UAlg
UNIVERSIDADE DO ALGARVE
DEPARTAMENTO DE CIÊNCIAS BIOMÉDICAS E MEDICINA

UNIVERSITY OF ALGARVE

DEPARTAMENTO DE CIÊNCIAS BIOMÉDICAS E MEDICINA

Identification of Regulatory Polymorphisms Associated with Breast Cancer Risk

Cátia Sofia Lourenço Rocha

Master Thesis

Biomedical Sciences

Supervisor

Professor Doctor Ana Teresa Maia

2014

UNIVERSITY OF ALGARVE

DEPARTAMENTO DE CIÊNCIAS BIOMÉDICAS E MEDICINA

Identification of Regulatory Polymorphisms Associated with Breast Cancer Risk

Cátia Sofia Lourenço Rocha

Master Thesis

Biomedical Sciences

Supervisor

Professor Doctor Ana Teresa Maia

2014

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 a todos que contribuíram de alguma forma para que esta dissertação fosse realizada. A qual dita o fim de mais uma etapa académica e que não seria possível sem o apoio e amizade de diversas pessoas. Não sendo viável citá-los a todos, há pessoas a quem não posso deixar de manifestar o meu apreço e gratidão.

À minha orientadora, Professora Doutora Ana Teresa Maia, um mero parágrafo não seria suficiente para demonstrar toda a consideração que lhe tenho. O meu sincero agradecimento pela disponibilidade manifestada e conhecimentos transmitidos em todo o processo de orientação científica. Pelo reconhecimento, incentivo e confiança depositada que contribuíram decisivamente para o resultado final que aqui apresento. Um obrigado por me ter ajudado a crescer em termos pessoais e profissionais.

À Joana Xavier, pela amizade construída ao longo deste projeto. Pelo profissionalismo e por todos os conhecimentos que partilhou comigo. Pelo tempo despendido na análise e cruzamento dos dados, o que tornou possível o grande avanço neste projeto.

Um agradecimento muito especial ao André Candelária, Sandra Miranda e Vera Sousa (estudantes do Mestrado Integrado em Medicina) e à Sofia Ferreira pela preciosa ajuda despendida no Laboratório, um grande obrigado também pelo companheirismo e boa disposição.

A todos os meus amigos, principalmente à Helena pelo apoio e amizade, por ter tornado estes dois anos cheios de boas recordações.

À minha irmã Susana e ao Roberto, ouvintes atentos de algumas dúvidas e inquietações, pela confiança e valorização entusiasta do meu trabalho. Obrigado pelas constantes manifestações de interesse e encorajamento.

O maior agradecimento é dirigido aos meus pais pelo constante sacrifício que fizeram para que eu pudesse chegar até aqui. Por acreditarem sempre em mim até quando eu própria duvidei. Por me incentivarem perante os desafios a fazer mais e melhor. Quero partilhar convosco a alegria de os conseguir vencer continuamente!

ABSTRACT

Breast cancer is the most common cancer among women and it is the most frequent cause of cancer death in women. In its aetiology are involved non-genetic and genetic factors. However, the current knowledge of breast cancer genetic risk cannot explain as much as two-thirds of familial cases. All common risk polymorphisms identified by GWAS and functionally analysed are cis-regulatory. Therefore, we **hypothesise that the remaining risk-associated common polymorphisms are likely to also be cis-regulatory**. The major challenges in the post-GWAS era include identification of the casual variant of risk-loci and to understand their link to disease aetiology. The present work aims to validate cis-regulatory variants for breast cancer, identified in a previous differential allelic expression (DAE) study. We also suggest a new approach to prioritise GWAS candidate risk loci for further functional characterization and validation.

Initially, DAE ratios for ten DAE and four non-DAE SNPs were validated in breast tissue samples from healthy controls, using allele-specific real-time PCR. We confirmed DAE in ten SNPs and non-DAE in one. Discordant results are being further analysed.

Subsequently, the DAE data was integrated with published breast cancer GWAS data according to genomic location and linkage disequilibrium (LD). This allowed us to identify several loci that contain both risk-associated SNPs and DAE SNPs, in strong LD. The two top candidate loci (17q22 and 12q24) were selected for functional characterisation. We have found evidences that support a cis-regulatory role for these risk-associated loci.

In this study we report a large overlap between GWAS and DAE data, confirming that cis-regulatory variants are indeed major players in breast cancer susceptibility, and that DAE studies are a good approach for identifying new susceptibility loci for breast cancer, including prioritising candidate GWAS for functional analysis.

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

RESUMO

O cancro da mama é uma das patologias oncológicas mais comuns, e causa de morte mais frequente entre as mulheres. Trata-se de uma doença complexa com fatores genéticos e não genéticos envolvidos na sua etiologia. Até ao momento, o conhecimento adquirido acerca do risco genético não explica mais do que dois terços dos casos de cancro familiares permanecendo ainda por explicar cerca de 65% a 70%. Com os recentes avanços na tecnologia de microarrays e nos estudos de associação realizados a nível de genoma inteiro (*genome-wide association studies*, GWAS), têm surgido evidências que sugerem que as variantes cis-reguladoras podem ser importantes para o risco de cancro da mama. Variantes cis-reguladoras são polimorfismos que regulam a expressão de genes. Podem-no fazer através da modificação de elementos reguladores como, por exemplo: afetando a ligação de fatores de transcrição em promotores e elementos intensificadores Essas variantes são frequentes na população e contribuem para a variabilidade intra e inter -espécies.

Com base nestas evidências, a nossa hipótese consiste em que a maioria dos polimorfismos comuns e associados a risco ainda por identificar podem ser também cis-reguladores. Deste modo, a procura para a restante suscetibilidade genética do cancro da mama deve centrar-se nas variantes com maior potencial cis-regulador. As variantes cis-reguladoras podem ser identificadas de diferentes formas, como por exemplo através de loci de características quantitativas de expressão (*expression quantitative trait loci*, eQTL) e análises de expressão diferencial alélica (*differential allelic expression*, DAE) em indivíduos heterozigóticos. DAE é uma abordagem que compara os níveis relativos de expressão dos dois alelos do mesmo gene em indivíduos heterozigóticos, utilizando um SNP transcrito (tSNP). Esta técnica tem revelado ser bastante eficaz, uma vez que compara os níveis de transcritos dos alelos dentro do mesmo contexto celular e de haplótipos, pelo que a influência de efeitos trans é eliminada (por exemplo concentração de fatores de transcrição no núcleo).

Este trabalho teve dois grandes objetivos: o primeiro, a validação da dos resultados de DAE nos polimorfismos cis-reguladores identificados num estudo anteriormente realizado pela Prof. Ana Teresa Maia; o segundo, a identificação de novos loci envolvidos no risco para o cancro da mama.

No estudo para validação da DAE, foi medida a quantidade relativa da expressão dos dois alelos de 10 SNPs, previamente identificados como apresentando DAE (rs2526935; rs10503416; rs10513376; rs7600326; rs6494466; rs10016; rs13265801; rs9250; rs8097892; rs1384) e de 4 SNPs sem DAE (rs2834653; rs710945; rs1477017 e rs10521). Foram utilizadas 18 amostras de tecido mamário de indivíduos saudáveis e os níveis da DAE foram determinados utilizando a técnica de PCR em tempo real específica alélica. As distribuições de DAE foram comparadas aplicando um Teste t de *Student* (t-test), o qual compara as médias da distribuição de rácios de DAE. No total dos 10 SNPs com DAE validámos 6 SNPs que eram consistentes com os resultados observados anteriormente. Apenas um de quatro SNPs sem evidências da DAE mostrou consistência com os resultados anteriores.

Para o estudo de identificação de novos loci foi aplicada uma abordagem inovadora, a qual consiste no cruzamento de dados para cancro da mama publicados nos GWAS com os nossos dados de DAE. Esse cruzamento de dados é feito de acordo com a localização cromossómica, distância física (janelas de ±250kb a partir da variante com DAE) e padrões de desequilíbrio de ligação (*linkage disequilibrium,* LD). Com este exercício pretendemos testar se esta abordagem pode ajuda a priorizar os loci candidatos de GWAS para validação e posteriores análises funcionais. Este cruzamento de dados permitiu identificar vários loci que contêm SNPs associados com risco para cancro da mama e SNPs com DAE. Para uma análise inicial selecionaram-se dois loci: 17q22 (*TOM1L1*/*COX11*/*STXBP4*) e 12q24 (*AACS*), com base nos cenários de DAE, LD entre o tSNP de DAE e o SNP do GWAS. Esses dois loci foram inicialmente analisados para potenciais elementos reguladores e evidência funcional (promotores, intensificadores e ligação de fatores de transcrição) nos locais onde se encontram os candidatos a variantes cis-reguladoras. Para essa análise foram utilizadas uma base de dados, RegulomeDB, e um navegador de informação genómica, Genome Browser,

que contêm informações sobre hipersensibilidade à desoxirribonuclease (DNAse), TFBS, e regiões promotoras, evidências obtidas em *estudos in silico* e/ou *in vitro* para variantes codificadoras e não codificadoras.

No locus 17q22, foram encontrados 12 SNPs em sobreposição com regiões que contêm marcadores para elementos funcionais, tais como promotores e intensificadores. Sugerindo que esses podem possivelmente ter um efeito funcional através da regulação da expressão do gene. Posteriormente, analisaram-se os SNPs para a estrutura de LD nessa região e para identificação dos haplótipos na população os quais podem ser responsáveis pelo aumento ou diminuição na expressão dos genes. Foram identificados quatro haplótipos comuns associados com diferenças nos níveis de expressão.

No locus 12q24, o tipo de análises realizadas foram as mesmas descritas para o outro locus. Para o locus 12q24 foram encontrados 13 SNPs em sobreposição com regiões contendo elementos reguladores o que sugere um possível efeito funcional na regulação da expressão dos genes. Todos SNPs foram também analisados para a estrutura de LD e identificação de haplótipos. Com base nessa análise, foram identificados quatro haplótipos comuns associados com as diferenças nos níveis de expressão. Essas variantes candidatas a serem cis-reguladoras foram selecionadas com base em análises *in silico*, utilizando ferramentas para a previsão de potenciais locais de ligação de fatores de transcrição. Posteriormente, 3 SNPs foram analisadas funcionalmente *in vitro*, numa linha celular de cancro da mama (HCC1954).

Como os genes presentes em cada um dos loci mostraram evidências para a presença de variantes cis-reguladoras, decidiu-se validar os níveis de DAE em tecido de mama normal e compará-las com os níveis de DAE no sangue. Para essa análise usou-se um tSNP para cada gene (rs17817901 no *COX11* e rs7138557 no *AACS*). Validaram-se os resultados da DAE, no tecido da mama, e para o *AACS* e observou-se que esses não são comparáveis com os níveis da DAE no sangue. Os nossos resultados não suportam, portanto, o uso de sangue em estudos de DAE em substituição para o tecido da mama, na aplicação futura em estudos de predisposição ao cancro da mama. Para averiguar se a DAE em ambos os loci está realmente

associada com o risco de cancro de mama, foi realizado um estudo de caso-controlo. Foi encontrada uma associação significativa neste estudo, o que sugere que DAE em ambos os loci deve ser futuramente explorado como instrumento de previsão de risco para cancro da mama.

Futuramente, uma análise mais aprofundada da regulação destes genes poderá levar também à compreensão da biologia de predisposição ao tumor e contribuir para o desenvolvimento de terapias futuras, especialmente na área da medicina personalizada.

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

INDEX OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

INDEX OF ANNEX

LIST OF ABREVIATIONS

AACS – acetoacetyl-CoA synthetase

COX11 – cytochrome-c oxidase assembly protein 11

DAE – differential allelic expression

DNA – deoxyribonucleic acid

dbSNP – database of single nucleotide polymorphism

FRET - fluorescent Resonance Energy Transfer

FGFR2 – fibroblast growth factor receptor 2

GWAS – genome wide association studies

LD – linkage disequilibrium

mRNA – messenger ribonucleic acid

RNA – ribonucleic acid

RNase – ribonuclease

SD – standard deviation

SNP – single nucleotide polymorphism

STXBP4 – syntaxin binding protein 4

TOM1L1 – target of myb1 (chicken)-like 1

cSNP – coding SNP

rSNP – regulatory SNP

tSNP – transcribed SNP

Ct – cycle threshold

CHAPTER 1 – INTRODUCTION

1.1 Cancer

Cancer is characterized by an abnormal growth and uncontrolled proliferation of cells, generated by the loss of response mechanisms to many of the signals controlling cellular growth and death. This loss of control results from an accumulation of genetic alterations that can occur in the germline, resulting in hereditary predispositions to cancer, or in somatic cells, resulting in sporadic tumours (Strachan & Read 1996; Garraway & Lander 2013).

These alterations will confer six fundamental properties that are acquired during the multistep progression of cancer and which are required for the development of a malignant tumour (Hanahan & Weinberg 2011):

1) Sustaining proliferative signalling;
2) Evading growth suppressor;
3) Resisting cell death;
4) Enabling replicative immortality;
5) Inducing angiogenesis;
6) Activating invasion and metastasis.

Genetic alterations can occur in three types of genes that are important in making a cell cancerous (Garraway & Lander 2013): oncogenes (genes that promote the cell growth and survival); tumour suppressor genes (that are involved in inhibition of cell growth and survival); and stability genes or caretakers (that are involved in mechanisms for maintaining DNA) (Vogelstein & Kinzler 2004; Ashworth et al. 2011).

The most common genetic alterations that result in oncogene activation are mutations, chromosomal translocations or gene amplifications, but other alterations can also occur in tumour suppressor genes such as deletions and allelic loss. In the case of caretaker genes, when mutated, lead to genomic instability and enhanced mutation acquisition.

In addition to genetic alterations, epigenetic events have been emerging as key mechanisms involved in the cancer development. These epigenetic events are defined as heritable changes in gene expression and chromatin structure without changes in DNA sequence. Epigenetic inheritance includes DNA methylation (hyper methylation and hypo methylation), histone modification, loss of imprinting (LOI) and relaxation of X-chromosome inactivation. These epigenetic changes also affect the same types of genes as genetic alterations (Sadikovic et al. 2008; You & Jones 2012).

1.2 Breast cancer

1.2.1 Epidemiology

Breast cancer is the second most common cancer in the world and the most frequent among women, with an worldwide estimation of 1.67 million new cancer cases diagnosed in 2012 (25% of all cancers). It is also the most frequent cause of cancer death in women (521 817deaths, 14.7% of all cancers) (http://globocan.iarc.fr/Default.aspx).

In Portugal, breast cancer is the most frequent cancer in woman with an incidence of 6088 cases per 100,000 people (85.6%). Approximately 4500 new cases of breast cancer are detected annually, and 1570 (18.4%) women die of this disease. It has been estimated that the prevalence within 5 years is of 24284 per 100,000 people (http://eu-cancer.iarc.fr/EUCAN/).

Nevertheless, breast cancer can affect both genders, but the male breast cancer incidence is much lower than for females, accounting for less than 1% of all breast carcinomas (about 349 cases of breast cancer) (http://www.cancerresearchuk.org/cancer-info/cancerstats/types/breast/incidence/uk-breast-cancer-incidence-statistics).

1.2.2 Histological, molecular and functional classification

Breast cancer is a genetically and clinically heterogeneous disease. There is a high degree of diversity between and within tumours as well as among patients (Kelsey & Berkowitz 1988; Polyak 2011). These diversities have served as the basis for disease classification. There are several approaches that can be used to classify breast cancer, namely histological, molecular and functional (Malhotra et al. 2010).

Histological classification is the most commonly used in the clinical practice and is based on the histological aspects of the primary lesion, having no molecular basis. At the histological level, breast cancer can be divided into two major groups: in situ carcinoma and invasive carcinoma (**Figure 1.1**). Based on growth patterns and cytological features the breast carcinoma in situ can be further sub-classified in ductal or lobular. Ductal carcinoma in situ (DCIS) is more common than lobular carcinoma in situ (LCIS). DCIS can be sub-classified based on the architectural features of the tumour in five subtypes: comedo, cribiform, micropapillary, papillary and solid. Similar to DCIS, invasive carcinomas are divided into several histological sub-types: infiltrating ductal, invasive lobular, ductal/lobular, mucinous (colloid), tubular, medullary and papillary carcinomas (**Figure 1.1**). Of these, infiltrating ductal carcinoma (IDC) is the most common accounting for 70-80% of all invasive lesions (Malhotra et al. 2010).

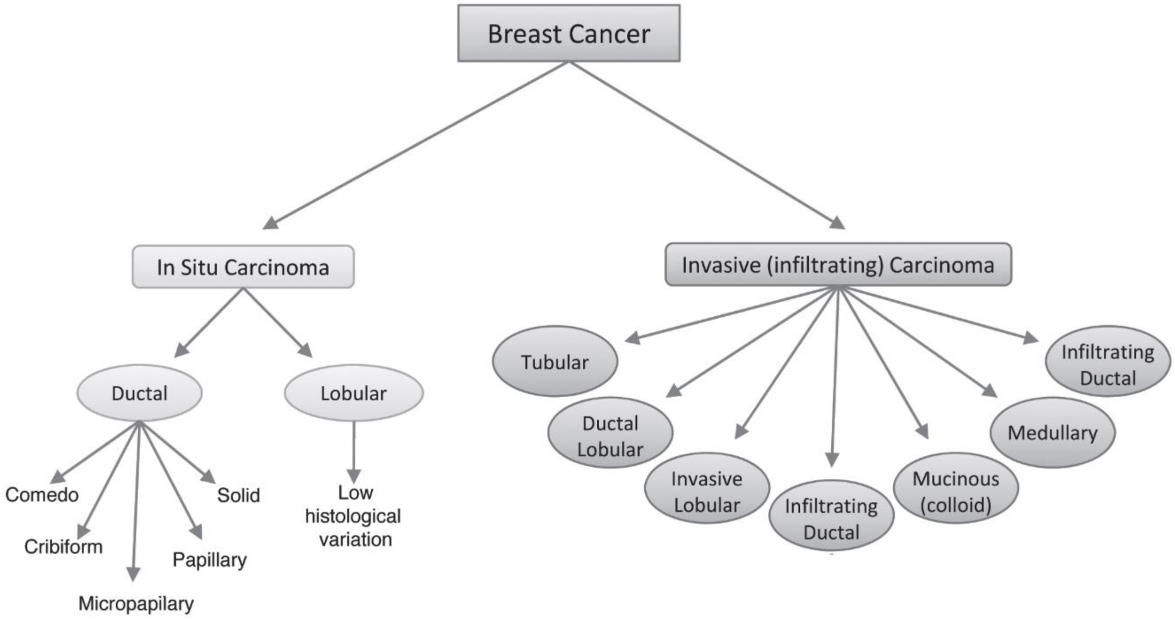


Figure 1.1 Histological classification of breast cancer. This classification is based on growth patterns and architectural features (Malhotra et al. 2010).

Molecular classification is based on gene expression profiling and combines molecular markers (such as expression of oestrogen receptor (ER), progesterone receptor (PR), v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2 (ErbB2 also known as Her2/neu)) and protein 53 (p53)) with the clinical outcome measures. At the molecular level, breast cancer can be classified as: claudin low, basal-like, ErbB2$^+$, normal breast like, luminal subtype A and luminal subtype B (**Figure 1.2**). Each of these subtypes exhibits differences in incidence, response to treatment, risk of disease progression and preferential sites of metastases. For example, patients with basal-like/triple-negative subtype (ER$^-$/PR$^-$/ErbB2$^-$) have a shorter survival. Furthermore, this classification allows the stratification of the ER$^+$ patients population into several subtypes, that shows differences in survival (Malhotra et al. 2010; Polyak 2011).

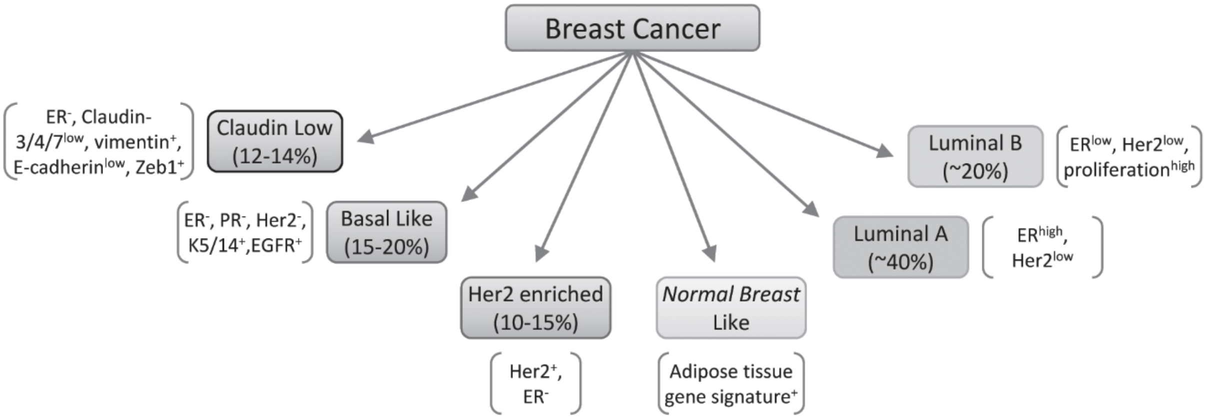

Breast Cancer
ER$^-$, Claudin-3/4/7^{low}, vimentin$^+$, E-cadherin^{low}, Zeb1$^+$
Claudin Low (12-14%)
ER$^-$, PR$^-$, Her2$^-$, K5/14$^+$,EGFR$^+$
Basal Like (15-20%)
Her2 enriched (10-15%)
Her2$^+$, ER$^-$
Normal Breast Like
Adipose tissue gene signature$^+$
Luminal A (~40%)
ER^{high}, Her2^{low}
Luminal B (~20%)
ER^{low}, Her2^{low}, proliferation^{high}

Figure 1.2 Molecular classification of breast cancer. This classification is based on gene expression profiling identified by microarray analysis (Malhotra et al. 2010).

It is important to clarify that combining both histological and molecular classifications results in significantly better predictive value than either one alone (Malhotra et al. 2010; Polyak 2011).

Concerning functional classification, this is yet an emerging area of research. This classification attempts to use one or more of the breast cancer stem cells (CSCs) markers (for example CD44/Lin) to quantify the percentage of CSCs in a patient's tumour. However, until now no specific molecular marker has been identified, because the markers used to identify normal mammary stem cells (MASCs) are the same as those used to identify CSCs. So, is still not possible to differentiate between the stem and the non-committed progenitor cells markers. Therefore, there is no conclusive markers to differentiate between stem and non-committed progenitor cells (Malhotra et al. 2010).

The development of these classification systems is of major importance since they represent a tool used in both prognosis and treatment.

1.2.3 Risk factors

Breast cancer is a complex disease with both genetic and non-genetic risk factors involved in its aetiology. The non-genetic factors are relatively well characterized and include lifestyle and environmental factors such as gender (feminine), age (65 years or more), obesity, early age ate menarche, late age at menopause, late age at first birth, use of hormones replacement therapy after menopause, alcohol consumption and radiation exposure(Oldenburg et al. 2007; Mavaddat et al. 2010).

The most important risk factor is the familiar history of disease indicating that genetic factors are important determinants of breast cancer risk (Antoniou & Easton 2006). Approximately 15-30% of breast cancer cases are attributed to hereditary factors. (Apostolou & Fostira 2013; Ghoussaini et al. 2013). However, the genetic factors are less well described than the non-genetic factors.

1.2.3.1 Breast cancer genetic susceptibility

The genetic components of breast cancer risk can be classified according to relative risk they confer and to the risk allele frequency (**Figure 1.3**). (Garcia-Closas & Chanock 2008; Apostolou & Fostira 2013; Ghoussaini et al. 2013).

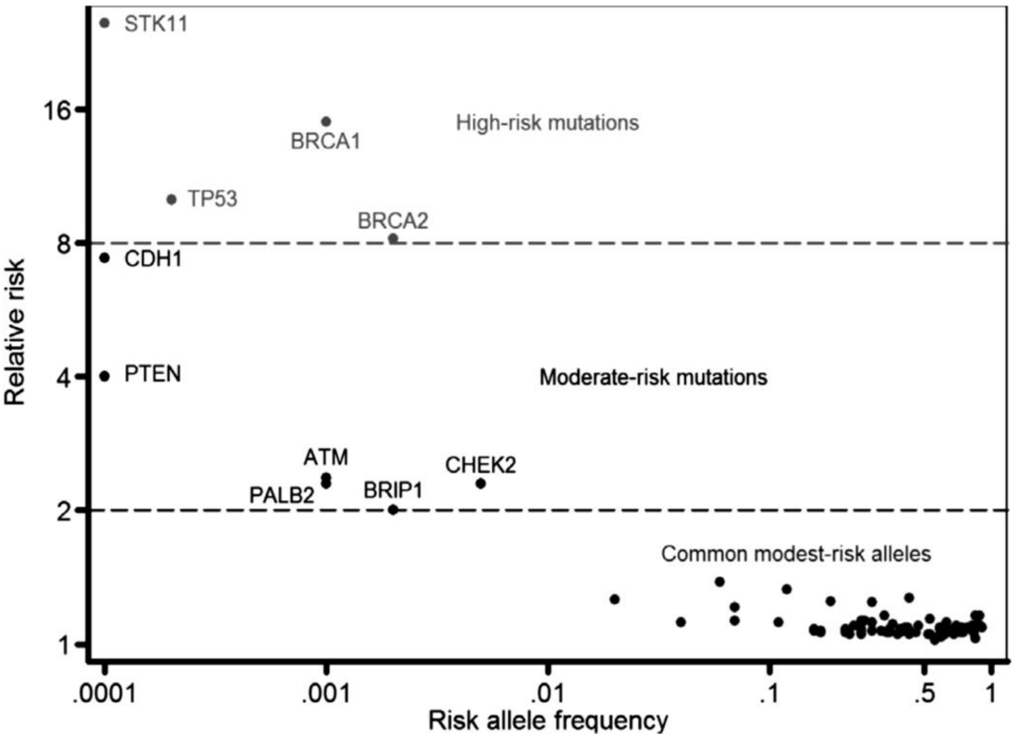


Figure 1.3 Characterization of breast cancer genetic susceptibility. This classification is based in relative risk (risk conferred) and risk allele frequency (Ghoussaini et al. 2013).

1.2.3.1.1 High-risk mutations

High risk mutations include mostly germline mutations in the genes *BCRA1* and *BCRA2*. (Walsh & King 2007), although *TP53* and *STK11*/*LKB1* mutations carriers also develop rare syndromes that predispose to cancer (Apostolou & Fostira 2013). These mutations are rare in the population but confer high lifetime risk of breast cancer (>50%) (**Figure 1.3**) (Ghoussaini et al. 2013) and explain approximately 20% to 25% of familial cases (Venkitaraman 2002; Cipollini & Tommasi 2004; Garcia-Closas & Chanock 2008; Mahdi et al. 2013) .Usually these types of mutations in familial cases are identified by twins studies, pedigree linkage analysis and studies of phenotypes associated with breast cancer risk (Antoniou & Easton 2006).

1.2.3.1.2 Moderate-risk mutations

These include mutations in *CDH1, PTEN, ATM, PALB2, BRIP1* and *CHECK2* genes that confer a lifetime risk of breast cancer of approximately 20% (**Figure 1.3**) and they account for <3% of the familial risk (Garcia-Closas & Chanock 2008; Apostolou & Fostira 2013; Ghoussaini et al. 2013; Mahdi et al. 2013).

1.2.3.1.3 Common low-susceptibility alleles

Susceptibility alleles are common in the population (minor frequency allele >5%) and these have been associated with a smaller decrease or increase in risk to breast cancer (relative risk of <1.5) (**Figure 1.3**). This increase in risk accounts for 10-20% of lifetime risk. Most of these common low risk alleles have been identified through genome wide association studies (GWAS) that have contributed with more than 70 new risk loci for breast cancer alone, including genes like *MAP3K1*, *FGFR2*, *LSP1*, *TNRC19*, *H19* and *CASP8* (Ghoussaini et al. 2013).

However, these only account for approximately 14% of the genetic component of breast cancer risk. Together with the previously known risk factors accounting for 15-30%, such as the rare high-risk mutations in *BRCA1* and *BRCA2* amongst others, there is still two thirds of the genetic burden to be determined (Ghoussaini et al. 2013).

Furthermore, the majority of variants identified in GWASes lie outside of genes, in either intronic and intergenic regions or gene deserts (Varghese & Easton 2010). It is therefore likely that protein altered functions is not the mode of action of these loci, as it was for the mutations in BRCA1 and BRCA2, for example.

1.3 Single nucleotide polymorphisms (SNPs)

Millions of genetic variations (polymorphisms) with a minor allele frequency >1% in the human population make an important contribution to disease (Buckland 2006). Common polymorphisms include tandem repeated segments (minisatellite and microsatellite), large (copy number variations) and small deletions/insertions/duplications, and single nucleotide polymorphisms (SNPs) (Wang et al. 2005; Chorley et al. 2008).

SNPs are changes of a single base in genomic DNA (gDNA) and several studies report that SNPs account for 90% of human sequence variations – occurring in 100-300 base pairs in the human genome (Brookes 1999; Wang et al. 2005; Chorley et al. 2008). The majority of SNPs are functionally neutral and some have functional effects. These SNPs with functional effects are associated with inter and intra-population diversity, susceptibility to diseases and individual response to therapeutic treatment or to environmental exposure (Wang et al. 2005).

SNPs occur throughout the genome in several sites such as coding regions (cSNPs) and non-coding regions (Gray et al. 2000; Mahdi et al. 2013). Variations in coding regions can be non-synonymous and synonymous, depending on whether they have or not an effect on phenotype at the level of protein sequence (Gray et al. 2000; Rockman & Wray 2002a). cSNPs have been extensively studied in disease since they can cause amino-acid codon alterations resulting in changes in the structure and biological properties of the encoded protein that may have important clinical consequences. Non-coding SNPs are mostly considered as non-functional, nevertheless, this type of alterations can includes those that can regulatory elements of the gene (rSNPs) (Gray et al. 2000; Rockman & Wray 2002a; Mahdi et al. 2013).

1.4 Cis-regulatory variants

Genetic, epigenetic and environmental factors contribute to gene expression control, acting in cis and trans (Stranger et al. 2007). Cis-regulatory variants are polymorphisms that regulate gene expression on the same chromosome and are mostly located immediately upstream or downstream of the gene, but can also be found in close proximity to the gene that they regulate, as well as in introns containing regulatory elements (Xiao & Scott 2011).

Cis-regulatory variants are frequent in the human genome affecting 20-60% of autosomal genes (Rockman & Wray 2002b) and are responsible for most of the phenotypic variability intra and inter species (Cheung et al. 2005; Stranger et al. 2007). These variants can be found in regulatory elements such as promoters and enhancers (**Figure 1.4**) and as well as in silencers and insulators (Jones & Swallow 2011; Pastinen et al. 2006).

The cis-regulatory variants contribute to the adaptation of populations to the environmental changes leading for example to the development of resistance (immune responses) and susceptibility to diseases (Lappalainen & Dermitzakis 2010; Jones & Swallow 2011; Worsley-Hunt et al. 2011; Vernot et al. 2012).

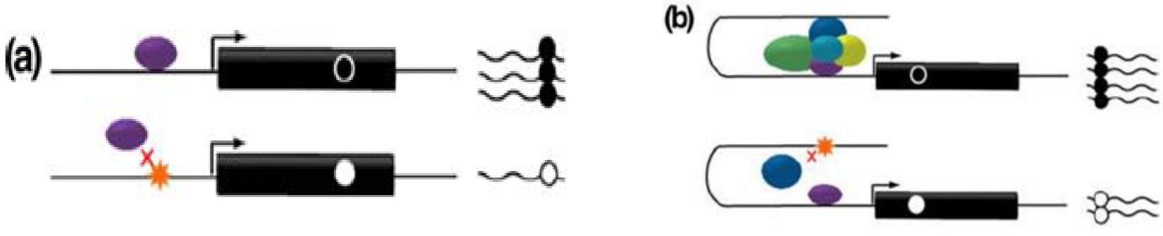


Figure 1.4 Allelic expression in heterozygous through cis-regulatory variants. SNP in a promoter **(a)** or enhancer **(b)** can affect the binding affinity of transcription factors altering the level of expression of the alleles (marker SNP is shown in black and white dots) (Jones & Swallow 2011).

However, the mechanisms by which these variants affect the phenotype are yet not totally understood. For example, mutations and polymorphisms in these elements can disrupt or increased the binding affinity of transcription factors (TF) (Jones & Swallow 2011) linked to one of the two alleles, thus altering the rate or efficiency of transcription and causing unequal levels allelic transcripts (Pastinen et al. 2006).

In fact, work performed by my supervisor Prof. Ana Teresa Maia and others has revealed that indeed most variants identified by GWAS for multiple cancers are cis-regulatory (**Table 1.1**) (Easton et al. 2007; Thomas et al. 2010; Fletcher et al. 2011; Chen et al. 2013; Michailidou et al. 2013; Turnbull et al. 2013).

Table 1.1 **Cis-regulatory SNPs associated with breast cancer risk (identified in GWAS).** In the table are shown the SNPs located locus/nearest gene, region, per allele odds ratio (OR), risk allele frequency and date when it was reported.

Best GWAS tag	Locus	Nearest Gene	Region	OR	Risk Allele Frequency	Date
rs3757318	6q25	*ESR1*	Intronic	1.16 (1.12-1.21)	0.07	2013
				1.30 (1.17-1.46)		2010
rs13281615	8q24	*MYC*	Intergenic	1.09 (1.07-1.12)	0.41	2013
rs1562430				1.16 (1.11-1.22	0.60	2011
rs2981582	10q26	*FGFR2*	Intronic	1.26 (1.23-1.30)	0.38	2007
rs2981579				1.43 (1.35-1.53)	0.42	2010
rs614367	11q13	*CCND1/FGFs* *MYEOV* *ORAOV1*	Intergenic	1.21 (1.18-1.24)	0.15	2013
				1.15 (1.10-1.20)		2010
rs12443621	16q12	*TOX3/LOC643714*	Intronic	1.11 (1.08–1.14)	0.46	2007
rs3803662				1.14 (1.10-1.18)	0.267	2013

As 66% genetic risk remains unidentified and all common risk polymorphisms identified by GWAS and functionally analysed were cis-regulatory, we **hypothesise that the remaining risk-associated common polymorphisms are likely to also be cis-regulatory**.

Cis-regulatory variants can be identified in different ways such as expression quantitative trait loci (eQTL) and differential allelic expression (DAE) analysis. eQTL analysis is an approach that looks for the association between variations in gene expression and individual genotypes (Serre et al. 2008; Nica et al. 2013). DAE is an approach that compares the relative expression of the two alleles from the same gene in an heterozygous individual (Serre et al. 2008). For the work developed under the context of this thesis we used the DAE analysis approach.

1.4.1 Differential allelic expression (DAE)

DAE identifies individual differences in the expression of the two copies of one gene (Serre et al. 2008). This approach compares the relative expression of the two transcribed alleles in complementary DNA (cDNA) from the same heterozygous samples using transcribed SNPs as allelic markers (marker SNP) (Pant et al. 2006; Serre et al. 2008; Xiao & Scott 2011; Jones & Swallow 2011).

This approach allows to control for environmental and trans-regulatory factors that affect the expression of both alleles, because each transcribed allele serves as an internal standard for the other (Verlaan et al. 2009; Jones & Swallow 2011; Xiao & Scott 2011). This approach can also detects epigenetic effects such as imprinting or random monoallelic expression (Jones & Swallow 2011; Verlaan et al. 2009).

1.5 Previous work – DAE map in normal breast tissue

In previous work, Maia et al performed a DAE scan of the entire genome in normal breast tissue samples using microarrays and obtained a whole genome map of the cis-regulatory SNPs in breast tissue (**Figure 1.5**).

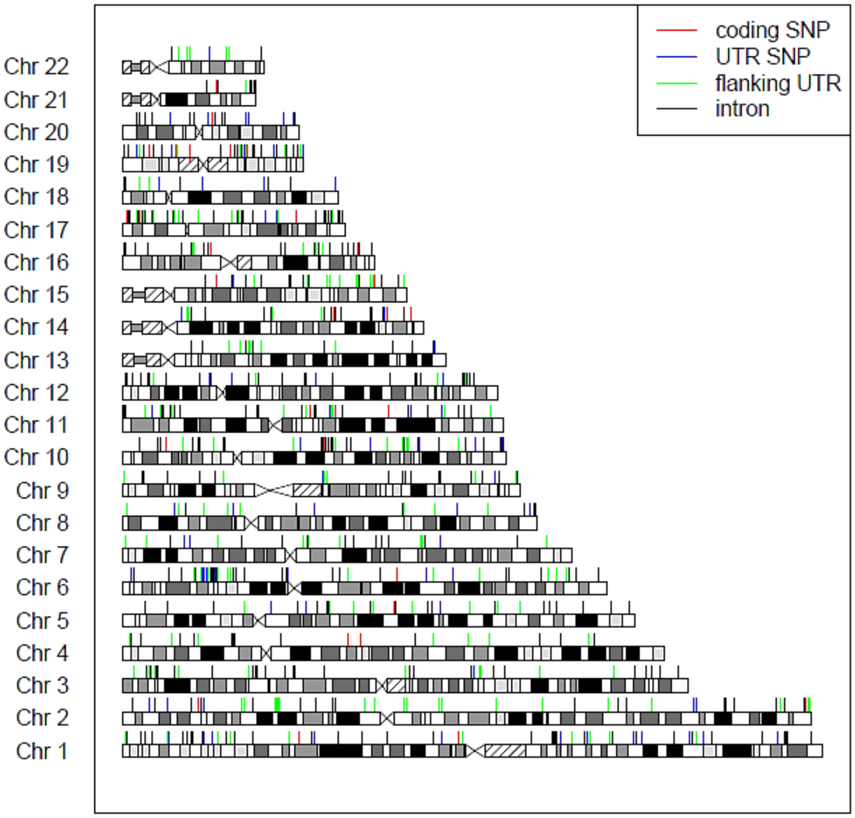


Figure 1.5 Map with the location of cis-regulatory SNPs across the genome. Results of the DAE scan in normal breast tissue samples using microarrays (Unpublished, Maia et al).

For this experiment (**Figure 1.6**) samples were genotyped and evaluated for allelic expression using Illumina Exon510S-Duo arrays. These contain approximately 500K SNPs with a predominance of coding SNPs.

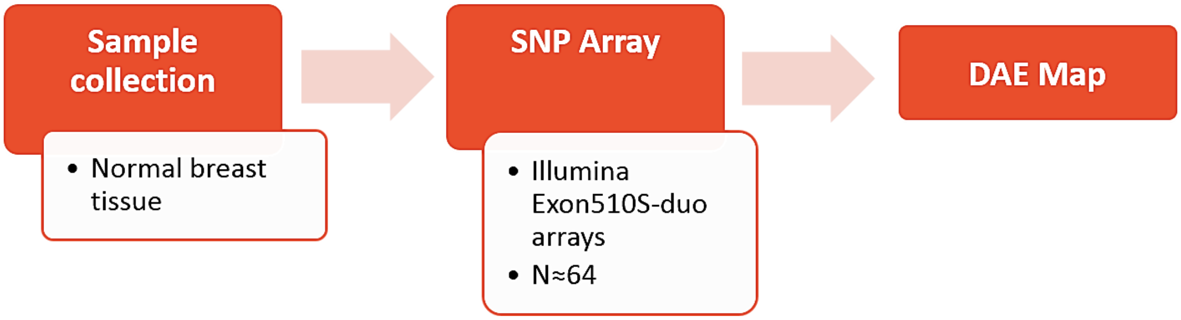


Figure 1.6 Schematic summary of the experimental design. This scheme shows the samples and arrays used to genotype and to evaluate allelic expression.

Maia and colleagues found that 7000 of approximately 33000 informative (21%) SNPs showed DAE. These mapped to about 4000 (26%) genes. Among the SNPs that showing DAE, approximately 6000 had a bidirectional distribution and 700 had a unidirectional distribution. They also identified five SNPs in two genes that showed allele-specific (mono-allelic expression) (**Table 1.2**).

Table 1.2 Total number of SNPs across the genome identified by Maia and colleagues.

	SNPs	Genes
Total number of SNPs	511350	
Informative SNPs	33816	16507
DAE	7011 (21%)	4258 (26%)
Bidirectional DAE	6268	3846
Unidirectional DAE	743	681
Allele-specific	5	2

The different distribution patterns of the DAE SNPs identified by Maia and colleagues was consistent with the scenario 1, 2 and 3 presented by Xiao and Scott (Xiao & Scott 2011). These different scenarios result from different levels of linkage disequilibrium (LD) (based in r^2 and D' measures) between the cSNP and rSNP.

LD is the non-random association of alleles at two or more loci reflecting haplotypes. The LD between two markers can be measured, being the most frequent measures the D' and r^2 that compare the observed frequencies of haplotypes with the frequencies expected (VanLiere & Rosenberg 2008). (Jorde 2000; Reich et al. 2001)

In Scenario 1, the rSNP is in complete LD ($r^2 \approx 1$ and $D' \approx 1$) with the cSNP and all heterozygotes for the cSNP show DAE. This happens because there is no recombination between the cSNP and the rSNP, and therefore all heterozygous individuals for the cSNP will also be heterozygous for the rSNP. Also only two haplotypes exist in the population (rc and RC, for example) (**Figure 1.7 A**). In Scenario 2, there is strong, but not complete, LD between the cSNP and the rSNP ($r^2<1$ and $D' \approx 1$). Therefore, the heterozygous individuals for the cSNP may be heterozygous or homozygous for one of the rSNP alleles, and there are three possible haplotypes in the population (for example, rc, RC, cR). So the distribution of DAE ratios seems unidirectional (from 0 upwards, or downwards) (**Figure 1.7 B**). In Scenario 3, $r^2<1$ and $D'<1$ the rSNP and cSNP are in linkage equilibrium, therefore four haplotypes exist in the population and we can observed a DAE distribution centred around 0: samples with one allele preferentially expressed, and others with the other allele being preferentially expressed. (**Figure 1.7 C**) (Xiao & Scott 2011).

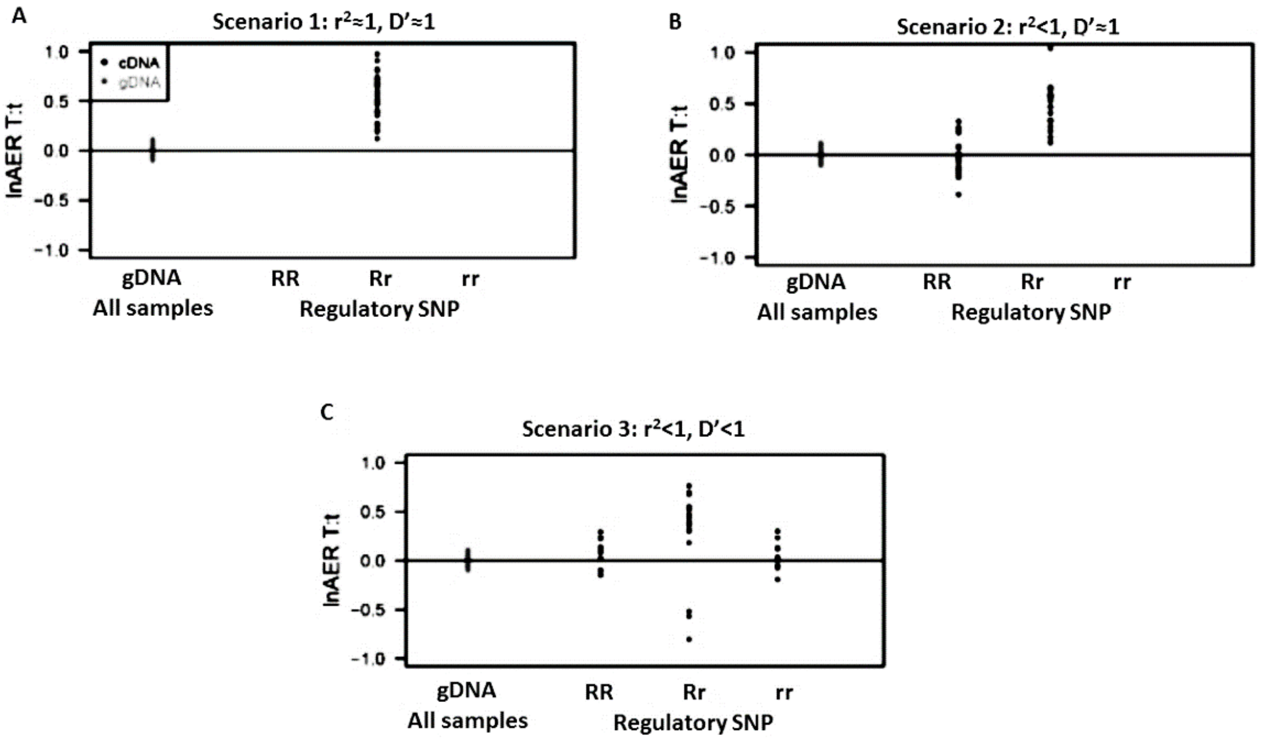


Figure 1.7 DAE scenarios for the different LD structures between the cSNP and rSNP. All samples are heterozygotes for the cSNP. SNP genotype for the samples is displayed on the x-axis. Percentage of allele of the transcribed SNP is displayed on the y-axis. Scenario 1, scenario 2 and scenario 3 represent the LD measures.

CHAPTER 2 - AIMS

The DAE map described before can be a useful tool in the search for cis-regulatory variants that affect genes associated with breast cancer susceptibility. It is also possible to cross the data in this map with the list of candidate SNPs obtained by GWAS in order to prioritise loci for functional validation.

Therefore, in my thesis project, I set up to:

1) Validate DAE polymorphisms found to be cis-regulators in a previous study using microarrays (unpublished data).

2) Identify new susceptibility loci for breast cancer by combining results of DAE mapping with results of breast cancer GWAS, and performing functional analysis.

CHAPTER 3 - MATERIALS AND METHODS

3.1 Samples studied

In this work we studied a total of 290 samples. Eighteen four samples were of normal breast tissue, were extracted from women submitted to reduction mastectomy, for reasons not related to cancer. Normal breast tissue was collected at Addenbrooke's Hospital, Cambridge, United Kingdom. A total of 150 samples were of Human B cells (blood) extracted from anonymous blood donors and 56 samples were extracted from cancer patients B cells (blood). These samples were collected in the Blood Centre at Addenbrooke's Hospital.

Blood and normal breast tissue samples were collected with approval from the Addenbrooke's Hospital Local Research Ethics Committee (REC reference 04/Q0108/21 and 06/Q0108/221, respectively).

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

3.2 Reverse transcription Polymerase Chain Reaction (RT-PCR)

For cDNA synthesis we used the SuperScript^{TM} III First-Strand Synthesis System for RT-PCR. In a first step this system synthesizes the first-strand cDNA from purified poly (A) and/or total RNA – selected RNA primed with oligo(dT), random primers. SuperScript^{TM} III Reverse Transcriptase include a RT Enzyme Mix. This enzyme is a version of M-MLV RT and is used to synthesize cDNA at a temperature between 42–60°C. This enzyme has advantages compared to other reverse transcriptase, because provides increased specificity, higher yields of cDNA, and more full-length products (http://www.lifetechnologies.com).

3.2.1 Preparation of samples by RT-PCR

We prepared cDNA from RNA extracted from 18 samples of normal breast tissue and 10 samples of blood B cells. cDNA was prepared from 0,5µg of total RNA per 20µl reaction using Reverse Transcriptase kit (SuperScript^{TM} III First-Strand Synthesis SuperMix for qRT-PCR, Invitrogen), according to the manufacturer's instructions. For run the RT-PCR reaction was used the BioRad C100 Touch™ Thermal Cycler. At the end, the reaction was diluted in a final volume of 100µl.

3.3 Real Time quantitative reverse transcription PCR (qRT-PCR)

Real time quantitative PCR allows detection and measurement of amplification products generated during each cycle of PCR process. Amplification is monitored by the detection and quantification of a fluorescent reporter signal. This amount of fluorescence is registered at each cycle and is measure during the exponential phase of the reaction because in this phase it is assumed that intensity of the signal increases in direct proportion to the amount of PCR product in the reaction (Green & Sambrook 2012). In the exponential phase it is also calculated the threshold line and the cycle threshold (Ct). Threshold line is the level of detection at which a reaction reaches a fluorescent intensity above the calculated baseline. Ct is the point in which occurs the intersection between the amplification curve and the threshold line (Heid et al. 1996; Green & Sambrook 2012) and is a relative measure of the concentration of the target product in the PCR reaction (**Figure 3.1**).

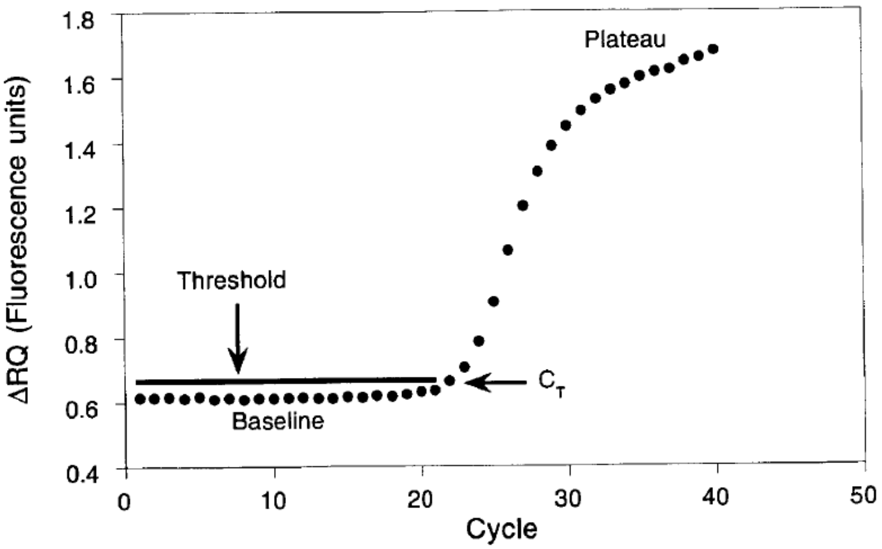


Figure 3.1 Real-time PCR detection of product in exponential phase. The standard deviation is determined from the data points collected from the base line of the amplification plot. Ct values are calculated by determining the point at which the fluorescence exceeds a threshold limit (usually 10 times the standard deviation of the base line) (Heid et al. 1996).

There are several types of detection chemistries utilized in qPCR: DNA-binding dyes (most common fluorophore used is SYBR Green), quenched dye primers (example: Amplifuor and LUX fluorogenic primers) and probe-based chemistries (most common used is TaqMan). In this study we used the TaqMan probes that contain a fluorophore reporter (in the 5'end of the probe) and a quencher (in 3'end of the probe) dye. While the probe remains intact, the reporter and quencher dyes are close. This proximity reduces the fluorescence emitted of the reporter dye through fluorescence resonance energy transfer (FRET). During the PCR, the probe anneals and when the Taq polymerase extends the primers, their 5' exonuclease activity degrade the probe separating the reporter and quencher. This process interrupts the FRET allowing the emission of fluorescence (Green, M. R. & Sambrook, 2012, M. J. Espy et al, 2006).

In this study we aim to quantify differential allelic expression and to genotype and therefore it was necessary detect the signal for each allele. For this, each qPCR reaction contains a primer pair targeting the region surrounding the marker SNP, two probes that differ by a single nucleotide and are complementary to each of the SNP alleles. The probes are labelled with different fluorochromes (HEX and FAM), designed to anneal specifically to either of the alleles of each SNP, generating two signals for each sample during the real-time PCR samples (**Figure 3.2**).

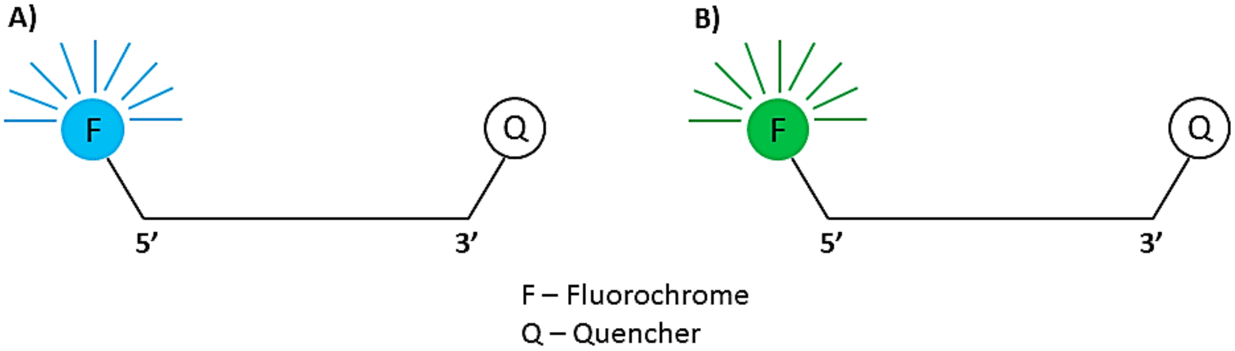


Figure 3.2 Representation of probes labelled with different fluorochromes. **A)** Probe labelled with HEX generating signal for one allele. **B)** Probe labelled with FAM generating signal for the other allele.

3.3.1 qRT-PCR to genotype all samples

Taqman® Genotyping assays were designed for 16 SNPs (rs7600326; rs10016; rs9250; rs10503416; rs13265801; rs10513376; rs1384; rs7138557; rs2526935; rs6494466; rs17817901; rs8097892; rs10521; rs710945; rs1477017; rs2834653) using the following criteria: SNP located within 100bp intronic flanking region and non-presence of other SNPs in amplification region.

For the amplification process was used approximately 5ng of genomic DNA in 5µl PCR reaction constituted by master mix (Kapa Probe Fast universal qPCR Kit (2x), Applied Biosystems™), assay (TaqMan® SNP Genotyping Assays (40x), Applied Biosystems™) and H_2O (DNase/Rnase free, gibco®by life technologies). To ensure a good quality of genotyping were included triplicates and no template controls (NTC).

The reaction mixture was initially incubated for 3 minutes at 95ºC (sufficient time for enzyme activation) and then submitted to 40 cycles including in each one denaturing (3 seconds at 95ºC) and annealing/extension (20 seconds at 60ºC) steps. The real time PCR was performed using the BioRad CFX384 Real-time System C100 Touch™ Thermal Cycler. The Ct values for each allele and for each sample were exported by Bio-RAD CFX Manager Software in the end of reaction.

3.3.2 qRT-PCR to quantification of differential allelic expression

Allele specific levels of gene expression were determined in normal breast tissue heterozygous cDNA samples (n=18), peripheral blood B cells heterozygous cDNA samples (n=12) and in blood heterozygous cDNA samples from cancer patients (n=12) using Taqman® technology in a 384-well plate.

Triplicate samples and NTCs were included in the analysis to ensure a good quality of reaction. To perform normalisation of DAE values both DNA and cDNA matching samples were included in the assays. Additionally for the case-control association study a standard curve was included in the assays, generated using a serial dilution (1, 1:2, 1:10, 1:20, 1:100, 1:1000 and 1:10000). This samples was from DNA and heterozygous for the quantified SNP, serving as a reference for the 50:50 allelic ratio.

All experiments contained replicates for each sample, and were repeated two times in different days. Real time PCR conditions were the same as used for genotyping and run on a BioRad CFX384 Real-time System C100 Touch™ Thermal Cycler. Ct values were obtained from Bio-RAD-CFX Manager Software by the same way used for genotyping. A gene was considered expressed if the PCR yielded Ct values lower than 40 cycles.

3.4 Statistical analysis

Data obtained by Real-time PCR were analysed on Microsoft® Excel® 2013 software. For each sample (DNA and cDNA) were calculated the mean of Ct values for HEX and FAM. The percentage of variation between replicates was calculated by dividing the standard deviation by the mean of the triplicates for each sample (%var=[SD/Mean]).

DAE values was determined by calculating the ratio of expression of one allele versus the other (as the Log_2 of the allelic-expression ratio) by following formula:

$$\boxed{\begin{array}{c}\text{DAE}=\text{Log}_2\,[(\text{Allele } 2^{\text{Hex}})/(\text{Allele } 1^{\text{FAM}})]\\ \text{in gDNA and cDNA}\end{array}}$$

The quantity of each allelic transcript was calculated from the Ct values using a threshold of 40. The normalised DAE values were calculated using the DAE values obtained for DNA and cDNA samples of the same individual by applying the formula:

$$\boxed{\text{Normalized DAE} = \text{DAE}_{\text{cDNA}} - \text{DAE}_{\text{gDNA}}}$$

The samples for which it was not possible to calculate the normalised DAE, were excluded from the analysis. We considered that there was DAE when the absolute normalised DAE values were above 0.58. One sample *t* S*tudent test* (t-test) was applied to compare the differences between the mean of the DAE distributions. *Variance-test* (*var*-test) was also to compare the differences between the variance of the DAE distributions. For both tests was used a confidence level of 0.05.

All the statistical analysis and graphics were performed using the R free software ((R Core Team 2014) (URL: http://www.R-project.org/)).

3.5 Selection of SNPs for DAE study validation

Data from different published GWA studies (Easton & Eeles 2008; Ahmed et al. 2009; Thomas et al. 2010; Kim et al. 2012; Michailidou et al. 2013; Turnbull et al. 2013; Couch et al. 2013; Low et al. 2013; Chen et al. 2013; Couch et al. 2014) was crossed with the DAE map and further SNPs were selected according to the following criteria.

1) Alignment of GWAS SNPs and the DAE map SNPs according to chromosome location;

2) Definition of clusters of loci with at least one GWAS associated SNP and one DAE SNP located in a window of ±250Kb

3) Analysis of LD patterns between the DAE SNP and the GWAS SNP in each cluster. Selection of clusters with at least one GWAS and one DAE SNP in LD ($r^2 \geq 0.8$).

3.6 Expression quantitative trait loci (eQTL) analysis

We performed eQTL analysis to test the associations between total level of expression of the genes (Illumina HT12 data) and genotype of a SNP (Affymetrix® SNP 6 data) using available data from the microarray experiments performed previously by my supervisor.

Two statistical tests were applied: analysis of variance (ANOVA) and Kruskal-Wallis test to assess if there were significant statistical differences ($p\text{-value}<0.05$) between the expression levels across the genotype groups.

All graphics and statistical analysis were performed using the R free software ((R Core Team 2014) (URL: http://www.R-project.org/)).

3.7 Linkage disequilibrium (LD) and haplotype analysis

SNP Annotation and Proxy Search (SNAP) web portal (Johnson et al. 2008, https://www.broadinstitute.org/mpg/snap/ldsearchpw.php) were used to test Pair-wise LD analysis and identify proxy (nearby SNPs in LD with a candidate SNP that can represent the signal from the candidate.) SNPs. These are pre-calculated based on phased genotype data using the Caucasian (CEU) population from the International HapMap Project (v3) and 1000 Genomes Pilot 1 projects. Proxy SNPs search was performed based on LD measurements ($r^2 \geq 0.8$) and distance limit of ±250kb.

HapMap is a catalogue of common human genetic variants. It contains information on location and population distribution of these genetic variants. HapMap release #27, phaseII+III, Feb09, on NCBI B36 assembly, dbSNP 126 (http://www.hapmap.ncbi.nlm.nih.gov/) was accessed to download the genotyping data of the area of interest around the *TOM1L1*/*COX11/STXBP4* (chromosome: 17; start: 50323kb and end: 50608kb) and in *AACS* locus (chromosome: 12, start: 12411kb and end: 12419kb). Haploview software (Barrett et al. 2005), http://www.broadinstitute.org/haploview) was used to analyse the pair-wise LD and haplotype structure (genotyping data from HapMap).

3.8 Analysis of regulatory potential

All SNPs and their proxies were examined for potential regulatory functions using UCSC Genome browser (http://genome.ucsc.edu/) and RegulomeDB free software ((Boyle et al. 2012)(URL: http://Regulomedb.org)).

RegulomeDB is a database with functional annotation of SNPs. The data sources are GEO (Gene Expression Omnibus), ENCODE Project Consortium (2012), NCBI Sequence Read Archive, and other resources (published literature). It is free and publicly accessible. In this database the non-coding query variants are classified into one of four categories with scores ranging from 1 to 6 (**Annex 1.1**). These scores are assigned according with the functional evidence of regulatory potential and the lower numbers represent the greatest evidence. These provide different types of information such as eQTL; chromatin immunoprecipitation sequencing (ChIP-seq); DNaseI hypersensitive sites (DHSs), chromatin interaction and TF-binding motifs. (Boyle et al. 2012)

UCSC Genome Browser, besides sequence data, contains several functional data such as: histone modifications, DNase hypersensitivity sites and transcription factors binding sites. The functional information from UCSC was complemented with additional data from collaborators (H3K4me1 ChIP-seq data from Dr. Jason Carroll, University of Cambridge) and freely available DNase hypersensitivity data (Dr Myles Brown laboratory in Harvard) (He et al. 2012). Histone modifications regulate chromatin structure and function by processes such methylation and acetylation. These types of modifications are markers of regulatory elements such as promoters and enhancers. DHSs reflect areas of chromatin open indicating the presence of active chromatin and have also been associated with regulatory elements such as: promoters, enhancers, silencers, insulators (Bannister & Kouzarides 2011; Hon et al. 2009; Wang et al. 2012). We used the version Human Feb. 2009 (GRCh/hg19) and version Human Mar. 2006 (NCBI36/hg18) Assembly from UCSC Genome Browser (that contain different and complementary features) to analyse the region covering our selected SNPs.

3.9 Transcription factor prediction

For the prediction of transcription factor binding sites (TFBS) we analysed 30 bps of DNA sequence around each SNPs for both alleles (**Annex 1.2**) using TRANSFAC® software (collaboration with Dr. Shamith Samarajiwa, University of Cambridge).

TRANSFAC is a professional database that analyses the binding affinity of a TF based in the core and matrix scores. These scores correspond to the quality of a match between the DNA sequence and the TF binding matrix and the core sequence of a matrix, respectively (Wingender et al. 2000).

Transcription factors were selected according to the following criteria: TF binding site overlap to the SNP location, difference between alleles in terms of presence/absence of binding and cut-off of ≥0.9 for both matrix and core scores.

3.10 Electrophoretic mobility shift assay (EMSA)

3.10.1 Description of the assay

This approach allows to detect protein-nucleic acid interactions. The method is based in the electrophoretic mobility of the protein-nucleic acid complex, through a non-denaturing polyacrylamide gel (Hellman & Fried 2007; Holden & Tacon 2011).

A double-strand oligonucleotide containing a putative or known binding sequence is labelled with a radioactive or fluorescent marker (Biotin) and then is added to nuclear extract allowing the formation of a DNA-protein complex. If protein binds to the labelled sequence, the DNA-protein complex migrates slowlery through the gel creating a shift compared to the unbound oligonucleotide. In the reaction can also be added a non-labelled sequence (competitive) which will test the binding specificity of a protein to the target sequence. In a last step, antibodies that recognize epitopes of the protein are added and generate a lower mobility of the DNA-protein complex resulting in a supershift on the gel (**Figure 3.3**) (Hellman & Fried 2007; Chorley et al. 2008; Holden & Tacon 2011).

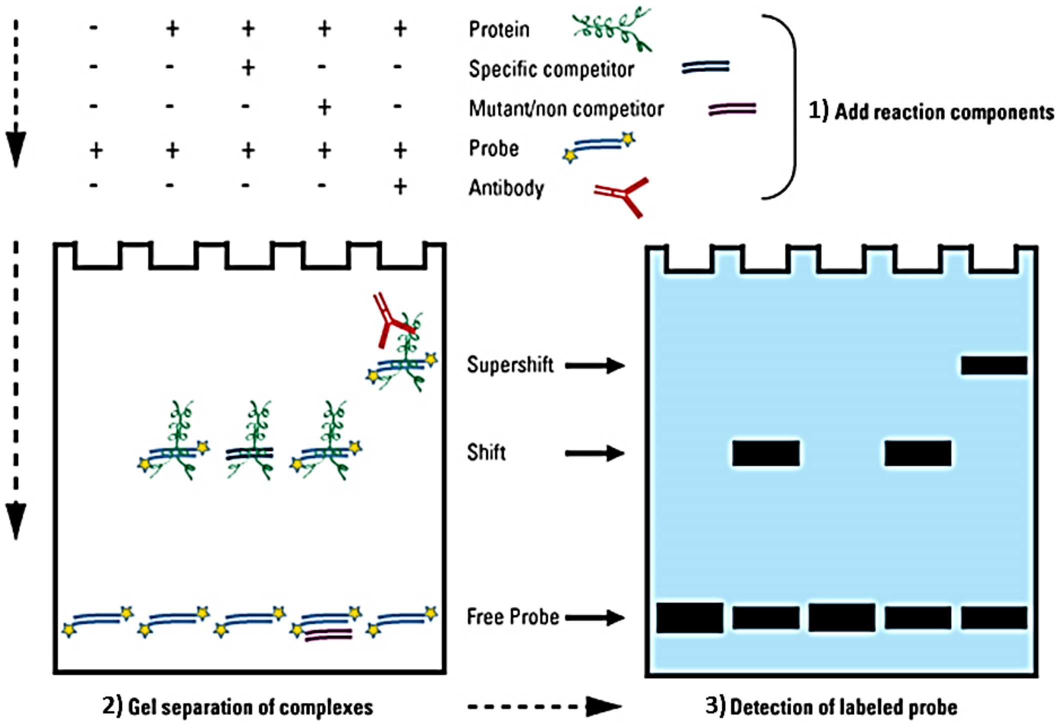

- + + + + Protein
- - + - - Specific competitor
- - - + - Mutant/non competitor
+ + + + + Probe
- - - - + Antibody
1) Add reaction components
Supershift
Shift
Free Probe
2) Gel separation of complexes
3) Detection of labeled probe

Figure 3.3 Illustration to EMSA method. The gel shift assay consists of three key steps: **1)** binding reactions; **2)** electrophoresis; **3)** probe detection (Thermo Scientific, URL: http://www.piercenet.com/method/gel-shift-assays-emsa).

3.10.2 Labelling Oligonucleotides

Oligonucleotides with 15bp of DNA sequence region surrounding the SNPs (rs7307700; rs12581512 and rs7133614) were designed (**Annex 1.3**) and two oligonucleotides previous reported were used as controls (FGFR2 and HMGI(Y)).

The first step was to label the complementary oligonucleotides separately that were then annealed. Labelling process and labelling efficiency (dot blot using hand spotting) was performed following the manufacturer's instructions (Biotin 3' End DNA Labelling Kit, Thermo Scientific). Chemiluminescent Nucleic Acid Detection Module (Thermo Scientific) was used to detect the spotted standards and double-strand oligonucleotide.

3.10.3 DNA-protein binding

EMSA was performed with Light®Shift Chemiluminescent EMSA Kit (Thermo Scientific) following the manufacturer's instructions. A 6% polyacrylamide gel (Bio-Rad) was used to run the reactions of EMSA.

Binding reactions were performed for each allele of the SNPs (double-strand oligonucleotide) and for Biotin-EBNA control DNA following the manufacturer's instructions with exception of the adding of complete Protease Inhibitors and dithiothreitol (DTT). At the end of electrophoresis, the binding reactions were transferred onto a nylon membrane in a Trans-Blot SD Semi-Dry Transfer Cell (Bio-Rad) at 20V for 40 minutes. Crosslink and detection were performed following the manufacturer's instructions.

CHAPTER 4 – RESULTS

4.1 Selection of top SNPs in DAE map based on p-value

Based on t-test p-values, ten of the most significant unidirectional DAE SNPs (rs2526935, rs10503416, rs10513376, rs7600326, rs6494466, rs10016, rs13265801, rs9250, rs8097892 and rs1384) and four non DAE SNPs (rs2834653, rs710945, rs1477017 and rs10521) obtained in DAE map using microarrays (**Annex 2.1** and **Annex 2.2**) (Unpublished, Maia et al) were chosen to be validate by Taqman qRT-PCR. Selected SNPs are located in either coding, introns or untranslated regions from different genes and serve as marker SNPs in genotyping and assessment of DAE ratios (**Table 4.1**).

Table 4.1 List of SNPs chosen to validate DAE map. In the table are shown the SNPs located gene, function (Status), chromosome (Chrom), minor allele frequency (MAF) in the CEU HapMap population and direction of Differential allelic expression (DAE) observed in the microarray study.

SNP	Gene	Chrom	Status	Alleles	MAF	DAE
rs2526935	*DPF3*	14	3'UTR	A/C	A = 0.437	Unidirectional DAE
rs10503416	*XKR6*	8	5'UTR	C/T	C = 0.310	Unidirectional DAE
rs10513376	*DAB2IP*	9	Flanking_5'UTR	C/T	C = 0.189	Unidirectional DAE
rs7600326	*CNRIP1*	2	Flanking_3'UTR	C/T	C = 0.434	Unidirectional DAE
rs6494466	*CSNK1G1*	15	Coding	A/G	A = 0.378	Unidirectional DAE
rs10016	*LOC92196*	2	3'UTR	A/G	A = 0.478	Unidirectional DAE
rs13265801	*WDYHV1*	8	Intron	A/G	G = 0.384	Unidirectional DAE
rs9250	*SENP6*	6	Coding	A/G	G = 0.165	Unidirectional DAE
rs8097892	*MPPE1*	18	Intron	C/T	T = 0.478	Unidirectional DAE
rs1384	*LYZ*	12	3'UTR	C/T	T = 0.473	Unidirectional DAE
rs2834653	*RUNX1*	21	Intron	A/C	C = 0.493	Non DAE
rs710945	*ZNF140*	12	Coding	A/G	A = 0.484	Non DAE
rs1477017	*MMP2*	16	Intron	A/G	G = 0.365	Non DAE
rs10521	*NOTCH1*	9	Coding	A/G	G = 0.417	Non DAE

UTR – untranslated region.

4.1.1 Genotyping of normal breast tissue

This 14 SNPs were first genotyped in 84 normal breast tissue samples in order to identify informative heterozygous samples. For all SNPs we identified the presence of the three genotype groups (as shown for rs8097892 in **Figure 4.1**), as expected since all SNPs were common in the studied population (reported frequencies>10% in the HapMap CEU population) (**Table 4.1**). The repeated samples used as controls gave consistent genotypes and the negative controls did not amplify (black dots in **Figure 4.1**). All genotyping results are represented as supplementary material in **Annex 2.3**.

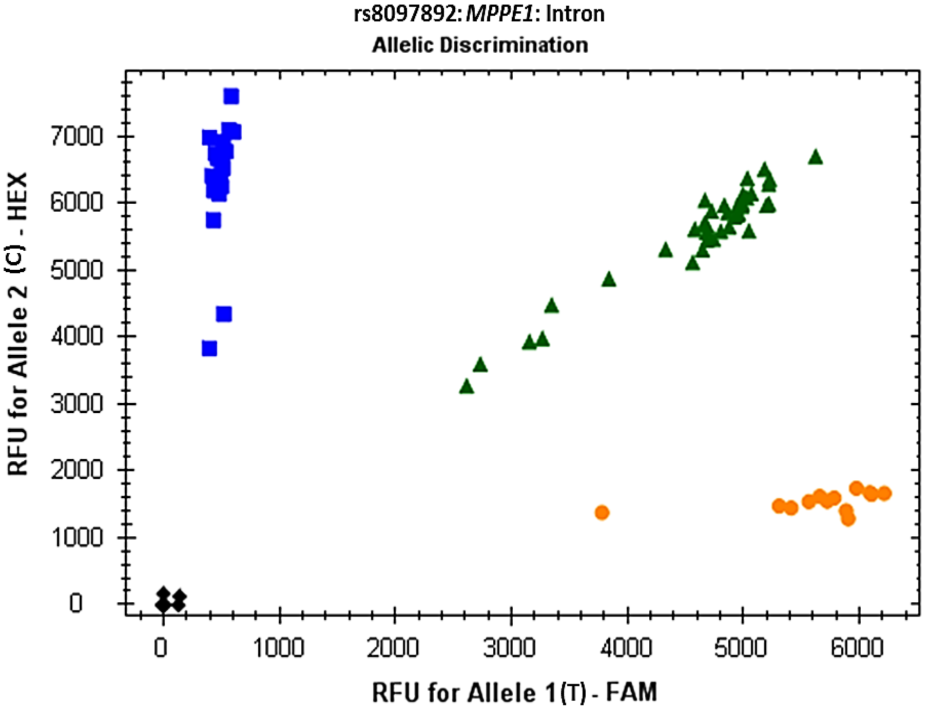


Figure 4.1 Genotyping results for rs8097892 in normal breast tissue DNA samples by qRT-PCR. The x-axis indicates the fluorescence intensity of Allele 1 emitted by probe FAM and the y axis indicates the fluorescence intensity of Allele 2 emitted by the probe HEX. The blue squares represent homozygous samples for Allele 2, orange circles represent samples homozygous for Allele 1 and the green triangles represent heterozygous samples. The black diamonds are representative of NTCs (no fluorescent signal).

The 14 SNPs showed heterozygosity values between 18-50% (**Table 4.2**) that is in agreement with the allelic frequency data described for the European population in the database dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi). Based on these results, we chose different sets of heterozygous samples for each SNP to be used in DAE quantification.

Table 4.2 Summary of genotyping results for DAE validation. In the table are shown the 14 genotyped SNPs and the percentages of undetermined genotypes, heterozigosity and allelic frequency for each SNP in the 84 normal breast tissue DNA samples genotyped.

SNP	Gene	Undetermined (%)	Heterozygosity (%)	Homozygosity	
				Allele 1 (%)	Allele 2 (%)
rs2526935	*DPF3*	13.33	48.33 (A/C)	11.67 (C)	26.67 (A)
rs10503416	*XKR6*	14.46	48.19 (C/T)	21.69 (T)	15.66 (C)
rs10513376	*DAB2IP*	25.30	31.33 (C/T)	40.96 (T)	2.41 (C)
rs7600326	*CNRIP1*	16.87	40.96 (C/T)	27.71 (T)	14.46 (C)
rs6494466	*CSNK1G1*	20.48	37.35 (A/G)	34.94 (G)	7.23 (A)
rs10016	*LOC92196*	14.46	44.58 (A/G)	4.82 (G)	36.14 (A)
rs13265801	*WDYHV1*	13.26	38.55 (A/G)	9.64 (G)	38.55 (A)
rs9250	*SENP6*	12.05	44.58 (A/G)	9.64 (G)	33.73 (A)
rs8097892	*MPPE1*	10.85	48.19 (C/T)	15.66 (T)	25.30 (C)
rs1384	*LYZ*	10.85	53.01 (C/T)	15.66 (T)	20.48 (C)
rs2834653	*RUNX1*	28.92	34.94 (A/C)	15.66 (C)	20.48 (A)
rs710945	*ZNF140*	24.10	38.55 (A/G)	28.92 (G)	8.43 (A)
rs1477017	*MMP2*	12.05	28.92 (A/G)	15.66 (G)	43.37 (A)
rs10521	*NOTCH1*	10.85	38.55 (A/G)	37.35 (G)	13.25 (A)

4.1.2 DAE analysis in normal breast tissue

Eighteen of the heterozygous matching cDNA samples were selected for quantification of allelic transcript levels in the 14 SNPs using allele-specific Taqman qRT-PCR. For all SNPs the samples replicates revealed a low percentage of variation ([mean allelic ratio/standard deviation], smaller than 1-5%). For each SNP a variable number of samples (1-17) failed quantification, in either cDNA or DNA samples, and were excluded from the analysis (**Table 4.3**).

Table 4.3 DAE average values for the validation SNPs in normal breast tissue. In the table are shown the 14 DAE SNPs, its gene location, the number of heterozygous in the successfully quantified samples, the DAE average values observed, the standard deviation (SD), the minimum and maximum values of DAE distribution (Min-Max) and the p-value for the one sample t-test.

SNP	Gene	Heterozygous with DAE	DAE average	SD	Min-Max	*t-test* (p-value)
rs2526935	*DPF3*	7/7	3.08	0.46	2.49 – 3.95	1.69E-06
rs10503416	*XKR6*	1/9	0.09	0.35	-0.33 – 0.85	2.19E-01
rs10513376	*DAB2IP*	4/8	-0.29	0.30	-1.03 – 0.25	6.10E-02
rs7600326	*CNRIP1*	6/13	-0.68	0.77	-2.19 – 0.40	4.24E-03
rs6494466	*CSNK1G1*	5/8	-0.05	1.05	-1.77 – 1.20	4.76E-01
rs10016	*LOC92196*	11/17	0.67	0.53	-0.53 – 1.45	2.08E-04
rs13265801	*WDYHV1*	4/8	-0.53	1.06	-1.91 – 1.85	2.52E-01
rs9250	*SENP6*	3/8	0.29	1.25	-0.38 – 2.78	5.91E-02
rs8097892	*MPPE1*	11/11	1.53	2.22	-4.10 – 4.54	4.22E-02
rs1384	*LYZ*	7/11	0.64	0.52	-0.22 – 1.60	2.66E-03
rs2834653	*RUNX1*	8/11	3.39	1.71	-0.31 – 4.03	5.31E-04
rs710945	*ZNF140*	8/8	-0.33	2.60	-2.62 – 3.44	9.52E-01
rs1477017	*MMP2*	3/10	-0.23	0.51	-0.97 – 0.84	2.86E-01
rs10521	*NOTCH1*	0/7	0.34	0.35	-0.18 – 0.92	5.01E-02

Concerning, the 10 SNPs that have shown DAE in previous data, we observed in all statistically significant DAE (mean DAE distribution smaller or greater than -0.58 or 0.58, respectively). Five SNPs (rs2526935, rs10503416, rs6494466, rs10016 and rs13265801) showed a consistent direction in the DAE distribution comparing with the previous results. Regarding, the four SNPs that did not show DAE in the original report, we observed significant DAE in three of these SNPs (rs2834653, rs710945, rs1477017) and no evidence of DAE in only one SNP (rs10521) (**Figure 4.2**).

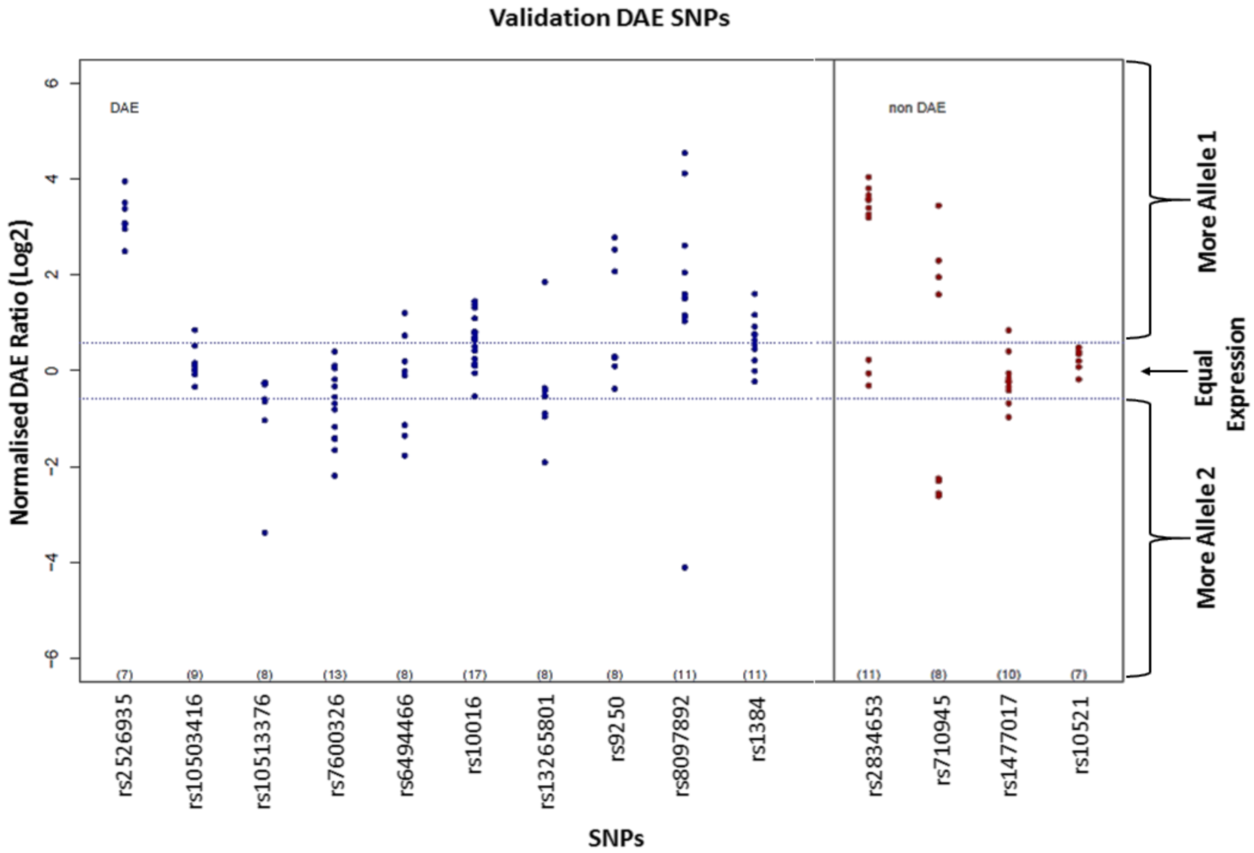


Figure 4.2 Results from the validation of the DAE map SNPs. x-axis indicates the name of SNPs and the y axis indicates the normalised DAE ratio obtained. Heterozygous individuals are represented as dots and are blue and red for the SNPs originally reported as having DAE or non DAE, respectively. The numbers in parentheses are the numbers of individuals in which the DAE was quantified in our validation. Dotted lines delimit the cut-off of preferential allelic expression ratio [$\log_2(1.5)=0.584$].

4.2 Crossing of GWAS and genomic DAE information

One of the main objectives from this thesis project is identify new susceptibility loci for breast cancer. To this end, we first collected the data on loci associated with breast cancer risk from all public available original GWAS and meta-analysis studies (Easton & Eeles 2008; Ahmed et al. 2009; Thomas et al. 2010; Kim et al. 2012; Chen et al. 2013; Couch et al. 2013; Low et al. 2013; Michailidou et al. 2013; Turnbull et al. 2013). These data were crossed with the DAE map established previously by Prof. Ana Teresa Maia, using both LD and physical distance as a criteria. Firstly, we selected GWAS SNPs which were within 250kb (on either side) of the DAE SNPs. Then, we selected from this list all DAE SNPs which were in strong LD ($r^2 \geq 0.8$) with the GWAS SNPs. The rationale behind the LD analysis was because both the DAE as the GWAS SNPs are markers for the actual cis-regulatory risk variant, with whom they should be in moderate to strong LD. Therefore, applying as a criteria to follow up genes/loci that have GWAS and DAE associated SNPs in LD we maximize our chances of identifying the true risk variant.

From this list we selected two regions to study: the 17q22 and the 12q24 loci. These regions were selected based on: DAE values and scenario consistent with the observed LD between the transcribed SNP (where the DAE was measured) and the GWAS associated SNP, and ranking of the GWAS associated variant in the original study and/or meta-analysis. After selecting these two loci, we searched for potential regulatory functions using the database RegulomeDB, Genome Browser and several histone modifications data files. Putative transcription factor binding sites (TFBS) were also searched within the sequences containing the candidate SNPs using Alibaba2 and TRANSFAC®.

4.2.1 Analysis of breast cancer risk locus 17q22 - *TOM1L1/COX11/STXBP4*

We found three SNPs with DAE in the 17q22 locus: rs7643 that is located in the 3'UTR of *COX11;* rs17817901 that is located in the intronic or 3'UTR region of *TOM1L1,* depending on the transcript and that also co-localizes to the 3'UTR region of *COX11* gene; and rs2628315 that is located in a intronic region of *STXBP4*,). DAE analysis showed that rs7643 and rs17817901 had a DAE pattern consistent with Scenario 1 (**Annex 3.1- A** and **B**, respectively), indicating that all are in complete LD ($r^2 \approx 1$, $D' \approx 1$) with the rSNP. rs2628315 shows a DAE pattern consistent with Scenario 2 ($r^2 < 1$, $D' \approx 1$) (**Annex 3.1 - C**).

In GWAS studies, rs6504950, an intronic variant in the *STXBP4* gene, has been consistently associated with breast cancer risk, having been ranked second (per-allele OR = 0.95, 95% CI = 0.92–0.97, P = 1.4E-08) in the report by Ahmed et al (Easton et al. 2007). rs6504950 has also been reported as being associated with breast cancer risk in several studies with the minor allele always being associated with decreased risk (meta-analysis OR = 0.92, 95% CI = 0.88–0.96, P = 1E-04) (Ahmed et al. 2009; Antoniou et al. 2011; Bhatti et al. 2010; Campa et al. 2011; Loizidou et al. 2011; Tang et al. 2012).

LD analysis showed that one DAE SNPs (rs17817901) is in strong, not complete, LD (r^2=0.830 and D'=1; distance=17.7 kb) with the GWAS associated SNP, suggesting that this DAE SNP and the GWAS associated SNP are stronger markers for the true disease risk variant(s).

4.2.1.1 Characterisation of regulatory landscape

Proxy SNPs (r^2>0.8 and D'=1) for rs17817901 and rs6504950 were searched and a total of 138 unique SNPs were identified (data not shown). Both rs1781901 and rs6504950 and their proxies were analysed for functional evidence on RegulomeDB. Six SNPs were chosen for further analysis (**Table 4.4**), there was eQTL evidence for *COX11* in monocytes (rs7222197, rs2628305, rs2787481, rs244317, rs12949538 and rs9902718), having one of these six SNPs also functional evidence in breast cancer cell lines (rs7222197). rs7222197, a proxy for the GWAS SNP rs6504950 (r^2=1 and D'=1), had ChIP-seq evidence for GATA3 binding in T47D cells (luminal ER^+ breast cancer cell line) and evidence of higher chromatin structure from FAIRE experiments in MCF-7 cells (another luminal ER^+ breast cancer cell lines). The remaining six did not have any data available in RegulomeDB were selected for further analysis: rs17817901 and rs7643 because they show DAE, rs6504950 because it was associated in several GWAS and rs12951898, rs12165058, rs3087650 because not having information available does not exclude the possibility of them still having a functional role.

Table 4.4 RegulomeDB scores for the selected SNPs in 17q22 locus.

TOM1L1/COX11/STXBP4	
Scores	**SNPs**
1f	rs7222197
	rs2628305
	rs2787481
	rs244317
6	rs12949538
	rs9902718
No data	rs12951898
	rs12165058
	rs3087650
	rs17817901
	rs7643
	rs6504950

Subsequently, this locus was analysed using the UCSC Genome Browser. Six SNPs (rs17817901, rs12949538, rs7222197, rs9902718, rs2628305 and rs6504950) were found to overlap regions containing possible regulatory elements (**Figure 4.3**). rs17817901 overlaps a region containing H3K4me1 (observed in BT474 cell lines, a ER$^+$ human breast cell line), a marker for enhancers and promoters. rs12949538 and rs6504950 were also found to overlap with a region containing both H3K4me1 and DNase hypersensitivity clusters (a marker for both promoters and enhancers, and protein binding). rs9902718 and rs2628305 lie in a region with evidence of H3K4me1 mark, but not in a breast cancer cell line. The rs7222197 overlaps a region containing the marks H3K4me1, H3K4me3 (marker for promoter) and H3K27Ac (marker marks active regulatory elements) and DNase hypersensitivity clusters, indicating a possibility of an active promoter (**Figure 4.3**). This data suggests that these six SNPs can possibly exert a functional effect by modulating the gene expression regulation.

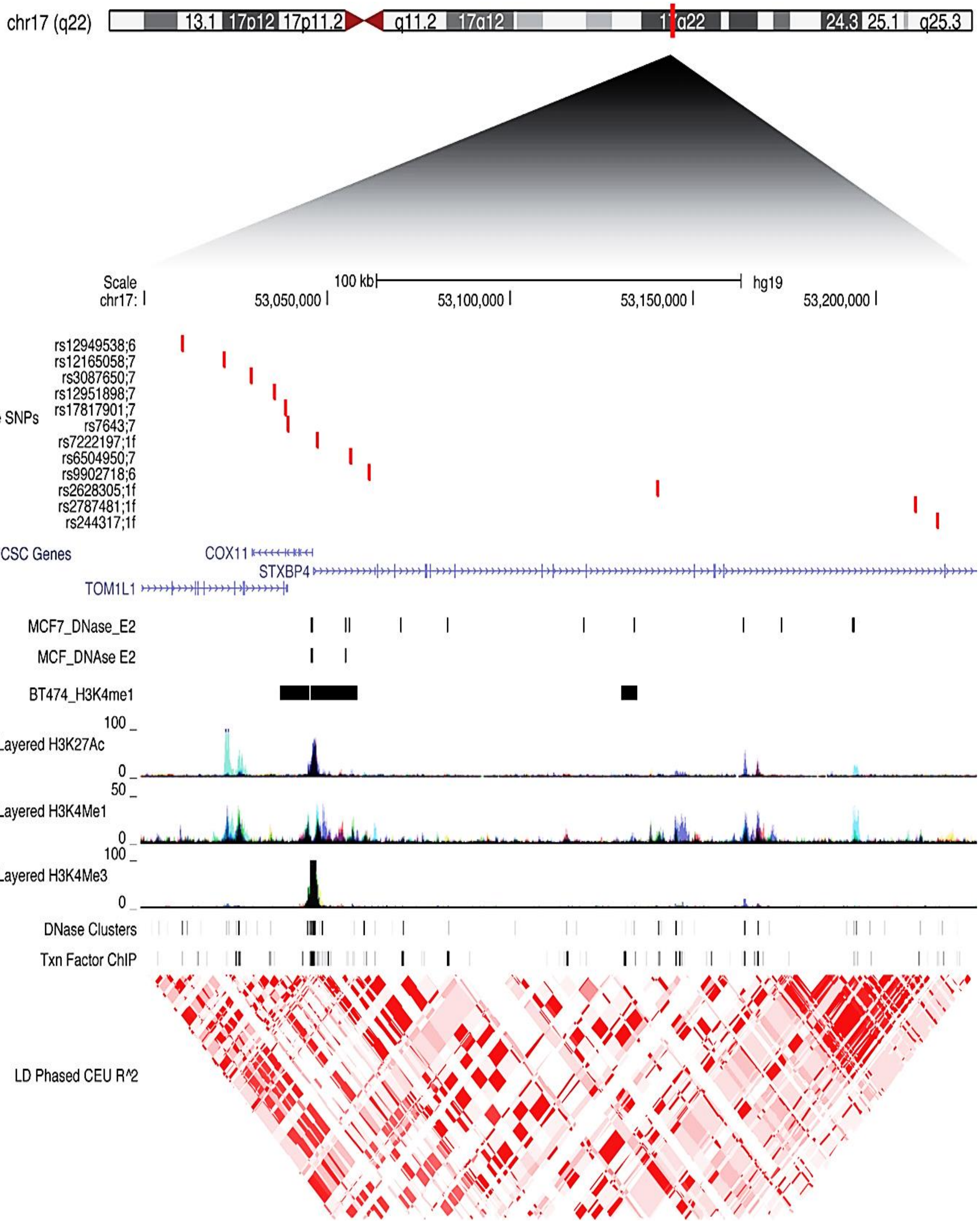


Figure 4.3 Genomic view of the 17q22 locus with functional regulatory evidence. From the top to the bottom of the figure are shown the candidate proxy SNPs, DAE and GWAS SNPs, the RefSeq genes mapped to the area of interest around the locus (TOM1L1, COX11 and STXBP4), information about additional data from collaborators (Dr. Jason Carroll, University of Cambridge and Dr Myles Brown (He et al. 2012), histone modifications, DNase clusters, transcription factors and LD structure according to the Genome Browser (http://genome.ucsc.edu/).

4.2.1.2 LD and haplotype analysis

Our data suggested that the cis-regulatory variation responsible for DAE in 12q24 locus can be located in the same haplotype block as rs17817901, rs7643 and rs2628315. Therefore, we further analysed LD in this region to define the haplotype structure in order to identify the haplotype block containing the risk cis-regulatory variants. In this exercise we aimed to identify the haplotypes responsible for decreased or increased gene expression. rs17817901 and rs6504950 had no genotype information available in HapMap, therefore we analysed two proxies in complete LD (r^2=1, D'=1) with each one of these SNPs. rs17817901 and rs6504950 in the haplotype analysis are thus represented by rs12936860 and rs2628315, respectively (**Figure 4.4**).

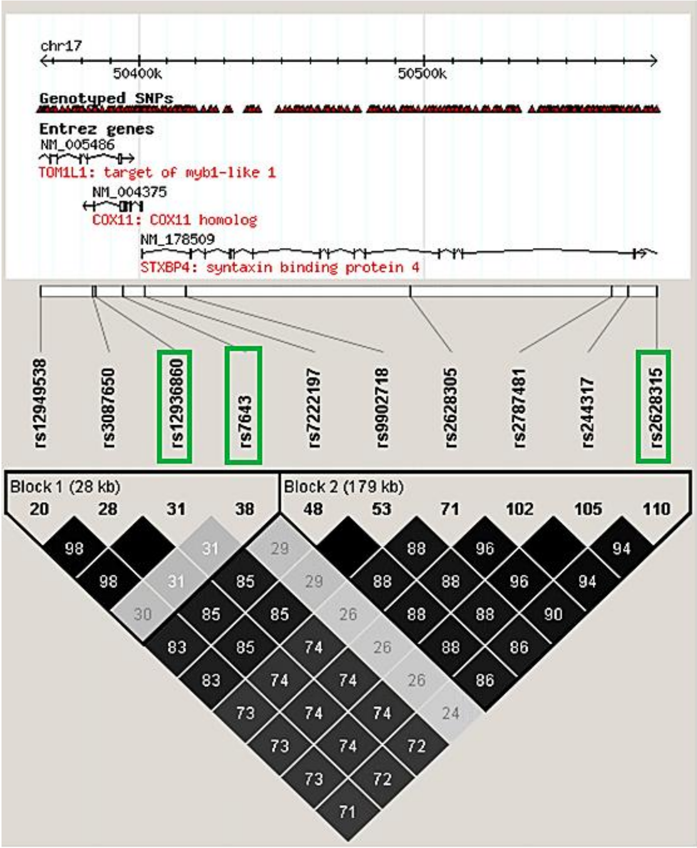


Figure 4.4 Linkage disequilibrium plot for the analysed SNPs in 17q22 locus in 30 CEU (CEPH population of Utah residents with ancestry from northern and western Europe). The SNP ID is displayed along the top of the diagram. This plot was obtained in Haploview using the r^2 colour scheme (black indicating r^2 =1, with different shades of grey indicating $0 < r^2 < 1$). In addition, values in the plot indicate r^2 values for pairwise comparisons between the SNPs. Blocks were defined using the confidence interval method. Black triangles denote the two haplotype blocks. The two SNPs used as marker SNPs for the rSNP in this analysis are shown in green. (http://www.broadinstitute.org/haploview).

Two main haplotype blocks were identified: Block 1 included three haplotypes and block 2 included two haplotypes in the European population. In block 1, the minor allele of rs12936860 (proxy for rs17817901) belongs to haplotype 2 and the minor allele of rs7643 belongs to haplotypes 2 and 3. In block 2, the minor allele of rs2628315 belongs to haplotype 5 (**Figure 4.5**).

In our DAE analysis SNPs rs17817901, rs7643 and rs2628315 always showed the minor alleles preferentially expressed (A allele). Therefore, the haplotypes in our population that are more likely to be associated with preferential expression are haplotype 2 and 3 in block 1 and haplotype 5 in block 2, and both haplotype 1 and 4 are associated with decreased expression (**Figure 4.5**).

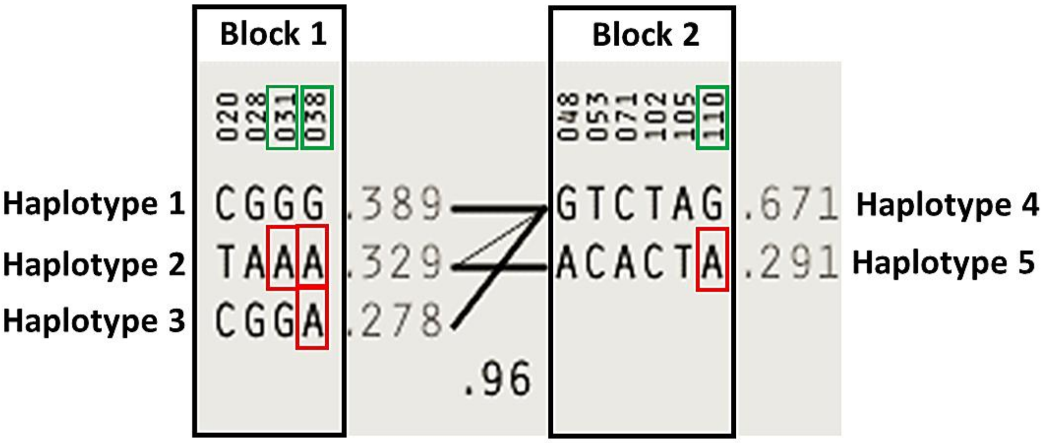


Figure 4.5 Haplotype blocks and haplotype frequencies in 17q22 locus. The haplotype frequencies are shown to the right of each haplotype. The SNP numbers in the top of the haplotypes correspond to those in the LD plot (**Figure 4.4**). The three SNPs used as marker SNPs for the rSNP in this analysis are shown in green. Haplotypes preferentially expressed are shown in red. This plot was obtained in Haploview using genotype information from 30 CEU (http://www.broadinstitute.org/haploview).

4.2.2 Analysis of breast cancer risk locus 12q24 – *AACS*

We found three SNPs with DAE (rs7138557, rs12581512 and rs2291248) located in the 12q24 locus, all located in introns of *AACS* (acetoacetyl-CoA synthetase). All three SNPs showed a DAE pattern consistent with Scenario 2 (**Annex 3.2 – A, B, C** respectively), indicating that they are in strong, not complete, LD with the rSNP (r^2<1, D'=1), as explained previously in section 1.5.

Easton et al (2007) reported an association with breast cancer risk at the same locus, tagged by rs7307700 (OR = 1.02; 95% CI = 0.99-1.05; p-value = 2E-03]. Another SNP, rs2291248, was associated by the same group in the same report but did not reach genome-wide significance and therefore was not reported (preliminary p-value = 7.49E-04, Dr D. Easton, personal communication). Interestingly, rs2291248 was identified both in GWAS and our DAE study.

LD analysis showed the DAE SNP rs7138557 is in LD with the GWAS hit rs7307700 (r^2=0.846 and D'=1, distance=0.4 Kb) and the DAE SNP rs12581512 is in LD with rs2291248, shared by DAE and GWAS (r^2=0.869 and D'=1, distance=2.9 Kb). This result suggests that these four DAE SNPs are markers for the true disease-causing variant(s).

4.2.2.1 Characterisation of regulatory landscape

We looked for proxy SNPs (r^2>0.8 and D'=1) for the four SNPs in LD, in order to identify the true cis-regulatory risk variant. A total of 123 proxy SNPs were identified (data not shown). On searching RegulomeDB we found functional evidence in breast cancer cell lines for 11 SNPs, and another two did not have any data (rs12581512 and rs7307700) (**Table 4.5**).

rs7133614, a proxy for the DAE SNP rs12581512 (r^2=0.965 and D'=1) had evidence of higher chromatin structure, from DNase-seq analysis, in T47D cells. rs7137742, a proxy for the DAE SNP rs12581512, had a binding motif evidence for two transcription factors, MAX and USF, from Footprinting experiments in MCF-7 cells. rs10846834, a proxy for the DAE SNP rs7138557 had ChIP-seq evidence for STAT3 binding in MCF10A-Er-Src cells, a normal breast epithelial ER^- cell line. rs7138790, rs35428999, rs7138557 and rs7304979 only had evidence of higher chromatin structure, from DNase-seq, in MCF-7 and T47D cells. rs7138405, had ChIP-seq evidence for MYC binding in MCF-7 cells and had evidence of histone modifications (H3k04me3 and H3k4me3) from ChIP-seq also in MCF-7 cells.

The two SNPs without functional evidence (rs12581512, rs7307700) on RegulomeDB were included in further analysis, as there was no information supporting their exclusion.

Table 4.5 RegulomeDB Scores for the selected SNPs in 12q24 locus.

AACS	
Scores	**SNPs**
1f	rs7133614
2a	rs7137742
2b	rs10846832
	rs58416336
	rs7138790
	rs35428999
3a	rs10846834
	rs7138405
	rs7304979
4	rs2291248
5	rs7138557
No data	rs12581512
	rs7307700

Subsequently, this locus was analysed using the UCSC Genome Browser. All SNPs were found to overlap regions containing binding sites for several transcription factors (**Figure 4.6**) and containing evidence of DNase hypersensitivity clustering. Some of these SNPs showed additional evidence of being located in a regulatory region: rs7138405, rs7137742, rs7138790 and rs10846834 overlap regions containing evidence of histone marks such as H3K4me1 (marking enhancers and promoters), H3K4me3 (marking promoters), and H3K27Ac (active regulatory elements); rs7138557, rs7307700, rs10846832 and rs58416336 also overlap regions containing evidence of one histone mark, such as H3K4me1; rs10846832 and rs58416336 had evidence of histone mark H3K4me1 observed in SK-BR-3 cell lines (human breast cell line that overexpresses the *Her2* gene) (**Figure 4.6**). This information suggests that all these 13 SNPs are putative cis-regulatory SNPs.

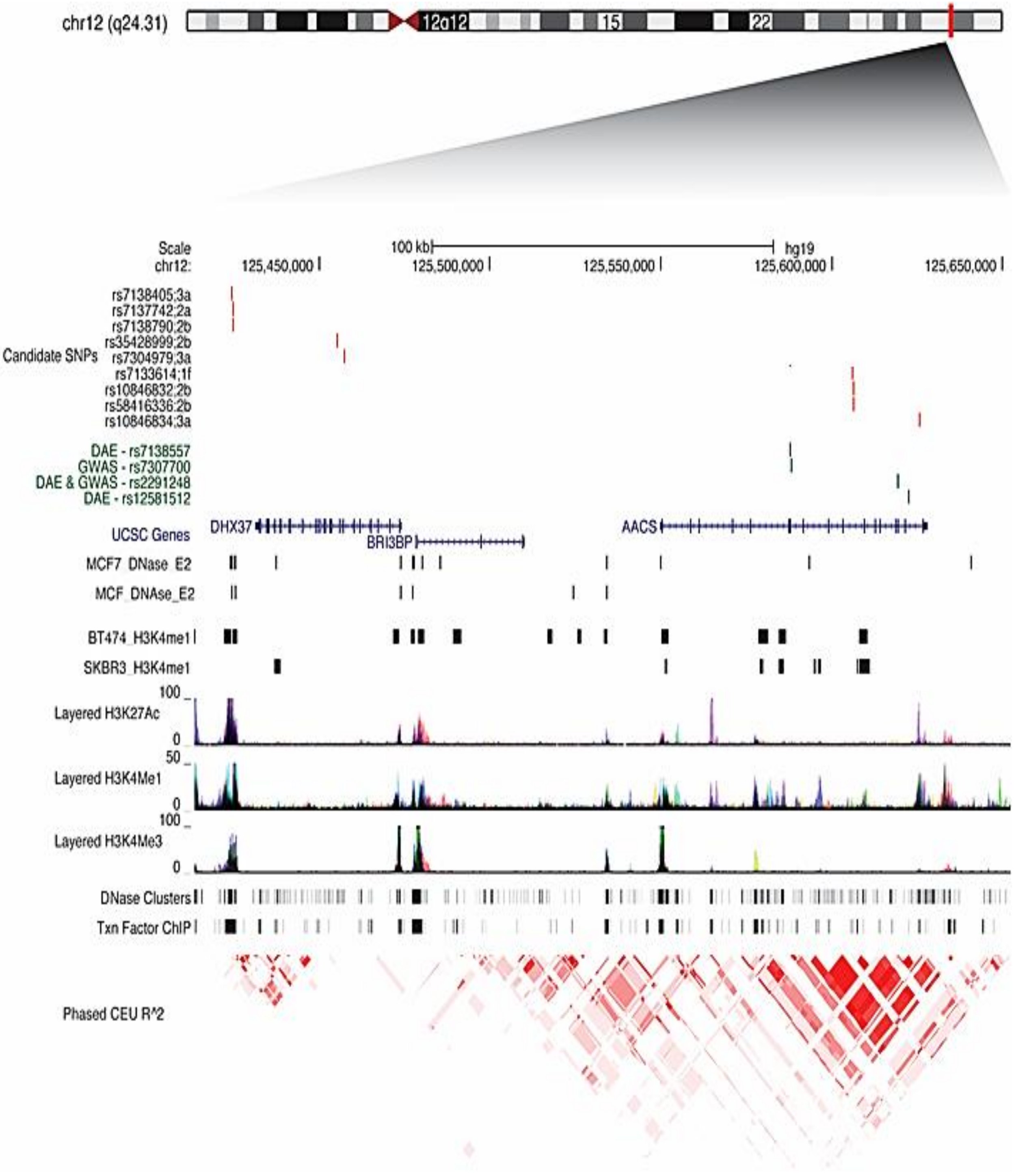


Figure 4.6 Genomic view of the 12q24 locus with functional regulatory. From the top to the bottom of the figure are shown the candidate proxy SNPs (red) and the DAE and GWAS SNPs (green), the RefSeq genes mapped to the area of interest around the locus (*AACS* gene), additional data from collaborators (Dr Jason Carroll, University of Cambridge and Dr Myles Brown (He et al. 2012), information about histone modifications, DNase clusters, transcription factors and LD structure according to the Genome Browser (http://genome.ucsc.edu/).

4.2.2.2 LD and haplotype analysis

Our data suggested that the cis-regulatory variation responsible for DAE in 12q24 locus can be located in the same haplotype block as rs7138557, rs12581512, rs7307700 and rs2291248 (**Figure 4.7**). Therefore, we looked at the LD structure in this region in order to identify the haplotypes responsible for decreased or increased gene expression.

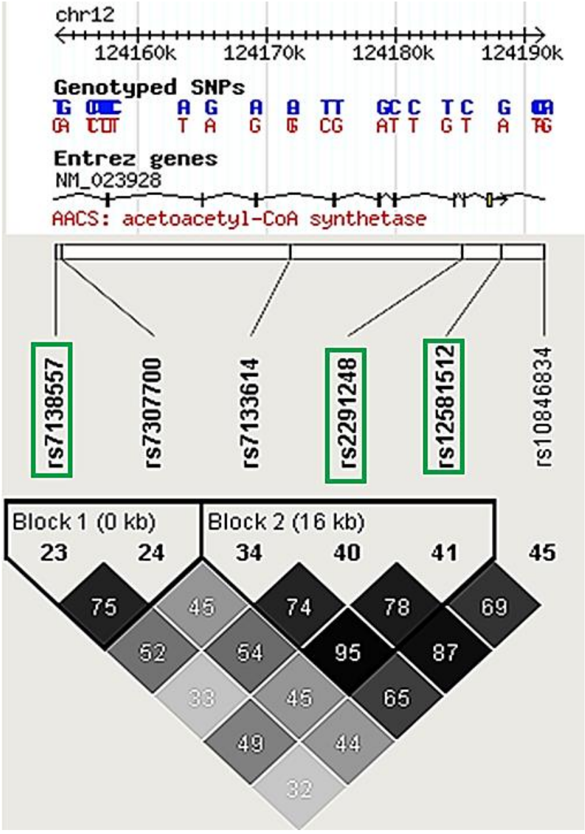


Figure 4.7 Linkage disequilibrium plot for the analysed SNPs in 12q24 locus in 30 CEU. The SNP ID is displayed along the top of the diagram. This plot was obtained in Haploview using the r2 colour scheme (black indicating $r^2 = 1$, with different shades of grey indicating $0 < r^2 < 1$). In addition, values in the plot indicate r^2 values for pairwise comparisons between the SNPs. Blocks were defined using the confidence interval method. Black triangles denote the two haplotype blocks. The two SNPs used as marker SNPs for the rSNP in this analysis are shown in green. (http://www.broadinstitute.org/haploview).

Two main haplotype blocks were identified: Block 1 included four haplotypes and Block 2 included three haplotypes in the European population (**Figure 4.8**). rs7138557 belongs to block 1 in which haplotypes 2 and 4 contain the minor allele. rs7307700 also belongs to block 1 in which haplotypes 2 and 3 contain the minor allele. rs12581512 is in block 2 and its minor allele is in haplotype 6. rs2291248 is also in block 2 and its minor allele corresponds to haplotypes 6 and 7.

Our DAE results showed that all minor alleles of the DAE SNPs were preferentially expressed. Therefore, in block 1 haplotype 2 is the most likely to be preferentially expressed, only because haplotype 4 is very rare, and could not justify the frequency of DAE we observed. In block 2, haplotypes 6 and 7 are the ones associated with preferential expression although haplotype 6 is probable contributing more to the DAE signal in our results because it is more frequent (**Figure 4.8**).

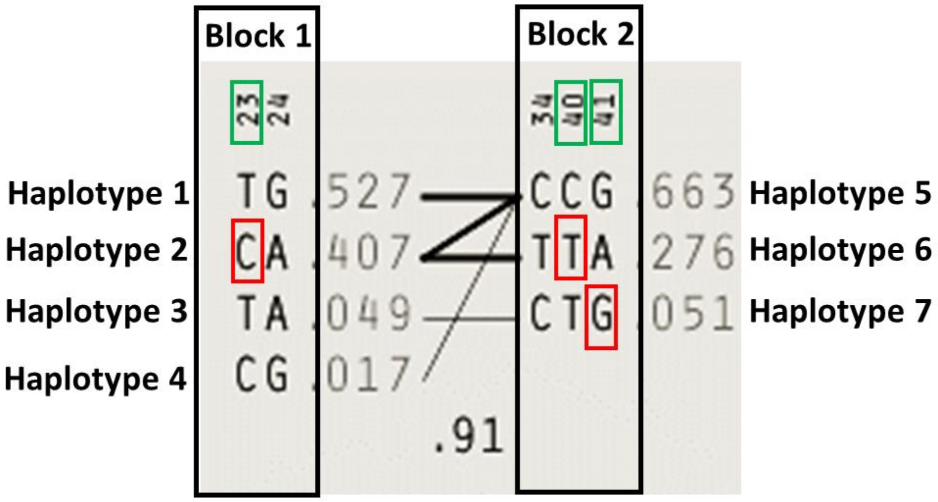


Figure 4.8 Haplotype blocks and haplotype frequencies in 12q24 locus. The haplotype frequencies are shown to the right of each haplotype. The SNP numbers in the top of the haplotypes correspond to those in the LD plot (**Figure 4.7**). The three SNPs used as marker SNPs for the rSNP in this analysis are shown in green. Haplotypes preferentially expressed are shown in red. This plot was obtained in Haploview using genotype information from 30 CEU (http://www.broadinstitute.org/haploview).

4.2.2.3 Prediction of transcription factor binding sites (TFBS)

Next our best 13 candidate SNPs (rs7133614, rs7137742, rs10846832, rs58416336, rs7138790, rs35428999, rs10846834, rs7138405, rs7304979, rs2291248, rs7138557, rs12581512 and rs7307700) were analysed to predict transcription factor binding sites (TFBS). Results for each SNP were selected based in the criteria described in section 3.9 (Materials and Methods). The results are presented in **Table 4.6**. In short, we selected the predictions with good core and matrix scores, and differences between the two alleles of each SNP, as described in Materials and Methods.

Table 4.6 TRANSFAC Results for all SNPs (continuation of the table on page 42). In the table the position of minor and common allele is represented in red. Minor allele is shown first and the common allele is shown second. Matrix identifier corresponds to the transcription factors that can binding in the sequence containing the allele.

SNP	Allele	Matrix identifier	Sequence	Core match	Matrix match
rs7133614	T	---------	---------	---------	---------
	C	V$GEN_INI3_B	gca**C**AGTC	0.996	0.962
		V$CAP_01	A**C**AGTcct	0.989	0.988
rs7307700	G	---------	---------	---------	---------
	A	V$HMGIY_Q6	GTA**A**Aag	0.964	0.906
rs12581512	A	V$HES1_Q2	taccc**A**CAAgctga	0.989	0.934
		V$CAP_01	CC**A**CAagc	0.964	0.959
	G	V$CBF_01	gatacCCC**G**Caagctg	0.982	0.95
rs7137742	A	---------	---------	---------	---------
	G	V$CMYC_02	agtcACGT**G**gtt	1	0.953
		V$STRA13_01	cagtcACGT**G**gtta	1	0.935
		V$USF_02	cagtCACGT**g**gtta	1	0.993
		V$TFE_Q6	tCACGT**g**g	0.95	0.959
		V$EBOX_Q6_01	agtcACGT**G**g	1	0.998
		V$CBF_01	agtcACGT**G**g	0.978	0.962
		V$AHRHIF_Q6	cACGT**G**gtt	0.982	0.955
		V$KID3_01	CGT**G**G	1	1
rs10846832	C	V$CBF_01	ctccctG**C**GGGgtcct	0.982	0.952
	G	V$CHCH_01	TG**G**GGg	0.986	0.986
		V$SPZ1_01	cctG**G**GGGgtcctgg	0.998	0.946
		V$LRF_Q2	cctG**G**GGGgtcctgg	1	0.983
rs58416336	T	V$CBF_01	tccctgG**T**GGGtcctg	0.96	0.947
		V$CAP_01	cctGG**T**GG	0.971	0.964
		V$KID3_01	GG**T**GG	1	1
	G	V$CHCH_01	TGG**G**Gg	0.986	0.986
		V$LRF_Q2	G**G**GGGtcct	1	0.983
		V$MOVOB_01	tgG**G**GGG	1	0.96
rs7138790	G	V$CBF_01	tccctGCC**G**Ctaatca	0.981	0.965
	C	V$CHCH_01	gCC**C**CT	0.985	0.978
rs35428999	A	V$GEN_INI3_B	cgcC**A**CTC	0.989	0.968
		V$MSX1_01	c**A**CTCActc	0.865	0.831
		V$CAP_01	CC**A**CTcac	0.984	0.982
		V$KID3_01	CC**A**CT	0.991	0.991
	G	V$CBF_01	tgcacGCC**G**Ctcactc	0.981	0.949
rs10846834	A	V$GEN_INI3_B	CACTGatt	0.984	0.944
		V$CAP_01	GCACTgat	0.987	0.98
	G	---------	---------	---------	---------

rs7138405	T	---------	---------	---------	---------
	C	V$AHRHIF_Q6	ctg**C**ACGTt	0.982	0.964
rs7304979	A	---------	---------	---------	---------
	G	V$CHCH_01	C**G**GGGc	1	0.993
		V$LRF_Q2	**G**GGGCcact	0.943	0.937
		V$AHRHIF_Q6	cgcCAC**G**Gg	0.93	0.91
rs2291248	T	V$EN1_01	G**T**GGTgt	0.926	0.906
		V$KID3_01	TG**T**GG	0.995	0.995
	C	---------	---------	---------	---------
rs7138557	C	V$USF_02	cttc**C**ACTTggggt	0.915	0.901
		V$GEN_INI3_B	ttc**C**ACTT	0.986	0.937
		V$LMO2COM_0	ttc**c**ACTTGggg	0.901	0.91
		V$EBOX_Q6_01	t**t**ccACTTGggg	0.98	0.949
		V$CBF_01	gatctTC**C**ACttgggg	0.972	0.959
		V$CAP_01	C**C**ACTtgg	0.984	0.978
	T	---------	---------	---------	---------

4.2.2.4 Identification of DNA-protein interaction

We have started in-vitro analysis of these predictions on three SNPs: rs7307700, rs12581512 and rs7133614. First, oligonucleotides containing the SNPs of interest, and control competition oligonucleotide (HMGI(Y)) for rs7307700 were labelled with Biotin.

Spot intensities of the seven double-strand oligonucleotides were compared with the Biotin-EBNA Control DNA within the kit to determine the labelling efficiency, which we found to be approximately 75% for all oligonucleotides (**Figure 4.9**). Based on this we used a concentration of 50nM for the EMSA analysis.

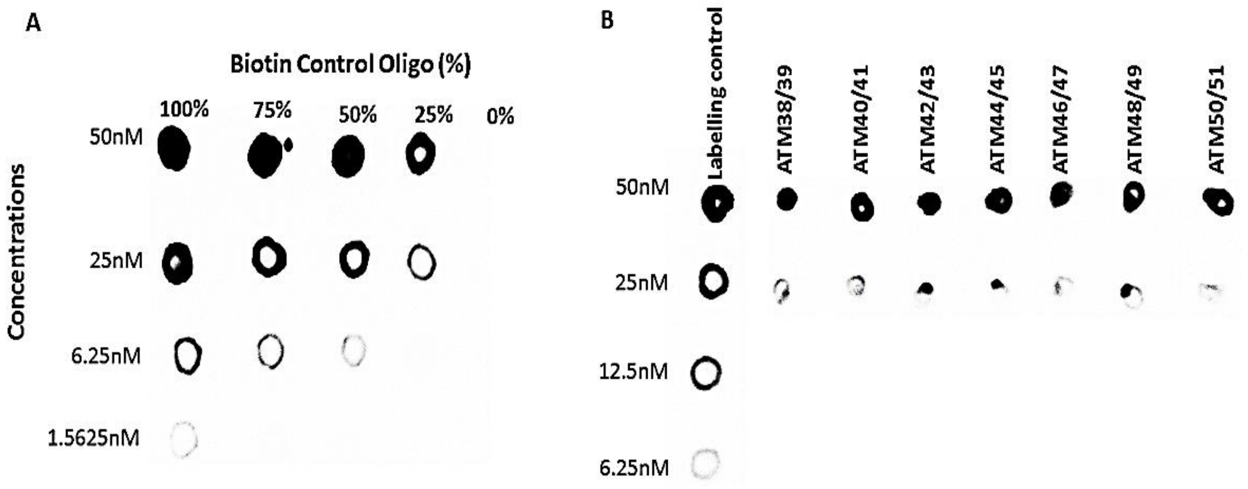


Figure 4.9 Labelling efficiency for Biotin-EBNA Control DNA and for all oligonucleotides annealed. The blot includes the dilution for each of the standards of the Procedures for Estimating Labelling Efficiency **(A)**, as well as the labelling control of the kit and seven different test oligonucleotides **(B)**, **ATM38/39** correspond to HMGI(Y), **ATM40/41** and **ATM42/43** correspond to G and A alleles (rs7307700), **ATM44/45** and **AMT46/47** correspond to A and G alleles (rs12581512) and **ATM48/49** and **ATM50/51** correspond to T and C allele (rs7133614).

EMSA was performed using nuclear protein extract previously prepared from HCC1954 breast cancer cell line and for the extraction was used the NE-PER™ Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific), following manufacturer's instructions. Biotin-EBNA Control DNA, Biotin-EBNA Control DNA+ EBNA Extract and HMGI(Y) and FGFR2 were also included as positive controls (**Figure 4.10**). HMGI(Y) (ATM38/39) and FGFR2 were included because these oligonucleotides have been reported to show binding in EMSAs using nuclear extracts from breast cell lines (Klein-Hessling et al. 1996; Reeves et al. 2001; Meyer et al. 2008).

We performed a first experiment in which we observed shift for the positive controls (**Figure 4.10**). Nevertheless, the signal was very weak and the shift was not complete (the majority of oligonucleotide was unbound). This result suggests that the protein extract were in good condition, but that the quantity of protein used was probably too low to form a strong binding between protein and DNA. Further experiments will be carried out to optimise these conditions. Afterwards, all 11 SNPs will be analysed by this method.

	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Biotin-EBNA Control DNA	+	+								
EBNA Extract		+								
FGFR2-13*			+							
ATM38/39*				+						
ATM40/41*					+					
ATM42/43*						+				
ATM44/45*							+			
ATM46/47*								+		
ATM48/49*									+	
ATM50/51*										+
NER-HCC1954			+	+	+	+	+	+	+	+

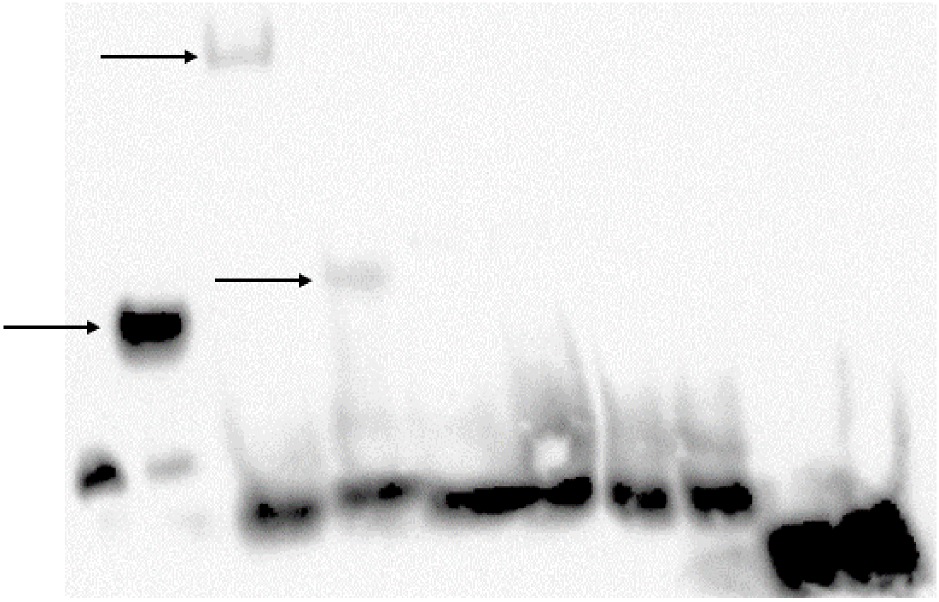


Figure 4.10 ***in vitro*** **DNA-protein binding studies**. This analysis was performed using HCC1954 nuclear extract. Biotin-EBNA Control DNA, Biotin-EBNA Control DNA+ EBNA Extract and FGFR2-13 and HMG I(Y) were included as positive control. rs7307700; rs12581512 and rs7133614 were the oligonucleotide previously chosen for the initial test. **ATM38/39** correspond to HMGI(Y), **ATM40/41** and **ATM42/43** correspond to G and A alleles (rs7307700), **ATM44/45** and **AMT46/47** correspond to A and G alleles (rs12581512) and **ATM48/49** and **ATM50/51** correspond to T and C allele (rs7133614). Arrow indicate specific bands for DNA-protein interactions. *****: represent the oligonucleotides labelled.

4.2.3 DAE validation of *COX11* and *AACS*

Since both *COX11* and *AACS* loci showed evidence functional regulatory potential in the *in silico* analysis we also decided to validate the DAE levels in normal breast tissue, for one SNP in each gene. We chose rs7138557 for *AACS* and rs17817901 for *COX11* (**Table 4.7**).

Table 4.7 SNPs in *COX11* and *AACS* used to assess DAE in breast tissue samples. In the table are shown the 2 located gene, function (Satus), chromosome (Chrom), minor allele frequency in the CEU HapMap population (MAF) and direction of DAE observed in the microarray study.

SNP	Gene	Chrom	Status	Alleles	MAF	DAE/scenario in microarrays
rs17817901	*COX11*	17	3'UTR	A/G	G = 0.165	Unidirectional (Scenario1)
rs7138557	*AACS*	12	Intron	C/T	C = 0.398	Unidirectional (Scenario2)

As explained previously, we first genotyped DNA samples from 84 normal breast tissue. The genotyping for both SNPs was well succeeded and 18 heterozygous samples were chosen for DAE assessment. All genotyping results are represented as supplementary material in **Annex 4.1**. SNPs showed a frequency of heterozygosity between 20-50% (**Table 4.8**). For rs17817901, the allelic frequencies observed were in agreement with what is described for the European population but for rs7138557 we observed a higher frequency of the C allele and this was reported has being the minor allele in the European population.

Table 4.8 Summary of genotyping results for the validation SNPs in breast tissue samples. In the table are shown the 2 genotyped SNPs and the percentages of undetermined genotypes, heterozigosity and allelic frequency for each SNP in the 84 normal breast tissue DNA samples genotyped.

SNP	Gene	Undetermined (%)	Heterozygosity (%)	Homozygosity	
				Allele 1 (%)	Allele 2 (%)
rs17817901	*COX11*	8.43	28.92 (A/G)	12.05 (G)	50.60 (A)
rs7138557	*AACS*	7.23	42.17 (C/T)	20.48 (T)	30.12 (C)

For rs17817901 (*COX11*) we did not observed DAE for any sample and therefore this result was not consistent with the original microarrays findings. For rs7138557 (*AACS*) we observed statistically significant DAE (p-value=8.76E-03) (**Table 4.9** and **Figure 4.11**), with a pattern consistent with Scenario 3, a pattern not consistent with the original microarrays findings.

Table 4.9 DAE average values for the validation SNPs in normal breast tissue. In the table are shown the 2 DAE SNPs, the number of heterozygous in the successfully quantified samples, the DAE average values observed, the standard deviation (SD), the minimum and maximum values of DAE distribution (Min-Max) and the p-value for the t-test.

SNP	Gene	Heterozygous with DAE	DAE mean	SD	Min-Max	*t-test* (p-value)
rs17817901	*COX11*	0/9	-0.16	0.16	-0.49 – -0.01	1.68E-02
rs7138557	*AACS*	4/11	0.03	0.72	-1.15 – 1.22	8.76E-03

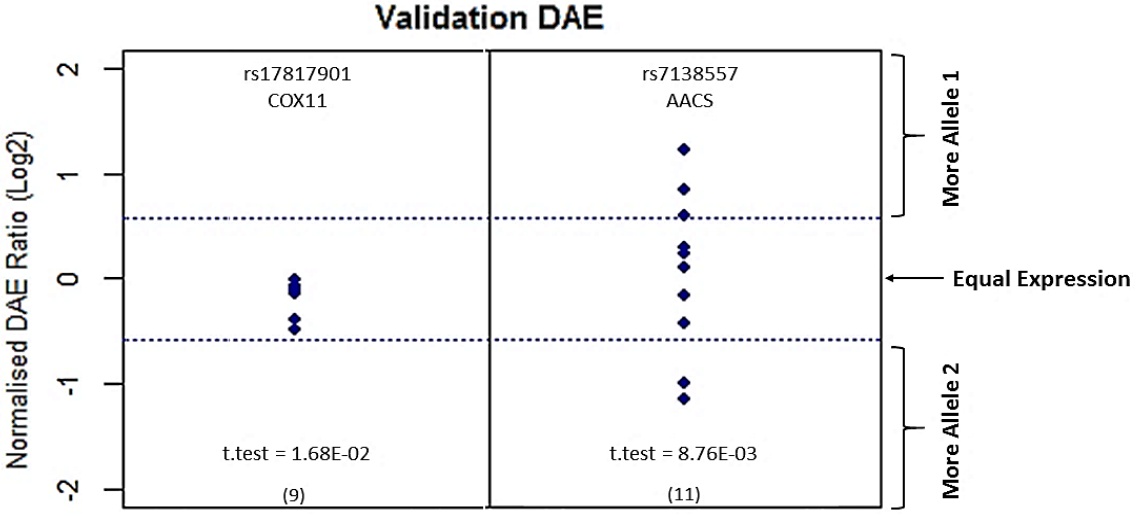


Figure 4.11 Results from DAE validation in *AACS* and *COX11*. x-axis indicates the name of SNPs and the y axis indicates the normalised DAE ratio. Heterozygous individuals are represented as blue dots. The numbers in parentheses are the numbers of individuals in which the DAE was quantified in our validation. Dotted lines delimit the cut-off of preferential allelic expression ratio [log2(1.5)=0.584].

4.2.4 Expression quantitative trait loci (eQTL) analysis

We also checked our samples of breast tissue and blood eQTL evidence in the two loci (17q22 and 12q24), as several variants seem to show eQTL in RegulomeDB. We used the data from previously microarray studies in breast already referred and also available data from blood for the same samples. For 17q22, we had genotyping data for SNPs rs244317, rs2628315, rs2628305 and rs2787481. For 12q24 we had genotyping data for rs7307700. For 17q22 and 12q24 loci, we found no significant eQTL evidence (**Annex 4.2** and **Annex 4.**, respectively). Therefore, no relationship was observed between the gene total expression levels with the SNPs genotype.

4.2.5 Comparison of DAE between breast tissue and blood

To understand whether DAE pattern is similar between blood and breast tissue and to see if it is possible to use blood as a suitable substitute for DAE quantification in breast, we performed DAE analysis in blood samples for *COX11* and *AACS*.

We began by genotyping rs17817901 (*COX11*) and rs7138557 (*AACS*) in 150 blood samples from healthy individuals (**Annex 4.1**) and found a frequency of heterozygosity of 41% and 38%, respectively (**Table 4.10**). For rs17817901, the allelic frequencies observed were in agreement with what is described for the European population and for rs7138557 was again discordant.

Table 4.10 Summary of genotyping results in blood. In the table are shown the 2 genotyped SNPs and the percentages of undetermined genotypes, heterozigosity and allelic frequency for each SNP in the 150 genotyped blood sample.

SNP	Gene	Undetermined (%)	Heterozygosity (%)	Homozygosity	
				Allele 1 (%)	Allele 2(%)
rs17817901	*COX11*	6.00	40.67 (A/G)	6.00 (G)	47.33 (A)
rs7138557	*AACS*	14.00	38.67 (C/T)	14.00 (T)	33.33 (C)

Subsequently, 18 heterozygous blood samples for both SNPs were chosen for DAE analysis. For rs17817901, two blood samples showed DAE but no breast samples and the DAE distribution showed a pattern consistent with Scenario 2. For rs7138557, four breast samples and one blood sample showed DAE, and the DAE distribution had a pattern consistent with Scenario 3 and 2, respectively (**Figure 4.12**).

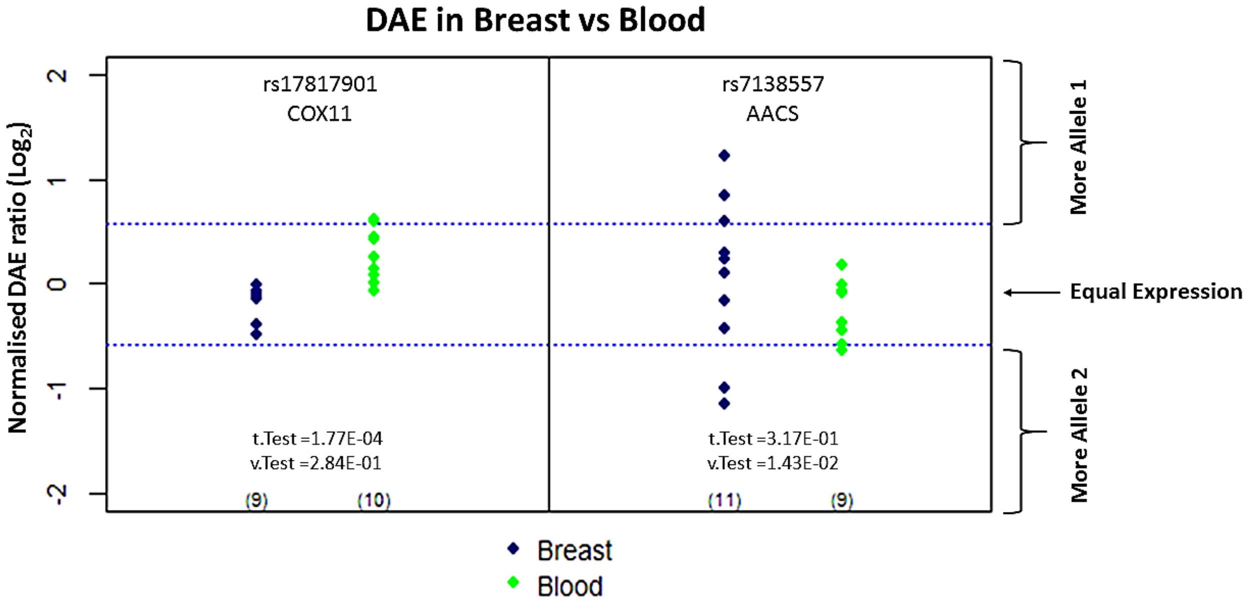


Figure 4.12 Comparison of DAE between normal breast tissue and blood from healthy individuals. Breast samples are represented as blue dots and blood samples are represented as green dots. Vertical lines separate each SNP. Dotted lines delimit the cut-off of preferential allelic expression ratio [$\log_2(1.5)=0.584$].

We found a significant difference between the two DAE distributions means for rs17817901 (p-value= 1.77E-04), but not for rs7138557 (p-value=3.17E-01). Nevertheless, detected significant differences between the DAE distribution variances for rs7138557 (p-value=1.43E-02) but not for rs17817901 (*p-value*= 2.84E-01) (**Figure 4.12** and **Table 4.11**). Therefore, this analysis suggests that DAE in breast tissue and blood is different.

Table 4.11 Comparison of normalised DAE between breast tissue and blood for *AACS* and *COX11*. In the table are shown the 2 DAE SNPs, the number of heterozygous in the successfully quantified samples, the DAE average values observed, the standard deviation (SD), the minimum and maximum values of DAE distribution (Min-Max).

SNP/Gene	Samples	Heterozygous with DAE	Mean DAE	SD	Min-Max	*t-test* (p-value)	*var-test* (p-value)
rs17817901 (*COX11*)	**Breast**	0/9	-0.16	0.16	-0.49 – -0.01	1.77E-04	2.84E-01
	Blood	2/10	0.29	0.24	-0.06 – 0.61		
rs7138557 (*AACS*)	**Breast**	4/11	0.03	0.72	-1.15 – 1.22	3.17E-01	1.43E-02
	Blood	1/9	-0.23	0.28	-0.64 – 0.18		

4.2.6 Case – control association study

To understand if DAE in both loci is truly associated with breast cancer risk, we performed a small case-control study.

Samples from healthy individuals (Controls) and from patients with breast cancer (Cases) were analysed for DAE in *COX11* and *AACS*. We began by genotyping rs17817901 and rs7138557 in 56 blood samples from cases of cancer (**Annex 4.**) since the controls had been already genotyped. Heterozygosity frequencies were 30% and 39% (**Table 4.12**), respectively. rs17817901 in concordance with European allelic frequency and rs7138557 was again discordant for the allelic frequency.

Table 4.12 Summary of genotyping results for the SNPs (in *COX11* and *AACS*) from blood cancer samples. In the table are shown the 2 genotyped SNPs and the percentages of undetermined genotypes, heterozigosity and allelic frequency for each SNP in the 56 genotyped blood cancer samples.

SNP	Gene	Undetermined (%)	Heterozygosity (%)	Homozygosity	
				Allele 1 (%)	Allele 2(%)
rs17817901	*COX11*	1.79	30.36 (A/G)	7.14 G (G)	60.71 (A)
rs7138557	*AACS*	14.00	38.67 (C/T)	14.00 (T)	33.33 (C)

Subsequently, 10 heterozygous samples from cases and controls were selected and analysed for DAE in both SNPs. For rs17817901, three controls showed DAE but no cases and the DAE distribution showed a pattern consistent with Scenario 2. For rs7138557, four controls and two cases showed DAE and the DAE distribution had a pattern consistent with Scenario 2 (**Figure 4.13**).

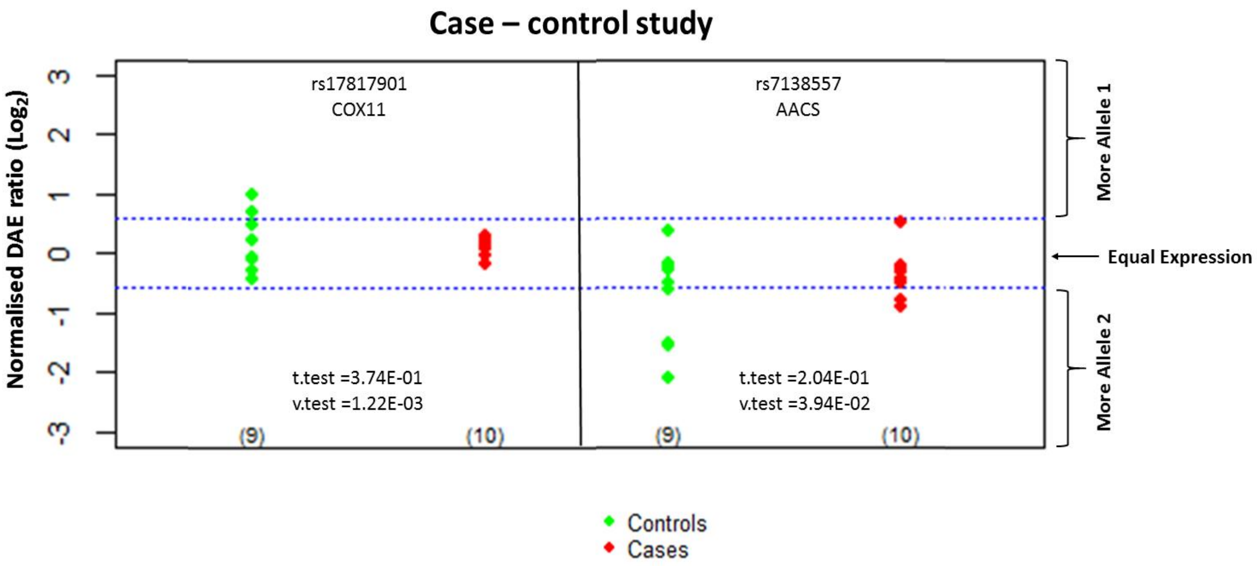


Figure 4.13 Case – control association study for rs7138557 (*AACS* gene) and rs17817901 (*COX11* gene) in blood. Controls are represented as green dots for cases are represented as red dots. Dotted lines delimit the cut-off of preferential allelic expression ratio [$\log_2(1.5)=0.584$].

A significant difference was found for the variances of the two populations for both SNPs (rs17817901: p-value= 1.22E-03 and rs7138557: *p-value*= 3.94E-02), but no significant difference was found for the DAE distribution means (rs17817901: p-value= 3.74E-01 and rs7138557: p-value= 2.04E-01) (**Figure 4.13** and **Table 4.13**).

Table 4.13 DAE distributions in controls (healthy individuals) and cases (patients with breast cancer). In the table are shown the 2 DAE SNPs, identification of the two groups of samples, the number of heterozygous in the successfully quantified samples.

SNP/Gene	Samples	Heterozygous with DAE	Mean DAE	SD	Min-Max	*t-test* (p-value)	*var-test* (p-value)
rs17817901 (*COX11*)	**Controls**	3/10	0.22	0.49	-0.41 – 1.00	3.74E-01	1.22E-03
	Cases	0/10	0.12	0.14	-0.17 – 0.29		
rs7138557 (*AACS*)	**Controls**	4/10	-0.46	0.76	-2.08 – 0.37	2.04E-01	3.94E-02
	Cases	2/10	-0.35	0.38	-0.90 – 0.51		

CHAPTER 5 – DISCUSSION

5.1 Validation of DAE microarray data using allele-specific Taqman

The first aim in this study was to validate the DAE results obtained previously by microarray technology. For this purpose, 14 candidate tSNPs were analysed for DAE in heterozygous samples from normal breast tissue, using allele-specific real-time PCR. These SNPs were located in either coding or non-coding regions, because total RNA (including unspliced primary transcripts) was used in the analysis and therefore it is possible to measure the DAE levels for SNPs located in introns.

According to the previous findings, we expected to observe DAE for 10 SNPs and no DAE for 4 SNPs. We confirmed that four SNPs (rs10503416, rs6494466, rs10016 and rs13265801) showed DAE pattern and scenario consistent with the previous results. According to the scenario, the regulatory variant (rSNP) is in strong LD with the assayed cSNP. Regarding the non-DAE SNPs, we observed that only one (rs10521) was consistent with the results in the previous work. Therefore, six out of 14 SNPs were validated. The remaining SNPs showed DAE distributions, but discordant patterns with the previous results. Therefore, the percentage of SNPs reported to have allelic imbalances in gene expression differs between approaches.

We need to confirm these results in order to understand the cause of disparity and to insure a good quality of validation. A possible solution can be to design new primers, to perform convectional PCR followed by Sanger sequencing, and to compare the sequencing traces between gDNA with the cDNA (semi-quantitative sequencing). Another will be to perform RNA-seq allelic expression quantification.

5.2 Crossing GWAS data and DAE data

The second aim of this project was to identify new susceptibility loci by crossing DAE data with published breast cancer GWAS top hits. We selected two loci 17q22 (*TOM1L1/COX11/STXBP4*) and 12q24 (*AACS*) to perform *in silico* analysis, in search of potential functional evidence in order to identify risk associated cis-regulatory variants.

In 17q22 we observed the presence of one DAE SNP (rs17817901) in strong LD with a GWAS associated SNP (rs6504950), suggesting that the rs17817901 is a marker for the cis-regulatory variant. This led us to investigate the region around these SNPs for histone modifications, DNase clusters and transcription factors binding sites. We observed that this is a very active region with multiple evidence for regulatory elements. Six investigated SNPs (proxies from the GWAS and DAE SNPs) overlap some of these elements and therefore are candidates to be cis-regulatory SNPs and risk-causing. We found that the minor allele of rs17817901, rs7643 and rs2628315 is associated with an increased expression. The next logical step will be to search for TFBS within the sequences containing the candidate rSNPs to understand if these alter the binding affinity of TFs and are likely to affect gene regulation. Further functional studies will follow. Nevertheless, we believe that there is already data supporting a cis-regulatory role for this breast cancer susceptibility locus.

In the 12q24 locus, we found three DAE SNP (rs7138557 and rs12581512) in strong LD with two GWAS associated SNP (rs7307700 and rs2291248, respectively). Interestingly, rs2291248 has both been identified in DAE studies and GWAS. Our in-silico analysis has produced 12 strong cis-regulatory candidates, which we are currently being tested by EMSA, for in vitro DNA-protein binding.

Of these, the common A allele of rs7307700 has one possible TFBS for HMGI(Y). HMGI-Y is a member of the mammalian high-mobility group I (HMGI) and has been shown to play an important role in the regulation of cell proliferation and differentiation (Klein-Hessling et al. 1996; Giannini et al. 2000; Reeves et al. 2001; Melillo & Pierantoni 2001). For rs12581512 both alleles had possible TFBS (minor A allele had two possible bindings for HES1 and CAP, and the common G allele had one possible binding for CBF). Hairy and enhancer of split homolog-1 (HES1) is involved in the maintenance of certain stem cells and progenitor cells. This TF has an important role in the tumourigenesis and some studies showed that it can modulate the therapeutic resistance in breast cancer (Murata et al. 2005; Gao et al. 2014). Catabolite activator protein (CAP) is involved in the active transcription of a several promoters, which directs operons involved in catabolite metabolism (signal for transcription initiation) (Zhou et al. 1993; Lawson et al. 2004).

CBF (also known as NF-Y) is a CCAAT-binding protein and is involved in the regulation of several genes and is also required for cell proliferation (Bhattacharya et al. 2003; Cagliari 2011). rs7133614 the common C allele had two possible TFBS for GEN_INI3_β and CAP. GEN_INI is a general initiator sequence (viral + cellular) and a component of the DNA replication machinery that recognizes and recruits factors necessary to initiate of replication (Stenlund 2003). All these TFs have been reported to be involved in cancer, which further supports the link between cis-regulation in this locus and susceptibility to breast cancer.

To investigate whether the sequence surrounding these candidate rSNPs can bind to the TF, the protein-DNA interactions were analysed through EMSA. We performed two experiments, but did not detect any shift, for our oligonucleotides. However, FGFR2-13 and HMGI(Y), used as positive controls, showed a faint shift. There are different possible explanations for these results: there is no interaction between TFs and our oligonucleotides, therefore these SNPs are not cis-regulatory; or the TF can bind to our oligonucleotides but the quantity of protein used was probably too low to form a strong binding. The last is the most plausible as the controls, which worked well before, show the need for optimisation. If the protein-DNA binding occurs we will use specific antibodies against the predicted TFs. This type of analysis indicates which variants candidates bind to TFs and how the allele change in the SNP affects the binding. Subsequently, chromatin immunoprecipitations will be carried out to study this interaction *in vivo*.

5.2.1 Validation of DAE in *COX11* and *AACS*

We used a marker SNP for each loci (rs17817901 in *COX11/TOM1L1* gene and rs7138557 in *AACS* gene) to validate the previous DAE results. We observed that *COX11 did* not show DAE and *AACS* showed a bidirectional DAE distribution. These results confirmed the presence of DAE in one gene but are discordant in terms of DAE pattern. The same discussion as before applies.

Our study comparing DAE between two different tissues (breast tissue and blood) showed that there are significant differences in the DAE distributions. Therefore, these data suggest that it is not feasible the use blood as surrogate for breast tissue for DAE quantification.

In the case-control study for both genes, we observed different DAE distributions between cases and controls. In *COX11 w*e observed that in the control samples there was preferential allelic expression and this can be associated with a protection to breast cancer in this group. In *AACS* we observed a preferential allelic expression of the same allele in both groups (cases and controls). In this case it may be inferred that individuals with more extreme DAE may be protected from breast cancer. Nevertheless, these results require further validation in a larger set of samples.

Ours results suggest that DAE can be a good method to perform association studies and understand the susceptibility of an individual to a complex disease.

Integrating GWAS data with our DAE map information proved to be a good approach for identifying new susceptibility loci for breast cancer, including prioritising candidate GWAS for functional analysis. We identified so far two risk-loci that could be under the influence of cis-regulatory variants. However, we have a long candidate list of loci that contain risk-associated SNPs and DAE SNPs. Therefore, further analysis is needed to understand if these loci could also be under the influence of cis-regulatory variants.

In the future, further analysis will contribute to a better understanding of the biology underlying breast cancer risk and contribute to the development of future therapies, especially in the personalized medicine area.

BIBLIOGRAPHY

Ahmed, S. et al., 2009. Newly discovered breast cancer susceptibility loci on 3p24q and 17q23.2. *Nature genetics*, 41(5), pp.585–590. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2748125/.

Antoniou, a C. & Easton, D.F., 2006. Models of genetic susceptibility to breast cancer. *Oncogene*, 25(43), pp.5898–905. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16998504 [Accessed January 7, 2014].

Antoniou, A.C. et al., 2011. Europe PMC Funders Group Common breast cancer susceptibility alleles and the risk of breast cancer for BRCA1 and BRCA2 mutation carriers : implications for risk prediction. *Cancer Research*, 70(23), pp.9742–9754.

Apostolou, P. & Fostira, F., 2013. Hereditary breast cancer: the era of new susceptibility genes. *BioMed research international*, 2013, p.747318. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3618918&tool=pmcentrez&rendertype=abstract.

Ashworth, A., Lord, C.J. & Reis-Filho, J.S., 2011. Genetic interactions in cancer progression and treatment. *Cell*, 145(1), pp.30–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21458666 [Accessed July 9, 2014].

Bannister, A.J. & Kouzarides, T., 2011. Regulation of chromatin by histone modifications. *Cell research*, 21(3), pp.381–95. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3193420&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Bhatti, P., Doody, M. & Rajaraman, P., 2010. Novel breast cancer risk alleles and interaction with ionizing radiation among US radiologic technologists. *Radiation Research*, 173(2), pp.214–224. Available at: http://www.rrjournal.org/perlserv/?request=get-abstract&doi=10.1667%2FRR1985.1 [Accessed September 22, 2014].

Boyle, A. et al., 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, 22((9)), pp.1790–1797. Available at: http://genome.cshlp.org/content/22/9/1790.short [Accessed August 31, 2014].

Brookes, a J., 1999. The essence of SNPs. *Gene*, 234(2), pp.177–86. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10395891.

Buckland, P.R., 2006. The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochimica et biophysica acta*, 1762(1), pp.17–28. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16297602 [Accessed May 30, 2014].

Campa, D. et al., 2011. Interactions between genetic variants and breast cancer risk factors in the breast and prostate cancer cohort consortium. *Journal of the National Cancer Institute*, 103(16), pp.1252–63. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3156803&tool=pmcentrez&rendertype=abstract [Accessed September 22, 2014].

Chen, F. et al., 2013. A genome-wide association study of breast cancer in women of African ancestry. *Human genetics*, 132(1), pp.39–48. Available at: http://link.springer.com/article/10.1007/s00439-012-1214-y [Accessed September 10, 2014].

Cheung, V. et al., 2005. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437(7063), pp.1365–1369. Available at: http://www.nature.com/nature/journal/v437/n7063/abs/nature04244.html [Accessed June 20, 2014].

Chorley, B.N. et al., 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutation research*, 659(1-2), pp.147–57. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2676583&tool=pmcentrez&rendertype=abstract [Accessed May 30, 2014].

Couch, F.J. et al., 2013. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS genetics*, 9(3), p.e1003212. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3609646&tool=pmcentrez&rendertype=abstract.

Couch, F.J. et al., 2014. NIH Public Access. , 45(4), pp.392–398.

Easton, D., Pooley, K. & Dunning, A., 2007. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148), pp.1087–1093. Available at: http://www.nature.com/nature/journal/v447/n7148/full/nature05887.html%3Freferer=www.clickfind.com.au7?message=remove&referer=www.clickfind.com.au7 [Accessed September 9, 2014].

Easton, D.F. & Eeles, R. a, 2008. Genome-wide association studies in cancer. *Human molecular genetics*, 17(R2), pp.R109–15. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18852198 [Accessed September 10, 2014].

Fletcher, O. et al., 2011. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *Journal of the National Cancer Institute*, 103(5), pp.425–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21263130 [Accessed September 4, 2014].

Garcia-Closas, M. & Chanock, S., 2008. Genetic susceptibility loci for breast cancer by estrogen receptor status. *Clinical Cancer Research*, 14(24), pp.8000–8009. Available at: http://clincancerres.aacrjournals.org/content/14/24/8000.short [Accessed June 22, 2014].

Garraway, L. a & Lander, E.S., 2013. Lessons from the cancer genome. *Cell*, 153(1), pp.17–37. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23540688 [Accessed July 12, 2014].

Ghoussaini, M., Pharoah, P.D.P. & Easton, D.F., 2013. Inherited genetic susceptibility to breast cancer: the beginning of the end or the end of the beginning? *The American journal of pathology*, 183(4), pp.1038–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23973388 [Accessed January 7, 2014].

Gray, I.C., Campbell, D. a & Spurr, N.K., 2000. Single nucleotide polymorphisms as tools in human genetics. *Human molecular genetics*, 9(16), pp.2403–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11005795.

Green, M.R. & Sambrook, J., 2012. *Molecular Cloning* Fourth. J. Inglis, A. Boyle, & A. Gann, eds., John Inglis.

Hanahan, D. & Weinberg, R. a, 2011. Hallmarks of cancer: the next generation. *Cell*, 144(5), pp.646–74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21376230 [Accessed November 6, 2013].

He, H.H. et al., 2012. Differential DNase I hypersensitivity reveals factor-dependent chromatin dynamics. *Genome research*, 22(6), pp.1015–25. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3371710&tool=pmcentrez&rendertype=abstract [Accessed September 12, 2014].

Heid, C. a et al., 1996. Real time quantitative PCR. *Genome Research*, 6(10), pp.986–994. Available at: http://www.genome.org/cgi/doi/10.1101/gr.6.10.986 [Accessed July 9, 2014].

Hellman, L.M. & Fried, M.G., 2007. Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein-Nuclec Acid Interaction. *NAt Protoc.*, 2(8), pp.1849–1861.

Holden, N.S. & Tacon, C.E., 2011. Principles and problems of the electrophoretic mobility shift assay. *Journal of pharmacological and toxicological methods*, 63(1), pp.7–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20348003 [Accessed July 19, 2014].

Hon, G., Wang, W. & Ren, B., 2009. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS computational biology*, 5(11), p.e1000566. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2775352&tool=pmcentrez&rendertype=abstract [Accessed September 22, 2014].

Johnson, A.D. et al., 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics (Oxford, England)*, 24(24), pp.2938–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2720775&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2014].

Jones, B.L. & Swallow, D.M., 2011. The impact of cis-acting polymorphisms on the human phenotype. *The HUGO journal*, 5(1-4), pp.13–23. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3238023&tool=pmcentrez&rendertype=abstract [Accessed January 30, 2014].

Jorde, L.B., 2000. Linkage Disequilibrium and the Search for Complex Disease Genes. *Genome Research*, 10(10), pp.1435–1444. Available at: http://www.genome.org/cgi/doi/10.1101/gr.144500 [Accessed September 10, 2014].

Kelsey, J.L. & Berkowitz, G.S., 1988. Breast Cancer Epidemiology Breast Cancer Epidemiology. , pp.5615–5623.

Kim, H. et al., 2012. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. *Breast cancer research : BCR*, 14(2), p.R56. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3446390&tool=pmcentrez&rendertype=abstract [Accessed September 10, 2014].

Klein-Hessling, S. et al., 1996. HMG I(Y) interferes with the DNA binding of NF-AT factors and the induction of the interleukin 4 promoter in T cells. *Proceedings of the National Academy of Sciences of the United States of America*, 93(26), pp.15311–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=26401&tool=pmcentrez&rendertype=abstract.

Lappalainen, T. & Dermitzakis, E.T., 2010. Evolutionary history of regulatory variation in human populations. *Human molecular genetics*, 19(R2), pp.R197–203. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20876617 [Accessed June 2, 2014].

Loizidou, M. a et al., 2011. Replication of genome-wide discovered breast cancer risk loci in the Cypriot population. *Breast cancer research and treatment*, 128(1), pp.267–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21210208 [Accessed September 22, 2014].

Low, S.-K. et al., 2013. Genome-wide association study of breast cancer in the Japanese population. *PloS one*, 8(10), p.e76463. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3797071&tool=pmcentrez&rendertype=abstract [Accessed September 10, 2014].

Mahdi, K., Nassiri, M. & Nasiri, K., 2013. Hereditary genes and SNPs associated with breast cancer. *Asian Pac. J. Cancer Prev*, 14, pp.3403–3409. Available at: http://www.apjcpcontrol.org/paper_file/issue_abs/Volume14_No6/3403-3409 5.24 Mohammad Mahdi Kooshyar.pdf [Accessed June 23, 2014].

Malhotra, G.K. et al., 2010. Histological, molecular and functional subtypes of breast cancers. *Cancer Biology & Therapy*, 10(10), pp.955–960. Available at: http://www.landesbioscience.com/journals/cbt/article/13879/ [Accessed July 17, 2014].

Mavaddat, N. et al., 2010. Genetic susceptibility to breast cancer. *Molecular oncology*, 4(3), pp.174–91. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20542480 [Accessed November 15, 2013].

Meyer, K.B. et al., 2008. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS biology*, 6(5), p.e108. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2365982&tool=pmcentrez&rendertype=abstract [Accessed September 21, 2014].

Michailidou, K. et al., 2013. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature …*, 45(4), pp.353–361. Available at: http://www.nature.com/ng/journal/v45/n4/abs/ng.2563.html [Accessed June 23, 2014].

Nica, A.C., Dermitzakis, E.T. & B, P.T.R.S., 2013. Expression quantitative trait loci : present and future Expression quantitative trait loci : present and future. , (May).

Oldenburg, R. a et al., 2007. Genetic susceptibility for breast cancer: how many more genes to be found? *Critical reviews in oncology/hematology*, 63(2), pp.125–49. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17498966 [Accessed November 13, 2013].

Pant, P.V.K. et al., 2006. Analysis of allelic differential expression in human white blood cells. , pp.331–339.

Pastinen, T., Ge, B. & Hudson, T.J., 2006. Influence of human genome polymorphism on gene expression. *Human molecular genetics*, 15 Spec No(1), pp.R9–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16651375 [Accessed May 29, 2014].

Polyak, K., 2011. Review series introduction Heterogeneity in breast cancer. *The Journal of Clinical Investigation*, 121(10), pp.3786–3788.

R Core Team, 2014. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. Available at: http://www.R-project.org/.

Reeves, R., Edberg, D. & Li, Y., 2001. Architectural transcription factor HMGI (Y) promotes tumor progression and mesenchymal transition of human epithelial cells. *Molecular and cellular biology*, 21(2), pp.575–594. Available at: http://mcb.asm.org/content/21/2/575.short [Accessed September 25, 2014].

Reich, D.E. et al., 2001. Linkage disequilibrium in the human genome. *Nature*, 411(6834), pp.199–204. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11346797.

Rockman, M. V & Wray, G. a, 2002a. Abundant raw material for cis-regulatory evolution in humans. *Molecular biology and evolution*, 19(11), pp.1991–2004. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12411608.

Rockman, M. V & Wray, G. a, 2002b. Abundant raw material for cis-regulatory evolution in humans. *Molecular biology and evolution*, 19(11), pp.1991–2004. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12411608.

Sadikovic, B. et al., 2008. Cause and consequences of genetic and epigenetic alterations in human cancer. *Current genomics*, 9(6), pp.394–408. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2691666&tool=pmcentrez&rendertype=abstract.

Serre, D. et al., 2008. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS genetics*, 4(2), p.e1000006. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2265535&tool=pmcentrez&rendertype=abstract [Accessed May 21, 2014].

Strachan, T. & Read, A.P., 1996. *Human Molecular Genetics* B. S. P. Limited, ed.,

Stranger, B. et al., 2007. Population genomics of human gene expression. *Nature …*, 39(10), pp.1217–1224. Available at: http://www.nature.com/ng/journal/v39/n10/abs/ng2142.html [Accessed June 20, 2014].

Tang, L. et al., 2012. Association of STXBP4/COX11 rs6504950 (G>A) polymorphism with breast cancer risk: evidence from 17,960 cases and 22,713 controls. *Archives of medical research*, 43(5), pp.383–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22863968 [Accessed September 21, 2014].

Thomas, G. et al., 2010. A multi-stage genome-wide association in breast cancer identifies two novel risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nature genetics*, 41(5), pp.579–584. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928646/.

Turnbull, C. et al., 2013. Europe PMC Funders Group Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature genetics*, 42(6), pp.504–507.

VanLiere, J. & Rosenberg, N., 2008. Mathematical properties of the r2 measure of linkage disequilibrium. *Theoretical population biology*, 74(1), pp.130–137. Available at: http://www.sciencedirect.com/science/article/pii/S0040580908000609 [Accessed September 21, 2014].

Verlaan, D., Ge, B. & Grundberg, E., 2009. Targeted screening of cis-regulatory variation in human haplotypes. *Genome …*, pp.118–127. Available at: http://genome.cshlp.org/content/19/1/118.short [Accessed June 20, 2014].

Vernot, B. et al., 2012. Personal and population genomics of human regulatory variation. *Genome research*, 22(9), pp.1689–97. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3431486&tool=pmcentrez&rendertype=abstract [Accessed June 2, 2014].

Vogelstein, B. & Kinzler, K.W., 2004. Cancer genes and the pathways they control. *Nature medicine*, 10(8), pp.789–99. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15286780 [Accessed July 11, 2014].

Walsh, T. & King, M.-C., 2007. Ten genes for inherited breast cancer. *Cancer cell*, 11(2), pp.103–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17292821 [Accessed November 7, 2013].

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. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16002116 [Accessed May 30, 2014].

Wang, Y.-M. et al., 2012. Correlation between DNase I hypersensitive site distribution and gene expression in HeLa S3 cells. *PloS one*, 7(8), p.e42414. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3416863&tool=pmcentrez&rendertype=abstract [Accessed September 22, 2014].

Wingender, E. et al., 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic acids research*, 28(1), pp.316–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102445&tool=pmcentrez&rendertype=abstract.

Worsley-Hunt, R., Bernard, V. & Wasserman, W.W., 2011. Identification of cis-regulatory sequence variations in individual genome sequences. *Genome medicine*, 3(10), p.65. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3239227&tool=pmcentrez&rendertype=abstract.

Xiao, R. & Scott, L.J., 2011. Detection of cis-acting regulatory SNPs using allelic expression data. *Genetic epidemiology*, 35(6), pp.515–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21769929 [Accessed May 6, 2014].

You, J.S. & Jones, P. a, 2012. Cancer genetics and epigenetics: two sides of the same coin? *Cancer cell*, 22(1), pp.9–20. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3396881&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

ANNEX 1

Annex 1.1 Scores assigned by RegulomeDB according the functional evidence

Category scheme	
Category	**Description**
Likely to affect binding and linked to expression of a gene target	
1a	eQTL + TF binding + matched TF motif + matched DNase footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding/DNase peak Likely to affect binding
2a	TF binding + matched TF motif + matched DNase footprint + DNase peak
2b	TF binding + any motif + DNase footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
Less likely to affect binding	
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
Minimal binding evidence	
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Motif hit

Annex 1.2 Sequences obtained in dbSNP for both alleles of each SNPs. The sequences were analysed in TRANSFAC. This sequence is in forward strand (FW). The minor and common alleles are represented in red. All sequences had a total of 61bp.

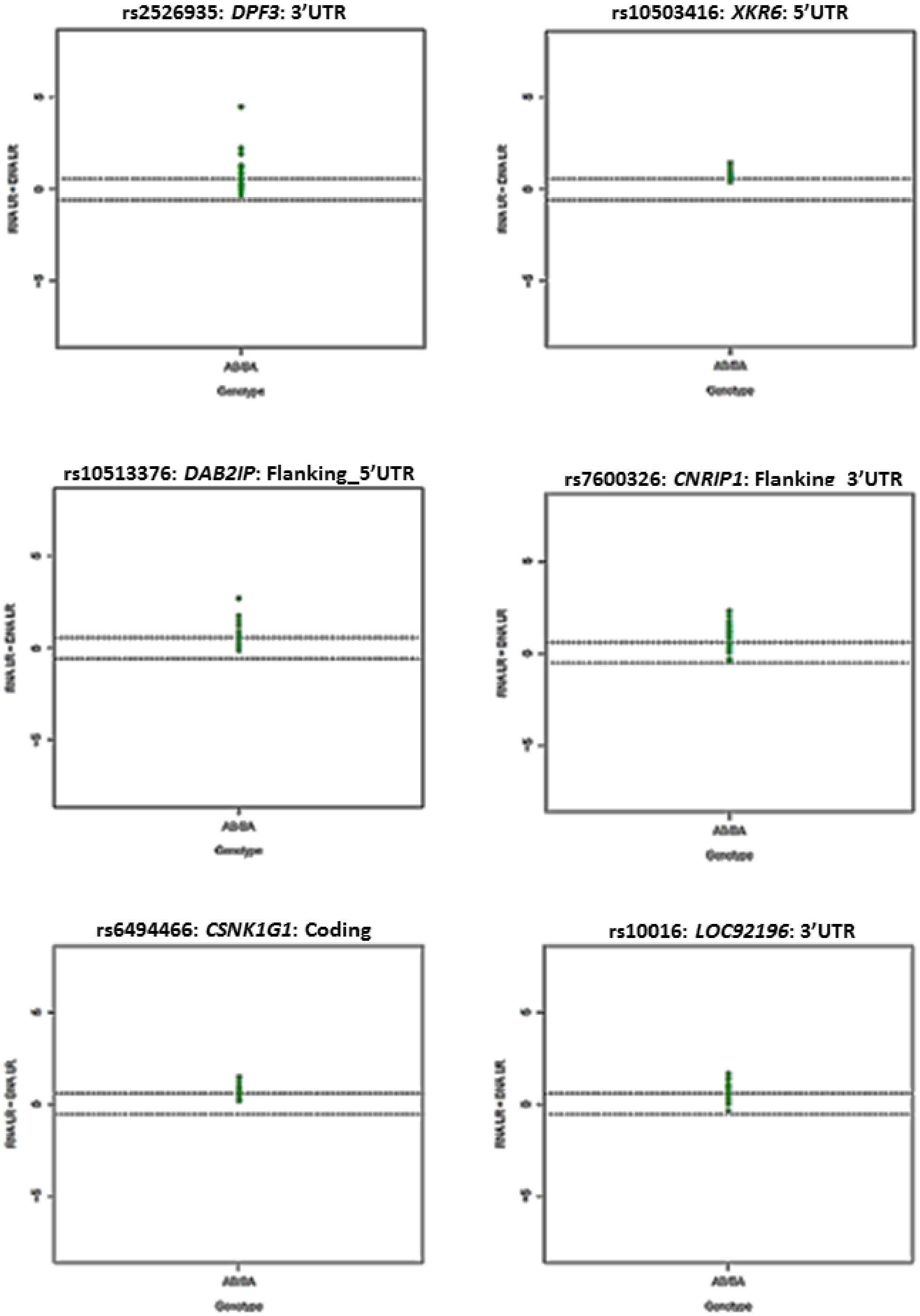
SNPs	Minor allele	Common allele	Sequence for the two alleles
rs7138557	C (FW)	T (FW)	TCCTGCGGGGAAGTAGGCCTCTGGATCTTCCACTTGGGGTCACTCAGAGAATTTTAGAAGT
			TCCTGCGGGGAAGTAGGCCTCTGGATCTTCTACTTGGGGTCACTCAGAGAATTTTAGAAGT
rs2291248	T (FW)	C (FW)	ACCCCTGGGACTTGCTTCACTCCAGAGCTGTGGTGTGGCCCTGACCTCTTCTCTCTTTCCA
			ACCCCTGGGACTTGCTTCACTCCAGAGCTGCGGTGTGGCCCTGACCTCTTCTCTCTTTCCA
rs12581512	A (FW)	G (FW)	CATTTGCATTGCTTACAAAACGGATACCCCACAAGCTGACAGAAGCTGCTGTGTGTGTGTG
			CATTTGCATTGCTTACAAAACGGATACCCCGCAAGCTGACAGAAGCTGCTGTGTGTGTGTG
rs7307700	G (FW)	A (FW)	ACATCTTCCTTGGCTTTTCTGACATTTGTAGAAGAATAAGCCACATGTTTTGTGTACCCAA
			ACATCTTCCTTGGCTTTTCTGACATTTGTAAAAGAATAAGCCACATGTTTTGTGTACCCAA
rs10846832	C (FW)	G (FW)	TGGACTACGTGCTCTGAGCCAAGCTCCCTGCGGGGTCCTGGGGGCCACACAGCAGCGACCA
			TGGACTACGTGCTCTGAGCCAAGCTCCCTGGGGGGTCCTGGGGGCCACACAGCAGCGACCA
rs10846834	A (FW)	G (FW)	GTAGCTAGGCACTTCTGAAGCTGTGTGTGCACTGATTCATTCACCCAGTGACTCACAGCCT
			GTAGCTAGGCACTTCTGAAGCTGTGTGTGCGCTGATTCATTCACCCAGTGACTCACAGCCT
rs58416336	T (FW)	G (FW)	GGACTACGTGCTCTGAGCCAAGCTCCCTGGTGGGTCCTGGGGGCCACACAGCAGCGACCAG
			GGACTACGTGCTCTGAGCCAAGCTCCCTGGGGGGTCCTGGGGGCCACACAGCAGCGACCAG
rs7133614	T (FW)	C (FW)	TTTTGTGGCAAGAGGCATGCGGGGCAGGCATAGTCCTCGTGATGCTGTCTCAGTGCCTCTG
			TTTTGTGGCAAGAGGCATGCGGGGCAGGCACAGTCCTCGTGATGCTGTCTCAGTGCCTCTG
rs7137742	A (FW)	G (FW)	CCGCCTTTTTCGGGCAAATTACAGTCACGTAGTTACAAAAGCTTGGAAGAGGACCCAGGCA
			CCGCCTTTTTCGGGCAAATTACAGTCACGTGGTTACAAAAGCTTGGAAGAGGACCCAGGCA
rs7138405	T (FW)	C (FW)	AGAGGCCCCTTTGAGTGGCTGGAGGCACTGTACGTTCCGAAGCTAAGGCACCTGAACTAGC
			AGAGGCCCCTTTGAGTGGCTGGAGGCACTGCACGTTCCGAAGCTAAGGCACCTGAACTAGC
rs7138790	G (FW)	C (FW)	GAAGGGGCTGCGCCAGCACATTTCCCTGCCGCTAATCACAAATGCCCTGGGCCCCTCCACC
			GAAGGGGCTGCGCCAGCACATTTCCCTGCCCCTAATCACAAATGCCCTGGGCCCCTCCACC
rs35428999	A (FW)	G (FW)	CAGACTTCTCCTCGCTCTGCAATGCACGCCACTCACTCCCTCCCTCATTCAGCCTTCCACT
			CAGACTTCTCCTCGCTCTGCAATGCACGCCGCTCACTCCCTCCCTCATTCAGCCTTCCACT
rs7304979	A (FW)	G (FW)	AGAGCCGTGCCTGGCTGGGGCTGTCGCCACAGGGCCACTACAAGGCAGGCCGTGGAGCAGG
			AGAGCCGTGCCTGGCTGGGGCTGTCGCCACGGGGCCACTACAAGGCAGGCCGTGGAGCAGG

Annex 1.3 Oligonucleotide sequences designed for EMSA. In the table are represented the three selected SNPs to analyse. Minor allele is shown first and the common allele is shown second. This table also show the sequences designed for both alleles of each SNP and the name the name we gave him (Primer).

SNP	Alleles	Strand	Sequence	Primer
rs7307700	G	FWD	TTCTGACATTTGTAGAAGAATAAGCCACATG	ATM40
		REV	CATGTGGCTTATTCTTCTACAAATGTCAGAA	ATM41
	A	FWD	TTCTGACATTTGTAAAAGAATAAGCCACATG	ATM42
		REV	CATGTGGCTTATTCTTTTACAAATGTCAGAA	ATM43
rs12581512	A	FWD	CAAAACGGATACCCCACAAGCTGACAGAAG	ATM44
		REV	CTTCTGTCAGCTTGTGGGGTATCCGTTTTG	ATM45
	G	FWD	CAAAACGGATACCCCGCAAGCTGACAGAAG	ATM46
		REV	CTTCTGTCAGCTTGCGGGGTATCCGTTTTG	ATM47
rs7133614	T	FWD	CATGCGGGGCAGGCATAGTCCTCGTGATGCTG	ATM48
		REV	CAGCATCACGAGGACTATGCCTGCCCCGCATG	ATM49
	C	FWD	CATGCGGGGCAGGCACAGTCCTCGTGATGCTG	ATM50
		REV	CAGCATCACGAGGACTGTGCCTGCCCCGCATG	ATM51

ANNEX 2

Annex 2.1 DAE SNPs reported in previous results obtained in microarray (Maia et al, unpublished). These are the 10 DAE SNPs that we chose to validate. In a x-axis indicates the genotype (A/B-heterozygous) and the y-axis indicates the normalised DAE ratio obtained. Dotted lines delimit the cut-off of preferential allelic expression ratio [$\log_2(1.5)=0.584$].



rs2526935: *DPF3*: 3'UTR

rs10503416: *XKR6*: 5'UTR

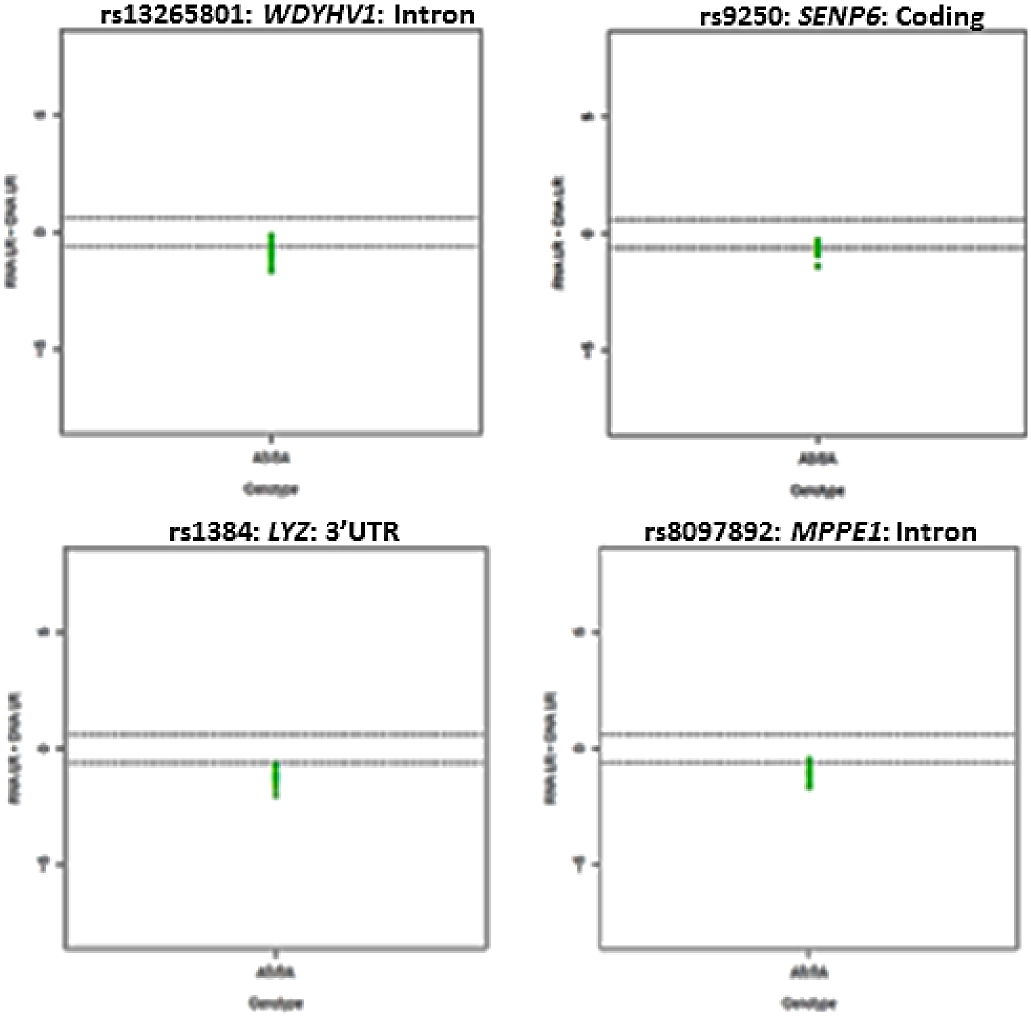
rs10513376: *DAB2IP*: Flanking_5'UTR

rs7600326: *CNRIP1*: Flanking_3'UTR

rs6494466: *CSNK1G1*: Coding

rs10016: *LOC92196*: 3'UTR

Annex 2.1 (Continuation of results of previous table – DAE SNPs reported in previous results obtained in microarray) (Maia et al, unpublished)



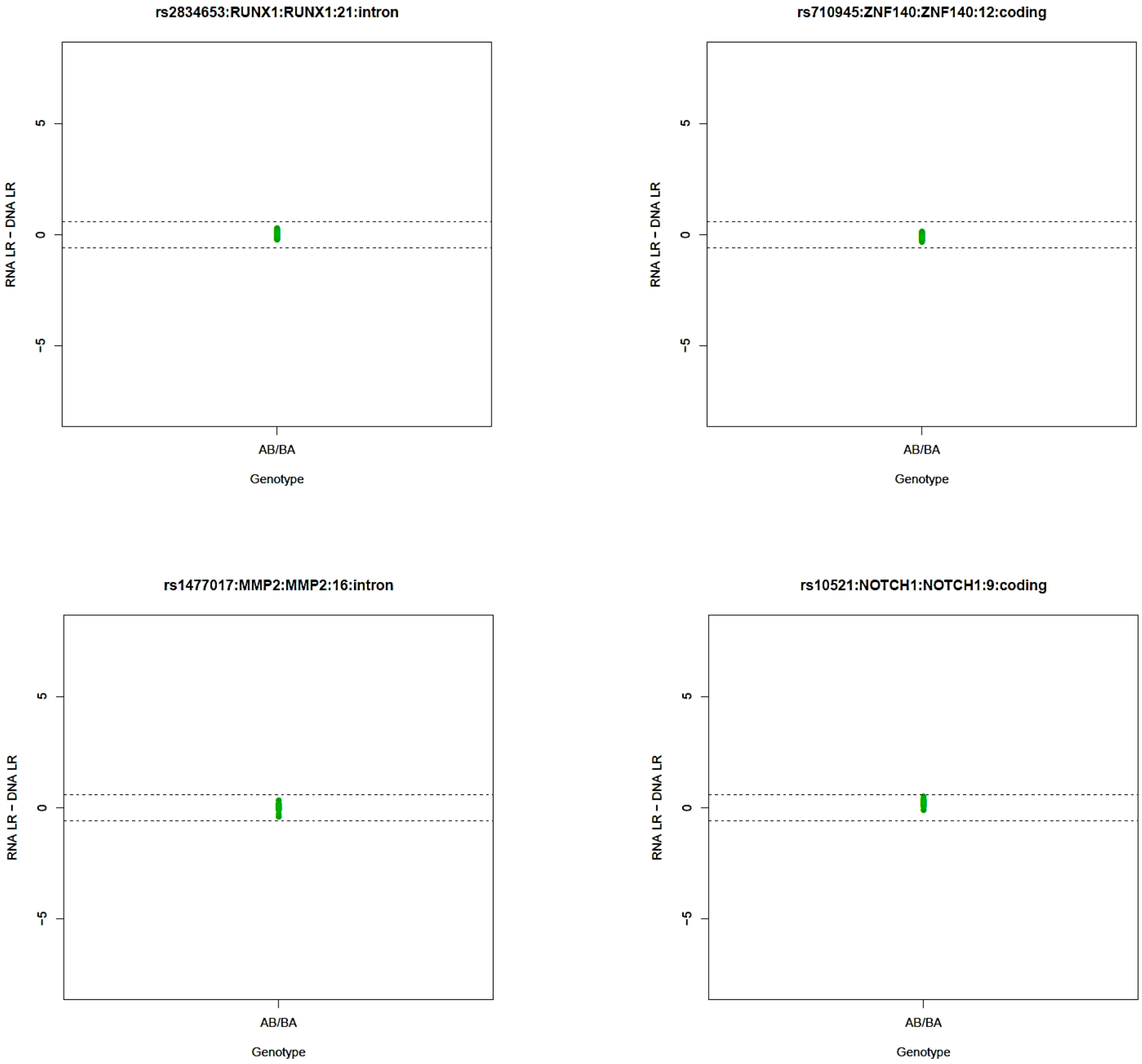
rs13265801: *WDYHV1*: Intron

rs9250: *SENP6*: Coding

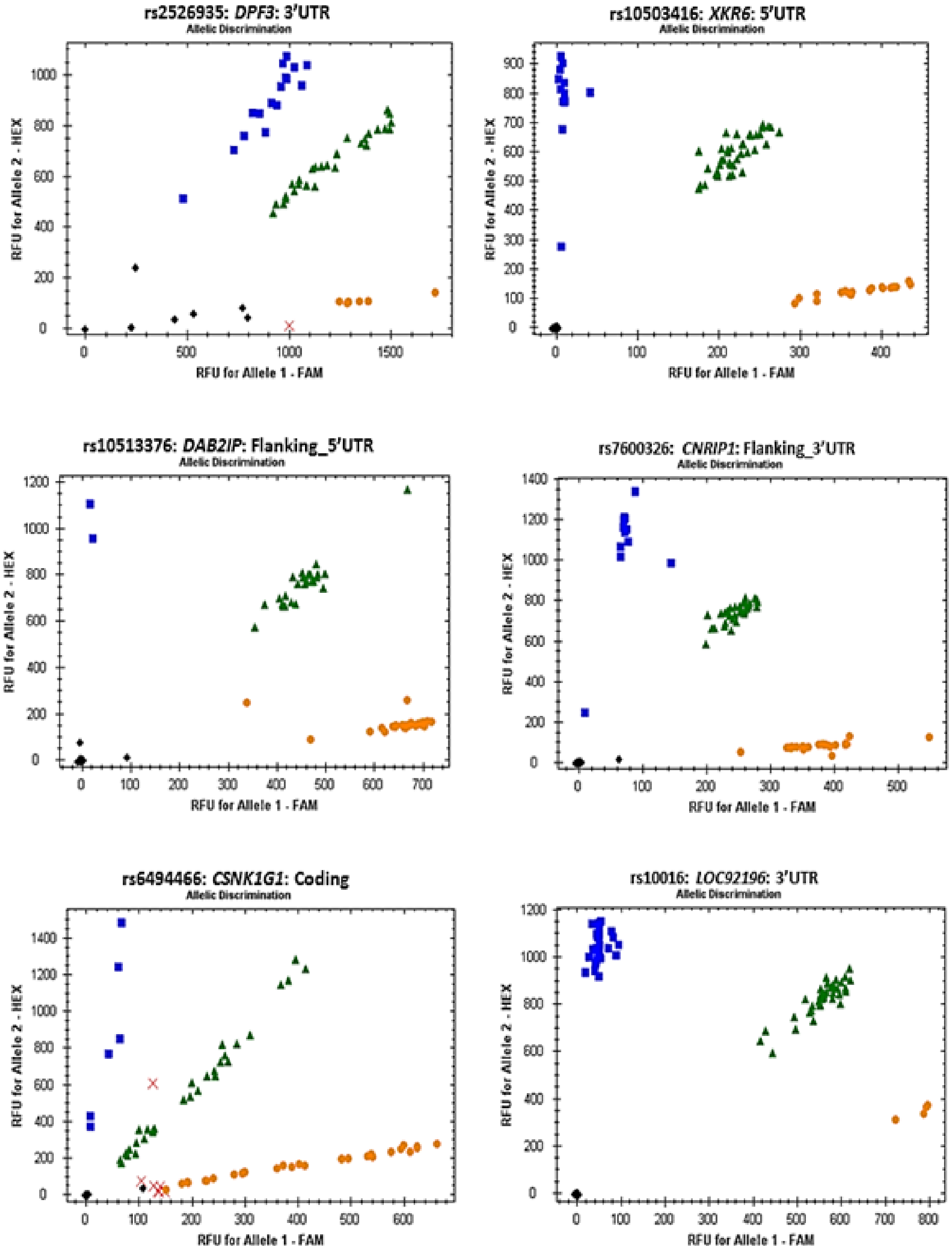
rs1384: *LYZ*: 3'UTR

rs8097892: *MPPE1*: Intron

Annex 2.2 Non DAE SNPs reported in previous results obtained in microarray (Maia et al, unpublished). These are the 4 DAE SNPs that we chose to validate. The x-axis indicates the genotype (A/B-heterozygous) and the y-axis indicates the normalised DAE ratio obtained. Dotted lines delimit the cut-off of preferential allelic expression ratio [$\log_2(1.5)=0.584$].



Annex 2.3 Genotyping of DNA from normal breast tissue samples by Taqman qRT-PCR. 10 DAE SNPs and 4 non DAE SNPs chose to validate. x-axis indicates the fluorescence intensity of Allele 1 emitted by probe FAM and the y axis indicates the fluorescence intensity of Allele 2 emitted by the probe HEX. The blue squares represent homozygous samples for Allele 2, orange circles represent samples homozygous for Allele 1 and the green triangles represent heterozygous samples. The black diamonds are representative of NTCs (no fluorescent signal) and red crosses are undeterminate samples.



rs2526935: *DPF3*: 3'UTR

rs10503416: *XKR6*: 5'UTR

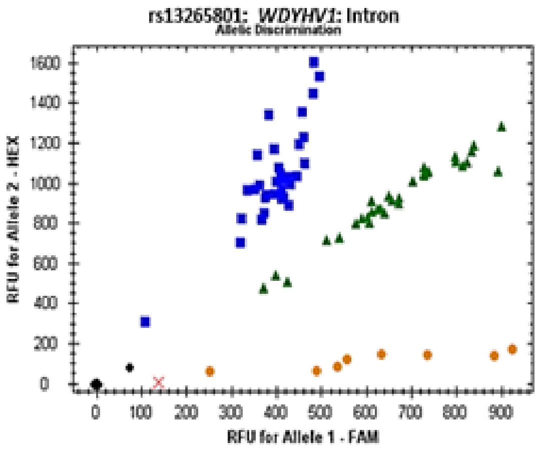
rs10513376: *DAB2IP*: Flanking_5'UTR

rs7600326: *CNRIP1*: Flanking_3'UTR

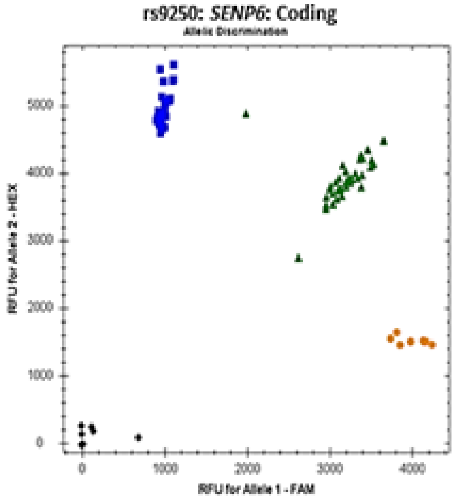
rs6494466: *CSNK1G1*: Coding

rs10016: *LOC92196*: 3'UTR

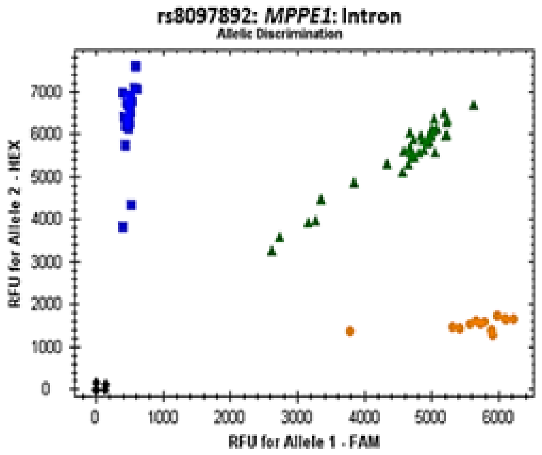
Annex 2.3 (Continuation of results of previous table – Genotyping of DNA samples from normal breast tissue by Taqman qRT-PCR).


rs13265801: *WDYHV1*: Intron

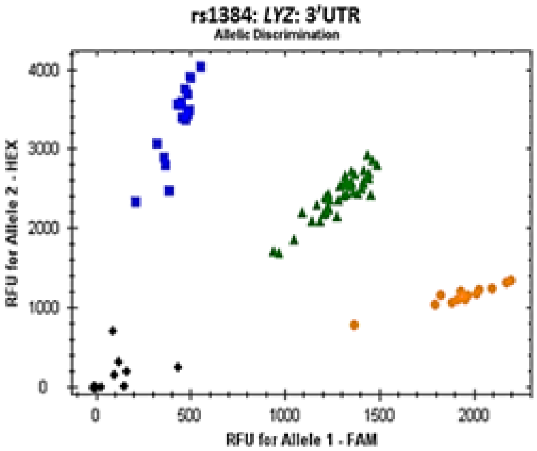
Allelic Discrimination


rs9250: *SENP6*: Coding

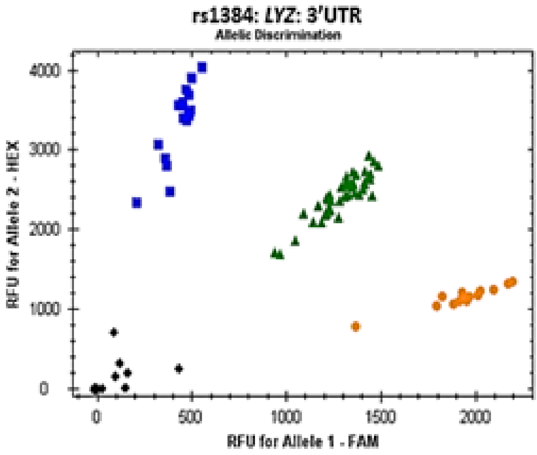
Allelic Discrimination


rs8097892: *MPPE1*: Intron

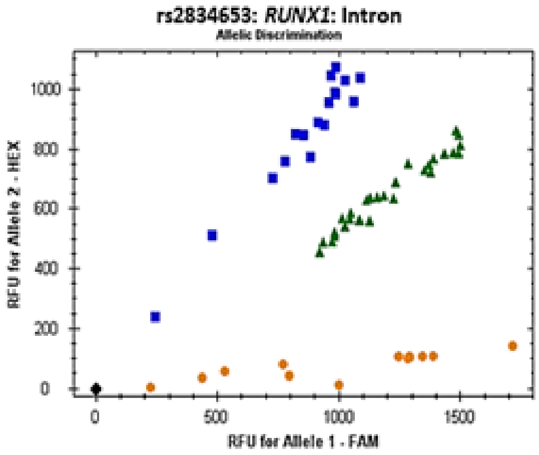
Allelic Discrimination


rs1384: *LYZ*: 3'UTR

Allelic Discrimination


rs1384: *LYZ*: 3'UTR

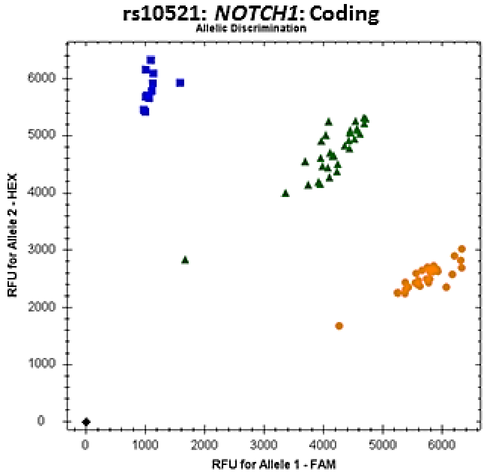
Allelic Discrimination


rs2834653: *RUNX1*: Intron

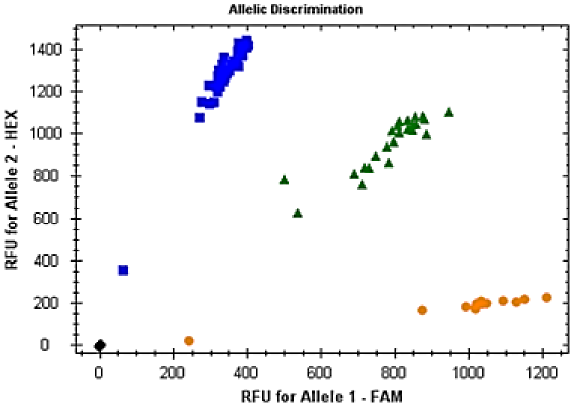
Allelic Discrimination

Annex 2.3 (continuing the above results – Genotyping of DNA samples from normal breast tissue by Taqman qRT-PCR).

rs10521: *NOTCH1*: Coding

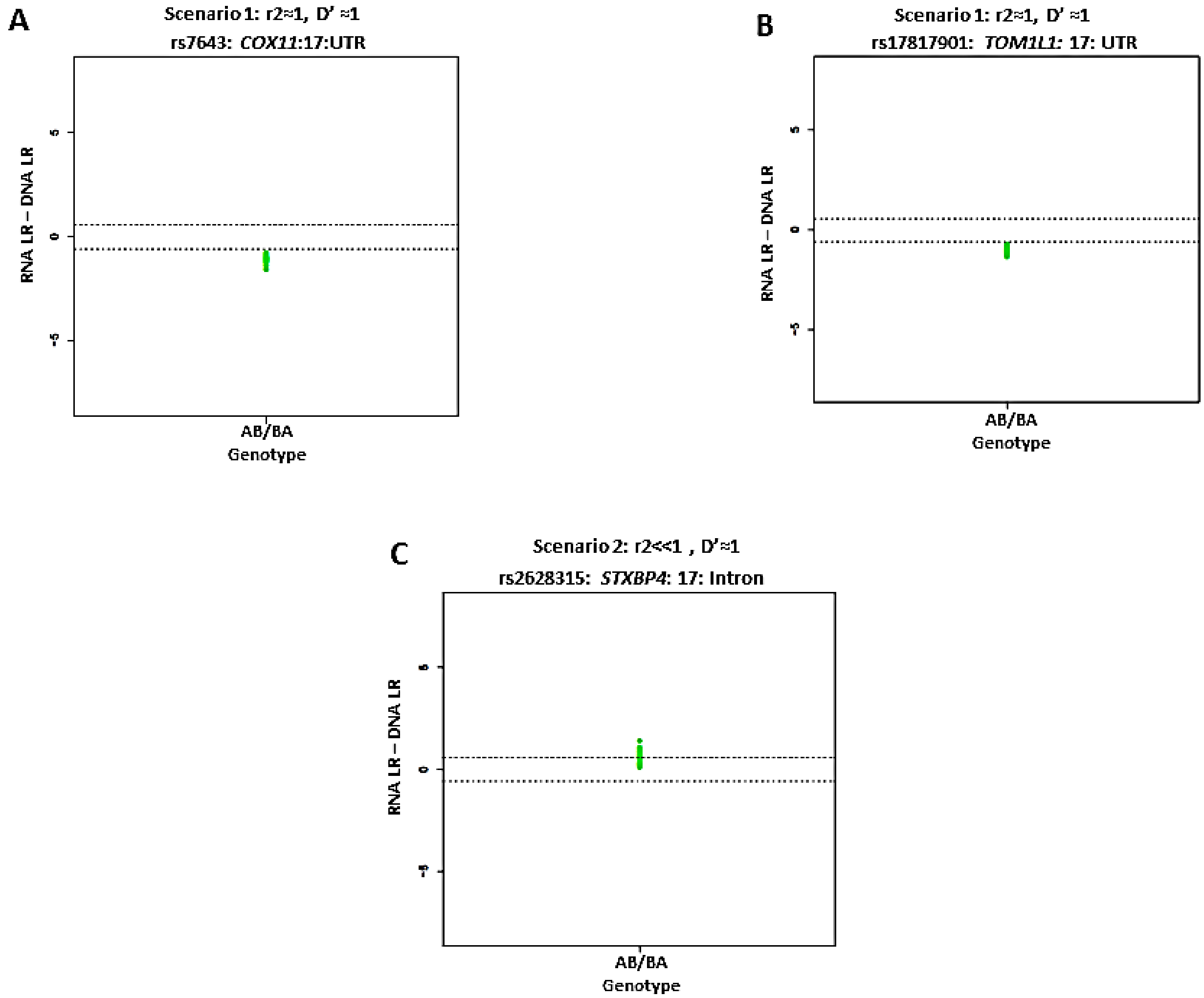


rs1477017: *MMP2*: Intron

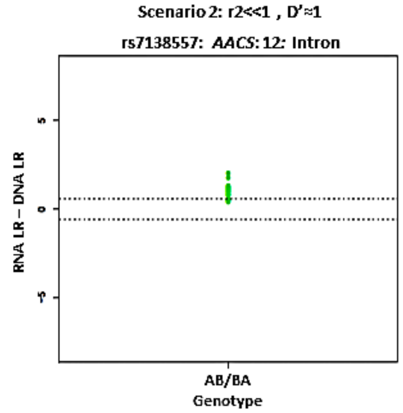


ANNEX 3

Annex 3.1 DAE scenarios observed for the 3 marker SNPs (*COX11*). The x-axis indicates the genotype (A/B-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].Scenario represent the linkage disequilibrium between the cSNP and rSNP. r^2 and D' are measures of linkage.

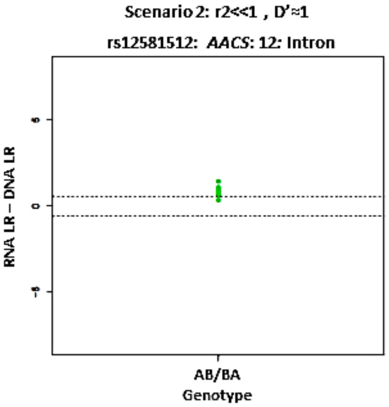


Annex 3.2 DAE scenarios observed for the 3 marker SNPs (*AACS*). The x-axis indicates the genotype (A/B-heterozygous) and the y-axis indicates the normalised DAE ratio obtained. Dotted lines delimit the cut-off of preferential allelic expression ratio [log2(1.5)=0.584].Scenario represent the linkage disequilibrium between the cSNP and rSNP. r^2 and D' represent the measures of linkage.



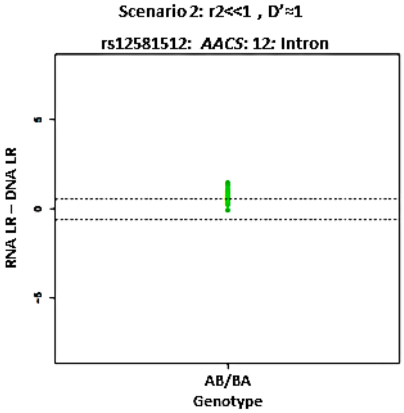
Scenario 2: r2<<1 , D'≈1

rs7138557: *AACS*: 12: Intron



Scenario 2: r2<<1 , D'≈1

rs12581512: *AACS*: 12: Intron

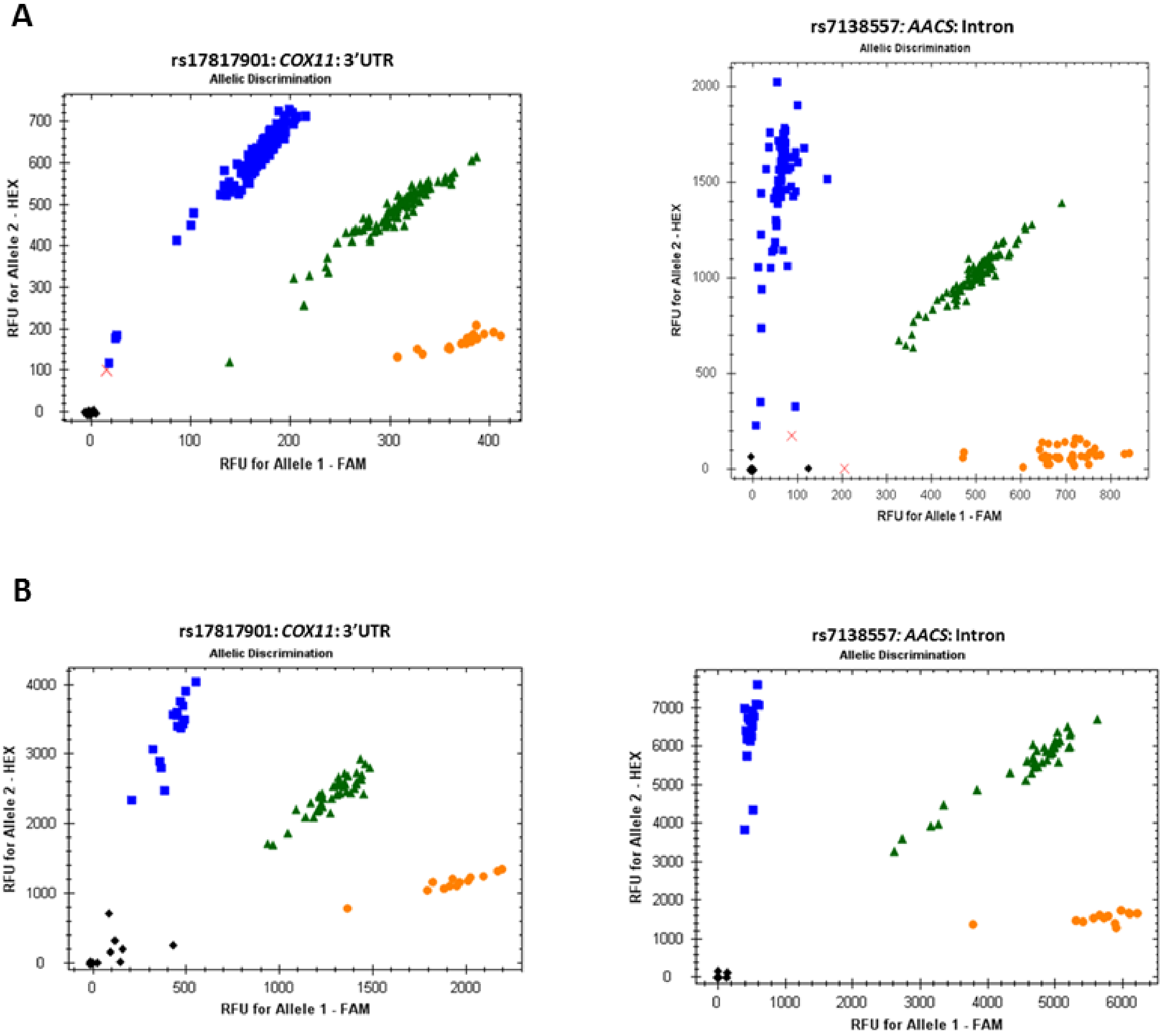


Scenario 2: r2<<1 , D'≈1

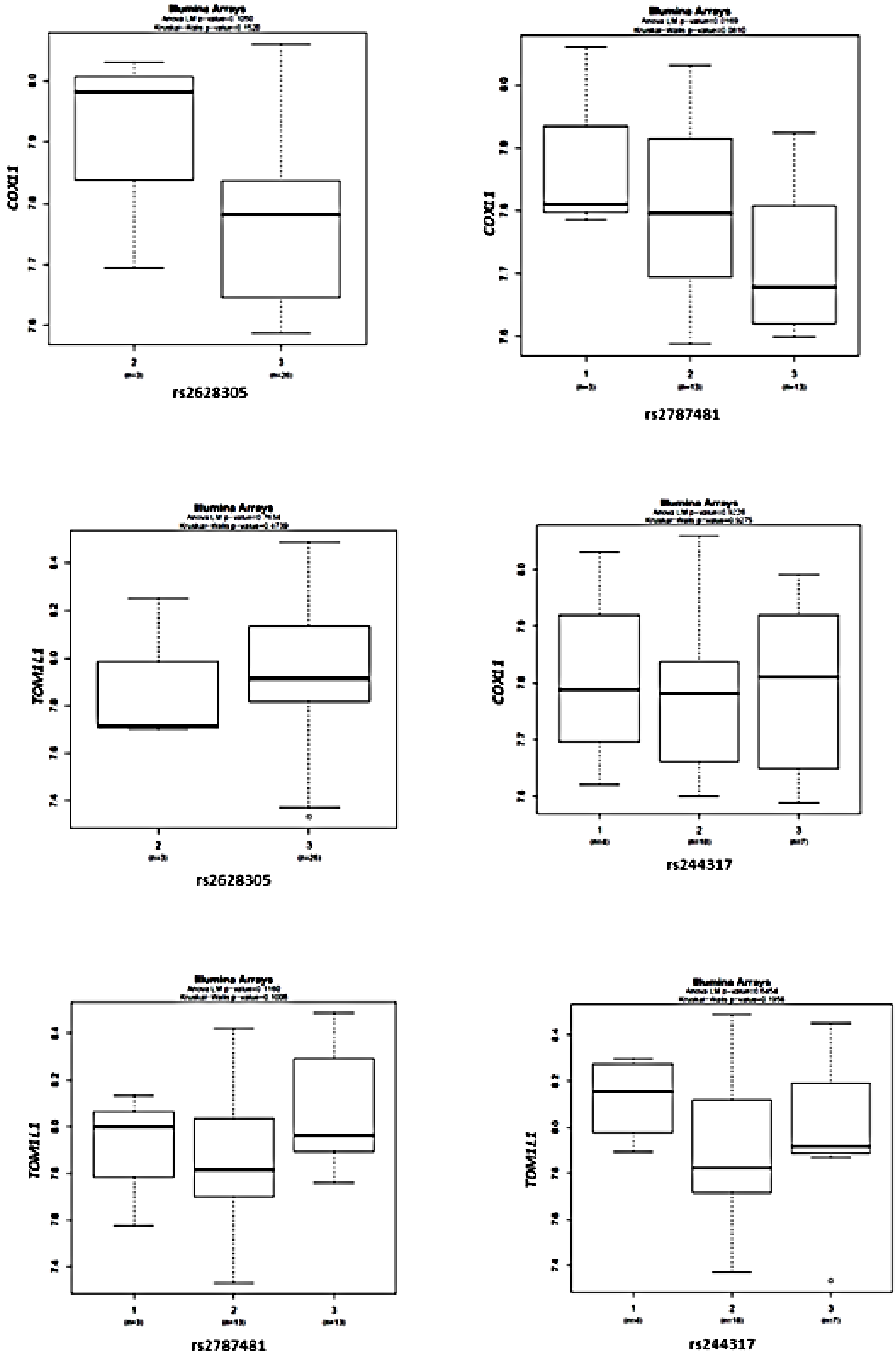
rs12581512: *AACS*: 12: Intron

ANNEX 4

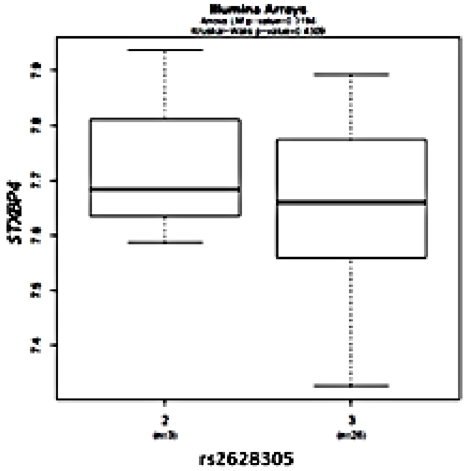
Annex 4.1 Genotyping of DNA from normal breast tissue (A) and blood samples (B) by Taqman qRT-PCR. Results of genotyping for the 2 DAE SNPs chose to validate in *COX11* and *AACS*.

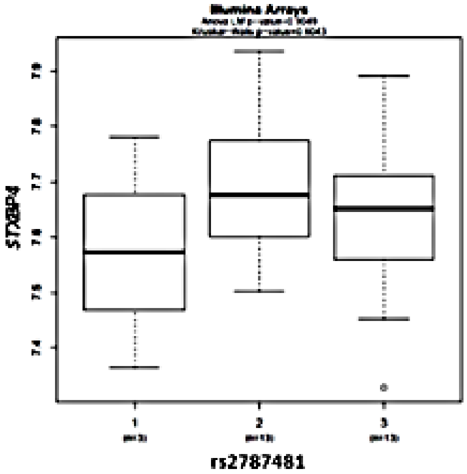


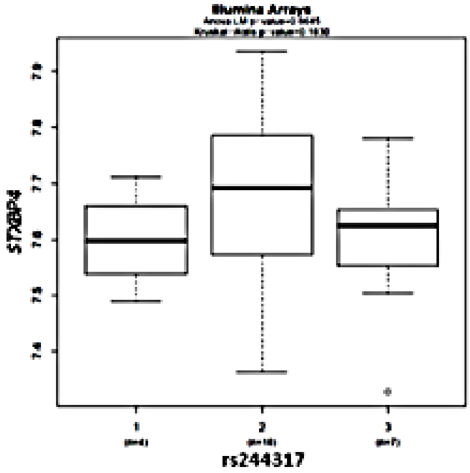
Annex 4.2 Correlation between *COX11, TOM1L1 and STXBP4* expression in normal breast and blood with the genotype of a SNP. The y-axis indicates the expression total level of expression of the genes and the x-axis indicate the genotype of a SNP.



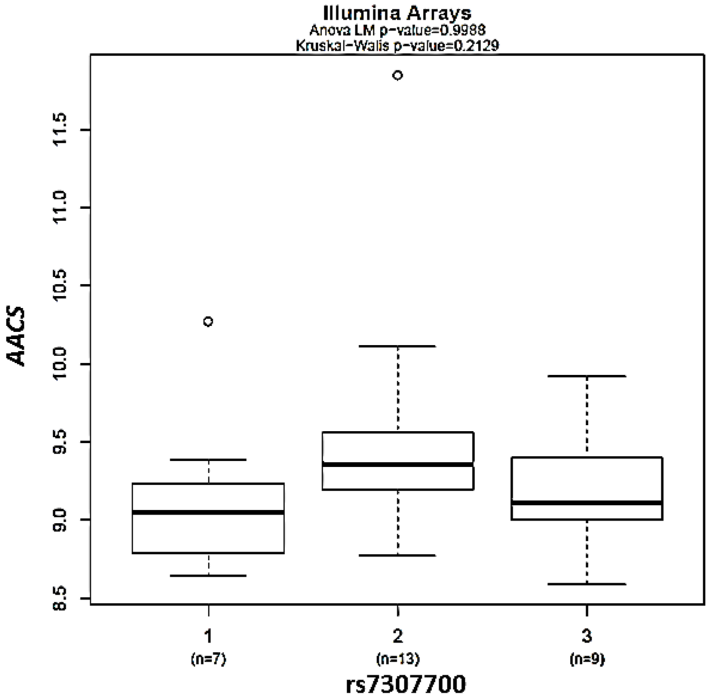
Annex 4.2 (continuing the above results – Correlation between *COX11, TOM1L1 and STXBP4* expression in normal breast and blood with the genotype of a SNP).



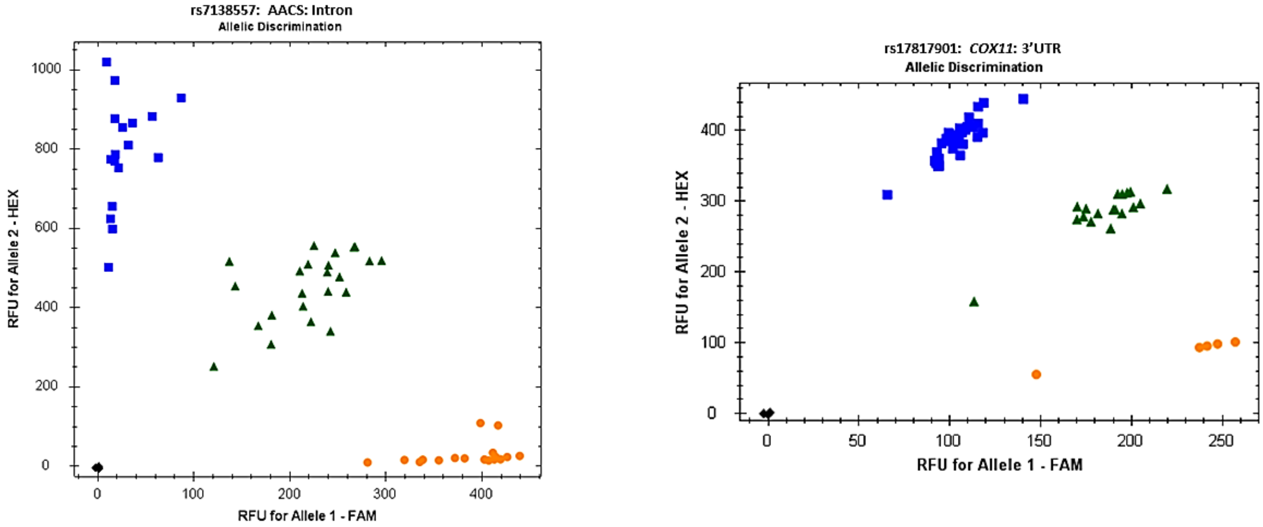




Annex 4.3 Correlation between *AACS* expression in normal breast and blood with the genotype of a SNP. The y-axis indicates the expression total level of expression of the genes and the x-axis indicate the genotype of a SNP.



Annex 4.4 Genotyping of DNA from patient cancer samples by Taqman qRT-PCR. Results of genotyping for the 2 DAE SNPs chose to validate in *COX11* and *AACS*.



rs7138557: AACS: Intron
Allelic Discrimination

rs17817901: *COX11*: 3'UTR
Allelic Discrimination