56]	1.65 [1.2-2.27]	Imputed
rs72953535	A	0.22	2.26E-06	1.27 [1.15-1.41]	1.28 [1.12-1.45]	1.57 [1.22-2.03]	1.17E-05	1.27 [1.14-1.41]	1.28 [1.11-1.47]	1.57 [1.2-2.06]	Imputed
rs7668383	C	0.16	2.29E-06	1.32 [1.18-1.48]	1.27 [1.1-1.47]	1.78 [1.3-2.43]	6.71E-07	1.35 [1.2-1.52]	1.32 [1.14-1.54]	1.91 [1.37-2.65]	Imputed
rs2217348	T	0.17	2.85E-06	1.3 [1.16-1.45]	1.29 [1.13-1.47]	1.6 [1.19-2.14]	1.80E-06	1.31 [1.17-1.47]	1.33 [1.15-1.53]	1.67 [1.23-2.27]	Imputed
rs28602756	C	0.22	4.29E-06	1.27 [1.15-1.41]	1.27 [1.11-1.45]	1.55 [1.2-2]	1.70E-05	1.26 [1.14-1.4]	1.28 [1.11-1.47]	1.55 [1.18-2.03]	Imputed
rs55771464	C	0.21	1.14E-05	1.26 [1.14-1.4]	1.24 [1.09-1.42]	1.59 [1.22-2.08]	7.82E-05	1.25 [1.12-1.39]	1.24 [1.08-1.43]	1.56 [1.17-2.08]	Imputed
rs1366689	C	0.15	2.17E-05	1.28 [1.14-1.44]	1.22 [1.06-1.4]	1.72 [1.27-2.34]	7.49E-06	1.3 [1.16-1.46]	1.25 [1.08-1.45]	1.88 [1.37-2.59]	Imputed
rs17023182	A	0.21	2.29E-05	1.25 [1.13-1.39]	1.24 [1.09-1.42]	1.56 [1.19-2.04]	1.52E-04	1.24 [1.11-1.38]	1.24 [1.08-1.43]	1.52 [1.14-2.02]	Imputed
rs4399964	A	0.15	3.89E-05	1.27 [1.13-1.42]	1.16 [1.01-1.32]	2.2 [1.56-3.11]	1.48E-05	1.3 [1.15-1.46]	1.18 [1.02-1.36]	2.38 [1.66-3.41]	Genotyped

Marker	MA	MAF	Corr/Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
rs12645918	T	0.21	4.25E-05	1.24 [1.12-1.38]	1.22 [1.07-1.39]	1.57 [1.2-2.06]	2.20E-04	1.23 [1.1-1.37]	1.22 [1.06-1.4]	1.54 [1.15-2.06]	Imputed
rs1429111	T	0.25	4.42E-05	1.23 [1.11-1.36]	1.16 [1.02-1.33]	1.55 [1.24-1.95]	7.08E-05	1.23 [1.11-1.36]	1.19 [1.03-1.38]	1.57 [1.24-2]	Imputed
rs12646693	T	0.21	6.54E-05	1.24 [1.12-1.38]	1.22 [1.07-1.4]	1.52 [1.16-1.99]	1.77E-04	1.23 [1.11-1.38]	1.24 [1.07-1.42]	1.51 [1.14-2.02]	Imputed
rs17023196	T	0.21	1.01E-04	1.23 [1.11-1.37]	1.23 [1.07-1.4]	1.49 [1.13-1.96]	2.27E-04	1.23 [1.1-1.37]	1.24 [1.08-1.43]	1.48 [1.1-1.98]	Imputed
rs58983705	C	0.20	1.36E-04	1.23 [1.11-1.37]	1.21 [1.06-1.39]	1.5 [1.14-1.98]	2.46E-04	1.23 [1.1-1.37]	1.24 [1.07-1.43]	1.5 [1.12-2.01]	Imputed
rs17023204	T	0.20	1.69E-04	1.23 [1.1-1.36]	1.21 [1.06-1.38]	1.49 [1.13-1.96]	3.52E-04	1.22 [1.1-1.37]	1.23 [1.06-1.41]	1.48 [1.11-1.99]	Imputed
rs12640442	C	0.20	1.79E-04	1.22 [1.1-1.36]	1.21 [1.06-1.39]	1.48 [1.13-1.95]	4.27E-04	1.22 [1.09-1.36]	1.23 [1.07-1.42]	1.45 [1.08-1.96]	Imputed
rs57997710	T	0.20	2.76E-04	1.22 [1.1-1.36]	1.2 [1.05-1.37]	1.5 [1.13-1.98]	5.50E-04	1.22 [1.09-1.36]	1.22 [1.06-1.4]	1.48 [1.1-2]	Imputed
rs61379585	G	0.09	4.36E-04	1.29 [1.12-1.49]	1.21 [1.03-1.42]	2.57 [1.46-4.53]	5.50E-04	1.3 [1.12-1.51]	1.23 [1.04-1.45]	2.63 [1.45-4.78]	Imputed
rs2174801	C	0.09	4.79E-04	1.3 [1.12-1.5]	1.21 [1.02-	2.59 [1.47-	5.35E-04	1.31 [1.12-1.52]	1.23 [1.03-	2.64 [1.45-	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
					1.42]	4.56]			1.46]	4.81]	
rs78865503	G	0.09	5.29E-04	1.29 [1.12-1.49]	1.2 [1.03-1.41]	2.57 [1.46-4.53]	6.68E-04	1.3 [1.12-1.51]	1.22 [1.03-1.44]	2.63 [1.45-4.78]	Imputed
rs17611755	T	0.11	8.15E-04	1.24 [1.09-1.41]	1.14 [0.99-1.33]	2.22 [1.44-3.42]	2.58E-03	1.23 [1.07-1.4]	1.14 [0.97-1.33]	2.18 [1.38-3.44]	Genotyped
rs937881	C	0.07	9.34E-04	1.31 [1.11-1.53]	1.21 [1.01-1.44]	2.78 [1.48-5.23]	1.44E-03	1.31 [1.11-1.55]	1.22 [1.02-1.47]	2.83 [1.44-5.58]	Imputed
rs2884222	G	0.12	2.11E-03	1.22 [1.07-1.38]	1.14 [0.99-1.32]	2.08 [1.33-3.24]	4.86E-03	1.21 [1.06-1.38]	1.14 [0.98-1.33]	2.01 [1.26-3.22]	Genotyped
rs7697539	C	0.20	2.44E-03	1.18 [1.06-1.32]	1.15 [1-1.32]	1.45 [1.1-1.93]	4.60E-03	1.18 [1.05-1.32]	1.16 [1-1.34]	1.45 [1.07-1.96]	Imputed
rs1354885	T	0.11	2.49E-03	1.22 [1.07-1.4]	1.13 [0.97-1.31]	2.13 [1.37-3.31]	6.00E-03	1.21 [1.06-1.39]	1.13 [0.96-1.32]	2.08 [1.3-3.31]	Imputed
rs1466985	T	0.08	2.72E-03	1.25 [1.08-1.44]	1.22 [1.04-1.43]	1.78 [0.97-3.26]	7.51E-03	1.23 [1.06-1.43]	1.21 [1.02-1.43]	1.75 [0.93-3.31]	Genotyped
rs62341517	G	0.11	2.73E-03	1.22 [1.07-1.39]	1.13 [0.97-1.31]	2.13 [1.37-3.3]	7.13E-03	1.21 [1.05-1.38]	1.12 [0.96-1.31]	2.07 [1.3-3.3]	Imputed
rs62341515	C	0.11	3.00E-03	1.22 [1.07-1.39]	1.12 [0.97-1.3]	2.13 [1.37-3.3]	8.58E-03	1.2 [1.05-1.37]	1.11 [0.95-1.31]	2.07 [1.3-3.3]	Imputed
rs1429109	C	0.2	3.17E-	1.18	1.15 [1-	1.41	5.08	1.18 [1.05-	1.16 [1-	1.42	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
		0	03	[1.06-1.31]	1.32]	[1.07-1.88]	E-03	1.32]	1.34]	[1.05-1.93]	
rs9942219	A	0.20	3.41E-03	1.18 [1.06-1.31]	1.15 [1.01-1.32]	1.41 [1.06-1.88]	5.34 E-03	1.18 [1.05-1.32]	1.16 [1.01-1.35]	1.41 [1.04-1.92]	Imputed
rs1429110	T	0.20	3.51E-03	1.17 [1.05-1.31]	1.16 [1.01-1.33]	1.36 [1.03-1.8]	4.10 E-03	1.18 [1.05-1.32]	1.18 [1.02-1.36]	1.38 [1.03-1.86]	Imputed
rs937880	A	0.08	3.79E-03	1.25 [1.07-1.45]	1.11 [0.93-1.31]	2.65 [1.58-4.43]	7.42 E-03	1.23 [1.06-1.44]	1.11 [0.93-1.33]	2.66 [1.54-4.58]	Imputed
rs75311641	C	0.11	3.99E-03	1.21 [1.06-1.38]	1.11 [0.96-1.29]	2.12 [1.37-3.3]	1.06 E-02	1.2 [1.04-1.37]	1.11 [0.94-1.3]	2.07 [1.3-3.29]	Imputed
rs11938488	G	0.08	4.15E-03	1.25 [1.07-1.45]	1.11 [0.94-1.32]	2.55 [1.52-4.29]	7.35 E-03	1.24 [1.06-1.45]	1.12 [0.94-1.35]	2.54 [1.46-4.41]	Imputed
rs11947583	A	0.11	4.28E-03	1.21 [1.06-1.38]	1.11 [0.95-1.29]	2.14 [1.38-3.33]	9.58 E-03	1.2 [1.05-1.38]	1.11 [0.94-1.3]	2.08 [1.31-3.32]	Imputed
rs17022714	T	0.11	4.40E-03	1.21 [1.06-1.38]	1.11 [0.95-1.29]	2.14 [1.38-3.33]	9.96 E-03	1.2 [1.04-1.38]	1.11 [0.94-1.3]	2.08 [1.31-3.32]	Imputed
rs113128727	C	0.11	5.30E-03	1.2 [1.06-1.37]	1.1 [0.95-1.28]	2.12 [1.37-3.3]	1.44 E-02	1.19 [1.03-1.36]	1.1 [0.93-1.29]	2.06 [1.29-3.28]	Imputed
rs59834135	C	0.11	5.62E-03	1.21 [1.06-1.38]	1.1 [0.94-1.28]	2.13 [1.37-3.31]	1.43 E-02	1.19 [1.04-1.37]	1.1 [0.93-1.29]	2.07 [1.3-3.29]	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
rs62341519	C	0.11	5.76E-03	1.2 [1.06-1.38]	1.12 [0.96-1.3]	2.02 [1.29-3.16]	1.65 E-02	1.18 [1.03-1.36]	1.11 [0.94-1.3]	1.94 [1.21-3.12]	Imputed
rs62341516	C	0.11	5.99E-03	1.2 [1.05-1.37]	1.1 [0.95-1.28]	2.12 [1.36-3.29]	1.60 E-02	1.18 [1.03-1.36]	1.09 [0.93-1.28]	2.06 [1.29-3.28]	Imputed
rs58378349	G	0.11	6.50E-03	1.2 [1.05-1.37]	1.1 [0.94-1.28]	2.12 [1.37-3.3]	1.51 E-02	1.19 [1.03-1.36]	1.09 [0.93-1.29]	2.06 [1.29-3.29]	Imputed
rs73853371	A	0.11	6.57E-03	1.2 [1.05-1.37]	1.1 [0.94-1.28]	2.13 [1.37-3.32]	1.54 E-02	1.19 [1.03-1.36]	1.09 [0.93-1.28]	2.07 [1.3-3.3]	Imputed
rs1828034	A	0.13	8.33E-03	1.18 [1.04-1.34]	1.11 [0.96-1.28]	1.83 [1.22-2.74]	6.04 E-03	1.2 [1.05-1.36]	1.13 [0.97-1.31]	1.88 [1.23-2.88]	Imputed
rs80077485	T	0.05	9.12E-03	0.75 [0.61-0.93]	0.74 [0.59-0.93]	0.76 [0.28-2.04]	9.09 E-03	0.74 [0.6-0.93]	0.71 [0.56-0.91]	0.83 [0.31-2.23]	Imputed
rs6822565	C	0.25	1.03E-02	1.13 [1.03-1.24]	1.14 [1.01-1.29]	1.26 [0.99-1.59]	1.23 E-02	1.14 [1.03-1.25]	1.17 [1.02-1.33]	1.23 [0.96-1.58]	Genotyped
rs6812093	T	0.25	1.18E-02	1.13 [1.03-1.24]	1.14 [1.01-1.29]	1.24 [0.98-1.58]	1.32 E-02	1.13 [1.03-1.25]	1.17 [1.02-1.33]	1.22 [0.95-1.57]	Imputed
rs2357604	C	0.11	1.22E-02	1.19 [1.04-1.36]	1.09 [0.94-1.27]	2.02 [1.29-3.16]	3.04 E-02	1.17 [1.02-1.34]	1.09 [0.92-1.28]	1.93 [1.2-3.11]	Imputed
rs1568136	T	0.25	1.51E-02	1.13 [1.02-	1.14 [1.01-	1.23 [0.97-	1.76 E-02	1.13 [1.02-1.25]	1.16 [1.02-	1.21 [0.94-	Imputed

Marker	MA	MAF	Corr/ Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
				1.24]	1.29]	1.56]			1.33]	1.56]	
rs1354884	A	0.11	1.74E-02	1.18 [1.03-1.34]	1.08 [0.92-1.26]	2.02 [1.29-3.16]	3.10 E-02	1.17 [1.01-1.34]	1.08 [0.92-1.27]	1.94 [1.2-3.12]	Imputed
rs7674137	G	0.19	2.31E-02	1.14 [1.02-1.28]	1.13 [0.98-1.3]	1.31 [0.96-1.79]	1.76 E-02	1.15 [1.03-1.29]	1.17 [1.01-1.35]	1.29 [0.93-1.78]	Imputed
rs1878404	A	0.18	2.51E-02	1.14 [1.02-1.28]	1.13 [0.98-1.3]	1.3 [0.95-1.78]	1.50 E-02	1.16 [1.03-1.3]	1.17 [1-1.35]	1.32 [0.95-1.84]	Imputed
rs9647489	C	0.11	2.57E-02	1.17 [1.02-1.34]	1.03 [0.87-1.21]	2.2 [1.43-3.38]	1.05 E-02	1.21 [1.05-1.39]	1.05 [0.88-1.25]	2.45 [1.57-3.83]	Imputed
rs6537487	C	0.19	3.15E-02	1.13 [1.01-1.27]	1.12 [0.98-1.29]	1.28 [0.94-1.74]	2.75 E-02	1.14 [1.02-1.28]	1.16 [1-1.34]	1.25 [0.9-1.73]	Imputed
rs6821368	T	0.19	3.62E-02	1.13 [1.01-1.26]	1.12 [0.97-1.28]	1.28 [0.94-1.73]	2.97 E-02	1.14 [1.01-1.28]	1.15 [1-1.33]	1.25 [0.91-1.73]	Imputed
rs1568137	C	0.09	3.82E-02	0.85 [0.73-0.99]	0.78 [0.65-0.92]	1.37 [0.79-2.39]	8.37 E-02	0.87 [0.74-1.02]	0.78 [0.66-0.94]	1.43 [0.81-2.53]	Imputed
rs369660577	C	0.08	3.84E-02	0.85 [0.72-0.99]	0.79 [0.66-0.94]	1.25 [0.68-2.31]	5.18 E-02	0.85 [0.72-1]	0.78 [0.65-0.94]	1.26 [0.67-2.36]	Imputed
rs376650129	A	0.08	3.84E-02	0.85 [0.72-0.99]	0.79 [0.66-0.94]	1.25 [0.68-2.31]	5.18 E-02	0.85 [0.72-1]	0.78 [0.65-0.94]	1.26 [0.67-2.36]	Imputed
rs13132657	A	0.0	4.24E-	0.86	0.77	1.37	7.50	0.87 [0.74-	0.78	1.37	Imputed

Marker	MA	MAF	Corr/Trend P-value	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	P Adjusted	OR [95% CI] Allelic	OR [95% CI] Heterozygote	OR [95% CI] Minor Homozygote	Genotyping
		9	02	[0.74-0.99]	[0.65-0.92]	[0.81-2.33]	E-02	1.01]	[0.66-0.94]	[0.79-2.37]	

*MA-Minor Allele, MAF - Minor Allele Frequency

This table represents the association of the fine-mapped SNPs with premenopausal breast cancer risk. The table represents 135 SNPs associated (p-value <0.05). The p-value and odds ratio with 95% confidence interval was estimated in both unadjusted and adjusted analysis. The analysis was adjusted for BMI. The unadjusted p-value was estimated using correlation trend test and p-value for the adjusted analysis was estimated using binary logistic regression. Odds ratio was estimated by assuming allelic and genotypic model.

Table A.3 List of tag SNPs genotyped from finemapped locus 4q31.22

Marker 1	Imputation and Genotyping concordance (r2)	Call Rate	Minor Allele	Major Allele	Minor Allele Frequency	Major Allele Frequency
rs1429142	0.99	1.00	C	T	0.18	0.82
rs2195469	0.93	1.00	T	C	0.04	0.96
rs28722867	0.98	1.00	T	C	0.02	0.98
rs1594082	0.97	1.00	G	T	0.18	0.82
rs28645698	0.96	0.99	C	T	0.19	0.81
rs6537474	0.94	0.99	T	C	0.05	0.95
rs6817192	0.89	0.99	A	G	0.05	0.95
rs73855101	0.88	0.99	T	C	0.04	0.96
rs2357607	0.97	0.99	T	A	0.06	0.94
rs1567180	0.96	0.99	C	T	0.05	0.95
rs11737828	0.95	0.99	A	G	0.01	0.99
rs2884222	0.92	0.99	C	T	0.12	0.88
rs1981987	0.89	0.99	G	C	0.04	0.96
rs4399964	0.95	0.99	A	G	0.15	0.85
rs11100939	0.92	0.99	C	T	0.50	0.50
rs17022364	0.85	0.99	A	T	0.05	0.95
rs1429133	0.93	0.99	G	A	0.21	0.79
rs17022379	0.96	0.99	G	A	0.02	0.98
rs11722693	0.93	0.99	C	T	0.41	0.59
rs9917863	0.91	0.99	C	G	0.13	0.87
rs10034043	0.92	0.99	A	C	0.40	0.60
rs72958286	0.80	0.99	A	G	0.02	0.98
rs1466985	0.94	0.99	T	A	0.09	0.91
rs4835072	0.85	0.99	C	A	0.05	0.95
rs2043702	0.93	0.99	C	G	0.38	0.62
rs7675774	0.92	0.99	G	A	0.42	0.58
rs4835370	0.94	0.99	C	T	0.16	0.84
rs10028838	0.95	0.99	G	T	0.40	0.60
rs78335024	0.93	0.99	G	C	0.05	0.95
rs1429130	0.90	0.99	C	T	0.05	0.95
rs2118258	0.92	0.99	T	C	0.15	0.85
rs6810798	0.97	0.99	A	G	0.19	0.81

Marker 1	Imputation and Genotyping concordance (r2)	Call Rate	Minor Allele	Major Allele	Minor Allele Frequency	Major Allele Frequency
rs4835408	0.89	0.99	A	G	0.11	0.89
rs17611755	0.92	0.99	T	C	0.11	0.89
rs4835362	0.94	0.99	A	G	0.18	0.82
rs150873193	0.86	0.99	A	G	0.03	0.97
rs143682942	0.86	0.99	T	G	0.04	0.96
rs17022600	1.00	0.99	G	A	0.01	0.99
rs4835084	0.93	0.99	A	T	0.30	0.70
rs1507500	0.86	0.99	C	T	0.11	0.89
rs1583003	0.96	0.99	T	C	0.23	0.77
rs6822565	0.95	0.99	C	T	0.25	0.75
rs28720373	0.96	0.99	T	C	0.19	0.81
rs6537450	0.93	0.99	A	G	0.40	0.60
rs2059904	0.96	0.99	C	T	0.18	0.82
rs11731096	0.91	0.99	G	A	0.29	0.71
rs9307838	0.95	0.98	G	A	0.25	0.75
rs7699439	0.89	0.98	T	C	0.29	0.71
rs6836670	0.96	0.98	G	A	0.18	0.82
rs4835456	0.92	0.98	A	T	0.14	0.86
rs10519886	0.93	0.97	A	G	0.24	0.76
rs4110	0.95	0.97	A	G	0.33	0.67
rs1346595	0.89	0.97	G	A	0.26	0.74
rs1429116	0.91	0.94	C	T	0.35	0.65
rs13134510	0.93	0.93	C	T	0.23	0.77
rs7669311	0.82	0.92	T	C	0.27	0.73

This table includes the 57 Tag SNPs selected from the Finemapped region which are imputed and genotyped in stage 1-4 samples. The Concordance is calculated between imputation and genotyping of stage 1 samples. All the SNPs had concordance r2>0.80. The callrate and allele frequencies are estimated based on the stage 1-4 samples.

Table A.4 Conditional Regression analysis

SNP	P conditioned rs1366691	P conditioned rs1429139	P conditioned rs12501429	P conditioned rs13134510
rs11735996	6.49E-02	6.33E-02	2.00E-01	1.82E-01
rs2217348	8.71E-02	1.28E-01	7.54E-02	5.11E-02
rs80077485	8.72E-02	9.69E-02	9.73E-02	7.04E-02
rs1366679	9.80E-02	9.05E-02	2.76E-01	1.70E-01
rs12511935	1.02E-01	9.48E-02	2.89E-01	1.81E-01
rs1579452	1.06E-01	1.26E-01	9.24E-02	8.32E-02
rs2562880	1.08E-01	8.18E-02	5.59E-02	1.97E-01
rs2562879	1.19E-01	8.45E-02	5.84E-02	1.85E-01
rs2562876	1.22E-01	8.60E-02	6.75E-02	1.87E-01
rs1429100	1.30E-01	1.53E-01	1.19E-01	7.94E-02
rs28612496	1.31E-01	1.71E-01	9.78E-02	1.18E-01
rs12500103	1.31E-01	1.17E-01	3.40E-01	2.07E-01
rs28406843	1.31E-01	1.71E-01	1.04E-01	1.09E-01
rs2174801	1.63E-01	1.32E-01	1.27E-01	2.32E-01
rs1583003	1.69E-01	1.49E-01	5.82E-01	1.38E-01
rs13120678	1.76E-01	8.72E-02	1.67E-01	9.89E-02
rs2714900	1.76E-01	1.14E-01	4.53E-02	1.95E-01
rs2714901	1.80E-01	1.18E-01	7.70E-02	1.69E-01
rs1878404	1.83E-01	2.53E-01	2.23E-01	2.02E-01
rs1594082	1.88E-01	2.54E-01	2.34E-01	1.18E-01
rs7674137	2.01E-01	3.05E-01	2.49E-01	2.14E-01
rs1864247	2.09E-01	1.09E-01	5.77E-02	2.18E-01
rs72953535	2.19E-01	3.22E-01	1.60E-01	4.56E-01
rs6537487	2.21E-01	3.54E-01	2.59E-01	2.53E-01
rs55771464	2.24E-01	3.16E-01	1.38E-01	5.35E-01
rs6821368	2.25E-01	3.38E-01	2.91E-01	2.80E-01
rs61379585	2.37E-01	1.96E-01	2.01E-01	2.54E-01
rs9647489	2.41E-01	2.03E-01	1.76E-01	3.14E-01
rs78865503	2.45E-01	2.09E-01	2.09E-01	2.79E-01
rs2435094	2.50E-01	1.65E-01	7.30E-02	2.52E-01
rs6812432	2.63E-01	4.83E-01	2.09E-01	4.08E-01
rs17023182	2.67E-01	3.48E-01	1.70E-01	6.14E-01
rs28602756	2.75E-01	3.87E-01	2.17E-01	3.76E-01
rs937881	2.78E-01	2.13E-01	2.31E-01	2.39E-01
rs2562874	3.13E-01	2.66E-01	1.78E-01	3.93E-01
rs1828034	3.21E-01	2.29E-01	1.75E-01	5.21E-01
rs1429106	3.33E-01	5.56E-01	2.76E-01	5.07E-01

SNP	P conditioned rs1366691	P conditioned rs1429139	P conditioned rs12501429	P conditioned rs13134510
rs7671190	3.34E-01	5.57E-01	2.76E-01	5.24E-01
rs1429105	3.35E-01	5.58E-01	2.85E-01	5.33E-01
rs11726718	3.39E-01	5.54E-01	2.53E-01	5.23E-01
rs1429141	3.44E-01	4.49E-01	3.07E-01	3.20E-01
rs11100960	3.50E-01	5.74E-01	2.69E-01	4.60E-01
rs13147231	3.55E-01	5.55E-01	2.44E-01	4.87E-01
rs6812819	3.56E-01	4.16E-01	3.07E-01	3.11E-01
rs11728738	3.58E-01	5.55E-01	2.63E-01	4.38E-01
rs2357778	3.65E-01	5.63E-01	2.81E-01	4.65E-01
rs2357779	3.65E-01	5.63E-01	2.81E-01	4.65E-01
rs2059904	3.68E-01	4.45E-01	2.52E-01	3.40E-01
rs13105529	3.74E-01	6.40E-01	2.97E-01	5.86E-01
rs12645918	3.93E-01	5.11E-01	2.75E-01	7.37E-01
rs1366689	3.93E-01	4.56E-01	2.37E-01	4.82E-01
rs9654228	4.09E-01	4.99E-01	3.40E-01	3.82E-01
rs1568137	4.15E-01	4.47E-01	3.61E-01	2.52E-01
rs2562869	4.19E-01	2.96E-01	2.61E-01	4.59E-01
rs369660577	4.21E-01	4.62E-01	3.35E-01	2.26E-01
rs376650129	4.21E-01	4.62E-01	3.35E-01	2.26E-01
rs10519886	4.25E-01	9.44E-01	8.00E-01	7.11E-01
rs7668383	4.27E-01	3.80E-01	3.60E-01	2.04E-01
rs1346595	4.47E-01	6.39E-01	2.87E-01	6.77E-01
rs2435095	4.50E-01	3.91E-01	2.61E-01	6.79E-01
rs11737107	4.61E-01	4.03E-01	2.88E-01	6.44E-01
rs12646693	4.77E-01	5.74E-01	4.21E-01	6.40E-01
rs1346594	4.78E-01	7.29E-01	3.27E-01	8.09E-01
rs28623525	4.81E-01	5.80E-01	3.34E-01	4.06E-01
rs2562871	4.89E-01	3.82E-01	2.75E-01	6.37E-01
rs2562873	4.91E-01	4.23E-01	2.75E-01	7.22E-01
rs1346598	5.03E-01	3.04E-01	2.03E-01	6.52E-01
rs11938488	5.10E-01	3.93E-01	4.75E-01	4.89E-01
rs2714905	5.11E-01	4.29E-01	3.21E-01	6.40E-01
rs17611755	5.12E-01	5.75E-01	5.33E-01	6.92E-01
rs13132657	5.15E-01	5.06E-01	4.27E-01	2.68E-01
rs17023141	5.16E-01	5.74E-01	3.88E-01	3.85E-01
rs57997710	5.24E-01	6.69E-01	3.66E-01	9.91E-01
rs1429112	5.43E-01	4.61E-01	3.49E-01	6.86E-01
rs1568136	5.44E-01	7.35E-01	6.90E-01	4.76E-01
rs17023196	5.44E-01	6.66E-01	4.03E-01	8.26E-01

SNP	P conditioned rs1366691	P conditioned rs1429139	P conditioned rs12501429	P conditioned rs13134510
rs2562882	5.46E-01	4.23E-01	2.55E-01	5.97E-01
rs6836670	5.46E-01	6.52E-01	5.49E-01	3.95E-01
rs2562875	5.49E-01	4.80E-01	3.36E-01	7.05E-01
rs6836562	5.49E-01	5.92E-01	4.25E-01	4.34E-01
rs1816280	5.50E-01	4.26E-01	2.57E-01	6.01E-01
rs6812093	5.56E-01	7.59E-01	7.21E-01	4.27E-01
rs58983705	5.64E-01	6.87E-01	4.23E-01	8.64E-01
rs2562870	5.72E-01	4.45E-01	3.43E-01	7.29E-01
rs6836525	5.74E-01	6.19E-01	4.32E-01	4.43E-01
rs17023204	5.79E-01	7.04E-01	4.55E-01	8.88E-01
rs6822565	5.83E-01	8.17E-01	7.33E-01	4.17E-01
rs1429111	5.86E-01	7.45E-01	5.70E-01	7.93E-01
rs2562877	5.97E-01	5.07E-01	3.95E-01	7.07E-01
rs2562878	5.97E-01	5.07E-01	3.95E-01	7.07E-01
rs937880	5.99E-01	4.90E-01	5.63E-01	5.41E-01
rs28645698	6.06E-01	7.20E-01	6.66E-01	6.21E-01
rs4593108	6.07E-01	6.97E-01	5.17E-01	5.41E-01
rs1975060	6.12E-01	4.74E-01	2.87E-01	6.60E-01
rs12640442	6.13E-01	7.41E-01	4.82E-01	9.49E-01
rs1429137	6.27E-01	7.12E-01	4.94E-01	5.40E-01
rs1560226	6.35E-01	5.01E-01	2.95E-01	6.30E-01
rs28720373	6.35E-01	7.48E-01	6.81E-01	4.17E-01
rs1864248	6.55E-01	4.62E-01	3.09E-01	7.62E-01
rs2884222	6.66E-01	7.00E-01	6.29E-01	9.44E-01
rs17022714	6.83E-01	6.00E-01	6.05E-01	8.56E-01
rs59834135	6.83E-01	5.48E-01	6.38E-01	8.81E-01
rs1354885	6.94E-01	6.10E-01	6.52E-01	7.52E-01
rs1429134	6.94E-01	5.64E-01	3.41E-01	8.58E-01
rs7667633	6.96E-01	7.89E-01	5.08E-01	6.28E-01
rs58378349	6.99E-01	6.15E-01	6.44E-01	9.16E-01
rs11947583	7.05E-01	6.00E-01	6.27E-01	8.54E-01
rs934146	7.13E-01	5.76E-01	3.55E-01	7.28E-01
rs1429109	7.24E-01	8.22E-01	5.89E-01	7.57E-01
rs1346600	7.28E-01	5.22E-01	3.59E-01	8.29E-01
rs9942219	7.39E-01	8.43E-01	6.10E-01	7.64E-01
rs6810798	7.41E-01	9.65E-01	7.51E-01	5.70E-01
rs7697539	7.51E-01	8.49E-01	5.74E-01	7.16E-01
rs75311641	7.53E-01	7.12E-01	6.85E-01	9.52E-01
rs62341515	7.69E-01	7.14E-01	7.21E-01	8.73E-01

SNP	P conditioned rs1366691	P conditioned rs1429139	P conditioned rs12501429	P conditioned rs13134510
rs62341517	7.73E-01	7.20E-01	7.05E-01	8.66E-01
rs73853371	7.75E-01	7.01E-01	6.82E-01	9.67E-01
rs113128727	7.93E-01	7.09E-01	7.30E-01	9.95E-01
rs62341516	8.00E-01	7.15E-01	7.36E-01	9.95E-01
rs1429142	8.30E-01	7.58E-01	7.16E-01	9.56E-01
rs62341519	8.47E-01	7.65E-01	8.59E-01	9.45E-01
rs1466985	8.51E-01	7.09E-01	6.54E-01	8.90E-01
rs2163012	8.57E-01	8.41E-01	5.57E-01	9.13E-01
rs12498595	8.98E-01	7.56E-01	7.76E-01	7.88E-01
rs1354884	8.99E-01	9.82E-01	9.54E-01	6.69E-01
rs4399964	9.04E-01	8.61E-01	8.24E-01	6.52E-01
rs2163011	9.21E-01	7.37E-01	7.84E-01	8.99E-01
rs2303839	9.34E-01	7.67E-01	5.17E-01	7.60E-01
rs1429110	9.39E-01	9.92E-01	7.69E-01	8.27E-01
rs1429133	9.46E-01	5.58E-01	4.06E-01	8.18E-01
rs2357604	9.99E-01	9.61E-01	9.70E-01	6.95E-01
rs12501429	NA	NA	NA	NA
rs13134510	NA	NA	NA	NA
rs1366691	NA	NA	NA	NA
rs1429139	NA	NA	NA	NA

The table represents the conditional regression analysis conditioned on the top 4 SNPs.

Corresponding p-value estimates from the regression analysis as indicated.

Table A.5 Potential functional causal variant predicted using likelihood ratio analysis

Marker	2 log Likelihood	Likelihood ratio	p-value	Variant
rs13134510	6123.128	Referent	>0.05	Potential functional causal variant
rs28645698	6121.273	1.855	>0.05	Excluded variants
rs2884222	6118.57	4.558	>0.05	Excluded variants
rs1594082	6118.408	4.72	>0.05	Excluded variants
rs1429133	6115.49	7.638	>0.05	Excluded variants
rs1466985	6113.879	9.249	0.05 to 0.01	Excluded variants
rs4399964	6112.56	10.568	<0.01	Potential functional causal variant
rs1583003	6104.122	19.006	<0.01	Potential functional causal variant
rs17611755	6103.807	19.321	<0.01	Potential functional causal variant
rs6822565	6099.999	23.129	<0.01	Potential functional causal variant
rs2059904	6090.463	32.665	<0.01	Potential functional causal variant
rs6810798	6087.017	36.111	<0.01	Potential functional causal variant
rs28720373	6073.079	50.049	<0.01	Potential functional causal variant
rs6812093	6032.227	90.901	<0.01	Potential functional causal variant
rs10519886	6022.739	100.389	<0.01	Potential functional causal variant
rs1346595	6019.228	103.9	<0.01	Potential functional causal variant
rs6836670	6006.46	116.668	<0.01	Potential functional causal variant
rs1568136	6003.462	119.666	<0.01	Potential functional causal variant
rs1346594	6001.358	121.77	<0.01	Potential functional causal variant
rs11947583	5983.166	139.962	<0.01	Potential functional causal variant
rs17022714	5979.445	143.683	<0.01	Potential functional causal variant
rs1354885	5974.907	148.221	<0.01	Potential functional causal

Marker	2 log Likelihood	Likelihood ratio	p-value	Variant
				variant
rs28623525	5967.853	155.275	<0.01	Potential functional causal variant
rs62341515	5959.886	163.242	<0.01	Potential functional causal variant
rs73853371	5954.522	168.606	<0.01	Potential functional causal variant
rs62341517	5954.399	168.729	<0.01	Potential functional causal variant
rs62341519	5936.181	186.947	<0.01	Potential functional causal variant
rs75311641	5931.761	191.367	<0.01	Potential functional causal variant
rs113128727	5930.281	192.847	<0.01	Potential functional causal variant
rs62341516	5924.165	198.963	<0.01	Potential functional causal variant
rs58378349	5921.051	202.077	<0.01	Potential functional causal variant
rs59834135	5905.393	217.735	<0.01	Potential functional causal variant
rs1429141	5899.825	223.303	<0.01	Potential functional causal variant
rs2357604	5899.179	223.949	<0.01	Potential functional causal variant
rs2217348	5897.983	225.145	<0.01	Potential functional causal variant
rs9654228	5890.799	232.329	<0.01	Potential functional causal variant
rs13132657	5882.017	241.111	<0.01	Potential functional causal variant
rs7671190	5876.471	246.657	<0.01	Potential functional causal variant
rs1429106	5873.905	249.223	<0.01	Potential functional causal variant
rs6812819	5870.066	253.062	<0.01	Potential functional causal variant
rs1429105	5869.766	253.362	<0.01	Potential functional causal variant
rs1354884	5860.684	262.444	<0.01	Potential functional causal variant
rs11735996	5858.13	264.998	<0.01	Potential functional causal variant

Marker	2 log Likelihood	Likelihood ratio	p-value	Variant
rs13105529	5856.449	266.679	<0.01	Potential functional causal variant
rs6812432	5853.793	269.335	<0.01	Potential functional causal variant
rs28406843	5850.483	272.645	<0.01	Potential functional causal variant
rs1568137	5847.867	275.261	<0.01	Potential functional causal variant
rs11726718	5836.949	286.179	<0.01	Potential functional causal variant
rs2163011	5835.072	288.056	<0.01	Potential functional causal variant
rs13147231	5826.961	296.167	<0.01	Potential functional causal variant
rs12498595	5808.559	314.569	<0.01	Potential functional causal variant
rs1366691	5803.278	319.85	<0.01	Potential functional causal variant
rs1346600	5801.051	322.077	<0.01	Potential functional causal variant
rs28612496	5799.913	323.215	<0.01	Potential functional causal variant
rs11728738	5791.917	331.211	<0.01	Potential functional causal variant
rs376650129	5790.732	332.396	<0.01	Potential functional causal variant
rs369660577	5790.732	332.396	<0.01	Potential functional causal variant
rs1864248	5790.543	332.585	<0.01	Potential functional causal variant
rs12511935	5780.621	342.507	<0.01	Potential functional causal variant
rs2357778	5777.205	345.923	<0.01	Potential functional causal variant
rs2357779	5777.205	345.923	<0.01	Potential functional causal variant
rs11100960	5772.671	350.457	<0.01	Potential functional causal variant
rs1429134	5766.16	356.968	<0.01	Potential functional causal variant
rs72953535	5764.977	358.151	<0.01	Potential functional causal variant
rs1828034	5753.122	370.006	<0.01	Potential functional causal variant

Marker	2 log Likelihood	Likelihood ratio	p-value	Variant
rs2562871	5752.659	370.469	<0.01	Potential functional causal variant
rs2562873	5747.017	376.111	<0.01	Potential functional causal variant
rs61379585	5744.806	378.322	<0.01	Potential functional causal variant
rs12500103	5744.61	378.518	<0.01	Potential functional causal variant
rs2435095	5739.167	383.961	<0.01	Potential functional causal variant
rs78865503	5738.491	384.637	<0.01	Potential functional causal variant
rs2562875	5733.313	389.815	<0.01	Potential functional causal variant
rs1366689	5727.215	395.913	<0.01	Potential functional causal variant
rs1366679	5726.927	396.201	<0.01	Potential functional causal variant
rs2174801	5722.277	400.851	<0.01	Potential functional causal variant
rs937880	5720.549	402.579	<0.01	Potential functional causal variant
rs1429112	5719.114	404.014	<0.01	Potential functional causal variant
rs2714905	5718.711	404.417	<0.01	Potential functional causal variant
rs2562870	5711.467	411.661	<0.01	Potential functional causal variant
rs1579452	5710.671	412.457	<0.01	Potential functional causal variant
rs11737107	5700.103	423.025	<0.01	Potential functional causal variant
rs12501429	5699.323	423.805	<0.01	Potential functional causal variant
rs2562877	5699.127	424.001	<0.01	Potential functional causal variant
rs2562878	5699.127	424.001	<0.01	Potential functional causal variant
rs17023141	5692.805	430.323	<0.01	Potential functional causal variant
rs1429100	5671.086	452.042	<0.01	Potential functional causal variant
rs11938488	5670.605	452.523	<0.01	Potential functional causal variant

Marker	2 log Likelihood	Likelihood ratio	p-value	Variant
rs2163012	5669.722	453.406	<0.01	Potential functional causal variant
rs1429139	5660.401	462.727	<0.01	Potential functional causal variant
rs1560226	5649.677	473.451	<0.01	Potential functional causal variant
rs1816280	5645.906	477.222	<0.01	Potential functional causal variant
rs1864247	5645.494	477.634	<0.01	Potential functional causal variant
rs2562882	5642.709	480.419	<0.01	Potential functional causal variant
rs12645918	5640.232	482.896	<0.01	Potential functional causal variant
rs1346598	5639.014	484.114	<0.01	Potential functional causal variant
rs7667633	5635.779	487.349	<0.01	Potential functional causal variant
rs2714900	5629.698	493.43	<0.01	Potential functional causal variant
rs2435094	5625.065	498.063	<0.01	Potential functional causal variant
rs1975060	5620.444	502.684	<0.01	Potential functional causal variant
rs1429142	5614.404	508.724	<0.01	Potential functional causal variant
rs13120678	5611.446	511.682	<0.01	Potential functional causal variant
rs2714901	5609.582	513.546	<0.01	Potential functional causal variant
rs28602756	5599.816	523.312	<0.01	Potential functional causal variant
rs2303839	5599.224	523.904	<0.01	Potential functional causal variant
rs937881	5592.426	530.702	<0.01	Potential functional causal variant
rs55771464	5587.225	535.903	<0.01	Potential functional causal variant
rs1429137	5583.346	539.782	<0.01	Potential functional causal variant
rs17023182	5572.162	550.966	<0.01	Potential functional causal variant
rs4593108	5568.512	554.616	<0.01	Potential functional causal variant

Marker	2 log Likelihood	Likelihood ratio	p-value	Variant
rs934146	5562.088	561.04	<0.01	Potential functional causal variant
rs6836525	5556.706	566.422	<0.01	Potential functional causal variant
rs6836562	5552.469	570.659	<0.01	Potential functional causal variant
rs2562879	5544.743	578.385	<0.01	Potential functional causal variant
rs17023196	5541.492	581.636	<0.01	Potential functional causal variant
rs2562880	5528.4	594.728	<0.01	Potential functional causal variant
rs2562874	5524.898	598.23	<0.01	Potential functional causal variant
rs17023204	5513.566	609.562	<0.01	Potential functional causal variant
rs2562876	5506.79	616.338	<0.01	Potential functional causal variant
rs12640442	5506.736	616.392	<0.01	Potential functional causal variant
rs57997710	5501.155	621.973	<0.01	Potential functional causal variant
rs2562869	5499.954	623.174	<0.01	Potential functional causal variant
rs12646693	5498.434	624.694	<0.01	Potential functional causal variant
rs58983705	5473.385	649.743	<0.01	Potential functional causal variant
rs7668383	5451.504	671.624	<0.01	Potential functional causal variant
rs1429110	5435.298	687.83	<0.01	Potential functional causal variant
rs1429111	5432.214	690.914	<0.01	Potential functional causal variant
rs1429109	5411.501	711.627	<0.01	Potential functional causal variant
rs9942219	5377.737	745.391	<0.01	Potential functional causal variant
rs7697539	5340.1	783.028	<0.01	Potential functional causal variant
rs80077485	5330.665	792.463	<0.01	Potential functional causal variant
rs6821368	5290.765	832.363	<0.01	Potential functional causal variant

Marker	2 log Likelihood	Likelihood ratio	p-value	Variant
rs6537487	5262.915	860.213	<0.01	Potential functional causal variant
rs7674137	5233.649	889.479	<0.01	Potential functional causal variant
rs1878404	5193.807	929.321	<0.01	Potential functional causal variant
rs9647489	5191.911	931.217	<0.01	Potential functional causal variant

Each associated SNPs were compared to the top associated SNP (rs13134510) and likelihood of potential causal variant is estimated. SNPs with likelihood ratio P-value <0.01 was considered significant and potential causal variant

Table A.6 Regulome Db scoring of associated SNPs

Coordinate	db SNP ID	Regulome DB score	Resources
chr4:148318047	rs7671190	1e	UCSC | ENSEMBL | dbSNP
chr4:148084304	rs4399964	1f	UCSC | ENSEMBL | dbSNP
chr4:148244801	rs2714900	1f	UCSC | ENSEMBL | dbSNP
chr4:148253323	rs2562873	1f	UCSC | ENSEMBL | dbSNP
chr4:148265187	rs1560226	1f	UCSC | ENSEMBL | dbSNP
chr4:148281811	rs1366691	1f	UCSC | ENSEMBL | dbSNP
chr4:148284103	rs1429139	1f	UCSC | ENSEMBL | dbSNP
chr4:148328867	rs17023196	1f	UCSC | ENSEMBL | dbSNP
chr4:148329645	rs17023204	1f	UCSC | ENSEMBL | dbSNP
chr4:148284372	rs6836670	2b	UCSC | ENSEMBL | dbSNP
chr4:148262343	rs2562880	3a	UCSC | ENSEMBL | dbSNP
chr4:148287359	rs9654228	3a	UCSC | ENSEMBL | dbSNP
chr4:148432439	rs1568136	3a	UCSC | ENSEMBL | dbSNP
chr4:148435009	rs6821368	3a	UCSC | ENSEMBL | dbSNP
chr4:148248189	rs2562871	4	UCSC | ENSEMBL | dbSNP
chr4:148272794	rs7668383	4	UCSC | ENSEMBL | dbSNP
chr4:148283139	rs7667633	4	UCSC | ENSEMBL | dbSNP
chr4:148287511	rs13134510	4	UCSC | ENSEMBL | dbSNP
chr4:148437511	rs6822565	4	UCSC | ENSEMBL | dbSNP
chr4:148056169	rs937880	5	UCSC | ENSEMBL | dbSNP
chr4:148062990	rs1354885	5	UCSC | ENSEMBL | dbSNP
chr4:148064828	rs11947583	5	UCSC | ENSEMBL | dbSNP
chr4:148077417	rs2357604	5	UCSC | ENSEMBL | dbSNP
chr4:148248745	rs2163011	5	UCSC | ENSEMBL | dbSNP
chr4:148267129	rs1975060	5	UCSC | ENSEMBL | dbSNP
chr4:148276042	rs12501429	5	UCSC | ENSEMBL | dbSNP
chr4:148276399	rs2059904	5	UCSC | ENSEMBL | dbSNP
chr4:148285766	rs6812819	5	UCSC | ENSEMBL | dbSNP
chr4:148289388	rs1429142	5	UCSC | ENSEMBL | dbSNP
chr4:148291580	rs28406843	5	UCSC | ENSEMBL | dbSNP
chr4:148310347	rs17023182	5	UCSC | ENSEMBL | dbSNP
chr4:148317339	rs1429105	5	UCSC | ENSEMBL | dbSNP
chr4:148317564	rs1429106	5	UCSC | ENSEMBL | dbSNP
chr4:148317855	rs11726718	5	UCSC | ENSEMBL | dbSNP

Coordinate	db SNP ID	Regulome DB score	Resources
chr4:148330068	rs7697539	5	UCSC | ENSEMBL | dbSNP
chr4:148330344	rs1429109	5	UCSC | ENSEMBL | dbSNP
chr4:148330379	rs1429110	5	UCSC | ENSEMBL | dbSNP
chr4:148375146	rs13132657	5	UCSC | ENSEMBL | dbSNP
chr4:148439600	rs7674137	5	UCSC | ENSEMBL | dbSNP
chr4:148066685	rs73853371	6	UCSC | ENSEMBL | dbSNP
chr4:148067083	rs62341516	6	UCSC | ENSEMBL | dbSNP
chr4:148067378	rs9647489	6	UCSC | ENSEMBL | dbSNP
chr4:148067562	rs113128727	6	UCSC | ENSEMBL | dbSNP
chr4:148067746	rs62341517	6	UCSC | ENSEMBL | dbSNP
chr4:148068052	rs75311641	6	UCSC | ENSEMBL | dbSNP
chr4:148068715	rs58378349	6	UCSC | ENSEMBL | dbSNP
chr4:148069855	rs59834135	6	UCSC | ENSEMBL | dbSNP
chr4:148073976	rs62341519	6	UCSC | ENSEMBL | dbSNP
chr4:148075339	rs1354884	6	UCSC | ENSEMBL | dbSNP
chr4:148090485	rs1828034	6	UCSC | ENSEMBL | dbSNP
chr4:148233472	rs2303839	6	UCSC | ENSEMBL | dbSNP
chr4:148236537	rs1346598	6	UCSC | ENSEMBL | dbSNP
chr4:148236731	rs10519886	6	UCSC | ENSEMBL | dbSNP
chr4:148241452	rs1346600	6	UCSC | ENSEMBL | dbSNP
chr4:148243616	rs1864248	6	UCSC | ENSEMBL | dbSNP
chr4:148245248	rs2435094	6	UCSC | ENSEMBL | dbSNP
chr4:148245634	rs1429134	6	UCSC | ENSEMBL | dbSNP
chr4:148247352	rs2562869	6	UCSC | ENSEMBL | dbSNP
chr4:148247481	rs2714901	6	UCSC | ENSEMBL | dbSNP
chr4:148253069	rs12498595	6	UCSC | ENSEMBL | dbSNP
chr4:148254878	rs2435095	6	UCSC | ENSEMBL | dbSNP
chr4:148259421	rs2562876	6	UCSC | ENSEMBL | dbSNP
chr4:148260069	rs2714905	6	UCSC | ENSEMBL | dbSNP
chr4:148260760	rs2562877	6	UCSC | ENSEMBL | dbSNP
chr4:148260895	rs2562878	6	UCSC | ENSEMBL | dbSNP
chr4:148262921	rs1429112	6	UCSC | ENSEMBL | dbSNP
chr4:148264672	rs2562882	6	UCSC | ENSEMBL | dbSNP
chr4:148268026	rs2163012	6	UCSC | ENSEMBL | dbSNP
chr4:148282108	rs1429137	6	UCSC | ENSEMBL | dbSNP
chr4:148288066	rs1429141	6	UCSC | ENSEMBL | dbSNP
chr4:148296024	rs1583003	6	UCSC | ENSEMBL | dbSNP

Coordinate	db SNP ID	Regulome DB score	Resources
chr4:148296556	rs1579452	6	UCSC | ENSEMBL | dbSNP
chr4:148296856	rs1429100	6	UCSC | ENSEMBL | dbSNP
chr4:148296903	rs12511935	6	UCSC | ENSEMBL | dbSNP
chr4:148297051	rs12500103	6	UCSC | ENSEMBL | dbSNP
chr4:148297179	rs1366679	6	UCSC | ENSEMBL | dbSNP
chr4:148313802	rs13105529	6	UCSC | ENSEMBL | dbSNP
chr4:148315662	rs11100960	6	UCSC | ENSEMBL | dbSNP
chr4:148315684	rs6812432	6	UCSC | ENSEMBL | dbSNP
chr4:148316576	rs2357778	6	UCSC | ENSEMBL | dbSNP
chr4:148318634	rs1346594	6	UCSC | ENSEMBL | dbSNP
chr4:148318729	rs1346595	6	UCSC | ENSEMBL | dbSNP
chr4:148320198	rs12645918	6	UCSC | ENSEMBL | dbSNP
chr4:148329350	rs58983705	6	UCSC | ENSEMBL | dbSNP
chr4:148329381	rs57997710	6	UCSC | ENSEMBL | dbSNP
chr4:148330513	rs1429111	6	UCSC | ENSEMBL | dbSNP
chr4:148330969	rs9942219	6	UCSC | ENSEMBL | dbSNP
chr4:148373899	rs376650129	6	UCSC | ENSEMBL | dbSNP
chr4:148373900	rs369660577	6	UCSC | ENSEMBL | dbSNP
chr4:148429410	rs6537487	6	UCSC | ENSEMBL | dbSNP
chr4:148437154	rs6812093	6	UCSC | ENSEMBL | dbSNP
chr4:148055431	rs11938488	No Data	UCSC | ENSEMBL | dbSNP
chr4:148056319	rs937881	No Data	UCSC | ENSEMBL | dbSNP
chr4:148058971	rs17611755	No Data	UCSC | ENSEMBL | dbSNP
chr4:148060450	rs61379585	No Data	UCSC | ENSEMBL | dbSNP
chr4:148061275	rs78865503	No Data	UCSC | ENSEMBL | dbSNP
chr4:148062555	rs62341515	No Data	UCSC | ENSEMBL | dbSNP
chr4:148065511	rs17022714	No Data	UCSC | ENSEMBL | dbSNP
chr4:148069539	rs2174801	No Data	UCSC | ENSEMBL | dbSNP
chr4:148243297	rs1864247	No Data	UCSC | ENSEMBL | dbSNP
chr4:148247367	rs2562870	No Data	UCSC | ENSEMBL | dbSNP
chr4:148254263	rs2562874	No Data	UCSC | ENSEMBL | dbSNP
chr4:148256734	rs1366689	No Data	UCSC | ENSEMBL | dbSNP
chr4:148256890	rs2562875	No Data	UCSC | ENSEMBL | dbSNP
chr4:148259751	rs11737107	No Data	UCSC | ENSEMBL | dbSNP
chr4:148260922	rs2562879	No Data	UCSC | ENSEMBL | dbSNP
chr4:148263828	rs1816280	No Data	UCSC | ENSEMBL | dbSNP
chr4:148268650	rs934146	No Data	UCSC | ENSEMBL | dbSNP

Coordinate	db SNP ID	Regulome DB score	Resources
chr4:148273396	rs13120678	No Data	UCSC | ENSEMBL | dbSNP
chr4:148278845	rs6836525	No Data	UCSC | ENSEMBL | dbSNP
chr4:148278890	rs6836562	No Data	UCSC | ENSEMBL | dbSNP
chr4:148281000	rs4593108	No Data	UCSC | ENSEMBL | dbSNP
chr4:148283783	rs17023141	No Data	UCSC | ENSEMBL | dbSNP
chr4:148290817	rs6810798	No Data	UCSC | ENSEMBL | dbSNP
chr4:148291241	rs28720373	No Data	UCSC | ENSEMBL | dbSNP
chr4:148291242	rs28623525	No Data	UCSC | ENSEMBL | dbSNP
chr4:148291437	rs28612496	No Data	UCSC | ENSEMBL | dbSNP
chr4:148293946	rs2217348	No Data	UCSC | ENSEMBL | dbSNP
chr4:148296165	rs11735996	No Data	UCSC | ENSEMBL | dbSNP
chr4:148315855	rs55771464	No Data	UCSC | ENSEMBL | dbSNP
chr4:148316668	rs2357779	No Data	UCSC | ENSEMBL | dbSNP
chr4:148319116	rs11728738	No Data	UCSC | ENSEMBL | dbSNP
chr4:148319184	rs28602756	No Data	UCSC | ENSEMBL | dbSNP
chr4:148319262	rs72953535	No Data	UCSC | ENSEMBL | dbSNP
chr4:148319554	rs13147231	No Data	UCSC | ENSEMBL | dbSNP
chr4:148326781	rs12640442	No Data	UCSC | ENSEMBL | dbSNP
chr4:148326866	rs12646693	No Data	UCSC | ENSEMBL | dbSNP
chr4:148334943	rs80077485	No Data	UCSC | ENSEMBL | dbSNP
chr4:148398228	rs1568137	No Data	UCSC | ENSEMBL | dbSNP
chr4:148440624	rs1878404	No Data	UCSC | ENSEMBL | dbSNP

The table represents the RegulomeDB scoring of the associated SNPs. The scores range from 1-6. The description of the scores include: 1a- eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak ; 1b -eQTL + TF binding + any motif + DNase Footprint + DNase peak ; 1c- eQTL + TF binding + matched TF motif + DNase peak; 1d-eQTL + TF binding + any motif + DNase peak; 1e-eQTL + TF binding + matched TF motif; 1f-eQTL + TF binding / DNase peak; 2a- TF binding + matched TF motif + matched DNase Footprint + DNase peak; 2b-TF binding + any motif + DNase Footprint + DNase peak; 2c-TF binding + matched TF motif + DNase peak; 3a-TF binding + any motif + DNase peak ; 3b-TF binding + matched TF motif ; 4-TF binding + DNase peak; 5-TF binding or DNase peak; 6-other.

Table A.7 Description of the RegulomeDB scoring of the associated SNPs in breast cancer cell lines

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)
rs1366691	1f	chr4:148281608..148281858/FOS/MCF10A-Er-Src/4ohtam_1um_12hr; chr4:148281590..148281840/FOS/MCF10A-Er-Src/4ohtam_1um_4hr		ARHGAP10/Lymphoblastoid;		chr4:148281200..148285000/Enhancers/ENCODE/HMEC Mammary Epithelial Primary Cells; chr4:148281200..148285000/Enhancers/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC); chr4:148253600..148282800/Quiescent/Low/Epithelial/Breast Myoepithelial Primary Cells
rs1429139	1f			ARHGAP10/Lymphoblastoid;	DNase-seq/ chr4:148284006..148284533/ Mcf7	chr4:148281200..148285000/Enhancers/ENCODE/HMEC Mammary Epithelial Primary Cells; chr4:148281200..148285000/Enhancers/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC) ; chr4:148282800..148285000/Enhancers/Epithelial/Breast Myoepithelial Primary Cells;
rs17023196	1f			ARHGAP10/Lymphoblastoid;		chr4:148328600..148329000/Enhancers/ENCODE/HMEC Mammary Epithelial Primary Cells; chr4:148309800..148330000/Weak transcription/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC); chr4:148326400.148329800 /Quiescent/Low/Epithelial/Breast Myoepithelial Primary Cells

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)

rs6836670	2b		Footprinting/chr4:148284371..148284386/HMGIY/Mcf7Hypoxlac;
			Footprinting/chr4:148284371..148284386/HMGIY/Mcf7;
					DNase-seq/chr4:148284006..148284533/Mcf7;
					DNase-seq/chr4:148284152..148284521/Mcf7;
					DNase-seq/chr4:148284157..148284475/Mcf7;
					DNase-seq/chr4:148284207..148284494/T47d;
					DNase-seq/chr4:148284260..148284430/Mcf7;
					DNase-seq/chr4:148284280..148284430/Mcf7;
					DNaseeq/chr4:148284244.148284396/Mcf7/Hypoxlac
						chr4:148281200..148285000/Enhancers/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC); hr4:148281200..148285000/Enhancers ENCODE/HMEC Mammary Epithelial Primary Cells; chr4:148282800..148285000/Enhancers EpithelialBreast Myoepithelial Primary Cells

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)
rs7667633	4	ChIP-seq/chr4:148282875..148283415/POLR2A/MCF10A-Er-Src/01pct/ENCODE ChIP-seq/chr4:148282928..148283244/STAT3/MCF10A-Er-Src/4ohtam_1um_36hr/ENCODE ChIP-seq/chr4:148282964..148283240/STAT3/MCF10A-Er-Src/01pct_4hr/ENCODE			DNase-seq/chr4:148282821..148283215/Mcf7//ENCODE DNase-seq/chr4:148282821..148283215/Mcf7/Ctcfshrna/ENCODE DNase-seq/chr4:148282826..148283183/Mcf7/Hypoxlac/ENCODE DNase-seq/chr4:148283125..148283275/Hmec//ENCODE	ChromHMM/chr4:148281200..148285000/Enhancers/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC) ChromHMM/chr4:148281200..148285000/Enhancers/ENCODE/HMEC Mammary Epithelial Primary Cells ChromHMM/chr4:148282800..148285000/Enhancers/Epithelial/Breast Myoepithelial Primary Cells

rs1560226	1f			ARHGAP10/Lymphoblastoid;		chr4:148262000..148272000/Weak transcriptionEpithelial/Breast variant Human Mammary Epithelial Cells (vHMEC);
						chr4:148262600..148272000/Quiescent/Low/ENCODE/HMEC Mammary Epithelial Primary Cells;
						chr4:148253600..148282800/Quiescent/Low/Epithelial/Breast Myoepithelial Primary Cells

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)
rs17023204	1f			ARHGAP10/Lymphoblastoid;		chr4:148309800..148330000/Weak transcription/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC); chr4:148329000..148330200/Weak transcription/ENCODE/HMEC Mammary Epithelial Primary Cells
rs2562873	1f		PWMchr4:148253323..148253337 NFATC1	ARHGAP10/Lymphoblastoid;		chr4:148238000..148255800/Weak transcription Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC); chr4:148253200..148253400/Quiescent/Low/Epithelial/Breast Myoepithelial Primary Cells/chr4:148209000..148261200 Quiescent/Low/ENCODE/HMEC Mammary Epithelial Primary Cells
rs2714900	1f			ARHGAP10/Lymphoblastoid; LSM6/ Lymphoblastoid		chr4:148238000..148255800/ Weak transcription/ Epithelial/ Breast variant Human Mammary Epithelial Cells (vHMEC); chr4:148209000..148261200Quiescent/Low/ENCODE/HMEC Mammary Epithelial PrimaryCells; chr4:148240600..148251400/Quiescent/Low/Epithelial/Breast Myoepithelial Primary Cells

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)
rs4399964	1f			ARHGAP10/Lymphoblastoid; LSM6/Lymphoblastoid		chr4:148079800..148096800/Quiescent/Low/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC); chr4:148080000..148138800/Quiescent/Low/Epithelial/Breast Myoepithelial Primary Cells; chr4:148080000..148205600/Quiescent/Low/HMEC Mammary Epithelial Primary Cells;
rs7671190	1e		chr4:148318041..148318060/HNF4	ARHGAP10/Lymphoblastoid		chr4:148309800..148328600 /Quiescent/Low/HMEC Mammary Epithelial Primary Cells ; chr4:148309800..148330000/Weak transcription/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC) ; chr4:148316200..148324600/Quiescent/Low/Epithelial/ Breast Myoepithelial Primary Cells
rs13134510	4					ChromHMM/chr4:148285000..148291200/Weak transcription/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC) ChromHMM/chr4:148285000..148291600/Quiescent/Low/Epithelial/Breast Myoepithelial Primary Cells ChromHMM/chr4:148285000..148290200/Weak transcription/ENCODE/HMEC Mammary Epithelial Primary Cells

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)
rs2562871	4					chr4:148238000..148255800/Weak transcription/Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC) REMC chr4:148240600..148251400/Quiescent/Low/Epithelial/Breast Myoepithelial Primary Cells
rs2562880	3a					chr4:148253600..148282800/Quiescent/Low/Epithelial Breast Myoepithelial Primary Cells; chr4:148262000..148272000/Weak transcription Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC);chr4:148261200..148262600 Enhancers ENCODE HMEC Mammary Epithelial Primary Cells
rs7668383	4	chr4:148272676..148272920/FOS/MCF10A-Er-Src/01pct chr4:148272676..148272926/FOS/MCF10A-Er-Src/4ohtam_1um_4hr chr4:148272652..148272932/STAT3/MCF10A-Er-Src/4ohtam_1um_12hr chr4:148272621..148272901/STAT3/			chr4:148272700..148272850 Hmec	

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)
		MCF10A-Er-Src/01pct_12hr chr4:148272720..148272848/FOS/MCF10A-Er-Src/4ohtam_1um_12hr chr4:148272734..148272837/FOS/MCF10A-Er-Src/4ohtam_1um_36hr				
rs9654228	3a		chr4:148287340..148287364 Gfi-1			chr4:148285000..148291200/Weak transcription/Epithelial Breast variant Human Mammary Epithelial Cells (vHMEC);chr4:148285000..148291600/Quiescent/Low Epithelial/Breast Myoepithelial Primary Cells; chr4:148285000..148290200/Weak transcription/ENCODE/ HMEC Mammary Epithelial Primary Cells

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)
rs1568136	3a		chr4:148432438..148432451/EN1			ChromHMM/chr4:148420000..148438800 /Weak transcription/Epithelial Breast variant Human Mammary Epithelial Cells (vHMEC) REMC ChromHMM/chr4:148420200..148439400/Quiescent/Low ENCODE/HMEC Mammary Epithelial Primary Cells/REMC ChromHMM/chr4:148422600..148433600/Weak transcription Epithelial/Breast Myoepithelial Primary Cells/REMC
rs6821368	3a		chr4:148435000..148435010/NF-AT chr4:148435008..148435021/SOX			ChromHMM/chr4:148420000..148438800/Weak transcription Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC)/REMC ChromHMM /chr4:148434400..148447400/Weak transcription Epithelial/Breast Myoepithelial Primary Cells/REMC ChromHMM/chr4:148420200..148439400/Quiescent/Low ENCODE/HMEC Mammary Epithelial Primary Cells/REMC

db SNP ID	Regulome DB score	Protein binding (Chip-seq)	Motif	eQTLs	DNAase hypersensitive site	Histone Modification (REMC, Regulome Db)
rs6822565	4					ChromHMM/chr4:148420000..148438800/Weak transcription Epithelial/Breast variant Human Mammary Epithelial Cells (vHMEC)/REMC ChromHMM/chr4:148434400..148447400/Weak transcription Epithelial/Breast Myoepithelial Primary Cells/REMC ChromHMM/chr4:148420200..148439400/Quiescent/Low ENCODE/HMEC Mammary Epithelial Primary Cells/REMC

This table represents the description of the RegulomeDb score for the associated SNPs with score from 1-4. The description includes transcription factor binding (ChIP-seq), changes in motif binding, eQTLs, hypersensitive sites, and histone modification in breast cell lines including, Human Mammary Epithelial Cells (HMEC), Breast variant Human Mammary Epithelial Cells (vHMEC), Breast Myoepithelial Primary Cells; MCF-7. Eqtl data was available based on Lymphoblastoid cell lines.

Table A.8 HaploReg analysis of the 19 putative functional SNPs in Human Mammary Cell lines

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
rs13134510											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)			H3K4me1_Enh						
E027	BRST.MYO	Breast Myoepithelial Primary Cells									
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									
rs1366691											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)	7_Enh	14_EnhA2	H3K4me1_Enh				DNase		
E027	BRST.MYO	Breast Myoepithelial Primary Cells		15_EnhAF							

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells	7_Enh	15_EnhAF	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro	DNase		
rs1429139											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)	7_Enh	14_EnhA2	H3K4me1_Enh				DNase		
E027	BRST.MYO	Breast Myoepithelial Primary Cells	7_Enh	14_EnhA2	H3K4me1_Enh						
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells	7_Enh	14_EnhA2	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro	DNase		
rs17023196											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)			H3K4me1_Enh						
E027	BRST.MYO	Breast Myoepitheli			H3K4me						

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
		al Primary Cells			1_Enh						
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells	7_Enh		H3K4me1_Enh						
rs7667633											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)	7_Enh	14_EnhA2	H3K4me1_Enh				DNase		
E027	BRST.MYO	Breast Myoepithelial Primary Cells	7_Enh	14_EnhA2	H3K4me1_Enh						
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells	7_Enh	13_EnhA1	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro	DNase		
	MCF10A-Er-Src									POL2 / STAT3	
rs7668383											
E028	BRST.HMEC.35	Breast variant Human Mammary		19_DNase	H3K4me1_Enh				DNase		

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
		Epithelial Cells (vHMEC)									
E027	BRST.MYO	Breast Myoepithelial Primary Cells		19_DNase							
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells		19_DNase					DNase		
rs7671190											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)									
E027	BRST.MYO	Breast Myoepithelial Primary Cells									
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									
rs2714900											

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)									
E027	BRST.MYO	Breast Myoepithelial Primary Cells			H3K4me1_Enh						
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									
rs2562873											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)									
E027	BRST.MYO	Breast Myoepithelial Primary Cells			H3K4me1_Enh						
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
rs1560226											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)									
E027	BRST.MYO	Breast Myoepithelial Primary Cells									
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									
rs17023204											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)			H3K4me1_Enh						
E027	BRST.MYO	Breast Myoepithelial Primary Cells			H3K4me1_Enh						
E119	BRST.HMEC	HMEC Mammary Epithelial			H3K4me1_En						

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
		Primary Cells			h						
rs6836670											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)	7_Enh	14_EnhA2	H3K4me1_Enh				DNase		
E027	BRST.MYO	Breast Myoepithelial Primary Cells	7_Enh	14_EnhA2	H3K4me1_Enh						
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells	7_Enh	14_EnhA2	H3K4me1_Enh		H3K27ac_Enh	H3K9ac_Pro	DNase		
	MCF10A-Er-Src	STAT3									
rs2562880											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)			H3K4me1_Enh						
E027	BRST.MYO	Breast Myoepithelial Primary			H3K4me1_En						

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
		Cells			h						
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells	7_Enh		H3K4me1_Enh						
rs9654228											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)			H3K4me1_Enh						
E027	BRST.MYO	Breast Myoepithelial Primary Cells									
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									
rs2562871											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)									

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
E027	BRST.MYO	Breast Myoepithelial Primary Cells									
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									
rs1568136											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)									
E027	BRST.MYO	Breast Myoepithelial Primary Cells									
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									
											EN1, CEBPB,PAX6
rs6821368											
E028	BRST.HMEC.35	Breast variant Human Mammary									

Epigenome ID	Mnemonic	Description	Chromatin states (Core 15-state model)	Chromatin states (25-state model using 12 imputed marks)	H3K4me1	H3K4me3	H3K27ac	H3K9ac	DNase	Bound Proteins	Regulatory motifs altered (Position Weighted Matrix)
		Epithelial Cells (vHMEC)									
E027	BRST.MYO	Breast Myoepithelial Primary Cells									
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									
											HDAC2,HOXA4,NF-AT,PAX4,POU2F2,POU3F2,SOX,ZFP187
rs6822565											
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)			H3K4me1_Enh						
E027	BRST.MYO	Breast Myoepithelial Primary Cells			H3K4me1_Enh			H3K9ac_Pro			
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells									

Table A.9 eQTL for the fine-mapped SNPs

dbSNP ID	eQTL		Source
	Gene/ Tissue	P-value	
rs7671190	ARHGAP10/Lymphoblastoid	-	HapMap
	EDNRA/ Heart - Left Ventricle	1.20E-07	GTEx

rs4399964	ARHGAP10/Lymphoblastoid;
	LSM6/Lymphoblastoid;
		-	HapMap

rs2714900	ARHGAP10/Lymphoblastoid;
	LSM6/Lymphoblastoid;
		-	HapMap

	EDNRA/ Heart - Left Ventricle	1.50E-09	GTEx
rs2562873	ARHGAP10/Lymphoblastoid;	-	HapMap
	EDNRA/ Heart - Left Ventricle	2.40E-09	GTEx
rs1560226	ARHGAP10/Lymphoblastoid;	-	HapMap
	EDNRA/ Heart - Left Ventricle	7.70E-09	GTEx
rs1366691	ARHGAP10/Lymphoblastoid;	-	HapMap
	EDNRA/ Heart - Left Ventricle	9.70E-09	GTEx
rs1429139	ARHGAP10/Lymphoblastoid;	-	HapMap
	EDNRA/ Heart - Left Ventricle	1.00E-08	GTEx
rs17023196	ARHGAP10/Lymphoblastoid;	-	HapMap
	EDNRA/ Testis	2.30E-05	GTEx
rs17023204	ARHGAP10/Lymphoblastoid;	-	HapMap
	EDNRA/ Testis	2.30E-05	GTEx

The table represents the eQTLs of the fine-mapped SNPs based on the data available from TCGA data analysis, HapMap[335] dataset and GTEx dataset. In each of the dataset, eQTL was estimated in different tissues.

Table A.10 List of genes (overlapped by CNV) that showed correlation with copy number specific gene dosage in breast tumor gene expression

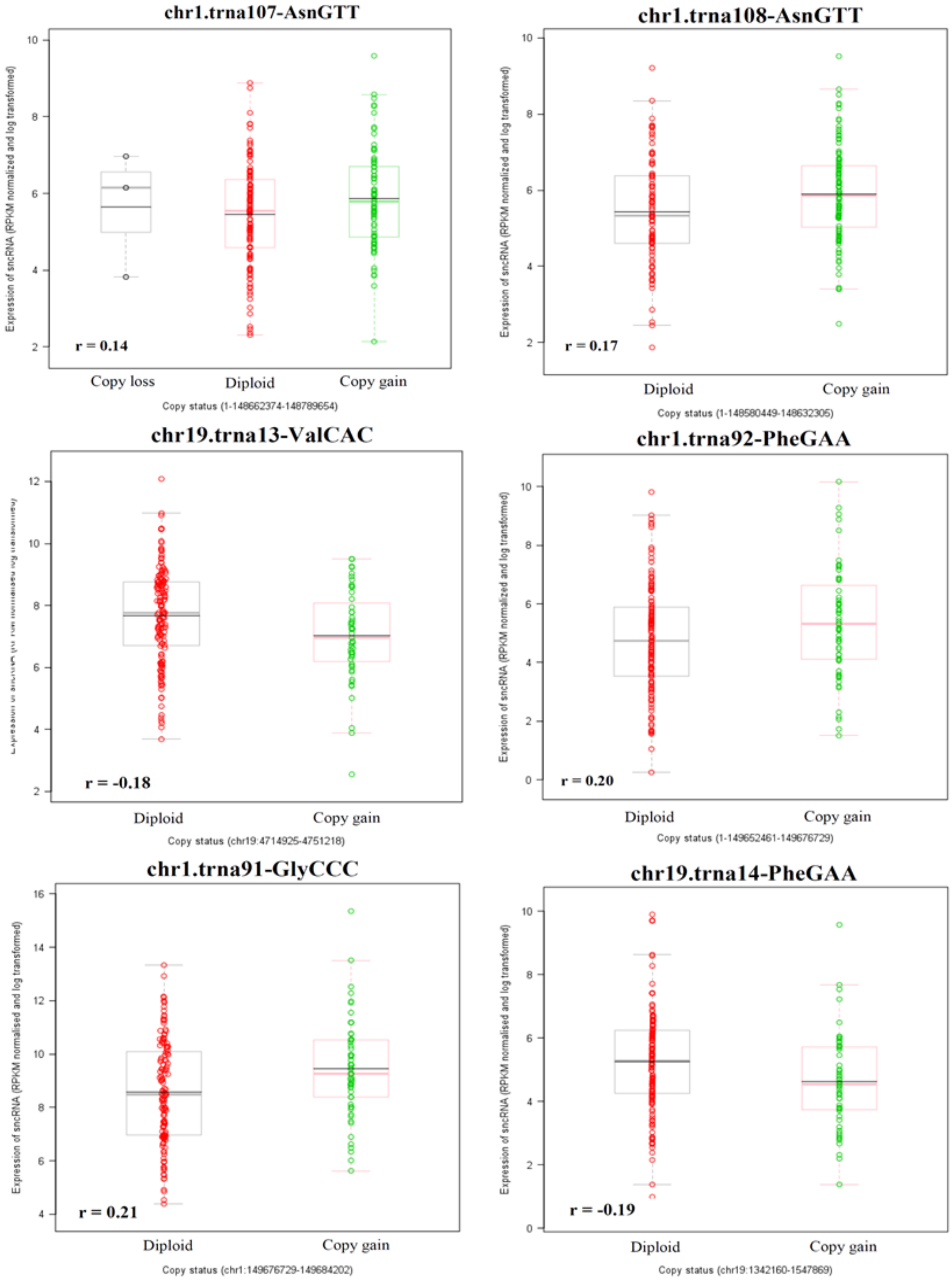
Copy Number region	Agilent gene Probe ID	Gene	Pearson correlation	p-value (correlation)
chr22:24323073-24329964	A_23_P357571	GSTT2	0.39	$1.44x10^{-04}$
chr22:24323073-24329964	A_23_P109427	GSTT2	0.37	$3.27x10^{-04}$
chr6:32482478-32487136	A_23_P45099	HLA-DRB5	0.35	$6.20x10^{-04}$
chr17:44662938-44720330	A_24_P221634	NSF	0.33	$1.64x10^{-03}$
chr4:69396464-69403345	A_23_P501624	UGT2B17	0.29	$5.16x10^{-03}$
chr7:143952878-143957098	A_23_P134566	OR2A7	0.24	$2.36x10^{-02}$
chr1:152572419-152575355	A_23_P405295	LCE3C	0.23	$2.64x10^{-02}$
chr1:196790951-196794804	A_24_P336510	CFHR1	0.23	$3.26x10^{-02}$
chr5:69313680-69784291	A_24_P126682	SMN2	0.21	$5.12x10^{-02}$
chr5:69313680-69784291	A_24_P935881	SERF1B	0.2	$5.75x10^{-02}$

This table shows correlations (r=0.2) between copy status and breast tumor tissue specific gene expression. The tumor gene expression profiles were derived from subset of patients (n=90) who also have their copy number profiles estimated.

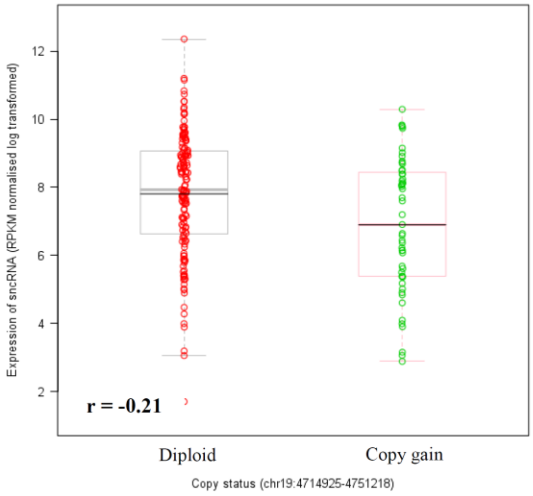
Table A.11 Breast tumor tissue specific expression of genes (overlapped by prognostic CNVs) and their association with prognosis

Gene	P-value	HR [95% CI]	Outcome
GSTM2	0.03	0.68 [0.48-0.95]	OS
SGCZ	0.11	1.29 [0.95-1.74]	OS
HLA_DRB5	0.13	0.78 [0.57-1.07]	OS
ZFP14	0.19	0.81 [0.59-1.11]	OS
LCE3C	0.03	0.73 [0.55-0.98]	RFS

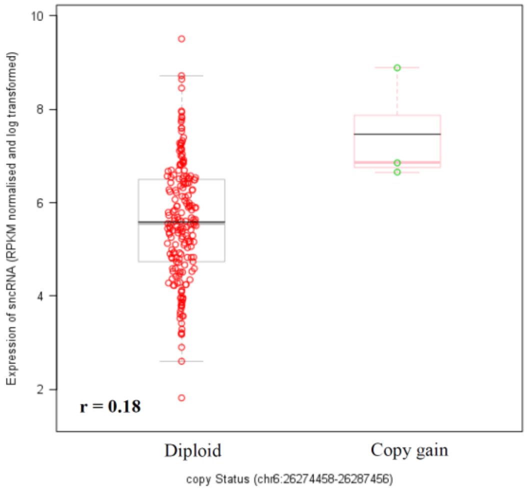
This table shows the gene that are overlapped by prognostic CNVs (as listed in Table 2 & 3) expressed at the breast tumor tissue and their association with prognosis. The tumor gene expression profiles were derived from subset of patients (n=90) who also have their copy number profiles estimated.


chr1.trna107-AsnGTT
Expression of sncRNA (RPKM normalized and log transformed)
10
8
6
4
2
r = 0.14
Copy loss
Diploid
Copy gain
Copy status (1-148662374-148789654)
chr1.trna108-AsnGTT
Expression of sncRNA (RPKM normalized and log transformed)
10
8
6
4
2
r = 0.17
Diploid
Copy gain
Copy status (1-148580449-148632305)
chr19.trna13-ValCAC
12
10
8
6
4
2
r = -0.18
Diploid
Copy gain
Copy status (chr19:4714925-4751218)
chr1.trna92-PheGAA
Expression of sncRNA (RPKM normalized and log transformed)
10
8
6
4
2
0
r = 0.20
Diploid
Copy gain
Copy status (1-149652461-149676729)
chr1.trna91-GlyCCC
Expression of sncRNA (RPKM normalised and log transformed)
16
14
12
10
8
6
4
r = 0.21
Diploid
Copy gain
Copy status (chr1:149676729-149684202)
chr19.trna14-PheGAA
Expression of sncRNA (RPKM normalised and log transformed)
10
8
6
4
2
0
r = -0.19
Diploid
Copy gain
Copy status (chr19:1342160-1547869)

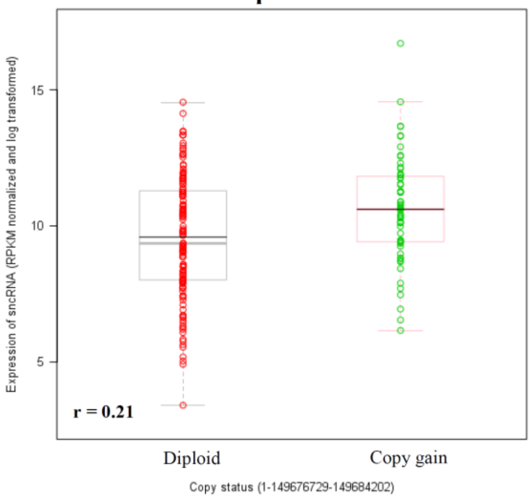
chr19.trna2-GlyTCC



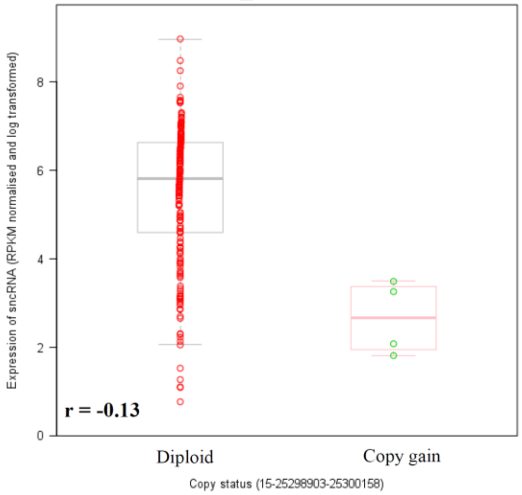
chr6.trna2-MetCAT



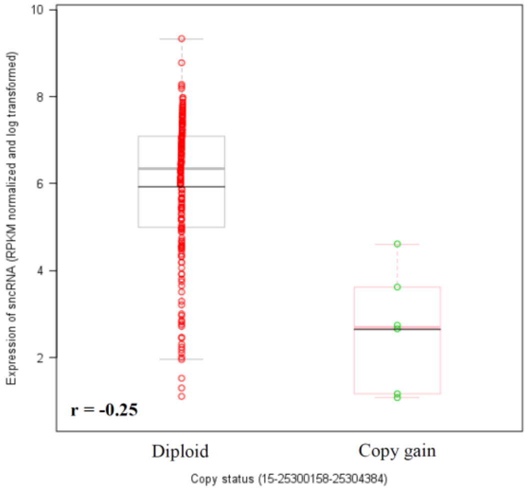
hsa-piR-20636



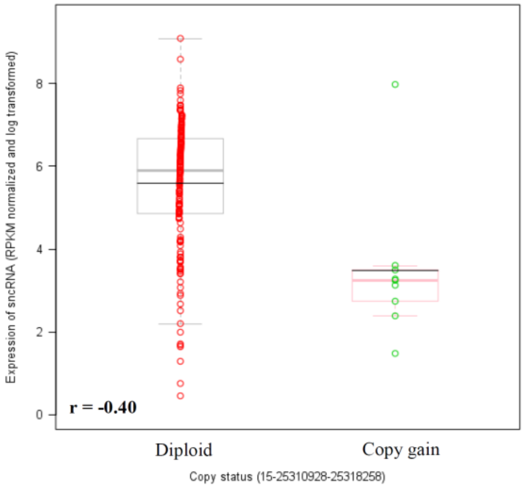
snoRNA_SNORD116-2-201



snoRNA_SNORD116-3-201



snoRNA_SNORD116-6-201



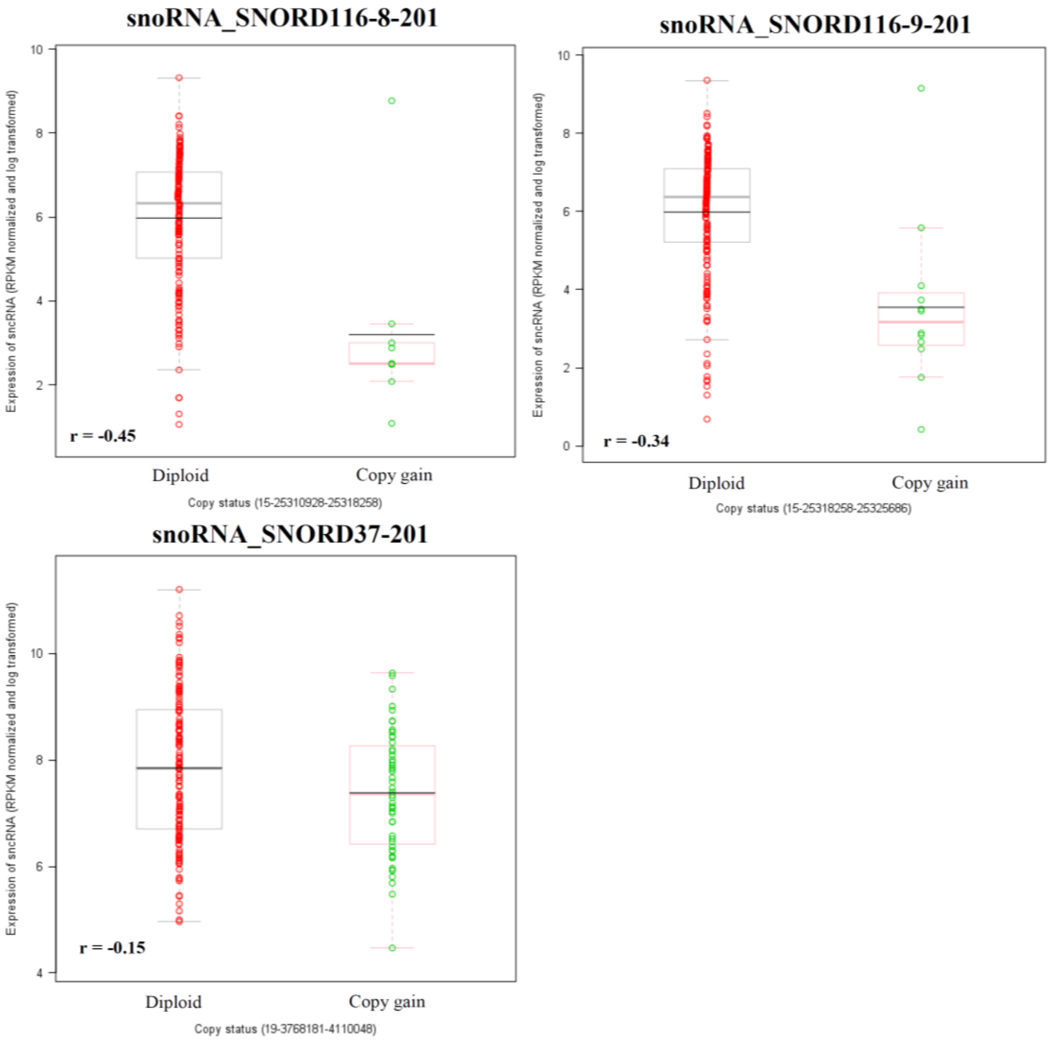


Figure A.2 Gene dosage analysis of CNV-sncRNAs

Gene Dosage for the embedded CNV-sncRNAs were estimated by correlating the germline copy status and sncRNA expression in breast tumor tissue data (HiSeq, n=198, RPKM normalized log transformed) using Pearson correlation. We observed significant positive and negative correlation among the correlated sncRNAs. Gray line in the plots represent the mean expression of sncRNA

Table A.12 NGS generated sequences and sncRNA annotations

Small RNA	Platform	Tissue type	No of samples	Total sncRNAs identified in tissues	No of sncRNAs retained after read count filtering	No of sncRNAs mapping to CNV regions
miRNA	HiSeq	Tumor	254	2235	445	10
miRNA	GA	Tumor	215	2068	360	7
miRNA	HiSeq	Adjacent normal	18	1616	484	12
miRNA	GA	Adjacent normal	13	1370	430	12
piRNA	HiSeq	Tumor	254	65074	168	1
piRNA	GA	Tumor	215	47695	147	1
piRNA	HiSeq	Adjacent normal	18	9325	187	1
piRNA	GA	Adjacent normal	13	4063	122	1
snoRNA	HiSeq	Tumor	254	1182	210	10
snoRNA	GA	Tumor	215	1001	201	10
snoRNA	HiSeq	Adjacent normal	18	665	218	11
snoRNA	GA	Adjacent normal	13	558	177	8
tRNA	HiSeq	Tumor	254	609	380	12
tRNA	GA	Tumor	215	597	364	8
tRNA	HiSeq	Adjacent normal	18	563	386	12
tRNA	GA	Adjacent normal	13	524	305	6

This table summarizes the result of sncRNA sequencing analysis, indicating the number of sncRNAs profiled in the tissue, number of sncRNA retained after read count filtering criteria (5 Read Counts (RC) in at least 50% of samples) and number of sncRNAs originating from the associated CNV regions. The results were summarized for tumor and adjacent normal tissues, as well as for the two sequencing platforms, Illumina HiSeq and Genome analyzer. There are no common samples between the two sequencing platforms.

Table A.13 List of 38 expressed sncRNAs (in TCGA dataset) embedded within the breast cancer associated CNVs

chr	start	stop	strand	Small RNA	Tissue	Platform
14	101513666	101513688	+	hsa-miR-539-5p	Adjacent normal	HiSeq, GA
14	101514286	101514307	+	hsa-miR-889-3p	Tumor, Adjacent normal	HiSeq, GA
14	101515947	101515969	+	hsa-miR-655-3p	Tumor, Adjacent normal	HiSeq, GA
14	101518795	101518817	+	hsa-miR-487a-5p	Adjacent normal	HiSeq, GA
14	101520653	101520675	+	hsa-miR-382-5p	Tumor, Adjacent normal	HiSeq, GA
14	101520689	101520710	+	hsa-miR-382-3p	Adjacent normal	HiSeq, GA
14	101521031	101521053	+	hsa-miR-134-5p	Tumor, Adjacent normal	HiSeq, GA
14	101521069	101521092	+	hsa-miR-134-3p	Tumor, Adjacent normal	HiSeq, GA
14	101521801	101521823	+	hsa-miR-485-3p	Tumor, Adjacent normal	HiSeq, GA
14	101522606	101522628	+	hsa-miR-323b-3p	Tumor, Adjacent normal	HiSeq, GA
14	101526106	101526128	+	hsa-miR-154-5p	Tumor, Adjacent normal	HiSeq, GA
14	101526142	101526164	+	hsa-miR-154-3p	Adjacent normal	GA
19	4445984	4446007	+	hsa-miR-4746-5p	Tumor	HiSeq
19	8454224	8454245	-	hsa-miR-4999-5p	Adjacent normal	HiSeq
1	149680248	149680274	+	hsa-piR-20636	Tumor, Adjacent normal	HiSeq, GA
15	25296624	25296719	+	snoRNA_SNORD116-1-201	Tumor, Adjacent normal	HiSeq, GA
15	25299357	25299452	+	snoRNA_SNORD116-2-201	Tumor, Adjacent normal	HiSeq, GA
15	25302007	25302102	+	snoRNA_SNORD116-3-201	Tumor, Adjacent normal	HiSeq, GA
15	25310173	25310269	+	snoRNA_SNORD116-6-201	Tumor, Adjacent normal	HiSeq, GA
15	25315579	25315674	+	snoRNA_SNORD116-8-201	Tumor, Adjacent normal	HiSeq, GA
15	253182	2531834	+	snoRNA_SNORD116-	Tumor, Adjacent	HiSeq,

chr	start	stop	strand	Small RNA	Tissue	Platform
	54	9		9-201	normal	GA
15	25325289	25325381	+	snoRNA_SNORD116-14-201	Tumor, Adjacent normal	HiSeq, GA
15	25326434	25326526	+	snoRNA_SNORD116-15-201	Tumor, Adjacent normal	HiSeq, GA
16	2012335	2012468	-	snoRNA_SNORA10-201	Tumor, Adjacent normal	HiSeq, GA
16	2012974	2013108	-	snoRNA_SNORA64-201	Tumor, Adjacent normal	HiSeq, GA
19	3982505	3982571	-	snoRNA_SNORD37-201	Tumor, Adjacent normal	HiSeq, GA
1	148598314	148598388	-	chr1.trna108-AsnGTT	Tumor, Adjacent normal	HiSeq, GA
1	148760356	148760430	-	chr1.trna107-AsnGTT	Tumor, Adjacent normal	HiSeq, GA
1	149608609	149608683	+	chr1.trna30-AsnGTT	Tumor, Adjacent normal	HiSeq
1	149664355	149664428	-	chr1.trna94-GluTTC	Tumor, Adjacent normal	HiSeq
1	149672905	149672977	-	chr1.trna92-PheGAA	Tumor, Adjacent normal	HiSeq
1	149680210	149680281	-	chr1.trna91-GlyCCC	Tumor, Adjacent normal	HiSeq
1	149684088	149684162	-	chr1.trna90-ValCAC	Tumor, Adjacent normal	HiSeq, GA
6	26286754	26286826	+	chr6.trna2-MetCAT	Tumor, Adjacent normal	HiSeq, GA
19	1383361	1383434	-	chr19.trna14-PheGAA	Tumor, Adjacent normal	HiSeq, GA
19	1383562	1383636	+	chr19.trna1-AsnGTT	Tumor, Adjacent normal	HiSeq
19	4724082	4724154	+	chr19.trna2-GlyTCC	Tumor, Adjacent normal	HiSeq, GA
19	4724647	4724720	-	chr19.trna13-ValCAC	Tumor, Adjacent normal	HiSeq, GA

This table lists the different classes of CNV-sncRNAs expressed in breast tumor and adjacent normal tissues, which were profiled using Illumina HiSeq and Genome analyzer platforms.

Table A.14 Gene Dosage analysis for CNV-sncRNAs

CNV region	Expressed CNV-sncRNAs	Pearson Correlation Coefficient (r)	p-value (correlation)	No of samples
1-148662374-148789654	chr1.trna107-AsnGTT	0.14	5.44E-02	198
1-148580449-148632305	chr1.trna108-AsnGTT	0.17	1.40E-02	198
1-149676729-149684202	chr1.trna91-GlyCCC	0.21	3.52E-03	198
1-149652461-149676729	chr1.trna92-PheGAA	0.20	3.90E-03	198
19-4714925-4751218	chr19.trna13-ValCAC	-0.18	1.06E-02	198
19-1342160-1547869	chr19.trna14-PheGAA	-0.19	6.07E-03	198
19-4714925-4751218	chr19.trna2-GlyTCC	-0.21	2.94E-03	198
6-26274458-26287456	chr6.trna2-MetCAT	0.18	1.20E-02	198
1-149676729-149684202	hsa-piR-20636	0.21	2.64E-03	198
15-25298903-25300158	snoRNA_SNORD116-2-201	-0.13	7.29E-02	198
15-25318258-25325686	snoRNA_SNORD116-9-201	-0.34	1.03E-06	198
15-25300158-25304384	snoRNA_SNORD116-3-201	-0.25	3.66E-04	198
15-25308383-25310928	snoRNA_SNORD116-6-201	-0.40	5.05E-09	198
15-25310928-25318258	snoRNA_SNORD116-8-201	-0.45	2.46E-11	198
19-3768181-4110048	snoRNA_SNORD37-201	-0.15	3.16E-02	198

Gene Dosage for the embedded CNV-sncRNAs were estimated by correlating the germline copy status and sncRNA expression data (HiSeq, n=198, RPKM normalized log transformed) using Pearson correlation. We observed significant positive and negative correlation among the correlated sncRNAs.

Table A.15 Gene targets for expressed CNV-miRNAs

miRNA	Copy status (no of samples)	No of Predicted and expressed targets	No of correlated targets	p-value	Pearson Correlation coefficient (r)	Correlated target genes considered for IPA analysis
hsa-miR-134-3p	Diploid (n=195)	4444	61	$<10^{-2}$	-0.20 to -0.27	*NAA40, TTF2, POLE,CDCA5,KDM2B,SETD8,ACP1,NCAPG2,TMED4,PGAM5,WDR77,DDX11,CDK5,GSG2,PTCD3,AGK,UBE2C,SRPK1,FARSB,SNRPD1,ELAVL1DLD,RAN, USP13,TBRG4,C18orf25,PLCXD1,NUDT19,ZNF131, TROAP,VPS33A,DUS4L,TRIP13,RBBP4,ANKRD45,C11orf48,MOV10,ZNF695,FAM64A,MRS2,NUF2,DOCK3,PPIL1,MAP4K2,KNTC1,FBXO41,RSPO4,ABCF2,ZSCAN16,KIAA1549,NCAPH,FBRSL1,ZNF76,ATAD3B,ULK3,FANCA,RNF165,ATP5F1,PFDN6,PSMG1,FAF1*
hsa-miR-134-5p	Diploid (n=195)	176	3	$<10^{-2}$	-0.20 to -0.22	*DPH2, NIPA1, EXD1*
hsa-miR-154-3p	Diploid (n=195)	23	0			
hsa-miR-323b-3p	Diploid (n=195)	2638	0			
hsa-miR-382-3p	Diploid (n=195)	202	2	$<10^{-2}$	-0.20 to -0.25	*OCIAD2, HMGN3*
hsa-miR-485-3p	Diploid (n=195)	389	6	$<10^{-2}$	-0.20 to -0.22	*C18orf25, AGAP3, PEX5, FXR2, POM121C, POM121*
hsa-miR-539	Diploid (n=195)	3082	0			
hsa-miR-655	Diploid (n=195)	805	3	$<10^{-2}$	-0.20 to -0.22	*VKORC1L1, DLD, WHSC1*
hsa-miR-889	Diploid (n=195)	4339	0			
hsa-miR-4746	Diploid (n=146)	699	25	$<10^{-2}$	-0.20 to -0.34	*NRIP2, TXNDC15,NISCH,MXD4,CLEC14A,CD34,APBB2,ZNF446,EDNRB,RAX2,PCDH1,CDH5,ADAMTS13,AQP1*

miRNA	Copy status (no of samples)	No of Predicted and expressed targets	No of correlated targets	p-value	Pearson Correlation coefficient (r)	Correlated target genes considered for IPA analysis
						,PALM,PDE11A,UNKL,F10,GIPR,PHF2,PDPK1,PHF1,LMX1B,NUDT16L1,AKAP12
hsa-miR-4746	Copy gain (n=52)	699	54	< 10^{-2}	-0.27 to -0.42	*KLF10,NISCH,PCDH1,NRIP2,ZNF407,SLC35E2,CLEC14A,PTGER3,GIGYF1,ZNF423,UBR1,TBC1D2B,CASZ1,IQSEC1,ADAMTS13,ADAMTSL1,CDH5,RAX2,CD34,LMX1B,ZBTB46,RPS6KA2,MXD4,ANKRD52,KBTBD11,CEP120,WDR81,SLC35E2,IGF1R,PDPK1,ERLIN2,EDNRB,LMTK2,MADD,PDE11A,MNT,ATOH8,CRX,CAMK2N1,CXorf23,CBX6,PHF2,KIAA1429,MOCS1,MGRN1,SERTAD1,SHOX,AQP1,ZHX3,ZBTB20,GLUL,PRDM2,KSR2,TMEM184A*

Table A.16 Ingenuity Pathway Analysis for the target genes regulated by CNV-miRNAs

miRNA name	Pathway	P-value	Target genes
hsa-miR-655 (Diploid)	Branched-chain α-keto acid Dehydrogenase Complex	5.62 E-04	DLD
	2-ketoglutarate Dehydrogenase Complex	7.08 E-04	DLD
	2-oxobutanoate Degradation I	7.08 E-04	DLD
	Glycine Cleavage Complex	8.51 E-04	DLD
	Acetyl-CoA Biosynthesis I (Pyruvate Dehydrogenase Complex)	9.77 E-04	DLD
	Isoleucine Degradation I	2.00 E-03	DLD
	Valine Degradation I	2.51 E-03	DLD
	TCA Cycle II (Eukaryotic)	3.24 E-03	DLD
	Super pathway of Methionine Degradation	4.47 E-03	DLD
hsa-miR-134-3p (Diploid)	Cell Cycle Control of Chromosomal Replication	4.79 E-03	CDK5, POLE
	Branched-chain α-keto acid Dehydrogenase Complex	1.10 E-02	DLD
	2-ketoglutarate Dehydrogenase Complex	1.35 E-02	DLD
	2-oxobutanoate Degradation I	1.35 E-02	DLD
	Glycine Cleavage Complex	1.62 E-02	DLD
	Acetyl-CoA Biosynthesis I (Pyruvate Dehydrogenase Complex)	1.91 E-02	DLD
	BER pathway	3.24 E-02	POLE
	NAD Phosphorylation and Dephosphorylation	3.47 E-02	ACP1
	Isoleucine Degradation I	3.72 E-02	DLD
	RAN Signaling	4.27 E-02	RAN
	Valine Degradation I	4.79 E-02	DLD

miRNA name	Pathway	P-value	Target genes
hsa-miR-4746 (Copy gain)	Growth Hormone Signaling	1.15 E-03	IGF1R, PDPK1, RPS6KA2
	Glutamine Biosynthesis I	2.48 E-03	GLUL
	FLT3 Signaling in Hematopoietic Progenitor Cells	1.89 E-02	PDPK1, RPS6KA2
	IGF-1 Signaling	2.85 E-02	IGF1R,PDPK1
	G-Protein Coupled Receptor Signaling	3.02 E-02	PDE11A,PDPK1,PTGER3
	NGF Signaling	3.55 E-02	PDPK1,RPS6KA2
	PTEN Signaling	3.55 E-02	IGF1R,PDPK1
hsa-miR-4746 (Diploid)	Cardiac β-adrenergic Signaling	1.12 E-02	AKAP12,PDE11A
	eNOS Signaling	1.41 E-02	AQP1,PDPK1
	Extrinsic Prothrombin Activation Pathway	1.86 E-02	F10
	Agranulocyte Adhesion and Diapedesis	2.04 E-02	CDH5,CD34
	RAR Activation	2.09 E-02	NRIP2,PDPK1
	cAMP-mediated signaling	2.82 E-02	AKAP12,PDE11A
	Intrinsic Prothrombin Activation Pathway	3.39 E-02	F10
	G-Protein Coupled Receptor Signaling	4.07 E-02	PDPK1,PDE11A
	Coagulation System	4.07 E-02	F10
	tRNA Splicing	4.47 E-02	PDE11A

This table represents the findings from the IPA; represented are the pathways significantly enriched at p-value <0.05. For the hsa-miR-4746, we performed the analysis based on the targets identified in each of the copy number groups (diploid and copy gain). For the other two miRNAs, we used the identified targets genes based on cases with diploid copy status.