JOCELINE JANICE CORREIA SILVA

Functional analysis of genetic variants associated with risk for breast cancer: 12q24, a candidate risk locus

UAlg
UNIVERSIDADE DO ALGARVE
DEPARTAMENTO DE CIÊNCIAS BIOMÉDICAS E MEDICINA

2016

JOCELINE JANICE CORREIA SILVA

Functional analysis of genetic variants associated with risk for breast cancer: 12q24, a candidate risk locus

Master in Oncobiology – Molecular Mechanisms of Cancer

Supervisor: Professor Doctor Ana Teresa Maia, PhD

Co-Supervisor: Doctor Joana Xavier, PhD

2016

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.

Agradecimentos

Este espaço é dedicado às pessoas que, de alguma forma, contribuíram para que esta dissertação fosse concluída.

Em primeiro lugar, gostaria de agradecer à minha orientadora, Professora Doutora Ana Teresa Maia, pela compreensão, orientação, por todo o conhecimento partilhado e sobretudo, pelo o incentivo e confiança, ao longo desta etapa. Obrigada pela oportunidade de pertencer a esta equipa.

Quero agradecer à minha co-orientadora, Doutora Joana Xavier, pela paciência, por todo o conhecimento partilhado e apoio sempre que precisei. Obrigada pela disponibilidade e compreensão mesmo nos momentos mais difíceis.

Um agradecimento muito especial ao Bernardo Almeida e à Iris Silva, pela ajuda preciosa no laboratório e pela amizade construída.

O meu maior agradecimento é para os meus pais, que nunca deixaram de acreditar em mim e sempre estiveram do meu lado, e ao meu irmão Henrique, que tanto sacrificou para que a conclusão desta etapa fosse possível.

Abstract

Common risk alleles identified through Genome-Wide Association Studies (GWAS) explain about 14% of familial breast cancer cases. However, GWAS do not identify causative variants in the risk loci and do not contribute to the understanding of risk mechanisms. All of the risk loci functionally analysed to date are cis-regulatory, i.e. polymorphisms that modify gene expression. Therefore, **we hypothesize that cis-regulation is a central mechanism in breast cancer susceptibility.**

Differential allelic expression (DAE) is the most robust method to identify the effect of cis-regulatory single nucleotide polymorphisms (SNPs). Our group established a whole-genome DAE map for normal breast tissue, which we integrated with the GWAS data, to identify risk loci with greater potential to be cis-regulatory. We identified 111 loci, with one of them in the 12q24 locus, containing an unpublished GWAS SNP, rs7307700, and 15 DAE SNPs.

We performed *in silico* analysis to characterize the regulatory potential of candidate cis-regulatory SNPs (rSNPs) in breast cell lines, and *in vitro* analysis by electrophoretic mobility shift assay (EMSA) to explore interactions between candidate rSNPs and candidate transcription factors (TFs). Three candidate rSNPs, rs10773145, rs10846834 and rs12302714, overlapped regulatory elements and DNase I hypersensitivity sites, and were associated with the DAE observed for two transcribed SNPs (or DAE SNPs), rs7301263 and rs12581512. The candidate SNPs rs10773145 and rs10846834 were both located within known c-FOS and STAT3 binding sites, but showed small allelic differences in the ChIP-seq data. Since there was no ChIP-seq data for rs12302714, we carried EMSA analysis. Although we detected DNA-protein binding for both alleles of this SNP, no allelic differences were detected. We also analysed candidate SNPs for microRNA binding and the results suggested that a microRNA have preferentially binding to the alleles of candidate rSNP rs12302714. These results indicate that the DAE observed might not be explained by differential binding of TFs at the three candidate rSNPs and might be due to other regulatory mechanisms, that require further exploration, such as splicing and microRNAs.

Keywords: breast cancer; single nucleotide polymorphisms; risk; cis-regulatory variants; differential allelic expression.

Resumo

O cancro da mama é uma das doenças oncológicas mais comuns, sendo a mais frequente causa de morte entre as mulheres. É estimado que uma em cada onze mulheres será diagnosticada com cancro da mama ao longo da sua vida. Trata-se de uma patologia complexa cuja etiologia pode ser devido a fatores genéticos e não genéticos. Estima-se que 5% a 10% dos casos de cancro da mama são devido a fatores genéticos, no entanto, o conhecimento atual acerca do risco hereditável não explica cerca 50% destes casos familiares. Recentes avanços tecnológicos, nomeadamente nos *microarrays* de genotipagem, e nos Estudos de Associação no Genoma Inteiro (*genome-wide association studies*, GWAS) permitiram identificar um grande número de variantes associadas a risco para cancro da mama. Os GWAS são estudos divididos por fases, que analisam variações no genoma inteiro, com o objetivo de descobrir fatores genéticos de risco de doenças comuns na população, como o cancro da mama.

As variantes cis-reguladoras são polimorfismos frequentes na população (>5% de frequência do alelo menos frequente na população), ao contrário das mutações (<1% de frequência na população). Estes polimorfismos têm a capacidade de regular a expressão de genes quando localizados em elementos reguladores, nomeadamente, promotores ou elementos intensificadores (*enhancer*), podendo afetar a ligação de fatores de transcrição e consequentemente, a regulação de determinado gene.

Atualmente, 94 *loci* de suscetibilidade para o cancro da mama foram identificados através de GWAS, que explicam apenas cerca de 14% do risco para esta patologia. Até à data, foram estudados funcionalmente 13 *loci*, e os resultados sugerem que os polimorfismos analisados tinham como mecanismo de atuação a cis-regulação. Adicionalmente, do 94 loci somente um se localiza numa região codificante, com todos os outros a localizarem-se em intrões, regiões intergénicas e regiões sem transcrição detetável (“gene deserts”). Finalmente, os GWAS para além dos 94 *loci* de risco validados, produziram longas listas de *loci* com significância estatística muito elevada, que necessitam de ser priorizados para estudos de validação.

Com base nestas evidências, a nossa hipótese é que a cis-regulação é um mecanismo importante para o risco do cancro da mama e que a maioria dos polimorfismos associados ao risco para o cancro da mama ainda por descobrir poderão ser também cis-reguladores.

Este trabalho foca-se nos polimorfismos de nucleótido único (SNPs) cis-reguladores e, entre outras abordagens, estes SNPs cis-reguladores (rSNP) podem ser identificados através da análise de *loci* de características quantitativas de expressão (*expression quantitative trait loci*, eQTL) e da análise de Expressão Alélica Diferencial (*differential allelic expression*, DAE). A análise de eQTL permite fazer uma associação entre SNPs e a variação de expressão total de determinado gene. No entanto, o nível de expressão total está sujeito a fatores em trans (tal como o nível de proteínas com função de fatores de transcrição), para além dos fatores em cis (alterações na sequência, tal como os SNPs). DAE é um dos possíveis efeitos observados na presença de rSNPs em elementos reguladores, dessa forma, a análise de DAE permite comparar os níveis relativos de expressão dos dois alelos do mesmo gene em indivíduos heterozigóticos, utilizando um SNP transcrito (tSNP ou DAE SNP). Esta abordagem não só indica qual o alelo a causar DAE, como elimina o efeito de fatores trans, pois compara os níveis de transcritos dos alelos individualmente no mesmo contexto celular e haplótipos.

Num trabalho anterior feito pela Prof. Ana Teresa Maia e colegas, desenvolveu-se um mapa de DAE em 64 amostras de tecido mamário normal, que informa quais genes estão sob a influência de rSNPs. O próximo passo será identificar os SNPs causadores de risco. Assim, os dados do mapa de DAE foram cruzados com os resultados publicados e não publicados de GWAS para cancro da mama. Este cruzamento de dados foi feito de acordo com a localização cromossómica, distância física (janelas de ±250kb entre o GWAS SNP e o DAE SNP) e padrões de desequilíbrio de ligação (*linkage disequilibrium,* LD) com o valor mínimo de r^2 = 0.4. Foram identificados 111 *loci* candidatos que contêm pelo menos um GWAS SNP e um DAE SNP e com forte potencial cis-regulador. Em 32 *loci* o GWAS SNP e o DAE SNP estavam em elevado LD, ou seja, os seus genótipos estavam fortemente associados. Como todos os *loci* estudados funcionalmente sugerem que o mecanismo causador de risco para o cancro da mama é a cis-

regulação, e como todos os *loci* identificados, com exceção a um, encontram-se em regiões não codificantes (sugerindo que estão localizados em regiões regulatórias), selecionámos para análise funcional o locus 12q24, não publicado, para testar se este locus encontra-se também sob influência de rSNPs e validar este locus para o risco de cancro da mama. O GWAS SNP neste locus não atingiu o valor estabelecido pelo GWAS para passar a fase III, talvez por não estar em elevado LD com o rSNP causal. Desta forma, iremos testar se a integração do nosso mapa de DAE com os dados do GWAS relativos ao cancro da mama é uma boa abordagem para priorizar *loci* ainda por validar, com maior probabilidade de estarem sob influência de variantes cis-reguladoras, e consequentemente, mais prováveis a estarem associados ao risco para o cancro da mama.

Este trabalho teve como objetivo: 1) validar um dos *loci* identificados, mas não validados, localizado na região 12q24, e confirmar a sua associação com o risco para o cancro da mama; 2) identificar e analisar funcionalmente as variantes com potencial a serem cis-reguladoras no locus 12q24; 3) testar se a nossa abordagem é um método eficaz para priorizar variantes candidatas a associados com risco.

Começou-se por analisar o nosso mapa de DAE nesta região. A região do locus 12q24 apresenta 15 DAE SNPs e um GWAS SNP, rs7307700, localizado no gene *AACS*. Para identificar e analisar possíveis variantes associadas ao risco e com potencial a serem rSNPs, foram feitas análises *in silico*. Os dados dos projetos HapMap e 1000 Genomes Project foram consultados para identificar os melhores candidatos a rSNPs em LD ≥ 0.4 com o GWAS SNP, sendo identificados 72 rSNPs candidatos. Para analisar estes candidatos, acedeu-se aos dados dos projetos ENCODE e Roadmap Epigenomics, que contêm informações sobre zonas de hipersensibilidade à desoxirribonuclease I (DHSs), imuno-precipitação da cromatina (ChIP-seq) para diversas modificações de histonas e fatores de transcrição, previsões alélicas de ligação de proteínas (PWM). No final desta análise, 12 rSNPs candidatos foram encontrados em sobreposição com DHSs e com regiões que contêm marcadores para elementos reguladores, com evidência de estarem ativos em linhas celulares mamárias,

sugerindo que esses podem ter um efeito funcional através da regulação da expressão de genes alvo.,

Para identificar as variantes que poderão estar a causar DAE no locus 12q24, testaram-se os níveis de expressão alélica dos 15 DAE SNPs com os genótipos dos 12 rSNPs candidatos. Dado o padrão de DAE demonstrado pelos DAE SNPs, pretendeu-se identificar os rSNP candidatos cujos homozigóticos não demonstrassem DAE nos DAE SNPs (i.e., SNPs transcritos), e cujos heterozigóticos apresentassem DAE nos DAE SNPs. Três dos 12 candidatos (rs10773145, rs10846834 e rs12302714) explicavam o DAE de dois DAE SNPs (rs12581512 e rs7301263). Para dois deles, rs10773145 e rs10846834, que se encontravam em completo LD um com o outro, existiam dados de ChIP-seq disponíveis que indicavam a ligação das proteínas STAT3 e c-FOS. No entanto, esses dados não revelavam diferenças de afinidade entre os alelos de cada SNP. Para o terceiro candidato, rs12302714, como não existiam dados de ChIP-seq, procedemos com ensaios *in vitro*. Os resultados de EMSA (*electrophoretic mobility shift assay*) sugeriram que, apesar de haver ligação de proteína, não existiam diferenças de afinidade para os alelos deste rSNP candidato. De acordo com estes resultados, é possível que estes três candidatos estejam a afetar o DAE observado nos DAE SNPs do gene *AACS* por outro mecanismo que não a ligação diferencial de fatores de transcrição em elementos reguladores. Outros mecanismos possíveis incluem diferenças alélicas de produção de transcritos alternativos (alelos a afetar o processo de *splicing*), ou de regulação por microRNAs.

De seguida, analisou-se se havia alguma previsão de ligação preferencial de microRNAs aos alelos dos 72 SNPs candidatos. Em 17 dos 72 SNPs (incluindo o SNP rs12302714) houve previsões de ligação microRNAs com preferência a um dos alelos comparativamente ao outro. Posteriormente, analisaram-se os genótipos dos candidatos rSNPs, DAE SNP e GWAS SNP para a estrutura de LD nessa região e para identificação dos haplótipos, nas 64 amostras de tecido normal da mama, que poderão ser responsáveis pelo aumento ou diminuição da expressão dos genes. Foram identificados seis haplótipos comuns, estando dois haplótipos associados a diferenças nos níveis de expressão. Estes resultados

sugerem que talvez seja o efeito acumulativo de dois ou mais rSNPs a causar o risco para cancro da mama e o DAE observado nos DAE SNPs no locus 12q24.

Em paralelo a este trabalho, um outro locus (5q14.2) foi funcionalmente analisado. Um dos candidatos rSNP identificados através da análise *in silico*, afeta diferencialmente a ligação de um fator de transcrição no gene *ATG10*, causando assim, DAE por cis-regulação. No entanto, o fator de transcrição que se liga preferencialmente a um dos alelos deste rSNP permanece por identificar.

Em suma, o cruzamento dos nossos dados de DAE com os dados de GWAS foi uma boa abordagem para priorizar *loci* não publicados dos GWASes que estão sob influência de cis-regulação, e com potencial para ser associado ao risco, para validação para o risco de cancro da mama. Futuramente, mais análises *in silico* e *in vitro* deverão ser feitas, de modo a entender que outro mecanismo de regulação poderá explicar o DAE observado no locus 12q24, e que fator de transcrição poderá estar a regular a expressão do gene *ATG10* (locus 5q14.2). Uma análise mais aprofundada da regulação destes genes poderá levar também à compreensão da biologia de predisposição ao cancro e contribuir para o desenvolvimento de terapias futuras, especialmente na área da medicina personalizada, baseada nos haplótipos que regem o DAE em cada indivíduo.

Palavras-chave: cancro da mama; suscetibilidade; polimorfismos de nucleóticos únicos; variantes cis-reguladoras; expressão alélica diferencial.

Index of contents

Index of figures

Index of tables

Index of annex

List of abbreviations

AACS – acetoacetyl-CoA synthetase

AC – adenocarcinoma

BC – breast cancer

bp – base pair

BRCA 1 / 2 – breast cancer 1 / 2

BR.H35 – breast variant HMEC

BR.MYO – breast myoepithelial primary

bZIP – basic leucine zipper

CCND1 – cyclin D1

c-FOS – FOS proto-oncogene

ChIP-seq – chromatin immunoprecipitation sequencing

CNA – copy number aberrations

D' – D prime

DAE – differential allelic expression

DAE SNP – transcribed single nucleotide polymorphism

DCIS – ductal carcinoma *in situ*

DHS – DNase I hypersensitive site

DNA – deoxyribonucleic acid

EGFR – epidermal growth factor receptor

EMSA – electrophoretic mobility shift assay

eQTL – expression quantitative trait loci

ER – oestrogen receptor

ERBB2 – erb-B2 receptor tyrosine kinase 2

FGF – fibroblast growth factor

FGFR – fibroblast growth factor receptor

FGFR2 – fibroblast growth factor receptor 2

GWAS – genome-wide association studies

HCC1954 - human mammary ductal carcinoma

HMEC – human mammary epithelial cells

HMF – human mammary fibroblasts

IDC – infiltrating ductal carcinoma

IGV – integrative genome viewer

IntClust – Integrative clustering

Kb – kilo-base

LCIS – lobular carcinoma *in situ*

LD - *linkage disequilibrium*

MAF – minor allele frequency

miR or miRNA – microRNA

MAP3K1 – mitogen-activated protein kinase kinase kinase 1

MDA-MB-231 – human mammary adenocarcinoma

MCF-7 – human mammary adenocarcinoma

MCF10A – human mammary epithelial

mRNA – messenger RNA

PR – progesterone receptor

PWM - protein weight matrix

r^2 – R square

RNA – ribonucleic acid

rSNPs – regulatory single nucleotide polymorphisms

SNAP – SNP annotation and proxy search

SNPs – single nucleotide polymorphisms

STAT3 – signal transducer and activator of transcription 3

T-47D – human mammary ductal carcinoma

tSNP or DAE SNP – transcribed single nucleotide polymorphism

TP53 – tumour protein 53

WHO – world health organization

1 Introduction

1.1 Cancer overview

1.1.1 Epidemiology

Cancer is among the diseases with the highest incidence in the world, leading not only to a reduction in the patient's quality of life but also to a socio-economic decline. It is estimated that one in four males and one in five females will have severe cancer in their lifetime (American Cancer Society, 2016). Every year, worldwide, fourteen million people are diagnosed with cancer and eight million die, and 1/3 of these deaths are thought to be preventable. In Europe alone, there were 3.45 million new cases of cancer (besides non-melanoma skin cancer) in 2012. Since cancer develops with age, these numbers tend to increase, as life expectancy becomes longer. Without further improvement at scientific and prevention level it is estimated that by 2030 death by cancer will increase 59% (Globocan, 2012). In Portugal, according to Globocan 2012, 49,174 people were diagnosed with cancer, with breast cancer being the most common malignant tumour, and the incidence is predicted to increase to 60,772 by 2030.

1.1.2 Aetiology

According to the World Health Organization (WHO), cancer corresponds to an uncontrolled growth and dissemination of cells. Normally, it is characterized by the accumulation of genetic mutations over time, in the same cell, mostly due to environmental factors that cause an abnormal and uncoordinated proliferation of the cell (**Figure 1.1.2.1**) (Jackson & Loeb 1998). In other words, the real problem in cancer is the uncontrolled ability of a single cell that carries a driver mutation to divide.

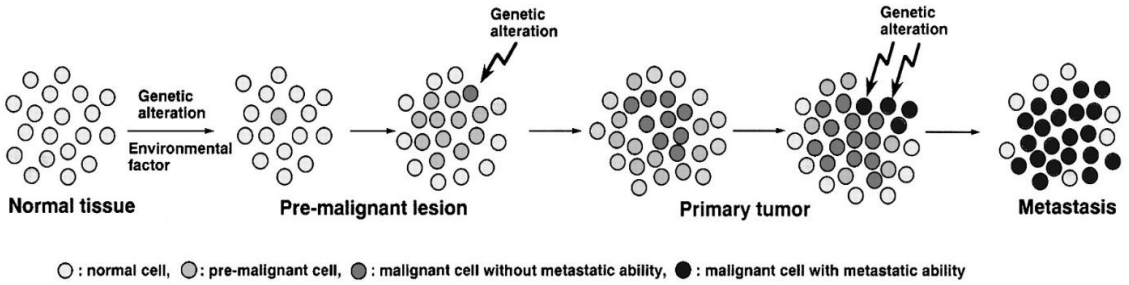


Figure 1.1.2.1 Carcinogenesis stages. Tumours are complex groups of cells, with high level of intra- and inter- heterogeneity. Each step reflects genetic changes that will lead to a cancer cell. It begins with alterations that will inactivate tumour suppressor genes and activate oncogenes, promoting uncontrolled proliferation of the mutated cell, leading further to metastasis. (Image taken from (Yokota 2000)).

Carcinogenesis is a multistep process which begins with the acquisition of somatic mutations in (proto-)oncogenes or tumour suppressor genes, where activation/up regulation or loss of function, respectively, causes hyperproliferation, blocking of differentiation and inhibition of cellular death (apoptosis) (Osborne 2004). (Proto-)oncogenes are involved in the normal growth of a cell, coding for proteins that stimulate cell growth, proliferation and regulate apoptosis. However, when mutated they become oncogenes that are constitutively activated, leading to abnormal cell proliferation, anomalous expression of growth factors and their receptors, such as fibroblast growth factor (FGF) and fibroblast growth factor receptor (FGFR), respectively. Other examples of oncogenes are *HER2*, *c-MYC*, *hTERT*, *EGFR*, *VEGFR* and *RAS*. Tumour suppressor genes are important for the delay of the cell division and DNA repair. Normally, tumour suppressor genes act by inhibiting cell growth, promoting apoptosis. The deregulation of these genes prevents abnormal cells to die. Some examples of tumour suppressor genes are *TP53*, *BRCA1*, *BRCA2*, *APC* and *RB1* (Lodish et al. 2000).

Tumour cells progressively acquire characteristics that allow them to continue proliferating and developing malignancy. These characteristics are called the *Hallmarks of Cancer*, and were first proposed in 2000 (Hanahan & Weinberg 2000):

- Sustaining proliferative signalling
- Evading growth suppressors

- Activating invasion and metastasis
- Enabling replicative immortality
- Inducing angiogenesis
- Resisting cell death

The same authors, in 2011, proposed four more new characteristics that are involved in the pathogenesis of cancer (Hanahan & Weinberg 2011). Since none of this new features have been validated, they are called emerging hallmarks:

- Deregulating cellular energetics
- Avoiding immune destruction
- Tumour-promoting inflammation
- Genome instability and mutation

Cancer cells have the ability to invade the surrounding tissues and metastasize to distant location, affecting almost any body part. Several types of cancers can be prevented if avoided exposition to common risk factors, such as tobacco and obesity. Furthermore, a significant percentage of cancers can be cured by surgery, radiotherapy or chemotherapy if detected early (World Health Organization, 2016).

1.2 Breast Cancer

1.2.1 Pathogenesis, histological and molecular subtypes

There are several risk factor that increase the susceptibility to develop breast cancer such as age, diet, genetics, familial history, infections and endocrine factors (endogenous and exogenous) (Abdulkareem 2013; Shah et al. 2014). The high heterogeneity, genetic instability and complexity makes the task to identify the biological mechanism that leads to breast cancer more challenging (Abdulkareem 2013). Different types of breast cancer have different aetio-pathogenesis. Morphologically, breast is essentially constituted by fat tissue and mammary glands. Mammary glands are composed by ducts and lobes, which have smaller sections named lobules. The majority of breast cancers are called carcinomas, and depending on the localization they can be called *in situ*

carcinoma – when localized in the region where it began -, or invasive carcinoma - if it spread to the surrounding tissues. The initiation of the *in situ* carcinomas may be in the lobules – lobular carcinoma – or in the ducts – ductal carcinoma, being the ductal carcinoma *in situ* (DCIS) significantly more common than lobular carcinoma *in situ* (LCIS) (National Cancer Institute, 2016). Histologically, unlike LCIS, DCIS and invasive carcinoma have intra-tumour histological differences and can be divided in five and seven subtypes, respectively:

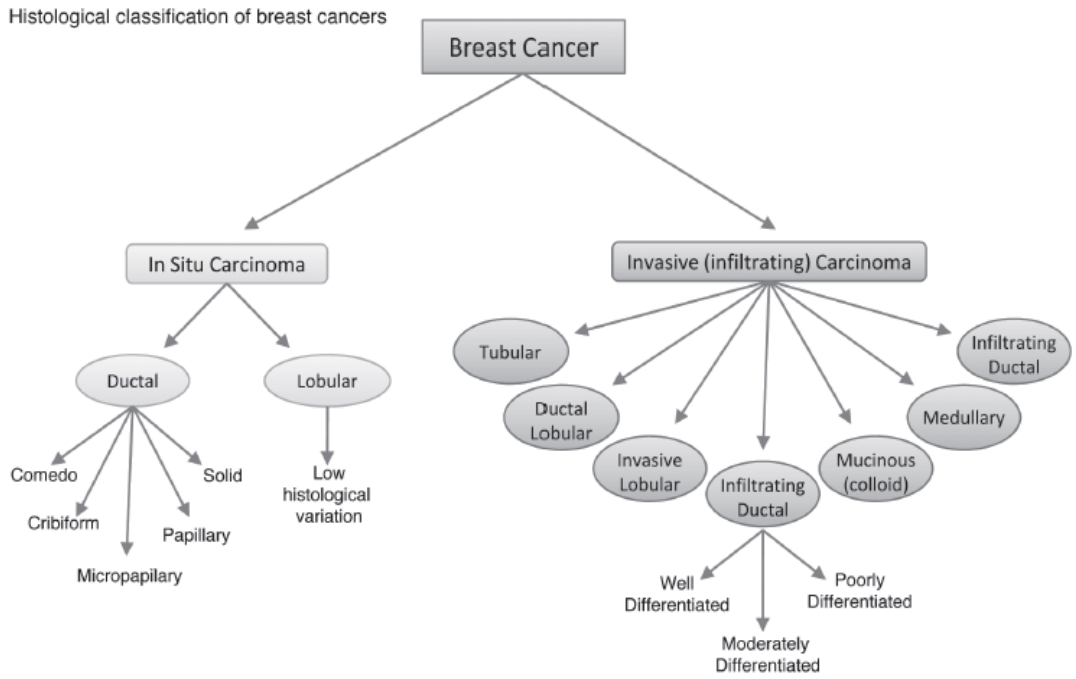


Figure 1.2.1.1 Histological classification of breast cancer. Breast cancer can be catalogued into different subtypes, according to histological features and growth patterns. This system is currently used by clinicians to categorize the heterogeneity found in breast cancer. (Image adapted from Malhotra et al. 2010).

Infiltrating ductal carcinoma (IDC) accounts for 70-80% of all invasive carcinomas. In the clinics, the pathologist analyses the nuclear pleomorphism, glandular/tubule formation and mitotic rate in IDC and ranks it according to grades: well differentiated (Grade 1), moderately differentiated (Grade 2) and poorly differentiated (Grade 3) (Malhotra et al. 2010).

Microarray analysis allowed investigators to understand and establish a molecular profile of gene expression in a tumour (Eroles et al. 2012). Molecular characteristics on cancer cells helped improved personalized medicine, since the

same type of tumour is different between people and within the tumour, culminating with differences in patient survival (Malhotra et al. 2010). Depending on the molecular type of tumour, we can predict the response to a directed treatment (Shah et al. 2014). Breast cancer can be divided in five major molecular subtypes (**Figure 1.2.1.2**) based on gene expression of the tumours: ER+ (oestrogen receptor positive)/Luminal A, Luminal B, Basal-like, ERBB2-enriched (or HER2) and Normal Breast-like (Sorlie et al. 2001; Sorlie et al. 2003).

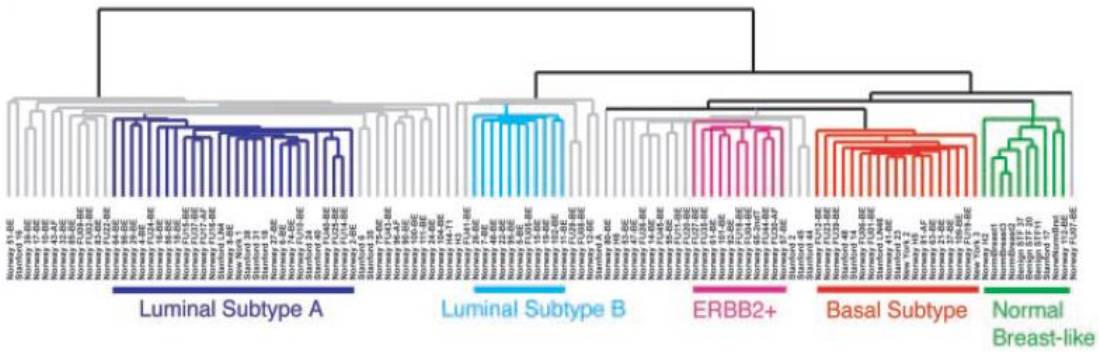


Figure 1.2.1.2 Molecular classification of breast cancer. Breast cancer present different subtypes, according to intrinsic molecular characteristics identified by microarray analysis of patient tumour specimens. Image taken from (Sorlie et al. 2003).

Luminal A is the most common subtype of breast cancer, representing 50-60% of total. It is characterized by an increase in *ER* (oestrogen receptor 1) expression and/or PR^+ (progesterone receptor)/$HER2^-$ status, GATA binding protein 3 and oestrogen-regulated *LIV-1*. It is also associated to low-grade tumours and good prognosis (Eroles et al. 2012; Malhotra et al. 2010; Sorlie et al. 2001).

Luminal B is less common than Luminal A, accounting for approximately 20% of all breast cancer, and it is known for having low levels of ER/PR receptors, being HER2 negative, with high levels of proliferation, and not having a good prognosis (Sorlie et al. 2001; Malhotra et al. 2010).

ERBB2 is an oncogene that encodes for the transmembrane tyrosine kinase growth receptor ERBB2 that is part of the human epidermal growth factor receptor (HER/EGFR/ERBB) family. This gene is overexpressed in 20-30% of all breast tumours and is involved in cell proliferation survival, cell motility, and invasion. ERBB2 positive tumours, where the expression of this gene is amplified, are more

aggressive and, therefore, present a poor prognosis (Shah et al. 2014; Sorlie et al. 2001; Perou et al. 2000).

Basal-like/triple-negative is characterized by the expression of keratin 5, 6 and 17, integrin beta, fatty acids, laminin and for the absence of ER, PR and HER2 expression. Accounting for 3-15% of all breast tumour, this subtype is associated with poor outcome due to the lack of treatment options (Sorlie et al. 2001; Perou et al. 2000; Badve et al. 2011; Sorlie et al. 2003).

Normal breast-like show similarities with normal breast tissue, expressing genes related to the adipose tissue, and other none-epithelial cell types (Sorlie et al. 2001).

More recently, a study has examined gene expression and copy number in 2,000 breast tumours, the METABRIC project (Molecular Taxonomy of Breast Cancer International Consortium), performing a new integrative clustering, based on gene expression and copy number data and the results suggested ten novel molecular subtypes, showed in **Table 1.2.1.1** (Curtis et al. 2012).

Table 1.2.1.1 Breast cancer subtypes classification according to the METABRIC project. This table was accomplished based on data presented by the METABRIC project (Curtis et al. 2012). IntClust, integrative clustering; CNA, copy number aberrations.

Subgroup	Characteristics
IntClust 1	17q23/20q cis-acting luminal B subgroup; relatively good outcome
IntClust 2	High-risk ER+ subgroup; characterized by amplification of 11q13/14, overexpressing genes like *CCND1* and *RSF1*, both previously linked to breast and ovarian cancer; associated with poor prognosis.
IntClust 3	Luminal A cases subgroup, enriched for histotypes that typically have good prognosis; characterized by low genomic instability.
IntClust 4	Includes both ER-positive and ER-negative cases; characterized by low levels of CNA and good prognosis.
IntClust 5	ERBB2-amplified subgroup; characterized by HER2 enrichment (ER-negative) cases and luminal (ER-positive) cases; low prognosis
IntClust 6	8p12 cis-acting luminal subgroup; characterized by amplification of 8p12

IntClust 7	Luminal A subgroup; characterized by 16p gain/16q loss and higher frequencies of 8q amplification
IntClust 8	Luminal A subgroup; characterized by 1q gain/16q loss, a common translocation event
IntClust 9	8q cis-acting/20q-amplified mixed subgroup
IntClust 10	Basal-like cancer enriched subgroup; characterized by high genomic instability and cis-acting alterations, namely, 5 loss/8q gain/10p gain/12p gain; good long-term outcome

These molecular characteristics are important biomarkers that can indicate the patient overall cancer outcome – prognosis biomarker – and the effect of a therapeutic intervention – predictive biomarker – in order to improve diagnosis and treatment for breast cancer (Oldenhuis et al. 2008).

1.2.2 Epidemiology

In 2012, nearly 1.67 million women were diagnosed with breast cancer worldwide, and almost 522,000 of these women died (ranking as the fifth cause of death worldwide), making breast cancer the most common cancer in women (Globocan 2012). In the same year, 463,800 European women were diagnosed with breast cancer, from which 131,200 died (Ferlay et al. 2013). In Portugal, out of 6,088 women diagnosed with breast cancer, 1,570 died. Unfortunately, these numbers have a tendency to increase and by 2050 it is estimated 3.2 million new cases per year worldwide. As expected, the incidence is higher in developed than in undeveloped countries, and this is related mostly to these countries' lifestyle. It is estimated that one in eight women will have breast cancer during their lifetime, in which 89% have more than 40 years. This change in incidence can be also due to an increase in population-based screening, which leads to an early detection and decrease in mortality (Youlden et al. 2012).

1.2.3 Aetiology

1.2.3.1 Environmental /Overall Risk factors

Although screening through mammography does not prevent cancer, it reduces the risk of dying from cancer. Finding cancer at early stages, while asymptomatic, makes the treatment easier and increases long-term survival (Centers for Disease Control and Prevention, 2016). There are some factors that can be avoided and that contribute to an increase risk of breast cancer, such as exposure to carcinogenic substances (e.g. alcohol, tobacco and red meat), oral contraceptives, give birth at older age (after 40s), obesity and exposure to radiation. Investing in protective factors such as physical exercise, keeping a healthy diet, a healthy weight and breastfeeding may decrease the risk of developing breast cancer (Youlden et al. 2012; McPherson et al. 1994). It is also known that breast cancer is strongly related with age, and in fact, this disease affects mostly elderly women after menopause (>54 years).

1.2.3.2 Genetic Susceptibility

Genetic susceptibility is an increased probability of an individual to develop a disease based on their genotype. This genetic inheritance can be triggered by environmental factors, normally at late age. Genetic susceptibility can be classified according to the relative risk that the genetic variant confer and their frequency in the population (**Figure 1.2.3.2.1**) (reviewed in Ghoussaini et al. 2013).

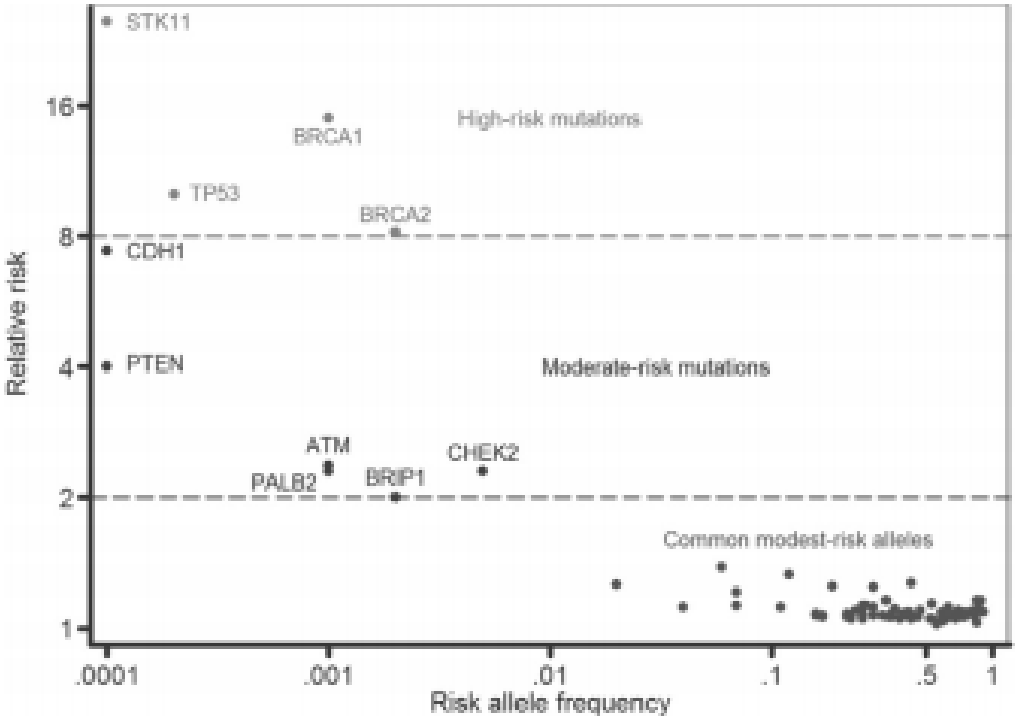


Figure 1.2.3.2.1 Breast cancer genetic susceptibility loci. The relative risks and risk allele frequency for each locus. Higher risk mutations have lower frequency on the population, while common modest-risk alleles confer only a small risk (Ghoussaini et al. 2013).

Inherited factors increase the probability of an individual having cancer due to mutations on the germline cells. About 5-10% of breast cancers are due to genetic predisposition, affecting mainly younger people (Gage et al. 2012). Comparing with the general population, individuals with at least one first-degree relative with breast cancer are two or more times more likely to develop breast cancer. Multiple-case families include a positive familial history with:

- At least three relatives, in the same side of the family, with breast or ovarian cancer
- At least, one first-degree relative
- At least one case per generation
- At least one first-degree relative diagnose at a younger age (<40) (McPherson et al. 1994)

Although those inherited autosomal dominant mutations represent a small amount of the causes for breast cancer, most of them have a significantly high

penetrance - meaning that the individuals that carry those genetic variants have high probability of expressing the phenotype.

1.2.3.2.1 High-risk mutations

Some tumour suppressor genes, such as *STK11/LKB1*, *BRCA1*, *BRCA2* and *TP53,* are involved in the repair of damaged DNA. When any of these genes acquire loss-of-function mutations the resulting protein will not be produced or function properly (Apostolou & Fostira 2013; Ripperger et al. 2009). These high-penetrance alleles increase the risk for developing breast cancer by 10- to 30-fold, through the direct effect of the mutation. Although mutations in these genes are rare in the population (have low frequency, <1%), they confer a significant lifetime risk for breast cancer (>50%) (**Figure 1.2.3.2.1**) (Ghoussaini et al. 2013; Fletcher & Houlston 2010). Twenty five percent of the familial cases of breast cancer are explained by high risk mutations, being 16% due to *BRCA1* and *BRCA2* germline mutations (Van Der Groep et al. 2011). In fact, multiple-case family studies have shown that by the age of 70, approximately 80% of the carriers of germline mutations in *BRCA1* and *BRCA2* genes would develop this type of cancer (Milne & Antoniou 2011). The studies and approaches that allowed the identification of these mutations were: i) linkage analysis, that provides statistical evidence of the contribution of a variant or gene in the disease aetiology within families (Ott et al. 2015; Aloraifi et al. 2015); ii) positional cloning, which helps identifying the causal genetic mutations of diseases with simple Mendelian inheritance (Puliti et al. 2007); and iii) DNA resequencing of candidate genes (Fletcher & Houlston 2010).

Indeed, there are studies that show the existence of differences in penetrance within the carrier families, suggesting that there are other factors, such as environmental and genetic modifiers, influencing the risk (Begg et al. 2009; Milne & Antoniou 2011; Ripperger et al. 2009). It has been described that polymorphisms can have an effect in genes, altering their expression, making them an important tool to predict the risk associated with cancer, and,

furthermore, this may lead to new therapeutic methodologies for breast cancer patients (Maia et al. 2012; Milne & Antoniou 2011).

1.2.3.2.2 Moderate-risk mutations

Genetics variants in *ATM*, *CHEK2*, *PALB2*, *BRIP1*, *PTEN* and *CDH1*, also involved in DNA repair, increase moderately the relative risk for breast cancer (approximately two-fold), conferring an higher probability of an individual to develop breast cancer in their lifetime of approximately ≥20% (**Figure 1.2.3.2.1**) (Ghoussaini et al. 2013; Hindorff et al. 2011). DNA resequencing of candidate genes for coding variation, using genetically enriched cases, allowed the identification of these variants (Fletcher & Houlston 2010).

1.2.3.2.3 Common low-susceptibility alleles and GWAS

These alleles are common in the population, with a minor allele frequency (MAF) of >5%, and confer a modest to low relative risk, corresponding to <1.5-fold and a lifetime risk of 10-20% (**Figure 1.2.3.2.1**) (Ghoussaini et al. 2013). These polymorphisms are usually found by genome-wide association studies (GWAS), that are a type of study that analyses DNA sequence variations through the entire human genome, aiming to identify genetic risk factors for diseases that are common in the population (Bush & Moore 2012; Knight 2014). For breast cancer, recent studies identified so far 94 loci, that explained about 14% of the inherited risk for breast cancer (Michailidou et al. 2015). Of the 94 loci, 13 were studied at a functional level and all suggested that these polymorphisms can modify the expression of genes in a allele-specific manner, namely *MAP3K1*, *CCDC170*, *ZNF365*, *CASP8*, *CCDN1*, *FGFR2*, *MDM4* and can modify breast cancer related genes, such as *BRCA1* and *BRCA2* (Ghoussaini et al. 2013; Michailidou et al. 2015; Ripperger et al. 2009; Maia et al. 2012; Glubb et al. 2015; Cai et al. 2011; Shephard et al. 2009; Wang et al. 2014; French et al. 2013; Meyer et al. 2008; Meyer et al. 2013; Wynendaele et al. 2010). Additionally, the underlying mechanisms of these susceptibility polymorphisms are still unresolved.

Single nucleotide polymorphisms (SNPs) are spread across the genome. GWAS use marker SNPs as genetic markers for a certain genomic region and allow the association between numerous SNPs and a specific phenotype or disease (Bush & Moore 2012), by using cases and controls samples. Under the assumption that when the GWAS SNP is not the causative genetic variant and that the actual causal SNP is in high *linkage disequilibrium* (LD) with the marker SNP, GWAS uses LD as a measurement to correlated the genotypes of two different SNPs (**Figure 1.2.3.2.3.1**). LD refers to when two or more markers on a chromosome are transmitted together within a population, during chromosome segregation in cell division, forming haplotypes. Thereupon, this non-random association between alleles at two or more loci, is commonly measured by two parameters - D prime (D') and R square (r^2) which compare the observed frequencies of co-occurrence for two alleles in a population with the frequencies expected if the two markers were independent (Bush & Moore 2012; Morton 2005). D' varies between zero and one, corresponding to *linkage equilibrium* when the recombination between two or more markers is elevated, and to *linkage disequilibrium*, when there is no recombination between the two markers. Coupled with D', high values of r^2 (also scaled between zero and one) indicate that the two markers transmit similar information. Therefore, it only takes one genotyped marker to find the allelic variation of the other (Bush & Moore 2012).

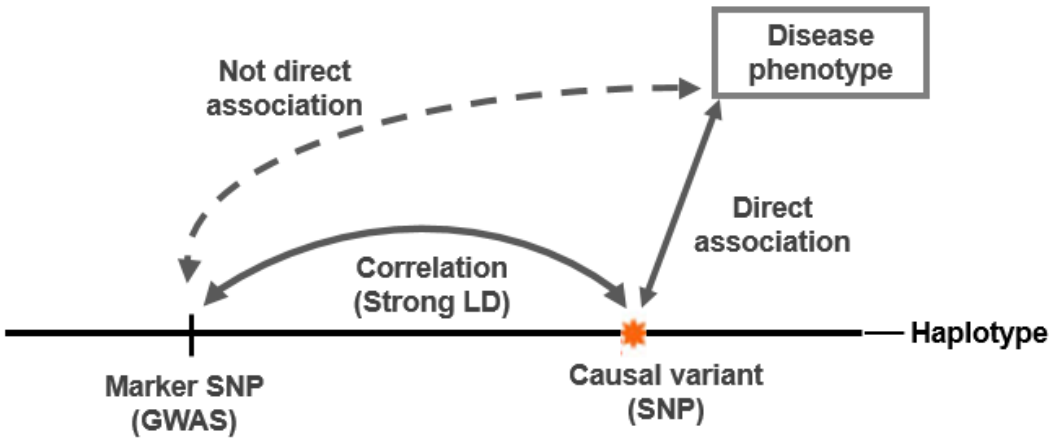


Figure 1.2.3.2.3.1 GWAS approach for identification of the causal SNP. A GWAS marker SNP in strong LD with a common causal variant, the true responsible for the phenotype, will report the causal SNP. Image obtained and adapted from (Balding 2006).

In other words, GWAS uses a marker SNPs that report the association with another SNP when both are in strong LD with each other. In this way, most of all genome sequence is covered only with a small portion of known markers SNPs. More loci associated with risk are yet to be identified, and is expected to be due to common-low susceptibility alleles (Galvan et al. 2010). However, the polymorphisms associated to a certain phenotype found in a GWAS, are rarely the individual causative polymorphism, and therefore it is necessary to improve the current approaches to be able to find the true cause of the observed phenotype and the intrinsic mechanism (Knight 2014; Consortium 2015).

1.3 Genetic Variation/Polymorphisms

Comparing DNA sequences from different individuals, it has been estimated that in every few hundred bases there is a genetic polymorphism. Considering that there are 3.2 billion nucleotides in the human genome and that they can be responsible for alterations in gene expression, it is of extreme importance to study these variations effect on differences in treatment response and disease risk (Stoneking 2001). Unlike mutations, these variations are common, with >1% of allele frequency in the population (Torkamani & Schork 2008).

Within DNA variations, the most common are SNPs (~90%), tandem repeat segments (minisatellite (10-100 bp) and microsatellite (1-6 bp)), and large and small duplications/deletions/insertions (Wang et al. 2005). SNPs are variations of one nucleotide in the DNA sequence with the ability to regulate gene expression. They can also be found in protein-coding regions, and depending on whether the SNP changes or not the encoded amino acid in the final protein it is classified as non-synonymous or synonymous, respectively (Torkamani & Schork 2008). Since SNPs are present in a significant proportion of human populations, they are considered common genetic variation. Also, most of the SNPs have two alleles and their frequency is represented by the minor allele frequency (MAF) rather than the more common allele frequency (the major allele) (Bush & Moore 2012).

1.3.1 Cis-Acting Regulatory Variants

Gene expression is regulated by environmental, epigenetic and genetic factors that act both in *cis* and in *trans* (Jones & Swallow 2011). Cis-regulation is the mechanism by which a variation in the DNA sequence affects the expression of a gene, by modulating the binding affinity of transcription factors, mRNA stability, methylation and splicing, for example. Meanwhile, trans-regulation relates to the effect of proteins regulating other gene expression, such as transcription factors. Variants in cis-regulatory elements, like promoters, enhancers, silencers and insulators can disrupt or enhance the binding affinity of transcription factors and can lead to unequal levels of transcription between the two alleles of a gene (**Figure 1.3.1.1**).

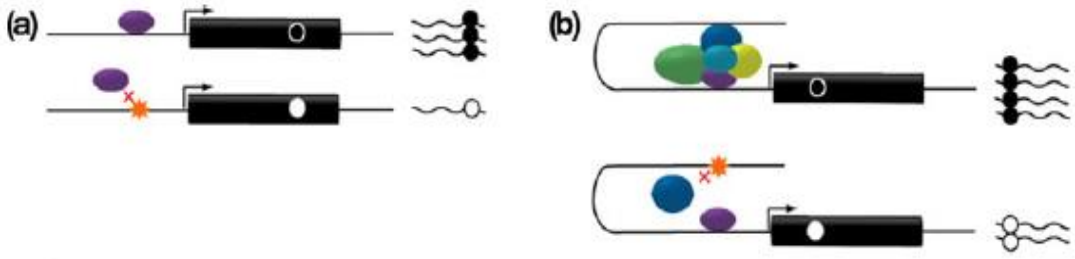


Figure 1.3.1.1 Cis-acting regulatory variation causing differential allelic expression. a) A variant in a proximal promoter may prevent transcription factor binding altering expression of the allelic transcript. Transcript SNPs (markers) (shown here as black/white circles) can be used to determine transcript ratios. b) A variant in a distal enhancer site may prevent complex binding and affect transcription levels. Image adapted from (Jones & Swallow 2011).

SNPs can be anywhere in the genome and may affect, for example, the binding affinity of proteins in an allele-specific manner. One of those examples, found through GWAS, is present in the risk-loci *FGFR2*, where it was shown in a functional analysis that the risk allele was associated with increased expression of this gene when compared to the common allele (Meyer et al. 2008). Another example is *MAP3K1*, whose risk alleles are also associated with increased gene expression (Glubb et al. 2015). MicroRNAs (miRNA) and methylation can also be affected by genetic alterations such as SNPs. miRNAs are small non-coding RNAs with approximately twenty two nucleotides, that regulate gene expression by binding to mRNA and preventing translation (synthesis of a protein) or by

promoting mRNA cleavage and destabilization (Liu et al. 2012). Roughly 3% of genes represent miRNA and 30% of coding genes can be affected by miRNA (Sassen et al. 2008). The presence of a SNP in the 3'UTR of a mRNA might affect the miRNA binding and, consequently, the mRNA translation (Liu et al. 2012). For example, it was shown that the target site of miR-125b, in the 3'UTR of the gene *BMPR1B*, which encode a kinase, contains a SNP (rs1434536) and that miR-125b differently binds to the C and T alleles of this SNP in breast cancer (Sætrom et al. 2009).

As well as miRNA, chemical modification of DNA and histones is also involved in gene expression regulation. DNA is packed around histone proteins (H1, H2A, H2B, H3, and H4 histones) forming the chromatin, that can suffer modifications such as methylation, acetylation, phosphorylation, ubiquitylation and sumolyation, specifically on lysine residues (K) of histones H3 and H4 (Hellman & Chess 2010; Ellis et al. 2009; Handy et al. 2011). The histone code hypothesis refer that the expression of the DNA information is partially regulated by these modifications. This epigenetic regulation can be complex since each histone can be modified simultaneously with different histone marks at multiple sites. Each histone has different number of lysine (that can be mono-, di- or tri-methylated or acetylated), arginine (that can be methylated) or threonine/serine/tyrosine (that can be phosphorylated). Therefore, it is probable that every nucleosome in a cell presents different modifications. In fact, in a recent study where they analysed 39 histone modifications in human CD4+ T cells, a group have shown that patterns of modifications can occur on the genome, and most of those modifications were associated with promoters and enhancers, suggesting a role of histone modifications in transcriptional regulation (Wang et al. 2008; Handy et al. 2011; Bannister & Kouzarides 2011). The DNA methylation usually occurs in CpG islands (areas in the DNA sequence rich in C (cytosine) and G (guanine) dinucleotides, found frequently in promoters), namely in the C nucleotide, and it might be associated with gene expression repression due to the steric inhibition of regulatory transcription complexes binding to DNA (Handy et al. 2011; Ellis et al. 2009).

Also, the accessibility to the chromatin for transcription factor binding varies according to the chromatin states, that can be open (euchromatin) or compact

(heterochromatin). Thus, the chromatin state is controlled by histone modifications. Since active regulatory elements are located in regions with open chromatin, that is, accessible to the transcription machinery, these DNA sites are highly sensitive to *DNase* I, an enzyme that digests the DNA strand (Jin et al. 2015). Therefore, if a SNP is present in a *DNase* I Hypersensitive site (DHS), it may cause differences in transcription factor binding between the two alleles and lead to different levels of expression (Schaub et al. 2012).

1.3.2 Differential Allelic Expression

As stated before, regulatory SNPs or rSNPs may lead to different levels of expression between the two alleles of a gene (Maia et al. 2012; Jones & Swallow 2011).

Currently, two approaches are used to detect these differences of expression, namely, expression quantitative trait loci (eQTL) and differential allelic expression (DAE). eQTL provide us information about overall expression (mRNA) of a gene, making association between markers of genetic variation with gene expression levels typically measured in tens or hundreds of individuals. One of the advantages of eQTL is that it allows the identification of new functional loci, through GWAS, without having previous knowledge of specific *cis* or *trans* regulatory regions. However, it does not inform us which allele is causing the difference on expression levels (**Figure 1.3.2.1**) (Pastinen 2010). On the other hand, DAE approach is an allele-specific study, and the differences in expression between the two alleles due to the presence of a rSNP, can be quantified in heterozygous individuals as a ratio of the expression of one allele compared with the other, using transcribed SNPs (DAE SNPs) as allelic markers (**Figure 1.3.1.1** and **Figure 1.3.2.1**). DAE also allows the elimination of environmental or trans-factors that can modify both alleles, since it is focused on the transcribed alleles individually (Pastinen 2010).

Indeed, our research group performed a whole-genome mapping of cis-regulation in normal breast samples using DAE, and these results suggest that approximately 87 % of genes expressed in normal breast tissue are affected by

regulatory SNPs (rSNPs) (Xavier *et al*, unpublished). Another study has used eQTL analyses with information from The Cancer Genome Atlas (TCGA), regarding gene expression in ER+ breast cancer. They conclude that 1.2% of gene expression variance was due to cis-acting SNP loci, which corresponded to 189 out of 15,732 tested genes (Li et al. 2013).

Therefore, here we focus on DAE studies, considering it is more accurate in detecting cis-regulatory loci and in mapping the causal regulatory variant (rSNP), enabling to identify which allele is conferring the up- or down-regulation of a specific gene.

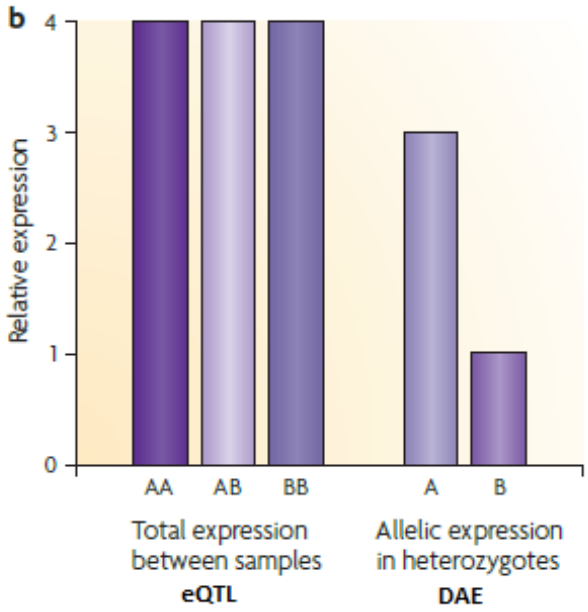


Figure 1.3.2.1 Differences between eQTL and DAE. In eQTL, the overall expression is equal even though the gene is being cis-regulated. In contrast, DAE shows the intrinsic difference between the alleles, in individuals heterozygous for a regulatory variant. Image adapted from (Pastinen 2010).

1.3.3 Previous work – DAE map in normal breast tissue

Previous work developed by Professor Ana Teresa Maia and her colleagues, consisted in a DAE scan of the entire genome. This was accomplished using microarrays (Illumina Exon510S-Duo arrays), and 64 normal breast tissue samples, which were genotyped and quantified for allelic expression. The result was a whole-genome map of cis-regulated genes (86.8% of the autosomal

genes), with 49,461 DAE SNPs located in 11617 genes, in breast tissue (**Figure 1.3.3.1**).

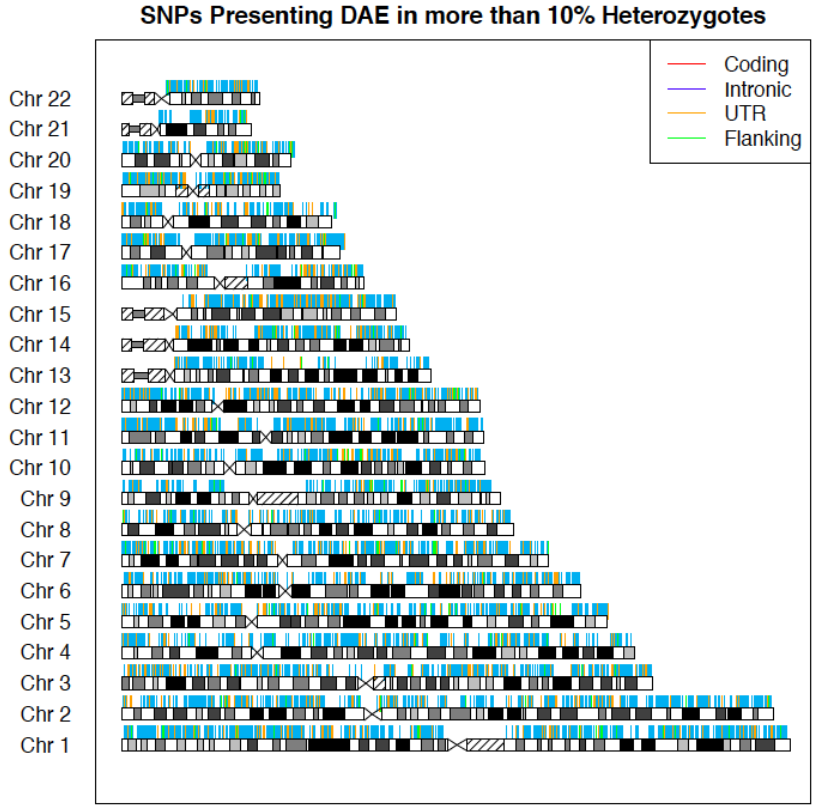


Figure 1.3.3.1 Global cis-regulation map of breast tissue. This map gives us information about which genes are being cis-regulated and, therefore, presenting differential allelic expression (Maia *et al*, unpublished).

The DAE measured in tSNPs (DAE SNPs) can be explained by the effect of an rSNP that differentially regulates their expression. Depending on the different levels of LD (measured by D' and r^2) between the rSNP and the tSNP, three patterns of DAE distribution can be seen in this map, consistent with the scenarios 1, 2 and 3 described by Xiao and Scott (Xiao & Scott 2011):

- Scenario 1, if the tSNP is in complete LD (r^2=1) with the rSNP. All heterozygous samples for the tSNP will show DAE (**Figure 1.3.3.2**. **A**) with the same allele being preferentially expressed.
- Scenario 2, when the LD between the tSNP and rSNP is not complete, but strong (r^2<1, D'=1). In this case, individuals heterozygous for the tSNP (Tt) might be homozygous (RR) or heterozygous (Rr) for the rSNP (**Figure 1.3.3.2**. **B**), and therefore some heterozygotes for the tSNP will show DAE

(those heterozygous for the rSNP) and some will not (homozygous for the rSNP)

- Scenario 3, when the tSNP and rSNP are in low LD (r^2<1, D'<1), it is possible to find four combinations of the heterozygous tSNP with the rSNP in the population (**Figure 1.3.3.2**. **C**). Therefore, some individuals will not display DAE (the ones homozygous RR or rr for the rSNP) and others will display preferential expression of one or another allele (Rr or rR for the rSNP).

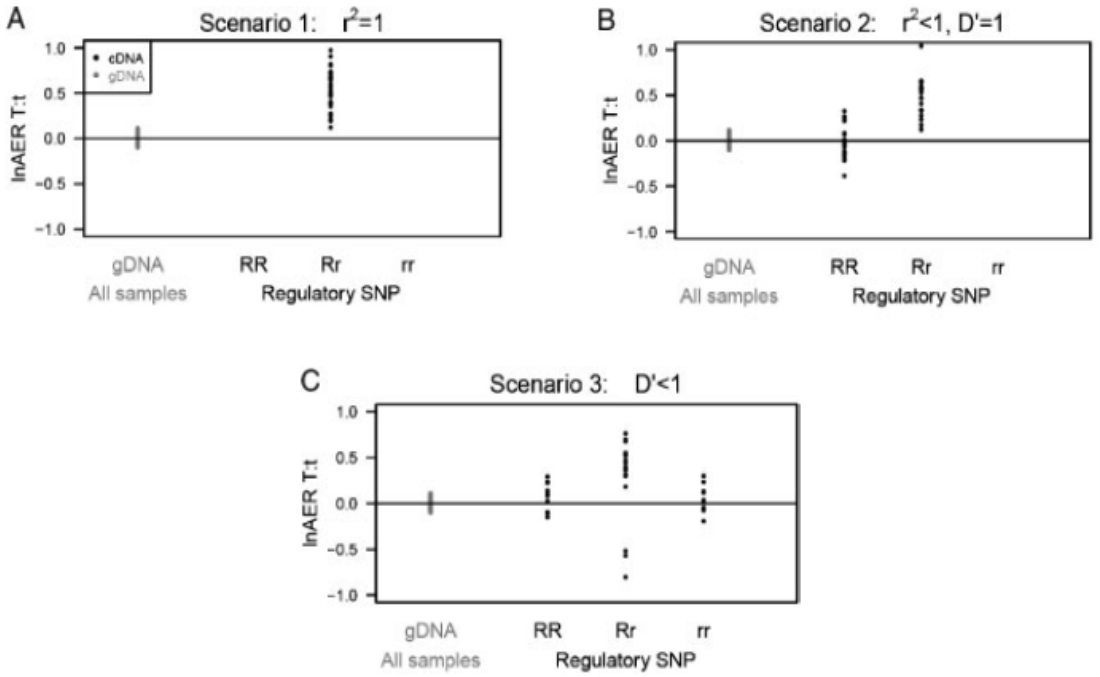


Figure 1.3.3.2 Patterns for different LD measurements between rSNP and a heterozygous DAE SNP. Image taken from (Xiao & Scott 2011).

Furthermore, with the aim to prioritize the best candidates for cis-acting regulatory SNPs in breast cancer, a member of our research group (Doctor Joana Xavier) cross-compared the DAE data (Maia *et al*, unpublished) with the published (94 loci associated with risk for breast cancer) and unpublished (reported in a GWAS late phase) GWAS data. This integration was done by identifying loci that had at least a GWAS SNP and a DAE SNP within 250kb away from each other and with a minimum LD of r^2 = 0.4. This generated a list of 111 clusters with strong cis-regulatory potential in breast tissue, where in 32 of them the GWAS SNP and the

DAE SNP are in high LD. In the end, one cluster was prioritized – 12q24 locus – that contains a GWAS SNP that did not pass the last phase of the GWAS (with a *p*-value = 0.002), since the threshold was $1x10^{-7}$, not being associated to breast cancer risk. Since GWASes have a produce a list of unpublished SNPs to validate, the integration of our DAE data with the GWAS data for breast cancer, was also a way to prioritize the loci with genes being cis-regulated by cis-regulatory variants, and therefore, more likely to be associated with breast cancer risk, for further validation studies. Therefore, we wanted to validate this unpublished locus to breast cancer risk, by identifying the cis-regulatory variants causing DAE, and further re-test the GWAS SNP with those candidate cis-regulatory variants, in order to associate the 12q24 locus to breast cancer risk. This way, the aim of this study was to validate an unpublished GWAS locus to confirm it as a new risk locus for breast cancer.

.

2 Hypothesis

One of the limitations of GWASes is that they are unable to identify the true causal variant or the mechanism conferring risk for breast. Also, all variants found so far associated with risk for breast cancer were located in non-coding regions, suggesting that breast cancer associated variants may be mainly located in regulatory elements. To this date, 13 of the 94 loci were studied at a functional level by other research groups, and all of these causative SNPs were cis-regulatory (Wynendaele et al. 2010; Ghoussaini et al. 2014; Gorbatenko et al. 2014; Quigley et al. 2014; Glubb et al. 2015; Wang et al. 2014; Hurtado 2013; Dunning et al. 2016; Cai et al. 2011; Meyer et al. 2008; French et al. 2013; Huijts et al. 2011; Long et al. 2010; Cowper-sal et al. 2012). Thus, we hypothesize that cis-regulation is an important mechanism, contributing to the risk for breast cancer, and that cis-acting variants are responsible for the DAE observed.

Additionally, GWAS have generated long lists of SNPs that were very close to reach genome-wide significance, and need to be validated to confirm their association with risk. Being DAE one of the effect observed in the presence of cis-regulatory SNPs, this makes it a powerful method to identify these SNPs. Therefore, integrating our DAE results with the GWAS unpublished and published data may be a powerful approach to prioritize loci to validate, with higher probability to be associated with risk for breast cancer and gene expression regulation.

3 Objective & Specific Aims

For this master thesis, our main objective was to test if our DAE studies are a powerful tool to prioritize unpublished GWAS loci for validation studies, in order to help identifying further risk loci associated with breast cancer. Our specific aims were:

1. To find new candidate cis-acting regulatory SNPs in the 12q24 locus;
2. To functionally analyse their regulatory potential;
3. To use them to validate the unpublished 12q24 locus and its association with breast cancer risk

4 Materials and methods

4.1 Study samples

In this work we studied a total of 290 samples. Eighty-four (84) samples were from normal breast tissue, extracted from women whose reduction mastectomy was performed for reasons not related to cancer. These normal breast tissue samples were collected at Addenbrooke's Hospital, Cambridge, United Kingdom. Additionally, 150 samples of Human B cells (blood) were extracted from anonymous blood donors and 56 samples were extracted from cancer patient B cells (blood), both obtained from Blood Centre at Addenbrooke's Hospital. All samples referred were acquired with the approval of the Addenbrooke's Hospital Local Research Ethics Committee (REC reference 04/Q0108/21 and 06/Q0108/221).

DNA and total RNA was previously extracted from all samples using a conventional SDS/proteinase K/phenol method and TRizol® method, respectively. All extraction procedures were done at the University of Cambridge and the extracted RNA was used for DAE analysis.

4.2 Cell lines

In vitro assays were made with nuclear extract from breast cancer cell lines, namely, T-47D (human mammary ductal carcinoma, oestrogen receptor positive (ER+)), HCC1954 (human mammary ductal carcinoma, an oestrogen receptor negative (ER-)), MCF-7 (human mammary adenocarcinoma, (ER+)) and MDA-MB-231 (human mammary adenocarcinoma, (ER-)) cell lines (**Table 4.2.1**). Nuclear extract from T-47D and HCC1954 cell lines were available in our stock. MCF-7 and MDA-MB-231 were cultured in DMEM medium at 37ºC and supplemented with penicillin/streptomycin to avoid contaminations and 10% foetal bovine serum, which is rich in growth factors, allowing the cells to grow, divide and survive. All cell lines were obtained from our collection.

The normal breast cell lines analysed *in silico* were HMEC (human mammary epithelial cells), HMF (human mammary fibroblasts), MCF10A (human mammary epithelial cells), BR.MYO (breast myoepithelial primary cells) and BR.H35 (breast variant HMEC).

Table 4.2.1 List of breast cancer cell lines analysed. IDC, invasive ductal carcinoma; AC, adenocarcinoma; DC, ductal carcinoma. ER/PR/ERBB2/TP53/EGFR status: ER/PR positivity, ERBB2/EGFR overexpression and TP53 mutational status and protein levels. WT, wild-type. Information from (Neve et al. 2006) and ATCC website (www.lgcstandards-atcc.org).

Cell line	Cell line Type	Tumour type	Origin of cells	ER	PR	*ERBB2* Amplification	*TP53* Mutation	*EGFR* Amplification
MCF-7	Human mammary adenocarcinoma	IDC	Metastatic site (pleural effusion)	Yes	Yes	No	+/- WT	Low (Xing et al. 2010; Mamot et al. 2003)
MDA-MB-231	Human mammary adenocarcinoma	AC	Metastatic site (pleural effusion)	No	No	No	Yes	Yes (high) (Xing et al. 2010)
T-47D	Human mammary ductal carcinoma	IDC	Metastatic site (pleural effusion)	Yes	Yes	No	Yes	–
HCC1954	Human mammary ductal carcinoma	DC	Primary stage IIA, grade 3 IDC	No	No	Yes	+/-	Yes (Henjes et al. 2012)

4.3 Linkage disequilibrium analyses and identification of proxy SNPs

The publicly available tool SNAP (SNP Annotation and Proxy Search) was used to measure the LD between SNPs and for the identification of proxy SNPs. Proxy SNPs are SNPs in high LD that report each other. Thus, when a candidate SNP is not available on a particular genotyping array, proxy SNPs in LD with that candidate SNP can report it, based on observed LD patterns in the International HapMap Project (HapMap) and 1000 Genome Project (Johnson et al. 2008; The 1000 Genomes Project Consortium 2012). Both projects were developed with the aim of identify and catalogue genetic variants with frequencies of at least 1% in the populations studied. We only use information of European (CEU) studies from 1000 Genome Project and a maximum limit distance between the SNPs of 500kb (kilo-base).

4.4 *In Silico* annotation of variants functional information

In silico annotation was performed in order to gather information regarding regulatory chromatin states, haplotype structures, DNase I Hypersensitive sites (DHSs), histone modifications, protein weight matrix previsions (PWM), microRNA binding predictions and chromatin immunoprecipitation sequencing (ChIP-seq), relatively to the candidate rSNPs.

For our analysis, histone modifications were used to identify regulatory elements, since they are responsible for controlling the accessibility for protein binding (**Table 4.4.1**) (Handy et al. 2011). We also analysed DHSs data, since it is a powerful method to identify transcriptional regulatory elements and the chromatin states (**Table 4.4.1**).

Table 4.4.1 Histone modifications. Examples of known histone modifications in breast tissue, where they are found genomically and some of their effect. Information gathered from Roadmap and UCSC Genome Browser in the scope of this work.

Modifications	Effect
H3K4me3	Active promotor
H3K4me1	Active and inactive enhancers – mostly intergenic regions
H3K4me2	Active promotor
H3K27me3	Repressive
H3K27ac	Active promotor and active enhancer
H3K36me3 and H3K79me2	Transcriptional repression
H3K9me3	Repeat repression
H3K9ac	Active mark

ChIP-seq consists in a combination of two techniques, which are chromatin immunoprecipitation and sequencing. In this method, DNA-protein complexes are precipitated with a specific antibody that recognizes the target protein followed by DNA sequencing, allowing the identification of protein-DNA binding sites (Mardis 2007). We gathered information regarding ChIP-seq in Haploreg v4.1 database and RegulomeDB.

PWM is a probabilistic model that provides predictions concerning transcription factor binding consensus sites in a certain DNA sequence (Chen et al. 2007). It

is helpful when no ChIP-seq data is available for a certain locus or SNPs position. Haploreg v4.1 was accessed to consult PWM results.

Besides these analyses, the following databases were accessed:

- Haploreg version 4.1 (Ward & Kellis 2011) – it was used to access information regarding regulatory elements at variants on haplotype blocks, such as candidate regulatory SNPs at disease-associated loci. Using genotyping information from the 1000 Genome Project to analyse the LD structure in each locus, Haploreg provides information about SNPs that are highly correlated with the candidate SNP. Also, it gives information about PWM, chromatin state, sequence conservation across mammals, regulatory motifs of SNPs and DHSs.
- UCSC Genome Browser (Kent et al. 2002) – it is a genomic browser that allows the visualisation of the genomic landscape of the candidate regulatory SNP concerning histone modifications, DHSs, chromatin state and transcription factor ChIP-seq. This interactive website gathers information from the Roadmap Epigenomics Project (Chadwick 2012) – a public resource of human epigenomic data such as DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in stem cells and *ex vivo* primary tissues – and the ENCODE Project (Encode Consortium 2012), regarding information about regulatory elements in the DNA, including ChIP-seq, DHSs and chromatin state.
- Integrative Genome Viewer (IGV) – it is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets (Robinson et al. 2011; Thorvaldsdóttir et al. 2013). Here, we analysed the results of several ChIP-seq experiments, to verify the intensity of the DNA-protein binding and the binding affinity of the protein towards the two alleles of a heterozygous SNP (comparing the reads for each allele).
- RegulomeDB (Boyle et al. 2012) – it is a website that collects information from Roadmap Epigenomics project and ENCODE project. Here, we analysed the candidate regulatory SNPs for known and predicted regulatory DNA elements including regions of DHSs, binding sites of transcription factors and promoter regions.

- Haploview software (Barrett et al. 2005) – it was used to analyse pair-wise LD and possible haplotype structure between the candidate SNPs, in our 64 samples of normal breast tissue.
- For microRNA analysis the following databases were accessed:
    - NCBI dbSNP (Sherry et al. 2001) – is a database of single nucleotide polymorphisms (SNPs) and multiple small-scale variations that include insertions/deletions, microsatellites. This database was used to obtain the sequence where each candidate SNP was located.
    - Ensembl (Yates et al. 2016) – is a genome browser with information, among others, sequence variation and transcriptional regulation. Ensembl tools include BLAST, BLAT, and the Variant Effect Predictor (VEP). This browser was used to find the location of the candidate SNPs in the gene in study.
    - miRBase database (Griffiths-Jones et al. 2008) – is a database of published miRNA sequences and annotation. Each entry in the miRBase Sequence database represents a predicted hairpin portion of a miRNA transcript (termed mir in the database). Here, we analysed the candidate SNPs for predicted microRNA binding sites.

4.5 DAE mapping analysis

DAE mapping analysis was done by plotting the DAE ratios (of the DAE tSNP) against the genotypes of the candidate rSNPs. Different statistical tests were used, depending on the DAE scenario observed. We used a t-test and Welch's-test to verify differences in DAE mean of the genotype groups, when variances were equal or unequal, respectively. We also used the F-test to verify differences in the variance between the homozygous and heterozygous groups for the candidate rSNP. We used permutation to correct the *p*-values. If the *p*-value is true, even when a 1000 permutation is applied (i.e. a 1000 combinations of the same samples) the outcome/results will be the same. With 1000 permutations the smallest possible *p*-value is 0.001, and the uncertainty *p*-value is 0.05. Genotype

imputation was performed by Doctor Joana Xavier, to cover more candidate regulatory SNPs, which consists in having reference haplotypes, obtained from HapMap and 1000 Genomes Project, and use it to predict genotypes at SNPs that were not directly assayed in individuals samples (imputation). In resume, the DAE mapping analysis was performed to analyse if the candidate rSNPs were associated with the DAE levels measured at the DAE SNPs.

The DAE ratio was calculated, both in cDNA (complementary DNA) and gDNA (genomic DNA), using the formula:

$$\mathbf{DAE} = log_2\left(\frac{Allele\ A}{Allele\ B}\right).$$

The gDNA was used to normalize the results, excluding other events in the DNA sequence, such as copy number variations, that might be also causing unequal levels of expression between the alleles of a gene, and in this work we only focus on cis-regulatory variants.

$$\mathbf{Normalized\ DAE} = \mathbf{DAE_{cDNA}} - \mathbf{DAE_{gDNA}}$$

Thus, this analysis results shows the DAE caused only by cis-regulation.

4.6 Polymerase Chain Reaction (PCR) for genotyping rs111549985

PCR is a technique that allows the amplification of a specific segment of the DNA. To perform this technique five components are necessary, namely the template DNA to amplify, primers to set up the beginning and ending of the fragment to be amplified, deoxynucleotides (dNTPs) that form the new strands of the PCR product, and DNA polymerase an enzyme that synthetizes the PCR product. It is composed by three phases (denaturation, annealing and extension) where temperature varies, and the final product will be billions of copies (amplification of each fragment: 2^n, n = number of cycles) of a specific DNA fragment, which can then be separated based on its size, using agarose gel electrophoresis. The agarose gel electrophoresis consists in adding the PCR product (DNA fragments) in an agarose gel and applying electric current. This way, it is possible to separate the DNA products on the basis of size, allowing the determination of the presence

and the size of the PCR product after the addition of a DNA stain, in the agarose gel when exposed to UV (ultraviolet) light (Lilit Garibyan 2013).

In this work, primers were designed to amplify the region containing the SNP rs111549985 (**Table 4.6.1**). The PCR was made using KAPA2G Fast ReadyMix PCR kit (from Kapa Biosystems) or KAPATaq HotStart (from Kapa Biosystems) to genotype 51 normal breast samples. Both master mixes were prepared according to manufacturer's instructions. The cycles set up was: 95ºC/3min (initial denaturation), 95ºC/15sec (denaturation), 60ºC-58ºC/15sec (annealing), 72ºC/2sec (extension) and 72ºC/0.3sec (final extension). Denaturation, annealing and extension steps were repeated for 30 cycles. The agarose gel was prepared in a final concentration of 1.5% in 0.5X TBE and the electrophoresis was carried for 40min at 100V. Four ng of genomic DNA from each normal breast tissue sample were used, and water was used as a negative control (a non-template control). Three µL of RedSafe (DNA stain) was added to agarose gel solution and we used Bio-Rad ChemiDoc to visualize the gel, to confirm the amplification and to verify the presence or not of contamination. The samples selected to sequence (by Sanger Sequencing) were then purified with Exo/SAP Go – PCR Purification kit (from GRiSP Research Solutions). The samples concentration was measured in a Nanodrop 2000c Spectrophotometer (Thermo Scientific) and then diluted to 80ng/µL and 50ng/µL.

SNP	Alleles	Strand	Sequence
rs111549985	C	FWD	GCTCTCCTCCCCCTGGCCCCGTCGCCCCGCCCTCGCC
		REV	GGCGAGGGCGGGGCGACGGGGCCAGGGGGAGGAGAGC
	G	FWD	GCTCTCCTCCCCCTGGCCGCGTCGCCCCGCCCTCGCC
		REV	GGCGAGGGCGGGGCGACGAGGCCAGGGGGAGGAGAGC

Table 4.6.1 Primers sequence designed for PCR. In the table is represented the selected SNP of the 5q14.2 locus to analyse. The common allele is shown first and the minor allele is shown second. For each allele of this SNP, the forward (FWD) and the reverse (REV) sequences were designed.

4.7 Nuclear protein extraction

The nuclear protein extraction was made using the NE-PER Nuclear and Cytoplasmic Extraction Reagent kit (Thermo Scientific) following the protocol stated by the manufacturer. Briefly, after scraping the cells in culture, two reagents were added to the cell pellet – CER (cytoplasmic extraction reagent) I and II – which disrupt the cell membrane, release the cytoplasmic contents, and leave the nucleus intact. NER (nuclear extraction reagent) was then added to lyse the nuclear membrane yielding the nuclear components. The concentration of the nuclear extract was measured using the Qubit 2.0 (Invitrogen by Life Technologies) spectrophotometer, according to the manufacturer's instructions. The nuclear extract was then used to test the protein binding affinity in the two alleles of the candidate rSNP rs12302714.

4.8 Electrophoretic Mobility Shift Assay (EMSA)

EMSA is an *in vitro* technique that allows the study of DNA and protein interactions. Under the observation that the electrophoretic mobility of DNA-protein is slower than the free DNA, due to their molecular weight, in a non-denaturing polyacrylamide gel, it is possible to obtain information regarding binding affinity (Hellman & Fried 2007).

This method consists in labelling a double-stranded oligonucleotide with a fluorescent marker, such as biotin or radioactive isotopes, and adding it to a nuclear extract. If proteins bind to the labelled sequence, the formed complex will migrate slowly through the polyacrylamide gel when compared to the free oligonucleotide, producing a specific band of higher molecular weight (shift) (**Figure 4.8.1**). Additionally, an unlabelled oligonucleotide (competition reaction) can also be added to assess the binding specificity between the protein and the oligonucleotide. The unlabelled oligonucleotide will be used in a much higher concentration than the labelled oligonucleotide, and therefore if the protein is specific to that sequence it will bind more to the unlabelled oligonucleotide in excess, and the result will be a weaker band or no band. In a final test, to confirm which protein is binding to the oligonucleotide, an antibody that recognizes a

specific target protein, will create a heavier complex with less mobility, forming a band with higher molecular weight in the gel (supershift) (**Figure 4.8.1**) (Hellman & Fried 2007; Chorley et al. 2008).

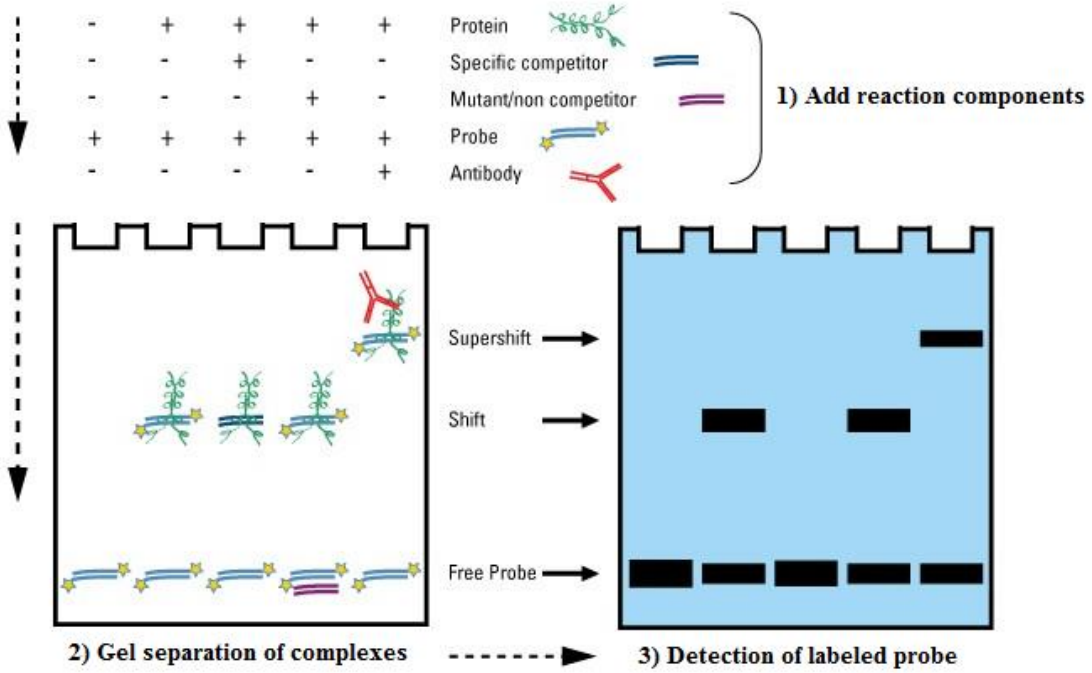


Figure 4.8.1 Illustration of EMSA technique. The gel shift assay consists in three major steps: 1) binding reactions; 2) electrophoresis; 3) probe detection (Thermo Scientific, URL: http://www.piercenet.com/method/gel-shift-assays-emsa).

4.8.1 Oligonucleotide Labelling and Detection

The oligonucleotides were designed for both alleles of the candidate rSNP rs12302714 (C>T, TATGACTAACCTTTTGTAAACGGGTTGTGAGAGGCTGGGAG and TATGACTAACCTTTTGTAAATGGGTTGTGAGAGGCTGGGAG). An oligonucleotide that covers a region in the *FGFR2* gene where Oct-1 and RUNX2 proteins bind, was used as positive control (Meyer et al. 2008). The first step was to label each complementary oligonucleotide separately, and then proceed to annealing (10 minutes at 80ºC, then overnight at room temperature). The labelling was made with a non-radioactive marker, biotin, a molecule with the ability to intercalate specifically in the 3' End DNA strands with the help of the enzyme terminal deoxynucleotidyl transferase (TdT). The labelling procedure

was executed following the manufacturer's instructions of Biotin 3' End DNA Labelling kit (from Thermo Scientific). To test the labelling efficiency, dot blot by hand spotting was performed. Before the detection process, the reactions were transferred from the gel in to a nylon membrane with the help of SD Semi-Dry Transfer Cell (from Bio-Rad), and then cross-linked using UV light. Further, the detection was made following the manufacturer's instructions of Chemiluminescent Nucleic Acid Detection Module (from Thermo Scientific), that consists in blocking the membrane with Blocking Buffer (to block unoccupied binding surfaces), wash it with 1X Wash Buffer (to remove any impurities, reducing background signal during visualization) and adding a chemiluminescent substrate – luminol – for horseradish peroxidase (HRP), that permits the visualization of the labelled oligonucleotides when exposed to UV light.

4.8.2 Protein-Nuclei Acid Binding and Competition Assay

All EMSAs were performed following the LightShift Chemiluminescent EMSA kit (from Thermo Scientific). The reaction for each allele of the SNP and for the positive control (all at 30nM concentration) contained 1X binding buffer to produce ionic conditions that allow binding DNA-protein (preventing the pH from changing), 10ng/µL poly(dI.dC) which is a sequence composed only by I and C nucleotides where unspecific proteins bind, 1X protease inhibitor to protect the proteins from the digestive function of proteases, 1mM DTT to stabilize enzymes and other proteins, 10µg of nuclear protein extract and buffer C, all in a final volume of 20µL. The binding buffer and buffer C were prepared according to the **Table 4.8.2.1**.

Table 4.8.2.1 Preparation of buffer C and 5X binding buffer, with a final volume of 1mL. BB, binding buffer.

Buffer C	
Component	**Final concentration (in 1 mL)**
Hepes, pH 7.9	20 mM
NaCl	400 mM

EDTA	1 mM
Glycerol	20%
H20	

5 X BB Buffer	
Component	**Final concentration (in 1 mL)**
Hepes, pH 7.4	100 mM
$ZnCl_2$	0.5 mM
Glycerol	50%
H20	

A 4 or 6% polyacrylamide gel was used to run the samples at 80V until the samples entered the lanes and then 120V for approximately 1h. The samples were then transferred to a nylon membrane, at 0.28A for 10min, cross-linked (exposure to UV-light two times, at 120mJ/cm2) and detected, as described previously.

The results that showed clear evidence of differences in protein binding affinity between the two alleles of a SNP were further analysed with an EMSA competition assay. The competitions were performed by adding unlabelled oligonucleotide in different concentrations (1X, 33X and 100X, relatively to the 30nM of the labelled oligonucleotide (considered 1X)) of the alleles of interest. For the alleles that continued to show protein binding with relatively high specificity, a supershift assay was performed by adding 2µL of specific antibodies (POL II and HMGA) at 200µg/0.1mL of concentration.

5 Results

5.1 Genomic view of the putative 12q24 risk locus for breast cancer

Our DAE map was combined with the GWAS published and unpublished breast cancer data, in order to identify the loci that have at least a GWAS SNP and a DAE SNP within 250kb away from each other and with a minimum LD of $r^2 = 0.4$. A list with 111 clusters with strong cis-regulatory potential in breast tissue was

generated, where in 32 clusters the GWAS SNP and the DAE SNP are in high LD.

For this thesis we chose the locus 12q24, which in the GWAS study from Easton and colleagues (Easton et al. 2007) reached a *p*-value = 0.002 with an OR = 1.04, 95% CI = 0.98–1.09). This level of significance was not sufficient for a genome-wide study (threshold p-value ≥ 10^{-7}), however it is high enough to suggest that if the region was analysed in more detail, and the true risk variants (if they exist) were identified, that the power to detect association with risk would be greatly improved. Thus, we aimed to test if the DAE method for identifying cis-regulatory variants was an efficient approach to prioritize the most promising candidates from unpublished GWAS lists.

We began by analysing the region where the GWAS SNP rs7307700 is located (the 12q24 locus, more specifically, intron seven of the *AACS* gene, **Figure 5.1.1**). According to our DAE map, there were 15 DAE SNPs that showed DAE in this locus, positioned in different genes, eight of which lie within the *AACS* gene and one in *UBC* gene (**Figure 5.1.2**). In this analysis, we focused on the region 250kb upstream and 250kb downstream starting from the GWAS SNP rs7307700 position.

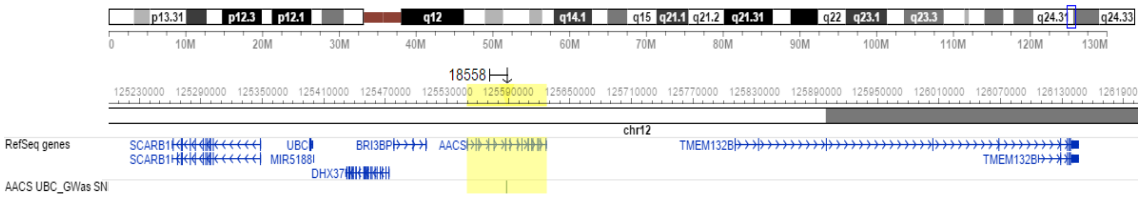


Figure 5.1.1 Genomic view of the GWAS SNP at the 12q24 locus. In the top panel is represented an ideogram of chromosome 12, with the locus containing the gene *AACS* represented by a blue box. In the middle panel, genes are represented in blue and the *AACS* gene area is shaded in yellow. In the lower panel is represented the GWAS SNP rs7307700. Image obtained from Roadmap Epigenomics.

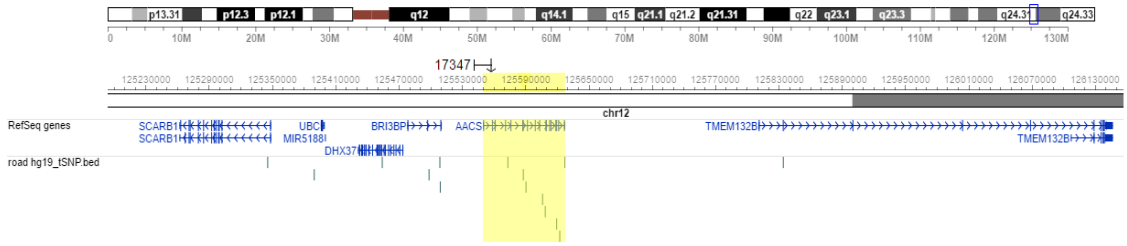


Figure 5.1.2 Genomic view of the tSNPs at the 12q24 locus. In the top panel is represented an ideogram of chromosome 12, with the locus containing the gene *AACS* represented by a blue box. In the middle panel, genes are represented in blue and the *AACS* gene area is shaded in yellow. In the lower panel are represented the DAE tSNPs. Image obtained from Roadmap Epigenomics.

5.2 Identification and analysis of candidate rSNPs in the 12q24 locus

To find candidate cis-acting regulatory SNPs in the 12q24 locus, which could be hypothetically associated with risk, we searched for SNPs in moderate LD ($r^2 \geq 0.4$ and D'≈1) with the GWAS marker SNP rs7307700. We chose this level of LD $r^2 \geq 0.4$ because we hypothesise that the GWAS SNP rs7307700 might not have passed the phase III of GWAS because it is not in high LD with the true causal variant. Seventy-two SNPs were met these criteria (**Annex 1.2**).

We next prioritized the candidates that were located in regulatory elements and DHS sites. Of the 72 SNPs, 36 SNPs were overlapping regulatory elements with evidence for being active in breast tissue and, of these, 12 SNPs were also located in DHS sites (**Table 5.2.1**, **5.2.2** and **Figure 5.2.1**).

Table 5.2.1 List of candidates rSNPs. Putative 12 rSNPs in moderate LD with the GWAS SNP, with evidence of being located in regulatory sites and DHS. Ref, reference allele; alt, alternative allele. MAF represents the frequency of the least common allele. These results were obtained by using SNAP tool, Haploreg v4 and RegulomeDB.

SNP	Alleles ref/alt	MAF
rs10846828	C/T	0.46
rs12302714	C/T	0.41
rs10846834	A/G	0.42
rs10773145	T/C	0.42
rs7133614	C/T	0.34
rs10773146	G/A	0.42
rs12578446	A/G	0.31
rs34151902	G/T	0.27
rs4765021	G/A	0.51
rs4622332	C/T	0.41
rs4559740	G/A	0.3

rs7294703	A/G	0.5

To assess the chromatin context of the 12 candidate rSNPs, we also looked for histone modifications. These candidate SNPs were located in regions harbouring histone marks H3K4me1, H3K4me3, which are indicate enhancers and promoters, respectively, and H3K27ac and H3K9ac, which are markers for active regulatory elements. **Figure 5.2.1** and **Table 5.2.2** show these results.

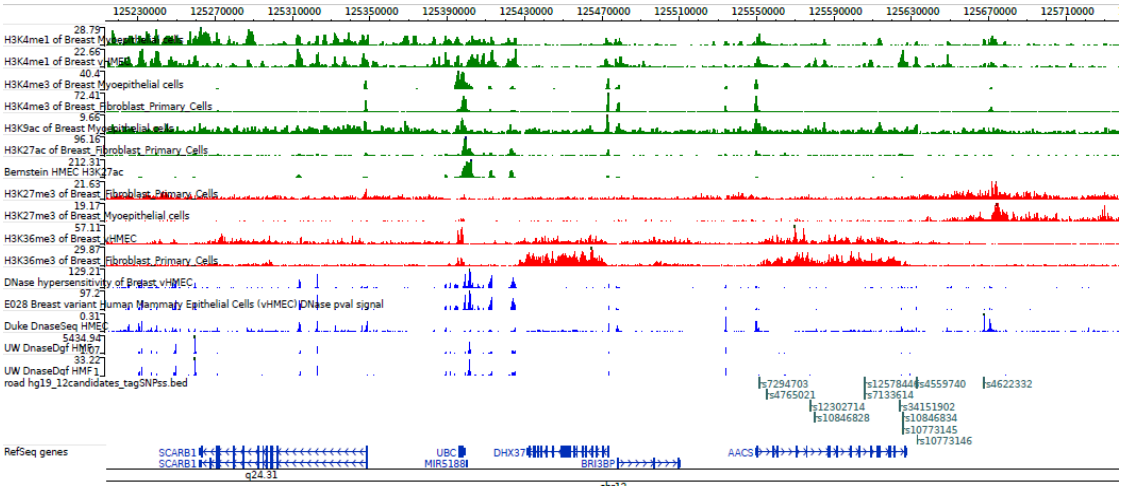


Figure 5.2.1 Genomic view of the 12 candidate rSNPs. In the top panel are represented the ChIP-seq data for a series of histone modifications (Green represents active marks, red is repressive marks) and DNase-seq experiments (in blue). In the middle panel are the 12 candidate rSNPs and in the lower panel are represented the genes in this region (blue). Image obtained from Roadmap Epigenomics.

Table 5.2.2 List of candidates SNPs in *AACS* gene. In this table is represented the 12 candidates rSNPs that were located in at least one promotor or one enhancer, with active histone marks in breast cell lines. T-47D (human mammary ductal carcinoma, oestrogen receptor positive (ER+)) and MCF-7 (human mammary adenocarcinoma, ER+) are breast cancer cell lines; HMEC (human mammary epithelial cells) and HMF (human mammary fibroblasts) are normal breast cell

lines. This table was made with results from Roadmap Epigenomics, Genome Browser, Haploreg v4.1 and RegulomeDB.

SNP	Promotor	Enhancer	H3K4me1	H3K27ac	H3K4me3	H3K9ac	DHS
rs10846828		Yes	Yes				T-47D
rs12302714		Yes	Yes				MCF-7, T-47D
rs10846834		Yes	Yes	Yes			HMEC, HMF, T-47D, MCF-7
rs10773145		Yes	Yes	Yes			HMEC, HMF, T-47D, MCF-7
rs7133614		Yes	Yes				HMEC, T-47D
rs10773146		Yes	Yes				HMEC
rs12578446		Yes	Yes				HMEC, T-47D
rs34151902		Yes	Yes				HMEC, T-47D
rs4765021	Yes					Yes	T-47D
rs4622332			Yes				T-47D
rs4559740		Yes	Yes	Yes		Yes	HMEC, T-47D
rs7294703		Yes	Yes		Yes	Yes	HMEC, HMF, T-47D

5.3 Mapping analysis

We next analysed the DAE distribution pattern of the 15 DAE SNPs located in this locus, in order to know their association with the candidate rSNPs genotypes (**Annex 1.1**). For example, the DAE SNP rs12581512 presents a DAE scenario 1 according to (Xiao & Scott 2011), with all except one heterozygote expressing more the A allele when compared to the G allele (**Figure 5.3.1 A**). This suggests that this SNP is most probably in high LD with the causal variant. rs7301263 shows a scenario 2 pattern of DAE, where some (but not all) heterozygotes preferentially express the A allele when compared to the C allele (**Figure 5.3.1 B**).

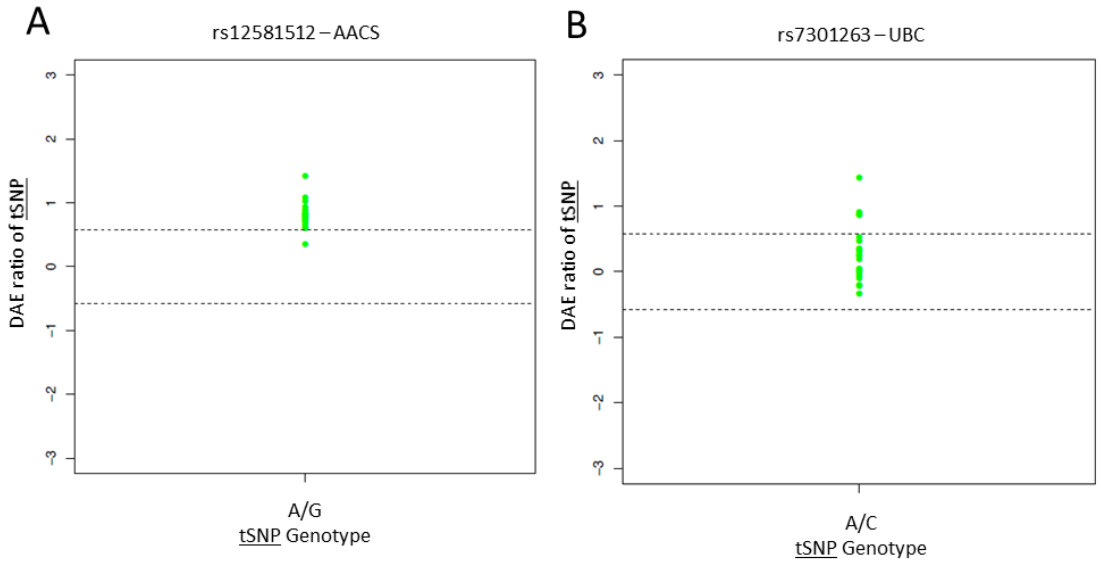


Figure 5.3.1 DAE distribution pattern of two DAE SNPs. In the top of each graphic is represented the DAE SNP and gene where is located. The x-axis represents the heterozygous genotypes of the two DAE SNPs meanwhile the y-axis represents the DAE ratio seen at these two DAE SNPs. (A) DAE ratio of the DAE SNP rs12581512 and (B) the DAE SNP rs7301263. Each green dot represents an individual heterozygous for the DAE SNP. The dotted lines are a threshold (0.58 to -0.58), defined by our group, for what we consider significant values (DAE ratio, log 2 (1.5) = 0.584). The samples observed in between the threshold have equal expression of both alleles, hence not causing a differential effect in expression. Above the dotted lines are the samples in which one allele is being differentially expressed and below are the samples in which the other allele is differentially expressed.

To test which candidates rSNPs could explain the DAE observed at the 12q24 locus, we analysed the association between genotypes at the 12 candidate rSNPs and the DAE measured at the 15 DAE SNPs, by plotting the distribution of DAE of each DAE SNP according to the candidate rSNP genotype. Genotyping data used in this analysis came from our genotyping experiments, the DAE experiments and imputation exercises. The candidate rSNPs are only considered potentially causal and responsible for the DAE if: (1) the homozygous samples for the rSNP do not show DAE; (2) the heterozygous samples for the rSNPs present DAE.

From this analysis, two candidate rSNPs showed greater potential to be the causal variant of DAE in *AACS* (DAE SNP rs12581512) – rs10773145 and rs10846834 – and one other candidate rSNP showed greater potential to be the causal variant of DAE in the *UBC* (DAE SNP rs7301263) - rs12302714 (**Figure 5.3.2 A and B**). No variant explained the DAE observed in any of the remaining

genes from the 12q24 locus. Only some heterozygous for the candidate rSNP rs12302714 display DAE. But more importantly, we observed that all homozygous individuals for rs12302714, show equal levels of transcription of the two alleles of the DAE tSNP rs7301263. For SNPs rs10773145 and rs10846834, the majority of heterozygous individuals show one allele more expressed compared to the other. The Welch's-test although not significant, was underpowered by the fact that there was only one homozygous sample for these SNPs, although, the DAE levels of the one homozygous sample and one heterozygous sample are near the cut-off.

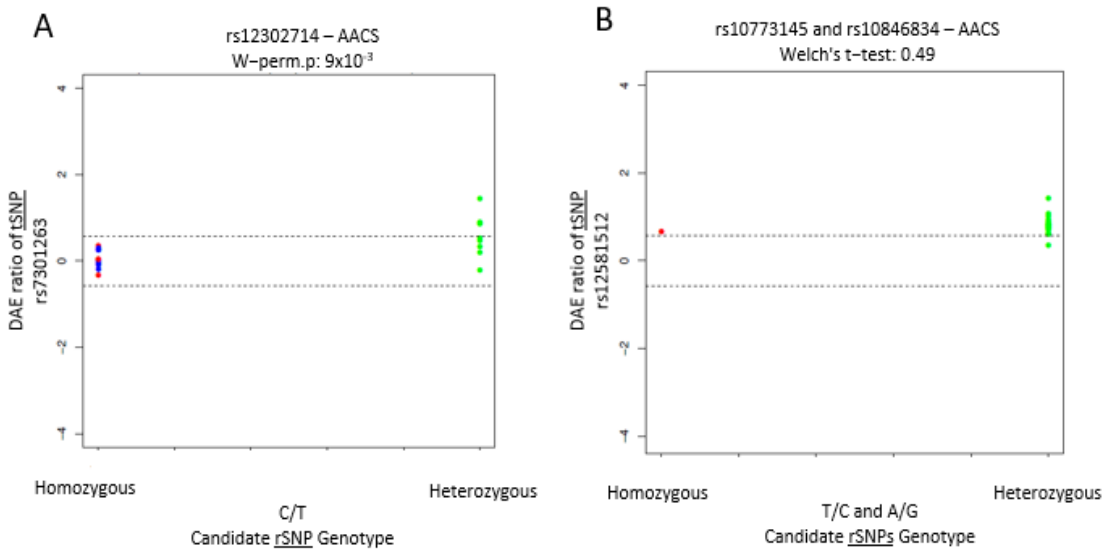


Figure 5.3.2 DAE mapping analysis for the candidate rSNPs rs12302714, rs10773145 and 10846834. In the top of each graphic is represented the candidate rSNP, gene where it is located and *p*-value for this test. The x-axis represents the genotypes of the candidates rSNPs and the y-axis represents the DAE ratio seen at the DAE SNPs rs7301263 and rs12581512. (A) rSNP rs12302714 with DAE SNP rs7301263, (B) rSNPs rs10773145 and rs10846834 with DAE SNP rs12581512. The dotted lines are a threshold (0.58 to -0.58), defined by our group, for what we consider significant values (DAE ratio, log 2 (1.5) = 0.584). The samples observed in between the threshold have equal expression of both alleles, hence not causing a differential effect in expression. Above the dotted lines are the samples in which one allele is being differentially expressed and below are the samples in which the other allele is differentially expressed.

5.4 *In silico* analysis of candidate rSNPs rs12302714, rs10773145 and 10846834

The rs12302714, rs10773145 and rs10846834 were located in the *AACS* gene (**Figure 5.2.1**). Since the three candidates rSNPs overlapped regulatory element regions, we searched for evidence of TF binding at their locations. No significant prediction of differences in allelic binding affinity was found for candidate rSNPs rs10773145 and rs12302714. For candidate rs10846834, PWM data analysis showed that several predicted TFs could have different allelic binding affinity (**Tables 5.4.1**, **5.4.2** and **5.4.3**).

Table 5.4.1 Predicted transcription factor binding for candidate rSNP rs10773145. TF, transcription factor. Information from Haploreg v4.1 database. The values represent PWM scores.

Predicted TF binding	Reference allele T	Alternative allele C
Sin3Ak-20	7.7	2.2

Table 5.4.2 Predicted transcription factor binding for candidate rSNP rs10846834. TF, transcription factor. Information from Haploreg v4.1 database. The values represent PWM scores.

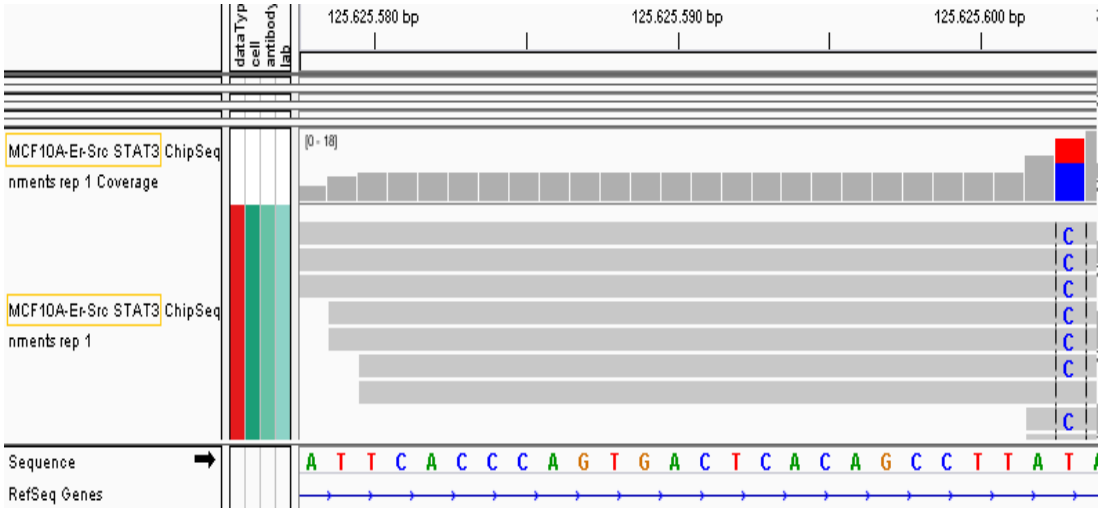
Predicted TF binding	Reference allele A	Alternative allele G
Gfi1	12.5	1.9
Maf	-1.1	10.9
NF-E2	6.3	12.9
Nrf-2	0.5	12.1

Table 5.4.3 Predicted transcription factor binding for candidate rSNP rs12302714. TF, transcription factor. Information from Haploreg v4.1 database. The values represent PWM scores.

Predicted TF binding	Reference allele C	Alternative allele T
Sox	11.8	11.1

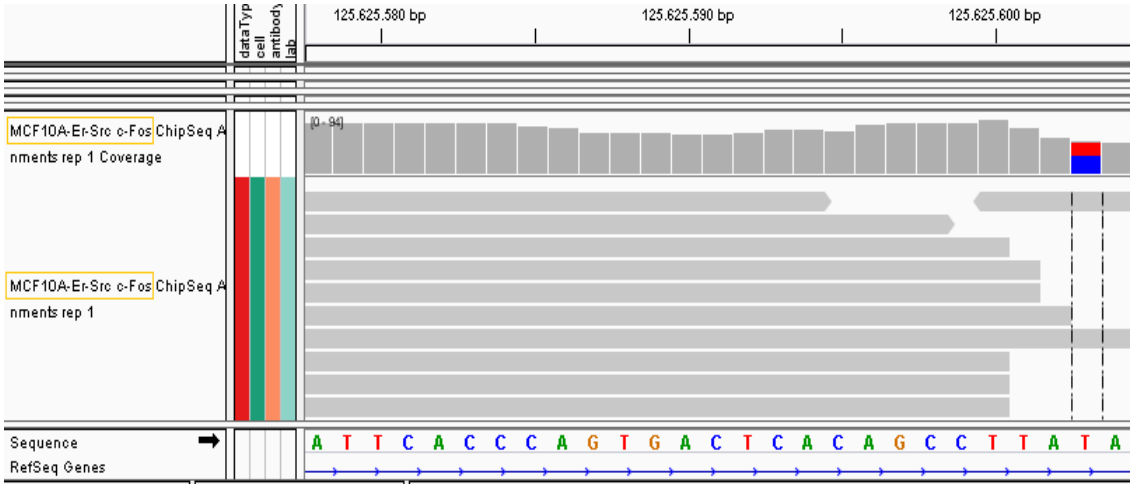
We next analysed ChIP-seq information for these three candidates. We found that rs10773145 and rs10846834, both located in an active enhancer at HMEC and BR.MYO cell lines (normal mammary cell lines), overlap a STAT3 and c-FOS proteins binding sites (**Figure 5.4.1** and **Figure 5.4.2**). However, for rs12302714 we were not able to find any evidence of TF binding in the ChIP-seq data available form ENCODE and Roadmap Epigenomics.

A



Total Reads Count: 15
C (Blue): 9 (60 %)
T (Red): 6 (40 %)

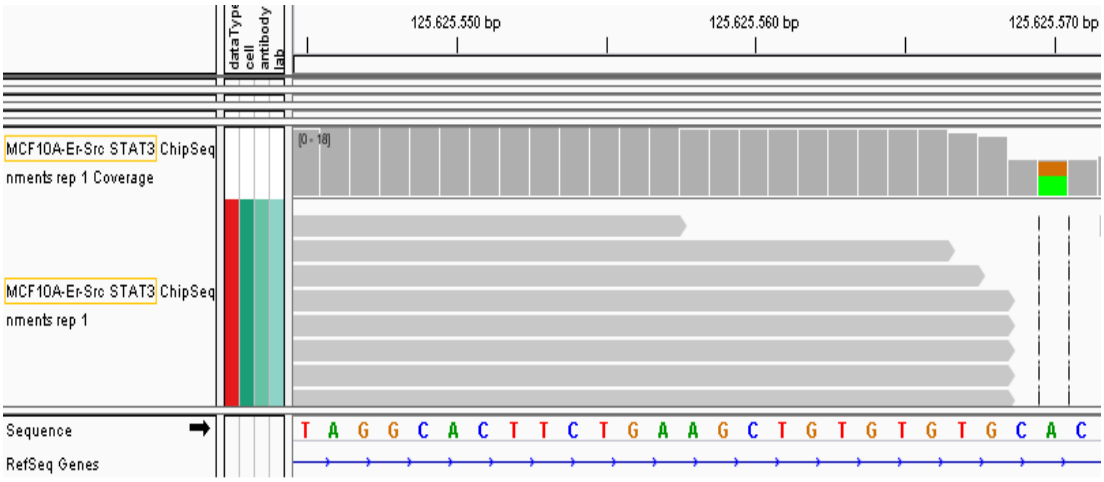
B



Total Reads Count: 48
C (Blue): 27 (56 %)
T (Red): 21 (44 %)

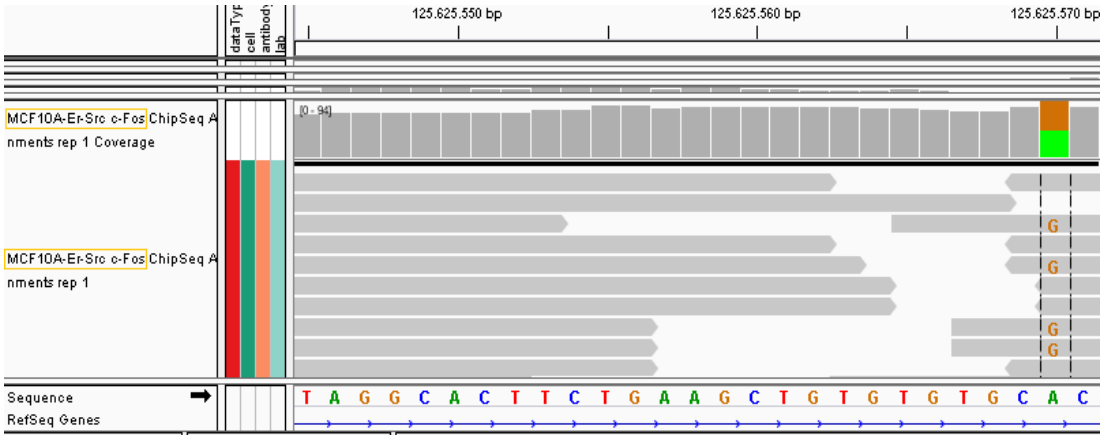
Figure 5.4.1 ChIP-seq results for (A) STAT3 and (B) c-FOS proteins in MCF10A cell line, at the candidate rSNP rs10773145. In the bottom of each image is represented the nucleotides, with adenine in green, thymine in red, cytosine in blue and guanine in orange. The horizontal bars correspond to the sequence reads of the experiment and the vertical bars to the frequency that the protein binds to each allele. The information was first obtained from Haploreg v4.1 and RegulomeDB and then visualised with IGV.

A



Total Reads Count: 9
A (Green): 5 (56 %)
G (Brown): 4 (44 %)

B



Total Reads Count: 94
A (Green): 44 (47 %)
G (Brown): 50 (53 %)

Figure 5.4.2 ChIP-seq results for (A) STAT3 and (B) c-FOS proteins in MCF10A cell line, at the candidate rSNP rs10846834. In the bottom of each image is represented the nucleotides, with adenine in green, thymine in red, cytosine in blue and guanine in orange. The horizontal bars correspond to the sequence reads of the experiment and the vertical bars to the frequency that the protein binds to each allele. The information was first obtained from Haploreg v4.1 and RegulomeDB and then visualised with IGV.

For rs10773145 and rs10846834 candidate rSNPs the ChIP-seq results show strong binding of c-FOS in MCF10A cell line, but with small differences of affinities

between the two alleles (56% (C) and 44% (T) for SNP rs10773145; 47% (A) and 53% (G) for SNP rs10846834). Regarding STAT3, the total number of reads (n ≤ 20) of each ChIP-seq experiment was not sufficient to draw any conclusions. Therefore, it is likely that these SNPs do not alter the binding affinity of the transcription factors and might be causing the DAE observed in the DAE SNP rs12581512 by other mechanisms.

Since there was no evidence of TF binding in the ChIP-seq data available, we decided to carry out an EMSA for the candidate rSNP rs12302714, to test if it could bind any other TF for which there was no ChIP-seq data available.

5.5 Analysis of the protein binding preferences in the candidate rSNP rs12302714

At the beginning of the EMSA experiments the oligonucleotide probes were labelled. This initial step included a labelling efficiency test, which was only done for the positive control (FGFR-13*). In **Figure 5.5.1**, it can be seen that this probe was labelled with 100% efficiency.

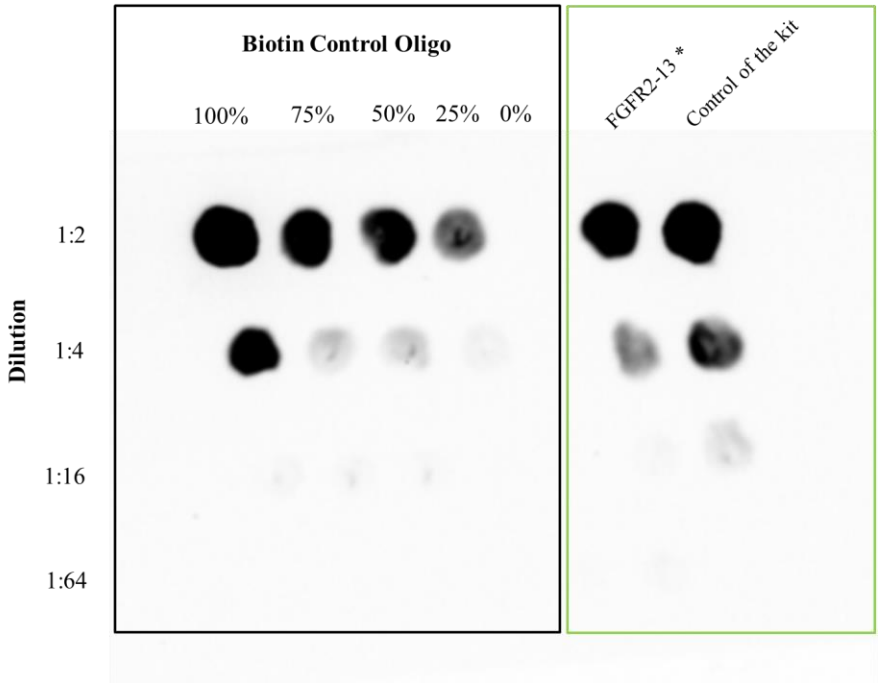


Figure 5.5.1 Determination of labelling efficiency for Biotin Control DNA and for positive control annealed FGFR2. The blot includes the dilution for each of the standards of the Procedures for Estimating Labelling Efficiency (option 2: Dot Blot by Hand Spotting) (inside black

box); as well as the labelling control of the kit and the positive control of nuclear extract FGFR2 (inside green box).

Then, three different EMSAs were performed using three nuclear extracts from the breast cancer cell lines MCF-7 (**Figure 5.5.2)**, T-47D (**Figure 5.5.3**) (both ER+) and HCC1954 (**Annex 2.1**) (ER-). As observed in Lane 1 of Figure 5.5.2, there is binding of Oct-1 and Runx2 to the control probe FGFR2-13*, which corresponds to the positive control of an rSNP regulating *FGFR2* expression. The binding seen in Lanes 2 and 3 is not allele specific (contains oligonucleotides corresponding to the C and T alleles for this candidate rSNP, respectively). This equal binding was confirmed by the competition assays with the unlabelled oligonucleotides (Lanes 4-7), in which all bands disappeared, meaning that both alleles can compete with each other. Although, if there were differential protein binding preference between the C and the T allele of rs12302714, we would see a weaker band for one allele and a stronger one for the other allele.

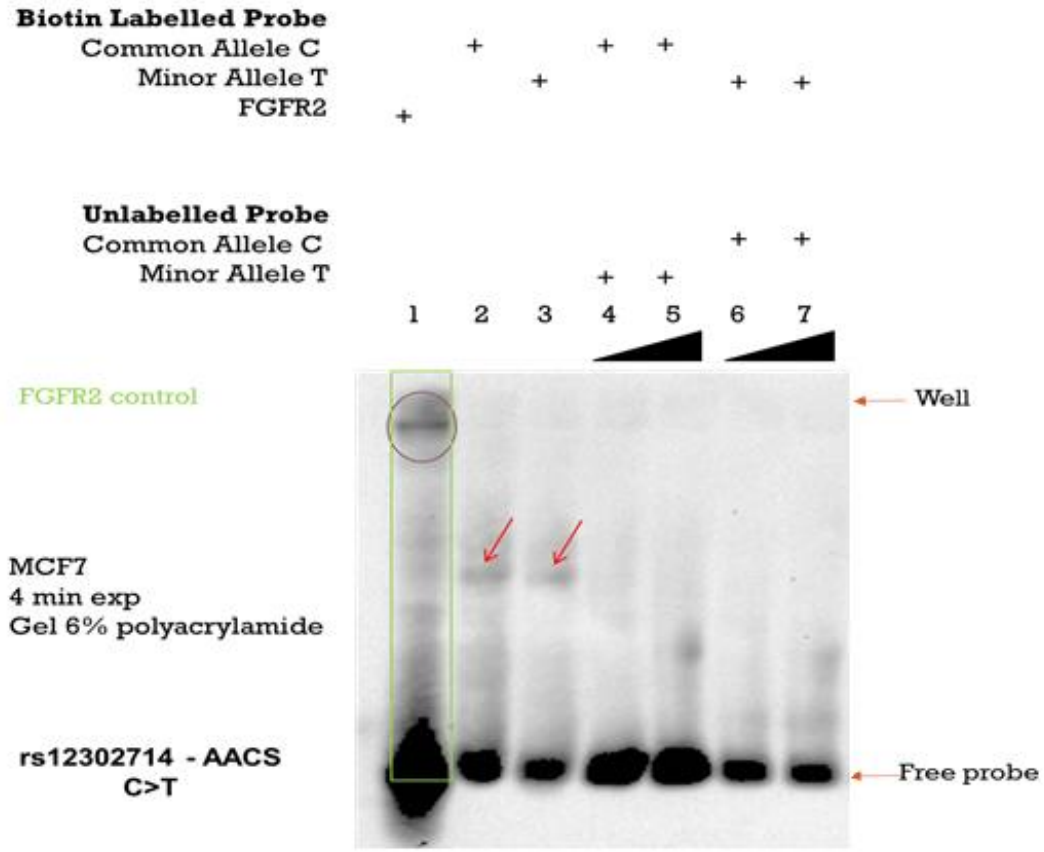


Figure 5.5.2 EMSA in vitro assay showing protein-nucleic acid interaction and competition binding studies. Nuclear extract from MCF-7 cell line were used. Lane 1 corresponds to the positive control, FGFR2 oligonucleotides, the band is identified by a circle. Lanes 2 and 3 (red

arrows) corresponds to labeled oligonucleotides containing the C and the T allele, respectively, of the SNP rs12302714. Lanes 4 and 5 contains labeled oligonucleotide with the C allele, while lanes 6 and 7 contais the labeled oligonucleotide containing the T allele. For competition, 30X higher concentrated unlabelled oligonucleotide was added to lanes 4 and 6, meanwhile to lanes 5 and 7, the unlabeled oligonucleatide was set 100X higher.

◢ represents the increase of oligonucleotide concentration.

The EMSA was repeated using a different cell line T-47D (**Figure 5.5.3**) and in the same conditions, showing similar results. Therefore, it is possible that this candidate rSNP is causing the DAE observed in the DAE SNP rs7301263 by other mechanisms rather than alterations of the binding of transcription factors.

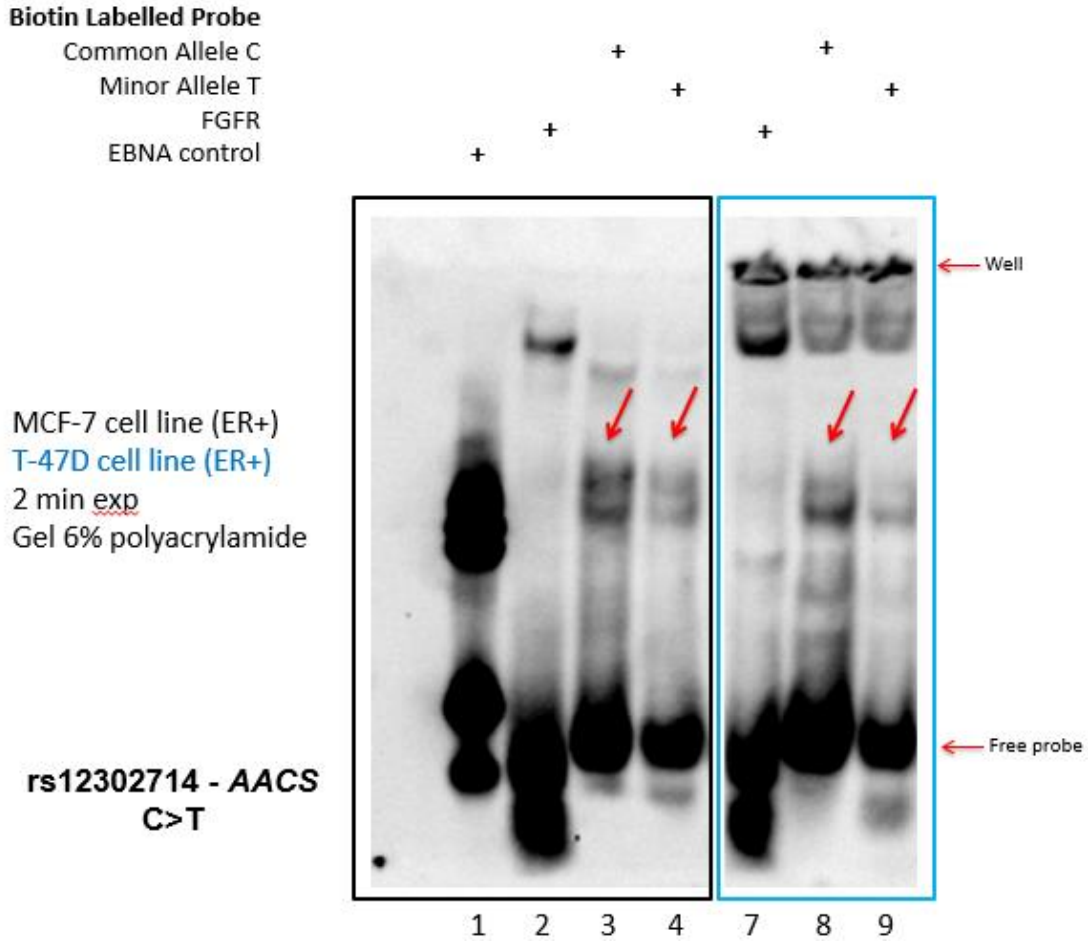


Figure 5.5.3 EMSA in vitro assay showing protein-nucleic acid interaction of candidate rSNP rs12302714 with two different nuclear extracts. Nuclear extract from MCF-7 (black box) and T-47D (blue box) cell lines were used. Lane 1 corresponds to EBNA control; lanes 2 and 7 contains the positive control FGFR2 oligonucleotides. Lanes 3, 4, 8 and 9 (red arrows) corresponds to labeled oligonucleotides containing the C and the T allele of the SNP rs12302714. It is important to say that, although it seems that there is more protein binding to the C allele,

those lanes (3 and 8) has more oligonucleotide pipetted that lanes 4 and 9 (which corresponds to the T allele).

5.6 *In silico* analysis of microRNAs binding

Since the results of ChIP-seq for the SNPs rs10846834 and rs10773145 and our EMSA results for SNP rs12302714 suggested that these candidate rSNPs might not be causing DAE by altering the binding affinity of transcription factors, we proceed to analyse if any of the initial 72 SNPs in LD with the GWAS SNP were causing DAE by altering the mRNA levels through differential allelic microRNAs regulation.

The results suggested that 17 out of the 72 SNPs (including rs12302714) could modify the binding affinity of microRNAs to their targeting sites on 3'UTR of mRNAs, and could therefore, be affecting the mRNA levels in an allele-specific manner (**Table 5.6.1**).

Table 5.6.1 List of SNPs predicted to be altering miRNA binding affinity. In this table is shown in the first column the 17 SNPs out of the 72 initial SNPs that alter miRNA binding affinity; in the second column is the alleles of each SNP; in the third column is the miRNA that is predicted to be binding differently to one allele compared to the other; and in the fourth column is the total number of miRNAs that binds to each SNP. This table was accomplished by consulting NCBI SNP, Ensembl and miRBase databases.

SNP	Alleles	miRNA binding differently	Total number of miRNA binding
rs12302714	C	-	11
	T	hsa-miR-4506	
rs7133614	C	hsa-miR-5708	8
	T	-	
rs12578446	A	hsa-miR-6747-5p	8
	G	-	
rs4765021	A	hsa-miR-6851-5p	4
	G	-	
rs4765217	G	hsa-miR-6866-5p	11
	T	-	
rs56394386	A	-	6
	G	hsa-miR-5095	
rs4765218	A	hsa-miR-4756-5p	12
		hsa-miR-4526	
	G	hsa-miR-3909	
rs7135489	A	hsa-miR-5694	3

	G	-	
rs6488989	A	-	16
	C	hsa-miR-3691-3p	
rs2018130	C	-	11
	T	hsa-miR-1972	
		hsa-miR-5708	
rs7955201	C	hsa-miR-346	16
	T	hsa-miR-3691-3p	
rs7138557	C	-	7
	T	hsa-miR-1245b-3p	
rs7137679	C	-	17
	T	hsa-miR-3613-3p	
rs900410	C	hsa-miR-3162-5p	12
	T	hsa-miR-4446-5p	
rs10846824	C	hsa-miR-6810-3p	4
		hsa-miR-6845-3p	
	T	-	
rs55999005	C	-	7
	G	hsa-miR-873-3p	
	T	-	
rs7953077	C	-	10
	T	hsa-miR-19a-3p	
rs34624329	G	-	24
	T	hsa-miR-550a-3p	
		hsa-miR-550b-2-5p	
		hsa-miR-550a-5p	
		hsa-miR-550a-3-5p	
rs12316499	A	-	22
	G	hsa-miR-3714	
		hsa-miR-6885-5p	
rs7398636	C	hsa-miR-1236-5p	20
		hsa-miR-3605-3p	
		hsa-miR-6798-3p	
		hsa-miR-6845-3p	
		hsa-miR-6729-3p	
		hsa-miR-4433a-3p	
		hsa-miR-513c-3p	
		hsa-miR-6891-5p	
		hsa-miR-4433b-5p	
	G	-	
rs41473449	A	-	2
	G	hsa-miR-638	
rs11058031	A	-	21
	T	hsa-miR-1303	

rs12303416	C	hsa-miR-665	22
	T	hsa-miR-5689	
rs58624919	A	hsa-miR-4298	7
	G	-	
rs10846822	C	-	16
	T	hsa-miR-1272	

5.7 LD structure and Haplotype analysis

As there are three candidate rSNPs that can explain the DAE observed in this locus, we analysed the haplotype block and LD structure of the region using the genotype information of our 64 normal samples, in order to identify the haplotype containing the potential cis-regulatory variant or variants responsible for the DAE observed in the *AACS* and *UBC* genes. In other words, we analysed the frequency of the possible recombination between the alleles of the GWAS SNP rs7307700, the candidate rSNPs (rs10773145, rs10846834 and rs12302714) and the DAE SNPs (rs12581512 and rs7301263).

Haplotype analysis showed that the region where the GWAS SNP, the candidate rSNPs and the DAE SNPs are positioned is divided in two haplotype blocks. The DAE SNP rs7301263 (located in the *UBC* gene) was not in high LD with the other SNPs. Meaning that random recombination may occur between the alleles of this SNP and the alleles of rs10773145, rs10846834, rs12302714, rs12581512 and rs7307700. Therefore, rs7301263 is not in a haplotype block, as we can see in **Figure 5.7.1**.

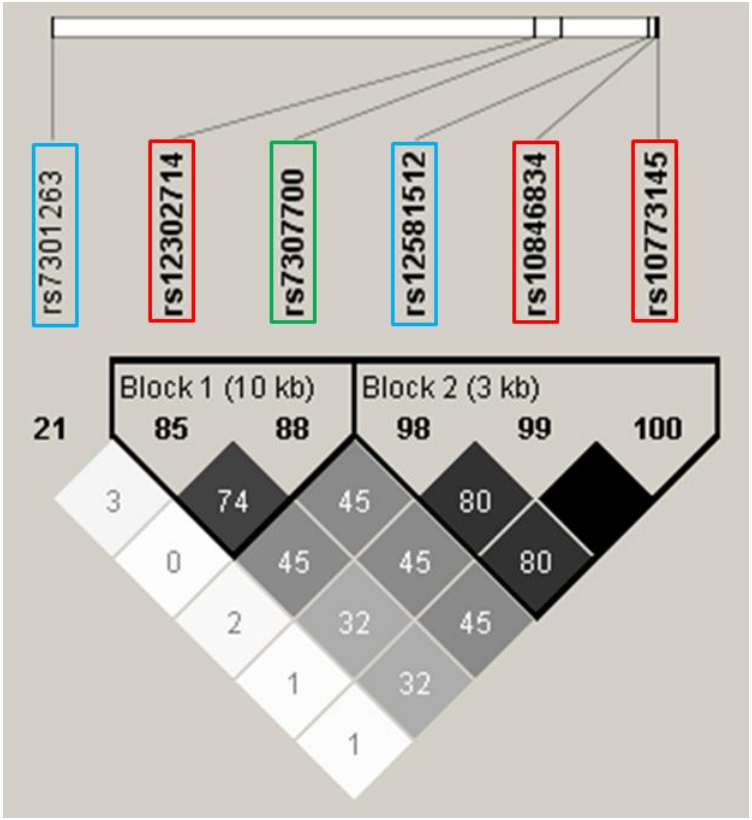


Figure 5.7.1 Linkage disequilibrium structure and haplotype blocks for the GWAS SNP (rs7307700, in green), the candidate rSNPs (rs12302714, rs10773145 and rs10846834, all in red) and the DAE SNPs (rs12581512 and rs7304293, in blue). In this LD plot, the blocks were defined using Confidence Intervals by Gabriel *et al* 2002. The SNPs are identified on top of the diagram. The r^2 squared colour scheme was chosen, where black represent $r^2=1$, the different shades of grey represent $0 < r^2 < 1$ and the r^2 values inside the plots indicate the pairwise LD between the SNPs. Black triangles indicate the two haplotype blocks. Plot obtain from Haploview tool.

Block 1 is about 10kb long and includes the candidate rSNP rs12302714 and the GWAS SNP rs7307700, meanwhile the block 2 (around 3 kb long) includes the DAE SNP rs12581512 and two candidate rSNPs, rs10773145 and rs10846834. Also, we observed that block 1 has 3 major haplotypes that account for 99.2% of the individuals (with some rarer not displayed in the figured), while block 2 has only 3 possible haplotypes, as shown in **Figure 5.7.2**.

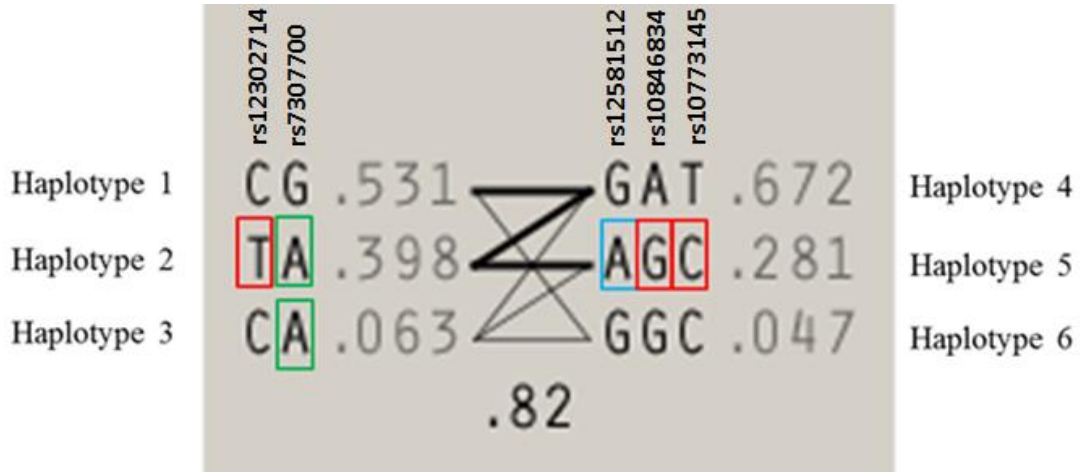


Figure 5.7.2 Blocks 1 and 2 and their respective haplotypes. Above are the rsID. The frequency of each haplotype is shown at the right side and the SNP rsID on top corresponds to those in LD in Figure 5.7.1. In red are the candidate rSNPs, in green is the GWAS SNP and in blue is the DAE SNP. The haplotype 1 is constituted by C and G nucleotides; haplotype 2 by T and A nucleotide; haplotype 3 by C and A nucleotides; haplotype 4 by G, A and T nucleotides; haplotype 5 by A, G; and C nucleotides and haplotype 6 by G, G and C nucleotides of the corresponding SNPs. Image obtained and adapted from Haploview tool.

The minor allele of the GWAS SNP rs7307700 (A allele, in green), which is associated with risk for breast cancer, is present in haplotypes 2 and 3. From our DAE data, the minor A allele of the DAE SNP rs12581512 (in blue), found in haplotype 5, is overexpressed when compared to the common G allele.

The minor alleles of the candidate rSNPs rs10846834 (G) and rs10773145 (C) are found in haplotype 5 in block 2 (frequency approximately 30%) and the minor allele of rs12302714 (T) is found in haplotype 2 in block 1 (frequency approximately 40%). Interestingly, both haplotypes are in high LD and comprise the GWAS SNP risk allele (A) and the preferentially expressed DAE SNP allele (A). Also, the results from IGV suggested that the minor alleles of rs10846834 (G) and rs10773145 (C) were associated with more binding affinity to STAT3 and c-FOS (**Figure 5.4.1** and **Figure 5.4.2**), and results of microRNA analysis suggested that the miRNA hsa-miR-4506 binds preferentially to the T allele of rs12302714. This way, the effect of two or more alleles may be causing the risk for breast cancer and the DAE.

A more detailed analysis of the haplotypes (**Table 5.7.1**), suggested that the haplotype 2, that also have the minor T of the candidate rSNP rs12302714, is more frequently transmitted with haplotype 5 (26.2% of frequency), which have the allele more expressed from the DAE SNP (A) and the G and C minor alleles

of two candidate rSNPs, both associated with more protein binding. However, haplotype 2 is also transmitted with haplotype 4, with 13.6% of frequency, that have the DAE SNP (G) less expressed allele and rSNPs (A and G) alleles associated with less protein binding.

The haplotype 3 is transmitted with haplotype 4, haplotype 6 (1.3% and 3.1% of frequency, respectively; both with the GWAS allele (G) not associated with risk, DAE SNP (G) less expressed allele and rSNPs (A and G) less associated with protein binding) and with haplotype 5 (1.9% of frequency).

Thus, the ratio of signal of the haplotypes more probably to be causing risk (haplotypes 2 / 5 and haplotypes 3 / 5) compared with the haplotypes not causing risk (haplotypes 2 / 4, 3 / 4 and 3 / 6) is 28.1% to 18%. Therefore, this might explain why the signal detected for GWAS was not significantly strong to associate the GWAS SNP with risk for breast cancer, since 18% of the signal detected is from haplotypes recombination not associated with neither DAE and cis-regulation.

Table 5.7.1 Blocks 1 and 2 and their respective haplotypes frequency recombination. The frequency of each haplotype in our 64 samples is shown below each block. In the column "Frequency of recombination (%)" is represented in percentage the frequency of recombination between block 1 (Haplotypes 1, 2 and 3) and block 2 (Haplotypes 4, 5 and 6). This table correspond to Figure 5.6.2 although more detailed. Table obtained and adapted from Haploview tool. Hap, haplotype.

BLOCK 1.	**Frequency of recombination (%)**		
Hap 1 (53.1%)	Hap 4 (51.5%)	Hap 5 (0%)	Hap 6 (01.6%)
Hap 2 (39.8%)	Hap 4 (13.6%)	Hap 5 (26.2%)	Hap 6 (0%)
Hap 3 (6.3%)	Hap 4 (1.3%)	Hap 5 (1.9%)	Hap 6 (3.1%)
BLOCK 2.			
Hap 4 (67.2%)			
Hap 5 (28.1%)			
Hap 6 (4.7%)			

5.8 EMSA for candidate rSNP rs111549985 of the 5q14.2 locus

During this master thesis project, in parallel, we conducted another functional analysis in the 5q14.2 locus, where we demonstrated differences in protein binding affinity between the two alleles of a SNP. It is included here to show an

example of an rSNP displaying preferential allelic binding affinity, which was not detected in the locus 12q24 (**Figure 5.7.1**).

In summary, the SNP associated with breast cancer risk in the GWAS was rs7707921 (OR = 0.94, *p*-value = $5x10^{-11}$). The 5q14.2 locus had three genes showing DAE, namely *ATG10*, *RPS23* and *ATP6AP1L*. After *in silico* analysis, one candidate rSNP located in *ATG10* gene showed a putative effect in transcription factor binding sites, rs111549985 (data not shown since was performed by another colleague). However, there was not enough functional information available to associate this SNP to the DAE observed in this locus So, we performed a PCR to genotype the candidate rSNP rs111549985 in 51 normal breast samples, in order to be possible to perform the DAE mapping analysis, which further indicated that this variant is associated with DAE levels at three SNPs, in the genes *ATG10* and *RPS23*.

Through EMSA *in vitro* assay, using MCF-7, MDA-MB-231 and HCC1954 cell lines, we verified that there is a protein binding to both alleles but preferentially to the C allele of rs111549985 rather than the G allele (Lanes 3 and 4), as shown in **Figure 5.7.1**. Therefore, rs111549985 is a good candidate to be the causal rSNP to the observed DAE in *ATG10* and *RPS23* and to be associated with the risk of breast cancer. This binding was confirmed with competitions assay using unlabeled oligonucleotides (Lanes 5-10). We further performed a supershift assay in order to identify the protein that was binding to rs111549985 using antibodies against POL II, E2F1 and c-MYC (the EMSA for the last two proteins is not shown in **Figure 5.7.1**), based on PWM analysis results. The results (Lanes 11-14) showed that none of these proteins are the one binding to rs111549985.

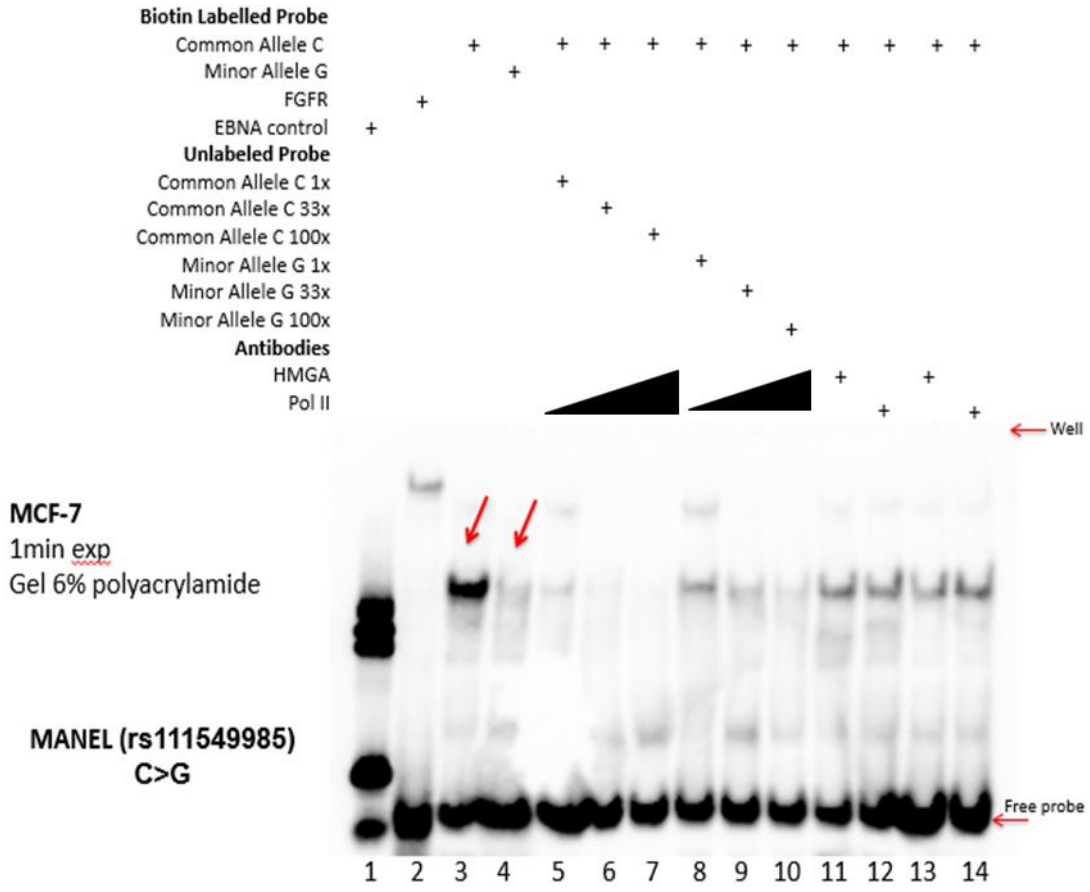


Figure 5.8.1 EMSA in vitro assay showing protein-nucleic acid interaction and competition binding studies for candidate rSNP rs111549985 (5q14.2 locus). Nuclear extract from MCF-7 cell line was used. Lane 1 corresponds to EBNA control; lane 2 contain the positive control FGFR2 oligonucleotides. Lanes 3 and 4 (red arrows) corresponds to labeled oligonucleotides containing the C and the G allele of the SNP rs111549985. For competition, 1X concentrated unlabelled oligonucleotide was added to lanes 5 and 8; 30X higher concentrated unlabelled oligonucleotide was added to lanes 6 and 9; 100X higher concentrated unlabelled oligonucleotide was added to lanes 7 and 10. To lanes 11 and 13, HMGA antibody was added as negative control, meanwhile in lanes 12 and 14 Pol II antibody was added.

6 Discussion and Conclusion

Prior studies have shown that inherited genetic variants contribute to the regulation of gene expression and may increase the risk of developing common diseases, such as breast cancer (Ghoussaini et al. 2013; Michailidou et al. 2015; Ripperger et al. 2009; Maia et al. 2012; Glubb et al. 2015; Cai et al. 2011; Shephard et al. 2009; Wang et al. 2014; French et al. 2013; Meyer et al. 2008; Meyer et al. 2013; Wynendaele et al. 2010). The fact that all 94 common low-penetrance loci identified by GWAS associated with risk for breast cancer were located in non-coding regions, intergenic region or gene deserts (except one loci that was located in a coding region), suggested that they are located in regulatory elements. Furthermore, since all loci studied at a functional level suggested that these genetic variants are conferring risk through cis-regulation, we hypothesize that cis-regulation is an important mechanism for breast cancer risk and that the genetic variants located in the remaining loci to analyse functionally are likely cis-regulatory. Since GWASes have a long list of unpublished SNPs to validate, we combined our DAE data, with unpublished and published GWAS data for breast cancer in order to prioritize the loci with genes being regulated by cis-regulatory variants, and therefore, more likely to be associated with breast cancer risk, for further validation studies. This way, the main objective of this study was to validate an unpublished GWAS locus to confirm it as a new risk locus for breast cancer, by first identifying the causal genetic variant(s) cis-regulating this locus and further performing a new association study to improve the GWAS SNP significance to breast cancer risk.

In previous work performed in our group the whole-genome DAE map was accomplished by using microarrays and comparing the expression of both alleles in 500K SNPs along the genome. Further, merging the DAE map with published and unpublished GWAS data, was generated a list of 111 clusters with strong cis-regulatory potential in breast tissue. Each cluster defined as having at least one GWAS SNP and one DAE SNP within 250kb. In 32 of the 111 clusters, the GWAS SNP and the DAE SNP were in strong LD. One cluster was located in the 12q24 locus, which was selected to be analysed under the course of this thesis, in order to test if we could validate the GWAS unpublished SNP (rs7307700) that did not

reach the phase III GWAS statistical significance threshold (p-value ≤ $1x10^{-5}$) with a p-value = 0.002, and further, associate this locus to breast cancer risk.

Besides the GWAS SNP rs7307700, the 12q24 locus contained also 15 SNPs displaying DAE. rs7307700 is located in the *AACS* gene, which encodes for an enzyme called acetoacetyl-CoA synthetase. Although, the physiological role of AACS is yet unclear in humans, some studies suggested its involvement in the regulation of lipid metabolism and the metabolism of ketone bodies (Schug et al. 2015). The ketone bodies are often used as an alternative resource of energy by tumour cells when undergoing starvation (Ohgami et al. 2003; Schug et al. 2015).

One possibility for this locus not passing the phase III significance threshold in the GWAS could be because it is not in high LD with the causal variant. As stated before, GWAS uses marker SNPs that report the association with other SNPs when they are in strong LD with each other. In this way, if the GWAS SNP rs7307700 is not in strong LD with the true causal variant, this may diminish the signal detected for rs7307700 during the association analysis, and as a result, the locus was not be associated with risk. We believe that by first looking for the possible causal regulatory variant, which evidence from other studies suggests might be a cis-regulatory SNP(s), and further analyse the correlation between the GWAS SNP and the causal rSNP genotypes, would be a possible way to improve the statistical power to detect association with risk at this locus.

Taking this into consideration, we chose to investigate SNPs that are in moderate to high LD with the GWAS SNP. Based on our DAE data, and on data available from regulatory genomic projects, we selected 12 candidate rSNPs that showed potential to be cis-acting regulatory variants. Since we had the genotypes of all 12 SNPs in 64 normal breast samples from the DAE study, we tested if any of these candidates could explain the DAE observed by performing a mapping analysis. In other words, we tested the genotypes of the 12 candidate rSNPs against the allelic expression of the 15 DAE SNPs, in order to see if when the candidates rSNPs were heterozygous, the samples presented DAE at the DAE SNPs, and when the rSNPs were homozygous the samples showed no DAE. The DAE SNP rs12581512 showed a DAE distribution pattern that suggest a complete LD between rs12581512 and the rSNP (or rSNPs), and the rs7301263

DAE distribution pattern suggested a strong, although incomplete LD between rs7301263 and the rSNP (or rSNPs). From this analysis, three of the initial 12 candidates SNPs, rs10846834, rs10773145 and rs12302714, were selected as candidates to be putatively regulating the *AACS* gene.

6.1 Analysis of candidates rSNPs rs10846834 and rs10773145

The candidate rSNPs rs10846834 and rs10773145 were strongly associated with the DAE SNP rs12581512, with the majority of individuals heterozygous for these two rSNPs showing DAE. Although the *p*-value was not significant (*p*-value = 0.485), we should note the lack of homozygotic samples for these SNPs, which weaken the statistical analysis. Therefore, we included these two candidate rSNPs in our analysis. However, for the heterozygous individuals for DAE SNP rs12581512 there was only one sample homozygous for the rSNPs rs10846834 and rs10773145. Therefore, to confirm if these two SNPs are truly associated with DAE levels, we need to increase the number of samples in the analysis to improve the statistical power.

The next step was to look for evidence of protein binding on these two candidates. For the rSNP rs10846834 there were predictions of three proteins possibly binding to its sequence (Nrf-2, Maf and Gfi1) with significant differences in allelic binding affinity. The protein Maf - encoded by the oncogene *Maf* - is part of the basic leucine zipper (bZIP) family of transcription factors, which have a basic domain capable of binding to the DNA and a bZIP domain to form heterodimers with specific transcription factors, such as NF-E2 and Nrf-2 (Kannan et al. 2012), which might explain the predicted binding of all of those proteins. PWM analysis suggested that Maf binds to the alternative G allele (PWM = 10.9), but not to the reference A allele (PWM = -1.1). We further looked on IGV for ChIP-seq experiments that could confirm the PWM results, but for these proteins there were no experiments on breast cell lines. Since there was no ChIP-seq experiments for these four proteins on breast cell lines, in the future we propose EMSA analysis to validate PWM results, particularly for Maf protein.

For the candidate rSNP rs10773145, there were predictions of Sin3Ak-20 protein binding, though with no significant allelic differences in binding affinity between the reference T allele and the alternative C allele.

Also, for both candidate rSNPs rs10846834 and rs10773145 we found in Haploreg v4.1 and in RegulomeDB, results of ChIP-seq experiments showing that these two SNPs are located in a region that has STAT3 and c-FOS proteins binding, in MCF10A cell line. We analysed these results on IGV, to verify the intensity of the DNA-protein binding and the binding affinity of STAT3 and c-FOS proteins towards the two alleles of the SNPs rs10846834 and rs10773145, by comparing the reads for each allele. Regarding STAT3 protein binding, both candidate rSNPs had less than twenty ChIP-seq reads count (a total 15 reads for rs10773145 and 9 reads for rs10846834), which can lead to uncertainties on the differences in allelic binding affinity between for each candidate rSNPs. Therefore, other functional studies are needed to confirm this binding. On the other hand, there was strong evidence of c-FOS protein binding at both SNPs, with a total of 48 reads for rs10773145 and 94 reads for rs10846834. However, it did not seem that the c-FOS protein had binding preference for either allele of SNP rs10773145 or SNP rs10846834. Nevertheless, it is known that c-FOS is a common co-factor of STAT3, as their DNA binding sites co-occur proximally together. STAT3 is a transcription factor that regulates gene expression, including the FOS proto-oncogene, involved in cell proliferation, differentiation and apoptosis, and therefore STAT3 is normally constitutively activated in cancer cells, playing a crucial role in carcinogenesis. c-FOS is a bZIP protein that dimerizes with proteins of the JUN family, forming the AP-1 transcription factor complex. This complex is also often regulated in cancer by STAT3, influencing tumour angiogenesis, inflammation and inhibition of apoptosis (Carpenter & Lo 2014; Xiong et al. 2014). A recent study in T-47D breast cell line, showed the cooperative transcriptional interaction among STAT3, c-FOS and c-JUN (AP-1) on the *CCND1* promoter (which the coding protein is essential for the cell cycle G1/S transition). It was shown that after drug stimulation of STAT3, this protein was recruited to the *CCND1* promoter along with c-FOS and c-JUN. However, in the absence of the AP-1 complex the STAT3 recruitment was abrogated (Díaz Flaqué et al. 2013). Still, there is very limited information regarding STAT3 and

c-FOS interaction, especially in breast cancer tissue, and the discrepancy in the number of reads in the ChIP-seq experiments for c-FOS and for STAT3 did not allow us to draw any conclusions. These results may also be influenced by this interaction or by the endogenous levels of STAT3 protein.

6.2 Analysis of candidate rSNP rs12302714

The candidate rSNP rs12302714 was significantly associated with the DAE levels at SNP rs7301263, since a portion of the heterozygous individuals for rs12302714 display DAE and all individuals homozygous for rs12302714 showed no DAE for tSNP rs7301263.

We next looked for evidence of protein binding on rs12302714. For this candidate rSNP, there was only prediction of Sox protein binding, but with no significant allelic differences in binding affinity predicted for its alleles.

Additionally, there was no ChIP-seq data available for this candidate rSNP, so, in order to test if there was any preferential binding affinity between the alleles of rs12302714 we performed an EMSA assay. The results suggested that there is a capability to bind protein (shift), although with no allelic differences in affinity, as both bands for the C and T alleles appeared with similar intensity. We also did a competition assay that confirmed previous evidence of a non-preferential protein binding of both alleles. Therefore, the candidate rSNP rs12302714 is probably not the responsible for a change in transcription factor binding that could cause the DAE observed on tSNP rs7301263 located in the *AACS* gene.

6.3 *In silico* analysis of microRNAs binding for the 72 SNPs

We looked for the causal rSNP in the *AACS* gene and we found three candidates rSNPs (rs10846834, rs10773145 and rs12302714) that explained the DAE observed in the SNPs rs12581512 and rs7301263, although after functional analysis the ChIP-seq results for rs10846834 and rs10773145, and performing EMSA in vitro assay for the SNP rs12302714, the results suggested that the DAE

observed was probably not due to cis-regulatory SNPs altering the binding affinity of transcription factors. Therefore, we searched for evidence of cis-regulatory variants acting in different ways, for example, altering miRNA binding. This analysis showed predictions that 17 SNPs (including rs12302714) could alter miRNA binding affinity. Normally, miRNAs are associated to gene silencing through post-transcriptional binding to their target site (frequently in the 3'UTR of mRNA sequence), affecting the translation and the mRNA stability (Liu et al. 2012; Humphreys et al. 2005). There are studies that showed that the presence of SNPs in miRNA target site may regulate gene expression in an allele-specific manner (Wynendaele et al. 2010). Thus, miRNA regulation may be the mechanism causing DAE in the *AACS* gene. The preferential binding predicted for miRNA hsa-miR-4506 to the T allele of the candidate rSNP rs12302714, may suggest the association of the A allele of this SNP with gene expression regulation, since hypothetically, the mRNA containing the T allele in the 3'UTR will be silenced. More functional analysis is needed to validate the miRNA results, such as, microRNA functional analysis, in order to confirm the difference in binding affinity between the alleles of the 17 SNPs, and further test exactly how it affects the *AACS* gene expression.

6.4 LD structure and Haplotype analysis for rs7307700, rs12581512, rs10846834 and rs10773145

In order to see if there is a haplotype more associated with risk for breast cancer we did a haplotype and LD structure analysis. The DAE SNP rs7301263 was not in high LD with the other SNPs, and therefore, was not in a haplotype block. The GWAS compared the frequency between cases and controls of rs7307700 minor A allele and identified an association with risk. Therefore, it is detecting the signal from haplotypes 2 and 3 that both have the GWAS risk allele. Both haplotypes contain the GWAS SNP A allele and, in combination with haplotype 5 (the one containing the DAE SNP allele A and the rSNPs alleles G and C), accounts for a total frequency of 28.1%. On the other hand, the haplotypes with the less expressed alleles but that had the GWAS A allele (haplotypes 2 / 4, 3 / 4 and 3 / 6) represent a total frequency of 18%. Therefore, maybe the risk signal detected

from the GWAS A allele was not significantly strong due to this small difference in the ratio of signal of the haplotypes carrying the preferentially expressed alleles. If these are the true causative of risk, then haplotypes 2 / 4, 3 / 4 and 3 / 6 could be masking the effect detected at rs7307700. This may suggest that the effect of two or more alleles may be causing the risk for breast cancer and the DAE.

Additionally, haplotype 5 possess the minor alleles G and C of the candidates rSNPs rs10846834 and rs10773145, respectively, which were associated with more binding affinity to STAT3 and c-FOS proteins, suggesting that this haplotype is associated with these transcription factors binding, and perhaps, with more expression of the *AACS* gene, when these variants are found together. The preferential binding predicted for miRNA hsa-miR-4506 to the T allele of rs12302714, found in haplotype 2 together with the GWAS risk allele, may suggest the association of the A allele of this candidate rSNP with DAE.

Therefore, more *in silico* and *in vitro* analysis are needed to try to explain the DAE in 12q24 locus, that could be due to other regulatory mechanisms not studied in this work, such as allele-specific splicing events, rather than transcription factor binding. Only after clarifying the mechanism behind the DAE present in 12q24 locus we will be able to understand if this locus could also be associated with risk for breast cancer, by re-testing the rSNPs for association with risk. Thus, more analysis are needed to understand the DAE observed in the *AACS* and *UBC* genes, although we believe that the *AACS* is more probable to be associated with risk for breast cancer, since the allele more expressed of the DAE SNP located in *AACS* is part of the haplotype that includes the GWAS risk allele, whereas the DAE SNP located in *UBC* seems not to be associated with risk. In summary, our approach, combining the data of our DAE map with the GWAS breast cancer data, revealed to be a fine method to prioritize unpublished locus under influence of cis-regulatory variants, associated with the DAE levels, to further validation studies.

In the future, performing an association study with these candidate cis-regulatory variants is fundamental to associate the 12q24 locus to breast cancer risk.

Further studies similar to this work will contribute to a better understanding of the biology underlying breast cancer risk, as well as contribute to future development of cancer prevention and treatment, improving personalized medicine.

7 **Bibliografia**

Abdulkareem, I.H., 2013. Aetio-pathogenesis of breast cancer. *Nigerian medical journal : journal of the Nigeria Medical Association*, 54(6), pp.371–5.

Aloraifi, F. et al., 2015. Gene analysis techniques and susceptibility gene discovery in non-BRCA1/BRCA2 familial breast cancer. *Surgical Oncology*, 24(2), pp.100–109.

Apostolou, P. & Fostira, F., 2013. Hereditary breast cancer: the era of new susceptibility genes. *Biomed Res Int*, 2013, p.747318.

Badve, S. et al., 2011. Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists. *Modern Pathology*, 24(2), pp.157–167.

Balding, D.J., 2006. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics*, 7(10), pp.781–91.

Bannister, A.J. & Kouzarides, T., 2011. Regulation of chromatin by histone modifications. *Cell research*, 21(3), pp.381–395.

Barrett, J.C. et al., 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2), pp.263–265.

Boyle, A.P. et al., 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, 22(9), pp.1790–1797.

Bush, W.S. & Moore, J.H., 2012. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12).

Cai, Q. et al., 2011. Genome-wide association study identifies breast cancer risk variant at 10q21.2: Results from the asia breast cancer consortium. *Human Molecular Genetics*, 20(24), pp.4991–4999.

Carpenter, R.L. & Lo, H.W., 2014. STAT3 target genes relevant to human cancers. *Cancers*, 6(2), pp.897–925.

Centers for Disease Control and Prevention, 2016. Available at: http://www.cdc.gov/cancer/breast/basic_info/screening.htm

Chadwick, L.H., 2012. The NIH Roadmap Epigenomics Program data resource Lisa. *Epigenomics* , 4(3), pp.317–324.

Chen, X., Guo, L. & Fan, Z., 2007. Learning Position Weight Matrices from Sequence and Expression Data. *Comput Syst Bioinform Conf.*, 6, pp.249–260.

Chorley, B.N. et al., 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: Current and developing technologies. , 659, pp.147–157.

Consortium, T.Gte., 2015. The Genotype-Tissue Expression (GTEx) pilot

analysis: Multitissue gene regulation in humans. , 348(May), pp.648–660.

Cowper-sal, R. et al., 2012. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics*, 44(11), pp.1191–1198.

Curtis, C. et al., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403), pp.346–52.

Díaz Flaqué, M.C. et al., 2013. Progesterone receptor assembly of a transcriptional complex along with activator protein 1, signal transducer and activator of transcription 3 and ErbB-2 governs breast cancer growth and predicts response to endocrine therapy. *Breast cancer research : BCR*, 15(6), p.R118.

Dunning, A.M. et al., 2016. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nature genetics*, 48(4), pp.374–386.

Easton, D.F. et al., 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148), pp.1087–1093.

Ellis, L., Atadja, P.W. & Johnstone, R.W., 2009. Epigenetics in cancer: targeting chromatin modifications. *Molecular cancer therapeutics*, 8(6), pp.1409–1420.

Encode Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Eroles, P. et al., 2012. Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. *Cancer Treatment Reviews*, 38(6), pp.698–707.

Ferlay, J. et al., 2013. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *European Journal of Cancer*, 49(6), pp.1374–1403.

Fletcher, O. & Houlston, R.S., 2010. Architecture of inherited susceptibility to common cancer. *Nature reviews. Cancer*, 10(5), pp.353–361.

French, J.D. et al., 2013. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *American Journal of Human Genetics*, 92(4), pp.489–503.

Gage, M., Wattendorf, D. & Henry, L.R., 2012. Translational advances regarding hereditary breast cancer syndromes. *Journal of surgical oncology*, 105(5), pp.444–51.

Galvan, A., Ioannidis, J. & Dragani, T.A., 2010. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends in Genetics*, 26(3), pp.132–141.

Ghoussaini, M. et al., 2014. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nature communications*, 4, p.4999.

Ghoussaini, M., Pharoah, P.D.P. & Easton, D.F., 2013. Inherited Genetic Susceptibility to Breast Cancer. *The American Journal of Pathology*, 183(4), pp.1038–1051.

Globocan previsions, 2012. Available at: http://globocan.iarc.fr/old/burden.asp?selection_pop=224900&Text-p=World&selection_cancer=290&Text-c=All+cancers+excl.+non-melanoma+skin+cancer&pYear=18&type=1&window=1&submit=%C2%A0Execute

Glubb, D.M. et al., 2015. Fine-Scale Mapping of the 5q11.2 Breast Cancer Locus Reveals at Least Three Independent Risk Variants Regulating MAP3K1. *The American Journal of Human Genetics*, 96(1), pp.5–20.

Gorbatenko, A. et al., 2014. Regulation and roles of bicarbonate transporters in cancer. *Frontiers in Physiology*, 5 APR(April), pp.1–15.

Griffiths-Jones, S. et al., 2008. miRBase: Tools for microRNA genomics. *Nucleic Acids Research*, 36(SUPPL. 1), pp.154–158.

Van Der Groep, P., Van Der Wall, E. & Van Diest, P.J., 2011. Pathology of hereditary breast cancer. *Cellular Oncology*, 34(2), pp.71–88.

Hanahan, D., 2000. The Hallmarks of Cancer. *Cell*, 100(1), pp.57–70.

Hanahan, D. & Weinberg, R.A., 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–74.

Handy, D., Castro, R. & Loscalzo, J., 2011. Epigenetic Modifications Basic Mechanisms and Role in Cardiovascular Disease. *Circulation*, 123(19), pp.2145–2156.

Hellman, A. & Chess, A., 2010. Extensive sequence-influenced DNA methylation polymorphism in the human genome. *Epigenetics & chromatin*, 3(1), p.11.

Hellman, L. & Fried, M., 2007. Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein- Nucleic Acid Interactions. *Nature protocols*, 2(8), pp.1849–1861.

Hindorff, L.A., Gillanders, E.M. & Manolio, T.A., 2011. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis*, 32(7), pp.945–954.

Huijts, P.E. a et al., 2011. Allele-specific regulation of FGFR2 expression is cell type-dependent and may increase breast cancer risk through a paracrine stimulus involving FGF10. *Breast cancer research*, 13(4), p.R72.

Humphreys, D.T. et al., 2005. MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. *Proceedings of the National Academy of Sciences of the United States of America*, 102(47), pp.16961–6.

Hurtado, A., 2013. A functional link between ER and the breast cancer SNP rs7716600. *J Proteomics Bioinform*, 6(8), p.7716600.

Jackson, a L. & Loeb, L. a, 1998. On the origin of multiple mutations in human cancers. *Seminars in cancer biology*, 8(6), pp.421–9.

Jin, W. et al., 2015. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature*, 528(7580), pp.142–6.

Johnson, A.D. et al., 2008. SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24), pp.2938–2939.

Jones, B.L. & Swallow, D.M., 2011. The impact of cis-acting polymorphisms on the human phenotype. *HUGO Journal*, 5(1–4), pp.13–23.

Kannan, M.B., Solovieva, V. & Blank, V., 2012. The small MAF transcription factors MAFF, MAFG and MAFK: Current knowledge and perspectives. *Biochimica et Biophysica Acta - Molecular Cell Research*, 1823(10), pp.1841–1846.

Kent, W.J. et al., 2002. The Human Genome Browser at UCSC. *Genome Research*, 12(6), pp.996–1006.

Knight, J., 2014. Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Medicine*, 6(10), p.92.

Li, Q. et al., 2013. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, 152(3), pp.633–641.

Lilit Garibyan, N.A., 2013. Research Techniques Made Simple : Polymerase Chain Reaction (PCR). *NIH Public Access*, 133(3), pp.1–8.

Liu, C. et al., 2012. MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC genomics*, 13, p.661.

Lodish, H. et al., 2000. Proto-Oncogenes and Tumor-Suppressor Genes. In *Molecular Cell Biology. 4th edition*.

Long, J. et al., 2010. Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. *PLoS genetics*, 6(6), p.e1001002.

Maia, A.-T. et al., 2012. Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. *Breast cancer research : BCR*, 14(2), p.R63.

Malhotra, G.K. et al., 2010. Histological, molecular and functional subtypes of breast cancers. *Cancer Biology and Therapy*, 10(10), pp.955–960.

Mardis, E.R., 2007. ChIP-seq: welcome to the new frontier. *Nat Methods*, 4(8), pp.613–614.

McPherson, K., Steel, C.M. & Dixon, J.M., 1994. ABC of breast diseases. Breast cancer--epidemiology, risk factors and genetics. *BMJ : British Medical Journal*, 309(6960), pp.1003–1006.

Meyer, K.B. et al., 2008. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS biology*, 6(5), p.e108.

Meyer, K.B. et al., 2013. Fine-scale mapping of the FGFR2 breast cancer risk locus: Putative functional variants differentially bind FOXA1 and E2F1. *American Journal of Human Genetics*, 93(6), pp.1046–1060.

Michailidou, K. 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), pp.373–80.

Milne, R.L. & Antoniou, a. C., 2011. Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers. *Annals of Oncology*, 22(SUPPL. 1), pp.11–17.

Morton, N.E., 2005. Review series Linkage disequilibrium maps and association mapping. *Conflict*, 115(6).

National Cancer Institute, 2016. Available at: http://www.cancer.gov/types/breast/patient/breast-treatment-pdq

Neve, R.M. et al., 2006. A collection of breast cancer cell lines for the study of functionally. *Cancer Cell*, 10(6), pp.515–527.

Nica, A.C. & Dermitzakis, E.T., 2013. Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 368(1620), p.20120362.

Ohgami, M. et al., 2003. Expression of acetoacetyl-CoA synthetase, a novel cytosolic ketone body-utilizing enzyme, in human brain. *Biochemical Pharmacology*, 65(6), pp.989–994.

Oldenhuis, C.N.A.M. et al., 2008. Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer*, 44(7), pp.946–953.

Osborne, C., 2004. Oncogenes and Tumor Suppressor Genes in Breast Cancer: Potential Diagnostic and Therapeutic Applications. *The Oncologist*, 9(4), pp.361–377.

Ott, J., Wang, J. & Leal, S.M., 2015. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*, 16(5), pp.275–284.

Pastinen, T., 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature reviews*, 11(August), pp.533–538.

Perou, C.M. et al., 2000. Molecular portraits of human breast tumours. *Nature*, 406(6797), pp.747–752.

Puliti, A. et al., 2007. Teaching molecular genetics : chapter 4 — positional cloning of genetic disorders. *Pediatriac Nephrology*, pp.2023–2029.

Quigley, D.A. et al., 2014. The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. *Molecular Oncology*, 8(2), pp.273–284.

Ripperger, T. et al., 2009. Breast cancer susceptibility: current knowledge and implications for genetic counselling. *European journal of human genetics : EJHG*, 17(6), pp.722–31.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24–26.

Sætrom, P. et al., 2009. A risk variant in an miR-125b binding site in BMPR1B is associated with breast cancer pathogenesis. *Cancer Research*, 69(18), pp.7459–7465.

Sassen, S., Miska, E.A. & Caldas, C., 2008. MicroRNA - Implications for cancer. *Virchows Archiv*, 452(1), pp.1–10.

Schaub, M.A. et al., 2012. Linking disease associations with regulatory information in the human genome. *Genome Research*, 22(9), pp.1748–1759.

Schug, Z.T. et al., 2015. Acetyl-CoA synthetase 2 promotes acetate utilization and maintains cancer cell growth under metabolic stress. *Cancer Cell*, 27(1), pp.57–71.

Shah, R., Rosso, K. & Nathanson, S.D., 2014. Pathogenesis, prevention, diagnosis and treatment of breast cancer. *World journal of clinical oncology*, 5(3), pp.283–98.

Shephard, N.D. et al., 2009. A breast cancer risk haplotype in the caspase-8 gene. *Cancer Research*, 69(7), pp.2724–2728.

Sherry, S.T. et al., 2001. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1), pp.308–311.

Sorlie, T. 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), pp.10869–74.

Sorlie, T. et al., 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14), pp.8418–23.

Stoneking, M., 2001. From the evolutionary past. *Nature*, 409(February).

The 1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422), pp.56–65.

Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), pp.178–192.

Torkamani, A. & Schork, N.J., 2008. Predicting functional regulatory polymorphisms. *Bioinformatics (Oxford, England)*, 24(16), pp.1787–1792.

Wang, X. et al., 2005. Single nucleotide polymorphism in transcriptional regulatory regions and expression of environmentally responsive genes. *Toxicology and Applied Pharmacology*, 207(2 SUPPL.), pp.84–90.

Wang, Y. et al., 2014. Evaluation of functional genetic variants at 6q25.1 and risk of breast cancer in a Chinese population. *Breast cancer research : BCR*, 16(4), p.422.

Wang, Z. et al., 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7), pp.897–903.

Ward, L.D. & Kellis, M., 2011. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(D1), pp.1–5.

World Health Organization, 2016. Available at: http://www.who.int/topics/cancer/en/.

Wynendaele, J. et al., 2010. An illegitimate microRNA target site within the 3′ UTR of MDM4 affects ovarian cancer progression and chemosensitivity. *Cancer Research*, 70(23), pp.9641–9649.

Xiao, R. & Scott, L.J., 2011. Detection of cis-acting regulatory SNPs using allelic expression data. *Genetic Epidemiology*, 35(6), pp.515–525.

Xiong, A. et al., 2014. Transcription factor STAT3 as a novel molecular target for cancer prevention. *Cancers*, 6(2), pp.926–957.

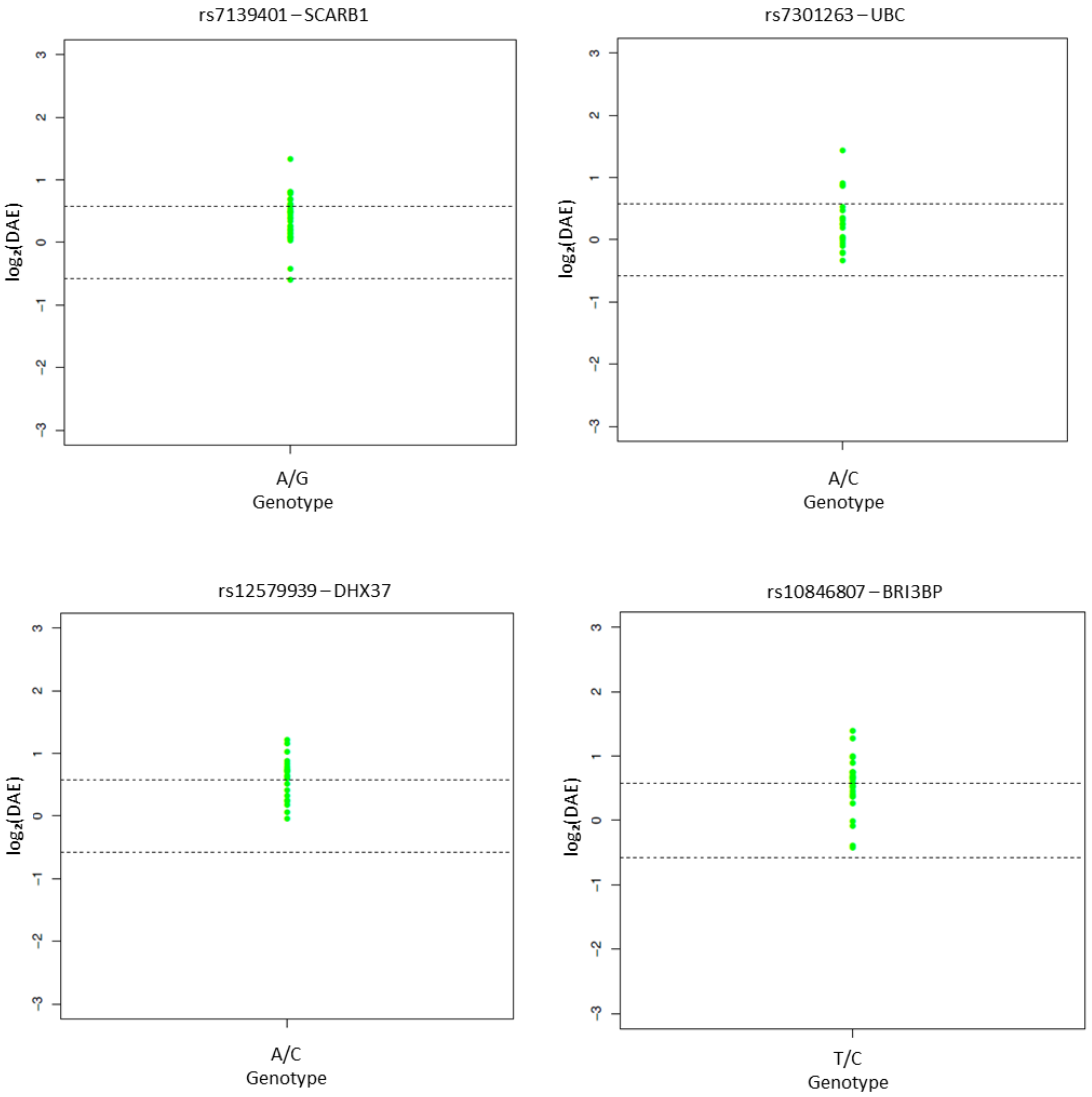
Yates, A. et al., 2016. Ensembl 2016. *Nucleic Acids Research*, 44(D1), pp.D710–D716.

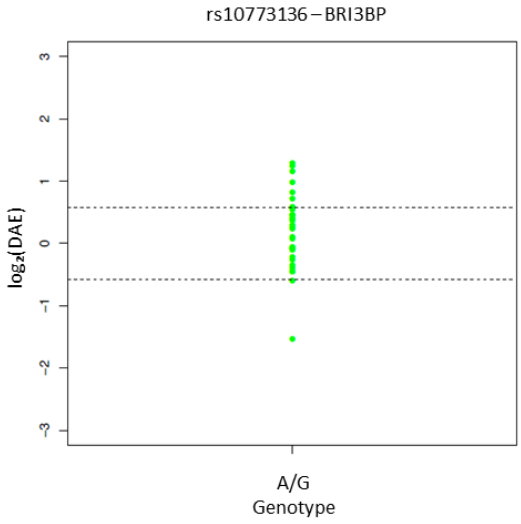
Yokota, J., 2000. Tumor progression and metastasis. *Carcinogenesis*, 21(3), pp.497–503.

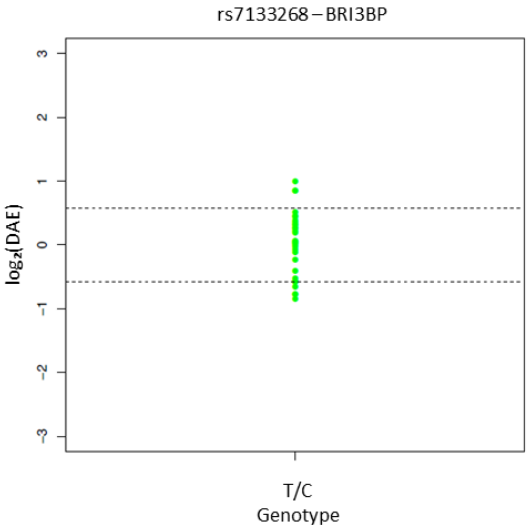
Youlden, D.R. et al., 2012. The descriptive epidemiology of female breast cancer: An international comparison of screening, incidence, survival and mortality. *Cancer Epidemiology*, 36(3), pp.237–248.

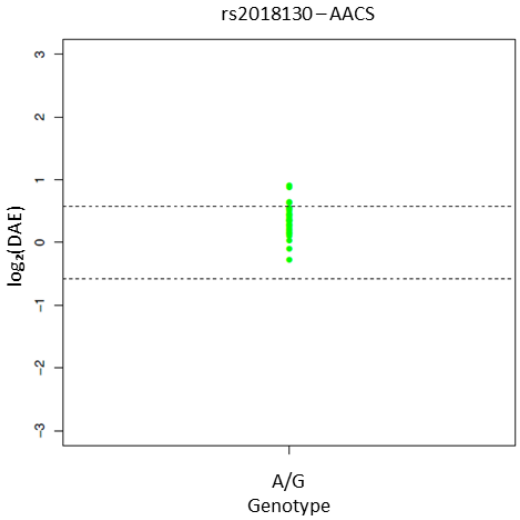
Annex 1

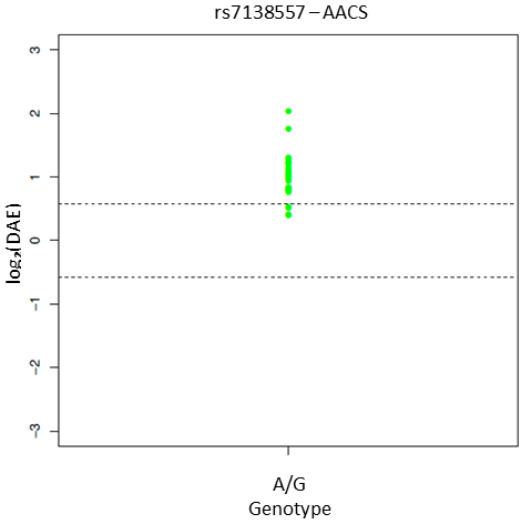
Annex 1.1 DAE SNPs reported in previous results obtained in microarray (Maia et al, unpublished). These are the 15 DAE SNPs that we chose to validate. The x-axis indicates the genotype (all heterozygous) and the y-axis indicates the normalised DAE ratio obtained. Dotted lines delimit the cut-off of preferential allelic expression ratio [log2(1.5) = 0.584].

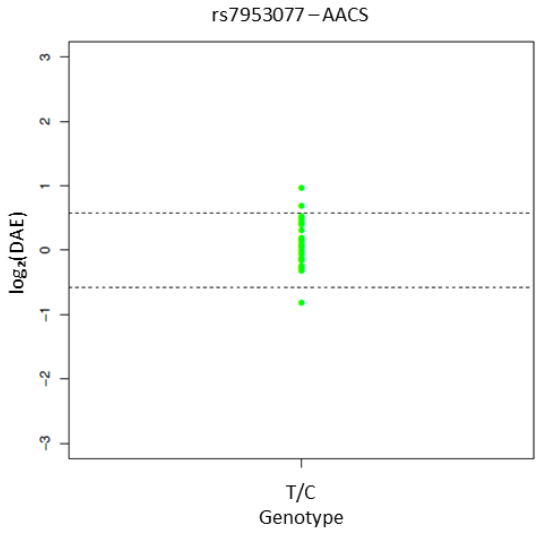


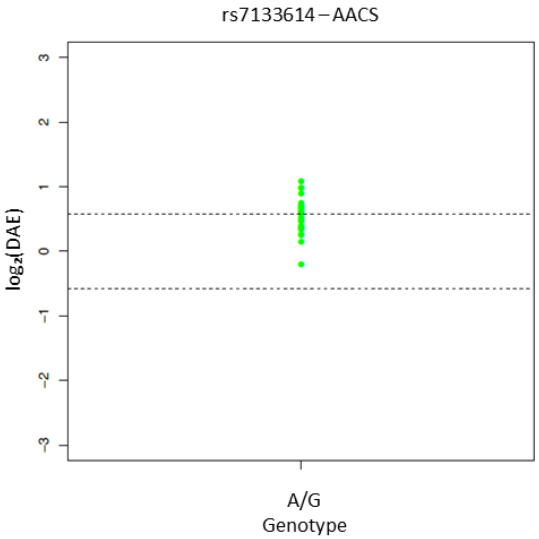

rs10773136 – BRI3BP
log2(DAE)
3
2
1
0
−1
−2
−3
A/G
Genotype

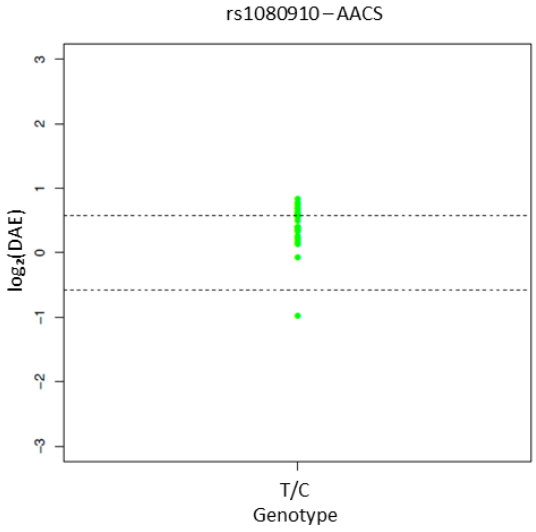

rs7133268 – BRI3BP
log2(DAE)
3
2
1
0
−1
−2
−3
T/C
Genotype

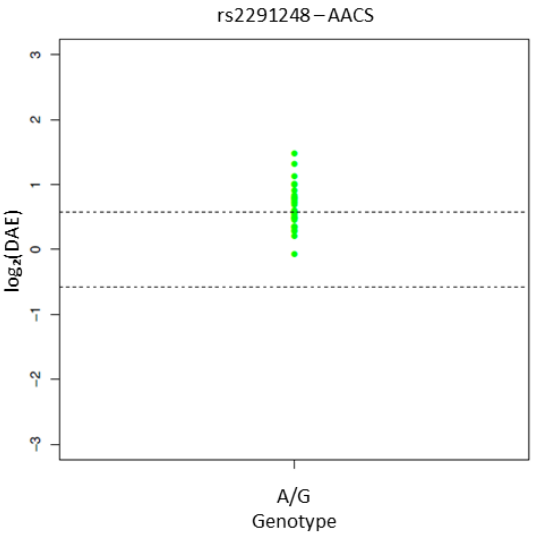

rs2018130 – AACS
log2(DAE)
3
2
1
0
−1
−2
−3
A/G
Genotype

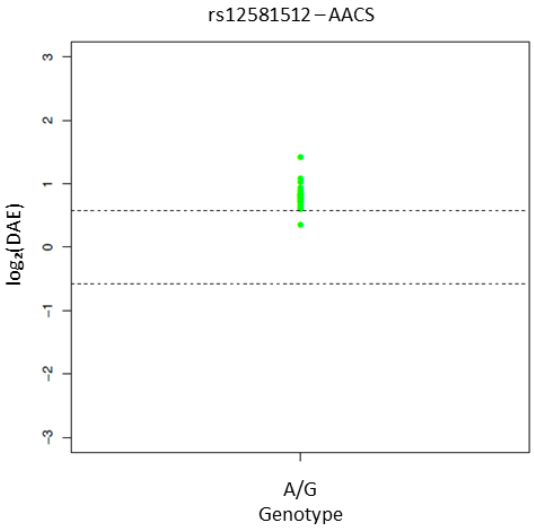

rs7138557 – AACS
log2(DAE)
3
2
1
0
−1
−2
−3
A/G
Genotype

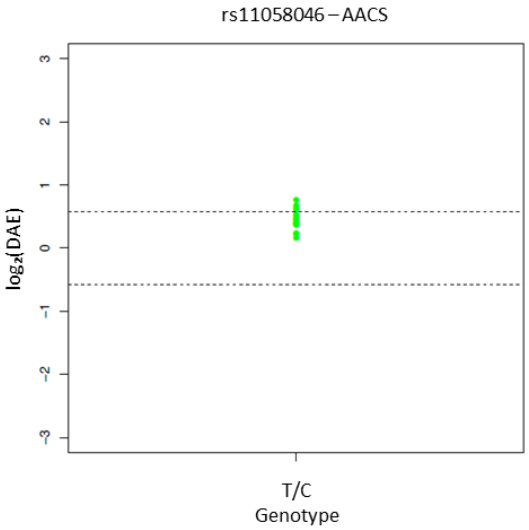

rs7953077 – AACS
log2(DAE)
3
2
1
0
−1
−2
−3
T/C
Genotype

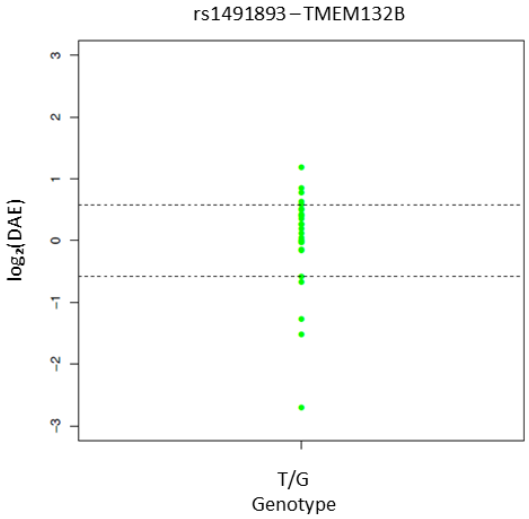

rs7133614 – AACS
log2(DAE)
3
2
1
0
−1
−2
−3
A/G
Genotype


rs1080910 – AACS
log2(DAE)
3
2
1
0
−1
−2
−3
T/C
Genotype


rs2291248 – AACS
log2(DAE)
3
2
1
0
−1
−2
−3
A/G
Genotype


rs12581512 – AACS
log2(DAE)
3
2
1
0
−1
−2
−3
A/G
Genotype


rs11058046 – AACS
log2(DAE)
3
2
1
0
−1
−2
−3
T/C
Genotype


rs1491893 – TMEM132B
log2(DAE)
3
2
1
0
−1
−2
−3
T/G
Genotype

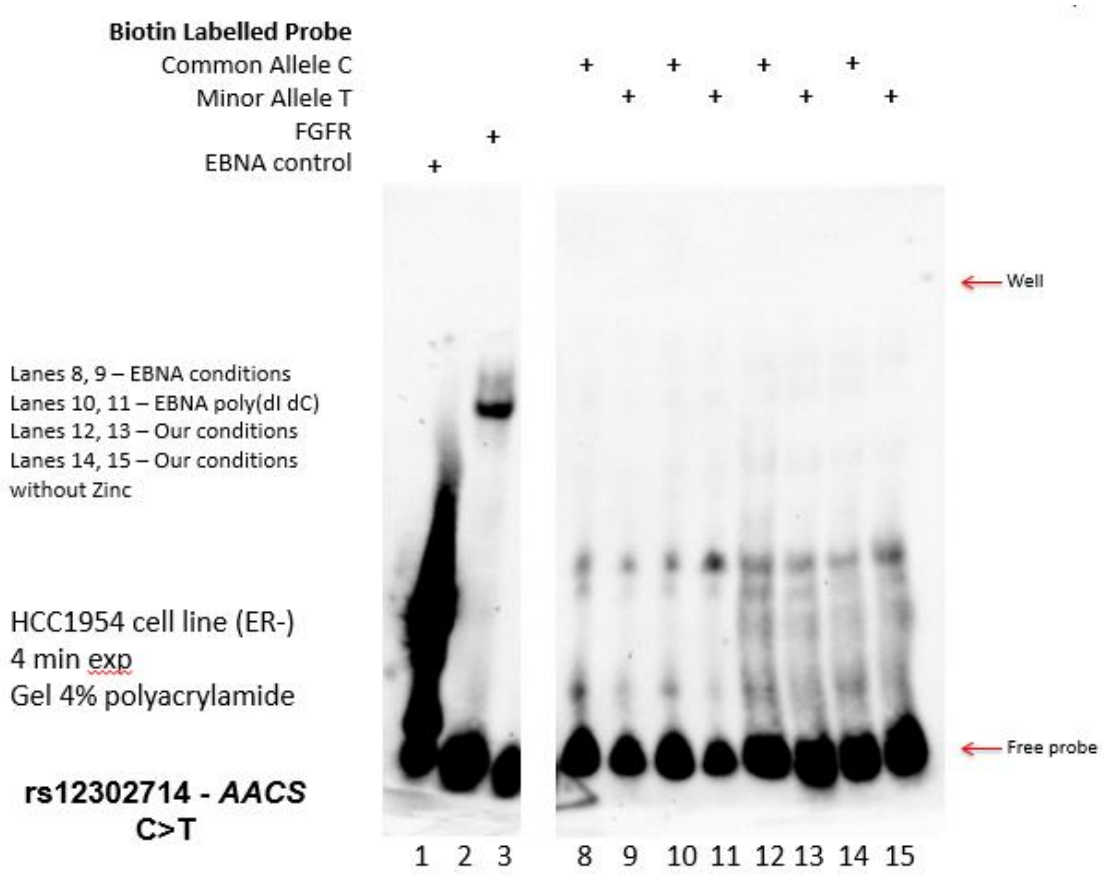
Annex 1.2 List of the 72 proxy SNPs in LD ≥ 0.4 with the GWAS SNP. 12 SNPs (in bold) were in at least one active promotor or enhancer and in DHS, in breast cell lines. The rest of the SNPs were excluded from our analysis.

SNP	Active enhancer or promotor in breast cell line	DNase I
rs10846828	Yes	Yes
rs12302714	Yes	Yes
rs10846834	Yes	Yes
rs10773145	Yes	Yes
rs7133614	Yes	Yes
rs10773146	Yes	Yes
rs12578446	Yes	Yes
rs34151902	Yes	Yes
rs4765021	Yes	Yes
rs4622332	Yes	Yes
rs4559740	Yes	Yes
rs7294703	Yes	Yes
rs4765217	Yes	Not in breast
rs56394386	Yes	Not in breast
rs4765218	Yes	Not in breast
rs7970937	Yes	Not in breast
rs12371384	Not in breast	Not in breast
rs7135489	Not in breast	No data
rs1384556	Not in breast	Yes
rs6488989	Yes	Not in breast
rs2018130	Not in breast	Not in breast
rs7955201	Yes	No data
rs7138557	Not in breast	Yes
rs7137679	Yes	No data
rs900410	Yes	Not in breast
rs10846824	Not in breast	No data
rs55999005	Not in breast	No data
rs7953077	Yes	Not in breast
rs10400509	Not in breast	No data
rs58416336	Yes	Not in breast
rs2291247	Not in breast	Not in breast
rs2291248	Not in breast	Not in breast
rs34624329	Not in breast	Not in breast
rs3751181	Yes	No data
rs12581512	Yes	Not in breast
rs12316499	Not in breast	Not in breast
rs10846829	Not in breast	Not in breast
rs7398636	Not in breast	Yes
rs7136220	Not in breast	Yes
rs7133006	Yes	Not in breast
rs7133864	Yes	Not in breast
rs1080910	Yes	Not in breast

rs34961756	Yes	Not in breast
rs41473449	Yes	Not in breast
rs35620656	Yes	Not in breast
rs12580221	Not in breast	Yes
rs11058031	Not in breast	Not in breast
rs2297478	Not in breast	Yes
rs12303572	Not in breast	Not in breast
rs57491100	Not in breast	Yes
rs57031290	Not in breast	No data
rs12303416	Not in breast	No data
rs12305181	Not in breast	Yes
rs35941060	Not in breast	Yes
rs58624919	Not in breast	Yes
rs900411	Not in breast	Not in breast
rs6488984	Not in breast	Not in breast
rs11058053	Yes	Not in breast
rs35933435	Yes	Not in breast
rs10773140	Not in breast	No data
rs7307545	Not in breast	No data
rs7315347	Yes	Not in breast
rs7954593	Not in breast	No data
rs10744191	Not in breast	Not in breast
rs10773142	Yes	Not in breast
rs4765214	Not in breast	No data
rs56255932	Yes	Not in breast
rs34107239	Yes	Not in breast
rs10846822	Not in breast	No data
rs1029075	Not in breast	Yes
rs7963307	Not in breast	No data
rs2343542	Not in breast	No data

Annex 2

Annex 2.1 EMSA for candidate rSNP rs12302714, showing no differences in binding affinity between the alleles, even in different binding conditions. Nuclear extract from HCC1954 cell line was used. Lane 1 corresponds to EBNA control; lane 2 contain the positive control FGFR-13 oligonucleotides. Lanes 8 and 9 corresponds to EBNA control conditions; lanes 10 and 11 is EBNA conditions but with less poly(dI dC); lanes 12 and 13 is our conditions (normal); and lanes 14 and 15 is our conditions but without zinc.



Annex 2.2.1 Different EMSA conditions used to test if: 1) the kit conditions are improved; 2) the concentrations of poly(dI.dC) could alter the binding process since more protein may bind to the poly(dI.dC) sequence instead of the oligonucleotide of interest; 3) zinc ions could affect the

binding process, since for the control kit EBNA no zinc is added to the reaction; 4) our conditions is more adequate for EMSA assay.

1) EBNA conditions

Component (initial concentration)	Final concentration (in 20µL)
10X BB	1X
Poly(dI.dC) (1µg/µL)	50ng/µL
50% Glycerol	2,50%
100mM MgCl2	5mM
1% NP-40	0,05%
H20	
Lysate	10µg

2) EBNA conditions poly(dIdC)

Component (initial concentration)	Final concentration (in 20µL)
10X BB	1X
Poly(dI.dC) (1µg/µL)	10ng/µL
50% Glycerol	2,50%
100mM MgCl2	5mM
1% NP-40	0,05%
H20	
Lysate	10µg

3) Our conditions without Zn

Component (initial concentration)	Final concentration (in 20µL)
5 x BB	1X
Poly(dI.dC) (1µg/µL)	10ng/µL
25x Prot. Inhibitor	1x
100mM DTT	1mM
H20	
Lysate	10µg
Buffer C	1X

4) Our conditions

Component (initial concentration)	Final concentration (in 20µL)
5 x BB	1X
Poly(dI.dC) (1µg/µL)	10ng/µL
25x Prot. Inhibitor	1x
100mM DTT	1mM
H20	
Lysate	10µg
Buffer C	1X