ANTÓNIO MANUEL LOURENÇO DE GÓIS

STUDY OF THE ASSOCIATION OF DIFFERENTIAL ALLELIC EXPRESSION WITH BREAST CANCER CLINICAL FEATURES

UAlg FMCB
UNIVERSIDADE DO ALGARVE
FACULDADE DE MEDICINA E CIÊNCIAS BIOMÉDICAS

2023

ANTÓNIO MANUEL LOURENÇO DE GÓIS

STUDY OF THE ASSOCIATION OF DIFFERENTIAL ALLELIC EXPRESSION WITH BREAST CANCER CLINICAL FEATURES

Master in Oncobiology – Molecular Mechanisms of Cancer

This work was done under the supervision of

Joana Xavier, Ph.D

Ana Teresa Maia, Ph.D

UAlg FMCB
UNIVERSIDADE DO ALGARVE
FACULDADE DE MEDICINA E CIÊNCIAS BIOMÉDICAS

2023

STUDY OF THE ASSOCIATION OF DIFFERENTIAL ALLELIC EXPRESSION WITH BREAST CANCER CLINICAL FEATURES

Declaração de autoria de trabalho

Declaro ser o autor deste trabalho escrito, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

António Manuel Lourenço de Góis

_____

“There is no such thing as a single-issue struggle because we do not lead single-issue lives.”

Audre Lorde

Acknowledgements

Quero especialmente agradecer à minha orientadora Professora Doutora Joana Xavier pela disponibilidade, paciência e ensinamentos bem como todas as propostas, prudências e recomendações providenciadas ao longo do projeto e de como o direcionar.

Um grande agradecimento à minha coorientadora Professora Doutora Ana Teresa Maia, por ter aconselhado este projeto e por me integrar no grupo onde recebi sugestões valiosas para a realização do projeto.

Gostaria de agradecer à colega de grupo Marinella Ghezzo, pelo trabalho desempenhado nos dados de expressão alélica fulcrais para realização desta tese, e sem a sua contribuição prévia esta tese não poderia ter sido realizada.

Agradeço também aos colegas Ramiro Magno e Isabel Duarte por providenciar ajuda e input em todas as etapas da minha aprendizagem da linguagem computacional R, ao providenciar lições e packages que me introduziram a como trabalhar neste meio, bem como vários datasets utilizados no projeto final.

Um grande agradecimento em particular ao colega André Duarte por me ajudar a adaptar no uso de R, interrompendo as suas obrigações repetidamente de modo a me esclarecer as minhas várias dúvidas, sempre de forma imediata e simpática.

E finalmente, agradeço à minha família por todo o apoio que me dá, mesmo nos momentos de maior inseguridade.

Abstract

Genome-wide association studies have identified thousands of low-risk variants associated with BC. Since most associated variants are in non-coding regions of the genome and can be in linkage disequilibrium with many others, it's difficult to pinpoint the true causal variant and its target gene.

Post-GWAS functional analysis have shown that altering the expression of genes by cis-acting variants is a common mechanism involved in risk for breast cancer (BC). Therefore, we hypothesize that germline cis-regulatory variants may determine/contribute to tumour clinical features. We propose to test the association of allelic expression (AE) ratios with BC tumours clinical features and patient's survival. Differences in allelic ratios distributions between tumours with different molecular profiles would indicate a different genetic background for cis-regulatory variants between these tumours. To perform this analysis, we used The Cancer Genome Atlas – Breast Cancer (TCGA-BRCA) normal-matched dataset, and with the R programming language, we complemented this data with allelic expression ratios in various SNPs previously calculated, resulting in a dataset of 105 patients with both clinical and AE data.

We identified 63 variants displaying significant effect sizes (|Hedges's G| $\geq$ 0.5, 95% interval not crossing 0) for ER statuses, 119 for PR status and 58 for HER2 status, some of them associated with two receptor statuses. Only one variant (rs3764859 in *COQ6* and *ENTPD5* genes) was found statistically associated with ER status after correction for multiple testing (q-value $\leq$ 0.1). Additionally, six of the variants were also associated with survival including rs2236225, a variant located at gene *MTHFD1* that showed a larger expression difference in both negative ER and PR populations compared to their positive counterparts.

We found evidence that cis-acting variants may determine BC clinical features and disease development, and therefore BC subtypes. However, these findings need to be replicated in an independent dataset to be validated.

Keywords:

Allelic Expression (AE) analysis – Breast Cancer – Cis-regulation – Tumours clinical features

Resumo

O cancro da mama feminino representa 11,7% de incidência de cancro a nível mundial, constituindo o cancro mais comum, reclamando 6,9% da mortalidade mundial associada com cancro. Em Portugal, o cancro da mama é a neoplasia mais comum nas mulheres (26,4%) sendo responsável pela maior fração de mortalidade feminina relacionada com cancro (15,5%). A biologia tumoral e quadro clínico fazem o cancro da mama uma doença heterogénea, que pode ser subdividida em diferentes subtipos de tumor da mama com perfis de expressão génica distintos. É também uma doença complexa que engloba fatores de risco ambientais e genéticos na sua etiologia. Estudos de associação do genoma completo (GWAS) já identificaram milhares de variantes de baixo risco associadas com cancro da mama, e inclusive, quatro subtipos específicos de tumores. Contudo, a maioria das variantes associadas estão em regiões não codificantes do genoma, e cada variante identificada pode estar em desequilíbrio de ligação (LD) com outras variantes, o que dificulta localizar verdadeira a variante causal e o seu gene alvo com precisão e confiança.

A importância e valor intrísico deste estudo consiste na identificação de variantes que podem estar associadas a cancro da mama e os seus subtipos. Estas associações acarretam implicações e ramificações importantes para o estudo da patologia em questão, uma vez que podem elucidar mecanismos de atuação da doença bem como a biologia de mesma. Esta melhor compreensão permite prever como e que o carcinoma evolui em populações de subtipos específicos e de que forma esta pode responder a terapia de acordo com as mutações adquiridas.

O cancro da mama, pode apresentar combinações específicas de caraterísticas moleculares e devido a essa peculiaridade, é possível agrupar os tumores em subtipos de cancro da mama que expressam e deixam de expressar uma gama de recetores, originando uma chave de expressão génica tumoral. O subtipo mais comum é identificado como Luminal A e carateriza-se pela maior expressão dos recetores de estrogénio (ER) e progesterona (PR) enquanto a expressão de recetor tipo 2 do fator de crescimento epidérmico humano (HER2) está reduzida. Por ordem de maior para menor incidência, o Luminal A (ER+/PR+/HER2-) é seguido por Luminal B (ER+/PR-/HER2+), Triplo Negativo (ER-/PR-/HER2-) e HER2-enriquecido (HER2+).

A maioria dos casos de cancro da mama são esporádicos, contudo cerca de 20%-30% apresentam hereditariedade, dos quais 5-10% apresentam uma forte componente herdada e um risco acrescido. Esta componente herdada tem diferentes graus de risco para o seu portador, que

pode chegar a um risco acrescido de 10 a 20 vezes, como no caso dos portadores de mutações patogénicas nos genes *BRCA1* e *BRCA2*, apesar de conjuntamente apresentarem uma frequência apenas de 0.4% em cancro da mama. Num grau intermédio, mutações em genes como *CHEK2* e *ATM* conferem o dobro do risco de cancro da mama, e por último, as componentes hereditárias de baixo risco (inferior a 1.5) são normalmente causadas por polimorfismos de nucleótido único (SNP) como o polimorfismo Pro919Ser no gene *BRIP1*. A introdução dos estudos de associação do genoma inteiro possibilitou a descoberta de centenas de variantes de baixo risco associadas com cancro da mama, e também associadas especificamente com subtipos moleculares de cancro da mama, que apresentam sobre ou subre expressão de recetores hormonais.

A análise funcional pós-GWAS das variantes associadas a risco demonstrou que a alteração da expressão génica por variantes cis-regulatórias constitui um mecanismo comum envolvido em risco para cancro da mama. Estas variantes regulam a expressão dos genes de um modo específico para cada um dos alelos do gene, estando geralmente associados a regiões regulatórias como promotores ou enhancers. Consequentemente, propusémos a hipótese que as variantes cis-regulatórias na linha germinal podem não só contribuir para risco mas também para determinar as características tumorais e quadro clínico do paciente. Para responder a esta pergunta, testámos a associação dos rácios de expressão alélica (AE) - um valor que mede diretamente o efeito de variantes cis-regulatórias – com caraterísticas clínicas de tumores da mama e sobrevivência dos pacientes. De modo a efetuar esta análise, foi utilizado o conjunto de dados The Cancer Genome Atlas – Breast Cancer (TCGA-BRCA) tecido normal-tumoral adjacente, para o qual possuímos dados de rácios de expressão alélica para vários polimorfismos de nucleótido único localizados em regiões transcritas de genes, assim como informação clínica para 105 indivíduos.

Testámos estatisticamente (recorrendo a linguagem de programação R) a associação entre rácios de expressão alélica (AE) medidos em 8039 SNPs expressos (aeSNPs) num número variável de indivíduos heterozigóticos para estas variantes, e a presença ou ausência dos recetores de estrogénio (ER), progesterona (PR) e sobre-expressão do fator de crescimento epidérmico humano 2 (HER2). Para esta análise aplicou-se o teste t de Student para variantes com rácios de AE que seguiam uma distribuição normal ou Mann-Whitney para variantes cujos rácios de AE apresentavam uma distribuição não normal, de acordo com os resultados do teste de normalidade Shapiro-Wilk. De forma a quantificar o tamanho de efeito (“effect size”) de AE entre tumores com presença ou ausência de receptores, utilizou-se o teste de Hedges’ G. A

análise de sobrevivência de acordo com os rácios de AE foi executada nos aeSNPs que apresentavam um tamanho de efeito significativo, usando um teste Two Stage Hazard Rate Comparison (TSHRC).

Identificámos 63 variantes que demonstravam tamanhos de efeito significantes de acordo com a presença/ausência de ER, 119 de PR e 58 sobre-expressão de HER2, e alguns destes, inclusive, associados com mais do que um recetor. Apenas uma variante (rs3764859 localizada nos genes *COQ6* e *ENTPD5*) apresentou-se estatisticamente associada com a presença/ausência do recetor de estrogénio após correção para testes múltiplos (q-value ≤ 0.1). Uma variante em particular - rs2236225, demonstrou uma elevada diferença na distribuição dos rácios de AE nas populações ER e PR negativas quando comparadas com a fração positiva e já foi previamente associada com mau prognóstico em mulheres na pré menopausa. Finalmente, 6 aeSNPs apresentavam valores de sobrevida livre de doença diferentes entre populações com rácios de expressão alélica opostos de acordo com o TSHRC.

Neste trabalho encontrámos evidências de que as variantes cis-regulatórias podem determinar caraterísticas de cancro da mama e do desenvolvimento da doença, e consequentemente de subtipos moleculares de cancro da mama. Contudo, estas descobertas precisam de ser replicadas de uma forma independente noutro conjunto de dados de forma a serem validadas. Outro tipo de estudos também será necessário para mapear as variantes cis-regulatórias cujos efeitos encontrámos associados com a biologia dos tumores, e revelar qual o seu mecanismo de regulação génica inerente. Este estudo demonstra o potencial de integrar análise de expressão alélica com informação clínica para melhor compreender a etiologia dos diferentes subtipos moleculaers de cancro da mama.

Palavras-chave:

Análise de Expressão Alélica (AE) – Cancro da Mama – Cis-regulação – Caraterísticas Clínicas Tumorais

Table of Contents

Index of Figures

Index of Tables

List of Abbreviations

AE – Allelic Expression

aeSNP – Allelic Expression Single Nucleotide Polymorphism

AHR – Aryl Hydrocarbon Receptor

AICR – American Institute for Cancer Research

ALT – Alternative Lengthening of Telomeres

APB – ALT-associated PML-NB

APOLD1 – Apolipoprotein L Domain Containing 1

ASR – Age Standardized Rate

ATM – Ataxia-Telangiesctasia Mutated Serine/Threonine Kinase

BC – Breast Cancer

BCA – Bias-corrected Accelerated confidence interval

BCAC – Breast Cancer Association Consortium

BC-MR – Breast Cancer Mortality Rate

Bet1 – Blocked Early in Transport 1 Gene

BMI – Body Mass Index

BORCS8 – BLOC-1 Related Complex Subunit 8

BRCA1 – Breast Cancer type 1

BRCA2 – Breast Cancer type 2

BRCT – BRCA1 C-terminus

BRIP1 – BRCA1-Interacting protein 1

CD8+ T-cells – Cluster of Differentiation 8 Thymus Cells

CD86 – Cluster of Differentiation 86

CHEK2 – Checkpoint Kinase 2

CIMBA – Consortium of Investigators of Modifiers of BRCA1/2

CK – Cytokeratin

CLIC4 – Chloride Intracellular Channel 4

COL4A1 – Collagen Type IV Alpha 1 Chain

COQ6 – Coenzyme Q6

CRAN – Comprehensive R Archive Network

DAE – Differential Allelic Expression

DALY – Disability-Adjusted Life Years

DCIS – Ductal Carcinoma *in Situ*

DDT – Dichloro-diphenyl-trichloroethane

DFS – Disease-Free Survival

DNA – Deoxyribonucleic Acid

DNAL4 – Dynein Axonemal Light Chain 4

DSB – Double Strand Break

ECM – Extracellular Matrix

EGF – Epidermal Growth Factor

EMT – Epithelial-Mesenchymal Transition

ENTPD5 – Ectonucleoside Triphosphate Diphosphohydrolase 5

eQTL – Expression Quantitative Trait Loci

ER – Estrogen Receptor

ER+ – Estrogen Receptor Positive

ER- – Estrogen Receptor Negative

ERBB2 – Erb-B2 Receptor Tyrosine Kinase 2 (HER2)

GC-content – Guanine-Cytosine Content

GLOBOCAN – Global Cancer Observatory

GRB7 – Growth Factor Receptor Bound Protein 7

GTEx – Genotype-Tissue Expression

GWAS – Genome-Wide Association Studies

Gy – Gray

HER2 – Human Epidermal Growth Factor Receptor 2

HER2+ – Human Epidermal Growth Factor Receptor 2 Positive

HER2- – Human Epidermal Growth Factor Receptor 2 Negative

HGNC – HUGO Gene Nomenclature Committee

HR – Hormone Receptor

iCOGS – Collaborative Oncological Gene-environment Study

ID – Identity

IDC – Infiltrating Ductal Carcinoma

IGF-1 – Insulin-like Growth Factor I

IL-2 – Interleukin 2

ILC – Invasive Lobular Carcinoma

KANK2 – KN Motif and Ankyrin Repeat Domains 2

LCIS – Lobular Carcinoma *in Situ*

LD – Linkage Disequilibrium

LIN52 – Lin-52 DREAM MuvB Core Complex Component

LINC – Linker of Nucleoskeleton and Cytoskeleton

LRP2 – Low Density Low Receptor Related Protein 2

LRP6 – Low Density Low Receptor Related Protein 6

LZTS1 – Leucine Zipper Tumour Suppressor 1

MEF2B – Myocyte Enhancer Factor 2B

METABRIC – Molecular Taxonomy of Breast Cancer International Consortium

METTL16 – Methyltransferase 16

MHC – major histocompatibility complex

MKI67 – Marker of Proliferation Ki-67

MRI – Magnetic Resonance Imaging

mRNA – Messenger Ribonucleic acid

MT1-MPP – Membrane Type-1 Metalloproteinase

MTHFD1 – Methylenetetrahydrofolate Dehydrogenase, Cyclohydrolase and Formyltetrahydrofolate Synthetase 1

MTMR6 – Myotubularin Related Protein 6

NCI – National Cancer Institute

NHGRI – National Human Genome Research Institute

NLRC5 – NLR Family CARD Domain Containing 5

OC – Oral Contraceptives

OR – Odds Ratio

PAM50 – Prediction Analysis of Microarray 50

PCB – Polychlorinated Biphenyl

PDE1B – Phosphodiesterase 1B

PIK3CA – Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha

P-LCIS – Pleomorphic Lobular Carcinoma *In Situ*

PML-NB – Promyelocytic Leukaemia Nuclear Bodies

PLEKHS1 – Pleckstrin Homology Domain Containing S1

PPP1R1A – Protein Phosphatase 1 Regulatory Inhibitor Subunit 1A

PR – Progesterone Receptor

PR+ – Progesterone Receptor Positive

PR- – Progesterone Receptor Negative

PBRM1 – Polybromo 1

PTBP3 – Polypyrimidine Tract Binding Protein 3

PTEN – Phosphatase and Tensin Homolog

RAD51 – RAD51 Recombinase

RNA – Ribonucleic Acid

RNAseq – Ribonucleic Acid Sequencing

ROS – Reactive Oxygen Species

rSNP – Regulatory Single Nucleotide Polymorphism

SASP – Senescence-Associated Secretory Phenotype

SBNO1 – Protein strawberry notch homolog 1

SEER – Surveillance, Epidemiology, and End Results

SGSM2 – Small G Protein Signalling Modulator 2

SLC44A2 – Solute Carrier Family 44 Member 2

SMIM4 – Small Integral Membrane Protein 4

SNP – Single Nucleotide Polymorphism

STK11 – Serine/Threonine Kinase 11

SUN2 – Sad1 and UNC84 Domain Containing 2

TBCD – Tubulin Folding Cofactor D

TCDD – Dioxin 2,3,7,8-TCDD

TCGA – The Cancer Genome Atlas

TERT – Telomerase Reverse Transcriptase

TGF-β – Transforming growth factor β

TNBC – Triple Negative Breast Cancer

TNS3 – Tensin 3

TP53 – Tumour Protein 53 Gene

TSHRC – Two Stage Hazard Rate Comparison

tSNP – Transcribed Single Nucleotide Polymorphism

UDPase – Uridine 5′-diphosphatase

US – United States

WCRF – The World Cancer Research Fund

WHO – World Health Organization

WNT – Wingless/Integrated Signalling Pathway

ZEB1 – Zinc Finger E-Box Binding Homeobox 1

ZNRF3 – Zinc and Ring Finger

Introduction

1 Introduction

1.1 Cancer

Cancer is a major cause of death worldwide, contributing to 1 in 6 deaths in 2020. Nevertheless, many cancers can be effectively treated when detected early through screening or early diagnosis. Many cancers are also preventable (30%-50% of all cancers are believed to be avoidable) by not partaking in risk behaviours such as smoking, lack of physical activity and low fruit and vegetable intake (Sung et al., 2021).

With an overall risk of 20% of developing cancer between the ages 0-74, it causes not only clinical but social and economic repercussions in a patient's life, and even though it is the second leading cause of death worldwide, the World Health Organization (WHO) deems cancer to be the most disruptive disease when it comes to Disability-Adjusted Life Years (DALYs), incumbering their patients and restricting their livelihood (Mattiuzzi & Lippi, 2019).

The invasion of other organs, better known as metastasis, is the main cause of cancer death when spread over the entire organism.

Invasion of neighbouring tissues by malignant cells might be what defines cancer as a disease, but to better understand and prevent invasion one needs to know which mechanisms these cells are able to employ for their rapid spread. The relentless study of this question resulted in the Hallmarks of Cancer (Figure 1.1) which encompass the capabilities of tumour cells to potentiate their growth and metastasis (Hanahan & Weinberg, 2011).

Cancer cells are capable of continuous and incessant proliferation by deregulating proliferative signals. This can be accomplished via autocrine stimulation by producing their own growth factors, by signalling nearby cells to provide said growth factors and by overexpressing cell surface receptor proteins making the cell more sensitive to growth signals. Activation of a downstream pathway of the receptor can cause the pathway to become independent of a ligand-receptor signal, being stimulated without the need for growth factors creating a constitutive activation (Hanahan & Weinberg, 2011).

Continuous proliferation is just as advantageous as the lack of barriers to cell growth. Cells have safety mechanisms to prevent unregulated growth via the aptly named tumour suppressor genes, that halt the cell-cycle progression in case of irregularities, further resuming the cycle when all fluctuation has been corrected or in case of irreversible damage leading to cell death. Cells also cease further proliferation when in contact with other cells in a process

named contact inhibition. But as these genes are called tumour suppressor, it means that their discovery was based on the fact that their inactivation results in tumorigenic tissue, and alas, these safeguard mechanisms can be ultimately evaded by cancer cells (Hanahan & Weinberg, 2011).

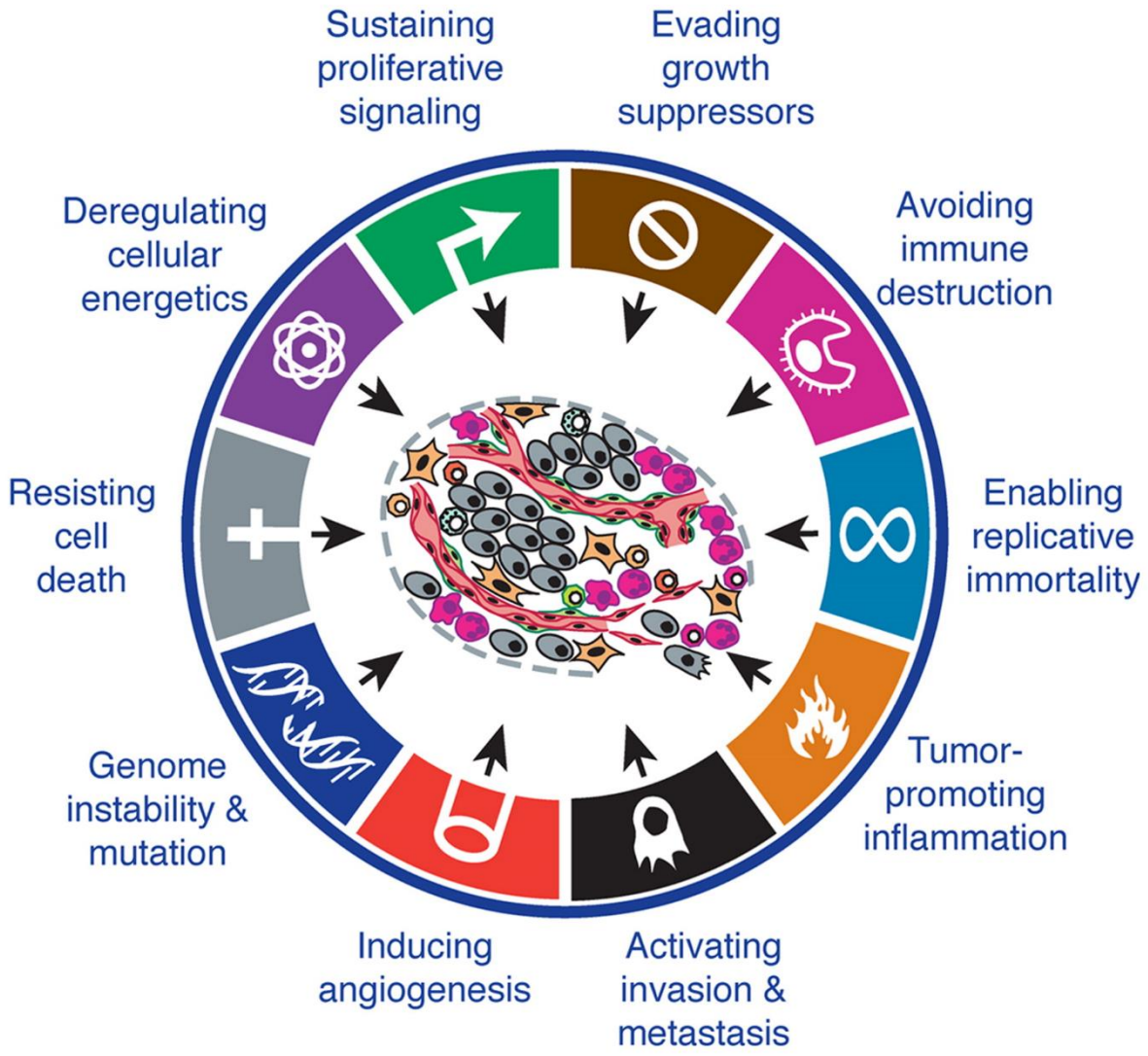

Figure 1.1: The Hallmarks of Cancer and Enabling Characteristics. This image depicts the eight distinct and complementary hallmarks of cancer that define the understanding of cancer mechanisms and its biology, alongside two crucial enabling characteristics: Genome instability & mutation and Tumour-promoting inflammation. Note: Adapted from "Hallmarks of cancer: the next generation" by D. Hanahan & R. Weinberg, Cell. 2011 Mar 4;144(5):646-74 (10.1016/j.cell.2011.02.013). Copyright © 2011 Elsevier Inc. All rights reserved.

Tumour cells can therefore avoid cell death not only by losing tumour suppressor genes, but also by expressing more antiapoptotic signals and conversely downregulating proapoptotic signals, and by degrading the extrinsically activated death pathway, avoiding cell death signals

from surrounding cells. Some tumour cells can also undergo autophagy in extreme stress caused by therapy lying in a dormancy state that can be reversed in the future, possibly originating late-stage tumours (Hanahan & Weinberg, 2011).

Besides avoiding death, cancer cells can also enable replicative immortality, while normal healthy cells have a limited number of replications at their disposal, also known as a senescent number. This number is thought to be related to the telomeres of a chromosome, segments that are present at the ends of all chromosomes being shortened with every cell division, whose main function is to protect the chromosomal integrity. Whilst healthy cells enter crisis and stop dividing when their telomeres are shortened to the point where the chromosome integrity is compromised, cancer cells can reactivate telomerase (Figure 1.2) - a DNA polymerase that lengthens the telomere repeats, a feature characteristic of immortal cell lines such as gametes and stem cells, effectively preventing telomere erosion and consequently avoiding a state of crisis (Hanahan & Weinberg, 2011).

While telomerase is more commonly associated with telomere conservation and elongation composing 85% of tumoral telomere lengthening, cancer cells can also rely on an additional telomere maintenance mechanism to ensure replicative immortality. Alternative lengthening of the telomeres (ALT) accounts for 15% of cancer telomere maintenance, via homologous recombination in the ALT-associated PML nuclear bodies (APB), although the mechanism is not fully understood (Chung et al., 2012). This alternative telomere maintenance pathway allows for the tumour to take advantage of telomere reconstruction and all the proliferative benefits that can be exploited, even while under the effect of a telomerase inhibitor therapy.

With all these mechanisms, cancer cells can effectively avoid death and proliferation restrictions. Nevertheless, they still require nutrients and oxygen for their various metabolic requirements to be met, and a way to dispose of the unneeded products of their metabolism. To address this difficulty, tumours often upregulate the expression of *VEGF* gene to enable angiogenesis in their vicinity, promoting the formation of new blood vessels through a process mainly associated with embryogenesis and tissue trauma. This process creates new atypical and misshaped blood vessels due to the incessant and erratic angiogenesis signals (Hanahan & Weinberg, 2011).

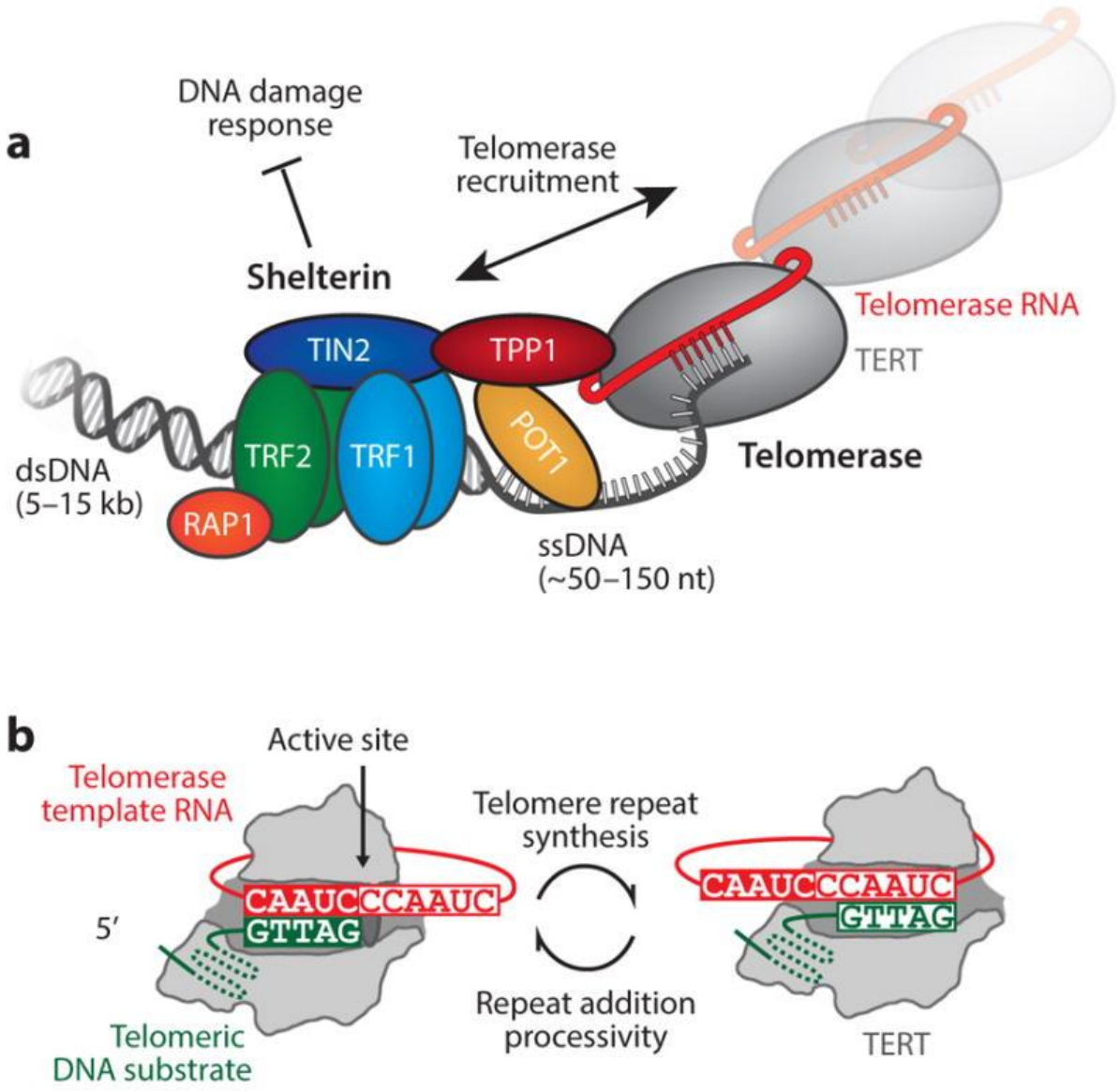

Figure 1.2: Telomerase catalytic cycle at the telomere. (A) To avoid false identification of double-stranded breaks at the telomere the shelterin complex binds via the TRF1/2 homodimers to the double-stranded telomere DNA alongside the recruit of the heterodimer POT1-TPP1. While POT1 binds to the single-stranded guanine-rich tail, TPP1 interacts with telomerase's N-terminal, recruiting the protein to the telomere, where (B) the telomerase ribonucleoprotein complex active site binds to the 3' end of the DNA substrate (in green) and aligns with the RNA template (in red) owing to base pairing, culminating in a hybrid RNA-DNA substrate required by telomerase reverse transcriptase (TERT) to additively elongate the telomere chain. Note: From "Single-Molecule Studies of Telomeres and Telomerase" by Joseph W. Parks & Michael D. Stone, 2017, Annu Rev Biophys. 2017 May 22; 46: 357–377. (10.1146/annurev-biophys-062215-011256). Copyright © 2017 The Authors. All rights reserved.

The formation of new blood vessels also facilitates the invasion of other tissues by the tumour cells. These pre-invasive cancer cells have been observed to gradually stop expressing E-cadherin, a protein involved in cell-to-cell adhesion and quiescence and start expressing pro-

invasive proteins like N-cadherin present in migratory cells, especially during embryogenesis. This process is described as Epithelial-Mesenchymal Transition (EMT) and it allows intravasation of invasive tumour cells into the blood vessels to distally extravasate from their point of origin in other tissues and settle micro-metastases, eventually colonizing new tumours in other parts of the organism (Hanahan & Weinberg, 2011).

Another interesting characteristic that cancer cells acquire, is the reprogramming of their own energy metabolism, by favouring glycolysis instead of oxidative phosphorylation in the mitochondria (even in the presence of oxygen), effectively creating a state of aerobic glycolysis (Figure 1.3). This phenomenon was later coined the Warburg effect in honour of Otto Heinrich Warburg's research (Warburg, 1931) and is somewhat perplexing since aerobic glycolysis is very ineffective at energy production when compared to its counterpart in normal cells. One theory that might explain this behaviour is that cancer cells prioritize glycolysis over aerobic phosphorylation so they would have access to glycolytic intermediates that can be redirected to numerous biosynthetic pathways, such as nucleoside and amino acid generation. This allows effectively trading energy efficiency for much needed building blocks, so to speak, in an unregulated malignant cell (Hanahan & Weinberg, 2011).

Highly immunogenic cancer cells are thought to evade an individual's immune system by secreting immunosuppressive factors such as TGF-β paralyzing immune response cells, as well as recruiting immunosuppressive regulatory T cells to mediate cytotoxic lymphocyte activity. Nevertheless, it has also been observed that these same highly immunogenic cancer cells are regularly eliminated by immunocompetent host through immunoediting. As a result, the less immunogenic variants are the true culprit of colonization in both sides of the spectrum (immunocompetent and immunocompromised) when it comes to immune capability of oneself, even though more research needs to be done to ascertain the mechanism tumours employ to evade the immune system and subsequent destruction (Hanahan & Weinberg, 2011).

The constant activity of the immune system in combating the cancer cells gives arise to a new problem - inflammation, an enabling characteristic that enhance tumour progression and tumorigenesis. These inflammatory cells often potentiate an environment that is highly mutagenic by releasing reactive oxygen species (ROS), growth factors, survival factors, angiogenic factors and overall reshaping the extracellular matrix (ECM). This serves the opposite purpose of allowing the tumour to accumulate even more highly aggressive mutations and paving the way for its metastasis (Hanahan & Weinberg, 2011).

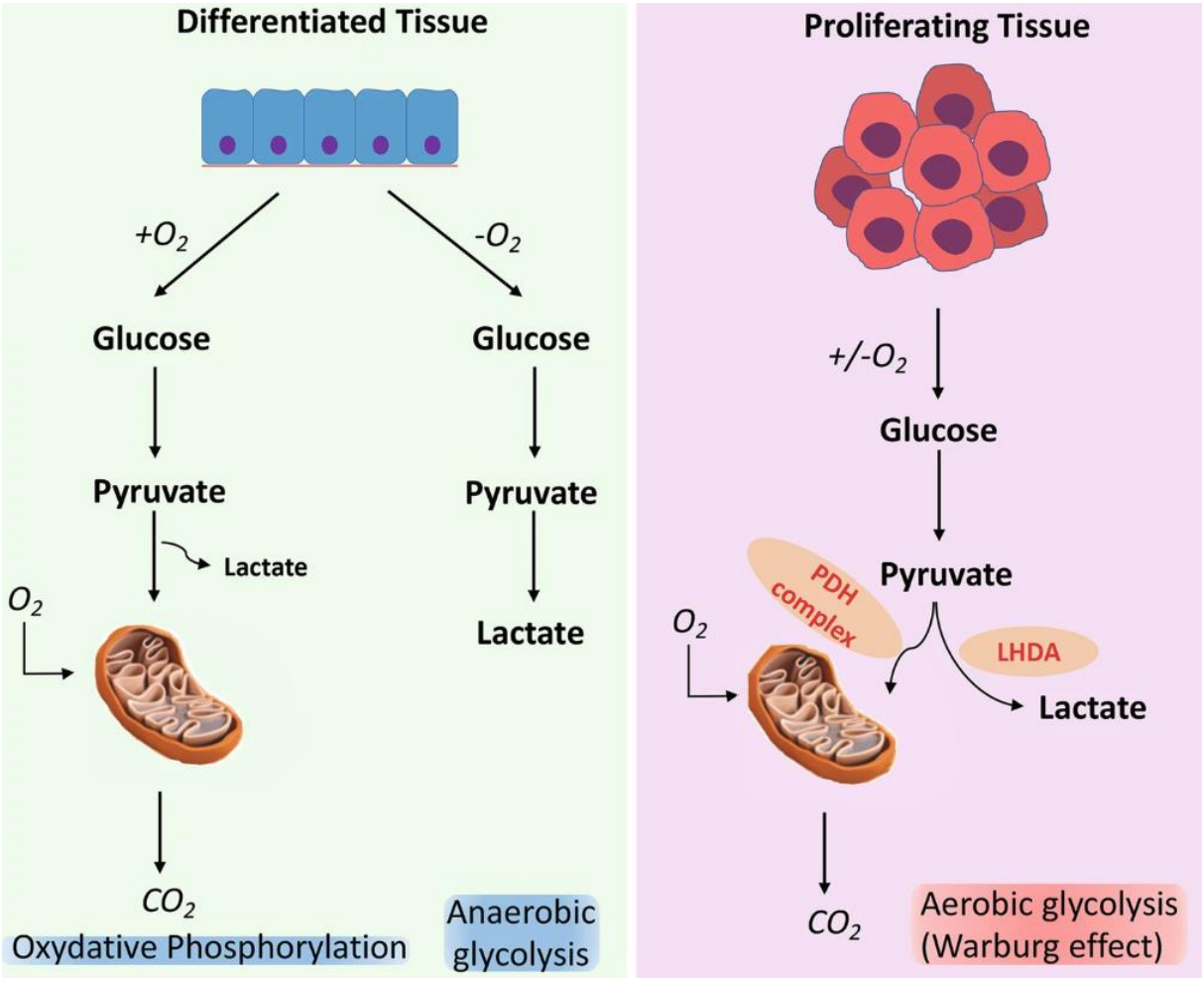

Figure 1.3: Normal cell metabolism and the Warburg Effect. Normal differentiated tissue cell metabolism in both presence (oxidative phosphorylation) and absence (anaerobic glycolysis) of oxygen is represented on the left side, glycolysis even with oxygen availability in proliferating tissue cells, a process named Warburg effect, can be observed in the right side. Note: From "Glucose Metabolism in Cancer: The Warburg Effect and Beyond" by S. Bose, C. Zhang, A. Lee, 2021, "The Heterogeneity of Cancer Metabolism" Second Edition page 5, from Advances in Experimental Medicine and Biology Series Volume 1311.

In its essence, most cancer hallmarks rely on one simple enabling premise: the accumulation of mutations and genome instability. These unstable genotypes often acquire mutations that grant a selective advantage to all the subsequent subclones when compared to neighbouring cells, even though there are failsafe mechanisms to detect and repair DNA damage. These compromised cell lines can become more sensitive to further mutagenic agents accelerating their mutation rate and consequent accumulation of cancer defining hallmarks (Hanahan & Weinberg, 2011).

As the cancer cells are relentless, so are researchers worldwide in a continuous effort to better understand it, and various new hallmarks and enabling characteristics, have been proposed since 2011.

One such is the ability to unlock phenotypic plasticity (Figure 1.4) where a cancer cell that originates from a well-differentiated normal cell can reverse this fully developed state and reacquire a certain state of potency, leading to redifferentiation into another completely distinct cell line than the one that originated it. Phenotypic plasticity can also occur through a process named transdifferentiation where cells in a specific differentiated phenotype, completely change morphologically to resemble cells from a distinct tissue, this ultimately equates cell lines that express other tissue-specific traits than the one they emerged from, compromising the tissue/organ's balance as a result. Progenitor cells can also suffer regulatory changes that can effectively lock them in the progenitor state by not proceeding through the differentiation timeline in a process known as blocked differentiation, effectively retaining the ability to differentiate into a broader selection of cell lines (Hanahan, 2022).

Non-mutational epigenetic reprogramming is also being considered an enabling characteristic of obtaining or facilitating the acquisition of cancer hallmarks by effectively reprogramming the genome epigenetically and ergo not by mutational instability as most frequently associated with cancer. Through methylation, critical segments of the genome can be silenced, especially the ones containing tumour suppressor genes essential in maintaining cell homeostasis. But the opposite can also occur, where silenced segments pertaining to oncogenes contribute to a more proliferative and growth-prone phenotype, one akin to a progenitor state. The methylation instability can also be perpetuated by the hypoxia microenvironments commonly known to occur in many tumours, by reducing the activity of demethylases leading to genome hypermethylation (Hanahan, 2022).

One proposed enabling characteristic investigates the microorganism composition of the microbiome in the context of population dynamics, constituting either a favourable environment for cancer development or the exact opposite and inhibiting its progression. Some bacterial environments can be rich in toxic molecules that could effectively damage the host's own cells, leading to DNA damage and consequently mutate said cells to acquire pro-growth and proliferation mutations.

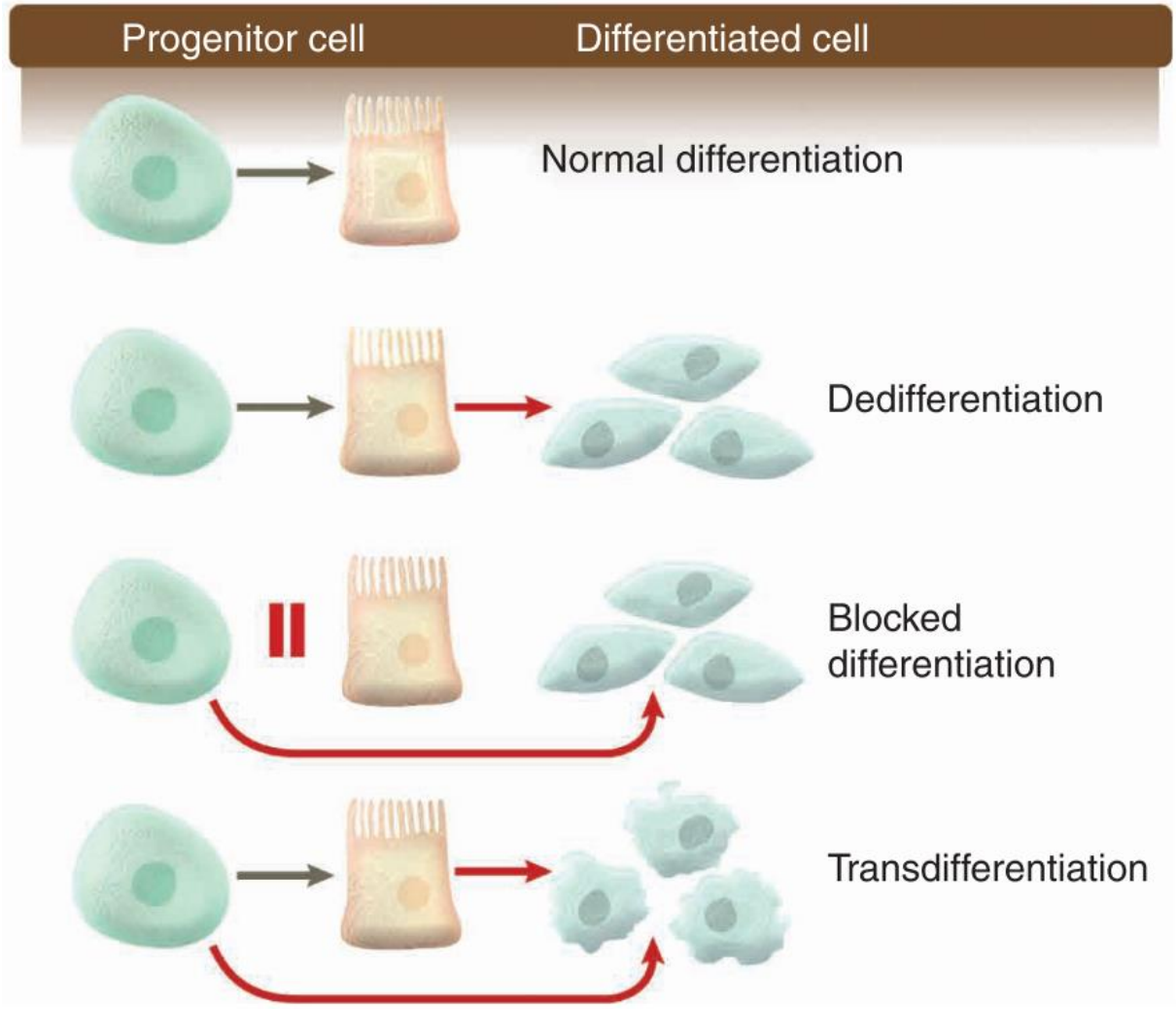

Figure 1.4: Phenotypic Plasticity. This cancer hallmark allows the tumour cells to swerve cell differentiation to their benefit by dedifferentiating back to progenitor states from a mature cell line, stop terminal differentiation of the progenitor cell lines to ensure potency and transdifferentiate a cell line to a completely different cellular lineage unlike the one of tissue origin. Note: From "Hallmarks of Cancer: New Dimensions" by D. Hanahan, Cancer Discov. 2022 Jan;12(1):31-46. (10.1158/2159-8290.CD-21-1059). Copyright © 2022 American Association for Cancer Research. All rights reserved.

Additionally, some bacteria can also bind to the host's cell by mimicking surface ligands, and when certain proliferative receptors are activated, this could result in a proliferation stimulative signal. Unregulated hostile microbiomes could result in tumour promoting inflammation (a distinct hallmark already referred to), as well as silence or actively combat the host's adaptive immune response. Commensal bacteria in homeostasis with the host can also enhance tumoral immune response by blocking T-cell immune checkpoint and enhancing its response (Hanahan, 2022).

The last hallmark proposed, can look quite counterintuitive: senescent cell tumour promotion. As stated in a previous cancer hallmark, avoiding cell senescence via telomerase reactivation, constitutes a useful ability in the cancer's arsenal to further proliferate. Nevertheless, some of these cells display a senescence-associated secretory phenotype (SASP), secreting signals that enhance tumoral growth and proliferation by enabling other cancer hallmarks. Some of these SASP cells can also revert their non-proliferative senescent state to growing cancer cells, being especially useful when avoiding therapeutic approaches that target malignant cells that are engaging in proliferative activity (Hanahan, 2022).

1.2 Breast cancer

1.2.1 Epidemiology

In 2020, there were over 19.3 million new cases of cancer worldwide, claiming more than 9.9 million lives, being female breast cancer the most diagnosed cancer accounting for 11.7% of new cases (2.3 million) (Sung et al., 2021). Although lung cancer remains the leading cause of cancer mortality, female breast cancer is responsible for 6.9% of all cancer related deaths (Sung et al., 2021).

China leads female breast cancer incidence and mortality according to GLOBOCAN 2020 data, while United States of America come second in incidence even though the incidence/mortality ratio is observably smaller than China's. India comes second in mortality worldwide despite having less 70,000 incidence cases than USA in the 2020 dataset (Figure 1.5).

Cancer is the second highest cause of mortality in the United States, and breast cancer is estimated to account for 15% of deaths related to cancer in women, being the leading cause of death by cancer in women ages 20 to 59 years (Siegel et al., 2020). Breast cancer epidemiology also varies ethnically, which can also be observed in the US population, where breast cancer incidence in white women is the highest of any ethnic group even though black women's BC mortality surpasses white women's (Siegel et al., 2020).

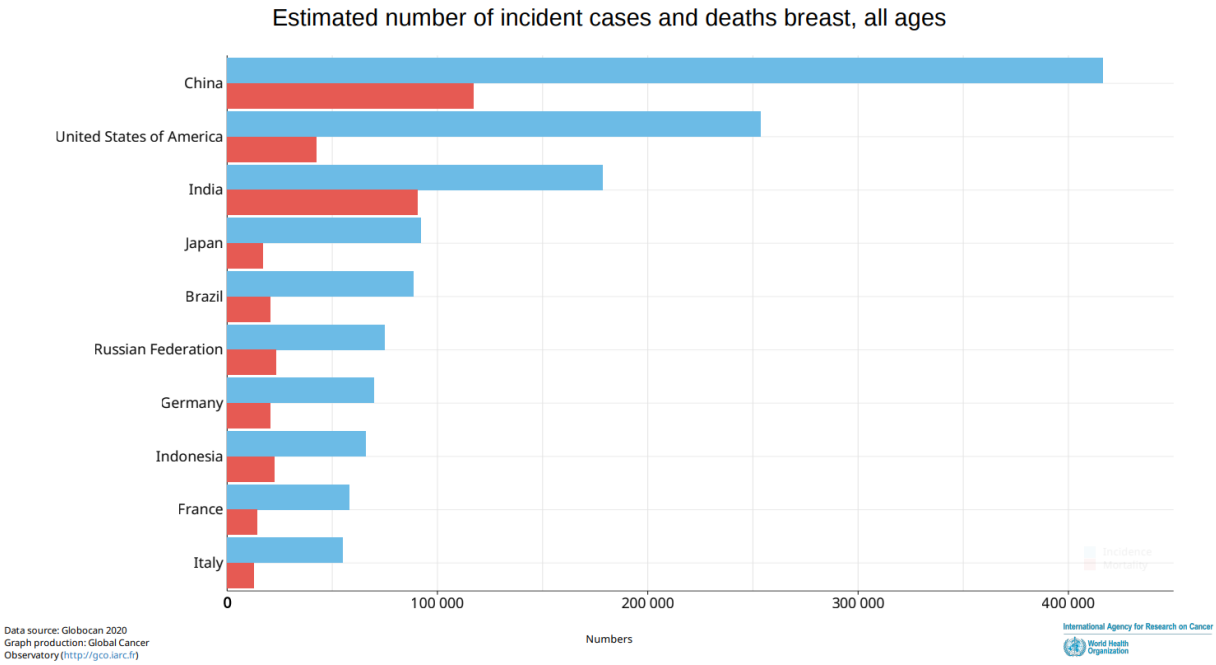

Figure 1.5: Breast cancer incidence and mortality in women of all ages worldwide. This bar chart summarizes the top 10 countries in the world in all age groups according to female breast cancer incidence numbers in blue, as well as display the respective female breast cancer related deaths in red. From: (Sung et al., 2021).

While the incidence gap between these two groups has been shortening, Hispanic women show an overall lower incidence of BC compared to white women although they are usually diagnosed at a later stage (Yedjou et al., 2019). This suggests that genetic background is not the only factor at play, and this inequality may be caused by socioeconomic disparity.

In the European continent, approximately 42 thousand women die of breast cancer every year and it is believed that the screening efforts have already contributed to saving 21680 lives on a yearly basis, equating to a hypothesized 34% breast cancer related mortality prevention. If the screening was extended to 100% of the European population, it is believed 12 thousand more yearly breast cancer deaths could be prevented (Zielonke et al., 2021).

Western European countries (Austria, Belgium, France, Germany, Ireland, Luxemburg, Netherlands, UK and Switzerland) display the highest breast cancer mortality reduction, in part due to the average 61.5% total breast cancer screening coverage, saving around 13 thousand lives (Zielonke et al., 2021).

Whilst northern Europe (Denmark, Estonia, Finland, Iceland, Latvia, Lithuania, Norway and Sweden) has the smallest populational dataset, they have on average 59% screening, the highest organized breast cancer screening (screening organized by national entities to allow for

an equal opportunity of breast cancer screening in a determined population). In contrast, eastern Europe (Bulgaria, Czechia, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia) has the lowest organized breast cancer screening, with an average of 39% (Zielonke et al., 2021).

Opportunistic breast cancer screening (screening performed after a recommendation of a health professional and outside any populational screening efforts) was lowest in northern European countries with an average of 5%, and highest in southern European countries (Cyprus, Greece, Italy, Malta, Portugal and Spain as well as the independent statistic of Gibraltar) with an average of 32% (Zielonke et al., 2021).

In 2020, breast cancer cases in Portugal represented 26.4% of new cancer cases in women, totalling 7041 individual cases and 1864 deaths and representing 15.5% (Figure 1.6) of all female cancer deaths (Sung et al., 2021). Between the years 2002-2013 the Portuguese region most affected by breast cancer deaths was Lisbon with 31.6%, also achieving the highest breast cancer mortality rate (BC-MR) of 33.87/100 000 inhabitants, followed by Alentejo with a BC-MR of 33.60/100 000 inhabitants (Gomes & Nunes, 2020). Lisbon's BC-MR can be justified due to the high participation rate in that region whilst the south lacks an organized screening program (Forjaz de Lacerda et al., 2018). If we analyse the Age Standardized Rate (ASR), the South represents the highest score with 155.8/100 000 cases in women per year, followed by the North with an incidence in women of 132.4/100 000. Estimates indicate however that the North will eventually overtake South's score, since its ASR increased by 3.6% during the 14-year study, compared to the 1.6% ASR increase of the current leader.

Organized breast cancer screening between 50 and 69 years of age in Portuguese women was only 33.8% in 2013, and while population awareness has since been bettered in more recent years, it is quite a low percentage, especially considering that organized breast cancer screening is 10% more effective in preventing breast cancer-related deaths in comparison to opportunistic screening. Additionally, it is believed that a further ~10% breast cancer-related deaths could be avoided with more screening programmes (Zielonke et al., 2021).

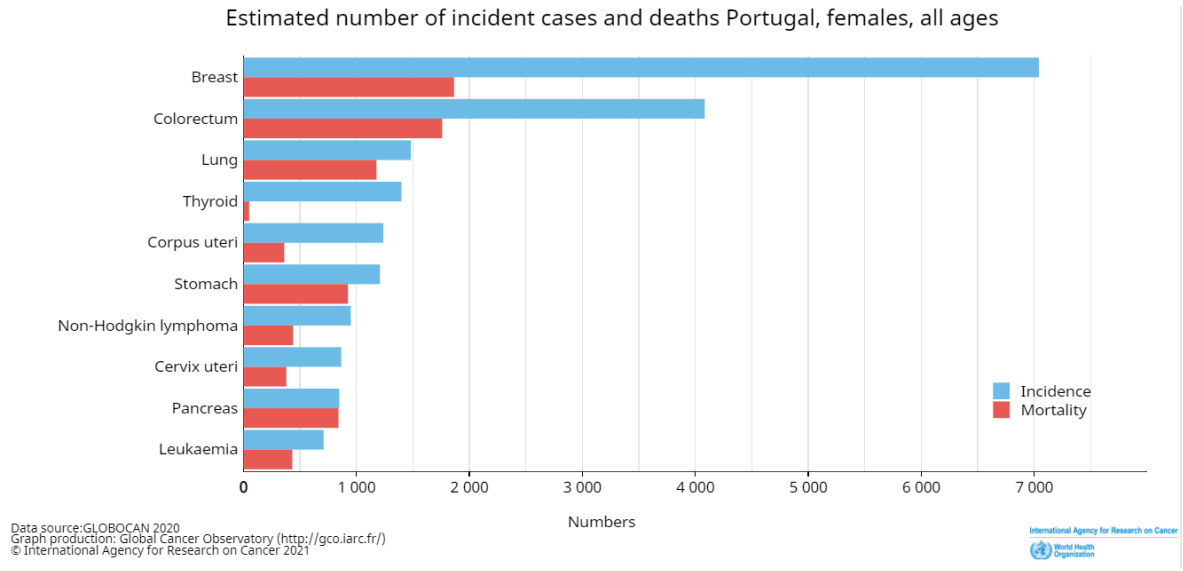

Figure 1.6: Breast cancer incidence and mortality in Portuguese females of all ages. This bar chart encompasses the impact of cancer in the female Portuguese population in all age groups, being measured by both incidence and mortality in numerical fashion. From: (Sung et al., 2021).

1.2.2 Classifying Different Types of Breast Cancer

Breast cancer is composed of various distinct features, both histological and biological, originating different clinical responses.

There are two main histological subtypes of breast tumours: *in situ* carcinoma and invasive carcinoma (Weigelt et al., 2010). *In situ* carcinomas can be either lobular carcinoma *in situ* (LCIS) (1.8 - 2.5% of all breast biopsies) (Wen & Brogi, 2018) or ductal carcinoma *in situ* (DCIS), the latter being the most common type and can be further subclassified as five histologically distinct types: Comedo, Cribriform, Micropapillary, Papillary and Solid. On the other hand, invasive carcinomas can be categorized as Infiltrating Ductal, Invasive Lobular, Mucinous, Tubular, Papillary and Medullary (Figure 1.7).

Ductal carcinomas *in* situ can be represented in 5 distinct formations, one of which is classified as Comedo (Figure 1.8 A) being characterized by irregular cellular size and shape as well as necrosis in the tissue interior (Ajisaka et al., 2002).

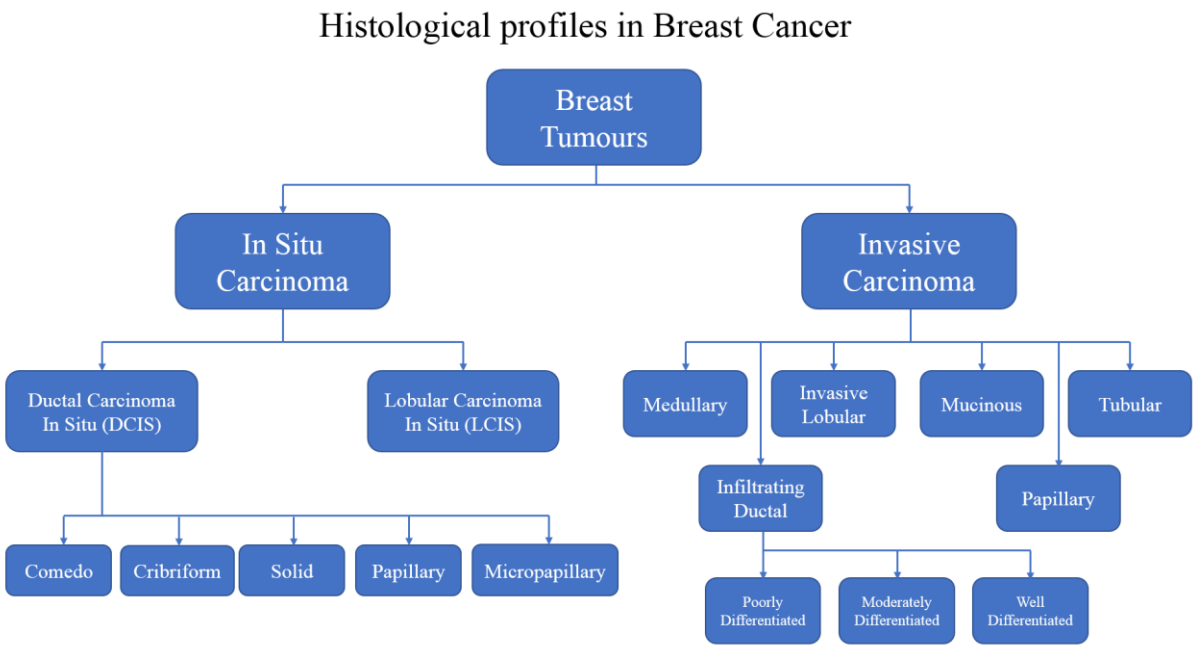

Figure 1.7: Breast Cancer Histological Subtypes. This chart stratifies the different breast cancer subtypes and their subsequent categorical subdivisions. Adapted from (Malhotra et al., 2010).

Cribriform breast carcinomas (Figure 1.8 B), as the name suggests, are identified by their signature multiples puncture-like spaces inside the mammary duct (Ajisaka et al., 2002) and are low-grade ductal carcinomas *in situ*. However, as of the third edition of the World Health Organization classification of breast tumours, cribriform carcinoma was defined as an invasive type (Cserni, 2020), being further redefined in the fifth edition as needing at least 90% cribriform constitution, grade 1, and a tubule formation score of 1 to be classified as invasive cribriform carcinoma (Demir et al., 2021).

With an incidence of up to 3.5% this type of carcinoma frequently expresses both estrogen and progesterone hormonal receptors in contrast to HER2. It has a better prognosis than invasive ductal carcinomas and its pure form (> 90% cribriform formation) also correlates to a better clinical outcome in contrast to mixed invasive cribriform carcinomas (Demir et al., 2021).

Solid DCIS formation (Figure 1.8 C) is represented by a mass of tumour cells that lack tubular formations granting it the solidified mass aspect.

Papillary carcinomas (Figure 1.8 D) are identified by their trademark papillae shaped projections inside the mammary duct composed of fibrovascular cores surrounded by the neoplastic cells (Cserni, 2020).

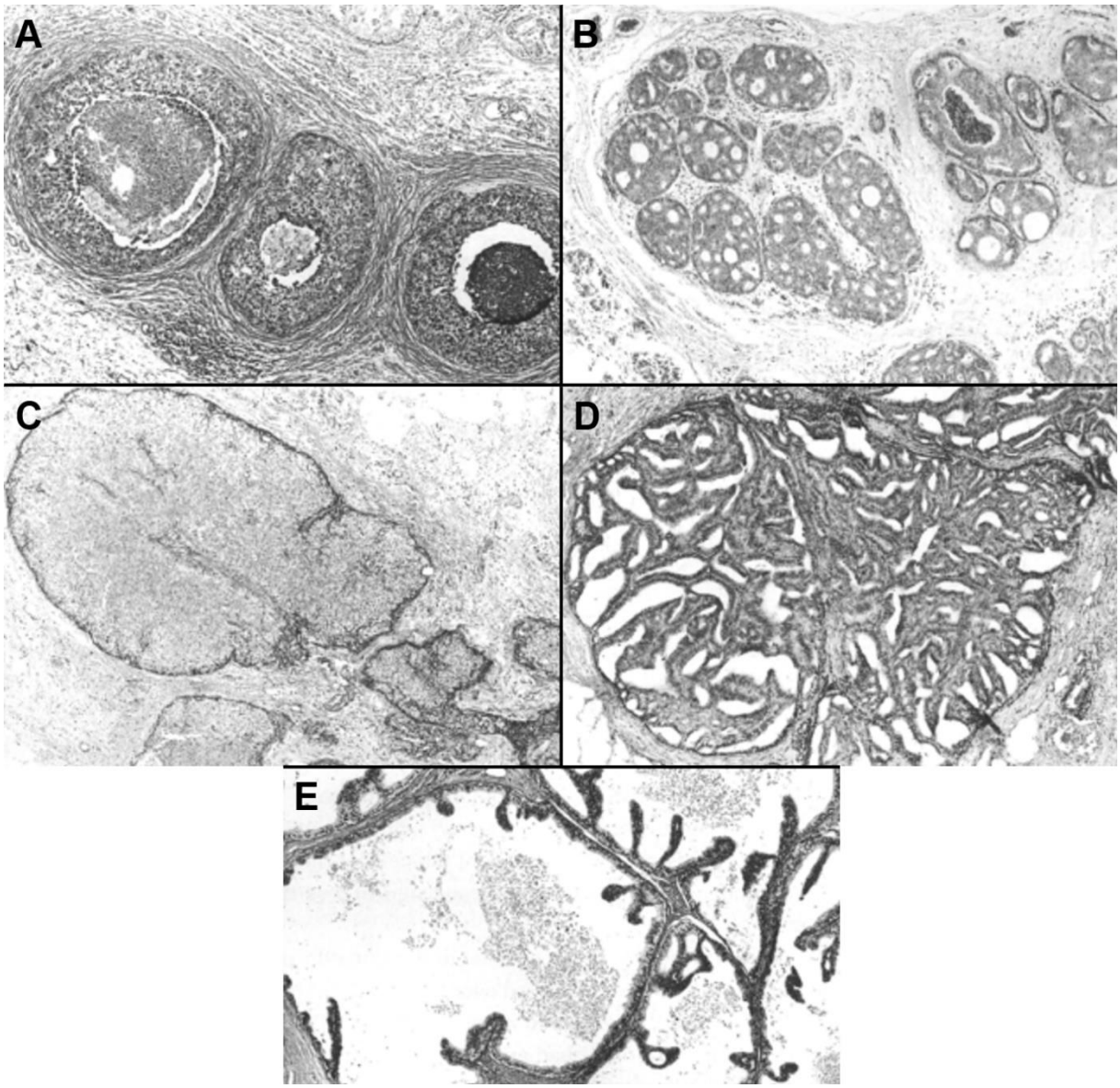

Figure 1.8: Ductal Carcinoma In Situ subtypes. In panel A we can verify the presence of the Comedo architecture with central necrosis, on B Cribriform with the typical round spaces, on C there is a distinct lack of tubular formations characteristic of the Solid pattern, D panel shows Papillary features and E corresponds the arch Micropapillary projections pattern. Haematoxylin and Eosin staining 40x magnification. Adapted from (Ajisaka et al., 2002).

They account for 0.5 – 1% of all breast cancers and are further subdivided in invasive and non-invasive. The non-invasive branch harbour papillary ductal carcinoma *in situ* which is a subtype of DCIS that usually coexists with other DCIS histological categories, being considered more of a pattern of growth (Cserni, 2020). Papillary DCIS is known to have palpable masses as well as nipple discharge, it has also a higher risk of recurrence than other

DCIS subfamilies and, therefore, mastectomy as a therapy is defended by authors in the field (Nuñez et al., 2020).

Invasive papillary carcinoma is a rare histological subtype that comprises 90% of the architecture of an invasive tumour. They present a luminal A expression profile and are more common in postmenopausal women identifiable by a palpable mass usually beneath the nipple and/or bloody discharge (Nuñez et al., 2020). The carcinoma histology shows papillae with fibrovascular cores without myoepithelial cells (Cserni, 2020).

Micropapillary (Figure 1.8 E) carcinomas in contrary to their nomenclature, are not smaller papillae structures than the previous category, it is instead the presence of epithelial outgrowths without the fibrovascular cores (Cserni, 2020). Micropapillary DCIS constitutes 6% of all diagnosed ductal carcinomas *in situ* and represent a risk to local microinvasion of the mammary duct, especially the tumours with a higher grade that express more HER2 receptor and show a higher proliferation rate (Castellano et al., 2009). The invasive variant comprises 2% of all invasive breast cancers, positively expressing ER, PR and HER2 they are associated with gain of 8q, 17q and 20q and deletions in 6q and 13q in addition to *MYC* and *CCND1* amplifications and *PIK3CA*, *TP53* and *GATA3* mutations (Jenkins et al., 2021). Patients usually show axillary lymph node metastasis, constituting a valuable prognosis predictor of this subtype alongside 50% micropapillary structure of the tumour. Mortality rates are usually 25% with a 70% of relapse (Nuñez et al., 2020).

Lobular carcinoma *in situ* is usually found while collecting biopsies from other lesions in mammary tissues and are hard to estimate their overall incidence, so the available metrics show a 2% presence in all breast biopsies and an incidence of 3.19/100000 women-years between the period of 1996 – 1998 (Wen & Brogi, 2018). It mainly manifests in late premenopausal women (~50 years of age) and can be further subdivided in three different forms:

- Classic LCIS (Figure 1.9 A) cells are oval as well as their nuclei, they are non-polarized and proliferate in a non-cohesive manner. Cytoplasm-wise, the cells have either minimal cytoplasm volume being identified as “type A” cells (Figure 1.9 B), or plentiful cytoplasmatic volume and categorized as “type B” cells (Figure 1.9 C) which are significantly bigger and more abundant in the classic subcategory. The hormonal receptor status (ER/PR) is mainly positive in the subtype with HER2 being negative (Wen & Brogi, 2018).

- Pleomorphic LCIS (P-LCIS) (Figure 1.9 D) are mainly observed in postmenopausal women, while presenting a solid growth pattern, nuclear pleomorphism and tissue necrosis this LCIS subtype can be mistaken by ductal carcinoma *in situ*, however unlike DCIS, pleomorphic LCIS are non-polar, architecturally misshapen and don't form secondary lumina, very mitotically active with a distinctive non-cohesive proliferation such as classic LCIS. Some authors subdivide P-LCIS in apocrine which display an acidophilic cytoplasm and are binuclear besides having a lower expression of ER/PR and in a few cases, overexpress HER2 in contrast to their non-apocrine counterparts. (Wen & Brogi, 2018).

- LCIS with necrosis also known as "florid" (Figure 1.9 E), is similar to classic LCIS in the cellular component, showing both type A and B cellular subtypes in its composition, although in contrary to the classic subtype, florid LCIS exhibits central necrosis bearing calcification and massive mammary acinar expansion (Wen & Brogi, 2018).

LCIS can be histologically similar to DCIS and to prevent a false categorization it is advised to recur to immunohistochemistry to verify a diagnosis, since there are critical differences between both tissues' expressed molecules. LCIS categories have been observed expressing aberrant and often not expressing E-cadherin comparing to DCIS that show membrane staining of this proliferation reducing protein as well as the p120 tyrosine kinase substrate that interacts with E-cadherin in neighbouring cell adhesion, who through staining can be mainly observed in the cytoplasm in LCIS, considering there is no functional E-cadherin in these subtypes (Wen & Brogi, 2018).

The lobular definition also extends to the invasive panorama, invasive lobular carcinoma shares morphological similarities to LCIS as well as the absence or aberrant expression of E-cadherin (Cserni, 2020). Mutations in *CDH1* are usually reported in this subtype alongside loss of 16q gain of 1q and amplifications in *CCND1* and *FGFR1* loci and the majority exhibit a luminal A expression pattern (ER+/PR+/HER2-) (Reed et al., 2021). This subtype is particularly common with 15% incidence of all cases, commonly observed to infiltrate the stroma in small numbers, positioning in a concentric manner around the duct with reduced rearrangement of the normal tissue (A. E. McCart Reed et al., 2015).

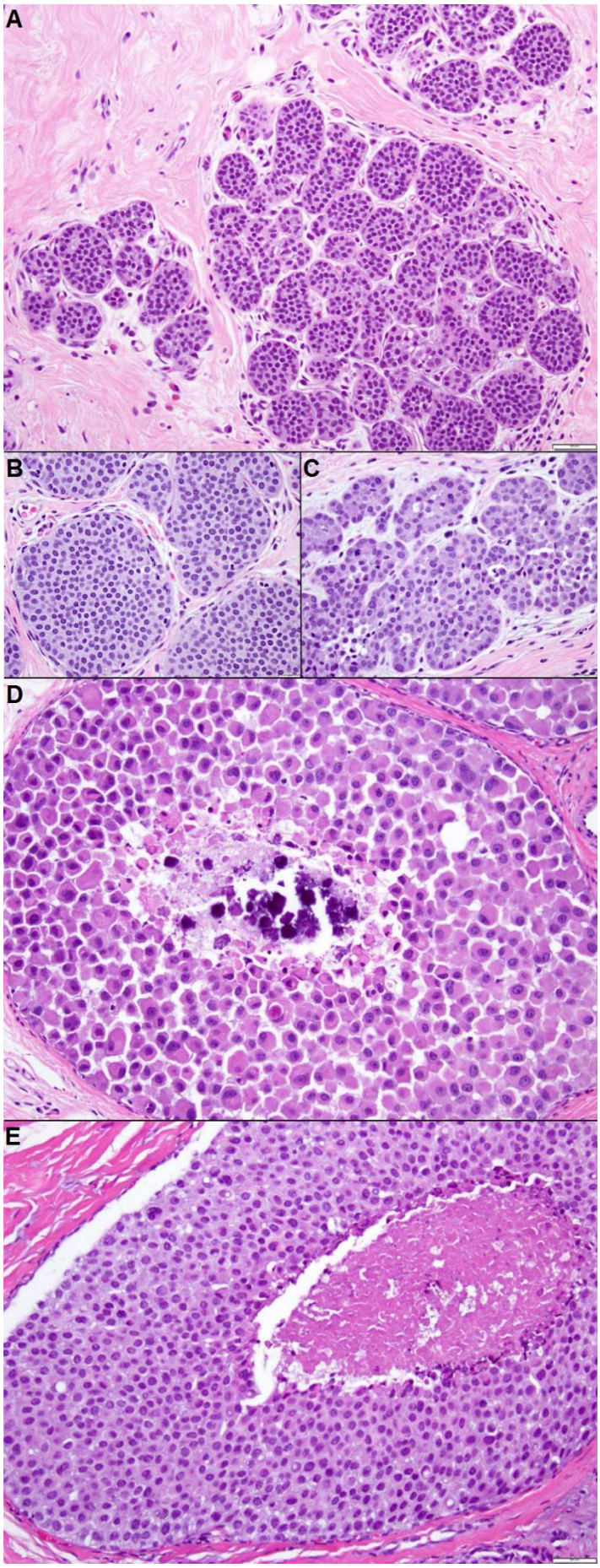

Figure 1.9: Lobular Carcinoma in Situ (LCIS) and it's subtypes. The Classical Lobular Carcinoma in Situ **(A, 200x magnification)** is characterized by dysheshive cells with few nuclear morphological abnormalities, and they still conserve a certain degree of spatial regularity. This classical subtype can be further deconstructed by the type of cells it is constituted:

- The smaller cells known as "Type A" cells **(B, 400x magnification)**, show dysheshive cells with non-polarized oval nuclei with a regular nuclear membrane and minimal cytoplasmic volume.
- The larger cells relative to the previous type are accordingly identified as "Type B" cells **(C, 400x magnification)** due to its larger uniformized nuclei and increased cytoplasmatic volume.

On the panel below **(D, 200x magnification)**, a Pleomorphic LCIS representation can be observed. This LCIS subtype is associated with oval cytoplasmic abundant, eccentric, and irregular nuclei cells whose chromatin is commonly observed in a coarse state. Pleomorphic LCIS is often associated with necrosis and calcification, which can in fact be perceived in this panel.

LCIS with necrosis also known as "florid" **(E, 200x magnification)** is indiscernible from Classic LCIS, even conserving the same cellular morphologies (Type A and Type B) being distinguished by massive acinar expansion and central necrosis. Adapted from (Wen & Brogi, 2018).

Mucinous breast carcinomas more commonly present in post-menopausal women, account for 3% of primary breast cancers. They are characterized by the increased presence of extracellular mucin where the malignant cell aggregates tend to drift. They are usually low-grade tumours as high-grade mucin producing carcinomas are classified as infiltrating ductal carcinoma. Carcinomas of the mucinous variety are usually ER and PR positive and therefore have a good prognosis as well as lacking major genetic and molecular instability when compared to the more aggressive counterparts. And, as a result, the most indicated therapy for these carcinomas is a mammary tissue conservative surgical removal rather than the more aggressive chemotherapy (Jenkins et al., 2021).

Tubular breast carcinomas account for 2% of invasive breast carcinomas and are mainly identifiable by their tubular structure around the milk duct of the breast, the point of its genesis. Like the previous carcinoma, tubular breast carcinomas also have high expression of both estrogen and progesterone receptors, usually following expression patterns of the luminal A subtype. When paired with the fact that they usually show a low capability of producing metastasis, and are mainly prevalent in post-menopausal women, this sub-family of carcinomas usually display a good prognosis, although they show genetic instability by loss of 16q, 8p and 3p while on the other hand gain of 1q, 16p and 11q (Jenkins et al., 2021).

Medullary breast carcinomas have a sizeable association with triple-negative breast cancer subtype (up to 17%) and germline mutations in the *BRCA1* gene (around 19%), with major incidence in premenopausal women with a mean age of 50 years with common bilateral manifestations in the case of familial propensity to breast cancer (Jenkins et al., 2021). Histologically, it is defined as a well-circumscribed carcinoma arranged in a large non-glandular sheet structure of poorly differentiated cells with minimal stroma and are usually identified as oval shaped masses (Jeong et al., 2012).

Infiltrating Ductal Carcinoma (IDC) is the most common variation of invasive breast carcinomas accounting for 70-80% of invasive cases. It is comprised of 3 distinct grades that indicate the level of differentiation, grade 1 being well differentiated and grade 3 corresponding to poorly differentiated (Malhotra et al., 2010).

1.2.3 Molecular Classification

Cancer is not a homogenous disease, and breast cancer is no different. No tumour is exactly the same as another, they are unique cases that behave and are composed of distinct

factors that are key to better understanding and treating them. This unlikeness is also dictated by a plethora of molecular features these tumours acquire, such as the expression of certain membrane receptors and/or lack of another.

The estrogen receptor can be subdivided into two different types, Erβ and Erα, the latter being more evidently associated with breast cancer development by being expressed in that same tissue and whose overstimulation can lead to tumoral growth (Wong et al., 2012). Tumours can benefit from the estrogen pathway by employing the hormonal activity inherent to this receptor in order to proliferate and stimulate the growth of neoplastic cells, as well as genotoxicity through cytochrome P450 mediated metabolism (Russo & Russo, 2006).

And similar to the estrogen receptor, the progesterone receptor has two predominant isoforms, PR-A and PR-B, which are both transcribed from the PGR gene with the main difference being the promoter that is implicated in the transcription, resulting in a 94 kDa molecule in the case of PR-A and 114 kDa PR-B, whose amino-acid increase is crucial in generating a transactivation effect that allows the activation specific target genes by PR-B and not PR-A (Trabert et al., 2020). In the case of a ligand activating the progesterone receptor, this molecule will now bind to DNA and regulate the transcription of proliferative genes such as growth factors and even ER in breast tissue (Trabert et al., 2020).

Both estrogen and progesterone receptors share a common purpose, to lead RNA pol II to genomic regions where their target genes reside and, in a normal setting, they ensure the orderly formation of the breast tissue and structures, mechanisms that can also be beneficial in propagating malignant cells. As a result, the analysis of this dichotomy's web of influence can help predict tumoral behaviour (Hilton et al., 2018). Hormone receptor (HR) positivity (ER+/PR+) is related to a 40% increase in disease-free survivability when compared to HR-negative (Prat et al., 2015).

Hormone receptors such as estrogen receptor (ER) and progesterone receptor (PR) are female ovarian hormones that play a critical role in breast development and function during and after puberty by regulating gene expression in breast tissue. As a result, they originate the complex structures present in the breast tissue, be it lobule development as well as creation of ductal passages.

It has also been observed that 70% to 80% of breast tumours express estrogen receptors, progesterone receptors, or even both (Hilton et al., 2018). This demonstrates the clear benefit these hormone receptors bring to the table especially for tumoral growth and proliferation. This is particularly important when they become hypersensitive to progesterone and estrogen growth

factors as discussed before. Albeit the presence of these hormonal receptors (positive status) in the cancer cells constitutes a good prognosis regarding breast cancer outcomes, due to the fact that malignant cells are subjected to receptor signalling to activate their downstream components (ultimately resulting in the regulation of growth associated genes as well as tumour suppressor genes), these receptors can also be targeted therapeutically to hinder tumour development.

On the other side of the coin, breast cancer cells that are hormone receptor-negative, may already have intrinsically activated these pathways (as previously mentioned) and as a result, no longer require receptor signalling to influence downstream gene regulation, eliciting a worse prognosis hence target therapy becomes much more difficult. Human epidermal growth factor receptor 2 (HER2) constitutes another receptor monitored in order to gage breast tumour prognosis, constituting another possible therapeutic target to prevent growth hormone dependnce by breast cancer cells (Hilton et al., 2018).

HER2 are receptors implicated in cell-to-cell communication as well as cell-to-stroma interactions via signal transduction that propagates external signalling to proliferative pathways that ultimately benefit tumoral growth such as PI3K/Akt and Ras. Amplification of the ERBB2 gene resulting in the overexpression of HER2 (also shown as HER2+), highly correlates to a worse breast cancer prognosis and is therefore one biomarker that can be used to better classify the tumour (Ross et al., 2003).

Molecularly speaking, breast tumours can be split into 5 major subtypes (Table 1.1) (Perou et al., 2000; Prat & Perou, 2011; Sørlie et al., 2001):

- Basal-like, with low expression of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor 2 (HER2), commonly called triple negative (ER-/PR-/HER2-);
- HER2-Enriched, suggestively overexpressing HER2 (HER2+) and consequently, an increase in transcribed proliferation genes downstream of this receptor (Prat et al., 2015);
- Luminal A with increased expression of ER and PR, without overexpression of HER2 (ER+/PR+/HER2-);
- Luminal B with increased expression of PR but less than Luminal A, and with a slight increase of HER2 expression (ER+, PR-, HER2+);

- Normal-like subtype, also being triple negative (ER-/PR-/HER2-) but unlike the basal-like subtype, these usually have a better prognosis and much slower growth and are therefore smaller than basal-like in addition to different genetic expression particularly in CK5/6 and EGFR, although genetic expression profiling allows for better differentiation between both subtypes since the normal-like subtype's genetic expression profile is more akin to normal breast stroma cells, hence the name (Z. Li et al., 2015).

There is an additional subtype of breast tumour named Claudin-low, distinguishable by the reduced expression of tight junction genes, such as Claudins, E-cadherin and Occludin, that result in an infiltrative phenotype (Herschkowitz et al., 2007). They are clinically represented as triple-negative (ER-/PR-/HER2-) invasive ductal carcinomas with propensity for epithelial-mesenchymal transition and therefore show a poor prognosis (Prat et al., 2010). A recent study explores the possibility of Claudin-low not being a breast cancer subtype but instead a phenotype, and the categorization of triple-negative variant is questionable due to the observation of estrogen receptor expression in Claudin-low, and therefore, not fitting the negative receptor criteria (Fougner et al., 2020). Furthermore, characteristic Claudin-low features such as low genomic instability and stromal infiltration are also observed in the other 5 breast cancer subtypes, indicating a potential binary system where the other profiles can be considered Claudin-low or non-Claudin-low in addition to their classification, or even a continuous grade of Claudin-low "likeness" that would also function additively and independently of the 5 perceived main breast cancer subtype classifications (Fougner et al., 2020).

To better identify the type of breast tumour in a more streamlined fashion, PAM50 was developed. It consists of a score integrating gene expression from 50 genes that allows molecular subtyping (Table 1.2), permitting an easier classification (Bernard et al., 2009).

PAM50 gene signature showed that, unlike Luminal A, the Luminal B subtype has a higher expression of genes responsible for cell cycle regulation such as *MKI67* and immune response like *IL-2* receptor and *CD86*, explaining the faster proliferation rates observed in Luminal B (Prat et al., 2013). The Luminal A subtype was characterized by a slightly higher PR expression and lower overall tumour development (tumour stage either T0 or T1 and histological grade 1), no significant differences were detected between both subtypes when it comes to ER expression (Prat et al., 2013).

PAM50 gene signature also showed that HER2-Enriched subtype is defined by the overexpression of proliferation genes in the 17q amplicon related to HER2 (*ERBB2/HER2* and *GRB7*), usually having mutated *TP53* (70% - 75% cases) and *PIK3CA* (40% cases) (Godoy-Ortiz et al., 2019).

Table 1.1: Breast Cancer Molecular Portraits According to Clinical Features. This table is composed of three different microarray datasets from: University Of North Carolina (UNC), Nederlands Kanker Instituut (NKI) and M.D. Anderson Cancer Center (MDACC). Clinical features consist in patient number, molecular trait prevalence in said microarray dataset, ER/PR/HER2 receptor statuses and combinations, lymph node-negative tumours, high histological grade tumours, tumours whose size surpass the 2 cm mark and pathological complete response (Pcr) rate.

A	Claudin-low			Basal-like			HER2-enriched			Luminal B			Luminal A			Normal-like		
	UNC	NKI	MDACC	UNC	NKI	MDACC	UNC	NKI	MDACC	UNC	NKI	MDACC	UNC	NKI	MDACC	UNC	NKI	MDACC
Num. Patients	37	21	18	73	42	15	39	49	28	62	69	27	99	84	37	10	30	8
Prevalence	12%	7%	14%	23%	14%	11%	12%	17%	21%	19%	23%	20%	31%	28%	28%	3%	10%	6%
ER+	12%	33%	22%	11%	19%	0%	36%	59%	29%	91%	100%	96%	91%	100%	97%	44%	93%	100%
PR+	23%	-	22%	6%	-	13%	30%	-	25%	53%	-	41%	74%	-	70%	22%	-	63%
HER2+	22%	-	6%	9%	-	13%	66%	-	71%	24%	-	15%	8%	-	11%	67%	-	25%
HER2-/ER-	70%	-	72%	82%	-	87%	25%	-	18%	8%	-	4%	6%	-	3%	13%	-	0%
HER2-/ER-/PR-	71%	-	61%	80%	-	73%	22%	-	14%	9%	-	4%	4%	-	3%	0%	-	0%
Node-	58%	48%	28%	63%	60%	20%	26%	47%	21%	44%	42%	33%	51%	58%	41%	33%	50%	25%
Grade 3	77%	38%	61%	88%	86%	93%	55%	61%	89%	62%	41%	46%	30%	13%	27%	63%	20%	50%
Tumor size > 2 cm	74%	38%	78%	77%	62%	80%	93%	57%	79%	85%	52%	96%	66%	36%	91%	89%	40%	88%
pCR	-	-	39%	-	-	73%	-	-	39%	-	-	19%	-	-	0%	-	-	0%

Note: From "Deconstructing the molecular portraits of breast cancer", by A. Prat & C. M. Perou, Mol Oncol. 2011 Feb; 5(1): 5–23 (10.1016/j.molonc.2010.11.003). Copyright © 2011 The Authors. Published by FEBS Press and John Wiley & Sons Ltd.

Despite having a lower expression of HER2, PR and ER, about 25% of basal-like breast carcinomas are not classified as triple-negative. It is the high expression of a specific cluster of genes, known as the basal cluster, that makes this subtype unique. Some of these genes include cytokeratins (CK 5, 6 ,14, 17), usually expressed in skin basal epithelial cells, hence the name (Perou & Borresen-Dale, 2011). Additionally, about 20% of basal-like breast carcinomas show germline or somatic mutations of either *BRCA1* or *BRCA2* (Muendlein et al., 2015; Villarreal-Garza et al., 2015), 80% show a high frequency of *TP53* mutations, followed by mutations at *PIK3CA* at 9% of the tumours (Koboldt et al., 2012).

Table 1.2: PAM50 Genetic Expression Profiles of the different breast cancer subtypes. This table provides additional information 6 breast cancer subtypes by including their commonly reported histological grade, KI67 protein expression levels used as a proliferation marker as well as genes expression fluctuations in PAM50 curated genes. Adapted from (Bernhardt et al., 2016).

Breast Cancer Subtype	Histological Grade	KI67	Genetic Expression Profile
Luminal A	Low	Low	Increased expression in ER regulated genes: *XBP1, BCL2, EsR1, ZIP6, PgR, SLC39A, FOXA1, LIV1, ERBB3, ERBB4, GATA3* and *TFF1*.
Luminal B	Medium	Average	Increased expression in ER regulated genes: *BCL2, FOXA1, PgR, GATA3, EsR1*. As well as increased expression in proliferation enabling genes: *CCNB1, CCND1, CCNE1, v-MYB* and *MYBL2.*
HER2-Enriched	Medium-High	High	*ERBB2* and *GRB7* amplification. Loss of *INPP4B* and *PTEN*, leading to PI3K pathway activation. Increased expression in proliferative the genes: *CCND1*, *CCNE1*, *MYBL2, ORC6L* and *BIRC5.*
Basal-like	High	High	Dysregulation of the following pathways: MAPK/AKT/PI3K, Ras/Raf and JAK/STAT. Increased expression of: *EGFR, cMYC, CCND1, CCNE1, FOXM1, CDC6, CDC20, ORC6L* and *BIRC5.*
Cloudin-low	-	Low	Loss of *claudins 3,4,7, CDH1 and E-cadherin,* tight junction proteins that are a known marker of non-proliferative cells. Expression of pro Epithelial-Mesenchymal Transition markers *SNAI1/2, TWIST1/2* and *ZEB2,* culminating in highly infiltrative behaviour.
Normal-like	Medium-High	High	Loss of *claudins 3,4,7 and E-cadherin,* tight junction proteins that are a known marker of non-proliferative cells as well as no detectable expression of *CK5 and EGFR* genes.

When it comes to incidence, Luminal A is the most common subtype with a frequency of 40% while its counterpart Luminal B reaches 10% of the total cases, basal-like is the second most frequent subtype with 25% incidence, leaving HER2-Enriched with a frequency of around 7% (remaining 18% were unclassified). Remarkably, it was observed a higher rate of basal-like cell carcinoma in African American population when compared to Non-African Americans, effectively doubling the number of cases (Carey et al., 2006; Perou & Borresen-Dale, 2011).

1.2.4 Aetiology

Breast cancer incidence can be affected by numerous variables, some of them environmental and behavioural, but also more intrinsic and inheritable (Figure 1.10).

Age at menarche is one such factor, being later age at menarche associated with a lower incidence of Triple Negative Breast Cancer (TNBC) (Dolle et al., 2009; Tamimi et al., 2012; Yang et al., 2007). The last event in women's fertile life - menopause, also influences breast cancer risk, being associated with increasing Luminal A and possibly TNBC incidence (Pollán et al., 2013; Tamimi et al., 2012; Xing et al., 2010; Yang et al., 2007; Reviewed in Barnard et al., 2015).

Behavioural risk relies solely on one's routines and decisions, making them theoretically avoidable even though they contribute to a myriad of pathologies, and breast cancer is no exception.

High Body Mass Index seems to have a protective effect on premenopausal women when it comes to Luminal A subtype, mainly in young adulthood, but an increased risk when it comes to the much more aggressive subtype TNBC. Concerning postmenopausal women, an increased BMI doesn't correlate with increased risk in any of the subtypes nor does it provide a protective factor (Gaudet et al., 2011; Millikan et al., 2008; Yang et al., 2007).

Sedentary life and intake of high-caloric foods lead to obesity, a state of increased adipose tissue inflammation, and a microenvironment where breast tumoral cells thrive (De Cicco et al., 2019).

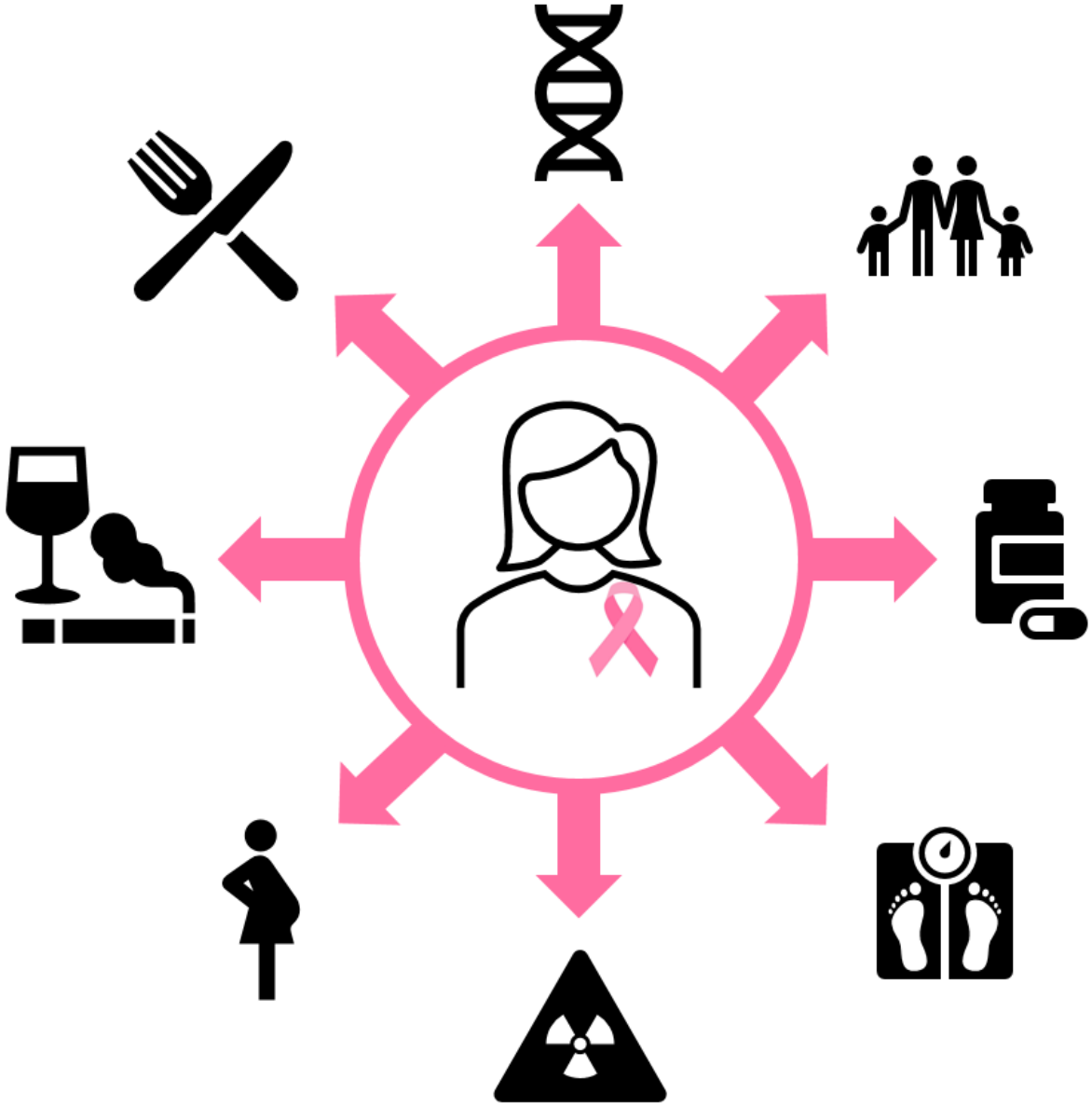

Figure 1.10: Breast Cancer Risk Factors. The risk of developing breast cancer has been correlated to a handful of risk factors, starting at the top and going clockwise: genetics and familial inheritance which are uncontrollable factors and the more lifestyle-oriented factors consumption of exogenous female hormones, health factors such as exercise and body mass index (BMI), exposure to ionizing radiation and mutagenic compounds, pregnancy and parity, alcohol and tobacco, and lastly diet.

Physical activity has been reported to be inversely proportional to breast cancer incidence providing a protective factor with an overall relative risk (ORR) of 0.87. While it slightly fluctuates between pre-menopausal (0.83 ORR) and post-menopausal (0.91 ORR) women, consistent physical activity is evidently associated with a lower breast cancer incidence, and women with a lifelong physical activity record show an overall relative risk of 0.81 (Xuyu Chen et al., 2019).

Dietary habits also influence breast cancer incidence: fruit and vegetable intake has been deemed a protective factor against BC with a relative risk of 0.87. The mechanism underlying said protection can be attributed to the antioxidant effect of the vitamin C and beta-carotene, effectively neutralizing reactive oxygen species and as a result mitigating DNA damage (Poorolajal et al., 2021).

The Mediterranean diet is especially beneficial to breast cancer prevention, in part due to its variety of both polyphenols and fibre. Polyphenols modulate breast tumour metastasis and proliferation by regulating interleukin 6 as well as inhibiting lipoxygenase, cyclooxygenase, and the transcription factor NF-κB activity, proteins that are overexpressed in tumour cells and regulate inflammatory cytokines. Fibre consumed in the diet can further control estrogen serum levels by binding them, fibre also increases insulin sensitivity and mitigate weight gain (De Cicco et al., 2019).

Consumption of red meat has a slight increase in breast cancer incidence of 6%, especially in women that exceed the recommended amounts of 350-500g portions of red meat weekly (Poorolajal et al., 2021). It is believed that red meat increases the circulating levels of endogenous estrogen, insulin-like growth factor 1, and pro-inflammatory cytokines (De Cicco et al., 2019). Processed meats are also detrimental to breast cancer prevention, representing an increased risk of 9% in breast cancer incidence in post-menopausal women (De Cicco et al., 2019).

Dietary fat consumption is another factor to consider, whilst since fat intake is associated with increased risk of developing BC of the positive receptor kind (ER+/PR+) in contrast to hormone receptor negative (ER-/PR-), suggesting signalling cascade manipulation as a means of pro-carcinogenesis process (De Cicco et al., 2019). High levels of dietary cholesterol have also been implicated in BC risk. Not all fats are detrimental to one's health, ω-3 fats have a protective effect and are considered healthier than the previously mentioned saturated fats (De Cicco et al., 2019).

Soy and its derivatives can be beneficial for women in breast cancer prevention, mainly due to their rich content in isoflavones, compounds that are structurally similar to endogenous human estrogen and, as a result, compete with the latter in estrogen receptor binding. Genistein, daidzein and glycitein are the main isoflavones present in soy products, and whilst some studies defend their anti-carcinogenic capabilities in inhibiting angiogenesis and inducing apoptosis, other studies point to their similarity to human endogenous estrogen as another form to facilitate

estrogen receptor activation in ER+ breast tumours and increase tumour proliferation (De Cicco et al., 2019).

When it comes to alcohol consumption it shows an increase in HER2-overexpression subtype risk as well as in Luminal A when compared to non-drinkers, showing an odds ratio (OR) growth with the subsequent increase of drinks per week (Tamimi et al., 2012; Trivers et al., 2009). The World Cancer Research Fund (WCRF) and the American Institute for Cancer Research (AICR) concluded in 2018 that alcohol consumption has a direct influence on breast cancer incidence for both premenopausal and postmenopausal women, especially the latter. It is also claimed that a 10-gram increase in daily alcohol intake resulted in a proportionately higher risk of developing breast cancer by 5% in premenopausal women and 9% in postmenopausal women (Freudenheim, 2020). In the European population's perspective, the same increase of 10% alcohol intake equates to a 4% increased risk for breast cancer.

Alcohol's influence on breast cancer risk seems to be more related to estrogen-positive tumours opposite to estrogen-negative, both in European and north American populations. Furthermore, it seems unlikely that beverage type has a significant influence over breast cancer risk, being majorly influenced by quantity over quality of liquor (Freudenheim, 2020).

Age of consumption is also a relevant factor to take into account: daily consumption of 10 grams of alcohol between the age at menarche and first pregnancy is associated with an increase of 11% in breast cancer incidence and, in the case of women between the ages of 30 – 55, the same 10 grams per day of alcohol is linked with an 8% BC risk increase. There is also a correlation between alcohol consumption and increased breast density, a factor that is strongly associated with an increase in BC risk (Freudenheim, 2020).

Alcohol's mechanism of influence over mammary tissue seems to rely on the oxidative stress its metabolism causes, as well as one of the products that come from said metabolism - acetaldehyde, a known carcinogen. DNA-methylation is also affected under the influence of alcohol (Freudenheim, 2020).

Invasive breast cancer is most common in current or former smokers when compared to never smokers, indicating tobacco has a noticeable influence on the development of breast carcinoma. The risk is highest in women who started smoking earlier, especially during the period of biological vulnerability, leading to the belief that it is due to the susceptibility of mammary tissue to genotoxic agents at that age.

Additionally, some studies defend a no dose-response relationship between smoking and breast cancer, being the risk solely dependent on the age of exposure itself (Gaudet et al., 2013), whilst other literature exposes women with specific *N-acetyltransferase 2* slow acetylation genotypes who have been observed to be more susceptible to breast cancer, reporting an increase of 1.44 relative risk in developing BC for women who smoke more than 20 pack/years (defined as the average packs smoked per day times the number of smoking years) and women with less than 20 pack/years incur in a 1.21 increased relative risk of developing breast cancer (Ambrosone et al., 2008).

Regarding oral contraceptives (OC), some studies have shown them to increase the risk for breast cancer, namely TNBC, whilst other studies suggest a reduction in the frequency of Luminal A breast carcinomas (Razzaghi et al., 2013). There is also evidence that oral contraceptive usage before the first full-term pregnancy or for longer periods (around 5 years) can influence the development of tumoral breast tissue specifically in BRCA1 and BRCA2 mutation carriers (Kanadys et al., 2021).

Hormonal replacement therapy, taken to reduce the affliction of postmenopausal symptoms, is associated with an increase of Luminal A subtype prevalence and it is a risk factor that should be considered when prescribing it (Saxena et al., 2010; Tamimi et al., 2012). Hormonal replacement therapy after menopause begins to show adverse effects after repeated estrogen use for 5 years, increasing the relative risk of developing breast cancer. In the case of hormonal replacement therapy administrated in the pre-menopause period, and/or with estradiol as an estrogen substitute after 15 years of recurrent dosage, the risk of developing BC increased to 2.2. Furthermore, in the extreme cases of patients with familial history of breast cancer, the risk rises to 3.4 (Steinberg et al., 1991).

The main mechanism of action in hormonal replacement therapy that likely contributes to breast cancer development stems from the hormonal exposure breast cancer cells, or tumoral cells that are hormone receptor-positive to be more precise, are subjected during the therapy regimen. This increase in proliferative behaviour is even more exaggerated in *BRCA1*-linked hormone-responsive breast cancer, being paramount to analyse the background of a hormonal replacement therapy candidate's familial history regarding breast cancer (Poorolajal et al., 2021).

Social decisions women make during their lifetime also influence BC's risk, and when it comes to pregnancy, it has been reported that women who had children experienced higher rates of TNBC but reduced incidence of Luminal A breast cancer when compared to nulliparity

women. This difference was further highlighted by the woman's age at the birth of her first child (Millikan et al., 2008; Xing et al., 2010). Breastfeeding duration was inversely proportional to the risk of Triple Negative, Luminal A and Luminal B breast cancers, meaning women that breastfed for a longer period would have a relatively decreased incidence of these subtypes of breast cancer (Tamimi et al., 2012; Trivers et al., 2009; Xing et al., 2010; Reviewed in Barnard et al., 2015.).

Environmental aggressors like ionizing radiation, have been correlated to breast cancer. This observation was performed by studying populations affected by abnormal Gray (Gy) rates like the Japanese atomic bomb survivors (0.41 Gy), tuberculosis patients exposed to a radiation-based therapy (0.79 Gy), and acute postpartum mastitis exposed to a mean cumulative dose of 2.47 Gy of therapy. These different ranges of radiation per population resulted in the observation of a linear increase in breast cancer with radiation below age 40. This is an indication that radiation in the early years of life is much more dangerous when it comes to breast cancer incidence, most likely due to the tissue's vulnerability in younger women (John & Kelsey, 1993).

Occupational exposure to mutation-inducing chemicals was also studied in breast cancer. These can be persistent endocrine-disrupting chemicals that alter mammary gland development and hormone responsiveness, such as polychlorinated biphenyls commonly used in electrical equipment, and organochlorine pesticides like Dichloro-diphenyl-trichloroethane (DDT), a widely known pesticide banned from many countries (Rodgers et al., 2018).

In the environment, we can also find a multitude of persistent organic pollutants, such as Dioxin 2,3,7,8-TCDD (or TCDD for short), a strong carcinogen that is an aryl hydrocarbon receptor (AHR) agonist. Its presence in adipose tissue is highly associated with an increased risk of lymph node metastasis, which is highly concerning in people with high BMI incurring since it confers an overall risk of 4.48 of developing breast cancer. TCDD is believed to induce EMT increasing migration and tumour invasion, as well as mitochondrial dysfunction, by inducing mitochondrial stress signalling. Although TCDD's influence in BC women is not unanimous, some observations concluded that this receptor antagonizes Erα signalling, which should in theory mitigate tumour proliferation by negating estrogen proliferation in susceptible tumours (Koual et al., 2020).

Polychlorinated biphenyls (PCBs) were banned from many countries in the 1980s due to health concerns, although due to their persistence in the environment, human exposure to this aromatic compound still occurs to this day. PCBs have been implicated in high-grade BC tumours and poor prognosis: 14 unique PCBs were associated with increased risk in BC recurrence and 27 PBCs were linked to the risk of death in ER-positive breast tumours (Koual et al., 2020). The mechanism of pathogenesis of PCBs in breast cancer consists in the increase of oxidative stress resulting in proinflammatory reactions, as well as VEGF overexpression and subsequent increase in hyperpermeability and trans-endothelial migration on cancer cells (Koual et al., 2020).

DDT is a synthetic compound of the organochlorine pesticide kind, that supports tumoral growth of the hormone dependent cancer cells by disrupting the estrogen-androgen balance. This occurs mainly by antagonizing the androgen portion of this dichotomy, which is responsible for the inhibition of uncontrolled cell growth in hormone-responsive prone cells, especially ER positive breast cancer (Koual et al., 2020).

1.2.5 Differential Gene Expression According to ER Status

Breast Cancer incidence rates show different patterns according to hormone (estrogen and progesterone) and HER2 receptor status (be it negative or positive) in women with age dissimilarity. Whilst hormone receptor positive and HER2 negative tumours are the most common in both younger and older women, hormone negative tumours take a bigger slice of the overall breast cancer cases in women under 50 years when compared to 65 or older women (recent trends in SEER Age-Adjusted Incidence Rates, 2000-2018). This age variance can be a consequence of several events in women's life, since age at menarche and age at first birth, parity and BMI (especially in post-menopausal women) can correlate to ER status.

This differential expression of hormone receptors is commonly associated with consistent "molecular portraits" (Perou et al., 2000) that determine tumorigenic biology and progression, consequently allowing for a better understanding of how risk correlates to molecular and morphological traits. Estrogen receptor status screening has even been used to predict overall breast cancer lifetime risk, although estrogen receptor-negative predictions are not as accurate as estrogen receptor-positive due to ER-negative disease being rarer than its counterpart and therefore harder to draw a reliable prediction (Mavaddat et al., 2019).

Determining these molecular portraits is especially useful since tumour response to treatment is influenced by its molecular characteristics (Shen et al., 2012).

1.2.6 Breast Cancer Genetics

In early genetic counselling, if a patient had a first-degree relative afflicted by either hereditary breast cancer or hereditary breast-ovarian cancer syndrome, their risk of developing breast cancer was estimated at 50%. This was before the discovery of well-known breast cancer genes *BRCA1* and *BRCA2* which led to a better insight into breast cancer evolution and incidence. Breast cancer cases can be branched into 3 different categories (Figure 1.11) (Lynch & Lynch, 1996):

- Sporadic Breast Cancer refers to individuals that do not have familial history of breast cancer for two generations accounting for both parental lineages (Up to 80%) (Lalloo & Evans, 2012).
- Familial Breast Cancer refers to patients with direct relatives (first and second-degree) affected by breast cancer, even though they lack key features of the next category (15-20%) (Lalloo & Evans, 2012).
- Hereditary Breast Cancer affects individuals with first to second-degree family members with BC showing an autosomal dominant inheritance pattern for cancer susceptibility, with younger incidence of breast cancer when compared to the Familial category, increased cancer bilaterality in breast tissue and hereditary breast-ovarian cancer syndrome which leads to a higher incidence of other primary tumours (5-10%) (Lalloo & Evans, 2012).

1.2.7 Hereditary Risk Factors

Genetic variants contributing to familial or hereditary breast cancer can be organized into three subgroups: High-risk variants, moderate penetrance, and low-penetrance alleles (Figure 1.12).

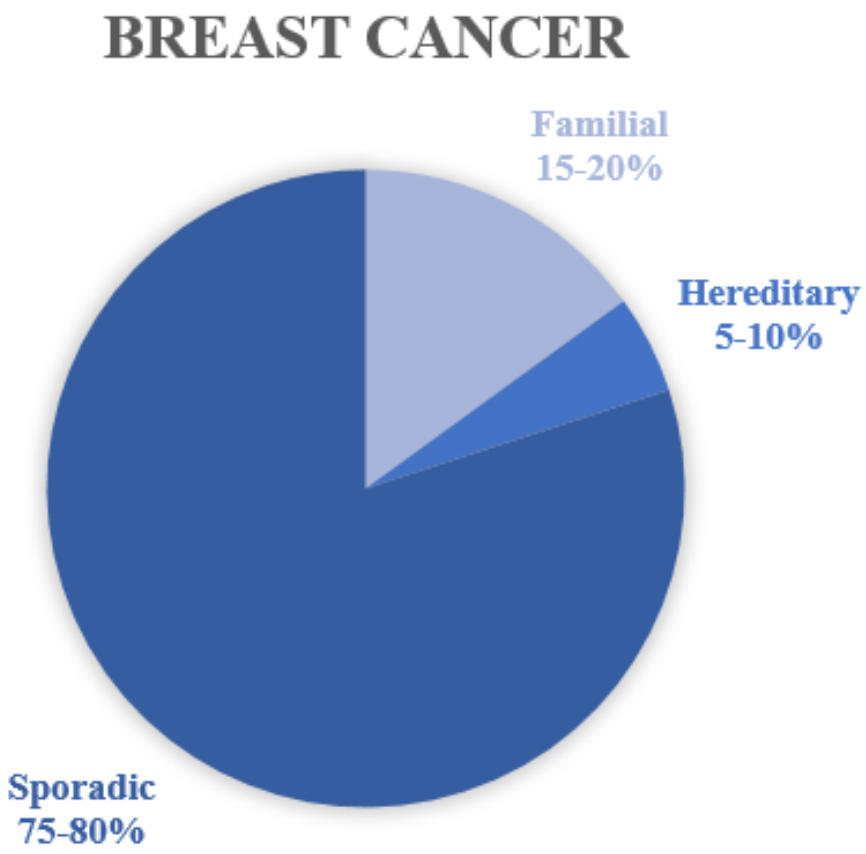

Figure 1.11: Breast Cancer Genetic distribution. A visual representation of breast cancer case origin distribution via pie-chart. Sporadic breast cancer is the most common (75-80% of all cases) represented in dark blue, followed by familiar cases (15-20%) coloured blue and lastly in light blue, hereditary BC cases being the least common (5-10%). Adapted from (Lalloo & Evans, 2012).

The first of which, mainly occurs on genes involved in rare hereditary cancer syndromes, therefore being identified in family studies. The first studies mapping the genes responsible for mendelian patterns of inheritance of breast cancer in some families, were performed by geneticists by tracing the mutation segregation through families (linkage analysis). Their associated relative risk is exceeded two-fold, and in particular cases such as high penetrance mutations in *BRCA1* and *BRCA2*, the risk of developing breast cancer increases from 10 to 30 times when compared to the general world population (Foulkes, 2008).

Moderate penetrance variants' relative risk fluctuates between two to three-fold over the normal population. These variants were mainly identified in four genes: *CHEK2, BRIP1, ATM* and *PALB2*. They are rare in most populations, and as a result, they are usually identified through resequencing of candidate genes in affected individuals of known breast cancer families (Foulkes, 2008).

And the third subgroup encompasses low penetrance alleles which are common in the overall population, but they usually confer a lower relative risk. Their frequency in breast cancer patients usually ranges from 15-40%. The previously described approaches for mapping high penetrance rare variants did not reach success in the identification of low-risk common variants, where a new method was developed that consisted in measuring common variants frequency in both affected and unaffected individuals from the same population, in other words, by performing massive genome-wide association studies (GWAS) (Foulkes, 2008) (Ghoussaini et al., 2013).

The three types of variants identified are more thoroughly explained in the upcoming section.

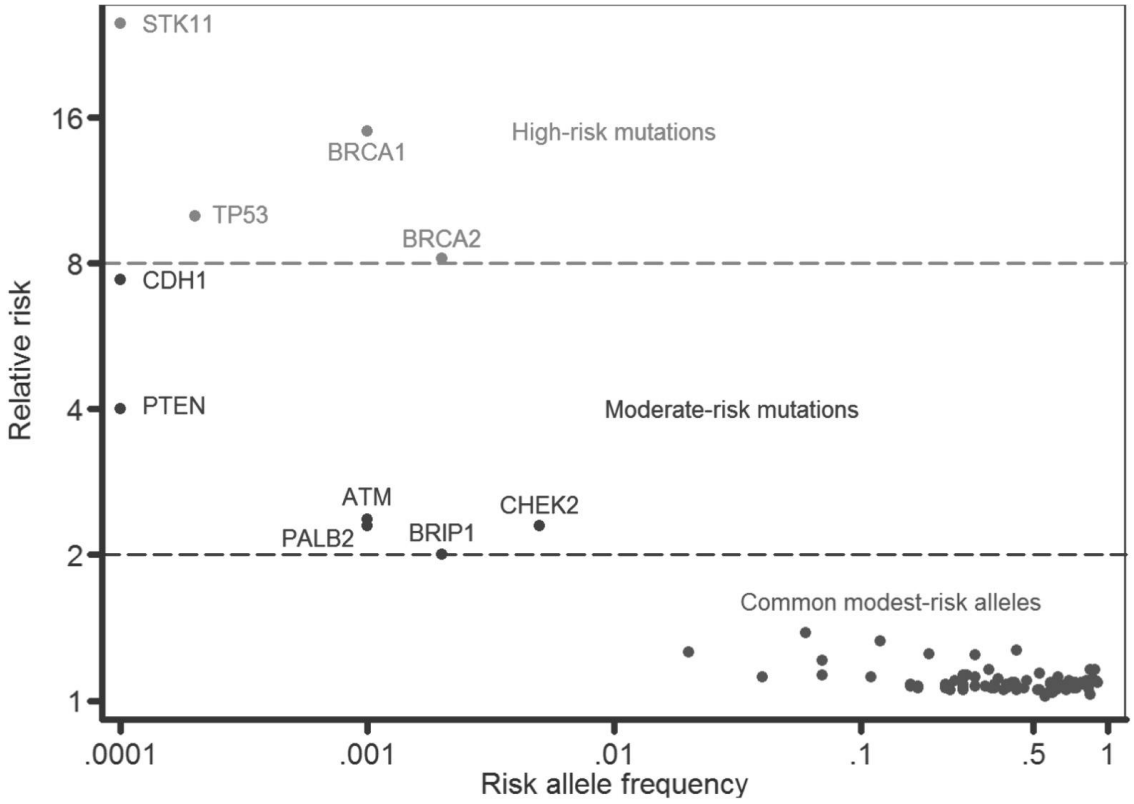

Figure 1.12: Risk Genes and variants in Breast Cancer. An arrangement of variants in genes according to their frequency in the overall population (x axis) and relative risk of developing breast cancer conferred in comparison to a "healthy" population (y axis). According to these statistics, three subgroups can be ascertained: A rare high-risk group of mutations, a low-frequency moderate-risk group of mutations and a common low-risk group of variants. Note: From "Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning?", by M. Ghoussaini, P. D. P. Pharoah & D. F. Easton, Am J Pathol. 2013 Oct;183(4):1038-1051 (10.1016/j.ajpath.2013.07.003).

1.2.7.1 High Penetrance alleles

High penetrance alleles, such as deleterious mutations in *BRCA1* and *BRCA2* that have a combined frequency of 0.4% but an overall breast cancer risk increase of 10 to 20-fold, attribute the highest lifetime risk of developing breast cancer, even though they are relatively rare in the general population (Lalloo & Evans, 2012). Both *BRCA1* and *BRCA2* are tumour suppressor genes involved in DNA repair, and therefore, cells with impaired function in these genes show double-strand break (DSB) DNA repair deficiency (Ellsworth et al., 2019). Located in 17q21 and comprising 24 exons, *BRCA1* pathogenic mutations are responsible for an increase of breast cancer lifetime risk between 60% to 85% and for other carcinomas namely ovarian cancer (40% to 60% lifetime risk) (Lalloo & Evans, 2012). *BRCA1* is essential in the repair of replication forks and transcriptional regulation in the event of DNA damage, and it is also involved in cell apoptosis and division (Ellsworth et al., 2019). *BRCA2* located at 13q12 and comprises 27 exons, it interacts with *RAD51*, a gene responsible for DSB DNA repair, conferring a lifetime breast cancer risk when harbouring pathogenic mutations between 40% - 85% (Ellsworth et al., 2019; Lalloo & Evans, 2012).

Disruptive mutations in *TP53* - a tumour-suppressor gene - can originate Li-Fraumeni syndrome which is associated with a plethora of tumours in different tissues, being breast cancer the most common malignancy in female carriers of the *TP53* mutation. Mutations in this gene are associated with very early onset breast tissue, with 5% of the cases being detected before 30 years of age. Even though Li-Fraumeni syndrome is associated with a small fraction of total breast cancer cases, carrying mutations in *TP53* represents a 56% to 90% lifetime risk of developing cancer (Apostolou & Fostira, 2013).

Cowden syndrome is an autosomal dominant disorder linked to malignant breast tumours and it is caused by germline mutations in *PTEN* tumour suppressor gene, originating multiple hamartomas (benign tumour-like malformations). Individuals affected by this syndrome have some form of disease (mainly mucocutaneous lesions) manifestation in their 20's and an overall 50% risk of developing breast cancer (Apostolou & Fostira, 2013).

STK11 (Serine Threonine Kinase) is a tumour suppressor gene involved in cell cycle regulation and apoptosis. Germline mutations in this gene cause Peutz-Jeghers syndrome, which in turn represents a risk of up to 85% of developing any cancer, namely breast cancer. Therefore, breast MRI (Magnetic Resonance Imaging) is prescribed in women as young as 25 who have developed this syndrome (Apostolou & Fostira, 2013).

1.2.7.2 Moderate Penetrance Genes

ATM is a gene that codes for a protein kinase involved in DNA repair and cell cycle regulation. This gene's name originates from the condition caused by both copies being mutated, ataxia-telangiectasia. Mutations in this gene confer a risk for breast cancer of 2.3, being more impactful in younger women (Ellsworth et al., 2019; Shiovitz & Korde, 2015).

CHEK2 is a gene that encodes a serine threonine kinase involved in DNA repair (Ellsworth et al., 2019) by interacting with BRCA1 and cell cycle regulation by stabilizing p53 during phase G2 (Shiovitz & Korde, 2015). The most common mutation is CHEK2*1100delC, a deletion present in 1-2% of the population, that increases breast cancer risk by two-fold in females and ten-fold in males. It is also interesting to denote that co-carriers of *CHEK2*, *BRCA1* and *BRCA2* mutations often show no additional risk of developing BC, which is thought to be caused by the similarity of effects these mutations cause in DNA repair.

BRIP1 mutations are responsible for a two-fold increase in breast cancer risk in women with familial history of breast cancer, even though they only account for less than 1% of total breast cancer cases. The protein encoded by this gene interacts with the BRCA1's BRCT domain (BRCA1 protein's C-terminus). *BRIP1's* most observed nefarious modifications are truncating mutations (Shiovitz & Korde, 2015).

1.2.7.3 Low Penetrance Alleles

With the introduction of Genome-Wide Association Studies (GWAS) hundreds of new common Single Nucleotide Polymorphisms (SNPs) were associated with breast cancer risk (Altshuler et al., 2008). GWASes stem from the knowledge of linkage disequilibrium (LD) which is the event of a specific allele in a SNP being simultaneously inherited alongside another precise allele in a nearby SNP in a particular population (Bush & Moore, 2012). LD is therefore a nonrandomized correlation of alleles in different loci, meaning loci within close proximity in the genome show a stronger LD whereas more distant loci in the chromosome are more likely to not be as closely associated, and as a result show a weaker LD between them (Visscher et al., 2012).

Over many generations, a sizeable population will slowly inch towards linkage equilibrium through chromosomal recombination events, meaning all possible haplotypes will show in equal proportions as a Hardy-Weinberg ratio (Bush & Moore, 2012).

The analysis of linkage disequilibrium can be beneficial to find SNPs associated with known causal variants, and with the achievement of the human genome sequencing milestone, mankind ushered into a new era where associations between variants and phenotypes could be performed within a whole-genome scope instead of being restricted to a couple of a hundred markers. This means one can perform an unbiased and hypothesis-free whole genome association analysis of marker variants with a trait of interest, without prior knowledge of the genomic location of the causal variant(s). Furthermore, there is no need to test the association of all known markers since due to the LD phenomena, some markers are highly correlated and therefore are redundant (Visscher et al., 2012).

GWASes have potentiated the universal knowledge of breast cancer by identifying around 173 common low-risk susceptibility variants, most of which of the SNP kind (Ahearn et al., 2022). These included early identified variants in *FGFR2*, *TOX3* gene and *MAP3K1* gene (Ahearn et al., 2022).

Some variants even show a specific association with either luminal or non-luminal cancers, which indicates a certain degree of correlation in terms of genomic features between subtypes (Zhang et al., 2020). In the study by Zhang and colleagues, 82 Breast Cancer Association Consortium (BCAC) studies encompassing women with European ancestry in 20 different countries alongside genotyping data from Collaborative Oncological Gene-environment Study (iCOGs) and OncoArrary, were analysed through LD score regression, in order to better understand the correlation between certain tumour clinical features (Figure 1.13). It could objectively be seen that luminal subtypes are highly similar to Luminal A to the same degree non-luminal subtypes are similar to triple-negative, although luminal A and triple-negative only share 0.46 genetic similarity (Zhang et al., 2020). In this study was also included a dataset originated from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA) composed of *BRCA1* mutation carriers, which effectively represented individuals with that specific breast cancer susceptibility. From BCAC/CIMBA *BRCA1* mutation-carriers meta-analysis there was a striking 0.83 genetic correlation between triple-negative disease and *BRCA1* mutation carriers. A few SNPs also emerged from this meta-analysis, such as rs17215231 and rs2464195 that were associated with triple-negative breast cancer (Zhang et al., 2020).

In addition, GWASes identified several loci specifically associated with ER-status (Michailidou et al., 2017), being the vast majority of identified SNPs related to ER-positive breast cancer. However, there are also variants specifically associated with ER-negative tumours as reported in a study that replicated the association of 10 known ER-negative disease SNPs as well as discovered 10 new ER-negative BC-associated SNPs in a dataset of European ancestry. Furthermore, five out of these 20 ER-negative SNPs, were also strongly associated with triple-negative disease risk (Milne et al., 2017). Other GWASes also pinpointed a myriad of variants specifically associated with other breast cancer tumour markers, such as HER2, and progesterone receptors (Zhang et al., 2020). Such examples include variants rs35054928, rs2981578 (both located in *FGFR2* gene) (Easton et al., 2007; Hunter et al., 2007) and rs13281615 (Easton et al., 2007), that have been associated with both luminal subtypes and HER2 enriched subtypes but not with triple negative disease. But a few variants such as rs4784227 (located in *TOX3* gene) (Easton et al., 2007; Stacey et al., 2007), rs62355902 (located in *MAP3K1* gene) (Easton et al., 2007) and rs2747652 (located in ESR1, in an enhancer region to be more precise, and this variant is peculiarly associated with *ESR1* reduced expression, consequently affecting the expression of the estrogen receptor) (Dunning et al., 2016) have been associated with the five main breast cancer molecular subtypes.

In order to identify which breast cancer risk variants showed heterogeneous associations according to ER, PR and HER2 status, Ahearn and colleagues tested the association of 173 previously reported BC associated variants with intrinsic molecular subtypes. They identified some variants displaying different effects depending on the tumours receptor status, such as rs6678914 and rs4577244 (Milne et al., 2017) which show opposing outcomes in ER-positive and ER-negative disease. Others vary on a by-tissue basis as rs17879961 (Fachal et al., 2020), which has been reported to be highly associated with increased risk in pancreatic ductal adenocarcinoma and chronic lymphocytic leukaemia even though it represents a considerably lower risk of developing ovarian cancer. This phenomenon indicates that the same biological pathways could distinctively influence different tumour types depending on the tissue or cell of origin, often creating antithetical effects in distinct cases (Ahearn et al., 2022).

Progesterone receptor-associated variants rs10759243 (X. Li et al., 2016; Michailidou et al., 2017), rs11199914 (Michailidou et al., 2013), rs10816625 (Orr et al., 2015) and rs72749841 (Michailidou et al., 2017) were also found associated with ER-positive breast cancers, showing the bias of the estrogen receptor when it comes to etiologic heterogeneity. One of the few variants solely associated with PR status is rs380366 (Broeks et al., 2011), and

the interesting fact about this variant is its high LD (r^2 = 0.78) with the aforementioned rs4784227, a variant located in *TOX3* gene, which is highly associated with PR status.

Association studies looking at HER2 over-expression have identified fewer variants solely associated with this receptor, but rs10995201 (Darabi et al., 2015) is one such variant, although it's strength of evidence is not as hard as observed for ER-associated variants (Ahearn et al., 2022).

In the study by Ahearn and colleagues, they also evaluated the association of variants with tumour grade and identified this tumour feature as an important determinant of etiologic heterogeneity. Variants rs17426269, rs11820646, and rs11571833 (Mazoyer et al., 1996) have been shown to be highly associated with grade 3 tumours, and the latter variant (rs11571833) is localized in the *BRCA2* coding region, being also responsible for a truncated conformation of the protein and highly associated with triple-negative breast cancer (Ahearn et al., 2022).

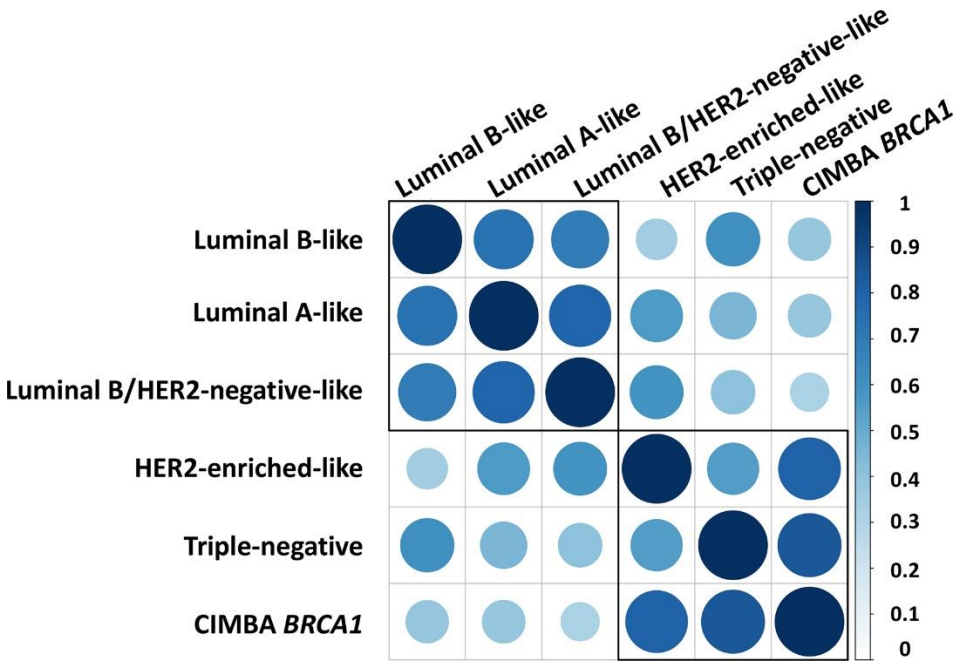

Figure 1.13: Genetic correlation between luminal and non-luminal subtype families and *BRCA1* mutation carriers gaged in LD-regression score. LD-regression score is represented by size and shade of blue circumferences, where dark blue corresponds to perfect genetic correlation (see gradient on the right side). Note: From "Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses", by H Zhang, T. U. Ahearn, J. Lecarpentier et al, Nat Genet. 2020 Jun;52(6):572-581 (10.1038/s41588-020-0609-2).

Since tumour grade results from multiple neoplastic characteristics, namely: nuclear pleomorphisms, mitotic count and the formation of cellular structures like glands and tubules,

the variants identified in this study as associated with tumour grade, can also be determinants of these characteristics and influence countless biological pathways implied in these mechanisms (Ahearn et al., 2022).One common characteristic of the variants found associated in GWAS is that they confer low risk (OR usually lower than 1.5) (Maxwell & Nathanson, 2013), and as expected, if an individual carries two risk alleles they have an increased risk when comparing to heterozygous for the same SNP.

This low risk conferred by common variants, instead of being originated from variants located in genes involved in DNA repair impairment, that is frequently involved in moderate and high risk, can also originate from a stimulation of growth and proliferation genes (Shiovitz & Korde, 2015).

The results from eQTL analysis can provide an incredible advantage in interpreting the GWAS outcome. However, mapping of true causal variants is not easy, since GWAS associated variants are primarily located in noncoding regions of the genome (either intergenic or intronic) and usually, GWAS hits for a certain disease encompass regions with many variants in linkage disequilibrium with each other, regions that may contain more than one gene, and frequently the hits fall completely outside of the coding region and congregate in regulatory elements being much harder to identify the risk gene and to draw causality (Albert & Kruglyak, 2015).

Interestingly, post-GWAS functional analysis have shown that altering expression of near, or long-range genes, is a common mechanism associated to risk for breast cancer (N. Li et al., 2018). *BRIP1's* (a gene that encodes for a BRCT interacting protein) Pro919Ser polymorphism, that confers an increased risk for BC in premenopausal women, is one such SNP (Shiovitz & Korde, 2015).

1.3 Cis-Regulation

As previously discussed, breast cancer molecular subtypes' uniqueness relies on the differential gene expression levels that result in different receptor combinations. Differences in expression levels can be influenced by heritable factors that can be grouped as either trans-acting factors or cis-acting factors. Whilst trans-acting mechanisms appear to be quantitatively more important, cis-acting variants account for 25% - 35% of different inter-individual gene expression, even though physiological feedback mechanisms might be diminishing the true impact of these variants (Pastinen & Hudson, 2004).

Cis-acting factors are defined by affecting a single allele's expression in a gene; therefore, they are commonly associated with promoters and enhancers resulting in a controlled influence on the gene in an allelic-specific manner. By contrast, trans-acting elements do not discriminate between alleles resulting in a wider influence over the gene by affecting both copies in the same magnitude (Pastinen et al., 2006) (Figure 1.14).

Different methods exist to detect the effect of cis-acting variants. Expression quantitative trait loci (eQTL) are sequence variants that are associated with the total expression level of singular or multiple genes (Figure 1.15), identifiable by the analysis of a genetically diverse population for the locus of interest (Albert & Kruglyak, 2015). These eQTLs can in part explain the genetic variance of a gene expression phenotype by influencing the expression levels of a given gene. Therefore, through standard eQTL analysis, genetic variants and gene expression levels' direct association can be tested. There is also plenty of evidence that gene regulation occurs in a cis specific manner and as a result, a great deal of genes have cis-eQTL influence (Nica & Dermitzakis, 2013).

In GWAS, genetic markers are examined for possible association with complex traits and if the association is confirmed, eQTL mapping is a commonly used approach to determine possible influence of the associated variants over quantitative expression levels of neighbour genes (Mehta et al., 2013).

A few steps are needed before preforming eQTL mapping, specifically two types of data must be collected from all the individuals that compose the studied population (Albert & Kruglyak, 2015):

- The genotype of each individual must be first determined, either by fully sequencing one's genome or by using SNP microarrays genotyping in specific genomic lengths, so that the variants of interest are identified and accounted for in every individual for the entirety of the population.
- The expression levels of the genes of interest must be measured through expression microarrays quantification or via RNA sequencing, effectively obtaining numerical values that uncover one's genetic expression profile.

With both these datasets in hand, individuals can then be assembled in separate groups depending on the specific combination of alleles they carry (genotype), logistic regression analysis can then be applied to compare total levels of gene expression between groups. This process can be repeated for every DNA variant in the genome for the same gene, with the

ambition to scan for every possible eQTL that can be influencing expression values (Albert & Kruglyak, 2015).

Therefore, GWAS don't identify the causal variant neither the target gene, but the identification of the associated variant of a candidate causal variant as an eQTL for a proximal or distal gene, can bridge the knowledge gap between GWAS variants and the biological implication they carry in disease development (Albert & Kruglyak, 2015).

When a GWAS hit is simultaneously implied as an eQTL for a certain gene, there is strong evidence of this gene's expression being highly correlated with the disease at hand. Furthermore, it is of common knowledge that GWAS hits in common diseases are prime locations for eQTL enrichment. This is also true for the reverse of this process: when it comes to eQTL studies in specific tissues implied in certain diseases such as breast tissue for breast cancer, the identification of numerous eQTL-GWAS pairs can provide strong evidence for a causal mechanism involved in disease progression (Albert & Kruglyak, 2015).

This spawned a new opportunity for researchers to validate their work, by prioritizing variants that have been previously identified as eQTLs in published datasets by other researchers, especially the ones located in regulatory regions of the genome (Albert & Kruglyak, 2015).

Regarding breast cancer, previous studies have analysed The Cancer Genome Atlas for possible eQTL activity, identifying ER-positive breast cancer association in 6,145 SNPs from which 5,893 are cis-acting eQTL loci as well as 1,359 target genes, of which 689 genes are solely regulated by a single cis-acting SNP locus. The remaining is influenced by more than one SNP ranging from 2 to 83 different loci applying direct influence. The genetic expression variation in the target genes provided by the cis-acting SNPs, fluctuates from 18.5% to 73.2% (Q. Li et al., 2013).

On the other hand, ER-negative breast cancer disease showed fewer cis-eQTL associations, bringing to light 380 relevant cis-acting SNPs in 179 targeted genes overlapping 43 significant target genes with the ER-positive population (Q. Li et al., 2013).

Previously identified 15 breast cancer associated loci from GWAS studies were also analysed in this study for possible association with gene expression levels, resulting in a significant outcome for 3 SNPs in particular regions of the genome: rs6721996 located in *IGFBP5* gene (2q35), rs4700485 located in C5orf35 (5q11) in the genome and lastly rs3803662 located in *TOX3* gene (16q12) (Q. Li et al., 2013).

Another more in-depth and wide study about the cis-eQTLs in breast cancer, surveyed data from not only The cancer Genome Atlas but also from both Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and Genotype-Tissue Expression (GTEx) as well (X. Guo et al., 2018). They analysed 159 lead variants previously associated with breast cancer in GWASes and already associated with expression alterations in other genes to possibly identify additional target genes for these SNPs. In METABRIC two datasets were analysed: 98 lead variants were matched with 222 target genes in tumoral tissue while in the normal tissue 72 lead variants were found associated with 161 distinct target genes. In TCGA (tumoral tissue) 216 target genes were found associated with 87 lead variants and lastly in GTEx project, 95 GWAS lead variants were identified to exert influence over 250 target genes (X. Guo et al., 2018). From the comparison of all three datasets 8 genes previously associated with carcinogenesis come to attention: *WNT3* heavily implied win the WNT pathway which has already been implied in oncogenesis due to its developmental ramifications; *CASP8* crucial in cell apoptosis and therefore inevitably linked to cancer progression in the eventuality of failure in executing cellular programmed death; *ESR1* the protein that encodes for the estrogen receptor and an overall clear dictator of breast cancer prognosis; *AKT1* involved in AKT/PIK3CA pathway that subsequently activates other pathways and that has been observed to be dysregulated in tumoral tissue; *POLR2L* a gene that encodes for a RNA polymerase II subunit and results in increased RNA polymerase activity, a trait immensely valuable to growing a proliferative tumours; *FGF10* a member of the fibroblast growth factor family that have vast implications in cellular activities from survival to mitogenic, as well as growth factor activity and chemoattractant activity, essential in tumour growth and invasion; *IGFBP5* relevant in smooth muscle migration and proliferation, in addition to the binding to insulin-like growth factor I (IGF-1) and consequently mediating the growth-promoting effects of IGFs, and *MUTYH,* whose mechanism of action lies on encoding a DNA glycosylase involved in oxidative damage repair in the DNA (Q. Li et al., 2013).

From the 159 lead variants, differential gene expression was analysed in both positive and negative ER status cell lines, by introducing the sequence with the functional SNP with the alternative allele suspected to alter the target gene expression in a luciferase reporter plasmid. Both fragments containing the alternative allele in rs11552449 (*DCLRE1B* gene) and rs7257932 (*SSBP4* gene) showed significantly reduced promoter activity in their respective target genes in both ER positive and ER negative cell lines in comparison to the reference allele (X. Guo et al., 2018).

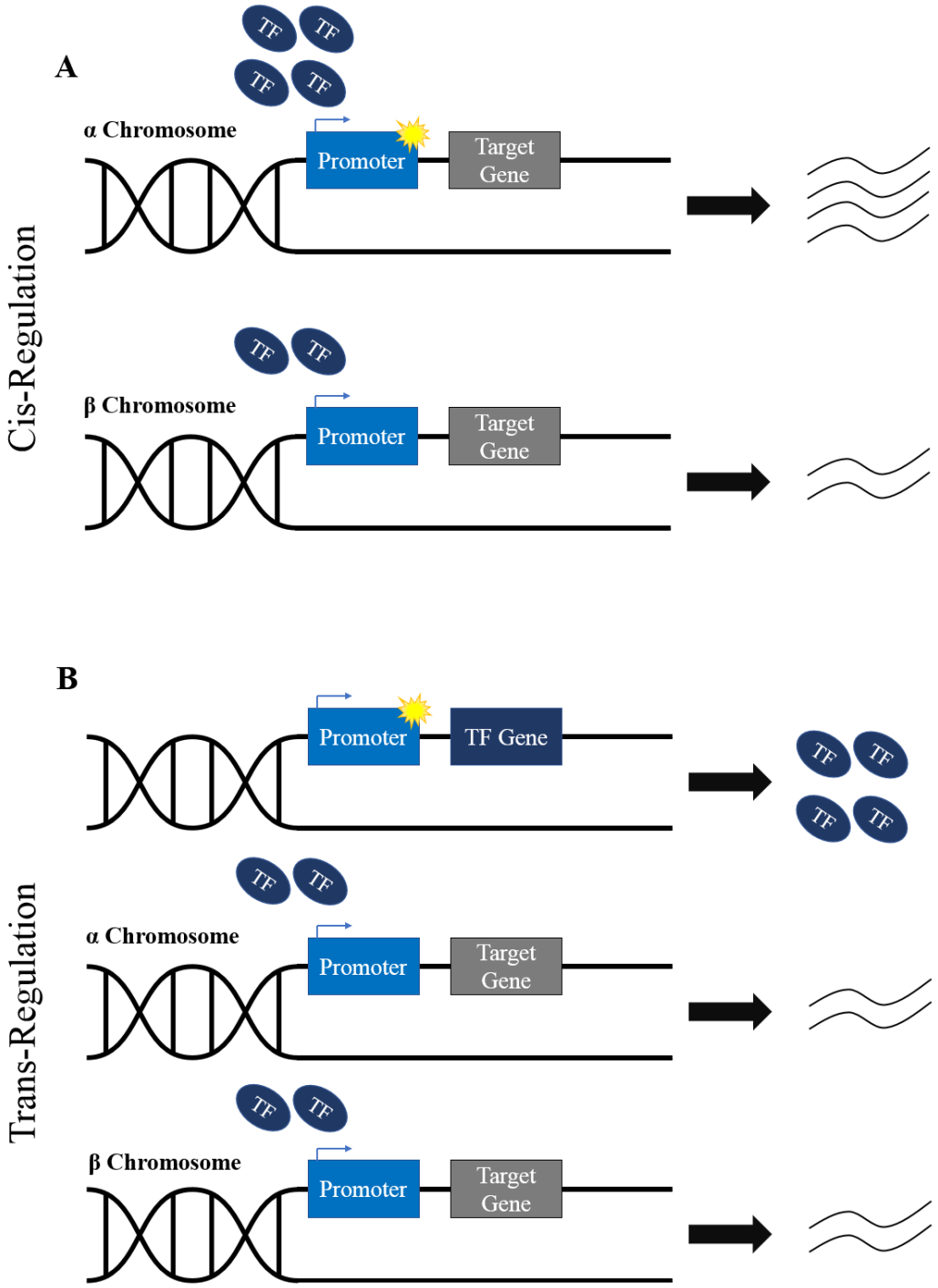

Figure 1.14: Cis-Regulation and Trans-Regulation. When a polymorphism (yellow splatter) affects the binding of transcription factors (TF) in a particular promoter of a gene in an allele-specific manner, the subsequent transcription of one of the alleles of the gene can be favoured over the other in the homologous chromosome, ultimately resulting in the two alleles being expressed at different levels (Cis-Regulation, **A**). On the other hand, when the polymorphism modulates the promoter region of the transcription factor gene resulting in the under expression or in overexpression (as in this example) leading to an increased presence of TFs, both alleles of the target gene are affected in the same magnitude, and their expression is therefore modulated equally (Trans-Regulation, **B**).

But the opposite effect could also be replicated in the case of the alternative alleles in SNPs rs3747479 (*MRPS30* gene) and rs73134739 (*ATG10* gene), which increased promoter activity in ER-positive and negative settings, when compared to the reference allele counterpart. Although in rs2236007 (*PAX9* gene), the alternative allele was associated with an increase in promoter activity for the ER-positive cell line, there wasn't a significant difference in activity between alternative and reference alleles in the case of ER-negative setting, reinforcing the need to evaluate both sides of the receptor status spectrum as independent diseases with intrinsically unique prognosis (Q. Li et al., 2013).

Many eQTLs are still left to be identified, some have low statistical power and are brushed off in small sample studies, but as the technology progresses and the interest in eQTL and collaboration increases, so do the sample sizes in future studies, allowing for the identification and cataloguing of new eQTLs with smaller but relevant effects in pathologies (Albert & Kruglyak, 2015).

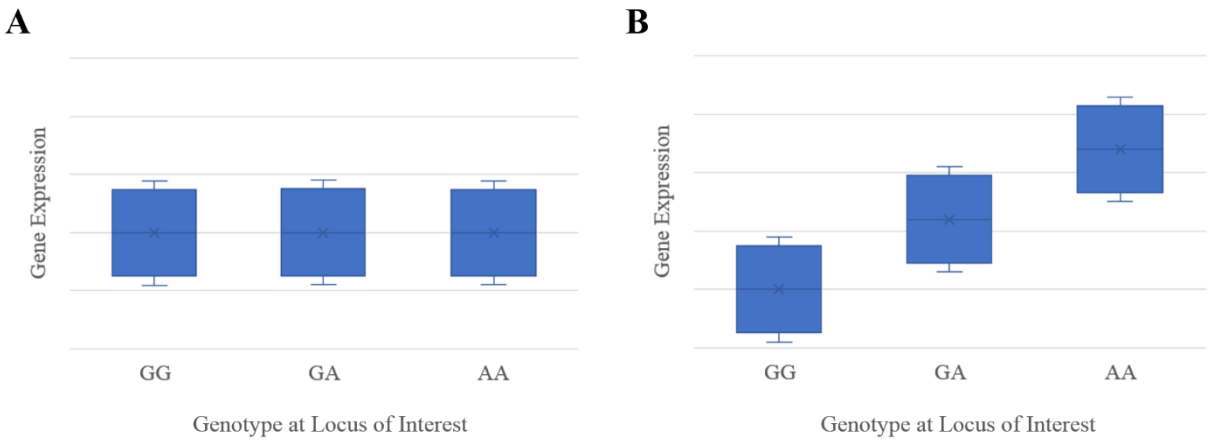

Figure 1.15: Expression Quantitative Trait Loci. The boxplots on the left (**A**) show equal gene expression for each distinct genotype and as a result it is not an eQTL, while the boxplots on the right side (**B**) have different levels of gene expression across the genotype variations, and therefore, is an eQTL.

Even though most studies performing cis-regulatory variant mapping focus on total gene expression analysis, the basis of regulating expression in transcripts due to variants localized in promoter/regulatory regions is allele specific.

An alternative approach to detect the presence of cis-acting regulatory variants (rSNPs) is to perform differential allelic expression (DAE) analysis of a particular gene (Figure 1.16), also known as allele-specific expression analysis (ASE).

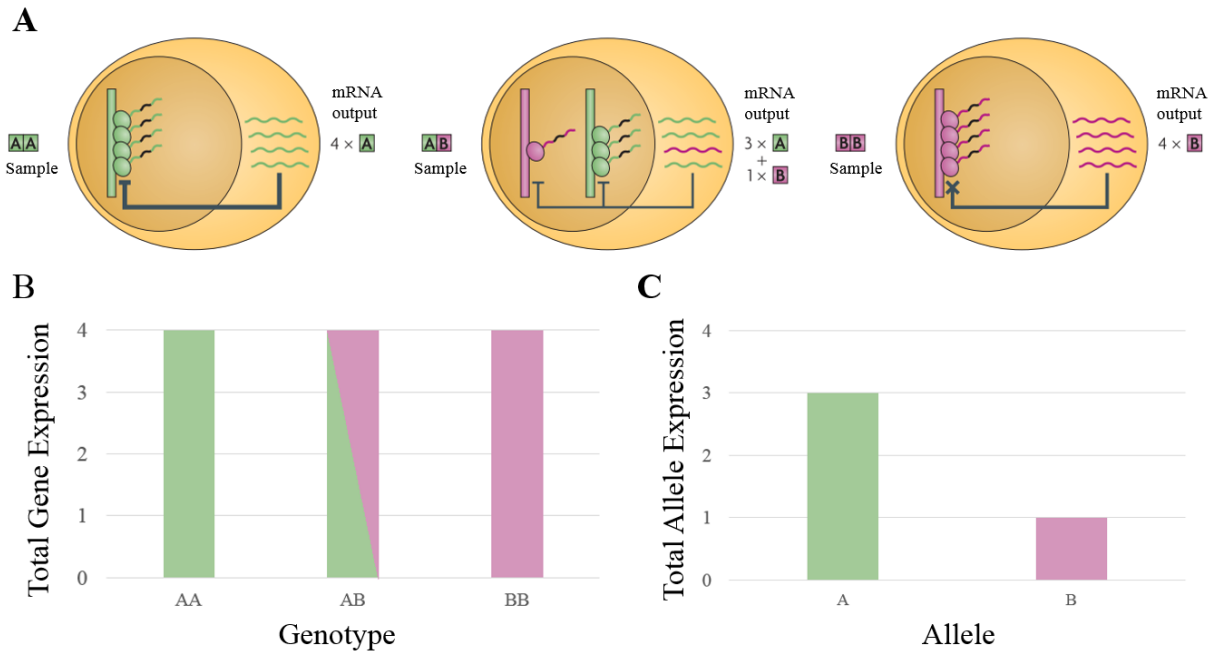

Figure 1.16: Cis-regulation in the presence of trans suppressive effects. If we consider that there is a cis-regulatory difference in expression between alleles A and B, analysing the total expression could be difficult to detect such phenomena due to negative-feedback mechanisms (**A**) since there is the same overall total gene expression in all three genotypes (**B**). However, when quantifying each allele's expression in the AB heterozygotes we can still verify differential allelic expression being the A allele more expressed than the B allele (**C**), attesting to the fact that cis-variation effects can be detected through independent allele expression measurement even in the presence of extrinsic suppressive effects. Adapted from (Pastinen, 2010).

In the earlier years of gene expression research, human genetic variation was in its majority preformed through expression quantitative trait loci mapping, although a better and more precise estimation of cis-regulatory effects could be provided by allele-specific methods (Pastinen, 2010).

Earlier studies from then showed promising results regarding differential allelic expression by demonstrating that 30% of the loci within an individual were affected by allele-specific differences when it came to transcripts, as well as the evidence of about 30% of the expressed genes to be under the influence of common cis-regulatory variants in the population. Applying this method to population studies, for a larger quantity of genes showed differential allelic expression was under the influence of previously unmapped common variants' allelic variation, leading to the belief that gene expression regulation is accomplished by either rare genetic variants or epigenetic effects (Pastinen, 2010).

To perform this analysis, one must study the mRNA levels of both alleles in individuals that are heterozygous for a transcribed SNP (tSNP) and observation of significant imbalances or unequal levels of expression of the two alleles will be an indication of a regulatory variant (rSNP) residing in the same locus (Xiao & Scott, 2011).

The DAE method has the advantage of having an internal standard, since both alleles are being compared with each other, they serve as a control for trans-regulatory and environmental factors, and as such, their expression would be equally affected by said variables (Xiao & Scott, 2011).

The need to use heterozygous samples can, in turn, reduce the number of samples available (Almlöf et al., 2012) and consequently, decrease the power of the study. It is also likely that trans effects interacting with cis regions cause difficulties in discerning cis interactivity from trans in this specific case (León-Novelo et al., 2018)

- Like eQTL analysis, DAE analysis can be applied to functional genomics analysis of loci of interest (as for loci identified as associated in GWAS) or alternatively to the whole genome, generating a genome-wide mapping of variants with cis-regulatory potential. Therefore, the application of DAE analysis can be separated in two different approaches: Polymorphism-directed approach: Characterized by querying individual variant sites for allele-specific function, yielding the advantage of only targeting heterozygous sites who can provide informative data to the allelic-expression analysis and also giving the possibility of analysing genomic DNA control samples with exactly the same allelic content, controlling for any inevitable biases in AE ratio measurements (Pastinen, 2010).
- Global approach: Through next-generation sequencing of the human transcriptome, it was possible to estimate allelic expression across the entire transcriptome. Since RNA-seq accumulates sequence reads from expressed transcripts, genes that show a sizeable expression in the tissue of interest and considerable genetic variation in their mRNA can be analysed at expressed polymorphic sites. These RNA-seq reads at the polymorphic sites are estimates of allelic abundance, and by applying simple statistical tests, any imbalances in expression can be detected in sites of interest. Nevertheless, in RNA-seq data it is important to account for a few caveats that could lead in error, such as clonal reads that need to be filtered and the required elevated read counts for a

determined polymorphic site, in order to provide a meaningful power to the observations this approach bestows (Pastinen, 2010).

The differential allelic expression method exposes the curious phenomena of allelic bias. SNP-targeted studies unravel that 20% of the variants show a 1.5-fold difference of expression between alleles, and this percentage raises to 30% for a 1.2-fold DAE (Pastinen, 2010).

eQTL alleles detected by RNA-seq and by allelic bias have also shown to be correlated, and in some cases significant differences in allelic expression was detected in genes without known eQTLs at the time (Pastinen, 2010).

By mapping allelic expression data and associating the differentially expressed alleles with disease association datasets, allows for possible correlation between DAE and disease progression, especially considering the alleles located in non-coding regions of the DNA that are presupposed to affect genetic expression in a cis-regulatory manner (Pastinen, 2010). Differential allelic expression approach in pathology could also determine eventual cases where disease associated genes could be enriched in the presence of specific cis-regulatory variants, helping unravel a possible lead in how the pathology evolves and behaves (Pastinen, 2010).

There have been examples of differential allelic expression measurement methods in identifying novel breast cancer risk loci in previous studies. Through this methodology it was possible to find 24 protein coding loci involved in major interaction networks affecting sex hormones and metalloproteinase pathways, two indubitably important tumorigenesis enabling fronts (Gao et al., 2012).

Furthermore, cancer causative genes have also been shown to display DAE: *ZNF331* (rs8110350 with a differential allelic expression ratio of 2.31, p-value: 0.001826) codes for a domain in transcriptional repressors and has been found to be inactive due to methylation in cancer cells; *USP6* (rs11658877 with a differential allelic expression ratio of 4.8, p-value: 0.001324) is expressed in many tumoral tissues and it codes for a deubiquitinase effectively cleaving ubiquitin tagging of proteins even its own; and also a gene heavily implicated in tumour progression *DMBT1* (rs11523871 with a differential allelic expression ratio of 2.03, p-value: 0.001676) that has been heavily implicated in tumour progression, it is believed to be a rather unique tumour suppressor due to a suspected role in cancer cell and immune system interaction and whose deletion main benefit tumour proliferation and progression in a known myriad of cancer tissues (Gao et al., 2012).

One study by Hamdi and colleagues tested for association with breast cancer in 313 variants that showed differential allelic expression in 175 genes involved in tumoral progression and proliferation enabling characteristics. These characteristics included DNA repair in both homologous recombination and inter-strand crosslink repair fashions, interaction with *BRCA1* and *BRCA2* breast and ovarian cancer susceptibility genes, controlling the cellular cycle or implicated in such, cellular programmed death, ubiquitination of proteins, known tumour suppressors, sexual steroid action as well as mammographic density. With the addition of genotyped data from iCOGS dataset, it was possible to evaluate the impact of those 313 variants in breast cancer risk (Hamdi et al., 2016). They identified a few SNPs of interest whose minor alleles were positively associated with higher risk of developing breast cancer: rs11099601 located at 4q21, rs656040 located at 11q13 and rs738200 located at 22q12.1. Additionally, rs11099601 was also associated with increased risk for estrogen receptor- positive and negative breast cancer, showing that variants with DAE can have an effect in tumours with antagonistic receptor statuses (Hamdi et al., 2016). One of the more interesting variants was rs656040, located in the *SNX32* 3'-UTR region upstream of its differential allelic expression targeted gene *MUS81*, a gene ultimately involved in DNA repair (Hamdi et al., 2016).

As previously mentioned, *BRCA1* and *BRCA2* mutation carriers have severely increased risk for breast cancer throughout their lifespan. Due to their relative importance in breast cancer aetiology, it is therefore imperative to study them thoroughly to better understand not only their mechanism of action but also their gene expression regulation and better contribute to risk management. By that line of thought, analysing DAE-associated variants specifically underlying the phenotypic variation in mutation carriers, could pinpoint variants associated with disease penetrance.

In another study by Hamdi and colleagues, breast cancer risk in *BRCA1* mutation carriers was highly associated with the following variants displaying DAE: rs6589007 and rs183459 located in *NPAT* gene and rs228592 located in *ATM* gene intron 11, all being located in chromosome 11 relatively close together in the genome, in a sort of hot spot genomic region for increased BC risk for *BRCA1* mutation carriers (Hamdi et al., 2017).

The *NPAT* gene's implication in increased breast cancer risk for *BRCA1* mutation carriers could rely on its association with cell cycle progression (reportedly phases G1 and S), as well as activation of transcription histones (histones H2A, H2B, H3 and H4) indicating a possible mechanism of involvement in tumoral progression. It is also important to note that,

germline mutations in the *NPAT* gene have already been implicated in Hodgkin Lymphoma (Hamdi et al., 2017).

The other highlighted gene *ATM* is a commonly known gene to be associated with breast cancer specifically, due to encoding a cell cycle checkpoint kinase implicated in double-strand DNA repair, which consequently impacts a variety of other well-known agents in tumorigenesis by phosphorylating downstream proteins such as p53, CHEK2 and unsurprisingly BRCA1. This gene has been also implied as an intermediate risk susceptibility gene in breast cancer and is, therefore, subject to extensive research in the field of breast cancer aetiology (Hamdi et al., 2017).

DAE in variants located in, or affecting, *CHEK2* (mentioned in the paragraph above), have also been targeted by researchers in the field of breast cancer disease. A particular study was performed with lymphoblastoid cell lines from high-risk breast cancer patients, with no prior mutation in either *BRCA1* or *BRCA2* identified and targeted two neighbouring variants located in *CHEK2*: rs2236141 and rs2236142 (Nguyen-Dumont et al., 2011). Even though no significant evidence of DAE was detected in rs2236141, four of the patients who were heterozygous for rs2236142 showed significant levels of DAE ratios (Nguyen-Dumont et al., 2011). These four individuals were then tested for the presence of CHEK2*1100delC mutation, that consists of a deletion that could effectively induce a premature termination codon leading to nonsense- mediated decay of the transcript and so it was the case. The fact that differential allelic expression was being caused by the deleterious mutation instead of a regulatory variant(s) affecting gene transcription, is an important display of how DAE can be generated by a multitude of mechanisms and lead to misinterpretations due to the complexity of processes that influence the perceived transcriptome (Nguyen-Dumont et al., 2011).

Still in the study by Hamdi and colleagues, variant rs6982040 located in the *PKRDC* gene intron 74 was observed to correlate with an increased breast cancer risk in *BRCA2* mutation carriers. The *PRKDC* gene encodes for the catalytic subunit of the DNA-dependant protein kinase (DNA-PK) and its main function is double-strand break repair, and for that reason, is a relevant player in tumour suppression, especially in individuals with *BRCA2* mutations that could naturally incur in genome instability (Hamdi et al., 2017).

Some of the previously mentioned variants have also been studied according to specific tumour features such as receptor status. Two examples are SNPs rs656040 and rs738200 whose correlation with estrogen receptor was investigated, but no significant differences when compared with its negative counterpart. Alternatively, rs110099601 was associated with

increased risk in both ER negative ((P = 4.08x10-4) and positive (P = 5.22x10-6) breast cancer (Hamdi et al., 2016).

A few variants displaying DAE in *BRCA1* mutation carriers showed association with estrogen receptor negative disease: rs11806633 located at *CDC42BPA* gene, rs6721310 located at *BRE* gene and rs2305354 located at *REV1* gene, although differences between hazard ratios in ER-positive and ER-negative breast cancer were not statistically significant. This may be in part due to the fact that the vast majority of *BRCA1* related cancers are triple-negative and therefore an overabundance of ER-negative cases tip the scales a bit too heavily (Hamdi et al., 2017).

In the case of *BRCA2* mutation carriers and specific estrogen receptor statuses, only rs9468322 located at *NKAPL* gene was associated with estrogen receptor positive disease, although once more, no meaningful statistical difference was detected between estrogen receptor positive and negative breast cancer hazard ratios in the population's survivability (Hamdi et al., 2017).

Fellow colleagues have also investigated how cis-regulatory variation influences *PIK3CA* expression in normal breast tissue. By obtaining different allelic ratios that independently measure the net mutant allele expression imbalance (RNA) and the mutant allele relative copy-number (DNA) the difference between the first and the latter ratios equated to the mutant allele expression imbalance derived from cis-regulatory variants (Correia et al., 2022). They found mutant allele expression imbalances to be common in breast tumours, in both METABRIC and TCGA datasets, as well as a sizeable portion of cis-regulatory effects influencing mutations in breast tumours (Correia et al., 2022). Furthermore, it was also evidenced that mutant allele-expression prompted by cis-regulatory variation was remarkably higher in tumours that were ER-negative, PR-negative tumours as well as HER2-positive tumours, which are categorically known as tumours displaying a worse prognosis (Correia et al., 2022). They also found six variants in *PIK3CA* in normal breast tissue that showed differential allelic expression. Mapping of variants with cis-regulatory potential in normal breast tissue followed by functional analysis pinpointed a variant with great potential of regulating PIK3CA expression by disrupting the binding motif of the transcription factor NF-YA to the gene promotor. (Correia et al., 2022). The findings reported in this work potentiated by differential allelic expression analysis have important clinical applications, whether by pinpointing a possible prognosis biomarker (based on AE ratios) or by providing evidence that patient 's harbouring somatic mutations may not express that mutations or express them in very

low levels and, therefore, will unlikely benefit from therapies specifically targeting the mutated protein. It also shows that studies effectively bridging the gap between research and possible clinical application (Correia et al., 2022).

With the present knowledge that: recent GWASes have identified variants associated exclusively with either positive or negative estrogen receptor status (Michailidou et al., 2015), showing a predisposition to form tumorigenic tissue with certain molecular characteristics, and that most breast cancer risk causal variants have been shown to be regulators of gene expression (Dunning et al., 2009), we hypothesize that cis-regulatory variants may determine clinical and molecular characteristics of the tumour.

Aims

2 Aims

The main objective of this study is to unveil the role of cis-regulatory factors in breast cancer tumours characteristics and patients' survival. To accomplish this goal, I pursued the following specific aims:

- Test the association of allelic expression (AE) ratios with BC clinical features and patient’s outcome using samples from normal-matched breast tissue from BC patients;
- Generate a list of AE associated candidate genes that can be further researched in their influence over breast cancer.

Materials and Methods

3 Materials and Methods

3.1 R Programming

The R programming language is a useful tool in the analysis of statistical data whether it be in the flexibility of test application or the ease of use when generating a visual representation of the data in question.

The R programming language was originally developed by Robert Gentleman and Ross Ihaka, whose first names share the same initial, and hence the name of this computer language, as well as a play on the original computer language that preceded R, named S programming language. This main original goal of the language was to teach introductory statistics at the University of Auckland but on February 29th 2000 the first public version was released (version 1.0) as a free open-source programming language (Ihaka, 2009).

R is useful for standard statistical and graphical analysis such as: parametric statistical modelling, multivariate analysis, nonparametric statistics, smoothing and nonparametric fitting, time series analysis as well as any additional functions provided by user developed packages (Ihaka, 2009).

To better archive all the R documents, files and executables, the Comprehensive R Archive Network (CRAN) was created, and now hosts a myriad of user created packages that can be easily implemented and offer a multitude of uses to the base R experience (CRAN, 2023).

This instrument's utility is boosted by the fact that it is open source, resulting in the development of a framework supported by many independent contributors ranging from quality-of-life implementations to statistical test functions and databases. The first few months of this master thesis were devoted to the acclimatization of this computer language, learning the syntax and importing data wrangling as well as visualisation.

The R language is an expression-based language, which means the user types language expressions in the R prompt and the R interpreter evaluates said expression and returns the computed values of the expression. As an example, the user types the expression *mean* enclosing the name of the dataset in brackets and the value of the mean in the dataset is printed (Ihaka, 2009).

However, R also has its limitations, and a few highlighted by one of its creators are: the restrictiveness inherent to the single thread of execution model, the potential huge demands

made on the system's resources, highlighting the memory specifically, and consequently the rather slow time operations might take especially in element-by-element computations (Ihaka, 2009).

All R related programming and learning was performed on R Studio (now known as Posit), a computer programme known as an integrated development environment which provides the user with a smoother experience in programming by bestowing easy and intuitive user interfaces alongside a multitude of tools and features integrated in one package to ease the learning curve and increase productivity.

The R Studio interface is divided in four corners (Figure 3.1):

- Top left corner is known as the source where the user code is written. After the user writes the code needed for a certain function or dataset analysis in this area, it can be saved as an R script, these scripts can then be opened by other users or accessed at a later date.
- In the bottom left corner, we can observe the console also called the terminal, which essentially and put simply, is where the outcome of our input is displayed by the program, showing results of calculations, dataset wrangling, function execution, and much more. It also presents any errors that might have occurred and also provide a possible solution. The user can also directly execute commands on the console, although it is a transient chain of commands that resets whenever the program is closed and it is highly discouraged to perform any meaningful commands that should be saved in a script, instead reserving this tool for quick alterations or troubleshooting.
- The top right corner is called environment or history, depending on which tab is displayed. The environment displays the currently loaded objects (functions, datasets, etc…) in the memory and is not subjective to a single script and can be carried over, therefore is encouraged to clear the environment after working on a script to avoid any conflicts in nomenclature between objects in different scripts that can result in deletion of the previous object or running code on unwanted datasets. The history tab is quite self-explanatory, it presents the previous entries printed on the console/terminal area, permitting a second window into the console history without needing to scroll the actual console to a previous state.

- Lastly, in the bottom right corner a multitude of useful tabs can be observed. The files tab allows for browsing the computer's files for any saved scripts, plots or objects to be loaded into the environment. Plots tab exhibits any plotting executed by commands allowing for a visual pre-display of the finalized plot in order to iron out any imperfections or otherwise unseen plotting mistakes. Packages shows all the downloaded packages to our client and provides a quick description of the package in question alongside its current version an whether its currently loaded or not, it also allows for a quick access to the package's designated help page which will be displayed in the help tab. There is also the option to use an additional dedicated viewer pane.

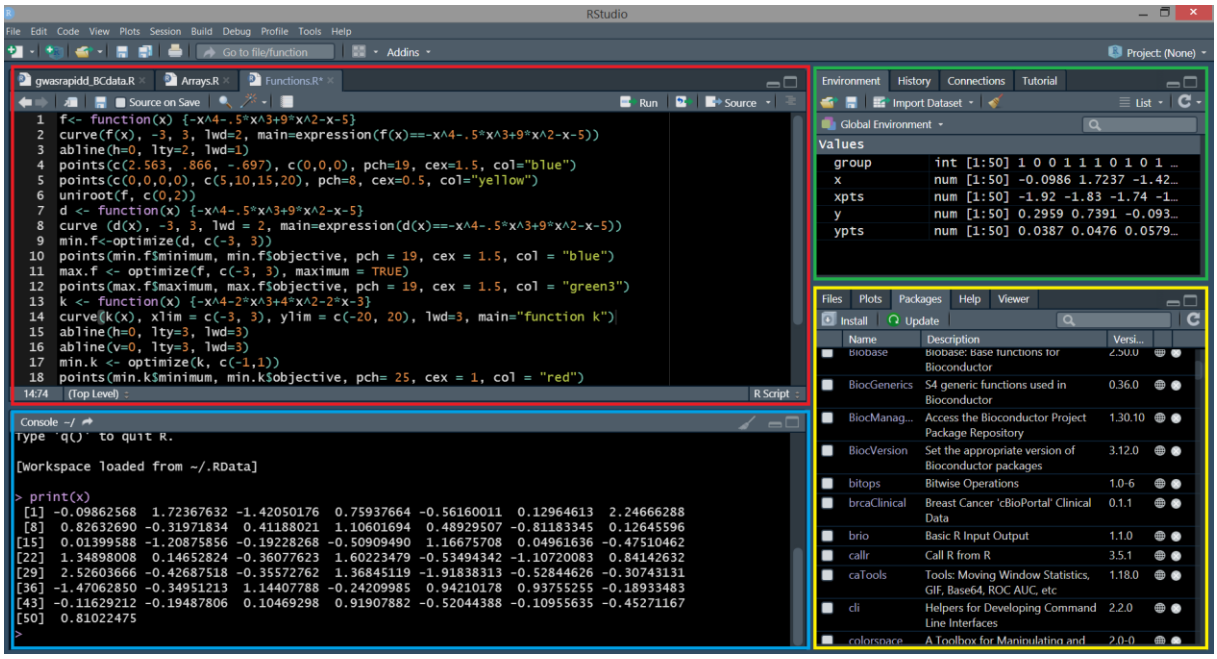

Figure 3.1: R Studio IDE user interface. In the top left corner (red) we can find the source panel displaying tabs with various open scripts and their content, bottom left (blue) is where the R studios console is located providing feedback to the user's inputs, top right (green) is currently displaying the objects loaded in the environment tab, but it is also where the console history can be reviewed, bottom right (yellow) is where we can access the files, plots, packages, help and viewer tabs.

3.2 Dataset

The data used in this thesis derived from The Cancer Genome Atlas (TCGA), a project that started in 2006 and results from the cooperation between National Cancer Institute (NCI)

and National Human Genome Research Institute (NHGRI). This project has catalogued more than 20000 primary cancer and matched normal samples spanning 33 different cancer types thus far, comprising genomic, epigenomic, transcriptomic and proteomic data. This dataset is comprised of a vast array of clinical data from patients with both healthy and tumoral tissue. The samples analysed in this study were normal matched samples (healthy tissue collected from a breast cancer patient) from breast invasive carcinoma (TCGA-BRCA). Three types of data were used:

(1) Patients' clinical data obtained from the cBioPortal via a dataset created by a fellow group member (Ramiro Magno). This dataset originated from the retrieval of all the studies that had the keyword “breast” through cBioPortalData R package;

(2) Genotyping data informing on heterozygous positions derived from microarrays genotyping data and previously processed by a fellow group member (Marinella Ghezzo);

(3) Allelic expression ratios at transcribed SNPs derived from RNA-Seq data. They were obtained by applying a log2 of the number of reads of the alternative allele by the number of reads of the reference allele. Therefore, positive numbers reflect more expression of the alternative allele, while negative numbers more expression of the reference allele. These data were previously calculated by a fellow group member (Marinella Ghezzo), resulting in data from 111 patients that were further trimmed down during this thesis work to 105 due to low per tile sequence quality, GC content and overexpressed sequences.

3.3 Filtering and annotation of RNA allelic data

The following analysis was performed in this project: The TCGA-BRCA clinical data was matched with the AE ratio data by patient ID resulting in a dataset comprised of 105 patients all women, with both clinical, genotyping and AE data (at expressed heterozygous variants, aeSNPs) The following filters were applied individually for each aeSNP at each patient: Total number of reads $\geq$ 11 and Phred-Scaled quality score $\geq$ 30. Finally, aeSNPs heterozygous for less than 10 samples (out of the 105 patient’s) were also excluded from the analysis. We also removed any variant that showed a Phred-Scaled quality score under 30 since this value represents 0.1% error margin (Figure 3.2).

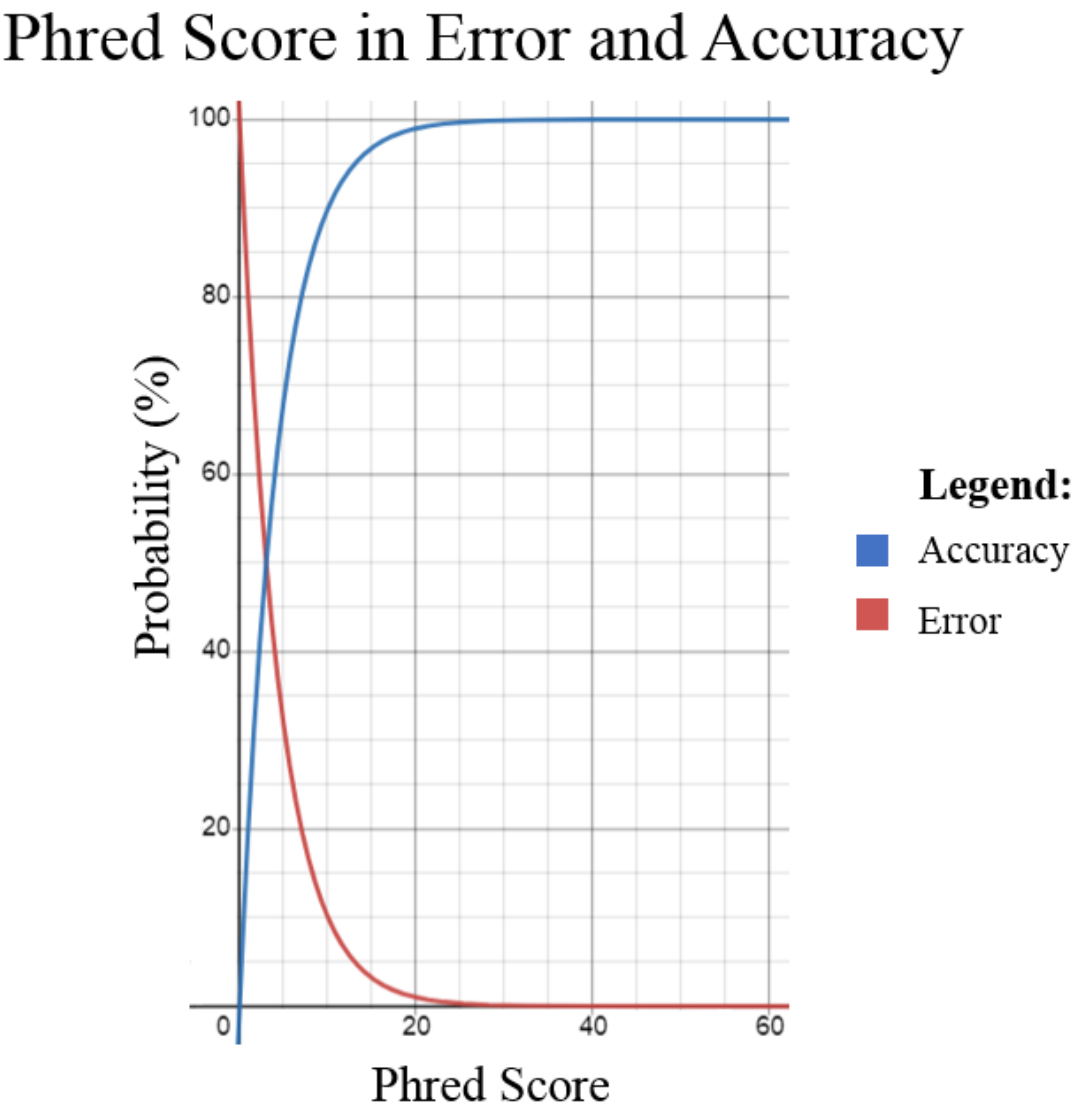

Figure 3.2: Percentile accuracy and error according to Phred-Scaled Quality Score. On the x axis it is represented the Phred-scaled quality score and the y axis displays the probability for either measurement in percentage. The blue line dictates how the percentage of accuracy a Phred score is entitled to whereas the red line indicates the opposing error percentage of the same Phred score value.

Annotation of genes (ensembl ID and HUGO Gene Nomenclature Committee (HGNC) symbol) where the aeSNPs analysed were located was retrieved using a BiomaRT extension in R ("biomartr" package version 1.0.2).

Table 3.1: Summary of studies and variants previously associated with risk for BC subtypes or survival. All studies identified by Gwasrappid according to the traits selected are displayed here. This package provides Study ID and Pubmed ID as well as publication date and author. Reported trait associated with the study and variants of interest identified in the study are also included.

Study ID	Pubmed ID	Publication date	Author	Reported Trait	Variants
GCST001373	22232737	2012-01-09	Shu XO	Breast cancer (survival)	rs3784099 rs9934948
GCST002727	25526632	2014-12-19	Rafiq S	Breast cancer (survival)	rs1728400 rs12358475 rs421379
GCST002861	25890600	2015-04-18	Guo Q	Breast cancer (survival)	rs148760487 rs2059614 rs7149859
GCST002903	25964295	2015-05-11	Khan S	Survival in breast cancer (ER positive)	-
GCST002902	25964295	2015-05-11	Khan S	Survival in breast cancer (ER negative)	-
GCST002901	25964295	2015-05-11	Khan S	Survival in endocrine treated breast cancer (ER positive)	rs8113308
GCST005106	29158497	2017-11-21	Kadalayil L	Breast cancer (survival)	-

3.4 Retrieval of GWAS hit variants associated with BC subtypes and outcomes

To retrieve variants previously associated with breast cancer subtypes or clinical outcome, we recurred to gwasrapidd (version 0.99.12), by searching for breast cancer studies using efo_trait = “breast carcinoma”. For all retrieved studies, we collected their reported traits and filtered them to include only traits that reported on receptor status or survival. Then, for each study ID the associated variants were retrieved (Table 3.1).

Using ensemblr (version 0.0.1) we searched the 1000 Genomes phase 3 EUROPEAN population for SNPs in moderate to strong LD ($r^2 \geq 0.4$) with the gwasrapidd retrieved SNPs. As a final step, we ran a BiomaRT search on every SNP (LD SNPs included) to obtain gene information.

3.5 Statistical Tests

We conducted a step-by-step statistical test flowchart on AE ratios data from normal-matched breast tissue samples (Figure 3.3). This flowchart included first measuring the effect size (by a Hedge's G test) between groups of samples according with their matched tumour samples receptors expression profile (ER, PR and HER2), in 8039 aeSNPs.

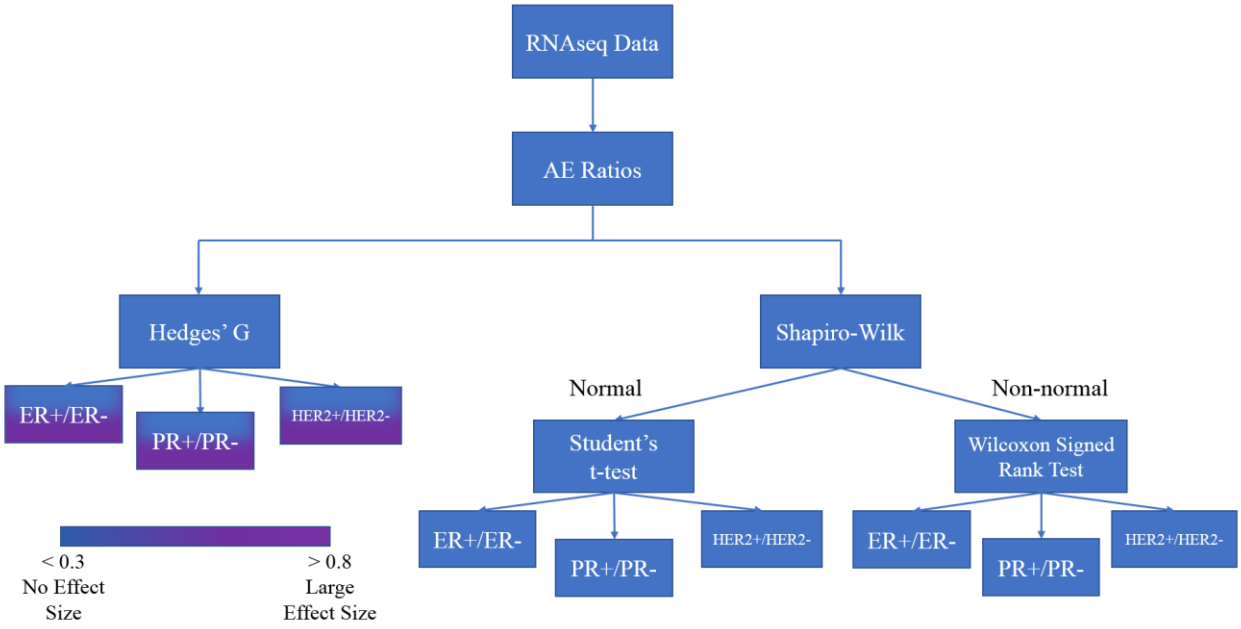

Figure 3.3: Data flowchart. Beginning with the RNAseq data transformation into easily usable AE ratios performed by a fellow colleague, the ratios would be separated into two distinct analyses with different requisites and therefore an unequal number of starting samples. Hedge's G was much stricter in its requirements and as a result had a lower input, the main goal was to quantify the difference between positive and negative receptor statuses in a set SNP. On the other hand, the other path led to a Shapiro-Wilk test, preformed to determine the data normality or lack of it, and have a more fitting test applied, Student's t-test on normality following data and Wilcoxon Rank-Sum Test. Both these tests would indicate whether the positive group of a receptor would be statistically different to its negative counterpart.

In parallel, a Shapiro-Wilk normality test was applied to assess whether the AE ratios at each aeSNP showed a normal distribution. Then, depending on the result from the Shapiro-Wilk test, a Student's t test or a Wilcoxon signed rank test was applied to determine if the AE ratios at groups of samples with different receptors expression profile (ER, PR and HER2) at their matched tumour samples showed the same mean AE ratios (null hypothesis) or if they were statistically significantly different.

3.5.1 Shapiro-Wilk test

To verify whether AE ratios followed normality or not, we implemented a Shapiro-Wilk test developed by Samuel Sanford Shapiro and Martin Wilk (Shapiro & Wilk, 1965), with the aim of identifying which data was fit for a parametric test (Student's T-test) or a non-parametric test (Wilcoxon Rank-Sum Test).

This test assesses data normality via the following formula:

$$W = \frac{(\sum a_i \, x_i)^2}{\sum (x_i \, - \, \mu)^2}$$

Where *W* score can be obtained by dividing the square sum of the a_i value (which represents the normal order statistics providing the best fit line for a normal dataset) times x_i by the sum of x_i minus the mean squared.

If the *W* score is equal to 1, this means that the dataset follows a normal distribution, and any deviation from 1 to be accepted as significant, must be inside the confidence interval agreed.

The null hypothesis states that the sample originated from a normally distributed population which is rejected when $p \leq 0.05$ with 95% confidence and values above this threshold equate to a distribution not unlike normal. This test was performed with shapiro.test function from the "stats" R package (version 3.6.2).

3.5.2 Student's t-test

With the parametric dataset isolated, the next step was to compare AE ratios between positive and negative receptor status populations for ER, PR and HER2 in each aeSNP. Since the separation of the samples into two subgroups would reduce each population's samples size, we decided to apply a Student's t-test which is ideal for groups under 30 samples, a test first published in 1908 by William Sealy Gosset under the pseudonym Student (Student, 1908). This test aims to answer if the mean difference between two populations is statistically significant enough, where the null hypothesis states that both populations are statistically equal and consequently the alternative hypothesis affirms that both populations are not statistically equal. This statistical tool can be further subdivided depending on the dataset in question: one-sample t-test, paired samples t-test and unpaired samples t-test. For this study we employed the latter, also being aptly named unpaired t-test to compare between independent positive and negative receptor populations to identify possible differences in the means between both groups and draw a plausible correlation (Mishra et al., 2019).

To answer the hypothesis, we first need to obtain a t-score:

$$t - score = \frac{Sample\ Means\ Difference}{Standard\ Error\ of\ Sample\ Means\ Difference}$$

Where the sample means difference can be obtained by subtracting one group's means to the other, and the Standard Error of Sample Means Difference (SE Difference):

$$SE\ Difference = SD\ Difference\ \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The SE difference can be obtained by multiplying the square root of the sum between the inverse of both populations size $\left(\frac{1}{n}\right)$ by the Standard Deviation Difference (SD Difference) between both samples, which in turn can be calculated by:

$$SD\ Difference = \sqrt{\frac{(n_1 - 1)\ \times\ SD_1^2 + \ (n_2 - 1)\ \times\ SD_2^2}{(n_1\ +\ n_2 - 2)}}$$

With the final calculated t-score it is now possible to obtain a p-value with the aid of a t-score/p-value calculator or a t-distribution table, by providing both t-score and degrees of freedom, rejecting the null hypothesis if said p-value is under the selected significance level (Whitley & Ball, 2002)

Only aeSNPs with at least 8 samples in each analysed group were considered, to guarantee comparison between two non-zero sizeable populations. We performed an unpaired two-sample Student's t-test with 95% confidence interval using the t.test function from "stats" package (version 3.6.2). False discovery rate (FDR) of 10% was used to correct p-values for multiple testing.

3.5.3 Wilcoxon Rank-Sum Test

In order to test the significance of any aeSNPs in different receptor status present in the non-parametric data we decided to apply the Wilcoxon Rank-Sum Test or Mann Whitney U test created by Henry Berthold Mann and Donald Ransom Whitney published in the *Annals of Mathematical Statistics* (Mann & Whitney, 1947). This test is especially ideal when the sample distribution is between the interval of 5-20 allowing the comparison of two non-parametric groups with different sample sizes.

Once more, we define the null hypothesis as the two population's means being equal and the alternative hypothesis states that the two population's means are not equal. The first step in a Wilcoxon Rank-Sum Test (Wilcoxon Mann-Whitney U) is to rank all the ratios from lowest to highest in both positive and negative populations, attributing the respective number in the order the results come in. In the next step, all the ranks are summed in their specific population (be it positive or negative) and with this rank sum (R) it is now possible to determine the Wilcoxon Mann-Whitney U score (U) of each group:

$$U_1 = \ n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \ R_1$$

And with a calculated U it is now possible to determine, with the aid of a table of critical values for U, if this score (when considering both group sizes and level of desired significance) is pronounced enough to declare both groups to be statistically different from each other.

The exact p-value can be obtained by singling out the exact rank combinations between groups that would equate to a significant difference in their values (such as ranks 1-8 being positive and ranks 9-16 being negative, given that it is two-sided), dividing the number of significant combinations by the total number of possible rank combinations, determined by the following equation:

$$p - value = \frac{Number\ of\ Significant\ Rank\ Combinations}{\frac{(n_1 + n_2)!}{n_1! \times n_2!}}$$

Only aeSNPs with AE ratios available for at least 8 samples in each group were tested for association. To perform the calculations, we used the wilcoxon.test() function in R, providing both groups x and y and making sure the paired parameter is set to FALSE to induce a Mann-Whitney U test as stated in the "stats" package documentation (version 3.6.2). False discovery rate (FDR) of 10% was used for multiple testing corrections.

3.5.4 Estimation of AE effect size at each aeSNP

To measure the effect size of AE ratios differences at each aeSNP according to the previously mentioned receptors, we applied the Hedges' G statistic, since it surpasses Cohen's D when sample sizes are below 20. This statistic was first introduced by Larry Vernon Hedges in *Distribution Theory for Glass's Estimator of Effect Size and Related Estimators* (Hedges, 1981).

In this test the effect size is called *g* and it can be calculated by dividing the difference between both groups' means by the pooled standard deviation (s_ρ), translated in the following equation:

$$g = \frac{\mu_1 - \mu_2}{s_\rho}$$

And by knowing the standard deviation (*s*) in each group as well as the sample size (*n*), we can then calculate s_p:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

With the Hedges' g it is now possible to evaluate the effect size between populations, a 0.3 score correlates to a low effect size, a 0.5 g score indicates medium effect size and any score above 0.8 is deemed a large effect size.

This test was performed in R with the function hedges_g() from dabestr package (version 0.3.0). Only aeSNPs with at least 10 samples in each analysed group were tested. AE ratios were considered to differ between two groups when the bias-corrected and accelerated (BCA) confidence interval did not cross 0, in order to ensure that the difference concerning both groups allelic expression was unequivocally one-sided. Furthermore, groups were considered to display large differences when $|g \geq 0.8|$.

3.5.5 Two-Stage Hazard Rate Comparison

AeSNPs whose AE ratios were found associated or differ between samples according with the patient's ER, PR, or HER2 receptors statuses, were selected to be tested for association with patient's survival. In order to compare disease free survival (DFS) in relation to AE ratios we divided the population for a specific aeSNP in two groups, one half with the highest AE ratios and therefore the highest expression of alternative allele and the other half with the lowest AE ratios that equate to a higher expression of reference allele.

To compare both populations' survival rates we employed the Two-Stage Hazard Rate Comparison (TSHRC) procedure (Figure 3.4). This procedure is flexible in solving cases were the hazard rates cross each other while also being capable of accounting for non-crossing different hazard rates, features that are usually exclusive in hazard rate comparison methods. This is accomplished by firstly applying a log-rank test which is the standard procedure to identify whether two hazard rates cross each other. Then, if they don't cross, or a crossing difference can be detected, the procedure ends.

On the other hand, if they cross and as a result the positive and negative differences are cancelled out, an independent (from log-rank test) second stage test specifically designed to identify crossing difference is applied, further classifying both hazard rates as identical or just crossing (Qiu & Sheng, 2008). This was accomplished in R with the twostage() function provided by "TSHRC" package (version 0.1-6).

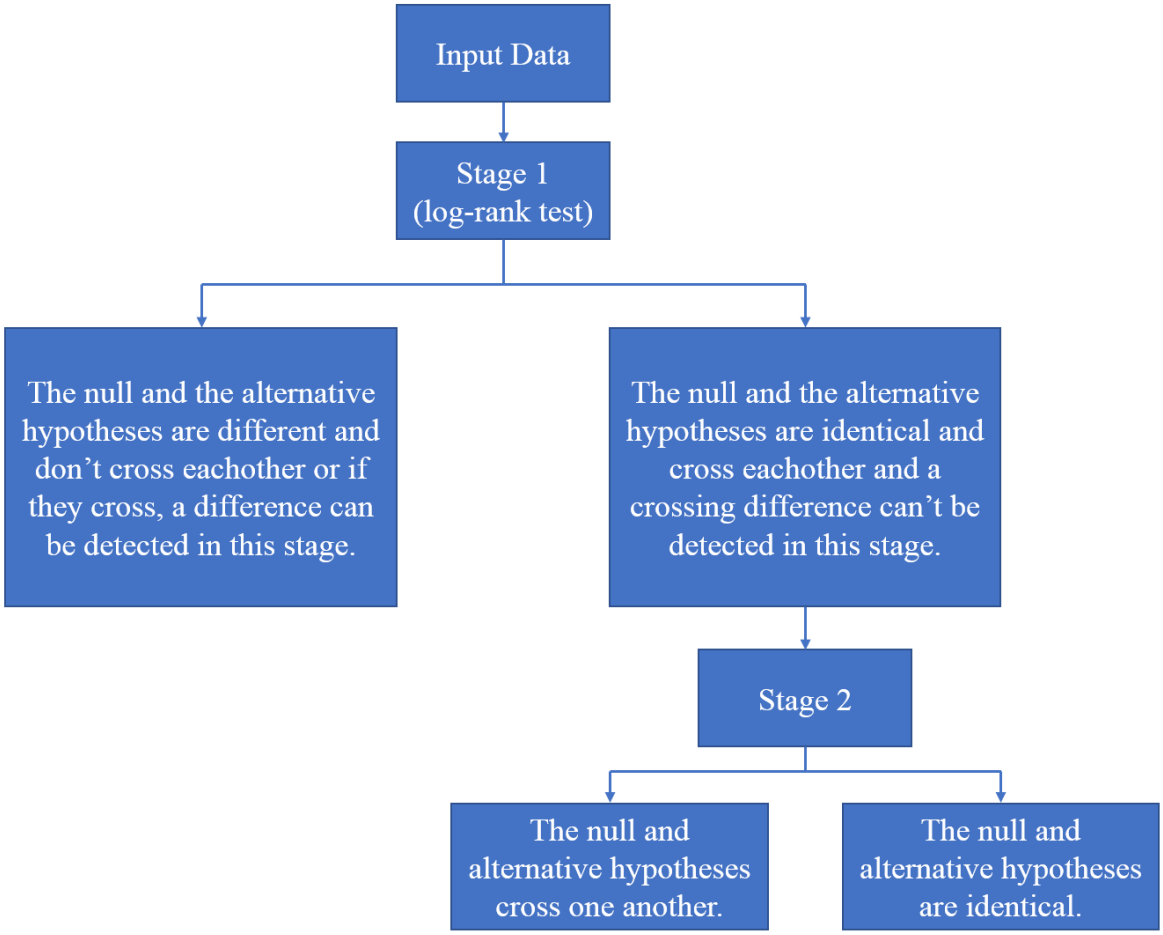

Figure 3.4: Flowchart exemplifying the Two Stage Hazard Rate Comparison method. TSHRC firstly preforms a log-rank test to detect possible crossings and differences between groups in other to confirm or reject the null hypothesis, leading to a stage 2 procedure unique to TSHRC to further scrutinize the data and maintain the null hypothesis or contest it. Adapted from (Qiu & Sheng, 2008).

3.5.6 Survival Curves

Kaplan-Meier survival curves were drawn in R Studio by first selecting the significant results from the Two Stage Hazard Rate Comparison test and through the R survival package (version 3.4-0) insert the survival time and vital status values from the relevant patients in the

survfit() command, where Disease-Free Survival was chosen as the time variable and Patient Status to indicate the vital status of the patient, the function output was then plotted via ()plot allowing for a visualization of the data.

In the Kaplan-Meier two curves are present, the blue curve represents the population with higher AE ratios in that SNP, translating to a higher expression of alternative allele whereas the red curve indicates lower expression of alternative allele and overall AE ratios, meaning a higher expression of reference allele in that SNP, with aim of indicating how would the differential allelic expression of both alleles reflect on patient Disease-Free Survival statistic.

3.5.7 Testing for three or more groups

In order to address clinical features with more than 2 variables (such as tumour grade and stage) we required more fitting statistical tests.

For the normal-like data according to Shapiro-Wilk test, we employed the one-way ANOVA statistical test in R studio with the aid of the anova() function from the car package (version 3.1-1), in an attempt to better understand how the AE ratios would reflect on these clinical features.

In the case of the non-parametric data, the Kruskal Wallis H test was preformed via kruskal.test() function available in the stats R package (version 3.6.2) being a preferred option to ANOVA in non-parametric data.

However, there wasn't enough patient data in non-binary clinical features, resulting in small or even empty groups being compared in each SNP, which led to the omission of this data from the final work.

3.5.8 Gardner-Altman plots

For a better visualization of the Hedges' G output data, we opted to display the test results in Gardner-Altman plots, being composed by:

- A swarmplot of all datapoints in the relevant group and in concordance to its distribution.

- A 95% confidence interval bootstrapped effect size being aligned to the negative receptor status group mean.

The Gardner-Altman plots were created in R Studio with the same package applied to find the Hefdge's G side effect (dabestr package version 0.3.0) by first creating a dabest object through the dabest() function followed by the function hedges_g() on that same dabest object of interest, finalizing with a standard plot() function specifying the groups being compared with color.column = receptor status.

To obtain a similar grouped plot display exhibited in this thesis we used the plot_grid() function from the cowplot R package (version 1.1.1) as referred by the developers in the dabestr's CRAN page.

Results

4 Results

4.1 Establishing the population dataset

To assess if cis-regulatory variants may determine clinical and molecular characteristics of breast tumours we analyse the association of AE with tumour characteristics in normal breast tissue samples from breast cancer patients. The dataset analysed comprised a subset of 105 patients with breast cancer from the TCGA project for which there was RNA-seq from normal-matched tissue available and previously processed to generate allelic counts for heterozygous germline SNPs (aeSNPs).

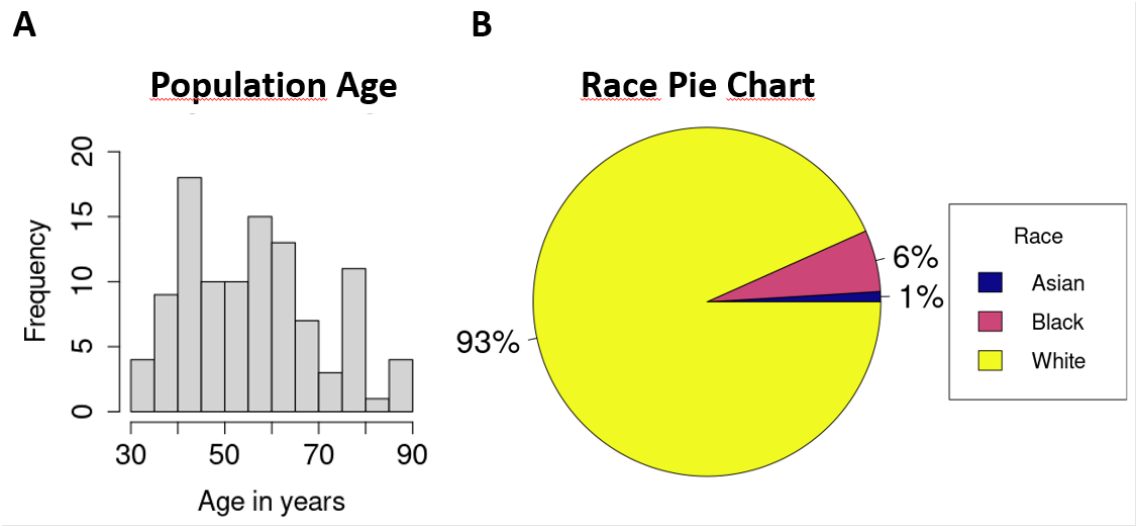

Figure 4.1: Distribution of Studied population Age and Race. A) histogram showing that the age of the individuals distributes between 30 to 90 years of age (in the x axis) with the bars representing ranges of ages of 5 years, In the y axis shows how many individuals are present in each 5-year range. **B**) The population's is mainly White followed by Black and Asian in that order (105 patients).

The analysed dataset is comprised of 105 patients mainly of white ethnic origin, spanning between the ages 55-65 (Figure 4.1). Concerning breast cancer subtypes, the most common is Luminal A followed by Luminal B (Figure 4.2 A), and accordingly, both progesterone and estrogen receptor statuses are mainly positive whilst HER2 status predominantly negative (Figure 4.2 B-D).

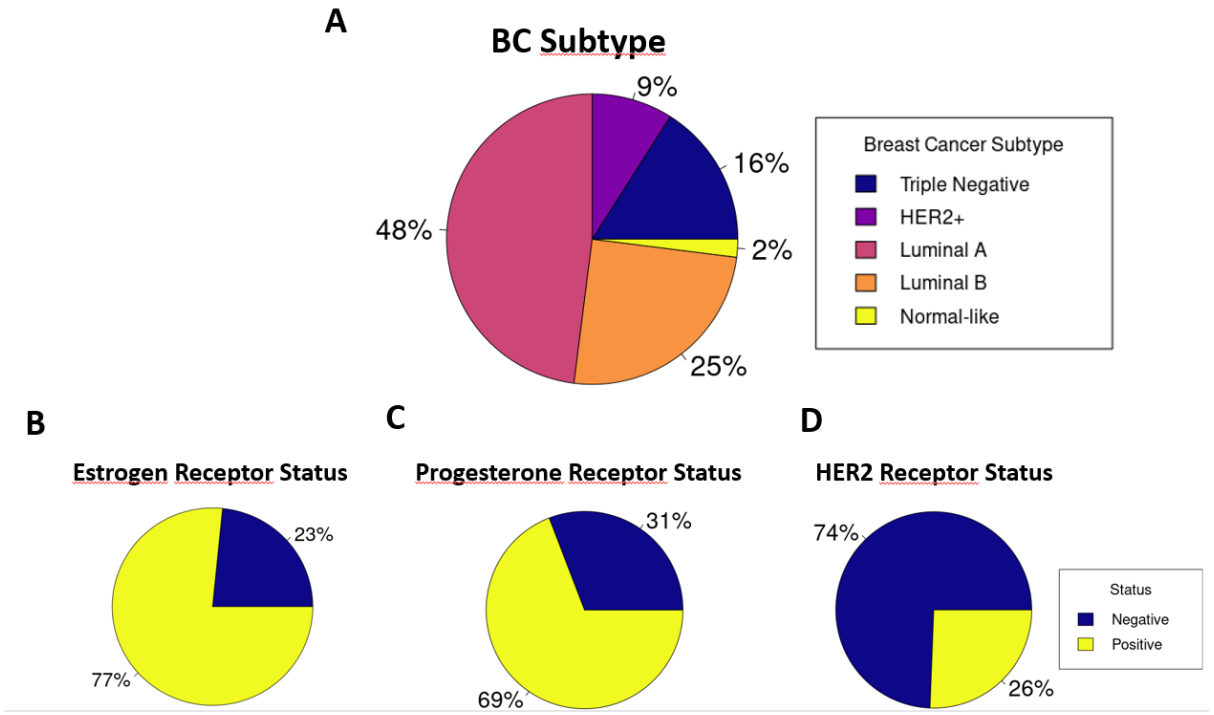

Figure 4.2: Sample characterisation according to breast Cancer Molecular Subtype (PAM50) and Receptor's Statuses. In **A**) it is represented a breast cancer subtype pie chart, where Luminal A is the most common and represented by magenta, followed by Luminal B in orange, Triple Negative in dark blue, HER2+ in purple and normal-like in yellow while also being the rarest. It is also represented the percentage of tumours according to estrogen **B**), progesterone **C**) and HER2 **D**) status. Yellow represents positive and dark blue represents negative tumours for that specific receptor.

Concerning the transcribed bi-allelic variants for which there were genotype information (aeSNPs), the vast majority were heterozygous in less than 5 samples (Figure 4.3 A), half of these aeSNPs didn't reach 11 total reads (Figure 4.3 B) and had a sub-optimal Phred-scaled quality score (Figure 4.3 C). Consequently, to obtain high-confidence AE ratios in a minimum number of samples, we ought to trim all the 64677 aeSNPs that had less than 10 heterozygous samples (Figure 4.3 D) and were under 11 total reads (Figure 4.3 E) as well as to exclude any aeSNPs that displayed a Phred-scaled quality score of less than 30 (Figure 4.3 F) which corresponds to a 0.1% margin of error. Since our next step would imply the use of statistical tests, we also removed every sample that showed an infinite allelic expression ratio value (indicative of mono-allelic expression of either the alternative or the reference allele), resulting in a final dataset composed of 8039 aeSNPs (Figure 4.4 A) annotated to 4942 genes.

These variants displayed AE ratios ranging between -3 and 3 although the majority of values were between 1 and –1 (corresponding to 2-fold differences between the expression of the two alleles (Figure 4.4 B).

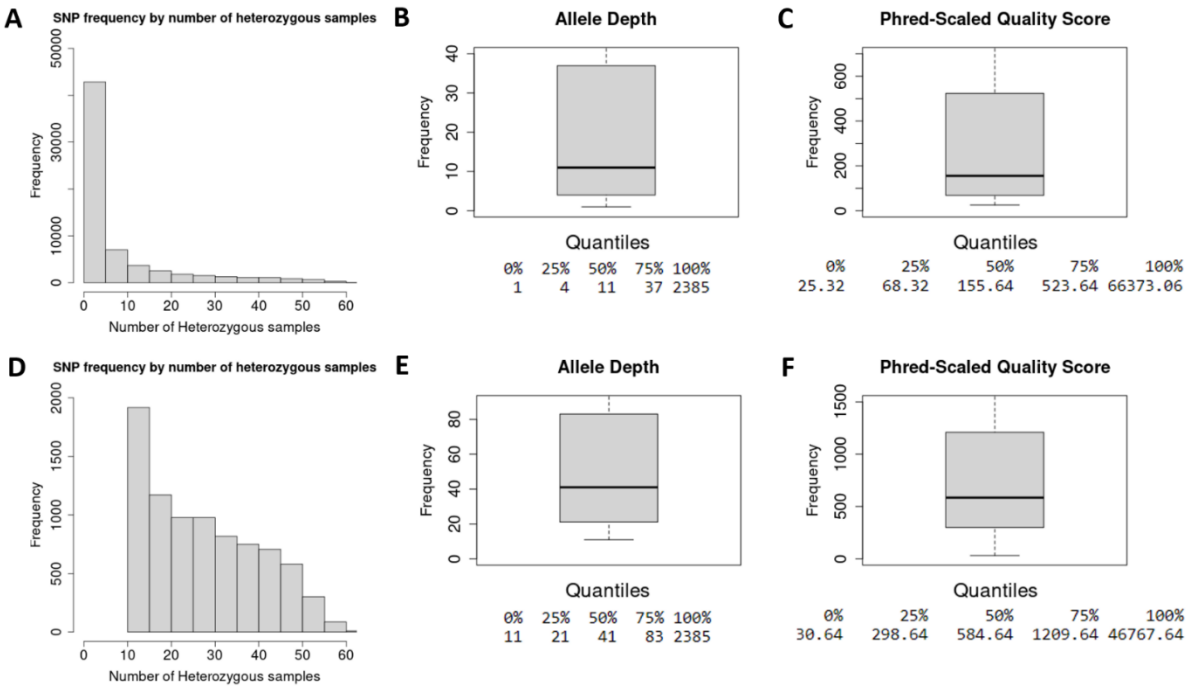

Figure 4.3: Dataset sample numbers, read depth and quality. This figure shows the before and after dataset filtering average sample number distribution by aeSNP **A**) and **D**), it also indicates how many reads each aeSNP in the dataset has compiling them into quantiles **B**) and **E**) as well as the average Phred-scaled quality score **C**) and **F**).

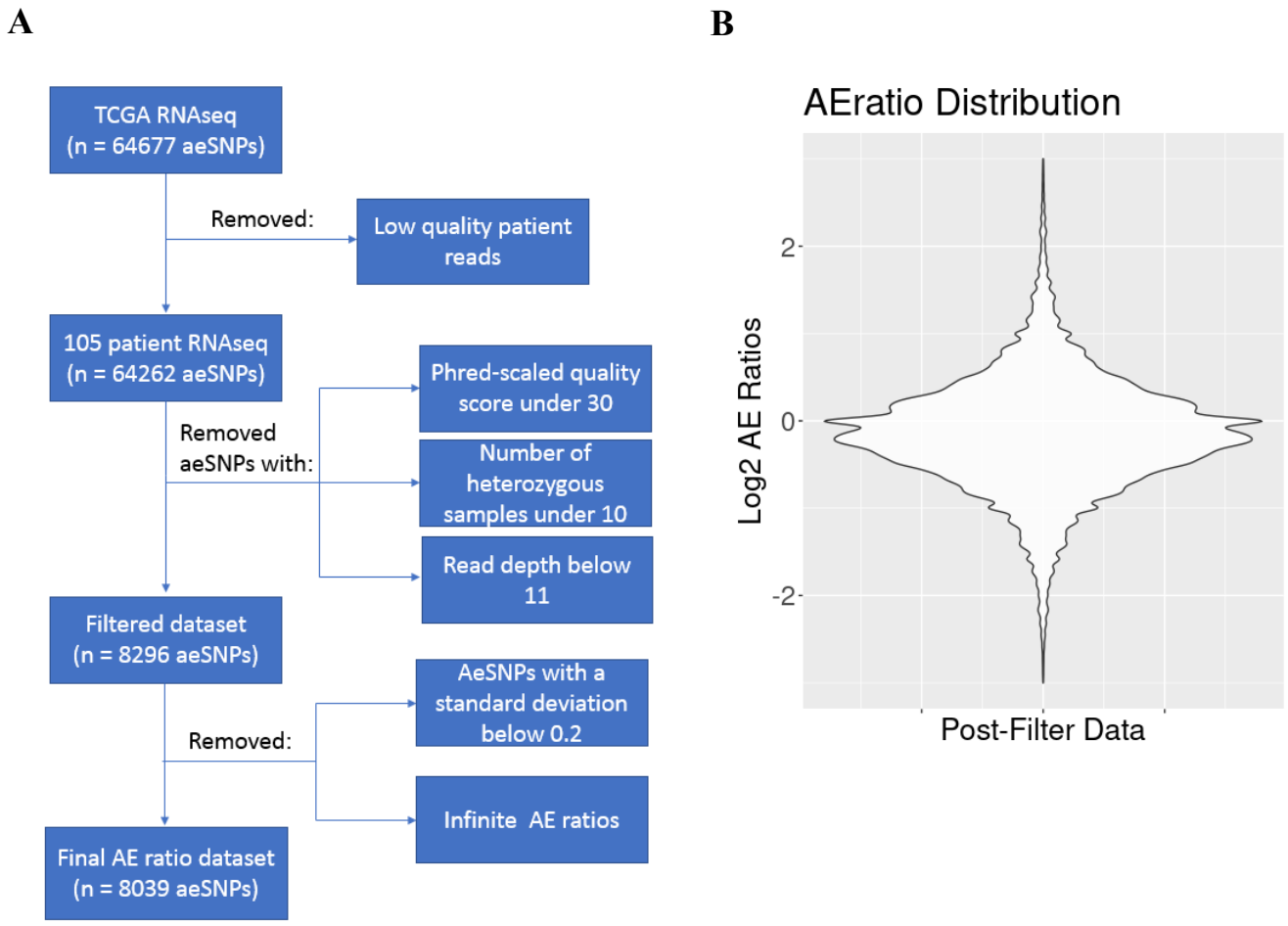

Figure 4.4: Filters applied to the RNA-seq data and final AEratio distribution. This image demonstrates the flowchart of all steps and filters previously described to reach the final AE ratio dataset used in tests and calculations **A**) as well as the distribution of the $\log_2$ AEratios in the dataset **B**).

4.2 Mathematical assessment of candidate risk aeSNPs

To correlate the clinical data and AE ratios at normal breast tissue, we ran statistical tests to mathematically predict an association between differential allelic expression and a possible clinical outcome. To correctly choose the statistical test applicable to our data we need to determine whether it meets certain assumptions, such as normality, therefore we firstly implemented a Shapiro-Wilk test to evaluate the data's normality. From that preliminary test we concluded that AE ratios at 6761 aeSNPs follow a normal distribution (Shapiro-Wilk test p-value > 0.05) whereas 1278 aeSNPs do not (Shapiro-Wilk test ≤ 0.05).

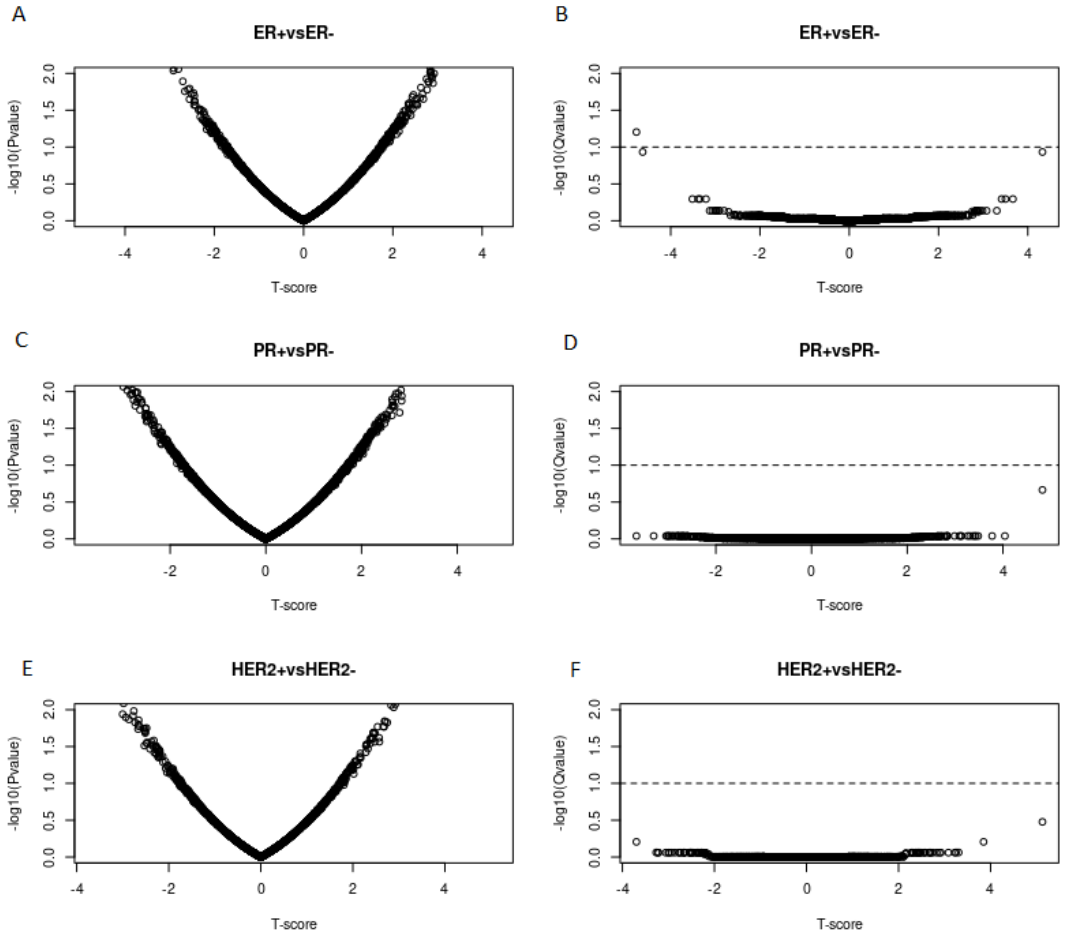

Figure 4.5: Summary of Student's t-test results for AE ratios association with receptor status. Are shown volcano plots displaying in the Y axis p-values and q-values for **A**) and **B**) ER status, **C**) and **D**) PR status, **E**) and **F**) HER2 status. It is also shown the T-score correspondent to the standard deviation from the mean, negative values tend toa higher alternative allele expression in positive receptor status whilst positive t-scores deviate to negative receptor status. A dashed horizontal line indicates the requisite for significance (q-value ≥ 0.1).

We employed a Student's t-test to the 6761 aeSNPs showing AE ratios normally distributed, to test for differences in mean AE ratios related to sample/tumour status in each receptor. We identified AE ratios at rs3764859 (one variant located at *COQ6* and *ENTPD5* genes) (Figure 4.5 A and 4.5 B) significantly associated with ER status after correcting for multiple testing (q-value ≤ 0.1). No significant association of AE ratios was found for PR status or HER2 status (Figure 4.5 C-F).

We analysed the association of AE ratios with receptor status in the remaining SNPs that follow a non-normal distribution using a Wilcoxon Rank-Sum test. These association results are summarized as volcano plots (Figure 4.6).

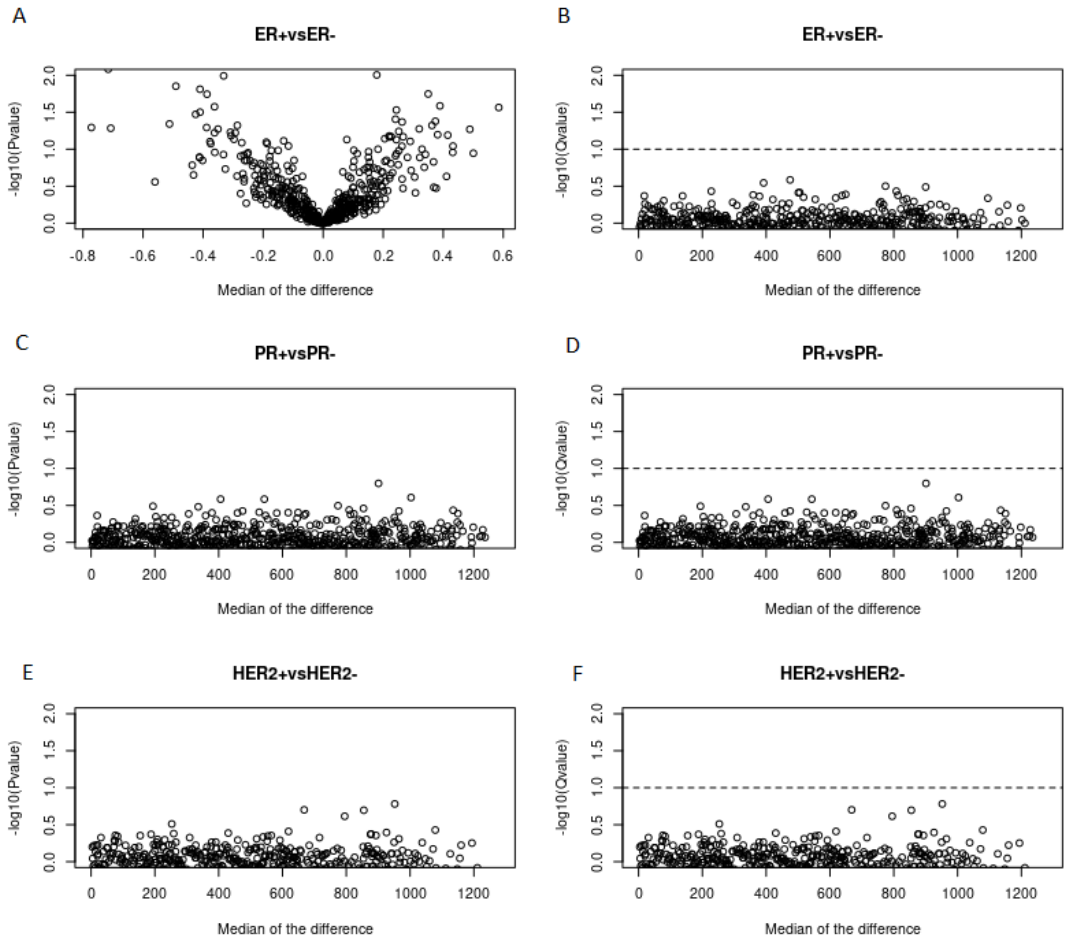

Figure 4.6: Summary of Wilcoxon Rank-Sum results for AE ratios association with receptor status. This figure exhibits volcano plots with both p-values and q-values in the Y axis inherent to: ER status **A**) and **B**), PR status **C**) and **D**), HER2 status **E**) and **F**). The median difference is portrayed by the X axis and the dashed horizontal line in qvalue plots represents the significance threshold (qvalue ≥ 0.1).

As evidenced by the plots, no significant AE association is found in this analysis, after correction for multiple testing (all q-values > 0.1).

4.3 Receptor status in differential allelic expression

In order to assess if the AE ratios at aeSNPs differ between samples according to the patient's tumours hormone receptors statuses, we implemented Hedge's G statistic to our data, regardless of normality. From those criteria 207 aeSNPs showed differences between groups: 63 for ER status, 119 for PR status and 55 for HER2 status, with some (38 aeSNPs in ER, 60 in PR and 33 in HER2) showing a large effect size (Effect Size ≥ 0.8) (Figure 4.7). For ER status the highest absolute effect sizes were obtained for rs43216, rs2251219 and rs2447097 located in *NLRC5*, *PBRM1-SMIM4* and *SGSM2* (from highest to lowest), respectively. Regarding PR status, the highest AE ratio differences were observed at rs7301360, rs10896 and rs884510 aeSNPs annotated to *APOLD1*, *BORCS8-MEF2B* and *PPP1R1A-PDE1B*, respectively. And finally, for HER2 status the highest Hedges' G were obtained at rs990626, rs3750163 and rs12452567, aeSNPs located in *LRP2*, *TNS3* and *METTL16*, correspondingly.

Figure 4.7: Confidence interval range by Hedges' G effect size in ER, PR and HER2. The confidence interval range shown in the y axis, was calculated by subtracting the upper bias-corrected accelerated (BCA) bootstrap by the lower limits and the x axis is comprised of the Hedges' G effect size. The blue dots correspond to samples that cross 0 somewhere between the upper and lower BCA range whereas the red dots do not.

AE ratios at some aeSNPs differ between samples according to more than one receptor status, resulting in the overlap of: 30 aeSNPs for ER and PR status and 3 aeSNPs for PR and HER2 status (Figure 4.8).

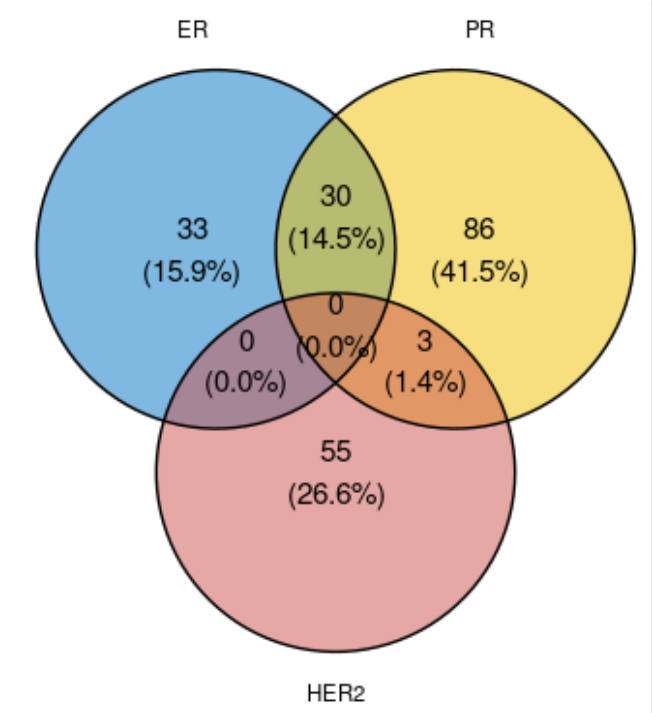

Figure 4.8: Venn diagram showing the overlap between Hedges' G significant AE ratios at aeSNPs by receptor. Estrogen receptor is represented in blue, PR in yellow and HER2 in pink while the ER/PR shared aeSNPs are coloured green, PR/HER2 orange and ER/HER2 in purple. The centre of the Venn diagram depicts the junction of all three receptors.

Regarding aeSNPs whose AE ratios differ according to both ER/PR (Figure 4.9 - 4.10) or HER2/PR status (Figure 4.11), 10 aeSNPS display large effect size regarding both receptors, of which three, located at *SBNO1*, *MTHFD1* and *COL4A1*, show higher expression of the alternative alleles in ER and PR negative tumours which are known to have a worse prognosis, than in ER and PR positive tumours. Regarding allelic expression at rs4312169 located at *MTMR6*, PR and HER2 negative samples shows higher expression of the alternative allele when compared with PR and HER2 positive samples (Table 4.1). We also compiled all large effect size aeSNP's data in all 3 receptors for easier consultation (Table 4.2 – 4.4).

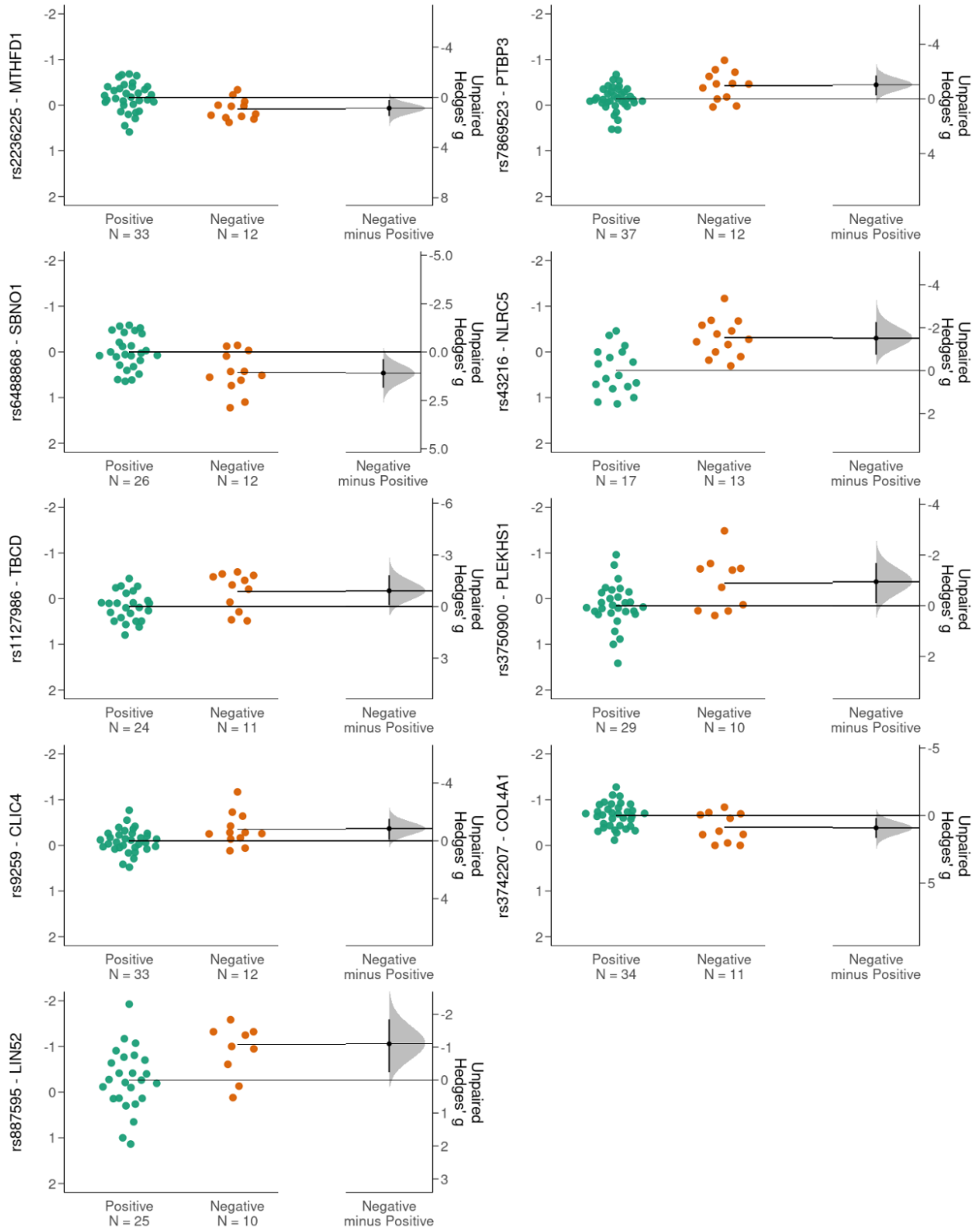

Figure 4.9: Gardner-Altman plots of 9 aeSNPs with an effect size higher than 0.8 in at least two receptors (ER-PR) according to ER status. The left y axis represents the AE ratio values and each point corresponds to that specific ratio in a single sample, the observations are divided in two groups: positive, which represents ER positive sample and negative correlating to an ER negative sample. The y axis present on the right-hand side shows the unpaired Hedges' G effect size (mean difference) as a point estimate and a vertical bar that indicates its 95% confidence interval. Note that negative AE ratios equate to higher reference allele expression.

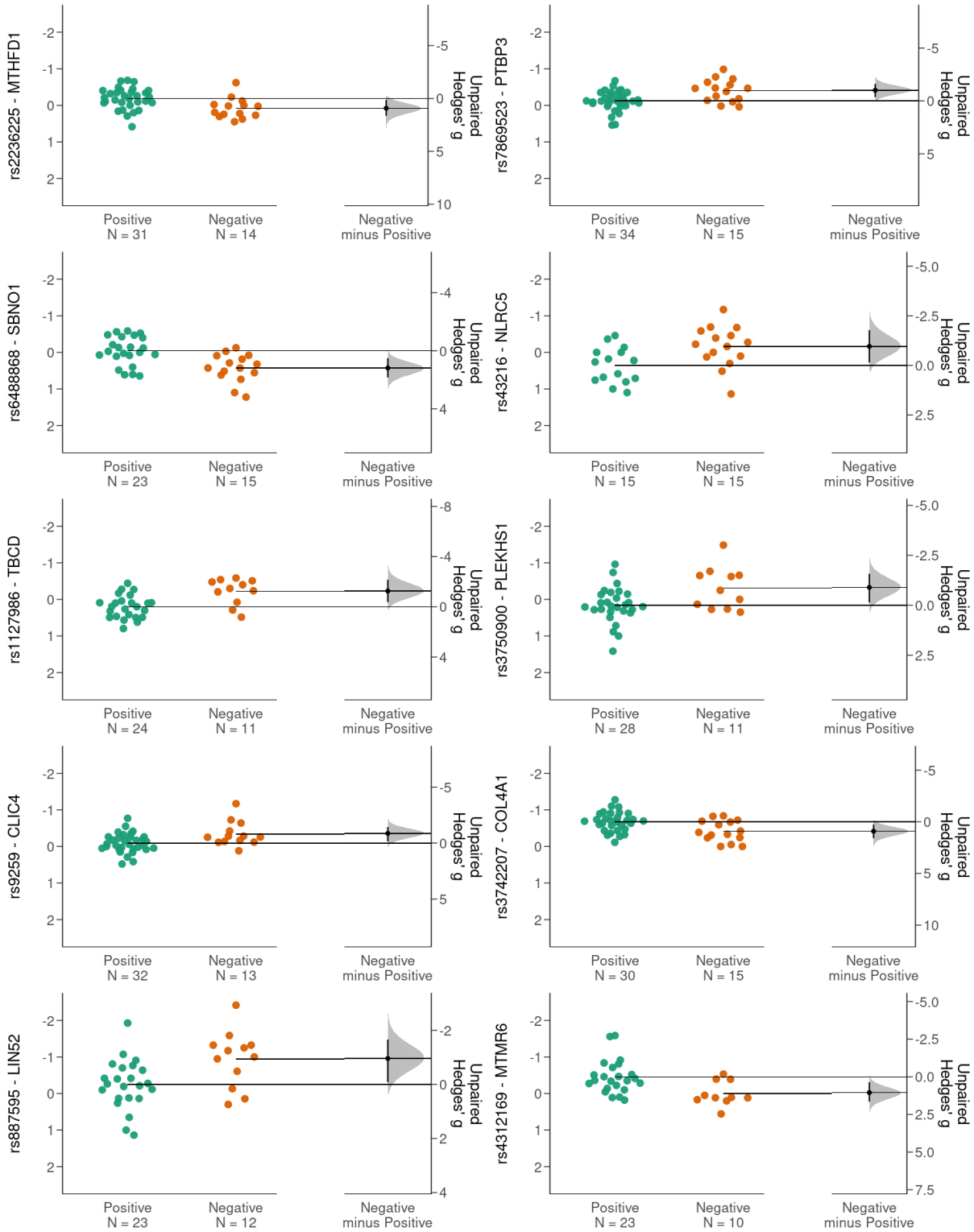

Figure 4.10: Gardner-Altman plots of 10 aeSNPs with an effect size higher than 0.8 in at least two receptors (ER-PR) according to PR status. The left y axis represents the AE ratio values and each point corresponds to that specific ratio in a single sample, the observations are divided in two groups: positive, which represents PR positive sample and negative correlating to a PR negative sample. The y axis present on the right-hand side shows the unpaired Hedges' G effect size (mean difference) as a point estimate and a vertical bar that indicates its 95% confidence interval. Note that negative AE ratios equate to higher reference allele expression.

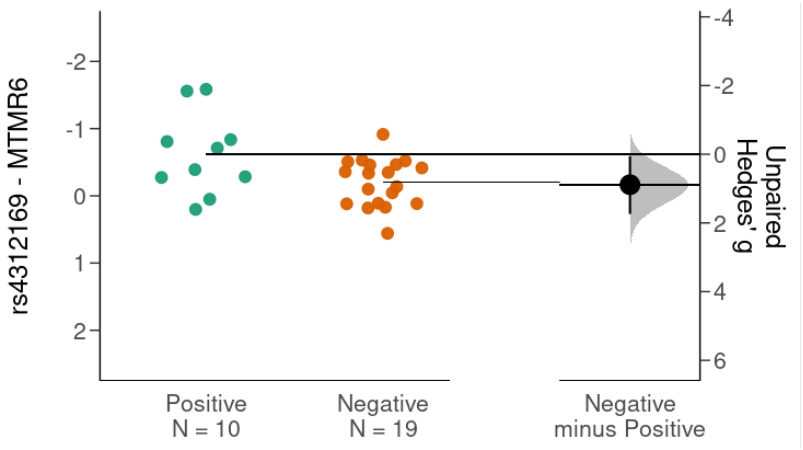

Figure 4.11: Gardner-Altman plots of an aeSNP with an effect size higher than 0.8 in at least two receptors (HER2-PR) according to HER2 status. The left y axis represents the AE ratio values and each point corresponds to that specific ratio in a single sample, the observations are divided in two groups: positive, which represents HER2 positive sample and negative correlating to an HER2 negative sample. The y axis present on the right-hand side shows the unpaired Hedges' G effect size (mean difference) as a point estimate and a vertical bar that indicates its 95% confidence interval. Note that negative AE ratios equate to higher reference allele expression.

Table 4.1: aeSNPs displaying large effect sizes associated with status at two receptors. This table compiles the aeSNP ID and correspondent gene in HGNC symbol format along with Hedges' G metrics (receptor negative mean minus receptor positive mean divided by pooled standard deviation) according to ER, PR and HER2 receptor status populations, together with reference and alternative alleles for each variant.

				Hedges' G Difference		
SNPID	**Gene**	**Reference Allele**	**Alternative Allele**	**PR**	**ER**	**HER2**
rs4312169	*MTMR6*	T	C	1.05	-	0.89
rs887595	*LIN52*	A	G	-0.96	-1.10	-
rs3742207	*COL4A1*	T	G	0.91	0.91	-
rs9259	*CLIC4*	G	C	-0.86	-0.85	-
rs3750900	*PLEKHS1*	C	T	-0.90	-0.94	-
rs2236225	*MTHFD1*	G	A	0.92	0.87	-
rs7869523	*PTBP3*	C	T	-1.01	-1.04	-
rs6488868	*SBNO1*	A	G	1.19	1.09	-
rs43216	*NLRC5*	A	G	-0.96	-1.51	-
rs1127986	*TBCD*	T	C	-1.26	-0.92	-

Table 4.2: 63 aeSNPs that fulfilled the Hedges' G criteria regarding ER status. This table comprehends associated gene name, both positive and negative sample sizes in ER, Hedges' G difference value, both upper and lower bias-corrected accelerated bootstrap limits and the reference and alternative allele for each aeSNP.

SNPID	Gene	ER Positive Samples	ER Negative Samples	Hedges' G Difference	Lower BCA bootsrap limit	Upper BCA bootstrap limit	REF	ALT
rs6488868	*SBNO1*	26	12	1.09	0.37	1.85	A	G
rs241402	-	38	10	1.04	0.35	1.71	T	C
rs3791979	*TNS1*	28	12	0.97	0.36	1.55	T	C
rs14038	*EFNA5*	36	11	0.94	0.44	1.43	G	A
rs3742207	*COL4A1*	34	11	0.91	0.18	1.65	T	G
rs7117111	*CUL5*	22	11	0.89	0.14	1.50	A	G
rs760482	*SUN2-DNAL4*	30	10	0.89	0.33	1.37	A	G
rs934005	*UACA*	28	10	0.89	0.23	1.51	G	A
rs4849167	*PSD4*	37	10	0.88	0.30	1.49	G	C
rs4865614	*SLC38A9*	26	10	0.88	0.21	1.48	A	G
rs2236225	*MTHFD1*	33	12	0.87	0.22	1.49	G	A
rs7116	*PRMT2*	25	12	0.85	0.18	1.46	C	G
rs473279	*MTF1*	31	11	0.84	0.15	1.57	T	C
rs1048077	*ENPP4*	23	10	0.82	0.16	1.47	A	G
rs3815348	*FAM126A*	24	14	0.82	0.14	1.47	T	C
rs4614	*VPS11*	31	10	0.82	0.14	1.48	A	G
rs11254468	*VIM*	28	11	0.81	0.06	1.40	C	T
rs2570	*ITPR2*	37	10	0.81	0.21	1.32	G	A
rs6694	-	26	11	0.80	0.10	1.42	T	A
rs1062108	*RRAGD*	31	10	0.76	0.07	1.43	A	G
rs4343	*ACE*	35	11	0.75	0.25	1.20	G	A
rs10756457	*MPDZ*	25	12	0.74	0.01	1.49	T	C
rs12034	*CXADR*	27	13	0.74	0.06	1.37	G	A
rs7187	*KANK2*	37	10	0.74	0.20	1.24	A	G
rs3828002	*SLC24A3*	29	13	0.71	0.11	1.33	G	A
rs726176	*SORBS1*	28	10	0.70	0.06	1.22	T	C
rs241601	*RNF24*	37	12	0.68	0.03	1.34	A	T
rs4687587	*CHDH*	27	10	0.68	0.02	1.37	A	G
rs3180467	*ADGRG1*	30	11	0.67	0.05	1.25	A	G
rs1053433	*KCTD12*	30	12	0.57	0.08	0.98	T	G

rs818708	*ALAD*	38	11	0.56	0.09	1.04	G	A
rs1046320	*WFS1*	35	11	-0.56	-1.06	-0.01	G	A
rs1770791	*OSBPL9-NRDC*	28	11	-0.62	-1.22	-0.01	A	G
rs3798577	*ESR1*	34	11	-0.65	-1.20	-0.04	T	C
rs9725887	*FAM20B*	34	11	-0.65	-1.14	-0.11	T	C
rs2184157	*MTG2*	25	10	-0.67	-1.29	-0.07	G	C
rs11887008	*MAP3K20*	26	12	-0.68	-1.24	-0.06	T	C
rs6718068	*FOXN2*	31	10	-0.68	-1.32	-0.07	C	T
rs729363	*FAM114A1*	30	10	-0.69	-1.30	-0.05	G	C
rs8137	*FMC1-LUC7L2*	27	11	-0.69	-1.18	-0.17	G	T
rs11544338	*FAM117B*	31	10	-0.70	-1.18	-0.19	T	C
rs41534051	*NSUN4*	23	10	-0.70	-1.21	-0.07	T	C
rs1135791	*SP110*	46	12	-0.71	-1.27	-0.12	A	G
rs1031327	*QDPR*	21	13	-0.72	-1.40	-0.04	G	A
rs999692	*OSGEP*	15	12	-0.80	-1.59	-0.04	T	C
rs1354091	*EIF4A2*	25	10	-0.81	-1.39	-0.13	T	G
rs12379	*LETMD1*	44	13	-0.82	-1.40	-0.22	G	A
rs2366894	*BIRC6*	36	10	-0.82	-1.55	-0.06	A	T
rs1045529	*ERI1*	31	10	-0.85	-1.36	-0.35	C	T
rs435382	*TMEM220*	36	11	-0.85	-1.65	-0.04	C	T
rs9259	*CLIC4*	33	12	-0.85	-1.52	-0.08	G	C
rs25654	*ANPEP*	33	11	-0.87	-1.58	-0.07	A	G
rs1127986	*TBCD*	24	11	-0.92	-1.82	-0.08	T	C
rs8327	*ARHGAP2*	24	12	-0.92	-1.67	-0.17	A	G
rs10923360	*TENT5C*	24	11	-0.93	-1.70	-0.22	C	T
rs3750900	*PLEKHS1*	29	10	-0.94	-1.68	-0.10	C	T
rs7869523	*PTBP3*	37	12	-1.04	-1.72	-0.26	C	T
rs1044708	*LANCL1*	24	10	-1.08	-1.86	-0.20	T	G
rs887595	*LIN52*	25	10	-1.10	-1.84	-0.24	A	G
rs2732018	*PI15*	30	11	-1.11	-1.71	-0.35	T	C
rs2447097	*SGSM2*	14	10	-1.17	-2.19	-0.13	T	G
rs2251219	*PBRM1-SMIM4*	28	11	-1.20	-1.88	-0.49	T	C
rs43216	*NLRC5*	17	13	-1.51	-2.26	-0.74	A	G

Table 4.3: 119 aeSNPs that fulfilled the Hedges' G criteria regarding PR status. This table comprehends associated gene name, both positive and negative sample sizes in PR, Hedges' G difference value, both upper and lower bias-corrected accelerated bootstrap limits and the reference and alternative allele for each aeSNP.

SNPID	Gene	PR Positive Samples	PR Negative Samples	Hedges' G Difference	Lower BCA bootsrap limit	Upper BCA bootstrap limit	REF	ALT
rs7301360	*APOLD1*	16	11	1.30	0.47	2.15	C	A
rs6488868	*SBNO1*	23	15	1.19	0.51	1.84	A	G
rs10519181	*TBC1D2B*	19	10	1.11	0.32	1.68	G	C
rs7328946	*KATNAL1*	12	10	1.10	0.16	2.11	G	T
rs4312169	*MTMR6*	23	10	1.05	0.36	1.64	T	C
rs3764859	*ENTPD5-COQ6*	27	12	1.01	0.30	1.61	C	A
rs6151588	*BNIP2*	22	10	0.96	0.15	1.67	G	A
rs1131356	*FLNB*	23	12	0.95	0.23	1.66	G	A
rs1054260	*THSD4*	21	16	0.94	0.26	1.57	C	T
rs6880	*TULP4*	28	12	0.94	0.25	1.56	G	C
rs11649450	*ZNF213*	13	14	0.92	0.08	1.77	G	A
rs2236225	*MTHFD1*	31	14	0.92	0.15	1.64	G	A
rs3742207	*COL4A1*	30	15	0.91	0.27	1.58	T	G
rs1292037	*VMP1*	20	11	0.89	0.11	1.58	T	C
rs7223	*RHBDD2*	22	10	0.89	0.12	1.68	G	A
rs1064261	*MTOR*	23	11	0.88	0.06	1.55	G	A
rs11780242	*LZTS1*	20	10	0.88	0.24	1.46	G	C
rs1657397	*WDR7*	21	13	0.87	0.10	1.62	T	C
rs1062108	*RRAGD*	26	15	0.83	0.25	1.35	A	G
rs1056836	*CYP1B1*	23	12	0.81	0.12	1.32	G	C
rs4020	*ZNF74*	27	11	0.81	0.10	1.44	C	A
rs4858798	*IP6K2*	33	12	0.81	0.15	1.44	G	A
rs234737	*PKNOX1*	25	12	0.80	0.07	1.46	C	T
rs3743773	-	26	10	0.80	0.03	1.47	G	A
rs4687587	*CHDH*	24	13	0.79	0.16	1.39	A	G
rs818708	*ALAD*	33	16	0.78	0.26	1.26	G	A
rs3826942	*PWWP3A*	22	12	0.77	0.07	1.20	G	A
rs9764	*NPY1R*	29	11	0.77	0.20	1.34	T	C
rs2072695	*KIF3C*	15	10	0.76	0.01	1.46	A	T

rs2570	*ITPR2*	32	15	0.76	0.22	1.30	G	A
rs3815348	*FAM126A*	21	17	0.74	0.04	1.43	T	C
rs1056513	*PATJ*	28	13	0.73	0.00	1.37	G	A
rs6801	*ST3GAL3*	23	12	0.73	0.11	1.39	C	G
rs1044431	*BTBD10*	19	12	0.72	0.04	1.35	C	T
rs2241960	*SLC38A1*	25	13	0.72	0.07	1.32	A	G
rs241601	*RNF24*	34	15	0.72	0.14	1.24	A	T
rs3791979	*TNS1*	27	13	0.72	0.10	1.27	T	C
rs760482	*SUN2-DNAL4*	26	14	0.72	0.16	1.21	A	G
rs8111801	*CHERP*	27	15	0.71	0.02	1.33	G	C
rs14073	*PRXL2A*	32	13	0.70	0.01	1.37	T	A
rs10448	*CENPBD1P1*	36	17	0.69	0.12	1.23	T	C
rs7187	*KANK2*	34	13	0.69	0.14	1.21	A	G
rs934005	*UACA*	26	12	0.69	0.01	1.31	G	A
rs10584	*ANPEP*	35	11	0.68	0.05	1.30	G	C
rs2297775	*GON4L*	20	13	0.68	0.06	1.16	T	C
rs12932018	*USP10*	23	10	0.67	0.00	1.28	A	G
rs10638	*WSB2*	30	15	0.65	0.04	1.25	A	G
rs2633852	*SUMF1*	28	15	0.64	0.02	1.20	A	G
rs7507	*MB21D2*	23	21	0.63	0.06	1.19	G	A
rs4237	*PITHD1*	30	16	0.62	0.05	1.12	G	A
rs2274064	*NCF2-SMG7*	25	10	0.61	0.02	1.19	T	C
rs1047799	*MPRIP*	31	15	0.56	0.00	1.07	G	C
rs1053433	*KCTD12*	30	12	0.55	0.04	0.95	T	G
rs4802261	*OPA3*	31	14	-0.58	-1.05	-0.07	A	G
rs11544338	*FAM117B*	28	13	-0.59	-1.10	0.00	T	C
rs3798577	*ESR1*	28	17	-0.59	-1.07	-0.02	T	C
rs41534051	*NSUN4*	22	11	-0.59	-1.16	-0.04	T	C
rs11729794	*MAML3*	24	10	-0.62	-1.17	-0.04	C	T
rs1545133	*POLR1B*	29	14	-0.66	-1.27	0.00	C	T
rs11887008	*MAP3K20*	24	14	-0.67	-1.19	-0.06	T	C
rs702681	*SETD9-MIER3*	29	13	-0.67	-1.30	-0.04	C	T
rs1800392	*WRN*	28	16	-0.68	-1.26	-0.04	G	T
rs2232504	*MON1B*	31	10	-0.68	-1.26	-0.01	T	C
rs1045529	*ERI1*	28	13	-0.69	-1.28	-0.03	C	T
rs1010156	*LOXL2*	32	11	-0.71	-1.27	-0.12	T	C
rs161941	*CHD1*	28	10	-0.71	-1.24	-0.14	C	T
rs2291853	*SH3BP5*	27	15	-0.71	-1.28	-0.08	G	A

rs8137	*FMC1-LUC7L2*	26	12	-0.71	-1.20	-0.14	G	T
rs4748047	*FRMD4A*	34	14	-0.72	-1.42	-0.05	A	C
rs1044708	*LANCL1*	21	13	-0.73	-1.44	0.00	T	G
rs10492987	*DDI2*	24	13	-0.73	-1.35	-0.12	A	G
rs1051038	*AGT-COG2*	30	11	-0.73	-1.32	-0.12	A	G
rs3193970	*SORBS1*	30	10	-0.73	-1.31	-0.11	T	C
rs701848	*PTEN*	27	11	-0.74	-1.25	-0.13	T	C
rs2372309	*GIHCG-ATP23*	37	16	-0.75	-1.31	-0.13	G	A
rs353291	-	29	12	-0.75	-1.35	-0.11	T	C
rs6468171	*TTI2-MAK16*	37	15	-0.75	-1.20	-0.27	A	G
rs1203974	*LUC7L*	22	11	-0.76	-1.45	0.00	G	A
rs2251219	*PBRM1-SMIM4*	25	14	-0.76	-1.48	-0.01	T	C
rs10923360	*TENT5C*	22	13	-0.77	-1.42	-0.06	C	T
rs3814127	*MVB12B*	16	13	-0.78	-1.45	-0.10	A	G
rs1309581	*CWC27*	38	13	-0.79	-1.25	-0.27	G	A
rs2286845	*BET1*	37	11	-0.79	-1.51	-0.07	A	G
rs15680	*LINC00680*	19	12	-0.80	-1.51	-0.02	C	G
rs2229869	*SOS2*	28	10	-0.81	-1.50	-0.04	G	A
rs4294650	*XPO4*	19	13	-0.81	-1.52	-0.05	C	T
rs909998	*BTBD9*	20	11	-0.81	-1.40	-0.14	C	T
rs12484060	*PPIL2*	29	12	-0.82	-1.48	-0.17	C	T
rs1483780	*ALDH7A1*	23	14	-0.82	-1.56	-0.04	C	T
rs12568	*DCAKD-NMT1*	12	11	-0.83	-1.47	-0.01	A	C
rs1860	*YPEL1*	23	11	-0.84	-1.49	-0.14	A	G
rs220079	*DLGAP4*	21	12	-0.84	-1.49	-0.10	G	A
rs2280076	*CCDC3*	23	12	-0.84	-1.34	-0.24	C	T
rs12451211	*SMG6-HIC1*	16	13	-0.85	-1.47	-0.12	G	A
rs9259	*CLIC4*	32	13	-0.86	-1.43	-0.15	G	C
rs11100790	*SMARCA5*	21	11	-0.87	-1.53	-0.15	T	C
rs11558511	*DCAF6*	21	11	-0.88	-1.65	-0.13	T	C
rs11886	*PELO-ITGA1*	22	11	-0.88	-1.67	-0.10	T	G
rs185200	*ARHGAP26*	14	12	-0.88	-1.65	-0.02	A	G
rs1135791	*SP110*	40	18	-0.89	-1.46	-0.28	A	G

rs2709416	*METTL21A*	21	10	-0.89	-1.35	-0.29	C	T
rs3750900	*PLEKHS1*	28	11	-0.90	-1.58	-0.14	C	T
rs11078677	*NLGN2*	28	11	-0.93	-1.66	-0.16	C	T
rs14243	*INPP5D*	19	11	-0.93	-1.69	-0.05	A	G
rs2288274	*ZNF135*	20	14	-0.93	-1.58	-0.20	G	A
rs1868844	*ATP6V1E2*	27	13	-0.95	-1.72	-0.20	G	A
rs43216	*NLRC5*	15	15	-0.96	-1.78	-0.14	A	G
rs887595	*LIN52*	23	12	-0.96	-1.66	-0.08	A	G
rs1045634	*SBF2*	24	14	-0.97	-1.63	-0.21	C	T
rs3814119	*LMX1B*	22	11	-0.97	-1.51	-0.39	T	C
rs1046528	-	17	11	-1.00	-1.84	-0.19	T	G
rs7869523	*PTBP3*	34	15	-1.01	-1.61	-0.36	C	T
rs1053007	*SLC44A2*	26	11	-1.02	-1.62	-0.36	A	G
rs10865994	*WNT5A*	27	14	-1.02	-1.54	-0.46	T	A
rs2228261	*THBS1*	14	10	-1.07	-1.71	-0.40	C	T
rs7944548	*RNH1*	25	11	-1.07	-1.75	-0.28	C	T
rs884510	*PPP1R1A-PDE1B*	17	12	-1.16	-1.87	-0.47	C	T
rs10896	*BORCS8-MEF2B*	26	11	-1.22	-2.04	-0.44	A	G
rs1127986	*TBCD*	24	11	-1.26	-2.16	-0.38	T	C

Table 4.4: 58 aeSNPs that fulfilled the Hedges' G criteria regarding HER2 status. This table comprehends associated gene name, both positive and negative sample sizes in HER2, Hedges' G difference value, both upper and lower bias-corrected accelerated bootstrap limits and the reference and alternative allele for each aeSNP.

SNPID	Gene	HER2 Positive Samples	HER2 Negative Samples	Hedges' G Difference	Lower BCA bootsrap limit	Upper BCA bootstrap limit	REF	ALT
rs990626	*LRP2*	11	11	1.32	0.31	2.08	G	A
rs3750163	*TNS3*	12	31	1.22	0.55	1.72	G	A
rs12452567	*METTL16*	11	17	1.21	0.47	1.93	A	G
rs7419	*GNG12*	10	22	1.09	0.06	1.92	C	T
rs4857302	*CRYBG3*	11	31	1.07	0.32	1.74	A	C
rs9460	*ADPGK*	10	32	1.01	0.44	1.53	G	A
rs9568	*CHP1*	10	22	1.01	0.14	1.83	C	A
rs2910344	*ZNF558*	11	24	1.00	0.26	1.74	T	C
rs3796700	*RNF150*	11	25	0.99	0.37	1.66	C	T

rs4151134	*FGF11*	12	15	0.97	0.06	1.81	C	T
rs8539	*HSPD1*	11	31	0.96	0.25	1.63	T	C
rs3737669	*GPR157*	10	22	0.94	0.09	1.74	G	A
rs2130882	*ARRDC4*	10	31	0.93	0.38	1.42	T	C
rs936227	*ULK3*	10	29	0.90	0.24	1.40	A	G
rs4312169	*MTMR6*	10	19	0.89	0.06	1.74	T	C
rs1045327	*TLK1*	10	21	0.87	0.17	1.55	T	C
rs8468	*RPS27L-LACTB*	11	25	0.87	0.05	1.63	C	T
rs2179129	*ZNRF3*	11	15	0.83	0.04	1.73	A	G
rs1044011	*DLC1*	11	21	0.81	0.17	1.42	C	A
rs971	*SMUG1*	10	30	0.81	0.14	1.41	T	C
rs6865	*HIGD1A*	11	24	0.77	0.09	1.37	T	G
rs1050847	*ZCCHC14*	10	30	0.76	0.07	1.37	C	T
rs6586	*TRMO*	11	30	0.76	0.17	1.36	C	T
rs9574551	*DCLK1*	15	21	0.72	0.06	1.36	T	C
rs1800380	*VWF*	11	23	0.68	0.05	1.29	C	T
rs1127313	*ADAR*	12	37	0.67	0.07	1.28	G	A
rs12049	*TMEM186-PMM2*	10	33	0.67	0.01	1.24	T	C
rs1050162	*MYH11-NDE1*	10	24	0.65	0.03	1.18	C	T
rs1052637	*DDX18*	12	27	0.65	0.06	1.22	C	G
rs1468542	*SACM1L*	11	29	0.63	0.05	1.18	A	T
rs6774933	*PARP14*	12	31	0.63	0.03	1.20	T	C
rs1048013	*CYP20A1*	11	33	0.61	0.10	1.02	C	T
rs2111212	*CORO1C*	12	29	0.60	0.03	1.11	A	C
rs2633852	*SUMF1*	12	27	-0.66	-1.18	-0.01	A	G
rs3130	*FGFBP2-PROM1*	11	31	-0.67	-1.25	-0.07	T	C
rs1051420	*ETS2*	13	27	-0.68	-1.33	-0.02	A	C
rs2332285	*DTX3L*	13	26	-0.68	-1.28	-0.05	G	A
rs1801131	*MTHFR*	10	24	-0.70	-1.30	-0.09	T	G
rs1065177	*RIF1*	11	19	-0.71	-1.27	-0.09	C	G
rs2069372	*RAB4A*	11	27	-0.72	-1.28	-0.12	C	T
rs2241335	*SLC35B4*	11	22	-0.72	-1.38	-0.04	C	T
rs1051009	*CXCL16*	11	25	-0.73	-1.30	-0.12	G	A
rs2274736	*PTPN21*	12	20	-0.74	-1.33	-0.04	A	G

rs3853640	*ARPIN-AP3S2*	10	30	-0.76	-1.21	-0.20	T	G
rs1873793	*SLC38A2*	11	21	-0.78	-1.33	-0.12	C	T
rs4682801	*FYCO1*	12	23	-0.81	-1.46	-0.09	G	T
rs2281791	*TBC1D12-HELLS*	10	27	-0.82	-1.57	-0.07	T	C
rs4758400	*FAM160A2*	12	17	-0.82	-1.56	-0.09	G	C
rs12197	*AP3S2*	12	25	-0.88	-1.56	-0.11	G	A
rs668556	*MACF1*	10	26	-0.89	-1.54	-0.03	G	C
rs1060442	*MED16*	10	32	-0.90	-1.49	-0.22	A	G
rs15710	*ESRP1*	10	21	-0.91	-1.48	-0.15	T	C
rs2257505	*SETD9*	11	16	-0.92	-1.73	-0.06	T	A
rs2286845	*BET1*	10	34	-0.92	-1.45	-0.40	A	G
rs3205166	*DDX58*	11	19	-0.92	-1.54	-0.24	T	G
rs328744	*RAB21*	10	29	-0.93	-1.64	-0.23	G	A
rs2241183	*BTBD11*	11	18	-0.95	-1.64	-0.15	A	G
rs6693451	*FBXO42*	12	20	-0.96	-1.61	-0.28	A	C

In order to assess if any of the genes showing large allelic effect sizes according to hormone receptor status, were previously associated with a particular sub-type of breast cancer, we perused the GWAS reported genes catalogue (Figure 4.12).

GWAS genes were defined as genes harbouring either a GWAS hit variant or variants in moderate/strong LD with them ($r^2 \geq 0.4$) (see Methods).

We did not detect any overlap between GWAS genes and the associated genes identified in our analysis. Additionally, we also looked at genes previously reported in the literature as interacting with the aforesaid receptors, but once more, no match was found.

4.4 Differential allelic expression in patient survival

To further investigate the impact of allelic expression imbalances in patient's prognosis, we tested if the aeSNPs whose AE ratios were found to differ according with the patient's tumours ER, PR or HER2 receptor statuses, were also associated with patient's disease-free survival.

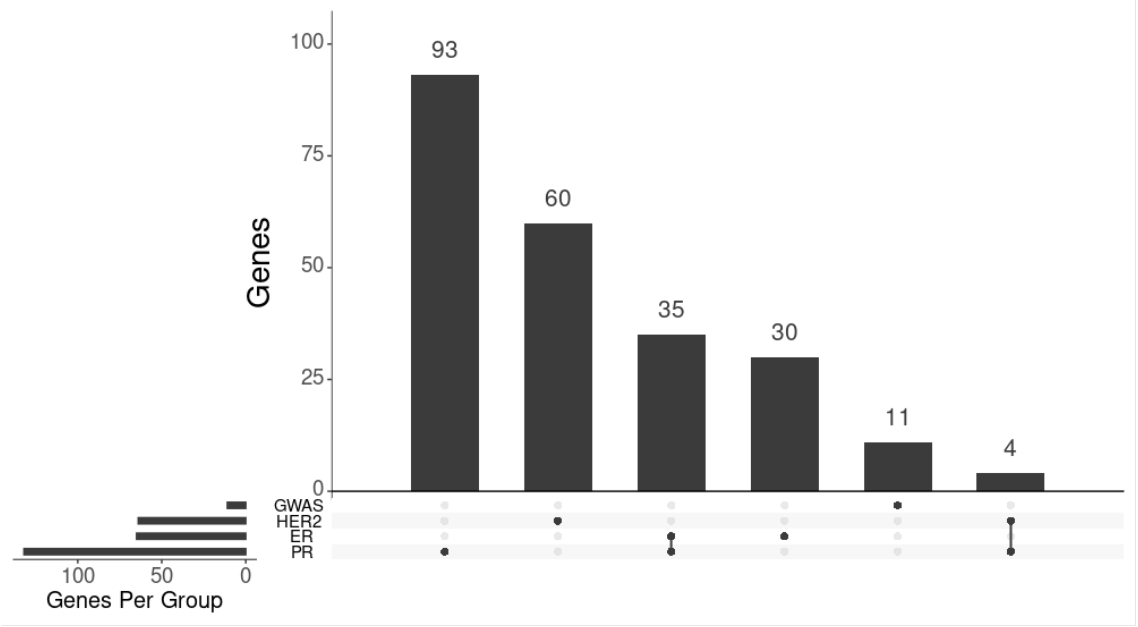

Figure 4.12: Upset plot comparing genes whose AE ratios were identified associated with ER, PR or Her2 receptor with previously reported GWAS Genes for BC. The y axis represents the number of genes in decreasing order distributed by the x axis different groups that encompass each individual receptor associated genes and GWAS genes together with any overlaps between them. The bottom half of this figure shows how many genes were present per group in a sideways bar plot to the left in conjunction with a dot-dash display of all the groups' intersecting genes.

We carried this analysis using a Two Stage Hazard Rate Comparison (TSHRC) and dividing the samples for each aeSNP in two groups based on the allelic expression ratio exhibited, resulting in one group favouring expression of the alternative allele (values more positive) and another group favouring the expression of the reference allele (values more negative) (Figure 4.13).

After running the stage one test (log-rank test) we identified 6 aeSNPs whose AE ratios were significantly different between populations (pvalue ≤ 0.05) and therefore relevant to patient survival: rs2179129 (*ZNRF3*), rs11780242 (*LZTS1*), rs1053007 (*SLC44A2*), rs760482 (*SUN2-DNAL4*), rs7187 (*KANK2*) and rs3795817 (1q42.13).

For the remaining aeSNPS whose populations were identical, or survival curves intercepted resulting in the inability of stage one test to detect any differences, the stage two test was applied to discern between the two possibilities, but no further aeSNPs of interest surfaced.

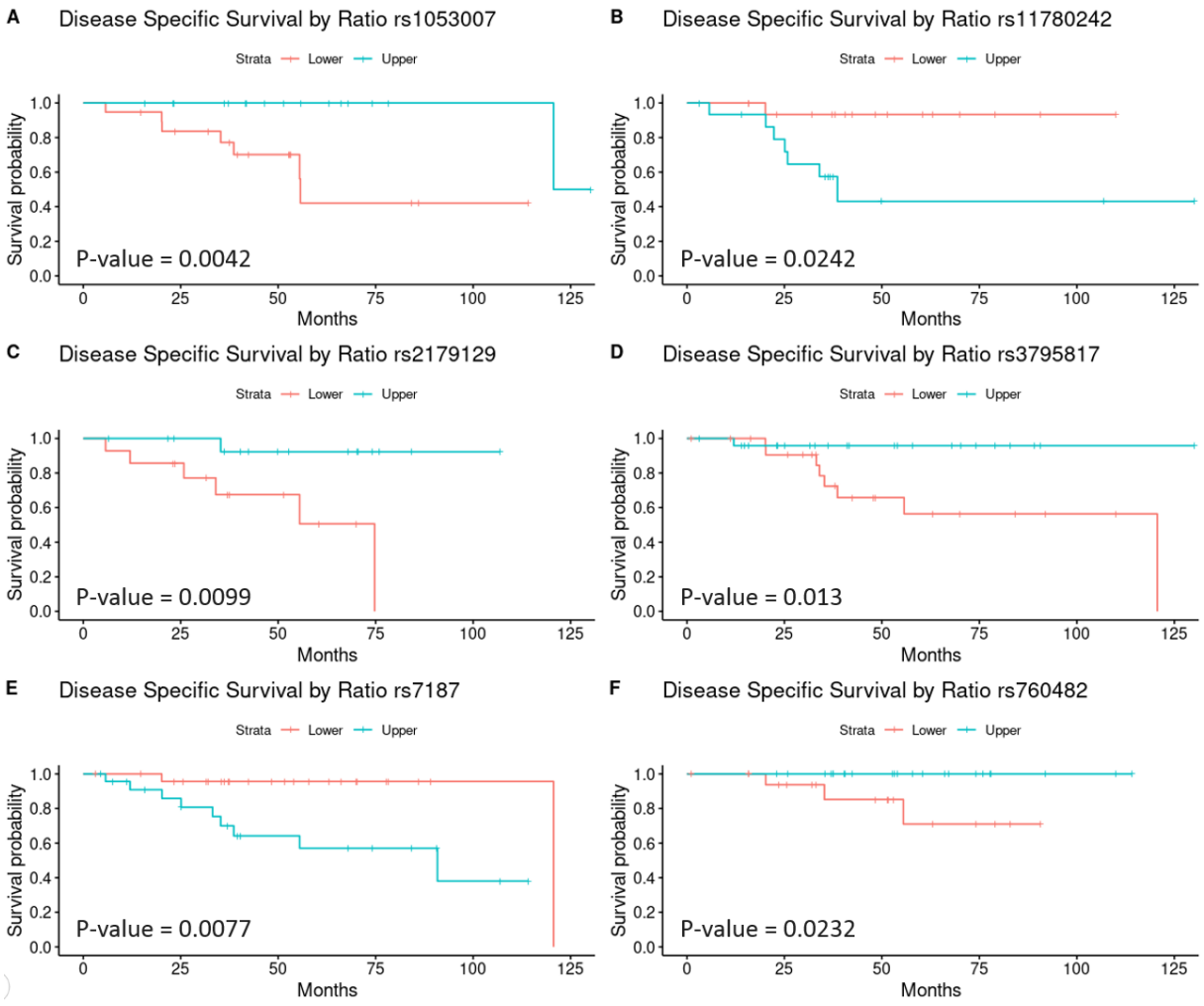

Figure 4.13: Kaplan-Meier curves of the 6 significant aeSNPs according to TSHRC test. This figure compiles all the Kaplan-Meier curves that show the correlation between disease-free survival in months and survival probability. The population is subdivided in two groups of individuals depending on AE ratio values (blue for samples that show higher AE ratio values indicating higher alternative allele expression and red for lower AE ratio values showing higher reference allele expression) for 6 aeSNPs that show a p-value under 0.05: rs1053007 **(A)** located in *SLC44A2* gene, rs11780242 **(B)** located in *LZTS1*, rs2179129 **(C)** located in *ZNRF*, rs3795817 **(D)** located in 1q42.13, rs7187 **(E)** located in *KANK2* and rs760482 **(F)** located in *SUN2-DNAL4*.

Discussion

5 Discussion

In this study we proposed to apply a redout of the effect of cis-regulatory variants – differential allelic expression – in order to investigate if these variants, at a germline level, could have an impact on breast cancer tumours characteristics and patients' survival. We found one variant (rs3764859) with AE ratios significantly associated with estrogen receptor status (q-value < 0.1).

Additionally, we identified 207 variants whose AE ratios display small to large differences (effect size ≥ 0.2) between samples from patients with tumours displaying different receptor statuses for ER, PR or HER2 receptors. Of these, 121 aeSNPs showed a large difference between them (≥ 0.8 effect size) and a further 10 showed a large difference according to two different receptors.

We also identified 6 aeSNPs whose AE ratios were associated with different time of survival. These observations of differences in AE ratios in normal-matched tissue, according to tumour biology, suggest the existence of cis-regulatory variants at germline tissue contributing to the tumour features.

Some of the variants whose AE ratios showed large differences between receptor status in normal tissue from patients with tumours, were in genes whose expression levels have been reported in the literature as altered in breast cancer.

One of these variants was rs7869523, found associated with both ER and PR statuses and located in the *PTBP3* gene. Previous studies indicated an increased expression of this gene in breast carcinoma when compared to normal breast tissue via the promotion of invasive growth and metastasis in breast cancer tumour cells by preventing the degradation of *ZEB1*'s mRNA, a regulatory transcription factor whose overexpression leads to epithelial mesenchymal transition (Hou et al., 2018).

rs9259 also showed a large AE ratio difference in ER and PR, located in *CLIC4*. This gene codifies for a protein that plays a role in angiogenesis and in creating a TGF-β rich cancer stroma which in turn mediates the conversion of primary human breast fibroblasts to activated myofibroblasts. These activated cancer stroma myofibroblasts possess the ability to remodel the extracellular matrix through synthesis and degradation create a tumoral stroma that enhances tumour invasiveness and progression (Shukla et al., 2013).

rs3742207 located in *COL4A1*, a gene that encodes collagen IV which composes most cellular basement membranes, was another variant showing large effect sizes. The knockout of this gene in invasive ductal carcinoma cells was shown to reduce tumoral proliferation and migration as well as differentiation (Jin et al., 2017).

rs43216 locates to *NLRC5,* a gene that has been implicated in transcriptional coactivation of MHC class I crucial in $CD8^+$ T cells identification and response to cancer cells, with some studies crediting the highest reduction of *NLRC5* to breast cancer (Yoshihama et al., 2016). Subsequent promotion of *NLRC5* expression in *SKB3* breast cancer cells that are MHC class I deficient (Zhao et al., 2019), may prove to be a mean to counteract tumorigenic immune system evasion.

rs2236225, also known as *MTHFD1* G1958A, has already been implicated in breast cancer studies and its alternative A allele, is a prognostic factor of poor clinical outcome in premenopausal breast cancer patients (Babyshkina et al., 2013) since their carriers had a significantly lowered progression-free survival. Interestingly, we also found this allele to be more expressed in patients with both negative ER and PR tumours, that are also considered to be more aggressive than ER and PR positive tumours.

rs2286845 located in the gene *BET1* which encodes for a protein that promotes transportation of MT1-MMP to the cell surface through endosomes, originating extracellular matrix degrading invadopodia in invasive breast cancer cells (Miyagawa et al., 2019).

rs3764859, that we found significantly associated with ER status, also showed a large effect size although only according to PR status. This variant is in the gene *ENTPD5*, which encodes for a UDPase targeted by mutant p53 via docking to Sp1 in *ENTPD5* promoter to stimulate calnexin/calreticulin-mediated N-glycosylated protein folding. This consequently increases the expression of prometastatic cell surface proteins, linking *ENTPD5* overexpression with the increase of p53 gain-of-function mutations (Vogiatzi et al., 2016).

Among the six aeSNPs whose AE ratios were associated with survival, two of them showed a large effect size regarding groups of tumours with different receptor statuses. rs760482 in particular, was localized in the *SUN2* gene, which encodes for a protein involved in the linker of nucleoskeleton and cytoskeleton (LINC) complex and has a suppressor role in breast cancer by inhibiting tumoral migration and progression while also promoting apoptosis (X. Chen et al., 2019).

We believe the TCGA-BRCA normal-matched was the optimal dataset to assess whether cis-regulatory variants could determine clinical and molecular characteristics of the tumour due to being able to observe germline variants in normal tissue and having the matched tumoral outcome allowing us to look at the of origin of the disease and therefore to investigate the read-out of the effect of germline cis-regulatory variants.

Even though we find this dataset to be suitable for our intended purpose, this does not imply our method is in no particular manner the only correct assessment of the hypothesis in question and reviewing this process considering other divulged breast cancer datasets is highly encouraged.

We also acknowledge that the nature of the study required heterozygous samples to gage the differential allelic expression meaning that our sample size was being heavily reduced to meet the criteria, and consequently, the statistical power of our conclusions was inevitably diminished.

The filters applied to meet the statistical tests conditions resulted in the removal of variants that did not have a significant number of positive and negative samples in the tested receptors, further reducing the study's genomic scope in breast cancer.

We also planned to test other clinical parameters, the majority of which non-binary such as tumour grade or nearby lymph node stage. But with the inclusion of 3 or more variables, the sample size for each one was severely shortened to the point where either the test was not carried out or no significant conclusion could be drawn.

We address other limitations as well, such as the analysis of a single dataset, and the fact that most of the samples are Caucasian and therefore our findings may not be suitable to all ethnic and genetic backgrounds, that have different incidence and clinical manifestations regarding breast cancer.

While we identified one variant whose AE ratios display statistical differences between ER+ and ER+ groups, and many others whose AE ratios display large effect size according to ER, PR or HER2 status, in the analysed dataset, these findings warrant further replication in an independent dataset.

We also encourage any future studies on the matter to consider including additional *in silico* functional analysis to identify the variants regulating allelic expression levels at each locus, the mechanisms by which they affect allelic expression, and to clarify their role in determining tumour biology.

It would also be important to analyse the association of cis-regulatory variants with tumour biology in other cancers, contributing to improve the knowledge on cancer aetiology.

Conclusion

6 Conclusion

This study constitutes as a catalogue of differential allelic expression according to ER, PR and HER2 status in the TCGA-BRCA project dataset, highlighting the variants with more relevant differences in one or more receptors, as well as coupling said SNPs with their respective gene(s) whose impact in breast cancer can already be found in the literature.

This work provides further research on the implications of cis-regulatory variants in allelic expression, by computing and compiling existing data to create a record of the interaction between DAE and clinical/molecular characteristics of the tumour. We also report on statistically significant variants located in genes implicated in a myriad of processes such as tumoral growth, metastasis, angiogenesis and suppression, whose mechanism of influence in the aforementioned genes could be further researched.

References

7 References

Ahearn, T. U., Zhang, H., Michailidou, K., Milne, R. L., Bolla, M. K., Dennis, J., Dunning, A. M., Lush, M., Wang, Q., Andrulis, I. L., Anton-Culver, H., Arndt, V., Aronson, K. J., Auer, P. L., Augustinsson, A., Baten, A., Becher, H., Behrens, S., Benitez, J., … Chatterjee, N. (2022). Common variants in breast cancer risk loci predispose to distinct tumor subtypes. *Breast Cancer Research*, *24*(1), 1–13. https://doi.org/10.1186/S13058-021-01484-X/TABLES/1

Ajisaka, H., Tsugawa, K., NoguchW, M., & Nonomura, A. (2002). Histological Subtypes of Ductal Carcinoma in situ of lhe Breast. In *Breast Cancer* (Vol. 9, Issue 1).

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

Almlöf, J. C., Lundmark, P., Lundmark, A., Ge, B., Maouche, S., Göring, H. H. H., Liljedahl, U., Enström, C., Brocheton, J., Proust, C., Godefroy, T., Sambrook, J. G., Jolley, J., Crisp-Hihn, A., Foad, N., Lloyd-Jones, H., Stephens, J., Gwilliam, R., Rice, C. M., … Syvänen, A. C. (2012). Powerful Identification of Cis-regulatory SNPs in Human Primary Monocytes Using Allele-Specific Gene Expression. *PLoS ONE*, *7*(12), 52260. https://doi.org/10.1371/journal.pone.0052260

Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic Mapping in Human Disease. *Science (New York, N.Y.)*, *322*(5903), 881. https://doi.org/10.1126/SCIENCE.1156409

Ambrosone, C. B., Kropp, S., Yang, J., Yao, S., Shields, P. G., & Chang-Claude, J. (2008). Cigarette smoking, N-acetyltransferase 2 genotypes, and breast cancer risk: pooled analysis and meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *17*(1), 15–26. https://doi.org/10.1158/1055-9965.EPI-07-0598

Apostolou, P., & Fostira, F. (2013). Hereditary breast cancer: The Era of new susceptibility genes. In *BioMed Research International* (Vol. 2013, p. 11). Hindawi Limited. https://doi.org/10.1155/2013/747318

Babyshkina, N., Malinovskaya, E., Nazarenko, M., Koval, M., Gervas, P., Potapova, O., Slonimskaya, E., & Cherdyntseva, N. (2013). The effect of folate-related SNPs on clinicopathological features, response to neoadjuvant treatment and survival in pre- and postmenopausal breast cancer patients. *Gene*, *518*(2), 397–404. https://doi.org/10.1016/J.GENE.2012.12.095

Barnard, M. E., Boeke, C. E., & Tamimi, R. M. (2015). Established breast cancer risk factors and risk of intrinsic tumor subtypes. In *Biochimica et Biophysica Acta - Reviews on Cancer* (Vol. 1856, Issue 1, pp. 73–85). Elsevier B.V. https://doi.org/10.1016/j.bbcan.2015.06.002

Bernard, P. S., Parker, J. S., Mullins, M., Cheung, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenhush, J. F., Stijleman, I. J., Palazzo, J., Matron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., & Perou, C. M. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, *27*(8), 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370

Bernhardt, S. M., Dasari, P., Walsh, D., Townsend, A. R., Price, T. J., & Ingman, W. V. (2016). Hormonal Modulation of Breast Cancer Gene Expression: Implications for Intrinsic Subtyping in Premenopausal Women. *Frontiers in Oncology*, *6*(NOV).

https://doi.org/10.3389/FONC.2016.00241

Bose, S., Zhang, C., & Lee, A. (2021). Glucose Metabolism in Cancer: The Warburg Efect and Beyond. In W. E. Crusio, D. Haidong, H. H. Radeke, N. Rezaei, & Junjie Xiao (Eds.), *The Heterogeneity of Cancer Metabolism* (Second Edi, pp. 3–15). Springer Nature Switzerland AG.

Broeks, A., Schmidt, M. K., Sherman, M. E., Couch, F. J., Hopper, J. L., Dite, G. S., Apicella, C., Smith, L. D., Hammet, F., Southey, M. C., Van't Veer, L. J., De Groot, R., Smit, V. T. H. B. M., Fasching, P. A., Beckmann, M. W., Jud, S., Ekici, A. B., Hartmann, A., Hein, A., … 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), 3289. https://doi.org/10.1093/HMG/DDR228

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, *8*(12), e1002822. https://doi.org/10.1371/JOURNAL.PCBI.1002822

Carey, L. A., Perou, C. M., Livasy, C. A., Dressler, L. G., Cowan, D., Conway, K., Karaca, G., Troester, M. A., Chiu, K. T., Edmiston, S., Deming, S. L., Geradts, J., Cheang, M. C. U., Nielsen, T. O., Moorman, P. G., Earp, H. S., & Millikan, R. C. (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *Journal of the American Medical Association*, *295*(21), 2492–2502. https://doi.org/10.1001/jama.295.21.2492

Castellano, I., Marchiò, C., Tomatis, M., Ponti, A., Casella, D., Bianchi, S., Vezzosi, V., Arisio, R., Pietribiasi, F., Frigerio, A., Mano, M. P., Ricardi, U., Allia, E., Accortanzo, V., Durando, A., Bussolati, G., Tot, T., & Sapino, A. (2009). Micropapillary ductal carcinoma in situ of the breast: an inter-institutional study. *Modern Pathology 2010 23:2*, *23*(2), 260–269. https://doi.org/10.1038/modpathol.2009.169

Chen, Xin, Chen, Y., Huang, H. M., Li, H. D., Bu, F. T., Pan, X. Y., Yang, Y., Li, W. X., Li, X. F., Huang, C., Meng, X. M., & Li, J. (2019). SUN2: A potential therapeutic target in cancer. In *Oncology Letters* (Vol. 17, Issue 2, pp. 1401–1408). Spandidos Publications. https://doi.org/10.3892/ol.2018.9764

Chen, Xuyu, Wang, Q., Zhang, Y., Xie, Q., & Tan, X. (2019). Physical Activity and Risk of Breast Cancer: A Meta-Analysis of 38 Cohort Studies in 45 Study Reports. *Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *22*(1), 104–128. https://doi.org/10.1016/J.JVAL.2018.06.020

Chung, I., Osterwald, S., Deeg, K. I., & Rippe, K. (2012). PML body meets telomere: The beginning of an ALTernate ending? *Nucleus*, *3*(3), 263. https://doi.org/10.4161/NUCL.20326

Correia, L., Magno, R., Xavier, J. M., de Almeida, B. P., Duarte, I., Esteves, F., Ghezzo, M., Eldridge, M., Sun, C., Bosma, A., Mittempergher, L., Marreiros, A., Bernards, R., Caldas, C., Chin, S. F., & Maia, A. T. (2022). Allelic expression imbalance of PIK3CA mutations is frequent in breast cancer and prognostically significant. *NPJ Breast Cancer*, *8*(1). https://doi.org/10.1038/S41523-022-00435-9

CRAN. (2023). *The Comprehensive R Archive Network*. https://cran.r-project.org/

Cserni, G. (2020). Histological type and typing of breast carcinomas and the WHO classification changes over time. *Pathologica*, *112*(1), 25. https://doi.org/10.32074/1591-951X-1-20

Darabi, H., McCue, K., Beesley, J., Michailidou, K., Nord, S., Kar, S., Humphreys, K., Thompson, D., Ghoussaini, M., Bolla, M. K., Dennis, J., Wang, Q., Canisius, S., Scott,

C. G., Apicella, C., Hopper, J. L., Southey, M. C., Stone, J., Broeks, A., … 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), 22–34. https://doi.org/10.1016/J.AJHG.2015.05.002

De Cicco, P., Catani, M. V., Gasperi, V., Sibilano, M., Quaglietta, M., & Savini, I. (2019). Nutrition and Breast Cancer: A Literature Review on Prevention, Treatment and Recurrence. *Nutrients*, *11*(7). https://doi.org/10.3390/NU11071514

Demir, S., Sezgin, G., Sari, A. A., Kucukzeybek, B. B., Yigit, S., Etit, D., Yazici, A., & Kucukzeybek, Y. (2021). Clinicopathological analysis of invasive cribriform carcinoma of the breast, with review of the literature. *Annals of Diagnostic Pathology*, *54*. https://doi.org/10.1016/J.ANNDIAGPATH.2021.151794

Dolle, J. M., Daling, J. R., White, E., Brinton, L. A., Doody, D. R., Porter, P. L., & Malone, K. E. (2009). Risk factors for triple-negative breast cancer in women under the age of 45 years. *Cancer Epidemiology Biomarkers and Prevention*, *18*(4), 1157–1166. https://doi.org/10.1158/1055-9965.EPI-08-1005

Dunning, A. M., Healey, C. S., Baynes, C., Maia, A. T., Scollen, S., Vega, A., Rodríguez, R., Barbosa-Morais, N. L., Ponder, B. A., Low, Y. L., Bingham, S., Haiman, C. A., Le Marchand, L., Broeks, A., Schmidt, M. K., Hopper, J., Southey, M., Beckmann, M. W., Fasching, P. A., … Chenevix-Trench, G. (2009). Association of ESR1 gene tagging SNPs with breast cancer risk. *Human Molecular Genetics*, *18*(6), 1131–1139. https://doi.org/10.1093/hmg/ddn429

Dunning, A. M., Michailidou, K., Kuchenbaecker, K. B., Thompson, D., French, J. D., Beesley, J., Healey, C. S., Kar, S., Pooley, K. A., Lopez-Knowles, E., Dicks, E., Barrowdale, D., Sinnott-Armstrong, N. A., Sallari, R. C., Hillman, K. M., Kaufmann, S., Sivakumaran, H., Marjaneh, M. M., Lee, J. S., … Edwards, S. L. (2016). Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nature Genetics*, *48*(4), 374–386. https://doi.org/10.1038/NG.3521

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C. S., Bowman, R., Meyer, K. B., Haiman, C. A., Kolonel, L. K., Henderson, B. E., Le Marchand, L., … Webb, P. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, *447*(7148), 1087. https://doi.org/10.1038/NATURE05887

Ellsworth, D. L., Turner, C. E., Ellsworth, R. E., & Li, C. J. (2019). A Review of the Hereditary Component of Triple Negative Breast Cancer: High- and Moderate-Penetrance Breast Cancer Genes, Low-Penetrance Loci, and the Role of Nontraditional Genetic Elements. *Journal of Oncology*, *2019*(4382606), 10. https://doi.org/10.1155/2019/4382606

Fachal, L., Aschard, H., Beesley, J., Barnes, D. R., Allen, J., Kar, S., Pooley, K. A., Dennis, J., Michailidou, K., Turman, C., Soucy, P., Lemaçon, A., Lush, M., Tyrer, J. P., Ghoussaini, M., Marjaneh, M. M., Jiang, X., Agata, S., Aittomäki, K., … Kraft, P. (2020). Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nature Genetics*, *52*(1), 56. https://doi.org/10.1038/S41588-019-0537-1

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*, 12–18. https://doi.org/10.1016/j.canep.2018.03.003

Fougner, C., Bergholtz, H., Norum, J. H., & Sørlie, T. (2020). Re-definition of claudin-low as a breast cancer phenotype. *Nature Communications 2020 11:1*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-15574-5

Foulkes, W. D. (2008). Inherited Susceptibility to Common Cancers. *New England Journal of Medicine*, *359*(20), 2143–2153. https://doi.org/10.1056/nejmra0802968

Freudenheim, J. L. (2020). Alcohol's Effects on Breast Cancer in Women. *Alcohol Research : Current Reviews*, *40*(2), 1–12. https://doi.org/10.35946/ARCR.V40.2.11

Gao, C., Devarajan, K., Zhou, Y., Slater, C. M., Daly, M. B., & Chen, X. (2012). Identifying breast cancer risk loci by global differential allele-specific expression (DASE) analysis in mammary epithelial transcriptome. *BMC Genomics*, *13*(1), 570. https://doi.org/10.1186/1471-2164-13-570

Gaudet, M. M., Gapstur, S. M., Sun, J., Ryan Diver, W., Hannan, L. M., & Thun, M. J. (2013). Active smoking and breast cancer risk: Original cohort data and meta-analysis. In *Journal of the National Cancer Institute* (Vol. 105, Issue 8, pp. 515–525). Oxford Academic. https://doi.org/10.1093/jnci/djt023

Gaudet, M. M., Press, M. F., Haile, R. W., Lynch, C. F., Glaser, S. L., Schildkraut, J., Gammon, M. D., Thompson, W. D., & Bernstein, J. L. (2011). Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger. *Breast Cancer Research and Treatment*, *130*(2), 587–597. https://doi.org/10.1007/s10549-011-1616-x

Ghoussaini, M., Pharoah, P. D. P., & Easton, D. F. (2013). Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning? *The American Journal of Pathology*, *183*(4), 1038–1051. https://doi.org/10.1016/J.AJPATH.2013.07.003

Godoy-Ortiz, A., Sanchez-Muñoz, A., Parrado, M. R. C., Álvarez, M., Ribelles, N., Dominguez, A. R., & Alba, E. (2019). Deciphering her2 breast cancer disease: Biological and clinical implications. *Frontiers in Oncology*, *9*(OCT), 1124. https://doi.org/10.3389/fonc.2019.01124

Gomes, I. A., & Nunes, C. (2020). Analysis of the breast cancer mortality rate in Portugal over a decade: Spatiotemporal clustering analysis. *Acta Medica Portuguesa*, *33*(5), 305–310. https://doi.org/10.20344/AMP.11749

Guo, Q., Schmidt, M. K., Kraft, P., Canisius, S., Chen, C., Khan, S., Tyrer, J., Bolla, M. K., Wang, Q., Dennis, J., Michailidou, K., Lush, M., Kar, S., Beesley, J., Dunning, A. M., Shah, M., Czene, K., Darabi, H., Eriksson, M., … Pharoah, P. D. P. (2015). Identification of novel genetic markers of breast cancer survival. *Journal of the National Cancer Institute*, *107*(5). https://doi.org/10.1093/JNCI/DJV081

Guo, X., Lin, W., Bao, J., Cai, Q., Pan, X., Bai, M., Yuan, Y., Shi, J., Sun, Y., Han, M. R., Wang, J., Liu, Q., Wen, W., Li, B., Long, J., Chen, J., & Zheng, W. (2018). A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *American Journal of Human Genetics*, *102*(5), 890–903. https://doi.org/10.1016/j.ajhg.2018.03.016

Hamdi, Y., Soucy, P., Adoue, V., Michailidou, K., Canisius, S., Lemaçon, A., Droit, A., Andrulis, I. L., Anton-Culver, H., Arndt, V., Baynes, C., Blomqvist, C., Bogdanova, N. V., Bojesen, S. E., Bolla, M. K., Bonanni, B., Borresen-Dale, A. L., Brand, J. S., Brauch, H., … 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), 80140. https://doi.org/10.18632/ONCOTARGET.12818

Hamdi, Y., Soucy, P., Kuchenbaeker, K. B., Pastinen, T., Droit, A., Lemaçon, A., Adlard, J.,

Aittomäki, K., Andrulis, I. L., Arason, A., Arnold, N., Arun, B. K., Azzollini, J., Bane, A., Barjhoux, L., Barrowdale, D., Benitez, J., Berthet, P., Blok, M. J., … Simard, J. (2017). Association of breast cancer risk in BRCA1 and BRCA2 mutation carriers with genetic variants showing differential allelic expression: identification of a modifier of breast cancer risk at locus 11q22.3. *Breast Cancer Research and Treatment*, *161*(1), 117. https://doi.org/10.1007/S10549-016-4018-2

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discovery*, *12*(1), 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/J.CELL.2011.02.013

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107. https://doi.org/10.2307/1164588

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K. E., Jones, L. P., Assefnia, S., Chandrasekharan, S., Backlund, M. G., Yin, Y., Khramtsov, A. I., Bastein, R., Quackenbush, J., Glazer, R. I., Brown, P. H., Green, J. E., Kopelovich, L., … Perou, C. M. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology*, *8*(5), R76. https://doi.org/10.1186/gb-2007-8-5-r76

Hilton, H. N., Clarke, C. L., & Graham, J. D. (2018). Estrogen and progesterone signalling in the normal breast and its implications for cancer development. *Molecular and Cellular Endocrinology*, *466*, 2–14. https://doi.org/10.1016/j.mce.2017.08.011

Hou, P., Li, L., Chen, F., Chen, Y., Liu, H., Li, J., Bai, J., & Zheng, J. (2018). PTBP3-Mediated Regulation of ZEB1 mRNA Stability Promotes Epithelial–Mesenchymal Transition in Breast Cancer. *Cancer Research*, *78*(2), 387–398. https://doi.org/10.1158/0008-5472.CAN-17-0883

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., … Chanock, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, *39*(7), 870. https://doi.org/10.1038/NG2075

Ihaka, R. (2009). *The R Project: A Brief History and Thoughts About the Future*. Stat.Auckland.Ac.Nz. https://www.stat.auckland.ac.nz/~ihaka/downloads/Massey.pdf

Jenkins, S., Kachur, M. E., Rechache, K., Wells, J. M., & Lipkowitz, S. (2021). Rare Breast Cancer Subtypes. *Current Oncology Reports*, *23*(5), 54. https://doi.org/10.1007/S11912-021-01048-4

Jeong, S. J., Lim, H. S., Lee, J. S., Park, M. H., Yoon, J. H., Park, J. G., & Kang, H. K. (2012). Medullary carcinoma of the breast: MRI findings. *AJR. American Journal of Roentgenology*, *198*(5). https://doi.org/10.2214/AJR.11.6944

Jin, R., Shen, J., Zhang, T., Liu, Q., Liao, C., Ma, H., Li, S., & Yu, Z. (2017). The highly expressed COL4A1 genes contributes to the proliferation and migration of the invasive ductal carcinomas. *Oncotarget*, *8*(35), 58172. https://doi.org/10.18632/ONCOTARGET.17345

John, E. M., & Kelsey, J. L. (1993). Radiation and other environmental exposures and breast cancer. *Epidemiologic Reviews*, *15*(1), 157–162. https://doi.org/10.1093/oxfordjournals.epirev.a036099

Kadalayil, L., Khan, S., Nevanlinna, H., Fasching, P. A., Couch, F. J., Hopper, J. L., Liu, J.,

Maishman, T., Durcan, L., Gerty, S., Blomqvist, C., Rack, B., Janni, W., Collins, A., Eccles, D., & Tapper, W. (2017). Germline variation in ADAMTSL1 is associated with prognosis following breast cancer treatment in young women. *Nature Communications*, *8*(1). https://doi.org/10.1038/S41467-017-01775-Y

Kanadys, W., Barańska, A., Malm, M., Błaszczuk, A., Polz-Dacewicz, M., Janiszewska, M., & Jędrych, M. (2021). Use of oral contraceptives as a potential risk factor for breast cancer: A systematic review and meta-analysis of case-control studies up to 2010. *International Journal of Environmental Research and Public Health*, *18*(9). https://doi.org/10.3390/IJERPH18094638/S1

Khan, S., Fagerholm, R., Rafiq, S., Tapper, W., Aittomäki, K., Liu, J., Blomqvist, C., Eccles, D., & Nevanlinna, H. (2015). Polymorphism at 19q13.41 Predicts Breast Cancer Survival Specifically after Endocrine Therapy. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *21*(18), 4086–4096. https://doi.org/10.1158/1078-0432.CCR-15-0296

Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Veizer, J., McMichael, J. F., Fulton, L. L., Dooling, D. J., Ding, L., Mardis, E. R., Wilson, R. K., Ally, A., Balasundaram, M., Butterfield, Y. S. N., Carlsen, R., Carter, C., Chu, A., Chuah, E., Chun, H. J. E., … Palchik, J. D. (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *490*(7418), 61–70. https://doi.org/10.1038/nature11412

Koual, M., Tomkiewicz, C., Cano-Sancho, G., Antignac, J. P., Bats, A. S., & Coumoul, X. (2020). Environmental chemicals, breast cancer progression and drug resistance. *Environmental Health*, *19*(1). https://doi.org/10.1186/S12940-020-00670-2

Lalloo, F., & Evans, D. G. (2012). Familial Breast Cancer. In *Clinical Genetics* (Vol. 82, Issue 2, pp. 105–114). John Wiley & Sons, Ltd. https://doi.org/10.1111/j.1399-0004.2012.01859.x

León-Novelo, L., Gerken, A. R., Graze, R. M., McIntyre, L. M., & Marroni, F. (2018). Direct testing for allele-specific expression differences between conditions. *G3: Genes, Genomes, Genetics*, *8*(2), 447–460. https://doi.org/10.1534/g3.117.300139

Li, N., Rowley, S. M., Thompson, E. R., McInerny, S., Devereux, L., Amarasinghe, K. C., Zethoven, M., Lupat, R., Goode, D., Li, J., Trainer, A. H., Gorringe, K. L., James, P. A., & Campbell, I. G. (2018). Evaluating the breast cancer predisposition role of rare variants in genes associated with low-penetrance breast cancer risk SNPs. *Breast Cancer Research*, *20*(1), 3. https://doi.org/10.1186/s13058-017-0929-z

Li, Q., Seo, J. H., Stranger, B., McKenna, A., Pe'Er, I., Laframboise, T., Brown, M., Tyekucheva, S., & Freedman, M. L. (2013). A novel eQTL-based analysis reveals the biology of breast cancer risk loci. *Cell*, *152*(3), 633. https://doi.org/10.1016/J.CELL.2012.12.034

Li, X., Zou, W., Liu, M., Cao, W., Jiang, Y., An, G., Wang, Y., Huang, S., & Zhao, X. (2016). Association of multiple genetic variants with breast cancer susceptibility in the Han Chinese population. *Oncotarget*, *7*(51), 85483. https://doi.org/10.18632/ONCOTARGET.13402

Li, Z., Ren, M., Tian, J., Jiang, S., Liu, Y., Zhang, L., Wang, Z., Song, Q., Liu, C., & Wu, T. (2015). The Differences in Ultrasound and Clinicopathological Features between Basal-Like and Normal-Like Subtypes of Triple Negative Breast Cancer. *PLoS ONE*, *10*(3). https://doi.org/10.1371/JOURNAL.PONE.0114820

Lynch, H. T., & Lynch, J. F. (1996). Breast cancer genetics: Family history, heterogeneity, molecular genetic diagnosis, and genetic counseling. *Current Problems in Cancer*, *20*(6), 329–365. https://doi.org/10.1016/s0147-0272(96)80010-9

Malhotra, G. K., Zhao, X., Band, H., & Band, V. (2010). Histological, molecular and functional subtypes of breast cancers. *Cancer Biology & Therapy*, *10*(10), 955–960. https://doi.org/10.4161/cbt.10.10.13879

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist.*, *18*(1), 50–60. https://doi.org/10.1214/AOMS/1177730491

Mattiuzzi, C., & Lippi, G. (2019). Current Cancer Epidemiology. *Journal of Epidemiology and Global Health*, *9*(4), 217. https://doi.org/10.2991/JEGH.K.191008.001

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T. H., Wang, Q., Bolla, M. K., Yang, X., Adank, M. A., Ahearn, T., Aittomäki, K., Allen, J., Andrulis, I. L., Anton-Culver, H., Antonenkova, N. N., Arndt, V., … Easton, D. F. (2019). Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *American Journal of Human Genetics*, *104*(1), 21–34. https://doi.org/10.1016/J.AJHG.2018.11.002

Maxwell, K. N., & Nathanson, K. L. (2013). Common breast cancer risk variants in the post-COGS era: a comprehensive review. *Breast Cancer Research 2013 15:6*, *15*(6), 1–17. https://doi.org/10.1186/BCR3591

Mazoyer, S., Dunning, A. M., Serova, O., Dearden, J., Puget, N., Healey, C. S., Gayther, S. A., Mangion, J., Stratton, M. R., Lynch, H. T., Goldgar, D. E., Ponder, B. A. J., & Lenoir, G. M. (1996). A polymorphic stop codon in BRCA2. *Nature Genetics*, *14*(3), 253–254. https://doi.org/10.1038/NG1196-253

Mehta, D., Heim, K., Herder, C., Carstensen, M., Eckstein, G., Schurmann, C., Homuth, G., Nauck, M., Völker, U., Roden, M., Illig, T., Gieger, C., Meitinger, T., & Prokisch, H. (2013). Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *European Journal of Human Genetics*, *21*(1), 48–54. https://doi.org/10.1038/ejhg.2012.106

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., Perkins, B. J., Czene, K., Eriksson, M., Darabi, H., Brand, J. S., Bojesen, S. E., Nordestgaard, B. G., Flyger, H., Nielsen, S. F., … 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), 373–380. https://doi.org/10.1038/ng.3242

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., Wang, Q., Dicks, E., Lee, A., Turnbull, C., Rahman, N., Fletcher, O., Peto, J., Gibson, L., Dos Santos Silva, I., … Easton, D. F. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics*, *45*(4), 353. https://doi.org/10.1038/NG.2563

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., Bolla, M. K., Wang, Q., Tyrer, J., Dicks, E., Lee, A., Wang, Z., Allen, J., Keeman, R., Eilber, U., … Easton, D. F. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature*, *551*(7678), 92. https://doi.org/10.1038/NATURE24284

Millikan, R. C., Newman, B., Tse, C. K., Moorman, P. G., Conway, K., Smith, L. V., Labbok, M. H., Geradts, J., Bensen, J. T., Jackson, S., Nyante, S., Livasy, C., Carey, L., Earp, H. S., & Perou, C. M. (2008). Epidemiology of basal-like breast cancer. *Breast Cancer Research and Treatment*, *109*(1), 123–139. https://doi.org/10.1007/s10549-007-9632-6

Milne, R. L., Kuchenbaecker, K. B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., Hui, S., Lemaçon, A., Soucy, P., Dennis, J., Jiang, X., Rostamianfar, A., Finucane, H., Bolla,

M. K., McGuffog, L., Wang, Q., Aalfs, C. M., Abctctb, I., Adams, M., … Simard, J. (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nature Genetics*, *49*(12), 1767. https://doi.org/10.1038/NG.3785

Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). Application of Student's t-test, Analysis of Variance, and Covariance. *Annals of Cardiac Anaesthesia*, *22*(4), 407. https://doi.org/10.4103/ACA.ACA_94_19

Miyagawa, T., Hasegawa, K., Aoki, Y., Watanabe, T., Otagiri, Y., Arasaki, K., Wakana, Y., Asano, K., Tanaka, M., Yamaguchi, H., Tagaya, M., & Inoue, H. (2019). MT1-MMP recruits the ER-Golgi SNARE Bet1 for efficient MT1-MMP transport to the plasma membrane. *The Journal of Cell Biology*, *218*(10), 3355. https://doi.org/10.1083/JCB.201808149

Muendlein, A., Rohde, B. H., Gasser, K., Haid, A., Rauch, S., Kinz, E., Drexel, H., Hofmann, W., Schindler, V., Kapoor, R., Decker, T., & Lang, A. H. (2015). Evaluation of BRCA1/2 mutational status among German and Austrian women with triple-negative breast cancer. *Journal of Cancer Research and Clinical Oncology*, *141*(11), 2005–2012. https://doi.org/10.1007/s00432-015-1986-2

Nguyen-Dumont, T., Jordheim, L. P., Michelon, J., Forey, N., McKay-Chopin, S., Sinilnikova, O., Le Calvez-Kelm, F., Southey, M. C., Tavtigian, S. V., & Lesueur, F. (2011). Detecting differential allelic expression using high-resolution melting curve analysis: application to the breast cancer susceptibility gene CHEK2. *BMC Medical Genomics*, *4*, 39. https://doi.org/10.1186/1755-8794-4-39

Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1620). https://doi.org/10.1098/RSTB.2012.0362

Nuñez, D. L., González, F. C., Ibargüengoitia, M. C., Corona, R. E. F., Villegas, A. C. H., Zubiate, M. L., Manjarrez, S. E. V., & Velasco, C. C. R. (2020). Papillary lesions of the breast: a review. *Breast Cancer Management*, *9*(4). https://doi.org/10.2217/BMT-2020-0028

Orr, N., Dudbridge, F., Dryden, N., Maguire, S., Novo, D., Perrakis, E., Johnson, N., Ghoussaini, M., Hopper, J. L., Southey, M. C., Apicella, C., Stone, J., Schmidt, M. K., Broeks, A., Van't Veer, L. J., Hogervorst, F. B., Fasching, P. A., Haeberle, L., Ekici, A. B., … Peto, J. (2015). Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. *Human Molecular Genetics*, *24*(10), 2966. https://doi.org/10.1093/HMG/DDV035

Parks, J. W., & Stone, M. D. (2017). Single-Molecule Studies of Telomeres and Telomerase. *Annual Review of Biophysics*, *46*, 357. https://doi.org/10.1146/ANNUREV-BIOPHYS-062215-011256

Pastinen, T. (2010). Insights Into Regulatory Variation. *Nature Publishing Group*, *11*(8), 533–538. http://dx.doi.org/10.1038/nrg2815

Pastinen, T., Ge, B., & Hudson, T. J. (2006). Influence of human genome polymorphism on gene expression. In *Human molecular genetics: Vol. 15 Spec No*. Hum Mol Genet. https://doi.org/10.1093/hmg/ddl044

Pastinen, T., & Hudson, T. J. (2004). Cis-acting regulatory variation in the human genome. In *Science* (Vol. 306, Issue 5696, pp. 647–650). Science. https://doi.org/10.1126/science.1101659

Perou, C. M., & Borresen-Dale, A. L. (2011). Systems biology and genomics of breast cancer. *Cold Spring Harbor Perspectives in Biology*, *3*(2), 1–17. https://doi.org/10.1101/cshperspect.a003293

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Ress, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergammenschlkov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., & Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–752. https://doi.org/10.1038/35021093

Pollán, M., Ascunce, N., Ederra, M., Murillo, A., Erdozáin, N., Alés-Martínez, J. E., & Pastor-Barriuso, R. (2013). Mammographic density and risk of breast cancer according to tumor characteristics and mode of detection: A Spanish population-based case-control study. *Breast Cancer Research*, *15*(1), 1–11. https://doi.org/10.1186/bcr3380

Poorolajal, J., Heidarimoghis, F., Karami, M., Cheraghi, Z., Gohari-Ensaf, F., Shahbazi, F., Zareie, B., Ameri, P., & Sahraei, F. (2021). Factors for the Primary Prevention of Breast Cancer: A Meta-Analysis of Prospective Cohort Studies. *Journal of Research in Health Sciences*, *21*(3), e00520. https://doi.org/10.34172/JRHS.2021.57

Prat, A., Cheang, M. C. U., Martín, M., Parker, J. S., Carrasco, E., Caballero, R., Tyldesley, S., Gelmon, K., Bernard, P. S., Nielsen, T. O., & Perou, C. M. (2013). Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal a breast cancer. *Journal of Clinical Oncology*, *31*(2), 203–209. https://doi.org/10.1200/JCO.2012.43.4134

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., He, X., & Perou, C. M. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research : BCR*, *12*(5), R68. https://doi.org/10.1186/BCR2635

Prat, A., Pineda, E., Adamo, B., Galván, P., Fernández, A., Gaba, L., Díez, M., Viladot, M., Arance, A., & Muñoz, M. (2015). Clinical implications of the intrinsic molecular subtypes of breast cancer. *Breast (Edinburgh, Scotland)*, *24 Suppl 2*, S26–S35. https://doi.org/10.1016/J.BREAST.2015.07.008

Qiu, P., & Sheng, J. (2008). A two-stage procedure for comparing hazard rate functions. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *70*(1), 191–208. https://doi.org/10.1111/j.1467-9868.2007.00622.x

Rafiq, S., Khan, S., Tapper, W., Collins, A., Upstill-Goddard, R., Gerty, S., Blomqvist, C., Aittomäki, K., Couch, F. J., Liu, J., Nevanlinna, H., & Eccles, D. (2014). A genome wide meta-analysis study for identification of common variation associated with breast cancer prognosis. *PloS One*, *9*(12). https://doi.org/10.1371/JOURNAL.PONE.0101488

Razzaghi, H., Troester, M. A., Gierach, G. L., Olshan, A. F., Yankaskas, B. C., & Millikan, R. C. (2013). Association between mammographic density and basal-like and luminal A breast cancer subtypes. *Breast Cancer Research*, *15*(5), R76. https://doi.org/10.1186/bcr3470

Reed, A. E. McCart, Kutasovic, J. R., Lakhani, S. R., & Simpson, P. T. (2015). Invasive lobular carcinoma of the breast: morphology, biomarkers and 'omics. *Breast Cancer Research : BCR*, *17*(1). https://doi.org/10.1186/S13058-015-0519-X

Reed, Amy E. M., Kalinowski, L., Simpson, P. T., & Lakhani, S. R. (2021). Invasive lobular carcinoma of the breast: the increasing importance of this special subtype. *Breast Cancer Research : BCR*, *23*(1). https://doi.org/10.1186/S13058-020-01384-6

Rodgers, K. M., Udesky, J. O., Rudel, R. A., & Brody, J. G. (2018). Environmental chemicals and breast cancer: An updated review of epidemiological literature informed by biological mechanisms. In *Environmental Research* (Vol. 160, pp. 152–182). Academic Press Inc. https://doi.org/10.1016/j.envres.2017.08.045

Ross, J. S., Fletcher, J. A., Linette, G. P., Stec, J., Clark, E., Ayers, M., Symmans, W. F.,

Pusztai, L., & Bloom, K. J. (2003). The Her-2/neu gene and protein in breast cancer 2003: biomarker and target of therapy. *The Oncologist*, *8*(4), 307–325. https://doi.org/10.1634/THEONCOLOGIST.8-4-307

Russo, J., & Russo, I. H. (2006). THE ROLE OF ESTROGEN IN THE INITATION OF BREST CANCER. *The Journal of Steroid Biochemistry and Molecular Biology*, *102*(1–5), 89. https://doi.org/10.1016/J.JSBMB.2006.09.004

Saxena, T., Lee, E., Henderson, K. D., Clarke, C. A., West, D., Marshall, S. F., Deapen, D., Bernstein, L., & Ursin, G. (2010). Menopausal hormone therapy and subsequent risk of specific invasive breast cancer subtypes in the California Teachers Study. *Cancer Epidemiology Biomarkers and Prevention*, *19*(9), 2366–2378. https://doi.org/10.1158/1055-9965.EPI-10-0162

Shapiro, S. S., & Wilk, ; M B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*, 591–611.

Shen, K., Song, N., Kim, Y., Tian, C., Rice, S. D., Gabrin, M. J., Symmans, W. F., Pusztai, L., & Lee, J. K. (2012). A Systematic Evaluation of Multi-Gene Predictors for the Pathological Response of Breast Cancer Patients to Chemotherapy. *PLoS ONE*, *7*(11), 1–9. https://doi.org/10.1371/journal.pone.0049529

Shiovitz, S., & Korde, L. A. (2015). Genetics of breast cancer: A topic in evolution. *Annals of Oncology*, *26*(7), 1291–1299. https://doi.org/10.1093/annonc/mdv022

Shu, X. O., Long, J., Lu, W., Li, C., Chen, W. Y., Delahanty, R., Cheng, J., Cai, H., Zheng, Y., Shi, J., Gu, K., Wang, W. J., Kraft, P., Gao, Y. T., Cai, Q., & Zheng, W. (2012). Novel genetic markers of breast cancer survival identified by a genome-wide association study. *Cancer Research*, *72*(5), 1182–1189. https://doi.org/10.1158/0008-5472.CAN-11-2561

Shukla, A., Edwards, R., Yang, Y., Hahn, A., Folkers, K., Ding, J., Padmakumar, V. C., Cataisson, C., Suh, K. S., & Yuspa, S. H. (2013). CLIC4 regulates TGF-β-dependent myofibroblast differentiation to produce a cancer stroma. *Oncogene 2014 33:7*, *33*(7), 842–850. https://doi.org/10.1038/onc.2013.18

Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(1), 7–30. https://doi.org/10.3322/caac.21590

Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A., Aben, K. K., Strobbe, L. J., Albers-Akkers, M. T., Swinkels, D. W., Henderson, B. E., Kolonel, L. N., Le Marchand, L., Millastre, E., Andres, R., … Stefansson, K. (2007). Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor–positive breast cancer. *Nature Genetics 2007 39:7*, *39*(7), 865–869. https://doi.org/10.1038/ng2064

Steinberg, K. K., Thacker, S. B., Smith, S. J., Stroup, D. F., Zack, M. M., Flanders, W. D., & Berkelman, R. L. (1991). A Meta-analysis of the Effect of Estrogen Replacement Therapy on the Risk of Breast Cancer. *JAMA*, *265*(15), 1985–1990. https://doi.org/10.1001/JAMA.1991.03460150089030

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1 (Mar, 1908)), 1–25.

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*, *0*(0), 1–41. https://doi.org/10.3322/caac.21660

Tamimi, R. M., Colditz, G. A., Hazra, A., Baer, H. J., Hankinson, S. E., Rosner, B., Marotti, J., Connolly, J. L., Schnitt, S. J., & Collins, L. C. (2012). Traditional breast cancer risk

factors in relation to molecular subtypes of breast cancer. *Breast Cancer Research and Treatment*, *131*(1), 159–167. https://doi.org/10.1007/s10549-011-1702-0

Trabert, B., Sherman, M. E., Kannan, N., & Stanczyk, F. Z. (2020). Progesterone and Breast Cancer. *Endocrine Reviews*, *41*(2), 320. https://doi.org/10.1210/ENDREV/BNZ001

Trivers, K. F., Lund, M. J., Porter, P. L., Liff, J. M., Flagg, E. W., Coates, R. J., & Eley, J. W. (2009). The epidemiology of triple-negative breast cancer, including race. *Cancer Causes and Control*, *20*(7), 1071–1082. https://doi.org/10.1007/s10552-009-9331-1

Villarreal-Garza, C., Weitzel, J. N., Llacuachaqui, M., Sifuentes, E., Magallanes-Hoyos, M. C., Gallardo, L., Alvarez-Gómez, R. M., Herzog, J., Castillo, D., Royer, R., Akbari, M., Lara-Medina, F., Herrera, L. A., Mohar, A., & Narod, S. A. (2015). The prevalence of BRCA1 and BRCA2 mutations among young Mexican women with triple-negative breast cancer. *Breast Cancer Research and Treatment*, *150*(2), 389–394. https://doi.org/10.1007/s10549-015-3312-8

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. https://doi.org/10.1016/J.AJHG.2011.11.029

Vogiatzi, F., Brandt, D. T., Schneikert, J., Fuchs, J., Grikscheit, K., Wanzel, M., Pavlakis, E., Charles, J. P., Timofeev, O., Nist, A., Mernberger, M., Kantelhardt, E. J., Siebolts, U., Bartel, F., Jacob, R., Rath, A., Moll, R., Grosse, R., & Stiewe, T. (2016). Mutant p53 promotes tumor progression and metastasis by the endoplasmic reticulum UDPase ENTPD5. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(52), E8433–E8442. https://doi.org/10.1073/pnas.1612711114

Warburg, O. (1931). The Metabolism of Tumours: Investigations from the Kaiser Wilhelm Institute for Biology, Berlin-Dahlem. *JAMA: The Journal of the American Medical Association*, *96*(23), 1982. https://doi.org/10.1001/jama.1931.02720490062043

Weigelt, B., Geyer, F. C., & Reis-Filho, J. S. (2010). Histological types of breast cancer: How special are they? In *Molecular Oncology* (Vol. 4, Issue 3, pp. 192–208). John Wiley and Sons Ltd. https://doi.org/10.1016/j.molonc.2010.04.004

Wen, H. Y., & Brogi, E. (2018). Lobular Carcinoma in Situ. *Surgical Pathology Clinics*, *11*(1), 123. https://doi.org/10.1016/J.PATH.2017.09.009

Whitley, E., & Ball, J. (2002). Statistics review 5: Comparison of means. *Critical Care*, *6*(5), 424. https://doi.org/10.1186/CC1548

Wong, E., Chaudhry, S., & Rossi, M. (2012). *Breast cancer | McMaster Pathophysiology Review*. http://www.pathophys.org/breast-cancer/

Xiao, R., & Scott, L. J. (2011). Detection of cis-acting regulatory SNPs using allelic expression data. *Genetic Epidemiology*, *35*(6), 515–525. https://doi.org/10.1002/gepi.20601

Xing, P., Li, J., & Jin, F. (2010). A case-control study of reproductive factors associated with subtypes of breast cancer in Northeast China. *Medical Oncology*, *27*(3), 926–931. https://doi.org/10.1007/s12032-009-9308-7

Yang, X. R., Sherman, M. E., Rimm, D. L., Lissowska, J., Brinton, L. A., Peplonska, B., Hewitt, S. M., Anderson, W. F., Szeszenia-Dąbrowska, N., Bardin-Mikolajczak, A., Zatonski, W., Cartun, R., Mandich, D., Rymkiewicz, G., Ligaj, M., Lukaszek, S., Kordek, R., & García-Closas, M. (2007). Differences in risk factors for breast cancer molecular subtypes in a population-based study. *Cancer Epidemiology Biomarkers and Prevention*, *16*(3), 439–443. https://doi.org/10.1158/1055-9965.EPI-06-0806

Yedjou, C. G., Sims, J. N., Miele, L., Noubissi, F., Lowe, L., Fonseca, D. D., Alo, R. A.,

Payton, M., & Tchounwou, P. B. (2019). Health and Racial Disparity in Breast Cancer HHS Public Access. *Adv Exp Med Biol*, *1152*, 31–49. https://doi.org/10.1007/978-3-030-20301-6_3

Yoshihama, S., Roszik, J., Downs, I., Meissner, T. B., Vijayan, S., Chapuy, B., Sidiq, T., Shipp, M. A., Lizee, G. A., & Kobayashi, K. S. (2016). NLRC5/MHC class I transactivator is a target for immune evasion in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(21), 5999–6004. https://doi.org/10.1073/pnas.1602069113

Zhang, H., Ahearn, T. U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., Jiang, X., O'Mara, T. A., Zhao, N., Bolla, M. K., Dunning, A. M., Dennis, J., Wang, Q., Ful, Z. A., Aittomäki, K., Andrulis, I. L., Anton-Culver, H., Arndt, V., Aronson, K. J., … García-Closas, M. (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature Genetics*, *52*(6), 572–581. https://doi.org/10.1038/S41588-020-0609-2

Zhao, M. Z., Sun, Y., Jiang, X. F., Liu, L., & Sun, L. X. (2019). Promotion on NLRC5 upregulating MHC-I expression by IFN-γ in MHC-I-deficient breast cancer cells. *Immunologic Research*, *67*(6), 497–504. https://doi.org/10.1007/S12026-019-09111-W

Zielonke, N., Kregting, L. M., Heijnsdijk, E. A. M., Veerus, P., Heinävaara, S., McKee, M., de Kok, I. M. C. M., de Koning, H. J., van Ravesteyn, N. T., Gredinger, G., De Brabander, I., Arbyn, M., Simoens, C., Martens, P., Candeur, M., Arbyn, M., Simoens, C., Burrion, J. B., Dimitrov, P., … Latinovic, R. (2021). The potential of breast cancer screening in Europe. *International Journal of Cancer*, *148*(2), 406. https://doi.org/10.1002/IJC.33204