FILIPA ESTEVES

UNVEILING THE ROLE OF CIS-REGULATORY VARIATION IN BREAST CANCER AETIOLOGY

UNIVERSIDADE DO ALGARVE

Faculdade de Medicina e Ciências Biomédicas

2021

FILIPA ESTEVES

UNVEILING THE ROLE OF CIS-REGULATORY VARIATION IN BREAST CANCER AETIOLOGY

PhD Programme in Mechanisms of Disease and Regenerative Medicine

Work developed under the supervision of:

Ana Teresa Maia, PhD

UNIVERSIDADE DO ALGARVE

Faculdade de Medicina e Ciências Biomédicas

2021

UNVEILING THE ROLE OF CIS-REGULATORY VARIATION IN BREAST CANCER AETIOLOGY

Declaration of authorship

I hereby declare that I am the author of this work, which is my original and unprecedented. The authors and publications consulted are duly cited in the text and are listed in the included references.

To my Mum

"The whole purpose of education is to turn mirrors into windows"

Sydney J. Harris

“Se um arbusto é amarelo, nunca será totalmente verde”

António Canário

Acknowledgments

A quem me apoiou e desafiou ao longo destes anos.

Em primeiro lugar, gostaria de agradecer à minha orientadora, Professora Doutora Ana Teresa Maia, por me ter recebido no seu grupo e permitido trabalhar neste projecto, tendo-me disponibilizado os meios para prosseguir e investir no meu interesse em Ciência. Obrigada pela oportunidade proporcionada para explorar novas áreas científicas, e por despertar o meu interesse pela Bioinformática. Obrigada pelo apoio, e por ter acreditado e confiado em mim durante este processo.

O meu obrigada aos meus colegas de grupo Joana Xavier, Ramiro Magno, Isabel Duarte, Marinella Ghezzo, Ana Fernandes e André Duarte, por todas as discussões científicas, pelas ideias e comentários, que sempre promoveram a minha evolução, bem como a do meu trabalho.

Agradeço à Direcção do ProRegeM, Doutor José Belo, António Jacinto, José Bragança e Gabriela Silva por me terem aceite neste Programa Doutoral, e proporcionado esta aprendizagem.

Obrigada à Juliana Machado e Catarina Martins pelos hashtags, e pelos momentos de descontração e riso, tão importantes em alguns momentos.

O meu muito obrigada à minha amiga Lizelle Correia pela amizade, pelo apoio e ânimo dados, pela partilha de muitos momentos em que as tuas palavras foram tão importantes. Obrigada pela ajuda, por me ouvires e me fazeres rir, mesmo em alturas de preocupação.

O meu obrigada à Inês Grenho, por ter estado “lá” ao longo deste último ano. Obrigada pela compreensão, pelos desabafos, pelas tonterias, pelos vídeos e imagens partilhadas que tanto alegraram momentos difíceis. Obrigada por tornares os dias mais leves.

À Flor por ter continuado a marcar presença ao longo destes anos, mesmo quando o tempo era pouco e a distância era muita.

Ao Billy por me ter proporcionado dias de reflexão, de introspecção e felicidade numa altura particularmente difícil.

Finalmente, o meu imenso agradecimento à minha pessoa, a minha Mãe. Obrigada por me estimulares a ser quem sou, e defender aquilo em que acredito. Obrigada por me motivares a sempre mais e melhor, e pelo modo como o fazes, sempre sem cobrança. Obrigada por todas as conversas e pelas inúmeras horas de troca de opiniões. Obrigada por, desde que me entendo por gente, me incentivares a procurar respostas. A minha curiosidade e interesse em aprender

e investigar, foram certamente estimulados pelas incontáveis vezes em que me fizeste procurar a resposta às infindáveis perguntas que te fazia. Obrigada Mãe, pela paciência, pelos muitos dias e noites em que estiveste sempre do meu lado. Obrigada por me aturares e me apoiares sempre, mesmo nos piores momentos de frustração. Obrigada pelo carinho, pela dedicação e pela força que sempre me deste incondicionalmente. Obrigada pelo amor incondicional e encorajamento, por seres a pessoa extraordinária que és e por sempre acreditares em mim. Obrigada por teres sempre estado lá para mim e por me acompanhares em mais este processo. Sem ti não teria sido possível.

List of Publications and Patent

Results from the present study contributed to the following publications:

Chapter IV. Identification of Target Genes and Causal Variants in Known Breast Cancer Risk Loci

1. Portuguese Patent Pending 2020 Request No 115881 – "METHODS OF ANALYSIS OF POLYMORPHISMS ASSOCIATED WITH CANCER AND USES THEREOF", submitted 02/11/2019
Inventors: Ana-Teresa Maia, **Filipa Esteves**, Joana Xavier.

2. **Filipa Esteves**, Joana M. Xavier, Anthony M. Ford, Cátia Rocha, Mel F. Greaves, Paul Pharoah, Carlos Caldas, Suet-Feung Chin, Ana-Teresa Maia. *Allelic expression of genes in the 17q22 locus predicts breast cancer risk.*
Manuscript in submission.
Author's contribution: Data collection, analysis, and interpretation of in-silico and in-vitro functional analysis, participated in manuscript drafting.

3. Lizelle Correia, Joana M. Xavier, Ramiro Magno, Bernardo P. de Almeida, **Filipa Esteves**, Isabel Duarte, Matthew Eldridge, Chong Sun, Astrid Bosma, Lorenza Mittempergher, Ana Marreiros, Rene Bernards, Carlos Caldas, Suet-Feung Chin and Ana-Teresa Maia. *Allelic expression imbalance of PIK3CA mutations is frequent in breast cancer and prognostically significant.*
Manuscript under review in *npj Breast Cancer*.
Author's contribution: Data collection, analysis, and interpretation of laboratory work for in-vitro validation of the candidate rSNP.

Chapter V. Pipeline Benchmarking for AE Analysis from RNA-Seq Data

1. **Filipa Esteves**, Ramiro Magno, Joana M. Xavier, Carlos Caldas, Suet-Feung Chin, Ana-Teresa Maia. *A pipeline benchmarking for allelic expression analysis from RNA-Seq data*.
Manuscript in preparation.
Author's contribution: Study design, data collection, analysis and interpretation, manuscript drafting.

Abstract

In 2020 female breast cancer moved from second to the most commonly diagnosed cancer worldwide. Although an estimated 30% of breast cancer cases are heritable or due to underlying genetic factors, approximately half of the familial risk for breast cancer still remains unknown. Since 2007, continuous efforts from genome-wide association studies (GWAS) and the Collaborative Oncological Gene-Environment Study (COGS) identified low-risk loci that explain up to 18% of the familial relative risk. But, most of the risk-associated variants identified by GWAS are not the true causal variants, and therefore, functional variants and the biological mechanisms underlying breast cancer susceptibility remain largely unknown. Since most of the variants identified by GWAS lie on non-coding genomic regions, risk-associated variants likely have a cis-regulatory function, as shown by several post-GWAS studies focusing on the identification of the causal variants. In this context, the main goal of this project was to develop and use a new and efficient approach to detect target genes and causal variants in known and new breast cancer predisposition loci.

Firstly, to address the aforementioned challenges, this work intended to Identify causal variants acting in candidate loci with strong cis-regulatory potential and association with published and unpublished GWAS. Allelic expression (AE) ratios were used as a quantitative variable in case-control association studies to understand how genetic variation can control gene expression and to identify cis-regulatory variants and their target genes. For the 17q22 locus two potential regulatory variants – rs17817901 and rs8066588 – altering a miRNA and a transcription factor binding site, respectively, were identified. Additionally, results showed that *STXBP4* and *COX11* are the most likely target genes in this locus. A significant association was found in normal breast tissue between the preferential expression of the reference alleles of two single nucleotide polymorphisms (SNPs) located on the 17q22 locus – rs17817901 (*TOM1L1/COX11*) and rs2628315 (*STXBP4*) –, and increased risk for breast cancer. This association was also observed in blood samples, which shows the possibility of using this approach in the screening of the general population for breast cancer risk. These results showed that integrating AE ratios as a quantitative variable in case-control association studies is a powerful approach to identify novel risk loci.

Next, to perform a genome-wide AE analysis from RNA-sequencing (RNA-seq) data, a comprehensive comparison of variant calling pipelines was conducted. Forty-two variant calling pipelines were systematically compared using data from a gold standard and a normal breast

tissue sample. This allowed establishing the most suitable analysis pipeline for further studies aiming at precise AE quantification using RNA-seq data.

Finally, this work aimed to identify new loci associated with breast cancer risk, using a genome-wide approach. RNA-seq data from 12 normal breast tissue samples of healthy women (controls) and 14 breast cancer patients (cases) was analysed and AE ratios were calculated genome-wide across 7,054 genetic variants. Eight candidate variants associated with breast cancer risk were identified, and for those, the previously proposed case-control association analysis using AE ratios was conducted. This identified *CDC16* as the strongest candidate new locus associated with breast cancer risk, with a predicted effect size of -1.83 [95% CI=-2.38, -1.14].

Results from this work provide further evidence that cis-regulatory variation plays a major role in breast cancer susceptibility and shows the power of integrating allelic expression data in cancer risk studies, particularly in identifying risk causal variants and their target genes. Furthermore, it presents a novel efficient approach to identify risk – case-control association analysis using AE ratios. Overall, besides providing important new knowledge on the biological mechanism underlying the risk of breast cancer, which will improve the identification and management of the population at risk, it also provides concepts and approaches that are applicable to other cancers and complex diseases.

Keywords: breast cancer; risk-associated loci; allelic expression; cis-regulation; causal variants; target genes.

Resumo

Em 2020, o cancro da mama passou a ser o cancro mais diagnosticado em todo o mundo. No total, estima-se que foram diagnosticados 2.261.419 (11,7% de todos os cancros), sendo que as 684.996 mortes decorrentes de cancro da mama (6,9% de todas as mortes por cancro), o colocou em quinto lugar como causa mais comum de morte por cancro.

O cancro da mama é uma doença multifactorial, com vários factores de risco ambientais e de estilo de vida, para além dos genéticos. Estes últimos são responsáveis por, aproximadamente, 30% de todos os casos de cancro da mama, sendo que 5-10% dos casos totais apresentam um padrão Mendeliano de heritabilidade. No entanto, as alterações responsáveis pelo risco genético não são completamente claras. Vários estudos levaram à identificação de variantes associadas ao risco para o cancro da mama, as quais são divididas em variantes de penetrância alta, moderada e baixa, consoante a sua frequência e risco associado. As variantes de alta penetrância têm uma frequência do alelo menor (MAF, do inglês *minor allele frequency*) na população geral abaixo dos 0,005 e conferem um risco relativo acima de 5. As variantes de penetrância moderada apresentam uma MAF de 000,5-0,01 e conferem um aumento de risco entre 2 a 4 vezes. As variantes de baixa penetrância são comuns na população, com uma MAF igual ou superior a 0,05 e aumentam o risco para cancro da mama menos de 1,5 vezes. Estas variantes são geralmente polimorfismos de nucleotídeos únicos (SNPs, do inglês *single nucleotide polymorphisms*). As variantes de alta e moderada penetrâncias explicam menos de 30% de todo o risco familiar para o cancro da mama, pelo que a restante fracção é provavelmente explicada por um grande número de variantes, que conferem, cada uma, um baixo risco. Este tipo de variantes tem sido maioritariamente identificado pelos estudos de associação do genoma completo (GWAS, do inglês *genome-wide association studies*). Estes estudos comparam a frequência dos SNPs entre pessoas com a doença (casos) e sem a doença (controlos), e partem do pressuposto que a variante causal se encontra num haplótipo, e que, consequentemente, uma variante que represente o haplótipo e que esteja em desequilíbrio de ligação (LD, do inglês *linkage disequilibrium*) com a variante causal, apresenta, por aproximação, uma associação com o fenótipo de interesse. No entanto, e a apesar de todos os esforços realizados pelos GWAS, desde 2007, na identificação de variantes comuns que possam explicar a percentagem de risco familiar desconhecida para o cancro da mama, os loci associados a risco por estes estudos representam apenas 18% do risco genético. Para além disso, a maioria das variantes dos GWAS não são as variantes funcionais causadoras do risco, permanecendo também desconhecidos os mecanismos biológicos responsáveis pela associação ao risco. Ainda,

a maioria das variantes identificadas pelos GWAS encontram-se em regiões não codificantes do genoma, o que sugere que as variantes causais associadas ao risco têm uma função cis-regulatória – regulam a expressão de genes perto e longe – o que tem sido corroborado por estudos funcionais subsequentes, cujo objectivo se focou na identificação das variantes causais.

Neste contexto, o presente estudo assenta na hipótese de que as variantes cis-regulatórias, por alterarem a expressão genética, desempenham um papel fundamental na susceptibilidade para o cancro mama. Assim, o objectivo central deste projecto foi desenvolver e utilizar uma abordagem nova e eficiente para detectar as variantes causais e os genes alvo em novos loci e em loci previamente associados ao risco para cancro da mama.

Inicialmente, este trabalho focou-se na identificação de variantes causais em loci com forte potencial cis-regulatório e previamente associados com risco por GWAS. Para isso, rácios de expressão alélica (AE, do inglês *allelic expression*) foram utilizados como variável quantitativa em estudos de associação caso-controlo em amostras de tecido mamário normal e em sangue, para validar as variantes cis-regulatórias candidatas, bem como os genes sob o seu efeito.

Dados de uma análise exploratória de expressão alélica diferencial em amostras de tecido mamário de mulheres saudáveis, identificou um forte potencial cis-regulatório em 12 loci associados anteriormente por GWAS a risco para cancro da mama, tendo o maior potencial sido verificado para os loci 1q32.1, 16q23.2, e 17q22. Para este último, foram identificadas duas possíveis variantes regulatórias – rs17817901 e rs8066588 – responsáveis por alterar, respectivamente, o local de ligação de um microRNA e de um factor de transcrição. Os genes *COX11* e *STXBP4* foram identificados como os mais prováveis genes alvo neste locus. Foi encontrada uma associação significativa em tecido mamário normal entre a expressão preferencial dos alelos de referência das variantes rs17817901 (*TOM1L1/COX11*) e rs2628315 (*STXBP4*), e um aumento de risco para o cancro da mama. Esta associação foi igualmente observada em amostras de sangue, demonstrando o potencial impacto da utilização desta abordagem no rastreio do risco de cancro da mama na população. Este trabalho mostra que integrar rácios de expressão alélica como uma variável quantitativa em estudos de associação caso-controlo, constitui uma abordagem poderosa na identificação de novos loci de risco.

Seguidamente, este trabalho também validou o SNP rs2699887 como variante regulatória do gene *PIK3CA*, por alteração da ligação do factor de transcrição NF-YA. As implicações funcionais desta variante incluem a modelação de resposta a fármacos específicos cujo alvo são as mutações no gene *PIK3CA*.

O objectivo seguinte deste projecto centrou-se na identificação de um pipeline adequado para a análise de expressão alélica a partir de dados de sequenciação de amostras de RNA (RNA-seq do inglês *RNA-sequencing*). Os dados de RNA-seq permitem quantificar a AE, através da contabilização do número de sequências (*reads*) que alinham em cada um dos alelos de indivíduos heterozigóticos. Dado que a determinação precisa de AE requer uma análise adequada dos dados de RNA-seq, e não existindo um único processo aplicável a todos os casos, este trabalho incluiu uma comparação detalhada de 42 sequências de ferramentas computacionais (*pipelines*).

Uma vez determinado o pipeline mais adequado para análise de dados de RNA-seq direccionada para a quantificação de AE, este trabalho teve também como objectivo identificar novos loci associados com risco para o cancro da mama. Dados de RNA-seq de amostras de tecido mamário normal de 12 mulheres saudáveis (controlos) e 14 mulheres com cancro da mama (casos) foram analisados, tendo sido quantificada a AE e determinados os rácios de AE em todo o genoma. Das 7.054 variantes genéticas para as quais os rácios de AE foram determinados, 353 apresentaram uma diferença significativa (p-value<0.05) entre os rácios de AE dos casos e dos controlos, sendo que para sete dessas variantes, essa diferença era superior a quatro vezes. A análise de seis destas sete variantes numa segunda fase de validação usando PCR em tempo-real para determinar AE num grupo de amostras maior, identificou a variante rs3211416 (*CDC16*), como associada a risco para cancro da mama, com um efeito previsto de pelo menos 1.83 [95% CI=-2.38, -1.14]. Os resultados mostram ainda que este risco estará associado ao alelo C do rs3211416 e à expressão mais elevada do gene *CDC16*, sugerindo assim um papel oncogénico para este gene no contexto do cancro da mama.

Os resultados deste trabalho fornecem evidências adicionais de que a variação cis-regulatória desempenha um papel crucial na suscetibilidade para o cancro de mama, e mostra o poder de integrar dados de expressão alélica em estudos de risco de cancro, particularmente na identificação de variantes causais de risco e seus genes alvo. Adicionalmente, apresenta uma nova abordagem altamente eficiente para identificar loci de risco – análise de associação de caso-controlo usando rácios de AE. No geral, além de gerar novo conhecimento sobre os mecanismos biológicos subjacentes ao risco de cancro de mama, o que irá melhorar a identificação e gestão da população em risco, também fornece conceitos e abordagens que são aplicáveis a outros cancros e doenças complexas.

Palavras-chave: cancro da mama; loci associados a risco; expressão alélica; cis-regulação; variantes causais; genes alvo.

Table of Contents

List of Figures

Chapter I. General Introduction

Chapter IV. Identification of Target Genes and Causal Variants in Known Breast Cancer Risk Loci

Chapter V. Pipeline Benchmarking for AE Analysis from RNA-Seq Data

Chapter VI. Identification of New Susceptibility Loci for Breast Cancer

Supplementary Material

List of Tables

Chapter I. General Introduction

Chapter IV. Identification of Target Genes and Causal Variants in Known Breast Cancer Risk Loci

Chapter V. Pipeline Benchmarking for AE Analysis from RNA-Seq Data

Chapter VI. Identification of New Susceptibility Loci for Breast Cancer

Supplementary Material

Abbreviations

AE	allelic expression
AEI	allelic expression imbalance
APC/C	anaphase-promoting complex/cyclosome
APOC1	apoliproprotein C1
APOC1P1	apoliproprotein C1 pseudogene 1
ARL4A	ADP ribosylation factor like GTPase 4A
ASR	age-standardised rates
AKT	protein kinase B alpha
ANKRD16	ankyrin repeat domain 16
ASE	allele-specific expression
ATM	ataxia telangiectasia mutated
BAM	Binary Alignment Map
BBCS	British Breast Cancer Study
BC	breast cancer
BCAC	Breast Cancer Association Consortium
BCL	per-cycle base call
BLM	bloom syndrome
bp	base-pair
BRCA1	breast cancer 1
BRCA2	breast cancer 2
BRIP1	BRCA1 interacting protein c-terminal helicase 1
C16orf46	chromosome 16 open reading frame 46
CASP8	caspase 8
CCDC86	coiled-coil domain containing 86
CCNG2	cyclin G2
CDC16	cell division cycle 16

CDH1	cadherin 1
CDKN2A	cyclin dependent kinase inhibitor 2A
CDKN2B	cyclin dependent kinase inhibitor 2B
CDKN2BAS	cyclin dependent kinase inhibitor 2B Antisense RNA
cDNA	complementary DNA
CEPH	Centre d'Etude du Polymorphisme Humain
CGEMS	Cancer Genetic Markers of Susceptibility
CHEK2	checkpoint kinase 2
ChIP	chromatin immunoprecipitation
ChIP-seq	ChIP-sequencing
CLPTM1L	cleft lip and palate transmembrane protein 1-like protein
CNA	copy number aberration
COGS	Collaborative Oncological Gene-environment Study
COX11	cytochrome C oxidase subunit 11
Ct	cycle threshold
CTCF	CCCT-binding factor
DAE	differential allelic expression
DCIS	ductal carcinoma in situ
DDSB	DNA double-strand break
DHS	DNase I hypersensitive sites
DNA	deoxyribonucleic acid
DNA-seq	DNA-sequencing
dNTP	deoxynucleotide
DP	depth
ds	double-stranded
DTT	dithiothreitol
ECRIC	Eastern Cancer Registration and Information Centre

EDTA	ethylenediamine tetraacetic acid
e.g.	*exempli gratia* (for example)
EGR1	early growth response 1
EMSA	electrophoretic mobility shift assay
ENA	European Nucleotide Archive
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantitative trait locus
ER	oestrogen receptor
ESR1	oestrogen receptor alpha
ESR2	oestrogen receptor beta
EU	European Union
FBXO18	F-box DNA helicase 1
FC	fold-change
FDR	false discovery rate
FGFR2	fibroblast growth factor receptor 2
FOS	fos proto-oncogene, AP-1 transcription factor subunit
GATK	Genome Analysis Toolkit
GCSH	glycine cleavage system protein H
gDNA	genomic DNA
GIAB	Genome In a Bottle
GSNAP	Genomic Short-read Nucleotide Alignment Program
GTEx	The Genotype-Tissue Expression project
GWAS	genome-wide association studies
H3K4me1	histone H3 lysine 4 mono-methylation
H3K4me3	histone H3 lysine 4 trimethylation
H3K9ac	histone H3 lysine 9 acetylation
H3K27ac	histone H3 lysine 27 acetylation

HCl	hydrochloric acid
HCS	HiSeq control software
HDI	human development index
HER2	human epidermal growth factor receptor 2
HGP	Human Genome Project
HGMA	high mobility group AT-hook 1
Hmec	Human mammary epithelial cell
HRT	hormone replacement therapy
iCOGS	Illumina iSelect genotyping array
IntClust	integrative clusters
JUN	jun proto-oncogene, AP-1 transcription factor subunit
Kb	kilobases
LCIS	lobular carcinoma in situ
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
LFS	Li-Fraumeni syndrome
lincRNAs	large intergenic non-coding RNAs
LRRN2	leucine-rich repeat neuronal 2
LSP1	lymphocyte specific protein 1
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MAF	minor allele frequency
MAP3K1	mitogen-activated protein kinase kinase kinase 1
MDM4	mouse double minute 4, human homolog of
miRNA	microRNA
min	minute
MQ	mapping quality
MRE11	meiotic recombination 11 homolog 1

mRNA	messenger RNA
MRPS30	mitochondrial ribosomal protein S30
mTOR	mammalian target of rapamycin
MYC	myc proto-oncogene protein
NBN	nibrin
NEK10	never in mitosis A-related kinase 10
NFYA	nuclear transcription factor Y subunit alpha
NRIP1	nuclear receptor interacting protein 1
NTC	no-template control
OR	odd ratio
PALB2	partner and localizer of BRCA2
PAM50	Prediction Analysis of Microarray 50
PCR	polymerase chain reaction
PDE4DIP	phosphodiesterase 4D interacting protein
PIK3	phosphatidylinositol-3-kinase
PIK3CA	phosphatidylinositol-3-kinase alpha
PIK3C2B	phosphatidylinositol-4-Phosphate 3-kinase catalytic subunit type 2 beta
POLR2A	RNA polymerase II subunit A
PR	progesterone receptor
PTEN	phosphatase and tensin homolog
PTHLH	parathyroid hormone-like protein
PWM	position weight matix
QC	quality control
QD	QualByDepth
qPCR	quantitative PCR
RECQL	recq like helicase
RNA	ribonucleic acid

RNA-seq	RNA-sequencing
RNPII	RNA polymerase II
rSNP	regulatory SNP
RTA	real time analysis
SAM	Sequence Alignment Map
SBS	sequencing by synthesis
SD	standard deviation
SDS	sodium dodecyl sulphate
SEARCH	Studies of Epidemiological and Risk factors in Cancer Heredity
sec	second
SNP	single nucleotide polymorphism
STAR	Spliced Transcripts Alignment to a Reference
STAT3	signal transducer and activator of transcription 3
STK11	serine/threonine kinase 11
STXBP4	syntaxin binding protein 4
TAE	tris-acetate-EDTA
TBC	Tre2-Bub2-Cdc16
TBC1D7	TBC family member 7
TBC1D12	TBC1 domain family member 12
TBE	tris-borate-EDTA
TCF7L2	transcription factor 7 like 2
TCGA	The Cancer Genome Atlas
TdT	deoxynucleotidyl transferase
TERT	telomerase reverse transcriptase
TF	transcription factor
TGFB1	transforming growth factor beta 1
TNBC	triple negative breast cancers

TNM	tumour–node–metastasis
TOM1L1	target of Myb1 like 1 membrane trafficking protein
TOP2A	DNA topoisomerase II alpha
TOX3	TOX high mobility group box family member 3
TP53	Tumour Suppressor 53
TSC	The SNP Consortium
tSNP	transcribed SNP
UK	United Kingdom
USA	United States of America
UTR	untranslated region
WES	whole exome sequencing
WGS	whole genome sequencing
WHO	World Health Organization
ZMIZ1	zinc finger MIZ-type containing 1
ZNF217	zinc finger protein 217
ZNF365	zinc finger protein 365

Chapter I

General Introduction

2.1. Cancer

Cancer is a complex and heterogenous disease, involving dynamic changes in the genome. It is characterized by the uncontrolled cell growth caused by abnormal changes in regulatory systems responsible for promoting and controlling normal cell proliferation and homeostasis (Hanahan & Weinberg, 2000). There are more than 200 types of cancer, which are usually identified by the name of the tissue from which the abnormal cells originated.

The existence of a set of six distinct and complementary capabilities acquired by normal cells, that together drive the appearance of cancer, was first suggested by Hanahan and Weinberg in 2000 (Hanahan & Weinberg, 2000). These six hallmarks of cancer comprised: sustaining proliferative signalling; evading growth suppressors; resisting apoptosis; enabling replicative immortality; inducing angiogenesis; and activating invasion and metastasis. Progress on cancer research, led to the reformulation of the hallmark capabilities, with the inclusion of two additional hallmarks – avoiding immune destruction and deregulating cellular energetics –, and two enabling characteristics – tumour-promoting inflammation, and genome instability and mutation – (Hanahan & Weinberg, 2011).

Among the cancer hallmarks, activating invasion and metastasis is the one that distinguishes cancer from benign tumours. By definition, metastasis is the dissemination of neoplastic cells to nearby or distant, discontiguous secondary sites, where they proliferate to establish an extravascular mass of incompletely differentiated cells (Welch D. R., 2006). The first attempt to define the hallmarks of metastasis identified four distinguishing features: motility and invasion; ability to modulate the secondary site or local microenvironments; plasticity; and stability to colonize secondary tissues (Welch & Hurst, 2019).

Data from the GLOBOCAN 2020 estimates suggest an increase of the global cancer burden to 19.3 million cases and 10 million cancer deaths. The rapid growth of cancer incidence and mortality reflects both aging and the growth of the population (Sung, et al., 2021). One of the primary prevention strategies, applied to several cancer types, including breast cancer (BC), is the routine screening aiming at the early detection of malignant or precursor lesions prior to the onset of symptoms, when the treatment is more likely to be effective (Shieh, et al., 2016). Increasing incidence rates do not necessarily reflect failure of primary prevention strategies in reducing cancer incidence, since early detection (tests or programs) can lead to a transient rise in incidence rates as subclinical cancer cases are discovered. On the other hand, also maintaining

this inflation is overdiagnosis, diagnosis of cancers that would not otherwise cause symptoms or death during the individual's lifespan (Bray, et al., 2018).

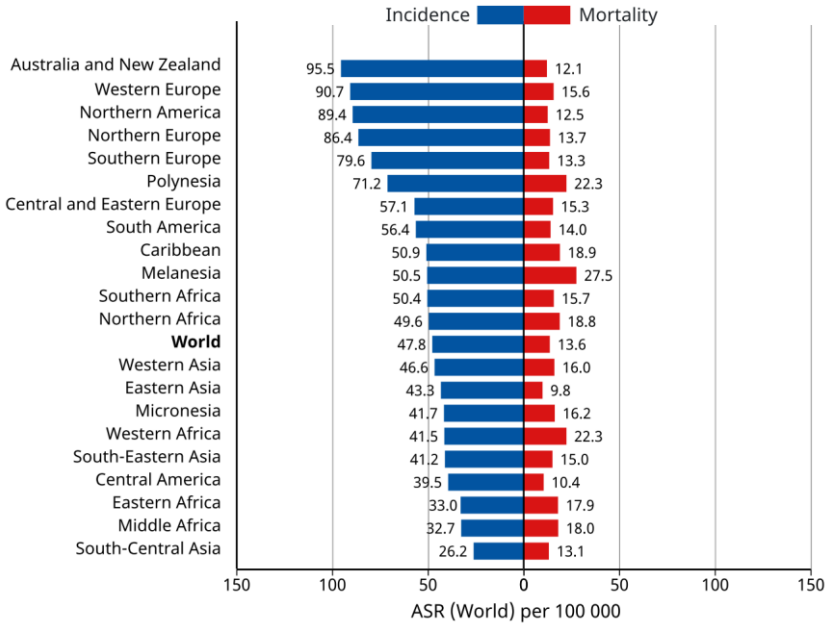
1.2. Breast Cancer

1.2.1. Epidemiology

In 2020, female breast cancer surpassed lung cancer and became the most commonly diagnosed cancer, with 2,261,419 (11.7%) new cases, also ranking fifth in mortality, with 684,996 (6.9%) deaths estimated worldwide (Sung, et al., 2021). Among females, breast cancer is the most commonly diagnosed cancer in the majority of the countries (159 of 185), accounting for 24.5% of overall new cancer cases, and the leading cause of cancer death in 110 countries, representing the highest percentage (15.5%) of cancer deaths in the world. Elevated incidence rates in countries with a high human development index (HDI, summary measure of average achievement considering life expectancy, education and gross national income per capita) (Figure 1.1A) are due to a higher prevalence of known risk factors. The rapid incidence increase in regions where the rates have been historically relatively low, such as South America, Africa, and Asia, are likely to reflect demographic factors combined with social and economic development. Conversely, the decrease in incidence observed in the beginning of the 21st century in several developed countries, including United States, Canada, the United Kingdom, France, and Australia, is partly attributed to the decline in the use of postmenopausalhormonal treatments (Bray, et al., 2018). It has been estimated that the annual worldwide incidence of breast cancer will reach approximately 3.2 million by 2050 (Hortobagyi, et al., 2005). Although male breast cancer can occur, men represent less than 1% of all patients (Stewart & Wild, 2014).

In Portugal, breast cancer incidence and mortality age-standardised rates (ASR) in 2020 were lower than the majority of European countries (Figure 1). GLOBOCAN statistics for Portugal, of 7,041 new cases diagnosed, show that breast cancer has the highest incidence rate among women (26.4%) and is second to colorectum cancer when considering both sexes (11.6%). BC ranked fifth in mortality, with 1,864 deaths. Those numbers translate into 70.8 new cases and 12.7 deaths per 100,000 persons (Figure 1.1B). A recent study (Forjaz de Lacerda, et al., 2018) demonstrated that between 1998 and 2011, breast cancer incidence increased in Portugal for all ages and across all regions. There were 71,545 cases diagnosed between the ages of 30 to 84 years. The south of the country presented the highest incidence, and the north the fastest, with forecasts showing that northern rates might rapidly surpass southern rates.

(A)



(B)

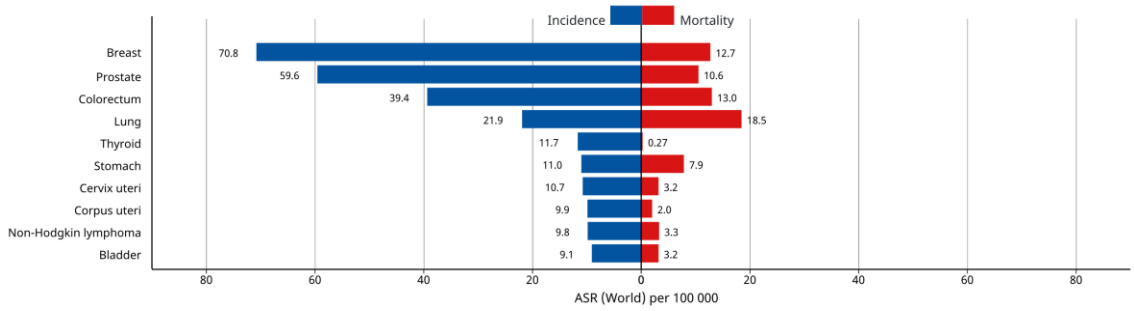


Figure 1.1. Bar charts of cancer incidence and mortality age-standardised (world) rates (ASR) in 2020. **(A)** Region-specific ASR for breast cancer. **(B)** ASR for the top 10 most common cancers in Portugal. Rates are shown in descending order of the world age-standerdised rate. Source: GLOBOCAN 2020.

1.2.2. Histological Classification

Breast cancer arises in any of the mammary gland cells. It is a clinically and morphologically heterogenous disease (Stingl & Caldas, 2007), with specific clinical courses and outcomes associated to a wide scope of morphological features, immunohistochemical profiles, and histopathological subtypes. In order to organize this heterogeneity, classification systems were established, aiming the improvement of clinical and pathological prognosis and selection of the most suitable therapy.

Histological classification of breast cancer considers histological grade (Elston & Ellis, 1991) and histological type (Ellis, et al., 1992). The Nottingham Grading System proposed in 1991 (Elston & Ellis, 1991) is the most widely used grading system, and assesses three features, including the proportion of tubule formation, the mitotic grade, and the degree of nuclear pleomorphism. The individual score of each feature is determined and the three values are added to grade cancer as poorly, moderately or well differentiated (low, intermediate and high grade, respectively).

On the other hand, histological types are defined by tumour morphological and cytological characteristics. Currently, over 20 histological subtypes are recognised by the World Health Organization (WHO) (Lakhani, Ellis, Schnitt, Tan, & van de Vivjer, 2019). Most breast cancers arise from epithelial cells (carcinomas) and can be generally categorized into in situ and invasive (infiltrating) carcinomas. Albeit the distinct terminology, the majority of breast carcinomas originate from the structural and functional unit of the breast named terminal duct lobular unit (Wellings, Jensen, & Marcum, 1975).

In situ carcinomas are preinvasive lesions in which the malignant epithelial cells are restricted to the ducts or lobules, and have not penetrated the ductal or lobular membrane. These preinvasive lesions can be subdivided into ductal carcinoma in situ (DCIS), and lobular carcinoma in situ (LCIS). DCIS and LCIS differ in the architectural and cytological features of the cells, in their distribution within the breast, in the associated respective risks of bilateral disease, and in their natural history (Stewart & Wild, 2014). DCIS is more common than LCIS, and can be subcategorised into comedo, cribiform, papillary, micropapillary and solid (Connolly, et al., 1995).

Histological grade and type provide complementary information, with grade being used to identify prognostic subgroups among special types of breast cancer, and information on these being used to tailor the best therapy for patients (Weigelt, Geyer, & Reis-Filho, 2010).

The histopathological analysis of breast cancers also provides information on stage of disease, determined from the size of the cancer and from the assessment of lymph nodes involvement. Both these parameters have prognostic value for patient management. The overall staging for the patient is determined with the American Joint Committee on Cancer staging system using the tumour–node–metastasis (TNM) classification. Histopathological assessment can provide additional prognostic data, such as lymphovascular permeation and degree of response to neoadjuvant therapy (Stewart & Wild, 2014).

Over the years, classification of breast cancer has gradually progressed from being only based on morphology to be more integrative, taking into account tissue-based biomarkers, besides the clinical features. These classification systems allow a more precise patient stratification based on relative risk of recurrence or progression and on therapy response, and incorporate molecular markers, such as oestrogen receptor (ER), progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2/ERbB2/neu). Since 1975 that steroid hormone receptors are known to play an important role in the biology of breast cancer (Horwitz, Pearson, & Segaloff, 1975). The assessment of ER, PR and HER2 status by immunohistochemistry on tissue sections, and the additional assessment of HER2 status by in situ hybridization, is currently routine practice for all primary breast cancers (Harris, et al., 2007; Wolff, et al., 2007; Hammond, et al., 2010; Rakha, et al., 2015; Wolff A. C., et al., 2018). ER and PR are nuclear hormone receptors found on breast cells that lead to cell growth upon receiving hormone signals. ER belongs to a group of nuclear hormone receptors that act as transcription factors, and given that it regulates PR expression, the last is considered indicative of a functional ER signalling (Walker, 2008). Depending if cancer cells express these hormone receptors or not, they are classified as ER-positive (ER+) or ER-negative (ER-), and PR-positive (PR+) or PR-negative (PR-). The presence of these hormone receptors indicates that cancer cells may receive hormone signals that promote their growth. Similarly, HER2 proteins are breast cells receptors that control cell growth, division and repair. The presence of too many copies *HER2* gene is known as *HER2* amplification, which can lead to HER2 protein overexpression, and ultimately to uncontrolled growth and proliferation of breast cells.

Although the proliferation index marker Ki67 encoded by the *MKI67* gene is not included in the list of required routine biological markers (Harris, et al., 2016), it is also a reliable indicator of the proliferative status of cancer cells (Ma, et al., 2008; Aleskandarany, et al., 2010; Yerushalmi, Woods, Ravdin, Hayes, & Gelmon, 2010). The Ki-67 is a nuclear protein expressed in proliferating cells and can be detected by immunohistochemistry.

Approximately 75%-80% of primary breast cancers are ER+PR+. Since the PR requires oestrogen and ER, ER+ tumours are commonly PR+, and ER- tumours are generally PR- (Cui, Schiff, Arpino, Osborne, & Lee, 2005). Previous reports show that only around 15% of ER+ cancers are PR- and about 3% of ER- cancers are PR+ (Anderson, Chatterjee, Ershler, & Brawley, 2002; Chu & Anderson, 2002; Colditz, Rosner, Chen, Holmes, & Hankinson, 2004; Dunnwald, Rossing, & Li, 2007; Onitilo, Engel, Greenlee, & Mukesh, 2009; Li, et al., 2020). HER2 is overexpressed in 12%-23% of breast cancer (Owens, Horten, & Da Silva, 2004; Yaziji, et al., 2004; Onitilo, Engel,

Greenlee, & Mukesh, 2009; Rakha & Green, 2017), and 10%-15% are triple (ER/PR/HER2) negative breast cancers (TNBC) (Onitilo, Engel, Greenlee, & Mukesh, 2009; Badve, et al., 2011).

Previous studies showed that ER- tumours are more frequently observed in younger women and are associated with worst survival. ER- status is also associated with race, with higher percentages of ER- tumours being observed in black women than in white women (Anderson, Chatterjee, Ershler, & Brawley, 2002; Dunnwald, Rossing, & Li, 2007; Li, et al., 2020). While the number of ER+ tumours increases with age, PR+ rate moves in the opposite direction, considering ER-PR+ patients over 40 years old (Dunnwald, Rossing, & Li, 2007).

Inhibition of ER through endocrine targeting, either directly with oestrogen antagonists or indirectly by blocking the conversion of androgens to oestrogen (e.g., aromatase inhibitors) is the main BC endocrine therapy. However, only approximately 50% of ER+ patients respond to hormone treatment (Early Breast Cancer Trialists' Collaborative Group, 1998). Previous reports indicated that breast cancers with an amplified *HER2* gene or with overexpression of the HER2 protein, are more aggressive and associated with a decrease in disease-free and overall survival compared to tumours that do not overexpress HER2 (Slamon, et al., 1987; Seshadri, et al., 1993). Several trials showed that the use of trastuzumab, a humanized monoclonal antibody that targets the extracellular domain of the HER2 protein, significantly improved disease-free and overall survival of HER+ patients by as much as 50% (Slamon, et al., 2001; Piccart-Gebhart, et al., 2005; Romond, et al., 2005; Joensuu, et al., 2006). Thus, inclusion of this therapy converted an aggressive tumour subtype into one with improved prognostic outcomes. In addition, it has been shown that HER2+ treated with trastuzumab had improved prognosis compared with HER2- patients (Dawood, Broglio, Buzdar, Hortobagyi, & Giordano, 2010). Finally, TNBC, characterised by not expressing ER, PR nor HER, is an aggressive subtype of breast cancer, with distinct metastatic patterns and lack of target therapies. It is more prevalent in young and African-American patients and is usually associated with a poor outcome (Badve, et al., 2011).

1.2.3. Molecular Classification

Progress in gene expression analysis, including high-throughput technologies, provided evidence of the heterogeneity of breast cancer at the molecular level. The first system for classifying tumours based on the gene expression profiles of the two distinct types of epithelial cells – basal (myoepithelial) cells and luminal epithelial cells – found in the human mammary gland (Taylor-Papadimitriou, et al., 1989; Ronnov-Jessen, Petersen, & Bissell, 1996) was developed in 2000

(Perou, et al., 2000). This study reported four main classes of breast cancer: luminal, HER2-overexpressing, basal-like, and normal-like. Subsequent studies suggested that the luminal class was heterogeneous relatively to the expression of some genes and outcome (Pusztai, et al., 2003; Sorlie, et al., 2003; Sotiriou, et al., 2003). Hence, luminal class was further subdivided into luminal A and luminal B, with five main molecular classes being recognized.

Luminal A is the most common molecular subtype, representing 45%-55% of invasive breast cancers (Bhargava, et al., 2009; Voduc, et al., 2010). These tumours are characterized for being ER+, PR+ or PR-, and Her2-, with high levels of ER and downstream transcriptional targets of ER, and low levels of expression of proliferation-related genes. Typically, luminal A tumours are low grade, and show the best prognosis. Luminal B subtype show low to moderate expression of ER and other luminal specific genes, and express high levels of proliferation-associated genes. Several studies show that this subtype tends to be higher grade and is associated with worst prognosis (Sorlie, et al., 2001; Bhargava & Dabbs, 2008; Rastelli & Crispino, 2008; Bhargava, et al., 2009; Al Tamini, Shawarby, Ahmed, Hassan, & AlOdaini, 2010; Habashy, et al., 2012). Clinically, luminal A tumours are likely to benefit from endocrine therapy alone but not from chemotherapy. On the other hand, luminal B patients show better responses to neoadjuvant chemotherapy (Ring, Smith, Ashley, Fulford, & Lakhani, 2004; Cheang, et al., 2009; Hugh, et al., 2009).

Other studies suggested additional luminal subclasses, including luminal C (Sotiriou, et al., 2003) and luminal N (Rakha, et al., 2014), but only luminal A and B remain recognised.

The HER2-overexpressing group is HER2+, ER- and PR-, and is likely to be high grade (Bhargava, et al., 2009; Al Tamini, Shawarby, Ahmed, Hassan, & AlOdaini, 2010). Although this subtype shows a poor prognosis (Sorlie, et al., 2001; Weigelt, et al., 2003), it demonstrates high sensitivity to neoadjuvant anti-HER2-targeted chemotherapy (Sorlie, et al., 2003; Rouzier, et al., 2005; Parker, et al., 2009).

The basal-like tumours express high levels of genes characteristic of breast basal epithelial cells, and do not express ER and most of other ER-related genes (Perou, et al., 2000), and also do not express PR nor HER2 genes (Bhargava, et al., 2009; Al Tamini, Shawarby, Ahmed, Hassan, & AlOdaini, 2010). This subtype is usually very aggressive, with high proliferation index, high grade, and poor prognosis (Carey, et al., 2004; Badve, et al., 2011). Since it shows a triple-negative phenotype, it does not respond to conventional target treatments. Nevertheless, basal-like tumours present high sensitivity to chemotherapy (Rouzier, et al., 2005).

The normal-like subtype also displays a triple-negative phenotype, but it does not cluster with basal-like tumours. It is characterized by high expression of genes characteristic of normal breast epithelial cells and adipose cells, and low expression of genes characteristic of luminal epithelial cells (Perou, et al., 2000). These tumours present good prognostic (Guedj, et al., 2012). However, this is a controversial cluster, and it has been suggested to be an artefact due to normal epithelial cell contamination (Peppercorn, Perou, & Carey, 2008), although other data indicates that this is a true subtype (Guedj, et al., 2012).

Gene expression profiling to assign breast tumours to specific intrinsic subtypes is not always reproducible (Weigelt, et al., 2010; Mackay, et al., 2011). PAM50 (Prosigna®), which stands for Prediction Analysis of Microarray 50, was developed in order to standardise BC classification and approved in 2012 in Europe and 2013 in the USA. Currently, PAM50 tests tumour samples from postmenopausal women for a group of 58 genes, and helps to predict the risk of recurrence of ER+ BC from 5-10 years after diagnosis after 5 years of hormonal therapy (Sestak, et al., 2015). This profiling test also show predictive value for benefit of adjuvant systematic therapy (Parker, et al., 2009; Krop, et al., 2017) and in predicting response to neoadjuvant therapy (Prat, et al., 2016).

Recently, several research groups used multiomic approaches, including exome sequencing, messenger RNA (mRNA) arrays, genomic DNA copy number arrays, DNA methylation arrays, and microRNA arrays, to achieve a more integrated molecular classification. One of these studies was performed by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis, et al., 2012). This study analysed 2,000 breast tumours and reported the existence of 10 molecular subtypes, designated integrative clusters (IntClust 1 to 10), each associated with specific copy number aberrations (CNAs) and gene expression changes.

The Cancer Genome Atlas (TCGA) program (Cancer Genome Atlas Network, 2012) analysed 466 primary breast cancers by genomic DNA copy number arrays, DNA methylation, exome sequencing, mRNA arrays, microRNA sequencing and reverse-phase protein arrays. This integrated analysis identified four main breast cancer classes.

1.2.4. Aetiology

Breast cancer is a multifactorial disease, and several genetic, lifestyle and environmental risk factors are known (Hulka & Stark, 1995; Keibl & Kristensen, 2016).

Breast cancer development is triggered by a gradual accumulation of mutations and epigenetic changes in mammary cells during lifetime (Polyak, 2007). Previous studies have identified numerous driver mutations that clarify the molecular complexity and clinical heterogeneity of breast cancer (Stephens, et al., 2012). Since tumorigenesis is characterised by progressive cell growth and proliferation, that converts normal cells into cancer cells, its most particular molecular events include an activation of the cell division cycle and evasion of apoptosis in cells with impaired maturation and abrogated senescence. The majority of breast cancer risk-associated genes code for proteins involved in DNA repair pathways controlling DNA integrity and in complex mechanisms of the DNA double-strand break (DDSB) repair, which promote the occurrence of mutations that trigger tumour development (Keibl & Kristensen, 2016).

Approximately 30% of breast cancer cases are caused by genetic risk, whereas sporadic cancers represent 70% of all cases, although alterations accounting for genetic risk are not completely clear (Figure 1.2) (Mucci, et al., 2016; Brewer, Jones, Schoemaker, Ashworth, & Swerdlow, 2017; Valencia, Samuel, Viscusi, Neumayer, & Aziz, 2017).

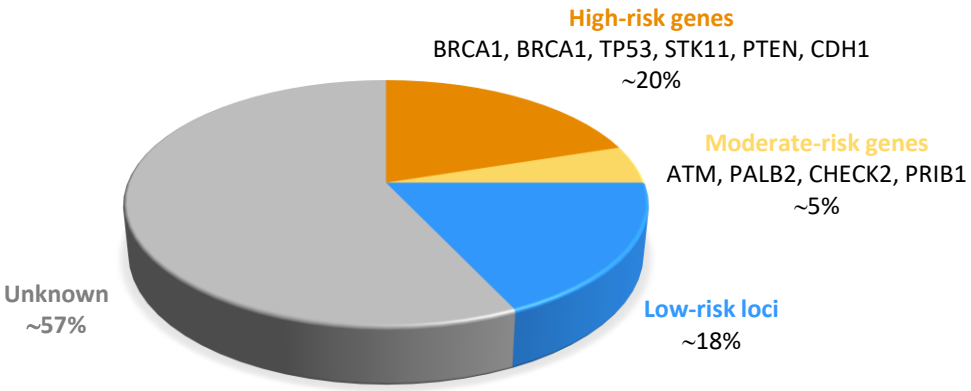


Figure 1.2. Schematic distribution of breast cancer according to genetic risk. Adapted from Wendt & Margolin, 2019.

1.2.4.1. Nongenetic Risk Factors

Race

Overall incidence of breast cancer is slightly lower in black women than in white women. However, black women are more likely to have a more aggressive type of cancer and to die (DeSantis, et al., 2015), regardless treatment, tumour characteristics, follow-up, and other risk

factors (Chlebowski, 2005; Albain, Unger, Crowley, Coltman, & Hershman, 2009; Iqbal, Ginsburg, Rochon, Sun, & Narod, 2015). In fact, by 2012 BC mortality rate was 42% higher in blacks than in whites. In addition, hormone receptor-negative and HER2+ BC is more frequent in black women than in women of other ethnicities (DeSantis, et al., 2015). On the other hand, Asian/Pacific Islander women have the lowest incidence of BC in the USA (DeSantis, et al., 2015), which might be explained by the reproductive patterns of these minorities, including greater number of children and younger age at first childbirth (Chlebowski, et al., 2005).

Age and Reproductive Factors

Breast cancer incidence and mortality increases with age (Bray, et al., 2018). Incidence doubles every 10 years until the menopause, when the rate slows dramatically (McPherson, Steel, & Dixon, 2000). The peak age at diagnosis is 60-70 years in Western countries and 40-50 years in Asian and African countries (Leong, 2010; Abdulrahman & Rahman, 2012).

Early menarche and late menopause increase breast cancer risk. It has been suggested that premenopausal and postmenopausal breast cancer risk is reduced 9% and 4%, respectively, for each additional year that menarche is postponed (Clavel-Chapelon & Gerber, 2002). Moreover, women that have a natural menopause after age 55 increased risk to develop BC than women that do so before age 45 (Bilimoria & Morrow, 1995).

Women that have their first child at a younger age and that have a higher number of children have decreased BC risk. Delayed pregnancy can increase the risk in 1.5% (Nelson, et al., 2012), and each birth can potentially reduce the risk in 7%-11% (Collaborative Group on Hormonal Factors in Cancer, 2002) (Ma, Bernstein, Pike, & Ursin, 2006). A previous report indicated that the risk of premenopausal and postmenopausal is increased by 5% and 3%, respectively, respectively, for each year that first full-term pregnancy is delayed (Clavel-Chapelon & Gerber, 2002). Long periods of breastfeeding are also associated with reduction in BC risk (Bray, et al., 2018).

Exogenous Hormones

The part of oral contraceptives in breast cancer risk is still not clear. Several reports suggested an increase in risk associated in women taking oral contraceptives and for 10 years after stopping taking it (Collaborative Group on Hormonal Factors in Breast , 1996) (Hunter, et al.,

2010). On the other hand, other studies, did not find association between oral contraceptive use and BC risk among women of 35 to 44 years of age (Marchbanks, et al., 2002), nor between ever-users compared with never-users (Beaber, et al., 2014). In this last study, however, use of oral contraceptives for 15 years or longer (compared with never-user) was associated with a 50% increased BC risk among all women.

Hormone replacement therapy (HRT) can be prescribed in to general formulations: combined oestrogen and progestin or oestrogen-only therapy. There is evidence that the addition of progestin to HRT enhances visibly breast cancer risk compared to oestrogen alone (Ross, Paganin-Hill, Wan, & Pike, 2000). The overall risk-benefit evaluation of replacement therapy with oestrogen alone tends for benefit, although with a slight increased risk of breast cancer (Ross, Pike, Henderson, Mack, & Lobo, 1989). Most studies suggest that using hormone replacement therapy for five years or longer periods after menopause can slightly increase BC risk by a factor of 1.02 for each year of use (McPherson, Steel, & Dixon, 2000). HRT was also associated with increased risk in white and Hispanic women, but not in black women. Moreover, higher risk was associated with low/normal body mass index and extremely dense breasts (Hou, et al., 2013).

Mammographic Density

High density of breast tissue is associated with an increased risk of developing breast cancer (Kerlikowske, et al., 2010). Highest levels of breast density were associated with a significant 3% increase in risk per 10 cm^3 of dense tissue (Duffy, et al., 2018). In 2013, mammographic density was strongly associated with all BC subtypes, especially large tumours and positive lymph nodes across all ages, and ER- tumours among women ages below 55 years, suggesting an important role of breast density in BC aggressiveness, particularly in younger women (Bertrand, et al., 2013).

Lifestyle

Obesity is associated with increased BC risk in postmenopausal women (Hunter & Willett, 1993) (Huang, et al., 1997), whereas among premenopausal women, it is associated with a reduced incidence (Ursin, Longnecker, Haile, & Greenland, 1995). Influence of high body mass index is most likely mediated through hormonal mechanisms, given its influence on the effects of exogenous and endogenous hormones. Anovulation associated with obesity has been suggested

has responsible for the decreased risk, while conversion of androgens to oestrogens in adipose tissue appears to influence the increased risk (Pike, 1990; Stewart & Wild, 2014).

Although there is a correlation between BC incidence and dietary intake in populations, the relation between fat intake and BC does not appear to be particularly strong or consistent (McPherson, Steel, & Dixon, 2000). Studies relating fat in the diet and BC risk often show conflicting results. Some studies suggest that high protein intake can increase the risk (Farvid, Cho, Chen, Eliassen, & Willett) by increasing the amount of insulin-like growth factor-1 (Levine, et al., 2014) or by increasing the intake of carcinogenic byproducts and exogenous hormones (Hamajima, et al., 2002), whilst others fail to find an association between red meat consumption and BC risk (Missmer, et al., 2002; Genkinger, Makambi, Palmer, Rosenberg, & Adams-Campbell, 2013).

Physical activity is another factor decreasing breast cancer risk. Previous reports indicate that practice of regular physical exercises during adolescence or adulthood, reduced BC risk among women 40 years of age or younger (Bernstein, Henderson, Hanisch, Sullivan-Halley, & Ross, 1994), and women at the age of 35 years had a 14% decrease in BC risk, which tends to be greater with more hours of exercise (McTiernan, et al., 2003).

It is clear that smoking negatively affects overall health and increases the risk of developing several cancers. Nevertheless, most studies did not find conclusive evidence of association between smoking and breast cancer risk (Terry & Rohan, 2002). On the other hand, other report showed that ever-smoking is associated with a significant 10% increase of relative risk. This effect appeared to be stronger for current smokers than for former smokers, besides being dose-dependent (National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health, 2014). Moreover, other data showed that smoking after diagnosis worsens prognosis (Passarelli, et al., 2016).

Although controversial, some epidemiological studies in the past indicated a consistent relationship between alcohol intake and breast cancer risk. Consumption of 35 to 44 g of alcohol per day was associated with an increased relative risk of breast cancer of 1.32, which increased to 1.46 in women who drank over 45 g per day. It was estimated that the BC risk increases 7% with each additional 10 g per day of alcohol consumed (Hamajima, et al., 2002). Other studies have indicated an association between alcohol intake and the risk of developing ER+ BC (Nasca, et al., 1994; Li, et al., 2003; Terry, et al., 2006). However, several other studies did not find association breast cancer and alcohol intake (Kuper, et al., 2000; Bessaoud & Daurès, 2008). Alcohol role in carcinogenesis can be related with alteration of carcinogens levels, like

acetaldehyde and benzene, present in alcohol or produced by its metabolism (Lachenmeier, Przybylski, & Rehm, 2012), as well as altered hormone levels (Purohit, 2000). In addition, alcohol can also supress immune function, inhibit DNA repair, increase cell proliferation, and promote cell invasion and migration (Terry, et al., 2006; Singletary & Gapstur, 2001).

1.2.4.2. Genetic Risk Factors

Genetic factors affecting the risk of breast cancer development are of particular importance. Germline genetics contribute significantly to risk, and an estimated 30% of breast cancers are heritable or due to underlying genetic factors (Lichtenstein, et al., 2000; Peto & Mack, 2000; Mucci, et al., 2016). Compared with women without a family history of breast cancer, women with a first-degree female relative with the disease are at 2-fold (Goldgar, Easton, Cannon-Albright, & Skolnick, 1994; Hemminki & Vaittinen, 1998; Collaborative Group on Hormonal Factors in Breast Cancer, 2001) or even 3.3-fold greater risk (Singletary E. S., 2003). Risk increases with the number of affected relatives, and an estimated 3.6-fold greater risk has been reported for women with two first-degree relatives with BC (Singletary E. S., 2003). It has been documented that, patients with at least one first-degree relative constitute 13% of the cases (Collaborative Group on Hormonal Factors in Breast Cancer, 2001). In parallel, twin studies reported that monozygotic twins have much higher BC risk than dizygotic twins (Lichtenstein, et al., 2000; Peto & Mack, 2000). Simulation studies have demonstrated that if familial relative risk were observed, environmental risk factors would need to confer very large risk ratios in order to confer even modest increases in the familial relative risk (Hopper & Carlin, 1992). Taken together these observations suggest that familial relative risk is a direct reflection of the genetic component of breast cancer (Peto & Mack, 2000; Easton, et al., 2007).

Several approaches have led to the identification of breast cancer susceptibility variants, which are generally divided into the categories of high-, moderate-, and low-penetrance variants, according to their frequency and conferred risk. The penetrance of BC susceptibility-associated variants is determined by the proportion of carriers that develop the disease during their lifetime. Genetic variants typically have two alleles, implying that there are two common base-pair (bp) possibilities for that specific genomic location within a population. The frequency by which each variant occurs is specified in terms of the less frequent allele, and thus is given by the minor allele frequency (MAF).

High-penetrance variants are very rare in the population with a MAF lower than 0.005 and confer a relative risk higher than 5. Moderate-penetrance variants are rare, with a MAF of 0.005-0.01 and confer a 2- to 4-fold increase in risk. Low-penetrance variants are common in the population with a MAF≥0.05 and confer a less than 1.5-fold increase in BC risk (Mavaddat, Antoniou, Easton, & Garcia-Closas, 2010). These common variants are generally single nucleotide polymorphisms (SNPs), which are single base-pair changes in the DNA sequence (The International HapMap 3 Consortium, 2010). SNPs are the most common form of genetic variation and occur normally in the human genome almost once in every 1,000 nucleotides on average, meaning that each individual genome carries approximately 3 million SNPs (Zhao, Fu, Hewett-Emmett, & Boerwinkle, 2003). Usually, common variants with MAF higher than 5% are designated as SNPs, as the term mutation is applied to rare genetic variants.

1.2.4.2.1. Linkage Studies: High-Penetrance Variants

Linkage analysis is used to map genetic loci through observations of related individuals. In this approach, families affected by a disease are genotyped using a collection of genetic markers across the genome, and the way how those markers segregate with the disease across multiple families is analysed. Early studies using linkage analysis have identified breast cancer susceptibility genes responsible for heritable breast cancer. These studies analysed families with multiple affected individuals and allowed the identification of highly penetrant risk genes, such as *BRCA1* (breast cancer 1) located on chromosome 17q21 (Miki, et al., 1994) and *BRCA2* (breast cancer 2) located on chromosome 13q12.3 (Wooster, et al., 1995). Mutations in these two genes account for around 5% of all BC cases and up to 15% of familial BC cases (Peto, et al., 1999; Antoniou & Easton, 2006). The familial relative risk of BC has been estimated to 11.4 for *BRCA1* and 11.7 for *BRCA2* (Easton, et al., 2015). In families with multiple cases, *BRCA1* and *BRCA2* were linked to breast cancer in 52% and 32% of families, respectively (Ford, et al., 1998). Deleterious mutations in these two genes confer a very high lifetime risk of breast cancer, with 65% of *BRCA1* carriers and 45% of *BRCA2* carriers, developing the disease by the age of 70 (Antoniou, et al., 2003). For women with familial history of BC, the risk increases to 85% and 84%, respectively (Easton, Ford, & Bishop, 1995; Ford, et al., 1998). *BRCA1* and *BRCA2* are tumour-suppressor genes that code for structurally unrelated proteins, which constitute a key component of large multiprotein complexes involved in DDSB repair (Ford, et al., 1998; Li & Greenberg, 2012). The multi-domain protein coded by *BRCA1* also functions in a number of cellular pathways relevant for genomic stability, including cell cycle checkpoint activation, transcriptional regulation, and

apoptosis (Roy, Chun, & Powell, 2011). Pathogenic variants found in *BRCA1* and *BRCA2* are located almost throughout all coding regions and are generally small deletions and insertions that result in truncated proteins. Usually, *BRCA* risk variants in patients with familial history are unique to each family. The frequency of these mutations may vary in different populations, and founder mutations may be observed within certain populations, as the case of three deleterious founder mutations described in the Ashkenazi Jews (King, Marks, Mandell, & Group, 2003). However, the level of penetrance conferred by these genes can vary, as different pathogenic variants show a varying spectrum of increased BC risk, as documented for example for the *BRCA1* missense variant R1699Q (Spurdle, et al., 2012), which increase the risk only mildly, although with a clear statistical significance, or the *BRCA2* truncating mutation c.K3326*, considered to be a low-risk variant (Michailidou, et al., 2013).

Most *BRCA1*-related breast cancers are basal-like tumours, and are characterised by the absence of expression of hormone receptors and no overexpression of HER2 (TNBC) (Foulkes, et al., 2003). *BRCA2*-related tumours have immunophenotype and gene expression profiles similar to sporadic breast cancers. Both *BRCA1* and *BRCA2* tumours exhibit high histological grade, with *BRCA1* tumours being often poorly differentiated and *BRCA2* tumours moderately or poorly differentiated (van der Groep, van der Wall, & van Diest, 2011).

Male carriers of *BRCA2* mutations have been estimated with 7% risk of developing the disease by the age of 80 (Thompson, Easton, & Breast Cancer Linkage Consortium, 2001).

There are other high-penetrance genes identified as part of inherited cancer syndromes that substantially increase breast cancer risk, and account for approximately 5% of the familial risk (Stratton & Rahman, 2008; Ghoussaini & Pharoah, 2009). Germline mutations in *TP53* (tumour suppressor 53) are found in the rare autosomal dominant disorder Li-Fraumeni syndrome (LFS) (Malkin, et al., 1990; Garber, et al., 1991). TP53 protein is involved in cell cycle regulation, DNA repair, apoptosis, cellular senescence, and metabolism. Somatic mutations in *TP53* were first identified in cancer types commonly observed in LFS families. Malkin and colleagues identified segregating germline mutations in exons 5 to 8, which contain a highly conserved DNA binding domain, in the affected members in five LFS families, consistent with an inherited dominant model, with germline missense mutations being more frequent (Malkin, et al., 1990). Female and male carriers have approximately 100% and 70% lifetime risk of cancer, respectively, with the higher percentage observed in females due to breast cancer (Chompret, et al., 2000). Breast cancer is the most common cancer affecting LFS individuals, and most carriers are diagnosed before 40 years, although there is a 49% overall lifetime risk of breast cancer development

before age 60 (Hwang, Lozano, Amos, & Strong, 2003). LFS patients are more frequently diagnosed with ductal carcinomas or ductal carcinomas in situ with ER+, PR+ and HER2+ (Masciari, et al., 2012). Increased risk of having *TP53* mutations is found in breast cancer patients younger than 30 that have a first- or second-degree relative affected with LFS-associated cancers. On the other hand, patients between the age 30 and 49 and without familial history of LFS-associated cancers, are not prone to present mutations in *TP53* (Gonzalez, et al., 2009).

Additional rare but high-penetrance genes are even less frequent, and include *STK11* (serine/threonine kinase 11), *CDH1* (cadherin 1) and *PTEN* (phosphatase and tensin homolog), each conferring a distinct clinical syndrome.

Germline deleterious mutations in the tumour suppressor gene *STK11* are associated with the Peutz-Jeghers syndrome, which is characterised by intestinal hamartomous polyps and mucocutaneous pigmentation (Tomlinson & Houlston, 1997; Hemminki, et al., 1998). The protein coded by *STK11* is important for cell cycle regulation and mediation of apoptosis. Peutz-Jeghers rare disorder is associated with 31% and 45% increased risk of breast cancer by the age of 60 and 70, respectively (Hearle, et al., 2006).

The *CDH1* gene encodes epithelial cadherin, a protein responsible for cell-to-cell adhesion, and functions as a suppressor of cell invasion (Takeichi, 1991). Germline mutations in *CDH1* promote hereditary diffuse gastric cancer, and carriers of truncating mutations have a relative risk of 6.6, and a cumulative risk of 39% of developing breast cancer by the age of 80 (Pharoah, Guilford, Caldas, & International Gastric Cancer Linkage Consortium, 2001). It has also been reported a cumulative risk of breast cancer development of 52% by the age 75 for female carriers of *CDH1* mutations (Kaurah, et al., 2007). This tumour suppressor gene is found to be mutated in invasive lobular carcinoma of the breast but not ductal breast cancer (Berx, et al., 1995). Deleterious mutations have been identified in women with bilateral lobular breast cancer without relatives with diffuse gastric cancer (Benusiglio, et al., 2013).

The *PTEN* gene encodes a multifunctional phosphatase with roles in the PI3K/AKT/mTOR signalling pathway. Hereditary heterozygous mutations in this gene cause Cowden syndrome, a rare multisystem disease (Nelen, et al., 1996) that increases lifetime risk of breast cancer in 25-50% (Pilarski, et al., 2013). A higher lifetime risk of 67-85% have been reported by other studies (Nieuwenhuis, et al., 2014).

1.2.4.2.2. Candidate Gene Resequencing: Moderate-Penetrance Variants

Further genetic linkage studies conducted in multiple cases breast cancer families that do not segregate *BRCA1* and *BRCA2* mutations, have failed to identify additional high-penetrance genes associated with breast cancer susceptibility (Antoniou & Easton, 2006). This suggested that effects similar to *BRCA1* and *BRCA2* are unlikely to exist, or it will be very rare. Thus, in order to elucidate hereditary causes for the disease, new research approaches were pursuit. Linkage analysis was combined with association studies, and further studies in families in which the number of cases was indicative of predisposition to breast cancer, focused on resequencing genes likely to increase BC risk based on their known cellular functions. Moderate-penetrance variants have been identified by resequencing candidate genes, such as those interacting with *BRCA1* and *BRCA2*, or acting in the same DNA repair pathways or other pathways important for breast cancer, including cell-cycle regulation, carcinogenesis, apoptosis, carcinogen metabolism and steroid hormone metabolism. This group includes the protein-truncating variants in *ATM* (ataxia telangiectasia mutated), *PALB2* (partner and localizer of BRCA2), *CHEK2* (checkpoint kinase 2), and *BRIP1* (BRCA1 interacting protein c-terminal helicase 1), and together these genes account for approximately 5% of the genetic component of breast cancer risk (Figure 1.2) (Stratton & Rahman, 2008; Ghoussaini & Pharoah, 2009).

The *PALB2* gene encodes a protein involved in nuclear localization and in maintaining genomic integrity through its role in scaffolding the BRCA1 and BRCA2 proteins during the DDSB repair process. PALB2 was shown to be crucial for key BRCA2 tumour suppressor functions (Xia, et al., 2006). *PALB2* was first indicated as a breast cancer susceptibility gene in 2007 (Rahman, et al., 2007). Biallelic mutations in *PALB2* cause Fanconi anaemia, have an estimated incidence of 1-2%, and confer a 2.3-fold higher risk for all women, and a relative risk of 3.0 in women under 50 of age (Rahman, et al., 2007) (Teo, et al., 2013). Ten *PALB2* truncating mutations were described (Rahman, et al., 2007) and an age-dependent trend in breast cancer risk was found among carriers (Antoniou, Foulkes, & Tischkowitz, 2014). This last study estimated an increased BC risk of 9.47 for *PALB2* mutation carriers compared with the general population, as well as a cumulative risk of 47.5% by the age 70. This suggests that *PALB2* could be considered a high-penetrance gene (Easton, et al., 2015). Women with PALB2 mutations with relatives diagnosed with BC showed a higher risk than women with no familial history. A higher incidence of *PALB2* mutations has been reported in male breast cancers, although it is likely to contribute to few familial cases (Adank, van Mil, Gille, Waisfisz, & Meijers-Heijboer, 2011).

The *CHEK2* gene codes for the CHK2 protein in response to DDSB or replicative stress activated by *ATM* (Matsuoka, Huang, & Elledge, 1998). CHEK2 plays an important role in phosphorylation of a number of downstream protein substrates, including TP53 and BRCA1 (Nevanlinna & Bartek, 2006). The *CHEK2* variants most commonly associated with breast cancer were identified in Li-Fraumeni patients, and include the truncating variant c.1100delC and the missense variant p.I157T (Bell, et al., 1999). Association of c.1100delC mutation with BC risk has been found in several populations, although it appears to be more frequent in populations from the Northern and Eastern Europe than in populations from Southern Europe and North America (Apostolou & Fostira, 2013; Schmidt, et al., 2016). This mutation was found to increase BC risk in noncarriers of *BRCA1* or *BRCA2* mutations (Meijers-Heijboer, et al., 2002). It was associated with a BC relative risk of 2.26, which is noticeably higher in carriers from a family with BC history. A meta-analysis of this variant has indicated that it significantly increases the risk of unselected BC, early onset breast cancer, and familial breast cancer. Heterozygous carriers from a familial BC group showed a cumulative risk around 37% (Weischer, Bojesen, Ellervik, Tybjaerg-Hansen, & Nordestgaard, 2008). Since carriers without diagnosed relatives may have a lifetime risk similar to the general population, it has been suggested that family history should be taken into account in risk prediction (Schmidt, et al., 2016). Carriers of c.1100delC mutation frequently develop tumours of luminal type, expressing ER and/or PR (Nagel, et al., 2012; Kriege, et al., 2014). In addition, this mutation has been associated with a worse prognosis, particularly for ER+ BC patients (Weischer, et al., 2012; Kriege, et al., 2014). Other two truncating mutations have been associated with similar breast cancer risk, the del5395 (Cybulski, et al., 2007), and the IVS2 + 1G>A (Bogdanova, et al., 2005). The missense mutation p.I157T was associated with a slight but significant increased risk (Kilpivaara, et al., 2004; Bogdanova, et al., 2005), particularly in lobular type breast tumours (Liu, Wang, Wang, & Wang, 2012).

The *ATM* gene encodes for a protein kinase involved in monitoring and DDSB repair, through downstream interaction with *TP53*, *BRCA1* and *CHEK2* (Ahmed & Rahman, 2006). Germline biallelic *ATM* mutations cause ataxia-telangiectasia (Renwick, et al., 2006), an autosomal recessive neurodegenerative disease characterised by cerebellar ataxia, immunological deficiency, hypersensitivity to ionizing radiation and increased risk of cancer. Heterozygotic carriers of *ATM* germline mutations are estimated to constitute 0.35-1% of the general population (Prokopcova, Kleibl, Banwell, & Pohlreich, 2007), to which an increased relative risk of 2.4 as been suggested (Renwick, et al., 2006). Other data indicates an overall relative risk of BC of 4.9 for women younger than age 50 years (Thompson, et al., 2005). A more recent study associated the *ATM* missense pathogenic variant pVal2424Gly or truncating mutations with a

significantly increased risk of BC with a penetrance that appears similar to that conferred by germline mutations in *BRCA2* (Goldgar, et al., 2011). These studies suggest a high BC risk in heterozygous carriers, while causing a milder form of ataxia-telangiectasia when homozygous.

The *BRIP1* encodes a protein that interacts with the C-terminus domain of *BRCA1*. *PRIB1* mutations account for less than 1% of all breast cancer cases. Biallelic mutations of *PRIB1* also cause Fanconi anaemia, and it was estimated that monoallelic carriers have a relative risk of breast cancer of 2.0, and a higher risk for early onset of the disease (Seal, et al., 2006).

Other candidate breast cancer susceptibility genes involved in DNA repair have been proposed, including genes in the *MRE11-RAD50-NBN* pathway. The MRE11-RAD50-NBN is an evolutionary conserved protein complex (the MRN complex), fundamental to maintain genomic integrity and for tumour suppression, since it is required for DNA strand processing during the DDSB repair and *ATM* activation. Carriers of heterozygous *NBN* mutations have been shown to have an increased risk of 3.1 (Bogdanova, et al., 2008), and a meta-analysis associated the truncating variant c.657del5 with a relative risk of 2.7 (Zhang, Zeng, Liu, & Wei, 2013), which was additionally associated with early onset breast cancer (Steffen, et al., 2006). Pathogenic germline mutations in the *MRE11* and *RAD50* have also been suggested as BC susceptibility genes, but their role remains unclear (Bartkova, et al., 2008) (Damiola, et al., 2014). Moreover, the RAD51 paralogs, RAD51B-RAD51C-RAD51D-XRCC2, that constitute the BCDX2 complex also involved in DDSB repair, have been associated with BC risk as well. Pathogenic hereditary variants have been identified in *RAD51C* and *RAD51D* genes, and in the *XR2CC* (*RAD51D* paralog) (Park, et al., 2012) (Couch, et al., 2015). Other study found common variation at the *RAD51B* to be significantly associated with familial breast cancer risk (Pelttari, et al.). Rare hereditary mutations in the helicase genes *RECQL* (Cybulski, et al., 2015) and *BLM* (Thompson, et al., 2012) were also associated with increased BC risk.

1.2.4.2.3. Association Studies: Low-Penetrance Variants

Even with the extensive efforts endeavoured with linkage analyses and resequencing of candidate genes, no further susceptibility genes for breast cancer were identified (Smith, et al., 2006). The fact that the previously described high- and moderate-penetrance variants account for less than 30% of breast cancer familial risk (Stratton & Rahman, 2008; Ghoussaini & Pharoah, 2009), suggests that the remaining familial clustering is likely to be conferred by a large number of variants, each conferring a low effect on breast cancer risk. If common genetic variants

influence disease, their effect size (or penetrance) must be small. Concurrently, if common variants have low penetrance, but the disease shows heritability, then disease susceptibility must be influenced by multiple common variants.

Identification of the third category of risk variants, including common variants frequent in the general population (frequency up to 50%) with modest levels of penetrance (relative risk less than 1.5), were first identified through candidate gene association (case-control) studies. However, the identification of low-penetrance variants has relied mainly on genome-wide association studies (GWAS). These studies scan most of the genome for potential susceptibility variants without the need of high-risk families or candidate genes. Contrarily to candidate gene association studies, which focused on candidate genes based on their potential involvement in carcinogenesis, GWAS search for common low risk variants without any assumption of biological function or location.

Association studies based on SNPs compare their frequencies between cases and controls, and can be performed either by direct testing of a SNP with functional consequence for association with a phenotype, or using a SNP as marker (or tag) for a block of correlated SNPs.

Candidate Gene Association Studies

Early association studies focused on a limited number of putative functional polymorphisms in candidate genes suspected to be important in carcinogenesis. This was a cheaper and simpler approach than the complete resequencing of candidate genes (Ghoussaini & Pharoah, 2009). These association studies used DNA-based assays, like gel-based methods to detect simple tandem-repeat polymorphisms and restriction fragment length polymorphisms, which were easier and cheaper to perform than protein-based methods. However, even though several positive associations were reported, none has been convincingly replicated (Pharoah, Dunning, Ponder, & Easton, 2004). The most probable reason is that most reports were false positives, occurring by chance, which can be quantified in terms of statistical significance. The levels of significance appropriate in other contexts (P=0.05 or P=0.01) can be highly misleading in association studies. Because the number of possible genetic polymorphisms is very large and the prior probability that any polymorphism will be associated with disease will be low, most polymorphisms achieving a modest level of statistical significance will be false positives. In addition, lack of confirmation by subsequent studies can also be due to lack of adequate statistical power, resulting in false negatives in the replication study. Moreover, population

heterogeneity regarding risk could also contribute to failure in replicating associations (Pharoah, Dunning, Ponder, & Easton, 2004).

In order to achieve the sample sizes necessary to detect more modest effects, avoiding false positives, the Breast Cancer Association Consortium (BCAC) was established in 2005. Analyses carried out by this consortium (Breast Cancer Association Consortium, 2006) and by others (Cox, et al., 2007) found evidence of an association with a moderate reduction in breast cancer risk for the common nonsynonymous coding variant D302H in the *CASP8* (caspase 8) gene, and this was the most convincing example that arose from the candidate gene association studies approach. *CASP8* is a regulator of apoptosis (Barnhart, Lee, Alappat, & Peter, 2003) and its impaired expression or function was reported to promote tumour formation (Fulda, 2009).

The identification of a large number of SNPs across the human genome, and their cataloguing by several projects funded by governments or industry (Thorisson & Stein, 2003; Varmus, 2003), has made it possible to move beyond association studies of specific candidate genes. Initial data was provided by the Human Genome Project (HGP) (Lander, et al., 2001) and The SNP Consortium (TSC) (Sachidanandam, et al., 2001), that together identified 1.42 million SNPs. Rapid advances in technology and quality control have permitted increasingly affordable and reliable genotyping of more and more SNPs in an individual genome at once, making genome-wide association studies possible.

Linkage Disequilibrium

Since most SNPs lie outside genes, they are not likely to alter gene structure or function, and thus are unlikely to be directly associated with a phenotype. Sets of nearby SNPs on the same chromosome are inherited in blocks. Each block is called a haplotype and can contain a large number of SNPs. The degree of inheritance or correlation between an allele of one SNP and an allele of other SNP within a population is given by the levels of linkage disequilibrium (LD). This term was created by population geneticists to mathematically describe changes in genetic variation within a population over time. Loci in LD are usually close, but the relation can vary. The first time a mutation (single nucleotide change, insertion/deletion, or structural alteration) causes the introduction of a variant into a population, this will be perfectly correlated (in perfect LD) with neighbouring variants. Over time meiotic recombination events within a family, over successive generations, will break apart chromosomal segments, breaking the correlation between variants, and consequently leading to LD decay until eventually all alleles in the

population are in linkage equilibrium or independent. LD decay occurs at an average rate of around 10^{-08} per bp per generation. The two most common pairwise measures of LD are D' and r^2 (Devlin, 1995; The International HapMap Consortium, 2005). All LD measures are ultimately related with the difference between the observed frequency of co-occurrence of two alleles and the expected frequency if the two alleles are independent. D' values go from 0 to 1. D' is set to be 0 if complete linkage disequilibrium exists, which implies frequent recombination between the two alleles and statistical independence under the principles of Hardy-Weinberg equilibrium. D' is set to be 1 in the absence of recombination, and decays only due to recombination or recurrent mutation (Daly, Rioux, Schaffner, Hudson, & Lander, 2001). LD is usually reported in term of r^2. The r^2 represents a statistical measure of correlation between the two SNPs, meaning that an r^2 of 1 indicates that two SNPs appeared on the same branch of the genealogy and remain undisrupted by recombination, and an r^2 less than 1 indicates that both SNPs arose on different branches, or that an initially strong correlation has been disrupted by crossing over. Thus, if a high value of r^2 is observed, the two SNPs carry similar information and only one of them needs to be genotyped to provide information on genetic variation.

The International HapMap Project was launched in 2002 to create a public, genome-wide database of common human sequence variation for medical genetic research (The International HapMap Consortium, 2003). This project mapped haplotype blocks in the human genome, aiming to guide the design and prioritisation of SNP genotyping assays for disease association studies.

The Phase I of the HapMap genotyped approximately 1.3 million SNPs from 270 individuals from four geographically diverse populations: 30 mother-father-adult child trios from the Yoruba in Ibadan, Nigeria (YRI); 30 trios of northern and western European ancestry living in Utah from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (CEU); 45 unrelated Han Chinese individuals in Beijing, China (CHB); and 45 unrelated Japanese individuals in Toki, Japan (JPT) (The International HapMap Consortium, 2005). In Phase II of the HapMap Project, a further 2.1 million SNPs were genotyped in the same individuals from Phase I (The International HapMap Consortium, 2007).

For Phase III, the HapMap Project collected and genotyped 1.6 million SNPs in an extended set of 1,184 samples from 11 populations, including all HapMap Phase I and II samples, along with further samples from the same four populations. In addition, samples from seven additional populations were also included: African ancestry in the southwestern USA (ASW); Chinese in metropolitan Denver, Colorado, USA (CHD); Gujarati Indians in Houston, Texas, USA (GIH); Luhya

in Webuye, Kenya (LWK); Maasai in Kinyawa, Kenya (MKK); Mexican ancestry in Los Angeles, California, USA (MXL); and Tuscans in Italy (Toscani in Italia, TSI) (The International HapMap 3 Consortium, 2010).

Overall, the International HapMap Project catalogued allele frequencies and LD patterns between nearby variants, across 11 populations for over 4.5 million SNPs. The HapMap genotype data provided LD information in which indirect association mapping relies on, since it makes unnecessary for the functional variant to be studied, so long as a variant in LD with it is measured. SNPs that are selected to capture the genetic variation in a haplotype block are called tag SNPs, in the sense that alleles of those SNPs tag the surrounding region of LD, suppressing the need to look for the patterns of genetic variation at all the SNPs in a given region that provide redundant information. Data from the HapMap Project can be used to estimate LD patterns within different populations (Conrad, et al., 2006; de Bakker, et al., 2006; Need & Goldstein, 2006), which can be then used to design and interpret GWAS.

The 1000 Genomes Project launched in 2008, has also been of great importance in the development of a comprehensive resource of human genetic variation across worldwide populations (1000 Genomes Project Consortium, 2010). This project ran between 2008 and 2015 and its aim was to extend previous substantial progress, including the one made by the HapMap Project, by discovering, genotyping, and providing haplotype information on all forms of human DNA polymorphism, including lower frequency alleles (down towards 0.1%) in several populations. The Pilot 1 of the 1000 Genomes Project sequenced 179 individuals from 4 populations and Phase 1 analysis mapped the genetic variation from 1,092 human genomes from 14 populations. The final data set (Phase 3) combined data regarding 2,504 individuals from 26 populations. After its completion in 2015, this project established the most comprehensive assessment of human genetic variation across global populations, having reconstructed the genomes of 2,504 individuals from 26 populations (Sudmant, et al., 2015). They characterised over 88 million variants, including 84.7 million SNPs, 3.6 million short insertions/deletions (indels), and 60,000 structural variants.

Genome-Wide Association Studies

GWAS approach is based on the premise that a causal variant is located on a haplotype, and therefore a marker (tag) variant in LD with the causal variant show (by proxy) an association with a trait of interest (Stranger, Stahl, & Raj, 2011). SNPs included in a GWAS chip are selected

considering that they tag the majority of common variants across the genome. Based on data from HapMap Project, most genetic common variants occurring in the CEU population can be tagged by 500,000 to 1,000,000 SNPs (The International HapMap Consortium, 2005).

One of the biggest challenges of GWAS is the need to recruit and genotype many subjects, suitable to detect the effects of common variants which are usually small (odds ratio, OR<1.5), while minimizing the risk of false positives. Thus, even though per-SNP genotyping costs are low, the large number of SNPs, together with the adequate sample size, can make a GWAS too expensive (Wang, Barratt, Clayton, & Todd, 2005, Thomas, 2006). To overcome this practical constraint, a multistage approach has been implemented by GWAS, reducing the amount of genotyping required, without sacrificing power. In this type of experimental design, in stage 1, the full set of SNPs is genotyped in a relatively small fraction of the samples, and a liberal p-value threshold is used to identify a subset of variants with the strongest evidence for association. In the second and, possibly, third stages, the SNPs identified in the previous stage are genotyped in the remaining set of samples that is larger or of a similar size. This approach has been shown to save more than half on genotyping costs, while distinguishing the true-positive associations identified in stage 1 from the false positives that would occur by chance (Hirschhorn & Daly, 2005; Wang, Thomas, Pe'er, & Stram, 2006).

Because GWAS involve hundreds of thousands of tests, very large test statistics (small p-values) are expected to occur by chance. Using standard thresholds for statistical significance (P<0.05 or P<0.01, meaning that 5% or 1% of the times, respectively, the null hypothesis is rejected, although it is true and a false positive is being detected) will result in a large number of statistically significant results, most of which are false positives. This is because the commonly used P<0.05 indicates the probability of detecting a false positive in a single statistical test, and in the case of GWAS hundreds of thousands to millions of tests are performed, each with its own false positive probability. In order to control false positive results in GWAS, corrections for multiple testing are made. One frequently used approach to correct for multiple testing is the Bonferroni correction, which adjusts the false positive rate (alpha value) from α=0.05 to α=(0.05/k), where k is the number of statistical tests conducted. This is the most conservative correction, because it assumes that each association test is independent, which is usually not true for GWAS SNPs due to LD. A recurrent alternative to adjust the alpha is to determine the false discovery rate (FDR) (Hochberg & Benjamini, 1990). This is an estimate of the proportion of significant results (usually at α=0.05) that are false positives. There are simple algebraic procedures using the ranked observed p-values that control the FDR when tests are independent

(Sabatti, Service, & Freimer, 2003) or regardless of the correlation pattern among tests (Benjamini & Yekutieli, 2001).

Permutation testing is another approach for establishing GWAS significance, and although somewhat computationally intensive, it is a straightforward method to create the empirical distribution of test statistics for a given dataset when the null hypothesis is true. In this test, phenotype of one individual of the dataset is randomly reassigned to another individual. Each random reassignment represents one possible sampling of individuals under the null hypothesis, and this process is repeated a predefined number of times (n) that will determine the empirical p-value within $1/n^{th}$ of a decimal place (Bush & Moore, 2012).

Contrary to association studies in candidate genes, GWAS in breast cancer and other multifactorial diseases have shown very promising results. Identification of unexpected loci pinpointed new potential pathways involved in breast carcinogenesis.

The establishment in 2005 of large collaborative consortiums, including the Cancer Genetic Markers of Susceptibility (CGEMS) project and the BCAC, increased the size and statistical power of GWAS to identify loci associated with breast cancer risk (Easton, et al., 2007; Hunter, et al., 2007).

In 2007, the first three breast cancer GWAS were published (Easton, et al., 2007; Hunter, et al., 2007; Stacey, et al., 2007). In one of these studies, Easton and colleagues (Easton, et al., 2007) identified five SNPs associated with BC risk ($P<10^{-07}$) located within genes or in LD blocks containing plausible causative genes – 10q26 (intron 2 of *FGFR2*, fibroblast growth factor receptor 2); 16q12 (an LD block that contains *TOX3*, TOX high mobility group box family member 3); 5q11 (including *MAP3K1*, mitogen-activated protein kinase kinase kinase 1); 11p15 (intron 10 of *LSP1*, lymphocyte specific protein 1); and rs13281615 on 8q24 (with no known genes) – in a three-stage GWAS. In the first stage 266,722 SNPs were analysed for 390 familial cases and 364 controls from the UK. For the second stage 12,711 SNPs, approximately 5% of those typed in stage 1, were selected based on the stronger evidence for association, and 10,405 SNPs were analysed in 3,990 invasive breast cancer cases and 3,916 controls from the Studies of Epidemiological and Risk factors in Cancer Heredity (SEARCH) (Pharoah, et al., 2007). In the third stage they tested 30 of the most significant SNPs in 22 additional case-control studies, comprising 21,860 cases of invasive breast cancer, 988 cases of carcinoma in situ and 22,578 controls. To identify additional loci, a subsequent study tested 800 promising associations from the second stage of this GWAS in a further two stages involving 37,012 cases and 40,069 controls from 33 studies in the CGEMS collaboration and BCAC (Ahmed, et al., 2009). They found strong

evidence for two additional susceptibility loci, and potential causative genes included *SLC4A7* and *NEK10* (never in mitosis A-related kinase 10) on 3p24 and *COX11* (cytochrome C oxidase subunit 11) on 17q22.

The second of the first three BC GWAS (Hunter, et al., 2007), was a two-stage study. Initially, 528,173 SNPs were genotyped in 1,145 postmenopausal women of European ancestry with invasive breast cancer and 1,142 controls. They identified four SNPs in intron 2 of *FGFR2* that were highly associated with breast cancer, and confirmed this association in the second stage of the study, using 1,776 affected individuals and 2,072 controls from three additional studies (Table 1.1). Four SNPs at other loci showed strong association in the first-stage of this GWAS, but were not associated in the replication studies.

Finally, Stacey and co-workers (Stacey, et al., 2007) genotyped 311,524 SNPs in 1,600 Icelandic individuals with breast cancer and 11,563 controls. Then, the ten variants showing the strongest evidence for BC association, were genotyped in an independent Icelandic group, including 584 cases and 1341 controls, and subsequently in four replication sets of non-Icelandic samples comprising 2350 cases and 4626 controls. Two SNPs, located on 2q35 (in which LD block no genes had been identified) and 16q12 (in which the associated SNP was located near the 5' end of *TOX3*) consistently associated with breast cancer (Table 1.1). Association of the SNP located on 2q35 was further confirmed for ER+, ER-, PR+ and PR- disease in another study using the BCAC (Milne, et al., 2009).

An additional three-stage GWAS carried out among Chinese women identified one SNP at 6q25, located upstream the *ESR1* (oestrogen receptor alpha) gene, associated with breast cancer risk. This association was replicated in an independent analysis comprising cases and controls of European ancestry from BCAC (Zheng, et al., 2009). Two other GWAS identified three more breast cancer risk-associated loci: 1p11 and 14q24 (*RAD51L1*) from a three-stage study in 9,770 cases and 10,799 controls in the CGEMS initiative (Thomas, et al., 2009); and 5p12 (*MRPS30*, mitochondrial ribosomal protein S30) from a study including 6,145 cases and 33,016 controls (Stacey, et al., 2008) (Table 1.1).

Moreover, 72 potential BC risk-associated SNPs that did not reach genome-wide significance in the UK2 and the British Breast Cancer Study (BBCS) GWAS, were analysed using samples from 41 studies in BCAC and 9 individual GWAS. This analysis resulted in the identification of three additional risk loci, including 12p11 (*PTHLH*, parathyroid hormone-like protein) associated with ER+ and ER- BC, and 12p24 and 21q21 (*NRIP1*, nuclear receptor interacting protein 1) associated with ER+ BC (Ghoussaini, et al., 2012) (Table 1.1). A two-stage GWAS that genotyped 528,886

SNPs in 3,659 cases with family history of the disease and 4,897 controls in stage 1, and 12,576 cases and 12,223 controls in the second stage, identified five additional susceptibility loci: 9p21 (*CDKN2A*, cyclin dependent kinase inhibitor 2A; *CDKN2B*, cyclin dependent kinase inhibitor 2B; *CDKN2BAS*, cyclin dependent kinase inhibitor 2B Antisense RNA); 10q21 (*ZNF365*, zinc finger protein 365); 10p15 (*ANKRD16*, ankyrin repeat domain 16, and *FBXO18*, F-box DNA helicase 1); 10q22 (*ZMIZ1*, zinc finger MIZ-type containing 1); 11q13 (without annotate genes, but flanked by six plausible causative genes) (Turnbull, et al., 2010) (Table 1.1). Similarly, Fletcher and colleagues identified a risk locus for breast cancer at 9q31 (gene desert) using 296,114 tagging SNPs in a three-stage association study including 1,694 cases and 2,365 controls in the first stage and totalling 11,880 cases and 12,487 controls in three independent validation stages (Fletcher, et al., 2011) (Table 1.1).

Therefore, in addition to *CASP8* and *TGFB1* (transforming growth factor beta 1) identified in the early candidate gene-association studies, the initial phase of GWAS including a limited number of cases and controls, and further analyses of these initial GWAS in larger sample sizes, identified 21 loci displaying genome-wide associations and conferring low risk to breast cancer (Lilyquist, Ruddy, Vachon, & Couch, 2018) (Table 1.1). Following similar genome-wide approaches identified four other risk loci, comprising 19p13 (Antoniou, et al., 2010), *TERT-CLPTM1L* (telomerase reverse transcriptase-cleft lip and palate transmembrane protein 1-like protein) (Haiman, et al., 2011), 6q14 and 20q11 (Siddiq, et al., 2012) (Table 1.1).

A collaborative EU funded genotyping initiative involving four consortia named Collaborative Oncological Gene-environment Study (COGS) was developed to study breast, ovarian and prostate cancer. A major component of this project was to genotype over 250,000 samples, for 211,155 SNPs using a custom Illumina iSelect genotyping array (iCOGS). In 2013, the first report using the iCOGS array was published reporting the evaluation of 199,961 SNPs in 52,675 cases and 49,436 controls from 52 studies participating in BCAC (Michailidou, et al., 2013). This study identified 41 new breast cancer susceptibility loci at genome-wide significance ($P<5\times10^{-08}$), and confirmed the existence of strong association with overall BC risk for 23 of the 27 previously established risk loci. Three of the remaining four loci showed weaker evidence for association, including a variant on *CASP8*, one on 10p15 and another on 19p13. The locus 20q11 shown to be associated with ER- BC was not selected for the iCOGS array. In this study, Michailidou and colleagues suggested that approximately 28% of familial risk could be explained by common variants selected for iCOGS, of which around 14% were explained by the 67 established loci (Table 1.1).

All taken together, by 2015, 79 breast cancer susceptibility loci had been published (Table 1.1), and 71 of these showed further evidence of association at $P<5\times10^{-08}$ for overall, ER+ or ER- disease risk. Four of the remaining eight variants showed slightly weaker evidence of association, and for the other four no association was observed (Michailidou, et al., 2015). Furthermore, this study identified 15 new BC susceptibility loci, increasing the number of known independent common risk loci for BC to 94, which were estimated to explain approximately 16% of the familial risk at that point.

In 2017 a new genotyping array, the OncoArray including 533,631 markers was developed by the OncoArray Consortium, in order to increase the number of variants influencing common cancers, in particular cancers of the breast, colon, lung, ovary and prostate (Amos, et al., 2017). The OncoArray identified of nine new loci associated with ER- BC risk after genotyping 21,468 ER- cases and 100,594 controls combined with 18,908 BRCA1 mutation carriers (9,414 with BC), all of European origin (Milne, et al., 2017). In addition, a genome-wide association study conducted in 122,977 cases and 105,974 controls of European ancestry (from BCAC, Biology and Risk of Inherited Variants in Breast Cancer Consortium, DRIVE, iCOGS and 11 other GWAS), and 14,068 cases and 13,104 controls of East Asian ancestry (from the OncoArray and iCOGS projects), identified 65 new loci (Table 1.1) that are associated with overall breast cancer risk at $P<5\times10^{-08}$. The newly identified BC risk loci were estimated to explain around 4% of the twofold familial relative risk of breast cancer and in total, common susceptibility variants identified through GWAS, were estimated to explain 18% of the familial relative risk (Michailidou, et al., 2017).

Table 1.1. Risk associated with overall breast cancer for 172 SNPs. Source: Lilyquist, Ruddy, Vachon, & Couch, 2018.

Locus	Locus discovery reference[a]	Genes	rs ID	Alleles	MAF[b]	Refined SNP reference[c]	Subtype[d]	Overall breast cancer risk (24)		
								OR[e] (95% CI)[f]	*P*[g]	Combined *P*[h]
2q33.1	(99)	*CASP8/ALS2CR12*	rs1830298	T/C	0.28	(100)	Overall	1.06 (1.04–1.08)	1.7×10^{-08}	1.9×10^{-16}
5q11.2	(12)	*MAP3K1*	rs62355902	A/T	0.16	(12)	Overall	1.18 (1.15–1.21)	8.5×10^{-42}	6.8×10^{-98}
8q24.21	(12)	—	rs13281615	A/G	0.41	(12)	Overall	1.11 (1.09–1.13)	5×10^{-28}	1.9×10^{-57}
11p15.5	(12)	*LSP1*	rs3817198	T/C	0.32	(12)	Overall	1.05 (1.03–1.07)	6.2×10^{-07}	9.9×10^{-19}
16q12.1	(12)	*TOX3*	rs4784227	C/T	0.24	(101)	Overall	1.23 (1.2–1.25)	7×10^{-88}	6.8×10^{-201}
10q26.13	(12)	*FGFR2*	rs2981578	T/C	0.47	(102)	Overall	1.23 (1.21–1.25)	1×10^{-114}	1.3×10^{-245}
10q26.13	(12)	*FGFR2*	rs35054928	G/GC	0.4	(102)	Overall	1.27 (1.25–1.3)	3.5×10^{-154}	2.3×10^{-322}
10q26.13	(12)	*FGFR2*	rs45631563	A/T	0.05	(102)	Overall	0.81 (0.78–0.85)	9.1×10^{-21}	7.3×10^{-37}
2q35	(46)	*IGFBP5*	rs4442975	G/T	0.5	(103)	ER+	0.89 (0.87–0.9)	3.6×10^{-40}	1.1×10^{-95}
2q35	(46)	*IGFBP5*	rs34005590	C/A	0.05	(104)	Overall	0.82 (0.79–0.86)	1.6×10^{-19}	3.2×10^{-41}
5p12	(13)	*FGF10/MRPS30*	rs10941679	A/G	0.25	(13)	ER+	1.15 (1.13–1.18)	6.2×10^{-43}	5.6×10^{-73}
3p24.1	(16)	*SLC4A7*	rs4973768	C/T	0.47	(16)	Overall	1.11 (1.09–1.13)	1.7×10^{-28}	4.8×10^{-57}
17q22	(16)	—	rs2787486	A/C	0.3	(105)	Overall	0.93 (0.91–0.94)	1.2×10^{-14}	5.6×10^{-29}
22q12.1	(106)	*CHEK2*	rs17879961	A/G	0.01	(22)	Overall	1.26 (1.11–1.42)	2.4×10^{-04}	9.7×10^{-09}
1p11.2	(107)	*EMBP1*	rs11249433	A/G	0.41	(107)	Overall	1.11 (1.09–1.13)	2.2×10^{-31}	1.8×10^{-52}
14q24.1	(107)	*RAD51B*	rs999737	C/T	0.23	(107)	Overall	0.91 (0.89–0.93)	1.1×10^{-18}	6.5×10^{-39}
19p13.11	(40)	—	rs67397200	C/G	0.3	(108)	*BRCA1*	1.03 (1.01–1.05)	4.2×10^{-03}	1.6×10^{-08}
9p21.3	(20)	*CDKN2A/CDKN2B*	rs1011970	G/T	0.16	(20)	Overall	1.07 (1.04–1.09)	1.4×10^{-07}	1×10^{-15}
10p15.1	(20)	*ANKRD16*	rs2380205	C/T	0.44	(20)	Overall	0.98 (0.96–0.99)	1.1×10^{-02}	1.7×10^{-04}
10q22.3	(20)	*ZMZ1*	rs704010	C/T	0.38	(20)	Overall	1.07 (1.05–1.09)	1.1×10^{-14}	1.7×10^{-35}
11q13.3	(20)	*CCND1*	rs554219	C/G	0.13	(97)	ER+	1.21 (1.18–1.24)	5.8×10^{-47}	5.8×10^{-47}
11q13.3	(20)	*CCND1*	rs75915166	C/A	0.06	(97)	ER+	1.28 (1.24–1.33)	4.1×10^{-42}	4.1×10^{-95}
6q25	(20)	*ESR1*	rs3757322	T/G	0.32	(93)	ER-	1.08 (1.06–1.1)	1.1×10^{-16}	3.3×10^{-41}
6q25	(20)	*ESR1*	rs9397437	G/A	0.07	(93)	ER-	1.17 (1.14–1.21)	6.3×10^{-21}	4.8×10^{-54}
6q25	(20)	*ESR1*	rs2747652	C/T	0.48	(93)	ER-	0.94 (0.92–0.960)	1.2×10^{-11}	1.3×10^{-26}
10q21.2	(109)	*ZNF365*	rs10995201	A/G	0.16	(110)	Overall	0.9 (0.88–0.92)	4.7×10^{-17}	1.6×10^{-51}
9q31.2	(21)	—	rs10816625	A/G	0.06	(111)	Overall	1.11 (1.07–1.15)	2.3×10^{-08}	5×10^{-18}
9q31.2	(21)	—	rs13294895	C/T	0.18	(111)	Overall	1.06 (1.03–1.08)	1.9×10^{-06}	6.5×10^{-17}
9q31.2	(21)	—	rs676256	T/C	0.38	(111)	Overall	0.91 (0.9–0.93)	1.9×10^{-21}	3.5×10^{-53}
5p15.33	(47)	*TERT*	rs10069690	C/T	0.26	(47)	ER-	1.06 (1.04–1.08)	2.5×10^{-08}	7.8×10^{-17}
12q24.21	(19)	*TBX3*	rs1292011	A/G	0.42	(19)	Overall	0.92 (0.9–0.94)	2.4×10^{-19}	4.4×10^{-39}
21q21.1	(19)	*NRIP1*	rs2823093	G/A	0.27	(19)	Overall	0.94 (0.92–0.96)	1.1×10^{-09}	1.5×10^{-20}
6q25.1	(112)	*TAB2*	rs9485372	G/A	0.19	(112)	Overall	0.96 (0.93–0.98)	7.7×10^{-05}	3.5×10^{-06}
6q14.1	(113)	—	rs17529111	T/C	0.22	(113)	ER-	1.02 (1–1.04)	4.2×10^{-02}	1.3×10^{-09}
20q11.22	(113)	*RALY*	rs2284378	C/T	0.32	(113)	Overall	1 (0.98–1.02)	7.9×10^{-01}	3.2×10^{-02}
5p15.33	(92)	*TERT*	rs3215401	A/AG	0.31	(92)	Overall	0.93 (0.91–0.95)	7.1×10^{-13}	1.1×10^{-20}
1q32.1	(48)	*LGR6*	rs6678914	G/A	0.41	(48)	Overall	1 (0.99–1.02)	7.3×10^{-01}	3×10^{-01}
1q32.1	(48)	*MDM4*	rs4245739	A/C	0.26	(48)	Overall	1.02 (1–1.04)	2.5×10^{-02}	1.3×10^{-04}
2p24.1	(48)	—	rs12710696	C/T	0.37	(48)	Overall	1.03 (1.01–1.04)	6.6×10^{-03}	1.3×10^{-08}
16q12.2	(48)	*FTO*	rs11075995	T/A	0.24	(48)	Overall	1.03 (1.01–1.06)	1.3×10^{-03}	8.7×10^{-09}
6p24.3	(114)	*TFAP2A*	rs9348512	C/A	0.33	(114)	*BRCA2*	1 (0.99–1.02)	6.4×10^{-01}	8×10^{-01}
22q13.1	(60)	*APOBEC3A/APOBEC3B*	chr22:39359355	I/D7	0.1	(60)	Overall	1.1 (1.07–1.14)	6.1×10^{-09}	4.9×10^{-12}
1p36.22	(22)	*PEX14*	rs616488	A/G	0.33	(22)	Overall	0.94 (0.93–0.96)	1.8×10^{-09}	5×10^{-20}
1p13.2	(22)	*DCLRE1B*	rs11552449	C/T	0.17	(22)	Overall	1.04 (1.01–1.06)	2.3×10^{-03}	4.6×10^{-11}
2q14.1	(22)	—	rs4849887	C/T	0.1	(22)	Overall	0.91 (0.88–0.94)	1.2×10^{-10}	6.9×10^{-20}
2q31.1	(22)	*DLX2-AS1*	rs2016394	G/A	0.47	(22)	Overall	0.95 (0.94–0.97)	3.1×10^{-07}	6.2×10^{-12}
2q31.1	(22)	*CDCA7*	rs1550623	A/G	0.15	(22)	Overall	0.95 (0.93–0.98)	8.8×10^{-05}	5.4×10^{-10}

(continued on the following page)

Table 1.1. Risk associated with overall breast cancer for 172 SNPs (continued). Source: Lilyquist, Ruddy, Vachon, & Couch, 2018.

Locus	Locus discovery reference[a]	Genes	rs ID	Alleles	MAF[b]	Refined SNP reference[c]	Subtype[d]	Overall breast cancer risk (24)		
								OR[e] (95% CI)[f]	P^g	Combined P^h
2q35	(22)	*DIRC3*	rs16857609	C/T	0.26	(22)	Overall	1.06 (1.04–1.09)	1.4×10^{-09}	1.8×10^{-25}
3p26.1	(22)	*EGPT/ITPR1*	rs6762644	A/G	0.38	(22)	Overall	1.05 (1.03–1.07)	1.6×10^{-08}	4×10^{-18}
3p24.1	(22)	*TGFBR2*	rs12493607	G/C	0.34	(22)	Overall	1.05 (1.03–1.07)	4.9×10^{-07}	6.9×10^{-14}
4q24	(22)	*TET2*	rs9790517	C/T	0.23	(22)	Overall	1.04 (1.01–1.06)	1.1×10^{-03}	5×10^{-11}
4q34.1	(22)	*ADAM29*	rs6828523	C/A	0.12	(22)	Overall	0.91 (0.88–0.93)	1.2×10^{-11}	1.8×10^{-25}
5q11.2	(22)	*RAB3C*	rs10472076	T/C	0.38	(22)	Overall	1.03 (1.01–1.04)	6.6×10^{-03}	9.6×10^{-09}
5q11.2	(22)	*PDE4D*	rs1353747	T/G	0.09	(22)	Overall	0.96 (0.93–0.99)	7.6×10^{-03}	4.1×10^{-09}
5q33.3	(22)	*EBF1*	rs1432679	T/C	0.43	(22)	Overall	1.08 (1.06–1.1)	2.9×10^{-17}	6.6×10^{-31}
6p25.3	(22)	*FOXQ1*	rs11242675	T/C	0.37	(22)	Overall	1 (0.98–1.02)	9.7×10^{-01}	1×10^{-04}
6p23	(22)	*RANBP9*	rs204247	A/G	0.44	(22)	Overall	1.04 (1.02–1.06)	6.4×10^{-05}	7.9×10^{-13}
7q35	(22)	*NOBOX/ARHGEF6*	rs720475	G/A	0.25	(22)	Overall	0.96 (0.94–0.98)	3×10^{-04}	1.2×10^{-11}
8p12	(22)	—	rs9693444	C/A	0.32	(22)	Overall	1.06 (1.04–1.08)	1.7×10^{-10}	1.6×10^{-21}
8q21.11	(22)	—	rs6472903	T/G	0.17	(22)	Overall	0.94 (0.92–0.96)	4.4×10^{-21}	4.4×10^{-21}
8q21.11	(22)	*HNF4G*	rs2943559	A/G	0.08	(22)	Overall	1.1 (1.07–1.14)	4.2×10^{-09}	4×10^{-24}
8q24.21	(22)	*MYC*	rs11780156	C/T	0.17	(22)	Overall	1.05 (1.03–1.08)	2.5×10^{-05}	1.1×10^{-13}
9q31.2	(22)	—	rs10759243	C/A	0.29	(22)	Overall	1.06 (1.04–1.08)	4.2×10^{-10}	2.2×10^{-18}
10p12.31	(22)	*DNAJC1*	rs7072776	G/A	0.29	(22)	Overall	1.05 (1.03–1.07)	2.7×10^{-07}	1.8×10^{-19}
10p12.31	(22)	*DNAJC1*	rs11814448	A/C	0.02	(22)	Overall	1.12 (1.06–1.19)	1.4×10^{-06}	6.1×10^{-18}
10q25.2	(22)	*TCFL2*	rs7904519	A/G	0.46	(22)	Overall	1.03 (1.01–1.05)	8.6×10^{-04}	1.5×10^{-13}
10q26.12	(22)	—	rs11199914	C/T	0.32	(22)	Overall	0.96 (0.94–0.98)	2.1×10^{-05}	6.5×10^{-12}
11q13.1	(22)	—	rs3903072	G/T	0.47	(22)	Overall	0.97 (0.95–0.97)	9.1×10^{-04}	2.3×10^{-12}
11q24.3	(22)	—	rs11820646	C/T	0.4	(22)	Overall	0.96 (0.94–0.98)	2.5×10^{-05}	2.1×10^{-14}
12p13.1	(22)	—	rs12422552	G/C	0.26	(22)	Overall	1.06 (1.04–1.08)	3.2×10^{-08}	3.6×10^{-15}
12q22	(22)	*NTN4*	rs17356907	A/G	0.3	(22)	Overall	0.91 (0.9–0.93)	6.6×10^{-20}	1×10^{-39}
13q13.1	(22)	*BRCA2*	rs11571833	A/T	0.01	(22)	Overall	1.35 (1.23–1.48)	4×10^{-10}	3.1×10^{-15}
14q13.3	(22)	*PAX9*	rs2236007	G/A	0.21	(22)	Overall	0.93 (0.91–0.95)	5.8×10^{-10}	4.2×10^{-21}
14q24.1	(22)	*RAD51B*	rs2588809	C/T	0.17	(22)	Overall	1.06 (1.03–1.08)	2.6×10^{-06}	6.3×10^{-14}
14q32.11	(22)	*CCDC88C*	rs941764	A/G	0.35	(22)	Overall	1.03 (1.02–1.05)	3.6×10^{-04}	8.2×10^{-13}
16q12.2	(22)	*FTO*	rs17817449	T/G	0.41	(22)	Overall	0.95 (0.93–0.96)	4.9×10^{-09}	2.5×10^{-21}
16q23.2	(22)	*CDYL2*	rs13329835	A/G	0.23	(22)	Overall	1.07 (1.05–1.09)	8.3×10^{-11}	8.8×10^{-27}
18q11.2	(22)	—	rs527616	G/C	0.38	(22)	Overall	0.97 (0.95–0.98)	2.8×10^{-04}	6.7×10^{-15}
18q11.2	(22)	*CHST9*	rs1436904	T/G	0.4	(22)	Overall	0.95 (0.94–0.97)	1.1×10^{-07}	9.9×10^{-15}
19p13.11	(22)	*ELL*	rs4808801	A/G	0.34	(22)	Overall	0.93 (0.91–0.95)	2×10^{-12}	4.7×10^{-28}
19q13.31	(22)	*KCCN4/LYPD5*	rs3760982	G/A	0.46	(22)	Overall	1.05 (1.03–1.07)	2.1×10^{-08}	1.4×10^{-16}
22q12.2	(22)	*EMID1*	rs132390	T/C	0.04	(22)	Overall	1.04 (0.99–1.09)	8.1×10^{-02}	1.2×10^{-08}
22q13.1	(22)	*MKL1*	rs6001930	T/C	0.1	(22)	Overall	1.12 (1.09–1.16)	5.7×10^{-16}	4.4×10^{-34}
1q32.1	(72)	*ZC3H11A*	rs4951011	A/G	0.16	(72)	Overall	1.04 (1.02–1.07)	5.4×10^{-04}	1.3×10^{-05}
5q14.3	(72)	*ARRDC3*	rs10474352	C/T	0.16	(72)	Overall	0.94 (0.92–0.97)	4.5×10^{-11}	4.5×10^{-11}
15q26.1	(72)	*PRC1*	rs2290203	G/A	0.21	(72)	Overall	0.94 (0.92–0.96)	1.8×10^{-07}	8.07×10^{-10}
3p14.1	(115)	*ATNX7*	rs1053338	A/G	0.14	(115)	Overall	1.05 (1.02–1.07)	5×10^{-04}	5.3×10^{-11}
7q21.2	(115)	*AKAP9*	rs6964587	G/T	0.39	(115)	Overall	1.03 (1.02–1.05)	2.2×10^{-04}	9×10^{-11}
7q34	(116)	—	rs11977670	G/A	0.43	(116)	Lobular	1.06 (1.04–1.08)	1.9×10^{-11}	1×10^{-16}
1q21.1	(23)	*RNF115*	rs12405132	C/T	0.37	(23)	Overall	0.97 (0.95–0.99)	1×10^{-03}	6.3×10^{-10}
1q21.2	(23)	*OTUD7B*	rs12048493	A/C	0.38	(23)	Overall	1.04 (1.02–1.06)	9.4×10^{-06}	8.6×10^{-14}
1q43	(23)	*EXO1*	rs72755295	A/G	0.03	(23)	Overall	1.15 (1.09–1.2)	8×10^{-08}	1.7×10^{-14}
3p21.31	(23)	—	rs6796502	G/A	0.1	(23)	Overall	0.92 (0.89–0.95)	2.5×10^{-08}	5.5×10^{-15}
5p15.1	(23)	—	rs13162653	G/T	0.45	(23)	Overall	0.99 (0.97–1.010)	1.8×10^{-01}	5.4×10^{-07}
5p13.3	(23)	—	rs2012709	C/T	0.48	(23)	Overall	1.02 (1–1.04)	2.1×10^{-02}	1.2×10^{-08}

(continued on the following page)

Table 1.1. Risk associated with overall breast cancer for 172 SNPs (continued). Source: Lilyquist, Ruddy, Vachon, & Couch, 2018.

Locus	Locus discovery reference[a]	Genes	rs ID	Alleles	MAF[b]	Refined SNP reference[c]	Subtype[d]	Overall breast cancer risk (24)		
								OR[e] (95% CI)[f]	*P*[g]	Combined *P*[h]
5q14.2	(23)	*ATG10*	rs7707921	A/T	0.25	(23)	Overall	0.96 (0.94–0.98)	1.2×10^{-04}	1.7×10^{-12}
6p22.1	(23)	—	rs9257408	G/C	0.41	(23)	Overall	1.02 (1–1.040)	5.5×10^{-02}	6.9×10^{-08}
7q32.3	(23)	*FLJ43663*	rs4593472	C/T	0.35	(23)	Overall	0.97 (0.95–0.99)	7.9×10^{-04}	1.8×10^{-11}
8p11.23	(23)	—	rs13365225	A/G	0.18	(23)	Overall	0.91 (0.89–0.93)	1.2×10^{-15}	1.4×10^{-20}
8q23.3	(23)	*LINC00536*	rs13267382	G/A	0.36	(23)	Overall	1.03 (1.01–1.05)	8.7×10^{-04}	1.6×10^{-11}
14q32.12	(23)	*RIN3*	rs11627032	T/C	0.25	(23)	Overall	0.96 (0.94–0.98)	1.6×10^{-04}	4.1×10^{-11}
17q11.2	(23)	*ATAD5*	rs146699004	GGT/G	0.27	(23)	Overall	0.97 (0.95–0.99)	1.3×10^{-03}	2×10^{-09}
17q25.3	(23)	—	rs745570	G/A	0.5	(23)	Overall	1.03 (1.01–1.05)	2.1×10^{-03}	3.9×10^{-10}
18q12.3	(23)	*SETBP1*	rs6507583	A/G	0.07	(23)	Overall	0.92 (0.89–0.96)	9.5×10^{-06}	2.2×10^{-12}
2p23.2	(10)	*WDR43*	rs4577244	C/T	0.23	(10)	ER-/*BRCA1*	1.01 (0.99–1.03)	2.4×10^{-01}	4.3×10^{-01}
12p11.22	(10)	—	rs7297051	C/T	0.24	(10)	ER-/*BRCA1*	0.89 (0.87–0.91)	2.9×10^{-27}	3×10^{-60}
13q22.1	(10)	—	rs6562760	G/A	0.24	(10)	ER-/*BRCA1*	0.95 (0.93–0.97)	8.6×10^{-06}	1.5×10^{-09}
1p36.13	(24)	*KLHDC7A*	rs2992756	C/T	0.49	(24)	Overall	1.06 (1.04–1.08)	1.3×10^{-11}	1.6×10^{-15}
1p34.2	(24)	—	rs4233486	T/C	0.36	(24)	Overall	0.97 (0.95–0.98)	2.3×10^{-04}	9.1×10^{-09}
1p34.2	(24)	*HIVEP3*	rs79724016	T/G	0.03	(24)	Overall	0.93 (0.88–0.97)	3.3×10^{-03}	3.5×10^{-08}
1p34.1	(24)	*PIK3R3, LOC101929626*	rs1707302	G/A	0.34	(24)	Overall	0.96 (0.95–0.98)	1.4×10^{-04}	3.0×10^{-08}
1p32.3	(24)	—	rs140850326	I/D9	0.49	(24)	Overall	0.97 (0.95–0.99)	3.4×10^{-04}	3.9×10^{-08}
1p22.3	(24)	—	rs17426269	G/A	0.15	(24)	Overall	1.05 (1.02–1.07)	1.7×10^{-04}	1.7×10^{-08}
1p12	(24)	—	rs7529522	T/C	0.23	(24)	Overall	1.06 (1.04–1.08)	1.6×10^{-08}	1.7×10^{-10}
1q22	(24)	*TRIM46*	rs4971059	G/A	0.35	(24)	Overall	1.05 (1.03–1.07)	3.9×10^{-08}	4.8×10^{-11}
1q32.1	(24)	*PHLDA3*	rs35383942	C/T	0.06	(24)	Overall	1.12 (1.08–1.17)	12×10^{-09}	3.8×10^{-13}
1q41	(24)	*ESRRG*	rs11117758	G/A	0.21	(24)	Overall	0.95 (0.93–0.97)	7.7×10^{-07}	3.9×10^{-09}
2p25.1	(24)	*GRHL1*	rs113577745	C/G	0.1	(24)	Overall	1.08 (1.05–1.11)	3.7×10^{-07}	3.9×10^{-10}
2p23.3	(24)	*ADCY3*	rs6725517	A/G	0.41	(24)	Overall	0.96 (0.94–0.98)	7.5×10^{-06}	2.9×10^{-12}
2q13	(24)	*BCL2L11*	rs71801447	CTTATGTT/C	0.06	(24)	Overall	1.09 (1.05–1.13)	7.7×10^{-06}	3.7×10^{-08}
2q36.3	(24)	—	rs12479355	A/G	0.21	(24)	Overall	0.96 (0.94–0.98)	4.7×10^{-04}	2.4×10^{-08}
3p13	(24)	*FOXP1*	rs6805189	T/C	0.48	(24)	Overall	0.97 (0.95–0.99)	3.3×10^{-04}	4.6×10^{-08}
3p12.1	(24)	*VGLL3*	rs13066793	A/G	0.09	(24)	Overall	0.94 (0.91–0.97)	1.5×10^{-04}	1.0×10^{-09}
3p12.1	(24)	*CMSS1, FILIP1L*	rs9833888	G/T	0.22	(24)	Overall	1.06 (1.04–1.08)	2.6×10^{-07}	5.2×10^{-10}
3q23	(24)	*ZBTB38*	rs34207738	CTT/C	0.41	(24)	Overall	1.06 (1.04–1.08)	1.4×10^{-09}	3.2×10^{-15}
3q26.31	(24)	—	rs58058861	G/A	0.21	(24)	Overall	1.06 (1.04–1.09)	1.6×10^{-08}	1.9×10^{-10}
4p14	(24)	—	rs6815814	A/C	0.26	(24)	Overall	1.06 (1.04–1.08)	6.1×10^{-08}	6.1×10^{-13}
4q21.23	(24)	*HELQ*	4:84370124	TA/TAA	0.47	(24)	Overall	1.04 (1.02–1.05)	1.7×10^{-04}	2.2×10^{-09}
4q22.1	(24)	*LOC105369192*	rs10022462	C/T	0.44	(24)	Overall	1.04 (1.02–1.06)	9.4×10^{-06}	1.6×10^{-09}
4q28.1	(24)	—	rs77528541	G/T	0.13	(24)	Overall	0.95 (0.92–0.97)	4.8×10^{-05}	1.4×10^{-09}
5p15.33	(24)	*AHRR*	rs116095464	T/C	0.05	(24)	Overall	1.06 (1.02–1.1)	2.6×10^{-03}	3.8×10^{-09}
5q11.1	(24)	—	rs72749841	T/C	0.16	(24)	Overall	0.93 (0.91–0.96)	8.5×10^{-06}	7.2×10^{-10}
5q11.1	(24)	—	rs35951924	A/AT	0.32	(24)	Overall	0.95 (0.93–0.97)	4.0×10^{-07}	1.3×10^{-11}
5q22.1	(24)	*NREP*	rs6882649	T/G	0.34	(24)	Overall	0.97 (0.95–0.99)	2.7×10^{-03}	3.7×10^{-09}
5q31.1	(24)	*HSPA4*	rs6596100	C/T	0.25	(24)	Overall	0.94 (0.92–0.96)	5.2×10^{-08}	7.7×10^{-09}
5q35.1	(24)	—	rs4562056	G/T	0.33	(24)	Overall	1.05 (1.03–1.07)	4.1×10^{-07}	4.7×10^{-10}
6p22.3	(24)	*ATXN1*	rs3819405	C/T	0.33	(24)	Overall	0.96 (0.94–0.97)	2.2×10^{-06}	1.7×10^{-08}
6p22.3	(24)	*CDKAL1*	rs2223621	C/T	0.38	(24)	Overall	1.04 (1.02–1.06)	1.0×10^{-04}	3.0×10^{-10}
6p22.2	(24)	—	rs71557345	G/A	0.07	(24)	Overall	0.92 (0.88–0.96)	8.4×10^{-05}	3.9×10^{-10}
6q14.1	(24)	—	rs12207986	A/G	0.47	(24)	Overall	0.97 (0.95–0.98)	2.0×10^{-04}	1.5×10^{-09}
6q23.1	(24)	*L3MBTL3*	rs6569648	T/C	0.24	(24)	Overall	0.94 (0.92–0.96)	4.8×10^{-08}	3.0×10^{-12}
7p15.3	(24)	*DNAH11, CDCA7L*	rs7971	A/G	0.35	(24)	Overall	0.96 (0.94–0.98)	1.4×10^{-05}	1.9×10^{-08}
7p15.1	(24)	*CREB5*	rs17156577	T/C	0.11	(24)	Overall	1.05 (1.02–1.08)	3.8×10^{-04}	4.3×10^{-09}

(continued on the following page)

Table 1.1. Risk associated with overall breast cancer for 172 SNPs (continued). Source: Lilyquist, Ruddy, Vachon, & Couch, 2018.

Locus	Locus discovery reference[a]	Genes	rs ID	Alleles	MAF[b]	Refined SNP reference[c]	Subtype[d]	Overall breast cancer risk (24)		
								OR[e] (95% CI)[f]	*P*[g]	Combined *P*[h]
7q21.3	(24)	—	rs17268829	T/C	0.28	(24)	Overall	1.05 (1.03–1.07)	1.3×10^{-06}	4.5×10^{-13}
7q22.1	(24)	*CUX1*	rs71559437	G/A	0.12	(24)	Overall	0.93 (0.91–0.96)	9.1×10^{-07}	5.1×10^{-12}
8q22.3	(24)	—	rs514192	T/A	0.32	(24)	Overall	1.05 (1.03–1.07)	3.7×10^{-06}	5.6×10^{-09}
8q23.1	(24)	*ZFPM3*	rs12546444	A/T	0.1	(24)	Overall	0.93 (0.91–0.96)	5.8×10^{-06}	7.5×10^{-11}
8q24.13	(24)	—	rs58847541	G/A	0.15	(24)	Overall	1.08 (1.05–1.1)	7.3×10^{-09}	5.5×10^{-13}
9q33.1	(24)	*ASTN2*	rs1895062	A/G	0.41	(24)	Overall	0.94 (0.92–0.95)	6.9×10^{-13}	1.1×10^{-14}
9q33.3	(24)	*LMX1B*	rs10760444	A/G	0.43	(24)	Overall	1.03 (1.02–1.05)	2.8×10^{-04}	9.1×10^{-09}
9q34.2	(24)	*ABO*	rs8176636	I/D10	0.2	(24)	Overall	1.03 (1.01–1.06)	3.2×10^{-03}	1.4×10^{-08}
10p14	(24)	—	rs67958007	TG/T	0.12	(24)	Overall	1.09 (1.06–1.12)	1.8×10^{-09}	1.7×10^{-10}
10q23.33	(24)	—	rs140936696	C/CAA	0.18	(24)	Overall	1.04 (1.02–1.07)	7.4×10^{-04}	4.2×10^{-08}
11p15	(24)	*PIDD1*	rs6597981	G/A	0.48	(24)	Overall	0.96 (0.94–0.97)	5.7×10^{-07}	1.4×10^{-12}
12q21.31	(24)	—	rs202049448	T/C	0.34	(24)	Overall	0.95 (0.93–0.97)	2.5×10^{-07}	2.7×10^{-08}
12q24.31	(24)	—	rs206966	C/T	0.16	(24)	Overall	1.05 (1.02–1.07)	2.7×10^{-04}	3.8×10^{-08}
14q32.33	(24)	*ADSSL1*	rs10623258	C/CTT	0.45	(24)	Overall	1.04 (1.02–1.06)	2.7×10^{-05}	2.3×10^{-08}
16q12.2	(24)	—	rs28539243	G/A	0.49	(24)	Overall	1.05 (1.03–1.07)	3.6×10^{-08}	9.1×10^{-15}
16q13	(24)	*AMFR*	rs2432539	G/A	0.4	(24)	Overall	1.03 (1.02–1.05)	3.1×10^{-04}	4.0×10^{-08}
16q24.2	(24)	—	rs4496150	C/A	0.25	(24)	Overall	0.96 (0.94–0.98)	3.4×10^{-05}	8.1×10^{-09}
17q21.2	(24)	*CNTNAP1*	rs72826962	C/T	0.01	(24)	Overall	1.2 (1.11–1.3)	5.1×10^{-06}	4.6×10^{-09}
17q21.31	(24)	*KANSL1*	rs2532263	G/A	0.19	(24)	Overall	0.95 (0.93–0.97)	4.7×10^{-06}	6.9×10^{-13}
18q12.1	(24)	*GAREM1*	rs117618124	T/C	0.05	(24)	Overall	0.89 (0.85–0.92)	4.5×10^{-08}	5.5×10^{-12}
19p13.13	(24)	*NFIX1*	rs78269692	T/C	0.05	(24)	Overall	1.09 (1.04–1.13)	3.9×10^{-05}	1.9×10^{-09}
19p13.12	(24)	—	rs2594714	G/A	0.23	(24)	Overall	0.97 (0.95–0.99)	6.7×10^{-03}	1.1×10^{-08}
19p13.11	(24)	*GATAD2A, MIR640*	rs2965183	G/A	0.35	(24)	Overall	1.04 (1.02–1.06)	9.6×10^{-06}	6.3×10^{-12}
19q13.22	(24)	*GIPR*	rs71338792	A/AT	0.23	(24)	Overall	1.05 (1.03–1.07)	8.1×10^{-06}	3.5×10^{-09}
20p12.3	(24)	*MCM8*	rs16991615	G/A	0.06	(24)	Overall	1.1 (1.06–1.14)	1.4×10^{-07}	1.9×10^{-09}
20q13.13	(24)	—	rs6122906	A/G	0.18	(24)	Overall	1.05 (1.03–1.07)	2.9×10^{-05}	2.5×10^{-10}
22q13.1	(24)	*PLA2G6*	rs738321	C/G	0.38	(24)	Overall	0.95 (0.93–0.97)	2.7×10^{-08}	1.0×10^{-13}
22q13.2	(24)	*XRCC6*	rs73161324	C/T	0.06	(24)	Overall	1.06 (1.02–1.09)	3.8×10^{-03}	2.0×10^{-09}
22q13.31	(24)	—	rs28512361	G/A	0.11	(24)	Overall	1.05 (1.02–1.08)	5.7×10^{-04}	2.3×10^{-08}

[a]Risk locus was originally discovered in the associated reference.
[b]MAF reported in OncoArray controls.
[c]Reference for the refined SNP of interest.
[d]Breast cancer subtype in which the refined SNP was evaluated.
[e]ORs reported for overall breast cancer OncoArray analysis.
[f]95% confidence interval (CI) reported for overall breast cancer OncoArray analysis.
[g]*P* values reported for overall breast cancer OncoArray analysis.
[h]Combined meta-analysis *P* values reported for overall breast cancer.

Fine-Mapping Studies

Although GWAS have led to the identification of over 170 variants associated with breast cancer risk in the past 15 years (Michailidou, et al., 2017), methodological limitations have challenged the identification of the true causal variants, their target genes nor the underlying functional mechanisms for the majority of those loci (Fachal & Dunning, 2015). In order to address these issues, fine-mapping studies have emerged in the post-GWAS era to provide more insights into the underlying biological mechanisms involved in risk-associated loci. Several fine-mapping approaches have been used to functionally understand which are the variants causing the effects detected by GWAS and through which mechanisms those variants act. An important feature of fine-mapping studies is the fact that cell-type and disease-specific context in taken into account. This is of high importance since the effects of non-coding causal variants can be very specific according to cell-type, tissue, context and disease (Zhao, et al., 2014). For instance, this specificity determines which are the active/repressed regulatory regions (promoters, enhancers) and which transcription factors are expressed, and thus which variants, genes and pathways are involved (Figure 1.3).

Despite the complexity associated with the characterisation of each GWAS-risk locus, 22 of the GWAS loci identified for breast cancer had been analysed in detail by 2018 (Meyer, et al., 2008; Udler, et al., 2009; Ahmadiyeh, et al., 2010; Udler, et al., 2010; Stacey, et al., 2010; Beesley, et al., 2011; Cai, et al., 2011; Bojesen, et al., 2013; French, et al., 2013; Meyer, et al., 2013; Ghoussaini, et al., 2014; Quigley, et al., 2014; Darabi, et al., 2015; Glubb, et al., 2015; Guo, et al., 2015; Lin, et al., 2015; Orr, et al., 2015; Darabi, et al., 2016; Dunning, et al., 2016; Ghoussaini, et al., 2016; Hamdi, et al., 2016; Horne, et al., 2016; Lawrenson, et al., 2016; Shi, et al., 2016; Sun, et al., 2016; Wyszynski, et al., 2016; Zeng, et al., 2016; Betts, et al., 2017; Helbig, et al., 2017; Michailidou, et al., 2017). Using several tools and strategies for fine-scale mapping and functional analysis, those studies led to a greater understanding of the mechanisms underlying breast cancer susceptibility. Although the majority of fine-mapping studies analyse a specific locus, a recent study performed fine-mapping of 150 of all genomic regions harbouring breast cancer risk variants from GWAS (Fachal, et al., 2020). In this work multiple, complementary analyses identified 205 variants associated with risk (p-value<10^{-6}) and 191 likely target genes. Identification of these new variants into account, increased the percentage ok known breast cancer familial risk in 6%.

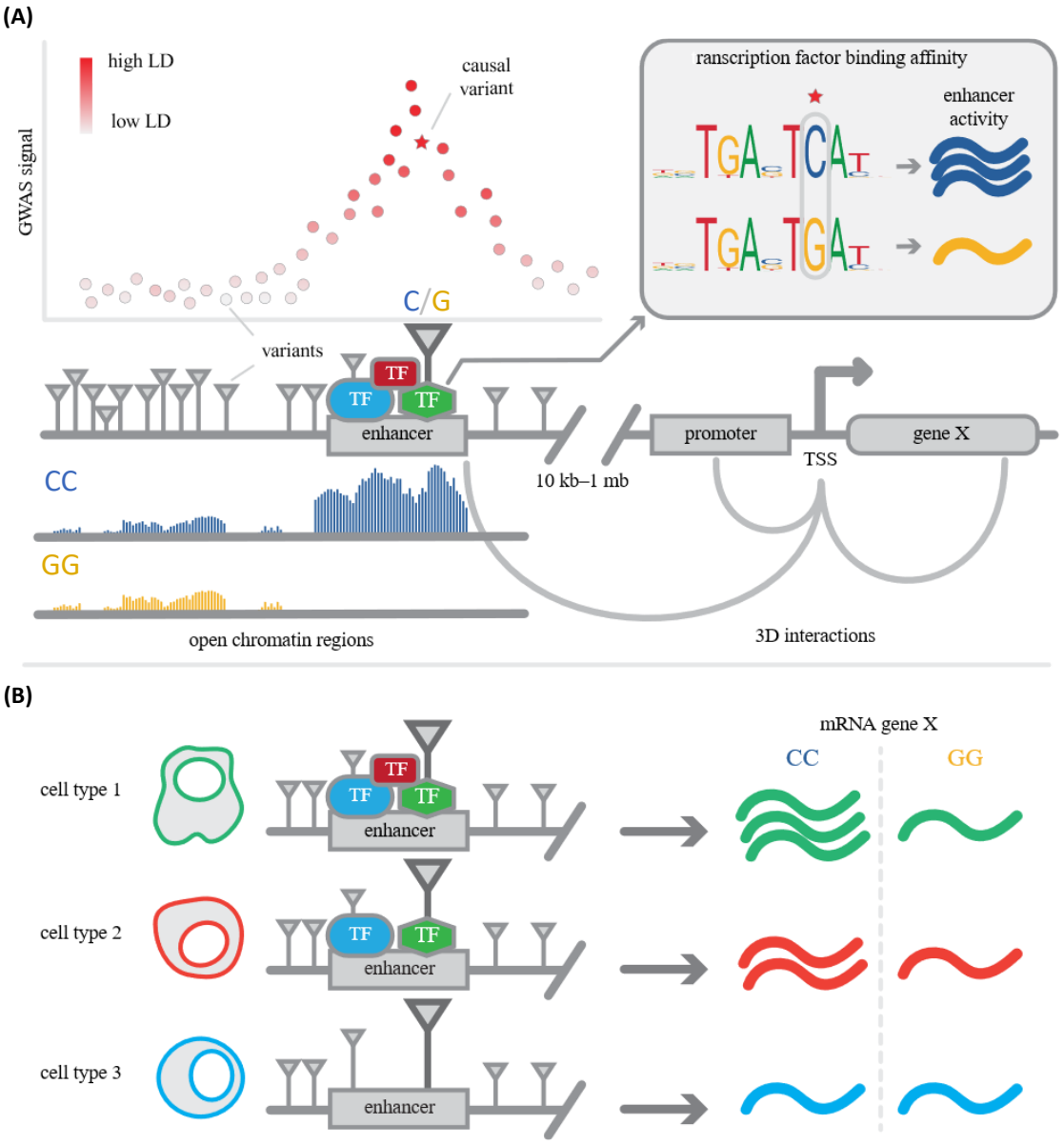


Figure 1.3. Illustration of a GWAS locus, where altered enhancer activity is depicted as an example of a mechanism of action of a causal variant. **(A)** Variants in different degree of LD with the GWAS variant are represented in the scatterplot. This example shows a causal variant (red star) located in an active enhancer, as shown by the open chromatin regions of the same locus (represented by the peak-density plot below the variant) The causal variant alters the binding of the green TF, which shows preferential binding to the C-allele, as shown by the higher number of blue transcripts comparing to the number of yellow transcripts associated with the G-allele. Grey arches connecting gene X to the promoter and enhancer represent 3D interactions, which indicate that the activity of the enhancer affected by the causal variant impacts the expression of gene X. **(B)** Different gene expression according to cell-type. The figure represents the effect that the causal variant has in three different cell-types. The levels of mRNA differ between cell-types if the homozygous genotype for the C-allele occurs. Due to the higher affinity to the green TF, mRNA levels decrease with the decrease in the TF availability. However, a homozygous genotype for the G-allele results in low but stable levels of mRNA regardless the type of cell. Adapted from Broekema, Bakker, & Jonkers, 2020.

1.3. Missing Heritability and Regulation of Gene Expression

Although the disease-associated SNPs identified by GWAS do not explain total heritability of complex diseases and traits (Korf, 2013), they have clinical implications, particularly when combined into polygenic risk scores. The odds ratio (OR) for a given tag SNP from GWAS is often less than 1.2, but the combined effects of a large number of risk loci may be large enough to be valuable in risk prediction and targeted prevention, including the stipulation of screening frequency (Muranen, et al., 2016). Polygenic scores from GWAS can be used to model the proportion of overall heritability from common variants (Park, et al., 2011). Previously described polygenic models (International Schizophrenia Consortium, et al., 2009; Yang, et al., 2010) estimate the total proportion of genetic variation that can be explained by all genotyped common SNPs, even if each one of these do not show individual significant association (after correcting for multiple testing).

On the other hand, GWAS have not been able to deliver direct evidence about the biological processes that link the associated variant to the phenotype. Adding to the fact that most of the common disease-associated variants are likely to have small effect sizes, the fact that most associations point to larger genomic regions of correlated variants represents a major challenge in interpreting GWAS data. Regions of strong LD tend to be large, and since the SNPs associated with the phenotype have been found to be in perfect LD with SNPs hundred kilobases away, the tag SNP is not necessarily the functional source of the association signal.

Adding to the difficulty of unequivocally identifying the causal variant for each locus, biological interpretation of GWAS data is impaired by the fact that around 90% of the disease-associated variants are not changing the protein structure, but are located in non-coding regions of the genome (Hindorff, et al., 2009; Pastinen, 2010; Schaub, Boyle, Kundaje, Batzoglou, & Snyder, 2012; Maurano, et al., 2012; Edwards, Beesley, French, & Dunning, 2013), and are over-represented in regulatory elements (Skelly, Ronald, & Akey, 2009; Maurano, et al., 2012; Gaffney, 2013; Battle & Montgomery, 2014). Thus, it is likely that the underlying mechanism linking the associated SNPs identified by GWAS to the phenotype is regulatory, meaning that those variants might affect disease risk by altering the genetic regulation of one or more target genes.

Regulatory variation can interfere with any of the steps along the gene expression cascade from DNA to protein, including transcription, and post-transcriptional processes like mRNA splicing, and mRNA degradation (Figure 1.4) (Zhang, et al., 2011; Gaffney, 2013; Richardson, et al., 2013;

Paraboschi, et al., 2014; Wang, et al., 2014; Pai, Pritchard, & Gilad, 2015). Studies that assessed the way that sequence variation affects the individual steps of gene expression, have characterized functional variants acting through transcription factor (TF) binding (Kasowski, et al., 2010; McDaniell, et al., 2010; Reddy, et al., 2012; Heinz, et al., 2013; Ding, et al., 2014; Tehranchi, et al., 2016), chromatin accessibility (Degner, et al., 2012; Lee, et al., 2013), DNA methylation (Bell, et al., 2011; Gutierrez-Arcelus, et al., 2013; Heyn, et al., 2013; Banovich, et al., 2014; McRae, et al., 2014; Kaplow, et al., 2015; Hannon, et al., 2016), alternative splicing (Montgomery, et al., 2010; Pickrell, Pai, Gilad, & Pritchard, 2010; Pickrell, et al., 2010; Lappalainen, et al., 2013; Hassan, Butty, Jensen, & Saeij, 2014), small RNAs (Parts, et al., 2012; Civelek, et al., 2013; Lappalainen, et al., 2013), large intergenic non-coding RNAs (lincRNAs) (Kumar, et al., 2013; Popadin, Gutierrez-Arcelus, Dermitzakis, & Antonarakis, 2013), RNA editing (Hassan, Butty, Jensen, & Saeij, 2014), and mRNA degradation (Pai, et al., 2012).

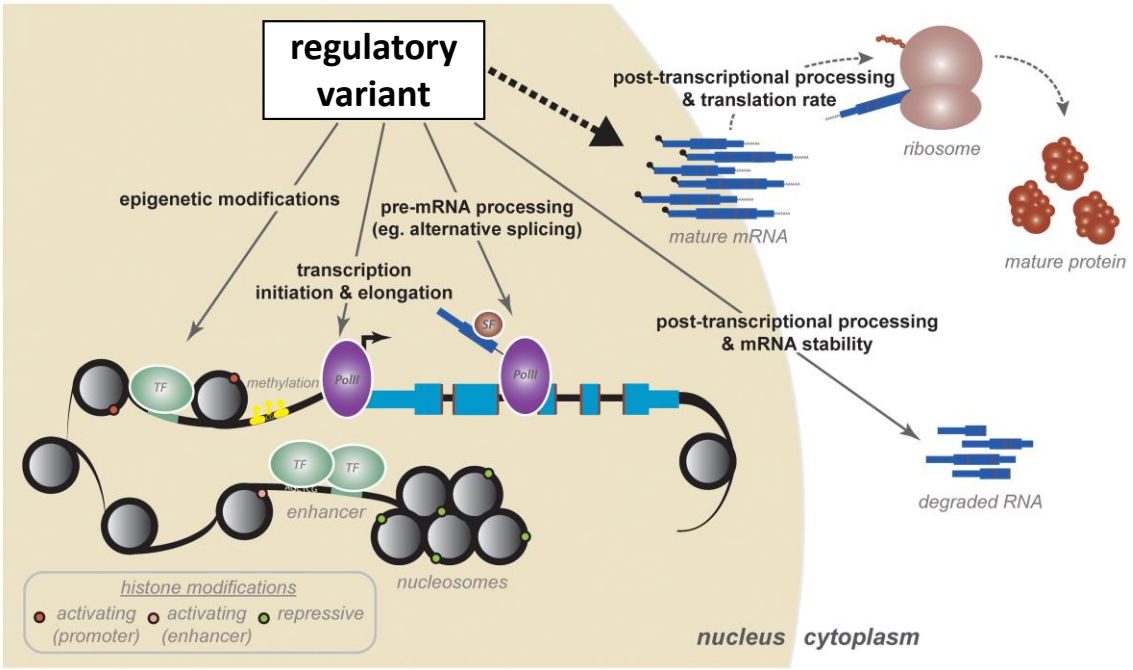


Figure 1.4. Cis-regulatory mechanisms by which a genetic variant can affect gene expression. Regulatory variants (rSNPs) can locate in different sites of the genome, affecting several steps of gene expression. rSNPs can alter histone modifications and transcription initiation, by altering DNA methylation, enhancer and/or promoter activity, and binding of transcription factors. The latest, is likely the strongest contributor to variation in mRNA levels. rSNPs can also affect transcription elongation (by PolII activity rates) and pre-mRNA processing, especially alternative splicing, which are major contributors to variation in gene expression levels and isoforms variety. rSNPs can impact post-transcriptional mRNA processing and stability, through mechanisms like miRNAs binding, or impaired polyadenylation. Adapted from Pai, Pritchard, & Gilad, 2015.

Previous studies on histone modification, TF binding and mRNA levels in human parent-offspring trios indicate that much of genetic variation affects TF binding (Göring, et al., 2007; Musunuru, et al., 2010; Kasowski, et al., 2013; Kilpinen, et al., 2013; Lee, et al., 2014; McVicker, et al., 2013). Variants located in TF binding sites are correlated with altered binding of the TF itself, and also with differential histone modifications, DNA methylation, and mRNA levels (Banovich, et al., 2014). This can be explained by the fact that the primary molecular change is the differential TF binding, which in turn, alters histone modifications, DNA methylation and mRNA expression. Nonetheless, only around a third of altered TF binding is due to the existence of genetic variants in the core of the TF binding motif (Kasowski, et al., 2013; Kilpinen, et al., 2013; McVicker, et al., 2013). TF binding can also be altered by variants outside core motifs, which can alter the local shape of DNA or influence the binding of other TFs in potential complexes. For example, the deletion of one nucleotide does not disrupt a TF binding motif, but puts two TF binding sites closer together, reducing their binding (Chang, et al., 2013).

Translating GWAS results into trait or disease insights involves clarifying molecular mechanisms, by which gene expression is influenced by genetic variants, and biological mechanisms, by which target genes influences a trait or disease. Finding these mechanisms entails to determine the number of association signals at a locus, to identify the candidate causal variant(s), and to identify the target gene(s).

Different approaches have been developed to identify variants that are likely to play a relevant biological role. Studies characterising gene regulation and regulatory variation have been able to assess the correlation between variation in different regulatory mechanisms and variation in mRNA levels, besides inferring the probable causal relationship between genetic variation, changes in regulatory interactions, and differences in gene expression levels.

Gene expression is controlled by genetic factors, which can act both in cis and in trans, epigenetic factors and environmental influences (Stranger, et al., 2007). Previous studies demonstrated that gene expression levels are strongly heritable and estimated that the average heritability for all human genes is approximately 0.25 (Monks, et al., 2004; Dixon, et al., 2007; Stranger, et al., 2007). Effects of genetic variation on gene expression have been studied in humans (Grundberg, et al., 2011; Majewski & Pastinen, 2011; Nica, et al., 2011; Trynka, et al., 2011; Grundberg, et al., 2012; Stranger, et al., 2012), mice (Tesson & Jansen, 2009; van Nas, et al., 2010), and other organisms (Keurentjes, et al., 2007). In order to identify candidate target genes, two approaches are commonly used to correlate gene transcription levels with genetic variants: expression quantitative trait loci (eQTLs) analysis (Schadt, et al., 2003; Lappalainen, et

al., 2013; GTEx Consortium, 2015), and allele-specific expression (ASE) – also termed allelic expression imbalance (AEI) or differential allelic expression (DAE) – analysis (Pickrell, et al., 2010; Sun, 2012; Hu, Sun, Tzeng, & Perou, 2015; van de Geijn, McVicker, Gilad, & Pritchard, 2015; Kumasaka, Knights, & Gaffney, 2016). Both methods use genotype and mRNA transcription profiles. Levels of mRNA can be quantified for thousands of genes simultaneously, by either using microarray expression data or by conducting RNA-sequencing (RNA-seq).

While eQTLs and ASE can be analysed using sets of non-diseased samples, since allele frequencies between cases and controls can differ, previous eQTL and ASE studies aiming to extract biological information from GWAS have used data from publicly databases, such as TCGA (Tomczak, Czerwińska, & Wiznerowicz, 2015) and The Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2015), to look for variants that are potentially relevant for cancer.

1.3.1. Analysis of expression Quantitative Trait Loci (eQTLs)

The concept of genetical genomics was introduced in 2001 to identify genes that are regulated by genetic variation (Jansen & Nap, 2001). In this study, genetic variants were correlated with intermediate molecular quantitative traits, including levels of gene expression protein and methylation, to identify quantitative trait loci (QTLs).

Analysis of eQTLs, that is, loci with total levels of expression regulated by specific genetic variation, was the first method used to extract biological information from GWAS. This approach assesses the functional effect of a given genetic variant by correlating genotyping data with the bi-allelic net effect in large populations.

eQTLs are usually divided into cis-eQTLs and trans-eQTLs. Classically, cis-eQTLs are defined as genetic variants that affect expression of a locus on the same DNA molecule and not the expression of the copy on the homologous chromosome, thus in an allele-specific fashion. Trans-eQTLs are due to variants that alter structure, function or expression of a diffusible factor, thus affecting gene expression through equal influence on both chromosomes, in an allele independent manner (Benzer, 1955). In eQTLs studies is commonly accepted that local associations, that is, loci located near the affected genes, likely act in a cis fashion. This led to local and distant (where the loci are located further away from the genes they affect) associations being often called cis and trans, respectively, in the literature (Rockman & Kruglyak, 2006). However, some advocate that the appropriate precise terminology should be used to

clearly distinguish relative position from mode of action (Albert & Kruglyak, 2015), which is also the case in the present study.

Although it was early observed that some eQTLs are located near the genes they affect (local eQTLs), these can act both in cis and in trans (Doss, Schadt, Drake, & Lusis, 2005; Ronald, Brem, Whittle, & Kruglyak, 2005; Rockman & Kruglyak, 2006). Similarly, distant eQTLs usually act in trans, but a trans-eQTLs can be located anywhere in the genome relative to the genes they regulate, and when located near a given gene, they will be local but not cis-acting eQTLs. Moreover, studies that report comparison between local and cis eQTLs suggest that the overlap is low (Hasin-Brumshtein, et al., 2014). In addition, the precise distance required for an eQTL to be considered local or distant is arbitrary, and it can refer to physical or genetic distance, and consequently it differs between studies (Brem, Yvert, Clinton, & Kruglyak, 2002; Monks, et al., 2004; Morley, et al., 2004; Göring, et al., 2007; Stranger, et al., 2007).

Several eQTLs studies (Morley, et al., 2004; Dixon, et al., 2007; Stranger, et al., 2007) used transformed cell lines, mainly Epstein-Barr immortalised lymphoblastoid cell lines (LCLs), and assessed differences between individuals, regardless of the inter-individual variability of gene expression that is not explained by the genetic cis-variation. Furthermore, initial eQTL studies based on combining genome-wide genotyping data with quantification of mRNA transcripts using microarray technology, showed limited accuracy inherent to those hybridisation-based gene expression assays. Thus, the high variability provided by environmental factors, as well as other genetic and epigenetic variability that is not accounted for, significantly reduces the statistical power for eQTL discovery.

1.3.2. Analysis of Allelic Expression

Allele-specific mapping approaches can also help to shed some lights on biological information obtainable from GWAS data, while addressing limitations of eQTL analysis. Cis-acting variation is predicted to regulate the expression of a large proportion of human genes (Dixon, et al., 2007; Göring, et al., 2007; Stranger, et al., 2007), contributing to both population-selective effects (Kudaravalli, Veyrieras, Stranger, Dermitzakis, & Pritchard, 2009; Torgerson, et al., 2009), and common disease signal (Nica, et al., 2010; Nicolae, et al., 2010; Gamazon, Nicolae, & Cox, 2011). Thus, cis-acting regulatory variants, or variants in LD with them, are valuable markers for association studies. Their identification broadens the understanding of variants influencing gene expression (Ge, et al., 2009; Cheung, et al., 2010; Pickrell, et al., 2010) and can lead to the

identification of causal SNPs and target genes (Cookson, Liang, Abecasis, Moffatt, & Lathrop, 2009; Fogarty, Xiao, Prokunina-Olsson, Scott, & Mohlke, 2010; Speliotes, et al., 2010).

While eQTLs approaches correlate genotypes and total gene expression across individuals, allele-specific approaches are based on the direct comparison of alleles in single individuals. Thus, allele-specific analysis requires smaller sample sizes and, isolated cis effects and more completely covers the genetic complexity, although only heterozygous individuals are informative, and data processing and analysis increases in complexity.

Trans-acting variants affecting gene expression influence equally both chromosomes, and thus, no imbalance of allelic expression is observed. On the other hand, cis-regulatory variants directly alter the expression of a gene in an allele-specific manner. Therefore, allelic expression (AE) displayed by heterozygotes for regulatory SNPs (rSNPs), is one of the hallmarks of cis-acting variation.

Differential allelic expression (DAE) is as a quantitative trait. Using ratios to detect the differences in the relative expression levels of transcripts carrying each allele within a heterozygous sample, isolates cis-acting effects, while controlling for trans-acting and environmental ones (Pastinen & Hudson, 2004; Verlaan, et al., 2009; Pastinen, 2010). Hence, this approach can improve the statistical power for detecting cis-variation over total expression alone (Sun, 2012), besides detecting epigenetic effects (Pollard, et al., 2008; Verlaan, et al., 2009). The extreme case of DAE is monoallelic expression, which refers to the exclusive expression of one of the alleles, while the other remains completely inactive. Chromosome X inactivation and parental imprinting are known examples of monoallelic expression, but a continuous spectrum of less evident DAE can originate from cis-regulatory variation.

The relative levels of mRNA transcripts carrying each allele can be measured by genotyping a transcribed SNP (tSNP) after reverse transcription of RNA to cDNA, or by RNA-seq. To date, several studies have analysed DAE (Yan, Yuan, Velculescu, Vogelstein, & Kinzler, 2002; Pastinen & Hudson, 2004; Verlaan, et al., 2009; Gregg, et al., 2010; Almlöf, et al., 2012; DeVeale, van der Kooy, & Babak, 2012; Grundberg, et al., 2012; Hasin-Brumshtein, et al., 2014; Baran, et al., 2015; Mohammadi, Castel, Brown, & Lappalainen, 2017), and led to the identification of hundreds to thousands of genes that show significant imbalance in the expression of the two alleles. Numerous studies compared the AE ratio in cDNA to the AE ratio in genomic DNA (gDNA) from the same samples, using gDNA as a reference for equal amounts of the two alleles, assuming that any technical bias in measurement of AE ratios is the same for both cDNA and gDNA (Bray, et al., 2004; Pant, et al., 2006; Campino, et al., 2008). If a single rSNP exists and it is in complete

LD with the tSNP, the observed AE imbalance will be consistently higher or lower than the gDNA levels and the mean AE ratios from cDNA and gDNA can be compared using a statistical test (Bray, et al., 2004; Campino, et al., 2008). On the other hand, if the rSNP is not in complete LD with the tSNP, the AE ratios will be distributed in a given range and will be dependent on the haplotypes including both the rSNP and tSNP present in the study sample (Xiao & Scott, 2011).

Currently, identification of candidate target genes through analysis of imbalance in allelic expression is typically assessed from RNA-seq data by directly counting the number of reads with each one of the alleles mapping to the candidate genes, within a certain distance from the GWAS SNP.

To date, most studies used unphased DAE analysis, meaning that the transcribed SNPs displaying allelic imbalance are not phased with the GWAS variant (Xiao & Scott, 2011; Li, et al., 2013; Sigurdsson, et al., 2016), and thereby it is not possible to infer if the risk-associated variant promotes or reduces the transcription of the target gene. In few studies using phased data (Conde, Bracci, Richardson, Montgomery, & Skibola, 2013), phasing accuracy has not been thoroughly evaluated. Although RNA-seq reads that span over two or more SNPs provide direct information on phase, the phasing accuracy tends to decline when the haplotypes extend to the candidate regulatory variant, which is often located in a non-coding region and hundred kilobases (kb) away from the target gene (Hu, Sun, Tzeng, & Perou, 2015).

A relevant constraint in studying cis-regulation is the fact that the transcriptome ultimately depends on the cell type and developmental stage. Previous reports showed tissue specificity of DAE in mice (Campbell, Kirby, Nemesh, Daly, & Hirschhorn, 2008) and between human brain regions (Verlaan, et al., 2009; Buonocore, et al., 2010; Gregg, et al., 2010; Almlöf, et al., 2012; Grundberg, et al., 2012). Thus, it is important to characterise cis-regulatory variation for individual tissue types.

The role of DAE in cancer has been studied previously, including in the breast (Chen, et al., 2008; Azzato, et al., 2010; Przytycki & Singh, 2020), colon (Yan, et al., 2002; Valle, et al., 2008), pancreas (Tan, et al., 2008), and ovary (Shen, Medico, & Zhao, 2011).

1.4. Context and Hypothesis

Breast cancer remains the most common cancer affecting women worldwide. Nevertheless, currently, approximately half of the familial risk for breast cancer remains unknown. Over the last decade, genome-wide association studies (GWAS) and the Collaborative Oncological Gene-Environment Study (COGS) made continuous efforts to identify common genetic variants that could explain the missing heritability of breast cancer. So far, low-risk loci identified by those studies explain up to 18% of the familial relative risk (Michailidou et al. 2017), and recently, a fine-mapping study of 150 of all genomic regions harbouring breast cancer risk variants from GWAS increased the percentage ok known breast cancer familial risk in 6%. (Fachal, et al., 2020). However, several factors have contributed to hinder the bridge from statistical associations between genotype and phenotype to the biology underlying breast cancer risk. First, given the complex haplotype structure of the genome, the genetic association detected from GWAS represent the linkage disequilibrium between a tag SNP and a causal variant, and the SNP with the strongest association in a given locus is not necessarily the most probable causal SNP. Thus, GWAS signals frequently do not identify the functional variants, neither the biological mechanisms underlying complex diseases. In addition, most of the variants identified by GWAS reside in intergenic and intronic regions, not impacting protein-coding regions. This indicates that risk-associated variants most likely have a cis-regulatory function, as further shown by subsequent functional studies searching for causal variants.

Preceding work carried out by Maia and collaborators conducted a genome-wide DAE survey in 64 samples of normal breast tissue of healthy women, using microarrays to genotype over 500K SNPs. According to that preliminary data, most of autosomal genes are affected by cis-regulatory variants. Approximately 44K of the 91,686 transcribed SNPs presented DAE in at least 10% of the heterozygous individuals and in a minimum of three. That suggests that almost 50% of SNPs show DAE and 85.6% of the genes expressed in breast tissue display cis-regulatory variants.

In this context, we hypothesise that cis-regulatory variants are major contributors to breast cancer susceptibility by altering gene expression.

Chapter II

Main Goal and Specific Aims

2. Main Goal and Specific Aims

Breast cancer is the most common cancer among women worldwide and it has a strong genetic risk component. Thus, knowledge of individual familial risk can improve early detection and prevention of the disease. However, although GWAS have identified over 170 loci associated with overall breast cancer risk ($P<5\times10^{-08}$) so far, approximately half of the breast cancer heritability remains unknown. In addition, GWAS variants tag a genomic region and are not necessarily the causal variants. Furthermore, the majority of the GWAS variants lie outside genes indicating that cis-regulation is the most likely mechanism contributing for risk. As the effect of cis-regulatory variation can be complex, with several variants acting via different mechanisms on a given target gene, understanding the underlying mechanism of each variant is challenging. Therefore, the main goal of this project was to develop a new and efficient approach to detect target genes and causal variants in known and new breast cancer predisposition loci.

To achieve the main goal, several tasks were conducted according to the following specific aims:

1) Identify causal variants with strong cis-regulatory potential and target genes in candidate loci:

A previous genome-wide study of DAE was performed with microarray genotyping and expression data (Affymetrix SNP6.0 and Exon510S Illumina platforms) of 64 normal breast tissue samples from healthy women. These DAE data provided a cis-regulation map on breast tissue and was crossed with data from published and unpublished GWAS, to identify candidate loci with both strong cis-regulatory potential and association with risk. Selection criteria for candidate loci for this aim included: transcribed SNPs (tSNPs) showing DAE (daeSNPs) and GWAS SNPs in strong linkage disequilibrium ($r^2\geq0.8$, $D'\approx1$), and located within ± 500kb of each other; and candidate genes with at least one daeSNP displaying preferential expression of the same allele in all heterozygotes tested, indicative of daeSNP and regulatory SNP (rSNP) in complete LD in the analysed data. Next, functional annotation from public databases was used to search for candidate cis-regulatory variants driving BC genetic susceptibility.

2) Develop a new approach to validate cis-regulated genes associated with known BC risk in breast tissue and blood samples:

Using allelic expression (AE) ratios as a quantitative variable in case-control association studies has increased statistical power compared to discrete variable analysis (like genotype) to find differences between patients and controls, and identify significant AE associations with risk. Based on previous data for AE distributions observed in controls, statistical power computation indicated that around 20 cases and 20 controls would suffice to identify significant (P=1.0×10^{-3}) differences between AE of cases and controls with 95% of power. Additionally, as AE ratios reflect transcript levels regulation, this approach has the advantage of simultaneously identifying the target gene and the biology underlying susceptibility (cis-regulation) in each locus. Therefore, this specific aim intended to assess the use of AE in case-control studies to detect the association of cis-regulation and risk to breast cancer, plus confirm or propose candidate target genes in these loci, as well as candidate causal variants. For this purpose, daeSNPs from the top candidate risk-associated loci from the Aim 1 were used.

3) Identification of a suitable pipeline for AE analysis from RNA-seq data:

Technological progress in high-throughput sequencing technologies, including RNA-sequencing (RNA-seq), facilitates AE quantification by directly assessing single nucleotide variant counts from both alleles in heterozygous individuals. As an accurate measurement of AE requires an optimally suited RNA-seq data analysis, and no single pipeline can be applied to all cases, for this specific aim a comprehensive comparison of variant calling pipelines for precise AE quantification using RNA-seq data was conducted.

4) Development of a novel genome-wide approach to discover DAE loci associated with breast cancer risk:

The identification of genomic variants through RNA-seq data analysis detects functionally important variants. Hence, using the new approach developed in Aim 2 with whole-transcriptome RNA-seq data, and using the pipeline for AE analysis from Aim 3, the last purpose of this work was to perform a genome-wide case-control study using AE ratios as a quantitative trait to identify new candidate DAE loci associated with breast risk.

Chapter III

General Materials and Methods

3.1. Breast Tissue Samples

Normal breast tissue from healthy controls was collected from reduction mammoplasties for reasons not related to cancer, and were pathologically analysed to ensure the absence of disease in the tissue. Normal breast tissue from breast cancer patients (normal-matched samples) were obtained from normal adjacent biopsies taken during

All samples were collected in the scope of other studies (Curtis, et al., 2012; Maia, et al., 2009). from the Tissue Bank at Addenbrooke's Hospital (Cambridge, UK) with approval from the Addenbrooke's Hospital Local Research Ethics Committee (REC reference 06/Q0108/221 and 07/H0308/161), and with free and fully informed consent of the donors.

Breast tissue samples were used for the identification of target genes and causal variants in known breast cancer risk loci, and target genes in new BC risk loci. Control normal breast tissue samples were also used for the previous DAE study.

3.2. Blood Samples

Human B cells (purified from white cell-reduction filters) from anonymous blood donors were collected from the Blood Centre at Addenbrooke's Hospital collected with approval from the Addenbrooke's Hospital Local Research Ethics Committee (REC reference 04/Q0108/21). Mononuclear cells were separated by density gradient centrifugation using Lymphoprep (Sigma, USA), according to the manufacturer's instructions. B cells were isolated from these samples by magnetic sorting using CD19 labelled magnetic check beads (Milteny Biotech, Germany) (Maia, et al., 2009).

Blood samples from breast cancer cases were obtained from patients drawn from Studies of Epidemiology and Risk factors in Cancer Heredity (SEARCH), a population-based study, with women diagnosed with invasive breast cancer in the UK selected from the Eastern Cancer Registration and Information Centre (ECRIC, formerly East Anglian Cancer Registry). All participants in the study provided informed consent, and the study was approved by the Eastern Multicentre Research Ethics Committee (Azzato, et al., 2010).

Lymphoblastoid cell lines derived from unrelated CEPH individuals were obtained from the Coriell Cell Repository. Cell lines were grown in RPMI 1640 with 10% FCS, supplemented with

penicillin, streptomycin and L-glutamine, at 37°C and 5% CO_2 (Invitrogen, USA) (Maia, et al., 2009).

Blood samples were used for the identification of target genes and causal variants in known breast cancer risk loci, and target genes in new BC risk loci. Control normal breast tissue samples were also used for the previous DAE study.

3.3. Nucleic Acid Isolation and Quality Assessment

DNA from normal breast tissue, human B cells and lymphoblastoid cell lines was extracted using a conventional SDS/proteinase K/phenol method. For those samples, total RNA was extracted using Qiazol (Invitrogen, USA) following manufacturer's instructions. The RNA was subsequently treated with DNase I, and repurified using acidic phenol-chloroform and ethanol precipitation (Maia, et al., 2009).

For RNA extraction from normal breast tissue, samples were soaked overnight in RNAlater-Ice® (Ambion, USA), and extraction was performed using QIAzol® (Invitrogen, USA) and the Precellys® 24 bead mechanism (Bertin Technologies, France) following the manufacturer's instructions. An additional centrifugation step prior to addition of chloroform to the lysate was performed to eliminate excessive fat (Maia, et al., 2009).

Total DNA and RNA were extracted from fresh frozen normal-matched breast samples from patients using the DNeasy Blood and Tissue Kit and the miRNeasy Kit (Qiagen, Crawley, UK) on the QIAcube (Qiagen) according to the manufacturer's instructions (Curtis, et al., 2012).

DNA and total RNA were extracted from SEARCH samples in the scope of a study from 2010 (Azzato, et al., 2010).

Nucleic acids were quantified with with Nanodrop 2000™ Spectophotrometer (ThermoFischer Scientific, USA). RNA quality was also assessed using the Agilent 2100 Bioanalyser Nanochip (Agilent Technologies, UK).

3.4. Sex Determination

Sex determination was carried out for blood samples from controls through PCR using two primers with identical sequences on the X and Y chromosomes that span the first intron of the amelogenin gene: AMELF – 5´-CTCTGATGGTTGGCCTCAAG-3'; AMELR – 5'-ACCTTGCTCATATTATACTTGACAAA-3' (Bailey, Affara, & Ferguson-Smith, 1992). Amplified fragments have different sizes due to difference in the length of introns on the X and Y chromosomes, which originates fragments of 542 bp and 358 bp, respectively. PCR amplifications were performed in a C1000 Touch™ Thermal Cycler (Bio-Rad Laboratories, Inc., USA), in a final volume of 10 μl with HotStart Taq *Plus* Master Mix Kit (Qiagen, Germany), 200 nM of each primer and 200 ng of DNA. PCR reactions included an initial enzyme activation step at 95 °C for 5 min, followed by 35 cycles of denaturation at 94 °C for 30 sec, annealing at 60 °C for 30 sec and extension at 72 °C for 1 min. After a final extension step at 72 °C for 10 min, reactions were stopped at 4 °C.

After electrophoresis in 1% agarose gel in 1X TAE buffer (pH 8.3) stained with Greensafe (NZYTech, Portugal), the PCR products were analysed under UV light with a ChemiDoc™ XRS+ Imaging System and using the ImageLab™ Software (Bio-Rad Laboratories, Inc., USA). The NZYDNA Ladder V (NZYTech, Portugal) was used as size marker.

3.5. Genotyping

Genomic DNA of breast and blood samples was genotyped for nine SNPs, with predesigned or Custom TaqMan® SNP Genotyping Assays (Applied Biosystems, USA).

These assays use TaqMan® 5´-nuclease chemistry for amplifying and detecting specific polymorphisms in purified genomic DNA samples. Each TaqMan® SNP Genotyping Assay contains sequence-specific forward and reverse primers to amplify the polymorphic sequence of interest, and two probes labelled with different fluorochromes: one VIC™-labelled probe to detect one allele sequence, and one FAM™-labelled probe to detect the other allele sequence. Each probe and primer set uniquely align with the genome, providing specificity for the allele of interest. The high binding stability of each probe-template complex allows the use of probes as short as 13 bases for allelic discrimination.

Genotyping experiments were performed in 384-well plates in a Bio-Rad CFX384 Real-Time System C1000 Touch™ Thermal Cycler (Bio-Rad Laboratories, Inc., USA). Reactions were prepared in a final volume of 5 μl with KAPA PROBE FAST qPCR Master Mix (2X) (KAPA Biosystems, USA) or TaqMan® Universal Master Mix II, with UNG (2X) (Applied Biosystems, USA), TaqMan® SNP Genotyping Assay (40X) (Applied Biosystems, USA), DNase/RNase-free water and 8 ng of DNA. Duplicates of randomly selected samples and at least two no-template controls (NTC) were included. Real-time PCR reactions included an enzyme activation step at 95 °C for 3 min, followed by 40 cycles of denaturation at 95 °C for 3 sec and annealing/extension 60 °C for 30 sec. Reactions with TaqMan® Universal Master Mix II, included an initial step for UNG activation at 50 °C for 2 min, followed by enzyme activation at 95 °C for 10 min, and 40 cycles of denaturation at 95 °C for 15 sec and annealing/extension 60 °C for 1 min. Results were analysed with the Bio-Rad CFX Manager 3.1 Software (Bio-Rad Laboratories, Inc., USA).

3.6. Quantification of Allelic Gene Expression

3.6.1. cDNA Synthesis and Preamplification

After genotyping, cDNA from heterozygotes for each SNP was synthesised with SuperScript® First-Strand Synthesis System (Invitrogen, USA). cDNA synthesis was performed in a final volume of 10 μl of a mixture containing 50 ng of total RNA, 1.25 μM of Oligo$(dT)_{20}$, 1.25 μM of random hexamers, 10 mM dNTP mix, 1X SSIV buffer [50 mM Tris-HCl (pH 8.3), 4 mM $MgCl^2$, 10 mM DTT, 50 mM KCl], 5 mM DTT, 20 units of RNaseOUT™ Recombinant RNase Inhibitor, 100 units of SuperScript® IV Reverse Transcriptase and nuclease-free water. Reactions with total RNA and primers were initially incubated at 65 °C for 5 min and then placed on ice for at least 1 min. Next, the remaining components were added and reverse transcription was conducted at 23 °C for 10 min, followed by incubation at 50 °C during 5 min. Reactions were inactivated at 72 °C for 15 sec. RNA was removed by addition of 1 unit of *E. coli* RNase H and incubation at 37 °C for 5 min.

Next, cDNA was preamplified in order to reduce the required initial cDNA amount prior to multi-target qPCR analysis. Target-specific preamplification was done with TaqMan® PreAmp Master Mix (2X) (Applied Biosystems, USA), pooled TaqMan® SNP Genotyping Assay (0.2X) (Applied Biosystems, USA) and 1.25 μl of cDNA. Thermal cycling conditions consisted of enzyme activation at 95 °C for 10 min, followed by 8 or 14 cycles of denaturation at 95 °C for 15 sec and

annealing/extension at 60 °C for 4 min. Finally, reactions were diluted 1:5 adding DNase/RNase-free water for a total volume of 25 μl.

3.6.2. Real-Time PCR

Preamplified cDNA or 10X dilutions of stock cDNA were used to perform case-control analysis by real-time PCR in 384-well plates in a Bio-Rad CFX384 Real-Time System C1000 Touch™ Thermal Cycler (Bio-Rad Laboratories, Inc., USA) or in BioMark™ HD 96.96 chips in a BioMark™ HD system (Fluidigm, USA).

Standard curves to access the efficiency of amplification for both alleles were prepared using serial dilutions with 100, 50, 25, 12.5, 2.5 and 1.25 ng of DNA from CEPH lymphoblastoid cell lines heterozygous for each SNP.

For reactions using Bio-Rad CFX384 Real-Time System C1000 Touch™ Thermal Cycler (Bio-Rad Laboratories, Inc., USA) the same protocol as for genotyping was used. Triplicates of each cDNA sample and at least 2 no-template controls (NTC) were included.

For reactions conducted in the BioMark™ HD system (Fluidigm, USA), six replicates of each cDNA sample and at least 2 no-template controls (NTC) were included in each experiment. For these, 10X Assays prepared with TaqMan® SNP Genotyping Assay (40X) (Applied Biosystems, USA) and Assay Loading Reagent (Fluidigm, USA). Sample pre-mix was prepared with TaqMan® Universal Master Mix II, with UNG (2X) (Applied Biosystems, USA) and GE Loading Reagent (20X) (Fluidigm, USA). Each cDNA sample (2.7 μl) was then added to an aliquot of pre-mix (3.3 μl). The assays and mixtures with pre-mix and samples were then loaded into the chips.

All experiments were repeated at least twice.

3.7. Statistical Analysis and AE Quantification in Case-Control Studies

Data obtained by real-time PCR were analysed on Microsoft® Excel® software. For each heterozygous sample for each SNP, the mean of Ct values from replicates was calculated for the reference allele (labelled with one fluorochrome) and the alternative allele (labelled with another fluorochrome). The percentage of variation between replicates was calculated as the ratio of the standard deviation (SD) by the triplicates mean Ct for each sample

(%var=[SD/Mean]). Triplicates with 10% or more variation were excluded from further analysis. The $\log_2$ allelic expression ratios were determined by calculating the delta Ct, corresponding to the ratio of expression of the minor allele by the reference allele.

Allelic expression ratios associated with breast cancer risk were found using the magnitude and direction of the differences between AE distributions in cases and controls (effect size), measured using the Hedges' *g* test. This statistic is suitable for small sample sizes (<20), which was the case of the sample group in this work. Hedges' *g* is the mean difference between two groups divided by the pooled standard deviation of both groups, taking into account a correction factor for small sample sizes (F):

$$Hedges'g = \left(\frac{\bar{X}_{Cases} - \bar{X}_{Controls}}{StDev}\right) \times F,$$

where the pooled SD is:

$$StDev = \sqrt{\frac{(n_{Cases} - 1)Var_{Cases} + (n_{Controls} - 1)Var_{Controls}}{n_{Cases} + n_{Controls} - 2}}$$

F employs the gamma function (Γ):

$$F = \frac{\Gamma\left(\frac{e}{2}\right)}{\sqrt{\frac{e}{2}} \times \Gamma\left(\frac{e-1}{2}\right)}$$

and *e* refers to the degrees of freedom:

$$e = n_{Cases} + n_{Controls} - 2$$

All estimation plots, and statistical tests were performed using the online tool available at www.estimationstats.com (Ho, Tumkaya, Aryal, Choi, & Claridge-Chang, 2019). The Hedges' *g* effect size is represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI).

Chapter IV

Identification of Target Genes and Causal Variants in Known Breast Cancer Risk Loci

4.1. Abstract

Breast cancer risk-associated variants identified by genome-wide association studies (GWAS) are mostly located in noncoding regions of the genome. This, together with the complex haplotype structure of the genome represents a challenge for the identification of the causal variants and understanding of the risk biological mechanisms. Moreover, it indicates that cis-regulatory variation is highly likely to play a relevant role in breast cancer susceptibility. To address these challenges, this work set out to develop a new approach to validate cis-regulated genes associated with known breast cancer risk in breast tissue and blood samples. To achieve this aim, allelic expression (AE) ratios were used as a quantitative variable in case-control association studies to identify cis-regulatory variants and their target genes.

Data from previous exploratory DAE analysis performed with breast tissue samples from healthy people showed that in 12 of the GWAS loci associated with breast cancer susceptibility, the transcribed SNPs showing DAE (daeSNPs) displayed preferential expression of the same allele in all heterozygotes tested. This was suggestive of complete LD between the daeSNPs and the regulatory SNPs (rSNPs) in the analysed data. In-silico functional analysis identified candidate SNPs with high regulatory potential in mammary tissue in three of these loci: 1q32.1, 16q23.2, and 17q22.

Further in-vitro analysis for the 17q22 locus identified two putative regulatory variants – rs17817901 and rs8066588 – that alter a miRNA and a transcription factor binding sites, respectively. Results from this work suggest that *STXBP4* and *COX11* are the most likely target genes for risk-associated variation in this locus. A significant association with risk for breast cancer was found for the preferential expression of the reference alleles of both rs17817901 (*TOM1L1/COX11*) and rs2628315 (*STXBP4*) in normal breast tissue. Interestingly, this association was also observed in blood samples, showing the potential impact of this approach in the screening of the general population for breast cancer risk. This work shows that integrating AE ratios as a quantitative variable in case-control association studies, is a powerful approach to identify novel risk loci.

Finally, in the scope of another study on allelic expression in *PIK3CA* in breast cancer, this work also validated the SNP rs2699887 as a candidate regulatory variant acting on *PIK3CA* by altering binding of the transcription factor NF-YA.

4.2. Introduction

Currently, the hundreds of common low-risk variants identified by genome-wide association studies (GWAS) and the Collaborative Oncological Gene-Environment Study (COGS) represent approximately 18% of the familial breast cancer risk (Michailidou, et al., 2017; Fachal, et al., 2020), taking the current total to about 50%. Conventional GWAS use single nucleotide polymorphisms (SNPs) to tag genetic variation in linkage disequilibrium (LD) blocks, and each block can harbour thousands of SNPs. Given the number of SNPs in LD, the identification of causal variants targeting genes in those loci, and the understanding of the functional mechanisms through which these variants influence susceptibility, remains a challenge. So far, published GWAS identified over 170 loci associated with BC susceptibility ($P<5\times10^{-08}$) (Easton, et al., 2007; Hunter, et al., 2007; Pharoah, et al., 2007; Stacey, et al., 2007; Stacey, et al., 2008; Ahmed, et al., 2009; Milne, et al., 2009; Thomas, et al., 2009; Zheng, et al., 2009; Antoniou, et al., 2010; Turnbull, et al., 2010; Fletcher, et al., 2011; Haiman, et al., 2011; Ghoussaini, et al., 2012; Siddiq, et al., 2012; Michailidou, et al., 2013; Michailidou, et al., 2015; Amos, et al., 2017; Milne, et al., 2017; Lilyquist, Ruddy, Vachon, & Couch, 2018), with the majority of GWAS variants located in intergenic and intronic regions of the genome.

Several studies have successfully integrated trait-associated variants at GWAS loci with the publicly available regulatory element datasets in disease-relevant cell types to guide identification of regulatory variants underlying disease susceptibility. Functional analyses of small proportion of GWAS variants have shown these to be cis-regulatory, providing evidence of their importance in breast cancer risk. However, the effect of such variation on each gene is often complex, with several variants acting in varying degrees via different mechanisms.

Identification of the target genes of cis-regulatory variants usually includes physical interaction studies (e.g., chromatin conformation capture, Baxter, et al., 2018), in-vitro assays to assess variation of protein binding at regulatory elements (e.g., band shifts and transfection assays, Meyer, et al., 2008), and more integrative approaches (Fachal, et al., 2020). However, these studies often lack direct validation of the regulation in-vivo in the complex human genomic context. As these variants regulate gene expression in an allele-specific manner, they can be detected by the imbalances they produce in the expression of the two alleles of autosomal genes in heterozygous individuals at transcribed SNPs (tSNPs), which can be measured as differential allelic expression (DAE) ratios. Analysis of DAE as a quantitative trait has increased statistical power compared to discrete variable analysis, like genotypes used by GWAS. Hence, DAE

analysis can be used to identify cis-regulatory variants, without the examination of large numbers of heterozygotes. Previous DAE analysis (Xavier, et al., 2016) with microarray genotyping and expression data of 64 normal breast tissue samples from healthy women, provided a whole-genome map of a cis-regulatory variation on breast tissue, which crossed with data from published and unpublished GWAS data, identified loci with both strong evidence of cis-regulation and association with risk. In 12 of these loci at least one variant was found to display preferential expression of the same allele in all heterozygotes tested. This indicates that the cis-regulatory variant acting on those genes is in strong to perfect LD with the tSNP where DAE was identified (daeSNP), which makes fine-mapping in those loci unnecessary and simultaneously makes them more amenable for functional characterization.

The initial purpose of this study was to identify target genes and causal variants in the aforementioned 12 risk-associated loci with strong cis-regulatory potential. Furthermore, using the top locus from the preceding analysis, 17q22, the present study intended to validate a new powerful and efficient approach to identify cis-regulated genes and causal variants associated with BC risk.

4.3. Materials and Methods

4.3.1. Identification of Candidate Cis-Regulated Genes in Known BC Risk Loci

Exploratory AE analysis was performed previously (Xavier, et al., 2016), with mRNA and DNA from 64 normal breast tissue samples, using Illumina Exon510S-Duo SNP microarrays. The sample filtering and normalization was performed as described formerly (Liu, et al., 2012) and 12 samples were removed from further analysis. After normalization and prior to AE analysis, extensive quality control for SNP expression and genotyping was performed as follows: 1) to avoid false positives due to low expression, SNPs with average $\log_2$ RNA intensity values <9.5 were discarded; 2) to verify allelic discrimination at RNA level, a two sample t-test was applied to compare RNA log ratios between heterozygous (AB) and homozygous groups (AA and BB), and only SNPs with $P \leq 0.05$ for all comparisons were further analysed; 3) to guarantee high quality genotyping data, SNPs with call rate <90%, Hardy-Weinberg equilibrium $P \leq 1.0 \times 10^{-05}$ and less than five heterozygotes were excluded; 4) SNPs mapping to multiple locations in the genome, flagged as suspected in dbSNP149 GRCh38p7 and located in sexual chromosomes were also discarded.

AE was measured in the filtered dataset of SNPs and samples, in a variable and an independent number of individuals who are heterozygous for a transcribed SNP (tSNP) with alleles A and B. DAE was defined as the logarithm of base two of the ratio between the levels of allele A transcript and the levels of allele B transcript (heterozygote ratio), normalized by the heterozygote ratio in genomic DNA (gDNA), on tSNP heterozygotes. DAE was defined as AE ratios greater than 0.58 or less than -0.58, and tSNPs were considered to have DAE (daeSNPs) when at least 10% of the heterozygotes and a minimum of three displayed DAE.

This AE analysis provided a whole-genome map of a cis-regulatory variation on breast tissue, which crossed with data from published and unpublished GWAS data, identified loci with both strong evidence of cis-regulation and association with risk

4.3.1.1. Proxies Identification

Proxy variants were identified for the transcribed SNPs (tSNPs) showing DAE (daeSNPs) at the selected breast cancer risk-associated loci. This was carried out with the tools SNP Annotation and Proxy Search (SNAP v2.2) (Johnson, et al., 2008) and HaploReg v4.1 (Ward & Kellis, 2012)

using a linkage disequilibrium (LD) threshold of $r^2 \geq 0.8$ for information from the CEU population. SNAP used pre-computed LD for all pairs of SNPs in Pilot 1 of the 1000 Genomes Project within 500 kb. HaploReg used pairwise LD computed for all SNPs within 250 kb in Phase 1 of the 1000 Genomes Project.

4.3.1.2. In-silico Regulatory Characterisation of Candidate Variants

Firstly, functional annotation of regulatory potential in normal breast cell lines (myoepithelial, epithelial, and fibroblasts) of daeSNPs and their proxies were assessed using HaploReg and RegulomeDB v1.1 (Boyle, et al., 2012) tools, and loci were visualised in the WashU Epigenome Browser from the Roadmap Epigenomics Project (Roadmap Epigenomics Consortium, 2015) and with the UCSC Genome Browser (Kent, et al., 2002).

Functional annotations used by HaploReg and RegulomeDB include chromatin marks from ChIP-seq data, chromatin state, accessibility (DNase-seq and FAIRE-seq data), and protein binding annotation from the Roadmap Epigenomics and Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium, 2011) projects, and the effect of SNPs on expression from eQTL studies from the Genotype-Tissue Expression (GTEx) project (Carithers & Moore, 2015). Annotation of the effect of SNPs on regulatory motifs provided by both tools is based on a library of Position Weight Matrixes (PWMs) to score each allele of a variant for its potential of binding to a transcription factor (TF). HaploReg also provides specific epigenomic information on marks from H3K4me1 and H3K27ac (used as enhancers), and from H3K4me3 and H3K9ac (used as promoters).

WashU Epigenome Browser and UCSC Genome Browser are online resources that include many assemblies and annotations of vertebrates and model organisms, to visualise, analyse and download data, such as cell type, assay type, epigenetic marks, and phenotype.

Identification of potential regulatory variants was first focused on the altered binding of transcription factors. Thus, priority was given to variants with evidence of being localized in regions of DNase I hypersensitive sites (DHS) and cis-regulatory elements. Variants were further considered based on the PWM scores and ChIP-seq data indicating a potential differential binding of TFs to the DNA sequences, including the reference and alternative alleles, in breast cell lines.

Afterwards, variants were prioritised according to computational allelic microRNA (miRNA) target predictions, previously conducted in our research group (Jacinta-Fernandes, Xavier, Magno, Esteves, & Maia, 2018). These predictions involved the adaptation of the TragetScan (Lewis, Shih, Jones-Rhoades, Bartel, & Burge, 2003) algorithm to perform allele-specific predictions of differential binding of miRNAs. Potential interactions between miRNAs and potential rSNPs in one of the loci under study were identified and selected for functional in-vitro analysis.

4.3.2. In-vitro and In-vivo Functional Analysis

4.3.2.1. Cell Lines

Human immortalized breast cancer cell lines MCF7 and HCC1954 were used to perform in-vitro functional analysis of the candidate regulatory variants by electrophoretic mobility shift assays (EMSAs). MCF7 was established in 1973 at the Michigan Cancer Foundation from the adenocarcinoma of a 69 years old Caucasian woman (Soule, Vazguez, Long, Albert, & Brennan, 1973). This cell line has characteristics specific to the mammary epithelium. MCF7 cells can process oestrogen, in the form of oestradiol, via oestrogen receptors in the cell cytoplasm. This makes the MCF7 cell line an oestrogen receptor (ER) positive control cell line. This cell line also expresses the progesterone receptor (PR) but not the human epidermal growth factor receptor 2 (HER2). The HCC1954 cell line was derived from a primary stage IIA, grade 3, invasive ductal carcinoma, with no lymph node metastases, from a 61 years old East Indian woman (Gazdar, et al., 1998). HCC1954 cells do not express ER or PR but overexpress HER2 protein.

Breast cancer cell lines T47D and SUM159 were used for in-vivo functional analysis of candidate regulatory variants by chromatin immunoprecipitation. T47D was isolated from the pleural effusion of a 54 years old woman with infiltrating ductal carcinoma (Keydar, et al., 1979). This cell line expresses ER, PR, but not HER2. The SUM159 cell line was isolated from a woman's anaplastic breast carcinoma (Forozan, et al., 1999). These cells do not express ER, PR or, HER2, which makes them triple negative (TN).

4.3.2.2. Cell Culture

MCF7 and T47D cells were cultured in Dulbecco's Modified Eagle Medium (DMEM), high glucose, pyruvate (Gibco™, Thermo Fisher Scientific Inc., USA), HCC1954 cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 Medium (Gibco™, Thermo Fisher Scientific Inc., USA), and SUM159 cells were cultured in Ham's F-12 medium (Gibco™, Thermo Fisher Scientific Inc., USA). DMEM and RPMI media were supplemented with 10% (v/v) foetal bovine serum (FBS, Gibco™, Thermo Fisher Scientific Inc., USA), and F-12 with 5% of FBS. All media were also supplemented with 100 U/ml penicillin, 100 μg/ml streptomycin, and 2 mM L-glutamine (all Gibco™, Thermo Fisher Scientific Inc., USA). To F-12 medium 0.05% of insulin (Sigma, USA) and 0.1% of hydrocortisone (1 μg/mL) (Sigma, USA) were also added. All cell lines were maintained at 37 °C in a humidified and 5% concentrated CO_2 atmosphere. The medium was renewed at least every two days and once cultures reached about 70% confluence cells were subcultured at a ratio of 1:2 or 1:3. For subculturing, the culture medium was removed and discarded, and the cell layers were washed with phosphate-buffered saline (PBS), pH 7.4 (Gibco™, Thermo Fisher Scientific Inc., USA). Cells were dissociated using 0.25% (w/v) trypsin-EDTA (Gibco™, Thermo Fisher Scientific Inc., USA), and trypsin was then inactivated with a complete growth medium. The cell suspension was centrifuged at 125 ×g for 5 min and the supernatant was discarded. The cell pellet was resuspended in a fresh complete growth medium, and aliquots of the cell suspension were added to new culture vessels.

4.3.2.3. Nuclear Extract Preparation

Nuclear protein extracts were prepared from MCF7 and HCC1954 cells using the NE-PER nuclear and cytoplasmic extraction kit (Thermo Fisher Scientific Inc., USA), according to the manufacturer's instructions.

Protein concentration from nuclear extracts was quantified with the Invitrogen™ Qubit™ Protein Assay Kit (Thermo Fisher Scientific Inc., USA), in an Invitrogen™ Qubit™ 2.0 Fluorometer (Thermo Fisher Scientific Inc., USA), following the manufacturer's instructions.

4.3.2.4. Synthetic Oligonucleotides

For further use in non-isotopic electrophoretic mobility shift assays, oligonucleotides were synthesised targeting the sequences including the reference and the alternative alleles of each candidate regulatory variant (Table 4.1). Single-stranded DNA was labelled with biotin using the Pierce™ Biotin 3' End DNA Labelling Kit (Thermo Fisher Scientific Inc., USA), and labelling efficiency was determined by dot blot (hand spotting) using the Chemiluminescent Nucleic Acid Detection Module (Thermo Fisher Scientific Inc., USA), following the manufacturer's procedures.

Table 4.1. Oligonucleotides used as labelled probes or unlabelled competitors in the EMSAs. The two alleles [ref/alt] of each SNP are indicated.

Gene	SNP	Sense strand (5' – 3') Antisense strand (5' – 3')
PIK3C2B	rs6692377	GCACTGGGCTCGGCGGGGCC**[G/A]**GGCTAACCCATCCAGGTCTG CAGACCTGGATGGGTTAGCC**[C/T]**GGCCCCGCCGAGCCCAGTGC
MDM4	rs3789052	GCATCTAGAGAGTGCTCTAC**[C/T]**TGATAACCATTGGGGAGATA TATCTCCCCAATGGTTATCA**[G/A]**GTAGAGCACTCTCTAGATGC
LOC105371692 (Intergenic *MDM4-LRRN2*)	rs12730457	AGGAGGAGGCTCCAGCCCAG**[C/A]**CTGGATCAGGGGAGGCTCCC GGGAGCCTCCCCTGATCCAG**[G/T]**CTGGGCTGGAGCCTCCTCCT
COX11	rs8066588	CATTTGCTCAAAACC**[C/T]**ACCTGTGATTTTCTTCC GGAAGAAAATCACAGGT**[G/A]**GGTTTTGAGCAAATG
	rs12945393	CAATTTCCAGAGTTACCCTT**[G/A]**GCACATAAACCACATGTGGT ACCACATGTGGTTTATGTGC**[C/T]**AAGGGTAACTCTGGAAATTG
STXBP4	rs9896044	GCATACTTACTTGTTTATTT**[C/G]**CTATTACAAGAACACTTTCT AGAAAGTGTTCTTGTAATAG**[G/C]**AAATAAACAAGTAAGTATGC
	rs9891865	ATTACTGGAATCCCTGTTGG**[C/T]**TGAGGGCCTTGCAAGTTTGT ACAAACTTGCAAGGCCCTCA**[G/A]**CCAACAGGGATTCCAGTAAT
	rs2628305	TTCTCCACCATTAAACT**[A/C]**TGACCAAGAAACAGCAG CTGCTGTTTCTTGGTCA**[T/G]**AGTTTAATGGTGGAGAA
	rs244353	AGGTTTCGATGATGACC**[G/A]**AAGGTCTTAGGCAACCA TGGTTGCCTAAGACCTT**[C/T]**GGTCATCATCGAAACCT
C16orf46	rs74878296	GTTGGGTTAGAGACTTGTGA**[C/G]**TAGCCTGGTCAACAGACTGT ACAGTCTGTTGACCAGGCTA**[G/C]**TCACAAGTCTCTAACCCAAC
PIK3CA	rs2699887	AGCGTGAGTAGAGCGCGGA**[C/T]**TGGCCGGTAGCGGGTGCGGTG CACCGCACCCGCTACCGGCC**[G/A]**ATCCGCGCTCTACTCACGCT

This nucleic acid labelling method uses the catalytic activity of the terminal deoxynucleotidyl transferase (TdT) to incorporate 1-3 biotinylated ribonucleotides (Biotin-11-UTP) onto the 3'-OH end of DNA (more efficiently in single-stranded DNA), which reduces the interference with hybridization or sequence-specific binding of proteins.

To prepare double-stranded probes, hybridization of complementary oligonucleotides was performed adding equal volumes of 100 nM labelled probes or 10 μM unlabelled oligonucleotides, followed by incubation at 80 °C for 10 min and overnight incubation at room temperature.

4.3.2.5. Electrophoretic Mobility Shift Assays

The interaction of proteins, such as TFs, and DNA in nuclear extracts were analysed by electrophoretic mobility shift assays (EMSAs). These assays allow the analysis of protein-DNA interactions, based on the different mobility that free DNA and DNA-protein complexes have (Hellman & Fried, 2007). Labelled linear DNA fragments containing specific binding sites are incubated with nuclear extracts containing DNA-binding proteins. The DNA-protein complexes are separated from the free DNA by nondenaturing polyacrylamide gel electrophoresis. Since the DNA-protein complexes have lower mobility than the unbound DNA, which migrates faster through the gel, the migration of labelled complexes is shifted and then detected by chemiluminescence. Additionally, the binding specificity of a protein to the target sequence can be evaluated with competition assays with unlabelled specific competitors, or protein specific antibodies. If the unlabelled DNA competes with the labelled DNA for the binding of the proteins, the labelled DNA-protein complexes will become a minority or will not form, and therefore will not be detected in the chemiluminescence assay. When using specific antibodies, if the prediction of the binding TF is correct, an even heavier, slower migrating- complex is formed, composed of antibody-protein-DNA resulting in a supershift on the gel.

All binding reactions for EMSAs were performed in a final volume of 20 μL. After optimization, 5 μg of nuclear extract was incubated with binding buffer [20 mM Hepes, pH 7.4, 5% glycerol, 10 ng/μL Poly (dI•dC), 1X protease inhibitor, 1 mM dithiothreitol (DTT)] at room temperature for 10 to 12 min. Next, 60 fmol of labelled probe were added to the binding mix and incubated for 15 min at room temperature. For competition reactions, 1-, 10-, 33-, or 100-fold excess of unlabelled oligonucleotides (Tables 4.1 and 4.2) were added in the initial incubation of nuclear

extract and binding buffer. For supershift experiments, 1 μg of specific antibody was added with the labelled oligonucleotides.

DNA-protein complexes and free probes were separated by nondenaturing polyacrylamide gel electrophoresis using 6% acrylamide/bisacrylamide (29:1) with 10% glycerol gels in 0.5X TBE (44.5 mM Tris, 44.5 mM boric acid, 1 mM EDTA, pH 8.3) for 10 min at 75 V, and 75 min at 120 V. Bound and free DNA were transferred to nylon membranes (Thermo Fisher Scientific Inc., USA), pre-equilibrated in 0.5X TBE for at least 10 min, with Trans-Blot Semi-Dry Transfer Cell (Bio-Rad Laboratories, Inc., USA), for 15 min at 290 mA. The membranes were UV crosslinked twice with 120 mJ/cm^2 in an UVP HL-HybriLinker™ Hybridization Oven/UV Crosslinker (Analytic Jena AG, Germany). Biotin-labelled complexes were detected by chemiluminescence using the Chemiluminescent Nucleic Acid Detection Module (Thermo Fisher Scientific Inc., USA), and visualised with a ChemiDoc™ XRS+ Imaging System and using the Image Lab™ Software (Bio-Rad Laboratories, Inc., USA). Analysis of DNA-protein complexes by EMSA was performed, for each candidate regulatory variant and cell line, at least twice.

Table 4.2. Consensus oligonucleotides used as unlabelled competitors in the EMSAs.

Gene	Sense strand (5' – 3') Antisense strand (5' – 3')
STAT3[1]	GATCCTTCTGGGAATTCCTAGATC GATCTAGGAATTCCCAGAAGGATC
GATA3[2]	CACTTGATAACAGAAAGTGATAACTCT AGAGTTATCACTTTCTGTTATCAAGTG
GATA3[3]	AATGTCCATCTGATAAGACG CGTCTTATCAGATGGACATT
E2F1[4]	ATTTAAGTTTCGCGCCCTTTCTCA TGAGAAAGGGCGCGAAACTTAAAT
NF-YA[5]	GACCGTACGTGATTGGTTAATCTCTT AAGAGATTAACCAATCACGTACGGTC

[1] sc-2571 (Santa Cruz Biotechnology, USA); [2] kindly provided by Dr. Chin from University of Cambridge; [3] Adomas, et al., 2014; [4] sc-2507 (Santa Cruz Biotechnology, USA); [5] Xu, et al., 2012.

4.3.2.6. Transfection and Luciferase Reporter Assays

To study potential gene regulation through the binding of miRNAs, reporter assays were conducted with the pmirGLO Dual-Luciferase miRNA Target Expression Vector (Promega, USA). This vector uses the dual-luciferase technology from Promega that includes two reporters: the firefly luciferase (*luc2*) which is the primary reporter that monitors mRNA regulation; and the *Renilla* luciferase (*hRluc-neo*) that acts as a control reporter for normalization and selection. The transcript stability and activity can be assessed by cloning only the miRNA binding sites or the entire 3' untranslated region (UTR) of the target gene at the 3' end of the firefly luciferase gene (*luc2*). Thus, miRNAs activity can be quantified through the expression of firefly luciferase. A reduction in its expression indicates the binding of endogenous or introduced miRNAs to the cloned miRNA target sequence.

Initially, sequences with 80 nucleotides harbouring 3 copies of the miRNA binding site with the reference allele or the alternative allele of the potential regulatory variant were synthesised and cloned into the pmirGLO Dual-Luciferase miRNA Target Expression Vector (Promega, USA). MCF7 cells were seeded in a 96-well plate (1.0×10^4 cells/well) and cultured for 24 h with an antibiotic-free complete medium. Next, transfections and co-transfections were performed with DharmaFECT DUO Reagent (Dharmacon, Inc., USA) following the manufacturer's instructions. Cells were transfected with empty vector, vector with the cloned constructs, and synthetic miRNAs mimics (miRIDIAN microRNA Human hsa-miR-194-5p - Mimic, C-300642-03-0002, and miRIDIAN microRNA Mimic Negative Control #1, CN-001000-01-05, Dharmacon). After 24h luciferase assays were performed with the Dual-GLO Luciferase Assay System (Promega, USA) following its protocol.

4.3.2.8. Chromatin Immunoprecipitation

Chromatin immunoprecipitation (ChIP) enables the analysis of proteins associated with specific genomic regions in living cells. Initially, the cells are fixed allowing protein-DNA interactions to be stabilised by a cross-linking agent, usually formaldehyde. Next, chromatin is extracted/recovered and sheared into small fragments that are subsequently precipitated with an antibody specific to the protein of interest (immunoprecipitation). If precipitated DNA is enriched in specific sequences, these are associated with the protein of interest in-vivo. Finally, the analysis of those genomic regions can be conducted by quantitative polymerase chain reaction (qPCR).

ChIP experiments were performed with the iDeal ChIP-seq kit for Transcription Factors (Diagenode, Belgium) according to the manufacturer's recommendations. Briefly, after fixation and lysis, chromatin was sheared for 12 cycles [30 seconds "ON", 30 seconds "OFF"] in the Bioruptor Pico (Diagenode, Belgium), preceded by chromatin shearing optimization. Magnetic immunoprecipitation was then conducted with the provided antibodies for positive (anti-CTCF, rabbit polyclonal antibody) and negative (rabbit IgG) controls, and the antibody against the protein of interest, NF-YA (sc-10779, Santa Cruz Biotechnology, USA). After elution, reverse cross-linking, and purification, the immunoprecipitated DNA was amplified using KAPA SYBR FAST Universal qPCR Kit (KAPA Biosystems, USA) with the provided positive (H19 imprinting control region) and negative (myoglobin exon 2) control primer pairs. To analyse the interaction between DNA and NF-YA by qPCR two primer pairs were used: one from another study (Shi, et al., 2015) that amplifies a 119 bp fragment of the *TOP2A* gene, and was used as a positive control for DNA - NF-YA interaction; and another designed to amplify a 114 bp region of the *PIK3CA* gene that includes the candidate regulatory variant (PIK3CA_Fw: 5'-GGACCCGATGCGGTTAGAG-3' and PIK3CA_Rev: 5'-GAGTCTCCGGCACCCACCC-3').

4.4. Results

4.4.1. Known breast cancer risk-associated loci 1q32.1, 16q23.2 and 17q22, have strong cis-regulatory potential

Integration of data from GWAS variants associated with breast cancer risk, with allelic expression data from breast tissue of healthy women, identified 32 loci associated with risk and with strong cis-regulatory potential. These loci included one or more GWAS variants in strong LD ($r^2 \geq 0.8$) with at least one tSNP with differential allelic expression (daeSNP). As previously mentioned, it is expected the cis-regulatory variants to be in moderate to strong LD with both the DAE and the GWAS SNPs. Moreover, daeSNPs with a specific pattern of allelic expression, where all individuals express preferentially the same allele, are likely to be in stronger LD with the regulatory variant(s) (rSNPs). From the 32 loci indicated above, 12 included this type of allelic expression pattern and, were thus considered potential candidates to harbour genetic variants responsible for cis-regulation in breast cancer and were functionally analysed in-silico.

In-silico functional analysis was performed for all 12 candidate loci (Figures 4.1, 4.5, 4.12D, and S4.1-S4.12), and identified three loci, 1q32.1, 16q23.2, and 17q22, with stronger cis-regulatory potential and association with breast cancer risk. Thus, these three loci were selected for analysis of functional annotations and prioritization of causal variants for experimental follow-up.

For those three loci, variants in high LD with the respective daeSNPs were identified and priority was given to those showing evidence of being simultaneously localized in DNase I hypersensitive sites and associated with histone marks found in active regulatory elements in mammary tissue.

4.4.2. Locus 1q32.1 has three candidate regulatory variants

The 1q32.1 *locus* spans a region including the genes *PIK3C2B* (phosphatidylinositol-4-Phosphate 3-kinase catalytic subunit type 2 beta), *MDM4* (mouse double minute 4, human homolog of, also known as *MDMX* and *HDMX*), and the intergenic region between *MDM4* and *LRRN2* (leucine-rich repeat neuronal 2).

The genetic variant rs11240762, one of the 12 daeSNPs located in the 1q32.1 locus (Figure 4.1) was found to be in moderate LD with the GWAS risk-associated variants rs2290854 (r^2=0.385), rs4245739 (r^2=0.331), and rs6682208 (r^2=0.357). The daeSNP rs11240762 showed a pattern of DAE in which all the individuals preferentially expressed the alternative C-allele (Figure 4.1B). Since this pattern is consistent with strong LD between the daeSNP and the rSNP, this locus was considered a potential candidate to harbour genetic variants responsible for cis-regulation in breast cancer, with the alternative C-allele of the rs11240762 being the one associated with decreased risk. Forty-five proxy SNPs in strong LD ($r^2 \geq 0.8$) with rs11240762 were identified. In-silico functional analysis showed that 28 out of the 44 variants had annotation for potential functional significance in mammary tissue, and from those, nine had functional annotations regarding histone modifications associated with regulatory marks and DHS (Figures 4.1A and 4.1C). After detailed screening and based on the strength of potential regulatory evidence, three causal candidates were found in strong LD with the daeSNP (rs11240762), and chosen for further analysis: rs6692377 (intronic of *PIK3C2B*, r^2=0.81), rs3789052 (intronic of *MDM4*, r^2=0.90) and rs12730457 (intergenic between *MDM4* and *LRRN2*, r^2=0.80).

For the variant rs6692377 ChIP-seq data from ENCODE show evidence of CTCF (CCCT-binding factor) binding in Hmec cells (primary human mammary epithelial cell), and binding of CTCF,

TCF7L2 (transcription factor 7 like 2), ZNF217 (zinc finger protein 217), MYC (myc proto-oncogene protein) and POLR2A (RNA polymerase II subunit A) in breast cancer MCF7 cells. This SNP was also found to overlap sequence corresponding to the binding motifs of the transcription factors EGR1 (early growth response 1) and ESR2 (oestrogen receptor beta). There was an allelic difference of at least 12 points for the PWM score for both TFs. The reference allele G was predicted to bind preferentially to EGR1, while the alternative allele A was generated a high binding score for ESR2.

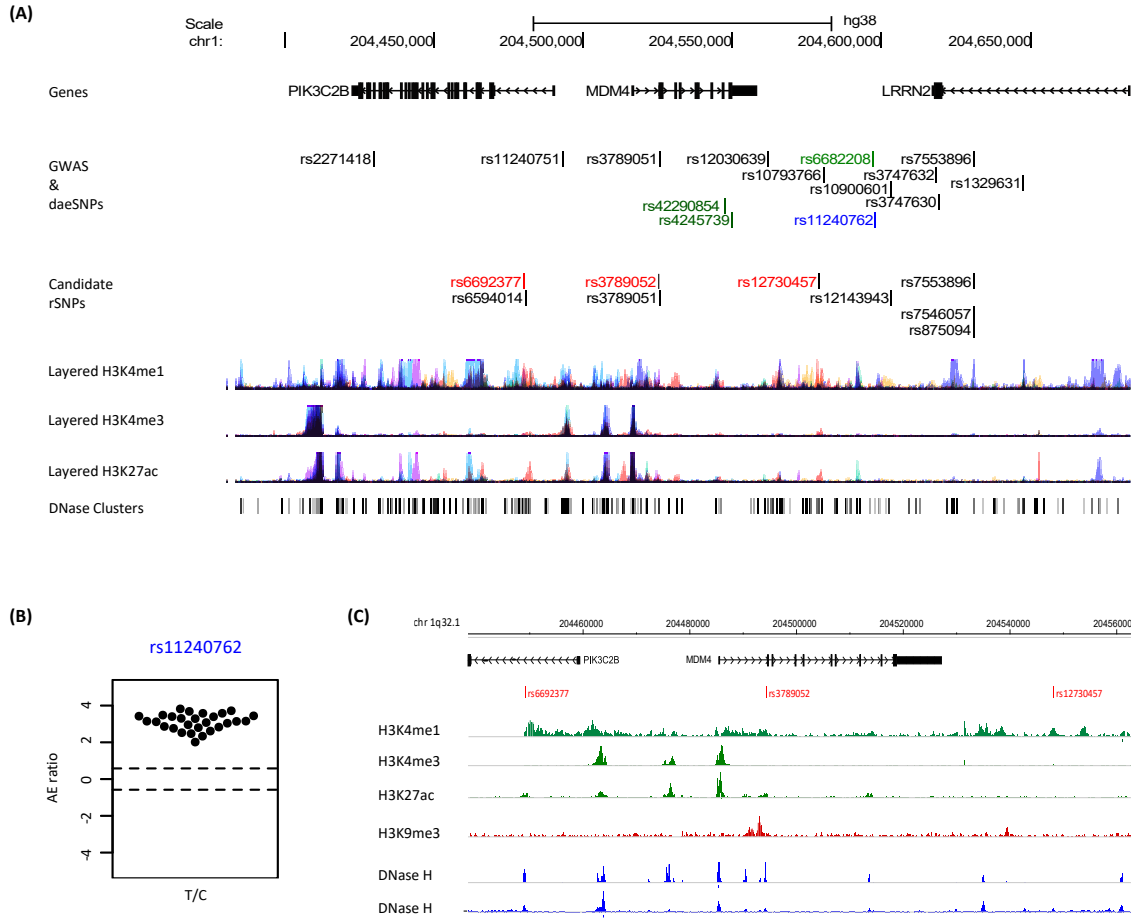


Figure 4.1. (A) Genomic landscape of the 1q32.1 locus with functional regulatory annotations on seven cell lines from ENCODE. The RefSeq genes are showed at the top. The breast cancer risk-associated SNPs from GWAS are indicated in green. The variant rs6682208 is both a GWAS SNP and a daeSNP. From the 12 transcribed SNPs with DAE (daeSNPs) found in this locus indicative of genes in this locus being under cis-regulation, the one with the specific pattern indicative of strong LD with the rSNP(s) is indicated in blue. The three SPNs with stronger regulatory potential are indicated in red. **(B)** DAE pattern of variant rs11240762, showing that all the analysed heterozygous individuals express preferentially the alternative C-allele (in healthy breast tissue). **(C)** Detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. Epigenetic marks for active or primed enhancers (H3K4me1), active enhancers (H3K27ac), and active promoters (H3K4me3) are shown, as well as repressive histone modifications (H3K9me3).

Besides overlapping regions of histones marks associated with enhancers and DHS, the variant rs3789052 also had evidence from ChIP-seq experiments of binding STAT3 (signal transducer and activator of transcription 3), FOS (fos proto-oncogene, AP-1 transcription factor subunit), and MYC in MCF10A-Er-Src cells (normal breast epithelial cell line). In addition, this SNP localizes in the binding site of the TF GATA3, originating over 12 points of difference in the PWM score of the two alleles (preferential binding of reference C, over the alternative T).

For the variant rs12730457 – reference allele C and alternative allele A – evidence of overlapping regions of histone marks associated with promoters and DHS was found. However, no predictions relative to PWM were available.

4.4.2.1. rs37899052 has the strongest cis-regulatory potential in locus 1q32.1

The initial in-vitro analysis involved the study of interactions between DNA and proteins. This intended to assess the previously described predictions of differential binding of TFs to DNA sequences harbouring the reference and alternative alleles of each of the selected SNPs.

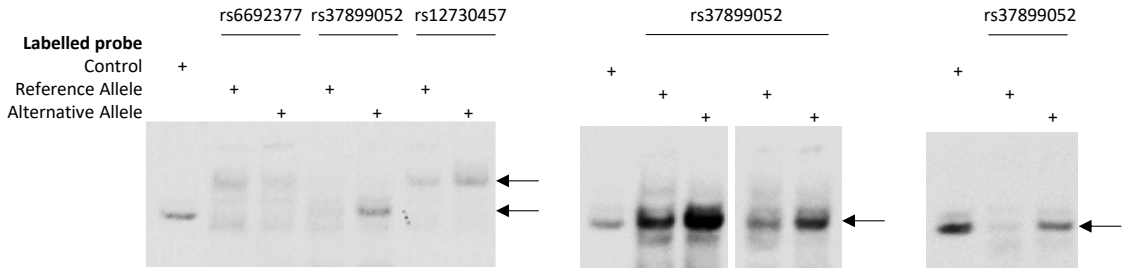


Figure 4.2. Analysis of DNA-protein binding of candidate rSNPs in 1q32.1. EMSAs were performed with protein nuclear extracts of HCC1954 cell line, and with probes containing the reference and alternative alleles of each potential rSNP. Candidate rSNPs include: rs6692377 (intronic in *PIK3C2B*), reference allele G, alternative allele A; rs3789052 (intronic in *MDM4*), reference allele C, alternative allele T; rs12730457 (5.3 kb from 3' of *MDM4*), reference allele C, alternative allele A. The black arrows indicate protein-oligonucleotide binding.

The EMSAs experiments in Figure 4.2 show that rs3789052 is the best candidate rSNP of the three in the 1q32.1 locus, as it presented a stronger allele-specific difference in DNA-protein binding interactions observed. Results showed a stronger interaction in the presence of the alternative T-allele, rather than in the presence of the reference C-allele.

Next, competition reactions were included in the EMSAs to identify the protein(s) binding preferentially to the alternative allele sequence. In these, specific unlabelled oligonucleotides for the reference and alternative alleles of the SNP were used. In addition, and according to ChIP-seq data and PWM predictions, unlabelled oligonucleotides known to bind STAT3 and GATA3 were also used for competition assays (Figures 4.3 and 4.4).

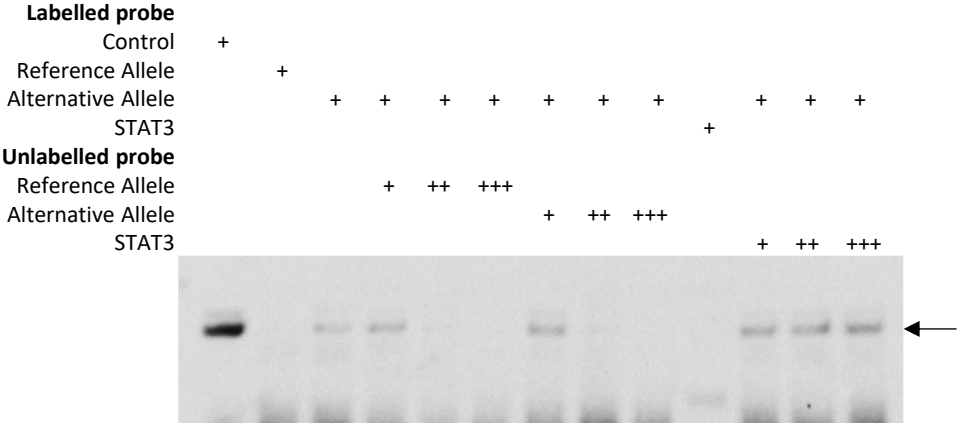


Figure 4.3. Analysis of DNA-protein binding of candidate rSNP rs3789052 (intronic in *MDM4* – reference allele C, alternative allele T) in 1q32.1. EMSAs were performed with protein nuclear extracts of the HCC1954 cell line, and probes containing the reference and alternative alleles of the variant rs3789052. Competition reactions were performed with 1-, 33- and 100-fold unlabelled probes. STAT3 consensus oligonucleotide sc-2571 (Santa Cruz Biotechnology, USA) was used. The black arrow indicates protein-oligonucleotide binding.

Surprisingly, the competition reactions using the alternative allele probe against the labelled reference probe were as strong as those of the alternative allele with itself. Additionally, competition with specific oligos for STAT3 binding was not observed (Figure 4.3), suggesting that another protein is involved in the DNA-protein interactions observed.

Competition assays with the alternative allele rs37899052 probe and specific oligonucleotides with the GATA3 consensus binding sequence also did not present evidence of GATA3 being the protein involved in the DNA-protein interactions observed (Figure 4.4).

Although these results validate that there are allelic differences in DNA-protein binding at the region of the candidate variants, the identity of the proteins involved remain to be identified.

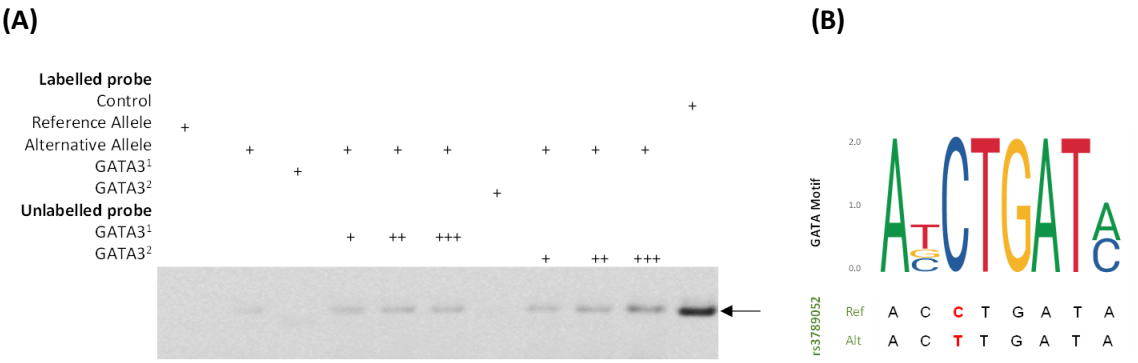


Figure 4.4. (A) Analysis of DNA-protein binding of candidate rSNP rs3789052 (intronic in *MDM4* – reference allele C, alternative allele T) in 1q32.1. EMSAs were performed with protein nuclear extracts of HCC1954 cell line, and with probes containing the reference and alternative alleles of the rs3789052. Competition reactions were performed with 1-, 33- and 100-fold unlabelled probes. [1]GATA3 oligonucleotide kindly provided by Dr. Chin from University of Cambridge. [2]GATA3 oligonucleotide sequence from other study (Adomas, et al., 2014). The black arrow indicates protein-oligonucleotide binding. **(B)** PWM predictions show that GATA transcription factor binds preferentially to the reference allele C of rs3789052.

4.4.3. In-silico functional analysis identifies one potential regulatory variant in locus 16q23.2

The 16q23.2 *locus* includes the *C16orf46* (chromosome 16 open reading frame 46) and the *GCSH* (glycine cleavage system protein H) genes.

From the 14 daeSNPs in this locus, the variant rs12444974, showed the pattern consistent with strong LD between the daeSNP and rSNP, with all the heterozygous individuals expressing preferentially the alternative A-allele (Figure 4.5B). Hence, this locus was also considered a potential candidate to harbour genetic variants responsible for cis-regulation in breast cancer, and proxy SNPs in strong LD ($r^2 \geq 0.8$) with rs12444974 were identified.

In the 16q23.2 locus, 112 SNPs were found to be in strong LD ($r^2 \geq 0.8$) with the daeSNP rs12444974 and 38 had annotation for potential functional significance in mammary tissue and/or breast cancer cell lines. From those, eight had functional annotations both for histone modifications associated with regulatory marks and DHS (Figures 4.5A and 4.5C). Analysis of the strength of potential regulatory evidence identified the variant rs74878296 (intronic in *C16orf46*) as the only SNP with strong potential to be the causal variant in high LD with the DAE SNP rs12444974 (r^2=0.82 and D'=0.95).

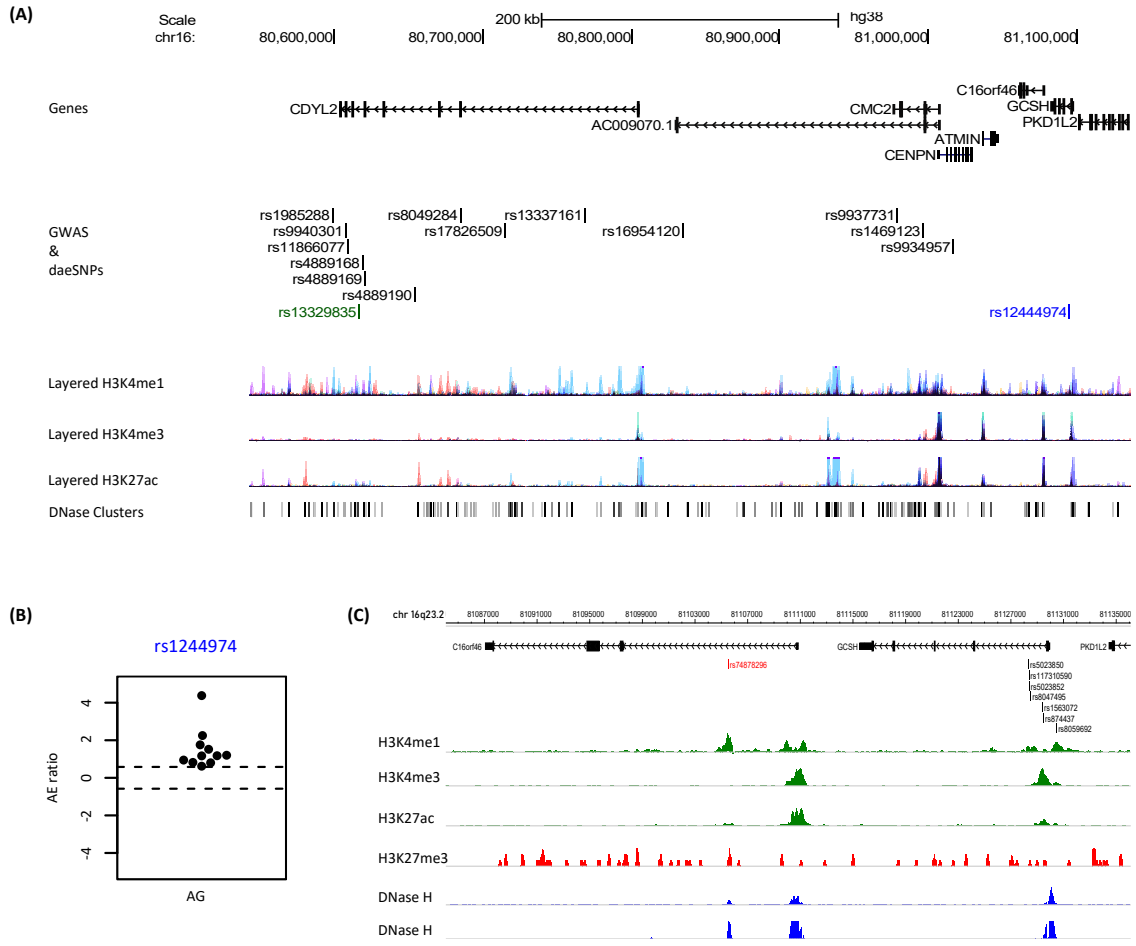


Figure 4.5. (A) Genomic landscape of the 16q23.2 locus with functional regulatory annotations on seven cell lines from ENCODE. The RefSeq genes are showed at the top. The breast cancer risk-associated SNP from GWAS is indicated in green. The transcribed SNP with DAE (daeSNP) with the specific pattern indicative of strong LD with the rSNP(s) is indicated in blue. **(B)** DAE pattern of variant rs12444974, showing that all the analysed heterozygous individuals express preferentially the alternative A-allele (in healthy breast tissue). **(C)** Detailed functional regulatory annotations in breast cell lines of the rSNP. The SNP with stronger regulatory potential is indicated in red. Epigenetic marks for active or primed enhancers (H3K4me1), active enhancers (H3K27ac), and active promoters (H3K4me3) are shown, as well as repressive histone modifications (H3K9me3).

Variant rs74878296 overlaps regions of histones marks associated with enhancers and DHS in breast epithelial cells, and PWM predictions showed that AP-1 (activating protein-1) transcription factor would preferentially bind to the alternative G-allele. Moreover, ChIP-Seq data also show evidence of FOS binding in MCF10A-Er-Src cells (normal breast epithelial cell line).

4.4.3.1. rs74878296 in locus 16q23.2 alters proteins binding

The in-vitro analysis involved the study of interactions between DNA and proteins by EMSA. This intended to assess the previously described predictions of differential binding of TFs to DNA sequences including the reference and alternative alleles of each of the selected SNPs.

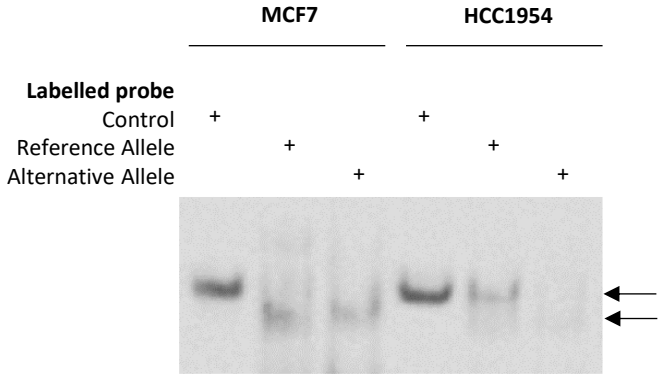


Figure 4.6. Analysis of DNA-protein binding of candidate rSNP rs74878296 (intronic in *C16orf46* – reference allele C, alternative allele G) in 16q23.2. EMSAs were performed with protein nuclear extracts of MCF7 and HCC1954 cell lines, and with probes containing the reference and alternative alleles. The black arrows indicate protein-oligonucleotide binding.

DNA-protein interactions were stronger in EMSAs conducted with HCC1954 protein nuclear extracts (Figure 4.6). Results show preferential binding of proteins to the reference C-allele of the candidate rSNP rs74878296 relatively to its alternative G-allele. Further analyses including competition reactions with unlabelled oligonucleotides for both alleles, confirmed the preferential binding of protein to the reference allele (Figure 4.7).

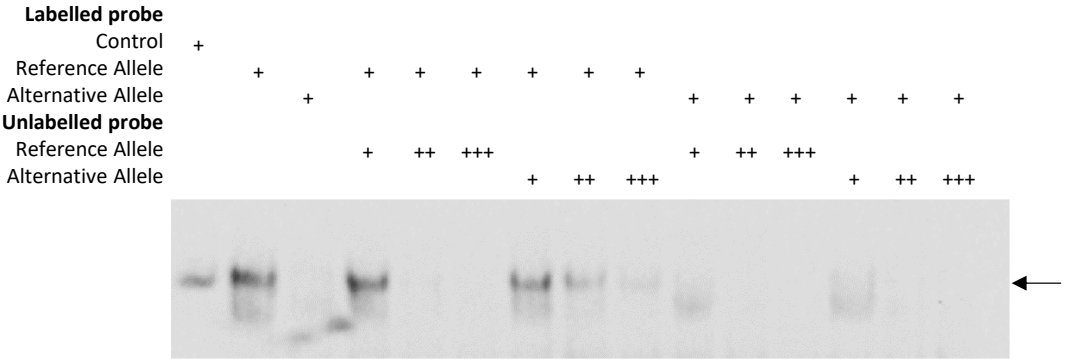


Figure 4.7. Analysis of DNA-protein binding of candidate rSNP rs74878296 (intronic in *C16orf46* – reference allele C, alternative allele G) in 16q23.2. EMSAs were performed with protein nuclear extracts of HCC1954 cell line, and with probes containing the reference and alternative alleles of the rs74878296. Competition reactions were performed with 1-, 33- and 100-fold unlabelled probes. The black arrow indicates protein-oligonucleotide binding.

According to PWM prediction, AP-1 would have a higher affinity with the alternative G-allele of rs74878296. However, considering that AP-1 is a heterodimer, different affinities are observed if FOS or JUN (jun proto-oncogene, AP-1 transcription factor subunit), or their variants are considered (Figure 4.8). Transcription factors FOS, and JUN are shown to have similar affinity for both alleles, whereas JUND has a slightly higher affinity for reference C-allele. Considering this, the available ChIP-seq results and the EMSAs results revealing stronger DNA-protein interaction in the presence of the reference allele, competition reaction with c-FOS specific antibody was performed (Figure 4.9). However, no supershift was observed in the presence of the c-FOS antibody, suggesting that c-FOS is not the protein binding preferentially to the reference C-allele.

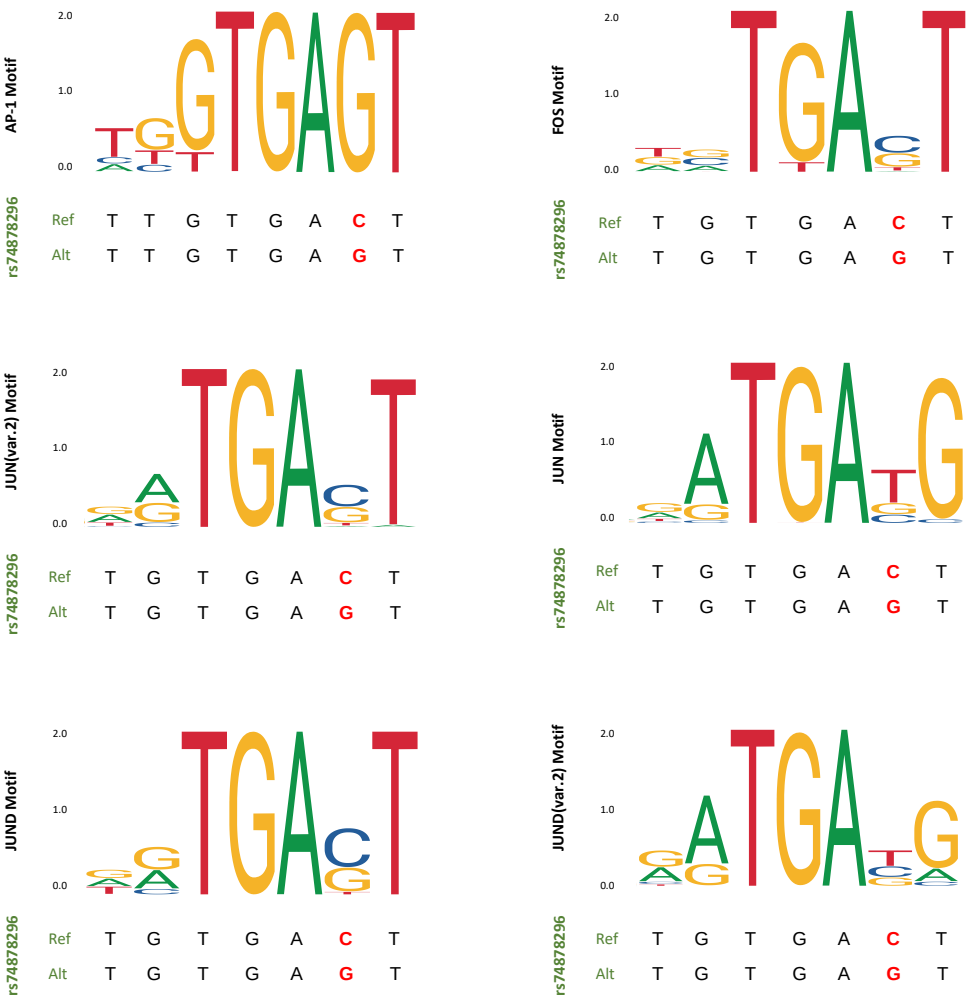


Figure 4.8. Position weight matrix profiles for transcription factor AP-1 and related proteins. Different components of the AP-1 heterodimer show different affinities for both alleles of rs74878296.

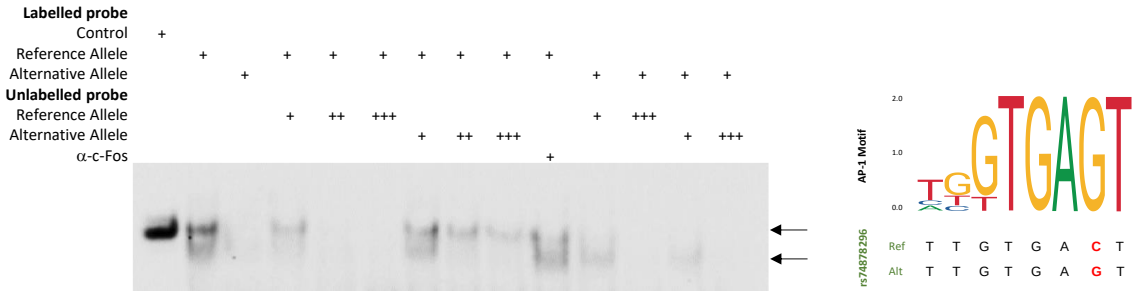


Figure 4.9. Analysis of DNA-protein binding of candidate rSNP rs74878296 (intronic in *C16orf46* – reference allele C, alternative allele G) in 16q23.2. EMSAs were performed with protein nuclear extracts of HCC1954 cell line, and with probes containing the reference and alternative alleles of the variant rs74878296. Competition reactions were performed with 1-, 33- and 100-fold unlabelled probes. c-FOS antibody sc-166940 (Santa Cruz Biotechnology, USA) was used. The black arrows indicate protein-oligonucleotide binding.

4.4.4. Allelic expression of genes in the 17q22 locus predicts breast cancer risk

4.4.4.1. Genetic variants regulate genes in 17q22 risk-locus in strong linkage disequilibrium with the lead risk-SNP

Firstly, we assessed whether *COX11* (cytochrome C oxidase subunit 11), *TOM1L1* (target of Myb1 like 1 membrane trafficking protein), or *STXBP4* (syntaxin binding protein 4) were under the control of cis-regulatory variants in normal breast tissue. For this, we calculated normalized allelic expression (AE) ratios at 22 tSNPs located in the three genes (Figure S4.19). Thirteen (59%) tSNPs showed significant deviations from equimolar allelic expression and were designated differentially allelic expressed SNPs – daeSNPs (Table 4.3). The observed differences between alleles reached a maximum of 2.15-fold at rs7643. We identified daeSNPs in all three genes in the locus, supporting that all are targets of *cis*-regulatory variation.

The patterns of the AE ratio distributions are indicative of the linkage disequilibrium (LD) between the regulatory variant and the transcribed variant where AE is measured (Xiao & Scott, 2011). Hence, to test if there is a link between cis-regulation and risk, we examined the pairwise LD between the daeSNPs and the locus lead risk-SNP rs2787486 and matched it to the AE ratio distribution patterns. Four daeSNPs showed marked preferential expression of the same allele in all heterozygous individuals tested and were in high LD ($r^2 \geq 0.6$) with rs2787486, suggesting that the same variant could confer risk and regulate gene expression levels.

One of these four daeSNPs, rs2628315, is in complete LD with the risk-variant and maps exclusively to the *STXBP4* gene (Table 4.3). At this variant, the allele preferentially expressed is associated with protection against breast cancer (Figure 4.10), suggesting that a higher predominance of the A allele is beneficial. Concordantly, the GTEx project reports rs2628315 as an eQTL (expression quantitative trait locus) for *STXBP4* expression in mammary tissue (p-value= 9.68E-07) (GTEx Consortium, 2013).

Another two daeSNPs, rs12936860 and rs17817901, map to a region shared by the *TOM1L1* and *COX11* genes and are in strong LD with rs2787486 (r^2=0.74 for both, Table 4.3). Of these two daeSNPs, rs17817901 showed the most significant differential allelic expression pattern, in which all heterozygotes preferentially expressed the alternative G allele. This AE ratio distribution is consistent with the daeSNP being in complete LD with the *cis*-regulatory variant (rSNP) creating the allelic effect (Xiao & Scott, 2011), which facilitates the mapping of the latter (Figure 4.10). Additionally, our results suggest that the preferential expression of the alternative G allele could correlate with the protective effect of the alternative C allele of rs2787486.

The fourth daeSNP in high LD with risk-variant rs2787486 is rs9899602 (r^2=0.66) that maps exclusively to *TOM1L1*. It showed preferential expression of the reference T allele, which is correlated to the risk-associated A allele of rs2787486 (Figure 4.10).

These results suggest that the differential allelic expression detected in all three genes could be associated with the risk of breast cancer and that all genes are candidate targets for the risk detected in the locus.

Table 4.3. Summary statistics for tSNPs at the 17q22 risk locus. MAF – alternative allele frequency; p.adjust – p-value adjusted for multiple testing with Bonferroni method; CIinf and CIsup – Inferior and superior limits of the 95% confidence interval (CI) of the mean.

						DAE - One-Sample t-Test			
rsID	**Gene**	**Reference Allele**	**Alternative Allele**	**MAF**	**LD with rs2787486 (r^2)**	**p.adjust**	**Mean**	**CIinf**	**CIsup**
rs9899602	*TOM1L1*	T	C	0.29	0.66	4.14E-06	0.410	0.298	0.522
rs9303360	*TOM1L1*	C	T	0.42	0.25	6.05E-01	-0.334	-0.629	-0.040
rs12165058	*TOM1L1*	C	T	0.46	0.22	1.00E+00	0.015	-0.089	0.119
rs12944690	*TOM1L1/COX11*	G	A	0.46	0.20	9.26E-04	-0.199	-0.284	-0.115
rs12936860	*TOM1L1/COX11*	G	A	0.27	0.74	1.53E-03	-0.209	-0.295	-0.122
rs17817865	*TOM1L1/COX11*	G	A	0.27	0.14	2.68E-12	0.350	0.296	0.404
rs2287136	*TOM1L1/COX11*	G	A	0.27	0.14	1.40E-01	0.098	0.030	0.166
rs17817901	*TOM1L1/COX11*	A	G	0.27	0.74	3.65E-21	-1.050	-1.110	-0.995
rs7643	*TOM1L1/COX11*	A	G	0.46	0.22	3.76E-25	-1.110	-1.170	-1.050
rs2541240	*COX11*	G	A	0.33	0.15	9.26E-02	-0.359	-0.594	-0.124
rs17817950	*COX11/STXBP4*	G	A	0.27	0.18	7.35E-01	0.277	0.024	0.531
rs12947685	*COX11/STXBP4*	T	C	0.27	0.18	1.00E+00	0.131	-0.047	0.309
rs8064882	*COX11/STXBP4*	C	T	0.27	0.18	8.27E-01	0.304	0.019	0.589
rs4794549	*COX11/STXBP4*	C	A	0.29	0.20	3.94E-02	0.282	0.116	0.449
rs1420526	*STXBP4*	T	C	0.26	0.00	1.96E-05	0.387	0.259	0.514
rs17818238	*STXBP4*	G	A	0.21	0.12	1.00E+00	-0.120	-0.280	0.409
rs8069999	*STXBP4*	G	A	0.19	0.12	3.45E-02	-0.159	-0.249	-0.069
rs2628318	*STXBP4*	G	A	0.42	0.27	3.72E-02	0.227	0.095	0.358
rs2628315	*STXBP4*	G	A	0.28	1.00	1.14E-06	0.617	0.464	0.770
rs244303	*STXBP4*	A	G	0.41	0.26	1.00E+00	0.028	-0.050	0.105
rs244301	*STXBP4*	T	C	0.41	0.26	4.31E-02	-0.115	-0.183	-0.046
rs244298	*STXBP4*	T	C	0.41	0.26	1.63E-07	-0.486	-0.607	-0.365

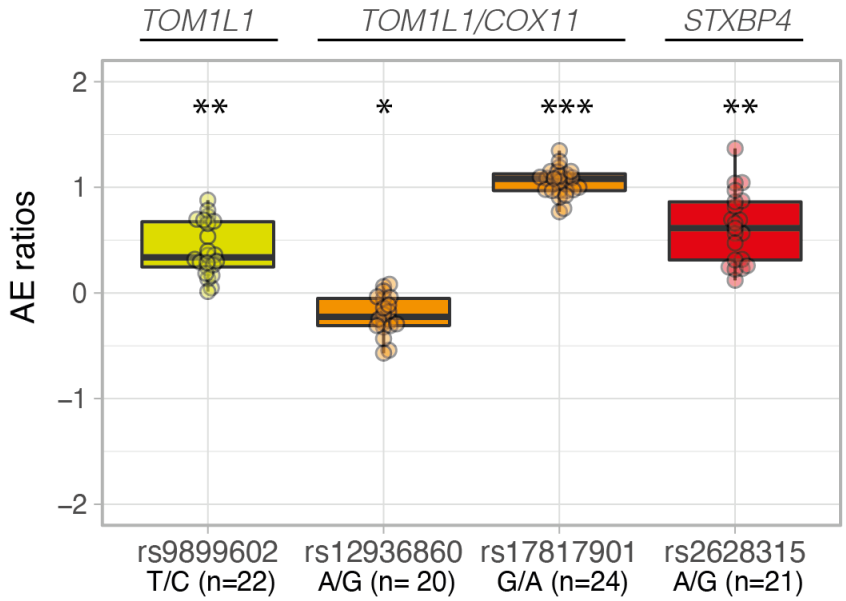


Figure 4.10. Genes in the 17q22 locus are under the effect of cis-regulatory variants genetically related to breast cancer risk variants. Boxplots of the allelic expression (AE) ratios for four variants in strong LD with lead risk-SNP rs2787486, located in the genes indicated above the graph. One-sample t-test for mean allelic expression equal to zero: * $P<10^{-2}$; ** $P<10^{-5}$; *** $P < 10^{-10}$; Boxplots and data points are coloured based on LD with rs2787486: yellow $r^2 \geq 0.6$, orange $r^2 \geq 0.7$, and red $r^2=1$.

4.4.4.2. AE ratios in normal breast tissue and blood associate with breast cancer risk

Next, we sought to discern between chance colocalization and a true association between allelic expression ratios and risk association. We hypothesized that if risk-causing variants are cis-regulating genes in the 17q22 locus, then the allelic expression ratios they generate should have distinct distributions in patients (cases) and healthy individuals (controls). Hence, we performed a case-control association analysis using AE ratios measured in the normal breast as a quantitative phenotype. We carried out this analysis for the three daeSNPs displaying the highest LD with the risk-associated variant, each localized in one of the genes in the locus. As rs12936860 and rs17817901 are in complete LD, we only analysed rs17817901. Data for the multiple experiments run for this analysis can be found in Supplementary Data.

Table 4.4. Association between AE ratios at two variants in 17q22 and Breast Cancer Risk statistics. n – number of samples; CIinf and CIsup – Inferior and superior limits of the 95% confidence interval (CI) of the Hedges' *g*; p.perm – Permuted Mann-Whitney test p-values.

	rsID	Controls (n)	Cases (n)	Hedges' g	CIinf	CIsup	p.perm
Breast Tissue	rs17817901	23	50	-0.486	-0.889	-0.147	0.0534
	rs2628315	19	42	-1.237	-1.928	-0.366	0.0002
	rs9899602	18	50	0.038	-0.593	0.756	0.889
Blood	rs17817901	14	24	-0.737	-1.431	-0.002	0.034
	rs2628315	12	23	-1.419	-2.156	-0.780	0

The daeSNP rs2628315, located in an intron of *STXBP4,* showed the largest effect size (*g*=-1.237) (Table 4.4, Figure 4.11), and the most significantly different AE ratio distributions. This result shows that the AE ratio distribution in the normal breast of cases is shifted towards the preferential expression of the reference G allele, the least expressed in controls. As rs2628315 and the risk-variant rs2787486 are in complete LD, this result suggests that increased risk is associated with the preferential expression of the reference allele of both variants.

The analysis of rs17817901 also revealed a shift in the distribution of AE ratios in cases towards the preferential expression of the reference A allele with an estimated effect size of *g*=-0.486 (Table 4.4, Figure 4.11). As rs17817901 is in strong LD with the risk-variant rs2787486, our results suggest that risk could be associated with a higher expression of the reference A allele of rs17817901. However, as rs17817901 locates in a genomic region shared by the *TOM1L1* and *COX11* genes, we considered both genes as candidate target genes for breast cancer.

However, the analysis of the daeSNP rs9899602 did not reveal any significant difference between the two populations, suggesting that *TOM1L1* might not be a target gene for the risk detected via the lead-SNP rs2787486 (Table 4.4, Figure S4.20). We note that rs9899602 is the daeSNP in weaker LD with the lead risk-SNP amongst the ones with significant DAE (Table 4.3). Integrated with the results obtained for rs17817901, this suggests that *COX11* is the most likely candidate from the two overlapping genes.

Based on the shared cis-regulation of breast cancer genes between breast tissue and blood (Maia, et al., 2009; Darabi, et al., 2016), we next examined the associations described above in blood samples from cases and controls. For the daeSNP rs2628315, we found a comparable effect size (*g*=-1.419) and a significant difference in the AE ratio distributions of the two groups,

with a concordant shift direction with that observed in breast tissue: preferential expression of the risk-associated G allele of rs2628315.

For the daeSNPs rs17817901, we found a larger effect size than was observed in breast tissue (*g*=-0.737), and in concordant direction – cases preferentially expressed the A- rs17817901 allele which is in strong LD with the risk-associated C-rs2787486 allele.

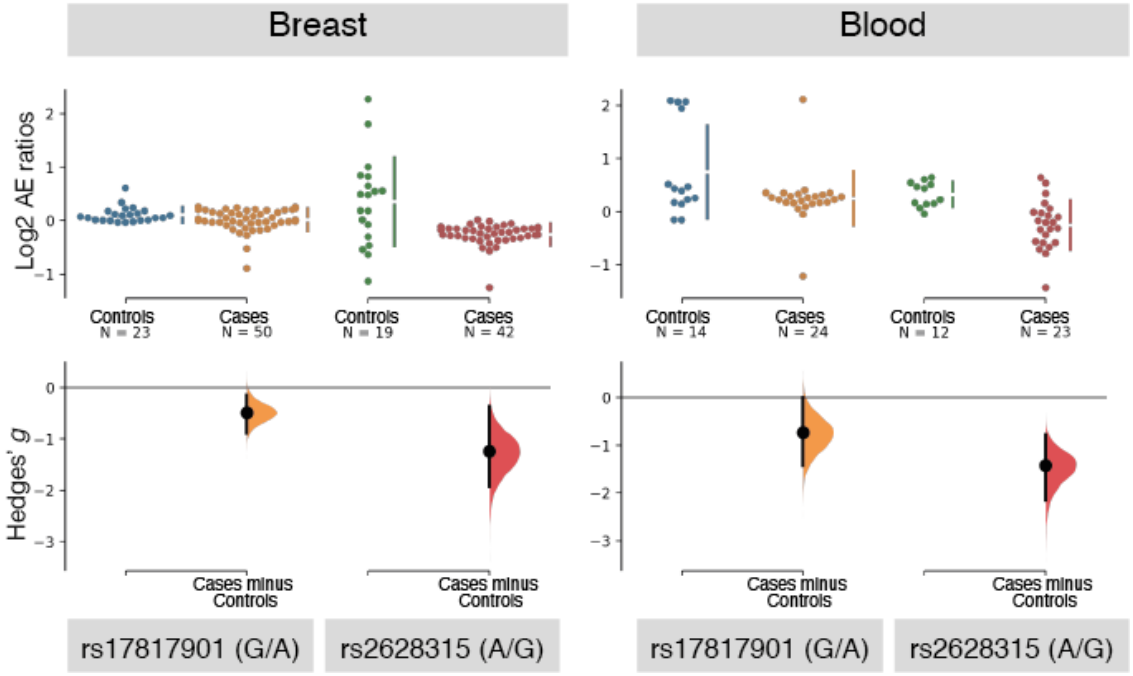


Figure 4.11. Case-control study using allelic expression ratios identifies risk in the 17q22 locus in breast tissue and blood samples. Cumming estimation plot of Hedges' g between breast cancer cases and controls for allelic expression ratios calculated at rs17817901 (ratio calculated as allele G by allele A) and rs2628315 (ratio calculated as allele A by allele G) in normal breast and blood. The heterozygous individuals for each indicated variant and tissue are plotted on the upper axes, with controls displayed in blue and cases in orange for rs17817901 and in green and red for rs2628315. The vertical lines next to the raw data correspond to the conventional mean ± standard deviation error bars, where the mean of each group is indicated as a gap in the line. The mean difference is plotted on the lower axes as a bootstrap sampling distribution (bootstrap n=5000). The mean differences are depicted as dots, and the 95% confidence intervals are indicated by the ends of the vertical error bars.

4.4.4.3. Functional analysis reveals three rSNPs in the locus 17q22

We next aimed at pinpointing the actual causing regulatory variant(s) (rSNPs) in the 17q22 locus. The pattern of AE ratio distribution is indicative of the LD between the rSNP(s) and the daeSNP where expression is measured (Xiao & Scott, 2011). The distribution of the normalized AE ratios

measured in breast tissue at rs17817901 (Figure 4.10) strongly suggests that the rSNP(s) generating this effect is(are) in strong to complete LD with rs17817901. Therefore, 106 variants in strong LD ($r^2 \geq 0.8$) with rs17817901 were analysed in-silico for known functional data and predictions of functionality, with the strongest candidates subsequently tested in-vitro (Table 4.1). These analyses identified three candidate rSNPs - rs17817901, rs8066588, and rs9891865- regulating the binding of transcription factors and a miRNA.

Data for rs17817901 from public databases indicated limited evidence for functionality on breast tissue. However, it showed that the variant overlaps an enhancer element active in T cells and is an eQTL for all three genes in the locus in various tissues (not breast or blood). Furthermore, rs17817901 maps to the shared 3'UTRs of the genes *TOM1L1* and *COX11*, and we predicted before that its alternative G allele generates a binding site for hsa-miR-194-5p (context score=-0.229) (Jacinta-Fernandes, Xavier, Magno, Lage, & Maia, 2020), an oncogenic miRNA expressed in breast (Yang, Xiao, & Zhang, 2018; Zhang, et al., 2014; Fernandes, et al., 2021). This prediction was validated in reporter assays using a mimic oligo of the oncogenic hsa-miR-194-5p, which showed decreased reporter activity for the alternative G allele (protective) compared to the A allele and the empty vector (Figure 4.12A). A comparable difference was observed for endogenous levels of the miRNA although non-significant.

Next, the candidate rSNP rs8066588, an intronic variant to *TOM1L1*, is in complete LD with rs17817901 and strong LD with the risk lead-SNP ($r^2=0.85$). According to the GTEx project, this variant is also an eQTL for *STXBP4* and an sQTL (splicing quantitative trait locus) for *COX11* in breast tissue. Also, we confirmed in-vitro that rs8066588 changes the binding motif of the transcription factor TCF3 (transcription factor 3, E2A family) (Figure 4.12B), by showing TCF3 preferential binding to the reference C-allele (Figure 4.12C, Figure S4.21). Moreover, this variant overlaps an active enhancer element in breast tissue and is associated with strong transcription and a DNAseI hypersensitive site in mammary cells (Figure 4.12D).

Finally, the candidate rSNP rs9891865, located in an *STXBP4* intron, is in strong LD with the daeSNP rs17817901 ($r^2=0.8$) and the risk lead-variant rs2787486 ($r^2=0.75$). This variant is indicated in the GTEx project as an eQTL to *STXBP4* and overlaps a DNAseI hypersensitive site in mammary cells (Figure 4.12D). The most significant allele-specific binding motif alteration was predicted for the transcription factor TCF12. Although we could detect in-vitro differences in the binding of proteins between the two alleles, we could not confirm the actual protein involved (Figure S4.22).

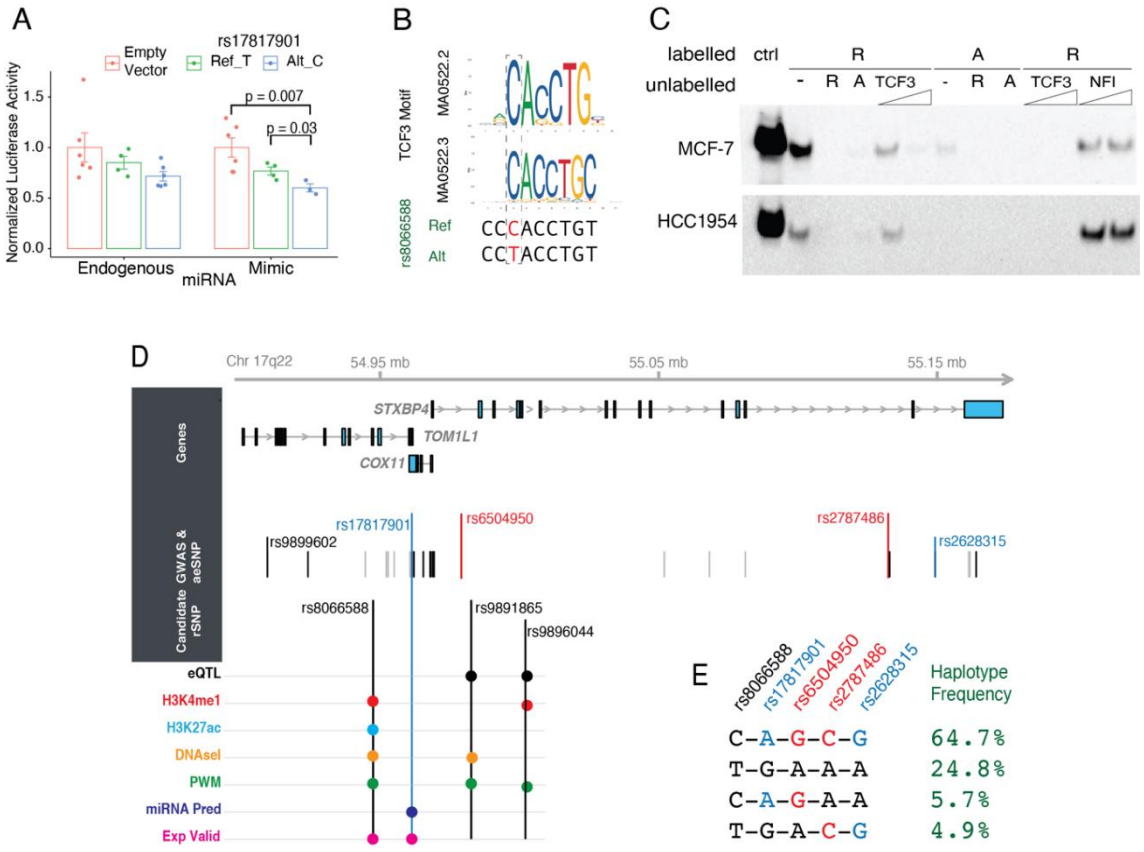


Figure 4.12. Functional characterization of candidate rSNPs related to breast cancer risk variants in the 17q22 locus. **(A)** Luciferase assays showed a more significant reduction of signal for the alternative G allele of rs17817901, both under endogenous miRNA conditions and using a mimic oligo of hs-miR-194-5p (y-axis shows the normalized luciferase Firefly/Renilla ratio) p-values indicated correspond to Welch's test). **(B)** The reference C allele of rs8066588 is predicted to generate a motif for the binding of TCF3. **C** - EMSA experiments using protein extracts from MCF-7 and HCC1954 cell lines show preferential binding of the reference C allele of rs8066588 (R – reference, A – alternative alleles); competition with oligo of known binding site forTCF3 competes with observed binding, which does not occur with negative control oligo (NFI binding motif). **(D)** Genomic landscape of 17q22 locus showing the RefSeq genes in the top panel; then the location of the variants in which AE ratios were measured (aeSNPs), with black indicating daeSNPs with significant differential allelic expression, in blue the risk-daeSNPs and in red the GWAS-SNPs; next is the location of candidate regulatory variants rSNPs and below the epigenetic, molecular and experimental evidence (eQTL, histone modifications, DNAseI hypersensitivity sites, PWM position weight matrix, miRNA prediction, and experimental validation). **(E)** Haplotypes constructed with genotyping data of normal breast samples from controls included in this study, with the frequency indicated. The colour scheme is as above, with risk-associated alleles for daeSNPs and GWAS-SNPs.

Overall, we found that cis-variants regulate all three genes in the 17q22 risk locus and that the population variability in allelic expression these genes present is associated with the risk of breast cancer. Furthermore, we identified and validated two cis-regulatory variants linked to the allelic expression observed in the three genes. Although the LD is high between the candidate rSNPs, the daeSNPs, and the risk-variant, these variants form four haplotypes (Figure 4.12E).

More importantly, two haplotypes in this region include the A-rs2787486 protective allele: the main haplotype (24.8%) in phase with preferentially expressed alleles in healthy controls, and another less common (5.7%) in phase with the preferentially expressed alleles in cases.

4.4.5. Candidate rSNP rs2699887 for *PIK3CA* alters binding of transcription factor NF-YA

In the scope of another study conducted in our group on allelic expression in *PIK3CA* (phosphatidylinositol-3-kinase alpha) in breast cancer, in-silico analysis identified the SNP rs2699887 as a candidate regulatory variant acting on *PIK3CA*. This section shows the work conducted for in-vitro validation of rs2699887 as the regulatory variant.

To search for the association of cis-regulatory variation and the expression of the *PIK3CA* gene, allelic expression analysis in normal breast tissue from 64 healthy women was assessed. This identified six SNPs with DAE in *PIK3CA*. Five of these six daeSNPs were in strong LD with each other, and for four of those (rs7636454, rs3960984, rs12488074, rs9838411), all samples expressed preferentially the same allele. Next, mapping analysis using genotype information from the SNPs located at ±250kb of each daeSNP, identified 273 candidate rSNPs. Further in-silico analysis identified the rs2699887, located in the first intron of *PIK3CA*, lying in a region rich in typical active promoter chromatin modification marks (H3K4me3, H3K9ac, H3K36me3), and open chromatin marks in breast mammary epithelial cells and breast myoepithelial cells, as the candidate rSNP with the strongest cis-regulatory potential. The significant association (permuted p-value=0.03) of this variant with the daeSNP rs12488074, indicated that the rs2699887 could explain the differential allelic expression observed in that *PIK3CA*.

Initial analysis of DNA-protein binding of rs2699887 showed differential binding of the reference and alternative alleles to proteins, with stronger DNA-protein interaction in the presence of the reference allele (Figure 4.13A). On the other hand, predictions of PWM indicated that the rs2699887 could alter the binding affinity of NF-YA (nuclear transcription factor Y subunit alpha), which would preferentially bind to the alternative T-allele (Figure 4.13B). Results from EMSAs where competition reactions were conducted, confirmed that prediction. Figure 4.13A shows that the NF-YA consensus nucleotide but not the E2F1 competes with the alternative allele for protein binding. These results were further confirmed with a second EMSA including increasing amounts of competing oligonucleotides, where specific DNA-protein binding was observed (Figure 4.14).

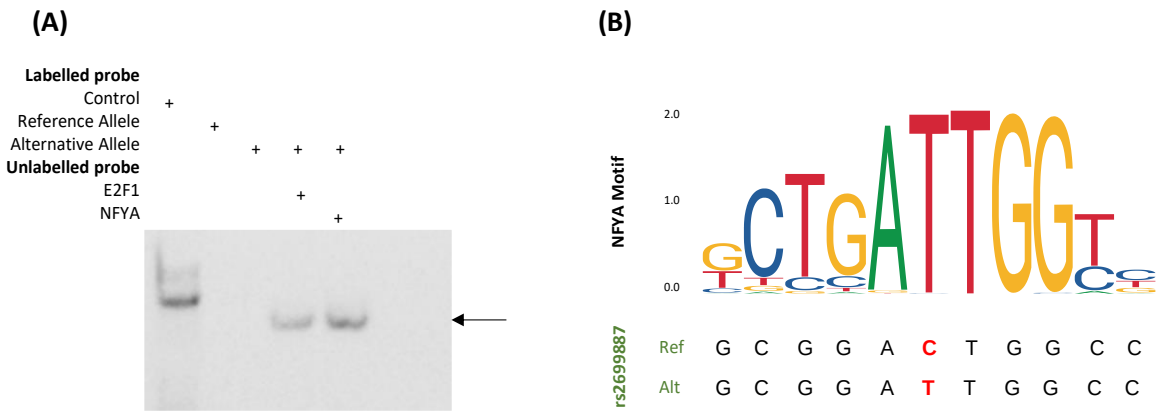


Figure 4.13. (A) Analysis of DNA-protein binding of candidate rSNP rs2699887 (intronic in *PIK3CA* – reference allele C, alternative allele T) in 3q26.32. EMSAs were performed with protein nuclear extracts of HCC1954 cell line, and with probes containing the reference and alternative alleles of the rs2699887. Competition reactions were performed with 100-fold unlabelled probes, including consensus binding sequences for E2F1 (sc-2507, Santa Cruz Biotechnology) and NF-YA (Xu, et al., 2012). The black arrow indicates protein-oligonucleotide binding. **(B)** PWM predictions show that NF-YA transcription factor binds preferentially to the alternative allele T of rs2699887.

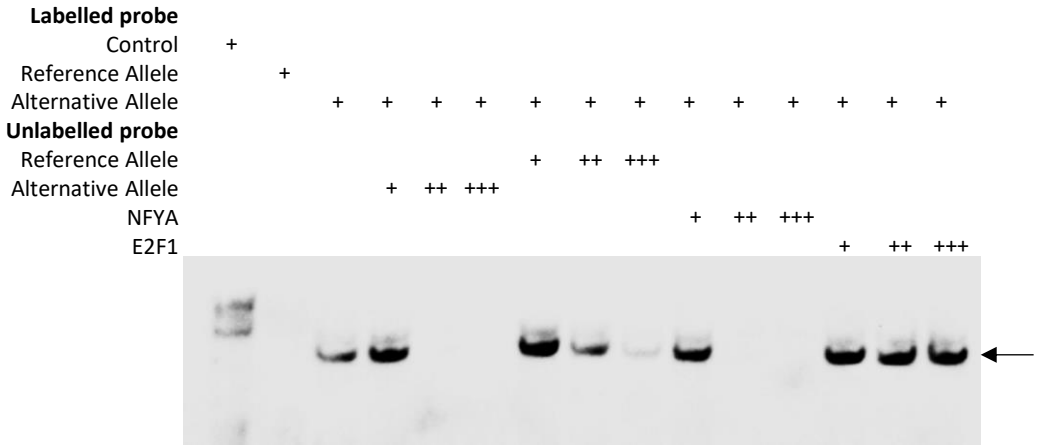


Figure 4.14. Analysis of DNA-protein binding of candidate rSNP rs2699887 (intronic in *PIK3CA* – reference allele C, alternative allele T) in 3q26.32. EMSAs were performed with protein nuclear extracts of HCC1954 cell line, and with labelled probes containing the reference and alternative alleles of the rs2699887. Competition reactions were performed with 1-, 33- and 100-fold unlabelled probes. Consensus oligonucleotides for E2F1 (sc-2507, Santa Cruz Biotechnology, USA) and NF-YA (Xu, et al., 2012) were used. The black arrow indicates protein-oligonucleotide binding.

Additionally, EMSAs including competition reactions with a NF-YA specific antibody were performed (Figure 4.15). Results confirmed that different affinities to NF-YA explained the differential protein binding observed for the two alleles of the rs2699887. Competition reactions conducted with the labelled probe for the alternative allele and the NF-YA specific antibody resulted in a supershift, due to a higher molecular weight of the DNA-protein-antibody complex. Specificity of antibody binding was confirmed with the addition of HMGA (high mobility group AT-hook 1) antibody, with which no interaction was observed.

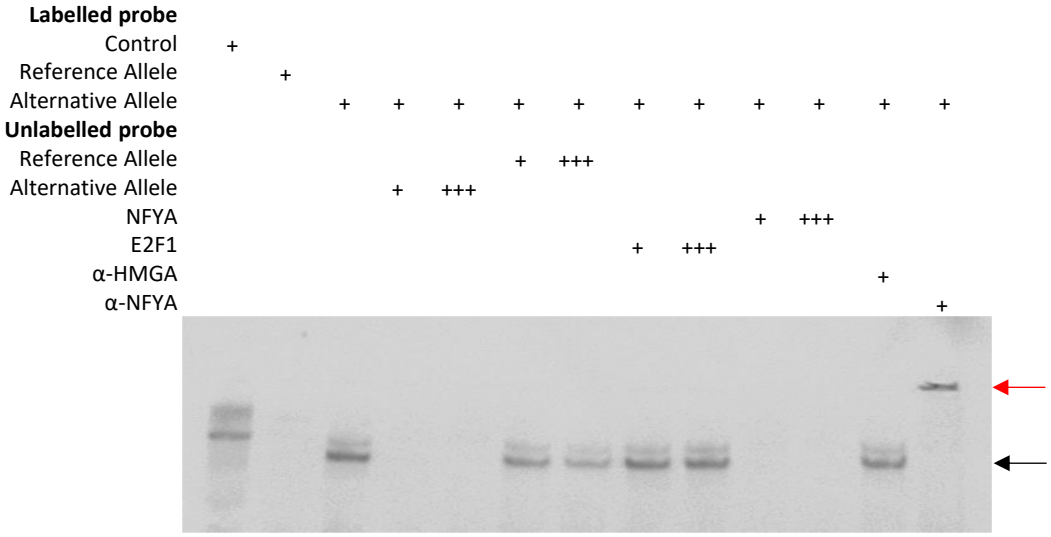


Figure 4.15. rs2699887 differentially binds transcription factor NF-YA in-vitro. EMSA analysis using protein extracts of breast cancer cell line HCC1954 and biotin-labelled oligonucleotides containing reference allele C or alternative allele T of rs2699887. Competition assays included 1- or 100-fold unlabelled oligonucleotides, which included consensus binding sequences for NF-YA (Xu, et al., 2012) and E2F1 (sc-2507, Santa Cruz Biotechnology, USA). Supershift assays were carried out using antibodies against NF-YA (sc-10779, Santa Cruz Biotechnology, USA) and HGMA (ab4078, abcam, UK). The black arrow indicates protein-oligonucleotide binding. The red arrow indicates antibody-protein-oligonucleotide binding.

4.4.5.2. In-vivo functional analysis of variant rs2699887 in *PIK3CA*

In-vivo functional analysis was performed to validate results from in-vitro and in-silico studies which showed that candidate rSNP rs2699887 for *PIK3CA* alters binding of transcription factor NF-YA. ChIP assays were performed to validate the altered binding of NF-YA between the two alleles in living cells heterozygous (C/T) for the SNP of interest.

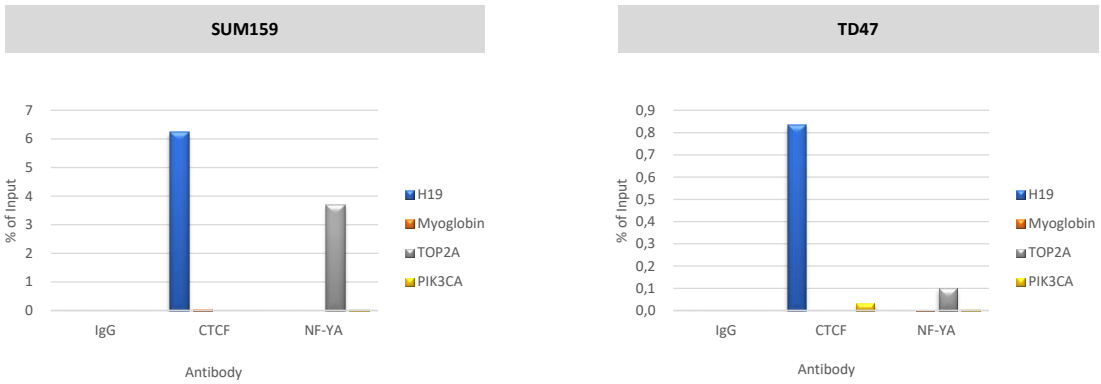


Figure 4.16. ChIP-qPCR analysis of candidate variant rs2699887 with SUM159 and T47D breast cancer cell lines. Sheared chromatin from 4 million cells, 0.5 µl of the positive control CTCF antibody and 1 µl of the negative IgG control, were used per IP. Primers for the H19 imprinting control region (positive control) and Myoglobin Exon 2 (negative control) were provided with the iDeal ChIP-seq kit. *TOP2A* primers amplify a 119 bp fragment (positive control for DNA::NF-YA interaction) and *PIK3CA* primers amplify a 114 bp region of the gene including the candidate rSNP. Since the amount of used input was 1% of the amount of sample used for ChIP, the recovery was calculated as follows: % of input =2^(C_{tinput}-Ct_{sample}).

ChIP-qPCR using control antibodies (anti-IgG and anti-CTCF) validated the assays, although percentages of recovery obtained with SUM159 cells were more concurrent with the expected than the ones from T47D cells (Figure 4.16). For the positive control antibody, 5% of recovery of the H19 control region was expected and over 6% was obtained. The recovery of the negative control target was lower than 0.5% as supposed. ChIP-qPCR with anti-NF-YA showed binding of this transcription factor to *TOP2A* as expected, but not to *PIK3CA*. Hence, these results did not validate the interaction between NF-YA and the candidate rSNP in living cells.

4.5. Discussion

Integration of GWAS data on risk-variants for breast cancer and data from a genome-wide regulatory map on breast tissue from healthy women, identified 32 loci associated with risk and with strong cis-regulatory potential. These loci included one or more GWAS variants in strong LD ($r^2 \geq 0.8$) with at least one tSNP with differential allelic expression. In particular, 12 of those loci presented a specific DAE pattern, where all the heterozygotes expressed preferentially only one allele, displaying higher potential to harbour genetic variants responsible for cis-regulation in breast cancer. In-silico functional analysis identified the 1q32.1, 16q23.2 and 17q22 loci as the ones with stronger cis-regulatory potential and association with breast cancer risk.

This work showed that the locus 1q32.1 harboured 12 daeSNPs in moderate to strong LD with the GWAS risk-associated variants rs2280854, rs4245739, and rs6682208, showing that this locus displays cis-regulatory variation and association with breast cancer risk. In-silico analysis revealed three potential causal variants in high LD with the daeSNP rs11240762. For this daeSNP the alternative C-allele is preferentially expressed in all the analysed heterozygous healthy women and is the one associated with breast cancer risk. *MDM4* and *PIK3C2B* genes harbour two of the strongest potential cis-regulatory variants: rs6692377 (intronic of *PIK3C2B*, r^2=0.81 with rs11240762); and rs3789052 (intronic of *MDM4*, r^2=0.90 with rs11240762). This indicated that these genes, which were previously associated with breast cancer, could be the target genes in 1q32.1.

MDM4 acts as a negative regulator of the tumour suppressor p53 and is thought to contribute to cancer development by inhibiting p53 transcriptional activity (Marine, Dyer, & Jochemsen, 2007). MDM4 protein is expressed in normal breast ductal epithelial cells (Haupt, et al., 2015), where malignant growth can originate. Imbalance in MDM4 levels in these cells is associated with breast cancer risk (Haupt, et al., 2017).

One genetic variant located in the 3'UTR of *MDM4*, rs4245739, was first associated with increased risk for ovarian cancer (Wynendaele, et al., 2010). The presence of the A-allele was associated with overexpression of MDM4, and patients homozygous for this allele who do not express oestrogen receptor showed a 4.2-fold (95%CI=1.2–13.5; *P*=0.02) increased risk of recurrence and 5.5-fold (95%CI=1.5–20.5; *P*=0.01) increased risk of tumour-related death. On the other hand, the C-allele creates a target site for miR-191, downregulating *MDM4* expression, thus delaying cancer progression and increasing chemosensitivity. Furthermore, two GWAS also associated rs4245739 with increased risk of breast and prostate cancers (Garcia-Closas, et al.,

2013; Eeles, et al., 2013). Additionally, downregulation of *MDM4* was associated with the C-allele of the rs4245739, by altering the binding of two miRNAs in prostate cancer (Stegeman, et al., 2015). Another gene from the 1q32.1 locus, the *PIK3C2B*, is part of the phosphoinositide 3-kinase (PI3K) family and encodes for the PI3K-C2β protein. PI3-kinases play an important role in signalling pathways involved in cell proliferation, cell survival and migration, intracellular protein trafficking, and oncogenic transformation (Cantley, 2002; Yardena S., et al., 2004; Liu, Cheng, Roberts, & Zhao, 2009). PI3K-C2β has been shown to promote resistance to tamoxifen in breast cancer cells (Iorns, Lord, & Ashworth, 2009), and several variants in the *PIK3C2B* were associated with prostate cancer risk (Koutros, et al., 2010). Finally, the third gene from 1q32.1 locus, the *LRRN2*, was found to be upregulated in lung cancer metastasis (Dat, et al., 2012), and amplified and overexpressed in malignant glial tumours (González-Tablas, et al., 2018). The protein encoded by *LRRN2* is part of the leucine-rich repeated superfamily and is a homolog of proteins that function as cell-adhesion molecules.

Results from in-vitro functional analysis by EMSAs further indicated that the SNP rs3789052 had the highest regulatory potential. This variant presented the strongest allele-specific difference in DNA-protein binding interactions, which were preferentially established in the presence of the alternative T-allele. PWM predictions pinpointed that STAT3, GATA3, FOS, and MYC could be the proteins binding to the T-allele of rs3789052. STAT3 accelerates cell proliferation and survival, enhances the angiogenesis, contributes to the promotion of metastasis, and induces immune evasion (Lee, Jeong, & Ye, 2019). GATA3 is a critical factor in the normal development and function of the mammary gland, and in breast cancer luminal cell differentiation (Takaku, Grimm, & Wade, 2015). GATA3 is likely to be a prognostic factor for ER+ patients (Afzaljavan, Sadr, Savas, & Pasdar, 2021), and loss of its expression is associated with poor prognosis (Mehra, et al., 2005; Liu, et al., 2016). Members of the FOS (c-Fos, FosB, Fra-1, and Fra-2) family dimerise with JUN proteins (c-Jun, JunB, or JunD) to form the AP-1 transcription factor, which binds to promoters of specific target genes altering their expression (Milde-Langosch, 2005). In addition, the heterodimer AP-1 also contains proteins from the ATF and MAF families (Shaulian & Karin, 2002), and has been shown to regulate breast cancer cell growth by leading to cell cycle progression and breast cancer cell proliferation (Shen, et al., 2008). The role of the proto-oncogenes c-Jun and c-Fos, in tumour development process, is well known. Several studies showed c-Fos to downregulate tumour suppressor genes (Bakin & Curran, 1999), promote invasive growth of cancer cells (Hu, et al., 1994), as well as invasive and metastatic growth in mammary epithelial cells (Fialka, et al., 1996). In breast cancer, c-Fos was also associated with decreased survival (Bland, Konstadoulakis, Vezeridis, & Wanebo, 1995). Activation of c-Jun was

associated with proliferation and angiogenesis in invasive breast cancer (Vleugel, Greijer, Bos, van der Wall, & van Diest, 2006). MYC is also involved in cell growth, proliferation, differentiation, and apoptosis (Dang, et al., 2006). Deregulation of MYC has been shown to contribute to breast cancer development and progression and was associated with poor outcomes (Xu, Chen, & Olopade, 2010).

Considering the roles of STAT3, GATA3, FOS, and MYC in breast cancer, preferential binding of these transcription factors to the alternative T-allele of rs3789052 could identify it as the risk-associated allele. However, results from in-vitro analyses conducted in this work did not allow identifying the protein binding to the alternative allele of the candidate rSNP. The performed assays discarded STAT3 and GATA3 as the involved proteins, but further studies should be conducted to assess the binding of FOS or MYC.

Concerning the 16q23.2 locus, little is known about the proteins encoded by the *C16orf46* and *GCSH* genes. To date, *C16orf46* has not been associated with cancer, but overexpression of GCSH protein was found in breast cancer tissue and breast cancer cell lines (Adamus, et al., 2018). GCSH is one of the four mitochondrial proteins that constitute the GCS multi-enzyme system that catalyses the degradation of glycine. Enzymes of glycine metabolism tend to be overexpressed in tumours since they allow continuous growth of cancer cells (Zhang, et al., 2012; Sun, Kim, Jung, & Koo, 2016).

Fourteen daeSNPs were found to be in moderate to strong LD with the GWAS risk variant rs13329835 in the 1q23.2 locus. The variant rs74878296 was the only SNP found in high LD with the daeSNP rs12444974 (r^2=0.82 and D'=0.95), for which all of the heterozygous individuals analysed expressed exclusively the alternative A-allele. In-vitro assays identified preferential protein binding to the reference C-allele of rs74878296. Although PWM predictions identified AP-1 as the probable transcription factor involved in that interaction, results from EMSAs performed with c-Fos antibody showed that this was not the protein involved. However, given the fact AP-1 is a heterodimer, different affinities are observed for different proteins that interact for AP-1 formation. Thus, as previously observed for the candidate causal variant in the 1q32.1 locus, more assays would be needed to unveil the identity of the protein binding to the DNA.

Relatively to the 17q22 locus, our work has revealed the power of integrating allelic expression data in cancer risk studies. By inspecting the distribution of AE ratios in normal tissue samples -

breast and blood - we showed that all three genes in the locus are *cis*-regulated by genetic variants but that *STXBP4* and *COX11* are the most likely target genes involved in the risk of breast cancer. Additionally, we present a novel approach to identifying risk - case-control association analysis using AE ratios, which proved robust when multiple cis-regulatory variants are involved in a complex risk genetic structure. Finally, the estimated effect sizes we report are large (detected in rs2628315) to medium (detected in rs17817901) and are independent of the sample size, unlike p-values.

Our findings undoubtedly confirm that cis-regulatory variation in locus 17q22 is involved in the risk of BC. Previous studies have pointed towards all three genes as candidate target genes, from fine mapping exercises (Darabi, et al., 2016), to chromatin conformation analysis (Baxter, et al., 2018), and studies integrating allelic expression data (the INQUIST algorithm indicates *STXBP4* and *COX11* as target genes at level 2 of confidence) (Fachal, et al., 2020). Here, we provide novel evidence that AE is associated with risk, the most direct indication of cis-regulatory variants control. On the one hand, we show that transcribed variants in all three genes show differential allelic expression and are in strong to complete LD with the lead-variant for risk in this locus; notably, all tSNPs with equimolar allelic expression were in weak to no LD with the risk lead-variant. On the other hand, we establish a direct association between differential allelic expression and disease risk.

The most significant association with breast cancer we found was for the AE ratios measured at rs2628315, in an intron of *STXBP4*. This variant is in complete LD with rs2787486, the strongest risk association reported in this locus (OR =0.92; 95%CI: 0.90–0.94; p-value =8.96×10^{-15}) (Darabi, et al., 2016). We found that the G- rs2628315 allele, proxy to the risk C- rs2787486 allele, is 1.5-fold more expressed in cases. Moreover, as *STXBP4* is lowly expressed in breast tissue, this result suggests that the G- rs2628315 allele is extremely lowly expressed in healthy tissue, and once it is upregulated, it increases the risk of cancer, which suggests an oncogenic role for *STXBP4* in breast cancer. This gene encodes the protein STXBP4 (Syntaxin Binding Protein 4) involved in glucose metabolism, vesicle, and insulin transport. Although there is no data on breast cancer, it binds ΔNp63 in lung cancer (N-terminally truncated isoform of p63), thus preventing its proteolysis, promoting growth, and blocking cell differentiation, in line with the oncogenic role that our data supports (Rokudai, et al., 2018; Bilguun, et al., 2020).

Moreover, we found an association between AE ratios measured at rs17817901, located in a genomic region shared by *TOM1L1* and *COX11,* and breast cancer. Patient samples more often preferentially expressed the reference A- rs17817901 allele, which is frequently linked to the

risk-associated C- rs2787486 allele. Because we found no association for the AE ratios measured at rs9899602, a variant mapping exclusively to the *TOM1L1* sequence, we believe that the association detected at rs17817901 is mostly due to cis-regulatory variation acting on *COX11*. rs17817901 is in strong LD with the risk lead-variant rs2787486 (r^2 = 0.74) but is in even stronger LD (r^2 = 0.85) with a previously associated variant rs6504950 (OR =0.95; 95%CI: 0.92–0.97; p-value =1.4 $\times 10^{-8}$) (Ahmed, et al., 2009). Like the findings for rs2628315 in *STXBP4*, we observed a shift from the controls preferentially expressing the protective G- rs17817901 allele to the patients preferentially expressing the risk A- rs17817901 allele. On average, we found that the patients express the risk-associated A-rs17817901 allele 2-fold more than the controls. Additionally, GTEx data shows that *COX11* is a highly expressed gene in breast tissue (higher than *STXBP4*). These data suggest that although *COX11* is already highly expressed, there could be an oncogenic advantage to have its expression further upregulated. *COX11* encodes for a mitochondrial membrane protein crucial for the assembly of an active cytochrome c oxidase complex, which in turn is linked to the metabolic changes that accompany tumour development (Krieg, et al., 2004; Carr, Maxfield, Horng, & Winge, 2005), supporting its oncogenic role.

The genetic control of expression is partially shared by tissues, particularly cis-QTLs, albeit resulting in different effect sizes in various tissues (GTEX Consortium, 2020; Kwan, et al., 2009; Adoue, et al., 2014). Our previous work established an overlap between cis-regulation of breast cancer genes in breast tissue and blood (Maia, et al., 2009). Here we sought to verify whether the risk association found for AE ratios measured in breast tissue were valid in blood, as the testing of this tissue in a future clinical setting greatly facilitates the translation of these results. We found a similar profile of association of AE ratios measured at rs2628315 and rs17817901; the effects were similar in direction and size. This result opens the possibility of carrying out a future genome-wide study for identifying other breast cancer risk-associated daeSNP in blood.

The use of normalized AE ratio distributions confers robustness to rSNP mapping purposes, as it isolates the effect of cis-regulatory variation (Pastinen & Hudson, 2004). Particularly in the case of rs17817901, we observed a very marked shift of the normalized AE ratio distribution from the equimolar allelic expression. This pattern indicates complete LD between the daeSNP and the variant(s) controlling expression and led us to identify two strong candidates rSNPs. The first is rs1781901 itself, predicted to alter the binding of the oncogenic miRNA hsa-miR-194-5p at the 3' UTR of *COX11*. We established that hsa-miR-194-5p binds preferentially to the alternative C-allele, corresponding to the protective haplotype. Hence, upregulation of the hsa-miR-194-5p expression inhibits the expression of the protective allele, leading to an increased risk of breast cancer. The second candidate rSNP is rs8066588, which overlays an active enhancer associated

with robust transcription and a DNaseI hypersensitive site in mammary cells. We found that rs8066588 modulates the TCF3 transcription factor binding, a positive factor of gene expression; TCF3 binds preferentially to the reference C-allele of rs8066588, upregulating the expression of the risk-associated allele. Inclusion of these candidate causal rSNPs and risk-daeSNPs in haplotype association studies will further validate our findings.

AE ratios have been used to detect association with disease, but not as a quantitative variable to compare cases and control (Valle, et al., 2008). Instead, Valle and colleagues calculated normalized ratios and set a cut-off to define samples with and without differential allelic expression, upon which they tested for differences in proportions in the two populations (Valle, et al., 2008). A caveat of this approach is that we and others have shown that the distribution of AE ratios differs between genes. Hence establishing a universal cut-off is difficult; statistically, it is less powerful than using a quantitative phenotype (AE ratios) to compare two groups. One further advantage of our approach is that we measured the effect size using the Hedges' *g*, a standardized mean difference method, independent of the sample size. Increasing the sample size will be an asset in future studies, but it should only tighten the 95% CI of the estimated effect size, not disprove it. However, one limitation of our approach is transversal to all studies of AE: only individuals heterozygous for the transcribed variants used for quantification are informative. Compared to the classical association studies using genotype frequencies for individual SNPs, besides the increased statistical power of using a quantitative phenotype, AE ratios report simultaneously on the effect of all the variants regulating a gene. As these rSNPs are not necessarily in complete/strong LD with each other, this is of particular importance in regions with more complex genetic architectures. This approach now requires testing in a genome-wide setting to confirm its full potential. Finally, using this approach, we identify the mechanisms underlying risk and the target genes, which has been challenging in post-GWAS studies.

In summary, our work shows that all genes in the risk locus 17q22 are under the control of cis-regulatory variants, supports that *STXBP4* and *COX11* are the most likely target genes of the risk-variants identified in previous GWAS, and establishes that AE ratios at two daeSNPs are strongly associated with risk of breast cancer. Our work also indicates that the disease is associated with a change of preferential expression from the protective allele (healthy controls) to the risk allele (patient samples) in both genes. Overall, we present a novel approach to studying the risk of cancer, applicable to other complex diseases, using AE expression ratios as a quantifiable phenotype in case-control studies.

Considering the study on allelic expression in *PIK3CA* in breast cancer, 273 candidate rSNPs acting on that gene were identified after mapping of daeSNPs and their proxies (Correia, et al., 2021). One strong candidate rSNP, rs2699887, was significantly associated with the DAE detected at rs12488074 and is an eQTL for PIK3CA in tumours from METABRIC, with the alternative T-allele associated with higher expression. Analysis of PWM suggested that NF-YA could have preferential binding affinity to the alternative T-allele, which was confirmed with in-vitro functional analyses by EMSAs. Thus, rs2699887 is possibly the main cis-regulatory variant controlling the expression of PIK3CA in the breast by altering the binding site of NF-YA transcription factor at the promoter of *PIK3CA*. NF-YA, is one of the three subunits of NF-Y, a ubiquitously expressed TF with high specificity to the regulatory element CCAAT box (de Silvio, Imbriano, & Mantovani, 1999). NF-YA was reported to function as an oncogene or suppressor affecting cell proliferation, metastasis, and tumorigenicity (De Amicis, et al., 2011; Belluti, et al., 2013; Dolfini & Mantovani, 2013; Dai, et al., 2015). Levels of NF-YA mRNA, but not NF-YB nor NF-YC, were shown to be increased in breast cancer cells compared to normal controls (Dolfini, Andrioletti, & Mantovani, 2019). That study also showed that a switch between the two isoforms of NF-YA occurs from normal to tumour cells. Therefore, binding of NF-YA to the alternative T-allele of causal variant rs2699887 is likely to increase expression of PIK3CA, which was also supported by the higher PIK3CA expression observed for heterozygotes comparing with the women homozygous for the reference allele (Correia, et al., 2021).

Further in-vivo analyses by ChIP did not confirm the binding of NF-YA in T47D and SUM159 breast cancer cells. This could be explained by the NF-YA antibody not being able to bind its target protein as this is obstructed by RNAPII (RNA polymerase II). Subsequent studies including exposure of cells to α-amanitin, a well-known specific inhibitor of RNAPII (Cochet-Meilhac & Chambon, 1974; C., 1982), can be conducted to test this hypothesis.

Chapter V

Pipeline Benchmarking for AE Analysis from RNA-Seq Data

5.1. Abstract

Common genetic variants are likely to play an essential role in breast cancer risk. Allelic expression (AE) is a powerful method to better understand how genetic variation can control gene expression and identify single nucleotide polymorphisms (SNPs) with cis-regulatory potential. Advances in high-throughput RNA-sequencing (RNA-seq) facilitate the quantification of AE by assessing single nucleotide variant counts from both alleles in heterozygous individuals. As an accurate measurement of AE requires an optimally suited RNA-seq data analysis and no single pipeline can be applied to all cases, this work intends to perform a comprehensive comparison of variant calling pipelines for precise AE quantification using RNA-seq data.

This study aimed the systematic comparison of forty-two variant calling pipelines for AE analysis, combining different trimming algorithms, alignment, and variant calling tools, using RNA-seq data from the sample NA12878 (from 1000 Genomes Project) as a gold standard, and one sample of normal breast tissue from an ongoing study on differential AE (DAE).

Results from this work showed that: the trimming algorithm Trimmomatic (most efficient removal of adapters and the highest percentage of low-quality bases); the aligner GSNAP (albeit more time and computationally consuming, mapped a higher number of properly paired reads, and identified more true positives, and less false negatives); and the variant caller GATK, which identified more variants, including true positives. The results obtained in this work showed significant divergence between the two variant callers, and suggest that variants called by both variant callers should be used in subsequent analysis.

The present study provides guidelines for RNA-seq analysis suitable for accurate detection of differentially expressed alleles and mapping of potential regulatory SNPs.

5.2. Introduction

Allelic expression (AE) analysis has been used as an alternative approach to expression quantitative trait loci (eQTL) analysis for mapping variants associated with altered gene expression (Yan, Yuan, Velculescu, Vogelstein, & Kinzler, 2002; Serre, et al., 2008; Milani, et al., 2009; Verlaan, et al., 2009; Heap, et al., 2010). AE quantifies the relative abundance of the two alleles of a transcript at heterozygous sites in a diploid individual. Since both alleles are expressed in the same cellular environment, unlike eQTL analysis that assesses total expression levels, AE analysis detects the effects of cis-regulatory variants minimizing the variation between samples originated by trans-acting factors (Pastinen & Hudson, 2004; Buckland, 2006; Adoue, et al., 2014). The imbalance between levels of transcripts harbouring each allele of the variant can be assessed by counting reads from RNA sequencing (RNA-seq) data (Verlaan, et al., 2009). In addition, the identification of genomic variants through RNA-seq data analysis detects functionally important variants. However, although AE is directly measured using RNA-seq, there is no single analysis pipeline optimally suited for it. AE quantification from RNA-seq data presents several challenges.

A general problem of high-throughput sequencing is the presence of adapter sequences that are ligated to the DNA or cDNA fragments of interest during library preparation. This happens when the fragments are shorter than the specified sequencing length. Accurate pre-processing of sequencing data before the alignment is of great importance to increase the quality and reliability of downstream analyses (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013). Removing adapters and low-quality bases improve correct mapping of the reads and influence variant calling. To address this question several trimming tools were developed. Nevertheless, the suitability of a trimming algorithm is highly dependent on the dataset and downstream analyses.

The most difficult caveat when analysing AE from RNA-seq data is the allelic mapping bias (Degner, et al., 2009; Stevenson, Coolon, & Wittkopp, 2013). Mapping the reads to a reference genome confer a lower probability of correct alignment of the reads carrying the non-reference allele. Reads with the alternative allele will display a higher number of mismatches and may fail to map uniquely. Several strategies can be used to reduce the effect of mapping bias on AE estimation.

One approach is the allele-specific software WASP (van de Geijn, McVicker, Gilad, & Pritchard, 2015), which can be used after mapping the reads with a selected aligner. However, when WASP is used, unknown variants in the genome are not considered and some bias is still observed (Castel, Levy-Moonshine, Mohammadi, Banks, & Lappalainen, 2015). In addition, WASP is time-

consuming and its stringent requirements exclude numerous reads (Miao, Alvarez, Pajukanta, & Ko, 2018). Another possible approach to overcome mapping bias is to align the reads to personalised genomes, by constructing a diploid genome where genotype information of the sample is incorporated into the reference genome (Rozowsky, et al., 2011). This requires prior information on the personal genomic variants. The variant-aware Genomic Short-read Nucleotide Alignment Program (GSNAP, Wu & Nacu, 2010) provides an alternative solution. This alignment algorithm inputs information on known SNPs and aligns the reads considering all possible combinations of reference and alternative alleles. The two approaches mentioned lastly are computationally heavier but are the most comprehensive and provide more bias reduction (Castel, Levy-Moonshine, Mohammadi, Banks, & Lappalainen, 2015).

Calling variants from RNA-seq data to obtain information from heterozygous sites on which AE analysis relies on, also represents a challenge (Piskol, Ramaswami, & Li, 2013). Several variant callers based on different algorithms have been developed, from which Genome Analysis Toolkit (GATK) (McKenna, et al., 2010) and SAMtools (Li, et al., 2009) are two of the most widely used. Previous studies have shown that there can be substantial differences in the sets of variants called with different pipelines (O'Rawe, et al., 2013; Yu & Sun, 2013; Pirooznia, et al., 2014). Accurate variant calls are critical for a correct assessment of AE and obtaining information of biological relevance. However, previous studies comparing the performance of variant callers used DNA-sequencing data (Liu, Han, Wang, Gelernter, & Yang, 2013; O'Rawe, et al., 2013; Yu & Sun, 2013; Pirooznia, et al., 2014; Hwang, Kim, Lee, & Marcotte, 2015; Ni, et al., 2015; Sandmann, et al., 2017) and the available variant callers still do not have optimised pipelines for RNA-seq as for DNA-seq data.

Although many tools are described to improve AE analysis (Rozowsky, et al., 2011; Piskol, Ramaswami, & Li, 2013; Soderlund, Nelson, & Goff, 2014; Castel, Levy-Moonshine, Mohammadi, Banks, & Lappalainen, 2015; van de Geijn, McVicker, Gilad, & Pritchard, 2015; Miao, Alvarez, Pajukanta, & Ko, 2018) to our knowledge there is not a comprehensive comparison of the possible pipelines combining open-source tools for all the steps of a typical workflow for high-throughput sequencing data analysis. Comparison studies are even sparser when considering the detection of variants from RNA-seq real data, and including a variant-aware aligner for the mapping step.

In this study, we performed the comparison of forty-two pipelines for variant calling using RNA-seq data from two samples: the NA12878, which has the only gold standard variant genotype data set publicly available; and one sample of normal breast tissue, which represents the real sample type that will be used for subsequent case-control studies to identify common variants

associated with breast cancer risk using AE analysis. The pipelines consisted of a combination of seven trimming tools – AdapterRemoval2 (Schubert, Lindgreen, & Orlando, 2016), Cutadapt (Martin, 2011), fastq-mcf from ea-utils (Aronesty, 2013), Flexbar (Dodt, Roehr, Ahmed, & Dieterich, 2012), SeqPurge (Sturm, Schroeder, & Bauer, 2016), Skewer (Jiang, Lei, Ding, & Zhu, 2014) and Trimmomatic (Bolger, Lohse, & Usadel, 2014) –, two aligners that map reads across splice junctions – the SNP-tolerant aligner GSNAP (Wu & Nacu, 2010) and STAR (Dobin, et al., 2013) –, and two variant callers – GATK (McKenna, et al., 2010) and SAMtools (Li H., 2011) –. Overall, the choice of tools was based on performance in previous studies, popularity, and relevancy to our data and AE analysis. Trimming algorithms were also chosen based on their ability to perform adapter trimming as well as filtering of low-quality bases and entire reads. Our results provide useful guidelines for similar data analyses from RNA-seq experiments.

5.3. Materials and Methods

5.3.1. Experiment Environment

All analyses were performed in a server with 2×20 Intel® Xeon® E5-2660 v2 @ 2.20GHz CPUs, 100 GB of RAM and, 2.5 T of disk storage, with Ubuntu 18.04.1 LTS (GNU/Linux 4.15.0-29-generic x86_64).

5.3.2. Datasets

Pipelines were evaluated on RNA-seq data from one gold standard sample and one sample of breast tissue from a healthy woman.

Paired RNA-seq reads (75 bp) from the GM12878 human lymphoblastoid cell line (1000 Genomes Project Consortium, 2010) were obtained from the European Nucleotide Archive (ENA) database (https://www.ebi.ac.uk/ena/browser/view/SRR1258218). This experiment (experiment accession SRX523286, run accession SRR1258218, sample accession SAMN02731489 or GSM1372331) included sequencing of 25,933,924 reads from cDNA libraries in an Illumina HiSeq 2000 instrument. Briefly, total RNA was extracted from LCLs with Trizol and its quality was assessed with the Agilent Bioanalyzer 2100. cDNA libraries were obtained from 1 µg of polyA purified mRNA using the TruSeq Preparation Kit and sequenced.

The breast tissue sample was collected in the scope of a previous study (Maia, et al., 2012) from the Tissue Bank at Addenbrooke's Hospital (Cambridge, UK), with approval from the Addenbrooke's Hospital Local Research Ethics Committee (REC reference 06/Q0108/221).

Total RNA from the breast tissue sample was previously obtained (Maia, et al., 2012), and RNA integrity and concentration were assessed with the Experion™ Automated Electrophoresis Station (Bio-Rad, USA). Preparation of mRNA libraries and sequencing was performed by Eurofins Genomics (Ebersberg, Germany). First, poly-A mRNA was isolated from 700 ng of total RNA, fragmented and subsequently used to obtain cDNA libraries using the Illumina TruSeq Stranded mRNA kit (Illumina Inc., San Diego, USA). Sequencing was then performed on an Illumina HiSeq 2500 instrument for 2 × 100 cycles (paired-end reads) using HiSeq Sequencing by Synthesis (SBS) v4 chemistry, generating 135,803,746 reads.

5.3.3. RNA-seq Pre-Processing

Before pre-processing, the quality of the data was assessed with the FastQC v0.11.7 application (Andrews, 2010) to ensure that the raw data was suitable for further analyses.

Trimming algorithms included in this work were chosen based on the performance in previous studies, popularity, and relevance to our data and AE analysis. In addition, the ability to perform adapter trimming as well as filtering of low-quality bases and entire reads were also considered. Seven different algorithms for pre-processing the data were compared, including AdapterRemoval v2.1.7 (Schubert, Lindgreen, & Orlando, 2016), Cutadapt v.1.18 (Martin, 2011), fasqt-mcf v.1.0.5 from ea-utils (Aronesty, 2013), Flexbar v.3.0.3 (Dodt, Roehr, Ahmed, & Dieterich, 2012), SeqPurge v.2018_06 (Sturm, Schroeder, & Bauer, 2016), Skewer v.0.2.2 (Jiang, Lei, Ding, & Zhu, 2014), and Trimmomatic v.0.36 (Bolger, Lohse, & Usadel, 2014) (Fig 1).

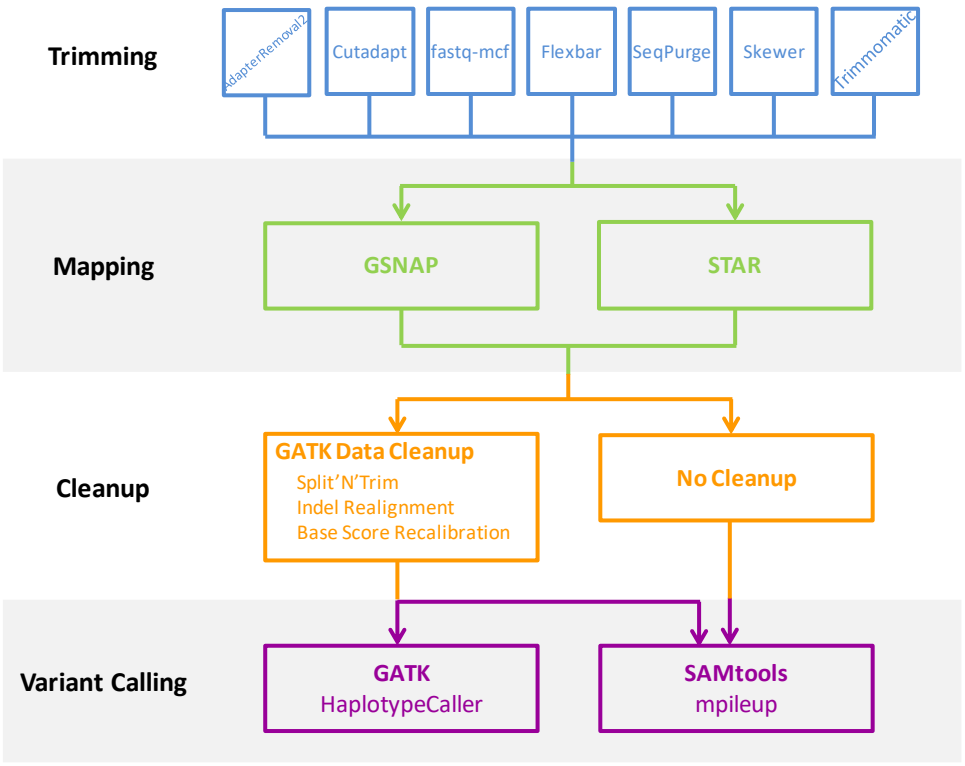


Figure 5.1. Summary diagram representing the 42 compared variant calling pipelines.

All tools can remove contaminant adapter sequences from the 3' ends of single-end (SE) or paired-end (PE) reads and perform trimming of low-quality bases and filtering of reads based on the resulting length. However, AdapterRemoval, Cudatapt, fastq-mcf, and Trimmomatic support trimming of low-quality bases from both 3' and 5' ends. Flexbar, SeqPurge, and Skewer only allow quality trimming for the 3' end.

All methods were run with similar parameters: single thread mode was used, with exception of fastq-mcf, for which the option is not available; besides adapters, low-quality bases (Phred quality score below 25, Q<25) were also trimmed; only reads with a minimum length of 25 bp after trimming were kept. Processing time and memory usage of all trimming algorithms were compared, as well as the number of adapters and low-quality bases remaining after trimming.

5.3.4. RNA-seq Mapping

Reads were aligned to the human reference genome GRCh38 with two different algorithms, including the Genomic Short-read Nucleotide Alignment Program (GSNAP v.2017-11-15, Wu & Nacu, 2010), and the Spliced Transcripts Alignment to a Reference (STAR v.2.6.0, Dobin, et al., 2013). For GSNAP alignments, the human dbSNP Build 150 (from Database of Single Nucleotide Polymorphisms) obtained from UCSC was used. For alignments performed with STAR, the 2-pass method was used. Both aligners take compressed fastq files as input and allow parallel processing. However, STAR output files are in BAM (Binary Alignment Map) format, and GSNAP outputs are in SAM (Sequence Alignment Map) format. For both algorithms, alignments were performed with multi-threading, using 10 threads.

The two algorithms can identify splicing within short reads, and specifically address several challenges of RNA-seq data alignment. Nevertheless, unlike STAR, GSNAP provides a SNP-tolerant alignment, in which minor alleles are not treated as mismatches to the reference genome, and thereby genotypes carrying those alleles are not penalized. This feature may be of high importance in the accurate measurement of allelic expression.

5.3.5. Variant Calling and Filtering

Alignments performed with the two tools were used for variant calling with Genome Analysis ToolKit (GATK v.3.8 and v.4.0.4.0, McKenna, et al., 2010), which uses some utilities from Picard v.2.18.3 (http://broadinstitute.github.io/picard), and with the SAMtools software package v.1.7 (Li H., 2011), including its two key components samtools and bcftools, which use htslib internally (library for reading/writing high-throughput sequencing data). Both variant callers use the Bayesian method for variant identification.

Until now GATK and SAMtools software do not have an optimized workflow for variant discovery from RNA-seq data. However, GATK provides some best practices recommendations for calling variants on RNA-seq data, to be run individually for each sample, based on their DNA-seq oriented workflow.

For whole genome sequencing (WGS) and whole exome sequencing (WES), SAMtools recommend improving the data with some GATK utilities, including realignment around indels and base recalibration before variant calling. But no specifications exist regarding RNA-seq data. Hence, for variant calling with SAMtools, two approaches were used: one in which variants were

called directly from the BAM file after sorting and indexing; and another in which variant calling was performed with the BAM file after processing with GATK utilities.

For both variant callers, GSNAP alignments in SAM format were initially converted into BAM files. Next, both GSNAP and STAR outputs were sorted and indexed. The subsequent step in a WGS or WES is to remove duplicates. Duplicates are marked as such if, reads 1 and 2 from 2 different pairs have the same starting position (the alignments have the same start and end) and have the same CIGAR string (the string that describes how the read aligns to the reference). Elimination of duplicates from DNA-seq data is a common practice but needs to be carefully done in RNA-seq analysis. Although duplicates can be technical, they can also be biological if different copies of the same sequence were randomly selected. To observe transcripts with low levels of expression, it is common to over-sequence high expressed transcripts, possibly creating a large number of duplicates. Therefore, although duplicates can be marked in RNA-seq data, usually they are not removed, which is even more relevant when performing analysis of allelic expression. Hence, for the present work, duplicates were marked but were not eliminated in the proceeding steps.

After marking the duplicates, for the workflows including GATK utilities, inaccuracies in reads splicing during alignments were corrected with the SplitNCigarReads tool, which was developed specially for RNA-seq data analysis. Then, reads were realigned around indels with IndelRealigner and the initial base quality scores were corrected with BaseRecalibrator. Finally, variants were called with HaplotypeCaller from GATK or with the 'mpileup' command from SAMtools.

Variants called with GATK were hard filtered according to their recommendations, including FisherStrand (FS) higher than 30.0, and QualByDepth (QD) lower than 2.0.

Variants from SAMtools were filtered based on their average mapping quality (MQ) and their raw read depth (DP). Only variants with MQ>19.0 and DP>9.0 were kept.

Regarding the identification of SNPs, only the bi-allelic SNPs were used for subsequent analyses.

5.3.6. Pipelines Performance

For both samples data, systematic comparison of all pipelines included the comparison of: time and memory usage; the number of reads mapped in pair and the number of singletons (mapped reads with unmapped pair); percentage of properly paired reads (both reads of the pair are

mapped on the same chromosome, at a good expected distance with consistent strand directions) and percentage of uniquely mapped reads (reads in proper pair that map to a single location in the genome); the number of SNPs and indels; and transition/transversion (Ti/Tv) ratios. Since transitions (interchanges of purines or pyrimidines) tend to occur more frequently in the genome than transversions (interchanges of a purine for pyrimidine bases), and sequencing errors tend to generate more transversions, the Ti/Tv ratio can be used to estimate the frequency of possible sequencing errors in SNP calling.

Finally, precision [TP/(TP+FP)], sensitivity [TP/(TP+FN] and the F-measure (the harmonic mean of precision and sensitivity with a value between 0, for the worst, and 1, for the perfect F-measure) were determined for the gold standard sample, using the high-confidence variant calls and bed files downloaded from the GIAB project (https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh38/) and the vcfeval command from RTG tools.

5.4. Results

5.4.1. RNA-Seq Pre-Processing

Initial quality assessment of the raw data performed with FastQC v0.11.7 application (Andrews, 2010) showed good quality overall.

For the performance comparison of the seven trimming tools, adapters and low-quality bases (Q<25) were trimmed, and reads shorter than 25 bp after trimming were removed.

Processing times required by each tool are presented in Table 5.1. Flexbar was the slowest trimming tool and fastq-mcf (the only tool for which multi-threading mode could not be disabled) was the fastest. Considering the computational memory, Skewer showed the lowest usage (3.2 MB), and Trimmomatic required the highest (1133.0 MB). Memory peaked at 13.0, 14.7, 352.2, 24.7 and 13.7 MB with AdapterRemoval2, Cutadapt, fastq-mcf, Flexbar and SeqPurge, respectively.

Table 5.1. Processing time (minutes) required for each pipeline. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

				Variant Calling					
	Trimming	Mapping		GATK		SAMtools		SAMtools after GATK data cleanup	
		GSNAP	STAR	GSNAP	STAR	GSNAP	STAR	GSNAP	STAR
Raw reads	N/A	630	20	1,734	2,312	183	141	1,084	1,779
Adapter removal v.2.1.7	15	506	17	1,844	1,779	175	130	1,289	1,287
Cutadapt v.1.18	21	457	14	1,649	1,543	168	119	1,224	1,109
fastq-mcf v.1.05	3	519	15	1,819	1,842	174	130	1,261	1,356
Flexbar v.3.0.3	25	506	15	1,630	1,743	120	127	1,148	1,319
SeqPurge v.2018_06	12	530	17	1,866	1,765	173	127	1,338	1,303
Skewer v.0.2.2	7	523	22	1,899	1,911	173	128	1,335	1,423
Trimmomatic v.0.36	23	467	19	1,754	1,756	173	127	1,288	1,309

(B)

				Variant Calling					
	Trimming	Mapping		GATK		SAMtools		SAMtools after GATK data cleanup	
		GSNAP	STAR	GSNAP	STAR	GSNAP	STAR	GSNAP	STAR
Raw reads	N/A	1,139	59	5,961	6,654	314	356	1,893	2,718
Adapter removal v.2.1.7	73	866	39	6,879	6,207	453	424	2,234	2,606
Cutadapt v.1.18	118	930	37	6,551	5,687	394	451	2,251	2,274
fastq-mcf v.1.05	8	881	42	6,798	6,378	466	451	2,049	2,441
Flexbar v.3.0.3	128	960	49	6,742	6,228	391	438	2,207	2,485
SeqPurge v.2018_06	68	1,220	48	6,890	6,702	415	443	2,172	2,692
Skewer v.0.2.2	33	1,000	43	7,076	6,807	400	448	2,275	2,689
Trimmomatic v.0.36	79	1,465	49	6,449	6,322	391	445	2,122	2,452

Table 5.2 shows the number of adapters, the total percentage of bases, and the percentage of low-quality bases (quality Phred score below 25, Q<25) that remained after trimming. Initially, there were 11,798 and 26,851 adapters present in the raw reads from the gold standard and the breast tissue samples, respectively. With exception of fastq-mcf, all trimming tools showed a good performance in removing the adapters, but only Trimmomatic completely removed them in both samples. Although Cutadapt did not trim all the existing adapters, it showed the higher stringency considering the number of trimmed bases. Since trimming algorithms only trim low-quality bases at the ends of the reads, bases with Q<25 that are present in the middle of the read were kept. More low-quality bases were trimmed with Cutadapt and Trimmomatic in the gold standard and breast tissue data, respectively.

Table 5.2. Trimming results of the seven tools tested on raw reads. **(A)** gold standard sample. **(B)** breast tissue sample**.** The remaining adapters refer to the first 20-mer of the adapters. The number of bases refers to bases from read1 and read2.

(A)

	Remaining adapters	**% Remaining bases (total)**	**% Remaining bases with Q<25**
Raw reads	11,798	100.0	17.0
AdapterRemoval v.2.1.7	32	83.2	9.9
Cutadapt v.1.18	1	**75.9**	**7.4**
fastq-mcf v.1.05	10	79.8	8.8
Flexbar v.3.0.3	3	82.3	10.0
SeqPurge v.2018_06	1	82.6	9.7
Skewer v.0.2.2	98	83.8	10.1
Trimmomatic v.0.36	**0**	80.5	8.8

(B)

	Remaining adapters	**% Remaining bases (total)**	**% Remaining bases with Q<25**
Raw reads	26,851	100.0	3.5
AdapterRemoval v.2.1.7	16	97.2	1.8
Cutadapt v.1.18	1	**96.4**	1.7
fastq-mcf v.1.05	26,477	97.4	1.9
Flexbar v.3.0.3	9	97.2	1.9
SeqPurge v.2018_06	**0**	97.4	1.9
Skewer v.0.2.2	33	97.4	1.9
Trimmomatic v.0.36	**0**	96.5	**1.6**

5.4.2. RNA-Seq Mapping

After pre-processing, reads were aligned to the human reference genome GRCh38 with STAR v.2.6.0 and GSNAP v.2017-11-15. Both aligners can map reads across splice junctions, but GSNAP is a variant-aware aligner, which made mapping with GSNAP a more time-consuming process than mapping with STAR (Table 5.1).

Figures 5.2 and Table S5.1 show the number of reads mapped in pair and the number of singletons. GSNAP mapped a higher number of reads in pair regardless of the trimming tool used for pre-processing the reads. Although Trimmomatic was the second tool to exclude more reads, it was the one for which more reads were mapped in pair with GSNAP. The number of reads

mapped in pair with STAR was concurrent with the number of reads after trimming, meaning that the higher the number of trimmed reads the higher the number of mapped reads. GSNAP also aligned fewer singletons than STAR.

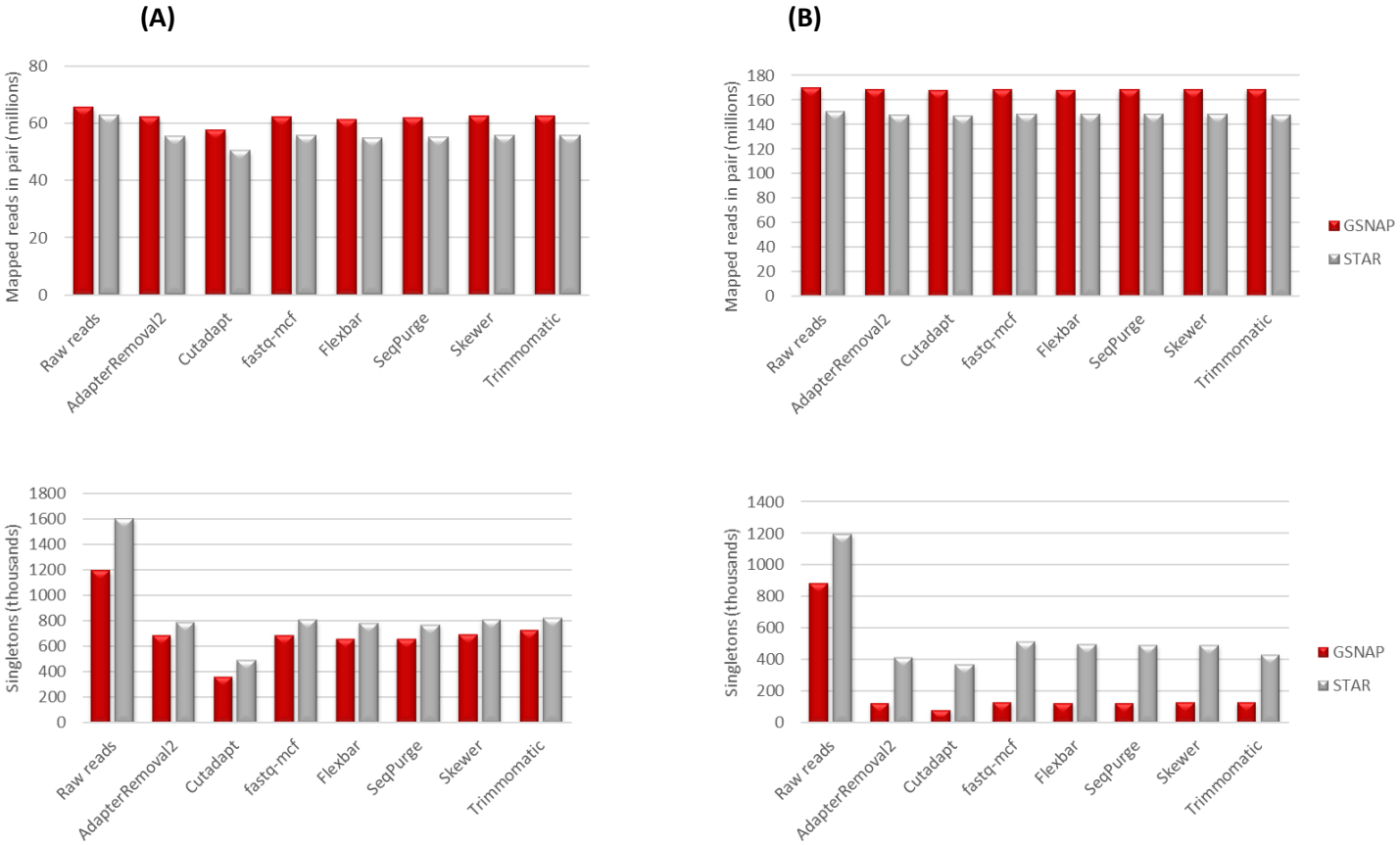


Figure 5.2. Mapped reads with GSNAP (red) and STAR (grey), after trimming with different tools. **(A)** gold standard sample. **(B)** breast tissue sample. Top: mapped reads in pair (millions). Bottom: singletons (mapped reads with unmapped pair) (thousands).

Regarding the percentage of properly paired reads and the uniquely mapped reads (Figure 5.3 and Table S5.2), GSNAP showed an equal or higher percentage of reads mapped in proper pairs for all the considered trimming methods. For each aligner, the highest number of properly paired reads was obtained with Cutadapt, which presented the same result as AdapterRemoval2, when considering STAR alignments. However, STAR showed more uniquely mapped reads regardless of the used trimming tool. Cutadapt was the algorithm that generated a higher percentage of uniquely mapped reads in both alignments, followed by Trimmomatic in breast tissue data and SeqPurge in gold standard data.

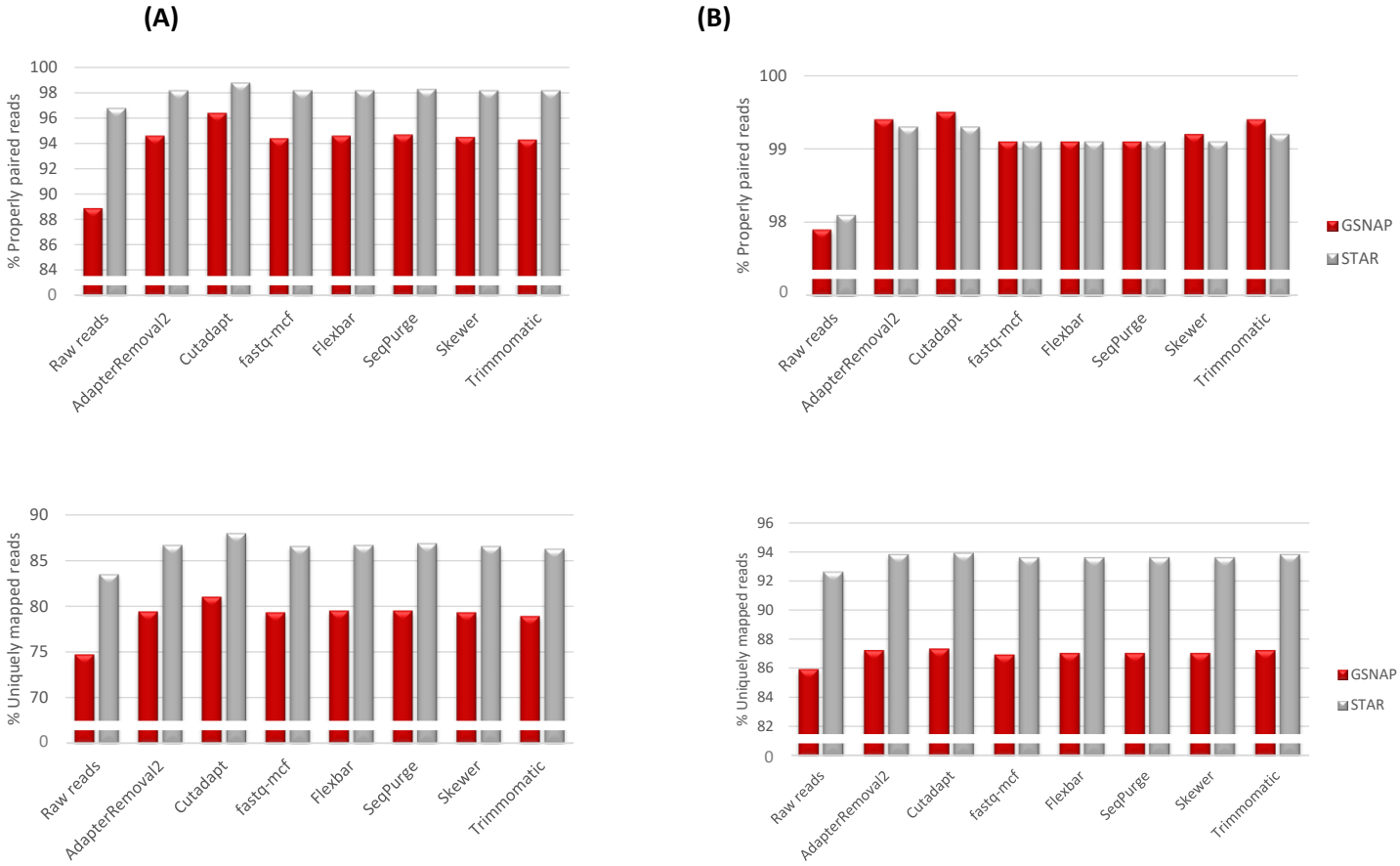


Figure 5.3. Percentage of properly paired reads (top) and percentage of uniquely mapped reads (bottom) obtained with GSNAP (red) and STAR (grey). **(A)** gold standard sample. **(B)** breast tissue sample.

5.4.3. Variant Calling and Filtering

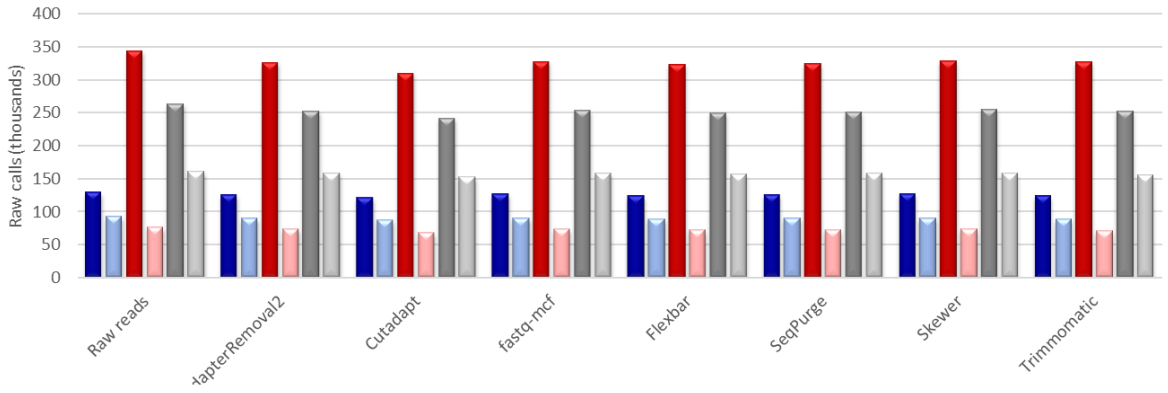
After variant calling and filtering, the first metric to be considered was the processing time required by each pipeline tested (Table 5.1). Given the less complex workflow, variant calling was faster using SAMtools than GATK. After mapping with GSNAP, variant calling with GATK and SAMtools was faster when trimming was performed with Trimmomatic for breast tissue data, and with Flexbar for the gold standard. Calling variants with STAR alignments was faster when reads were trimmed with Cutadapt (GATK) or AdapterRemoval2 (SAMtools) for breast tissue, and with Cutadapt (GATK and SAMtools) for the gold standard. When variant calling was performed with SAMtools after improvement with GATK utilities, time requirements increased relatively to not performing data cleanup but remained lower than the ones observed when variants were called by GATK.

Next, more variants were identified from GSANP alignments (Figures 5.4 and 5.5, and Tables S5.3-S5.4) regardless of the workflow used for variant calling. A lower number of variants were identified when pre-processing with Cutadapt, except for breast tissue data when calling was performed with SAMtools after improvement with GATK utilities. For both alignments, and both samples, when the GATK workflow was used, more variants were called after trimming the reads with Skewer. The same was verified for the gold standard data, after applying the SAMtools workflow in alignments from GSNAP and STAR. However, for the breast tissue data, a higher number of variants were called with SAMtools from reads trimmed with Trimmomatic or fastq-mcf, after GSNAP or STAR mapping, respectively.

After variant calling, variants were filtered according to the specifications previously described. The percentage of filtered variants was higher when trimming was not performed. Considering all pipelines, up to 2.0% or 4.3% more variants were filtered from raw data than from trimmed data from the gold standard or breast tissue, respectively. Results in Figure 5.5 and Tables and S5.3-S5.4 show that although filtering reduced the number of variants identified by each pipeline, it did not alter the proportion of variants identified before filtering. Thus, a higher number of calls were obtained when reads were mapped with GSNAP, and even more, when calling was performed with GATK. For the gold standard, many more variants were kept from GATK calling (up to 79% and 89% using GSNAP or STAR) than SAMtools (13% and 29%, respectively).

For breast tissue, the percentage of kept variants after filtering with GSNAP alignments (~63% and ~50% for GATK and SAMtools, respectively) was similar to those kept with STAR alignments (~67% and ~62%, respectively)

(A)



(B)

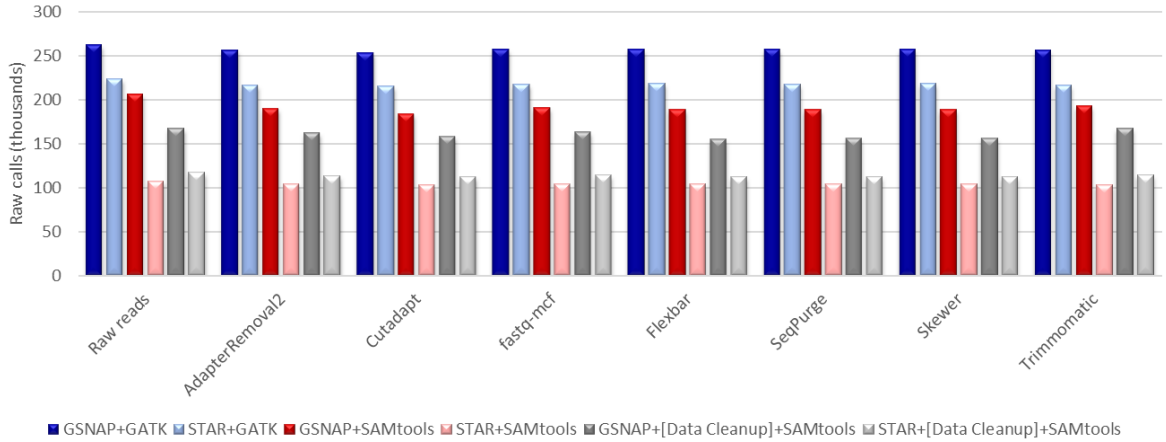
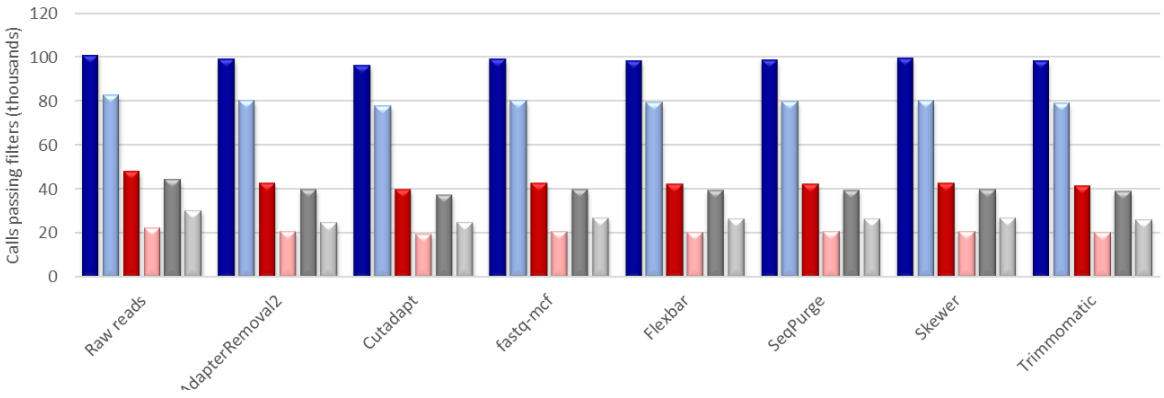


Figure 5.4. Number of raw variants (thousands) called with each pipeline. **(A)** gold standard sample. **(B)** breast tissue sample.

Using SAMtools after GATK data cleanup, decreased the number of detected variants, and more importantly, resulted in a higher percentage of variants failing to pass the filters (83.18-84.67% for GSNAP and 81.15-84.07% for STAR in gold standard, and 80.87-84.18% for GSNAP and 80.49-84.43% for STAR in breast tissue).

(A)



(B)

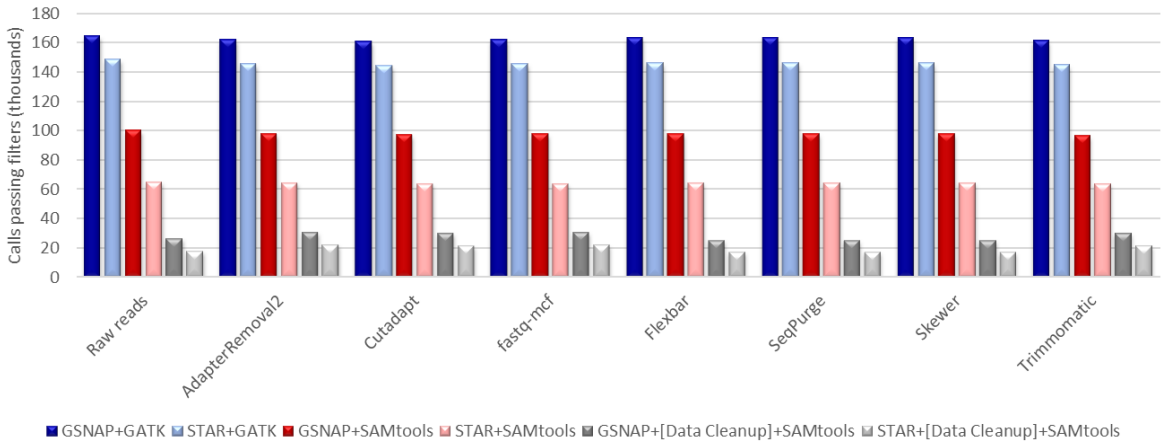


Figure 5.5. Number of variants (thousands) passing filters called with each pipeline. **(A)** gold standard sample. **(B)** breast tissue sample.

The overall concordance of variants (SNPs and indels) identified by GATK and SAMtools, obtained with the pipelines including each trimming tool, showed a maximum variation of 0.25% and 0.43% for breast tissue and gold standard data, respectively. Thus, the percentages of overall variants shown in Figure 5.6 refer to the mean of called variants using all trimming tools.

The overlap of filtered variants generated by each variant caller (Figure 5.6) demonstrated that a higher number of variants were shared when mapping was performed with GSNAP. This was expected given that more variants were identified by both callers when GSNAP was used. It was also evident that SAMtools had a higher fraction of shared variants. For GSNAP mappings, 71.12%-72.18% (gold standard) and 68.46%-69.40% (breast tissue) of variants from SAMtools

overlap variants from GATK, but less than half of GATK variants (28.68%-29.32% for gold standard, and 41.47%-41.74% for tissue sample) are common to SAMtools. This difference was larger after mapping with STAR, where 23.94%-24.53% of GATK variants and 97.00%-97.30% of SAMtools variants were shared for the gold standard data, and 35.91%-36.20% of GATK variants and 81.74%-81.83% of SAMtools variants were shared for the breast tissue data.

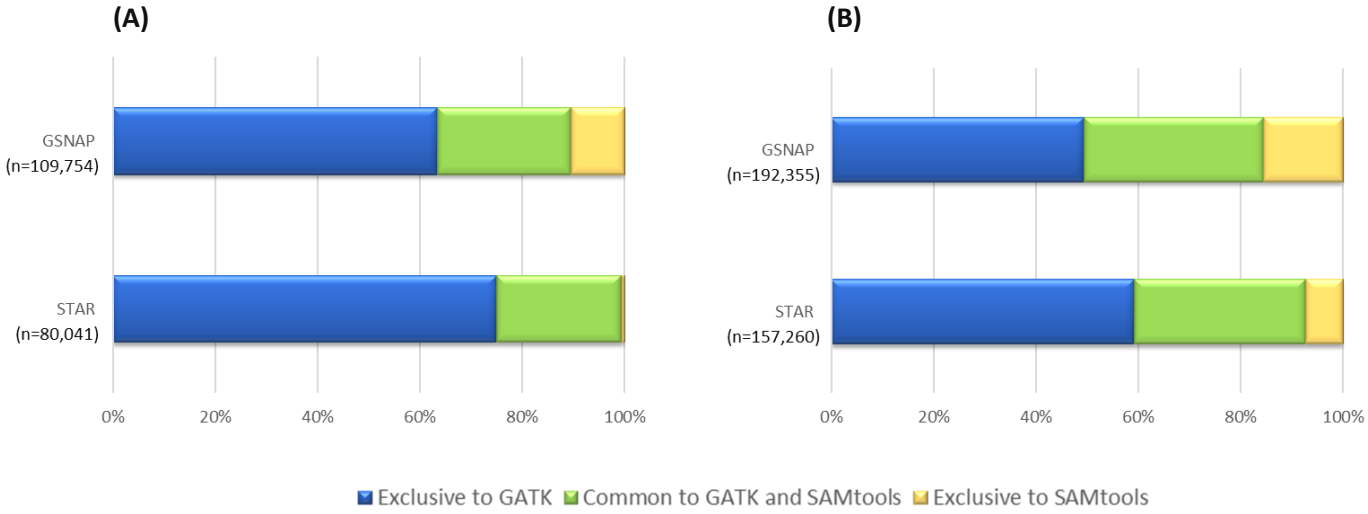


Figure 5.6. Overall variants (SNPs and indels) identified by GATK and SAMtools after mapping with GSNAP and STAR. **(A)** gold standard sample. **(B)** breast tissue sample. The represented number of variants refers to the mean of called variants using each trimming tool.

The same comparison was performed with variants called by SAMtools after data cleanup with GATK tools (Figure S5.1). Similar results were observed, with a lower percentage of GATK variants being shared considering results from tissue sample data (up to 11.19% after GSNAP mapping and up to 10.54% after STAR mapping). On the other hand, for that sample, around 59% (GSNAP) and 69% (STAR) of the variants identified by SAMtools after data cleanup were shared with GATK. These differences were smaller for variants called with SAMtools after data cleanup using the gold standard data. In this case, up to 67% (GSNAP) and 70% (STAR) of variants from GATK and SAMtools, respectively, were shared.

Subsequently, the number of known SNPs (dbSNPs) and novel SNPs identified with all the pipelines were compared (Figures 5.7 and 5.8, and Tables S5.5-S5.9). In Figures 5.7 and 5.8 is represented the mean of called variants using all trimming tools (between which the maximum variation was 1.46% considering the gold standard, and 0.61% considering the tissue sample).

Results obtained with GATK (Table S5.5) and SAMtools (Table S5.6) show that more SNPs, either known or novel, were identified after mapping with GSNAP (Figure 5.7) than with STAR (Figure 5.8). Regardless of the aligner used, a higher number of novel SNPs were identified with GATK (novel SNPs represented up to 20% and 35% of all the SNPs identified with GATK for the gold standard and the breast tissue sample, respectively). Known SNPs represented over 99.50% of the SNPs called by SAMtools, regardless of the aligner or the sample.

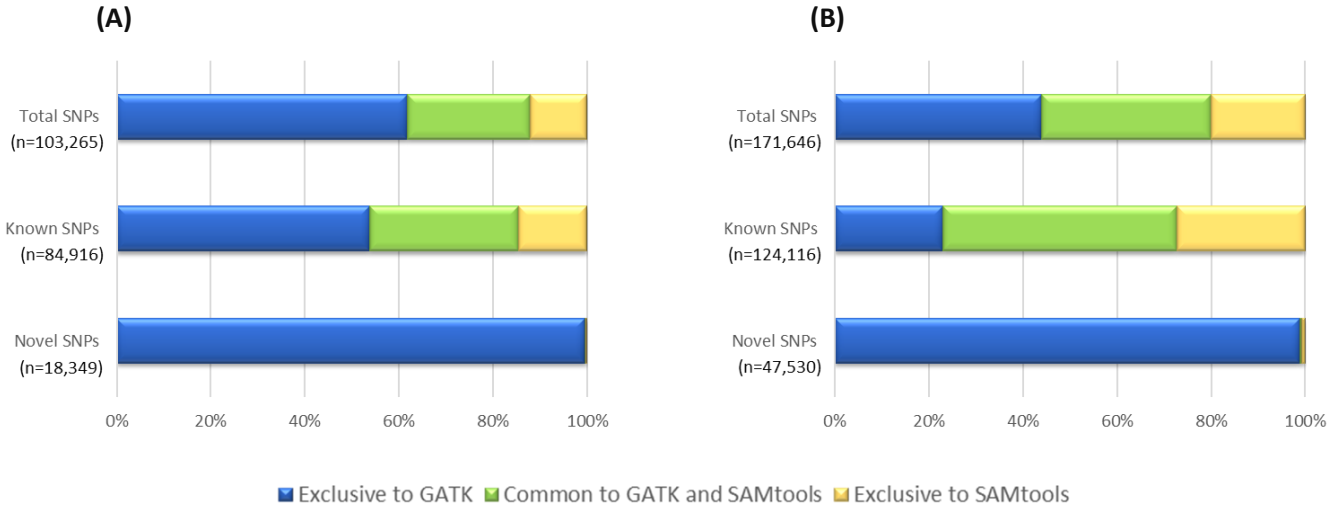


Figure 5.7. SNPs identified by GATK and SAMtools after mapping with GSNAP. **(A)** gold standard sample. **(B)** breast tissue sample. Total SNPs – all SNPs (known and novel). Known SNPs – SNPs found in dbSNP150. Novel SNPs – SNPs not found in dbSNP150. The represented number of SNPs refers to the mean of called variants using each trimming tool.

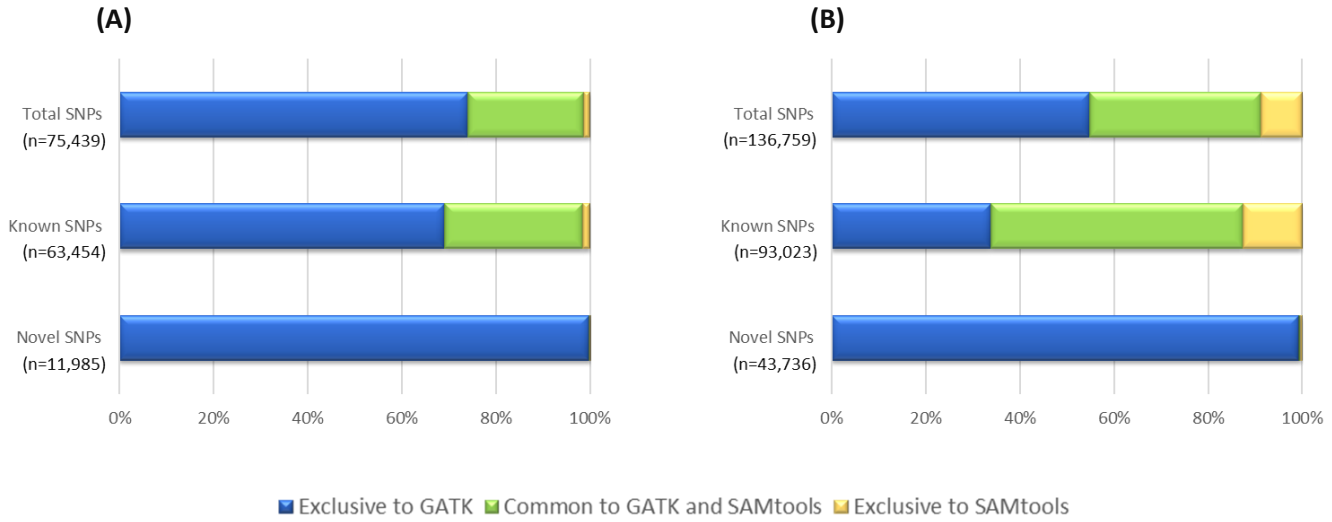


Figure 5.8. SNPs identified by GATK and SAMtools after mapping with STAR. **(A)** gold standard sample. **(B)** breast tissue sample. Total SNPs – all SNPs (known and novel). Known SNPs – SNPs found in dbSNP150. Novel SNPs – SNPs not found in dbSNP150. The represented number of SNPs refers to the mean of called variants using each trimming tool.

After overlapping the SNPs identified by both callers, more shared SNPs were obtained with GSNAP (Table S5.7). In this case, regarding results from the breast tissue data, and depending on whether the data were trimmed or not, 61,517 (using Cutadapt) to 62,439 (using raw reads) SNPs were detected by both GATK and SAMtools. This represents about 45% of the total SNPs called by GATK and 64% of those called by SAMtools. When mapping was performed with STAR, the lowest and highest number of overlapping SNPs was 49,621 (using Trimmomatic) and 50,344 (using raw reads), respectively. These corresponded to 39% of SNPs identified by GATK, and to 80% of the calls from SAMtools. For gold standard data aligned with GSNAP, percentages of common SNPs represented approximately 30% of total SNPs called by GATK and 64% of total SNPs from SAMtools. Considering common calls obtained after STAR alignments, those represented ~80% of total SNPs called by GATK, and 94% of total SNPs from SAMtools.

For breast tissue data, around 10,000 more SNPs were called by SAMtools after data cleanup with GATK, using GSNAP compared to STAR (Figure S5.2 and Table S5.8). Considering only common variants called by GATK and SAMtools after data cleanup, 2,000 more SNPs were identified with GSNAP (Table S5.9). For gold standard data, more SNPs were identified using this pipeline than for the breast tissue data.

Next, the Ti/Tv ratio was used to estimate the frequency of possible sequencing errors in SNP calling. The Ti/Tv ratios of the known SNPs are very similar between aligners and variant callers, and are in the expected range (2.0-2.1 for the whole human genome and 3.0-3.3 for the human exome), demonstrating the quality of the called variants and indicating that the called SNPs are very likely true polymorphisms (Tables 5.3 and S5.10). Novel SNPs identified in this work showed either similar or higher Ti/Tv ratios (Table 5.4) than known SNPs (Table 5.3), except for some of the novel SNPs called by SAMtools after data cleanup (Table S5.11).

Table 5.3. Ti/Tv ratios of known SNPs called by GATK and SAMtools. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	GATK	SAMtools	Overlap	GATK	SAMtools	Overlap
Raw reads	2.50	2.58	2.63	2.45	2.54	2.55
AdapterRemoval v.2.1.7	2.50	2.59	2.63	2.45	2.56	2.58
Cutadapt v.1.18	2.51	2.59	2.64	2.45	2.59	2.59
fastq-mcf v.1.05	2.50	2.58	2.63	2.45	2.57	2.58
Flexbar v.3.0.3	2.51	2.58	2.63	2.45	2.57	2.58
SeqPurge v.2018_06	2.50	2.58	2.63	2.45	2.57	2.57
Skewer v.0.2.2	2.50	2.59	2.63	2.45	2.57	2.58
Trimmomatic v.0.36	2.50	2.56	2.62	2.45	2.57	2.58

(B)

	GSNAP			STAR		
	GATK	SAMtools	Overlap	GATK	SAMtools	Overlap
Raw reads	2.43	2.47	2.47	2.38	2.48	2.44
AdapterRemoval v.2.1.7	2.43	2.48	2.47	2.38	2.48	2.44
Cutadapt v.1.18	2.43	2.49	2.48	2.39	2.48	2.44
fastq-mcf v.1.05	2.43	2.48	2.47	2.38	2.48	2.44
Flexbar v.3.0.3	2.43	2.48	2.47	2.38	2.48	2.45
SeqPurge v.2018_06	2.43	2.48	2.47	2.38	2.48	2.44
Skewer v.0.2.2	2.43	2.48	2.47	2.38	2.48	2.44
Trimmomatic v.0.36	2.43	2.49	2.47	2.38	2.48	2.44

Table 5.4. Ti/Tv ratios of novel SNPs called by GATK and SAMtools. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	GATK	SAMtools	Overlap	GATK	SAMtools	Overlap
Raw reads	4.95	3.00	2.42	6.79	4.00	0.00
AdapterRemoval v.2.1.7	4.96	3.36	3.50	7.28	5.67	0.00
Cutadapt v.1.18	5.01	3.47	3.14	7.51	8.00	0.00
fastq-mcf v.1.05	4.97	3.32	3.38	7.29	6.00	0.00
Flexbar v.3.0.3	4.98	3.32	2.78	7.40	4.50	0.00
SeqPurge v.2018_06	4.97	3.04	3.38	7.41	3.40	0.00
Skewer v.0.2.2	4.94	3.13	3.00	7.27	4.50	0.00
Trimmomatic v.0.36	4.93	3.65	3.25	7.32	2.80	0.00

(B)

	GSNAP			STAR		
	GATK	SAMtools	Overlap	GATK	SAMtools	Overlap
Raw reads	4.83	2.52	3.10	5.01	2.52	3.81
AdapterRemoval v.2.1.7	5.03	2.69	2.94	5.37	2.95	4.86
Cutadapt v.1.18	5.18	2.79	3.02	5.50	2.94	4.67
fastq-mcf v.1.05	5.01	2.70	2.92	5.33	2.89	4.81
Flexbar v.3.0.3	4.80	2.68	2.98	5.10	2.97	4.86
SeqPurge v.2018_06	4.82	2.73	3.10	5.12	2.99	4.64
Skewer v.0.2.2	4.81	2.68	3.02	5.11	3.01	4.81
Trimmomatic v.0.36	5.03	2.59	2.98	5.43	3.15	4.71

Finally, the known and novel indels detected with GATK and SAMtools are shown in Figures 5.9 and 5.10 and Tables S5.12-S5.13. Results in Figures 5.9 and 5.10 represent the mean of called indels using all trimming tools (between which the maximum variation for gold standard data was up to 0.32% and 0.70% for GSNAP and STAR, and for breast tissue data was up to 0.21% and 0.27% for GSNAP and STAR, respectively). In every case, more indels were identified using GATK.

For calls from GSNAP mappings of gold standard data, known indels represented around 82% and 79% of all indels, from GATK and SAMtools, respectively. For STAR alignments, higher percentages were observed in calls from GATK (~91%), and from SAMtools (~92%). For calls from GSNAP mappings of breast tissue data, known indels represented about 64% and 85% of all indels, from GATK and SAMtools, respectively. For STAR alignments, comparable percentages were observed in calls from GATK (~65%), but the fraction of known indels was lower with SAMtools (~73%). The number of total indels detected with SAMtools after using GATK data cleanup was much lower (Table S14). For those pipelines, known indels from gold standard data represented about 87% for GSNAP and 89% for STAR. For the breast tissue data, these values were similar for GSNAP (89%) and lower (73%) for STAR.

The number of indels shared by GATK and SAMtools was reduced by the lower amount of indels called by SAMtools (Table S5.15). As for SNPs, more indels were called from raw reads both with GSNAP and STAR, and the percentages of known indels relatively to all indels, was about 97% (gold standard) and 92% (breast tissue), and over 99% (gold standard) and 95% (breast tissue), respectively.

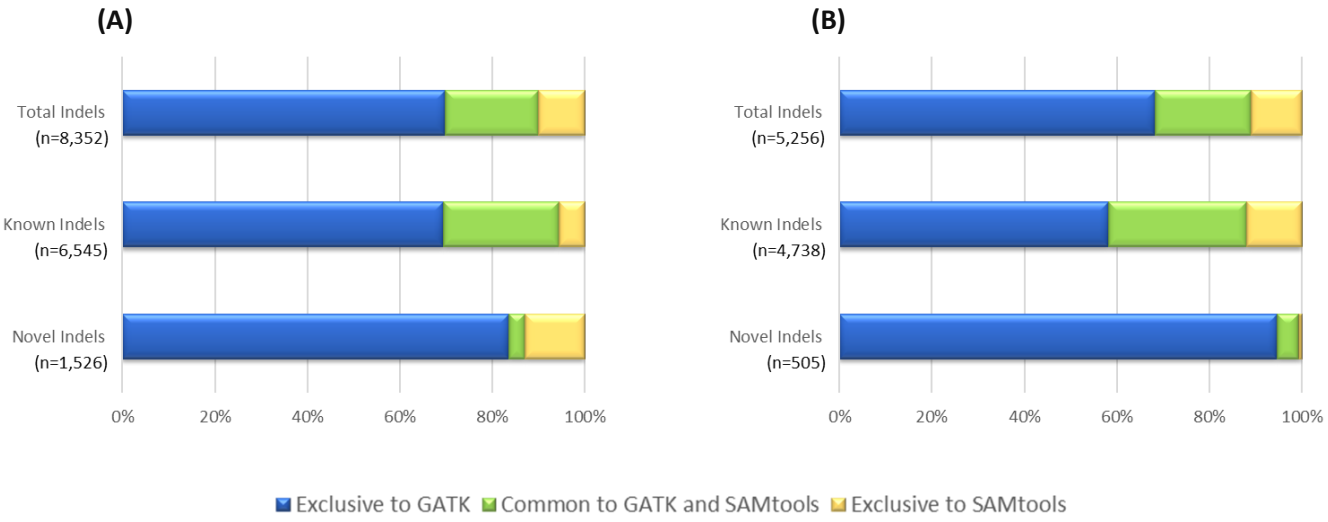


Figure 5.9. Indels identified by GATK and SAMtools after mapping with GSNAP. **(A)** gold standard sample. **(B)** breast tissue sample. Total Indels – all indels (known and novel). Known Indels – indels found in dbSNP150. Novel Indels – indels not found in dbSNP150. The represented number of indels refers to the mean of called indels using each trimming tool.

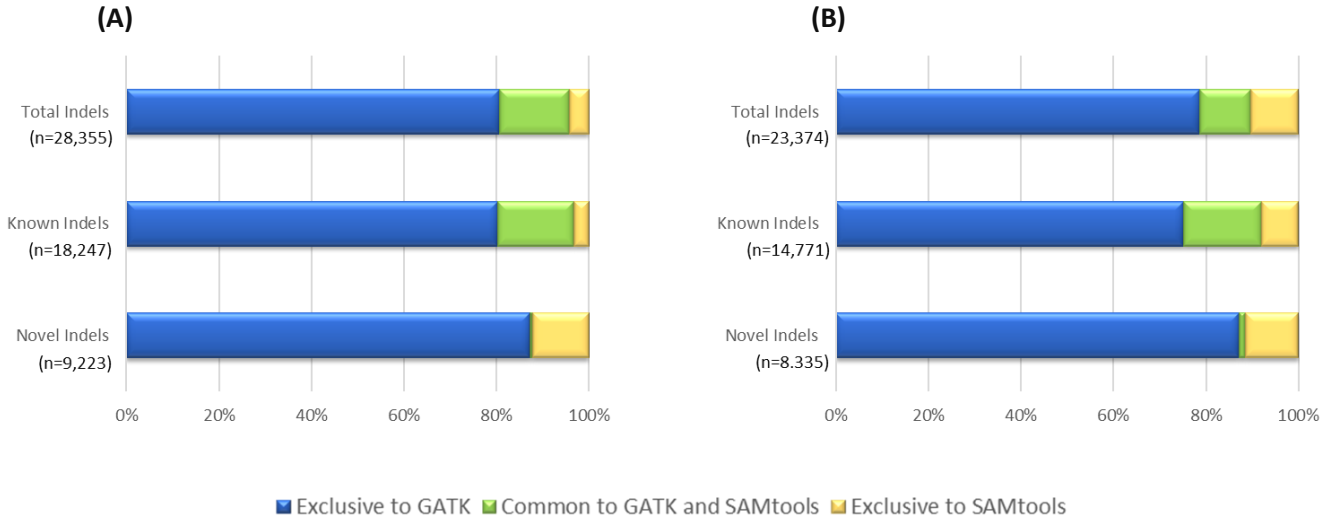


Figure 5.10. Indels identified by GATK and SAMtools after mapping with STAR. **(A)** gold standard sample. **(B)** breast tissue sample. Total Indels – all indels (known and novel). Known Indels – indels found in dbSNP150. Novel Indels – indels not found in dbSNP150. The represented number of indels refers to the mean of called indels using each trimming tool.

Regarding all identified indels, the number of overlapping known indels from GATK and SAMtools after data cleanup ranged from 97% (gold standard) and 96% (breast tissue) for GSNAP to 99% (gold standard) and 98% (breast tissue) for STAR (Figure S5.3 and Table S5.16).

For the gold standard data, the performance of all variant calling pipelines was also assessed, by determining precision, sensitivity, and the F-measure (Tables 5.5-5.7). Overall, alignments with GSNAP identified more true positives (TP) and fewer false negatives (FN) than alignments from STAR, either with GATK or SAMtools. However, this also translated into a higher number of false positives obtained with GSNAP mappings. Thus, slightly higher precision and lower sensitivity were obtained for calls from STAR alignments. Finally, an equal (with GATK) or lower (with SAMtools) value for the F-measure was observed in calls from reads mapped with STAR, relatively to reads mapped with GSNAP. Levels of precision improved considering variants called by both variant callers, GATK and SAMtools.

Table 5.5. Precision and Sensitivity of variants called with GATK. TP – true positive, FP – false positive, FN – false negative, precision=TP/(TP+FP), sensitivity=TP/(TP+FN), F-measure=(2TP)/(FN+FP+2TP). F-measure is the harmonic mean of precision and sensitivity with a value between 0, for the worst, and 1, for the perfect F-measure.

	GSNAP						STAR					
	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure
Raw reads	20,137	2,426	42,552	89,2	32,1	0,5	20,050	1,692	42,639	92,2	32,0	0,5
AdapterRemoval v.2.1.7	20,055	2,435	42,634	89,2	32,0	0,5	19,948	1,591	42,741	92,6	31,8	0,5
Cutadapt v.1.18	19,814	2,358	42,875	89,4	31,6	0,5	19,746	1,553	42,943	92,7	31,5	0,5
fastq-mcf v.1.05	20,055	2,428	42,634	89,2	32,0	0,5	19,936	1,584	42,753	92,6	31,8	0,5
Flexbar v.3.0.3	19,990	2,404	42,699	89,3	31,9	0,5	19,888	1,577	42,801	92,7	31,7	0,5
SeqPurge v.2018_06	20,035	2,429	42,654	89,2	32,0	0,5	19,929	1,576	42,760	92,7	31,8	0,5
Skewer v.0.2.2	20,059	2,442	42,630	89,1	32,0	0,5	19,958	1,584	42,731	92,6	31,8	0,5
Trimmomatic v.0.36	19,958	2,415	42,731	89,2	31,8	0,5	19,824	1,549	42,865	92,8	31,6	0,5

Table 5.6. Precision and Sensitivity of variants called with SAMtools. TP – true positive, FP – false positive, FN – false negative, precision=TP/(TP+FP), sensitivity=TP/(TP+FN), F-measure=(2TP)/(FN+FP+2TP). F-measure is the harmonic mean of precision and sensitivity with a value between 0, for the worst, and 1, for the perfect F-measure.

	GSNAP						STAR					
	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure
Raw reads	14,933	2,125	47,760	87,5	23,8	0,4	12,218	363	50,471	97,1	19,5	0,3
AdapterRemoval v.2.1.7	14,280	1,812	48,414	88,7	22,8	0,4	12,061	318	50,628	97,4	19,2	0,3
Cutadapt v.1.18	13,844	1,659	48,848	89,3	22,1	0,4	11,731	297	50,958	97,5	18,7	0,3
fastq-mcf v.1.05	14,286	1,818	48,408	88,7	22,8	0,4	12,029	315	50,660	97,4	19,2	0,3
Flexbar v.3.0.3	14,211	1,807	48,483	88,7	22,7	0,4	11,991	317	50,698	97,4	19,1	0,3
SeqPurge v.2018_06	14,249	1,779	48,445	88,9	22,7	0,4	12,037	309	50,652	97,5	19,2	0,3
Skewer v.0.2.2	14,299	1,821	48,395	88,7	22,8	0,4	12,070	316	50,619	97,4	19,3	0,3
Trimmomatic v.0.36	14,082	1,752	48,612	88,9	22,5	0,4	11,950	286	50,739	97,7	19,1	0,3

Table 5.7. Precision and Sensitivity of common variants from GATK and SAMtools. TP – true positive, FP – false positive, FN – false negative, precision=TP/(TP+FP), sensitivity=TP/(TP+FN), F-measure=(2TP)/(FN+FP+2TP). F-measure is the harmonic mean of precision and sensitivity with a value between 0, for the worst, and 1, for the perfect F-measure.

	GSNAP						STAR					
	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure
Raw reads	14,447	603	48,241	96,0	23,0	0,4	11,876	141	50,813	98,8	18,9	0,3
AdapterRemoval v.2.1.7	13,891	538	48,797	96,3	22,2	0,4	11,757	129	50,932	98,9	18,8	0,3
Cutadapt v.1.18	13,479	503	49,209	96,4	21,5	0,4	11,441	123	51,248	98,9	18,3	0,3
fastq-mcf v.1.05	13,894	541	48,794	96,3	22,2	0,4	11,724	129	50,965	98,9	18,7	0,3
Flexbar v.3.0.3	13,821	527	48,867	96,3	22,0	0,4	11,687	124	51,002	99,0	18,6	0,3
SeqPurge v.2018_06	13,855	544	48,833	96,2	22,1	0,4	11,735	124	50,954	99,0	18,7	0,3
Skewer v.0.2.2	13,901	545	48,787	96,2	22,2	0,4	11,762	125	50,927	98,9	18,8	0,3
Trimmomatic v.0.36	13,700	533	48,988	96,3	21,9	0,4	11,654	130	51,035	98,9	18,6	0,3

5.5. Discussion

In this work, 42 variant calling pipelines for RNA-seq data were compared, through the combination of seven trimming tools, two aligners and, two variant callers. The analyses were conducted with paired-end data from a gold standard sample and one sample of breast tissue from a healthy woman. Using RNA-seq data to identify genomic variants facilitates the discovery of variants with functional relevance. It also allows AE quantification in heterozygous SNPs and consequently the identification of cis-regulated genes and mapping of the functional variant. However, calling variants from RNA-seq data to retrieve allele counts in heterozygous sites raises several challenges mainly regarding mapping bias. This study intended to compare the performance of different pipelines to identify the most suitable for the intended analysis.

For comparison of the results from each pipeline, several metrics were considered in each step. Assessment of the number of adapters kept after trimming showed that only Trimmomatic was able to remove all the adapters. From the seven trimming tools compared, Trimmomatic also removed a higher percentage of low-quality bases in breast tissue data, while demanding intermediate time. Excluding fastq-mcf, for which it was not possible to disable the option of multi-threading (which explains the time difference relative to the other tools), Skewer had the best performance regarding time and memory but failed to remove several adapters and low-quality bases. Trimmomatic required a higher memory usage, but the reason behind it is not clear. However, this did not hinder the usage of this tool as part of a pipeline for variant calling. On the other hand, the lower number of low-quality bases discarded by Flexbar, SeqPurge, and Skewer can be explained by the fact that these trimmers only allow quality trimming for the 3' end.

To evaluate the read alignment by GSNAP and STAR, after trimming with each trimming tool, the numbers of mapped reads in pair and singletons were compared. GSNAP mapped more reads in pair and fewer singletons than STAR. Regarding these metrics, the best results were observed after pre-processing the data with Trimmomatic. The percentage of properly paired reads was similar for both aligners, but STAR produced around 6% more uniquely mapped reads. Decreased mapping of reads to unique locations is expected in a SNP-tolerant alignment as reads can be mapped to additional genomic regions, as described by GSNAP developers (Wu & Nacu, 2010). GSNAP accounts not only for the mappable genomic regions in the reference genome but also the ones in the non-reference genome. Thus, the lower percentage of uniquely mapped reads obtained with GSNAP may be explained by reads carrying alternative alleles that may fail to map uniquely.

These results are consistent with another benchmarking study (Engström, et al., 2013). Moreover, GSNAP required considerable computational resources, as previously reported (Castel, Levy-Moonshine, Mohammadi, Banks, & Lappalainen, 2015; Grant, et al., 2011; Engström, et al., 2013).

Given the less complex workflow, variant calling was faster using SAMtools than GATK. Overall comparison of variants called before and after trimming showed very similar results for all the trimmers. In addition, concordant calls from GATK and SAMtools revealed that up to 2% fewer calls were made after pre-processing the reads. This may indicate that pre-processing of sequencing data with very good quality, as was the case in this work, may not be crucial. Nevertheless, since the difference between called variants from raw and trimmed reads is not significant, and given the importance of trimming showed previously (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013), trimming the data would be recommended.

More variants were filtered after calling with SAMtools – up to 87% and 71% for gold standard data, and around 49% or 38% for breast tissue data, with GSNAP or STAR, considering all trimmers –, than with GATK – up to 22% or 11% (gold standard), and 37% or 33% (breast tissue), with GSNAP or STAR –. Comparison of the variants called with each pipeline in this work showed that, similarly to previous studies (Liu, Han, Wang, Gelernter, & Yang, 2013; O'Rawe, et al., 2013; Yu & Sun, 2013; Baes, et al., 2014; Ni, et al., 2015), more variants were called when GATK was used. For the gold standard data and breast sample data, and considering the calls passing the filters, GATK identified, respectively, about 52% and 40% (mapping with GSNAP), or 73% and 56% (mapping with STAR) more variants than SAMtools. The impact of aligners on the number of detected variants was also noteworthy. Relatively to mapping with STAR, mapping with GSNAP identified 18% (gold standard) or 10% (breast tissue) more variants with GATK, and 53% (gold standard) or 35% (breast tissue) more variants with SAMtools.

Up to date, different studies using WGS and WES data have shown contradictory results showing either concordance (Liu, Han, Wang, Gelernter, & Yang, 2013; Hwang, Kim, Lee, & Marcotte, 2015; Ni, et al., 2015) or discrepancy (O'Rawe, et al., 2013; Yu & Sun, 2013; Cornish & Guda, 2015) between the variants called by different variant callers. As in the later studies, results from the present work showed significant divergence between the variants called by the two variant callers. Results from Yu and Sun (Yu & Sun, 2013) showed that 41.1% of known SNPs and 17.4% of novel SNPs were shared by four pipelines. Comparing five different pipelines O'Rawe and colleagues found overall concordances of 59.6% and 11.4% between known and novel SNPs, and 43.4% and 4.7% between known and novel indels, respectively (O'Rawe, et al., 2013). Results

from this study showed the same tendency. Regarding the gold standard data, about 31% or 29% (GSNAP or STAR) of all known SNPs, and 25% or 16% (GSNAP or STAR) of known indels were identified by both callers. For the breast tissue data, approximately 50% or 53% (GSNAP or STAR) of all detected known SNPs, and 30% or 17% (GSNAP or STAR) of known indels, were shared by the two callers. The discrepancy increased when considering novel variants. In this case, only less than 0.4% of novel SNPs and less than 4.6% of novel indels were common to GATK and SAMtools, regardless of the aligner or the sample.

To estimate the quality of the detected SNPs, Ti/Tv ratios were estimated. Ti/Tv ratios of known SNPs were highly similar, demonstrating a good performance for all the compared pipelines. Usually, novel SNPs tend to show a lower Ti/Tv due to the presence of false positives, a higher transition ratio at lower frequency variation, and reduced transitions due to sequencing context bias (DePristo, et al., 2011). Novel SNPs identified in this work showed either similar or higher Ti/Tv ratios, except for some of the novel SNPs called by SAMtools after data cleanup, for which the Ti/Tv ratios were lower than expected.

Respecting precision and sensitivity, mapping with GSNAP allowed the identification of more true positives, although more false positives have been also identified relatively to STAR. However, whether the identification of more true positives compensates for the detection of a slightly higher number of false positives, depends on the user's objective, and the advantages and disadvantages should be weighted accordingly. In addition, although more false positives were identified with GSNAP, this aligner also led to the identification of fewer false negatives, and better F-measures, which indicates a favourable balance between precision and sensitivity.

In this work, and considering the subsequent study aiming at the identification of variants for AE analysis, the detection of more true positives would be favoured. Considering the substantial level of divergence in variants identified with different calling pipelines, it has been suggested to focus on variants detected by different pipelines simultaneously (Liu, Han, Wang, Gelernter, & Yang, 2013), thus reducing the false positive rate. Since the identification of cis-regulatory variants by AE analysis relies on accurate detection of single variants, and since none of the variant callers compared in this study are optimised for RNA-seq data, and do not allow yet joint calling from multiple samples, the two callers should be used to find common variants for further analyses.

Taken together, these results suggest that Trimmomatic removes more adapters and low-quality bases, mapping with GSNAP allows the detection of a higher number of variants, including true positives, and calling should be performed with GATK and SAMtools. The intersection of the

common variants should be used for subsequent analysis. The overall results show that running the data cleanup with tools from GATK before calling with mpileup, as indicated by SAMtools workflows for WGS and WES sequencing, does not improve the obtained results when variant calling is performed with mpileup command from SAMtools after mapping.

Nevertheless, it should be taken into consideration that even for variant calling performed with GATK, some utilities from SAMtools remain useful for several tasks.

Chapter VI

Identification of New Susceptibility Loci for Breast Cancer

6.1. Abstract

Using allelic expression (AE) ratios in association studies to identify new risk loci for breast cancer, presents several advantages to classical studies using genotypes as it requires a smaller number of samples and reports on the effect of all risk variants within a locus. In this work, RNA-seq data from 12 normal breast tissue samples of healthy women (controls) and 14 breast cancer patients (cases) was analysed using the pipeline previously established in chapter V. AE was quantified and AE ratios were calculated genome-wide across 7,054 informative heterozygous genetic variants. From these, 353 showed statistically significant AE differences (p-value<0.05) between cases and controls, of which eight presented $|\log_2(\text{fold-change})|$ greater than two. Validation using real-time PCR to quantify AE and in a larger sample set revealed a new risk-associated variant, rs3211416 (*CDC16*), with a predicted effect size of -1.83 [Hedges' *g*, 95%CI=-2.38, -1.14]. Furthermore, this association was not impacted by the ER status of the tumours of the patients. The case-control study with AE ratios was also performed with blood samples but did not reveal a significant association with breast cancer risk.

Case-control association studies from this work show the advantage of using a quantitative variable (AE), which directly identifies the effects of cis-regulatory variation, and consequently, the target gene(s), while showing regulation of transcript levels as the biological mechanism underlying susceptibility. Results from this work show the potential of using our approach for further case-control studies with data from public databases, which will increase its statistical power to identify risk loci for breast cancer or other complex diseases.

6.2. Introduction

Allelic expression (AE) analysis has become a tool of interest to unravel genome function, as it quantifies the variation of expression between the two alleles in diploid individuals at heterozygous sites. This approach has the advantage of assessing expression within the same individual, cellular environment, and trans-acting factors, which improves the detection of the effects of cis-regulatory variants (Skelly, Ronald, & Akey, 2009; Pastinen, 2010). Thus, identification of differential allelic expression (DAE) in transcribed variants indicates the existence of one or more cis-regulatory variants altering gene expression. Moreover, analysis of AE ratios, which are a continuous variable, has high statistical power to detect cis-regulatory effects even with small sample sizes (Maia, et al., 2009; Fogarty, Xiao, Prokunina-Olsson, Scott, & Mohlke, 2010; Fontanillas, et al., 2010).

Advances in high-throughput RNA sequencing (RNA-seq) allow access to the transcriptional landscape, including heterozygous variants lying in expressed genes, and to date, several studies have integrated RNA-seq data to perform genome-wide AE analysis (Heap, et al., 2010; Castel, Levy-Moonshine, Mohammadi, Banks, & Lappalainen, 2015; Edsgärd, et al., 2016). The presence of DAE in a given heterozygous transcribed SNP, is indicated by the significant departure from a 1:1 ratio of RNA-seq reads from the two alleles. Therefore, RNA-seq data can be used to detect and quantify AE in cases and controls, and the differences between both groups can lead to the identification of susceptibility loci for disease (Valle, et al., 2008). Additionally, the magnitude of the difference between the two groups indicates the effect size of the detected association, in this case, at a given SNP. Measurement of the effect size has the advantage of being independent of sample size, unlike significance tests (Sullivan & Feinn, 2012).

In this work, we performed a two-phase case-control study using AE ratios as a quantitative phenotype. In the first phase, we used RNA-seq data in a small group of samples to identify the candidates showing more association potential with risk. Then, in a second phase, we used real-time PCR to carry the case-control association analysis in a larger set, like what is described in Chapter IV, both in breast tissue and blood.

6.3. Materials and Methods

6.3.1. RNA Samples

Twenty-seven RNA samples extracted from normal breast tissue were used in this work. These included 12 samples from healthy women submitted to reduction mastectomy for reasons other than cancer, collected at the Tissue Bank at Addenbrooke's Hospital (Cambridge, UK) (Maia, et al., 2012). The remaining 15 RNA samples were extracted from healthy tissue of breast cancer patients (normal matched samples) within the METABRIC project (Curtis C., et al., 2012). All samples were collected with the approval of Addenbrooke's Hospital Local Research Ethics Committee (REC reference 06/Q0108/221 and 07/H0308/161).

Total RNA was previously extracted using QIAzol® (QIAGEN, Germany) following the manufacturer's instructions. The RNA was subsequently treated with DNaseI, and repurified using acidic phenol-chloroform and ethanol precipitation (Maia, et al., 2012).

RNA concentration was initially evaluated with Nanodrop 2000™ Spectophotrometer (ThermoFischer Scientific, USA) and with Qubit 2.0 Fluorometer (ThermoFischer Scientific, USA) using the RNA HS Assay.

RNA integrity and concentration were assessed through analysis of its length distribution, using the electrophoretic method on a microfluidic platform with the Experion™ Automated Electrophoresis Station (Bio-Rad, USA). This system generates the RNA quality indicator (RQI), which is a quantitative integrity assessment. It also automatically generates concentration data, visual electropherogram data, and ribosomal RNA ratios. The RQI is based on a numbering system between 1 (totally degraded RNA) and 10 (fully intact RNA). This score has different designations and is created by different algorithms depending on the platform/technology used. The most common score is the RNA integrity number (RIN) obtained with the Bioanalyzer (Agilent Technologies, USA), which also uses microfluidic separation technology. An analogue score is the RNA quality number (RQN) from the Fragment Analyzer™ Automated CE System (Advanced Analytical Technologies, Inc., USA). This system incorporates a fluorescence-based, parallel capillary electrophoresis that also provides information on nucleic acid concentration and size distribution.

RNA samples with an RQI above seven were selected for NGS. Samples were properly conditioned and shipped with dry ice.

6.3.2. Libraries Preparation

At least 400 ng of RNA from each sample was sent for commercial next-generation sequencing (NGS) at Eurofins Genomics (Munich, Germany). An initial quality control (QC) step was carried out by the company. RQN was analysed by capillary electrophoresis performed with the Fragment Analyzer™ Automated CE System (Advanced Analytical Technologies, Inc., USA). Libraries were then prepared using the Illumina TruSeq Stranded mRNA kit. Briefly, the mRNA molecules containing polyA were purified using poly-T oligo attached magnetic beads. Subsequently, the polyA RNA was fragmented and primed with random hexamers for the first strand cDNA synthesis. Spurious DNA-dependent synthesis was prevented with the addition of Actinomycin D. The RNA template was then removed and a replacement strand (second strand cDNA) was synthesised by incorporating dUTP in place of dTTP, generating double-stranded cDNA (substitution of dTTP by dUTP quenches the second strand during amplification since the polymerase activity is blocked past this nucleotide). Next, blunt-ended double-stranded cDNA was isolated with magnetic beads, and a single 'A' nucleotide was added to the 3' ends, allowing a complementary overhang for ligation to the corresponding single 'T' nucleotide on the 3' end of the adapters. After ligation of indexing adapters, DNA fragments with adapter molecules on both ends were enriched by PCR. Indexed DNA libraries were normalised to the same concentration and pooled in equal volumes.

6.3.3. RNA-sequencing

Sequencing of the libraries was performed on an Illumina HiSeq 2500 instrument for 2×100 cycles (paired-end reads) using HiSeq Sequencing by Synthesis (SBS) v4 chemistry and the HiSeq Control Software (HCS) v2.2.58. Primary analysis of sequencing data, including base calling and quality checking (Q-scoring) was performed with the software application Real Time Analysis (RTA) v1.18.64. The bcl2fastq Conversion Software v1.8.4 from the CASAVA software suit was used to demultiplex the data and to convert the per-cycle base call (BCL) files, generated at the end of the sequencing runs, into FASTQ file formats for downstream analysis. The quality scale used was Phred+33 (Sanger/Illumina1.9).

6.3.4. RNA-sequencing Analysis

All data processing and analysis were carried out at Universidade do Algarve. Firstly, the quality of RNA-seq data was assessed with the FastQC v0.11.7 application (Andrews, 2010). Raw data were then pre-processed with Trimmomatic v.0.36 (Bolger, Lohse, & Usadel, 2014) to trim adapter sequences and filter low-quality bases from reads. These were then aligned to the human reference genome GRCh38 with the Genomic Short-read Nucleotide Alignment Program (GSNAP) v.2017-11-15 (Wu & Nacu, 2010), using the human dbSNP Build 150 (Bethesda (MD): National Center for Biotechnology Information, s.d.) obtained from UCSC. Alignments were performed with multi-threading, using 10 threads. Next, variant calling was performed with the Genome Analysis ToolKit (GATK v.3.8 and v.4.0.4.0) (McKenna, et al., 2010), which uses some utilities from Picard v.2.18.3 (http://broadinstitute.github.io/picard), and with the SAMtools software package v.1.7 (Li H., 2011), including its two key components samtools and bcftools. Variants called with GATK were hard filtered, using FisherStrand (FS) higher than 30.0, and QualByDepth (QD) lower than 2.0. Variants from SAMtools were filtered based on their average mapping quality (MQ) and their raw read depth (DP), and variants with MQ>19.0 and DP>9.0 were kept. Common variants from both variant callers were identified with BEDTools v.2.25.0 (Quinlan & Hall, 2010) and used for further analyses.

6.3.5. AE Ratios Analysis from RNA-seq Data

Analysis of read counts and AE ratios [AE = $\log_2$ (number of read counts from alternative allele / number of read counts from reference allele)] was performed with RStudio Server v.1.4.1103 (RStudio Team, 2010). Only heterozygote bi-allelic genetic variants with 15 reads or more for each allele, from at least three controls and three cases, were kept. Variants from sexual chromosomes were also excluded. To test if the variances of the AE ratios from cases and controls were equal (null hypothesis) an F-test was performed, and subsequently, a t-test identified the variants with significant differences (p-value<0.05) between cases and controls. From those, SNPs with a $|\log_2(\text{fold-change})|$ larger than two between AE ratios of controls and cases were considered to identify the best candidate loci associated with breast cancer risk.

6.3.6. PCR analysis and sequencing of the variant rs757142894

Amplification of the sequence harbouring the variant rs757142894 was carried out for DNA samples from controls through PCR with the primers APOC1P1_Fw – 5'-AGACAGTGGGGATGGAGATT -3' and APOC1P1_Rev - 5'-CAGTTTGCAGGGTTCTTAGG-3' designed in this work. These primers amplify DNA fragments from the *APO1C1* pseudogene (806 bp) and the *APOC1* gene (498 bp). PCR amplifications were performed in a C1000 Touch™ Thermal Cycler (Bio-Rad Laboratories, Inc., USA), in a final volume of 50 μl with colourless GoTaq Flexi Buffer (Promega, USA), 2 mM of $MgCl_2$, 200 nM of dNTPs, 500 nM of each primer, 1.25 U of GoTaq G2 Flexi DNA Polymerase (Promega, USA), and 200 ng of DNA. PCR reactions included an initial enzyme activation step at 95 °C for 3 min, followed by 35 cycles of denaturation at 95 °C for 30 sec, annealing at 50 °C for 30 sec and extension at 72 °C for 30 sec. After a final extension step at 72 °C for 5 min, reactions were stopped at 10 °C.

Electrophoresis was conducted in a 1% agarose gel in 1X TAE buffer (pH 8.3), which was stained with Greensafe (NZYTech, Portugal). The NZYDNA Ladder V (NZYTech, Portugal) was used as a size marker. The PCR products were analysed under UV light with a ChemiDoc™ XRS+ Imaging System and using the ImageLab™ Software (Bio-Rad Laboratories, Inc., USA). Sanger sequencing was performed at NZYTech, Genes & Enzymes (Lisbon, Portugal), and sequences were analysed with the BioEdit Sequence Alignment Editor (Hall, 2011).

6.4. Results

6.4.1. RNA samples preparation and sequencing

RNA-sequencing was successful for 26 of the 27 RNA samples of healthy breast tissue (12 from controls and 14 from breast cancer patients). The remaining sample (from a patient) did not show the same performance during the clustering analysis and was discarded. Namely, only about 21 million reads were obtained for this sample, whilst an average of 50 to 60 million reads (100 to 120 million paired-end reads) were obtained for all others (Table S6.1).

6.4.2. RNA-seq data analysis

Two fastq files per sample (one from each read end) were initially assessed for RNA-seq data quality check with the FastQC v0.11.7 application. Data from all samples passed this quality control step and were used in the case-control study.

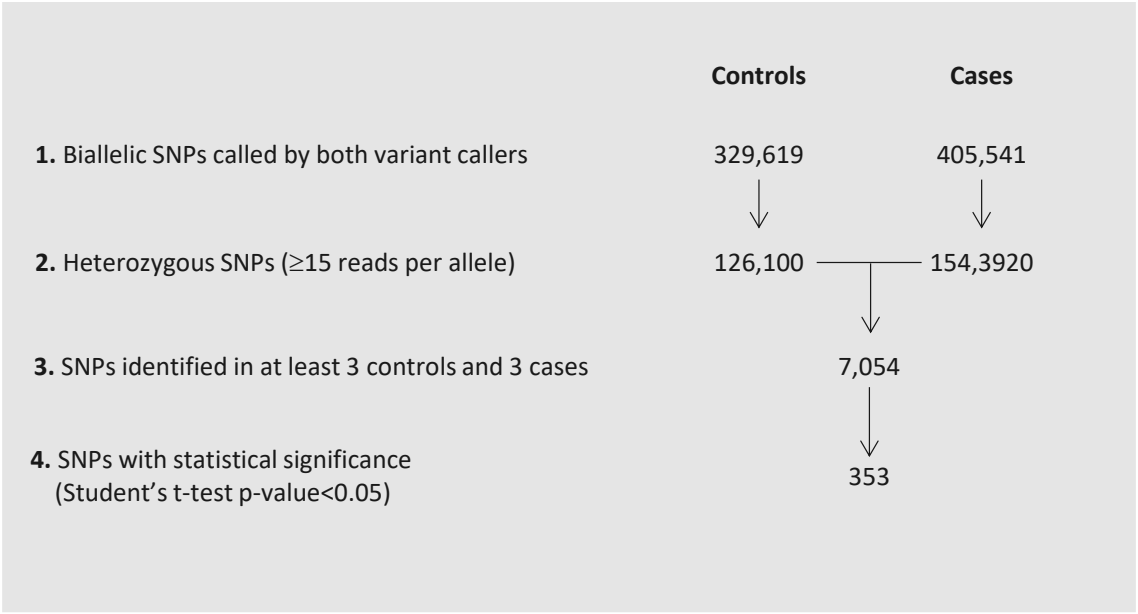


Figure 6.1. Filtering process followed to identify genetic variants with potentially different cis-regulation between breast cancer patients and healthy women.

RNA-seq data from all samples were analysed with the pipeline established in the previous chapter, and 329,619 and 405,541 biallelic SNPs were identified from controls and cases, respectively (Figure 6.1). From those, only the heterozygous variants with at least 15 reads for each allele and found in at least three controls and three cases were used in further analysis. In total, 7,054 SNPs fitted those criteria, and 353 SNPs showed a difference (uncorrected p-value<0.05) between the allelic expression ratios from cases and controls (Figure 6.1 and Figure 6.2).

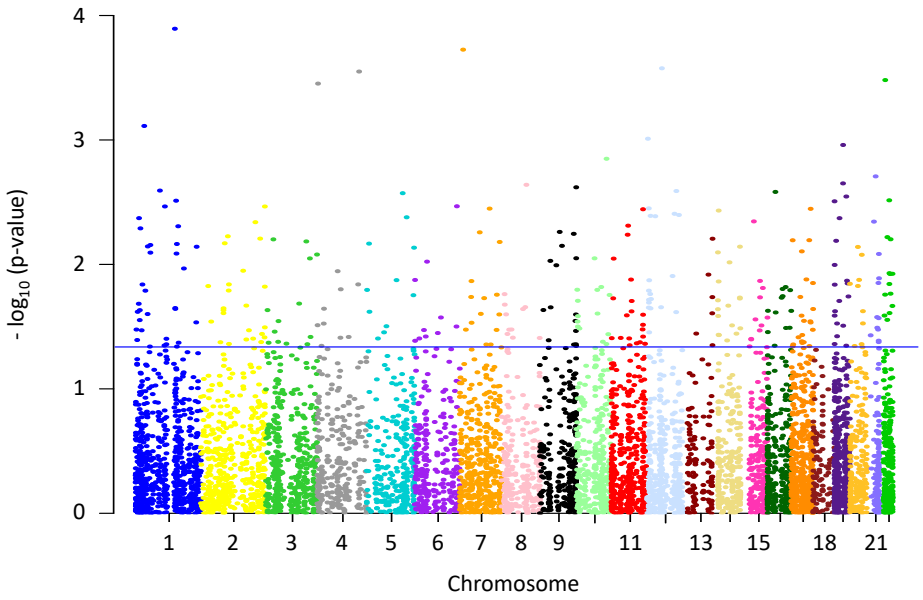


Figure 6.2. Manhattan plot showing the genetic variants identified across all autosomes after analysis of AE ratios from the case-control study using RNA-seq data from 14 cases and 12 controls. Each dot represents a SNP. The level of association of each SNP is represented as $-\log_{10}$ (p-value) in the y-axis, and the SNPs' position is represented in the x-axis. The blue line represents the threshold for statistical significance (p-value<0.05), above which there are 353 SNPs.

6.4.3. Case-control study based on RNA-seq data analysis identified eight genetic variants with potential association with breast cancer risk

To identify new candidate risk-associated loci for breast cancer, the fold-change (FC) between the allelic expression ratios of cases and controls was determined. From the genetic variants for which differences in AE ratios between cases and controls were significant (uncorrected p-value<0.05), 49 showed a $|\log_2(FC)|>1.5$ (Figure S6.1) and eight a $|\log_2(FC)|>2$ (Figure 6.3 and Figure 6.4). These variants located in eight different genes and proceeded to the next analysis step: rs34169189 (*PDE4DIP*, phosphodiesterase 4D interacting protein), rs62449782 (*ARL4A*, 11ADP ribosylation factor like GTPase 4A), rs11545332 (*DDX11*, DEAD/H-box helicase), rs530963 (*CCDC86*, coiled-coil domain containing 86), rs757142894 (*APOC1P1*, apoliproprotein C1 pseudogene 1), rs3211416 (*CDC16*, cell division cycle 16), rs11724432 (*CCNG2*, cyclin G2), and rs2281791 (*TBC1D12*, TBC1 domain family member 12).

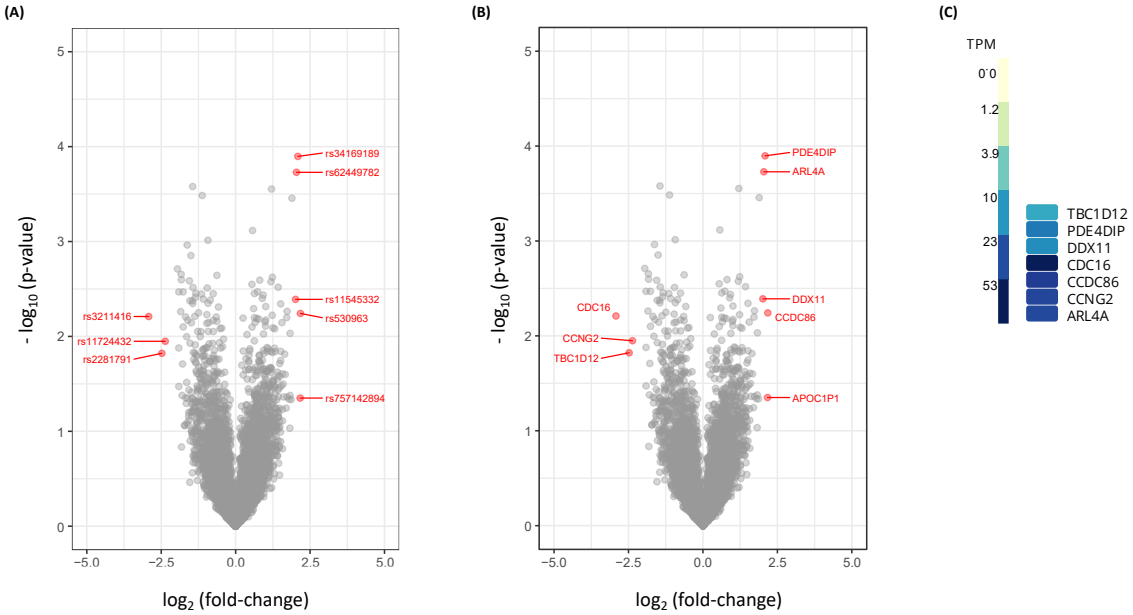


Figure 6.3. Volcano plots showing statistical significance [-$\log_{10}$ (p-value)] against fold-change (FC) of genetic variants from case-control study using RNA-seq data. In red are represented the eight genetic **(A)** variants and **(B)** genes with statistically significant difference between allelic expression (AE) ratios from breast cancer patients and controls (p-value<0.05) and with a $|\log_2(FC)|>2$. Positive x-values indicate up-regulated variants/genes, and negative x-values indicate down-regulated variants/genes. **(C)** expression from seven out of the eight genes in transcripts per million (TPM) in breast tissue (GTEx Consortium, 2015).

Expression data for these genes in breast tissue was retrieved from the GTEx database (GTEx Consortium, 2015), and seven of the eight genes were found to be highly expressed (Figure 6.3C). There was no data available to assess the expression of the *APOC1P1* gene. The eight SNPs showed a large effect size (Figure S6.2).

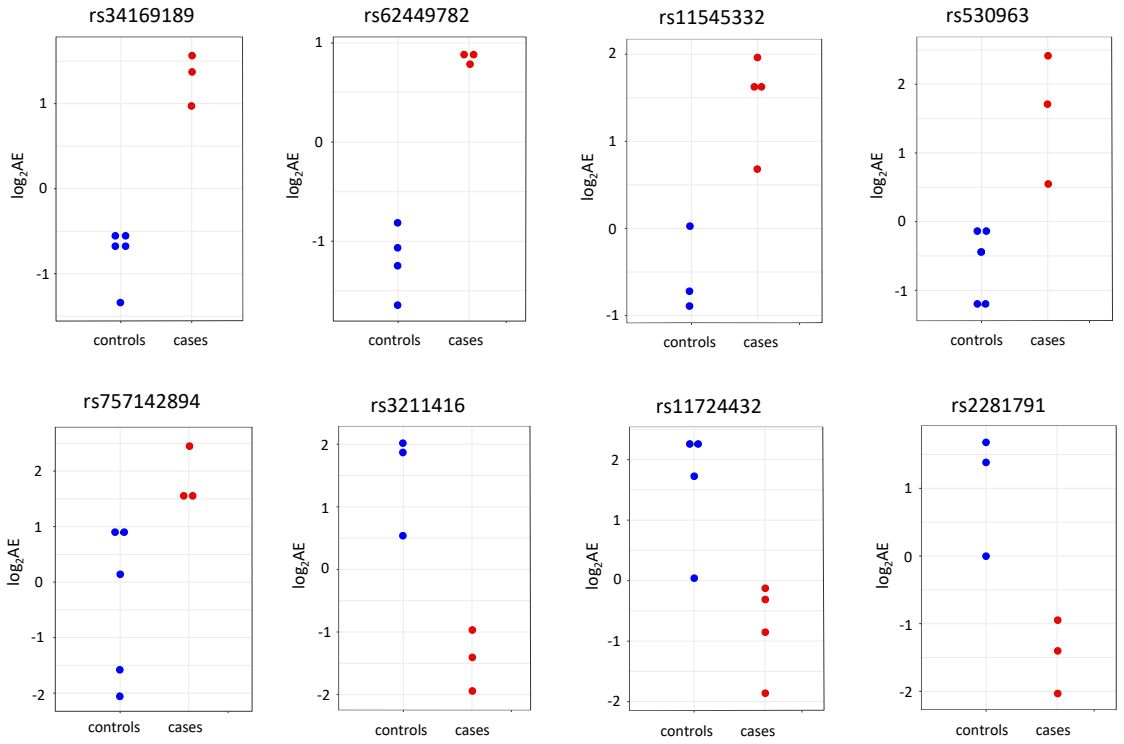


Figure 6.4. Plots showing the difference between the $\log_2$ of allelic expression (AE) ratio (AE=read counts of alternative allele / read counts of reference allele) of the eight genetic variants with statistically significant difference (p-value<0.05) between AE ratios from breast cancer patients (red dots) and controls (blue dots) and with a $|\log_2(FC)|>2$.

6.4.4. Variant rs757142894 was excluded as a candidate for association with breast cancer risk

The minor allele frequency (MAF) of the eight candidate variants was assessed, and a large discordance was identified between the observed and published MAF of rs757142894 (published MAF=0.000043, observed MAF=0.15). To understand this discordance, amplification of the DNA sequence harbouring the SNP was performed in DNA samples from controls. This resulted in the amplification of two fragments of different lengths. After Sanger sequencing, it was found that one of the fragments was from the *APOC1* (Apoliproprotein C1) gene and the other from the *APOC1P1*. Both genes are very similar at the 50 bases surrounding the rs757142894 position, with only two different bases. Thus, all reads harbouring this sequence are prone to misalignment and this variant was discarded from further analysis.

6.4.5. Case-control studies using digital real-time PCR show no association of rs530963 (*CCDC86*) and rs11724432 (*CCNG2*) with breast cancer risk

Next, case-control analyses using real-time PCR were performed for the six remaining candidates. The variant rs34169189 in *PDE4DIP* is located at the beginning of the first open reading frame of the gene (nucleotide 30), leaving insufficient space to design primers for the TaqMan™ SNP Genotyping Assay. For the variant rs530963, located on *CCDC86*, not only the AE ratios distributions showed opposite shifts in the two experiments conducted with digital PCR with the first set of cases, but also did not show significant differences between cases and controls (Figure S6.3 and Table 6.1). This was not due to differences between ER+ and ER- cases (Figure S6.4), although ER+ cases always showed higher effect size than ER- cases (g=-0.49 vs. g=-0.04 and g=0.38 vs. g=-0.24). Similar results were obtained for the variant rs11724432, located on *CCNG2* (Figures S6.3 and S6.5), for which even lower effect sizes were obtained (Table 6.1). Thus, for both rs530963 and rs11724432 no further analysis was performed.

6.4.6. Case-control studies using real-time PCR did not show an association of rs62449782 (*ARL4A*) and rs2281791 (*TBC1D12*) with breast cancer risk

Analysis of the SNP rs62449782, located on *ARL4*, revealed a significant difference between cases and controls when considering the second experiment conducted with the first set of cases, and the third experiment using the second group of cases, with a size effect of g=-0.65 and g=-0.67, respectively. The difference between AE ratios from the two populations was even higher if ER+ cases were isolated (Figures 6.5 and S6.6, Table 6.1). Given that results from the first experiment (first set of cases) did not show significant differences, a fourth experiment was performed with real-time PCR. Results were similar to the first experiment (Figure S6.7) albeit with a smaller effect size (g=-0.40). Although results showed the preferential expression of the reference allele in cases in all experiments, the inconsistencies between the results, with the measured effect size varying from significant to non-significant, suggesting that this variant is unlikely to be associated with breast cancer risk.

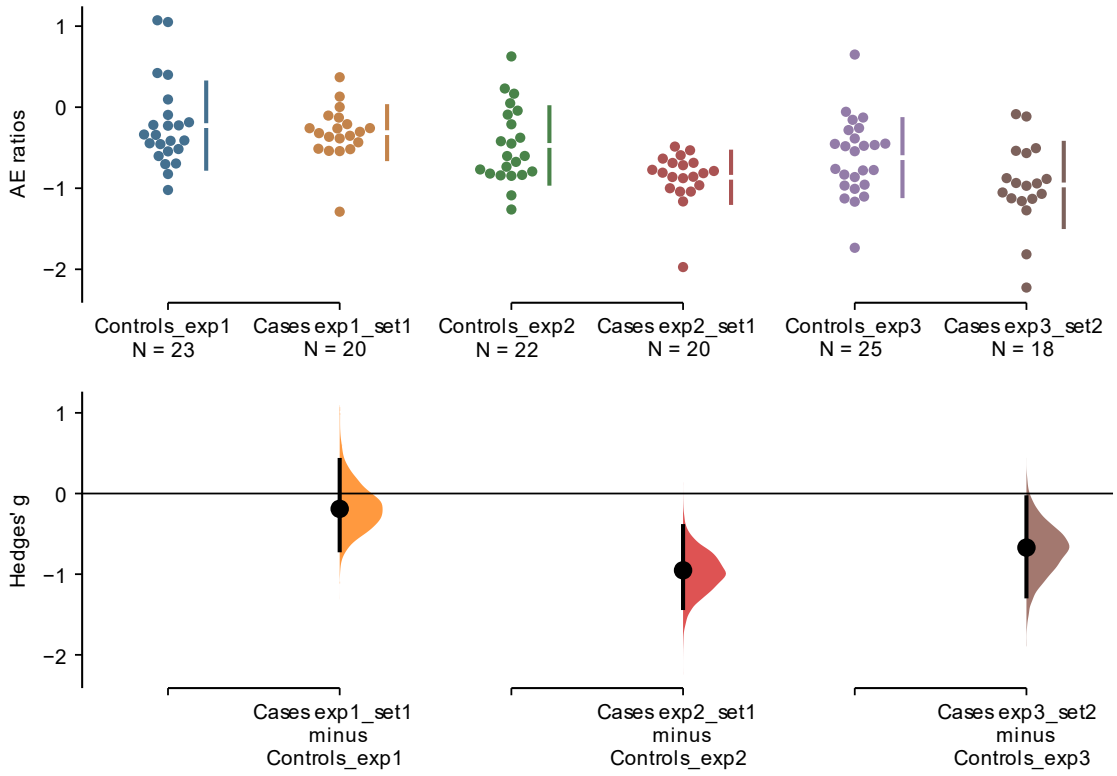


Figure 6.5. Case-control study using allelic expression (AE) ratios measured at rs62449782 (T/G, *ARL4A* gene) in normal breast tissue with the Biomark™ HD system (Fluidigm). In the upper y-axis dots represent the heterozygous individuals for the SNP, and the vertical lines correspond to the conventional mean ± standard deviation error bars, where the mean is indicated by the gap in the line. Each pair of controls-cases represents an experiment. The same set of controls was used for the three experiments (exp1, exp2, and exp3). One set of cases (set1) was used for experiments 1 and 2, and the second set of cases (set2) was used for the third experiment. The number of samples used in each case-control experiment is indicated (N). In the lower y-axis is a Cumming plot, with the Hedges' *g* effect size represented as the mean difference between cases and controls (black dots), and with the bootstrap (5000 resamples) 95% confidence interval (95% CI) indicated by the vertical black bars.

The variant rs2281791, located on *TBC1D12*, revealed a shift in the distribution of AE ratios in cases towards the preferential expression of the alternative allele in the first experiment, but in subsequent experiments, the reference allele was preferentially expressed (Figure 6.6 and S6.8, Table 6.1). Hence, even though large effect sizes (*g* values ranging from -1.02 to 1.12) were observed for this SNP, either considering all cases or separating them by ER status, results do not undoubtably support the association of this variant with breast cancer susceptibility.

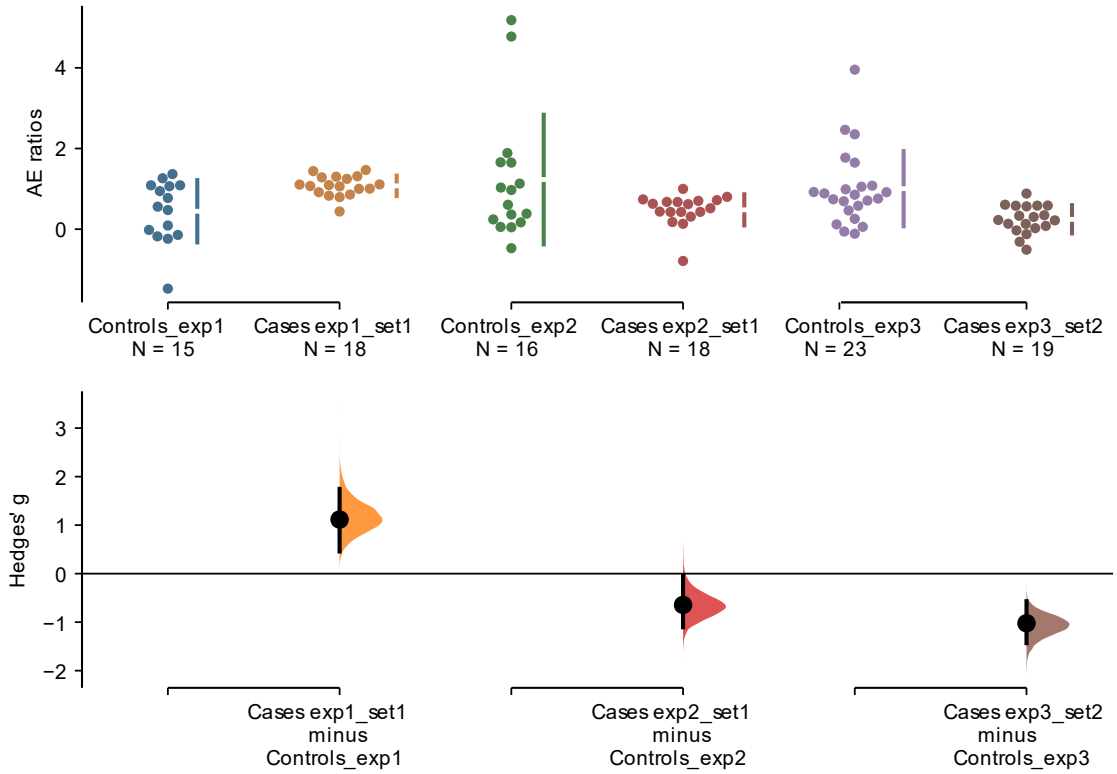


Figure 6.6. Case-control study using allelic expression (AE) ratios measured at rs2281791 (T/C, *TBC1D12* gene) in normal breast tissue with the Biomark™ HD system (Fluidigm). In the upper y-axis dots represent the heterozygous individuals for the SNP, and the vertical lines correspond to the conventional mean ± standard deviation error bars, where the mean is indicated by the gap in the line. Each pair of controls-cases represents an experiment. The same set of controls was used for the three experiments (exp1, exp2, and exp3). One set of cases (set1) was used for experiments 1 and 2, and the second set of cases (set2) was used for the third experiment. The number of samples used in each case-control experiment is indicated (N). In the lower y-axis is a Cumming plot, with the Hedges' *g* effect size represented as the mean difference between cases and controls (black dots), and with the bootstrap (5000 resamples) 95% confidence interval (95% CI) indicated by the vertical black bars.

Table 6.1. Association between AE ratios and breast cancer risk in breast tissue at the six candidate variants from the case-control study with RNA-seq data. exp – experiment; set1 – first set of cases; set2 – second set of cases; n – number of samples; CIinf and CIsup – inferior and superior limits of the 95% confidence interval (CI) of the Hedges' *g*; p-value – Mann-Whitney test p-value.

			All cases					ER+				ER-			
rsID	**Gene**	**Experiment**	**controls (n)**	**cases (n)**	**Hedges' *g***	**[95% CI CIinf CIsup]**	**p-value**	**cases (n)**	**Hedges' *g***	**[95% CI CIinf CIsup]**	**p-value**	**cases (n)**	**Hedges' *g***	**[95% CI CIinf CIsup]**	**p-value**
rs530963	CCDC86	Fluidigm_exp1_set1	15	10	-0.26	[-1.20, 0.59]	0.21	5	-0.49	[-1.54, 0.34]	0.14	4	-0.04	[-1.10, 1.07]	0.73
		Fluidigm_exp2_set1	17	13	0.29	[-0.51, 0.94]	0.50	6	0.38	[-0.90, 1.17]	0.28	5	-0.24	[-1.41, 0.29]	0.31
rs11724432	CCNG2	Fluidigm_exp1_set1	16	20	5.77×10^{-3}	[-0.73, 0.74]	0.86	11	-0.02	[-0.79, 0.68]	0.86	7	0.03	[-0.78, 0.68]	0.97
		Fluidigm_exp2_set1	15	19	6.82×10^{-2}	[-0.78, 0.67]	0.65	10	0.07	[-0.64, 0.75]	0.98	7	-0.26	[-1.28, 0.56]	0.40
rs62449782	ARL4	Fluidigm_exp1_set1	23	20	-0.19	[-0.71, 0.46]	0.86	11	-0.19	[-0.80, 0.40]	0.63	7	-0.02	[-0.49, 0.88]	0.81
		Fluidigm_exp2_set1	22	20	-0.95	[-1.42, -0.40]	7.89×10^{-3}	11	-1.09	[-1.64, -0.50]	4.44×10^{-3}	7	-0.50	[-0.93, 0.03]	0.52
		Fluidigm_exp3_set2	25	18	-0.67	[-1.23, -0.02]	2.43×10^{-2}	10	-0.99	[-1.74, -0.31]	1.12×10^{-2}	7	-0.49	[-1.19, 0.39]	0.14
		CFX384_set1	23	20	-0.40	[-0.92, 0.19]	0.34	11	-0.44	[-1.20, 0.22]	0.44	7	-0.29	[-1.02, 0.29]	0.52
rs2281791	TBC1D12	Fluidigm_exp1_set1	15	18	1.12	[0.42, 1.72]	9.74×10^{-3}	11	0.97	[0.36, 1.60]	3.33×10^{-2}	5	0.92	[0.42, 1.50]	5.48×10^{-2}
		Fluidigm_exp2_set1	16	18	-0.65	[-1.10, -0.04]	0.42	11	-0.61	[-1.06, -0.07]	0.34	5	-0.44	[-0.81, 0.08]	0.84
		Fluidigm_exp3_set2	23	19	-1.02	[-1.45, -0.58]	7.09×10^{-4}	15	-0.91	[-1.32, -0.48]	3.11×10^{-3}	4	-1.10	[-1.73, -0.48]	1.85×10^{-2}
11545332	DDX11	Fluidigm_exp1_set1	24	19	1.27	[0.61, 1.87]	4.92×10^{-4}	10	1.37	[0.63, 2.03]	1.82×10^{-3}	6	1.21	[0.21, 1.94]	1.83×10^{-2}
		Fluidigm_exp2_set1	23	19	1.46	[0.66, 2.21]	1.22×10^{-4}	10	1.73	[0.81, 2.59]	7.03×10^{-4}	6	1.42	[0.06, 2.65]	9.02×10^{-3}
		Fluidigm_exp3_set2	24	21	0.45	[-0.21, 1.00]	0.12	13	0.30	[-0.47, 0.99]	0.44	7	0.71	[-0.03, 1.33]	8.46×10^{-2}
		CFX384_set1	25	19	1.57	[0.95, 2.15]	1.45×10^{-5}	10	1.51	[0.71, 2.11]	3.22×10^{-4}	6	1.63	[0.60, 2.30]	1.50×10^{-3}
		CFX384_set2	25	33	-0.27	[-0.83, 0.30]	0.16	21	-0.36	[-0.95, 0.24]	0.12	11	-0.08	[-0.76, 0.40]	0.49
		CFX384_set1+set2	25	52	0.36	[-0.07, 0.84]	0.31	31	0.22	[-0.30, 0.73]	0.66	17	0.51	[-0.09, 1.09]	0.27
rs3211416	CDC16	Fluidigm_exp1_set1	15	21	-0.84	[-1.46, -0.16]	5.30×10^{-2}	13	-0.93	[-1.64, -0.22]	5.30×10^{-2}	7	-0.62	[-1.48, 0.52]	0.26
		Fluidigm_exp2_set1	16	21	-1.23	[-2.06, -0.20]	3.99×10^{-4}	13	-1.14	[-1.99, -0.05]	1.27×10^{-3}	7	-1.08	[-2.13, -0.21]	1.47×10^{-2}
		Fluidigm_exp3_set2	23	22	0.34	[-0.37, 0.84]	0.30	16	0.35	[-0.37, 0.82]	0.22	5	-0.15	[-0.30, 0.60]	1
		CFX384_exp1_set1	21	21	-2.03	[-2.66, -1.33]	5.56×10^{-7}	13	-2.14	[-2.88, -1.42]	8.00×10^{-6}	7	-0.52	[-2.34, -0.85]	8.30×10^{-4}
		CFX384_exp1_set2	21	28	-1.59	[-2.20, -0.80]	3.21×10^{-6}	20	-1.54	[-2.18, -0.82]	1.18×10^{-5}	7	-1.20	[-1.90, -0.54]	3.83×10^{-3}
		CFX384_exp1_set1+set2	21	49	-1.89	[-2.45, -1.12]	3.57×10^{-8}	33	-1.85	[-2.39, -1.19]	1.74×10^{-7}	14	-1.48	[-2.18, -0.77]	7.08×10^{-5}
		CFX384_exp2_set1	21	21	-1.97	[-2.59, -1.33]	5.17×10^{-7}	13	-2.05	[-2.77, -1.39]	5.70×10^{-6}	7	-1.45	[-2.18, -0.83]	9.10×10^{-4}
		CFX384_exp2_set2	21	28	-1.49	[-2.09, -0.84]	1.16×10^{-5}	20	-1.43	[-2.03, -0.79]	3.35×10^{-5}	7	-1.19	[-1.86, -0.58]	4.16×10^{-3}
		CFX384_exp2_set1+set2	21	49	-1.83	[-2.38, -1.14]	9.04×10^{-8}	33	-1.77	[-2.31, -1.10]	3.36×10^{-7}	14	-1.45	[-2.08, -0.78]	8.14×10^{-5}
		CFX384_exp3_set1	21	21	-2.04	[-2.72, -1.34]	1.20×10^{-6}	13	-2.15	[-2.87, -1.41]	1.30×10^{-5}	7	-1.48	[-2.26, -0.82]	4.21×10^{-3}
		CFX384_exp3_set2	21	28	-1.63	[-2.29, -0.86]	4.30×10^{-6}	20	-1.60	[-2.28, -0.86]	1.25×10^{-5}	7	-1.21	[-1.90, -0.49]	5.80×10^{-3}
		CFX384_exp3_set1+set2	21	49	-1.96	[-2.57, -1.16]	6.82×10^{-8}	33	-1.93	[-2.52, -1.15]	2.42×10^{-7}	14	-1.47	[-2.17, -0.74]	1.32×10^{-4}

6.4.7. Case-control studies using real-time PCR show a potential association of rs11545332 in *DDX11* and rs3211416 in *CDC16* with breast cancer risk

For case-control studies performed with breast tissue, the SNP rs11545332, located on *DDX11* showed a large effect size with a Hedges' *g* over 1.2 (Table 6.1) for experiments conducted with the first set of cases, either using digital or real-time PCR (Figure 6.7). Analysis of AE ratios showed a preferential expression of the alternative allele A in the normal breast of cases, which is the least expressed allele in controls. Conversely, results with the second set of cases were not concurrent, with an effect size lower than 0.5, in both platforms. In fact, for the real-time experiment the effect size was in the opposite direction (*g*=-0.27). Since the single experiment conducted by real-time PCR included both sets of cases, results were also analysed considering all cases together. In this case, the observed effect size was small (*g*=0.36). Separation of cases by ER status did not show an association (Figure 6.8 and S6.6 and Table 6.1). Considering these results, the association of rs11545332 with breast cancer risk was not confirmed and no further experiments were conducted.

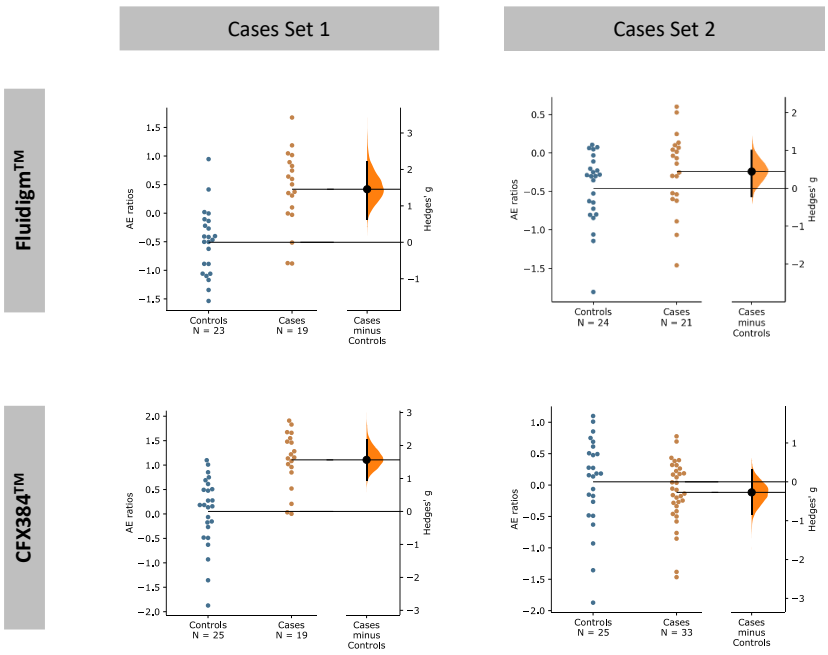


Figure 6.7. Case-control study using allelic expression (AE) ratios measured at rs11545332 (G/A, *DDX11* gene) in normal breast tissue with the digital PCR system (Biomark™ HD system, Fluidigm) and with the real-time PCR system (CFX384 Real-Time system, BioRad). In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNP (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows one experiment per set of cases (set 1 and set 2) for each PCR system.

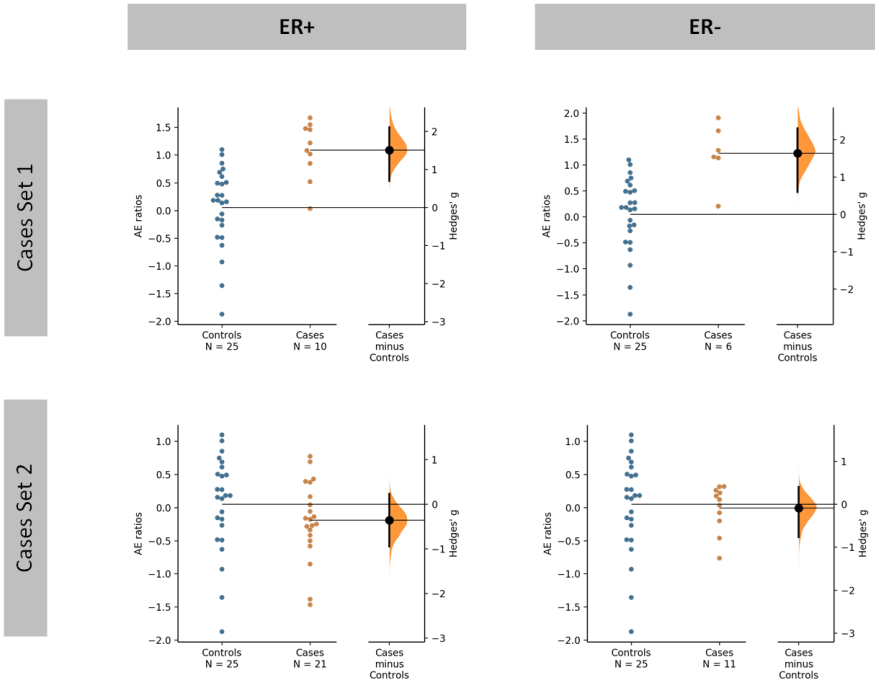


Figure 6.8. Case-control study by oestrogen receptor (ER) status from cases using allelic expression (AE) ratios measured at rs11545332 (G/A, *DDX11* gene) in normal breast tissue. Experiments were conducted with the real-time PCR system (CFX384 Real-Time system, BioRad). In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNP (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows one experiment per set of cases (set 1 and set 2).

Case-control studies performed with breast tissue using rs3211416, located on *CDC16*, revealed a shift in the distribution of AE ratios in cases towards the preferential expression of the reference allele C in all experiments performed by Fluidigm PCR and real-time PCR (Figures 6.9 and S6.8-S6.11).

Except for experiment 3 with set 2 of cases conducted with Fluidigm, the observed effect sizes ranged from *g*=-0.84 (Fluidigm experiment 1, set 1) to *g*=-2.04 (CFX384 experiment 3, set 1), indicating an association of the reference C-allele of rs3211416 with susceptibility for breast cancer (Table 6.1). Analysis of the AE ratios distribution according to ER status from cases showed all distributions being very similar, with slightly smaller effects sizes for ER- women (Figure 6.10 and S6.8).

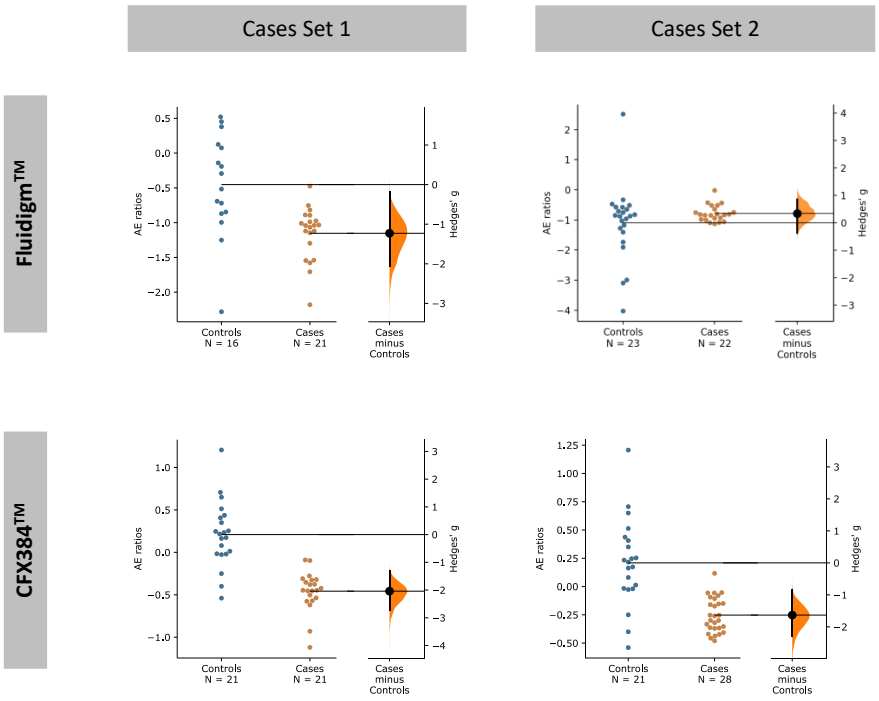


Figure 6.9. Case-control study using allelic expression (AE) ratios measured at rs3211416 (C/T, *CDC16* gene) in normal breast tissue with the digital PCR system (Biomark™ HD system, Fluidigm) and with the real-time PCR system (CFX384 Real-Time system, BioRad). In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNP (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows one experiment per set of cases (set 1 and set 2) for each PCR system.

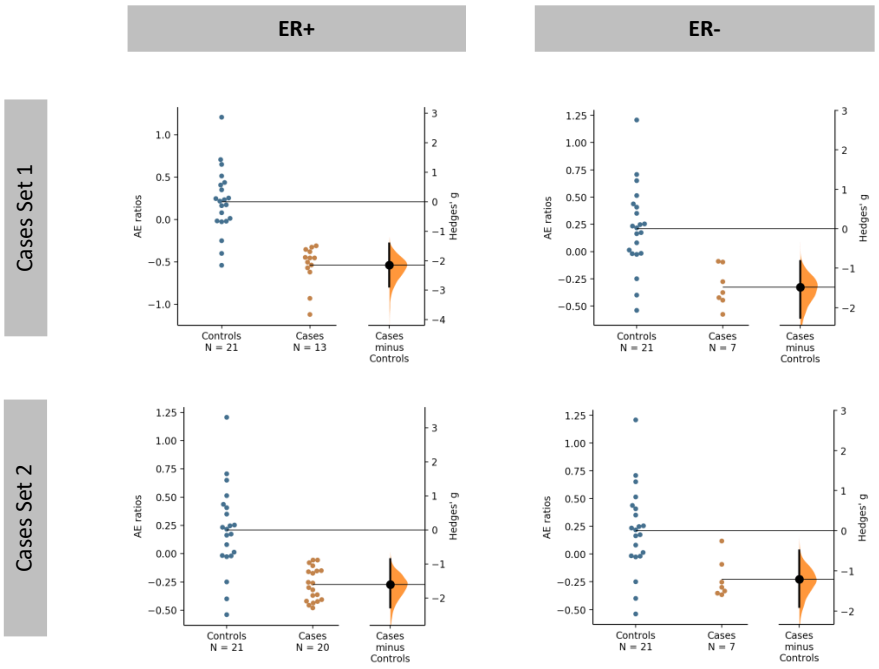


Figure 6.10. Case-control study by oestrogen receptor (ER) status from cases using allelic expression (AE) ratios measured at rs3211416 (C/T, *CDC16* gene) in normal breast tissue. Experiments were conducted with the real-time PCR system (CFX384 Real-Time system, BioRad). In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNP (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows one experiment per set of cases (set 1 and set 2).

Based on the evidence of cis-regulation of genes in breast cancer common to breast tissue and blood (Maia, et al., 2009; Darabi, et al., 2016), the association of AE at rs11545332 (*DDX11*) and the rs3211416 (*DCD16*) was examined in blood. Unlike the results in chapter IV, no significant difference was observed for the distribution of AE ratios in blood (Figure 6.11 and Table 6.2).

Table 6.2. Association between AE ratios and breast cancer risk in blood at 2 candidate variants from the case-control study with RNA-seq data. exp – experiment; set1 – the first set of cases; set2 – the second set of cases; n -number of samples; CIinf and CIsup – inferior and superior limits of the 95% confidence interval (CI) of the Hedges' *g*; p-value – Mann-Whitney test p-values.

			All cases				
rsID	Gene	Experiment	n controls	n cases	Hedges' *g*	[95% CI CIinf CIsup]	p-value
11545332	DDX11	CFX384_exp1	12	24	-0.32	[-1.05, 0.34]	0.29
		CFX384_exp2	11	23	-0.24	[-0.92, 0.51]	0.54
rs3211416	CDC16	CFX384_exp1	10	22	-0.43	[-1.21, 0.39]	0.31
		CFX384_exp1	10	22	-0.29	[-0.97, 0.48]	0.28
		CFX384_exp3	10	22	-0.26	[-0.97, 0.49]	0.48

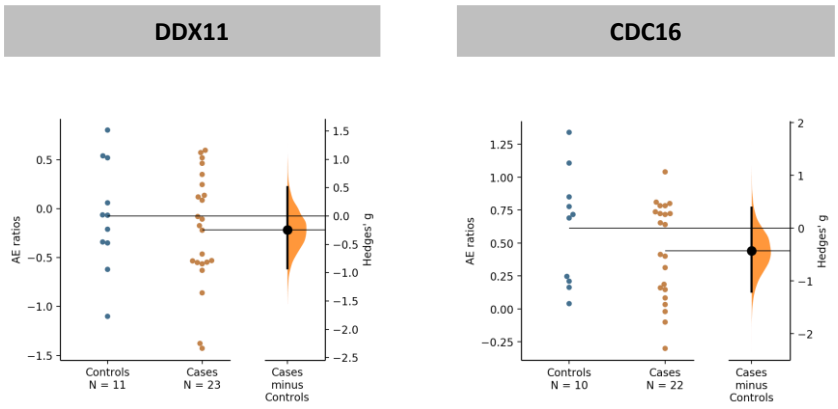


Figure 6.11. Case-control study using allelic expression (AE) ratios measured at rs11545332 (G/A, *DDX11* gene) and at rs3211416 (C/T, *CDC16* gene) in blood. Experiments were conducted with the real-time PCR system (CFX384 Real-Time system, BioRad). In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows one experiment per set of cases (set 1 and set 2).

6.5. Discussion

This work represents an innovative design to study the genetic breast cancer susceptibility, consisting of a genome-wide association analysis of AE measured in normal breast tissue. RNA-seq data were obtained, processed, and analysed for 14 cases and 12 control samples. A stepwise approach was used to determine AE ratios from RNA-seq read counts (100 bp paired-end). Only heterozygous bi-allelic genetic variants with 15 reads or more for each allele, from at least three controls and three cases, were analysed, resulting in 7,054 informative genetic variants. One new breast cancer risk locus was identified on 13q34 through the significantly different AE of *CDC16* association between cases and controls.

Different numbers of read counts have been reported for AE quantification in previous studies, ranging from 10 to 50 (Heap, et al., 2010; Castel, Levy-Moonshine, Mohammadi, Banks, & Lappalainen, 2015; Harvey, et al., 2015; Edsgärd, et al., 2016). Thus, and since accounting for genes with a lower number of mapped reads tackles the possible associated bias of the higher number of reads being absorbed by highly expressed genes (Finotello & Di Camillo, 2015), 15 reads from each allele, was the minimum number established for this study. Additionally, as these samples were newly sequenced, genotyping information was not available for all, and to eliminate possible genotype calling errors, the requirement to identify reads for both alleles

assures that all informative variants are indeed heterozygous. The minimum number of samples (three controls and three cases) to consider a variant for AE analysis was based on the "rule of three", according to which if *N* samples are tested and no effect is found, it is reasonable to estimate that the probability of occurrence is less than 3/*N* with a 95% confidence interval (considering a probability *P* of not finding the effect in *N* trials is 0.05) (Louis, 1981).

Variants located in the X-chromosome were also excluded to address X-chromosome inactivation bias, through which only one copy of the X chromosome is expressed. This results in a specific case of allelic expression where only the allele from the activated chromosome is transcribed.

The variants with significant differences (Student's t-test p-value<0.05) between cases and controls and with a $|\log_2(\text{fold-change})|>2$ between AE ratios from controls and cases were considered to rank candidate loci for the second phase of the association study. Using the fold-change between the expression of the alternative allele to the reference allele is a suitable measurement of cis-regulatory effect size (Mohammadi, Castel, Brown, & Lappalainen, 2017), and combining fold-change and statistical significance is considered to find more biologically meaningful sets of genes (McCarthy & Smyth, 2009). Additionally, the main reason for a design consisting of two phases of association study resulted primarily from a recognised limitation of this work: the statistical power to identify true differences. Power calculations indicated that to achieve the limit of statistical significance with correction for multiple testing (considering 7,054 tests p-correct<$7\text{x}10^{-6}$), with 95% confidence and a $|\log_2(\text{fold-change})|>2$, a minimum of five heterozygous individuals were needed in each compared group. As this number was unachievable in a small sample set for variants in Hardy-Weinberg equilibrium, we used this first phase to shortlist the variants to be tested in a larger sample size in the second phase with real-time PCR.

The first phase of this study identified eight candidate loci for association with breast cancer risk: rs34169189 (*PDE4DIP)*, rs62449782 (*ARL4A)*, rs11545332 (*DDX11)*, rs530963 (*CCDC86)*, rs757142894 (*APOC1P1)*, rs3211416 (*CDC16)*, rs11724432 (*CCNG2)*, and rs2281791 (*TBC1D12)*. These candidates also presented a large effect size considering Hedges' *g* (*g*>1.5), which was the measure for effect size used in subsequent case-control studies in breast tissue and blood by real-time PCR. The variant rs757142894 was excluded as a candidate after genotyping analysis confirmed incorrect mapping of reads between the *APOC1P1* pseudogene and the *APO1C* gene. The incorrect mapping of these reads can be explained by the use of a SNP-tolerant algorithm for the alignment of RNA-seq reads, which treats minor alleles as matches, rather than

mismatches to a reference sequence. To minimise false positive results such as this, a filtering step validating the observed MAF of each variant will be added to the RNA-seq data analysis pipeline for future studies.

Also, the results for the variant rs34169189 were not validated due to technical issues related to its genomic location at the beginning of the first open reading frame (nucleotide 30) of the *PDE4DIP* gene, which prevented the design of adequate primers and probes for AE quantification with real-time PCR. However, the potential association of this SNP with breast cancer susceptibility can be studied by sequencing the sequences including this variant after amplification by RT-PCR. Although not a straightforward process, it is possible to perform quantitative data analysis from Sanger sequencing using peak height or area ratios to measure allele proportions.

In the second phase of this study, RNA from breast tissue was analysed by digital real-time PCR to make the best use of the small amount of RNA available from each sample for rs62449782, rs11545332, rs530963, rs3211416, rs11724432, and rs2281791. Normal breast tissue is a difficult type of sample to have access to, particularly from healthy women. In addition, extraction of nucleic acids from normal breast tissue can be very challenging, given the fatty nature of the tissue, with yields lower than for other types of tissues (Mee, et al., 2011; McDonough, et al., 2019). Therefore, performing multiplex pre-amplification of cDNA also enabled an increase in the quantity of specific cDNA targets, and the number of possible real-time PCR reactions, without introducing amplification bias to the sample and without compromising the sensitivity of the PCR analysis. Additionally, the inclusion of a calibration curve with serial dilutions of DNA samples from CEPH cell lines heterozygous for the SNPs of interest, ensured the efficiency of the subsequent PCR reactions and the correct quantification of AE. Firstly, a set of controls and a set of cases were analysed in two separate experiments by digital PCR, for which six replicates were used. Two of the variants showed discordant results between experiments: rs530963 (*CCDC86*) and rs2281791 (*TBC1D12*). For these variants, analysis of the distribution of AE ratios showed opposite directions of the effect size in each one of the two experiments. In addition, the effect size of the SNP rs530963 was not significant (g=-0.26, 95%CI=-1.20-0.59, p-value=0.21; and g=0.29, 95%CI=-0.51-0.94, p-value 0.50), leading to the exclusion of this variant from other experiments. rs2281791 (*TBC1D12*) was still analysed with a second set of cases but results continued inconsistent, and this gene and variant were not validated as associated with risk. The variant rs11724432 (*CCNG2*) was also excluded based on low g values with 95% confidence intervals crossing the 0 line.

Next, the second set of cases was used to test the remaining three SNPs with concurrent results from the first two experiments: rs62449782 (*ARL4*); rs11545332 (*DDX11*); and rs3211416 (*CDC16*). This confirmed the results obtained with the first set, with exception of rs3211416 (*CDC16*), for which the direction of the effect size shifted. It is noteworthy that, although one experiment using this second set of cases was performed by digital PCR, it was reported that the pre-amplified cDNA was in suboptimal conditions upon arrival for testing. Thus, these results were taken into consideration with caution, and results from case-control experiments using the first set of cases prevailed over the one with the second set of cases. Further experiments with real-time PCR, did not validate the association of rs62449782 (*ARL4*) and rs11545332 (*DDX11*) with breast cancer risk.

Finally, the differences in the distribution of AE ratios between normal breast samples from cases and controls of the SNP rs3211416 identified the *CDC16* gene as a strong candidate for association with risk for breast cancer. The predicted effect size is g=-1.83 (95% CI=-2.38, -1.14), and no association with ER status was found. Since cases express preferentially the reference C-allele and eQTL in normal breast tissue show an increase in *CDC16* expression associated with the same allele, the risk C-allele could be associated with *CDC16* functioning as an oncogene. CDC16 is a subunit of the anaphase-promoting complex/cyclosome (APC/C) whose primary function is to promote cell cycle progression, by tagging for degradation proteins that inhibit specific cell cycle transitions (Barford, 2011). Recently, CDC16 has been associated with melanoma as part of Tre2-Bub2-Cdc16 (TBC) family proteins. Interference of TBC family member 7 (TBC1D7), was associated with inhibition of proliferation, migration, and invasion of melanoma cells (Tang, et al., 2020). CDC27, another component of APC/C has also been associated with alterations in cell cycle and mitosis, as well as cancer pathogenesis and prognosis (Kazemi-Sefat, et al., 2021). Further functional analysis of the SNP rs3211416, may clarify what are the variant(s) and biological mechanisms responsible for the allelic expression imbalances.

Overall, case-control studies presented here show the advantage of using a quantitative variable in association studies, with a quantifiable phenotype closely related to the most reported mechanisms associated with risk: cis-regulatory variation. Also, it evidences the importance of using the tissue of origin of the disease, as gene expression is functionally informative but also very tissue-specific. On the one hand, it improves the chance of pinpointing precisely the target genes responsible for the disease; on the other hand, this represents a challenge as normal breast tissue from healthy women is a very limited source. Moreover, the approach previously established, presents the additional benefit of direct identification of the target gene, while showing that regulation of transcript levels underlies biologically the onset of the disease.

Subsequent functional analyses are needed to identify the true causal variant(s) associated with the *CDC16* gene, and the cis-regulatory mechanisms responsible for the association of this locus with breast cancer risk. The approach used in this work has the potential to be applied to additional case-control studies with RNA-seq data from public databases, like GTEx and TCGA, which will increase its statistical power to identify risk loci for breast cancer.

Chapter VII

General Discussion and Conclusion

7.1. General Discussion

Although breast cancer is currently the most diagnosed cancer in the world (Sung, et al., 2021), with approximately 30% of all cases being heritable or due to genetic risk (Lichtenstein, et al., 2000; Peto & Mack, 2000; Mucci, et al., 2016), half of these remain unexplained with currently known genetic risk factors (Wendt & Margolin, 2019). Given that high- and moderate-penetrance variants have a large impact on the disease risk (Mavaddat, Antoniou, Easton, & Garcia-Closas, 2010), the proportion of risk they account for is already known, but explains only about 25% of familial risk, as they are rare in the population. Thus, since 2007, genome-wide association studies have conducted extensive efforts to identify the remaining genetic risk, likely to be conferred by common variants with a small effect size (low-penetrance variants). Nevertheless, even with an increasing number of variants and individuals included in GWAS, to date, those large collaborative consortiums, identified common variants associated with risk that explain a further 18% of the familial risk (Michailidou, et al., 2017). Subsequent fine-mapping studies using integrated approaches identified new common variants associated with breast cancer, increasing the percentage ok known familial risk to 24% (Fachal, et al., 2020). This shows that the GWAS era was successful in associating hundreds of loci with the disease, but GWAS do not deliver a functional understanding of the biological mechanisms underlying the risk or the true causal variant(s) responsible for that effect and their target genes. Since GWAS rely on using one variant to represent all others in LD with it in the same haplotype, they are not able to distinguish between the statistically associated variants, leaving the identification of the functional variant still dependent on subsequent validation. Interestingly, about 90% of the GWAS variants are in non-coding regions of the genome (Hindorff, et al., 2009; Pastinen, 2010; Schaub, Boyle, Kundaje, Batzoglou, & Snyder, 2012; Maurano, et al., 2012; Edwards, Beesley, French, & Dunning, 2013) and are enriched in regulatory elements (Skelly, Ronald, & Akey, 2009; Maurano, et al., 2012; Gaffney, 2013; Battle & Montgomery, 2014). This suggests that the causal variants acting on the risk-associated loci from GWAS are likely altering the genetic regulation of one or more target genes.

Mostly, analysis of expression quantitative trait loci (Schadt, et al., 2003; Lappalainen, et al., 2013; GTEx Consortium, 2015), and analysis of allelic expression (Meyer, et al., 2008; Pickrell, et al., 2010; Sun, 2012; Hu, Sun, Tzeng, & Perou, 2015; van de Geijn, McVicker, Gilad, & Pritchard, 2015; Kumasaka, Knights, & Gaffney, 2016) are used to study genetic factors controlling gene expression. Since the expression of numerous human genes is predicted to be regulated by cis-acting variants (Dixon, et al., 2007; Göring, et al., 2007; Stranger, et al., 2007), their identification

can pinpoint causal variants and target genes. Contrarily to eQTLs that compare total gene expression across genotype groups, and are influenced by trans-acting factors, allelic expression isolates cis-effects by comparing both alleles in diploid heterozygotes.

Considering all this, the main goal of the present study was to develop a new and efficient approach to detect risk loci, target genes, and causal variants via the integration of allelic expression data in association studies.

A previous genome-wide analysis of differential allelic expression in breast tissue from healthy women, previously conducted in our group, provided a map of cis-regulation in the tissue of origin of breast cancer, and the starting point for the first aim of this work: identification of causal variants and target genes in candidate loci. This work showed that the integration of allelic expression data to map cis-regulatory variants in known risk loci is a powerful approach to identify causal variants and their target genes. In particular, transcribed SNPs for which the analysed heterozygous exhibit a specific pattern of allelic expression, with all the individuals expressing preferentially the same allele, are extremely helpful for the identification of causal variants since this AE pattern is indicative of high linkage disequilibrium with such variants. This first line of work initially identified three loci – 1q32.1, 16q23.2, and 17q22 –, associated with risk and with strong cis-regulatory potential. Further in-vitro analysis undoubtedly confirmed that cis-regulatory variation in locus 17q22, spanning and regulating the genes *TOM1L1*, *COX11* and *STXBP4*, is involved in the breast cancer risk. Two of the three genes from this locus, *COX11*, and *STXBP4*, had been previously suggested as candidate target genes, after fine-mapping studies (Darabi, et al., 2016), chromatin conformation analysis (Baxter, et al., 2018), and studies integrating allelic expression (Fachal, et al., 2020). The present work established a direct association between breast cancer risk and differential allelic expression of *COX11* and *STXBP4*. Case-control association analysis using the AE ratios measured in normal breast tissue showed the largest effect size (g=-1.237) for the daeSNP rs2628315, located in an intron of *STXB4*, which is in complete LD with the locus lead risk-variant rs2787486 identified in a previous study (Darabi, et al., 2016). Results show that AE ratio distribution in the normal breast from breast cancer patients, favours the expression of the reference G allele, suggesting that this allele, as well as the reference allele of rs2787486, are associated with increased risk. Besides, the distribution of AE ratios in cases shifted towards the preferential expression of the reference allele A of the daeSNP rs17817901, a variant shared by *TOM1L1* and *COX11* genes, although with a smaller effect size (g=-0.486). This variant is in strong LD with the risk-variant rs2787486, indicating that also both these genes are candidate targets associated with breast cancer. On the other hand, analysis of the daeSNP rs9899602, located on the *TOM1L1* gene, did not show

significant differences between cases and controls, suggesting *STXBP4* and *COX11* are the most likely target genes. Further case-control analysis in blood samples showed the same association between the expression of the reference alleles of rs2628315 and rs17817901 and increased breast cancer risk.

Thus, this work presented a novel approach to identify causal variants and novel risk-loci: integration of AE ratios as a quantitative variable in case-control association studies. A similar approach was presented by others (Valle, et al., 2008), but in that study, an AE cut-off value was defined to categorise cases and controls as showing differential allelic expression or not. Since the distribution of AE ratios differs between genes, establishing that cut-off is not only difficult but also loses statistical power when compared to the approach used in the present work. Moreover, in our approach, the effect size is measured with the Hedges' *g*, a standardised mean difference method, independent of the sample size. Thus, although a larger sample size than the one used in this work, would be an advantage, it should only tighten the confidence interval of the estimated effect size. A caveat of our approach is that it can only include individuals heterozygous for the transcribed variants, where the allelic expression is quantified. However, using AE ratios identifies all the cis-regulatory effects acting on a gene, which is of particular importance in complex genomic regions, and simultaneously identifying the mechanisms underlying risk and the target genes.

Once we established our approach, we next sought to test it in a genome-wide setting. The imbalance between the two alleles of a transcribed variant can be measured by counting reads from RNA-seq (Verlaan, et al., 2009). Thus, to test our approach we designed a case-control study using RNA-seq data. However, AE quantification from RNA-seq data presents several challenges. There is no single pipeline optimally suited for it, and the biggest caveat in AE analysis from RNA-seq data is the allelic mapping bias (Degner, et al., 2009; Stevenson, Coolon, & Wittkopp, 2013). Reads carrying the alternative allele will present more mismatches than the reads carrying the reference allele, when mapping them to a reference genome, and may fail to map uniquely. Although there are several tools described to improve AE analysis (Rozowsky, et al., 2011; Piskol, Ramaswami, & Li, 2013; Soderlund, Nelson, & Goff, 2014; Castel, Levy-Moonshine, Mohammadi, Banks, & Lappalainen, 2015; van de Geijn, McVicker, Gilad, & Pritchard, 2015; Miao, Alvarez, Pajukanta, & Ko, 2018), there is not a comprehensive comparison between pipelines using open-source tools for all the steps of a typical workflow for RNA-seq data analysis, and even less including a variant-aware aligner. So, the next step in this work was to compare forty-two pipelines for variant calling and subsequent AE analysis. Paired-end RNA-

seq data from two samples were used: NA12878 which is the only sample with a gold standard variant genotype data set publicly available; and a sample of normal breast tissue. Pipeline' comparison included the combination of seven trimming tools, two aligners, and two variant callers. Results showed that Trimmomatic was the only trimming tool able to remove all the adapters from RNA-seq data from both samples and removed more low-quality bases in the breast tissue sample. GSNAP, the variant-aware aligner included in this work, mapped a higher number of reads in pair and fewer singletons, with the best results being obtained after pre-processing with Trimmomatic. Both aligners, GSNAP and STAR, mapped a similar number of properly paired reads, but STAR mapped more reads at unique genomic locations. Since GSNAP is a SNP-tolerant aligner, when mapping to a reference genome, it can align reads harbouring alternative alleles to more than one genomic site, which is an advantage to bridge the mapping bias usually observed towards the reference allele, particularly when the aim is to precisely quantify allelic expression. Considering the variant callers compared, GATK identified a higher number of variants, including known and novel SNPs and indels, than SAMtools. After mapping gold standard data with GSNAP and STAR, GATK called two and four times more variants than SAMtools, respectively. For breast tissue data, these numbers increased to six (GSNAP) and eight times (STAR) more variants identified by GATK relatively to SAMtools. Moreover, the higher number of variants identified by GATK was not due to low-quality calls, since filtering of those still provided more variants using GATK. Although mapping with GSNAP identified more false positives, it also identified more true positives, fewer false negatives, and better F-measures than STAR, indicating a positive balance between precision and sensitivity. This balance should always be weighted according to the aim of the study. In this case, since the final goal was to identify new breast cancer risk loci through AE analysis of RNA-seq data, the detection of more true positives and fewer false negatives was favoured. To reduce the number of false positives, previous studies (Liu, Han, Wang, Gelernter, & Yang, 2013), suggested focusing only on the variants called simultaneously by several variant callers. According to the results obtained from this work, the established workflow for RNA-seq data analysis included: trimming with Trimmomatic, mapping with GSNAP, and variant calling with both GATK and SAMtools. Only variants common to both variant callers were considered suitable for further identification of cis-regulatory variants by AE analysis.

Finally, with the novel approach of case-control association analysis using AE ratios to identify risk, and the established RNA-seq data analysis pipeline for AE quantification, this work aimed at identifying new breast cancer risk loci. In a two-phase study, RNA-seq data analysis from normal breast tissue of 14 cases and 12 controls identified 7,054 informative genetic variants.

Then, using real-time PCR to quantify AE and using a larger number of RNA samples from normal breast tissue samples, identified a new risk-associated variant rs3211416 (*CDC16*). Cases preferentially expressed the reference C-allele of rs3211416 and controls the alternative T-allele. Concordantly, the C-allele is shown by eQTL to be associated with higher expression of *CDC16* in normal breast tissue, which supports the oncogenic function of *CDC16*, as previously reported in other types of tumours. Subsequent functional analyses, like the ones performed in this study for other loci, will identify the functional variant(s) associated with risk acting on this locus.

7.2. Conclusion

In conclusion, this work contributed to elucidate the role of cis-regulatory variation in breast cancer risk, by integrating bioinformatics with experimental laboratory work.

The main contributions of this work fall in two categories: technical/analytical and biological/clinical.

Technical/Analytical contributions:

1. Provided compelling evidence supporting the power of integrating allelic expression data in functional analysis of risk loci, leading to the identification of causal variants and target genes.
2. Provided an optimised and accurate pipeline for AE analysis from RNA-seq data.
3. Piloted the use of case-control association analysis using AE as a quantitative trait to improve the identification of risk associated with complex diseases, with the advantages of directly identify target genes while also uncovering the underlying biological mechanism.

Biological/Clinical contributions:

4. Identified several cis-regulatory variants responsible for the differential allelic expression observed in several genes associated with risk for breast cancer (*COX11*, *STXBP4*, *MDM4*, *PIK3C2B*) and its clinical features (*PIK3CA*).
5. Established COX11 and STXBP4 as target genes of the cis-regulatory variants rs17817901 and rs8066588 in the 17q22 locus, previously associated with cancer risk.

6. Identified two new risk biomarkers for breast cancer – AE ratios at rs17817901 and rs2628315 – with great potential for integration in clinical screening tests.
7. Identified a new risk locus, 13q34, via the association of allelic expression of the gene *CDC16* in normal breast tissue, in a genome-wide association analysis of AE.

Overall, findings presented here support the hypothesis that cis-regulatory variants play a major role in breast cancer susceptibility and have the potential to be applicable in future studies to identify causal variants and target genes in known and new risk loci for breast cancer or other complex diseases. The greater implications of this type of work are that of helping to improve population risk stratification and pre-disease management, as well as contributing towards developing future preventive therapies, especially with respect to personalised medicine.

References

1000 Genomes Project Consortium, A. G. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), pp. 1061-1073. doi:10.1038/nature09534

Abdulrahman, G. O., & Rahman, G. A. (2012). Epidemiology of Breast Cancer in Europe and Africa. *Journal of Cancer Epidemiology, 2012*. doi:10.1155/2012/915610

Adamus, A., Müller, P., Nissen, B., Kasten, A., Timm, S., Bauwe, H., . . . Engel, N. (2018). GCSH antisense regulation determines breast cancer cells' viability. *Scientific Reports, 8*(1), p. 15399. doi:10.1038/s41598-018-33677-4

Adank, M. A., van Mil, S. E., Gille, J. J., Waisfisz, Q., & Meijers-Heijboer, H. (2011). PALB2 Analysis in BRCA2-like Families. *Breast Cancer Research and Treatment, 127*(2), pp. 357-362. doi:10.1007/s10549-010-1001-1

Adomas, A. B., Grimm, S. A., Malone, C., Takaku, M., Sims, J. K., & Wade, P. A. (2014). Breast tumor specific mutation in GATA3 affects physiological mechanisms regulating transcription factor turnover. *BMC Cancer, 14*, p. 278. doi:10.1186/1471-2407-14-278

Adoue, V., Schiavi, A., Light, N., Almlöf, J. C., Lundmark, P., Ge, B., . . . Pastinen, T. (2014). Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. *Molecular Systems Biology, 10*(10), p. 754. doi:10.15252/msb.20145114

Afzaljavan, F., Sadr, A. S., Savas, S., & Pasdar, A. (2021). GATA3 somatic mutations are associated with clinicopathological features and expression profile in TCGA breast cancer patients. *Scientific Reports, 11*(1), p. 1679. doi:10.1038/s41598-020-80680-9

Ahmadiyeh, N., Pomerantz, M. M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., . . . Freedman, M. L. (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proceedings of the National Academy of Sciences of the United States of America, 107*(21), pp. 9742–9746. doi:10.1073/pnas.0910668107

Ahmed, M., & Rahman, N. (2006). ATM and breast cancer susceptibility. *Oncogene, 25*, pp. 5906-5911. doi:10.1038/sj.onc.1209873

Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R., . . . Easton, D. F. (2009). Newly Discovered Breast Cancer Susceptibility Loci on 3p24 and 17q23.2. *Nature Genetics, 41*(5), pp. 585-590. doi:10.1038/ng.354

Al Tamini, D. M., Shawarby, M. A., Ahmed, A., Hassan, A. K., & AlOdaini, A. A. (2010). Protein expression profile and prevalence pattern of the molecular classes of breast cancer: a Saudi population based study. *BMC Cancer, 10*, p. 223. doi:10.1186/1471-2407-10-223

Albain, K. S., Unger, J. M., Crowley, J. J., Coltman, C. A., & Hershman, D. L. (2009). Racial Disparities in Cancer Survival Among Randomized Clinical Trials Patients of the

Southwest Oncology Group. *Journal of the National Cancer Institute, 101*(14), pp. 984-992. doi:10.1093/jnci/djp175

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews. Genetics, 16*(4), pp. 197-212. doi:10.1038/nrg3891

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews. Genetics, 16*(4), pp. 197-212. doi:10.1038/nrg3891

Aleskandarany, M. A., Rakha, E. A., Macmillan, R. D., Powe, D. G., Ellis, I. O., & Green, A. R. (2010). MIB1/Ki-67 labelling index can classify grade 2 breast cancer into two clinically distinct subgroups. *Breast Cancer Research and Treatment, 127*(3), pp. 591-599. doi:ff10.1007/s10549-010-1028-3

Almlöf, J. C., Lundmark, P., Lundmark, A., Ge, B., Maouche, S., Göring, H. H., . . . Syvänen, A. C. (2012). Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. *PLoS One, 7*(12), p. e52260. doi:10.1371/journal.pone.0052260

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 Epidemiology, Biomarkers & Prevention, 26*(1), pp. 126-135. doi:10.1158/1055-9965.EPI-16-0106

Anderson, W. F., Chatterjee, N., Ershler, W. B., & Brawley, O. W. (2002). Estrogen receptor breast cancer phenotypes in the Surveillance, Epidemiology, and End Results database. *Breast Cancer Research and Treatment, 76*(1), pp. 27-36. doi:10.1023/a:1020299707510

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*. Obtido de http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Antoniou, A. C., & Easton, D. F. (2006). Models of Genetic Susceptibility to Breast Cancer. *Oncogene, 25*(43), pp. 5898-5905. doi:10.1038/sj.onc.1209879

Antoniou, A. C., Foulkes, W. D., & Tischkowitz, M. (2014). Breast-Cancer Risk in Families with Mutations in PALB2. *The New England Journal of Medicine, 371*(17). doi:10.1056/NEJMoa1400382

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. *42*(10), pp. 885-892. doi:10.1038/ng.669

Antoniou, A., Pharoah, P. D., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., . . . Easton, D. F. (2003). Average Risks of Breast and Ovarian Cancer Associated With BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies. *American journal of Human Genetics, 72*(5), pp. 1117-1130. doi:10.1086/375033

Apostolou, P., & Fostira, F. (2013). Hereditary Breast Cancer: The Era of New Susceptibility Genes. *BioMed Research International, 2013*, p. 747318. doi:10.1155/2013/747318

Aronesty, E. (2013). Comparison of Sequencing Utility Programs. *Open Bioinformatics Journal, 7*, pp. 1-8.

Azzato, E. M., Lee, A. J., Teschendorff, A., Ponder, B. A., Pharoah, P., Caldas, C., & Maia, A. T. (2010). Common germ-line polymorphism of C1QA and breast cancer survival. *British Journal of Cancer, 102*(8), pp. 1294–1299. doi:10.1038/sj.bjc.6605625

Badve, S., Dabbs, D. J., Schnitt, S. J., Baehner, F. L., Decker, T., Eusebi, V., . . . Reis-Filho, J. S. (2011). Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Modern Pathology, 24*, pp. 157-167. doi:10.1038/modpathol.2010.200

Baes, C. F., Dolezal, M. A., Koltes, J. E., Bapst, B., Fritz-Waters, E., Jansen, S., . . . Gredler, B. (2014). Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics, 15*(1), p. 948. doi:10.1186/1471-2164-15-948

Bailey, D. M., Affara, N. A., & Ferguson-Smith, M. A. (1992). The X-Y homologous gene amelogenin maps to the short arms of both the X and Y chromosomes and is highly conserved in primates. *Genomics, 14*(1), pp. 203-205. doi:doi: 10.1016/s0888-7543(05)80310-6

Bakin, A. V., & Curran, T. (1999). Role of DNA 5-methylcytosine transferase in cell transformation by fos. *Science, 283*(5400), pp. 387–390. doi:10.1126/science.283.5400.387

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., . . . Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genetics, 10*(9), p. e1004663. doi:10.1371/journal.pgen.1004663

Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E. K., Rivas, M. A., . . . Lappalainen, T. (2015). The landscape of genomic imprinting across diverse adult human tissues. *Genome Research, 25*(7), pp. 927-936. doi:10.1101/gr.192278.115

Barford, D. (2011). Structural insights into anaphase-promoting complex function and mechanism. *Philosophical Transactions of the Royal Society of London, 366*(1584), pp. 3605–3624. doi:10.1098/rstb.2011.0069

Barnhart, B. C., Lee, J. C., Alappat, E. C., & Peter, M. E. (2003). The Death Effector Domain Protein Family. *Oncogene, 22*(53), pp. 8634-8644. doi:10.1038/sj.onc.1207103

Bartkova, J., Tommiska, J., Oplustilova, L., Aaltonen, K., Tamminen, A., Heikkinen, T., . . . Bartek, J. (2008). Aberrations of the MRE11-RAD50-NBS1 DNA Damage Sensor Complex in Human Breast Cancer: MRE11 as a Candidate Familial Cancer-Predisposing Gene. *Molecular Oncology, 2*(4), pp. 296-316. doi:10.1016/j.molonc.2008.09.007

Battle, A., & Montgomery, S. B. (2014). Determining causality and consequence of expression quantitative trait loci. *Human Genetics, 133*(6), pp. 727–735. doi:10.1007/s00439-014-1446-0

Baxter, J. S., Leavy, O. C., Dryden, N. H., Maguire, S., Johnson, N., Fedele, V., . . . Fletcher, O. (2018). Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nature Communications, 9*(1), p. 1028. doi:10.1038/s41467-018-03411-9

Bayraktar, S., Thompson, P. A., Yoo, S. Y., Do, K. A., Sahin, A. A., Arun, B. K., . . . Brewster, A. M. (2013). The relationship between eight GWAS-identified single-nucleotide polymorphisms and primary breast cancer outcomes. *The Oncologist, 18*(5), pp. 493–500. doi:10.1634/theoncologist.2012-0419

Beaber, E. F., Malone, K. E., Tang, M. T., Barlow, W. E., Porter, P. L., Daling, J. R., & Li, C. I. (2014). Oral Contraceptives and Breast Cancer Risk Overall and by Molecular Subtype Among Young Women. *Cancer Epidemiology, Biomarkers & Prevention, 23*(5), pp. 755-764. doi:10.1158/1055-9965.EPI-13-0944

Beesley, J., Pickett, H. A., Johnatty, S. E., Dunning, A. M., Chen, X., Li, J., . . . Consortium, O. C. (2011). Functional polymorphisms in the TERT promoter are associated with risk of serous epithelial ovarian and breast cancers. *PlOS One, 6*(9), p. e24987. doi:10.1371/journal.pone.0024987

Bell, D. W., Varley, J. M., Szydlo, T. E., Kang, D. H., Wahrer, D. C., Shannon, K. E., . . . Haber, D. A. (1999). Heterozygous Germ Line hCHK2 Mutations in Li-Fraumeni Syndrome. *Science, 286*(5449), pp. 1518-2531. doi:10.1126/science.286.5449.2528

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., . . . Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology, 12*(1), p. R10. doi:10.1186/gb-2011-12-1-r10

Belluti, S., Basile, V., Benatti, P., Ferrari, E., Marverti, G., & Imbriano, C. (2013). Concurrent inhibition of enzymatic activity and NF-Y-mediated transcription of Topoisomerase-IIα by bis-DemethoxyCurcumin in cancer cells. *Cell Death & Disease, 4*(8), p. e756. doi:10.1038/cddis.2013.287

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics, 29*(4), pp. 1165-1188.

Benusiglio, P. R., Malka, D., Rouleau, E., De Pauw, A., Buecher, B., Noguès, C., . . . Caron, O. (2013). CDH1 Germline Mutations and the Hereditary Diffuse Gastric and Lobular Breast Cancer Syndrome: A Multicentre Study. *Journal of Medical Genetics, 50*(7), pp. 486-489. doi:10.1136/jmedgenet-2012-101472

Benzer, S. (1955). FINE STRUCTURE OF A GENETIC REGION IN BACTERIOPHAGE. *Proceedings of the National Academy of Sciences of the United States of America, 41*(6), pp. 344-354. doi:10.1073/pnas.41.6.344

Bernstein, L., Henderson, B. E., Hanisch, R., Sullivan-Halley, J., & Ross, R. K. (1994). Physical exercise and reduced risk of breast cancer in young women. *Journal of the National Cancer Institute, 86*(18), pp. 1403-1408. doi:10.1093/jnci/86.18.1403

Bertrand, K. A., Tamimi, R. M., Scott, C. G., Jensen, M. R., Pankratz, V. S., Visscher, D., . . . Vachon, C. .. (2013). Mammographic density and risk of breast cancer by age and tumor characteristics. *Breast Cancer Research, 15*, p. R104. doi:10.1186/bcr3570

Berx, G., Cleton-Jansen, A. M., Nollet, F., de Leeuw, W. J., van de Vijver, M., Cornelisse, C., & van Roy, F. (1995). E-cadherin Is a Tumour/Invasion Suppressor Gene Mutated in Human Lobular Breast Cancers. *The EMBO Journal, 14*(24), pp. 6107-6115.

Bessaoud, F., & Daurès, J. P. (2008). Patterns of Alcohol (Especially Wine) Consumption and Breast Cancer Risk: A Case-Control Study Among a Population in Southern France. *Annals of Epidemiology, 18*(7), pp. 467-475. doi:10.1016/j.annepidem.2008.02.001

Bethesda (MD): National Center for Biotechnology Information, N. L. (s.d.). *Database of Single Nucleotide Polymorphisms (dbSNP)*. Obtido de http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/snp150.txt.gz

Betts, J. A., Moradi Marjaneh, M., Al-Ejeh, F., Lim, Y. C., Shi, W., Sivakumaran, H., . . . French, J. D. (2017). Long Noncoding RNAs CUPID1 and CUPID2 Mediate Breast Cancer Risk at 11q13 by Modulating the Response to DNA Damage. *American Journal of Human Genetics, 101*(2), pp. 255–266. doi:10.1016/j.ajhg.2017.07.007

Beute, B. J., Kalisher, L., & Hutter, R. V. (1991). Lobular carcinoma in situ of the breast: clinical, pathologic, and mammographic features. *American Journal of Roentgenology, 157*(2), pp. 257-265. doi:10.2214/ajr.157.2.1853802

Bhargava, R., & Dabbs, D. J. (2008). Luminal B breast tumors are not HER2 positive. *Breast Cancer Research, 10*(5), p. 404. doi:10.1186/bcr2134

Bhargava, R., Striebel, J., Beriwal, S., Flickinger, J. C., Onisko, A., Ahrendt, G., & Dabbs, D. J. (2009). Prevalence, Morphologic Features and Proliferation Indices of Breast Carcinoma Molecular Classes Using Immunohistochemical Surrogate Markers. *International journal of clinical and experimental pathology, 2*(5), pp. 444-445.

Bilguun, E. O., Kaira, K., Kawabata-Iwakawa, R., Rokudai, S., Shimizu, K., Yokobori, T., . . . Nishiyama, M. (2020). Distinctive roles of syntaxin binding protein 4 and its action target, TP63, in lung squamous cell carcinoma: a theranostic study for the precision medicine. *BMC Cancer, 20*(1). doi:10.1186/s12885-020-07448-2

Bilimoria, M. M., & Morrow, M. (1995). The Woman at Increased Risk for Breast Cancer: Evaluation and Management Strategies. *CA: A Cancer Journal for Clinicians, 45*(5), pp. 263-278. doi:10.3322/canjclin.45.5.263

Bland, K. I., Konstadoulakis, M. M., Vezeridis, M. P., & Wanebo, H. J. (1995). Oncogene protein co-expression. Value of Ha-ras, c-myc, c-fos, and p53 as prognostic discriminants for

breast carcinoma. *Annals of Surgery, 221*(6), pp. 706-720. doi:10.1097/00000658-199506000-00010

Bogdanova, N., Enssen-Dubrowinskaja, N., Feshchenko, S., Lazjuk, G. I., Rogov, Y. I., Dammann, O., . . . Dörk, T. (2005). Association of Two Mutations in the CHEK2 Gene With Breast Cancer. *International Journal of Cancer, 116*(2), pp. 263-266. doi:10.1002/ijc.21022

Bogdanova, N., Feshchenko, S., Schürmann, P., Waltes, R., Wieland, B., Hillemanns, P., . . . Dörk, T. (2008). Nijmegen Breakage Syndrome Mutations and Risk of Breast Cancer. *International Journal of Cancer, 122*(4), pp. 802-806. doi:10.1002/ijc.23168

Bojesen, S. E., Pooley, K. A., Johnatty, S. E., Beesley, J., Michailidou, K., Tyrer, J. P., . . . Dunning, A. M. (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature Genetics, 45*(4), pp. 371–384e3842. doi:10.1038/ng.2566

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), pp. 2114–2120. doi:10.1093/bioinformatics/btu170

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), pp. 1790-1797. doi:10.1101/gr.137323.112

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A cancer Journal for Clinicians, 68*(6), pp. 394-424. doi:10.3322/caac.21492

Bray, N. J., Moskvina, V., Buxbaum, J. D., Dracheva, S., Haroutunian, V., Williams, J., . . . O'Donovan, M. C. (2004). Allelic expression of APOE in human brain: effects of epsilon status and promoter haplotypes. *Human Molecular Genetics, 13*(22), pp. 2885–2892. doi:10.1093/hmg/ddh299

Breast Cancer Association Consortium. (2006). Commonly Studied Single-Nucleotide Polymorphisms and Breast Cancer: Results From the Breast Cancer Association Consortium. *Journal of the National Cancer Institute, 98*(19), pp. 1382-1396. doi:10.1093/jnci/djj374

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science, 296*(5568), pp. 752-755. doi:10.1126/science.1069516

Brewer, H. R., Jones, M. E., Schoemaker, M. J., Ashworth, A., & Swerdlow, A. J. (2017). Family history and risk of breast cancer: an analysis accounting for family structure. *Breast cancer research and treatment, 165*(1), pp. 193-200. doi:10.1007/s10549-017-4325-2

Broekema, R. V., Bakker, O. B., & Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *10*(1), p. 190221. doi:10.1098/rsob.190221

Broeks, A., Schmidt, M. K., Sherman, M. E., Couch, F. J., Hopper, J. L., Dite, G. S., . . . Garcia-Closas, M. (2011). Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. *Human Molecular Genetics, 20*(16), pp. 3289–3303. doi:10.1093/hmg/ddr228

Buckland, P. R. (2006). The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochimica et Biophysica Acta, 1762*(1), pp. 17–28. doi:10.1016/j.bbadis.2005.10.004

Buonocore, F. H., Oladimeji, P. B., Jeffries, A. R., Troakes, C., Hortobagyi, T., Williams, B. P., . . . Bray, N. J. (2010). Effects of cis-regulatory variation differ across regions of the adult human brain. *Human Molecular Genetics, 19*(22), pp. 4490-4496. doi:10.1093/hmg/ddq380

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide Association Studies. *PLoS Computational Biology, 8*(12), p. e1002822. doi:10.1371/journal.pcbi.1002822

C., L. L. (1982). Photoreactivation of amanitin-inhibited RNA polymerase II. *The Journal of Biological Chemistry, 257*(4), pp. 1577–1578.

Cai, Q., Long, J., Lu, W., Qu, S., Wen, W., Kang, D., . . . Zheng, W. (2011). Genome-wide Association Study Identifies Breast Cancer Risk Variant at 10q21.2: Results From the Asia Breast Cancer Consortium. *Human Molecular Genetics, 20*(24), pp. 4991-4999. doi:10.1093/hmg/ddr405

Cai, Q., Wen, W., Qu, S., Li, G., Egan, K. M., Chen, K., . . . … Zheng, W. (2011). Replication and functional genomic analyses of the breast cancer susceptibility locus at 6q25.1 generalize its importance in women of chinese, Japanese, and European ancestry. *Cancer Research, 71*(4), pp. 1344–1355. doi:10.1158/0008-5472.CAN-10-2733

Campbell, C. D., Kirby, A., Nemesh, J., Daly, M. J., & Hirschhorn, J. N. (2008). A survey of allelic imbalance in F1 mice. *Genome Research, 18*(4), pp. 555–563. doi:10.1101/gr.068692.107

Campino, S., Forton, J., Raj, S., Mohr, B., Auburn, S., Fry, A., . . . Kwiatkowski, D. P. (2008). Validating discovered Cis-acting regulatory genetic variants: application of an allele specific expression approach to HapMap populations. *PLoS One, 3*(12), p. e4105. doi:10.1371/journal.pone.0004105

Cancer Genome Atlas Network. (2012). Comprehensive Molecular Portraits of Human Breast Tumours. *Nature, 490*(7418), pp. 61-70. doi:10.1038/nature11412

Cantley, L. C. (2002). The phosphoinositide 3-kinase pathwa. *Science, 296*(5573), pp. 1655–1657. doi:10.1126/science.296.5573.1655

Carey, L. A., Perou, C. M., Dressler, L. G., Livasy, C. A., Geradts, J., Cowan, D., . . . Millikan, R. C. (2004). Race and the poor prognosis basal breast tumor (BBT) phenotype in the population-based Carolina Breast Cancer Study (CBCS). *Journal of Clinical Oncology, Suppl*, p. Abstr 9510. doi:10.1200/jco.2004.22.90140.9510

Carithers, L. J., & Moore, H. M. (2015). The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and biobanking, 13*(5), pp. 307–308. doi:10.1089/bio.2015.29031.hmm

Carr, H. S., Maxfield, A. B., Horng, Y. C., & Winge, D. R. (2005). Functional analysis of the domains in Cox11. *The Journal of Biological Chemistry, 280*(24), pp. 22664–22669. doi:10.1074/jbc.M414077200

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology, 16*(1), p. 195. doi:10.1186/s13059-015-0762-6

Chang, J., Zhou, Y., Hu, X., Lam, L., Henry, C., Green, E. M., . . . Fraser, H. B. (2013). The molecular mechanism of a cis-regulatory adaptation in yeast. *PLoS Genetics, 9*(9), p. e1003813. doi:10.1371/journal.pgen.1003813

Cheang, M. C., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., . . . Nielsen, T. O. (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *JNCI: Journal of the National Cancer Institute, 101*(10), pp. 736-750. doi:10.1093/jnci/djp082

Chen, X., Weaver, J., Bove, B. A., Vanderveer, L. A., Weil, S. C., Miron, A., . . . Godwin, A. K. (2008). Allelic imbalance in BRCA1 and BRCA2 gene expression is associated with an increased breast cancer risk. *Human Molecular Genetics, 17*(9), pp. 1336–1348. doi:10.1093/hmg/ddn022

Cheung, V. G., Nayak, R. R., Wang, I. X., Elwyn, S., Cousins, S. M., Morley, M., & Spielman, R. S. (2010). Polymorphic Cis- and Trans-Regulation of Human Gene Expression. *PLoS Biology, 8*(9), p. e1000480. doi:10.1371/journal.pbio.1000480

Chlebowski, R. T., Chen, Z., Anderson, G. L., Rohan, T., Aragaki, A., Lane, D., . . . Prentice, R. (2005). Ethnicity and Breast Cancer: Factors Influencing Differences in Incidence and Outcome. *Journal of the National Cancer Institute, 97*(6), pp. 439-448. doi:10.1093/jnci/dji064

Chompret, A., Brugières, L., Ronsin, M., Gardes, M., Dessarps-Freichey, F., Abel, A., . . . Feunteun, J. (2000). P53 germline mutations in childhood cancers and cancer risk for carrier individuals. *British Journal of Cancer, 82*(12), pp. 1932-1937. doi:10.1054/bjoc.2000.1167

Chu, K. C., & Anderson, W. F. (2002). Rates for breast cancer characteristics by estrogen and progesterone receptor status in the major racial/ethnic groups. *Breast Cancer Research and Treatment, 74*(3), pp. 199-211. doi:10.1023/a:1016361932220

Civelek, M., Hagopian, R., Pan, C., Che, N., Yang, W. P., Kayne, P. S., . . . Lusis, A. J. (2013). Genetic regulation of human adipose microRNA expression and its consequences for metabolic traits. *Human Molecular Genetics, 22*(15), pp. 3023–3037. doi:10.1093/hmg/ddt159

Clavel-Chapelon, F., & Gerber, M. (2002). Reproductive Factors and Breast Cancer Risk. Do They Differ According to Age at Diagnosis? *Breast Cancer Research and Treatment, 72*(2), pp. 107-115. doi:10.1023/a:1014891216621

Cochet-Meilhac, M., & Chambon, P. (1974). Animal DNA-dependent RNA polymerases. 11. Mechanism of the inhibition of RNA polymerases B by amatoxins. *Biochimica et Biophysica Acta, 353*(2), pp. 160-184. doi:10.1016/0005-2787(74)90182-8

Colditz, G. A., Rosner, B. A., Chen, W. Y., Holmes, M. D., & Hankinson, S. E. (2004). Risk Factors for Breast Cancer According to Estrogen and Progesterone Receptor Status. *JNCI: Journal of the National Cancer Institute, 96*(3), pp. 218-228. doi:10.1093/jnci/djh025

Collaborative Group on Hormonal Factors in Breast . (1996). Breast Cancer and Hormonal Contraceptives: Collaborative Reanalysis of Individual Data on 53 297 Women With Breast Cancer and 100 239 Women Without Breast Cancer From 54 Epidemiological Studies. *Lancet, 347*(9017), pp. 1713-1727. doi:10.1016/s0140-6736(96)90806-5

Collaborative Group on Hormonal Factors in Breast Cancer. (2001). Familial Breast Cancer: Collaborative Reanalysis of Individual Data From 52 Epidemiological Studies Including 58,209 Women With Breast Cancer and 101,986 Women Without the Disease. *Lancet, 358*(9291), pp. 1389-1399. doi:10.1016/S0140-6736(01)06524-2

Collaborative Group on Hormonal Factors in Cancer. (2002). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50,302 women with breast cancer and 96,973 women without the disease. *Lancet, 360*(9328), pp. 187-195. doi:10.1016/S0140-6736(02)09454-0

Conde, L., Bracci, P. M., Richardson, R., Montgomery, S. B., & Skibola, C. F. (2013). Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. *American Journal of Human Genetics, 92*(1), pp. 126-130. doi:10.1016/j.ajhg.2012.11.009

Connolly, J. L., Fechner, R. L., Livolsi, V. A., Page, D. L., Patchefsky, A. A., & Silverberg, S. G. (1995). Recommendations for the Reporting of Breast Carcinoma. *AMERICAN JOURNAL OF CLINICAL PATHOLOGY, 104*(6), pp. 614-619. doi:10.1016/S0046-8177(96)90060-X

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., & Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics, 38*(11), pp. 1251-1260. doi:10.1038/ng1911

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews. Genetics, 10*(3), pp. 184–194. doi:10.1038/nrg2537

Cornish, A., & Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. *BioMed Research International, 2015*, p. 456479. doi:10.1155/2015/456479

Correia, L., Xavier, J. M., Magno, R., de aAlmeida, B. P., Esteves, F., Duarte, I., . . . Maia, A. (2021). Allelic expression imbalance of PIK3CA mutations is frequent in breast cancer and prognostically significant. *bioRXiv*. doi:10.1101/2021.03.05.434137

Couch, F. J., Hart, S. N., Sharma, P., Toland, A. E., Wang, X., Miron, P., . . . Fasching, P. A. (2015). Inherited Mutations in 17 Breast Cancer Susceptibility Genes Among a Large Triple-Negative Breast Cancer Cohort Unselected for Family History of Breast Cancer. *Journal of Clinical Oncology, 33*(4), pp. 304-311. doi:10.1200/JCO.2014.57.1414

Cox, A., Dunning, A. M., Garcia-Closas, M., Balasubramanian, S., Reed, M. W., Pooley, K. A., . . . Consortium, B. C. (2007). A Common Coding Variant in CASP8 Is Associated With Breast Cancer Risk. *Nature Genetics, 39*(3), pp. 352-358. doi:10.1038/ng1981

Cui, X., Schiff, R., Arpino, G., Osborne, C. K., & Lee, A. V. (2005). Biology of Progesterone Receptor Loss in Breast Cancer and Its Implications for Endocrine Therapy. *Journal of Clinical Oncology, 23*(30), pp. 7721-7735. doi:10.1200/JCO.2005.09.004

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), pp. 346–352. doi:10.1038/nature10983

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), pp. 346-352. doi:10.1038/nature10983

Cybulski, C., Carrot-Zhang, J., Kluźniak, W., Rivera, B., Kashyap, A., Wokołorczyk, D., . . . Akbari, M. R. (2015). Germline RECQL Mutations Are Associated With Breast Cancer Susceptibility. *Nature Genetics, 47*(6), pp. 643-636. doi:10.1038/ng.3284

Cybulski, C., Wokołorczyk, D., Huzarski, T., Byrski, T., Gronwald, J., Górski, B., . . . Lubiński, J. (2007). A Deletion in CHEK2 of 5,395 Bp Predisposes to Breast Cancer in Poland. *Breast Cancer Research and Treatment, 102*(1), pp. 119-122. doi:10.1007/s10549-006-9320-y

Dai, C., Miao, C. X., Xu, X. M., Liu, L. J., Gu, Y. F., Zhou, D., . . . Lu, G. X. (2015). Transcriptional activation of human CDCA8 gene regulated by transcription factor NF-Y in embryonic stem cells and cancer cells. *The Journal of Biological Chemistry, 290*(37), pp. 22423–22434. doi:10.1074/jbc.M115.642710

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics, 29*(2), pp. 229-232. doi:10.1038/ng1001-229

Damiola, F., Pertesi, M., Oliver, J., Le Calvez-Kelm, F., Voegele, C., Young, E. L., . . . Tavtigian, S. V. (2014). Rare Key Functional Domain Missense Substitutions in MRE11A, RAD50, and NBN Contribute to Breast Cancer Susceptibility: Results From a Breast Cancer Family

Registry Case-Control Mutation-Screening Study. *Breast Cancer Research, 16*(3), p. R58. doi:10.1186/bcr3669

Dang, C. V., O'Donnell, K. A., Zeller, K. I., Nguyen, T., Osthus, R. C., & Li, F. (2006). The c-Myc target gene network. *Seminars in Cancer Biology, 16*(4), pp. 253–26. doi:10.1016/j.semcancer.2006.07.014

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*, p. 32512. doi:10.1038/srep32512

Darabi, H., McCue, K., Beesley, J., Michailidou, K., Nord, S., Kar, S., . . . Chenevix-Trench, G. (2015). Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. *American Journal of Human Genetics, 97*(1), pp. 22-34. doi:10.1016/j.ajhg.2015.05.002

Dat, l., Matsuo, T., Yoshimaru, T., Kakiuchi, S., Goto, H., Hanibuchi, M., . . . Katagiri, T. (2012). Identification of genes potentially involved in bone metastasis by genome-wide gene expression profile analysis of non-small cell lung cancer in mice. *International Journal of Oncology, 40*(5), pp. 1455–1469. doi:10.3892/ijo.2012.1348

Dawood, S., Broglio, K., Buzdar, A. U., Hortobagyi, G. N., & Giordano, S. H. (2010). Prognosis of Women With Metastatic Breast Cancer by HER2 Status and Trastuzumab Treatment: An Institutional-Based Review. *Journal of Clinical Oncology, 28*(1), pp. 92-98. doi:10.1200/JCO.2008.19.9844

De Amicis, F., Giordano, F., Vivacqua, A., Pellegrino, M., Panno, M. L., Tramontano, D., . . . Andò, S. (2011). Resveratrol, through NF-Y/p53/Sin3/HDAC1 complex phosphorylation, inhibits estrogen receptor alpha gene expression via p38MAPK/CK2 signaling in human breast cancer cells. *FASEB Journal: official publication of the Federation of American Societies for Experimental Biology, 25*(10), pp. 3695–3707. doi:10.1096/fj.10-178871

de Bakker, P. I., Burtt, N. P., Graham, R. R., Guiducci, C., Yelensky, R., Drake, J. A., . . . Altshuler, D. (2006). Transferability of Tag SNPs in Genetic Association Studies in Multiple Populations. *Nature Genetics, 38*(11), pp. 1298-1303. doi:10.1038/ng1899

de Silvio, A., Imbriano, C., & Mantovani, R. (1999). Dissection of the NF-Y transcriptional activation potential. *Nucleic Acids Research, 27*(13), pp. 2578–2584. doi:10.1093/nar/27.13.2578

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics, 25*(24), pp. 3207–3212. doi:10.1093/bioinformatics/btp579

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., . . . Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature, 482*(7385), pp. 390–394. doi:10.1038/nature10808

Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One, 8*(12), p. e85024. doi:10.1371/journal.pone.0085024

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics, 43*(5), pp. 491–498. doi:10.1038/ng.806

DeSantis, C. E., Fedewa, S. A., Sauer, A. G., Kramer, J. L., Smith, R. A., & Jemal, A. (2015). Breast Cancer Statistics, 2015: Convergence of Incidence Rates Between Black and White Women. *CA: A Cancer Journal for Clinicians, 66*(1), pp. 31-42. doi:10.3322/caac.21320

DeVeale, B., van der Kooy, D., & Babak, T. (2012). Critical evaluation of imprinted gene expression by RNA-Seq: a new perspective. *PLoS Genetics, 8*(3), p. e1002600. doi:10.1371/journal.pgen.1002600

Devlin, N. R. (1995). A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. *Genomics, 29*(2), pp. 311-322. doi:10.1006/geno.1995.9003

Ding, Z., Ni, Y., Timmer, S. W., Lee, B. K., Battenhouse, A., Louzada, S., . . . Birney, E. (2014). Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genetics, 10*(11), p. e1004798. doi:10.1371/journal.pgen.1004798

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., . . . Cookson, W. O. (2007). A genome-wide association study of global gene expression. *Nature Genetics, 39*(10), pp. 1202–1207. doi:10.1038/ng2109

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), pp. 15-21. doi:10.1093/bioinformatics/bts635

Dodt, M., Roehr, J. T., Ahmed, R., & Dieterich, C. (2012). FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology, 1*(3), pp. 895–905. doi:10.3390/biology1030895

Dolfini, D., & Mantovani, R. (2013). Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y? *Cell Death and Differentiation, 20*(5), pp. 676–685. doi:10.1038/cdd.2013.13

Dolfini, D., Andrioletti, V., & Mantovani, R. (2019). Overexpression and alternative splicing of NF-YA in breast cancer. *Scientific Reports, 9*(1), p. 12955. doi:10.1038/s41598-019-49297-5

Doss, S., Schadt, E. E., Drake, T. A., & Lusis, A. J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research, 15*(5), pp. 681-691. doi:10.1101/gr.3216905

Duffy, S. W., Morrish, O. W., Allgood, P. C., Black, R., Gillan, M. G., Willsher, P., . . . Gilbert, F. J. (2018). Mammographic density and breast cancer risk in breast screening assessment cases and women with a family history of breast cancer. *European Journal os Cancer*, pp. 48-56. doi:10.1016/j.ejca.2017.10.022

Dunning, A. M., Michailidou, K., Kuchenbaecker, K. B., T. D., French, J. D., Beesley, J., . . . Edwards, S. L. (2016). Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nature Genetics, 48*(4), pp. 374–386. doi:10.1038/ng.3521

Dunnwald, L. K., Rossing, M. A., & Li, C. (2007). Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast Cancer Research, 9*(1), p. R6. doi:10.1186/bcr1639

Early Breast Cancer Trialists' Collaborative Group. (1998). Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. *Lancet, 351*(9114), pp. 1451-1467. doi:10.1016/S0140-6736(97)11423-4

Easton, D. F., Ford, D., & Bishop, D. T. (1995). Breast and ovarian cancer incidence in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. *American Journal of Human Genetics, 56*(1), pp. 265-271.

Easton, D. F., Pharoah, P. D., Antoniou, A. C., Tischkowitz, M., Tavtigian, S. V., Nathanson, K. L., . . . Foulkes, W. D. (2015). Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *The New England Journal of Medicine, 372*. doi:10.1056/NEJMsr1501341

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., . . . Ponder, B. A. (2007). Genome-wide Association Study Identifies Novel Breast Cancer Susceptibility Loci. *Nature, 447*(7148), pp. 1087-1093. doi:10.1038/nature05887

Eccles, S. A., Aboagye, E. O., Ali, S., Anderson, A. S., Armes, J., Berditchevski, F., . . . Thompson, A. M. (2013). Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Research, 15*(5), p. R92. doi:10.1186/bcr3493

Edsgärd, D., Iglesias, M. J., Reilly, S. J., Hamsten, A., Tornvall, P., Odeberg, J., & Emanuelsson, O. (2016). GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Scientific Reports, 6*, p. 21134. doi:10.1038/srep21134

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. *American Journal of Human Genetics, 93*(5), pp. 779–797. doi:10.1016/j.ajhg.2013.10.012

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. *American Journal of Human Genetics, 93*(5), pp. 779-797. doi:10.1016/j.ajhg.2013.10.012

Eeles, R. A., Olama, A. A., Benlloch, S., Saunders, E. J., Leongamornlert, D. A., Tymrakiewicz, M., . . . Easton, D. F. (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nature Genetics, 45*(4), pp. 385–391e3912. doi:10.1038/ng.2560

Ellis, I. O., Galea, M., Broughton, N., Locker, A., Blamey, R. W., & Elston, C. W. (1992). Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. *Histopathology*, pp. 479-489. doi:10.1111/j.1365-2559.1992.tb01032.x

Elston, C. W., & Ellis, I. O. (1991). Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology, 19*(5), pp. 403-410. doi:10.1111/j.1365-2559.1991.tb00229.x

ENCODE Project Consortium. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biology, 9*(4), p. e1001046. doi:10.1371/journal.pbio.1001046

Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Rätsch, G., . . . Consortium, R. (2013). Engström, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Rätsch, G., Goldman, N., Hubbard, T. J., Harrow, J., Guigó, R., Bertone, P., & RGASP Consortium. *Nature Methods, 10*(12), pp. 1185–1191. doi:10.1038/nmeth.2722

Fachal, L., & Dunning, A. M. (2015). From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Current opinion in genetics & development, 40*, pp. 32-41. doi:10.1016/j.gde.2015.01.004

Fachal, L., Aschard, H., Beesley, J., Barnes, D. R., Allen, J., Kar, S., . . . Dunning, A. M. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Natures Genetics, 52*(1), pp. 56-73. doi:10.1038/s41588-019-0537-1

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., . . . Iggo, R. (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene, 24*(29), pp. 4660-4671. doi:10.1038/sj.onc.1208561

Farvid, M. S., Cho, E., Chen, W. Y., Eliassen, A. H., & Willett, W. C. (s.d.). Dietary protein sources in early adulthood and breast cancer incidence: prospective cohort study. *BMJ, 348*, p. g3437. doi:10.1136/bmj.g3437

Fernandes, R. C., Toubia, J., Townley, S., Hanson, A. R., Dredge, B. K., Pillman, K. A., . . . Selth, L. A. (2021). Post-transcriptional Gene Regulation by MicroRNA-194 Promotes Neuroendocrine Transdifferentiation in Prostate Cancer. *Cell Reports, 34*(1), p. 108585. doi:10.1016/j.celrep.2020.108585

Fialka, I., Schwarz, H., Reichmann, E., Oft, M., Busslinger, M., & Beug, H. (1996). The estrogen-dependent c-JunER protein causes a reversible loss of mammary epithelial cell polarity involving a destabilization of adherens junctions. *The Journal of Cell Biology, 132*(6), pp. 1115–1132. doi:10.1083/jcb.132.6.1115

Finotello, F., & Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics, 14*(2), pp. 130–142. doi:10.1093/bfgp/elu035

Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., . . . Peto, J. (2011). Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study. *Journal of the National Cancer Institute, 103*(5), pp. 425-435. doi:10.1093/jnci/djq563

Fogarty, M. P., Xiao, R., Prokunina-Olsson, L., Scott, L. J., & Mohlke, K. L. (2010). Allelic expression imbalance at high-density lipoprotein cholesterol locus MMAB-MVK. *Human Molecular Genetics, 19*(10), pp. 1921–1929. doi:10.1093/hmg/ddq067

Fontanillas, P., Landry, C. R., Wittkopp, P. J., Russ, C., Gruber, J. D., Nusbaum, C., & Hartl, D. L. (2010). Key considerations for measuring allelic expression on a genomic scale using high-throughput sequencing. *Molecular Ecology, 19 Suppl 1*(Suppl 1), pp. 212–227. doi:10.1111/j.1365-294X.2010.04472.x

Ford, D., Easton, D. F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., . . . Zelada-Hedman, M. (1998). Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families. The Breast Cancer Linkage Consortium. *American Journal of Human Genetics, 62*(3), pp. 676-689. doi:10.1086/301749

Forjaz de Lacerda, G., Kelly, S. P., Bastos, J., Castro, C., Mayer, A., Mariotto, A. B., & Anderson, W. F. (2018). Breast cancer in Portugal: Temporal trends and age-specific incidence by geographic regions. *Cancer Epidemiology, 54*, pp. 12-18. doi:10.1016/j.canep.2018.03.003

Forozan, F., Veldman, R., Ammerman, C. A., Parsa, N. Z., Kallioniemi, A., Kallioniemi, O. P., & Ethier, S. P. (1999). Molecular cytogenetic analysis of 11 new breast cancer cell lines. *British Journal of Cancer, 81*(8), pp. 1328–1334. doi:10.1038/sj.bjc.6695007

Foulkes, W. D., Stefansson, I. M., Chappuis, P. O., Bégin, L. R., Goffin, J. R., Wong, N., . . . Akslen, L. A. (2003). Germline BRCA1 Mutations and a Basal Epithelial Phenotype in Breast Cancer. *Journal of the National Cancer Institute, 95*(19), pp. 1482-1485. doi:10.1093/jnci/djg050

Freedman, M. L., Monteiro, A. N., Gayther, S. A., C. G., Risch, A., Plass, C., . . . Mills, I. G. (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics, 43*(6), pp. 513-518. doi:10.1038/ng.840

French, J. D., Ghoussaini, M., Edwards, S. L., Meyer, K. B., Michailidou, K., Ahmed, S., . . . Dunning, A. 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), pp. 489–503. doi:10.1016/j.ajhg.2013.01.002

Fulda, S. (2009). Caspase-8 in Cancer Biology and Therapy. *Cancer Letters, 281*(2), pp. 128-133. doi:10.1016/j.canlet.2008.11.023

Gaffney, D. J. (2013). Global properties and functional complexity of human gene regulatory variation. *PLoS Genetics, 9*(5), p. e1003501. doi:10.1371/journal.pgen.1003501

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 Genetics, 7*(2), p. e1001292. doi:10.1371/journal.pgen.1001292

Garber, J. E., Goldstein, A. M., Kantor, A. F., Dreyfus, M. G., Fraumeni Jr, J. F., & Li, F. P. (1991). Follow-up Study of Twenty-Four Families With Li-Fraumeni Syndrome. *Cancer Research, 51*(22), pp. 6094-6097.

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 loc. *Nature Genetics, 45*(4), pp. 392-398e3982. doi:10.1038/ng.2561

Gazdar, A. F., Kurvari, V., Virmani, A., Gollahon, L., Sakaguchi, M., Westerfield, M., . . . Shay, J. W. (1998). Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *International Journal of Cancer, 78*(6), pp. 766-774. doi:10.1002/(sici)1097-0215(19981209)78:6<766::aid-ijc15>3.0.co;2-l

Ge, B., Pokholok, D. K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D. J., . . . Pastinen, T. (2009). Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nature Genetics, 41*(11), pp. 1216–1222. doi:10.1038/ng.473

Genkinger, J. M., Makambi, K. H., Palmer, J. R., Rosenberg, L., & Adams-Campbell, L. L. (2013). Consumption of Dairy and Meat in Relation to Breast Cancer Risk in the Black Women's Health Study. *Cancer Causes & Control, 24*(4), pp. 675-684. doi:10.1007/s10552-013-0146-8

Ghoussaini, M., & Pharoah, P. D. (2009). Polygenic susceptibility to breast cancer: current state-of-the-art. *Future Oncology, 5*(5), pp. 689-701. doi:10.2217/fon.09.29

Ghoussaini, M., Edwards, S. L., Michailidou, K., Nord, S., Cowper-Sal Lari, R., Desai, K., . . . Group, A. O. (2014). Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nature Communications, 4*, p. 4999. doi:10.1038/ncomms5999

Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., . . . Easton, D. F. (2012). Genome-wide Association Analysis Identifies Three New Breast Cancer Susceptibility Loci. *Nature Genetics, 44*(3), pp. 312-318. doi:10.1038/ng.1049

Ghoussaini, M., French, J. D., Michailidou, K., Nord, S., Beesley, J., Canisus, S., . . . Edwards, S. L. (2016). Evidence that the 5p12 Variant rs10941679 Confers Susceptibility to Estrogen-Receptor-Positive Breast Cancer through FGF10 and MRPS30 Regulation. *American Journal of Human Genetics, 99*(4), pp. 903–911. doi:10.1016/j.ajhg.2016.07.017

Glubb, D. M., Maranian, M. J., Michailidou, K., Pooley, K. A., Meyer, K. B., Kar, S., . . . French, J. D. (2015). Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *American Journal of Human Genetics, 96*(1), pp. 5-20. doi:10.1016/j.ajhg.2014.11.009

Goldgar, D. E., Easton, D. F., Cannon-Albright, L. A., & Skolnick, M. H. (1994). Systematic Population-Based Assessment of Cancer Risk in First-Degree Relatives of Cancer Probands. *Journal of the National Cancer Institute, 86*(21), pp. 1600-1608. doi:10.1093/jnci/86.21.1600

Goldgar, D. E., Healey, S., Dowty, J. G., Da Silva, L., Chen, X., Spurdle, A. B., . . . Chenevix-Trench, G. (2011). Rare Variants in the ATM Gene and Risk of Breast Cancer. *Breast Cancer Research, 13*(4), p. R73. doi:10.1186/bcr2919

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. *Journal of Clinical Oncology, 27*(8), pp. 1250-1256. doi:10.1200/JCO.2008.16.6959

González-Tablas, M., Crespo, I., Vital, A. L., Otero, Á., Nieto, A. B., Sousa, P., . . . Tabernero, M. D. (2018). Prognostic stratification of adult primary glioblastoma multiforme patients based on their tumor gene amplification profiles. *Oncotarget, 9*(46), pp. 28083–28102. doi:10.18632/oncotarget.25562

Göring, H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. A., . . . Blangero, J. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics, 39*(10), pp. 1208-1216. doi:10.1038/ng2119

Grant, G. R., Farkas, M. H., Pizarro, A. D., Lahens, N. F., Schug, J., Brunk, B. P., . . . Pierce, E. A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics, 27*(18), pp. 2518–2528. doi:10.1093/bioinformatics/btr427

Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G. P., Haig, D., & Dulac, C. (2010). High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science, 329*(5992), pp. 643-648. doi:10.1126/science.1190830

Grundberg, E., Adoue, V., Kwan, T., Ge, B., Duan, Q. L., Lam, K. C., . . . Pastinen, T. (2011). Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. *PLoS Genetics, 7*(1), p. e1001279. doi:e1001279

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., . . . Multiple Tissue Human Expression Resource (MuTHER). (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics, 44*(10), pp. 1084-1089. doi:10.1038/ng.2394

GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics, 45*(6), pp. 580-585. doi:10.1038/ng.2653

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*(348), pp. 648–660. doi:10.1126/science.1262110

GTEX Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissue. *Science, 369*(6509), pp. 1318–1330. doi:10.1126/science.aaz1776

Guedj, M., Marisa, L., de Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., . . . Theillet, C. (2012). A refined molecular taxonomy of breast cancer. *Oncogene, 31*(9), pp. 1196-1206. doi:10.1038/onc.2011.301

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. *Cancer Epidemiology, Biomarkers & Prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 24*(11), pp. 1680–1691. doi:10.1158/1055-9965.EPI-15-0363

Gutierrez-Arcelus, M. L., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., . . . Dermitzakis, E. T. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife, 2*, p. e00523. doi:10.7554/eLife.00523

Habashy, H. O., Powe, D. G., Abdel-Fatah, T. M., Gee, J. M., Nicholson, R. I., Green, A. R., . . . Ellis, I. O. (2012). A review of the biological and clinical characteristics of luminal-like oestrogen receptor-positive breast cancer. *Histopathology, 60*, pp. 854-863. doi:10.1111/j.1365-2559.2011.03912.x

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), pp. 1210-1214. doi:10.1038/ng.985

Hall, T. (2011). BioEdit: An important software for molecular biology. *GERF Bulletin of Biosciences, 2*(1), pp. 60-61.

Hamajima, N., Hirose, K., Tajima, K., Rohan, T., Calle, E. E., Heath, C. W., . . . Cancer, C. G. (2002). Alcohol, tobacco and breast cancer--collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *British Journal of Cancer, 87*(11), pp. 1234-1245. doi:10.1038/sj.bjc.6600596

Hamdi, Y., Soucy, P., Adoue, V., Michailidou, K., Canisius, S., Lemaçon, A., . . . Simard, J. (2016). Association of breast cancer risk with genetic variants showing differential allelic expression: Identification of a novel breast cancer susceptibility locus at 4q21. *Oncotarget, 7*(49), pp. 80140–80163. doi:10.18632/oncotarget.12818

Hammond, M. H., Hayes, D. F., Dowsett, M., Allred, D. C., Hagerty, K. L., Fitzgibbons, S. B., . . . Wolff, A. C. (2010). American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer. *Journal of Clinical Oncology, 28*(16), pp. 2784-2795. doi:10.1200/JCO.2009.25.6529

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell, 100*(1), pp. 57-70. doi:10.1016/S0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell, 144*(5), pp. 646-674. doi:10.1016/j.cell.2011.02.013

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., . . . Mill, J. (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature Neuroscience, 19*(1), pp. 48–54. doi:10.1038/nn.4182

Harris, L. N., Ismaila, N., McShane, L. M., Andre, F., Collyar, D. E., Gonzalez-Angulo, A. M., . . . Hayes, D. F. (2016). Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. *Journal Of Clinical Oncology, 34*(10), pp. 1134-1150. doi:10.1200/JCO.2015.65.2289

Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Somerfield, S. T., . . . Bast Jr, R. C. (2007). American Society of Clinical Oncology 2007 Update of Recommendations for the Use of Tumor Markers in Breast Cancer. *Journal of Clinical Oncology, 25*(33), pp. 5287-5312. doi:10.1200/JCO.2007.14.2364

Harvey, C. T., Moyerbrailean, G. A., Davis, G. O., Wen, X., Luca, F., & Pique-Regi, R. (2015). QuASAR: quantitative allele-specific analysis of reads. *Bioinformatics, 31*(8), pp. 1235–1242. doi:10.1093/bioinformatics/btu802

Hasin-Brumshtein, Y., Hormozdiari, F., Martin, L., van Nas, A., Eskin, E., Lusis, A. J., & Drake, T. A. (2014). Allele-specific expression and eQTL analysis in mouse adipose tissue. *BMC Genomics, 15*(1), p. 471. doi:10.1186/1471-2164-15-471

Hassan, M. A., Butty, V., Jensen, K. D., & Saeij, J. P. (2014). The genetic basis for individual differences in mRNA splicing and APOBEC1 editing activity in murine macrophage. *Genome Research, 24*(3), pp. 377–389. doi:10.1101/gr.166033.113

Haupt, S., Buckley, D., Pang, J. M., Panimaya, J., Paul, P. J., Gamell, C., . . . Haupt, Y. (2015). Targeting Mdmx to treat breast cancers with wild-type p53. *Cell Death Disease, 6*(7), p. e1821. doi:10.1038/cddis.2015.17

Haupt, S., Vijayakumaran, R., Miranda, P. J., Burgess, A., Lim, E., & Haupt, Y. (2017). The role of MDM2 and MDM4 in breast cancer development and prevention. *Journal of Molecular Cell Biology, 9*(1), pp. 53-61. doi:10.1093/jmcb/mjx007

Heap, G. A., Yang, J. H., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., . . . Plagnol, V. (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Human Molecular Genetics, 19*(1), pp. 122–134. doi:10.1093/hmg/ddp473

Heap, G. A., Yang, J. H., Downes, K., Healy, B. C., Hunt, K. A., Bockett, N., . . . Plagnol, V. (2010). Genome-wide analysis of allelic expression imbalance in human primary cells by high-

throughput transcriptome resequencing. *Human Molecular Genetics, 19*(1), pp. 122–134. doi:10.1093/hmg/ddp473

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. *Clinical Cancer Research, 12*(10), pp. 3209-3215. doi:10.1158/1078-0432.CCR-06-0083

Heinz, S., Romanoski, C. E., Benner, C., Allison, K. A., Kaikkonen, M. U., Orozco, L. D., & Glass, C. K. (2013). Effect of natural genetic variation on enhancer selection and function. *Nature, 503*(7477), pp. 487–492. doi:10.1038/nature12615

Helbig, S., Wockner, L., Bouendeu, A., Hille-Betz, U., McCue, K., French, J. D., . . . Beesley, J. (2017). Functional dissection of breast cancer risk-associated TERT promoter variants. *Oncotarget, 8*(40), pp. 67203–67217. doi:10.18632/oncotarget.18226

Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols, 2*(8), pp. 1849-1861. doi:10.1038/nprot.2007.249

Hemminki, A., Markie, D., Tomlinson, I., Avizienyte, E., Roth, S., Loukola, A., . . . Aaltonen, L. A. (1998). A Serine/Threonine Kinase Gene Defective in Peutz-Jeghers Syndrome. *Nature, 391*(6663), pp. 184-187. doi:10.1038/34432

Hemminki, K., & Vaittinen, P. (1998). Familial Breast Cancer in the Family-Cancer Database. *International Journal of Cancer, 77*(3), pp. 386-391. doi:10.1002/(sici)1097-0215(19980729)77:3<386::aid-ijc13>3.0.co;2-6

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., . . . Esteller, M. (2013). DNA methylation contributes to natural human variation. *Genome Research, 23*(9), pp. 1363–1372. doi:10.1101/gr.154187.112

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America, 106*(23), pp. 9362–9367. doi:10.1073/pnas.0903103106

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America, 106*(23), pp. 9362–9367. doi:10.1073/pnas.0903103106

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetis, 6*(2), pp. 95-108. doi:10.1038/nrg1521

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics, 6*(2), pp. 95–108. doi:10.1038/nrg1521

Ho, J., Tumkaya, T., Aryal, S., Choi, H., & Claridge-Chang, A. (2019). Moving beyond P values: data analysis with estimation graphics. *Nature Methods, 16*(7), pp. 565-566. doi:10.1038/s41592-019-0470-3

Hochberg, Y., & Benjamini, Y. (1990). More Powerful Procedures for Multiple Significance Testing. *Statistics in Medicine, 9*(7), pp. 811-818. doi:10.1002/sim.4780090710

Hopper, J. L., & Carlin, J. B. (1992). Familial Aggregation of a Disease Consequent Upon Correlation Between Relatives in a Risk Factor Measured on a Continuous Scale. *American Journal of Epidemiology, 136*(9), pp. 1138-1147. doi:10.1093/oxfordjournals.aje.a116580

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), p. e0160316. doi:10.1371/journal.pone.0160316

Hortobagyi, G. N., Salazar, J. d., Pritchard, K., Amadori, D., Haidinger, R., Hudis, C. A., . . . Albain, K. S. (2005). The global breast cancer burden: variations in epidemiology and survival. *Clinical Breast Cancer, 6*(5), pp. 391-401. doi:10.3816/cbc.2005.n.043

Horwitz, K. B., Pearson, O. H., & Segaloff, A. (1975). Predicting Response to Endocrine Therapy in Human Breast Cancer: A Hypothesis. *Science, 189*(4204), pp. 726-727. doi:10.1126/science.168640

Hou, N., Hong, S., Wang, W., Olopade, O. I., D. J., & H. D. (2013). Hormone Replacement Therapy and Breast Cancer: Heterogeneous Risks by Race, Weight, and Breast Density. *105*(18), pp. 1365-1372. doi:10.1093/jnci/djt207

Hu, E., Mueller, E., Oliviero, S., Papaioannou, V. E., Johnson, R., & Spiegelman, B. M. (1994). Targeted disruption of the c-fos gene demonstrates c-fos-dependent and -independent pathways for gene expression stimulated by growth factors or oncogenes. *The EMBO Journal, 13*(13), pp. 3094–3103.

Hu, Y. J., Sun, W., Tzeng, J. Y., & Perou, C. M. (2015). Proper Use of Allele-Specific Expression Improves Statistical Power for cis-eQTL Mapping with RNA-Seq Data. *Journal of the American Statistical Association, 110*(511), pp. 962–974. doi:10.1080/01621459.2015.1038449

Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., . . . Perou, C. M. (2006). The Molecular Portraits of Breast Tumors Are Conserved Across Microarray Platforms. *BMC Genomics, 7*, p. 96. doi:10.1186/1471-2164-7-96

Huang, Z., Hankinson, S. E., Colditz, G. A., Stampfer, M. J., Hunter, D. J., Manson, J. E., . . . Willett, W. C. (1997). Dual effects of weight and weight gain on breast cancer risk. *JAMA, 278*(17), pp. 1407-1411.

Hugh, J., Hanson, J., Cheang, M. C., Nielsen, T. O., Perou, C. M., Dumontet, C., . . . Vogel, C. (2009). Breast Cancer Subtypes and Response to Docetaxel in Node-Positive Breast Cancer: Use

of an Immunohistochemical Definition in the BCIRG 001 Trial. *Journal of Clinical Oncology, 27*(8), pp. 1168-1176. doi:10.1200/JCO.2008.18.1024

Hulka, B. S., & Stark, A. T. (1995). Breast Cancer: Cause and Prevention. *Lancet, 346*(8979), pp. 883-887. doi:10.1016/s0140-6736(95)92713-1

Hunter, D. J., & Willett, W. C. (1993). Diet, body size, and breast cancer. *Epidemiologic Reviews, 15*(1), pp. 110-132. doi:0.1093/oxfordjournals.epirev.a036096

Hunter, D. J., Colditz, G. A., Hankinson, S. E., Malspeis, S., Spiegelman, D., Chen, W., . . . Willett, W. C. (2010). Oral Contraceptive Use and Breast Cancer: A Prospective Study of Young Women. *Cancer Epidemiology, Biomarkers & Prevention, 19*(10), pp. 2496-2502. doi:10.1158/1055-9965.EPI-10-0747

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), pp. 870-874. doi:10.1038/ng2075

Hwang, S. J., Lozano, G., Amos, C. I., & Strong, L. C. (2003). Germline p53 Mutations in a Cohort With Childhood Sarcoma: Sex Differences in Cancer Risk. *American Journal of Human Genetics, 72*(4), pp. 975-983. doi:10.1086/374567

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports, 5*, p. 17875. doi:10.1038/srep17875

International Schizophrenia Consortium, Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., . . . Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature, 460*(7256), pp. 748–752. doi:10.1038/nature08185

Iorns, E., Lord, C. J., & Ashworth, A. (2009). Parallel RNAi and compound screens identify the PDK1 pathway as a target for tamoxifen sensitization. *The Biochemical Journal, 417*(1), pp. 361–37. doi:10.1042/BJ20081682

Iqbal, J., Ginsburg, O., Rochon, P. A., Sun, P., & Narod, S. A. (2015). Differences in Breast Cancer Stage at Diagnosis and Cancer-Specific Survival by Race and Ethnicity in the United States. *JAMA, 313*(2), pp. 165-173. doi:10.1001/jama.2014.17322

Jacinta-Fernandes, A., Xavier, J. M., Magno, R., Esteves, F., & Maia, A.-T. (2018). Uncovering miRNA-mediated cis-regulation in common cancers. *Pulmonology Journal, 24*(SC1), p. 2.

Jacinta-Fernandes, A., Xavier, J. M., Magno, R., Lage, J. G., & Maia, A. T. (2020). Allele-specific miRNA-binding analysis identifies candidate target genes for breast cancer risk. *NPJ Genomic Medicine, 5*, p. 4. doi:10.1038/s41525-019-0112-9

Jansen, R. C., & Nap, J. P. (2001). Genetical Genomics: The Added Value From Segregation. *Trends in genetic, 17*(7), pp. 388-391. doi:10.1016/s0168-9525(01)02310-1

Jiang, H., Lei, R., Ding, S. W., & Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics, 15*, p. 182. doi:10.1186/1471-2105-15-182

Joensuu, H., Kellokumpu-Lehtinen, P.-L., Bono, P., Alanko, T., Kataja, V., Asola, R., . . . Isola, J. (2006). Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer. *The New England Journal of Medicine, 354*(8), pp. 809-820. doi:10.1056/NEJMoa053028

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., & de Bakker, P. I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics, 24*(24), pp. 2938-2939. doi:10.1093/bioinformatics/btn564

Kaplow, I. M., MacIsaac, J. L., Mah, S. M., McEwen, L. M., Kobor, M. S., & Fraser, H. B. (2015). A pooling-based approach to mapping genetic variants associated with DNA methylation. *Genome Research, 25*(6), pp. 907–917. doi:10.1101/gr.183749.114

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., . . . Snyder, M. (2010). Variation in transcription factor binding among humans. *Science, 328*(5975), pp. 232–235. doi:10.1126/science.1183621

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., . . . Snyder, M. (2013). Extensive variation in chromatin states across humans. *Science, 342*(6159), pp. 750–752. doi:10.1126/science.1242510

Kaurah, P., MacMillan, A., Boyd, N., Senz, J., De Luca, A., Chun, N., . . . Huntsman, D. (2007). Founder and Recurrent CDH1 Mutations in Families With Hereditary Diffuse Gastric Cancer. *JAMA, 297*(21), pp. 2360-2372. doi:10.1001/jama.297.21.2360

Kazemi-Sefat, G. E., Keramatipour, M., Talebi, S., Kavousi, K., Sajed, R., Kazemi-Sefat, N. A., & Mousavizadeh, K. (2021). The importance of CDC27 in cancer: molecular pathology and clinical aspects. *Cancer Cell International, 21*(1), p. 160. doi:10.1186/s12935-021-01860-9

Keibl, Z., & Kristensen, V. N. (2016). Women at high risk of breast cancer: Molecular characteristics, clinical presentation and management. *The Creast, 28*, pp. 136-144. doi:10.1016/j.breast.2016.05.006

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research, 12*(6), pp. 996-1006. doi:10.1101/gr.229102

Kerlikowske, K., Cook, A. J., Buist, D. S., Cummings, S. R., Vachon, C., Vacek, P., & Miglioretti, D. L. (2010). Breast Cancer Risk by Breast Density, Menopause, and Postmenopausal Hormone Therapy Use. *Journal of Clinical Oncology, 28*(24), pp. 3830-3837. doi:10.1200/JCO.2009.26.4770

Keurentjes, J. J., Fu, J., Terpstra, I. R., Garcia, J. M., van den Ackerveken, G., Snoek, L. B., . . . Jansen, R. C. (2007). Regulatory network construction in Arabidopsis by using genome-

wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America, 104*(5), pp. 1708–1713. doi:10.1073/pnas.0610429104

Keydar, I., Chen, L., Karby, S., Weiss, F. R., Delarea, J., Radu, M., . . . Brenner, H. J. (1979). Establishment and characterization of a cell line of human breast carcinoma origin. *European Journal of Cancer, 15*(5), pp. 659–670. doi:10.1016/0014-2964(79)90139-7

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., . . . Dermitzakis, E. T. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*(342), pp. 744–747. doi:10.1126/science.1242463

Kilpivaara, O., Vahteristo, P., Falck, J., Syrjäkoski, K., Eerola, H., Easton, D., . . . Nevanlinna, H. (2004). CHEK2 Variant I157T May Be Associated With Increased Breast Cancer Risk. *International Journal of Cancer, 111*(4), pp. 543-547. doi:10.1002/ijc.20299

King, M. C., Marks, J. H., Mandell, J. B., & Group, N. Y. (2003). Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2. *302*(5645), pp. 643-646. doi:10.1126/science.1088759

Korf, B. R. (2013). Integration of genomics into medical practice. *Discovery medicine, 16*(89), pp. 241–248.

Koutros, S., Schumacher, F. R., Hayes, R. B., Ma, J., Huang, W. Y., Albanes, D., . . . Berndt, S. I. (2010). Pooled analysis of phosphatidylinositol 3-kinase pathway variants and risk of prostate cancer. *Cancer Research, 70*(6), pp. 2389–2396. doi:10.1158/0008-5472.CAN-09-3575

Krieg, R. C., Knuechel, R., Schiffmann, E., Liotta, L. A., Petricoin, E. F., & Herrmann, P. C. (2004). Mitochondrial proteome: Cancer-altered metabolism associated with cytochrome c oxidase subunit level variation. *Proteomics, 4*(9), pp. 2789–2795. doi:10.1002/pmic.200300796

Kriege, M., Hollestelle, A., Jager, A., Huijts, P. E., Berns, E. M., Sieuwerts, A. M., . . . Seynaeve, C. (2014). Survival and Contralateral Breast Cancer in CHEK2 1100delC Breast Cancer Patients: Impact of Adjuvant Chemotherapy. *British Journal of Cancer, 111*(5), pp. 1004-1013. doi:10.1038/bjc.2014.306

Krop, I., Ismaila, N., Andre, F., Bast, R. C., Barlow, W., Collyar, D. E., . . . Stearns, V. (2017). Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. *Journal of Clinical Oncology, 35*(24), pp. 2838-2847. doi:10.1200/JCO.2017.74.0472

Kudaravalli, S., Veyrieras, J. B., Stranger, B. E., Dermitzakis, E. T., & Pritchard, J. K. (2009). Gene expression levels are a target of recent natural selection in the human genome. *Molecular Biology and Evolution, 26*(3), pp. 649–658. doi:10.1093/molbev/msn289

Kumar, V., Westra, H. J., Karjalainen, J., Zhernakova, D. V., Esko, T., Hrdlickova, B., . . . Wijmenga, C. (2013). Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genetics, 9*(1), p. e1003201. doi:10.1371/journal.pgen.1003201

Kumasaka, N., Knights, A. J., & Gaffney, D. J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics, 48*(2), pp. 206–213. doi:0.1038/ng.3467

Kuper, H., Ye, W., Weiderpass, E., Ekbom, A., Trichopoulos, D., Nyrén, O., & Adami, H. O. (2000). Alcohol and Breast Cancer Risk: The Alcoholism Paradox. *British Journal of Cancer, 83*(7), pp. 949-951. doi:10.1054/bjoc.2000.1360

Kwan, T., Grundberg, E., Koka, V., Ge, B., Lam, K. C., Dias, C., . . . Majewski, J. (2009). Tissue Effect on Genetic Control of Transcript Isoform Variation. *PLoS Genetics, 5*(8), p. e1000608. doi:10.1371/journal.pgen.1000608

Lachenmeier, D. W., Przybylski, M. C., & Rehm, J. (2012). Comparative Risk Assessment of Carcinogens in Alcoholic Beverages Using the Margin of Exposure Approach. *International Journal of Cancer, 131*(16), pp. E995–E1003. doi:10.1002/ijc.27553

Lakhani, S. R., Ellis, I. O., Schnitt, S. J., Tan, P. H., & van de Vivjer, M. J. (Eds.). (2019). *WHO Classification of Tumours of the Breast* (5th ed.). Lyon, France: International Agency for Research on Cancer.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . International Human Genome Sequencing Consortium. (2001). Initial Sequencing and Analysis of the Human Genome. *Nature, 409*(6822), pp. 860-921. doi:10.1038/35057062

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., . . . Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature, 501*(7468), pp. 506–511. doi:10.1038/nature12531

Lawrenson, K., Kar, S., McCue, K., Kuchenbaeker, K., Michailidou, K., Tyrer, J., . . . Gayther, S. A. (2016). Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus. *Nature Communications, 7*, p. 12675. doi:10.1038/ncomms12675

Lee, H., Jeong, A. J., & Ye, S. K. (2019). Highlighted STAT3 as a potential drug target for cancer therapy. *BMB Reports, 52*(7), pp. 415–423. doi:10.5483/BMBRep.2019.52.7.152

Lee, K., Kim, S. C., Jung, I., Kim, K., Seo, J., Lee, H. S., . . . Choi, J. K. (2013). Genetic landscape of open chromatin in yeast. *PLoS Genetics, 9*(2), p. e1003229. doi:10.1371/journal.pgen.1003229

Lee, M. N., Ye, C., Villani, A. C., Raj, T., Li, W., Eisenhaure, T. M., . . . Hacohen, N. (2014). Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science, 343*(7175), p. 1246980. doi:10.1126/science.1246980

Leong, S. P. (2010). Is Breast Cancer the Same Disease in Asian and Western Countries? *34*(10), pp. 2308-2324. doi:10.1007/s00268-010-0683-1

Levine, M. E., S. J., Brandhorst, S., Balasubramanian, P., Cheng, C. W., Madia, F., . . . Longo, V. D. (2014). Low protein intake is associated with a major reduction in IGF-1, cancer, and overall mortality in the 65 and younger but not older population. *Cell Metabolism, 19*(3), pp. 407-417. doi:10.1016/j.cmet.2014.02.006

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell, 115*(7), pp. 787-798. doi:10.1016/s0092-8674(03)01018-3

Li, C. I., Malone, K. E., Porter, P. L., Weiss, N. S., Tang, M. T., & Daling, J. R. (2003). The Relationship Between Alcohol Use and Risk of Breast Cancer by Histology and Hormone Receptor Status Among Women 65-79 Years of Age. *Cancer Epidemiology, Biomarkers & Prevention, 12*(10), pp. 1061-1066.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), pp. 2987–2993. doi:10.1093/bioinformatics/btr509

Li, H. H., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., . . . 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), pp. 2078–2079. doi:10.1093/bioinformatics/btp352

Li, M. L., & Greenberg, R. A. (2012). Links Between Genome Integrity and BRCA1 Tumor Suppression. *Trends in Biochemical Sciences, 37*(10), pp. 418-424. doi:10.1016/j.tibs.2012.06.007

Li, Q., Seo, J. H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., . . . Freedman, M. L. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell, 152*(3), pp. 633-641. doi:10.1016/j.cell.2012.12.034

Li, Q., Yang, J., Yu, Q., Wu, H., Liu, B., Xiong, H., . . . Liao, Z. (2013). Associations between single-nucleotide polymorphisms in the PI3K-PTEN-AKT-mTOR pathway and increased risk of brain metastasis in patients with non-small cell lung cancer. *Clinical Cancer Research, 19*(22), pp. 6252–6260. doi:10.1158/1078-0432.CCR-13-1093

Li, Y., Yang, D., Yin, X., Zhang, X., Huang, J., Wu, Y., . . . Ren, G. (2020). Clinicopathological Characteristics and Breast Cancer – Specific Survival of Patients With Single Hormone Receptor–Positive Breas tCancer. *JAMA Network Open, 3*(1), pp. 1-15. doi:10.1001/jamanetworkopen.2019.18160

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. *The New England Journal of Medicine, 343*(2), pp. 78-85. doi:10.1056/NEJM200007133430201

Lilyquist, J., Ruddy, K. J., Vachon, C. M., & Couch, F. J. (2018). Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. *Cancer Epidemiology, Biomarkers & Prevention, 27*(4), pp. 380-394. doi:10.1158/1055-9965.EPI-17-1144

Lin, W. Y., Camp, N. J., Ghoussaini, M., Beesley, J., Michailidou, K., Hopper, J. L., . . . Cox, A. (2015). Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. *Human Molecular Genetics, 24*(1), pp. 285–298. doi:10.1093/hmg/ddu431

Liu, C., Wang, Y., Wang, Q. S., & Wang, Y. J. (2012). The CHEK2 I157T Variant and Breast Cancer Susceptibility: A Systematic Review and Meta-Analysis. *Asian Pacific Journal of Cancer Prevention , 13*(4), pp. 1355-1360. doi:10.7314/apjcp.2012.13.4.1355

Liu, J., Prager-van der Smissen, W. J., Look, M. P., Sieuwerts, A. M., Smid, M., Meijer-van Gelder, M. E., . . . Martens, J. W. (2016). GATA3 mRNA expression, but not mutation, associates with longer progression-free survival in ER-positive breast cancer patients treated with first-line tamoxifen for recurrent disease. *Cancer Letters, 376*(1), pp. 104-109. doi:10.1016/j.canlet.2016.03.038

Liu, P., Cheng, H., Roberts, T., & Zhao, J. J. (2009). Targeting the phosphoinositide 3-kinase pathway in cancer. *Nature Reviews Drug Discovery, 8*, pp. 627-644. doi:doi.org/10.1038/nrd2926

Liu, R., Maia, A. T., Russell, R., Caldas, C., Ponder, B. A., & Ritchie, M. E. (2012). Allele-specific expression analysis methods for high-density SNP microarray data. *Bioinformatic, 28*(8), pp. 1102–1108. doi:10.1093/bioinformatics/bts089

Liu, X., Han, S., Wang, Z., Gelernter, J., & Yang, B. Z. (2013). Variant callers for next-generation sequencing data: a comparison study. *PLoS One, 8*(9), p. e75619. doi:10.1371/journal.pone.0075619

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), p. e1001002. doi:10.1371/journal.pgen.1001002

Louis, T. A. (1981). Confidence Intervals for a Binomial Parameter after Observing no Successes. *The American Statistician, 35*(3), p. 154.

Ma, H., Bernstein, L., Pike, M. C., & Ursin, G. (2006). Reproductive factors and breast cancer risk according to joint estrogen and progesterone receptor status: a meta-analysis of epidemiological studies. *Breast Cancer Research, 8*(4), p. R43. doi:10.1186/bcr1525

Ma, X.-J., Salunga, R., Dahiya, S., Wang, W., Carney, E., Durbecq, V., . . . Sgroi, D. (2008). A Five-Gene Molecular Grade Index and HOXB13:IL17BR Are Complementary Prognostic Factors in Early Stage Breast Cancer. *Clinical Cancer Research, 9*, pp. 2601-2608. doi:10.1158/1078-0432.CCR-07-5026

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 Research, 45*(D1), pp. D896–D901. doi:10.1093/nar/gkw1133

Mackay, A., Wigelt, B., Grigoriadis, A., Kreike, B., Natrajan, R., A'Hern, R., . . . Reis-Filho, J. S. (2011). Microarray-based Class Discovery for Molecular Classification of Breast Cancer: Analysis of Interobserver Agreement. *Journal of the National Cancer Institute, 103*(8), pp. 662-673. doi:10.1093/jnci/djr071

Maia, A. T., Antoniou, A. C., O'Reilly, M., Samarajiwa, S., Dunning, M., Kartsonaki, C., . . . Ponder, B. A. (2012). Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. *Breast Cancer Research, 14*(2), p. R63. doi:10.1186/bcr3169

Maia, A.-T., Antoniou, A. C., O'Reilly, M., Samarajiwa, S., Dunning, M., Kartsonaki, C., . . . Ponder, B. A. (2012). Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. *Breast Cancer Research, 14*(2), p. R63. doi:10.1186/bcr3169

Maia, A.-T., Spiteri, I., Lee, A. J., O'Reilly, M., Jones, L., Caldas, C., & Ponder, B. A. (2009). Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. *Breast Cancer Research, 11*(6), p. R88. doi:doi:10.1186/bcr2458

Majewski, J., & Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics, 27*(2), pp. 72-79. doi:10.1016/j.tig.2010.10.006

Malkin, D. (1994). Germline p53 Mutations and Heritable Cancer. *Annual Review of Genetics, 28*, pp. 443-465. doi:10.1146/annurev.ge.28.120194.002303

Malkin, D., Li, F. P., Strong, L. C., Fraumeni Jr, J. F., Nelson, C. E., Kim, D. H., . . . Tainsky, M. A. (1990). Germ Line p53 Mutations in a Familial Syndrome of Breast Cancer, Sarcomas, and Other Neoplasms. *Science, 250*(4895), pp. 1233-1238. doi:10.1126/science.1978757

Marchbanks, P. A., McDonald, J. A., Wilson, H. G., Folger, S. G., Mandel, M. G., Daling, J. R., . . . Weiss, L. K. (2002). Oral Contraceptives and the Risk of Breast Cancer. *The New England Journal of Medicine, 346*(26), pp. 2025-2032. doi:10.1056/NEJMoa013202

Marine, J. C., Dyer, M. A., & Jochemsen, A. G. (2007). MDMX: from bench to bedside. *Journal of Cell Science, 120*(Pt 3), pp. 371–378. doi:10.1242/jcs.03362

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal, 17*(1), pp. 10-12. doi:10.14806/ej.17.1.200

Masciari, S., Dillon, D. A., Rath, M., Robson, M., Weitzel, J. N., Balmana, J., . . . Garber, J. E. (2012). Breast Cancer Phenotype in Women With TP53 Germline Mutations: A Li-Fraumeni Syndrome Consortium Effort. *Breast Cancer Research and Treatment, 133*(3), pp. 1125-1130. doi:10.1007/s10549-012-1993-9

Matsuoka, S., Huang, M., & Elledge, S. J. (1998). Linkage of ATM to Cell Cycle Regulation by the Chk2 Protein Kinase. *Science, 282*(5395), pp. 1893-1897. doi:10.1126/science.282.5395.1893

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., . . . Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science, 337*(6099), pp. 1190–1195. doi:10.1126/science.1222794

Mavaddat, N., Antoniou, A. C., Easton, D. F., & Garcia-Closas, M. (2010). Genetic Susceptibility to Breast Cancer. *Molecular Oncology, 4*(3), pp. 174-191. doi:10.1016/j.molonc.2010.04.011

McCarthy, D. J., & Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics, 25*(6), pp. 765–771. doi:10.1093/bioinformatics/btp053

McDaniell, R., Lee, B. K., Song, L., Liu, Z., Boyle, A. P., Erdos, M. R., . . . Birney, E. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. *Science, 328*(5975), pp. 235-239. doi:10.1126/science.1184655

McDonough, S. J., Bhagwate, A., Sun, Z., Wang, C., Zschunke, M., Gorman, J. A., . . . Cunningham, J. M. (2019). Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods. *PLoS One, 14*(4), p. e0211400. doi:10.1371/journal.pone.0211400

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), pp. 1297–1303. doi:10.1101/gr.107524.110

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20*(9), pp. 1297–1303. doi:10.1101/gr.107524.110

McKeown, P. C., Fort, A., & Spillane, C. (2014). Analysis of genomic imprinting by quantitative allele-specific expression by Pyrosequencing(®). *Methods in Molecular Biology, 85*(104), pp. 85–104. doi:10.1007/978-1-62703-773-0_6

McPherson, K., Steel, C. M., & Dixon, J. M. (2000). ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. *BMJ (Clinical research ed.), 321*(7261), pp. 624-628. doi:10.1136/bmj.321.7261.624

McRae, A. F., Powell, J. E., Henders, A. K., Bowdler, L., Hemani, G., Shah, S., . . . Montgomery, G. W. (2014). Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biology, 15*(5), p. R73. doi:10.1186/gb-2014-15-5-r73

McTiernan, A., Kooperberg, C., White, E., Wilcox, S., Coates, R., Adams-Campbell, L. L., . . . Study, W. H. (2003). Recreational Physical Activity and the Risk of Breast Cancer in

Postmenopausal Women: The Women's Health Initiative Cohort Study. *JAMA, 290*(10), pp. 1331-1336. doi:10.1001/jama.290.10.1331

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., . . . Pritchard, J. K. (2013). dentification of genetic variants that affect histone modifications in human cells. *Science, 342*(6159), pp. 747–749. doi:10.1126/science.1242429

Mee, B. C., Carroll, P., Donatello, S., Connolly, E., Griffin, M., Dunne, B., . . . Gaffney, E. F. (2011). Maintaining Breast Cancer Specimen Integrity and Individual or Simultaneous Extraction of Quality DNA, RNA, and Proteins from Allprotect-Stabilized and Nonstabilized Tissue Samples. *Biopreservation and Biobanking, 9*(4), pp. 389–398. doi:10.1089/bio.2011.0034

Mehra, R., Varambally, S., Ding, L., Shen, R., Sabel, M. S., Ghosh, D., . . . Kleer, C. G. (2005). Identification of GATA3 as a breast cancer prognostic marker by global gene expression meta-analysis. *Cancer Research, 65*(24), pp. 11259–11264. doi:10.1158/0008-5472.CAN-05-2495

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., . . . CHEK2-Breast Cancer Consortium. (2002). Low-penetrance Susceptibility to Breast Cancer Due to CHEK2(*)1100delC in Noncarriers of BRCA1 or BRCA2 Mutations. *Nature Genetics, 31*(1), pp. 55-59. doi:10.1038/ng879

Meyer, K. B., Maia, A. T., O'Reilly, M., Teschendorff, A. E., Chin, S. F., Caldas, C., & Ponder, B. A. (2008). Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biology, 6*(5), p. e108. doi:10.1371/journal.pbio.0060108

Meyer, K. B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S. L., French, J. D., . . . Easton, D. F. (2013). American Journal of Human Genetics. *Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1, 93*(6), pp. 1046–1060. doi:10.1016/j.ajhg.2013.10.026

Miao, Z., Alvarez, M., Pajukanta, P., & Ko, A. (2018). ASElux: an ultra-fast and accurate allelic reads counter. *Bioinformatics, 34*(8), pp. 1313–1320. doi:10.1093/bioinformatics/btx762

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., . . . Easton, D. F. (2015). Genome-wide Association Analysis of More Than 120,000 Individuals Identifies 15 New Susceptibility Loci for Breast Cancer. *Nature Genetics, 47*(4), pp. 373-380. doi:10.1038/ng.3242

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), pp. 353-361. doi:10.1038/ng.2563

Michailidou, K., Lindström, 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), pp. 92-94. doi:10.1038/nature24284

Michailidou, K., Lindström, 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), pp. 92-94. doi:10.1038/nature24284

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., . . . Skolnick, M. H. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science, 266*(5182), pp. 66-71. doi:10.1126/science.7545954

Milani, L., Lundmark, A., Nordlund, J., Kiialainen, A., Flaegstad, T., Jonmundsson, G., . . . Syvänen, A. C. (2009). Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome Research, 19*(1), pp. 1-11. doi:10.1101/gr.083931.108

Milde-Langosch, K. (2005). The Fos family of transcription factors and their role in tumourigenesis. *European Journal of Cancer, 41*(16), pp. 2449–2461. doi:10.1016/j.ejca.2005.08.008

Milne, R. L., Benítez, J., Nevanlinna, H., Heikkinen, T., Aittomäki, K., Blomqvist, C., . . . Breast Cancer Association Consortium. (2009). Risk of Estrogen Receptor-Positive and -Negative Breast Cancer and Single-Nucleotide Polymorphism 2q35-rs13387042. *Journal of the National Cancer Institute, 101*(14), pp. 1012-1018. doi:10.1093/jnci/djp167

Milne, R. L., Kuchenbaecker, K. B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., . . . Simard, J. (2017). Identification of Ten Variants Associated With Risk of Estrogen-Receptor-Negative Breast Cancer. *Nature Genetics, 49*(12), pp. 1767-1778. doi:10.1038/ng.3785

Missmer, S. A., Smith-Warner, S. A., Spiegelman, D., Yaun, S. S., Adami, H. O., Beeson, W. L., . . . Hunter, D. J. (2002). Meat and Dairy Food Consumption and Breast Cancer: A Pooled Analysis of Cohort Studies. *31*(1), pp. 78-85. doi:10.1093/ije/31.1.78

Mohammadi, P., Castel, S. E., Brown, A. A., & Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Research, 27*(11), pp. 1872-1884. doi:10.1101/gr.216747.116

Mohammadi, P., Castel, S. E., Brown, A. A., & Lappalainen, T. (2017). Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Research, 27*(11), pp. 1872–1884. doi:10.1101/gr.216747.116

Monks, S. A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., . . . Schadt, E. E. (2004). Genetic Inheritance of Gene Expression in Human Cell Lines. *American Journal of Human Genetics, 75*(6), pp. 1904-1105. doi:10.1086/426461

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*(7289), pp. 773–777. doi:10.1038/nature08903

Morgese, F., Soldato, D., Pagliaretta, S., Giampieri, R., Brancorsini, D., Torniai, M., . . . Berardi, R. (2017). Impact of phosphoinositide-3-kinase and vitamin D3 nuclear receptor single-

nucleotide polymorphisms on the outcome of malignant melanoma patients. *Oncotarget, 8*(44), pp. 75914–7592. doi:10.18632/oncotarget.18304

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature, 430*(7001), pp. 743-747. doi:10.1038/nature02797

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature, 430*(7001), pp. 743–747. doi:10.1038/nature02797

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods, 5*(7), pp. 621–628. doi:0.1038/nmeth.1226

Mucci, L. A., Hjelmborg, J. B., Harris, J. R., Czene, K., Havelick, D. J., Scheike, T., . . . Kaprio, E. (2016). Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA, 315*(1), pp. 68-76. doi:10.1001/jama.2015.17703

Muranen, T. A., Mavaddat, N., Khan, S., Fagerholm, R., Pelttari, L., Lee, A., . . . Nevanlinna, H. (2016). Polygenic risk score is associated with increased disease risk in 52 Finnish breast cancer familie. *Breast Cancer Research and Treatment, 158*(3), pp. 463–469. doi:10.1007/s10549-016-3897-6

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., . . . Rader, D. J. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature, 466*(7307), pp. 714–719. doi:10.1038/nature09266

Nagel, J. H., Peeters, J. K., Smid, M., Sieuwerts, A. M., Wasielewski, M., de Weerd, V., . . . Meijers-Heijboer, H. (2012). Gene Expression Profiling Assigns CHEK2 1100delC Breast Cancers to the Luminal Intrinsic Subtypes. *Breast Cancer Research and Treatment, 132*(2), pp. 439-448. doi:10.1007/s10549-011-1588-x

Nasca, P. C., Liu, S., Baptiste, M. S., Kwon, C. S., Jacobson, H., & Metzger, B. B. (1994). Alcohol Consumption and Breast Cancer: Estrogen Receptor Status and Histology. *American Journal of Epidemiology, 140*(11), pp. 980-988. doi:10.1093/oxfordjournals.aje.a117205

National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. (2014). *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General.* Atlanta (GA). Retrieved March 30, 2020, from https://www.ncbi.nlm.nih.gov/books/NBK179276/

Need, A. C., & Goldstein, D. B. (2006). Genome-wide Tagging for Everyone. *Nature Genetics, 38*(11), pp. 1227-1228. doi:10.1038/ng1106-1227

Nelen, M. R., Padberg, G. W., Peeters, E. A., Lin, A. Y., van den Helm, B., Frants, R. R., . . . Eng, C. (1996). Localization of the Gene for Cowden Disease to Chromosome 10q22-23. *Nature Genetics, 13*(1), pp. 114-116. doi:10.1038/ng0596-114

Nelson, D., H., Zakher, B., Cantor, A., Fu, R., Griffin, J., . . . Miglioretti, D. L. (2012). Risk Factors for Breast Cancer for Women Age 40 to 49: A Systematic Review and Meta-analysis. *Annals of Internal Medicine, 156*(9), pp. 635-648. doi:10.1059/0003-4819-156-9-201205010-00006

Nevanlinna, H., & Bartek, J. (2006). The CHEK2 Gene and Inherited Breast Cancer Susceptibility. *Oncogene, 25*(43), pp. 5912-5919. doi:10.1038/sj.onc.1209877

Ni, G., Strom, T. M., Pausch, H., Reimer, C., Preisinger, R., Simianer, H., & Erbe, M. (2015). Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken. *BMC Genomics, 16*, p. 824. doi:10.1186/s12864-015-2059-2

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics, 6*(4), p. e1000895. doi:10.1371/journal.pgen.1000895

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., . . . MuTHER Consortium. (2011). he architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics, 7*(2), p. e1002003. doi:10.1371/journal.pgen.1002003

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics, 6*(4), p. e1000888. doi:10.1371/journal.pgen.1000888

Nieuwenhuis, M. H., Kets, C. M., Murphy-Ryan, M., Yntema, H. G., Evans, D. G., Colas, C., . . . Vasen, H. F. (2014). Cancer Risk and Genotype-Phenotype Correlations in PTEN Hamartoma Tumor Syndrome. *13*(1), pp. 57-63. doi:Cancer Risk and Genotype-Phenotype Correlations in PTEN Hamartoma Tumor Syndrome

Onitilo, A. A., Engel, J. M., Greenlee, R. T., & Mukesh, B. N. (2009). Breast Cancer Subtypes Based on ER/PR and Her2 Expression: Comparison of Clinicopathologic Features and Survival. *Clinical Medicine & Research, 7*(1/2), pp. 4-13. doi:10.3121/cmr.2009.825

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., . . . Lyon, G. J. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine, 5*(3), p. 28. doi:10.1186/gm432

Orr, N., Cooke, R., Jones, M., Fletcher, O., Dudbridge, F., Chilcott-Burns, S., . . . Swerdlow, A. (2011). Genetic variants at chromosomes 2q35, 5p12, 6q25.1, 10q26.13, and 16q12.1 influence the risk of breast cancer in men. *PLoS Genetics, 7*(9), p. e1002290. doi:10.1371/journal.pgen.1002290

Orr, N., Dudbridge, F., Dryden, N., Maguire, S., Novo, D., Perrakis, E., . . . Group, A. O. (2015). Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. *Human molecular genetics, 24*(10), pp. 2966–2984. doi:10.1093/hmg/ddv035

Owens, M. A., Horten, B. C., & Da Silva, M. M. (2004). HER2 amplification ratios by fluorescence in situ hybridization and correlation with immunohistochemistry in a cohort of 6556 breast cancer tissues. *Clinical Breast Cancer, 5*(1), pp. 63-69. doi:10.3816/cbc.2004.n.011

Pai, A. A., Cain, C. E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J. B., . . . Gilad, Y. (2012). The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS Genetics, 8*(10), p. e1003000. doi:10.1371/journal.pgen.1003000

Pai, A. A., Pritchard, J. K., & Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS Genetic, 11*(1), p. e1004857. doi:10.1371/journal.pgen.1004857

Pant, P. V., Tao, H., Beilharz, E. J., Ballinger, D. G., Cox, D. R., & Frazer, K. A. (2006). Analysis of allelic differential expression in human white blood cells. *Genome Research, 16*(3), pp. 331–339. doi:10.1101/gr.4559106

Paraboschi, E. M., Rimoldi, V., Soldà, G., Tabaglio, T., Dall'Osso, C., Saba, E., . . . Asselta, R. (2014). Functional variations modulating PRKCA expression and alternative splicing predispose to multiple sclerosis. *Human Molecular Genetics, 23*(25), pp. 6746–6761. doi:10.1093/hmg/ddu392

Park, D. J., Lesueur, F., Nguyen-Dumont, T., Pertesi, M., Odefrey, F., Hammet, F., . . . Southey, M. C. (2012). Rare Mutations in XRCC2 Increase the Risk of Breast Cancer. *American Journal of Human Genetics, 90*(4), pp. 734-739. doi:10.1016/j.ajhg.2012.02.027

Park, J. H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., . . . Chatterjee, N. (2011). Distribution of Allele Frequencies and Effect Sizes and Their Interrelationships for Common Genetic Susceptibility Variants. *Proceedings of the National Academy of Sciences of the United States of America, 108*(44), pp. 18026–18031. doi:10.1073/pnas.1114759108

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Davies, T. V., . . . Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology, 27*(8), pp. 1160-1167. doi:10.1200/JCO.2008.18.1370

Parts, L., Hedman, Å. K., Keildson, S., Knights, A. J., Abreu-Goodger, C., van de Bunt, M., . . . Lindgren, C. M. (2012). xtent, causes, and consequences of small RNA expression variation in human adipose tissue. *PLoS Genetics, 8*(5), p. e1002704. doi:10.1371/journal.pgen.1002704

Passarelli, M. N., Newcomb, P. A., Hampton, J. M., Trentham-Dietz, A., Titus, L. J., Egan, K. M., . . . Willett, W. C. (2016). Cigarette Smoking Before and After Breast Cancer Diagnosis: Mortality From Breast Cancer and Smoking-Related Diseases. *Journal of Clinical Oncology, 34*(12), pp. 1315–1322. doi:10.1200/JCO.2015.63.9328

Pastinen, T. (2010). Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews. Genetics, 11*(8), pp. 533-538. doi:10.1038/nrg2815

Pastinen, T., & Hudson, T. J. (2004). Cis-acting regulatory variation in the human genome. *Science, 306*(5696), pp. 647-650. doi:10.1126/science.1101659

Pastinen, T., & Hudson, T. J. (2004). Cis-acting regulatory variation in the human genome. *Science, 306*(5696), pp. 647–650. doi:10.1126/science.1101659

Pelttari, L. M., Khan, S., Vuorela, M., Kiiski, J. I., Vilske, S., Nevanlinna, V., . . . Nevanlinna, H. (s.d.). RAD51B in Familial Breast Cancer. *PLoS One, 11*(5), p. e0153788. doi:10.1371/journal.pone.0153788

Peppercorn, J., Perou, C. M., & Carey, L. A. (2008). Molecular Subtypes in Breast Cancer Evaluation and Management: Divide and Conquer. *Cancer Investigation, 26*(1), pp. 1-10. doi:10.1080/07357900701784238

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*, pp. 747-752. doi:doi.org/10.1038/35021093

Peto, J., & Mack, T. M. (2000). High Constant Incidence in Twins and Other Relatives of Women With Breast Cancer. *Nature Genetics, 26*(4), pp. 411-414. doi:10.1038/82533

Peto, J., Collins, N., Barfoot, R., Seal, S., Warren, W., Rahman, N., . . . Stratton, M. R. (1999). Prevalence of BRCA1 and BRCA2 Gene Mutations in Patients With Early-Onset Breast Cancer. *Journal of the National Cancer Institute, 91*(11), pp. 943-949. doi:10.1093/jnci/91.11.943

Pharoah, P. D., Dunning, A. M., Ponder, B. A., & Easton, D. F. (2004). Association Studies for Finding Cancer-Susceptibility Genetic Variants. *Nature Reviews Cancer, 4*(11), pp. 850-860. doi:10.1038/nrc1476

Pharoah, P. D., Guilford, P., Caldas, C., & International Gastric Cancer Linkage Consortium. (2001). Incidence of Gastric Cancer and Breast Cancer in CDH1 (E-cadherin) Mutation Carriers From Hereditary Diffuse Gastric Cancer Families. *Gastroenterology, 121*(6), pp. 1348-1353. doi:10.1053/gast.2001.29611

Pharoah, P. D., Tyrer, J., Dunning, A. M., Easton, D. F., Ponder, B. A., & SEARCH Investigators. (2007). Association between Common Variation in 120 Candidate Genes and Breast Cancer Risk. *3*(3), p. e42. doi:10.1371/journal.pgen.0030042

Piccart-Gebhart, M. J., Procter, M., Leyland-Jones, B., Goldhirsch, A., Untch, M., Smith, I., . . . Gelbert, R. D. (2005). Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *The New England Journal of Medicine, 353*(16), pp. 1659-1672. doi:10.1056/NEJMoa052306

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., . . . Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature, 464*(7289), pp. 768–772. doi:10.1038/nature08872

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., . . . Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature, 464*(7289), pp. 768-772. doi:10.1038/nature08872

Pickrell, J. K., Pai, A. A., Gilad, Y., & Pritchard, J. K. (2010). Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genetics, 6*(12), p. e1001236. doi:10.1371/journal.pgen.1001236

Pike, M. C. (1990). Reducing cancer risk in women through lifestyle-mediated changes in hormone levels. *Cancer Detection and Prevention, 14*(6), pp. 595-607.

Pilarski, R., Burt, R., Kohlman, W., Pho, L., Shannon, K. M., & Swisher, E. (2013). Cowden Syndrome and the PTEN Hamartoma Tumor Syndrome: Systematic Review and Revised Diagnostic Criteria. *ournal of the National Cancer Institute, 105*(21), pp. 1607-1616. doi:10.1093/jnci/djt277

Pirooznia, M., Kramer, M., Parla, J., Goes, F. S., Potash, J. B., McCombie, W. R., & Zandi, P. P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human Genomics, 8*(1), p. 14. doi:10.1186/1479-7364-8-14

Piskol, R., Ramaswami, G., & Li, J. B. (2013). Reliable identification of genomic variants from RNA-seq data. *American Journal of Human Genetics, 93*(4), pp. 641–651. doi:10.1016/j.ajhg.2013.08.008

Pollard, K. S., Serre, D., Wang, X., Tao, H., Grundberg, E., Hudson, T. J., . . . Frazer, K. (2008). A genome-wide approach to identifying novel-imprinted genes. *Human Genetics, 122*(6), pp. 625-634. doi:10.1007/s00439-007-0440-1

Polyak, K. (2007). Breast Cancer: Origins and Evolution. *The Journal of Clinical Investigation, 117*(11), pp. 3155-3163. doi:10.1172/JCI33295

Popadin, K., Gutierrez-Arcelus, M., Dermitzakis, E. T., & Antonarakis, S. E. (2013). Genetic and epigenetic regulation of human lincRNA gene expression. *American Journal of Human Genetics, 93*(6), pp. 1015–1026. doi:10.1016/j.ajhg.2013.10.022

Prat, A., Galván, P., Jimenez, B., Buckingham, W., Jeiranian, H. A., Schaper, C., . . . Alba, E. (2016). Prediction of Response to Neoadjuvant Chemotherapy Using Core Needle Biopsy Samples With the Prosigna Assay. *Clinical Cancer Research, 22*(3), pp. 560-566. doi:10.1158/1078-0432.CCR-15-0630

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., . . . Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research, 12*(5), p. R68. doi:10.1186/bcr2635

Prokopcova, J., Kleibl, Z., Banwell, C. M., & Pohlreich, P. (2007). The Role of ATM in Breast Cancer Development. *104*(2), pp. 121-128. doi:10.1007/s10549-006-9406-6

Przytycki, P. F., & Singh, M. (2020). Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. *Cell Systems, 10*(2), pp. 193–203.e4. doi:10.1016/j.cels.2020.01.002

Purohit, V. (2000). Can Alcohol Promote Aromatization of Androgens to Estrogens? A Review. *Alcohol, 22*(3), pp. 123-127. doi:10.1016/s0741-8329(00)00124-5

Pusztai, L., Ayers, M., Stec, J., Clark, E., Hess, K., Stivers, D., . . . Symmans, W. F. (2003). Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors. *Clinical Cancer Research, 9*(7), pp. 2406-2415.

Quigley, D. A., Fiorito, E., Nord, S., Van Loo, P., Alnæs, G. G., Fleischer, T., . . . Kristensen, V. (2014). The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. *Molecular Oncology, 8*(2), pp. 273–284. doi:10.1016/j.molonc.2013.11.008

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), pp. 841-842. doi:10.1093/bioinformatics/btq033

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. *Nature Genetics, 39*(2), pp. 165-167. doi:10.1038/ng1959

Rakha, E. A., & Green, A. R. (2017). Molecular classification of breast cancer: what the pathologist needs to know. *Pathology, 49*(2), pp. 111-119. doi:10.1016/j.pathol.2016.10.012

Rakha, E. A., Pinder, S. E., Bartlett, J. M., Ibrahim, M., Starczynski, J., Carder, P. J., . . . Ellis, I. O. (2015). Updated UK Recommendations for HER2 assessment in breast cancer. *Journal of Clinical Pathology, 68*(2), pp. 93-99. doi:10.1136/jclinpath-2014-202571

Rakha, E. A., Soria, D., Green, A. R., Lemetre, C., Powe, D. G., Nolan, C. C., . . . Ellis, I. O. (2014). Nottingham Prognostic Index Plus (NPI+): A Modern Clinical Decision Making Tool in Breast Cancer. *British Journal of Cancer, 110*(7), pp. 1688-1697. doi:10.1038/bjc.2014.120

Rastelli, F., & Crispino, S. (2008). Factors Predictive of Response to Hormone Therapy in Breast Cancer. *Tumori, 94*(3), pp. 370-383.

Reddy, T. E., Gertz, J., Pauli, F., Kucera, K. S., Varley, K. E., Newberry, K. M., . . . Myers, R. M. (2012). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research, 22*(5), pp. 860–869. doi:10.1101/gr.131201.111

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. *Nature Genetics, 38*(8), pp. 873-875. doi:10.1038/ng1837

Richardson, K., Nettleton, J. A., Rotllan, N., Tanaka, T., Smith, C. E., Lai, C. Q., . . . Ordovas, J. M. (2013). Gain-of-function lipoprotein lipase variant rs13702 modulates lipid traits through disruption of a microRNA-410 seed site. *American Journal of Human Genetics, 92*(1), pp. 5-14. doi:10.1016/j.ajhg.2012.10.020

Ring, A. E., Smith, I. E., Ashley, S., Fulford, L. G., & Lakhani, S. R. (2004). Oestrogen Receptor Status, Pathological Complete Response and Prognosis in Patients Receiving Neoadjuvant Chemotherapy for Early Breast Cancer. *British Journal of Cancer, 91*(12), pp. 2012-2017. doi:10.1038/sj.bjc.6602235

Rivandi, M. M., & Hollestelle, A. (2018). Elucidating the Underlying Functional Mechanisms of Breast Cancer Susceptibility Through Post-GWAS Analyses. *Frontiers in Genetics, 9*, p. 280. doi:10.3389/fgene.2018.00280

Roadmap Epigenomics Consortium. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), pp. 317-330. doi:10.1038/nature14248

Rockman, M. V., & Kruglyak, L. (2006). Genetics of global gene expression. *Nature Reviews. Genetics, 7*(11), pp. 862-872. doi:10.1038/nrg1964

Rokudai, S., Li, Y., Otaka, Y., Fujieda, M., Owens, D. M., Christiano, A. M., . . . Prives, C. (2018). STXBP4 regulates APC/C-mediated p63 turnover and drives squamous cell carcinogenesis. *Proceedings of the National Academy of Sciences of the United States of America, 115*(21), pp. E4806–E4814. doi:10.1073/pnas.1718546115

Romond, E. H., Perez, E. A., Bryant, J., Suman, V. J., Geyer, C. E., Davidson, N. E., . . . Wolmark, N. (2005). Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *The New England Journal of Medicine, 353*(16), pp. 1673-1684. doi:10.1056/NEJMoa052122

Ronald, J., Brem, R. B., Whittle, J., & Kruglyak, L. (2005). Local Regulatory Variation in Saccharomyces Cerevisiae. *PLoS Genetics, 1*(2), p. e25. doi:10.1371/journal.pgen.0010025

Ronnov-Jessen, L., Petersen, O. W., & Bissell, M. J. (1996). Cellular Changes Involved in Conversion of Normal to Malignant Breast: Importance of the Stromal Reaction. *Physiological Reviews, 76*(1), pp. 69-125. doi:10.1152/physrev.1996.76.1.69

Ross, R. K., Paganin-Hill, A., Wan, P. C., & Pike, M. C. (2000). Effect of Hormone Replacement Therapy on Breast Cancer Risk: Estrogen Versus Estrogen Plus Progestin. *92*(4), pp. 328-332. doi:10.1093/jnci/92.4.328

Ross, R. K., Pike, M. C., Henderson, B. E., Mack, T. M., & Lobo, R. A. (1989). Stroke Prevention and Oestrogen Replacement Therapy. *Lancet, 8636*(505), p. 505. doi:10.1016/s0140-6736(89)91411-6

Rouzier, R., Perou, C. M., Symmans, W. F., Ibrahim, N., Cristofanili, M., Anderson, K., . . . Pusztai, L. (2005). Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy. *11*(16), pp. 5678-5685. doi:10.1158/1078-0432.CCR-04-2421

Roy, R., Chun, J., & Powell, S. N. (2011). BRCA1 and BRCA2: Different Roles in a Common Pathway of Genome Protection. *Nature Reviews. Cancer, 12*(1), pp. 68-78. doi:10.1038/nrc3181

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., . . . Gerstein, M. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology, 7*, p. 522. doi:10.1038/msb.2011.54

RStudio Team. (2010). *RStudio: Integrated Development for R. RStudio, PBC, Boston, MA*. Obtido de http://www.rstudio.com/

Russnes, H. G., Lingjaerde, O. C., Borresen-Dale, A.-L., & Caldas, C. (2017). Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *The American Journal of Pathology, 187*(10), pp. 2152-2162. doi:10.1016/j.ajpath.2017.04.022

Sabatti, C., Service, S., & Freimer, N. (2003). False Discovery Rate in Linkage and Association Genome Screens for Complex Disorders. *Genetics, 164*(2), pp. 829-833.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., . . . International SNP Map Working Group. (2001). A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms. *Nature, 409*(6822), pp. 928-933. doi:10.1038/35057149

Sandmann, S., de Graaf, A. O., Karimi, M., van der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., & Dugas, M. (2017). Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports, 7*, p. 43169. doi:10.1038/srep43169

Saunier, E. F., & Akhurst, R. J. (2006). TGF Beta Inhibition for Cancer Therapy. *Current Cancer Drug Targets, 6*(7), pp. 565-578. doi:10.2174/156800906778742460

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., . . . Friend, S. H. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature, 422*(6929), pp. 297–302. doi:10.1038/nature01434

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research, 22*(9), pp. 1748–175. doi:10.1101/gr.136127.111

Schmidt, M. K., Hogervorst, F., van Hien, R., Cornelissen, S., Broeks, A., Adank, M. A., . . . Easton, D. F. (2016). Age- And Tumor Subtype-Specific Breast Cancer Risk Estimates for CHEK2*1100delC Carriers. *Journal of Clinical Oncology, 34*(23), pp. 2750-2760. doi: 10.1200/JCO.2016.66.5844

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes, 9*, p. 88. doi:10.1186/s13104-016-1900-2

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. *Nature Genetics, 38*(11), pp. 1239-1241. doi:10.1038/ng1902

Serre, D., Gurd, S., Ge, B., Sladek, R., Sinnett, D., Harmsen, E., . . . Hudson, T. J. (2008). Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genetics, 4*(2), p. e1000006. doi:10.1371/journal.pgen.1000006

Seshadri, R., Firgaira, F. A., Horsfall, D. J., McCaul, K., Setlur, V., & Kitchen, P. (1993). Clinical significance of HER-2/neu oncogene amplification in primary breast cancer. The South Australian Breast Cancer Study Group. *11*(10), pp. 1936-1942. doi:10.1200/JCO.1993.11.10.1936

Sestak, I., Cuzick, J., Dowsett, M., Lopez-Knowles, E., Filipits, M., Dubsky, P., . . . Gnant, M. (2015). Prediction of late distant recurrence after 5 years of endocrine treatment: a combined analysis of patients from the Austrian breast and colorectal cancer study group 8 and arimidex, tamoxifen alone or in combination randomized trials using the PAM50 risk. *Journal of Clinical Oncology, 33*(8), pp. 916-922. doi:10.1200/JCO.2014.55.6894

Shaulian, E., & Karin, M. (2002). AP-1 as a regulator of cell life and death. *Nature Cell Biology, 4*(5), pp. E131–E136. doi:10.1038/ncb0502-e131

Shen, J., Medico, L., & Zhao, H. (2011). Allelic imbalance in BRCA1 and BRCA2 gene expression and familial ovarian cancer. *Cancer Epidemiology, Biomarkers & Prevention, 20*(1), pp. 50–56. doi:10.1158/1055-9965.EPI-10-0720

Shen, Q., Uray, I. P., Li, Y., Krisko, T. I., Strecker, T. E., Kim, H. T., & Brown, P. H. (2008). The AP-1 transcription factor regulates breast cancer cell growth via cyclins and E2F factors. *Oncogene, 27*(3), pp. 366-377. doi:10.1038/sj.onc.1210643

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. *International Journal of Cancer, 139*(6), pp. 1303–1317. doi:10.1002/ijc.30150

Shi, Z., Chiang, C. I., Labhart, P., Zhao, Y., Yang, J., Mistretta, T. A., . . . Mori-Akiyama, Y. (2015). Context-specific role of SOX9 in NF-Y mediated gene regulation in colorectal cancer cells. *Nucleic Acids Research, 43*(13), pp. 6257–6269. doi:10.1093/nar/gkv568

Shieh, Y., Eklund, M., Sawaya, G. F., Black, W. C., Kramer, B. S., & Esserman, L. J. (2016). Population-based screening for cancer: hope and hype. *Nature Reviews Clinical Oncology, 13*(9), pp. 550-565. doi:10.1038/nrclinonc.2016.50

Siddiq, A., Couch, F. J., Chen, G. K., Lindström, S., Eccles, D., Millikan, R. C., . . . Vachon, C. M. (2012). A Meta-Analysis of Genome-Wide Association Studies of Breast Cancer Identifies Two Novel Susceptibility Loci at 6q14 and 20q11. *Human Molecular Genetics, 21*(14), pp. 5373-5384. doi:10.1093/hmg/dds381

Sigurdsson, M. I., Saddic, L., Heydarpour, M., Chang, T. W., Shekar, P., Aranki, S., . . . Muehlschlegel, J. D. (2016). Allele-specific expression in the human heart and its

application to postoperative atrial fibrillation and myocardial ischemia. *Genome Medicine, 8*(1), p. 127. doi:10.1186/s13073-016-0381-1

Silverstein, M. J., Poller, D. N., Waisman, J. R., Colburn, W. J., Barth, A., Gierson, E. D., . . . Slamon, D. J. (1995). Prognostic classification of breast ductal carcinoma-in-situ. *Lancet, 345*, pp. 1154-1157. doi:10.1016/s0140-6736(95)90982-6

Singletary, E. S. (2003). Rating the Risk Factors for Breast Cancer. *Annalas of Surgery, 237*(4), pp. 474-482. doi:10.1097/01.SLA.0000059969.64262.87

Singletary, K. W., & Gapstur, S. M. (2001). Alcohol and Breast Cancer: Review of Epidemiologic and Experimental Evidence and Potential Mechanisms. *JAMA, 286*(17), pp. 2143-2151. doi:10.1001/jama.286.17.2143

Sinn, H.-P., & Kreipe, H. (2013). A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. *Breast Care (Basel, Switzerland), 8*(2), pp. 149-154. doi:10.1159/000350774

Skelly, D. A., Ronald, J., & Akey, J. M. (2009). Inherited variation in gene expression. *Annual Review of Genomics and Human Genetics, 10*, pp. 313–332. doi:0.1146/annurev-genom-082908-150121

Slamon, D. J., Clark, G. M., Wong, S. G., Levin, W. J., Ullrich, A., & McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science, 235*(4785), pp. 177-182. doi:10.1126/science.3798106

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., . . . Norton, L. (2001). Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *The New England Journal of Medicine, 344*(11), pp. 783-792. doi:10.1056/NEJM200103153441101

Smith, P., McGuffog, L., Easton, D. F., Mann, G. J., Pupo, G. M., Newman, B., . . . Stratton, M. R. (2006). A Genome Wide Linkage Search for Breast Cancer Susceptibility Genes. *Genes Chromoses Cancer, 45*(7), pp. 646-655. doi:10.1002/gcc.20330

Soderlund, C. A., Nelson, W. M., & Goff, S. A. (2014). Allele Workbench: transcriptome pipeline and interactive graphics for allele-specific expression. *PLoS One, 9*(12), p. e115740. doi:10.1371/journal.pone.0115740

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., . . . Borresen-Dale, A.-L. (2001). Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses With Clinical Implications. *98*(19), pp. 10869-10874. doi:10.1073/pnas.191367098

Sorlie, T., Robert, T., Parker, J., Hastie, T., Marron, J. S., Nobel, A., . . . Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America, 100*(14), pp. 8418-8423. doi:10.1073/pnas.0932692100

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., . . . Liu, E. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences of the United States of America, 100*(18), pp. 10393-10298. doi:10.1073/pnas.1732912100

Soule, H. D., Vazguez, J., Long, A., Albert, S., & Brennan, M. (1973). A human cell line from a pleural effusion derived from a breast carcinoma. *Journal of the National Cancer Institute, 51*(5), pp. 1409-1416. doi:10.1093/jnci/51.5.1409

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., . . . Loos, R. J. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics, 42*(11), pp. 937–948. doi:10.1038/ng.686

Spurdle, A. B., Whiley, P. J., Thompson, B., Feng, B., Healey, S., Brown, M. A., . . . Consortium, E. (2012). BRCA1 R1699Q Variant Displaying Ambiguous Functional Abrogation Confers Intermediate Breast and Ovarian Cancer Risk. *Journal of Medical Genetics, 49*(8), pp. 525-532. doi:10.1136/jmedgenet-2012-101037

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), pp. 865-869. doi:10.1038/ng2064

Stacey, S. N., Manolescu, A., Sulem, P., Thorlacius, S., Gudjonsson, S. A., Jonsson, G. F., . . . Stefansson, K. (2008). Common Variants on Chromosome 5p12 Confer Susceptibility to Estrogen Receptor-Positive Breast Cancer. *Nature Genetics, 40*(6), pp. 703-706. doi:10.1038/ng.131

Stacey, S. N., Sulem, P., Zanon, C., Gudjonsson, S. A., Thorleifsson, G., Helgason, A., . . . Stefansson, K. (2010). Ancestry-shift refinement mapping of the C6orf97-ESR1 breast cancer susceptibility locus. *PLoS Genetic, 6*(7), p. e1001029. doi:10.1371/journal.pgen.1001029

Steffen, J., Nowakowska, D., Niwińska, A., Czapczak, D., Kluska, A., Piatkowska, M., . . . Paszko, Z. (2006). Germline Mutations 657del5 of the NBS1 Gene Contribute Significantly to the Incidence of Breast Cancer in Central Poland. *International Journal of Cancer, 119*(2), pp. 472-475. doi:10.1002/ijc.21853

Stegeman, S., Moya, L., A., S. L., Spurdle, A. B., Clements, J. A., & Batra, J. (2015). A genetic variant of MDM4 influences regulation by multiple microRNAs in prostate cancer. *Endocrine-Related Cancer, 22*(2), pp. 265-276. doi:10.1530/ERC-15-0013

Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., . . . Stratton, M. R. (2012). The Landscape of Cancer Genes and Mutational Processes in Breast Cancer. *Nature, 486*(7403), pp. 400-404. doi:10.1038/nature11017

Stevenson, K. R., Coolon, J. D., & Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC Genomics, 14*, p. 536. doi:10.1186/1471-2164-14-536

Stewart, B. W., & Wild, C. P. (Eds.). (2014). *World Cancer Report 2014.* Lyon, France: International Agency for Research on Cancer.

Stingl, J., & Caldas, C. (2007). Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nature Reviews Cancer, 7*, pp. 791-799. doi:10.1038/nrc2212

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., . . . Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science, 315*(5813), pp. 848–853. doi:10.1126/science.1136678

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., . . . Dermitzakis, E. T. (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genetics, 8*(4), p. e1002639. doi:10.1371/journal.pgen.1002639

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., . . . Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics, 39*(10), pp. 1217-1224. doi:10.1038/ng2142

Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics, 187*(2), pp. 367-383. doi:10.1534/genetics.110.120907

Stratton, M. R., & Rahman, N. (2008). The Emerging Landscape of Breast Cancer Susceptibility. *Nature Genetics, 40*(1), pp. 17-22. doi:10.1038/ng.2007.53

Sturm, M., Schroeder, C., & Bauer, P. (2016). SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. *BMC Bioinformatics, 17*, p. 208. doi:10.1186/s12859-016-1069-7

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., . . . Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature, 526*(7571), pp. 75-81. doi:10.1038/nature15394

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education, 4*(3), pp. 279–282. doi:10.4300/JGME-D-12-00156.1

Sun, W. (2012). A statistical framework for eQTL mapping using RNA-seq data. *Biometrics, 68*(1), pp. 1-11. doi:10.1111/j.1541-0420.2011.01654.x

Sun, W. Y., Kim, H. M., Jung, W. H., & Koo, J. S. (2016). Expression of serine/glycine metabolism-related proteins is different according to the thyroid cancer subtype. *14*(1), p. 168. doi:10.1186/s12967-016-0915-8

Sun, Y., Ye, C., Guo, X., Wen, W., Long, J., Gao, Y. T., . . . Cai, Q. (2016). Evaluation of potential regulatory function of breast cancer risk locus at 6q25.1. *Carcinogenesis, 37*(2), pp. 163–168. doi:10.1093/carcin/bgv170

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians, 71*(3), pp. 209–249. doi:10.3322/caac.21660

Takaku, M., Grimm, S. A., & Wade, P. A. (2015). GATA3 in Breast Cancer: Tumor Suppressor or Oncogene? *Gene Expression, 16*(4), pp. 163–168. doi:10.3727/105221615X14399878166113

Takeichi, M. (1991). Cadherin Cell Adhesion Receptors as a Morphogenetic Regulator. *Science, 251*(5000), pp. 1451-1455. doi:10.1126/science.2006419

Tan, A. C., Fan, J. B., Karikari, C., Bibikova, M., Garcia, E. W., Zhou, L., . . . Chakravarti, A. (2008). Allele-specific expression in the germline of patients with familial pancreatic cancer: an unbiased approach to cancer gene discovery. *Cancer Biology & Therapy, 7*(1), pp. 135–144. doi:10.4161/cbt.7.1.5199

Tang, L., Peng, C., Zhu, S. S., Zhou, Z., Liu, H., Cheng, Q., . . . Chen, X. P. (2020). Tre2-Bub2-Cdc16 Family Proteins Based Nomogram Serve as a Promising Prognosis Predicting Model for Melanoma. *Frontiers in Oncology, 10*, p. 579625. doi:10.3389/fonc.2020.579625

Taylor-Papadimitriou, J., Stampfer, M., Bartek, J., Lewis, A., Boshell, M., Lane, E. B., & Leigh, I. M. (1989). Keratin Expression in Human Mammary Epithelial Cells Cultured From Normal and Malignant Tissue: Relation to in Vivo Phenotypes and Influence of Medium. *94*(Pt 3), pp. 403-413.

Tehranchi, A. K., Myrthil, M., Martin, T., Hie, B. L., Golan, D., & Fraser, H. B. (2016). Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell, 165*(3), pp. 730–741. doi:10.1016/j.cell.2016.03.041

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 Research, 15*(1), p. R17. doi:10.1186/bcr3392

Terry, M. B., Zhang, F. F., Kabat, G., Britton, J. A., Teitelbaum, S. L., Neugut, A. I., & Gammon, M. D. (2006). Lifetime Alcohol Intake and Breast Cancer Risk. *Annals of Epidemiology, 16*(3), pp. 230-240. doi:10.1016/j.annepidem.2005.06.048

Terry, P. D., & Rohan, T. E. (2002). Cigarette Smoking and the Risk of Breast Cancer in Women: A Review of the Literature. *Cancer Epidemiology, Biomarkers & Prevention, 11*(10 Pt 1), pp. 953–971.

Tesson, B. M., & Jansen, R. C. (2009). eQTL analysis in mice and rats. *Methods in Molecular Biology, 573*, pp. 285–309. doi:10.1007/978-1-60761-247-6_16

The International HapMap 3 Consortium. (2010). Integrating Common and Rare Genetic Variation in Diverse Human Populations. *Nature, 467*(7311), pp. 52-58. doi:10.1038/nature09298

The International HapMap Consortium. (2003). The International HapMap Project. *Nature, 426*, pp. 789-796. doi:10.1038/nature02168

The International HapMap Consortium. (2004). Integrating ethics and science in the International HapMap Project. *Nature reviews. Genetics, 5*, pp. 467-475. doi:10.1038/nrg1351

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature, 437*, pp. 1299-1320. doi:10.1038/nature04226

The International HapMap Consortium. (2007). A Second Generation Human Haplotype Map of Over 3.1 Million SNPs. *Nature, 449*(7164), pp. 851-861. doi:10.1038/nature06258

Thomas, D. C. (2006). Are We Ready for Genome-Wide Association Studies? *Cancer Epidemiology, Biomarkers & Prevention, 15*(4), pp. 595-598. doi:10.1158/1055-9965.EPI-06-0146

Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., . . . Hunter, D. J. (2009). A Multistage Genome-Wide Association Study in Breast Cancer Identifies Two New Risk Alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nature Genetics, 41*(5), pp. 579-584. doi:10.1038/ng.353

Thompson, D., Duedal, S., Kirner, J., McGuffog, L., Last, J., Reiman, A., . . . Easton, D. F. (2005). Cancer Risks and Mortality in Heterozygous ATM Mutation Carriers. *ournal of the National Cancer Institute, 97*(11), pp. 813-822. doi:10.1093/jnci/dji141

Thompson, D., Easton, D., & Breast Cancer Linkage Consortium. (2001). Variation in Cancer Risks, by Mutation Position, in BRCA2 Mutation Carriers. *American Journal of Human Genetics, 68*(2), pp. 410-419. doi:10.1086/318181

Thompson, E. R., Doyle, M. A., Ryland, G. L., Rowley, S. M., Choong, D. Y., Tothill, R. W., . . . Campbell, I. G. (2012). Exome Sequencing Identifies Rare Deleterious Mutations in DNA Repair Genes FANCC and BLM as Potential Breast Cancer Susceptibility Alleles. *PLoS Genetics, 8*(9), p. e1002894. doi:10.1371/journal.pgen.1002894

Thorisson, G. A., & Stein, L. D. (2003). The SNP Consortium Website: Past, Present and Future. *Nucleic Acids Research, 31*(1), pp. 124-127. doi:10.1093/nar/gkg052

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology, 19*(1A), pp. A68–A77. doi:10.5114/wo.2014.47136

Tomlinson, I. P., & Houlston, R. S. (1997). Peutz-Jeghers Syndrome. *Journal of Medical Genetics, 34*(12), pp. 1007-1011. doi:10.1136/jmg.34.12.1007

Torgerson, D. G., Boyko, A. R., Hernandez, R. D., Indap, A., Hu, X., White, T. J., . . . Clark, A. G. (2009). Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. *PLoS Genetics, 5*(8), p. e1000592. doi:10.1371/journal.pgen.1000592

Toriola, A. T. (2013). Trends in Breast Cancer Incidence and Mortality in the United States: Implications for Prevention. *Breast Cancer Research and Treatment, 138*(3), pp. 665-673. doi:10.1007/s10549-013-2500-7

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. *Nature Genetics, 43*(12), pp. 1193–1201. doi:10.1038/ng.998

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., . . . Easton, D. F. (2010). Genome-wide Association Study Identifies Five New Breast Cancer Susceptibility Loci. *Nature Genetics, 42*(6), pp. 504-507. doi:10.1038/ng.586

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), pp. 2507–2515. doi:10.1093/hmg/ddq122

Udler, M. S., Meyer, K. B., Pooley, K. A., Karlins, E., Struewing, J. P., Zhang, J., . . . Collaborators, S. (2009). FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Human Molecular Genetics, 18*(9), pp. 1692–1703. doi:10.1093/hmg/ddp078

Ursin, G., Longnecker, M. P., Haile, R. W., & Greenland, S. (1995). A meta-analysis of body mass index and risk of premenopausal breast cancer. *Epidemiology, 6*(2), pp. 137-141. doi:10.1097/00001648-199503000-00009

Valencia, O. M., Samuel, S. E., Viscusi, R. K., Neumayer, L. A., & Aziz, H. (2017). The Role of Genetic Testing in Patients With Breast Cancer: A Review. *JAMA Surgery, 152*(6), pp. 589–594. doi:10.1001/jamasurg.2017.0552

Valle, L., Serena-Acedo, T., Liyanarachchi, S., Hampel, H., Comeras, I., Li, Z., . . . de la Chapelle, A. (2008). Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science, 321*(5894), pp. 1361–1365. doi:10.1126/science.1159397

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods, 12*(11), pp. 1061–1063. doi:10.1038/nmeth.3582

van de Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods, 12*(11), pp. 1061–1063. doi:10.1038/nmeth.3582

van der Groep, P., van der Wall, E., & van Diest, P. J. (2011). Pathology of Hereditary Breast Cancer. *Cellular Oncology, 34*(2), pp. 71-88. doi:10.1007/s13402-011-0010-3

van Nas, A., Ingram-Drake, L., Sinsheimer, J. S., Wang, S. S., Schadt, E. E., Drake, T., & Lusis, A. J. (2010). Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics, 185*(3), pp. 1059–1068. doi:10.1534/genetics.110.116087

Varmus, H. (2003). Genomic Empowerment: The Importance of Public Databases. *Nature Genetics, 35 Suppl 1*, p. 3. doi:10.1038/ng1186

Verlaan, D. J., Ge, B., Grundberg, E., Hoberman, R., Lam, K. C., Koka, V., . . . Pastinen, T. (2009). Targeted screening of cis-regulatory variation in human haplotypes. *Genome Research, 19*(1), pp. 118-127. doi:10.1101/gr.084798.108

Verlaan, D. J., Ge, B., Grundberg, E., Hoberman, R., Lam, K. C., Koka, V., . . . Pastinen, T. (2009). Targeted screening of cis-regulatory variation in human haplotypes. *Genome Research, 19*(1), pp. 118-127. doi:10.1101/gr.084798.108

Vleugel, M. M., Greijer, A. E., Bos, R., van der Wall, E., & van Diest, P. J. (2006). c-Jun activation is associated with proliferation and angiogenesis in invasive breast cancer. *Human Pathology, 37*(6). doi:10.1016/j.humpath.2006.01.022

Voduc, K. D., Cheang, M. C., Tyldesley, S., Gelmon, K., Nielsen, T. O., & Kennecke, H. (2010). Breast cancer subtypes and the risk of local and regional relapse. *Journal of Clinical Oncology, 28*(10), pp. 1684-1691. doi:10.1200/JCO.2009.24.9284

Walker, R. A. (2008). Immunohistochemical markers as predictive tools for breast cancer. *Journal of Clinical Pathology, 61*(6), pp. 689-696. doi:10.1136/jcp.2006.041830

Wang, D., Poi, M. J., Sun, X., Gaedigk, A., Leeder, J. S., & Sadee, W. (2014). Common CYP2D6 polymorphisms affecting alternative splicing and transcription: long-range haplotypes with two regulatory variants modulate CYP2D6 activity. *Human Molecular Genetics, 23*(1), pp. 268–278. doi:10.1093/hmg/ddt417

Wang, H., Thomas, D. C., Pe'er, I., & Stram, D. O. (2006). Optimal Two-Stage Genotyping Designs for Genome-Wide Association Scans. *Genetic Epidemiology, 30*(4), pp. 356-368. doi:10.1002/gepi.20150

Wang, W. Y., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide Association Studies: Theoretical and Practical Concerns. *Nature Reviews. Genetics, 6*(2), pp. 109-118. doi:10.1038/nrg1522

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*(Database issue), pp. D930–D934. doi:10.1093/nar/gkr917

Weigelt, B., Geyer, F. C., & Reis-Filho, J. S. (2010). Histological types of breast cancer: How special are they? *Molecular Oncology, 4*(3), pp. 192-208. doi:10.1016/j.molonc.2010.04.004

Weigelt, B., Glas, A. M., Wessels, L. F., Witteveen, A. T., Peterse, J. L., & van´t Veer, L. J. (2003). Gene Expression Profiles of Primary Breast Tumors Maintained in Distant Metastases. *Proceedings of the National Academy of Sciences of the United States of America, 100*(26), pp. 15901-15905. doi:10.1073/pnas.2634067100

Weigelt, B., Mackay, A., A'Hern, R., Natrajan, R., Tan, D. S., Dowsett, M., . . . Reis-Filho, J. S. (2010). Breast Cancer Molecular Profiling With Single Sample Predictors: A

Retrospective Analysis. *Lancet Oncology, 11*(4), pp. 339-349. doi:10.1016/S1470-2045(10)70008-5

Weischer, M., Bojesen, S. E., Ellervik, C., Tybjaerg-Hansen, A., & Nordestgaard, B. G. (2008). CHEK2*1100delC Genotyping for Clinical Assessment of Breast Cancer Risk: Meta-Analyses of 26,000 Patient Cases and 27,000 Controls. *Journal of Clinical Oncology, 26*(4), pp. 542-548. doi:10.1200/JCO.2007.12.5922

Weischer, M., Nordestgaard, B. G., Pharoah, P., Bolla, M. K., Nevanlinna, H., Van't Veer, L. J., . . . Giles, G. G. (2012). CHEK2*1100delC Heterozygosity in Women With Breast Cancer Associated With Early Death, Breast Cancer–Specific Death, and Increased Risk of a Second Breast Cancer. *Journal of Clinical Oncology, 30*(35), pp. 4308–4316. doi:10.1200/JCO.2012.42.7336

Welch, D. R. (2006). Do we need to redefine a cancer metastasis and staging definitions? *Breast Disease, 26*, pp. 3-12. doi:10.3233/bd-2007-26102

Welch, D. R., & Hurst, D. R. (2019). Defining the Hallmarks of Metastasis. *Cancer Research*. doi:10.1158/0008-5472.CAN-19-0458

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), pp. 661–678. doi:10.1038/nature05911

Wellings, S. R., Jensen, H. M., & Marcum, R. G. (1975). An atlas of subgross pathology of the human breast with special reference to possible precancerous lesions. *Journal of the National Cancer Institute, 55*(2), pp. 231-273. doi:10.1093/jnci/55.2.231

Wendt, C., & Margolin, S. (2019). Identifying breast cancer susceptibility genes - a review of the genetic background in familial breast cancer. *Acta Oncologica, 58*(2), pp. 135–146. doi:10.1080/0284186X.2018.1529428

Wolff, A. C., Hammond, E. H., Allison, K. H., Harvey, B. E., Mangu, P. B., Bartlett, J. M., . . . Dowsett, M. (2018). Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Archives of Pathology & Laboratory Medicine, 142*(11), pp. 1364-1382. doi:10.5858/arpa.2018-0902-SA

Wolff, A. C., Hammond, M. H., Schwartz, J. N., Hagerty, K. L., Allred, D. C., Dowsett, R. J., . . . Hayes, D. F. (2007). American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer. *Journal of Clinical Oncology, 25*(1), pp. 118-145. doi:10.1200/JCO.2006.09.2775

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), pp. 989-992. doi:10.1038/378789a0

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics, 26*(7), pp. 873–881. doi:10.1093/bioinformatics/btq057

Wynendaele, J., Böhnke, A., Leucci, E., Nielsen, S. J., Lambertz, I., Hammer, S., . . . Bartel, F. (2010). An illegitimate microRNA target site within the 3' UTR of MDM4 affects ovarian cancer progression and chemosensitivity. *Cancer Research, 70*(23), pp. 9641–9649. doi:10.1158/0008-5472.CAN-10-0527

Wyszynski, A., Hong, C. C., Lam, K., Michailidou, K., Lytle, C., Yao, S., . . . Cole, M. D. (2016). An intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression. *Human Molecular Genetics, 25*(17), pp. 3863–3876. doi:10.1093/hmg/ddw223

Xavier, J., Russell, R., Almeida, B. P., Rosli, N., Rocha, C., Samarajiwa, S., . . . Maia, A.-T. (2016). Abstract A31: Integrative differential allelic expression analysis efficiently reveals the biology underlying risk to breast cancer. *Proceedings of the AACR Special Conference on Advances in Breast Cancer Research*, *14*, p. A31. doi:10.1158/1557-3125.ADVBC15-A31

Xia, B., Sheng, Q., Nakanishi, K., Ohashi, A., Wu, J., Christ, N., . . . Livingston, D. M. (2006). Control of BRCA2 Cellular and Clinical Functions by a Nuclear Partner, PALB2. *Molecular Cell, 22*(6), pp. 719-729. doi:10.1016/j.molcel.2006.05.022

Xiao, R., & Scott, L. J. (2011). Detection of cis-acting regulatory SNPs using allelic expression data. *Genetic Epidemiology, 35*(6), pp. 515–525. doi:10.1002/gepi.20601

Xu, H., Fu, J., Ha, S. W., Ju, D., Zheng, J., Li, L., & Xie, Y. (2012). The CCAAT box-binding transcription factor NF-Y regulates basal expression of human proteasome genes. *Biochimica et Biophysica Acta, 1823*(4), pp. 818-825. doi:10.1016/j.bbamcr.2012.01.002

Xu, J., Chen, Y., & Olopade, O. I. (2010). MYC and Breast Cancer. *Genes & Cancer, 1*(6), pp. 629–640. doi:0.1177/1947601910378691

Yan, H., Dobbie, Z., Gruber, S. B., Markowitz, S., Romans, K., Giardiello, F. M., . . . Vogelstein, B. (2002). Small changes in expression affect predisposition to tumorigenesis. *Nature Genetics, 30*(1), pp. 25–26. doi:10.1038/ng799

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science, 297*(5584), p. 1143. doi:10.1126/science.1072545

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science, 297*(5584), p. 1143. doi:10.1126/science.1072545

Yang, F., Xiao, Z., & Zhang, S. (2018). Knockdown of miR-194-5p inhibits cell proliferation, migration and invasion in breast cancer by regulating the Wnt/β-catenin signaling pathway. *International Journal of Molecular Medicine, 42*(6), pp. 3355–3363. doi:0.3892/ijmm.2018.3897

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., . . . Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics, 42*(7), pp. 565–569. doi:10.1038/ng.608

Yardena S., Z. W., B., A., S., N., P., J., S., S., Y., H., . . . V., V. E. (2004). High frequency of mutations of the PIK3CA gene in human cancers. *Science, 304*(5670), p. 554. doi:10.1126/science.1096502

Yaziji, H., Goldstein, L. C., Barry, T. S., Werling, R., Hwang, H., Ellis, G. K., . . . Gown, A. M. (2004). HER-2 Testing in Breast Cancer Using Parallel Tissue-Based Methods. *JAMA, 291*(16), pp. 1972-1977. doi:10.1001/jama.291.16.1972

Yerushalmi, R., Woods, R., Ravdin, P. M., Hayes, M. M., & Gelmon, K. A. (2010). Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncology, 11*(2). doi:10.1016/S1470-2045(09)70262-1

Yu, K.-D., Shen, Z.-Z., & Shao, Z.-M. (2009). The immunohistochemically "ER-negative, PR-negative, HER2-negative, CK5/6-negative, and HER1-negative" subgroup is not a surrogate for the normal-like subtype in breast cancer. *Breast Cancer Research and Treatment, 118*(3), pp. 661-663. doi:10.1007/s10549-009-0522-y

Yu, X., & Sun, S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics, 14*, p. 274. doi:10.1186/1471-2105-14-274

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), p. 64. doi:10.1186/s13058-016-0718-0

Zhang, G., Zeng, Y., Liu, Z., & Wei, W. (2013). Significant Association Between Nijmegen Breakage Syndrome 1 657del5 Polymorphism and Breast Cancer Risk. *Tumour Biology, 34*(5), pp. 2753-2757. doi:10.1007/s13277-013-0830-z

Zhang, J., Zhao, C. Y., Zhang, S. H., Yu, D. H., Chen, Y., Liu, Q. H., . . . Zhu, M. H. (2014). Upregulation of miR-194 contributes to tumor growth and progression in pancreatic ductal adenocarcinoma. *Oncology Reports, 31*(3), pp. 1157–1164. doi:10.3892/or.2013.2960

Zhang, L., Liu, Y., Song, F., Zheng, H., Hu, L., Lu, H., . . . Chen, K. (2011). Functional SNP in the microRNA-367 binding site in the 3'UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification. *Proceedings of the National Academy of Sciences of the United States of America, 108*(33), pp. 13653–13658. doi:0.1073/pnas.1103360108

Zhang, W. C., Shyh-Chang, N., Yang, H., Rai, A., Umashankar, S., Ma, S., . . . Lim, B. (2012). Glycine decarboxylase activity drives non-small cell lung cancer tumor-initiating cells and tumorigenesis. *Cell, 148*(1-2), pp. 259–272. doi:10.1016/j.cell.2011.11.050

Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., . . . … Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Natures, 507*(7493), pp. 455-461. doi:10.1038/nature12787

Zhao, Z., Fu, Y. X., Hewett-Emmett, D., & Boerwinkle, E. (2003). Investigating Single Nucleotide Polymorphism (SNP) Density in the Human Genome and Its Implications for Molecular Evolution. *Gene, 312*, pp. 201-213. doi:10.1016/s0378-1119(03)00670-x

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), pp. 324-328. doi:10.1038/ng.318

Supplementary Material

Chapter IV. Identification of Target Genes and Causal Variants in Known Breast Cancer Risk Loci

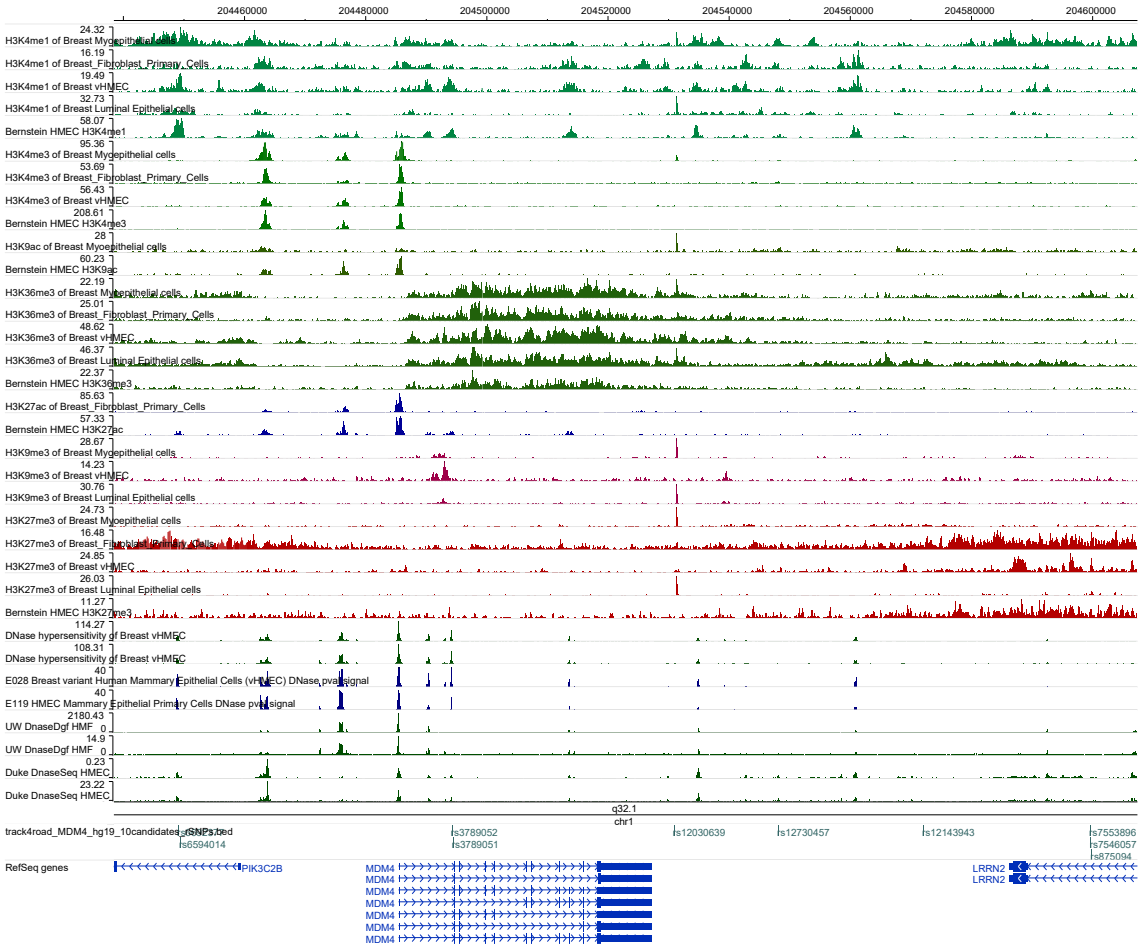


Figure S4.1. Genomic landscape of the 1q32.1 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 3 GWAS SNPs (rs2290854, rs4245739 and rs6682208) associated with breast cancer risk, and 12 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The variant rs6682208 is both a GWAS SNP and a daeSNP. The daeSNP rs11240762 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 44 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs11240762. From those, 28 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. This figure shows the 9 variants with strongest regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

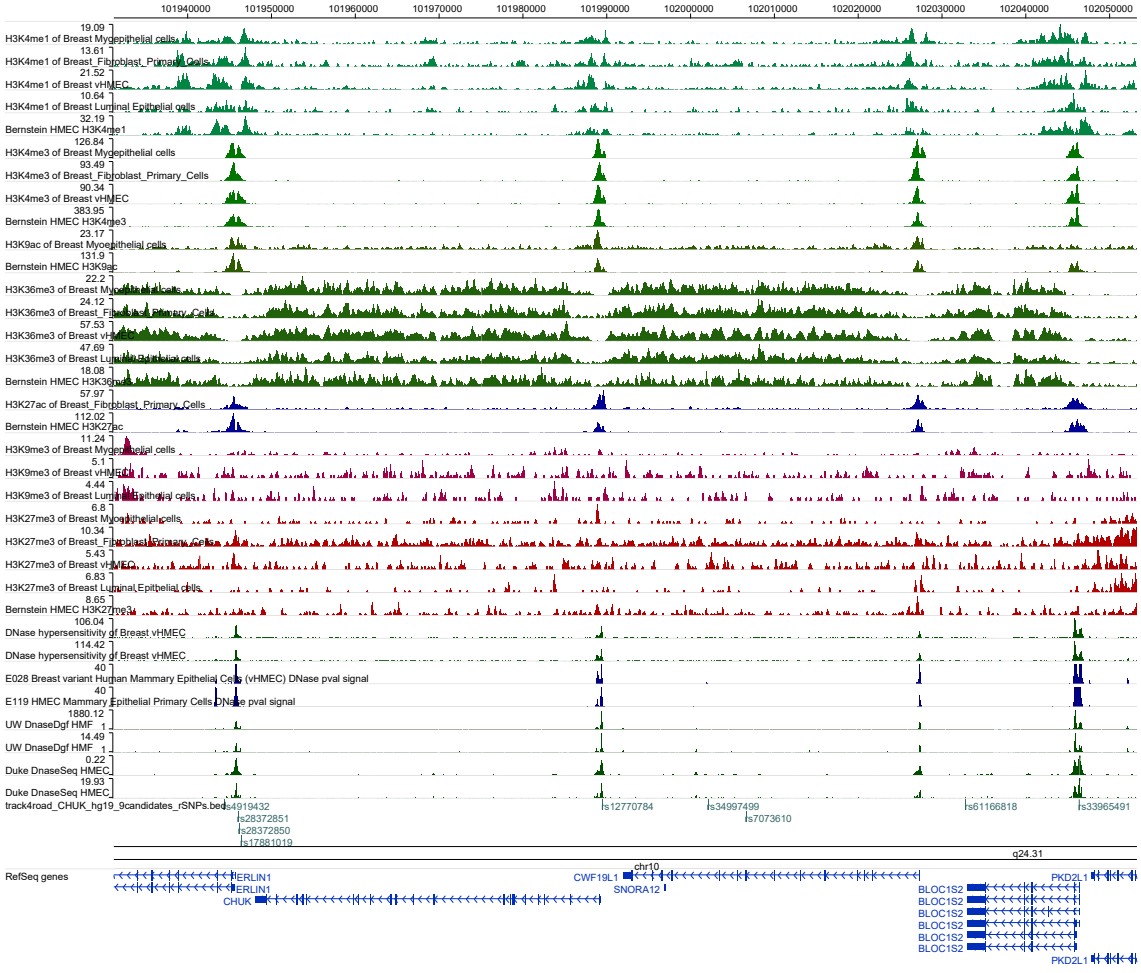


Figure S4.2. Genomic landscape of the 10q24.31 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs2298075) associated with breast cancer risk, and 15 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs3180413 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 108 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs3180413. From those, 59 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. This figure shows the 9 variants with strongest regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

Figure S4.3. Genomic landscape of the 14q32.12 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs941764) associated with breast cancer risk, and 8 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs12886328 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 27 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs12886328. From those, 6 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. All those 6 SNPs had strong regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast, and are shown in the figure. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

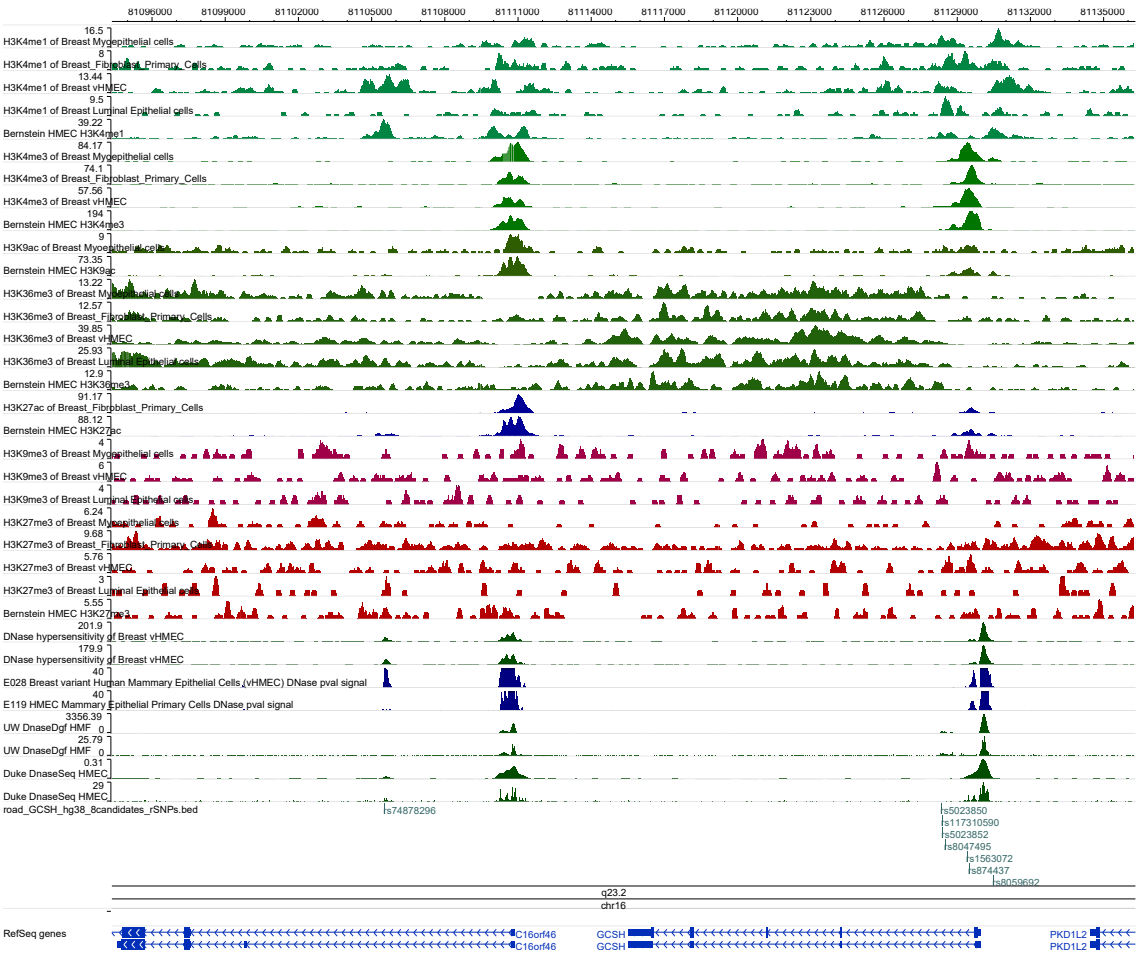


Figure S4.4. Genomic landscape of the 16q23.2 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs13329835) associated with breast cancer risk, and 14 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs12444974 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 112 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs12444974. From those, 38 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. This figure shows the 8 variants with strongest regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

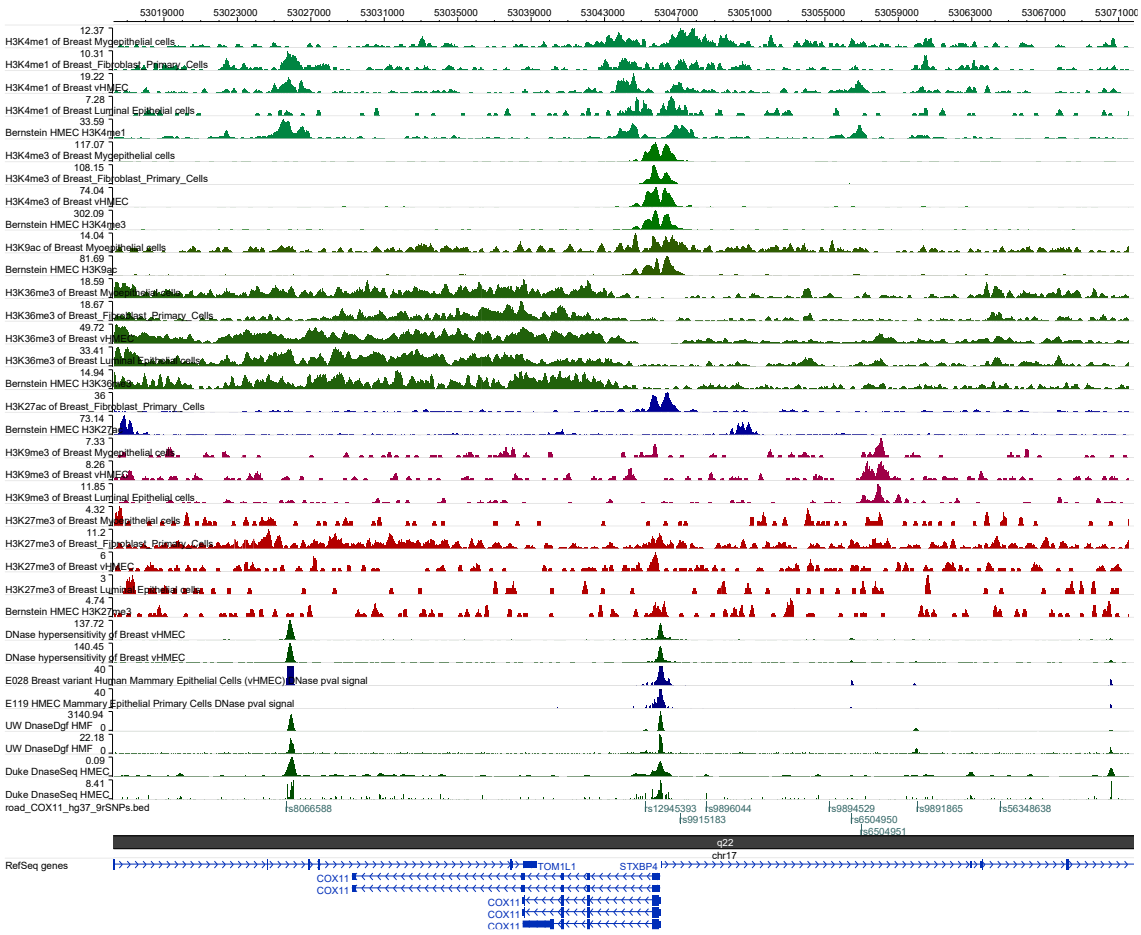


Figure S4.5. Genomic landscape of the 17q22 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs6504950) associated with breast cancer risk, and 22 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs17817901 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 106 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs17817901. From those, 66 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. This figure shows the 9 variants with strongest regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

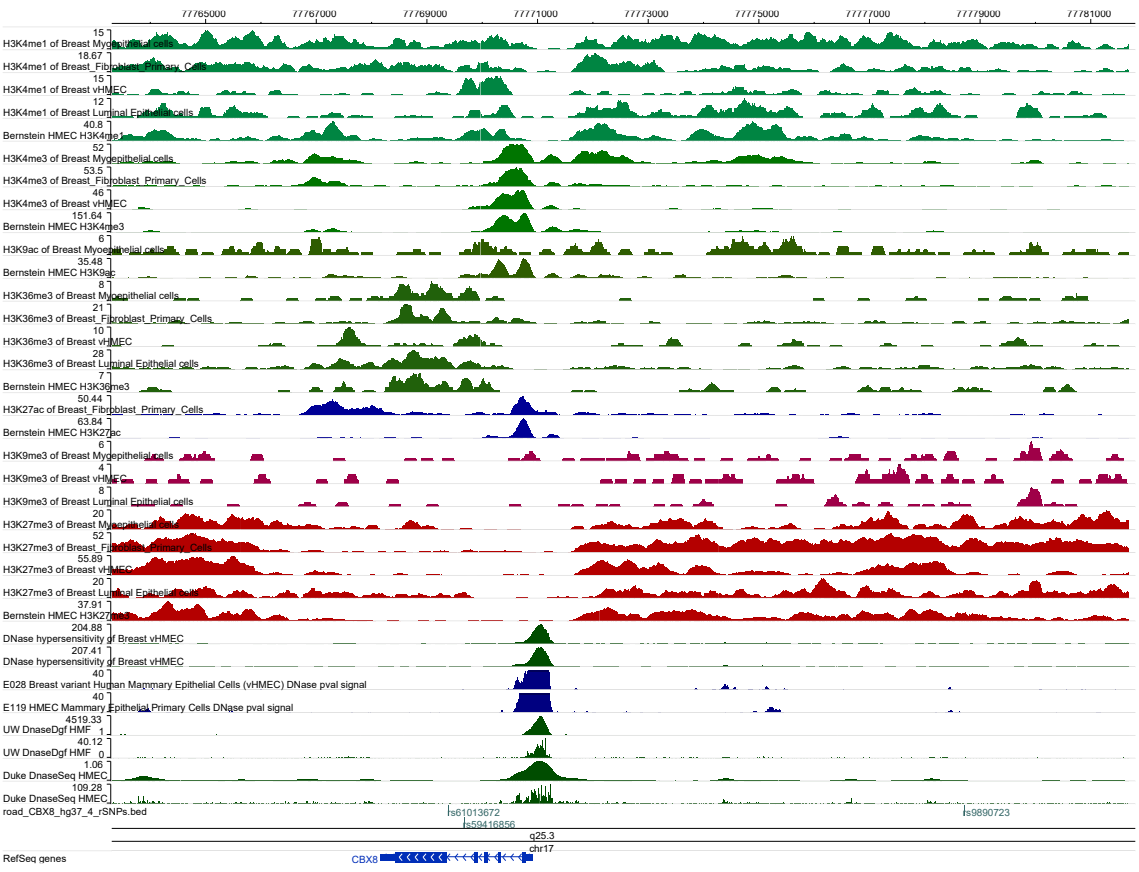


Figure S4.6. Genomic landscape of the 17q25.3 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs745570) associated with breast cancer risk, and 7 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs9897277 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 4 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs9897277. From those, 3 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. All those 3 SNPs had strong regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast, and are shown in the figure. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

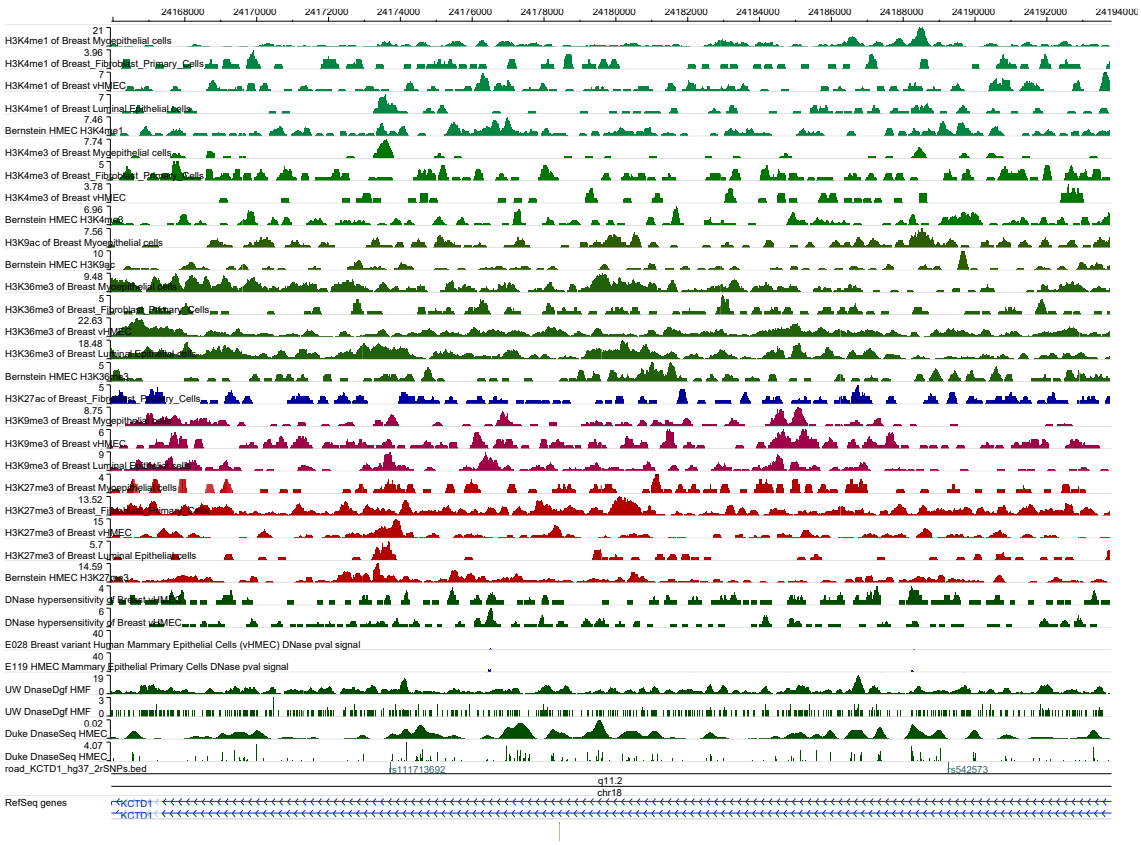


Figure S4.7. Genomic landscape of the 17q25.3 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs527616) associated with breast cancer risk, and 6 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs2438413 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 2 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs2438413. Those 2 SNPs had strong regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast, and are shown in the figure. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

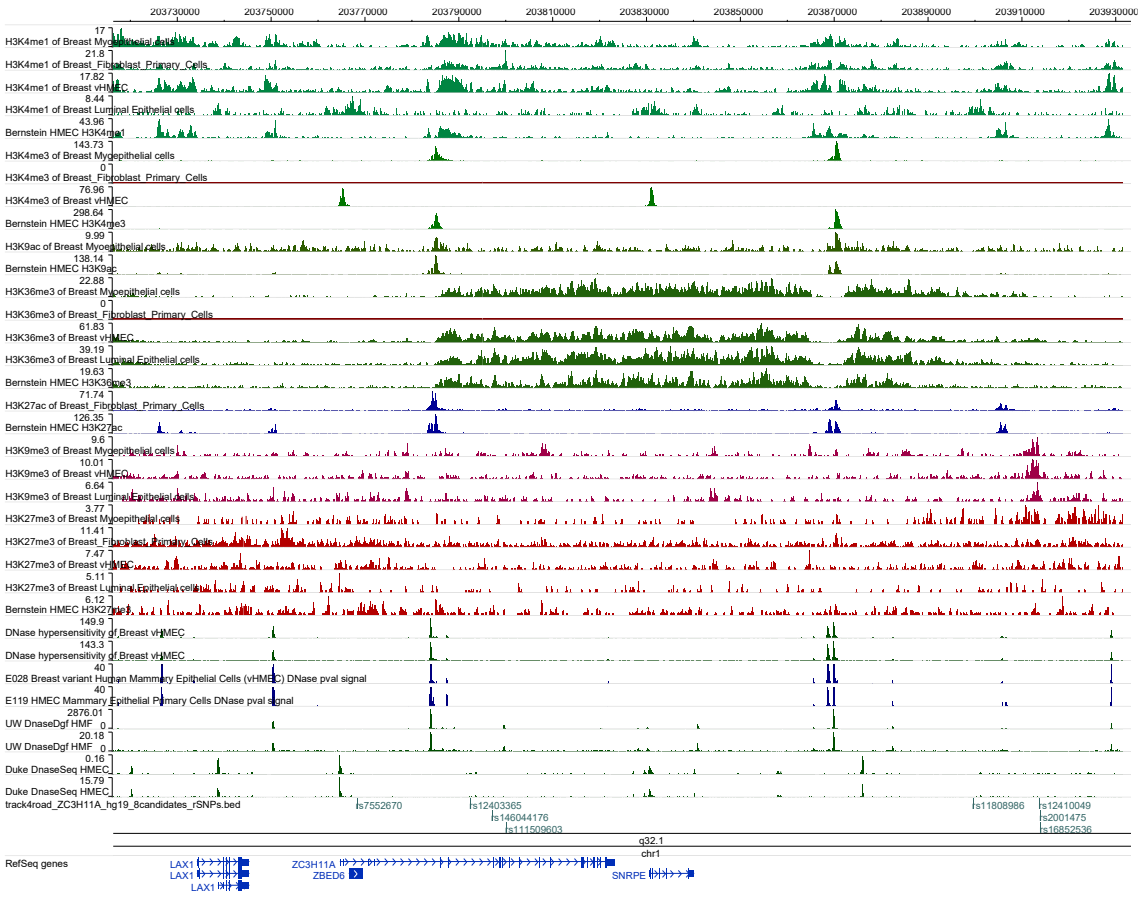


Figure S4.8. Genomic landscape of the 1q32.1 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 2 GWAS SNPs (rs4951011 and rs59867004) associated with breast cancer risk, and 14 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs7532505 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 60 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs7532505. From those, 23 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. This figure shows the 8 variants with strongest regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

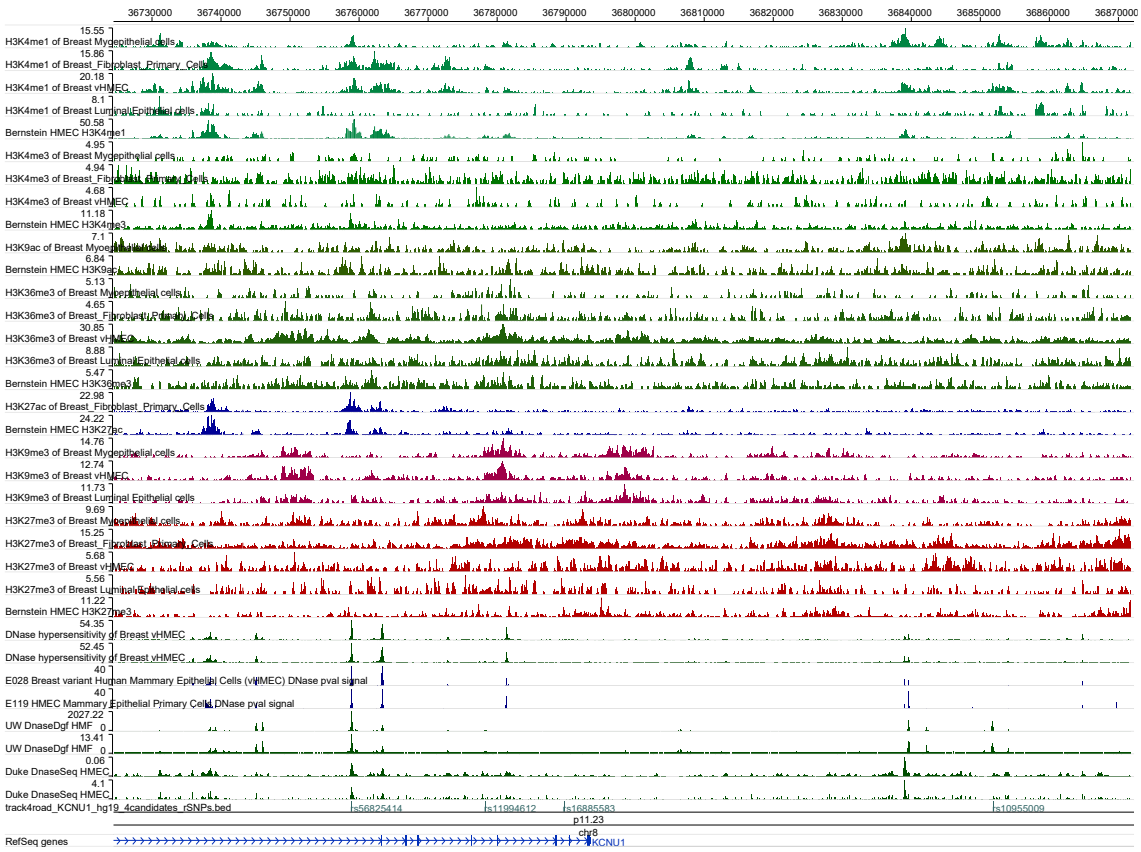


Figure S4.9. Genomic landscape of the 8q11.23 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 2 GWAS SNPs (rs13365225 and rs4286946) associated with breast cancer risk, and 2 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs10101721 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 43 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs10101721. From those, 17 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. This figure shows the 4 variants with strongest regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

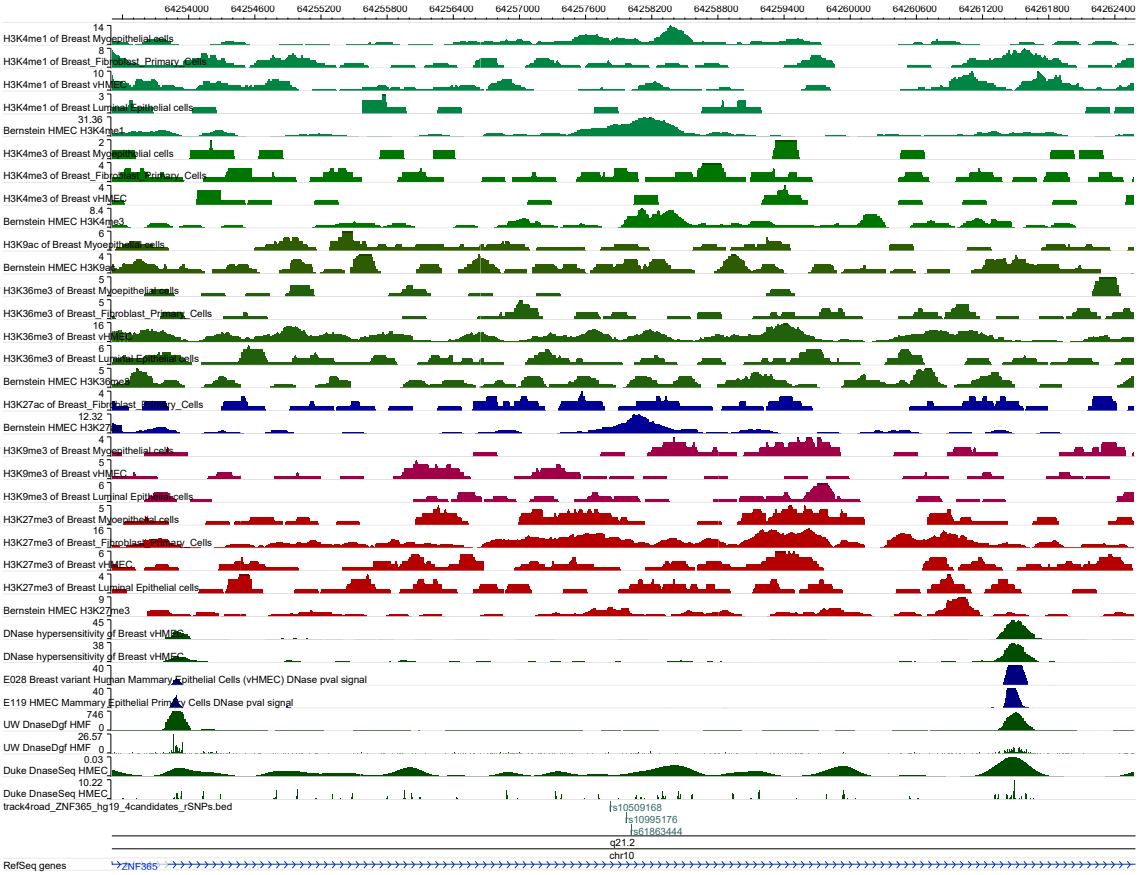


Figure S4.10. Genomic landscape of the 10q21.2 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs10822013) associated with breast cancer risk, and 1 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The only daeSNP in this locus, rs2393886, showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 4 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs2393886. From those, 3 variants had strong regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast, and are shown in the figure. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

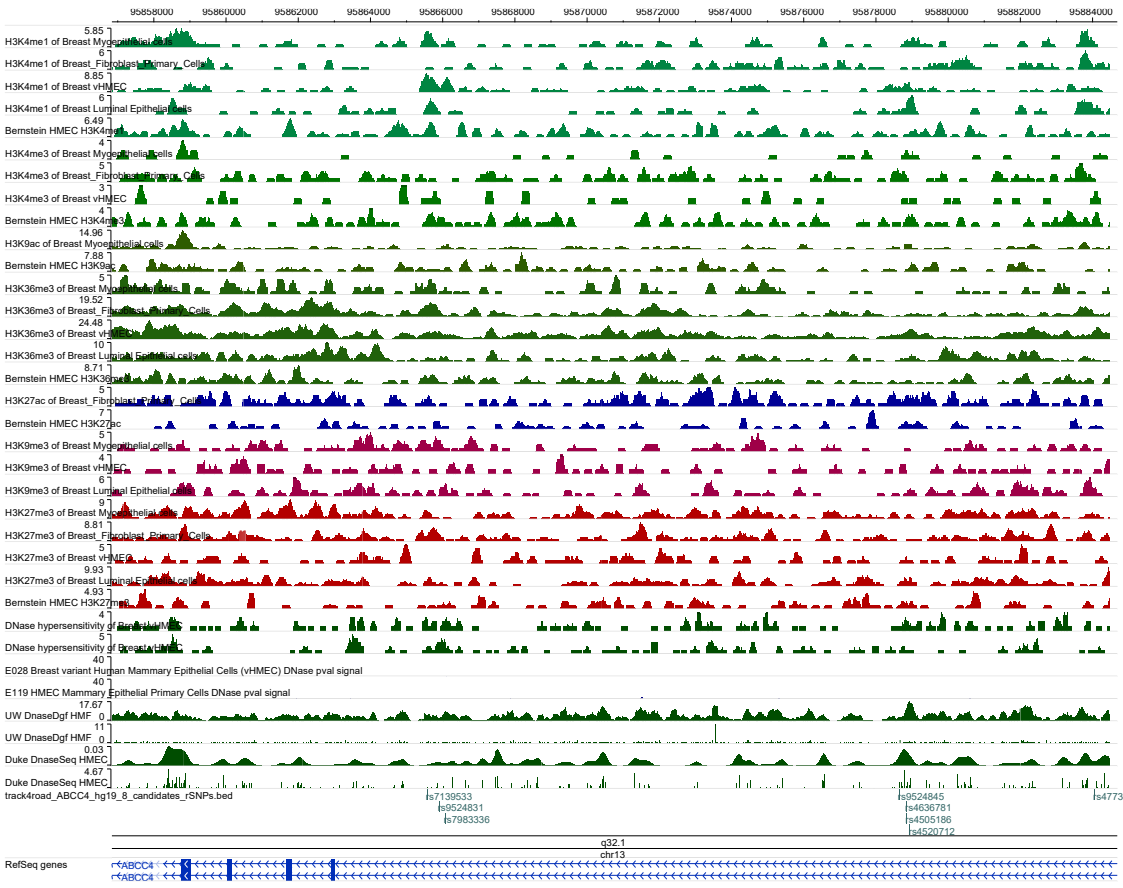


Figure S4.11. Genomic landscape of the 13q32.1 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs1926657) associated with breast cancer risk, and 1 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The only daeSNP in this locus, rs7988494, showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 83 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs7988494. From those, 8 variants had strong regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast, and are shown in the figure. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

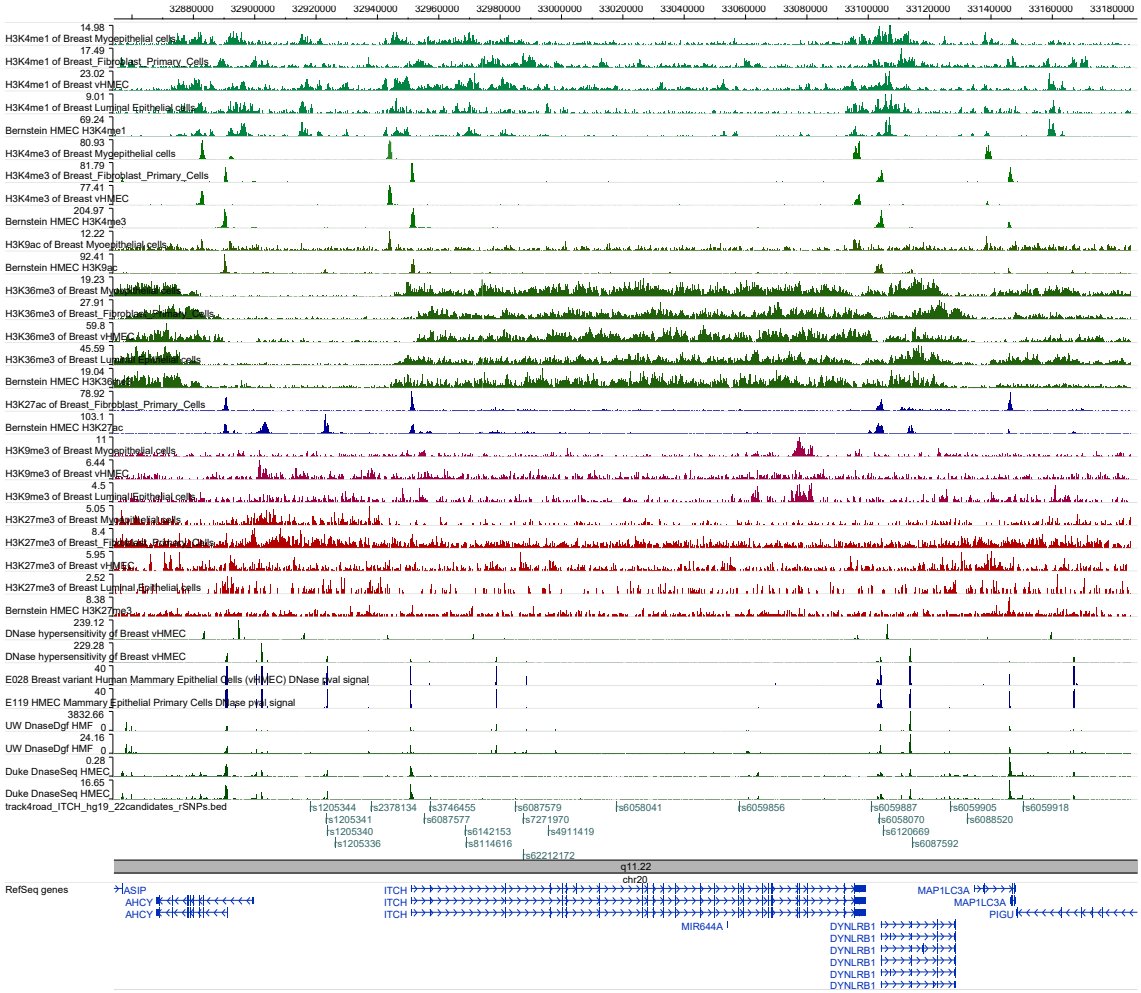


Figure S4.12. Genomic landscape of the 20q11.22 locus with detailed functional regulatory annotations in breast cell lines of the top candidate rSNPs. The RefSeq genes are showed at the bottom. In this locus there are 1 GWAS SNPs (rs2284378) associated with breast cancer risk, and 28 transcribed SNPs with DAE (daeSNPs) indicative of genes in this locus being under cis-regulation. The daeSNP rs6059860 showed a specific pattern indicative of strong LD with the rSNP(s), with all the analysed heterozygous individuals expressing preferentially the same allele (in healthy breast tissue). In-silico functional analysis of this locus identified 165 variants in strong LD ($r^2 \geq 0.8$) with the daeSNP rs6059860. From those, 88 variants had evidence of regulatory potential in normal breast cells and/or breast cancer cells. This figure shows the 22 variants with strongest regulatory potential (candidate rSNPs), meaning that this SNPs had evidence of being localized in regions of DNase I hypersensitive sites (DHS) and in cis-regulatory elements in breast. The image was obtained with the WashU EpiGenome Browser v40.6, and genomic positions refer to the human genome GRCh37 (hg19).

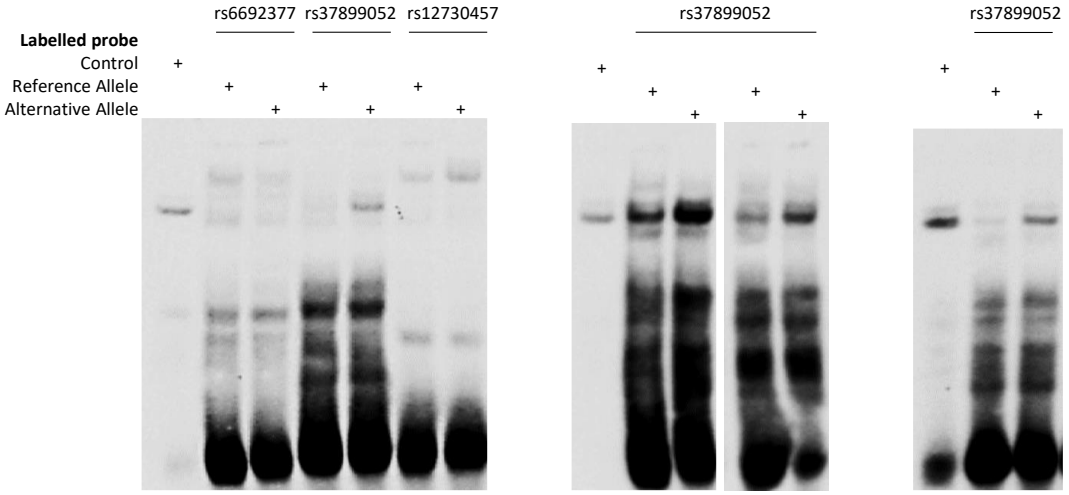


Figure S4.13. Complete gel from Figure 4.2. Analysis of DNA-protein binding of candidate rSNPs in 1q32.1.

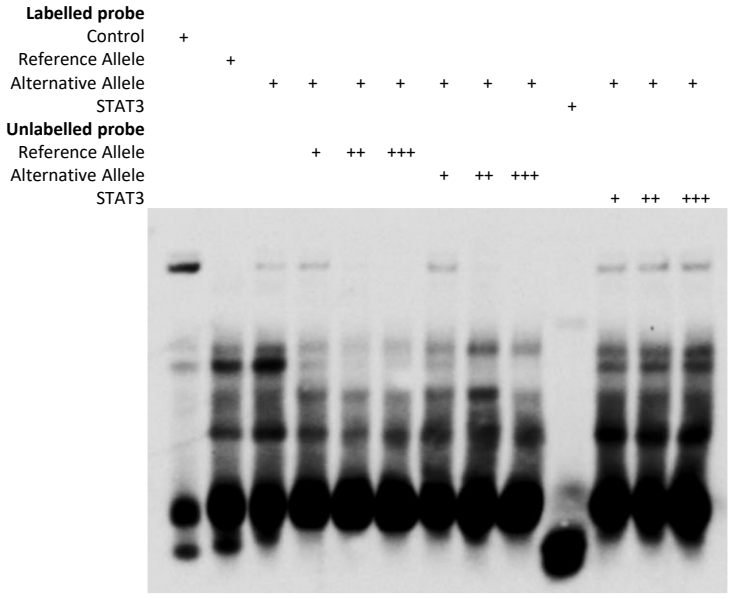


Figure S4.14. Complete gel from Figure 4.3. Analysis of DNA-protein binding of candidate rSNP rs3789052 (intronic in MDM4 – reference allele C, alternative allele T) in 1q32.1.

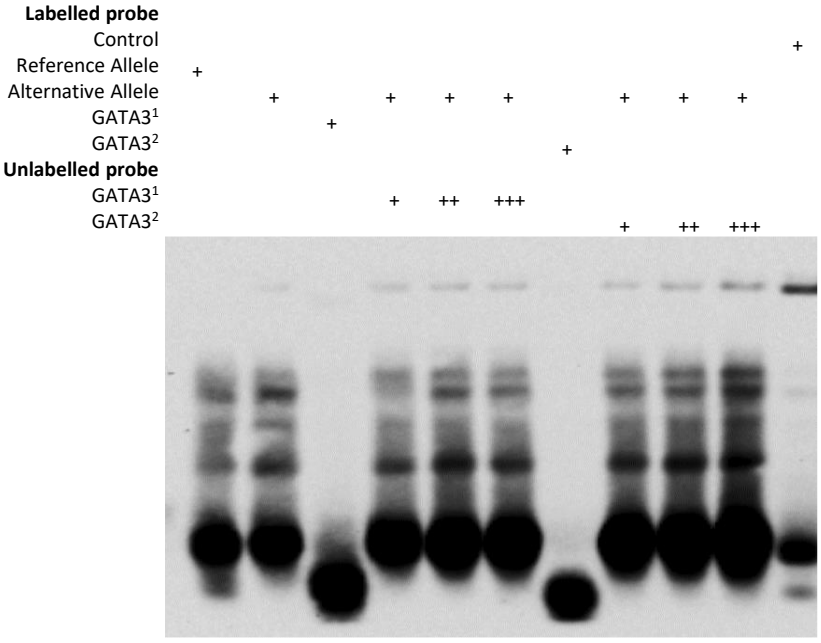


Figure S4.15. Complete gel from Figure 4.4. Analysis of DNA-protein binding of candidate rSNP rs3789052 (intronic in MDM4 – reference allele C, alternative allele T) in 1q32.1.

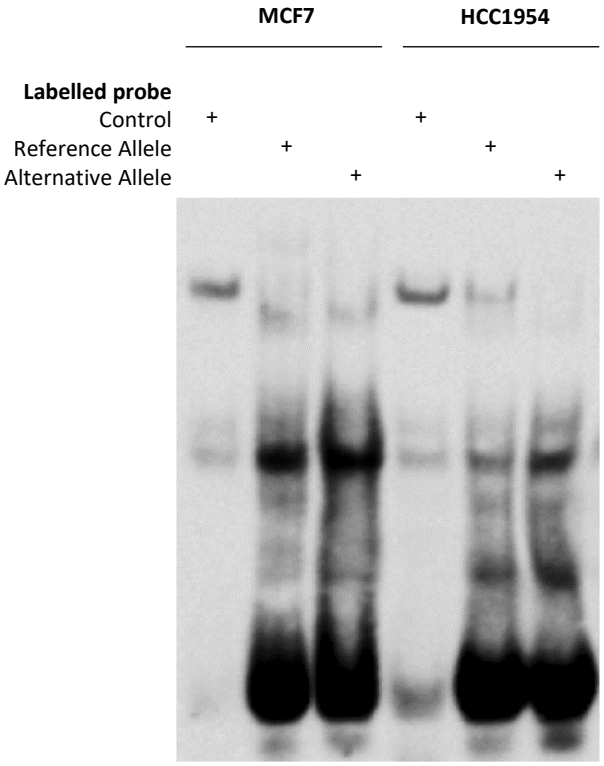


Figure S4.16. Complete gel from Figure 4.6. Analysis of DNA-protein binding of candidate rSNP rs74878296 (intronic in C16orf46 – reference allele C, alternative allele G) in 16q23.2.

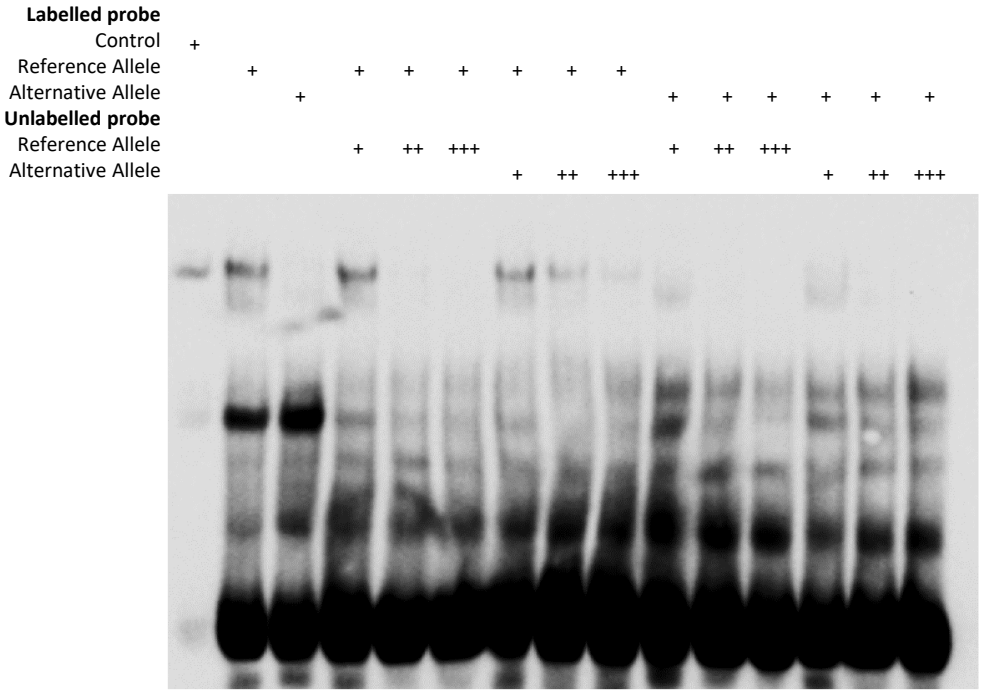


Figure S4.17. Complete gel from Figure 4.7. Analysis of DNA-protein binding of candidate rSNP rs74878296 (intronic in C16orf46 – reference allele C, alternative allele G) in 16q23.2.

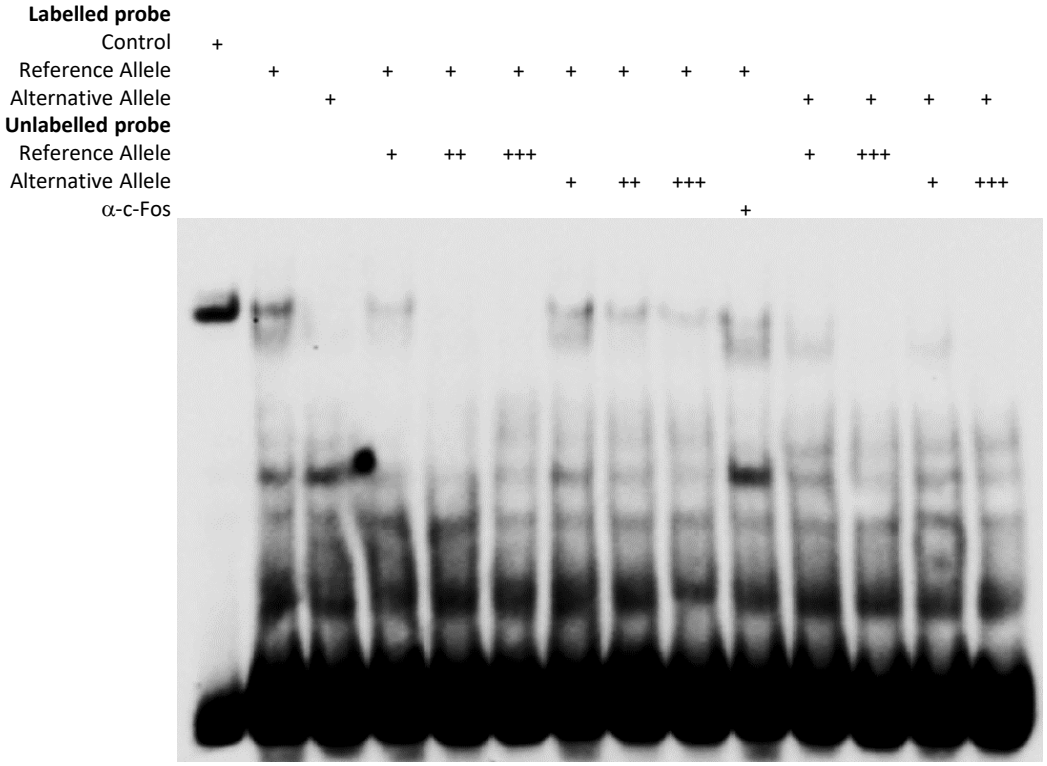


Figure S4.18. Complete gel from Figure 4.9. Analysis of DNA-protein binding of candidate rSNP rs74878296 (intronic in *C16orf46* – reference allele C, alternative allele G) in 16q23.2.

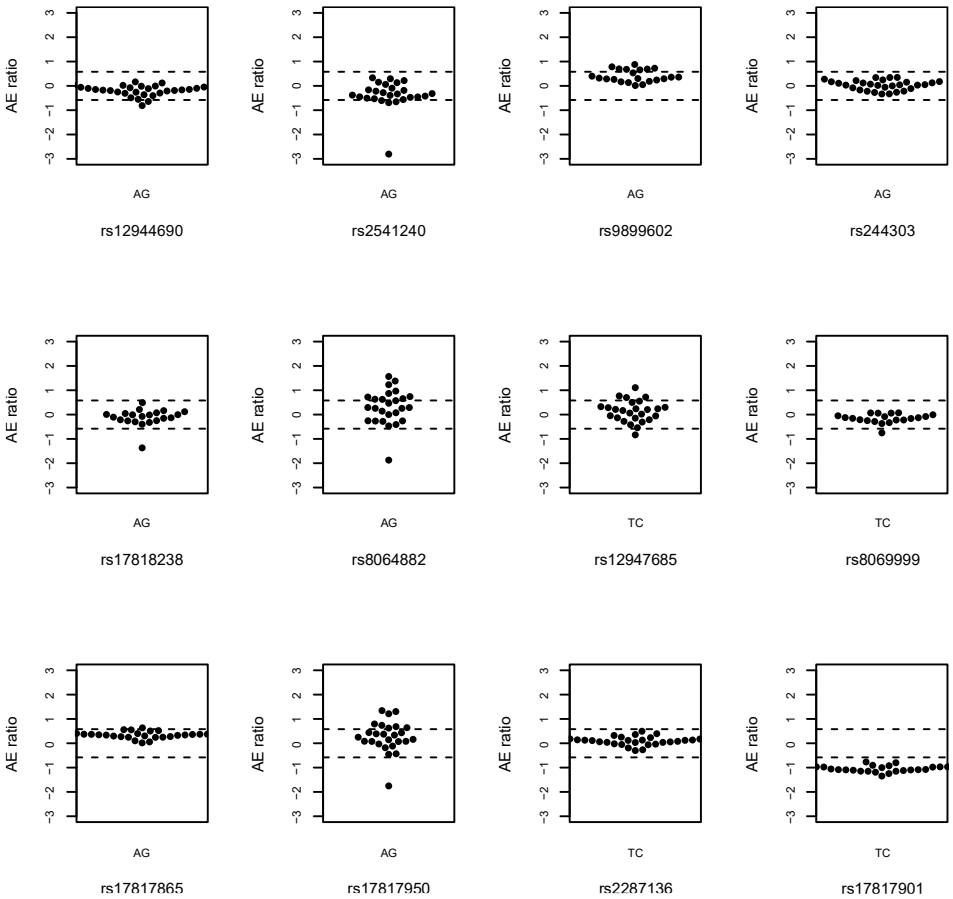

AE ratio
−3 −2 −1 0 1 2 3
AG
rs12944690
rs2541240
rs9899602
rs244303
rs17818238
rs8064882
TC
rs12947685
rs8069999
rs17817865
rs17817950
rs2287136
rs17817901
rs244301
TC
rs244298
TG
rs4794549
rs2628315
rs2628318
rs1420526
rs12936860
rs7643
rs9303360
rs12165058

Figure S4.19. Allelic expression analysis of 22 variants in the 17q22 risk locus. Plots of AE ratios (y-axis) calculated for heterozygous individuals (dots) for each variant indicated in the x-axis. The alleles for each variant are indicated below the graphs in the order they were used to calculate the AE ratios: e.g., AB, AE ratio ≈ log2 (allele A/ allele B). Dotted lines indicate 1.5-fold difference between alleles (absolute AE ratio = 0.58).

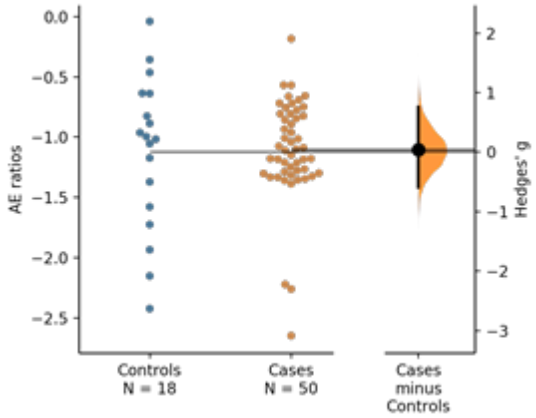


Figure S4.20. Case-control study using allelic expression ratios measured at rs9899602 in the 17q22 risk locus in breast tissue samples. Gardner-Altman estimation plot of Hedges' g between breast cancer cases and controls for allelic expression ratios calculated at rs9899602 (ratio calculated as allele T by allele C) in normal breast tissue. The heterozygous individuals for both groups are plotted on the left axes, with controls displayed in blue and cases in orange. The mean difference is plotted on the floating axes on the right as a bootstrap sampling distribution (bootstrap n=5000). The mean difference is depicted as a dot, and the 95% confidence interval is indicated by the ends of the vertical error bar.

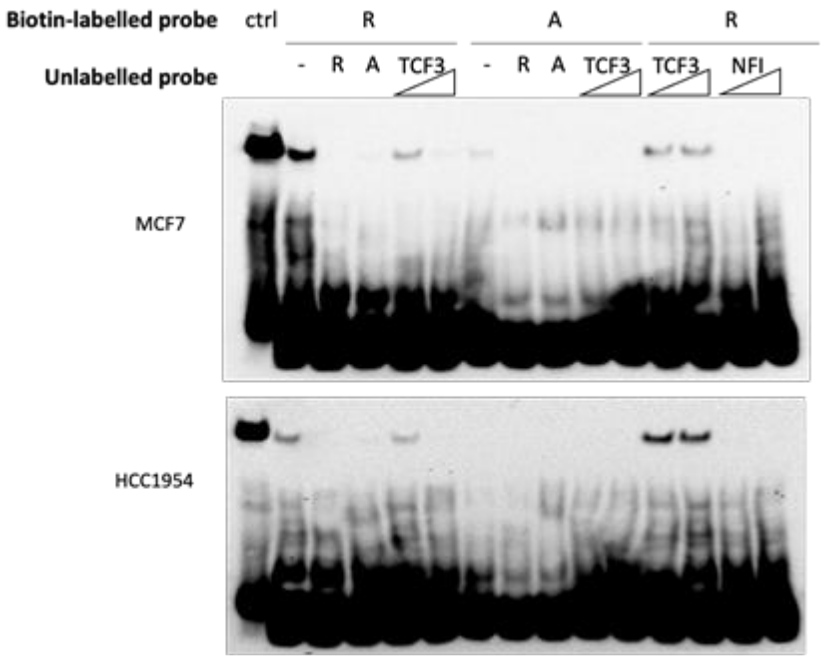


Figure S4.21. Complete gel images for EMSA experiments for rs8066588 using protein extracts from MCF-7 and HCC1954 cell lines show preferential binding of the reference C allele (R – reference, A – alternative alleles). Competition with an oligo of known binding site for TCF3 competes with observed binding, which does not occur with negative control oligo (NFI binding motif).

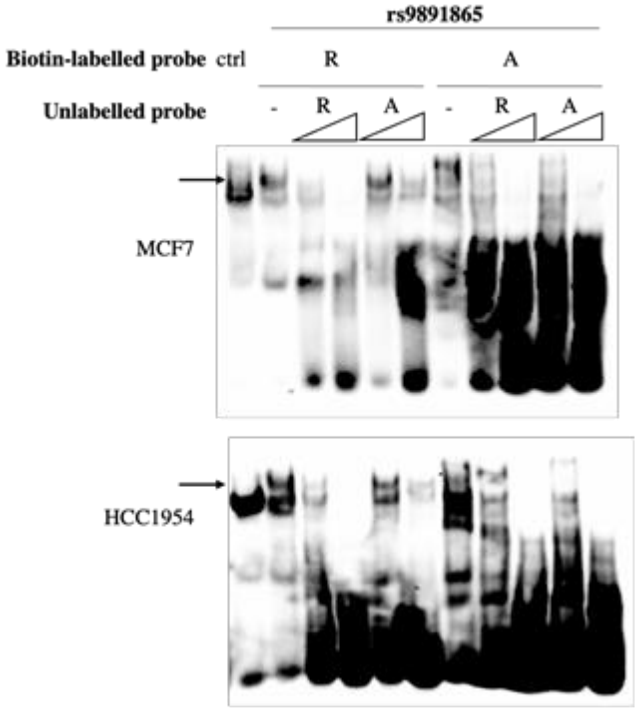


Figure S4.22. EMSA experiments for rs9891865 using protein extracts from MCF-7 and HCC1954 cell lines show differential allelic binding (R – reference, A – alternative alleles). Competition with unlabeled oligos of both alleles show that the high specificity of the binding of the reference allele. Side arrows indicate the position of the reference allele strongest specific band.

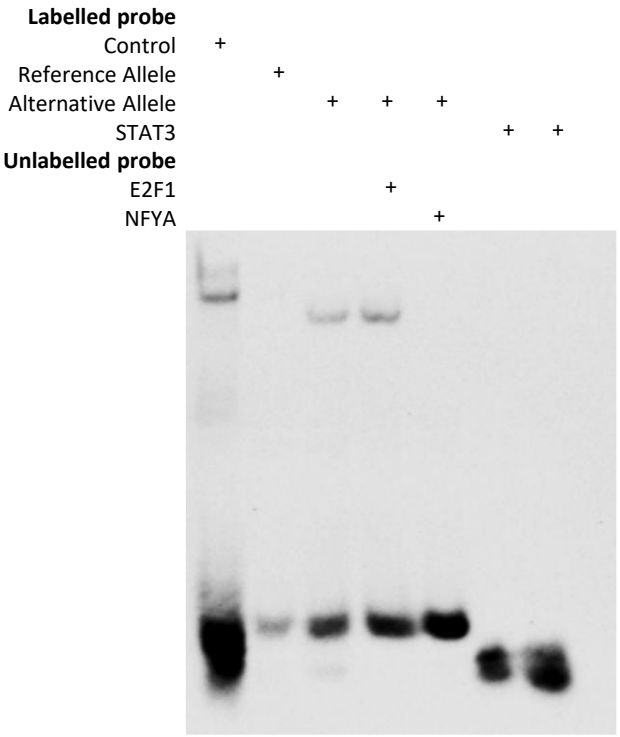


Figure S4.23. Complete gel from Figure 4.13. Analysis of DNA-protein binding of candidate rSNP rs2699887 (intronic in *PIK3CA* – reference allele C, alternative allele T) in 3q26.32.

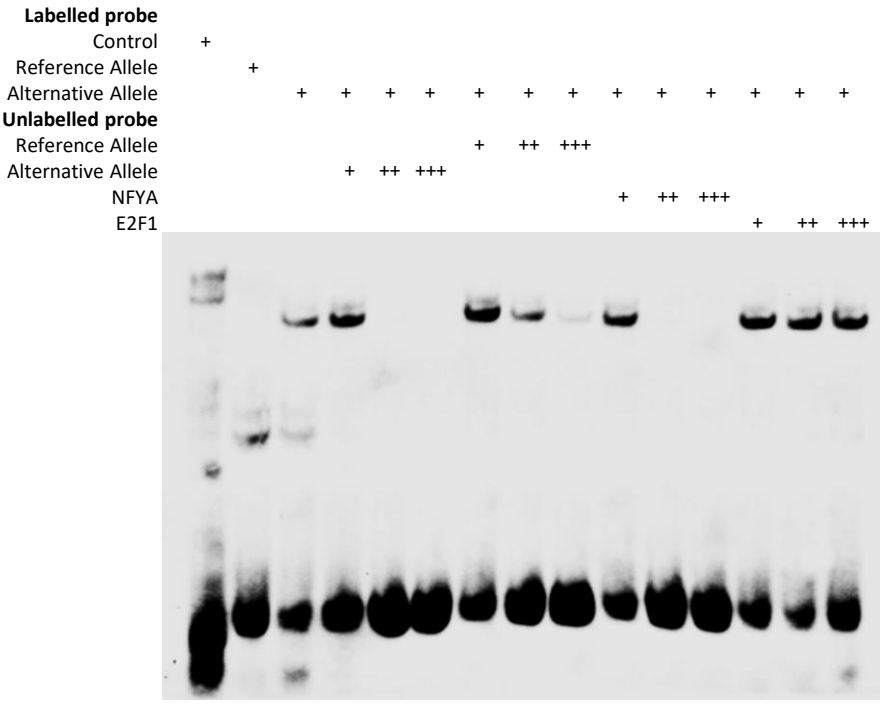


Figure S4.24. Complete gel from Figure 4.14. Analysis of DNA-protein binding of candidate rSNP rs2699887 (intronic in *PIK3CA* – reference allele C, alternative allele T) in 3q26.32.

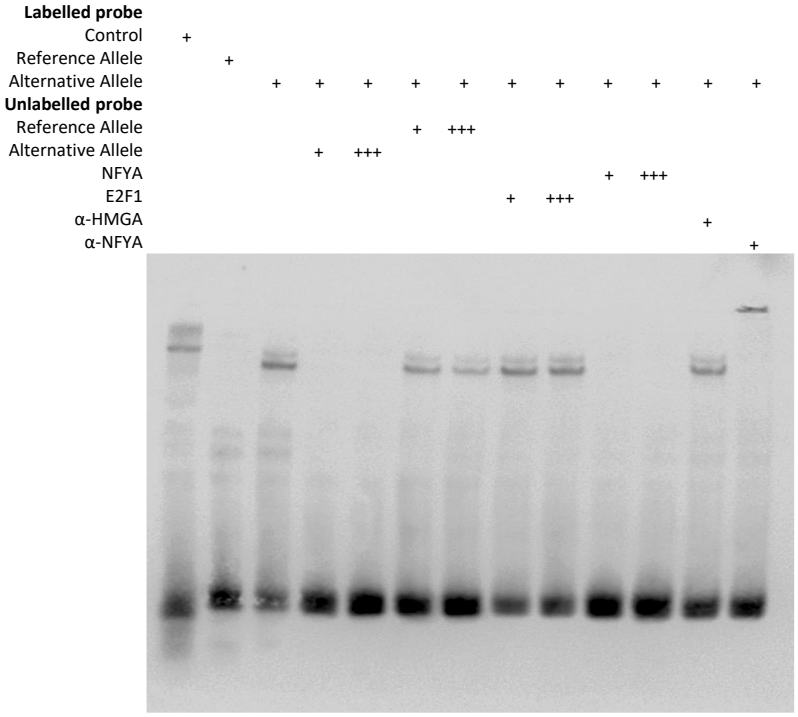


Figure S4.25. Complete gel from Figure 4.15. rs2699887 differentially binds transcription factor NF-YA in-vitro. EMSA analysis using protein extracts of breast cancer cell line HCC1954 and biotin-labeled oligonucleotides containing reference allele C or alternative allele T of rs2699887.

Chapter V. Pipeline Benchmarking for AE Analysis from RNA-Seq Data

Table S5.1. Alignment results considering the number of reads mapped with and without a mapped pair. **(A)** gold standard sample. **(B)** breast tissue sample. The highest and lowest numbers obtained when trimming was performed are highlighted.

(A)

	Initial reads	Mapped reads in pair		Singletons	
		GSNAP	STAR	GSNAP	STAR
Raw reads	51,867,848	65,345,768	62,737,684	1,202,774	1,603,878
AdapterRemoval v.2.1.7	45,994,060	62,154,836	55,470,513	686,878	793,869
Cutadapt v.1.18	**41,908,482**	**57,369,585**	**50,514,855**	**367,150**	**493,797**
fastq-mcf v.1.05	46,065,238	62,105,049	55,548,837	691,712	814,041
Flexbar v.3.0.3	45,367,712	61,204,019	54,692,778	664,150	788,409
SeqPurge v.2018_06	45,764,644	61,891,532	55,172,875	661,121	772,616
Skewer v.0.2.2	**46,228,506**	62,372,357	**55,757,390**	696,655	810,834
Trimmomatic v.0.36	46,040,848	**62,457,027**	55,756,788	**730,126**	**830,107**

(B)

	Initial reads	Mapped reads in pair		Singletons	
		GSNAP	STAR	GSNAP	STAR
Raw reads	135,803,746	170,587,292	150,915,342	883,505	1,196,605
AdapterRemoval v.2.1.7	132,856,772	168,787,678	148,051,437	129,114	420,031
Cutadapt v.1.18	**131,977,352**	**167,823,006**	**147,103,769**	**83,936**	**371,580**
fastq-mcf v.1.05	133,261,384	168,645,915	148,878,764	132,971	**518,646**
Flexbar v.3.0.3	133,045,104	168,474,404	148,633,635	129,821	502,404
SeqPurge v.2018_06	**133,265,682**	168,759,651	**148,884,160**	130,468	498,025
Skewer v.0.2.2	133,235,346	168,756,232	148,798,861	131,007	496,840
Trimmomatic v.0.36	132,820,490	**168,887,338**	148,052,881	**134,876**	436,661

Table S5.2. Alignment results considering the percentage of properly paired reads and the percentage of uniquely mapped reads. **(A)** gold standard sample. **(B)** breast tissue sample. The highest numbers obtained when trimming was performed are highlighted.

(A)

	Initial reads (M)	% Properly paired reads		% Uniquely mapped reads	
		GSNAP	STAR	GSNAP	STAR
Raw reads	51,868	88.9	96.8	74.7	83.5
AdapterRemoval v.2.1.7	45,994	94.6	98.2	79.4	86.7
Cutadapt v.1.18	41,908	**96.4**	98.8	**81.0**	**88.0**
fastq-mcf v.1.05	46,065	94.4	98.2	79.3	86.6
Flexbar v.3.0.3	45,368	94.6	98.2	79.5	86.7
SeqPurge v.2018_06	45,765	94.7	98.3	79.5	86.9
Skewer v.0.2.2	46,229	94.5	98.2	79.3	86.6
Trimmomatic v.0.36	46,041	**94.3**	98.2	**78.9**	**86.3**

(B)

	Initial reads (M)	% Properly paired reads		% Uniquely mapped reads	
		GSNAP	STAR	GSNAP	STAR
Raw reads	135,804	97.9	98.1	85.9	92.6
AdapterRemoval v.2.1.7	132,857	99.4	**99.3**	87.2	93.8
Cutadapt v.1.18	**131,977**	**99.5**	**99.3**	**87.3**	**93.9**
fastq-mcf v.1.05	133,261	99.1	99.1	86.9	93.6
Flexbar v.3.0.3	133,045	99.1	99.1	87.0	93.6
SeqPurge v.2018_06	**133,266**	99.1	99.1	87.0	93.6
Skewer v.0.2.2	133,235	99.2	99.1	87.0	93.6
Trimmomatic v.0.36	132,820	99.4	99.2	87.2	93.8

Table S5.3. Raw variants called through each pipeline. **(A)** gold standard sample. **(B)** breast tissue sample. The highest and lowest numbers of called variants when trimming was performed are highlighted.

(A)

	GATK		SAMtools		SAMtools after GATK data cleanup	
	GSNAP	STAR	GSNAP	STAR	GSNAP	STAR
Raw reads	129,766	93,001	344,991	76,657	263,871	160,834
AdapterRemoval v.2.1.7	125,423	89,627	327,070	72,937	253,098	157,423
Cutadapt v.1.18	**120,771**	**86,872**	**309,726**	**68,541**	**242,544**	**152,180**
fastq-mcf v.1.05	125,885	89,638	327,998	73,160	254,003	157,799
Flexbar v.3.0.3	124,161	88,937	323,602	72,409	250,232	156,813
SeqPurge v.2018_06	124,817	89,284	325,510	72,687	252,306	157,547
Skewer v.0.2.2	**125,989**	**89,994**	**329,061**	**73,235**	**256,024**	**158,381**
Trimmomatic v.0.36	124,297	88,181	327,389	71,005	252,967	155,325

(B)

	GATK		SAMtools		SAMtools after GATK data cleanup	
	GSNAP	STAR	GSNAP	STAR	GSNAP	STAR
Raw reads	262,546	222,942	206,385	106,513	167,083	116,966
AdapterRemoval v.2.1.7	255,625	216,575	189,142	103,422	161,809	113,230
Cutadapt v.1.18	**252,880**	**214,742**	**183,690**	**102,584**	157,670	112,090
fastq-mcf v.1.05	256,528	217,470	190,232	**103,860**	162,667	**113,751**
Flexbar v.3.0.3	256,621	217,653	188,060	103,498	**155,007**	**111,792**
SeqPurge v.2018_06	256,828	217,581	188,511	103,503	155,449	111,929
Skewer v.0.2.2	**256,991**	**217,666**	188,653	103,528	155,595	111,906
Trimmomatic v.0.36	255,591	215,875	**193,008**	103,031	**166,815**	113,560

Table S5.4. Variants passing filters called through each pipeline. **(A)** gold standard sample. **(B)** breast tissue sample. The highest and lowest numbers of called variants obtained when trimming was performed are highlighted.

(A)

	GATK		SAMtools		SAMtools after GATK data cleanup	
	GSNAP	STAR	GSNAP	STAR	GSNAP	STAR
Raw reads	100,576	82,458	48,047	22,550	44,395	30,325
AdapterRemoval v.2.1.7	98,913	80,133	42,555	20,887	39,757	**25,072**
Cutadapt v.1.18	**95,981**	**77,819**	**39,872**	**19,760**	**37,399**	25,114
fastq-mcf v.1.05	99,024	80,014	42,625	20,842	39,807	**27,092**
Flexbar v.3.0.3	98,116	79,393	42,330	20,724	39,429	26,748
SeqPurge v.2018_06	98,674	79,799	42,279	20,779	39,459	26,736
Skewer v.0.2.2	**99,286**	**80,264**	**42,793**	**20,944**	**39,881**	27,073
Trimmomatic v.0.36	98,115	78,980	41,421	20,394	38,773	26,095

(B)

	GATK		SAMtools		SAMtools after GATK data cleanup	
	GSNAP	STAR	GSNAP	STAR	GSNAP	STAR
Raw reads	164,247	148,773	99,934	65,074	26,435	18,213
AdapterRemoval v.2.1.7	162,082	145,327	97,859	64,328	**30,584**	22,095
Cutadapt v.1.18	**160,883**	**144,489**	96,886	63,884	30,167	21,845
fastq-mcf v.1.05	162,130	145,746	**97,968**	64,184	30,567	**22,139**
Flexbar v.3.0.3	162,959	146,133	97,820	64,368	**25,543**	**17,479**
SeqPurge v.2018_06	162,988	146,091	97,903	64,440	25,575	17,490
Skewer v.0.2.2	**163,036**	**146,120**	97,948	**64,470**	25,579	17,509
Trimmomatic v.0.36	161,648	145,035	**96,594**	**63,705**	30,509	21,896

Table S5.5. Number of SNPs called by GATK after mapping raw and trimmed data with GSNAP and STAR. **(A) g**old standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	Total	dbSNPs	Novel SNPs	Total	dbSNPs	Novel SNPs
Raw reads	92,853	73,807	19,046	77,094	64,249	12,845
AdapterRemoval v.2.1.7	91,297	72,887	18,410	75,041	62,909	12,132
Cutadapt v.1.18	88,773	71,192	17,581	72,917	61,357	11,560
fastq-mcf v.1.05	91,411	72,938	18,473	74,938	62,843	12,095
Flexbar v.3.0.3	90,575	72,341	18,234	74,345	62,370	11,975
SeqPurge v.2018_06	91,093	72,724	18,369	74,727	62,756	11,971
Skewer v.0.2.2	91,629	73,062	18,567	75,143	62,979	12,164
Trimmomatic v.0.36	90,705	72,307	18,398	74,027	62,123	11,904

(B)

	GSNAP			STAR		
	Total	dbSNPs	Novel SNPs	Total	dbSNPs	Novel SNPs
Raw reads	138,525	90,413	48,112	127,243	82,158	45,085
AdapterRemoval v.2.1.7	136,926	90,068	46,858	124,459	81,170	43,289
Cutadapt v.1.18	136,234	89,904	46,330	123,979	80,996	42,983
fastq-mcf v.1.05	136,981	90,073	46,908	124,883	81,336	43,547
Flexbar v.3.0.3	137,886	90,230	47,656	125,294	81,325	43,969
SeqPurge v.2018_06	137,901	90,274	47,627	125,257	81,342	43,915
Skewer v.0.2.2	137,916	90,240	47,676	125,272	81,332	43,940
Trimmomatic v.0.36	137,537	90,379	47,158	124,562	81,251	43,311

Table S5.6. Number of SNPs called by SAMtools after mapping raw and trimmed data with GSNAP and STAR. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	Total	dbSNPs	Novel SNPs	Total	dbSNPs	Novel SNPs
Raw reads	44,896	44,780	116	21,385	21,365	20
AdapterRemoval v.2.1.7	39,935	39,839	96	19,854	19,834	20
Cutadapt v.1.18	37,506	37,421	85	18,800	18,782	18
fastq-mcf v.1.05	40,009	39,914	95	19,807	19,786	21
Flexbar v.3.0.3	39,726	39,631	95	19,676	19,654	22
SeqPurge v.2018_06	39,712	39,619	93	19,743	19,721	22
Skewer v.0.2.2	40,165	40,070	95	19,899	19,877	22
Trimmomatic v.0.36	38,946	38,853	93	19,423	19,404	19

(B)

	GSNAP			STAR		
	Total	dbSNPs	Novel SNPs	Total	dbSNPs	Novel SNPs
Raw reads	98,302	97,697	605	62,718	62,394	324
AdapterRemoval v.2.1.7	96,319	95,766	553	62,025	61,729	296
Cutadapt v.1.18	95,407	94,873	534	61,624	61,340	284
fastq-mcf v.1.05	96,425	95,867	558	61,881	61,585	296
Flexbar v.3.0.3	96,288	95,729	559	62,059	61,761	298
SeqPurge v.2018_06	96,366	95,807	559	62,138	61,843	295
Skewer v.0.2.2	96,412	95,852	560	62,169	61,872	297
Trimmomatic v.0.36	95,269	94,731	538	61,440	61,158	282

Table S5.7. Number of SNPs shared by GATK and SAMtools after mapping raw and trimmed data with GSNAP and STAR. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	Total	dbSNPs	Novel SNPs	Total	dbSNPs	Novel SNPs
Raw reads	29,410	29,369	41	19,896	19,889	7
AdapterRemoval v.2.1.7	27,260	27,224	36	18,825	18,818	7
Cutadapt v.1.18	25,930	25,901	29	17,880	17,874	6
fastq-mcf v.1.05	27,281	27,246	35	18,788	18,781	7
Flexbar v.3.0.3	27,078	27,044	34	18,665	18,659	6
SeqPurge v.2018_06	27,133	27,098	35	18,748	18,740	8
Skewer v.0.2.2	27,364	27,328	36	18,879	18,872	7
Trimmomatic v.0.36	26,582	26,548	34	18,479	18,472	7

(B)

	GSNAP			STAR		
	Total	dbSNPs	Novel SNPs	Total	dbSNPs	Novel SNPs
Raw reads	62,439	62,234	205	50,344	50,219	125
AdapterRemoval v.2.1.7	61,834	61,637	197	50,039	49,916	123
Cutadapt v.1.18	61,517	61,328	189	49,783	49,664	119
fastq-mcf v.1.05	61,818	61,622	196	49,903	49,781	122
Flexbar v.3.0.3	61,834	61,639	195	50,069	49,946	123
SeqPurge v.2018_06	61,869	61,672	197	50,150	50,026	124
Skewer v.0.2.2	61,874	61,677	197	50,164	50,042	122
Trimmomatic v.0.36	61,598	61,403	195	49,621	49,501	120

Table S5.8. Number of SNPs called by SAMtools with GATK data cleanup after mapping raw and trimmed data with GSNAP and STAR. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	SNPs	dbSNPs	Novel SNPs	SNPs	dbSNPs	Novel SNPs
Raw reads	41,403	41,316	87	28,082	28,028	54
AdapterRemoval v.2.1.7	37,283	37,216	67	25,087	25,047	40
Cutadapt v.1.18	35,133	35,079	54	23,344	23,310	34
fastq-mcf v.1.05	37,331	37,265	66	25,122	25,082	40
Flexbar v.3.0.3	36,971	36,909	62	24,813	24,775	38
SeqPurge v.2018_06	37,010	36,946	64	24,797	24,761	36
Skewer v.0.2.2	37,389	37,324	65	25,103	25,063	40
Trimmomatic v.0.36	36,390	36,329	61	24,273	24,235	38

(B)

	GSNAP			STAR		
	SNPs	dbSNPs	Novel SNPs	SNPs	dbSNPs	Novel SNPs
Raw reads	26,050	25,976	74	17,170	17,110	60
AdapterRemoval v.2.1.7	30,114	30,026	88	20,764	20,674	90
Cutadapt v.1.18	29,696	29,614	82	20,554	20,472	82
fastq-mcf v.1.05	30,094	30,009	85	20,807	20,717	90
Flexbar v.3.0.3	25,176	25,110	66	16,451	16,394	57
SeqPurge v.2018_06	25,212	25,142	70	16,464	16,406	58
Skewer v.0.2.2	25,210	25,141	69	16,478	16,421	57
Trimmomatic v.0.36	30,063	29,966	97	20,582	20,501	81

Table S5.9. Number of SNPs called by both GATK and SAMtools with GATK data cleanup after mapping raw and trimmed data with GSNAP and STAR. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	SNPs	dbSNPs	Novel SNPs	SNPs	dbSNPs	Novel SNPs
Raw reads	28,512	28,475	37	24,311	24,297	14
AdapterRemoval v.2.1.7	26,569	26,536	33	22,277	22,264	13
Cutadapt v.1.18	25,279	25,251	28	20,953	20,944	9
fastq-mcf v.1.05	26,565	26,532	33	22,293	22,280	13
Flexbar v.3.0.3	26,315	26,284	31	22,043	22,031	12
SeqPurge v.2018_06	26,365	26,333	32	22,071	22,059	12
Skewer v.0.2.2	26,577	26,544	33	22,283	22,271	12
Trimmomatic v.0.36	25,876	25,846	30	21,673	21,663	10

(B)

	GSNAP			STAR		
	SNPs	dbSNPs	Novel SNPs	SNPs	dbSNPs	Novel SNPs
Raw reads	13,959	13,947	12	11,788	11,782	6
AdapterRemoval v.2.1.7	16,746	16,733	13	14,252	14,241	11
Cutadapt v.1.18	16,629	16,616	13	14,175	14,164	11
fastq-mcf v.1.05	16,728	16,714	14	14,276	14,265	11
Flexbar v.3.0.3	13,763	13,756	7	11,415	11,411	4
SeqPurge v.2018_06	13,765	13,757	8	11,416	11,412	4
Skewer v.0.2.2	13,768	13,760	8	11,423	11,419	4
Trimmomatic v.0.36	16,690	16,677	13	14,186	14,177	9

Table S5.10. Ti/Tv ratios of known SNPs called by GATK and SAMtools after GATK data cleanup. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	GATK	SAMtools with data cleanup	Overlap	GATK	SAMtools with data cleanup	Overlap
Raw reads	2.50	2.64	2.64	2.45	2.54	2.61
AdapterRemoval v.2.1.7	2.50	2.63	2.64	2.45	2.55	2.61
Cutadapt v.1.18	2.51	2.62	2.65	2.45	2.58	2.63
fastq-mcf v.1.05	2.50	2.62	2.63	2.45	2.56	2.61
Flexbar v.3.0.3	2.51	2.63	2.64	2.45	2.56	2.62
SeqPurge v.2018_06	2.50	2.62	2.64	2.45	2.57	2.61
Skewer v.0.2.2	2.50	2.63	2.64	2.45	2.56	2.61
Trimmomatic v.0.36	2.50	2.61	2.63	2.45	2.57	2.62

(B)

	GSNAP			STAR		
	GATK	SAMtools with data cleanup	Overlap	GATK	SAMtools with data cleanup	Overlap
Raw reads	2.43	2.58	2.65	2.38	2.51	2.64
AdapterRemoval v.2.1.7	2.43	2.55	2.66	2.38	2.52	2.62
Cutadapt v.1.18	2.43	2.55	2.67	2.39	2.52	2.63
fastq-mcf v.1.05	2.43	2.56	2.66	2.38	2.51	2.62
Flexbar v.3.0.3	2.43	2.55	2.67	2.38	2.52	2.64
SeqPurge v.2018_06	2.43	2.56	2.67	2.38	2.53	2.63
Skewer v.0.2.2	2.43	2.56	2.67	2.38	2.53	2.63
Trimmomatic v.0.36	2.43	2.57	2.67	2.38	2.53	2.63

Table S5.11. Ti/Tv ratios of novel SNPs called by GATK and SAMtools after GATK data cleanup. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	GATK	SAMtools with data cleanup	Overlap	GATK	SAMtools with data cleanup	Overlap
Raw reads	4.95	2.78	2.08	6.79	3.91	0.00
AdapterRemoval v.2.1.7	4.96	4.15	2.67	7.28	5.67	0.00
Cutadapt v.1.18	5.01	4.40	2.50	7.51	7.50	0.00
fastq-mcf v.1.05	4.97	4.08	2.67	7.29	5.67	0.00
Flexbar v.3.0.3	4.98	4.17	2.44	7.40	5.33	0.00
SeqPurge v.2018_06	4.97	4.33	3.00	7.41	8.00	0.00
Skewer v.0.2.2	4.94	4.00	2.67	7.27	5.67	0.00
Trimmomatic v.0.36	4.93	4.08	2.33	7.32	6.60	0.00

(B)

	GSNAP			STAR		
	GATK	SAMtools with data cleanup	Overlap	GATK	SAMtools with data cleanup	Overlap
Raw reads	4.83	1.96	2.00	5.01	2.33	2.00
AdapterRemoval v.2.1.7	5.03	2.38	1.60	5.37	2.33	4.50
Cutadapt v.1.18	5.18	2.73	2.25	5.50	2.42	2.67
fastq-mcf v.1.05	5.01	2.27	1.33	5.33	2.21	4.50
Flexbar v.3.0.3	4.80	2.00	0.75	5.10	2.56	3.00
SeqPurge v.2018_06	4.82	1.92	1.00	5.12	2.87	3.00
Skewer v.0.2.2	4.81	1.88	1.00	5.11	2.56	3.00
Trimmomatic v.0.36	5.03	2.46	1.60	5.43	2.52	3.50

Table S5.12. Number of indels called by GATK. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	7,716	6,357	1,358	5,358	4,886	472
AdapterRemoval v.2.1.7	7,607	6,256	1,349	5,086	4,643	443
Cutadapt v.1.18	7,202	5,936	1,264	4,897	4,455	442
fastq-mcf v.1.05	7,604	6,260	1,342	5,070	4,628	442
Flexbar v.3.0.3	7,531	6,209	1,320	5,042	4,608	434
SeqPurge v.2018_06	7,571	6,222	1,347	5,066	4,619	447
Skewer v.0.2.2	7,648	6,288	1,358	5,115	4,662	453
Trimmomatic v.0.36	7,400	6,072	1,326	4,947	4,503	444

(B)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	26,141	16,762	9,351	21,794	14,281	7,495
AdapterRemoval v.2.1.7	25,532	16,275	9,231	21,114	13,700	7,398
Cutadapt v.1.18	25,000	15,847	9,128	20,735	13,379	7,340
fastq-mcf v.1.05	25,524	16,266	9,232	21,109	13,697	7,396
Flexbar v.3.0.3	25,447	16,216	9,206	21,075	13,652	7,406
SeqPurge v.2018_06	25,456	16,221	9,210	21,071	13,654	7,400
Skewer v.0.2.2	25,494	16,246	9,223	21,083	13,661	7,405
Trimmomatic v.0.36	24,438	15,509	8,905	20,697	13,343	7,335

Table S5.13. Number of indels called by SAMtools. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	3,116	2,449	306	1,161	1,065	79
AdapterRemoval v.2.1.7	2,595	2,044	258	1,029	946	68
Cutadapt v.1.18	2,340	1,855	228	956	883	60
fastq-mcf v.1.05	2,592	2,041	261	1,031	949	68
Flexbar v.3.0.3	2,579	2,034	256	1,044	965	64
SeqPurge v.2018_06	2,538	2,020	245	1,032	957	62
Skewer v.0.2.2	2,604	2,051	263	1,041	961	66
Trimmomatic v.0.36	2,451	1,945	245	967	896	59

(B)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	10,040	8,464	551	5,305	3,933	1,107
AdapterRemoval v.2.1.7	9,136	7,747	508	5,037	3,708	1,073
Cutadapt v.1.18	8,813	7,501	468	4,866	3,568	1,058
fastq-mcf v.1.05	9,133	7,743	509	5,036	3,707	1,073
Flexbar v.3.0.3	9,123	7,738	505	5,036	3,706	1,074
SeqPurge v.2018_06	9,116	7,731	503	5,029	3,701	1,072
Skewer v.0.2.2	9,142	7,753	505	5,041	3,717	1,069
Trimmomatic v.0.36	8,335	7,150	401	4,832	3,513	1,077

Table S5.14. Number of indels called by SAMtools after GATK data cleanup. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	2,972	2,570	141	2,238	1,989	200
AdapterRemoval v.2.1.7	2,458	2,150	110	1,960	1,753	174
Cutadapt v.1.18	2,257	1,992	99	1,765	1,593	147
fastq-mcf v.1.05	2,461	2,154	117	1,964	1,755	177
Flexbar v.3.0.3	2,442	2,140	108	1,930	1,726	173
SeqPurge v.2018_06	2,434	2,128	112	1,933	1,734	170
Skewer v.0.2.2	2,475	2,164	114	1,965	1,757	176
Trimmomatic v.0.36	2,368	2,078	106	1,817	1,629	157

(B)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	2,118	1,878	88	2,209	1,616	527
AdapterRemoval v.2.1.7	2,511	2,244	101	2,790	2,021	669
Cutadapt v.1.18	2,458	2,201	97	2,715	1,967	654
fastq-mcf v.1.05	2,511	2,244	100	2,796	2,028	669
Flexbar v.3.0.3	1,960	1,755	75	2,129	1,558	510
SeqPurge v.2018_06	1,956	1,752	73	2,126	1,558	507
Skewer v.0.2.2	1,961	1,756	75	2,135	1,563	511
Trimmomatic v.0.36	2,404	2,165	94	2,717	1,965	667

Table S5.15. Number of indels shared by GATK and SAMtools. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	1,913	1,852	61	877	874	3
AdapterRemoval v.2.1.7	1,714	1,659	55	798	795	3
Cutadapt v.1.18	1,604	1,550	54	754	752	2
fastq-mcf v.1.05	1,713	1,658	55	799	796	3
Flexbar v.3.0.3	1,697	1,643	54	811	808	3
SeqPurge v.2018_06	1,703	1,650	53	803	801	2
Skewer v.0.2.2	1,716	1,662	54	808	805	3
Trimmomatic v.0.36	1,651	1,594	57	757	755	2

(B)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	6,195	5,751	441	2,771	2,645	126
AdapterRemoval v.2.1.7	5,958	5,526	430	2,621	2,500	121
Cutadapt v.1.18	5,819	5,393	425	2,542	2,426	116
fastq-mcf v.1.05	5,952	5,520	430	2,619	2,498	121
Flexbar v.3.0.3	5,952	5,521	429	2,616	2,495	121
SeqPurge v.2018_06	5,954	5,525	427	2,615	2,492	123
Skewer v.0.2.2	5,958	5,525	431	2,625	2,503	122
Trimmomatic v.0.36	5,609	5,206	401	2,505	2,394	110

Table S5.16. Number of shared indels by both GATK and SAMtools after GATK data cleanup. **(A)** gold standard sample. **(B)** breast tissue sample.

(A)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	2,028	1,970	58	1,547	1,538	9
AdapterRemoval v.2.1.7	1,813	1,759	54	1,398	1,389	9
Cutadapt v.1.18	1,710	1,656	54	1,286	1,278	8
fastq-mcf v.1.05	1,813	1,758	55	1,401	1,391	10
Flexbar v.3.0.3	1,797	1,744	53	1,380	1,371	9
SeqPurge v.2018_06	1,801	1,745	56	1,387	1,379	8
Skewer v.0.2.2	1,820	1,766	54	1,400	1,391	9
Trimmomatic v.0.36	1,757	1,700	57	1,323	1,313	10

(B)

	GSNAP			STAR		
	Total	dbIndels	Novel Indels	Total	dbIndels	Novel Indels
Raw reads	1,239	1,189	49	901	886	15
AdapterRemoval v.2.1.7	1,501	1,447	53	1,138	1,114	24
Cutadapt v.1.18	1,491	1,435	56	1,110	1,086	24
fastq-mcf v.1.05	1,504	1,450	53	1,134	1,111	23
Flexbar v.3.0.3	1,203	1,155	48	853	841	12
SeqPurge v.2018_06	1,198	1,154	44	849	837	12
Skewer v.0.2.2	1,203	1,157	46	851	839	12
Trimmomatic v.0.36	1,455	1,411	43	1,093	1,071	22

Table S5.17. Precision and Sensitivity of variants called with SAMTools with GATK data cleanup. TP – true positive, FP – false positive, FN – false negative, precision=TP/(TP+FP), sensitivity=TP/(TP+FN), F-measure=(2TP)/(FN+FP+2TP). F-measure is the harmonic mean of precision and sensitivity with a value between 0, for the worst, and 1, for the perfect F-measure.

	GSNAP						STAR					
	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure
Raw reads	14,680	1,790	48,011	89,1	23,4	0,4	14,713	885	47,979	94,3	23,5	0,4
AdapterRemoval v.2.1.7	14,078	1,535	48,614	90,2	22,5	0,4	14,101	731	48,592	95,1	22,5	0,4
Cutadapt v.1.18	13,657	1,424	49,033	90,6	21,8	0,4	13,639	639	49,053	95,5	21,8	0,4
fastq-mcf v.1.05	14,077	1,539	48,615	90,1	22,5	0,4	14,106	735	48,587	95,0	22,5	0,4
Flexbar v.3.0.3	13,984	1,522	48,708	90,2	22,3	0,4	14,009	710	48,683	95,2	22,3	0,4
SeqPurge v.2018_06	14,010	1,532	48,682	90,1	22,3	0,4	14,032	708	48,662	95,2	22,4	0,4
Skewer v.0.2.2	14,071	1,541	48,621	90,1	22,4	0,4	14,098	729	48,525	95,1	22,6	0,4
Trimmomatic v.0.36	13,882	1,481	48,810	90,4	22,1	0,4	13,878	669	48,814	95,4	22,1	0,4

Table S5.18. Precision and Sensitivity of common variants from GATK and SAMTools with GATK data cleanup. TP – true positive, FP – false positive, FN – false negative, precision=TP/(TP+FP), sensitivity=TP/(TP+FN), F-measure=(2TP)/(FN+FP+2TP). F-measure is the harmonic mean of precision and sensitivity with a value between 0, for the worst, and 1, for the perfect F-measure.

	GSNAP						STAR					
	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure	TP	FP	FN	Precision (%)	Sensitivity (%)	F-measure
Raw reads	14,276	578	48,413	96,1	22,8	0,4	14,058	262	48,631	98,2	22,4	0,4
AdapterRemoval v.2.1.7	13,748	515	48,941	96,4	21,9	0,4	13,517	236	49,172	98,3	21,6	0,4
Cutadapt v.1.18	13,339	489	49,349	96,5	21,3	0,3	13,093	203	49,596	98,5	20,9	0,3
fastq-mcf v.1.05	13,745	521	48,944	96,3	21,9	0,4	13,516	237	49,173	98,3	21,6	0,4
Flexbar v.3.0.3	13,656	514	49,033	96,4	21,8	0,4	13,420	226	49,269	98,3	21,4	0,4
SeqPurge v.2018_06	13,681	522	49,008	96,3	21,8	0,4	13,453	228	49,236	98,3	21,5	0,4
Skewer v.0.2.2	13,735	526	48,954	96,3	21,9	0,4	13,507	236	49,182	98,3	21,5	0,4
Trimmomatic v.0.36	13,561	513	49,128	96,4	21,6	0,4	13,320	215	49,369	98,4	21,2	0,3

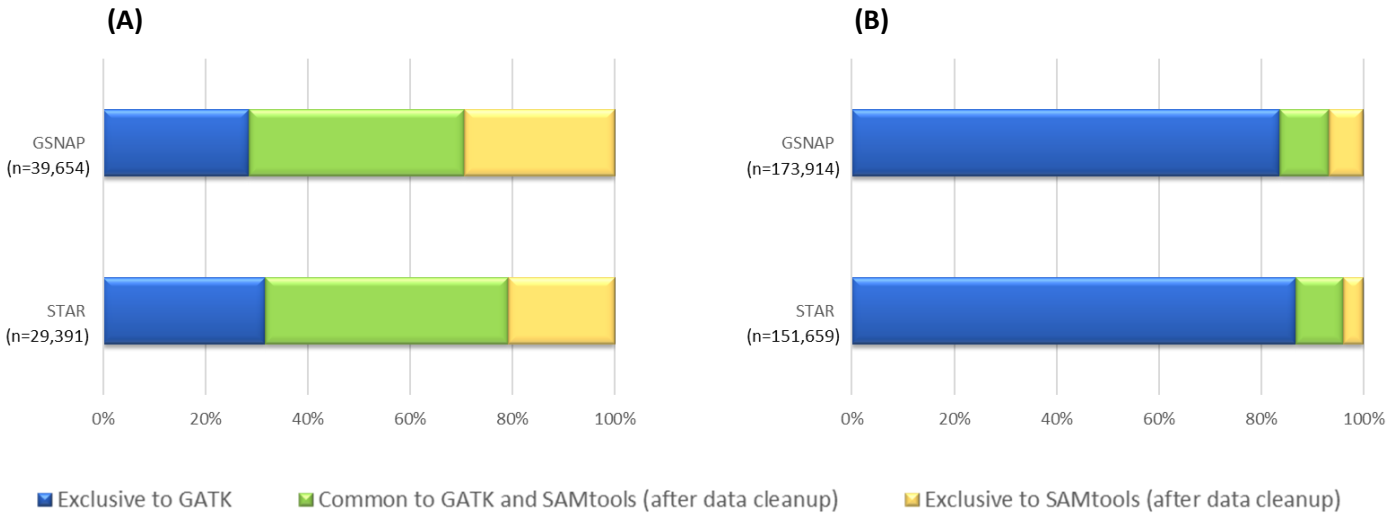


Figure S5.1. Overall variants (SNPs and indels) identified by GATK and SAMtools after GATK data cleanup following mapping with GSNAP and STAR: **(A)** gold standard; **(B)** breast tissue sample. The represented number of variants refers to the mean of called variants using each trimming tool.

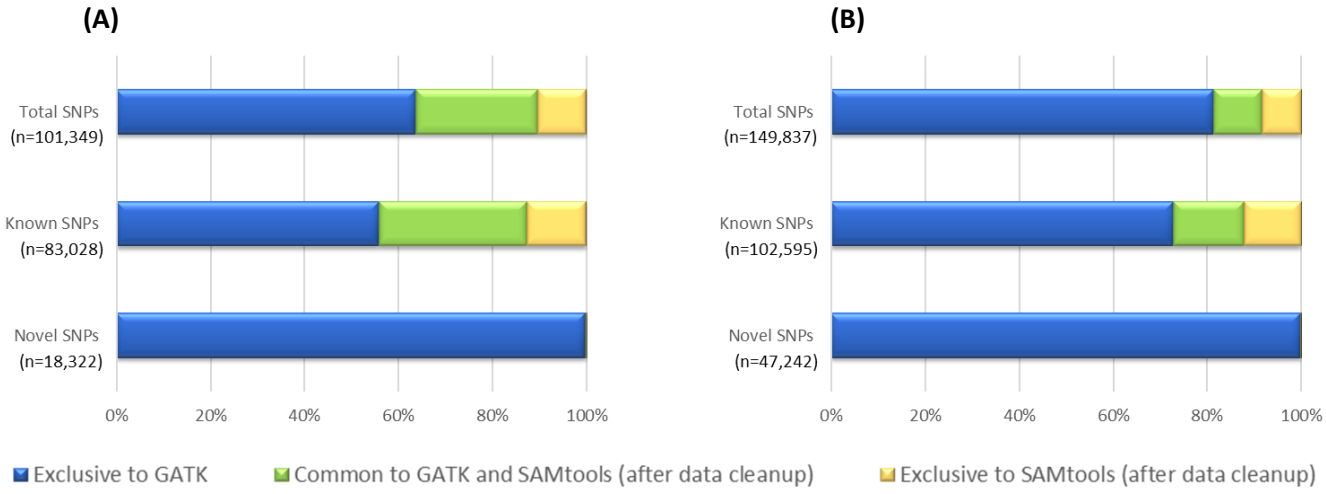


Figure S5.2. SNPs identified by GATK and SAMtools after GATK data cleanup following mapping with GSNAP: **(A)** gold standard; **(B)** breast tissue sample. Total SNPs - all SNPs (known and novel). Known SNPs - SNPs found in dbSNP150. Novel SNPs - SNPs not found in dbSNP150. The represented number of SNPs refers to the mean of called variants using each trimming tool.

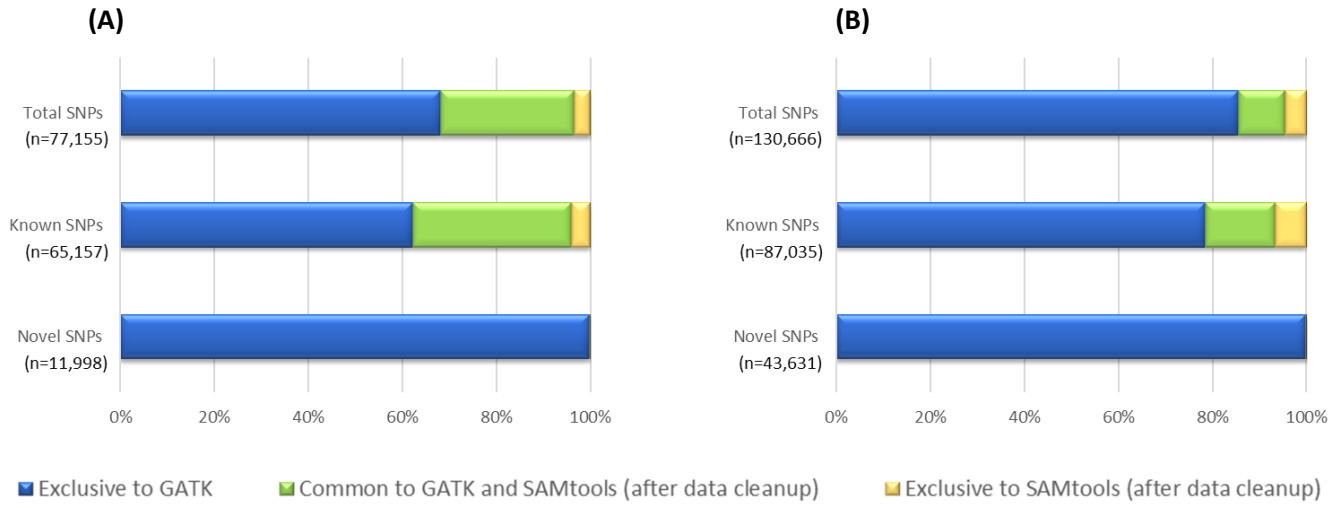


Figure S5.3. SNPs identified by GATK and SAMtools after GATK data cleanup following mapping with STAR: **(A)** gold standard; **(B)** breast tissue sample. Total SNPs - all SNPs (known and novel). Known SNPs - SNPs found in dbSNP150. Novel SNPs - SNPs not found in dbSNP150. The represented number of SNPs refers to the mean of called variants using each trimming tool.

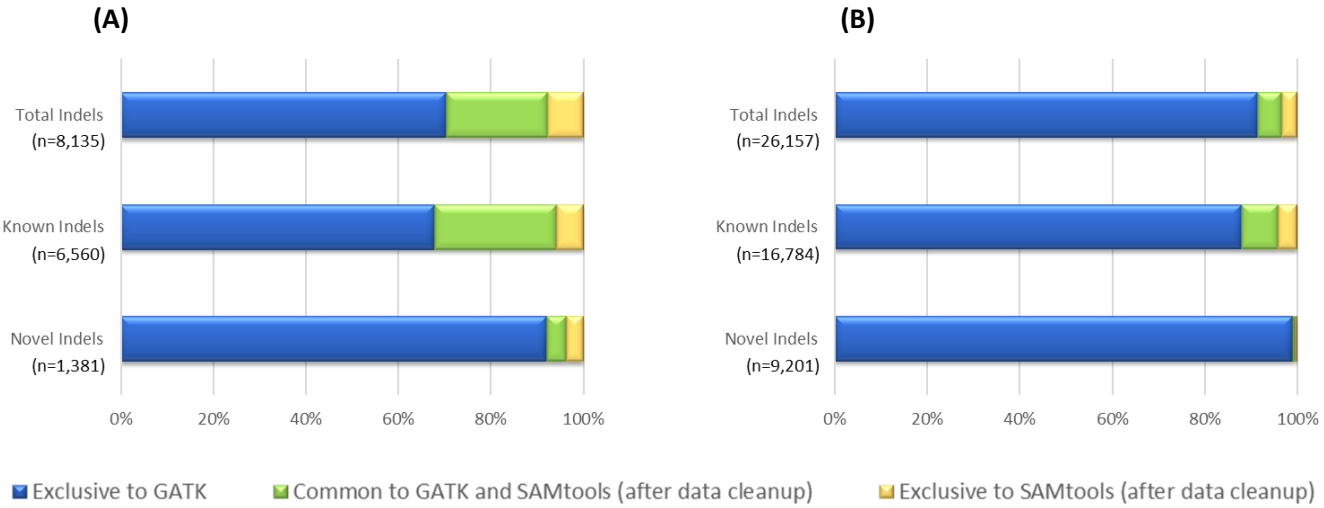


Figure S5.4. Indels identified by GATK and SAMtools after GATK data cleanup, following mapping with GSNAP: **(A)** gold standard; **(B)** breast tissue sample. Total Indels - all indels (known and novel). Known Indels - indels found in dbSNP150. Novel Indels - indels not found in dbSNP150. The represented number of indels refers to the mean of called variants using each trimming tool.

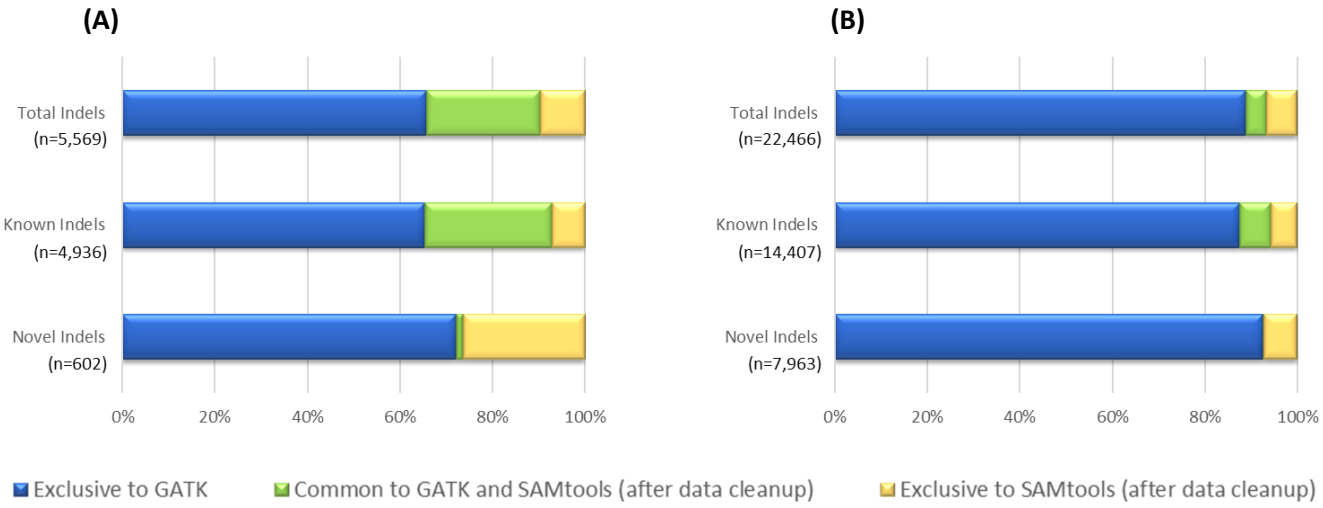


Figure S5.5. Indels identified by GATK and SAMtools after GATK data cleanup, following mapping with STAR: **(A)** gold standard; **(B)** breast tissue sample. Total Indels - all indels (known and novel). Known Indels - indels found in dbSNP150. Novel Indels - indels not found in dbSNP150. The represented number of indels refers to the mean of called variants using each trimming tool.

Chapter VI. Identification of New Susceptibility Loci for Breast Cancer

Table S6.1. Number of reads obtained for each sample (paired-end reads). CBr – breast tissue samples from controls (healthy women). NMBr – normal-matched breast tissue samples from breast cancer patients.

Sample	Number of reads (paired-end)
CBr_1	144,994,868
CBr_2	116,051,194
CBr_3	107,143,246
CBr_4	135,803,746
CBr_5	108,762,208
CBr_6	113,009,960
CBr_7	114,180,014
CBr_8	111,020,632
CBr_9	110,028,616
CBr_10	142,775,670
CBr_11	100,309,500
CBr_12	113,631,676
NMBr_1	112,711,790
NMBr_2	186,664,912
NMBr_3	124,642,784
NMBr_4	137,295,484
NMBr_5	170,296,276
NMBr_6	138,006,366
NMBr_7	118,910,890
NMBr_8	127,985,674
NMBr_9	93,756,892
NMBr_10	110,932,024
NMBr_11	134,660,924
NMBr_12	123,206,430
NMBr_13	130,668,812
NMBr_14	131,218,580
NMBr_42	21,116,794

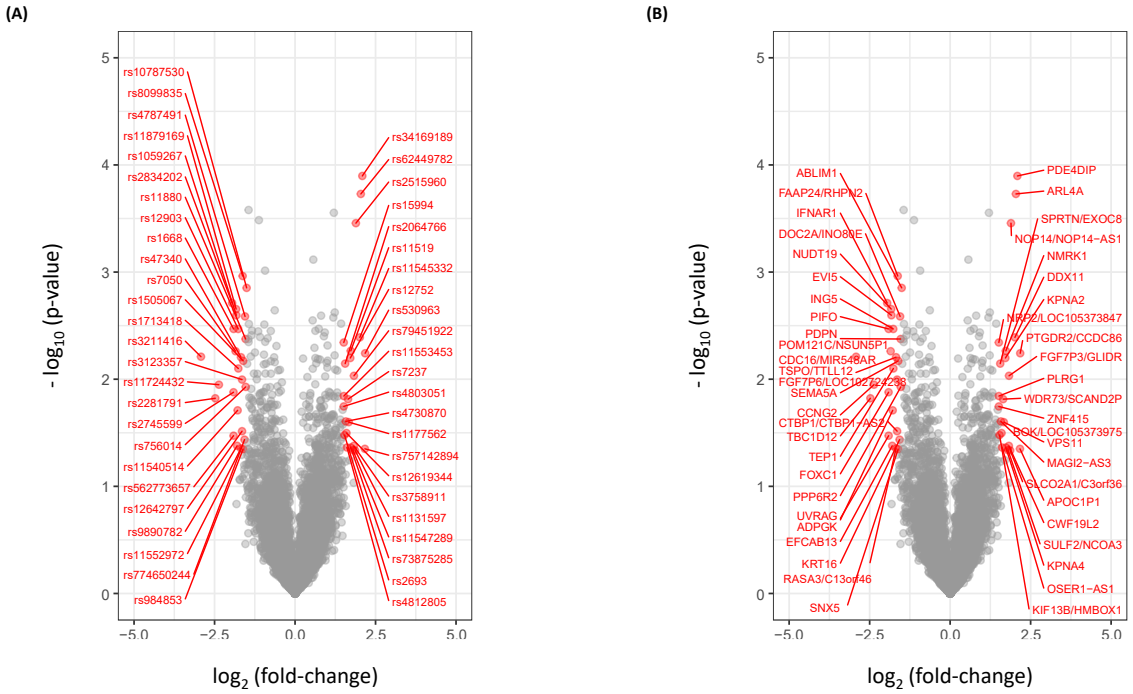


Figure S6.1. Volcano plot showing statistical significance [-$\log_{10}$ (p-value)] against fold-change (FC) of genetic variants from the case-control study using RNA-seq data. In red are represented the 49 genetic **(A)** variants and **(B)** genes with statistically significant difference between allelic expression (AE) ratios from breast cancer patients and controls (p-value<0.05) and with a $|\log_2(FC)|>1.5$. Positive x-values indicate up-regulated variants/genes, and negative x-values indicate down-regulated variants/genes.

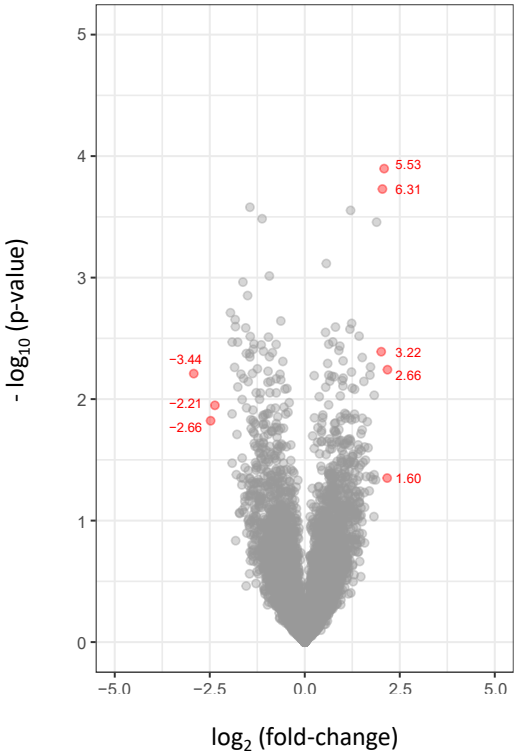


Figure S6.2. Volcano plot showing statistical significance [-$\log_{10}$ (p-value)] against fold-change (FC) of genetic variants case-control study using RNA-seq data. In red are the Hedges' g values of the 8 genetic variants with statistically significant difference between allelic expression (AE) ratios from breast cancer patients and controls (p-value<0.05) and with a $|\log_2(FC)|>2$. Positive x-values indicate up-regulated variants/genes, and negative x-values indicate down-regulated variants/genes.

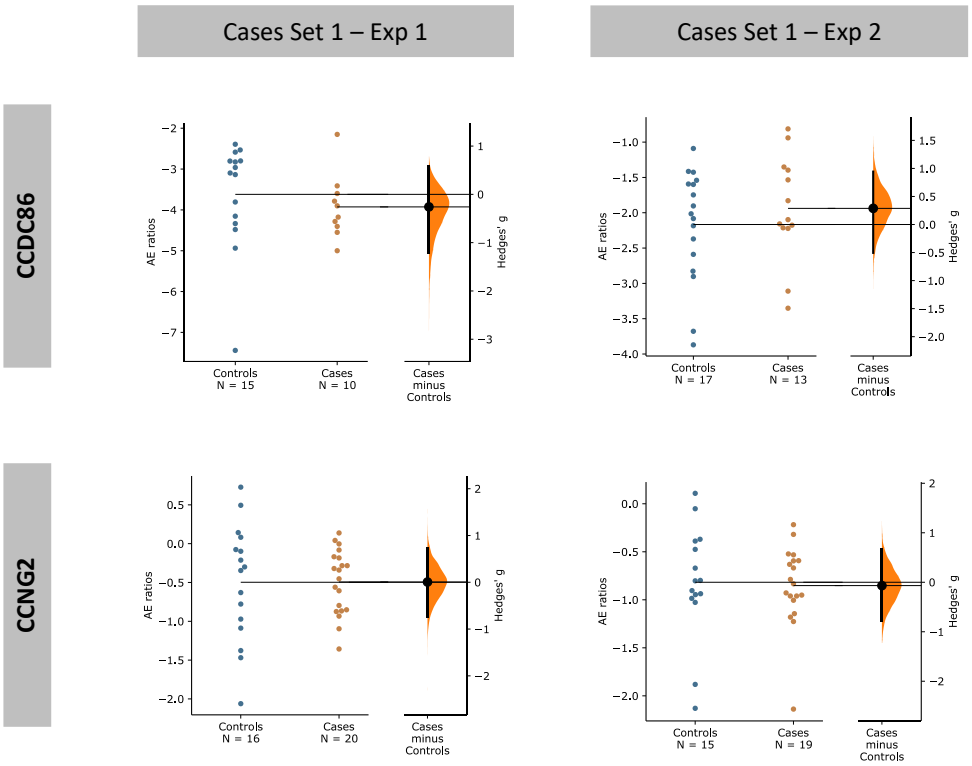


Figure S6.3. Case-control study using allelic expression (AE) ratios measured at rs530963 (C/A, *CCDC86* gene) and at rs11724432 (T/G, *CCNG2* gene) in normal breast tissue. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the two experiments conducted with the digital PCR system (Biomark™ HD system, Fluidigm) using the first set of cases (set 1).

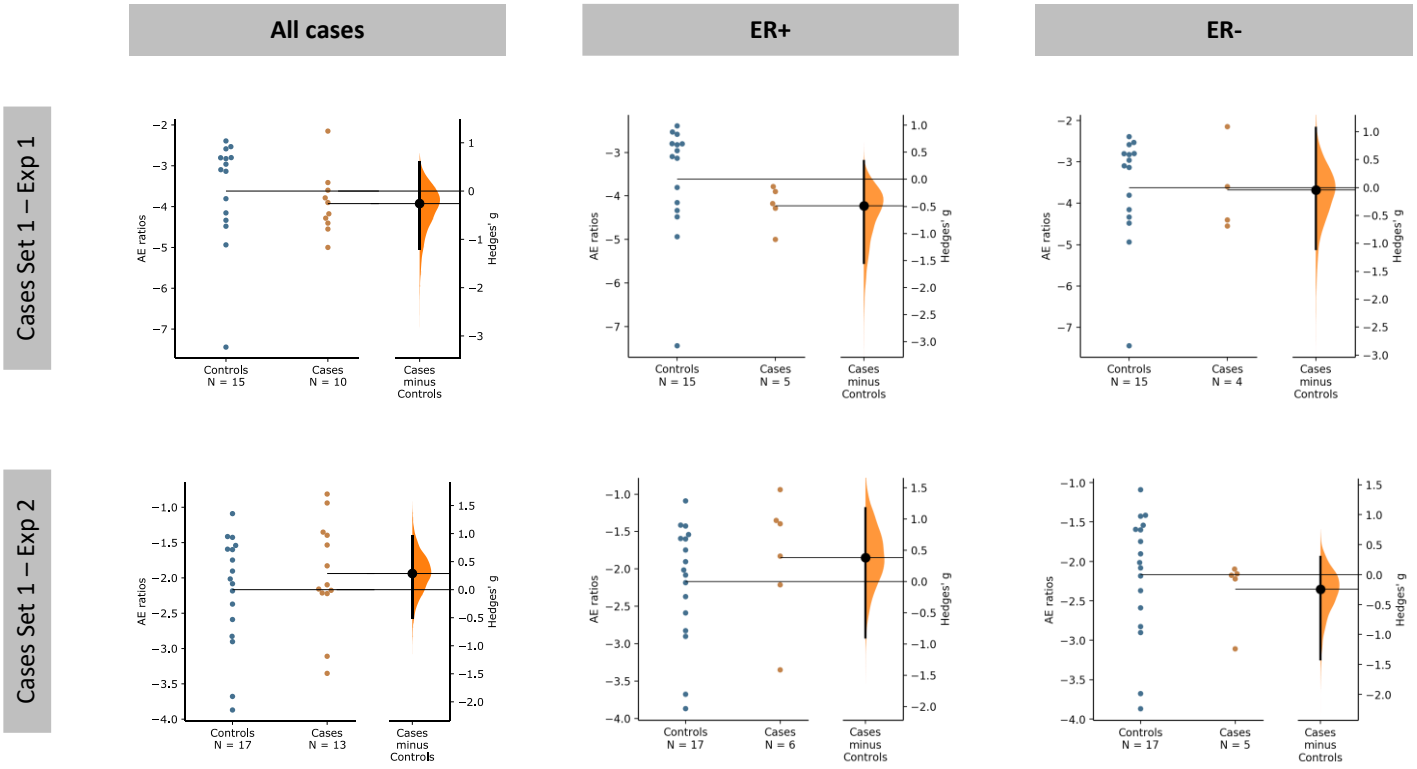


Figure S6.4. Case-control study using allelic expression (AE) ratios measured at rs530963 (C/A, *CCDC86* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the two experiments conducted with the digital PCR system (Biomark™ HD system, Fluidigm) using the first set of cases (set 1).

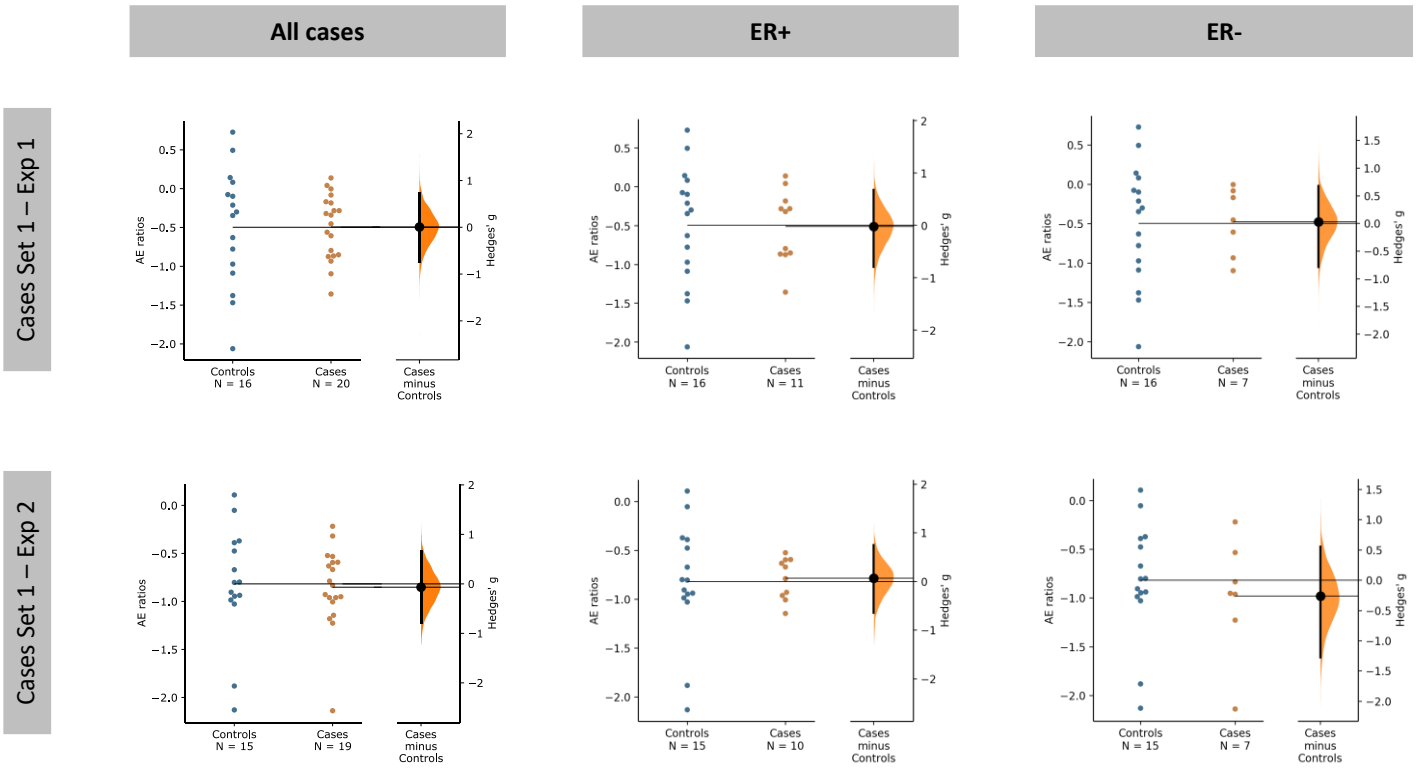


Figure S6.5. Case-control study using allelic expression (AE) ratios measured at rs11724432 (T/G, *CCNG2* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the two experiments conducted with the digital PCR system (Biomark™ HD system, Fluidigm) using the first set of cases (set 1).

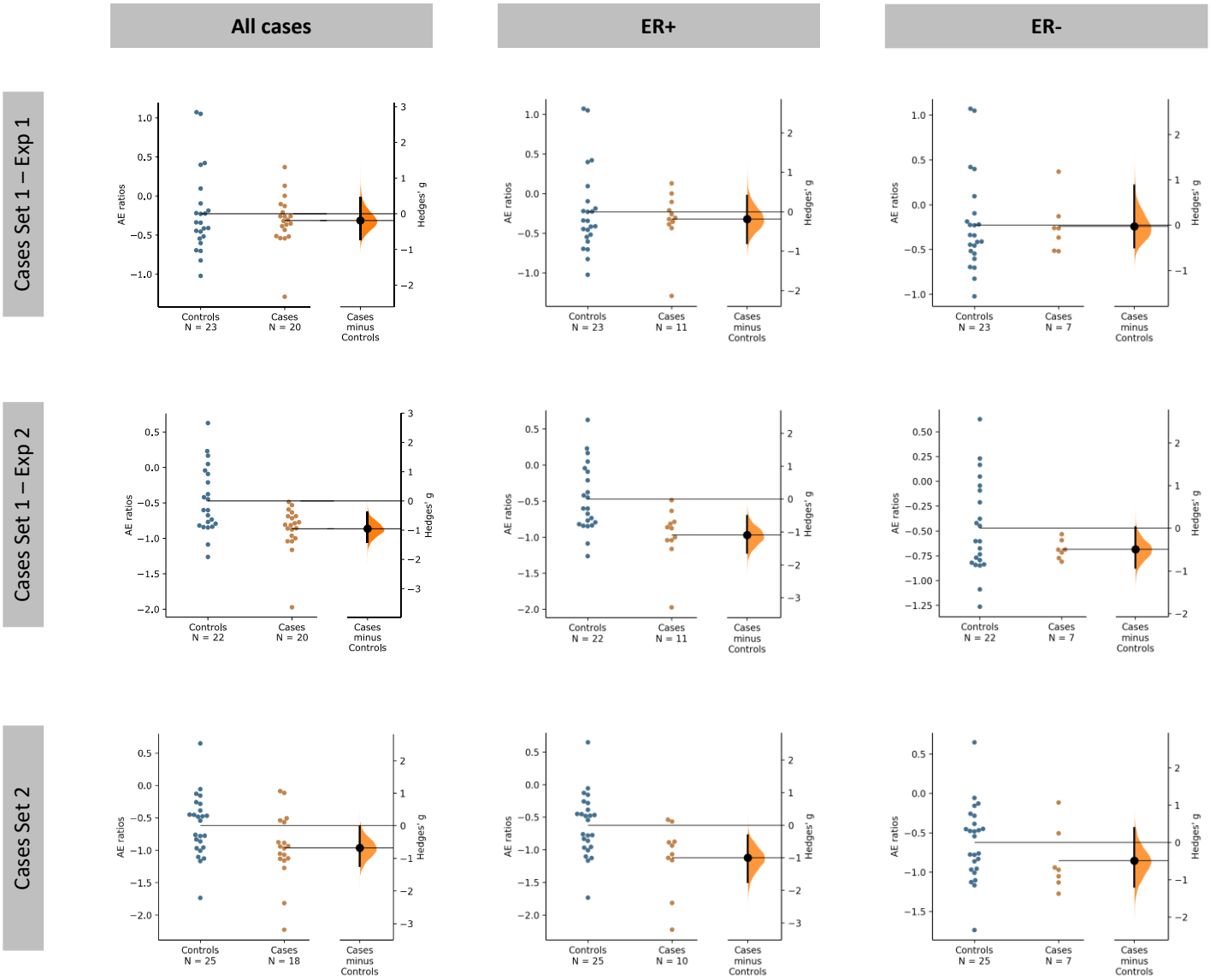


Figure S6.6. Case-control study using allelic expression (AE) ratios measured at rs62449782 (T/G, *ARL4* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the three experiments conducted with the digital PCR system (Biomark™ HD system, Fluidigm) using the first set of cases (set 1 – exp 1 and exp 2) and the second set of cases (set 2 - exp 3).

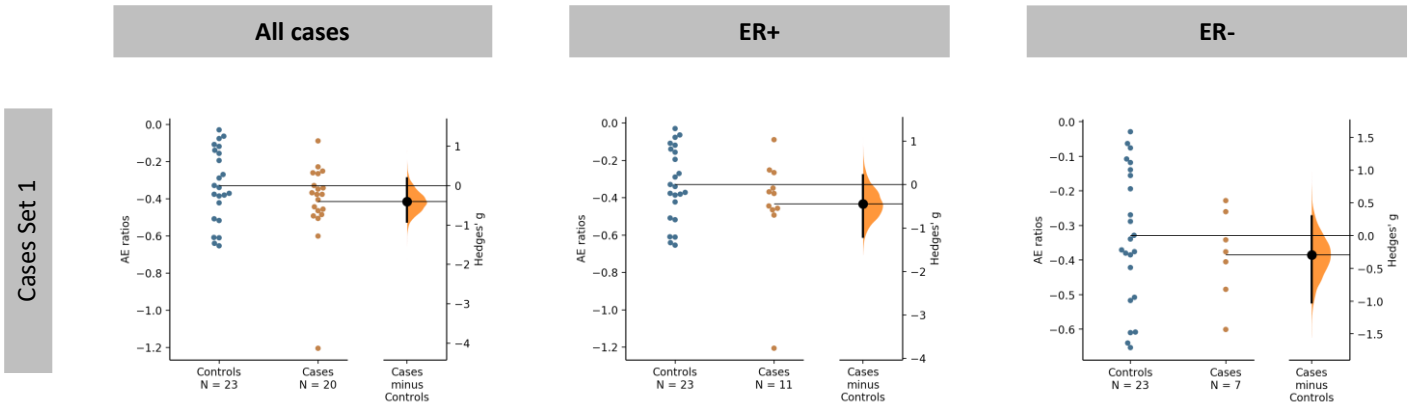


Figure S6.7. Case-control study using allelic expression (AE) ratios measured at rs62449782 (T/G, *ARL4* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the experiment conducted with the real-time PCR system (CFX384 Real-Time system, BioRad) using the first set of cases (set 1).

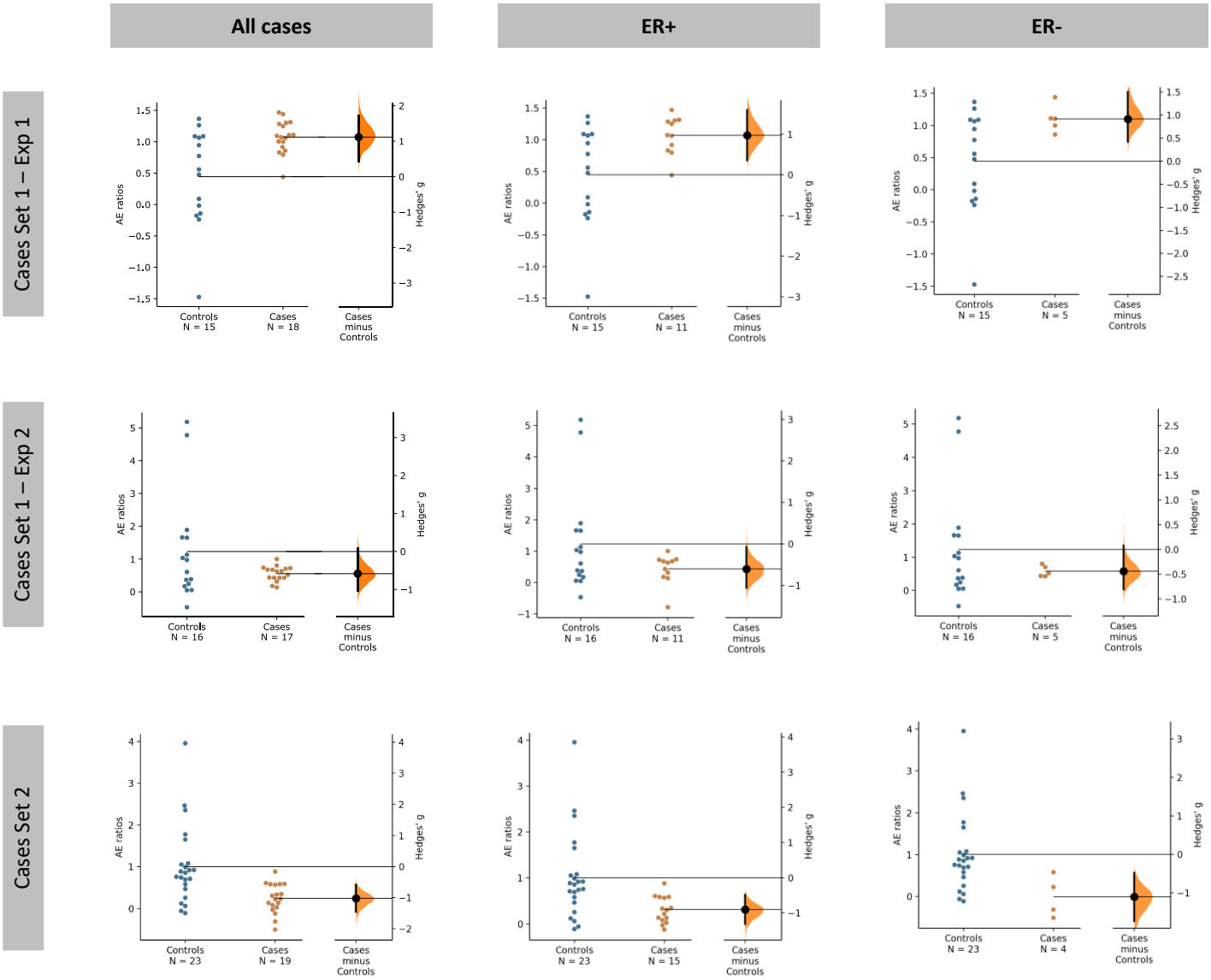


Figure S6.8. Case-control study using allelic expression (AE) ratios measured at rs2281791 (T/C, *TBC1D12* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the three experiments conducted with the digital PCR system (Biomark™ HD system, Fluidigm) using the first set of cases (set 1 – exp 1 and exp 2) and the second set of cases (set 2 - exp 3).

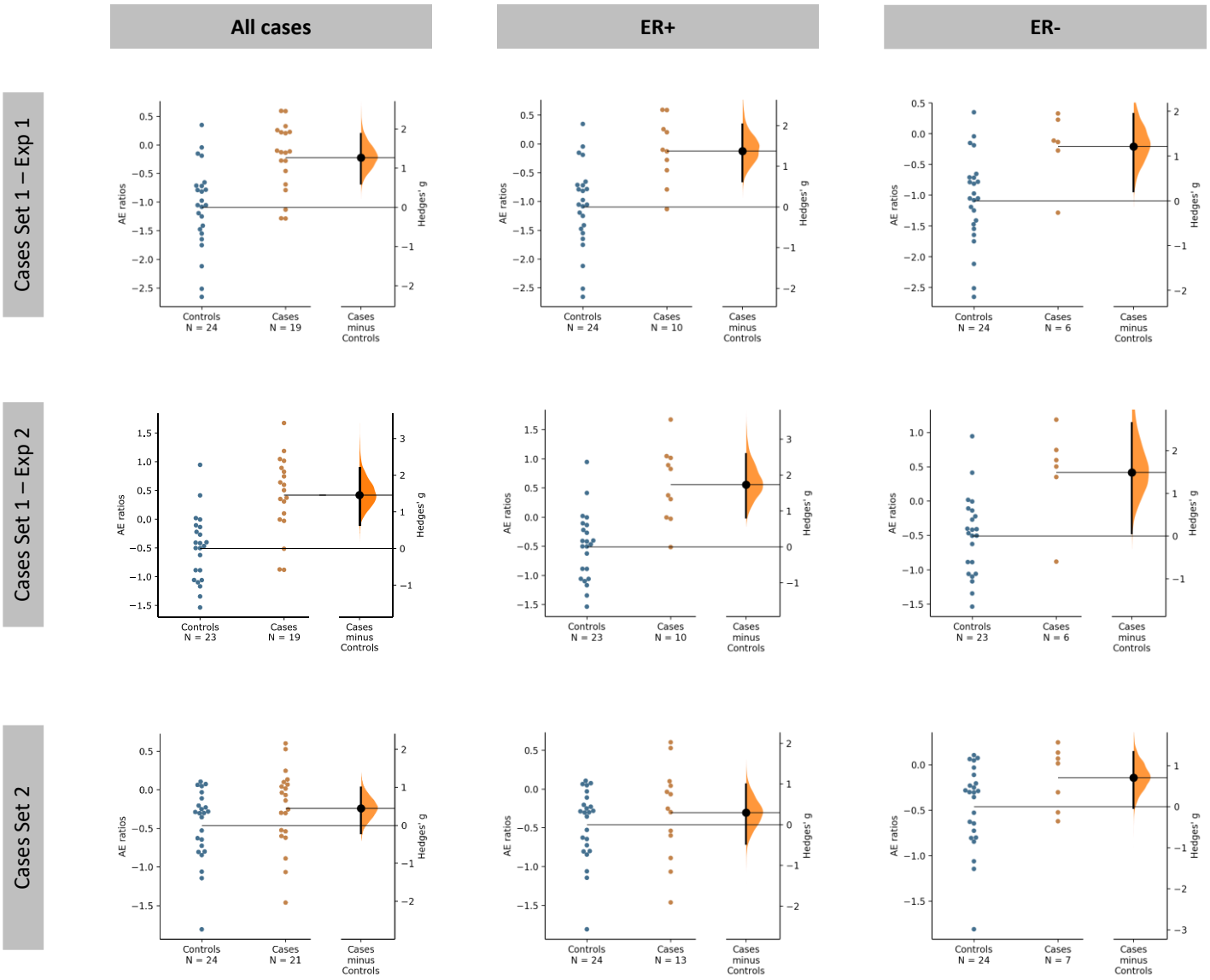


Figure S6.9. Case-control study using allelic expression (AE) ratios measured at rs11545332 (G/A, *DDX11* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the three experiments conducted with the digital PCR system (Biomark™ HD system, Fluidigm) using the first set of cases (set 1 – exp 1 and exp 2) and the second set of cases (set 2 - exp 3).

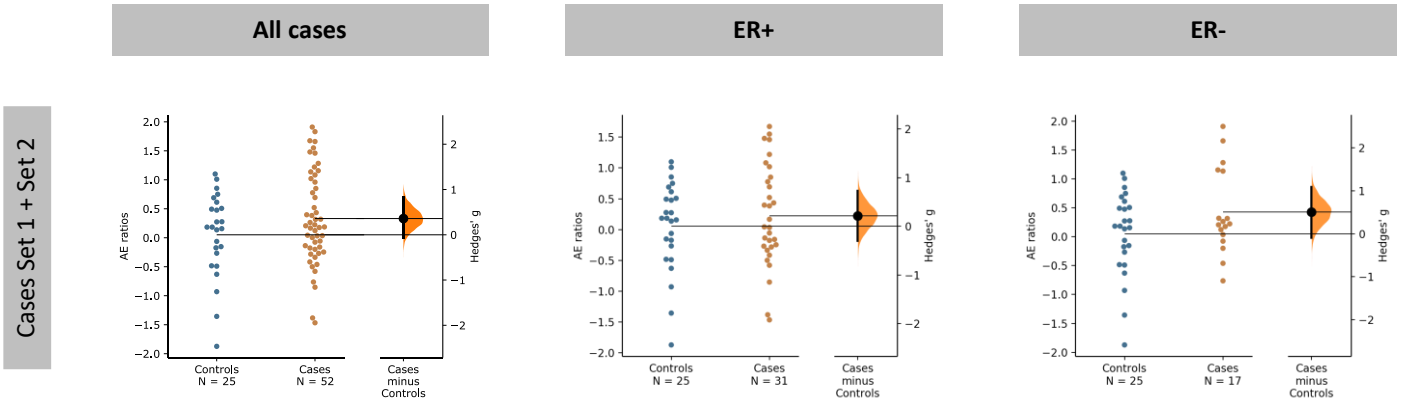


Figure S6.10. Case-control study using allelic expression (AE) ratios measured rs11545332 (G/A, *DDX11* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the results from the experiment conducted with the real-time PCR system (CFX384 Real-Time system, BioRad) using samples from both set of cases (set 1 and set 2).

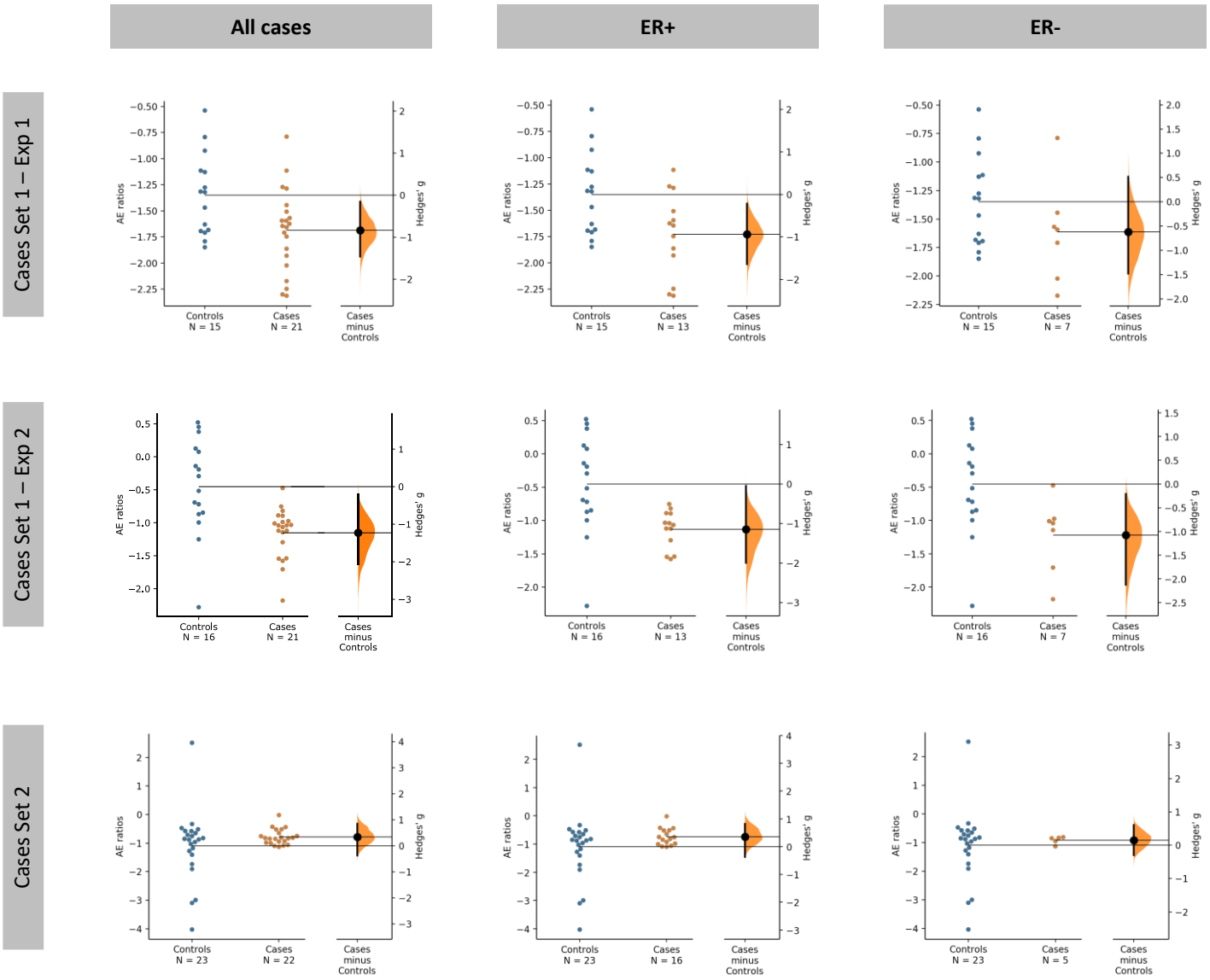


Figure S6.11. Case-control study using allelic expression (AE) ratios measured at rs3211416 (C/T, *CDC16* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the three experiments conducted with the digital PCR system (Biomark™ HD system, Fluidigm) using the first set of cases (set 1 – exp 1 and exp 2) and the second set of cases (set 2 - exp 3).

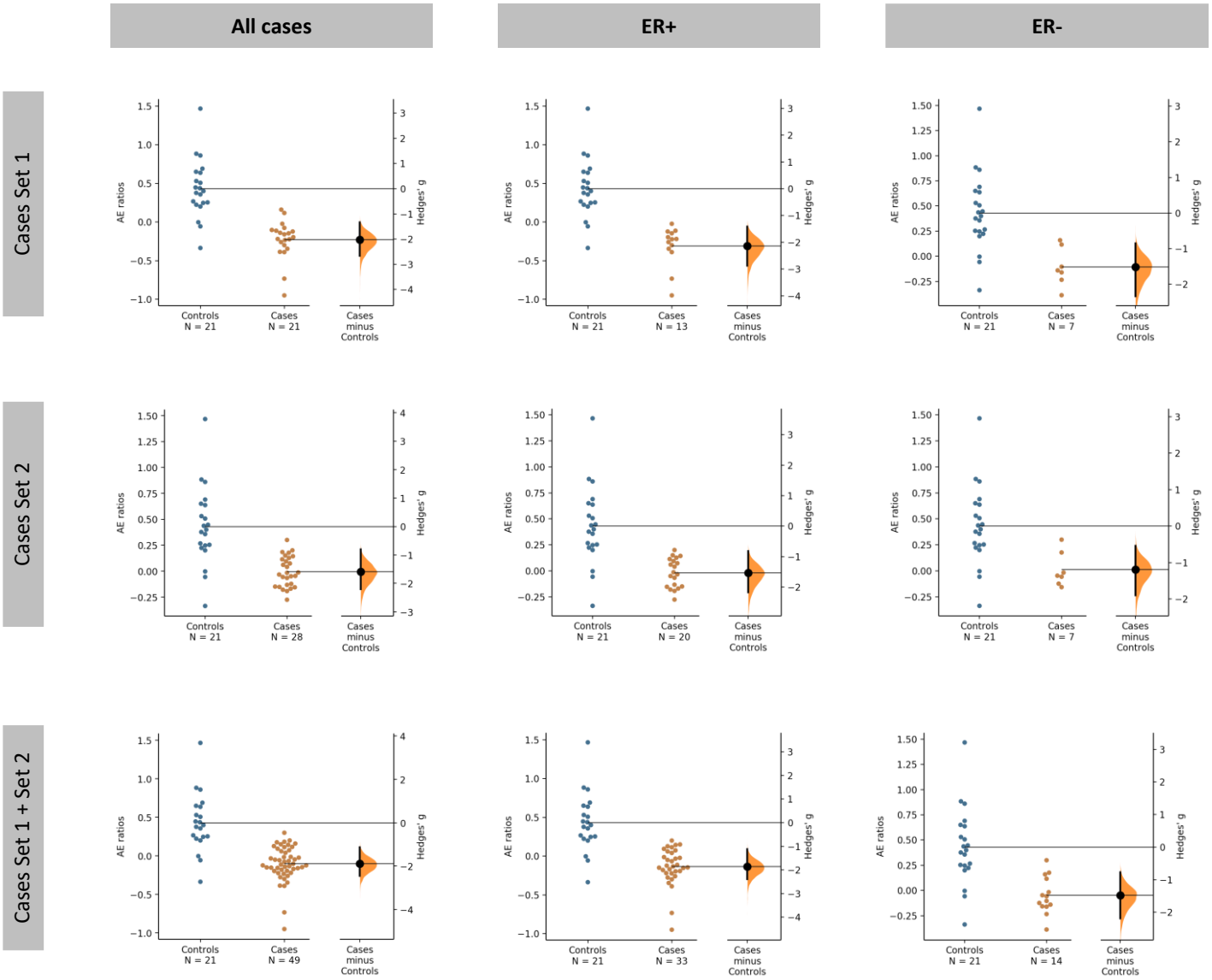


Figure S6.12. Case-control study using allelic expression (AE) ratios measured at rs3211416 (C/T, *CDC16* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the results from the first experiment conducted with the real-time PCR system (CFX384 Real-Time system, BioRad) using samples from both set of cases (set 1 and set 2).

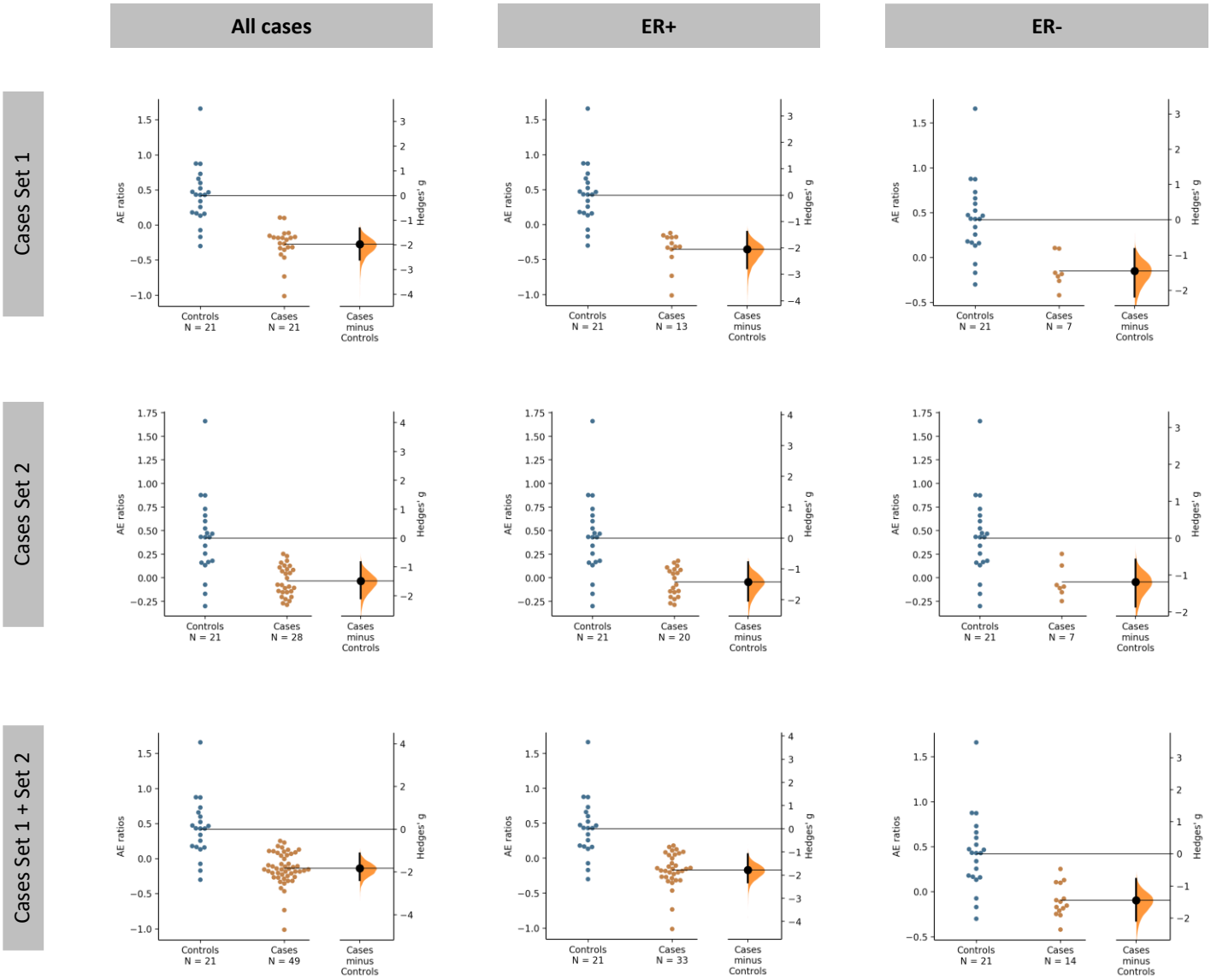


Figure S6.13. Case-control study using allelic expression (AE) ratios measured at rs3211416 (C/T, *CDC16* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the results from the second experiment conducted with the real-time PCR system (CFX384 Real-Time system, BioRad) using samples from both set of cases (set 1 and set 2).

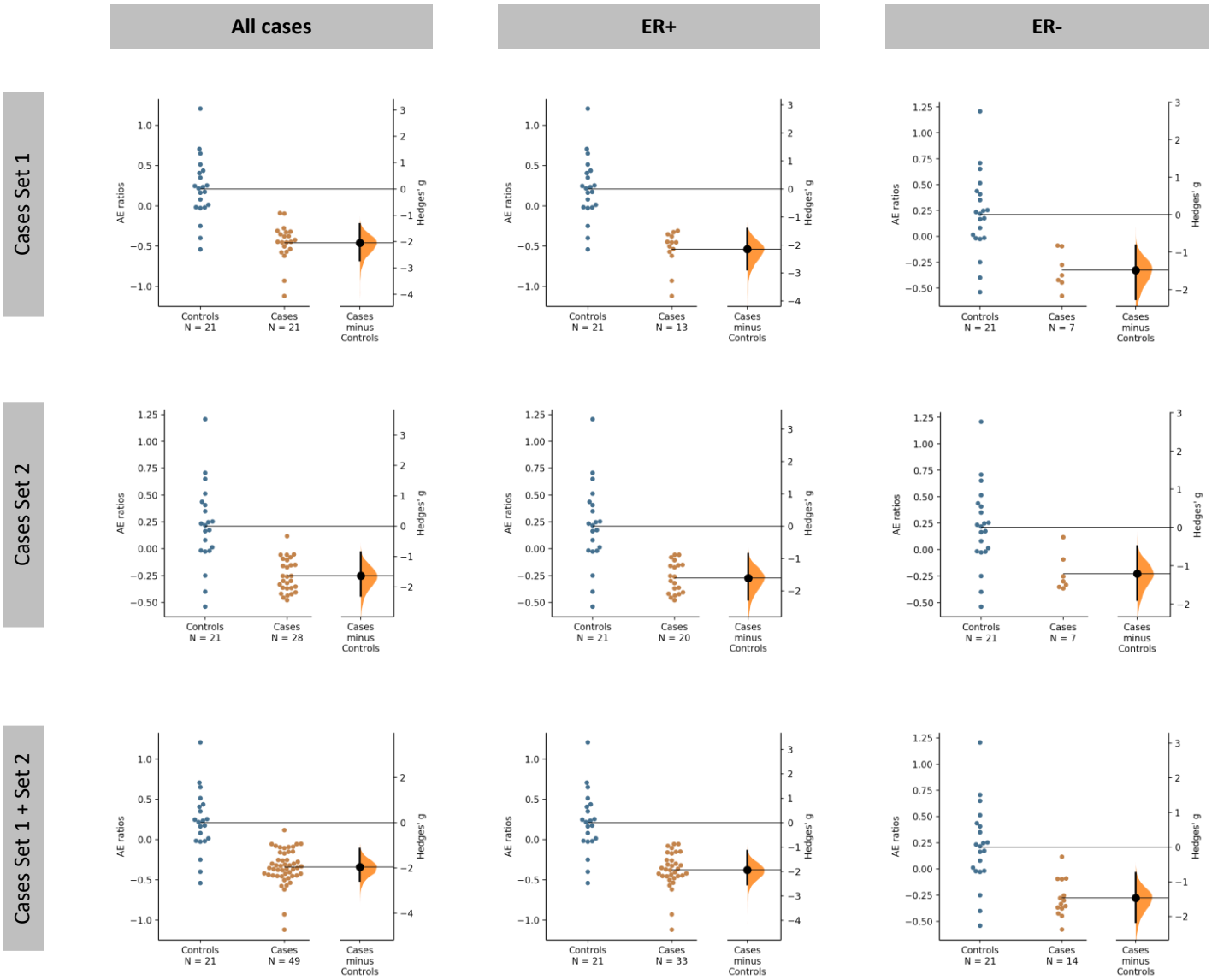


Figure S6.14. Case-control study using allelic expression (AE) ratios measured at rs3211416 (C/T, *CDC16* gene) in normal breast tissue, considering oestrogen receptor (ER) status from cases. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the results from the third experiment conducted with the real-time PCR system (CFX384 Real-Time system, BioRad) using samples from both set of cases (set 1 and set 2).

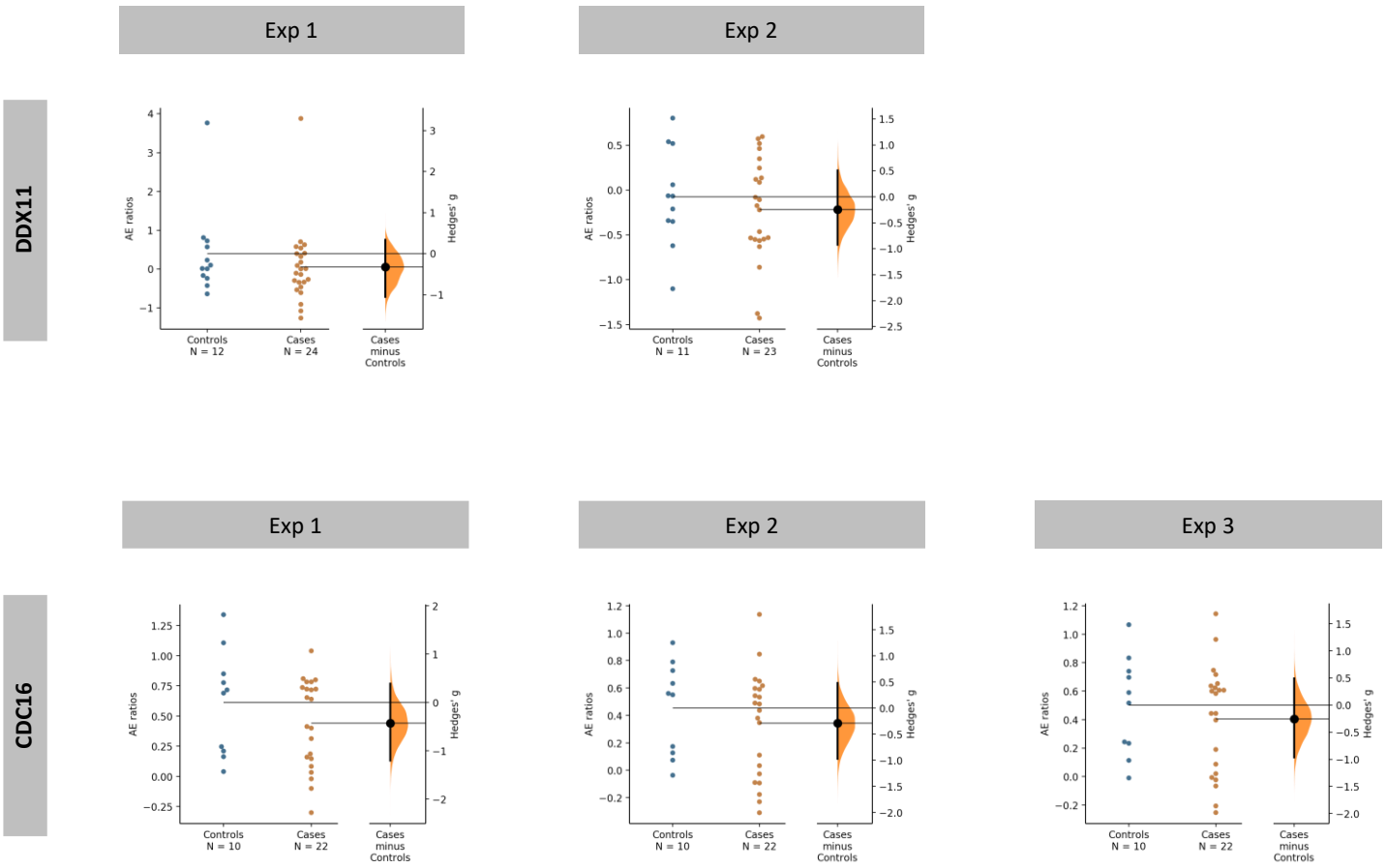


Figure S6.15. Case-control study using allelic expression (AE) ratios measured at rs11545332 (G/A, *DDX11* gene) and at rs3211416 (C/T, *CDC16* gene) in blood. In the left y-axis are indicated the AE ratios from heterozygous individuals for the SNPs (blue dots for controls and orange dots for cases). The number of samples used in each case-control experiment is indicated (N). In the right y-axis is the Hedges' *g* effect size represented as a bootstrap (5000 resamples) 95% confidence interval (95% CI), and aligned with the mean of the cases group (black dot). The figure shows the experiments conducted for each SNP with the real-time PCR system (CFX384 Real-Time system, BioRad).