CATARINA PINTO MARTINS

IDENTIFYING NOVEL GENES ASSOCIATED WITH BREAST CANCER SUSCEPTIBILITY USING DIFFERENTIAL ALLELIC EXPRESSION RATIOS

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2018

CATARINA PINTO MARTINS

IDENTIFYING NOVEL GENES ASSOCIATED WITH BREAST CANCER SUSCEPTIBILITY USING DIFFERENTIAL ALLELIC EXPRESSION RATIOS

Master in Oncobiology – Molecular Mechanisms of Cancer

This work was done under the supervision of

Ana Teresa Maia, Ph.D

Joana Xavier, Ph.D

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2018

IDENTIFYING NOVEL GENES ASSOCIATED WITH BREAST CANCER SUSCEPTIBILITY USING DIFFERENTIAL ALLELIC EXPRESSION RATIOS

Declaração de autoria de trabalho

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

_____

Catarina Pinto Martins

However difficult life may seem, there is always something you can do, and succeed at. It matters that you don't just give up.

Stephen Hawking

ACKNOWLEDGEMENTS

Em primeiro lugar, quero agradecer à minha orientadora Professora Doutora Ana Teresa Maia pela oportunidade de me integrar no seu laboratório, por todo o apoio, força e dedicação prestada, por ser um dos pilares fundamentais para a realização deste trabalho, por toda a calma transmitida e boa energia, por ter acreditado em mim, um enorme obrigada.

Quero também agradecer à minha coorientadora, Doutora Joana Xavier, por todo o seu contributo para a realização deste trabalho, por toda a motivação transmitida, por toda a paciência, por todos os seus ensinamentos, por todos os risos e também desesperos, por toda a dedicação e força transmitida, nada seria o mesmo sem ti.

Gostaria também de agradecer ao meu colega de grupo Ramiro Magno por toda a paciência e sabedoria transmitida, por todos os bons momentos partilhados. À minha colega de grupo Filipa Esteves um grande obrigada por todo o companheirismo e calma transmitida, és incrível.

À Juliana Machado, amiga de todos os dias, de todas as horas, de todos os minutos e segundos um muito obrigada. Guardo todos os nossos risos, todos os choros, guardo todos os momentos no coração. À Rita Ferreira um enorme obrigada por me fazer sempre libertar um sorriso e pelo apoio dado.

Um grande obrigado por toda a força e amizade à Ana Fernandes, Mariana Jordão, Ricardo Carvalho e Rodrigo Mota, sem vocês nada seria igual.

Às minha amigas/o desde sempre e para sempre Patrícia Couto, Ana Luísa, Sara Neves, Sara Ferreira, Cláudia Moutinho e César Moutinho eternamente grata por fazerem parte da minha vida.

Por fim quero agradecer à minha família por serem os pilares fundamentais da minha vida. Pai e Mãe, sem vocês nada disto seria possível, obrigada por toda a força, ensinamentos e amor. Mana obrigada por toda a amizade, amor e conselhos, és a melhor.

ABSTRACT

Breast Cancer (BC) is the most common cancer among women worldwide. However, the current knowledge of BC susceptibility only accounts for half of the familial cases. The few functional studies performed for genome-wide association studies (GWAS) loci revealed a role for cis-regulatory variation, suggesting that risk variants may be acting by regulating gene expression levels. Therefore, we hypothesise that the most efficient approach to tackle BC missing heritability is to focus susceptibility studies on variants with greater cis-regulatory potential. Hereby, we present an innovative approach to genetic association studies, using a quantifiable readout of the effect of cis-regulatory variants — differential allelic expression (DAE).

To identify candidate risk genes for our study, we selected Single Nucleotide Polymorphisms (SNPs) weakly associated with BC risk in GWAS and in the iCOGS consortium and identified their proxy SNPs. The resulting 591 candidate risk variants were located in 92 different genes, of which 41 had evidence of being cis-regulated in a DAE study of normal breast tissue. The clinical impact of these genes was assessed, for a diverse list of clinical variables (differential expression analysis, FDR $\leqslant$ 1% and absolute fold-change $\geqslant$1.5). A final list of 18 risk candidates cis-regulated and with clinical impact genes was identified.

OCIAD1 and *GRHL2* genes were selected to perform case-control association studies using DAE values. DAE of *OCIAD1* was significantly associated with BC risk (p-value=0.002 and 0.008, in two independent experiments using blood samples), while DAE of *GRHL2* needs further validation of association (p-value=0.014 and 0.096, in two independent experiments in breast tissue).

This project proved that association studies using DAE as a quantifiable variable, together with the whole pipeline used to select the candidate genes, is an efficient approach to detect novel risk genes for BC.

Keywords

breast cancer • risk • cis-regulation • *OCIAD1* • *GRHL2*

RESUMO

O cancro da mama é o tipo de cancro mais diagnosticado em mulheres, tanto em países desenvolvidos como em países em vias de desenvolvimento, representando 25% de todos os cancros mais diagnosticados. É uma doença caracterizada pelo crescimento anormal de células da mama conduzindo à formação do tumor. Esta neoplasia pode ser classificada segundo a sua morfologia e histologia básica, mas também a nível molecular. Trata-se de uma doença complexa, com uma componente genética e ambiental. Os fatores de risco para desenvolvimento de cancro da mama podem ser modificáveis, como é o caso da diminuição do contacto com radiação ionizante, consumo de hormonas femininas exógenas, entre outros, mas também existem fatores não modificáveis como é o caso da genética e da herança familiar. Cerca de 10-30% dos cancros da mama estão relacionados com fatores hereditários e dentro destes, entre 5-10% dos casos têm uma forte componente herdada. Os alelos genéticos podem ser categorizados de acordo com o risco e com a sua frequência na população, em alelos de alto risco, como é o caso de mutações nos genes *BRCA1*/*BRCA2,* alelos de risco moderado, como as mutações em *ATM* e *BRIP1* e em alelos que conferem baixo risco na população, como o caso dos polimorfismos genéticos identificados nos genes *FGFR2* e *TOX3*. No entanto cerca de metade da componente de risco familiar para o cancro da mama permanece por identificar.

Os estudos de associação do genoma (GWAS) permitiram a identificação de muitos alelos de risco sem conhecimento prévio da posição ou função do gene. Estes estudos em cancro da mama revelaram que os polimorfismos de variante única (SNPs) associados a risco para a doença estão presentes maioritariamente em regiões não codificantes e os poucos estudos funcionais realizados nestes loci mostraram que variantes cis-reguladoras causavam risco através da regulação da expressão genética. As variantes em cis alteram a síntese dos transcritos de forma específica para cada alelo, podendo estar localizadas em regiões promotoras do gene, *enhancers*, bem como a centenas de kilobases (kb) de distância. A medição do rácio de expressão do RNA entre os dois alelos permite a deteção direta do efeito destas variantes, designando-se por análise de expressão alélica diferencial (DAE).

Posto isto, sugere-se que para identificar novas variantes associadas a risco para cancro da mama deve-se focar os estudos em variantes cis-regulatórias e propõe-se utilizar rácios de DAE como medida quantitativa em estudos de associação genética.

Para selecionar os genes candidatos para serem testados nos estudos de associação usando níveis de DAE usou-se a seguinte abordagem: primeiro identificaram-se 608 variantes candidatas a conferir risco para cancro da mama, uma vez que mostraram evidência de estarem associadas com risco para cancro da mama, mas não atingiram significância estatística nos GWAS. Estas variantes e os seus proxy SNPs ($r^2 \geq 0.8$) estavam localizadas em 92 genes diferentes. Visto que se pretendia selecionar genes para validação no laboratório que tivessem cis-regulados, filtrou-se os 92 genes candidatos a risco com os dados de DAE do grupo (que indicavam quais os genes com evidência de cis-regulação), ficando-se com 41 genes cis-regulados e com evidência de poderem contribuir para risco de desenvolver cancro da mama. Com o objetivo de se selecionarem genes com possível impacto clínico, realizaram-se análises de associação estatística entre expressão genética no genoma inteiro (derivada de *microarrays* de ácido desoxirribonucleico complementar (cADN) e diferentes variáveis clínicas (recetor de estrogénio, recetor de progesterona, grau, entre outras). Com esta análise de expressão diferencial verificou-se que 18 dos genes anteriores possuíam impacto clínico em pelo menos uma das análises consideradas (*p-value*≤0,01 (1% False Discovery Rate correction) e |*Fold Change*|≥ 1,5) e foram estabelecidos os genes finais de interesse. Destes 18 genes, selecionaram-se dois para a realização do estudo caso-controlo, com base nas evidências de DAE das análises efetuadas no grupo anteriormente, na significância estatística para o risco observada nas primeiras fases dos GWAS, no impacto clínico verificado nas análises de expressão diferencial, na frequência do alelo mais raro do SNP localizado nesse gene), na expressão total dos transcritos no tecido saudável da mama, bem como na presença de *expression quantitative trait* loci para o SNP transcrito (tSNP) de cada gene candidato. Após esta análise o gene *OCIA domain containing 1* (*OCIAD1*), e nomeadamente o seu tSNP rs9997920 (alelos C/T (minor)), assim como o *Grainyhead Like Transcription Factor 2* (*GRHL2*) e o seu tSNP rs6989650 (alelos C/T (minor)) foram selecionados para realização dos estudos

caso-controlo, tendo sido o *OCIAD1* testado em tecido da mama e no sangue e o *GRHL2* apenas em mama.

Para cada estudo caso-controlo foi incorporada uma curva de calibração com uma amostra heterozigótica de modo a quantificar-se de forma precisa o DAE e uma curva standard com amostras heterozigóticas com os dois alelos em diferentes proporções para servir de controlo positivo. Os estudos de associação revelaram que no gene *OCIAD1* no sangue existe uma diferença significativa entre os níveis de DAE nos casos e dos controlos (p-value= 0.002 e 0.008 em duas experiências independentes), tendo o alelo menos frequente do rs9997920 uma maior expressão nos casos do que nos controlos. No tecido mamário não se verificou esta diferença nos níveis de DAE do *OCIAD1* (*p-value*= 0.4 e 0.07 para a segunda experiência), o que nos leva a concluir que apenas o sangue poderá ser usado para inferir o risco conferido por este gene. No entanto, e uma vez que os coortes populacionais do sangue são diferentes dos do tecido, este estudo deveria ser repetido usando sangue e tecido das mesmas pessoas. O estudo de associação genética para o *GRHL2* revelou que este é um potencial gene de risco para cancro da mama uma vez que apresentava níveis de DAE diferentes entre casos e controlos no tecido da mama, mas apenas numa das experiências realizadas (=0,014 e 0.096 para a segunda experiência). Verificou-se que o alelo comum do rs6989650 está mais expresso nos casos do que nos controlos o que nos leva a concluir que este possivelmente é o alelo de risco, mas mais estudo têm que ser feitos para garantir que realmente a presença desta variante leva a risco para o desenvolvimento da doença.

Este estudo permitiu a identificação de potenciais novos genes associados a suscetibilidade para cancro da mama, e gerou uma lista de novos genes candidatos para serem testados no futuro para associação com cancro da mama, através da análise de DAE. Futuras repetições das experiências têm que ser realizadas para garantir que o *OCIAD1* e o *GRHL2* são genes de risco para cancro da mama. Ao confirmarem-se, mais estudos deverão ser feitos de modo a identificar as variantes causais bem como o mecanismo pelo qual estarão a provocar o risco.

Palavras-chave

Cancro da mama • risco • cis-regulação • *OCIAD1* • *GRHL2*

CONTENT

INDEX OF FIGURES

INDEX OF TABLES

INDEX OF ANNEXES

LIST OF ABBREVIATIONS

ATM – Ataxia Telangiectasia mutated

AE – Allelic Expression

BC – Breast Cancer

BCAC – Breast Cancer Association Consortium

BMI – Body Mass Index

BRCA1 – Breast Cancer type 1

BRCA2 – Breast Cancer type 2

BRIP1 – BRCA1 – Interacting protein 1

CDH1 – Cadherin 1 gene

cDNA – complementary DNA

CEPH – Centre d'étude du polymorphism humain

CHB – Han Chinese in Beijing, China

CHD – Chinese in Metropolitan Denver, Colorado

CHEK2 – Checkpoint Kinase 2

COGS – Collaborative Oncological Gene-environment Study

DAE – Differential Allelic Expression analysis

DCIS – Ductal Carcinoma in Situ

DNA – Deoxyribonucleic Acid

dNTPs – deoxynucleotide triphosphates

EGA – European Genome-phenome Archive

EMT- Epithelial-mesenchymal transition

ER – Oestrogen Receptor

ER^- - Oestrogen Receptor Negative

ER+ - Oestrogen Receptor Positive

eQTL – expression quantitative trait loci

FC – Fold Change

FDR – False Discovery Rate

GRHL2 – GrainyHead Like Transcription Factor 2

GTEx – Genotype-Tissue Expression Project

GWAS – Genome-wide association studies

HapMap – The Haplotype Map Project

HER2 $^{+}$ - HER2 amplified

HER2 $^{-}$ - HER2 not amplified

iCOGS – illumina Collaborative Oncological Gene-environment study

JPT – Japanese in Tokyo, Japan

Kb - kilobases

LCIS – Lobular Carcinoma in Situ

LCLs – Lymphoblastoid cell lines

LD – Linkage Disequilibrium

LN – Lymph node

MAF – Minor allele frequency

METABRIC – Molecular Taxonomy of Breast Cancer International Consortium

mRNA – messenger RNA

NB – Normal Breast

NCBI-dbSNP – dnSNP from National Center for biotechnology information database

NM- Normal-matched

NTC – No Template Control

OCIAD1 – OCIA Domain Containing 1

Onekgpilot – 1000 genomes pilot 1

PALB2 – Partner and localizer of BRCA2

PCR – Polymerase Chain Reaction

PR – Progesterone Receptor

PR^- - Progesterone Receptor Negative

PR+ – Progesterone Receptor Positive

PTEN – Phosphatase and Tensin Homolog gene

RFU – Relative Fluorescent Unit

RNA – Ribonucleic Acid

rSNPs – regulatory variants

RT – Reverse transcriptase

RT-PCR – Reverse Transcriptase Polymerase Chain Reaction

RT-qPCR – Real-Time Quantitative Polymerase Chain Reaction

SNP – Single Nucleotide Polymorphisms

STK11 – Serine/Threonine kinase 11 gene

TNBCs – Triple Negative Breast Cancers

TP53 – Tumour Protein 53 gene

tSNP – transcribed SNP

YRI – Yoruba in Ibadan, Nigeria

3'UTR – 3' Untranslated Region

CHAPTER I

Introduction

1. INTRODUCTION

1.1. BREAST CANCER

1.1.1. Epidemiology

Cancer is a public health problem counting 14.1 million new cases and 8.2 million deaths occurring worldwide every year. The most common cancer worldwide is lung cancer, followed by breast cancer (BC) (Ferlay et al. 2015). However, in women, BC is the most common diagnosed cancer in developed and developing countries, with approximately 1.7 million new cases, representing 25% of all cancers (Tao et al. 2015; Torre et al. 2015; Ferlay et al. 2015). BC is also the fifth cause of death worldwide, the most frequent cause of cancer death in women in developing countries, and the second cause of death by cancer in developed countries (**Figure 1.1**) (Ferlay et al. 2015; Gómez-Flores-Ramos et al. 2017).

Based on data from the American Cancer Society, one in eight women in the United States will have BC in her lifetime. The prediction is that by 2050 there will be approximately 3.2 million new cases per year of female BC (Tao et al. 2015).

In Portugal, female BC is the most common cancer with an estimated 6088 new cases each year, being also the leading cause of cancer mortality in women, with 1570 deaths in 2012. These represent, respectively, 30% of all cancer cases and 16% of all cancer deaths (Forjaz de Lacerda et al. 2018). These numbers also reflect the magnitude of new cases per year of BC, its impact on society worldwide and the urgency for improving prevention and treatments (Tao et al. 2015).

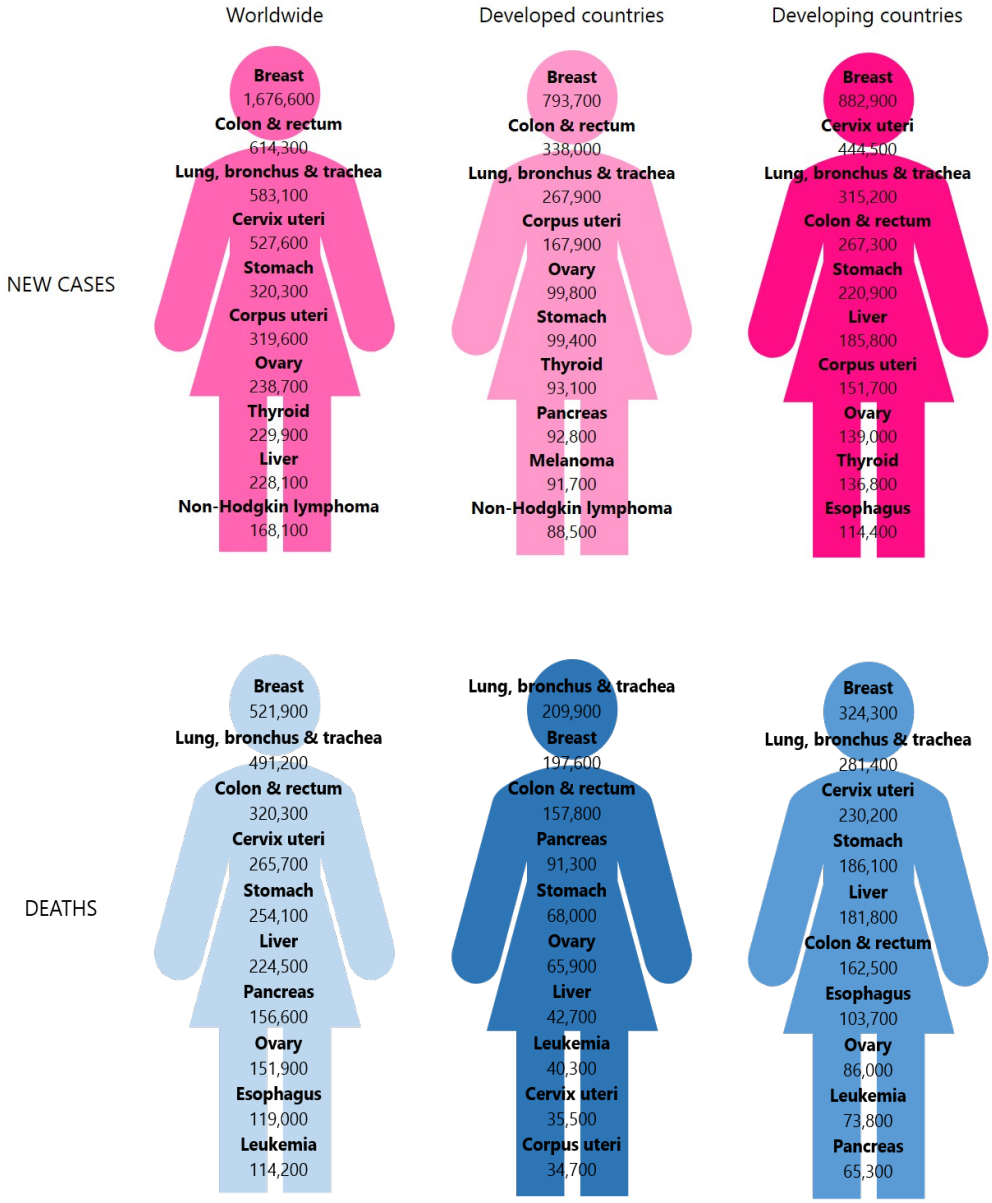


Figure 1.1 - Epidemiology of Breast Cancer. Representation of the estimated new cases and deaths for female cancer worldwide, in developed and developing countries. Cancer types are ordered according to prevalence, from the most to the less common (Adapted from Torre et al. 2015)

1.1.2. Morphological and Molecular Classification of Breast Tumours

Breast tumours, characterised by the abnormal growth of breast cells to form a malignant neoplasm (Apostolou and Fostira 2013), are genetically and clinically diverse in their natural history and in their response to treatment (Stingl

and Caldas 2007; Malhotra et al. 2010). A portion of this biological diversity can be explained by changes in gene expression (C. M. C. M. Perou et al. 2000) but also to other factors such as tumour microenvironment, cell signalling, among others (Natrajan et al. 2016).

Breast cancer can be classified based on tumour morphology and basic histology (Malhotra et al. 2010; Tao et al. 2015). There are two major types of BC: In situ Carcinoma and Invasive/Infiltrating Carcinoma. The In Situ carcinoma can be Ductal (DCIS) or Lobular (LCIS); Invasive carcinoma can be Tubular, Ductal/Lobular, Lobular, Infiltrating Ductal, Mucinous, Medullary or Papillary **(Figure 1.2)**. Regarding to In Situ Carcinoma, DCIS is the most common, with DCIS being further subclassified relative to its architectural characteristics into five subtypes: Comedo, Cribiform, Micropappilary, Papillary and Solid; LCIS presents low histological variation and is not further subdivided. The Infiltrating Ductal Carcinoma is the most common in the population, accounting for 70-80% of all invasive lesions (Malhotra et al. 2010).

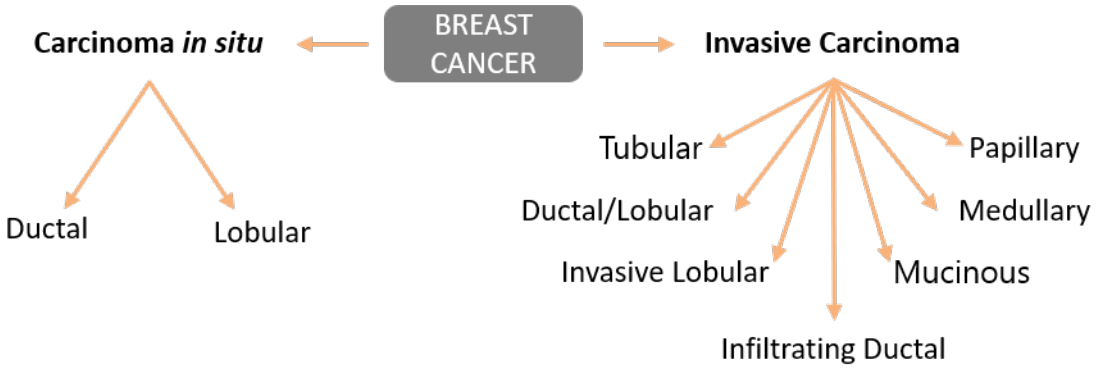


Figure 1.2 – Histological Breast Cancer classification.

Breast tumours can also be molecularly classified, where gene expression profiling is combined with molecular markers, such as hormone receptors (C. M. C. M. Perou et al. 2000). The first such classification was the PAM50 classification, which is based on an algorithm that uses the expression level of 50 genes, to identify the five major intrinsic subtypes: Luminal A (oestrogen receptor (ER) positive- ER+/progesterone receptor (PR) positive- PR+/*HER2* not amplified or not overexpressed- HER2-), Luminal B (ER+/progesterone receptor negative- PR-

?/*HER2* amplified or overexpressed- HER2+), Basal-like (oestrogen receptor negative- ER-/PR-/Her2-'triple-negatives'), HER2-enriched (ER-/Her2+) and Normal-like (C. M. C. M. Perou et al. 2000; Sorlie et al. 2001; C. M. Perou and Borresen-Dale 2011).

In general, in terms of hormonal receptor expression, tumours that express the oestrogen receptor (ER+) are more common, smaller, with a lower grade and lymph node negative when compared to those that do not express oestrogen receptor (ER-) (Anderson et al. 2002; Tao et al. 2015).

The luminal tumours are defined by expression of hormonal genes such as *ESR1*, *PGR* (C. M. C. M. Perou et al. 2000). Luminal A tumours normally have a higher expression of ER and a lower expression of HER2, as well as lower expression of proliferation-associated genes, like *MKI67* (Sorlie et al. 2001; Sorlie et al. 2003; Hu et al. 2006; C. M. Perou and Borresen-Dale 2011). Luminal B tumours usually have higher proliferation rates, tend to have mutations in the *TP53* gene, and normally show a lower expression of ER- responsive genes (C. M. Perou and Borresen-Dale 2011). The most common luminal subtype is A, representing about 40% of all breast tumours, whilst the B subtype comprises approximately 10% of all breast tumours(C. M. Perou and Borresen-Dale 2011).

There are subtypes with low expression of hormone receptors and their regulated genes that are HER2-enriched, basal-like (histologically comprised of more basal/myoepithelial cells) and claudin-low (claudin is a protein associated with tight-junctions). The HER2 enriched subtype is rather infrequent, accounting for 10% of all BC and normally shows a higher expression of HER2 and other genes The basal-like subtype, corresponding to 10-25% of all tumours, is defined by the lower expression of characteristic genes of luminal BC, low expression of HER2, high expression of genes involved in proliferation and a higher expression of a cluster of genes named basal cluster (C. M. Perou and Borresen-Dale 2011). Usually, these tumours are defined as triple-negative BC (TNBCs) representing 75% of basal-like tumours (C. M. Perou and Borresen-Dale 2011). Basal-like tumours also present a higher frequency of *TP53* mutations (approximately 80%) (Koboldt et al. 2012) and are common in patients with germline BRCA1 mutations or of African ancestry (Morris et al. 2007; Dent et al. 2007; Koboldt et al. 2012).

Claudin-low subtype is defined by the low expression of genes involved in tight junctions and cell to cell adhesion (C. M. Perou and Borresen-Dale 2011). Relatively to prognosis, both HER2 and basal-like subtypes presents a significantly poorer outcome when compared with luminal and normal-like subtypes (Sorlie et al. 2001), and that HER2 and basal-like are also related with advanced stage (Iwase et al. 2010).

These classification systems are constantly updated and improved to support development of novel treatments and to help improve disease prognosis prediction.

1.1.3. Aetiology

Risk factors to develop BC can be inherited (discussed in the next subchapter), histopathological or environmental (Sauter 2018).

Some of the risk factors can be modified, such as diet, exercise, tobacco and alcohol consumption, female hormones (exogenous), ionizing radiation, pregnancy and breastfeeding (Key, Verkasalo, and Banks 2001; Key et al. 2003; Sauter 2018). Other risk factors, like age of menarche, menopause, anthropometry, family history and genetic factors, cannot be modified (**Figure 1.3**) (Key, Verkasalo, and Banks 2001).

Pregnancy and nursing are known to decrease the risk to BC (Layde et al. 1989; Ewertz et al. 1990; Tao et al. 2015; Torre et al. 2015). Women who have had at least on child present a reduction in risk to the disease around 25%, but it is also known that protection is higher the younger the age at first pregnancy, as well as on first breastfeeding (Layde et al. 1989; Ewertz et al. 1990).

It has also been observed that the older the woman at menarche, the lower the risk for BC. Relatively to menopause, those who enter it at a later age are at a higher risk of disease than those who terminate menstruating earlier, with a risk increasing of about 3% for each year older at menopause (Key, Verkasalo, and Banks 2001). However, increased used of post-menopausal hormones contribute to the steady rise in incidence of ER/PR positive BC, because these hormones are able to promote growth and proliferation (Colditz 2007). It is also relevant that

women who take oral contraceptives currently are at a higher risk, approximately 25% (Calle et al. 1996).

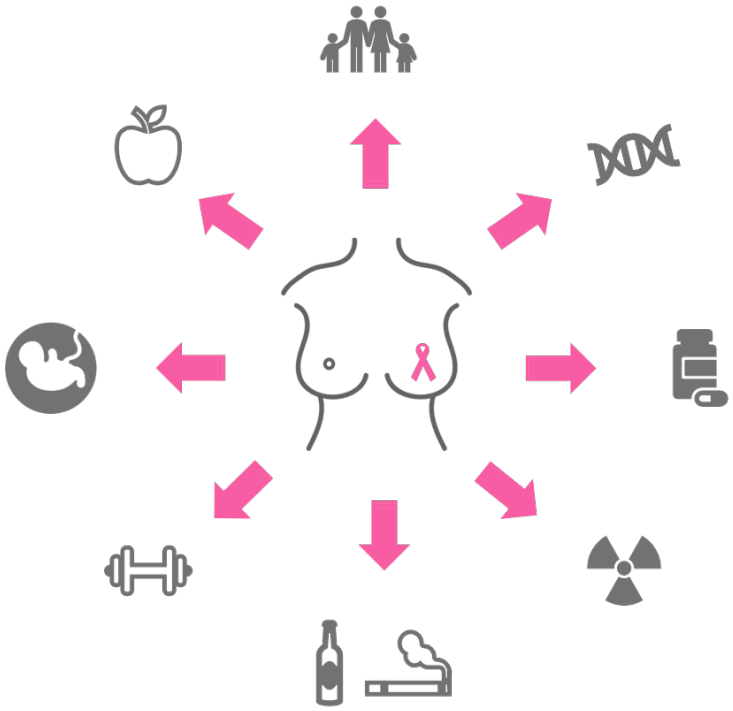


Figure 1.3 - Risk factors for BC development. Representation of the principal risk factors, including the modifiable ones (such as pregnancy, physical activity, alcohol and tobacco consumption, contact with ionizing radiation, consumption of female exogenous hormones and diet) and the non-modifiable factors (such as genetics and familial inheritance).

Another risk factor for BC development is ionizing radiation. An extensive follow-up of many populations exposed to radiation showed that the breast is highly sensitive to the radiation effects (Boice Jr. et al. 1979).

It was also demonstrated that adult height has a weak positive relation with BC risk (Hunter and Willett 1993). Obesity is a common problem in many cancers such as BC, endometrial cancer, ovary cancer, among others (Sauter 2018). The strongest relation between nutrition and BC is for relative body weight, as indicated by body mass index (BMI) (Key et al. 2003). Studies have revealed that women with a high BMI and with more adipose tissue have an increment of aromatase, which is the responsible enzyme to catalyse the conversion of androstenedione to oestrogen, which can be converted to oestradiol. So, it has been proposed that the higher the BMI the higher the concentration of free

oestradiol, which increases the risk to BC (Key, Verkasalo, and Banks 2001; Key et al. 2003).

Epidemiological studies demonstrated also a positive association between smoking and BC in current alcohol drinkers (Sauter 2018), and that alcohol consumption is related with a moderate increment in the risk for the disease (Hamajima et al. 2002).

Physical activity has been associated with a lower risk to develop BC (Key, Verkasalo, and Banks 2001) and was related with energy intake and partly determines BMI in many studies. It was found in epidemiological studies that women more physically active have a reduction of risk to BC around 40%, when compared with sedentary women (Key et al. 2003).

There is a nutritional hypothesis that says that obesity and an elevated intake of meat, dairy products, fat and alcohol probably increases the risk of BC, and that a high intake of fibres, fruits, vegetables, antioxidants and phytoestrogens reduces BC risk (Key et al. 2003).

1.2. FAMILIAL BREAST CANCER

The observation of cancer clustering in families and the increased cancer susceptibility in individuals with some genetically determined syndromes, first revealed the existence of a genetic component to BC risk. Nevertheless, familial aggregation can be attributed both to shared genes and to shared physical environments and lifestyles (Key, Verkasalo, and Banks 2001). Over time, the knowledge of the BC heritability has increased significantly as represented in **Figure 1.4** (Eccles et al. 2013).

Approximately 10-30% BC cases are related with hereditary factors and only 5-10% have a robust inherited component, with identified high deleterious mutations transmitted in an autosomal dominant manner (Newman et al. 1988; Claus, Risch, and Thompson 1991; Apostolou and Fostira 2013; Rich et al. 2015; Gómez-Flores-Ramos et al. 2017).

Genetic alleles can be categorized according to their relative risk (high, moderate and low penetrance alleles) (Ghoussaini, Pharoah, and Easton 2013) and

to the risk allele frequency (**Figure 1.5**) (Apostolou and Fostira 2013; Ghoussaini, Pharoah, and Easton 2013). High penetrant alleles confer a disease relative risk higher than 5, and intermediate-penetrant alleles confer a relative risk around 1.5-5. Low penetrant loci present a relative risk of about 1.5 (Apostolou and Fostira 2013; Gómez-Flores-Ramos et al. 2017).

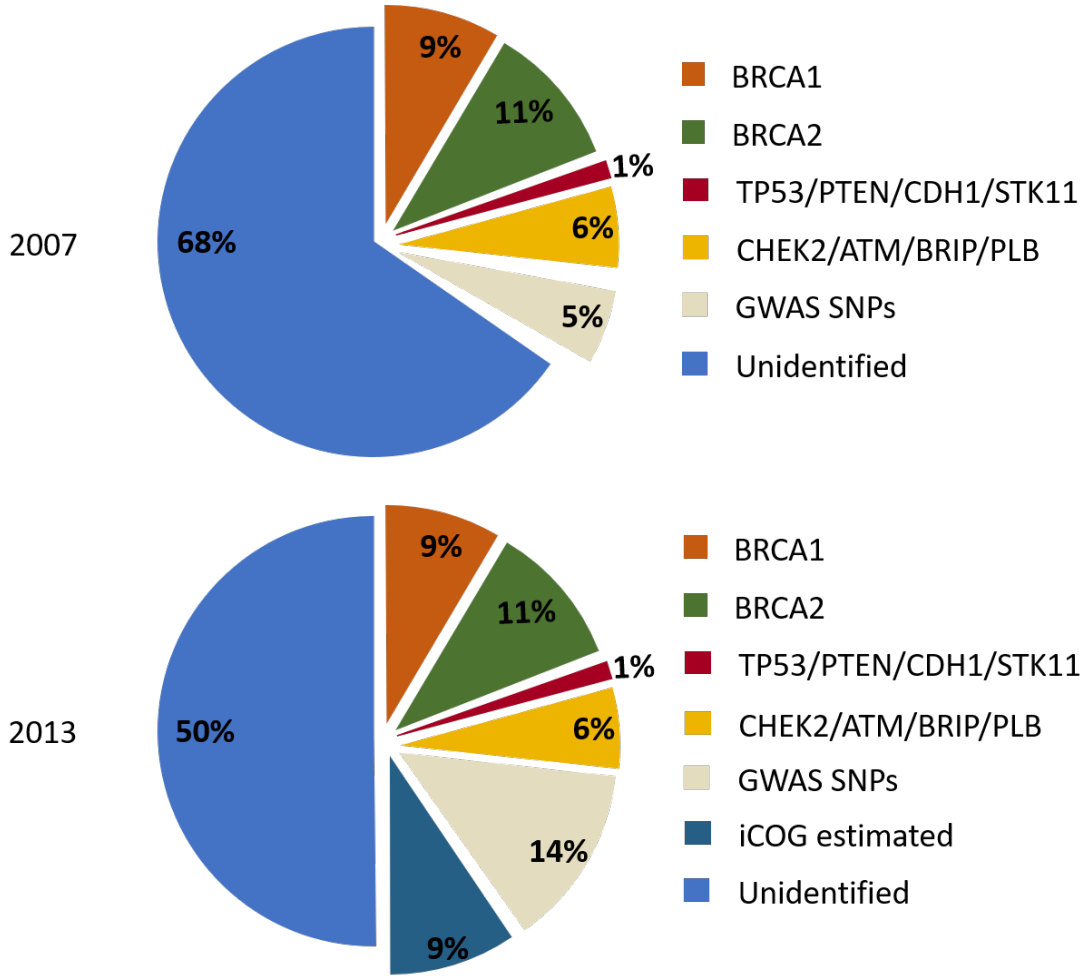


Figure 1.4 - Familial BC genetics. Percentage of heritability attributed to different genetic factors and unidentified fraction, in 2007 and 2013. Adapted from Eccles et al. 2013.

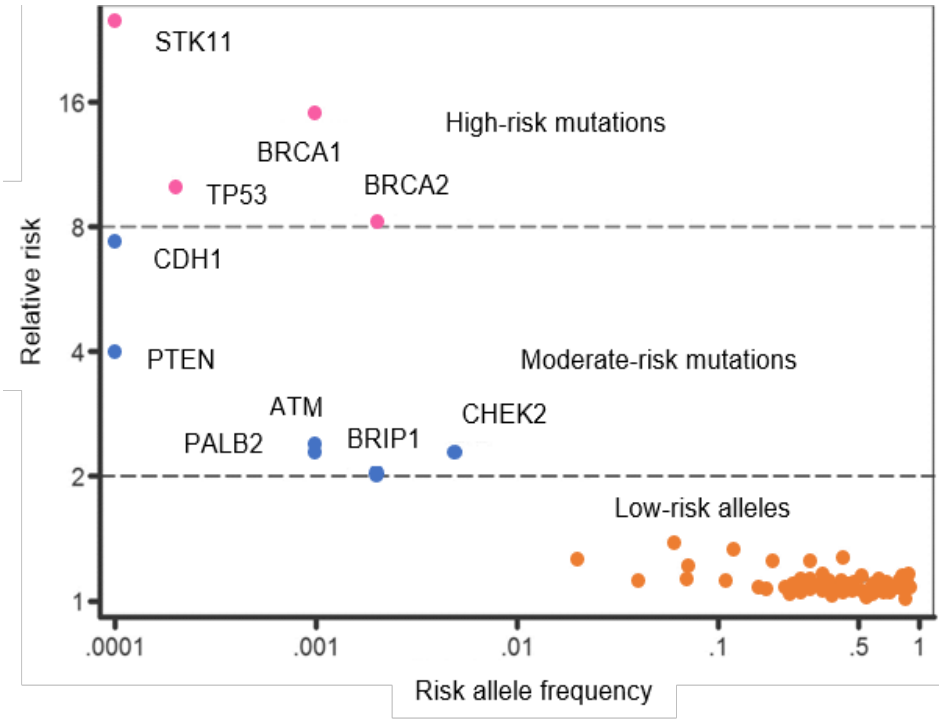


Figure 1.5 - Genetic risk loci identified for BC. The known susceptibility alleles for BC are stratified by relative risk and risk allele frequency into high, moderate and low risk. Adapted from Ghoussaini, Pharoah, and Easton 2013.

1.2.1. High risk mutations

Family-based linkage analysis and positional cloning were used to identify high risk mutations in *BRCA1/2* genes, that are two tumour suppressor genes involved in DNA repair (Miki et al. 1994; Wooster et al. 1995). These are rare but can cause a high risk of BC, about 10-30 fold increase in carriers compared with non-carriers (Antoniou et al. 2003). These deleterious alleles also confer risk to other cancers, like ovarian and prostate cancers (King 2003).

Other high risk penetrant alleles were defined as the mutations in the *Tumour Protein p53* (*TP53*) gene, the *Phosphatase and Tensin homolog* (*PTEN*) gene, the *Serine/threonine kinase 11* (*STK11*) gene and the *Cadherin 1* (*CDH1*) gene (Gonzalez et al. 2009; Li et al. 1997; Hearle et al. 2006; Pharoah, Guilford, and Caldas 2001; Key, Verkasalo, and Banks 2001). The germline mutations in *TP53* predispose to the Li-Fraumeni cancer syndrome, which includes childhood sarcomas and brain tumours and, also, early-onset BC (Gonzalez 2009). Mutations in *PTEN* lead to Cowden disease (disorder with multiple hamartomas), where BC is a dominant feature (Li et al. 1997; Apostolou and Fostira 2013). The germline

mutations in *CDH1* carry a higher susceptibility of lobular BC and women carriers of these mutations have a risk of 40-54% of developing the disease in their lifetime (Kangelaris and Gruber 2007).

1.2.2. Moderate risk mutations

Through association studies and resequencing of candidate genes, moderate penetrant alleles were identified. Certain mutations in genes like *Checkpoint kinase 2* (*CHEK2*), *Partner and localizer of BRCA2* (*PALB2*), *Ataxia-Telangiectasia mutated* (*ATM*), *BRCA1- interacting protein 1* (*BRIP1*), among others, confer an increased BC risk of 2-4 fold (Causeway 2004; Thompson et al. 2005; Rafnar et al. 2011; Ghoussaini, Pharoah, and Easton 2013; Apostolou and Fostira 2013), accounting for 25% of the total familial risk (Easton 1999; Ghoussaini, Pharoah, and Easton 2013).

1.2.3. Low risk variants

Common BC susceptibility loci in the general population were associated with an increased or decreased risk to the disease (Ghoussaini, Pharoah, and Easton 2013; Apostolou and Fostira 2013; Shiovitz and Korde 2015). Studies of low-penetrant variants are focused on polymorphisms that can be relevant to cancer biology. Polymorphisms are characterized by the existence of two or more variants at significant frequencies (>1%) in the population (Pharoah et al. 2004) and can be tandem repeated segments (minisatellite and microsatellite), large (copy number variations) and small deletions/insertions/duplications, as well as single nucleotide polymorphisms (SNPs), the most common in our genome (X. Wang et al. 2005).

The first low risk loci were identified by case-control association studies in candidate genes such as the *CASP8* gene (Cox et al. 2007). With the improvement of genotyping technology, the genome wide association studies (GWAS, explained in detail below) were performed, where thousands of associations were tested simultaneously for BC risk (MacArthur et al. 2017; Ghoussaini, Pharoah, and Easton 2013). Examples of some loci that were identified by this approach and

associated with low risk (< 1.5) are *FGFR2, TOX3, MAP3K1, LSP1*, among others (Easton et al. 2007).

However, despite all the risk alleles already identified, a large proportion of the familial risk is still unaccounted for (Ghoussaini, Pharoah, and Easton 2013).

1.3. GENOME-WIDE ASSOCIATION STUDIES CONTRIBUTION

Population association studies aimed to identify patterns of polymorphisms that differ among individuals with different diseases states and could thus represent the effects of risk or protective alleles (Balding 2006). Genetic case-control association studies became common as an approach to investigate new susceptibility loci that underlie complex diseases, as is BC. The identification of a variant/marker associated with illness status can indicates the possible presence on the genome of a nearby causal risk locus (Rosenberg and VanLiere 2009) (**Figure 1.6**).

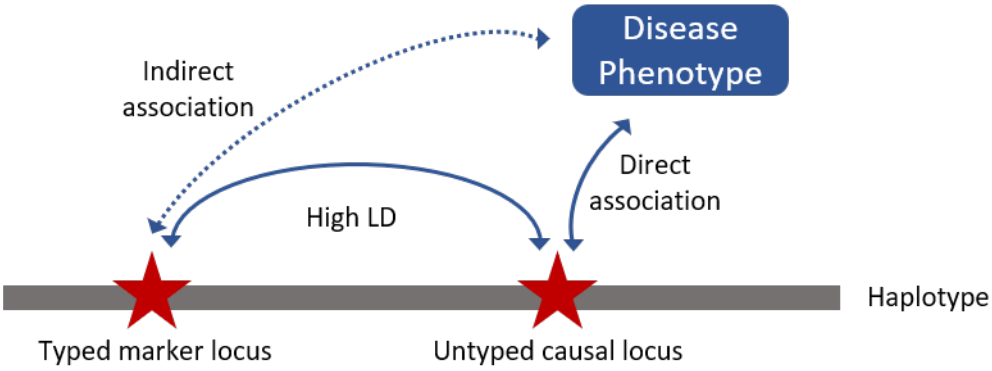


Figure 1.6 - Representation of variants causing the disease phenotype. This haplotype is composed by two markers: the genotyped marker, that is the known variant in the haplotype, that does not have any direct association with phenotype and the ungenotyped marker that is in high Linkage Disequilibrium with the typed causal marker and is the causal variant. The signal of association is captured by the ungenotyped marker (indirect association). Adapted from Balding 2006.

These studies can be classified into several types, as candidate polymorphism (focusing on an individual polymorphism), candidate gene (where

several SNPs within a gene are studied), fine mapping (studies of a candidate region with hundreds of SNPs and several genes) and genome-wide (studies common SNPs throughout the genome) (Balding 2006).

GWAS allow the identification of risk alleles without prior knowledge of position or function. They rely on the testing of the association of thousands to millions of variants across the genome with a trait. The Haplotype Map Project (HapMap) at the beginning of 2005 genotyped about 1.1 million of SNPs across four populations (Thorisson et al. 2005) and with the technological advances were created platforms that allowed hundreds of thousands of SNPs to be analysed simultaneously in association studies (Kruglyak and Nickerson 2001). Currently, it is estimated that about 7 million common SNPs exist in the human genome with a minor allele frequency (MAF) of at least 5% (Kruglyak and Nickerson 2001).

The first GWAS for BC was performed by Easton and colleagues (Easton et al. 2007), and numerous subsequent GWASes have identified hundreds of new risk loci (MacArthur et al. 2017). GWAS are usually performed in two to three phases, differentiated by the number/origin of samples and/or SNPs to be tested. The first phase is characterized by more genotyped and tested SNPs (using genotyping microarrays) in hundreds/thousands of cases and controls samples. The selected most associated SNPs proceed to a second phase (replication) and are tested in a larger number of samples representative of cases and controls. In phase III (validation), the most significative SNPs from phase II are tested in an even larger number of samples, in the order of tens of thousands of individuals that many times involves datasets from different populations/countries (Easton et al. 2007; Consortium et al. 2007).

The way SNPs are selected for a GWAS is based on the concept of linkage disequilibrium (LD), the non-random co-occurrence of alleles at two loci (Pritchard and Przeworski 2001; Slatkin 2008; VanLiere and Rosenberg 2008). As human genome recombination tends to happen at distinct blocks, neighbouring polymorphisms are often strongly correlated with each other (Kruglyak and Nickerson 2001). LD is measured normally based on comparisons of the observed frequencies of haplotypes and the frequencies of the alleles comprising the various haplotypes. Regarding biallelic markers, r^2 is one of the most used measures (Hill and Robertson 1968; VanLiere and Rosenberg 2008), corresponding to the square

of the correlation coefficient for the presence or absence of a particular allele at the first locus and the presence or absence of another allele at the second locus. This statistic is used in power calculations for the ability to detect a disease risk locus (VanLiere and Rosenberg 2008). In the Caucasian population, high LD blocks may vary in length from a few kilobases (kb) to >300 kb (Gabriel et al. 2002; Phillips et al. 2003; Allen-Brady and Camp 2005). High LD regions have redundant information and can thus be minimized in smaller subsets of tag SNPs (Johnson et al. 2001), where these tag SNPs identify all common haplotypes within the high LD region (Allen-Brady and Camp 2005). These are the SNPs commonly selected for a GWAS.

More recently, the 1000 Genomes Project (http://www.internationalgenome.org/), aimed at discovering, genotyping and giving precise haplotypes information for worldwide human Deoxyribonucleic Acid (DNA) polymorphisms (Altshuler et al. 2010). It discovered a large number of variants, which allowed a better coverage of the genome in GWAS, leading to the identification of further associations of low frequency and rare variants to disease risk (Manolio et al. 2009).

Additionally, a consortium of GWAS studies was formed, *Illumina* Collaborative Oncological Gene-environment Study (iCOGS), in which hundreds of thousands of samples have been analysed, in the hope of not only detecting the genetic portion of risk, but also some gene-environment risk factors (Couch et al. 2013).

Overall, GWASes have revealed that disease-associated SNPs occur more frequently in non-coding regions (80% of loci), like promoters, intragenic and intergenic regions (**Figure 1.7**) (Gusev et al. 2014; Corradin and Scacheri 2014).

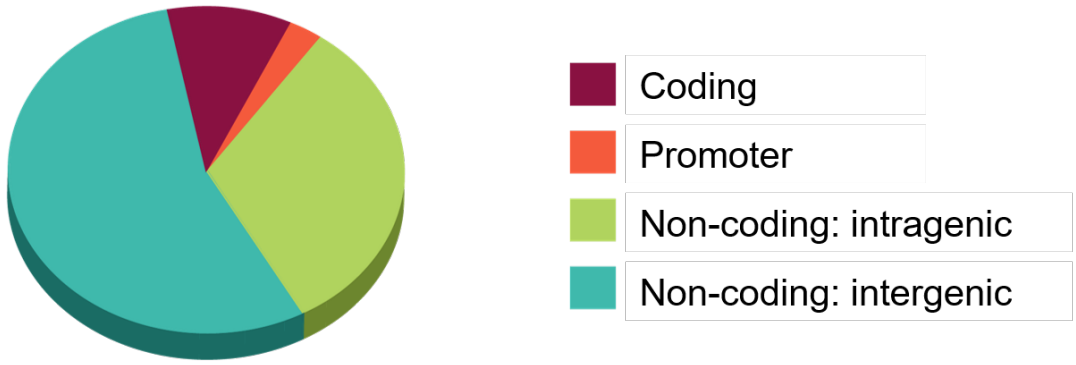


Figure 1.7 - Annotation of BC GWAS variants according to location in the genome. Representation of the BC disease-associated variants that lie in coding-regions (red), in promoter regions (orange), in non-coding regions like intragenic (green) and intergenic (Caribbean green). Adapted from Corradin and Scacheri 2014.

1.4. CIS-REGULATORY VARIATION ON GENE EXPRESSION

The few existing functional studies of the risk loci identified by GWASes in breast cancer, have suggested that the variants included in them are regulating gene expression, i.e. are cis-regulatory (Meyer et al. 2008; A. M. Dunning et al. 2009).

Ribonucleic Acid (RNA) levels are influenced by genetic regulatory elements residing within and outside of the, epigenetic modifications and environmental alterations (Gilad, Rifkin, and Pritchard 2008; V.G. Cheung and Spielman 2009; Pastinen 2010). The quantity of a given RNA allele is regulated both by cis-acting factors, like DNA polymorphisms in the flanking DNA sequence of the gene, and trans-acting factors, that are themselves controlled by other genetic and environmental factors of the cell (Pastinen and Hudson 2004).

Cis-acting variants alter transcript synthesis in an allele-specific manner and are commonly located in regulatory elements such as promoters and enhancers but can also be found hundreds of kb away (Pastinen, Ge, and Hudson 2006). Trans-acting factors regulate both alleles of the gene equally and can be located on the same or in a different chromosome (**Figure 1.8**) (Monks et al. 2004; Vivian G Cheung et al. 2005; Xiao and Scott 2011).

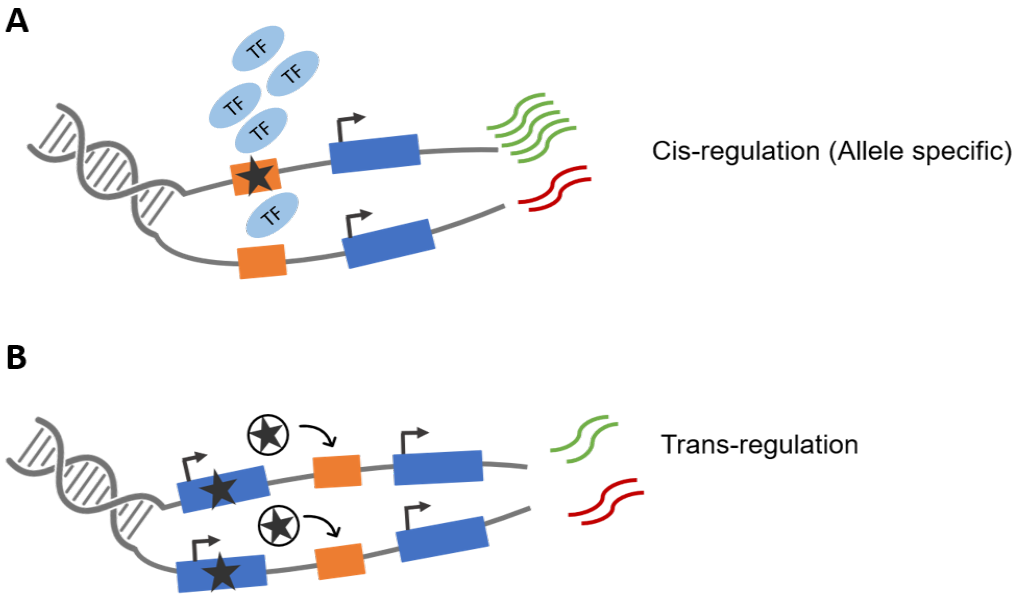


Figure 1.8 - Cis-regulation vs Trans-regulation. Cis-regulation affects genes in an allele-specific manner. When a polymorphism is residing in an allele of the gene (star) may influence the binding of Transcription Factors (TFs), affecting the expression of the gene in an allele-specific way (A). Trans-regulation affects both alleles equitably because the trans-factor (in this case the protein with the mutation) regulates the promoter region of other gene, increasing equitably the expression of both alleles (B).

Although heritable expression differences resulting from trans-acting factors seem to be quantitatively more important, cis-acting variants may be responsible for 25-35% of inter-individual differences in gene expression (Pastinen and Hudson 2004). The identification of cis-acting regulatory variants (rSNPs) contributes to an understanding of variants that alter local gene expression (Ge et al. 2009; Xiao and Scott 2011), and can help with the characterisation of causative variants in regions identified by GWAS (Xiao and Scott 2011), which, as previously mentioned, are in their majority non-coding. To detect the effect of cis-acting variants there are two main methods: expression quantitative trait loci (eQTLs) analysis and Differential Allelic Expression (DAE) analysis.

1.4.1. Expression quantitative trait loci (eQTL)

eQTLs are sequence DNA variants that are associated with the expression level of a given gene. They are identified by measuring the total gene expression in groups of genetically distinct genotyped individuals (**Figure 1.9 B-C**) (Rockman and Kruglyak 2006; Albert and Kruglyak 2015).

eQTLs can be classified according to their location regarding the gene or genes they influence, as local or distant eQTLs (Albert and Kruglyak 2015). Local-eQTLs can alter the gene expression of a gene by two distinct manners. They can act in cis and influence the gene expression in an allele-specific manner (Rockman and Kruglyak 2006; Albert and Kruglyak 2015) or can act in trans, where trans-eQTLs are due to polymorphisms that affect the structure, function or expression of a diffusible factor. Trans-eQTLs do not lead to a different allele expression in heterozygous individuals as the cis-eQTLs, because the diffusible factor is equally accessible to both alleles of a gene. Distant eQTLs are characterised as loci that are positioned further away from the genes they influence and normally only act in trans (Albert and Kruglyak 2015).

eQTL mapping associates genotypes and gene expression levels (Brem et al. 2002; Schadt et al. 2003; Vivian G. Cheung et al. 2003), however for most eQTLs the causal variant is still unknown. However, eQTLs studies contribute to the knowledge of the spatial distribution of regulatory variants in the genome (Veyrieras et al. 2008), to the temporal specificity of the effect of regulatory elements on gene expression, as well as to the magnitude of the expression changes related with cis or trans variation (Stranger et al. 2007; Göring et al. 2007; Pai, Pritchard, and Gilad 2015).

Nonetheless, RNA levels are largely affected by trans-factors and expression quantitative trait loci analysis does not have the capability to remove these effects to isolate cis-factors, because it measures total gene expression (**Figure 1.9 A-C)**. So the direct assessment of cis-regulatory variation needs allele-specific approaches, such as differential allelic expression analysis (DAE) (Pastinen and Hudson 2004; Pastinen 2010).

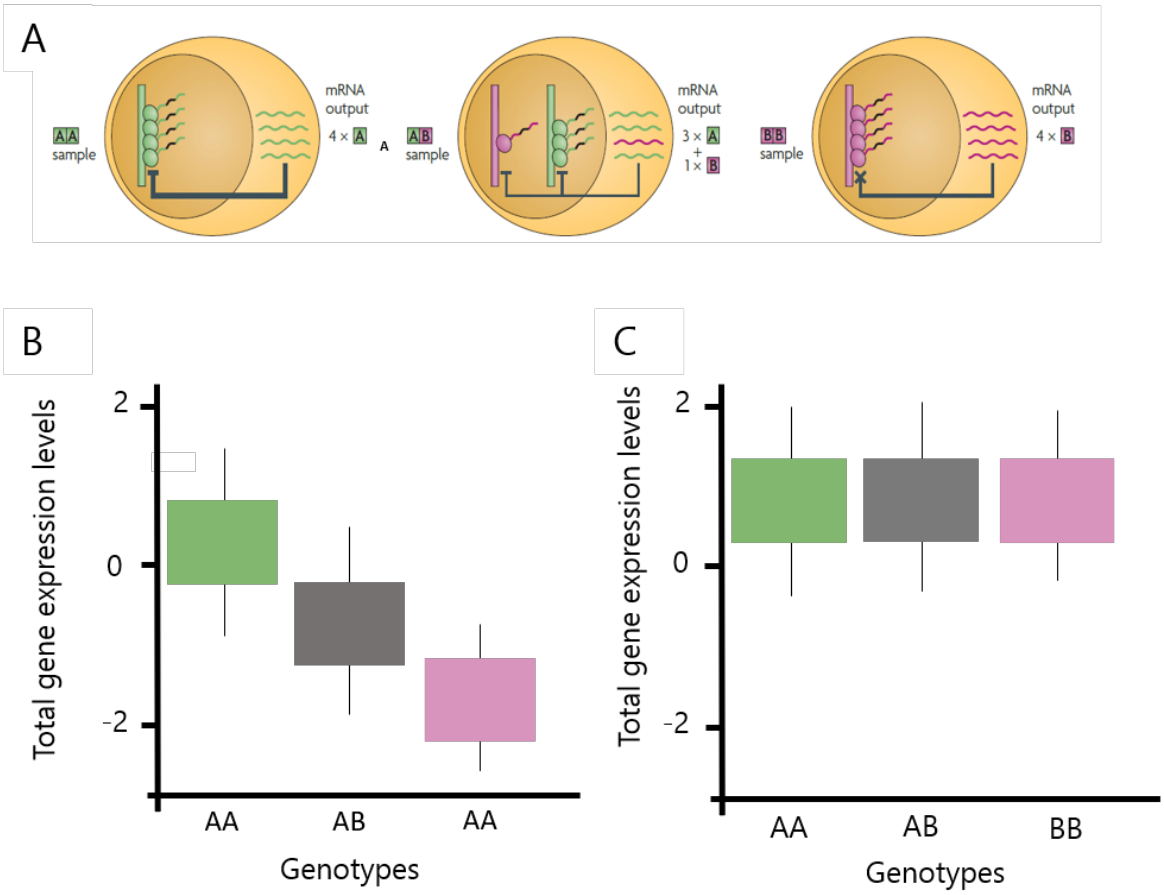


Figure 1.9 - eQTLs representation in the presence and absence of trans-acting factors. In **A** is represented three possible genotypes for a sample (AA, AB and BB) in a population and it is assumed that a true cis-regulatory difference occurs between A and B alleles and the interaction between the trans-acting factor and alleles A and B is absence. Duo to a homeostatic feedback mechanism (bold arrow) the activity of the A allele is increased but the overall expression output in the AA homozygote is powerfully repressed. In BB homozygotes, intrinsically, there is a lower expression of the regulated transcript and therefore negative feedback is down (crossed arrow). Regarding to AB heterozygotes, they have intermediate levels of negative feedback (light arrow). So, if total gene expression is quantified beyond the three genotypes (eQTL), the negative feedback decreases the variance between individuals and the cis-acting variant is not detected (**C**). However, if there is not any trans-acting factor/negative feedback influencing the gene-expression the cis-acting variant effect can be detected (**B**).

1.4.2. Differential Allelic Expression analysis

RNA allelic specific expression measurements enable the direct detection of cis-acting variants' effect (Pastinen and Hudson 2004). Heterozygous individuals are needed to measure allelic expression so that alleles of a variant can be distinguished and quantified separately (**Figure 1.10**). Another advantage on using differential allelic expression analysis is that it eliminates the environmental or trans-acting factors effects, which are altering gene expression or DNA-protein interactions. The trans-acting effects are excluded because when the ratio of the

expression levels of both alleles in the same individual is calculated, the trans factors that affect both alleles cancel each other (Pastinen 2010; Xiao and Scott 2011).

The allele-specific expression of a transcript can be uncovered by *in vitro* and *in vivo* methods that measure the cumulative effects on several cellular processes. The measurement of allelic expression is commonly performed by reverse transcription (RT) to obtain complementary DNA (cDNA) from tissues or cell lines, it requires the existence of a transcribed polymorphism, to which allele-specific probes can be designed (Ge et al. 2009; Milani et al. 2009; Pastinen 2010). Imbalanced allelic expression is detected when the allelic ratio deviates from 50:50, i.e. the gene displays DAE, also called allelic expression imbalance (Pastinen 2010; Xiao and Scott 2011).

Studies revealed that this approach has improved sensitivity to identify cis-rSNPs when compared with eQTL studies and shows an eightfold decrease in sample size necessary to accomplish the same statistical power of as eQTL mapping (Almlöf et al. 2012). Bing Ge et al. with allelic expression mapping in lymphoblastoid cell lines (LCLs) demonstrated that over 30% of all loci have significant DAE, where the cis-rSNPs explained more than 50% of the population variance in allelic expression (Ge et al. 2009).

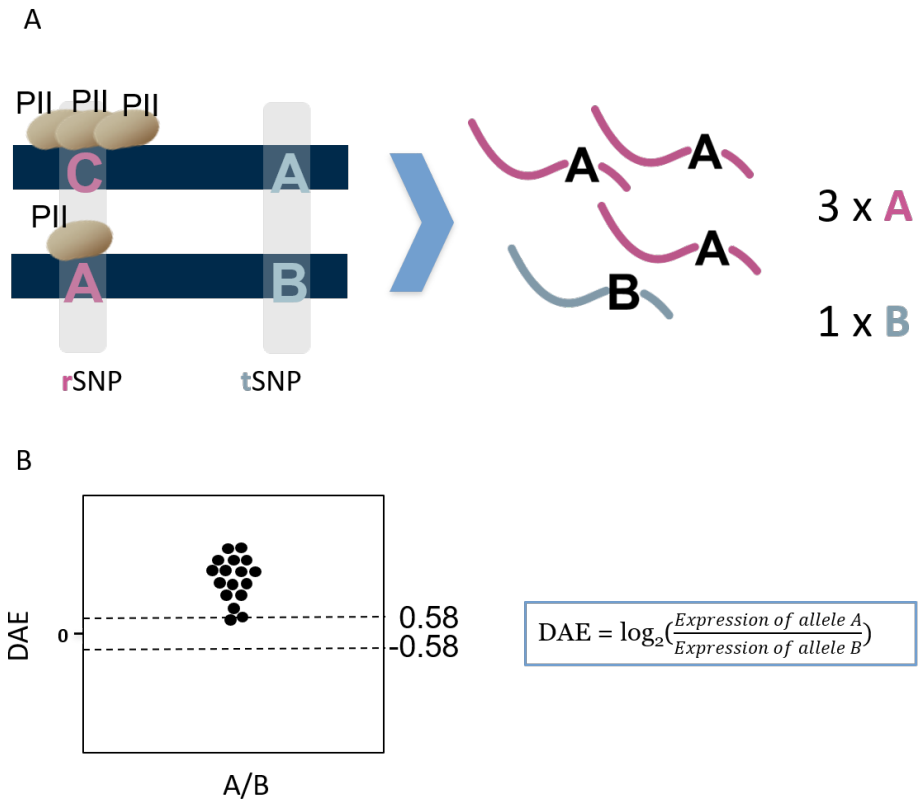


Figure 1.10 - Differential Allelic Expression measurement. In **A** is shown the effect of an rSNP on the transcription rate of a gene, which reflects on different amounts of allelic transcripts, which can be detected via a transcribed SNP (tSNP), in a heterozygous individual. The C allele of the rSNP is causing an increase of the transcription rate of the A allele. This increase can be quantified by calculating DAE ratios as represented in the formula showed in **B.** In the plot the x axis indicates the heterozygote genotype for the tSNP, and the y axis the values of the DAE ratios. Each dot on the plot represents a single heterozygous individual.

After all, the **risk-associated variants** previously functionally analysed were shown to be **cis-regulatory**, so we hypothesise that cis-regulation is a major risk mechanism and to identify further risk variants we should focus our studies in cis-regulatory variants. It is important to notice that **GWAS** have identified a large number of loci associated with BC risk, but have **also identified many others which shown some association with risk** (p-value <0.05 and >5x10^{-8}, depending of GWAS phase), but still **require validation**. However, this validation is expensive and difficult due to number of loci to be validated and the number of samples required to perform these studies (expected effect sizes are small, and hence required numbers for validation are extremely large).

Since DAE is a most powerful approach to identify risk cis-regulated variants we propose to use **DAE ratios as a quantitative trait in association**

studies. This approach will have increased statistical power when compared with the current GWAS set-up, which is based on genotypes frequencies (discrete variables), to find inferences between patients and controls.

CHAPTER II

Aim

AIM

The general aim of this study is **to identify new genes associated with BC risk**. To accomplish this goal, we believe we should focus our studies on cis-acting regulatory variants.

As such, we propose the following specific tasks:

1 - To select candidate loci with some evidence of possible association with BC risk based on: whether they harbour genes with evidence of being cis-regulated, i.e. genes showing DAE, in normal breast tissue from healthy individuals; and whether they show further association with clinical variables, such as tumour grade, BC molecular subtypes, ER status, HER2 status and PR status, etc.

2 - To validate the association of the top candidates, by performing a case-control association study using DAE as the quantitative trait.

CHAPTER III

Materials and Methods

3. MATERIALS AND METHODS

3.1. PROGRAMMING IN R

R is a programming language and an environment that provides a variety of statistical techniques and graphical interfaces, as well as programmatic commands. This programming language is highly vast, and its environment has the capability to interact with other repositories where there are freely available R packages (bunch together data, code, documentation and tests to share in an easy way with others). In this work, all statistical analyses, graphics and tables editing were performed using *R studio* (R Core Team 2017). Most of the time dedicated to this dissertation was to learn how to programme in R, namely the basics of the language and also statistical and package implementation. To make it easier to program in R (version 3.4.4), an integrated development environment (IDE) for R: *R studio* (R *studio* version 1.1.447) was used, which includes a code editor, debugging & visualization tools, and was ran under platform x86_64-pc-linux_gnu. All most of the plots were generated using *ggplot2* (version 2.21.) package (Wickham 2016) and venn diagrams were performed using *vennDiagram* (version 1.6.20) package (Chen and Boutros 2011).

3.2. RETRIEVAL OF CANDIDATE GWAS VARIANTS

3.2.1. Data characterisation

Breast cancer GWAS and corresponding risk variants were retrieved from NHGRI-EBI Catalog of published genome-wide association studies (MacArthur et al. 2017), accessed on 23/04/2018, available at www.ebi.ac.uk/gwas, using the traits "Breast Cancer", "Breast Cancer (early onset)", "Breast cancer (male)" and "cancer", and a p-value threshold $\leq 5\times10^{-5}$. Additionally, these studies were manually curated in order to retrieve variants that were found associated with BC risk in the first phases, but not in the last ones, of GWASes with more than one phase. These SNPs were considered to be good candidates for being further studied, although they were not considered associated with BC risk. From now on we will refer to these SNPS as the "candidate risk SNPs/variants".

3.2.2. Linkage Disequilibrium analysis

In order to identify proxy SNPs (SNPs in high LD) with the candidate risk SNPs, *rsnps* R package was used, a programmatic interface to several online 'SNP datasets', like: *"OpenSNP"*, *"dbSNP from National Center for Biotechnology Information"* database - *NCBI-dbSNP*, and *Broad Institute SNP Annotation and Proxy Search* (Chamberlain, Ushey, and Zhu 2016). To get the proxy SNPs we applied an $r^2 \geq 0.8$ limit with the index SNP (in a window of 500 kb each side of the index SNP) and we chose the reference population most similar to the one where the GWAS was performed. Namely, for GWASes performed in the European population, we chose the CEU (Utah residents with Northern and Western European ancestry from the Centre d'Etude du Polymorphism Humain (CEPH) collection) from the 1000 Genomes Project pilot 1 (*onekgpilot*). Regarding the GWAS performed in the Asian population, we used the JPT (Japanese in Tokyo, Japan), CHB (Han Chinese in Beijing, China) and CHD (Chinese in Metropolitan Denver, Colorado) from HapMap 3 release 2 (*hapmap3r2*). The *onekgpilot for YRI (Yoruba in Ibadan, Nigeria)* was used as reference population to identify proxy SNPs within the African population.

3.2.3. SNPs annotation

Using the R package *biomaRt* (version 2.34.2; version 92; (Durinck et al. 2005; Durinck et al. 2009)), the *ensembl* genes ids where the candidate risk variants plus their proxies mapped, were retrieved, as well as for the GWAS established risk variants SNPs, via *getBM* function. *biomaRt* can output different information according to the arguments that are provided in the parameter attributes of getBM function. The attributes chosen were: "ensembl_gene_stable_id", "refsnp_id", "consequence_type_tv" (that give us information about localization of the variant, if it is upstream of the gene, in an intron, downstream of the gene, etc.).

The *Ensembl* browser enables access to genomic annotation from different species, including vertebrate genomes that supports comprehensive annotation of different attributes such as sequence variation, among others (Zerbino et al. 2018). The *biomaRt* package has also the ability to export custom sets of data from the

Ensembl browser, enabling the users to uniformize their data (Durinck et al. 2005; Durinck et al. 2009).

3.2.4. iCOGS association study

The Collaborative Oncological Gene-environment Study (COGS), is a European Union-funded project, integrating four consortia to drive an exhaustive investigation of the genetics of hormone-related cancers: Breast Cancer Association Consortium (BCAC), Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome (PRACTICAL), Ovarian Cancer Association Consortium (OCAC) and The Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). Two of the main aims of COGS are to identify the variants associated with susceptibility to these cancers and the risks associated with the variants identified. The samples from the four consortia (over 150000 samples) were genotyped using the Illumina Custom Infinium array COGS (iCOGS). This array included over 200000 variants, including replicates to a great number of suggestive associations from GWAS, as well as others to investigate a large variety of phenotypes (Couch et al. 2013).

Data from BC iCOGS (BCAC) association study was downloaded from the website: http://bcac.ccge.medschl.cam.ac.uk/bcacdata/icogs-complete-summary-results/, where there was information about Europeans, Asians, and Africans ("ICOGS Complete Summary Results" 2018; Michailidou et al. 2013; Guo et al. 2015; Michailidou et al. 2015).

The data of interest were from the European people where SNPs with p-value ≤ 0.05 and $\geq 5x10^{-5}$ were extracted.

3.3. DIFFERENTIAL ALLELIC EXPRESSION ANALYSIS

3.3.1. Dataset

Genome-wide Differential Allelic Expression (DAE) analysis was done previously in Professor Ana-Teresa Maia's group, using microarrays technology on 64 normal breast tissue samples. Here, DNA and total RNA were run on Illumina Exon510S-Duo arrays (genotyping microarrays) and DAE was measured in a

filtered dataset of SNPs, in a variable and independent number of individuals who were heterozygous for a transcribed SNP (tSNP) with alleles A and B, allowing allele-specific measurements. The DAE ratio was defined as the logarithm of base two between the allele A transcript expression level and the allele B transcript expression level (heterozygote ratio), normalized by the heterozygote ratio in genomic DNA (gDNA), on tSNP heterozygotes:

$$DAE = log2(\frac{cDNA\ heterozygotes\ ratio}{gDNA\ heterozygote\ ratio})$$

DAE was defined as DAE ratios greater than 0.58 or less than -0.58 and SNPs were considered to be differential allelic expressed when at least 10% of the heterozygotes and four samples displayed DAE

This study resulted in a whole genome map of the cis-regulatory genes in normal breast tissue (Xavier et al. 2016).

3.4. GENE EXPRESSION ANALYSES

3.4.1. Dataset characterisation

Data from Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) were used to perform gene expression analyses (Curtis et al. 2012). These data were obtained by our group upon request to the European Genome-phenome Archive (EGA) (Lappalainen et al. 2015). It consists on an assembled collection of over 2000 clinically annotated fresh frozen breast cancer tissue, that have passed initial selection norms from tumour banks in United Kingdom (UK) and Canada (Pereira et al. 2016; Curtis et al. 2012). All DNA and RNA of patient specimens were isolated and hybridized using the Affymetrix SNP 6.0 and Illumina HT-12 v3 platforms respectively (Curtis et al. 2012). The tumours were primary invasive breast carcinomas for which there are demographics, clinical, and pathological information, and these data were obtained using the website cBioPortal (Gao et al. 2013; Cerami et al. 2012). Regarding the patients, all ER+ and/or lymph node (LN) negative did not receive any kind of chemotherapy treatment, while ER- and LN negative patients receive.

These samples gave rise to two different datasets. The first dataset, named discovery dataset, consisted on DNA and RNA with quality profiles available for 997 female tumours. The second cohort, composed of 995 female tumour samples, were denominated validation dataset. A total of 1992 tumour cases were used, with the tumour subtypes distribution described in **Table 3.1**. The adjacent normal tissue (named normal-matched, NM) with high quality RNA is composed by 144 samples. The matrices of gene expression data were already normalised for the log2 intensities (Curtis et al. 2012).

Table 3. 1 – METABRIC tumour subtypes summary

Clinical variable		Number of cases (%)
ER status	Pos	**1498** (77%)
	Neg	**439** (23%)
PR status	Pos	**1040** (53%)
	Neg	**940** (47%)
HER2 amplification	Pos	**247** (12%)
	Neg	**1733** (87%)
Grade	I	**169** (9%)
	II	**771** (41%)
	III	**952** (50%)

Pos- positive; Neg- Negative

3.4.2. Illumina probes quality filter

Each matrix of gene expression was composed of 48,803 *Illumina* probes (Curtis et al. 2012). Before starting the differential expression analysis, we applied a probe quality filter using the R package *illuminaHumanv3.db* from the Bioconductor repository. We used the annotation file *illuminaHumanv3PROBEQUALITY* and kept the probes classified as "Perfect", that are the ones uniquely matching the target transcript and the "Good" which are the ones that provide considerably signal. As previous suggested, we excluded the "no match" probes (the ones that do not match any genomic region or transcript) and the "Bad" (the ones matching repeated sequences, intergenic or intronic

regions, or that did not have the capability of provide a specific signal for any transcript) (M. Dunning, Lynch, and Eldridge 2015; Barbosa-Morais et al. 2010).

3.4.3. Differential Expression analyses

Technologies for gene expression analysis, like microarrays or RNA Sequencing (RNA-Seq), are central in molecular biology research because of their contribute to the knowledge of the transcriptional activity in diverse populations of cells and different tissues (Smyth 2004; Ritchie et al. 2015)**.** These techniques are crucial to identify and compare gene expression changes regarding to a treatment condition or some phenomena of interest. The application of these technologies, to measure gene expression, are recurrent in cancer to identify pathogenic features (Ritchie et al. 2015). One possible application of the microarray technology is to identify differentially expressed genes (Sartor et al. 2006), however such tasks are challenging because the measured expression levels usually do not follow a normal distribution and do not have a dependent and identical distribution among genes (Smyth 2004).

To assess the statistical significance of the results it is essential to apply statistical tests (Sartor et al. 2006). Gene expression studies are complex (Ritchie et al. 2015) and, considering data for each gene transcript individually is statistically ineffective (Jain et al. 2003; Sartor et al. 2006). When these studies involve a minor number of biological replicates it can even lead to statistical problems (Ritchie et al. 2015) such as low reliability in the variance estimates (Jain et al. 2003; Sartor et al. 2006). In order to solve these problems, there was the need to use specialized statistical techniques to obtain the best outcome of each data set (Ritchie et al. 2015), like the application of the hierarchical Bayesian model, which is based on Robbins-type empirical Bayes theory and on the hierarchical model of Lönnstedt and Speed (2002) (Lonnstedt and Speed 2002; Efron 2003; Smyth 2004; Sartor et al. 2006).

Empirical Bayes is a statistical technique that allows its users substantial gains in performance, and was therefore the approach applied for the microarrays differential expression analysis in this thesis. This technique assumes a hierarchical model for the genewise variances where the *prior (*probability of the

hypothesis previously to looking to the data) distribution is estimated from the marginal distribution of the observed data and not on previous knowledge (Efron et al. 2001; Smyth 2004; Phipson et al. 2016).

Gordon Smyth (2004) developed the model of Lönnstedt and Speed (2002) into a practical approach for overall microarray data with arbitrary numbers of treatments and RNA samples, and turned the approach adaptable to both single channel and two-color microarrays. The approach of Smyth (2004) fits an independent linear model for each gene separately, instead of a single linear model for an entire microarray experiment (Kerr, Martin, and Churchill 2000). Therefore, the variances are assumed to be different between genes and the estimators of variance (standard errors) are moderated across genes allowing for different levels of variability between genes and between samples (Smyth 2004; Ritchie et al. 2015).

After the linear model was established, a moderated t-statistic was applied. The t statistic has advantages over other statistics (e.g. the posterior odds (B-statistics)) because it depends only on residual variances and on degrees of freedom, whereas the B-statistics depends on several hyperparameters. Furthermore, the moderated t does not require any knowledge of the proportion of the genes that are differentially expressed, and it does not imply any expectations around the magnitude of differential expression, and making statistical conclusions more reliable when the number of samples is low (Ritchie et al. 2015).

3.4.3.1. Statistical association analyses between gene expression and clinical variables

The differential expression analyses took into account diverse clinical variables. We tested: tumours vs normal-matched (NM), oestrogen receptor (ER) positive vs normal-matched, ER negative vs normal-matched, ER positive vs ER negative, progesterone (PR) positive vs normal-matched, PR negative vs normal-matched, PR positive vs PR negative, HER2 positive vs normal-matched, HER2 negative vs normal-matched, HER2 positive vs HER2 negative, as well as differential expression between tumours with distinct grade classification (1-3). All the analyses were performed with the *limma* package, except when considering the grade classification, where the Kruskal-Wallis rank sum test was applied.

3.4.3.2. limma package

limma uses linear regression models and empirical Bayes methods to assess differentially expressed genes. The linear regression models were implemented to decompose the independent effect of each phenotype on gene expression (*lmFit* function) and genes were ranked by the moderated-t statistic (*eBayes* function) for differential expression. For each gene this statistic was calculated by the ratio between the base 2 logarithm of the fold-change ($\log_2$(Fold Change)) of expression and the moderated standard error of the log2 of expression per samples. Genes were considered differentially expressed when the adjusted p-value for multiple testing was ≤ 0.01 (1% False Discovery Rate - FDR). A threshold for the effect size, which gives us the biological magnitude, was also applied at FC of 1.5.

3.4.3.3. Kruskal-Wallis rank sum test

The *Kruskal-Wallis* rank sum test is a non-parametric test applied when we want to compare distributions between more than two groups of samples. We applied this to test if the samples from patients with different tumour grade classifications had the same distributions (Kruskal and Wallis 1952). The function *Kruskal.test* provided by the *stats* package implemented in R was used. We also corrected the p-value for multiple testing using the FDR (False Discovery Rate) method, and genes with adjusted p-value ≤ 0.01 (1% FDR) were considered to be differential expressed between tumours with different grade.

3.4.4. Multiple testing correction

Statistical tests sometimes involve testing of the same hypothesis a lot of subsequent/simultaneous times. This increases the chances of obtaining significant results just by chance, so more false-positive cases can arise. It is important to correct p-values for multiple testing to adjust the statistic according to the number of tests performed (Noble 2009).

There are different types of multiple tests corrections, where the Bonferroni correction is the most widely used, although it is very conservative, and therefore, the FDR correction, which controls for the type I errors (false-positive cases), is the most commonly used in microarrays analysis (Noble 2009).

We used the FDR method developed by Benjamini and Hochberg, which multiplies the nominal p-values by the number of tests performed and next divide it by its position taking into account the nominal p-values for the smallest to largest (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001).

Multiple testing correction can be implemented in R using the function *p.adjust* provided by the stats R package (R Core Team 2017) or can be incorporated in some functions from packages like *limma* (Ritchie et al. 2015).

3.5. PRIORITIZATION OF GENES AND VARIANTS FOR THE CASE-CONTROL ASSOCIATION STUDY

Genes that were both associated with clinical features (in the gene expression association studies), showed DAE in normal breast tissue and were weakly/mildly associated with breast cancer risk in the first phases of GWAS, were further analysed in order to select candidate genes for the case-control association study using DAE ratios. Information for expression in normal breast tissue and in blood were retrieved from the Genotype-Tissue Expression project (GTEx) (https://gtexportal.org/home/), as well as information for the presence of eQTLs for the gene ("GTEx Portal" 2018). Information of minor allele frequency (MAF) in the European population for each SNP gene was obtained using the *NCBI-dbSNP* (https://www.ncbi.nlm.nih.gov/snp) ("Home - SNP - NCBI" 2018; Sherry 2001). Information for the DAE values in normal breast tissue of genes at a given tSNP was obtained using data from our group. Two genes were selected for analysis in thecase-control association studies using DAE ratios: *OCIA Domain Containing 1* gene *(OCIAD1)* and *GrainyHead Like Transcription Factor 2* gene *(GRHL2)*.

3.6. GENETIC ASSOCIATION STUDIES

3.6.1. Samples characterisation

DNA and RNA were extracted from 170 samples of white cell-reduction filters from anonymous blood donors (used as controls in the association study in blood) and a total of 56 blood samples from breast cancer patients (cases in the association study in blood) at Addenbrooke's Hospital. DNA and RNA were previously extracted from all samples through SDS/proteinase K/phenol and TRizol methods, respectively, at the University of Cambridge. We used 56 cDNA samples from blood cancer donors (with unknown genotypes), previously prepared in the group using the SuperScript III First-Strand Synthesis System for Reverse Transcriptase - Polymerase Chain Reaction (RT-PCR) (Invitrogen), and 13 cDNA blood donors samples, previously synthetized in the group by the same method referred above. Both cDNAs were prepared aiming at a final concentration equivalent to at 10ng initial RNA per μL of reaction.

Normal breast tissue (controls in the association study in breast tissue) was extracted previously at Addenbrooke's Hospital, from 64 women submitted to reduction mastectomy (without reasons related with cancer). RNA extraction were performed at University of Cambridge. A total of 45 cDNA samples at 10 ng/uL were used.

Normal-matched tissue samples (cases in the association study in breast tissue, NM), samples of normal adjacent tissue of patients with cancer, were received from collaborators from the METABRIC project. DNA and RNA samples were retrieved at BC Cancer Research Centre in Vancouver and University of Cambridge. A total of 49 samples were used in this project.

Lymphoblastoid cell lines derived from unrelated CEPH individuals were acquired from the Coriell Cell Repository. DNA was previously extracted in at University of Cambridge by a conventional SDS/proteinase K/phenol method and total RNA was collected using Qiazol (Invitrogen, Carlsband, CA, USA) following manufacturer's instructions. RNA was later treated with *DNase*I and repurified with acidic phenol-chloroform and ethanol precipitation. A total of 20 DNA samples were studied.

All samples were collected in compliance with ethics guidelines and regulations. Blood samples from both cancer patients and healthy donors, as well as normal-breast samples from healthy women were obtained with approval from Addenbrooke's Hospital Local Research Ethics Committee (REC reference 04/Q0108/221, respectively). NM samples (i.e. adjacent free tumour tissue of women with breast cancer) were collected with the approval from the ethics committees in Cambridge and Vancouver, which are the two sites responsible for the molecular analysis of the samples (REC ref 07/H0308/161; REC ref 12/EE/0484; REC ref 07/Q0106/63).

3.6.2. DNA quantification

DNA from 20 CEPH samples was quantified using NanoDrop and diluted to 40ng/μL, in a total volume of 100μL. The remaining DNA samples used in this project were already quantified.

3.6.3. cDNA preparation using RT-PCR

cDNA was synthesized from purified poly(A) and total RNA for 49 normal-matched samples and for 7 normal breast samples, using the SuperScript First-Strand Synthesis System for RT-Polymerase Chain Reaction (RT-PCR) (Invitrogen). The first-strand cDNA synthesis reaction is catalysed by SuperScript II Reverse Transcriptase, which is an enzyme engineered to reduce the activity of RNase H (which degrades messenger RNA (mRNA) during the reaction) producing a greater full-length cDNA. This enzyme exhibits a higher thermal stability and can be used at temperatures above of 50°C. The reactions were prepared for a final volume of 10 μL (5 μL per step). The first step included the addition of RNA (1ng-5μg), 0.5 mM dNTP mix, 2.5 ng/μL of Random hexamers, 0.025 μg/μL oligo(dt) and DEPC-treated water. The reactions were incubated at 65°C for 5 minutes (min) using the BioRad C100 Touch Thermal Cycler and then placed on ice for at least 1 min. The second step of RT was done using RT buffer at 2X, MgCl2 at 10mM, DTT at 0.02, RNaseOUT at 4 (U/μL) and 0.5 μL of Superscript II RT. The reactions were incubated at room temperature for 10 min, followed by running at BioRad C100 Touch Thermal Cycler for 50 min at 42°C and 15 min at 70°C. Than

the reactions were chilled on ice and added 0.5 uL of RNase H and incubated again for 20 min at 37ºC. At the end, each reaction was diluted to a final volume of 100 µL.

3.6.4. Real-Time quantitative Polymerase Chain Reaction (RT-qPCR)

The RT-qPCR is considered a gold standard for quantitative data analysis in molecular medicine, microbiology, biotechnology and diagnostics and became one of the methods of election for the messenger RNA (mRNA) quantification (Nolan, Hands, and Bustin 2006).

PCR is a genomic cloning technique that allows a selective and automated amplification of DNA (Innis et al. 1988) and PCR and reverse transcriptase (responsible for the conversion of mRNA to cDNA) permitted the study of little quantities of DNA/cDNA (Heid et al. 1996; Murphy et al. 1990).

Each PCR assay requires:

- **DNA template**, the DNA sample to be amplified;
- **Primers**, these are present in the reaction and specify the DNA to be amplified. Primers are short fragments of DNA with a concrete sequence complementary to the mark DNA that is to be detected, amplified and serve like an extension point for the polymerase act;
- **Nucleotides**, more specifically deoxynucleoside triphosphates (dNTPs), that include the four bases that are in DNA and function like building blocks for the DNA polymerase to generate the PCR product;
- **DNA polymerase**, a key enzyme to form the PCR product, because the enzyme is the DNA polymerase III thermostable, where the temperature does not deregulate its action and has the ability to links individual nucleotides together (Garibyan and Avashia 2013; Nolan, Hands, and Bustin 2006). One of the most used enzymes is Thermus aquaticus (*Taq*) DNA polymerase that can be employed in a PCR and has 5'-3' exonuclease activity (Holland et al. 1991);
- **Magnesium chloride (MgCl2)**, a cofactor for DNA polymerase;
- **Buffer**, used to ensure the physiological conditions to the DNA polymerase action, as pH and ionic concentrations (Nolan, Hands, and Bustin 2006);

The PCR reaction comprises three steps: denaturation, annealing and extension that are controlled by temperature changes and repeated cyclically (**Figure 3.1**), between 35 to 40 cycles. Denaturation process is characterised by the separation of the double strand of DNA in two single strands, using the maximum temperature that polymerase can support, 95ºC. Then the primers have the ability to hybridize with the single strands of DNA, what is called of annealing. Lastly, the DNA polymerase at a reaction temperature of about 60 degrees makes the extension of DNA products. With the repetition of these steps the number of DNA products doubles at each cycle (Garibyan and Avashia 2013; Nolan, Hands, and Bustin 2006).

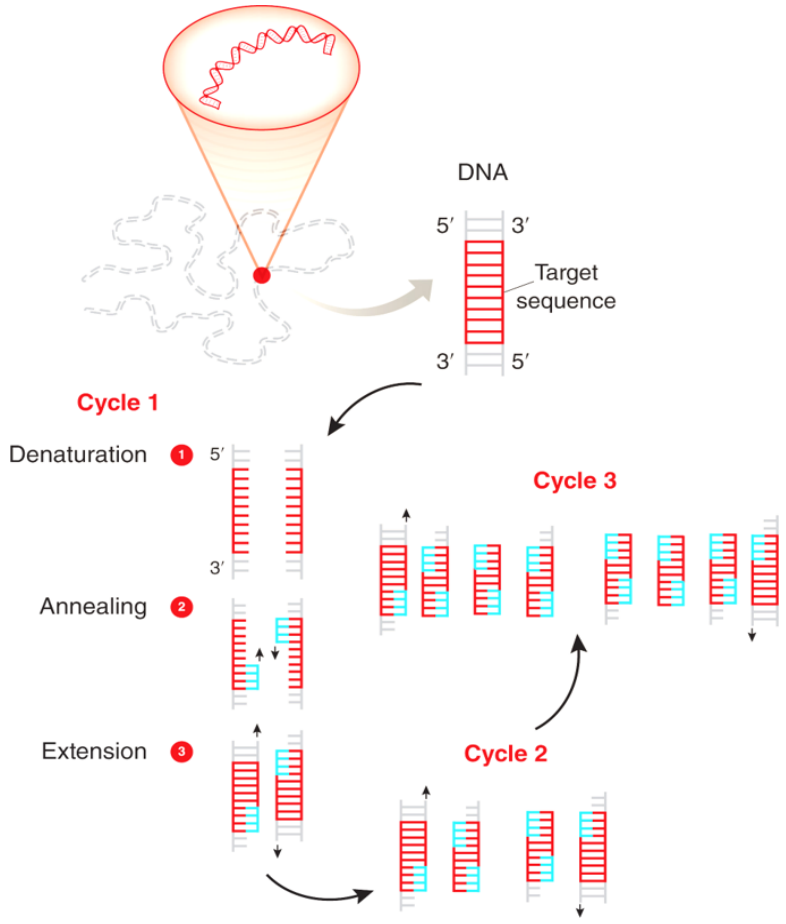


Figure 3.1 - Polymerase Chain Reaction. PCR is composed by 3 principal steps: Denaturation, Annealing and Extension. As represented in the figure, to each cycle the DNA double, i.e. if we have 2 DNA strands in the end of the cycle we stayed with 4 DNA strands. Adapted from Garibyan and Avashia 2013.

Real-Time quantitative PCR (RT-qPCR) occurs at real time, i.e. uses fluorophores to combine the amplification and detection steps of the PCR while it is being synthesized (Nolan, Hands, and Bustin 2006; VanGuilder, Vrana, and Freeman 2008; Garibyan and Avashia 2013). The individual reactions are characterised by the PCR cycle at which the fluorescence cross a defined threshold, a parameter known as the threshold cycle (C_t) (Nolan, Hands, and Bustin 2006; Schmittgen and Livak 2008).

Regarding to methods used to quantify and detect RT-qPCR products, they can be by fluorescent dyes (intercalate the DNA double-stranded nonspecifically, like SYBR Green technology) or by sequence specific DNA probes (fluorescent labeled reports, as Taqman™ technology)(Innis et al. 1988; Nolan, Hands, and Bustin 2006; Garibyan and Avashia 2013). In this study we used allele-specific Taqman™ probes that are considered more sensitive and specific, because they have specific sequence oligonucleotides. These probes in the 5' end contain a fluorophore reporter and in the 3´end contain a quencher (**Figure 3.2**). At the end of each PCR cycle, when the fluorophore is excited by a laser, if the quencher is on its side, it absorbs all fluorescence emitted and the level of solution fluorescence is low; if the DNA polymerase approaches the probe, which is downstream of the primer, cleaves it allowing the fluorophore to separate from the quencher and, accordingly, fluorescence levels increase (Holland et al. 1991; Heid et al. 1996; VanGuilder, Vrana, and Freeman 2008).

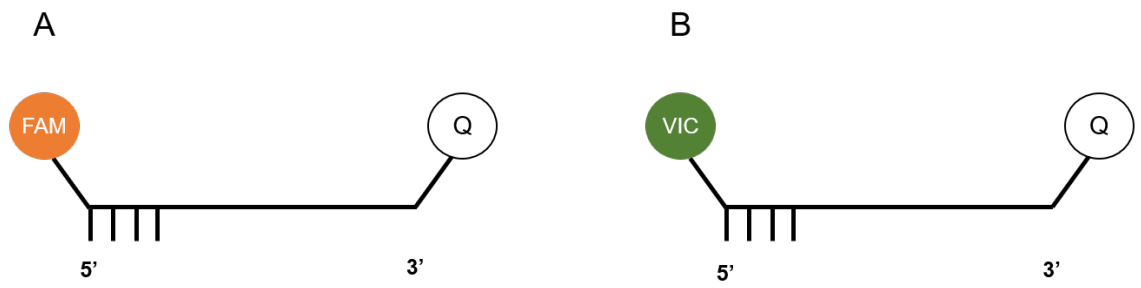


Figure 3.2 - Taqman™ probes design. This figure represents probes labelled with respectively fluorophores. In **A** it is represented probe labelled with FAM fluorophore that generates signal for one specific allele and Q corresponds to the Quencher. In **B** it is described the probe labelled with VIC fluorophore producing signal for the other allele.

The quantification of amplified product can be absolute, when is determined the copy number of the products of interest, usually by the ratio of PCR signal to a standard curve (it is done by performing a serial dilution of a known amount of pure RNA, or can be relative (Heid et al. 1996; Schmittgen and Livak 2008; VanGuilder, Vrana, and Freeman 2008). The relative quantification reports the alteration in expression of the target gene relative to another reference gene, group such as an untreated control or a housekeeping gene (Livak and Schmittgen 2001; Schmittgen and Livak 2008). In our study an absolute quantification was used.

It is important to notice that in the RT-qPCRs performed it was included a standard curve on every plate for the precise quantification in each experiment.

We applied the RT-qPCR to genotype DNA samples and to quantify allelic gene expression in cDNA from heterozygous individuals. Each qPCR reaction contained a primer pair targeting the region surrounding the marker SNP, two probes that differ by a single nucleotide and are complementary to each allele of the gene. The probes have different fluorochromes (VIC and FAM), generating two distinguished signals during the RT-qPCR. The two assays used for each gene (**Annex A**) were for rs9997920 from *OCIAD1* and for rs6989650 from *GRHL2*. For rs9997920 the minor allele in the CEU population (T) is labelled with FAM that corresponds to allele 1 in RT-qPCR graphics and the more frequent allele (C) is labelled with VIC fluorophores that corresponds to allele 2 in RT-qPCR graphics. This tSNP is located in an intronic region. Regarding to rs6989650 from *GRHL2*, the minor allele was labelled with FAM that corresponds to allele 1 (T) in RT-qPCR graphics and the other allele (C), more frequent in European population, is labelled with VIC that corresponds to allele 2 in RT-qPCR graphics. This tSNP is located in a 3' untranslated region (3' UTR) (**Table 3.2**).

Table 3.2 - tSNPs location and corresponding alleles labelling in the Taqman™ probes

Gene	tSNP	FAM allele (1)	VIC allele (2)	Location
OCIAD1	rs9997920	T (minor allele)	C (major allele)	Intron
GRHL2	rs6989650	T (minor allele)	C (major allele)	3' UTR

To perform the RT-qPCR, 384-wells plates (Hard- Shell PCR Plates 384-Well CLR/WHT, Bio-Rad) were used and were ran in a BioRad CFX384 Real-Time System C100 Touch Thermal Cycler. The reactions mixtures were incubated for 3 min until 95ºC for the activation of the enzyme and consequently submitted to 40 cycles with denaturation step for 3 seconds at 95ºC and annealing/extension for 30 seconds at 60ºC. The allelic discrimination charts were exported by Bio-RAD CFX Manager Software, at the end of the RT-qPCR. All experiments contained between 1-3 No Template Controls (NTCs), where the amplification should not occur.

3.6.5. Genotyping

cDNA from normal-matched samples (49), normal breast (NB) samples (45), blood breast cancer samples (56) and DNA from CEPH samples (20) was genotyped using 5' exonuclease Taqman™ technology (Applied Biosystems). For the remaining samples studied by RT-qPCR, the genotypes for the SNPs of interest were already known and there was no need to genotype. Between 12,5 ng and 50 ng of DNA/cDNA was used in a 5µl RT-qPCR reaction constituted by Kapa Probe Fast universal qPCR Kit(2x) (Applied Biosystems), assay with two primers and FAM and VIC labelled probes as described above (Taqman™ SNP Genotyping Assays (40X), Applied Biosystems) and H2O (DNase/RNase free, gibco by Life technologies). NTCs were included in experiment (1-3).

Genotyping was done in 384-wells plates (Hard-Shell PCR Plates 384- Well CLR/WHT, Bio-Rad) using BioRad CFX384 Real-Time System C100 Touch Thermal Cycler. The allelic discrimination charts were exported by Bio-RAD CFX Manager Software, at the end of RT-qPCR.

3.6.6. Quantification of differential allelic gene expression

Allele specific levels of gene expression, or DAE ratios were quantified in heterozygous samples using Taqman™ technology (Applied Biosystems) and RT-qPCR, as described in the genotyping section. The assay for rs9997920 (*OCIAD1*)

was ran both on blood tissue samples (13 blood donors samples and 26 blood BC samples) and on breast tissue samples (22 normal breast samples from healthy controls and 24 normal-matched samples). The assay for rs6989650 (*GRHL2*) was ran only on breast tissue samples (15 normal and 12 normal-matched samples). Each experiment was run twice independently. A calibration curve was performed using a serial dilution of a heterozygous CEPH DNA sample for each SNP of interest (serial diluted samples). This curve served as a reference for the 50:50 allelic ratio as represented in the **Table 3.3**. Another curve, a standard positive control curve - series of standard curve, was performed using different homozygous samples for my SNPs (**Table 3.4**) with different proportions of allele A and B (for each SNP) to evaluate the AE ratios obtained versus the expected ones, as represented in **Table 3.5** In this way, the quantity of each allele in the different samples was extrapolated from the linear regression equation, as explained below.

Table 3.3 - Serial dilution of a heterozygous sample.

Serial Dilution	Dilution factor	Final Concentration of DNA (ng/uL)
ST6	1	10
ATM1	1:2	5
ATM2	1:10	1
ATM3	1:20	0.5
ATM4	1:100	0.1
ATM5	1:200	0.05
ATM6	1:1000	0.01

Table 3.4 - Serial dilution for homozygous samples of my SNPs

Serial Dilution	Dilution Factor	Final Concentration of DNA (ng/uL)
1	1	10
2	1:2	5
3	1:4	2.5
4	1:8	1.25
5	1:16	0.625

C_t values were obtained from Bio-Rad CFX Manager Software 3.1. The threshold used for both reading channels of FAM and VIC were of 260 relative fluorescent units (RFU).

Data obtained by RT-qPCR were analysed on Microsoft Excel 2016 software. Mean and percentage of variation (%var=[SD/Mean]) between replicates were calculated. Linear regression (*linest* function) for Taqman™ calibration curves was performed, as well as the graphical output. The efficiency of RT-qPCR of each allele was calculated using the followed formula:

$$E = 2^{-m}$$

where E corresponds to efficiency and m to the slope of the linear regression for the heterozygous sample for each allele. The quantity of each allele was calculated using the followed equation (that was obtained from the linear regression equation):

$$Q = 2^{m*Ct,i+b}$$

where Q is the quantity, m is the slope of the linear regression and b is the intercept of the same linear regression established by allele in the heterozygous

samples serial diluted and C_t,i is the cycle when the amplification reaches the threshold.

DAE ratios were calculated using the followed formula:

$$DAE = log2(\frac{Q(Allele\ 2}{Q(Allele\ 1})$$

Table 3.5 - Volumes of serial dilutions of each homozygous samples used to form heterozygous mixes with different proportions of allele 1 and allele 2 for rs9997920 and rs6989650.

Reaction (Hetmix)	Volume of AA homozygous samples (uL)	Volume of BB homozygous samples (uL)	Proportion of 2/1
Hetmix 1 (ST1)	10 uL of dilution 1	-	1x/0x
Hetmix 2 (ST2)	5 uL of dilution 1	5 uL of dilution 5	1x/16x
Hetmix 3 (ST3)	5 uL of dilution 1	5 uL of dilution 4	1x/8x
Hetmix 4 (ST4)	5 uL of dilution 1	5 uL of dilution 3	1x/4x
Hetmix 5 (ST5)	5 uL of dilution 1	5 uL of dilution 2	1x/2x
Hetmix 6 (ST6)	5 uL of dilution 1	5 uL of dilution 1	1x/1x
Hetmix 7 (ST7)	5 uL of dilution 2	5 uL of dilution 1	2x/1x
Hetmix 8 (ST8)	5 uL of dilution 3	5 uL of dilution 1	4x/1x
Hetmix 9 (ST9)	5 uL of dilution 4	5 uL of dilution 1	8x/1x
Hetmix 10 (ST10)	5 uL of dilution 5	5 uL of dilution 1	16x/1x
Hetmix 11 (ST11)	-	10 uL of dilution 1	0x/1x

3.6.7. Statistical analyses

Graphical outputs of case-control studies were generated using *beeswarm* R package ("CRAN - Package Beeswarm" 2018). A Levene's test *(levene.test* from *lawstat R* package version 3.2) was applied to test for homogeneity of variances between DAE ratios from cases and controls, prior to the application of the test to analyse differences in mean values. The null hypothesis is that the homogeneity of

variances is equal between the two groups. So if the p-value is higher than 0.05 we cannot reject the null hypothesis and we can assume equal variances between the groups that we are testing, so we can apply a parametric test (Levene, Olkin, and Hotelling 1960). The t-test was applied when the variances between groups were equal and the Welch test when the variances were unequal. The null hypothesis of the t-test is that the means of the two groups are equal. Therefore, if the p-value was lower or equal to 0.05 we rejected the null hypothesis (B. L. Welch 1947; Kim 2015). The *t.test* function from *stats* R package was implemented, changing the argument var.equal = TRUE or FALSE depending on the result from the Levene's test (R Core Team 2017)

CHAPTER IV

Results

4. RESULTS

4.1. RETRIEVAL OF CANDIDATE RISK SNPS

In order to identify candidate risk variants to be further tested for association with breast cancer risk, we began by identifying variants in the literature (published GWAS) that have shown some evidence of being associated with breast cancer risk, although they did not achieve GWAS significance. With this analysis we identified six hundred and eight candidate risk SNPs using data of phase I, II or III from BC GWAS studies (**Annex B**). Since GWAS studies only test a subset of variants present in the microarray, using a haplotype-tagging SNP approach, we also retrieved the proxy SNPs for each index SNP, obtaining in the end a total of 7,429 candidate risk variants/SNPs.

To ensure that the candidate risk SNPs were not in genes already associated with BC risk, we queried the GWAS Catalog for BC-associated genes (p-value $\leq 1 \times 10^{-5}$), and we removed them from our analysis. So, we proceeded the analyses with 3928 candidate risk variants annotated to 614 *ensembl* ids.

In order to strengthen the evidence that our candidates risk SNPs are indeed good candidates to be further tested for association with BC risk, we filtered them based on the iCOGS GWAS data for BC from the European population. This iCOGS dataset tested 14,061,818 SNPs and we first filtered this data to stay with SNPs with a p-value higher than 5×10^{-5} and lower than 0.05. This resulted in 948,036 SNPs with marginal evidence of conferring risk to breast cancer. Then we crossed the 3928 candidate risk variants with the iCOGS SNPs of interest and identified 591 common candidate risk SNPs, as represented in **Figure 4.1** located to 92 different genes (number of different *ensembl ids*).

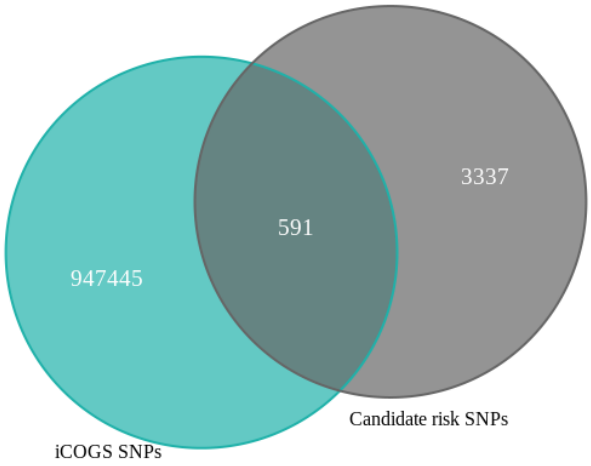


Figure 4.1- *Venn diagram showing the* candidate risk SNPs (591) resulting from the overlap of iCOGS SNPs marginally associated with BC risk (948,036) and candidate risk variants from the GWAS studies identified via GWAS Catalog (3,928).

4.2. CIS-REGULATED CANDIDATE GENES ASSOCIATED WITH BC RISK

Since our goal was to select candidate genes to be tested in case-control association studies using DAE ratios, it is important that the genes to be tested show differential DAE, i.e., evidence of being cis-regulated, in normal breast tissue. In order to identify cis-regulated genes that have candidate risk variants located in them or near (in LD), we filtered the 92 genes where the candidate risk variants were located, according to evidence of having DAE in normal breast tissue (data from the group), as shown in **Figure 4.2** We proceed the analyses with 41 genes (432 candidate risk SNPs).

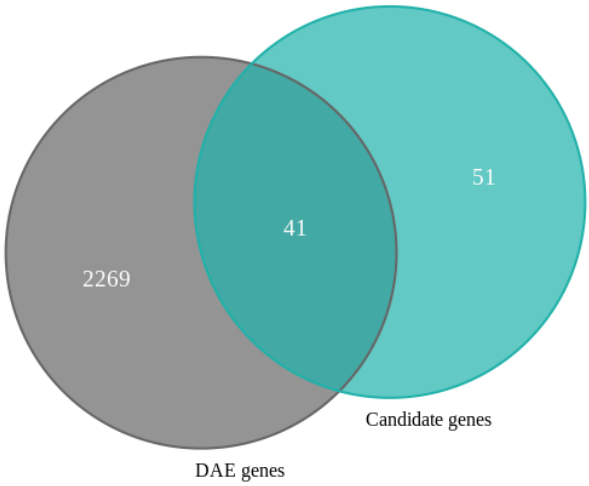


Figure 4.2 - Venn diagram of overlapping 13,570 genes showing evidence of cis-regulation (DAE genes) with 92 candidate risk genes (Candidate Genes) identified through Ensembl IDs attributed to variants retrieved from GWAS data for BC risk with p-value >5x10^{-5} and <0.05 (including those in high LD, at r^2>0.8). Unique numbers are shown for all areas of the diagram.

4.3. CIS-REGULATED GENES ASSOCIATED WITH BC RISK AND WITH CLINICAL IMPACT

In order to identify genes with clinical impact, differential expression analyses were performed using expression and clinical data from the METABRIC project. The data are composed by 1992 tumour cases, with a long clinical history, and 144 normal-matched samples, which were tested for correlation against several clinical variables (see **Table 3.1** in the Material and Methods section). We started the analyses using information on 48803 Illumina probes (corresponding to 21409 genes), which were further shortlisted to 34362 good quality probes (18886 genes), after excluding 13475 bad and 966 no match probes. We performed the analyses of differential expression and considered genes to be differentially expressed when significant for at least one of the clinical variables (FDR p-value < 0.01 and |FC| ≥ 1.5), being them tumours vs normal-matched, ER status, PR status, HER2 status and grade.

In **Table 4.1** shows the number of *Ensembl* genes found to be differentially expressed between clinical sub-groups. In total, 10599 genes were identified as differentially expressed for at least one clinical variable. Next, we crossed this

information with that of the genes with evidence of being cis-regulated, and found 2012 genes in common. This result is depicted in **Figure 4.3.**

Table 4. 1 - Number of genes differentially expressed according to clinical context.

Differential Expression analysis	Differential expressed genes/total genes
Tumours vs Normal-matched	2,685/18,886
ER+ vs Normal-matched	2,727/18,886
ER- vs Normal-matched	3,191/18,886
ER+ vs ER-	914/18,886
HER2+ vs Normal-matched	3,283/18,886
HER2- vs Normal-matched	2,646/18,886
HER2+ vs HER2-	251/18,886
PR+ vs Normal-matched	2,720/18,886
PR- vs Normal-matched	2,808/18,886
PR+ vs PR-	322/18,886
Grade	9,786/18,886

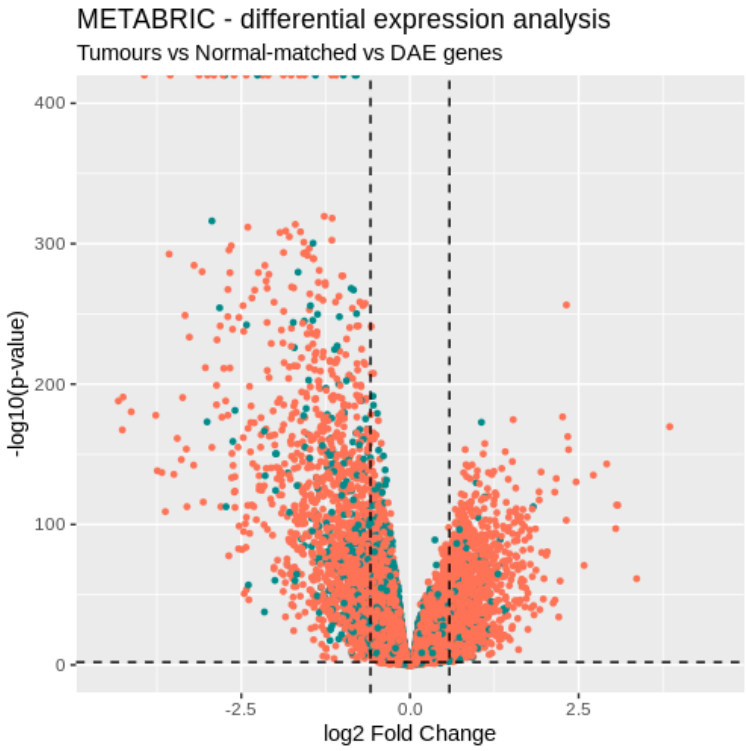


Figure 4.3 - Volcano plot for differential gene expression analysis for tumours vs normal matched tissue taking in account evidence for cis-regulation of genes. The x axis displays the log2 FC, representing the biological magnitude of the difference, and the y axis displays the $-\log_{10}$ of the FDR corrected p-value, representing the test's significance. Dotted lines represent the thresholds applied for establishing statistical significance for the differential expression of a gene, set at p-value = 0.01 (1% FDR) and of FC = 1.5. Points in cyan blue correspond to genes with evidence of being cis-regulated, and orange points correspond to genes without this evidence in breast tissue.

As we are interested in selecting cis-regulated genes with potential association with BC risk to BC, and with a clinical impact, for experimental validation, we crossed the previously selected 41 genes (see Figure 4.2) with the 10599 genes that were differentially expressed genes in at least one clinical variable. In the end, we identified 18 genes with significant clinical impact and potentially associated with BC risk, with evidence of being cis-regulated (**Figure 4.4, Table 4.2**).

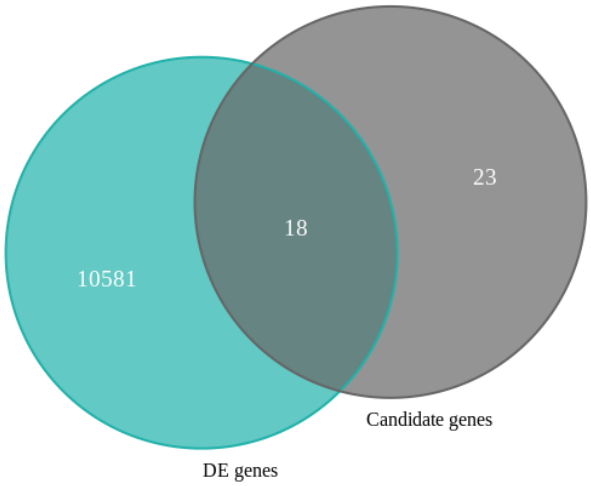


Figure 4. 4 - Venn diagram showing cis-regulated genes associated with BC risk and with clinical impact resulting from the overlap of candidate cis-regulated genes associated with BC risk (41) and genes that are considered with differential gene expression at least one clinical variable (10,599).

Table 4. 2 - Final list of candidate risk genes genes, grouped by significant clinical analysis. In bold are shown the two genes selected for the in case-control association study using AE ratios, used as validation of their association with risk. Continued on next page

Differential Expression analysis	Significant Genes
Tumours vs normal-matched	***GRHL2***, *MECOM, NCALD, RMI2, FUT8, MAFF, PCCA, PECR*
ER+ vs normal-matched	***GRHL2***, *NCALD, RMI2, FUT8, MAFF, MECOM, PCCA, PECR*
ER- vs normal-matched	***GRHL2***, *NCALD, RMI2, MAFF, MECOM, PCCA, PECR, SYK*
ER+ vs ER-	*FUT8*
PR+ vs normal-matched	***GRHL2***, *NCALD, RMI2, MAFF, MECOM, PCCA, PECR*
PR- vs normal-matched	***GRHL2***, *NCALD, RMI2, FUT8, MAFF, MECOM, PCCA, PECR, SYK*
PR+ vs PR-	-
HER2+ vs normal-matched	***GRHL2***, *NCALD, RMI2, FUT8, MAFF, MECOM, PCCA, PECR, SYK*
HER- vs normal-matched	***GRHL2***, *NCALD, RMI2, FUT8, MAFF, MECOM, PCCA, PECR*

Table 4.2 - continued

HER2+ vs HER2-	-
Grade	***OCIAD1***, ***GRHL2***, *RMI2, CHMP4B, CLEC16A, FRYL, FUT8, LYPD5, MECOM, MREG, MYL3, NUP107, PCCA, SYK, TGM5*

4.4. PRIORITIZATION OF GENES AND VARIANTS FOR THE CASE-CONTROL ASSOCIATION STUDY

As it was unfeasible to validate all 18 candidate genes for association with BC risk, they were compared based on the relevant variables (risk to BC, DAE, and clinical impact), in order to select the two best candidates for the association studies in the lab. One of the genes, MYL3, was eliminated immediately because, after a fine curation, it was found that the candidate variant rs6796502 had multiple entries in the GWAS Catalog and already associated with BC risk according to one of the studies. Regarding the remaining 17 genes (represented in **Table 4.3**), firstly we looked at the p-value of the candidate risk SNPs and odds ratio in the original reporting study, as well as in the iCOGS study, in order to select the genes with the lowest p-value. We also looked at the MAF in the European population in order to exclude possible rarer variants. The total expression of the transcripts in normal breast tissue, as well as in blood (data from GTEx project), was also taken into consideration to select genes with higher expression. The in-house evidence of differential allelic expression of each gene in normal breast tissue was also considered, as well as the clinical impact of the gene on BC. We also looked for the presence of eQTLs, consulting GTEX project, for the tSNP of each candidate gene.

Table 4. 3 - 17 final candidate cis-regulated genes, with potential to be associated with risk to BC and associated with clinical impact. Continued on the next page

Gene with DAE and clinical impact	Ensembl ID	Candidate riskSNP	Cytoband	p-value GWAS study	Odds ratio GWAS study**	p-value iCOGS	Odds ratio iCOGS	MAF in Europeans	Study
FUT8	ENSG00000033170	rs76389600	14q23.3	$7.76x10^{-4}$ (I&II)	0.79 [0.69-0.91]	0.027	1.11	0.03	Low, Siew-Kee 2013
CLEC16A	ENSG00000038532	rs998592	16p13.3	$8x10^{-1}$(III)	0.98 [0.89-1.08]	0.02	0.97	0.42	Thomas, Gilles 2009
FRYL	ENSG00000075539	rs6843340	4p11	$9x10^{-4}$(III)	0.99 [0.94-1.05]	0.01	1.02	0.45	Easton, Douglas 2007
OCIAD1	ENSG00000109180	rs6843340	4p11	$9x10^{-4}$(III)	0.99 [0.94-1.05]	0.02	1.02	0.45	Easton, Douglas 2007
TGM5	ENSG00000104055	rs6493076	15q15.2	$9.79x10^{-1}$(II)	0.89 [0.75-1.04]	0.018	0.95	0.1	Sehrawat, Badan 2011
CHMP4B	ENSG00000101421	rs6059504	20q11.22	$9.31x10^{-3}$(I&II)	0.91*	0.02	0.97	0.22	Antoniou, Antonis 2010
MAFF	ENSG00000185022	rs5756968	22q13.1	$1X10^{-2}$(I&II)	0.89 [0.82-0.96]	0.018	1.03	0.41	Long, Jirong 2010
PECR	ENSG00000115425	rs4672790	2q35	$2.49x10^{-3}$(II)	0.80-0.95*	0.022	0.98	0.43	Gaudet, Mia 2010
MREG	ENSG00000118242	rs4672790	2q35	$2.49x10^{-3}$(I&II)	0.80-0.95*	0.022	0.98	0.43	Gaudet, Mia 2010
RMI2	ENSG00000175643	rs4451969	16p13.13	$1.6X10^{-1}$ (I&II)	1.07 [0.99-1.17]	0.002	1.04	0.43	Long, Jirong 2010
NUP107	ENSG00000111581	rs2546513	12q15	$6.08x10^{-1}$(II)	0.89 [0.80-1.00]	0.04	0.97	0.32	Sehrawat, Badan 2010

Table 4.3 - Continued

GRHL2	ENSG00000083307	rs2387620	8q22.3	$5.9x10^{-4}$(I, II & III)	1.05 [1.00-1.11]	0.006	1.03	0.32	Fletcher, Olivia 2011
GRHL2	ENSG00000083307	rs2211914	8q22.3	$4.96x10^{-5}$(I, II&III)	1.08 [1.02-1.14]	0.02	1.02	0.32	Fletcher, Olivia 2011
MECOM	ENSG00000085276	rs1918964	3q26.2	$1.77x10^{-2}$(I&II)	1.07*	0.03	1.02	0.42	Antoniou, Antonis 2010
LYPD5	ENSG00000159871	rs17725531	19q13.31	$2.56x10^{-2}$(I&II)	-	0.002	1.06	0.12	Palomba, Grazia 2015
SYK	ENSG00000165025	rs12553524	9q22.2	$2.03x10^{-2}$(I&II)	0.82-0.98*	0.01	1.03	0.23	Gaudet, Mia 2010
PCCA	ENSG00000175198	rs1112044	13q32.3	$2x10^{-2}$(I&II)	0.91 [0.84-0.99]	0.03	1.02	0.44	Long, Jirong 2010

I&II is the combined p-value between phase I and II of GWAS; III is the p-value on phase III; II is the p-value in phase II of the GWAS and I, II & III is concerning to the combined p-value on three phases. * is relative to the hazard ratio. ** Odds ratio at 95% confidence interval.

Two candidates were selected to be validated experimentally, based on the description above: *OCIA Containing Domain 1* (*OCIAD1*) gene, with DAE tSNP: rs9997920; and *Grainyhead Like Transcription Factor 2* (*GRHL2*) gene, with DAE tSNP: rs6989650.

rs6843340 is mapped to the 4p11 locus, in a large intron of the *FRYL* gene 5' to the *OCIAD1* gene (**Figure 4.5**), and its minor allele (T) has a frequency of 0.45 in the European population (HapMap CEU data). This SNP was genotyped by Easton et al as part of the phase 3 on the first GWAS for BC but was reported as not being significantly associated with BC after this final phase (it was associated in phase I and II with a combined p-value of 0.004, but not in phase III alone, with p-value of 0.97) (Easton et al. 2007). Subsequently, in the iCOGS study it had a p-value of 0.011, which we considered to be a good p-value to proceed with our studies.

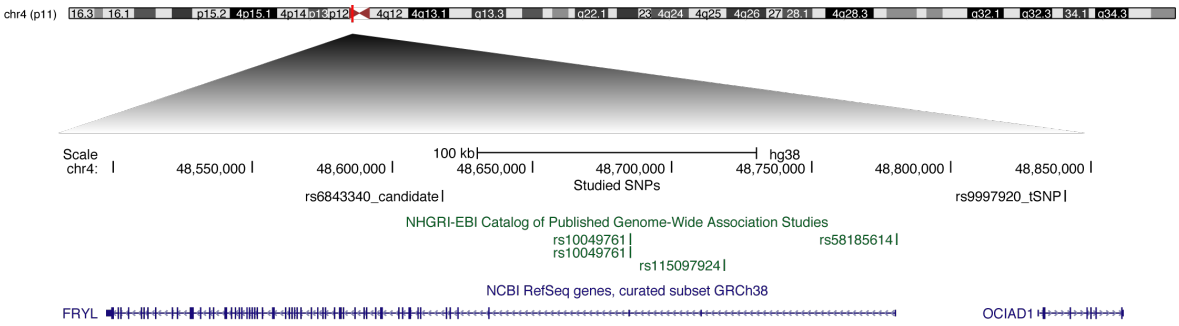


Figure 4.5 - Candidate 4p11 genomic locus. It shows the position of the GWAS candidate risk variant rs6843340 and the DAE tSNP rs9997920 (both in black). In green are shown the locations of SNPs associated with risk to diseases other than BC, from the GWAS Catalog. In blue are shown the positions of the RefSeq genes *FRYL* and in *OCIAD1*.

The *OCIAD1* gene is expressed quite widely across tissues, but it is less expressed in blood than in breast tissue, the two tissues in which we carried the subsequent case-control analysis ("GTEx Portal" 2018) (**Figure 4.6**). In the clinical analysis, its expression was correlated with Grade (**Figure 4.7 A)**, more specifically lower expression of *OCIAD1* correlates with higher BC Grade.

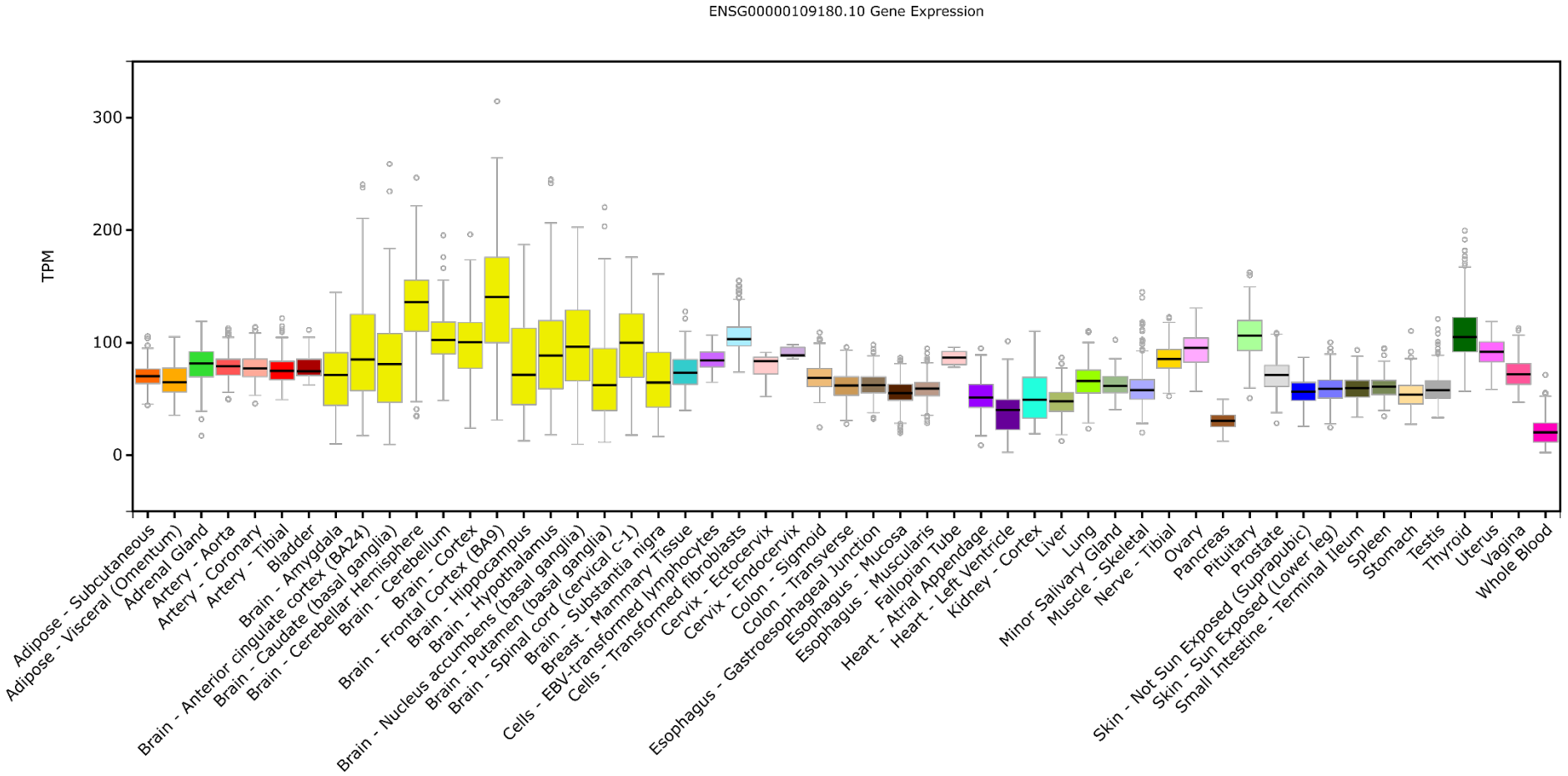


Figure 4.6 - *OCIAD1* gene expression levels across tissues. The y axis corresponds to transcripts per million expressed and in the x axis it is represented the tissues. From "GTEx Portal" 2018.

Whole-genome DAE data generated in our group (Xavier et al, unpublished) revealed that in this locus *OCIAD1* is cis-regulated, and that the minor T allele of the tSNP rs9997920 is less expressed that the common C allele (**Figure 4.7B**). Interestingly, although rs9997920 is not an eQTL for the expression of *OCIAD1* in breast tissue, it is so in whole blood, where the TT genotype group is also the one with lower expression, in concordance with our DAE data ("GTEx Portal" 2018). In the DAE data was also observed that this tSNP (rs9997920) has a high heterozygosity frequency and more samples displaying DAE than the *FRYL* tSNPs, which is crucial to measure DAE in the future case-control study. Therefore, this was the tSNP selected for the next phase of our study.

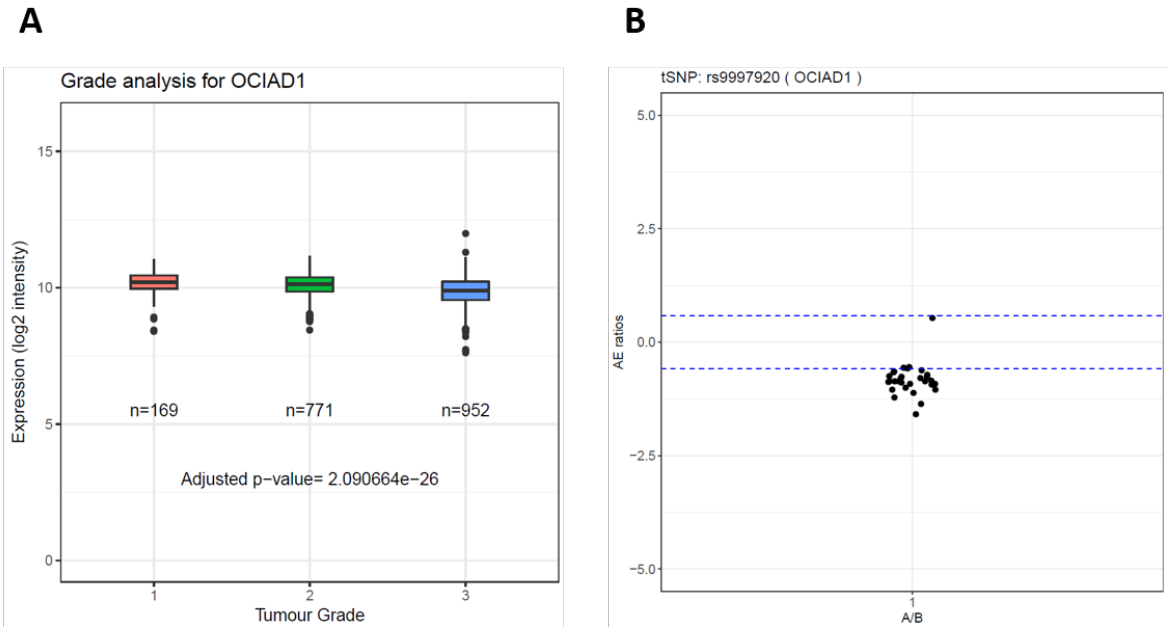


Figure 4.7 - OCIAD1 clinical and functional characterisation. In figure **A**, the x-axis represents the grade groups (1-3) and in y-axis it is indicated the expression values in log2 intensities. The graphic shows the results for the grade analysis at probe ILMN_1700306 from OCIAD1. In figure **B** it is represented a AE analysis plot (performed by Xavier, Joana) for a tSNP in OCIAD1 (rs9997920) where it is possible to observe that most of the samples (black dots) display DAE. In the y axis are the AE ratios and in x axis are the heterozygous samples, where the allele A corresponds to T allele from OCIAD1 and the allele B indicates the C allele from OCIAD1**.**

Regarding to rs2387620 is mapped to 8q22 locus, in a intron variant of *NCALD* gene (**Figure 4.8**), and its minor allele (A) has a frequency of 0.47 in European population (HapMap CEU data). This SNP was genotyped by Fletcher et al as part of phase III but was reported as not being significantly associated with

BC (Fletcher et al. 2011). It was associated in phase I with a p-value of 7.48 x 10^{-3} and in phase II with a p-value of 1.18 x 10^{-1}, but not in phase III with a combined p-value 5.90 x 10^{-4}. Consequently, in the iCOGS study it had a p-value of 0.006, which we considered to be a good p-value to proceed with our studies. The other candidate risk SNP for GRHL2 study is rs2211914 that is mapped to 8q22 locus too (**Figure 4.8**), in an intron variant of *GRHL2* gene, and its minor allele (C) has a frequency of 0.40 in European population (HapMap CEU). Subsequently in the iCOGS study it had a p-value of 0.02. The LD between this risk candidate SNPs were verified, and they in moderate LD with each other (r^2=0.59) which means that although they are not highly correlated their effect is not independent.

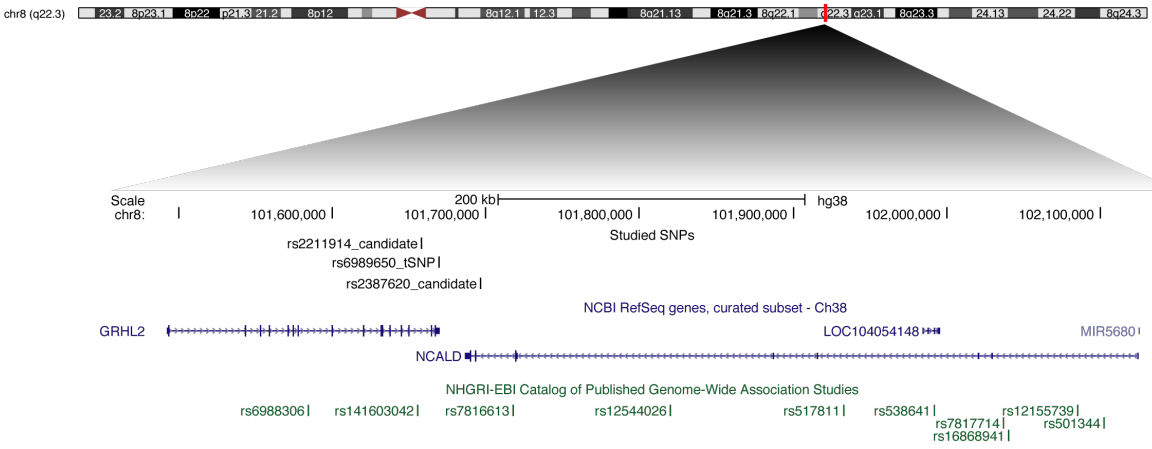


Figure 4.8 - Candidate 8q22 genomic locus. It shows the position of the GWAS candidate risk variants rs2211914 and rs2387620 and the DAE tSNP rs6989650 (all in black). In blue are shown the positions of the RefSeq genes, including *NCALD* and *GRHL2*. In green are shown the locations of SNPs associated with risk to diseases other than BC, retrieved from the GWAS Catalog.

The *GRHL2* gene is lowly expressed across tissues, but in breast it shows some notorious expression, in blood it does not show expression, so this gene only can be tested in breast (**Figure 4.9**). In the clinical analysis, its expression was correlated with tumours in general, ER status, PR status, HER2 status and grade, where overall *GRHL2* is overexpressed in tumours vs normal-matched (**Figure 4.10- B-F**).

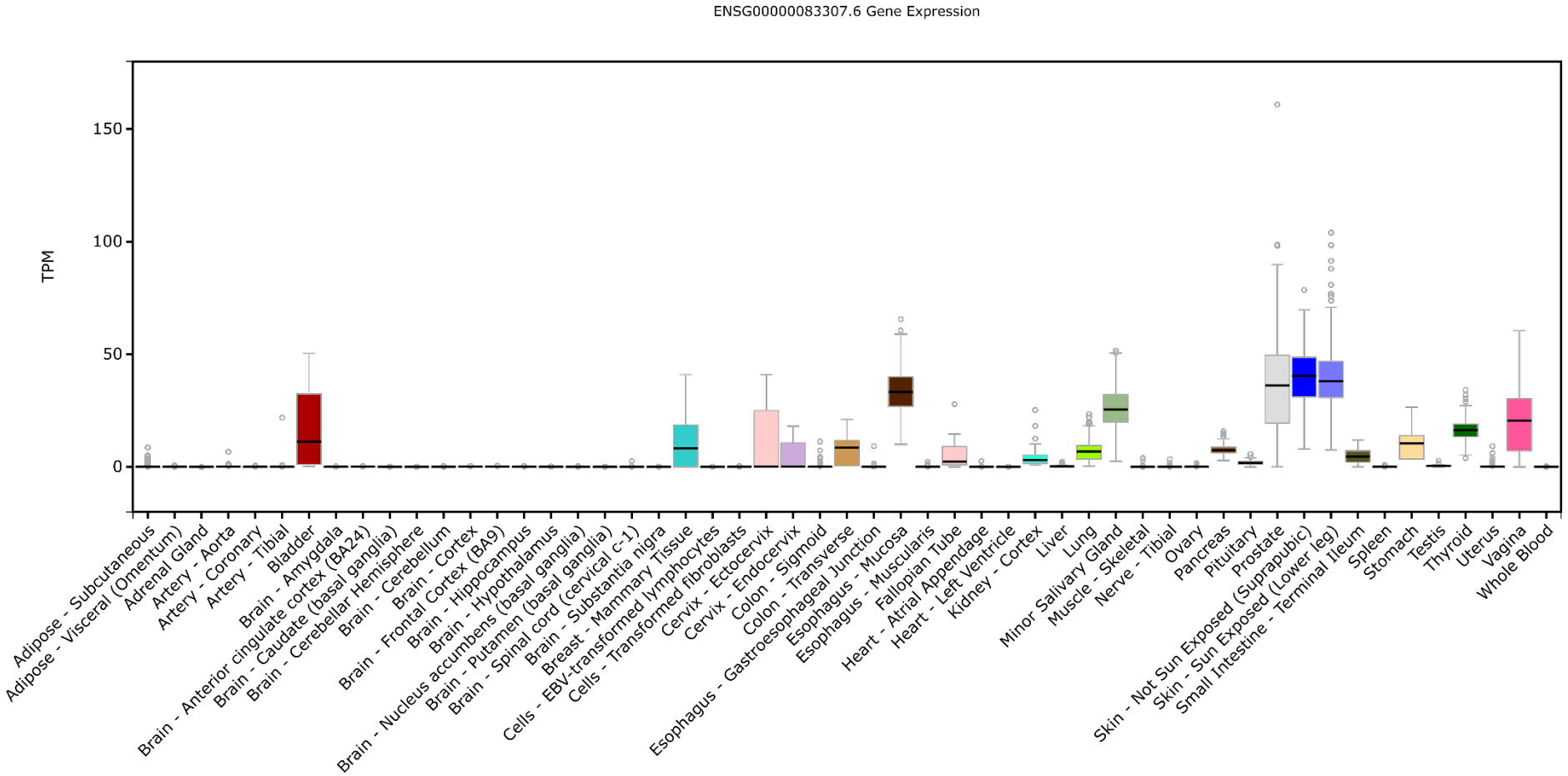


Figure 4.9 – *GRHL2* gene expression levels across tissues. The y axis corresponds to transcripts per million expressed and in the x axis is represented the tissues From ("GTEx Portal" 2018)

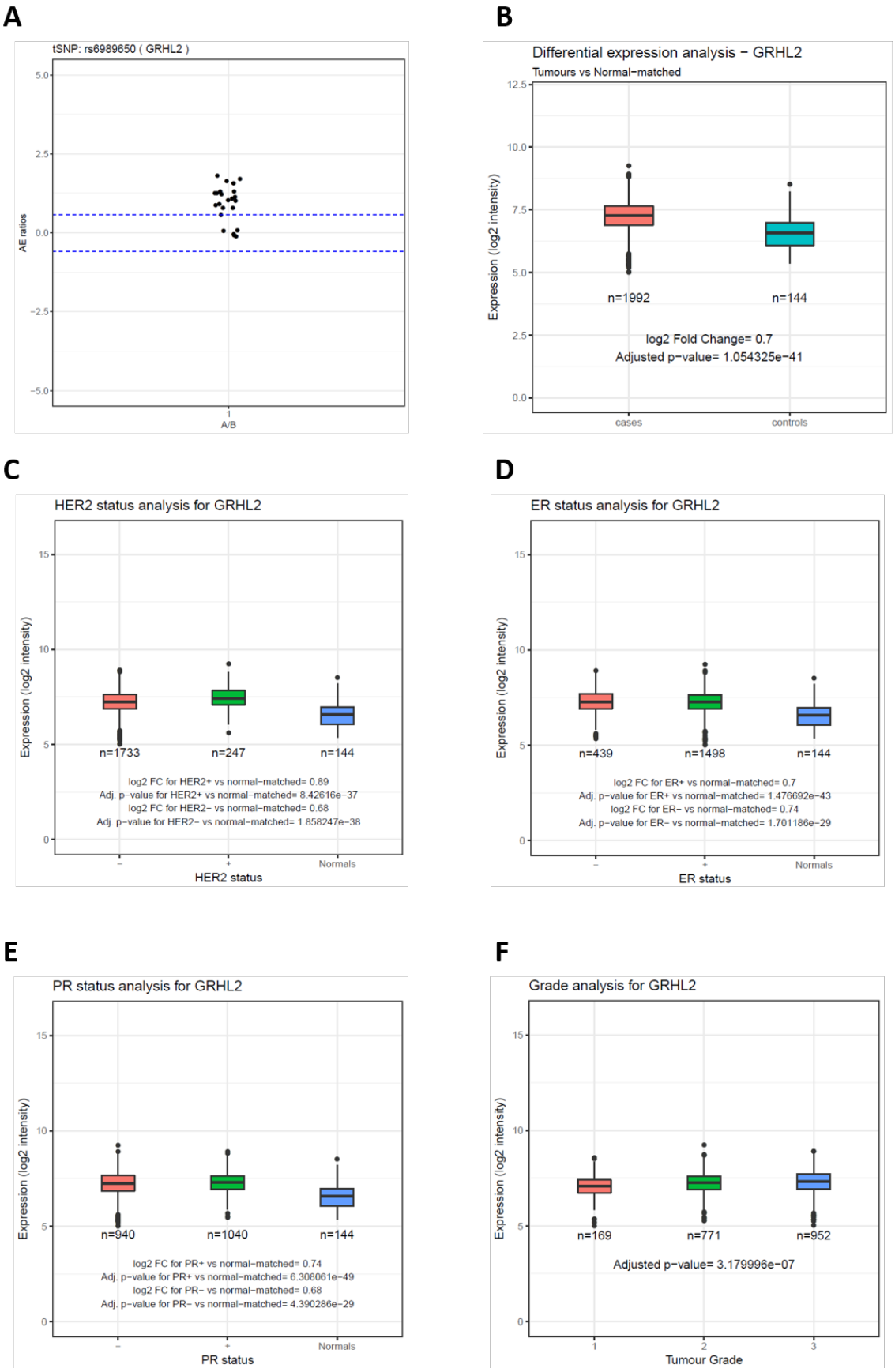


Figure 4.10 - GRHL2 clinical and functional characterisation. In **A** it is represented an AE analysis plot (performed by Xavier, Joana) for a tSNP in GRHL2 (continued on the next page)

Figure 4.10 (continued) - (rs6989650) where is possible to observe that most of the samples (black dots) display DAE. In the y axis are the DAE ratios and the x axis are the heterozygous samples and in, where the A allele corresponds to (T) and the B allele represents (**C**). In **B - F** x-axis represents the clinical variables (Tumours, ER status, PR status, HER2 status, Grade status) and NM controls and y axis indicates the expression values in log2 intensities. In **B** the graphic shows tumours vs NM analysis at probe ILMN_2060145 from *GRHL2*. In **C** it is represented the differential expression analysis between HER2 cases vs NM for GRHL2 (ILMN_2060145); Figure **D** presents the analyses of ER+ cases vs NM and ER- cases vs NM for GRHL2 gene. In **E** it is demonstrated the differential expression analyses for PR status cases vs NM. In **F** it is shown the grade analysis for GRHL2 gene.

Whole-genome DAE data generated in our group (Xavier et al. 2016, unpublished) revealed that this locus GRHL2 is cis-regulated and composed by a lot of tSNPs, but one of the most interesting to study was rs6989650 (**Figure 4.10 A**) because it has high heterozygosity frequency, and more samples displaying DAE. This tSNP has the major allele (C) more expressed than the minor (T) (**Figure 4.10 A**). This SNP is not an eQTL for the expression of *GRHL2* in breast tissue ("GTEx Portal" 2018). So, this was the tSNP selected for the next phase of our study.

4.5. GENOTYPING OF BLOOD, BREAST TISSUE SAMPLES AND CEPH SAMPLES

To perform DAE analysis, and subsequently the association studies, we first needed to identify heterozygous samples for the tSNPs of both genes.

Firstly, 20 CEPH samples were genotyped for each tSNP (rs9997920 and rs6989650) in order to identify homozygous samples for both alleles of the two SNPs to be selected for the construction of calibration curve and series standard curves (**Annex C**).

The tSNPs rs9997920 and rs6989650 were genotyped in 49 normal-matched breast tissue (cases) and in 45 breast tissue samples (controls). For both SNPs we identified the presence of the expected three genotype groups (**Figure 4.11**), as the minor allele frequency for both SNPs in high. The negative controls were not amplified (**Figure 4.11),** indicating that there was no contamination during the PCR reaction. We identified 26 heterozygous samples for rs9997920 in normal-matched samples and 19 heterozygous samples for the same SNP in normal breast samples.

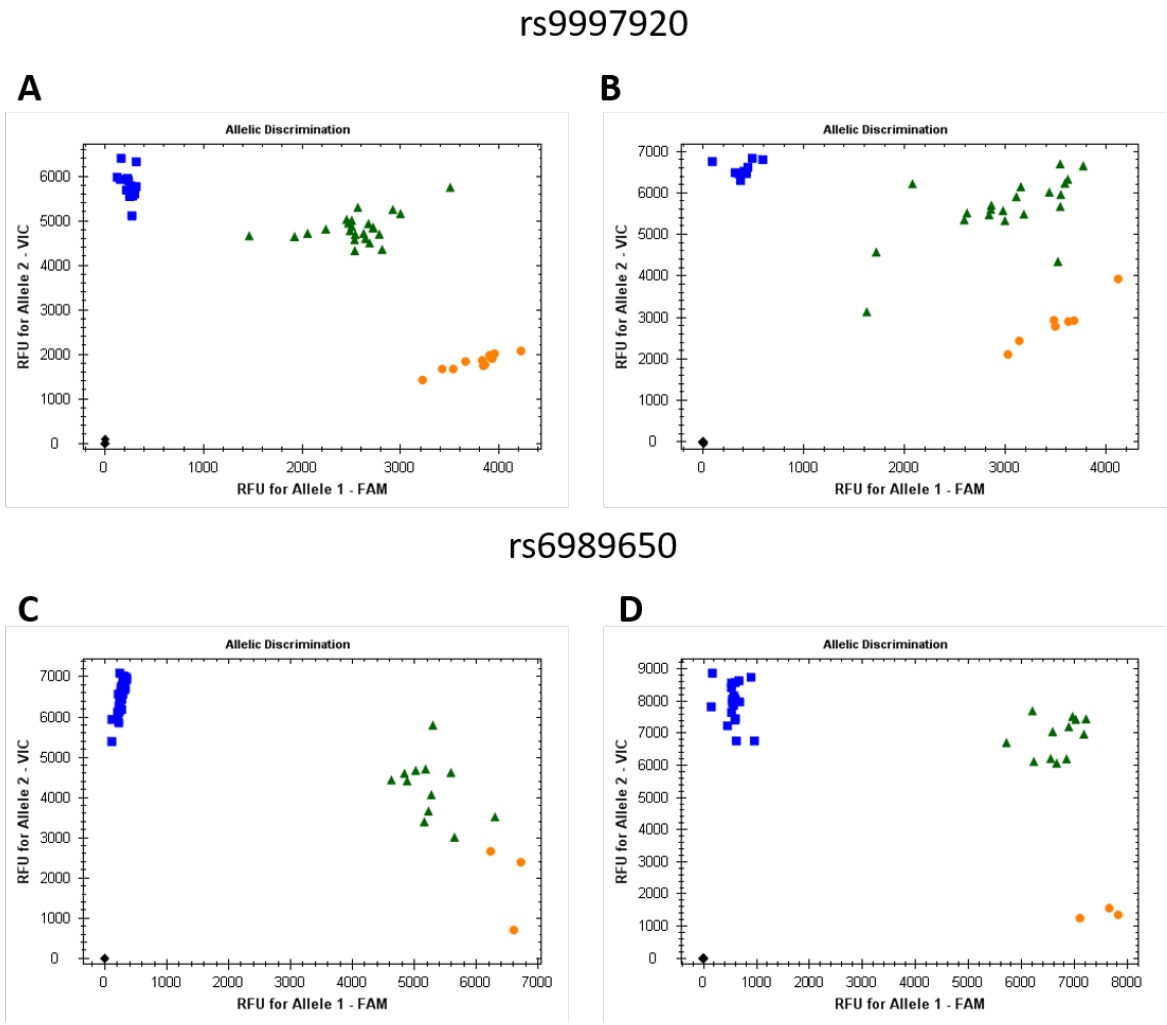


Figure 4.11 - Genotyping for rs9997920 and rs6989650 in breast tissue. The x-axis represents the fluorescence intensity for Allele 1, emitted by probe FAM and the y-axis indicates the fluorescence intensity for Allele 2, emitted by the probe VIC. The orange circles represent the homozygous samples for Allele 2, the blue squares indicate the homozygous samples for Allele 1, the green triangle represent heterozygous samples and the black diamonds represent the NTC samples (without fluorescent signal as expected). In **A** and **B** are displayed the genotyping results for rs9997920, in the *OCIAD1* gene, in normal-matched and normal breast tissues, respectively, where Allele 1 corresponds to the T allele of rs9997920 and the Allele 2 corresponds to the C allele. In **C** and **D** are the genotyping results for rs6989650, in the *GRHL2* gene, in normal-matched samples and breast tissue samples, respectively, where Allele 1 indicates the T allele of rs6989650 and Allele 2 indicates C allele of rs6989650.

Regarding rs6989650, we identified 12 heterozygous samples in normal breast tissue and 12 heterozygous normal-matched samples.

As *OCIAD1* is also expressed in blood and we could perform the case-control study in these tissues, rs9997920 was also genotyped in blood cancer samples of patients with BC (**Annex D**). There was no need to genotype samples from healthy

blood donors as the genotypes were already available for rs9997920 in these samples (26 heterozygous samples).

4.6. QUANTIFICATION OF DIFFERENTIAL ALLELIC GENE EXPRESSION TO PERFORM CASE-CONTROL STUDIES

In the last part of this work, we tried to validate the association of variants with BC risk, by performing a case-control study using DAE as a quantifiable variable. For this study, as mentioned before, only heterozygous samples are informative, and from the experiments described above, we identified 26 normal-matched and 22 normal breast heterozygous samples for rs9997920 (*OCIAD1*), and 12 normal-matched and 15 normal breast heterozygous samples for rs6989650 (*GRHL2*). For rs9997920 (*OCIAD1*), we also identified 26 heterozygous blood samples of cancer patients and 13 heterozygous samples of healthy blood controls, from genotyping analysis previously carried in our group.

As we want to precisely quantify allelic expression ratios in order to perform the case-control association studies, in all the experiments we included a calibration curve consisting on a serial diluted heterozygous CEPH sample (serial diluted samples). This heterozygous sample was composed of two homozygous samples for different alleles (one homozygous for allele 1 and the other for allele 2 for each one of the two SNPs under study) mixed in equimolar proportions (50:50). This curve was used in each qPCR run to do an extrapolation of individual allelic quantities. Because the efficiency of amplification of each allele of the SNP is different, this was also a crucial step to calculate the RT-qPCR efficiency for each allele separately.

All experiments were run twice and all samples were tested in triplicate in each run. These triplicates revealed a very low percentage of variation, smaller than 5%. For both SNPs, in all qPCRs performed in breast tissue, two control samples were excluded from the analyses because the cDNA did not amplify.

Regarding the study of DAE of *OCIAD1* in blood samples, all the heterozygous serial diluted samples were used to perform the linear regression for each allele (**Figure 4.12 A-B, Annex E**) in both runs. In the first run the RT-qPCR efficiency for allele 1 was of 83.81% and for allele 2 was of 78.398% (with

correlation r^2 of 0.99 for both). The second run had 69.492% efficiency of amplification for allele 1 and 60.947% for allele 2 (with correlation r^2 of 0.97 and 0.99, respectively).

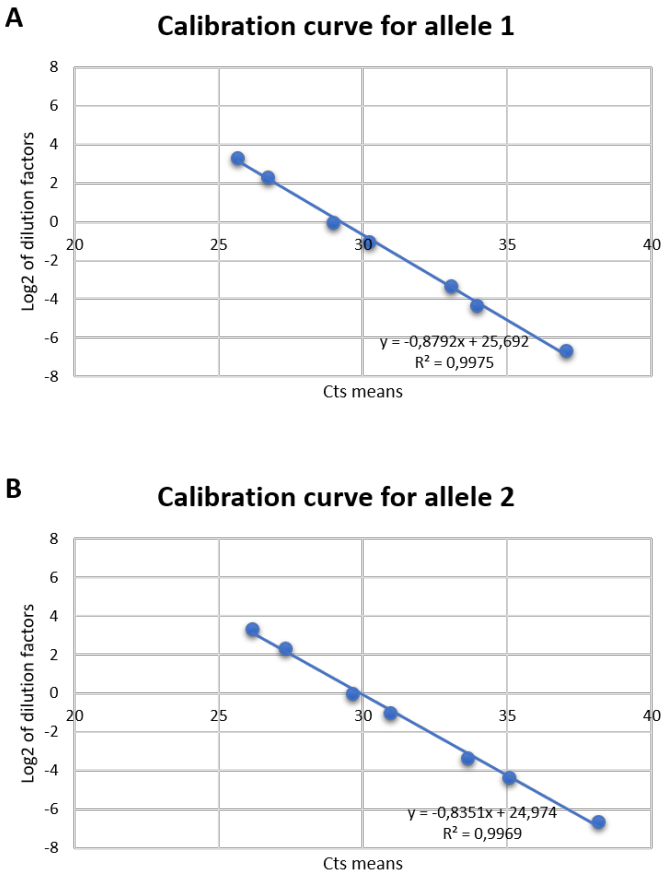


Figure 4.12 - Calibration curve for the OCIAD1 blood RT-qPCR in the first run. These figures show the calibration curve for each one of the alleles of a heterozygous sample for rs9997920 of OCIAD1 gene in the blood study. In **A** and **B** are shown the standard curves for allele 1 (T allele) and allele 2 (C allele), respectively, showing for both the corresponding linear regression equation.

To further ensure our ability to accurately measure allelic expression ratios, we performed a series of standard curves, by mixing two CEPH homozygous samples with different proportions of allele 1 and allele 2, for each one of the two SNPs under study (Hetmixes samples). The DAE ratios were calculated for all experiments (observed ratios) and compared with the expected one given the different proportions prepared.

From **Figure 4.13** it is observed that the observed DAE ratios are similar to the expected, with the two values highly correlated (R^2= 0.97 for the first run of *OCIAD1* in blood). To perform the correlation analysis Hetmix 1 and Hetmix 11 were removed because the expected ratio is minus and plus infinite, since both of them are homozygous samples. For all the Hetmixes performed for *OCIAD1* both in blood tissue and in breast tissue the DAE ratios observed were coincident with the expected ones (**Annex G**), giving us confidence to proceed to the case-controls studies. It is worth pointing that the observed DAE ratios started differing from the expected ones for expected DAE values greater than 2 or lower than -2. This means that our ability to measure DAE is stronger inside the interval DAE ratios= [2; -2].

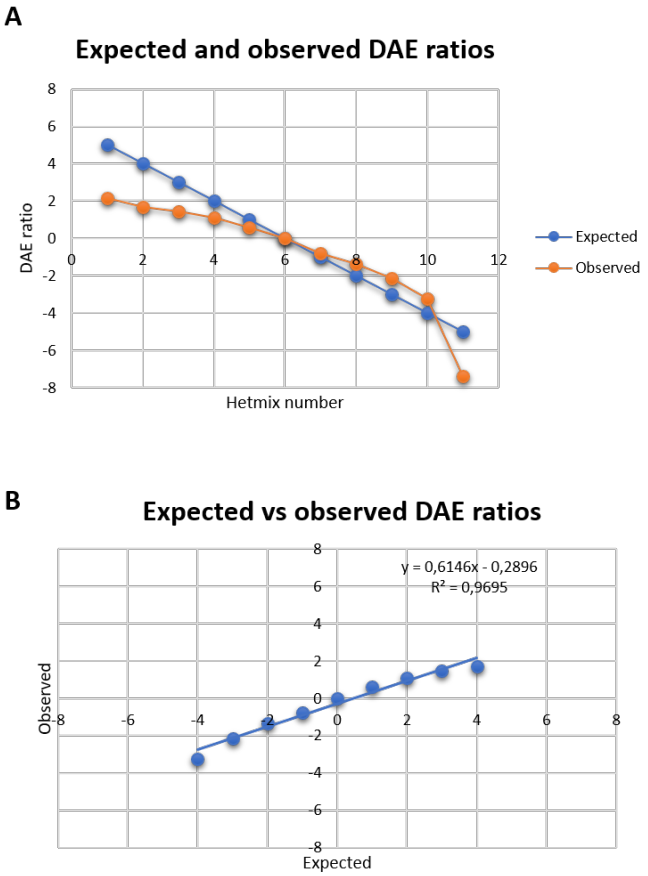


Figure 4.13 - Hetmixes DAE ratios for rs9997920 (OCIAD1) at the first RT-qPCR in blood tissue. In A it is represented in blue the DAE expected ratios and in orange the measured AE ratio. The points at the end of the expected ratio represents +∞/-∞ (1/0= ∞) and once this point was not possible to insert in the graphic the 5/-5 values were chosen. In B is shown the correlation graphic between observed and expected DAE ratios (samples with AE ratio of ∞ were removed).

Regarding the study of *GRHL2*, in breast tissue samples, in the first run all heterozygous serial diluted samples were used for the elaboration of the calibration curve, whilst in the second run five samples were used, as some samples did not amplify. In the first run the efficiency of the RT-qPCR was of 91,55% for allele 1 and of 86,8% for allele 2 (with correlation r^2 of 0.99 for both alleles) (**Figure 4.14**). In the second run, it was observed an efficiency of 75,6% for allele 1 and of 79,7% for the allele 2 (with correlation r^2 of 99% for both alleles) (**Annex H**).

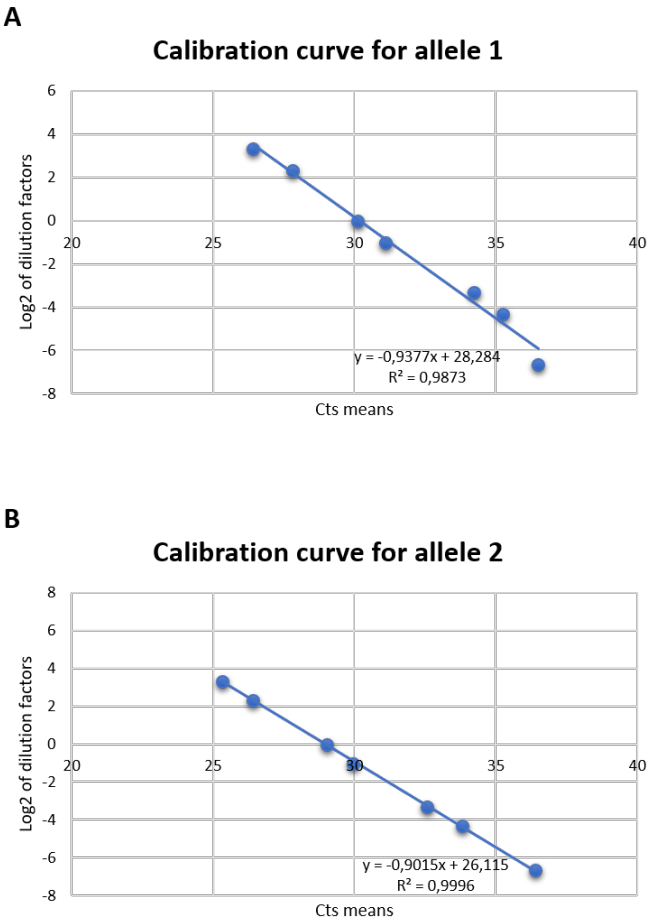


Figure 4.14 - Calibration curve for the GRHL2 tissue RT-qPCR in the first run. These figures show the calibration curve for each one of the alleles of a heterozygous sample for rs6989650 of *GRHL2* gene in the blood study. In A and B are shown the calibration curves for allele 1 (FAM, C allele) and allele 2 (VIC, T allele), respectively, showing for both the corresponding linear regression equation.

Once more, to ensure that the allelic ratios measured were accurate the Hetmix standard positive control curve for rs6989650 GRHL2 gene were performed and the DAE ratio was calculated for all experiments and compared

with the expected one given the different proportions prepared. As we can see in the **Figure 4.15 A-B**, the obtained DAE are similar to the expected, being the two values highly correlated (correlation R^2 = 0.98 for the first run of GRHL2 in tissue, and correlation R^2= 0.89 for the second run of *GRHL2*).

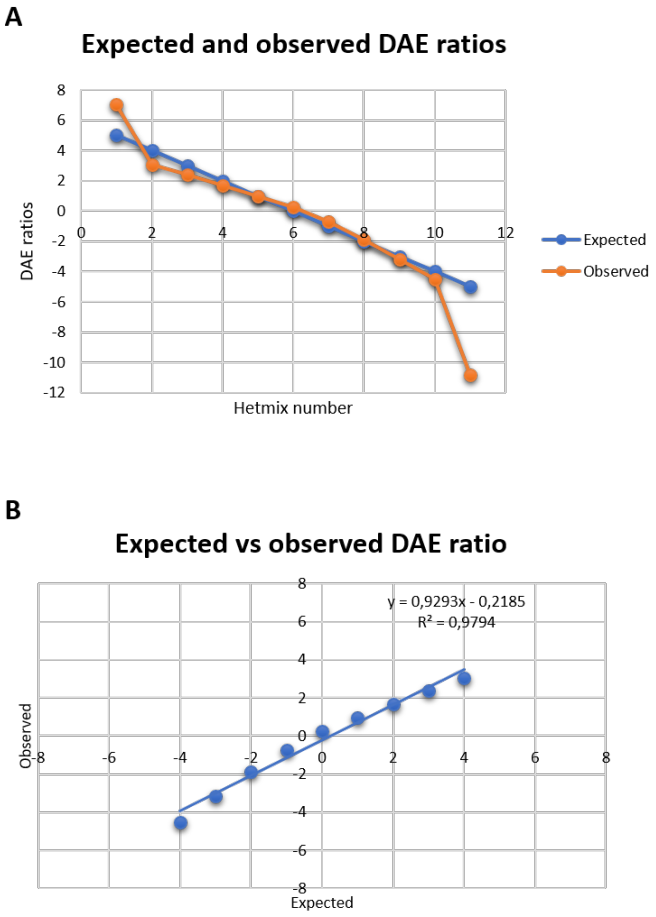


Figure 4.15 - Hetmixes DAE ratios for rs6989650 (*GRHL2*) at the first RT-qPCR in breast tissue. In **A** it is represented in blue the DAE expected ratios and in orange the observed DAE ratio. The points at the end of the expected ratio represents +∞/-∞ (1/0= ∞) and once this point was not possible to insert in the graphic the 5/-5 values were chosen. In **B** is shown the correlation graphic between observed and expected DAE ratios (samples with DAE ratio of ∞ were removed).

To perform the correlation analysis Hetmix 1 and Hetmix 11 were removed because the expected ratio plotted in **Figure 4.15 and Annex I** was not correct since both of them are homozygous samples and therefore the DAE ratio should take infinitive values. For all the Hetmixes performed for *GRHL2* in tissue the DAE ratios observed were coincident with the expected one (**Figure 4.15, Annex I**), giving us confidence to proceed to the case-controls measurements. As mentioned

above for *OCIAD1*, the DAE ratios observed starting to differ from the expected ones for ratios greater than 2 or lower than -2.

4.7. BC CASE-CONTROL STUDIES FOR OCIAD1 AND GRHL2 GENES

In the last part of this project, DAE ratios were calculated for cases and controls samples, for both SNPs in the indicated tissue samples, and the association studies using DAE ratios as a quantitative trait were performed. Regarding the case-control study for *OCIAD1* in blood, rs9997920 was tested in the first run in 26 blood samples from BC patients (cases) and in 13 blood samples from healthy donors (controls). In the second run, the respective number of samples was 26 cases and 8 controls (the cDNA for the other five samples was no longer available). The homogeneity of variances between cases and controls was tested and no significant differences were found, both in the first run and in the second (Levene test, p-value=0.2 and p-value=0.53, or the first and second runs respectively). A *t*-test was then applied, and in both runs significant differences in the average DAE ratios, between cases and controls, were found (p-value= 0.002 and 0.008 for the first and second runs, respectively). These results suggest that DAE ratios of *OCIAD1* in blood are associated with BC risk. The results for the first run are presented in **Figure 4.16 A** and for the second run in **Annex J.**

The case-control study for *OCIAD1* in breast tissue was performed in two independent runs, both with 24 cases of NM samples from BC patients and 22 controls from breast tissue samples of healthy donors. The DAE was quantified just for 20 controls because two samples did not amplify. Before performing the case-control association studies, the homogeneity of the measured DAE ratios distribution variances were tested, and significant differences were found (Levene's test, p-value=0.02 and p-value=0.0002, on the first and second runs respectively). Therefore, to test differences in the mean DAE values between cases and controls a Welch t-test was applied, and that did not find significant differences (p-value= 0.4 and 0.07 for the first and second runs, respectively). The association study results for the first run are shown in **Figure 4.16 B** and for the second run in **Annex K.**

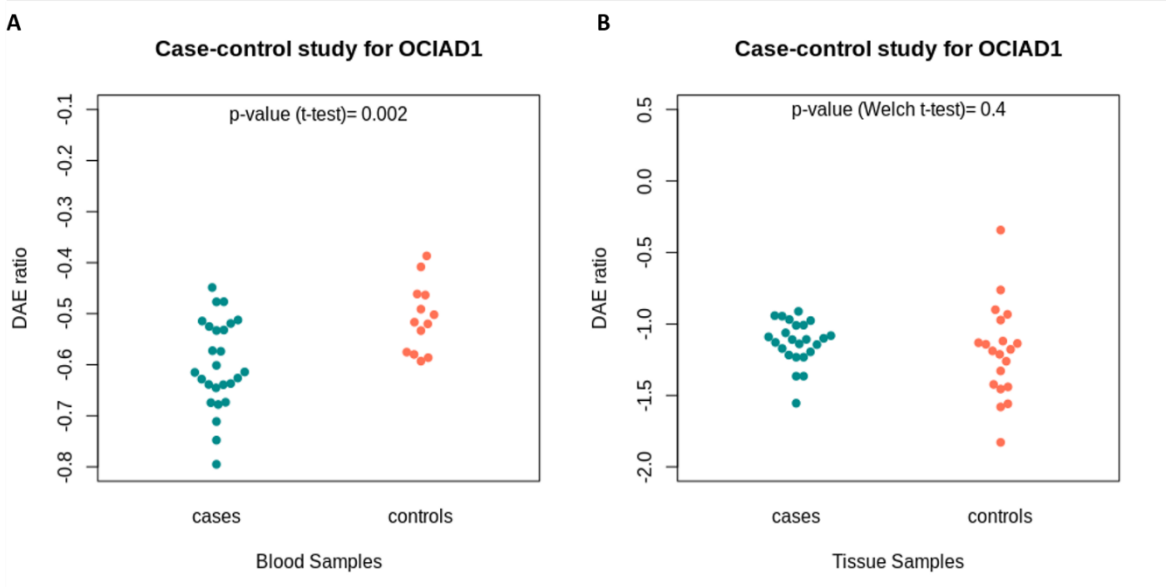


Figure 4. 16 - Case-control association results using DAE for *OCIAD1* (rs9997920) in blood and breast tissue samples. On the x-axis are represented the samples studied and on the y axis the DAE ratios. All dots represent the DAE measured for heterozygous individuals for rs9997920. Cases are shown in cyan blue and the controls in orange. In **A** are represented the DAE values obtained for OCIAD1 in blood samples (first run); In **B** are the DAE values obtained for *OCIAD1* in breast tissue samples (first run).

The case-control studies for *GRHL2* were performed using 12 NM samples from BC patients and 15 controls from breast tissue samples of healthy donors. It is important to refer that the DAE values were quantified only for 13 of the 15 controls, because two samples did not amplify. The homogeneity of variances of the DAE ratio distributions was tested and confirmed between the groups (Levene's test, p-value= 0.39 and 0.5, respectively for first and second runs). Regarding to case-control studies a significant difference between the mean DAE of the two groups was found, suggesting an association with risk for BC in the first run (t-test p-value=0.014) (**Figure 4.17**). The second run did not show the same, with the two distributions (p-value= 0.096) (**Annex L)**.

The overall results for these studies are resumed in **Table 4.4**.

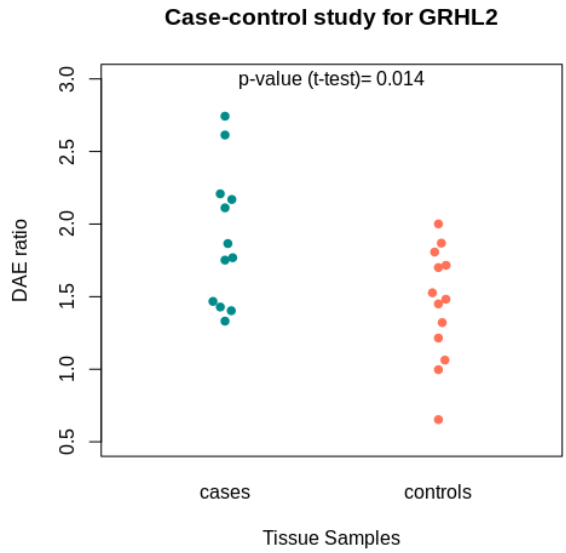


Figure 4. 17- Case-control association study using DAE for GRHL2 (rs6989650) in breast tissue samples. On the x-axis are represented the samples studied and on the y axis the DAE ratios. All dots represent the DAE measured for heterozygous individuals for rs6989650 in the first experiment. Cases are shown in cyan blue and the controls in orange.

Table 4. 4 – Results for the case-control association studies using DAE ratios

Gene	SNP	Tissue	Run	t.test (p-value)
OCIAD1	rs9997920	Blood	First	**0.002**
OCIAD1	rs9997920	Blood	Second	**0.008**
OCIAD1	rs9997920	Breast	First	0.4
OCIAD1	rs9997920	Breast	Second	0.07
GRHL2	rs6989650	Breast	First	**0.014**
GRHL2	rs6989650	Breast	Second	0.096

CHAPTER V

Discussion

5. DISCUSSION

Our aim in this study was to identify new genes associated with BC risk, via focusing on variants with cis-acting regulatory potential. We started by performing *in silico* analysis to select candidate genes to perform the case-control association studies using DAE ratios. After studying genes from GWAS, not yet associated with BC risk, that were cis-regulated in breast tissue and showed clinical impact, two genes were selected to be tested for association with BC risk. We found a significant association between BC risk and DAE values for *OCIAD1* in blood and DAE values for *GRHL2* in breast tissue, although the results from this last gene were not replicated in an independent experiment and therefore need further validation.

This innovative approach of using DAE values in a case-control study, in which the cases consisted of NM or blood from BC patients and the controls of NB tissue or blood from healthy donors, allowed the detection of differences in the effect of cis-acting variants regulating the expression of *OCIAD1* and *GRHL2,* which can help to identify women at higher risk of BC.

This association allowed the establishment that women with lower values of DAE for *OCIAD1* are at a higher risk of developing BC. Nevertheless, by how much (effect size) and the mechanism behind this association are unknown and should be studied in the future. One hypothesis is that in the genetic background of women with higher risk there could be rare variants contributing to disease development. So, to detect these risk variants it is important to have samples from the populations at risk, to characterize the differences in comparison with the healthy population, for example by doing deep-sequencing.

5.1. DAE OF *OCIAD1* IS ASSOCIATED WITH RISK

OCIAD1 encodes a protein that is overexpressed in several carcinomas (Shetty, Kalamkar, and Inamdar 2018; Sengupta et al. 2008), but its normal expression and function are not yet well known. The genetic modulation of *OCIAD1* allowed the discovery that its presence aids in maintenance of the pluripotent state, and if its levels are reduced it promotes differentiation of stem cells. This gene encodes the OCIAD1 protein that is involved in the regulation of the energetic metabolism in human pluripotent cells, in the electron transport

chain proteins and, also downregulates the oxidative phosphorylation, among other processes (Shetty, Kalamkar, and Inamdar 2018).

DAE at *OCIAD1* (measured in rs9997920) was tested in case-control association studies and significant association was verified with BC risk. The study of DAE at *OCIAD1* involved blood and breast tissue. DAE at this gene was significantly associated with risk to BC, in blood but not in breast tissue. Although these experiments were repeated twice, they will be repeated in order to strengthen our confidence in the results. One reason for repetition regarding the study in blood, is that the number of controls was not the same. So, we should repeat this experiment in the way of preserve always the same number of controls to give us even more confidence in the results. Another thing that should be improved in this experiment is to increase the number of both controls and cases, to increase the power to reach a stronger statistical significance. Relative to the *OCIAD1* experiment in breast tissue (24 cases & 22 controls), we did not find a significant association.

Regarding to significant association results obtained in blood tissue, we can infer the lower levels of *OCIAD1* DAE, measured in blood, may indicate women at higher risk of developing BC. This is in accordance with eQTL analysis, available at the GTEx database ("GTEx Portal" 2018), where in blood tissue rs9997920 is an eQTL for *OCIAD1*, but not in breast tissue. In spite of this, studies to establish the size effect have to be performed in order to identify the limits of DAE from which women are at risk and also to determine how much of the risk in the population is due to this new risk locus.

It was also observed that all of the samples displayed DAE values with a similar magnitude to what was observed before, using microarrays technology in this same tissue (see Figure 4.7B), validating the DAE values obtained for this gene in our previous study (lower expression of the minor allele T for both tSNPs studied).

The cis-acting variants regulating *OCIAD1* gene expression in breast seem to confer a similar pattern of DAE as in blood tissue but might have some differences which explain the fact that no association was detected in breast. It should be noted that the total levels of expression are also different in the two tissues. One more hypothesis is that this gene, *OCIAD1*, might be associated, for

example, with risk of metastasis in breast cancer patients. Investigators showed that *OCIAD1* is overexpressed in metastatic ovarian cancer tissues and it is important for the recurrence of disease (Sengupta et al. 2008; C. Wang et al. 2010). In our analysis the lower gene expression of this gene was correlated with higher tumour grade, which is also associated with greater propensity for metastasis. So, this gene can be conferring higher risk for breast tissue involving metastasis development too (Miki et al. 1994; Wooster et al. 1995).

It is important to validate further our findings in order to use them to identify women at risk for BC, using a DAE test in blood in the future. This is a long process, that can take decades, as the first GWAS loci identified in 2007 have not yet been included in risk testing. Additionally, we believe it is highly important to carry functional studies to fully understand the mechanism and identify the causal SNP (or SNPs) conferring the risk for BC in this locus.

5.2. *GRHL2* MIGHT BE A NOVEL RISK GENE FOR BC

Grainyhead transcription factors are highly conserved and work as key regulators of epithelial differentiation, organ development and skin barrier formation (Ming et al. 2018). *GRHL2*, a member of this family, is a wound healing regulatory transcription factor and suppresses epithelial mesenchymal transition (EMT), a process involved for invasion and metastasis. The gene *GRHL2* is considered a potential tumour suppressor gene in breast cancer (Cieply et al. 2013). Some studies demonstrated that *GRHL2* regulates epithelial alterations in pancreatic cancer progression (Nishino et al. 2017) and that suppresses metastasis in non-small cell lung cancer (Pan et al. 2017).

DAE of *GRHL2* (measured at rs6989650) was associated with risk for breast cancer in the first experiment performed, but not in the replication with the same samples (p-value= 0.014/0.096). These studies were performed in 13 controls (normal breast of healthy donors) samples and in 12 cases (NM) samples. It is important to notice that the second run had some problems. When the calibration curve was calculated, two samples had to be removed because they were not amplified. This could have influenced the results, so this experiment has

to be repeated to confirm if *GRHL2* is really a novel gene associated with BC risk or not. As for the case of *OCIAD1*, in GRHL2 we also observed a similar pattern of DAE using rs6989650, as was observed previously in the group using microarrays (Figure 4.10A), higher expression of the minor T allele.

Again, the presence of DAE indicates the existence of regulatory variants acting on gene expression, that might be the ones conferring risk. The analysed tSNP (rs6989650) was in LD, although weak, with two GWAS SNPs: rs2387620 (r^2= 0.28 e D'=0.79) and rs2211914 (r^2= 0.33 e D'=0.99) from Fletcher et al 2011 GWAS study. The GWAS SNP rs2387620 are located in *NCALD*, the neighbouring gene (16 kb downstream of *GRHL2*), which also showed DAE in our previous study. However, none of the tSNPs in *NCALD* is in LD with the GWAS SNPs. This lead us to think that although the risk variants were not in *GRHL2*, this might be the target gene by the causal variants, which we believe to have a cis-regulatory role.

Regarding the results from the differential expression analyses *GRHL2* was overexpressed in tumours vs normal, ER+, ER-, PR+, PR-, HER2 amplified and not amplified cases and in grade analyses when compared with NM samples. Other studies revealed that *GRHL2* is downregulated specifically in the claudin-low BC subtype and in basal cell lines of BC (Cieply et al. 2012) when compared with different tumour subtypes. In our analyses *GRHL2* might be overexpressed because the METABRIC project data is enriched in ER+ cases (1498 cases), which are mostly of luminal subtype, and the effect of *GRHL2* may be biased by this difference in type distribution. But we have not tested this. It is also important to refer that the METABRIC project did not have any cases of Grade 4, which has more propensity to metastasize and to suffer epithelial-mesenchymal-transition, where *GRHL2* probably will be downregulated. In the future, it would be interesting to study a subset of NM samples enriched from patients with basal tumours and claudin low subtype and verify their levels of *GRHL2* DAE.

5.3. RT-QPCR FOR QUANTIFICATION OF DAE AND CANDIDATE GENES ASSOCIATION STUDIES BENEFITS

Individual loci can be queried by RT-qPCR method in order to identify DAE and consequently identify the presence of cis-acting variants acting on genes with DAE (Bray et al. 2003). Previous reports showed that using imbalances of allelic expression in heterozygous samples allows a much greater power of detection of cis-acting variants (Yan et al. 2002; Pant et al. 2006; Bjornsson et al. 2008; Serre et al. 2008). Maia et al in 2009 showed that DAE could be used to discover cis-regulatory variants associated with BC susceptibility. To assess DAE, the allele-specific transcript levels were quantified using RT-qPCR and the ratios of one allele by the other were calculated. They showed the feasibility of using DAE in blood as quantitative trait in future genetic association studies for risk to BC (Maia et al. 2009).

In this study RT-qPCR was performed using allele-specific Taqman™ technology. It is important to notice that RT-qPCR quantification depends on the efficiency of PCR using fluorophores, so it is important to take this into account in a precise quantification. All experiments must contain a serial diluted calibration curve of heterozygous samples, in order to obtain the respective linear regression for each allele and calculate the quantity expressed by each allele from the specific curve. In our study, we verified that the allele labelled with FAM showed higher efficiency of amplification than the allele labelled with VIC, for both assays. In this work two homozygous samples were mixed in a 50:50 proportion to form the calibration curve but in the future we can just serial diluted a heterozygous sample.

As positive controls for all experiments performed, a series of standards curves with mixtures of different proportions of each allele of the SNP under study, were used to verify if we could observe the expected DAE ratios by extrapolating the allelic quantities with the previously described serial diluted curves. The accuracy was very high when the DAE values were between 2 and -2, and low for more extreme DAE values. The less the quantity of an allele the more difficult it was to observe ratios close to the expected ones. The values of the expected and the observed ratios were well correlated, so we could trust in this method for quantification of allelic expression.

So, the case-control studies were performed, and we verified that with this technique it was possible to identify genes associated with BC risk, because the ratios of DAE between cases and controls were significantly different, for example in the first run of *GRHL2* in breast tissue, and for *OCIAD1* in blood, as mentioned above.

There is at least one published study that uses allele specific expression (ASE) and case-control studies to identify risk genes. This study performed case-control association studies using samples already known to display germline ASE of *TGFBR1,* and using RT-qPCR they showed it conferred an increased risk for colorectal cancer (Valle et al. 2008). However, they used ASE as a status to compare frequency of samples with and without ASE in cases vs controls.

Our approach is different and innovative because it quantifies DAE in heterozygous samples and uses it as a quantitative measure to compare cases and controls, in opposition to current GWAS studies that use discrete variables, i.e. genotypes.

There are other options to quantify DAE to validate our results, as the Sanger sequencing (Ge et al. 2009) or by next generation sequencing using RNA-Seq approaches (Montgomery et al. 2010). The Sanger sequencing was the first method of sequencing discovered, nowadays it is mostly used for validation of DAE experiments, because it is time consuming and low-throughput. Microarrays technology was the most used solution for gene expression profiling, because it is high-throughput and relatively low-cost, however it has some disadvantages, such as rely on prior knowledge about the genome for probe design, imperfect hybridization of probes, among others (Finotello and Di Camillo 2014). The best method used nowadays is RNA-Seq, because it is a technique that captures regulatory variation both in a small number of samples as in a large number, allowing integrated analysis of the genome. However, this technique is expensive and requires an extensive analysis of data/reads obtained, and the analysis pipeline is not well established regarding how it should be used for DAE measurements (Castel et al. 2015). So, the RT-qPCR is a great option for DAE quantification because it is a simple technique, low cost and efficient for genotyping (Molicotti, Bua, and Zanetti 2014), particularly if we have a low number of candidates such as our study.

In fact, in view of our results, all the other 15 candidate risk SNPs should be studied by the same approach in the future.

5.4. RETRIEVAL OF PROXIES SNPS

GWAS association studies were used to retrieve all variants with weak association with BC risk, needing validation. The selection of the candidate risk variants was done based on the phases of each study, however a conservative criterion could be applied and only candidate risk variants present in phase II or in phase III could have been included. We wanted to capture the maximum good risk candidates, so we used risk candidates from all GWAS studies phases.

As recombination tends to occur in the human genome at preferential sites, distinct linkage disequilibrium blocks are formed, and neighbouring polymorphisms, like SNPs, are often strongly genetically correlated with each other (Kruglyak and Nickerson 2001). Therefore, the proxies SNPs for the candidate risk variants were retrieved using the *rsnps* package, as referred in the subchapter 3.2.2. As association studies were performed in different populations, the proxies SNPs were retrieved according to the study population (CEU). The *rsnps* package has some disadvantages. The reference studies used are from HapMap project and from phase 1 of the 1000 Genome Project, and nowadays data from phase 3 of 1000 Genome Project is available. In the future we should update this analysis using data just from the phase 3 of the 1000 Genomes Project.

5.5. EXPRESSION OF CIS-REGULATED GENES IN BREAST TISSUE

DAE is a hallmark of cis-regulation. Our candidate risk genes were filtered using evidence of DAE in normal breast tissue, which provides direct evidence that these genes are being cis-regulated in breast tissue. Studies proved that some genes exhibit tissue-specific expression (Zhu et al. 2016; Patient 1990). For example, *GRHL2* is highly expressed in breast tissue but less expressed in blood ("GTEx Portal" 2018), so we could not test its DAE also in blood.

DAE data from previous analyses in the group, as mentioned in subchapter 3.3, was defined as the heterozygous allelic expression ratio in cDNA, normalized by the heterozygous allelic ratio in gDNA. A sample was considered to have DAE if the $\log_2$ of the allelic expression ratios were greater than 0.58 or smaller than -0.58. SNPs were considered to display DAE when at least 10% of the heterozygotes and four different samples displayed DAE (Xavier et al. 2016). In that study, the data was normalized by gDNA because to serve as a reference for equal levels of the two transcripts alleles. This assumes that any technical bias included in DAE measurement for the cDNA is the same for the gDNA (Xiao and Scott 2011).

In our study, DAE quantification was performed without considering the gDNA, because we did not have gDNA for all samples. But our standard curves show that our calculations were accurate and allowed us to correct for technical bias. However, one difference is that when correcting for gDNA we are also removing the effects of possible CNV (copy number variants) affecting gene expression. Therefore, our DAE measures might be affected by CNVs, and not only by cis-regulatory variants, what might be useful if it is this join effect that impact breast cancer risk. Either way, this seems not to be the case for both *OCIAD1* and *GRHL2,* since the DAE values measured in this work were in agreement with the values obtained before with the microarray analysis (Xavier et al. 2016), indicating that copy number effect is not notably in these genes.

5.6. DIFFERENTIAL EXPRESSION ANALYSES

Whole genome statistical association analyses between gene expression and clinical variables were performed with the goal of generating an easy consulting database of results for Professor Ana Teresa Maia's group. In the Curtis et al work, the investigators performed DE analyses taking into account (co-factor in regression model) the samples origin (Vancouver or Cambridge) (Curtis et al. 2012). However, since we did not have access to this metadata information, we could not take this factor into account in our analysis. In the future we should request this information to be able to remove/adjust the analyses for population effects. Another DE analyses could have been done, like considering PAM50 classification, stage among others.

The R package *limma* was used in our analysis as well as in the Curtis et al study. This package is one of the most used for microarray analyses and has been recently updated by increasing the robustness of the hyperparameter estimation procedure (Phipson et al. 2016). In the original method, used in Curtis and in our analyses, genewise linear models are fitted to the log-expression values and the variances are extracted. With the new approach, a special attention to variances is given when they are exceptionally large or small. The genes corresponding to extreme variances will be indicated as outliers (genes with large variances: "hypervariable genes"). It was demonstrated that with this new approach in certain genomic datasets, a small number of outlier genes can have an unwanted influence on the estimation of variances, decreasing the effectiveness of the empirical Bayes' differential expression approach (Phipson et al. 2016). Therefore, analyses using this robust parameter could be performed in the future in the way of comparing if the identified differential expressed genes with the first method have a substantial difference when compared to the ones obtained with the new approach.

In this work the differential expression analyses helped us to prioritize the genes for validation in the laboratory. The genes have to be differentially expressed in at least one clinical variable. The two candidate genes selected (*OCIAD1* and *GRHL2*) were associated with clinical impact as referred above. The threshold defined for the biological magnitude (FC) was of 1.5 because we seek to identify novel genes associated with BC risk, and not genes that have huge differences in expression. Other studies, like of Patterson et al 2006, used FC thresholds of 1.5, 2 or 4 to satisfy a statistical significance level of p-value < 0.01 or < 0.05 to identify differential expressed genes (Patterson et al. 2006). In other studies, it was defined that genes are differentially expressed if they show a FC of at least 1.5 and also satisfy a criterion of p-value<0.05, after adjustment for multiple testing (Peart et al. 2005; Raouf et al. 2008). Thus the parameters used in our clinical association analyses (FC of 1.5 and p-value of 0.01 after FDR correction) are in concordance to published studies.

The *limma* package is also suitable for RNA-Seq analyses (Ritchie et al. 2015). It would be interesting to perform these analyses using, for example data from the Tumour Cancer Genome Atlas (TCGA),and verify what is the percentage of concordance between the differentially expressed genes obtained using the two

different datasets (that have different percentages of tumour subtypes). The METABRIC project does not have available whole genome expression data obtained by RNA-Seq technology, just data for some targeted genes (Bernard et al 2016).

CHAPTER VI

Conclusions and Future Perspectives

6. CONCLUSIONS AND FUTURE PERSPECTIVES

In conclusion, we showed the cis-regulated genes *OCIAD1* and *GRHL2,* are candidate risk loci for BC, and have clinical impact. We also show that case-control association studies using DAE values is a powerful approach to identify new genes associated with BC risk and supports the use of this approach in more candidate genes ou even genome-wide

In silico functional analysis should follow in order to detect the cis-variant(s) affecting gene expression. To date, the underlying mechanisms by which cis-acting genetic variation impacts gene expression differences and disease risk remains poorly understood. These cis-acting variants might be affecting binding of TFs, binding of micro RNA (miRNA) or altering the splicing mechanism.

Our work contributed to the improvement of the understanding of BC aetiology. Since we identified novel genes associated with BC risk, and one of them is a risk biomarker in blood tissue, this gene is an easy candidate to be incorporated in future blood genetic tests. The identification of risk genes for BC remains emergent since the identification of women at higher risk and their improved management is of increase importance to decrease aggressiveness of disease at diagnosis and improve survival

CHAPTER VII

References

7. REFERENCES

Albert, Frank W., and Leonid Kruglyak. 2015. "The Role of Regulatory Variation in Complex Traits and Disease." *Nature Reviews Genetics* 16 (4). Nature Publishing Group: 197–212. doi:10.1038/nrg3891.

Allen-Brady, Kristina, and Nicola J. Camp. 2005. "Characterization of the Linkage Disequilibrium Structure and Identification of Tagging-SNPs in Five DNA Repair Genes." *BMC Cancer* 5 (Ld): 1–10. doi:10.1186/1471-2407-5-99.

Almlöf, Jonas Carlsson, Per Lundmark, Anders Lundmark, Bing Ge, Seraya Maouche, Harald H.H. Göring, Ulrika Liljedahl, et al. 2012. "Powerful Identification of Cis-Regulatory SNPs in Human Primary Monocytes Using Allele-Specific Gene Expression." *PLoS ONE* 7 (12). doi:10.1371/journal.pone.0052260.

Altshuler, David L., Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Francis S. Collins, et al. 2010. "A Map of Human Genome Variation from Population-Scale Sequencing." *Nature* 467 (7319): 1061–73. doi:10.1038/nature09534.

Anderson, William F., Nilanjan Chatterjee, William B. Ershler, and Otis W. Brawley. 2002. "Estrogen Receptor Breast Cancer Phenotypes in the Surveillance, Epidemiology, and End Results Database." *Breast Cancer Research and Treatment* 76 (1): 27–36. doi:10.1023/A:1020299707510.

Antoniou, A., P.D.P. Pharoah, S. Narod, H.A. Risch, J.E. Eyfjord, J.L. Hopper, N. Loman, et al. 2003. "Average Risks of Breast and Ovarian Cancer Associated with BRCA1 or BRCA2 Mutations Detected in Case Series Unselected for Family History: A Combined Analysis of 22 Studies." *The American Journal of Human Genetics* 72 (5): 1117–30. doi:10.1086/375033.

Apostolou, P, and F Fostira. 2013. "Hereditary Breast Cancer: The Era of New Susceptibility Genes." *Biomed Res Int* 2013: 747318. doi:10.1155/2013/747318.

B. L. Welch, B.A. 1947. "THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARIANCES ARE INVOLVED." *Biometrika Trust* 34: 28–35.

Balding, David J. 2006. "A Tutorial on Statistical Methods for Population Association Studies." *Nature Reviews Genetics* 7 (10): 781–91. doi:10.1038/nrg1916.

Barbosa-Morais, Nuno L., Mark J. Dunning, Shamith A. Samarajiwa, Jeremy F.J. Darot, Matthew E. Ritchie, Andy G. Lynch, and Simon Tavaré. 2010. "A Re-Annotation Pipeline for Illumina BeadArrays: Improving the Interpretation of Gene Expression Data." *Nucleic Acids Research* 38 (3). doi:10.1093/nar/gkp942.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society B*. doi:10.2307/2346101.

Benjamini, Yoav, and Daniel Yekutieli. 2001. “The Control of the False Discovery Rate in Multiple Testing under Dependency.” *Annals of Statistics* 29 (4): 1165–88. doi:10.1214/aos/1013699998.

Bjornsson, Hans T, Thomas J Albert, Christine M Ladd-acosta, Roland D Green, Michael A Rongione, Christina M Middle, Rafael A Irizarry, Karl W Broman, and Andrew P Feinberg. 2008. “SNP-Specific Array-Based Allele-Specific Expression Analysis,” 771–79. doi:10.1101/gr.073254.107.1.

Boice Jr., J D, C E Land, R E Shore, J E Norman, and M Tokunaga. 1979. “Risk of Breast Cancer Following Low-Dose Radiation Exposure.” *Radiology* 131 (3): 589–97.

Bray, Nicholas J, Paul R Buckland, Michael J Owen, and Michael C O’Donovan. 2003. “Cis-Acting Variation in the Expression of a High Proportion of Genes in Human Brain.” *Human Genetics* 113 (2): 149–53. doi:10.1007/s00439-003-0956-y.

Brem, Rachel B., Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. 2002. “Genetic Dissection of Transcriptional Regulation in Budding Yeast.” *Science* 296 (5568): 752–55. doi:10.1126/science.1069516.

Calle, E. E., C. W. Heath, H. L. Miracle-McMahill, R. J. Coates, J. M. Liff, S. Franceschi, R. Talamini, et al. 1996. “Breast Cancer and Hormonal Contraceptives: Collaborative Reanalysis of Individual Data on 53 297 Women with Breast Cancer and 100 239 Women without Breast Cancer from 54 Epidemiological Studies.” *Lancet* 347 (9017): 1713–27. doi:10.1016/S0140-6736(96)90806-5.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. “Tools and Best Practices for Data Processing in Allelic Expression Analysis.” *Genome Biology* 16 (1). Genome Biology: 1–12. doi:10.1186/s13059-015-0762-6.

Causeway, Worts. 2004. “CHEK2*1100delC and Susceptibility to Breast Cancer: A Collaborative Analysis Involving 10,860 Breast Cancer Cases and 9,065 Controls from 10 Studies.” *The American Journal of Human Genetics* 74 (6): 1175–82. doi:10.1086/421251.

Cerami, Ethan, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, et al. 2012. “The CBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data.” *Cancer Discovery* 2 (5). American Association for Cancer Research: 401–4. doi:10.1158/2159-8290.CD-12-0095.

Chamberlain, Scott, Kevin Ushey, and Hao Zhu. 2016. “Rsnps: Get ‘SNP’ (‘Single-Nucleotide’ ’Polymorphism’) Data on the Web Version 2.0.” Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/rsnps/index.html.

Chen, Hanbo, and Paul C. Boutros. 2011. “VennDiagram: A Package for the Generation of Highly-Customizable Venn and Euler Diagrams in R.” *BMC Bioinformatics* 12 (1). BioMed Central Ltd: 35. doi:10.1186/1471-2105-12-35.

Cheung, V.G., and R.S. Spielman. 2009. “Genetics of Human Gene Expression:

Mapping DNA Variants That Influence Gene Expression." *Nat Rev Genet* 10 (9): 595–604. doi:10.1038/nrg2630.Genetics.

Cheung, Vivian G., Laura K. Conlin, Teresa M. Weber, Melissa Arcaro, Kuang Yu Jen, Michael Morley, and Richard S. Spielman. 2003. "Natural Variation in Human Gene Expression Assessed in Lymphoblastoid Cells." *Nature Genetics* 33 (3): 422–25. doi:10.1038/ng1094.

Cheung, Vivian G, Richard S Spielman, Kathryn G Ewens, Teresa M Weber, Michael Morley, and Joshua T Burdick. 2005. "Mapping Determinants of Human Gene Expression by Regional and Genome-Wide Association." *Nature* 437 (7063): 1365–69. doi:10.1038/nature04244.Mapping.

Cieply, Benjamin, Joshua Farris, James Denvir, Heide Ford, and Steven M. Frisch. 2013. "Epithelial-Mesenchymal Transition and Tumor Suppression Are Controlled by a Reciprocal Feedback Loop between ZEB1 and Grainyhead-like-2." *Cancer Research* 73 (20): 6299–6309. doi:10.1021/nl061786n.Core-Shell.

Cieply, Benjamin, Philip Riley IV, Phillip M. Pifer, Joseph Widmeyer, Joseph B. Addison, Alexey V. Ivanov, James Denvir, and Steven Frisch M. 2012. "SUPPRESSION OF THE EPITHELIAL-MESENCHYMAL TRANSITION BY GRAINYHEAD-LIKE-2." *Cancer R* 72 (9): 2440–53. doi:10.1007/s10555-012-9359-7.Regulation.

Claus, E B, N Risch, and W D Thompson. 1991. "Genetic Analysis of Breast Cancer in the Cancer and Steroid Hormone Study." *American Journal of Human Genetics* 48 (2): 232–42. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1683001&tool=pmcentrez&rendertype=abstract.

Colditz, Graham A. 2007. "Decline in Breast Cancer Incidence Due to Removal of Promoter: Combination Estrogen plus Progestin." *Breast Cancer Research* 9 (4): 3–5. doi:10.1186/bcr1736.

Consortium, International Multiple Sclerosis Genetics, Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL de Bakker, et al. 2007. "Risk Alleles for Multiple Sclerosis Identified by a Genomewide Study." *The New England Journal of Medicine* 357 (851–862): 1–3. doi:10.1056/NEJMp1002530.

Corradin, Olivia, and Peter C. Scacheri. 2014. "Enhancer Variants: Evaluating Functions in Common Disease." *Genome Medicine* 6 (10): 1–14. doi:10.1186/s13073-014-0085-3.

Couch, Fergus J, Xianshu Wang, Lesley McGuffog, Andrew Adam Lee, Curtis Olswold, Karoline B Kuchenbaecker, Penny Soucy, et al. 2013. "Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk." *PLoS Genet* 9 (3): e1003212. doi:10.1371/journal.pgen.1003212\rPGENETICS-D-12-02260 [pii].

Cox, Angela, Alison M. Dunning, Montserrat Garcia-Closas, Sabapathy Balasubramanian, Malcolm W.R. Reed, Karen A. Pooley, Serena Scollen, et al. 2007. "A Common Coding Variant in CASP8 Is Associated with Breast Cancer Risk." *Nature Genetics* 39 (3): 352–58. doi:10.1038/ng1981.

"CRAN - Package Beeswarm." 2018. Accessed September 23. https://cran.r-project.org/web/packages/beeswarm/index.html.

Curtis, Christina, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, et al. 2012. "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups." *Nature*, 1–7. doi:10.1038/nature10983.

Dent, Rebecca, Maureen Trudeau, Kathleen I. Pritchard, Wedad M. Hanna, Harriet K. Kahn, Carol A. Sawka, Lavina A. Lickley, Ellen Rawlinson, Ping Sun, and Steven A. Narod. 2007. "Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence." *Clinical Cancer Research* 13 (15): 4429–34. doi:10.1158/1078-0432.CCR-06-3045.

Dunning, Alison M, Catherine S Healey, Caroline Baynes, Ana-teresa Maia, Nuno L Barbosa-morais, Bruce A J Ponder, Ana Vega, et al. 2009. "Association of ESR1 Gene Tagging SNPs with Breast Cancer Risk." *Human Molecular Genetics* 18 (6): 1131–39. doi:10.1093/hmg/ddn429.

Dunning, M, A Lynch, and M Eldridge. 2015. "IlluminaHumanv3.Db: Illumina HumanHT12v3 Annotation Data (Chip IlluminaHumanv3)." *R Package Version 1.26.0*.

Durinck, Steffen, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. 2005. "BioMart and Bioconductor: A Powerful Link between Biological Databases and Microarray Data Analysis." *Bioinformatics* 21 (16): 3439–40. doi:10.1093/bioinformatics/bti525.

Durinck, Steffen, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. 2009. "Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package BiomaRt." *Nature Protocols* 4 (8): 1184–1191. doi:10.1038/nprot.2009.97.

Easton, Douglas F., Karen A. Pooley, Alison M. Dunning, Paul D. P. Pharoah, Deborah Thompson, Dennis G. Ballinger, Jeffery P. Struewing, et al. 2007. "Genome-Wide Association Study Identifies Novel Breast Cancer Susceptibility Loci." *Nature* 447 (7148): 1087–93. doi:10.1038/nature05887.

Easton, Douglas F. 1999. "How Many More Breast Cancer Predisposition Genes Are There?" *Breast Cancer Research* 1 (1): 14. doi:10.1186/bcr6.

Eccles, Suzanne A., Eric O. Aboagye, Simak Ali, Annie S. Anderson, Jo Armes, Fedor Berditchevski, Jeremy P. Blaydes, et al. 2013. "Critical Research Gaps and Translational Priorities for the Successful Prevention and Treatment of Breast Cancer." *Breast Cancer Research* 15 (5): 1. doi:10.1186/bcr3493.

Efron, Bradley. 2003. "Robbins, Empirical Bayes and Microarrays." *Annals of Statistics* 31 (2): 366–78. doi:10.1214/aos/1051027871.

Efron, Bradley, Robert Tibshirani, John D. Storey, and Virginia Tusher. 2001. "Empirical Bayes Analysis of a Microarray Experiment." *Journal of the American Statistical Association* 96 (456): 1151–60. doi:10.1198/016214501753382129.

Ewertz, Marianne, Stephen W. Duffy, Hans□Olov □O Adami, Gunnar Kvåle, Eiliv Lund, Olav Meirik, Anders Mellemgaard, Irma Soini, and Hrafn Tulinius.

1990. “Age at First Birth, Parity and Risk of Breast Cancer: A Meta□analysis of 8 Studies from the Nordic Countries.” *International Journal of Cancer* 46 (4): 597–603. doi:10.1002/ijc.2910460408.

Ferlay, Jacques, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. 2015. “Cancer Incidence and Mortality Worldwide : Sources, Methods and Major Patterns in GLOBOCAN 2012.” *International Journal of Cancer* 136. doi:10.1002/ijc.29210.

Finotello, Francesca, and Barbara Di Camillo. 2014. “Measuring Differential Gene Expression with RNA-Seq: Challenges and Strategies for Data Analysis.” *Briefings in Functional Genomics* 14 (2): 130–42. doi:10.1093/bfgp/elu035.

Fletcher, Olivia, Nichola Johnson, Nick Orr, Fay J. Hosking, Lorna J. Gibson, Kate Walker, Diana Zelenika, et al. 2011. “Novel Breast Cancer Susceptibility Locus at 9q31.2: Results: Of a Genome-Wide Association Study.” *Journal of the National Cancer Institute* 103 (5): 425–35. doi:10.1093/jnci/djq563.

Forjaz de Lacerda, Gonçalo, Scott P. Kelly, Joana Bastos, Clara Castro, Alexandra Mayer, Angela B. Mariotto, and William F. Anderson. 2018. “Breast Cancer in Portugal: Temporal Trends and Age-Specific Incidence by Geographic Regions.” *Cancer Epidemiology* 54 (October 2017): 12–18. doi:10.1016/j.canep.2018.03.003.

Gabriel, Stacey B, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Brendan Blumenstiel, John Higgins, Matthew Defelice, et al. 2002. “The Structure of Haplotype Blocks in the Human Genome.” *Science* 296 (5576): 2225–29.

Gao, Jianjiong, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, et al. 2013. “Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the CBioPortal.” *Science Signaling* 6 (269). American Association for the Advancement of Science: pl1. doi:10.1126/scisignal.2004088.

Garibyan, Lilit, and Nidhi Avashia. 2013. “Polymerase Chain Reaction.” *Journal of Investigative Dermatology* 133 (3). Elsevier Masson SAS: 1–4. doi:10.1038/jid.2013.1.

Ge, Bing, Dmitry K. Pokholok, Tony Kwan, Elin Grundberg, Lisanne Morcos, Dominique J. Verlaan, Jennie Le, et al. 2009. “Global Patterns of Cis Variation in Human Cells Revealed by High-Density Allelic Expression Analysis.” *Nature Genetics* 41 (11). Nature Publishing Group: 1216–22. doi:10.1038/ng.473.

Ghoussaini, Maya, Paul D.P. Pharoah, and Douglas F. Easton. 2013. “Inherited Genetic Susceptibility to Breast Cancer: The Beginning of the End or the End of the Beginning?” *American Journal of Pathology* 183 (4). American Society for Investigative Pathology: 1038–51. doi:10.1016/j.ajpath.2013.07.003.

Gilad, Yoav, Scott A. Rifkin, and Jonathan K. Pritchard. 2008. “Revealing the Architecture of Gene Regulation: The Promise of EQTL Studies.” *Trends in Genetics* 24 (8): 408–15. doi:10.1016/j.tig.2008.06.001.Revealing.

Gómez-Flores-Ramos, Liliana, Rosa María Álvarez-Gómez, Cynthia Villarreal-

Garza, Talia Wegman-Ostrosky, and Alejandro Mohar. 2017. “Breast Cancer Genetics in Young Women: What Do We Know?” *Mutation Research/Reviews in Mutation Research* 774 (22). Elsevier: 33–45. doi:10.1016/j.mrrev.2017.08.001.

Gonzalez, Kelly D., Katie A. Noltner, Carolyn H. Buzin, Dongqing Gu, Cindy Y. Wen-Fong, Vu Q. Nguyen, Jennifer H. Han, et al. 2009. “Beyond Li Fraumeni Syndrome: Clinical Characteristics of Families with P53 Germline Mutations.” *Journal of Clinical Oncology* 27 (8): 1250–56. doi:10.1200/JCO.2008.16.6959.

Göring, Harald H.H., Joanne E. Curran, Matthew P. Johnson, Thomas D. Dyer, Jac Charlesworth, Shelley A. Cole, Jeremy B.M. Jowett, et al. 2007. “Discovery of Expression QTLs Using Large-Scale Transcriptional Profiling in Human Lymphocytes.” *Nature Genetics* 39 (10): 1208–16. doi:10.1038/ng2119.

“GTEx Portal.” 2018. Accessed September 23. https://gtexportal.org/home/.

Guo, Qi, Marjanka K. Schmidt, Peter Kraft, Sander Canisius, Constance Chen, Sofia Khan, Jonathan Tyrer, et al. 2015. “Identification of Novel Genetic Markers of Breast Cancer Survival.” *Journal of the National Cancer Institute* 107 (5): 1–9. doi:10.1093/jnci/djv081.

Gusev, Alexander, S Hong Lee, Benjamin M Neale, Gosia Trynka, B. J. Vilhjalmsson, H. Finucane, H. Xu, et al. 2014. “Regulatory Variants Explain Much More Heritability than Coding Variants across 11 Common Diseases.” *Genomics*, no. Ld: 0–21. doi:10.1101/004309.

Hamajima, N., K. Hirose, K. Tajima, T. Rohan, E. E. Calle, C. W. Heath, R. J. Coates, et al. 2002. “Alcohol, Tobacco and Breast Cancer - Collaborative Reanalysis of Individual Data from 53 Epidemiological Studies, Including 58 515 Women with Breast Cancer and 95 067 Women without the Disease.” *British Journal of Cancer* 87 (11): 1234–45. doi:10.1038/sj.bjc.6600596.

Hearle, Nicholas, Valérie Schumacher, Fred H. Menko, Sylviane Olschwang, Lisa A. Boardman, Johan J.P. Gille, Josbert J. Keller, et al. 2006. “Frequency and Spectrum of Cancers in the Peutz-Jeghers Syndrome.” *Clinical Cancer Research* 12 (10): 3209–15. doi:10.1158/1078-0432.CCR-06-0083.

Heid, CA, K Livak, J Stevens, and PM Williams. 1996. “Real Time Quantitative PCR.” *Genome Research* 6: 986–94. doi:10.1101/gr.6.10.986.

Hill, W. G., and Alan Robertson. 1968. “Linkage Disequilibrium in Finite Populations.” *TAG Theoretical and Applied Genetics* 38 (6): 226–31. doi:10.1007/bf01245622.

Holland, Pamela M, Richard D Abramson, Robert Watson, and David H Gelfand. 1991. “Detection of Specific Polymerase Chain Reaction Product by Utilizing the 5’ → 3’ Exonuclease Activity of Thermus Aquaticus DNA Polymerase.” *Proceedings of the National Academy of Sciences of the United States of America* 88 (16): 7276. doi:10.1073/pnas.88.16.7276.

“Home - SNP - NCBI.” 2018. Accessed September 23. https://www.ncbi.nlm.nih.gov/snp.

Hu, Zhiyuan, Cheng Fan, Daniel S. Oh, J. S. Marron, Xiaping He, Bahjat F. Qaqish,

Chad Livasy, et al. 2006. “The Molecular Portraits of Breast Tumors Are Conserved across Microarray Platforms.” *BMC Genomics* 7: 1–12. doi:10.1186/1471-2164-7-96.

Hunter, D J, and W C Willett. 1993. “Diet, Body Size, and Breast Cancer.” *Epidemiologic Reviews* 15 (1): 110–32. http://www.ncbi.nlm.nih.gov/pubmed/8405195.

“ICOGS Complete Summary Results.” 2018. Accessed June 12. http://bcac.ccge.medschl.cam.ac.uk/bcacdata/icogs-complete-summary-results/.

Innis, M. A., K. B. Myambo, D. H. Gelfand, and M. A. Brow. 1988. “DNA Sequencing with Thermus Aquaticus DNA Polymerase and Direct Sequencing of Polymerase Chain Reaction-Amplified DNA.” *Proceedings of the National Academy of Sciences* 85 (24): 9436–40. doi:10.1073/pnas.85.24.9436.

Iwase, Hirotaka, Junichi Kurebayashi, Hitoshi Tsuda, Tomohiko Ohta, Masafumi Kurosumi, Kazuaki Miyamoto, Yutaka Yamamoto, and Takuji Iwase. 2010. “Clinicopathological Analyses of Triple Negative Breast Cancer Using Surveillance Data from the Registration Committee of the Japanese Breast Cancer Society.” *Breast Cancer* 17 (2): 118–24. doi:10.1007/s12282-009-0113-0.

Jain, Nitin, Jayant Thatte, Thomas Braciale, Klaus Ley, Michael O’Connell, and Jae K. Lee. 2003. “Local-Pooled-Error Test for Identifying Differentially Expressed Genes with a Small Number of Replicated Microarrays.” *Bioinformatics* 19 (15): 1945–51. doi:10.1093/bioinformatics/btg264.

Johnson, G C, L Esposito, B J Barratt, a N Smith, J Heward, G Di Genova, H Ueda, et al. 2001. “Haplotype Tagging for the Identification of Common Disease Genes.” *Nature Genetics* 29 (october): 233–37. doi:10.1038/ng1001-233.

Kangelaris, Kirsten N., and Stephen B. Gruber. 2007. “Clinical Implications of Founder and Recurrent CDH1 Mutations in Hereditary Diffuse Gastric Cancer.” *Journal of the American Medical Association* 297 (21): 2410–11. doi:10.1001/jama.297.21.2410.

Kerr, M. Kathleen, Martin Martin, and Gary A. Churchill. 2000. “Analysis of Variance for Gene Expression Microarray Data.” *J. Comput. Biol.* 7(6) (6): 819.

Key, Timothy J., Naomi E. Allen, Elizabeth A. Travis, and Spencer and Ruth C. 2003. “Nutrition and Breast Cancer.” *The Breast* 33 (1): 412–16. doi:10.3109/09637487909143346.

Key, Timothy J, Pia K Verkasalo, and Emily Banks. 2001. “Epidemiology of Breast Cancer” 44 (0): 133–40. doi:10.1016/S1470-2045(00)00254-0.

Kim, Tae Kyun. 2015. “T-Test as a Parametric Statistic.” *Korean Journal of Anesthesiology* 68 (6): 540–46. doi:10.4097/kjae.2015.68.6.540.

King, M.-C. 2003. “Breast and Ovarian Cancer Risks Due to Inherited Mutations in BRCA1 and BRCA2.” *Science* 302 (5645): 643–46. doi:10.1126/science.1088759.

Koboldt, Daniel C., Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, et al. 2012. "Comprehensive Molecular Portraits of Human Breast Tumours." *Nature* 490 (7418): 61–70. doi:10.1038/nature11412.

Kruglyak, Leonid, and Deborah A. Nickerson. 2001. "Variation Is the Spice of Life." *Nature Genetics* 27 (3): 234–36. doi:10.1038/85776.

Kruskal, William H., and W. Allen Wallis. 1952. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association* 47 (260). Taylor & Francis, Ltd.American Statistical Association: 583. doi:10.2307/2280779.

Lappalainen, Ilkka, Jeff Almeida-King, Vasudev Kumanduri, Alexander Senf, John Dylan Spalding, Saif Ur-Rehman, Gary Saunders, et al. 2015. "The European Genome-Phenome Archive of Human Data Consented for Biomedical Research." *Nature Genetics* 47 (7): 692–95. doi:10.1038/ng.3312.

Layde, Peter M., Linda A. Webster, Andrew L. Baughman, Phyllis A. Wingo, George L. Rubin, and Howard W. Ory. 1989. "The Independent Associations of Parity, Age at First Full Term Pregnancy, and Duration of Breastfeeding with the Risk of Breast Cancer." *Journal of Clinical Epidemiology* 42 (10): 963–73. doi:10.1016/0895-4356(89)90161-3.

Levene, H, II Olkin, and H Hotelling. 1960. "Robust Tests for Equality of Variances." *Contributions to Probability and Statistics; Essays in Honor of Harold Hotelling*, no. November 2013: 78–92.

Li, Jing, Clifford Yen, Danny Liaw, Katrina Podsypanina, Shikha Bose, Steven I. Wang, Janusz Puc, et al. 1997. "PTEN, a Putative Protein Tyrosine Phosphatase Gene Mutated in Human Brain, Breast,and Prostate Cancer." *Science* 275: 1943–47. doi:10.1126/science.275.5308.1943.

Livak, Kenneth J., and Thomas D. Schmittgen. 2001. "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2-ΔΔCT Method." *Methods* 25 (4): 402–8. doi:10.1006/meth.2001.1262.

Lonnstedt, Ingrid, and Terry Speed. 2002. "Replicated Microarray Data." *Statistical Sinica* 12 (12): 31–46.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. "The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog)." *Nucleic Acids Research* 45 (D1): D896–901. doi:10.1093/nar/gkw1133.

Maia, Ana-Teresa, Inmaculada Spiteri, Alvin J X Lee, Martin O'Reilly, Linda Jones, Carlos Caldas, and Bruce A J Ponder. 2009. "Extent of Differential Allelic Expression of Candidate Breast Cancer Genes Is Similar in Blood and Breast." *Breast Cancer Research : BCR* 11 (6): R88. doi:10.1186/bcr2458.

Malhotra, Gautam K., Xiangshan Zhao, Hamid Band, and Vimla Band. 2010. "Histological, Molecular and Functional Subtypes of Breast Cancers." *Cancer Biology and Therapy* 10 (10): 955–60. doi:10.4161/cbt.10.10.13879.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the

Missing Heritability of Complex Diseases." *Nature* 461 (7265). Nature Publishing Group: 747–53. doi:10.1038/nature08494.

Meyer, Kerstin B., Ana Teresa Maia, Martin O'Reilly, Andrew E. Teschendorff, Suet Feung Chin, Carlos Caldas, and Bruce A.J. Ponder. 2008. "Allele-Specific up-Regulation of FGFR2 Increases Susceptibility to Breast Cancer." *PLoS Biology* 6 (5): 1098–1103. doi:10.1371/journal.pbio.0060108.

Michailidou, Kyriaki, Jonathan Beesley, Sara Lindstrom, Sander Canisius, Joe Dennis, Michael J. Lush, Mel J. Maranian, et al. 2015. "Genome-Wide Association Analysis of More than 120,000 Individuals Identifies 15 New Susceptibility Loci for Breast Cancer." *Nature Genetics* 47 (4): 373–80. doi:10.1038/ng.3242.

Michailidou, Kyriaki, Per Hall, Anna Gonzalez-Neira, Maya Ghoussaini, Joe Dennis, Roger L Milne, Marjanka K Schmidt, et al. 2013. "Large-Scale Genotyping Identifies 41 New Loci Associated with Breast Cancer Risk." *Nature Genetics* 45 (4): 353–61. doi:10.1038/ng.2563.

Miki, Yoshio, Jeff Swensen, Donna Shattuck-eidens, P Andrew Futreal, Keith Harshman, Sean Tavtigian, Qingyun Liu, et al. 1994. "Strong Candidate for the Breast and Ovarian Cancer Susceptibility Gene BRCA1." *Science* 266: 66–71. doi:10.1126/science.7545954.

Milani, Lili, Anders Lundmark, Jessica Nordlund, Lili Milani, Anders Lundmark, Jessica Nordlund, Anna Kiialainen, et al. 2009. "Reveal Regulation of Gene Expression by CpG Site Methylation Allele-Specific Gene Expression Patterns in Primary Leukemic Cells Reveal Regulation of Gene Expression by CpG Site Methylation," 1–11. doi:10.1101/gr.083931.108.

Ming, Qianqian, Yvette Roske, Anja Schuetz, Katharina Walentin, Ibraim Ibraimi, Kai M. Schmidt-Ott, and Udo Heinemann. 2018. "Structural Basis of Gene Regulation by the Grainyhead/CP2 Transcription Factor Family." *Nucleic Acids Research* 46 (4). Oxford University Press: 2082–95. doi:10.1093/nar/gkx1299.

Molicotti, Paola, Alessandra Bua, and Stefania Zanetti. 2014. "Cost-Effectiveness in the Diagnosis of Tuberculosis: Choices in Developing Countries." *Journal of Infection in Developing Countries* 8 (1): 24–38. doi:10.3855/jidc.3295.

Monks, S A, A Leonardson, H Zhu, P Cundiff, P Pietrusiak, S Edwards, J W Phillips, A Sachs, and E E Schadt. 2004. "Genetic Inheritance of Gene Expression in Human Cell Lines." *American Journal of Human Genetics* 75 (6): 1094–1105. doi:10.1086/426461.

Montgomery, S B, M Sammeth, M Gutierrez-Arcelus, R P Lach, C Ingle, J Nisbett, R Guigo, and E T Dermitzakis. 2010. "Transcriptome Genetics Using Second Generation Sequencing in a Caucasian Population." *Nature* 464 (7289): 773–77. doi:10.1038/nature08903.Transcriptome.

Morris, Gloria J., Sashi Naidu, Allan K. Topham, Fran Guiles, Yihuan Xu, Peter McCue, Gordon F. Schwartz, et al. 2007. "Differences in Breast Carcinoma Characteristics in Newly Diagnosed African-American and Caucasian Patients: A Single-Institution Compilation Compared with the National Cancer Institute's Surveillance, Epidemiology, and End Results Database." *Cancer*

110 (4): 876–84. doi:10.1002/cncr.22836.

Murphy, Lizabeth Deutsch, Cynthia Elaine Herzog, Jonathan Brett Rudick, Antonio Tito Fojo, and Susan Elaine Bates. 1990. “Use of the Polymerase Chain Reaction in the Quantitation of Mdr-1 Gene Expression.” *Biochemestry* 29: 10351–56.

Natrajan, Rachael, Heba Sailem, Faraz K. Mardakheh, Mar Arias Garcia, Christopher J. Tape, Mitch Dowsett, Chris Bakal, and Yinyin Yuan. 2016. “Microenvironmental Heterogeneity Parallels Breast Cancer Progression: A Histology–Genomic Integration Analysis.” *PLoS Medicine* 13 (2): 1–19. doi:10.1371/journal.pmed.1001961.

Newman, B, M a Austin, M Lee, and M C King. 1988. “Inheritance of Human Breast Cancer: Evidence for Autosomal Dominant Transmission in High-Risk Families.” *Proceedings of the National Academy of Sciences of the United States of America* 85 (9): 3044–48. doi:10.1073/pnas.85.9.3044.

Nishino, Hitoe, Shigetsugu Takano, Hideyuki Yoshitomi, Kensuke Suzuki, Shingo Kagawa, Reiri Shimazaki, Hiroaki Shimizu, Katsunori Furukawa, Masaru Miyazaki, and Masayuki Ohtsuka. 2017. “Grainyhead-like 2 (GRHL2) Regulates Epithelial Plasticity in Pancreatic Cancer Progression.” *Cancer Medicine* 6 (11): 2686–96. doi:10.1002/cam4.1212.

Noble, William S. 2009. “How Does Multiple Testing Correction Work?” *Nature Biotechnology* 27 (12). Nature Publishing Group: 1135–37. doi:10.1038/nbt1209-1135.

Nolan, Tania, Rebecca E. Hands, and Stephen A. Bustin. 2006. “Quantification of MRNA Using Real-Time RT-PCR.” *Nature Protocols* 1 (3): 1559–82. doi:10.1038/nprot.2006.236.

Pai, Athma A., Jonathan K. Pritchard, and Yoav Gilad. 2015. “The Genetic and Mechanistic Basis for Variation in Gene Regulation.” *PLoS Genetics* 11 (1). doi:10.1371/journal.pgen.1004857.

Pan, Xiang, Rong Zhang, Caifeng Xie, Mingxi Gan, Sheng Yao, Yubin Yao, Jiangbo Jin, et al. 2017. “GRHL2 Suppresses Tumor Metastasis via Regulation of Transcriptional Activity of Rhog in Non-Small Cell Lung Cancer.” *American Journal of Translational Research* 9 (9): 4217–26.

Pant, P V Krishna, Heng Tao, Erica J Beilharz, Dennis G Ballinger, David R Cox, and Kelly A Frazer. 2006. “Analysis of Allelic Differential Expression in Human White Blood Cells,” 331–39. doi:10.1101/gr.4559106.2.

Pastinen, Tomi. 2010. “Genome-Wide Allele-Specific Analysis: Insights into Regulatory Variation.” *Nature Reviews Genetics* 11 (August): 533–38.

Pastinen, Tomi, Bing Ge, and Thomas J. Hudson. 2006. “Influence of Human Genome Polymorphism on Gene Expression.” *Human Molecular Genetics* 15 Spec No (1): 9–16. doi:10.1093/hmg/ddl044.

Pastinen, Tomi, and Thomas J. Hudson. 2004. “Cis-Acting Regulatory Variation in the Human Genome.” *Science* 306 (5696): 647–50. doi:10.1126/science.1101659.

Patient, Roger K. 1990. “Control of Gene Expression: Tissue-Specific Expression.” *Current Opinion in Biotechnology* 1: 151–58.

Patterson, Tucker A., Edward K. Lobenhofer, Stephanie B. Fulmer-Smentek, Patrick J. Collins, Tzu Ming Chu, Wenjun Bao, Hong Fang, et al. 2006. “Performance Comparison of One-Color and Two-Color Platforms within the MicroArray Quality Control (MAQC) Project.” *Nature Biotechnology* 24 (9): 1140–50. doi:10.1038/nbt1242.

Peart, M. J., G. K. Smyth, R. K. van Laar, D. D. Bowtell, V. M. Richon, P. A. Marks, A. J. Holloway, and R. W. Johnstone. 2005. “Identification and Functional Significance of Genes Regulated by Structurally Different Histone Deacetylase Inhibitors.” *Proceedings of the National Academy of Sciences* 102 (10): 3697–3702. doi:10.1073/pnas.0500369102.

Pereira, Bernard, Suet Feung Chin, Oscar M. Rueda, Hans Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, et al. 2016. “The Somatic Mutation Profiles of 2,433 Breast Cancers Refines Their Genomic and Transcriptomic Landscapes.” *Nature Communications* 7 (May). doi:10.1038/ncomms11479.

Perou, C.M. Charles M., Therese Sørile, M.B. Michael B. Eisen, Matt van de Rijn, Stefanie S.S. Jeffrey, C.A. Christian A. Ress, J.R. Jonathan R. Pollack, et al. 2000. “Molecular Portraits of Human Breast Tumours.” *Nature* 406 (6797): 747–52. doi:10.1038/35021093.

Perou, Charles M., and Anne Lise Borresen-Dale. 2011. “Systems Biology and Genomics of Breast Cancer.” *Cold Spring Harbor Perspectives in Biology* 3 (2): 1–17. doi:10.1101/cshperspect.a003293.

Pharoah, Paul D.P., Parry Guilford, and Carlos Caldas. 2001. “Incidence of Gastric Cancer and Breast Cancer in CDH1 (E-Cadherin) Mutation Carriers from Hereditary Diffuse Gastric Cancer Families.” *Gastroenterology* 121 (6): 1348–53. doi:10.1053/gast.2001.29611.

Pharoah, Paul D P, Alison M. Dunning, Bruce A J Ponder, and Douglas F. Easton. 2004. “Association Studies for Finding Cancer-Susceptibility Genetic Variants.” *Nature Reviews Cancer* 4 (11): 850–60. doi:10.1038/nrc1476.

Phillips, M. S., R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding, M. A. Donaldson, J. F. Studebaker, et al. 2003. “Chromosome-Wide Distribution of Haplotype Blocks and the Role of Recombination Hot Spots.” *Nature Genetics* 33 (3): 382–87. doi:10.1038/ng1100.

Phipson, Belinda, Stanley Lee, Ian J. Majewski, Warren S. Alexander, and Gordon K. Smyth. 2016. “Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression.” *Annals of Applied Statistics* 10 (2): 946–63. doi:10.1214/16-AOAS920.

Pritchard, Jonathan K., and Molly Przeworski. 2001. “Linkage Disequilibrium in Humans: Models and Data.” *The American Journal of Human Genetics* 69 (1): 1–14. doi:10.1086/321275.

R Core Team. 2017. “R: The R Project for Statistical Computing.” *R Foundation for Statistical Computing, Vienna, Austria*. https://www.r-project.org/.

Rafnar, Thorunn, Daniel F. Gudbjartsson, Patrick Sulem, Aslaug Jonasdottir, Asgeir Sigurdsson, Adalbjorg Jonasdottir, Soren Besenbacher, et al. 2011. "Mutations in BRIP1 Confer High Risk of Ovarian Cancer." *Nature Genetics* 43 (11). Nature Publishing Group: 1104–7. doi:10.1038/ng.955.

Raouf, Afshin, Yun Zhao, Karen To, John Stingl, Allen Delaney, Mary Barbara, Norman Iscove, et al. 2008. "Transcriptome Analysis of the Normal Human Mammary Cell Commitment and Differentiation Process." *Cell Stem Cell* 3 (1): 109–18. doi:10.1016/j.stem.2008.05.018.

Rich, Thereasa A., Ashley H. Woodson, Jennifer Litton, and Banu Arun. 2015. "Hereditary Breast Cancer Syndromes and Genetic Testing." *Journal of Surgical Oncology* 111 (1): 66–80. doi:10.1002/jso.23791.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47. doi:10.1093/nar/gkv007.

Rockman, Matthew V., and Leonid Kruglyak. 2006. "Genetics of Global Gene Expression." *Nature Reviews Genetics* 7 (11): 862–72. doi:10.1038/nrg1964.

Rosenberg, Noah A., and Jenna M. VanLiere. 2009. "Replication of Genetic Associations as Pseudoreplication Due to Shared Genealogy." *Genetic Epidemiology* 33 (6): 479–87. doi:10.1016/j.immuni.2010.12.017.Two-stage.

Sartor, Maureen A., Craig R. Tomlinson, Scott C. Wesselkamper, Siva Sivaganesan, George D. Leikauf, and Mario Medvedovic. 2006. "Intensity-Based Hierarchical Bayes Method Improves Testing for Differentially Expressed Genes in Microarray Experiments." *BMC Bioinformatics* 7: 1–17. doi:10.1186/1471-2105-7-538.

Sauter, Edward R. 2018. "Breast Cancer Prevention: Current Approaches and Future Directions." *European Journal of Breast Health*, no. 4: 64–71. doi:10.5152/ejbh.2018.3978.

Schadt, Eric E., Stephanie A. Monks, Thomas A. Drake, Aldons J. Lusis, Nam Che, Veronica Colinayo, Thomas G. Ruff, et al. 2003. "Genetics of Gene Expression Surveyed in Maize, Mouse and Man." *Nature* 422 (6929): 297–302. doi:10.1038/nature01434.

Schmittgen, Thomas D, and Kenneth J Livak. 2008. "Analyzing Real-Time PCR Data by the Comparative CT Method." *Nature Protocols* 3 (6): 1101–8. doi:10.1038/nprot.2008.73.

Sengupta, Saubhik, Chad M. Michener, Pedro Escobar, Jerome Belinson, and Ram Ganapathi. 2008. "Ovarian Cancer Immuno-Reactive Antigen Domain Containing 1 (OCIAD1), a Key Player in Ovarian Cancer Cell Adhesion." *Gynecologic Oncology* 109 (2): 226–33. doi:10.1016/j.ygyno.2007.12.024.

Serre, David, Scott Gurd, Bing Ge, Robert Sladek, Donna Sinnett, Eef Harmsen, Marina Bibikova, et al. 2008. "Differential Allelic Expression in the Human Genome: A Robust Approach to Identify Genetic and Epigenetic Cis-Acting Mechanisms Regulating Gene Expression." *PLoS Genetics* 4 (2). doi:10.1371/journal.pgen.1000006.

Sherry, S. T. 2001. “DbSNP: The NCBI Database of Genetic Variation.” *Nucleic Acids Research* 29 (1): 308–11. doi:10.1093/nar/29.1.308.

Shetty, Deeti K., Kaustubh P. Kalamkar, and Maneesha S. Inamdar. 2018. “OCIAD1 Controls Electron Transport Chain Complex I Activity to Regulate Energy Metabolism in Human Pluripotent Stem Cells.” *Stem Cell Reports* 11 (1). ElsevierCompany.: 128–41. doi:10.1016/j.stemcr.2018.05.015.

Shiovitz, Stacey, and L. A. Korde. 2015. “Genetics of Breast Cancer: A Topic in Evolution.” *Annals of Oncology* 26 (7): 1291–99. doi:10.1093/annonc/mdv022.

Slatkin, Montgomery. 2008. “Linkage Disequilibrium - Understanding the Evolutionary Past and Mapping the Medical Future.” *Nature Reviews Genetics* 9 (6): 477–85. doi:10.1038/nrg2361.Linkage.

Smyth, Gordon K. 2004. “Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments.” *Statistical Applications in Genetics and Molecular Biology* 3 (1): 1–25. doi:10.2202/1544-6115.1027.

Sorlie, T., R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, et al. 2003. “Repeated Observation of Breast Tumor Subtypes in Independent Gene Expression Data Sets.” *Proceedings of the National Academy of Sciences* 100 (14): 8418–23. doi:10.1073/pnas.0932692100.

Sorlie, T, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, et al. 2001. “Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (19): 10869–74. doi:10.1073/pnas.191367098.

Stingl, John, and Carlos Caldas. 2007. “Molecular Heterogeneity of Breast Carcinomas and the Cancer Stem Cell Hypothesis.” *Nat. Rev. Cancer* 7 (10): 791–99. doi:10.1038/nrc2212.

Stranger, Barbara E., Alexandra C. Nica, Matthew S. Forrest, Antigone Dimas, Christine P. Bird, Claude Beazley, Catherine E. Ingle, et al. 2007. “Population Genomics of Human Gene Expression.” *Nature Genetics* 39 (10): 1217–24. doi:10.1038/ng2142.

Tao, Zi Qi, Aimin Shi, Cuntao Lu, Tao Song, Zhengguo Zhang, and Jing Zhao. 2015. “Breast Cancer: Epidemiology and Etiology.” *Cell Biochemistry and Biophysics* 72 (2). Springer US: 333–38. doi:10.1007/s12013-014-0459-6.

Thompson, Deborah, Silvia Duedal, Jennifer Kirner, Lesley McGuffog, James Last, Anne Reiman, Philip Byrd, Malcolm Taylor, and Douglas F. Easton. 2005. “Cancer Risks and Mortality in Heterozygous ATM Mutation Carriers.” *Journal of the National Cancer Institute* 97 (11): 813–22. doi:10.1093/jnci/dji141.

Thorisson, Gudmundur A., Albert V. Smith, Lalitha Krishnan, and Lincoln D. Stein. 2005. “The International HapMap Project Web Site.” *Genome Research* 15: 1592–93. doi:10.1038/nature02168.

Torre, Lindsey A., Freddie Bray, Rebecca L. Siegel, Jacques Ferlay, Joannie Lortet-

tieulent, and Ahmedin Jemal. 2015. “Global Cancer Statistics, 2012.” *CA: A Cancer Journal of Clinicians*. 65 (2): 87–108. doi:10.3322/caac.21262.

Valle, Laura, Tarsicio Serena-acedo, Sandya Liyanarachchi, Heather Hampel, Zhongyuan Li, Qinghua Zeng, Hong-tao Zhang, et al. 2008. “Germline Allele-Specific Expression of TGFBR1 Confers an Increased Risk of Colorectal Cancer” 321 (5894): 1361–65. doi:10.1126/science.1159397.Germline.

VanGuilder, Heather D., Kent E. Vrana, and Willard M. Freeman. 2008. “Twenty-Five Years of Quantitative PCR for Gene Expression Analysis.” *BioTechniques* 44 (5): 619–26. doi:10.2144/000112776.

VanLiere, Jenna M., and Noah A. Rosenberg. 2008. “Mathematical Properties of R2 Measure of Linkage Disequilibrium” 74 (1): 130–37. doi:10.1016/j.tpb.2008.05.006.Mathematical.

Veyrieras, Jean Baptiste, Sridhar Kudaravalli, Su Yeon Kim, Emmanouil T. Dermitzakis, Yoav Gilad, Matthew Stephens, and Jonathan K. Pritchard. 2008. “High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation.” *PLoS Genetics* 4 (10). doi:10.1371/journal.pgen.1000214.

Wang, Chunyan, Chad M Michener, Jerome L Belinson, Susan Vaziri, Ram Ganapathi, and Saubhik Sengupta. 2010. “Role of the 18:1 Lysophosphatidic Acid-Ovarian Cancer Immunoreactive Antigen Domain Containing 1 (OCIAD1)-Integrin Axis in Generating Late-Stage Ovarian Cancer.” *Molecular Cancer Therapeutics* 9 (6): 1709–18. doi:10.1158/1535-7163.MCT-09-1024.

Wang, Xuting, Daniel J. Tomso, Xuemei Liu, and Douglas A. Bell. 2005. “Single Nucleotide Polymorphism in Transcriptional Regulatory Regions and Expression of Environmentally Responsive Genes.” *Toxicology and Applied Pharmacology* 207 (2 SUPPL.): 84–90. doi:10.1016/j.taap.2004.09.024.

Wickham, Hadley. 2016. *Ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wooster, R, G Bignell, J Lancaster, S Swift, S Seal, J Mangion, N Collins, S Gregory, C Gumbs, and G Micklem. 1995. “Identification of the Breast Cancer Susceptibility Gene BRCA2.” *Nature* 378 (6559): 789–92. doi:10.1038/378789a0.

Xavier, Joana, Roslin Russell, Bernardo P. Almeida, Nordiana Rosli, Catia Rocha, Shamith Samarajiwa, Suet-Feung Chin, Carlos Caldas, Bruce AJ Ponder, and Ana-Teresa Maia. 2016. “Abstract A31: Integrative Differential Allelic Expression Analysis Efficiently Reveals the Biology Underlying Risk to Breast Cancer.” *Molecular Cancer Research* 14 (2 Supplement). American Association for Cancer Research: A31–A31. doi:10.1158/1557-3125.ADVBC15-A31.

Xiao, Rui, and Laura Scott. 2011. “Detection of Cis-Acting Regulatory SNPs Using Allelic Expression Data.” *Genetic Epidemiology* 35 (6): 515–25. doi:10.1016/j.chemosphere.2012.12.037.Reactivity.

Yan, Hai, Weishi Yuan, Victor E. Velculecu, Bert Vogelstein, and Kenneth W. Kinzler. 2002. “Allelic Variation in Human Gene Expression.” *Science* 297

(5584).

Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. “Ensembl 2018.” *Nucleic Acids Research* 46 (D1): D754–61. doi:10.1093/nar/gkx1098.

Zhu, Xiangyuan, Kenli Li, Ahmad Salah, Stinus Lindgreen, Marcel Martin, Ben Langmead, Steven L Salzberg, et al. 2016. “Parallel Computation in Biological Sequence Analysis.” *Parallel and Distributed Systems, IEEE Transactions On* 9 (4): 283–94. doi:10.1186/gb-2013-14-4-r36.

Annexes

ANNEXES

Annex *A* - Details of Taqman™ SNP Genotyping Assays (Applied Biosystems by Thermo Fisher Scientific) for the 2 SNPs studied (rs9997920 from *OCIAD1* and rs6989650 from *GRHL2* gene)

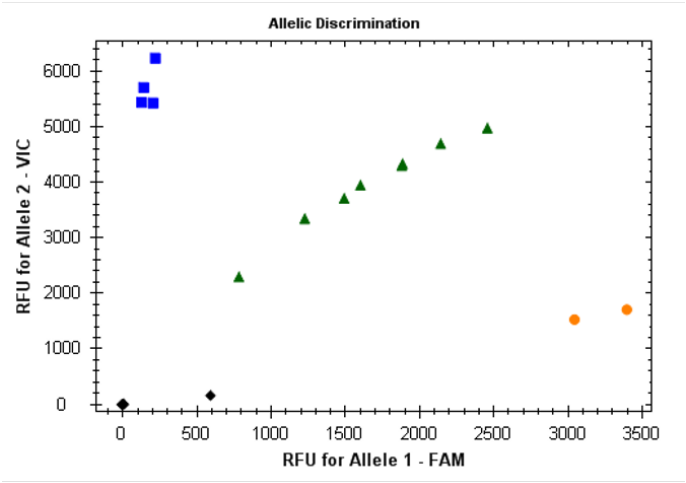
Assay ID	Rs Number	Part Number	Product	Reporter/Quencher	Volume (uL)	Formulation	Number of reactions (5 uL reaction size)
C___7932003_10	rs9997920	4351379	Taqman™ SNP Genotyping Assays, Human, SM	Allele C: VIC/MGB-NFQ	188	40x	1,5
C__11849073_10	rs6989650			Allele T:FAM/MGB-NFQ			

Annex B - List of candidate risk SNPs.

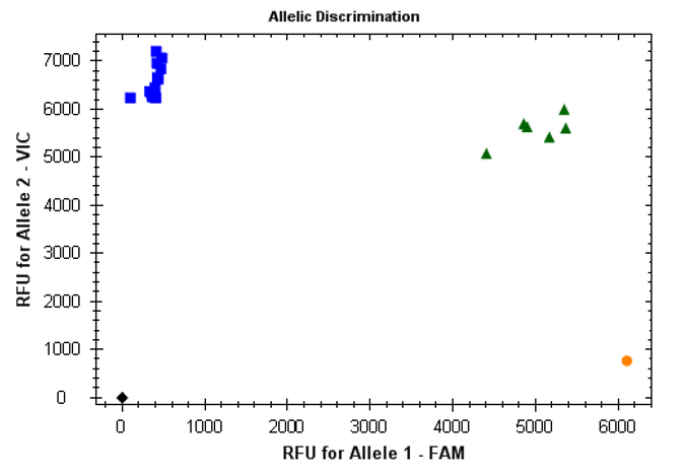
https://drive.google.com/open?id=1J7jdlPZvZkVoNncUSEEpIHErmwvzhMiP

Annex c - Genotyping of DNA CEPH samples by Taqman™ RT-qPCR. Results of genotyping for 2 SNPs studied in *OCIAD1* gene (rs9997920) and GRHL2 gene (rs6989650). The x-axis represents the fluorescence intensity for allele 1, emitted by probe FAM and the y-axis indicates the fluorescence intensity for allele 2. The blue squares represent the homozygous samples for allele 1, the green triangle represent heterozygous samples and the black diamonds represent the NTC.

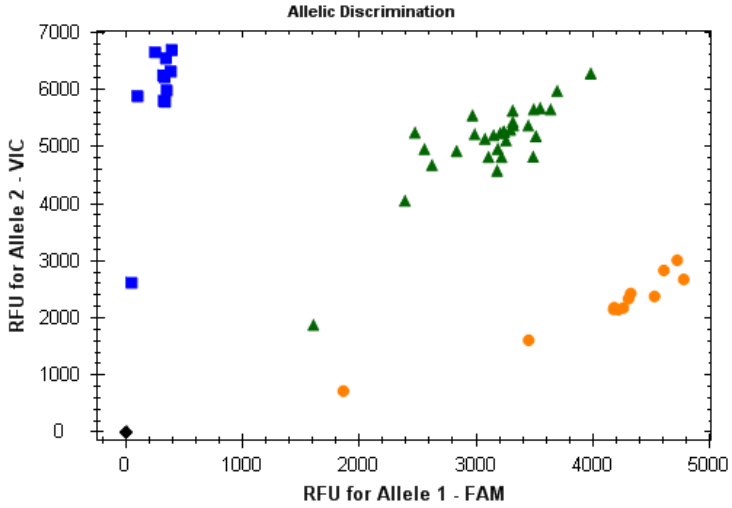
rs9997920



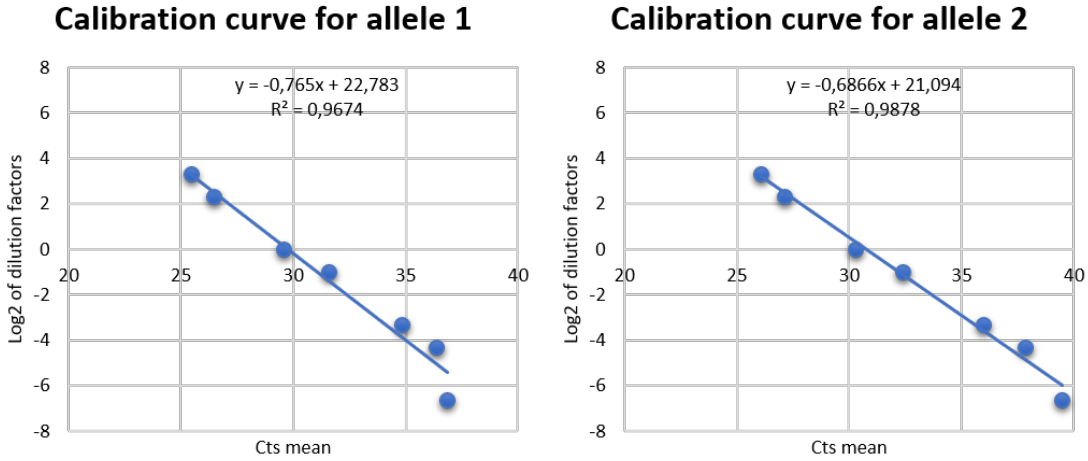
rs6989650



Annex D - Genotyping of cDNA from blood cancer samples of patients with BC by Taqman™ RT-qPCR for the rs9997920 from *OCIAD1* gene. The x-axis represents the fluorescence intensity for allele 1, emitted by probe FAM and the y-axis indicates the fluorescence intensity for allele 2. The blue squares represent the homozygous samples for allele 1, the green triangle represent heterozygous samples and the black diamonds represent the NTC.

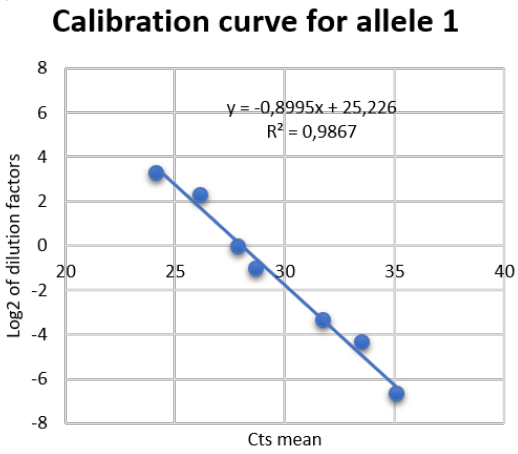


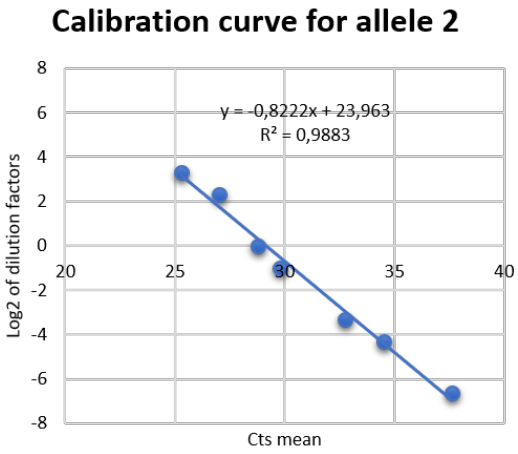
Annex E - Calibration curve for OCIAD1 blood tissue RT-qPCR in the second run. These figures show the calibration curve for each one of the alleles of a heterozygous sample for rs9997920 of OCIAD1 gene in the blood tissue samples.



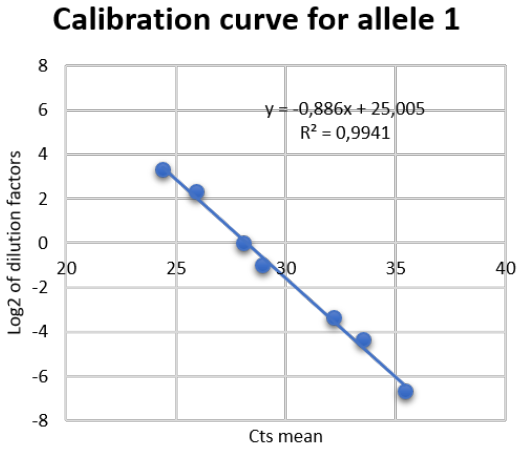
Annex F. Calibration curve for OCIAD1 breast tissue RT-qPCR in first and second run. In **A** is represented the calibration curve for each one of the alleles of a heterozygous sample for rs9997920 in first run. **B** shows the same for the second run..

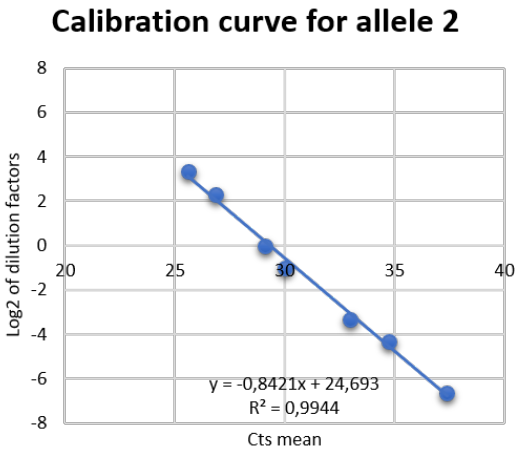
A





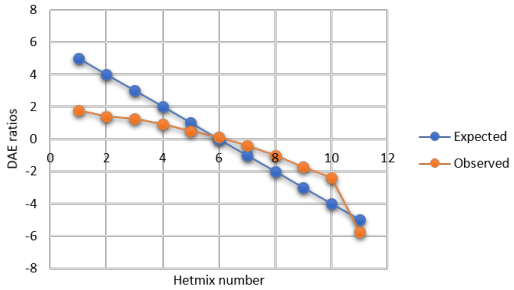
B

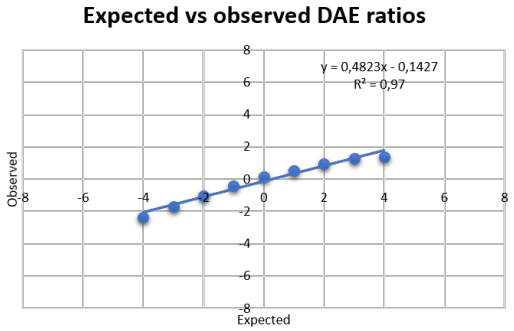




Annex G. Hetmixes DAE ratios for rs9997920 (*OCIAD1*). In **A** is represented Hetmixes DAE ratios for *OCIAD1* in second run RT-qPCR in blood tissue and respective correlation. In **B** is showed the Hetmixes DAE ratios for *OCIAD1* in first RT-qPCR in breast tissue and respective correlation. In **C** is demonstrated the Hetmixes DAE ratios for *OCIAD1* in the second RT-qPCR for breast tissue and respective correlation (continued on the next page)

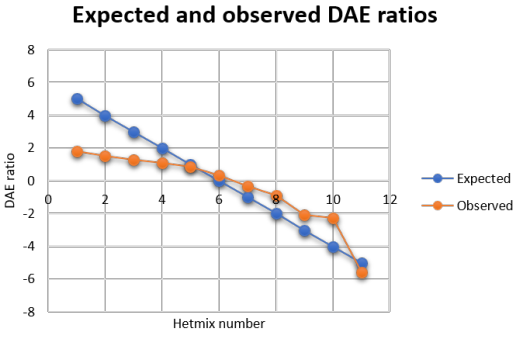
A

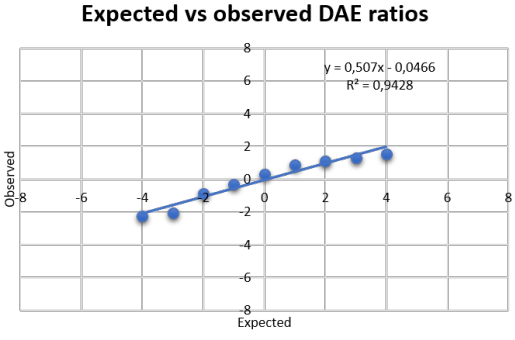




Annex G – continued

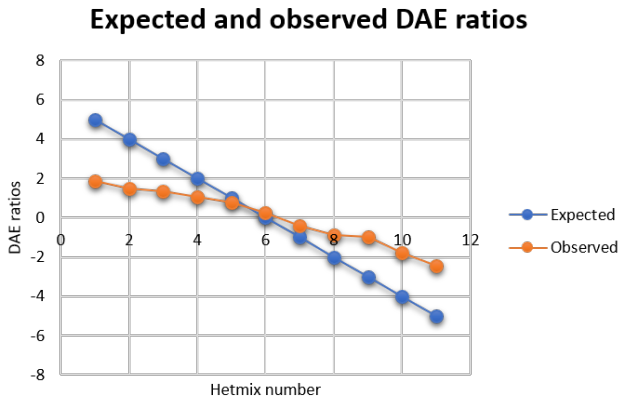
B

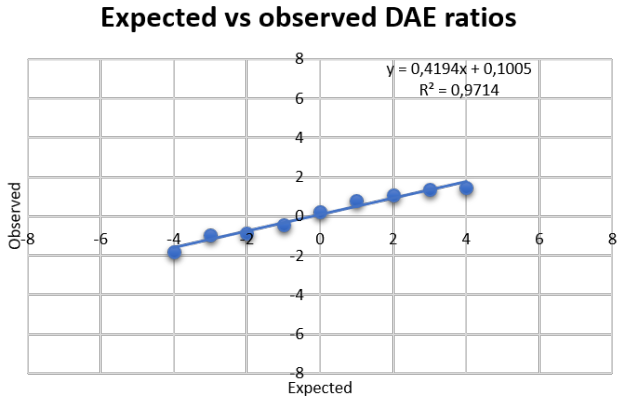




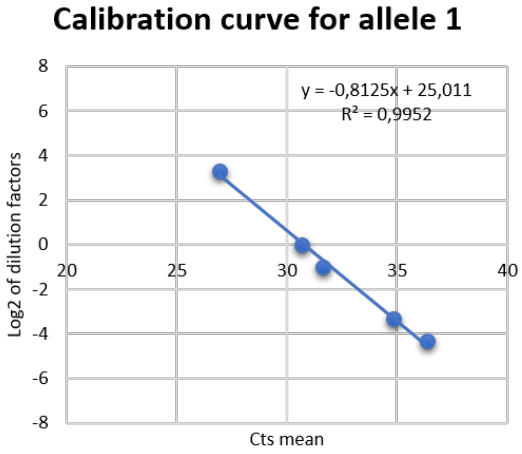
Annex G – continued

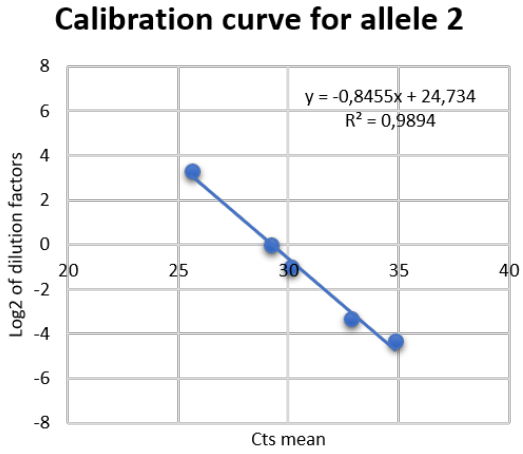
C



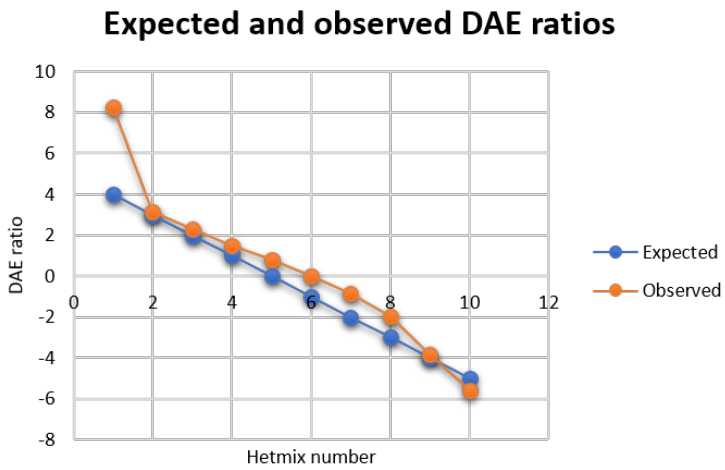


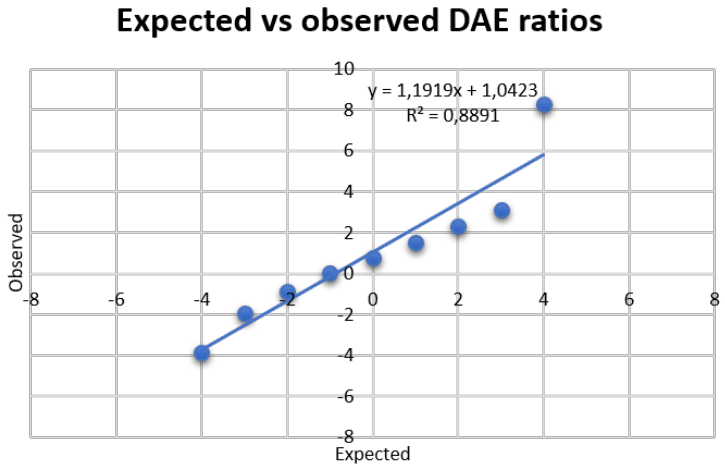
Annex H. Calibration curve for *GRHL2* breast tissue RT-qPCR in second run. These figures show the calibration curve for each one of the alleles of a heterozygous sample for rs6989650 od *GRHL2* gene in the breast tissue samples.



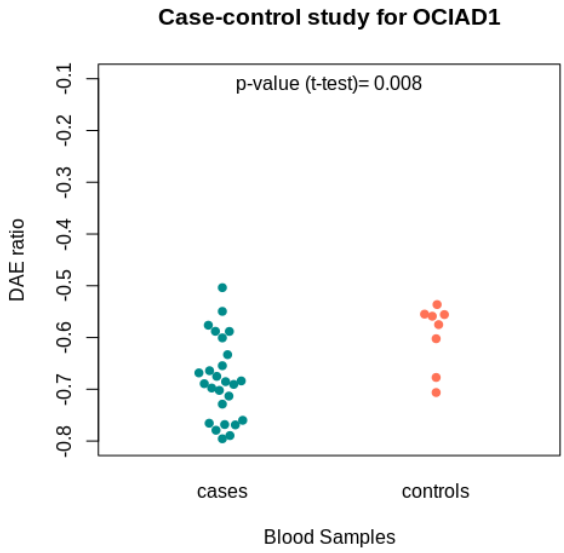


Annex I - Hetmixes DAE ratios for rs6989650 from *GRHL2*. This representation shows Hetmixes DAE ratios for *OCIAD1* in second run RT-qPCR in blood tissue and respective correlation.

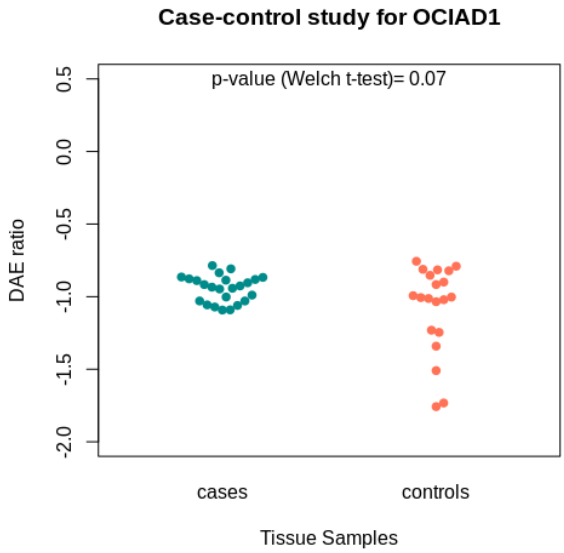




Annex J - Case-control association results using DAE ratios for *OCIAD1* (rs9997920) in blood tissue (second run).



Annex K - Case-control association study using DAE ratios for *OCIAD1* in breast tissue (second run).



Annex L - Case-control association study using DAE ratios for *GRHL2* in breast tissue (second run).