UNIVERSIDADE DO ALGARVE

DEPARTAMENTO DE CIÊNCIAS BIOMÉDICAS E MEDICINA

Study of Influence of Regulatory Polymorphisms of Expression in Development of Breast Cancer

Joana Margarida Teixeira Lopes

Tese

Mestrado em Ciências Biomédicas

Supervisor: Professora Doutora Ana Teresa Maia

Co-Supervisor: Doutora Natércia Conceição

2013

Study of Influence of Regulatory Polymorphisms of Expression in Development of Breast Cancer

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

No decorrer de mais uma etapa na vida académica, foram muitas as pessoas que me apoiaram e incentivaram. Por essa razão, desejo expressar os meus sinceros agradecimentos:

À Professora Doutora Ana Teresa Maia, orientadora desta tese, por me ter concedido a oportunidade de participar neste projecto. Agradeço o apoio, a partilha dos conhecimentos que muito contribuíram para a realização deste projecto e pelo sentido de responsabilidade que me incutiu em todas as fases desta tese.

À Doutora Natércia Conceição que sempre me incentivou a ir mais longe. Muito obrigado pela amizade, pelo profissionalismo e pela total disponibilidade demonstrada ao longo deste projecto.

À Professora Doutora Leonor Cancela pela ajuda e disponibilidade que sempre demonstrou.

Ao grupo EDGE, em especial à Iris e à Cindy por toda a ajuda, paciência e disponibilidade demonstrada.

A todos os professores, funcionários e colegas do DCBM e CBME pelo apoio prestado.

A todos os meus amigos que sempre me aconselharam, apoiaram e incentivaram para nunca desistir.

Por último e não menos importante à minha família, em especial aos meus pais pelo seu apoio incondicional, amizade e incentivo demonstrado. Obrigado por serem modelos de coragem e pela total ajuda na superação dos obstáculos que foram surgindo pois sozinha nada disto teria sido possível. Obrigado pelo amor incondicional. A eles, dedico todo este trabalho.

Obrigado a todos.

ABSTRACT

The human genome has millions of genetics variants that can affect gene expression. These variants are known as cis-regulatory variants and are responsible for intra-species phenotypic differences and individual susceptibility to disease. One of the diseases affected by cis-regulatory variants is breast cancer. Breast cancer is one of the most common cancers, with approximately 4500 new cases each year in Portugal. Breast cancer has many genes mutated and *TP53* has been shown to be relevant for this disease. *TP53* is one of the most commonly mutated genes in human cancer and it is involved in cell cycle regulation and apoptosis. Previous work by Maia et al has shown that *TP53* has differential allelic expression (DAE), which suggests that this gene may be under the influence of cis-regulatory variants. Also, its DAE pattern is totally altered in breast tumours with normal copy number. We hypothesized that cis-regulatory variants affecting *TP53* may have a role in breast cancer development and treatment.

The present work aims to identify the cis-regulatory variants playing a role in *TP53* expression, using in silico, in vitro and in vivo approaches. By bioinformatic tools we have identified candidate cis-regulatory variants and predicted the possible transcription factor binding sites that they affect. By EMSA we studied DNA-protein interactions in this region of *TP53*.

The *in silico* analysis allowed us to identified three candidate cis-regulatory SNPs which may affect the binding of seven transcription factors. However, the EMSA experiments have not been conclusive and we have not yet confirmed whether any of the identified SNPs are associated with gene expression control of TP53. We will carry out further experiments to validate our findings.

RESUMO

As células normalmente crescem, dividem-se e morrem de uma forma controlada mas ao longo da vida surgem alguns danos no ADN. De forma a manter a função normal do organismo estes danos são reparados ou a célula entra em apoptose. Quando estes mecanismos falham pode ocorrer acumulação de mutações, o que leva à formação do cancro.

O cancro não é uma doença única e homogénea, há sim uma grande variedade de cancros. O cancro da mama é um dos cancros mais comuns e em 2012, foi o quarto cancro com maior incidência em Portugal. Em cada ano existem 4500 novos casos e 1500 mulheres morrem com esta doença. No mundo, o cancro da mama representa cerca de 16% de todos os cancros femininos e 22.9% dos cancros invasivos na mulher. Esta doença é caraterizada por alterações nas células da mama que se tornam anormais e multiplicam-se descontroladamente levando à formação de um tumor. Estas alterações podem ser devido a fatores de risco, embora estes não sejam ainda bem compreendidos. Estes fatores incluem: idade, sexo, genética, história familiar, hormonas, contraceptivos, tumores benignos, entre outros.

Existem muitos genes cujas mutações contribuem para a susceptibilidade do cancro da mama. Os genes frequentemente mutados são: *BRCA1*, *BRCA2*, *ATM*, *TP53*, *CHECK2*, *PTEN*, *CDH1*, *STK11*, entre outros. Mutações da linha germinal raras nos genes *BRCA1* e *BRCA2* estão associadas a um alto risco e são responsáveis por aproximadamente 20% dos casos familiares. Mutações da linha germinal no gene *TP53* são mais raras mas têm um risco muito elevado. Mutações somáticas no gene *TP53* são encontradas também no cancro da mama em cerca de 15-20% dos tumores.

O *TP53* é um gene supressor de tumores que codifica uma proteína designada p53. A proteína p53 é um fator de transcrição e responde a vários stresses celulares através da regulação da expressão dos seus genes alvo. Esta proteína pode induzir paragem do ciclo celular, apoptose, senescência, reparação do DNA, ou alterações no metabolismo. O gene *TP53* é um dos genes mais frequentemente mutado nos cancros. É um importante regulador homeostático, atuando na reparação do DNA, no controlo negativo do crescimento, em diversas vias de regulação do ciclo celular e apoptose. O *TP53* tem muitos transcritos e

isoformas que resultam do splicing alternativo e da utilização de promotores e/ou codões de iniciação da tradução alternados. Este gene localiza-se no braço pequeno do cromossoma 17 na posição 13.1 e é constituído por 11 exões. Frequentemente as mutações somáticas no *TP53* são acompanhadas por perda de heterozigotia, indicando que o alelo normal residual é perdido em tumores.

Milhões de variantes genéticas ocorrem no genoma humano e são designadas por polimorfismos. Estas variantes genéticas representam aproximadamente 1% do genoma. Existem algumas variantes que levam a numerosas diferenças fenotípicas e são responsáveis pela variabilidade intra-espécies. Os polimorfismos mais frequentes são os polimorfismos de um nucleótido (SNP) e constituem cerca de 90% de todas as variações conhecidas. Os SNPs localizados na região codificante de um gene podem levar a alterações de aminoácidos. Isto pode induzir a mudança de polaridade da proteína, alteração da configuração secundária e terciária da proteína, fosforilação inadequada, entre outras consequências funcionais. Os SNPs localizados na região não codificante são na maior parte considerados não funcionais. Todavia, este tipo de alterações pode afetar os elementos reguladores dos genes tais como promotores, amplificadores ou silenciadores, que podem estar próximo do gene, mas podem também estar à distância de centenas de kilobases do gene que regulam. Estes elementos reguladores podem coletivamente modular o tempo, magnitude e especificidade celular da expressão de um gene.

SNPs cis-regulatórios são as variantes genéticas que afectam a expressão de um gene. Podem atuar através de diferentes mecanismos afetando por exemplo os sítios de ligação dos factores de transcrição (TFBS), modicações de histonas e metilação do DNA. O mecanismo mais comum é a modificação dos TFBSs.

Uma área importante da pesquisa médica é a compreensão do papel dos SNPs cis-regulatórios na doença. Como os SNPs cis-regulatórios originam um desiquilíbrio na expressão dos alelos de um gene, a análise da expressão alélica diferencial (DAE) é o método de excelência para estudá-los. Para estudar DAE compara-se a expressão relativa de dois alelos num indivíduo heterozigótico. A DAE é responsável pela variabilidade intra-espécie e pode levar ao aparecimento da susceptibilidade para doenças complexas. A variação nos elementos cis-regulatórios é comum no genoma humano e a DAE foi estimado a afetar 20-50% dos genes, dependendo do método usado. Para identificar DAE, recentes estudos de genómicos de associação (GWAS) têm se focado na interpretação de SNPs codificantes ou

outros SNPs em regiões transcritas. O nosso grupo está principalmente interessado em estudar DAE na susceptibilidade do cancro da mama. Anteriormente, foi observado que o *TP53* apresenta DAE e todos os indivíduos heterozigóticos expressam mais do mesmo alelo, no sangue e no tecido de mama. Isto indica que o SNP utilizado neste estudo está em completo linkage disequilibrium (LD) com a(s) variante(s) cis-regulatória(s). Como o sangue e o tecido de mama normais expressam o mesmo alelo, nós podemos concluir que o *TP53* tem variantes cis-regulatórias.

O objetivo deste estudo foi o de investigar se SNPs cis-regulatórios do *TP53* têm algum efeito no cancro da mama. Para isso, nós propusemos mapear os SNPs cis-regulatórios que afetam a expressão do gene *TP53* e prever TFBSs nesta região.

Inicialmente, selecionámos o SNP (rs1042522) que previamente demonstrou expressão alélica diferencial no sangue e tecido de mama normais. As nossas análises *in silico* mostraram que rs1042522 está em completo LD com outro SNP (rs1642785). Isto levou-nos a investigar a região entre estes dois SNPs para modificações de histonas, zonas de hipersensibilidade à DNaseI e sítios de ligação de factores de transcrição. Descobrimos que neste intervalo há uma região muito ativa, com a possibilidade de existência de elementos reguladores. Outros quatro SNPs sobrepõem-se a este elementos e são candidatos para ser SNPs cis-regulatórios. Então analisámos os TFBSs para seis SNPs cis-regulatórios. Cada um destes SNPs tem dois alelos diferentes que são analisados individualmente para diferentes ligações de factores de transcrição (TF). Desses, três foram identificados como possíveis TFBSs que são expressos em tecido de mama, com diferenças para os dois alelos correspondentes.

Para estudar se na sequência à volta destes SNPs cis-regulatórios candidatos pode realmente ligar-se algum TF, analisámos interações proteína-DNA através da técnica EMSA, usando oligonucleótidos contendo os SNPs de interesse. Realizámos várias experiências mas não detetámos nenhuma alteração, correspondendo a uma interação proteína-DNA positiva. No entanto, admitimos que o nosso extrato nuclear possa não estar funcional, dado que controlo positivo (FGFR2-13) usado para avaliar a qualidade dos extratos não mostrou uma alteração como era esperado. Existem duas possíveis explicações: (1) não existir realmente interação entre os TFs e os nossos oligonucleótidos, assim sendo estes SNPs não são cis-regulatórios, e (2) o TF pode ligar aos nossos oligonucleótidos mas os extratos estão em boas condições. Iremos realizar mais experiências para resolver esta questão. Se a segunda

hipótese for confirmada, vamos confirmar se estes SNPs são ocupados *in vivo* pelos TFs respectivos através da técnica de imunoprecipitação da cromatina (ChIP).

O objectivo central deste trabalho era estudar as variações cis-regulatórias do gene *TP53* em cancro da mama. Não nos foi possível ainda confirmar se os SNPs existentes na região candidata são realmente cis-regulatórios e regulam a expressão do *TP53* originando DAE. Como tal prosseguiremos os nossos estudos para clarificar a sua existência e função. Para confirmar o sue envolvimento em cancro da mama iremos também comparar os resultados finais entre tecido da mama normal e tumoral.

TABLE OF CONTENTS

Agradecimentos **3**

Abstract **4**

Resumo **5**

Table of Contents **9**

Table of Tables **8**

Table of Figures **9**

Abreviations **10**

1. Introduction **13**

1.1 Cancer and Breast cancer 15

1.2 *TP53* gene 17

1.3 Cis-regulatory SNPs 19

1.4 Differential Allelic Expression 20

1.5 Aims 23

2 Materials and Methods **24**

2.1 Cell lines and cell culture 24

2.2 Nucleic acid extraction and processing 25

2.3 Quantification of DNA and RNA 27

2.4 SNPs and haplotype analyses 27

2.5 *in silico* prediction of transcription factor binding sites 27

2.6 Mobility shift DNA – binding assay 28

2.6.1 General description of the assay 28

2.6.2 Preparation of Nuclear Extracts 28

2.6.3 Determination of total protein concentration 29

2.6.4 Preparation of DNA probes 29

2.6.5 Identification of DNA-protein binding reaction 31

2.7 Genotyping 32

3 Results **34**

3.1 *in silico* analysis 34

3.2 Transcription factor binding sites results 37

3.3 *in vitro* analysis 38
3.3.1 Identification of DNA-protein binding..... 38
3.4 Genotyping analysis..... 41
Discussion..... 43
References..... 45

TABLE OF TABLES

Table 2.1 – Primers designed for EMSA. 30
Table 2.2- Binding reaction used for each SNP 31
Table 2.3- Primers for the region around the selected SNPs. 32
Table 3.1 - SNPs selected for analysis of TFBS. 38
Table 3.2 – Analysis of genotype from different cell lines.. 41

TABLE OF FIGURES

Figure 1.1 - Incidence of cancers in Portugal in 2012. 16
Figure 1.2 - *TP53* gene structure and the different transcript variants.. 19
Figure 1.3 - Differential allelic expression analysis. 21
Figure 1.4 - Comparison of DAE in blood and breast tissue in different genes.. 21
Figure 1.5 - DAE analysis of *TP53* in control, normal breast and tumours.. 22
Figure 3.1- LD plot for the region around *TP53*.. 35
Figure 3.2 - Haplotype blocks and haplotype frequencies in *TP53*..... 36
Figure 3.3- Diagram of the *TP53* gene..... 37
Figure 3.4 - Level of labelling from all primers and Biotin-EBNA Control DNA..... 39
Figure 3.5 - *in vitro* DNA-protein binding studies.. 40
Figure 3.6 - Results of sequencing of heterozygous SNPs in GM12878 cell line.. 42

ABREVIATIONS

A – Adenine

C – Cytosine

cDNA – Complementary deoxyribonucleic acid

CEPH – Centre d'Étude du Polymorphisme Humain

ChIP – Chromatin immunoprecipitation

CO_2 – Carbon dioxide

DAE – Differential allelic expression

DMEM – Dulbecco's modified Eagle medium

DMSO – Dimethyl sulfoxide

DNA – Deoxyribonucleic acid

DNase – Desoxirribonuclease

DTT - Dithiothreitol

ER+ - Oestrogen receptor-positive

EDTA – Ethylenediaminetetraacetic acid

EMSA – Electrophoretic mobility shift assay

FBS – Foetal bovine serum

G – Guanine

GWAS – Genome-wide association studies

KCl – Potassium chloride

LD – Linkage Disequilibrium

LFS – Li-Fraumeni syndrome

M – Molar

mM – Milimolar

MgCl – Magnesium chloride

mRNA – Messenger ribonucleic acid

NaCl – Sodium chloride

ng – Nanograme

nm - Nanometres

NP40 – Nonyl phenoxypolyethoxylethanol

N_2 - Nitrogen

p – Short arm of the chromosome

PBS – Phosphate buffered saline

PCR – Polymerase chain reaction

PI – Protease inhibitor

RNA – Ribonucleic acid

RNase – Ribonuclease

rpm – Rotations per minute

RPMI – Roswell Park Memorial Institute

SNP – Single nucleotide polymorphism

T - Thymine

TAE – Tris-acetate-EDTA buffer

TBE –Tris-borate-EDTA buffer

TE – Tris-EDTA buffer

TF – Transcription factor

TFBS – Transcription factor binding site

V - Volts

1. INTRODUCTION

1.1 Cancer and Breast cancer

Normal cells grow, divide and die controllably. During a person's lifetime some damage occurs in the cell. In order to keep the normal function of the organism this damage can either be repaired or the cell can enter apoptosis. When these mechanisms fail, there can be an abnormal growth and/or division of the cells with further accumulation of mutations that can lead to cancer formation (Weinberg, 2007; Hanahan et al., 2011).

Cancer is not a single disease, there are in fact many different kinds of cancers. Most cancers are designated according to the organ in which they start, such as lung, breast, colon, skin, among others, and/or type of cells in which they start; for example epithelial cells originate carcinomas, tumours of the connective tissues are called sarcomas and hematopoietic cells generate lymphomas or leukaemias. All cancers can be classified as in situ/non-invasive or invasive. In situ/non-invasive means that cancer has not yet invaded other tissues of the affected organ. Invasive means it has spread to other tissues of the affected organ (Strachan et al., 2010). Cancer cells can also spread to others parts of the body through the blood or lymph systems and settle in one or several organs. Over time, the cancer cells take the place of normal cells and give rise to metastases (Levin, 1913; Weinberg, 2007; Scully et al., 2012).

Breast cancer is one of the most common cancers. In 2012, breast cancer was the fourth cancer with the highest incidence in Portugal (Figure 1.1). Each year there are 4500 new cases and 1500 women die of this disease. In the world, breast cancer accounts for 16% of all female cancers and 22.9% of invasive cancers in women (Jemal et al., 2010; http://www.eu-cancer.iarc.fr/EUCAN/).

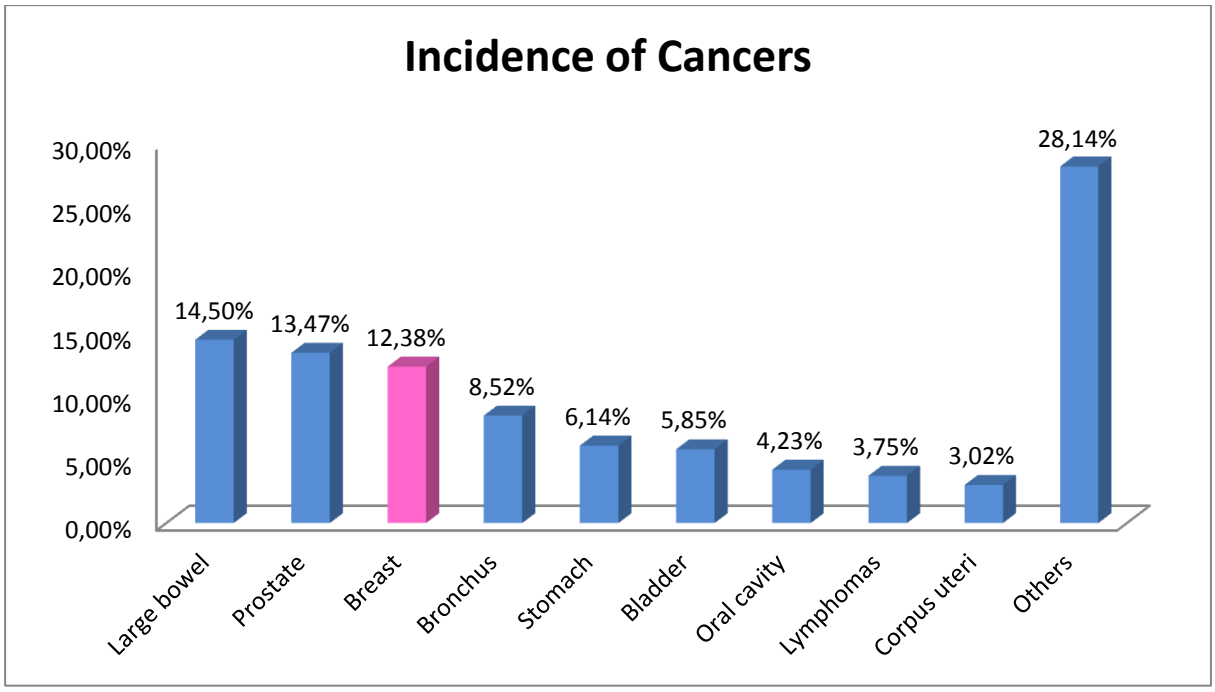

Figure 1.1 - Incidence of cancers in Portugal in 2012 (http://www.eu-cancer.iarc.fr/EUCAN/).

This disease is characterized by alterations in breast cells, which become abnormal and multiply uncontrollably to form a tumour. These alterations can be due to risk factors, although not all are yet well understood. These factors include: age, sex, genetics, familial history, hormones, contraceptive pill, age of motherhood or childless, age of menarche and menopause, ethnic group, benign tumour, dense breast tissue, among others (Kelsey et al., 1988).

Breast cancer can begin in different areas of the breast, originating two mains types of breast cancer: ductal carcinoma and lobular carcinoma. Ductal carcinoma is the most common type of breast cancer and originates in the milk ducts. Lobular carcinoma is the second most common type of breast cancer and initiates in the lobules that produce milk and empty out into the ducts (Russnes et al., 2011). Breast cancer can also metastasize to many tissues throughout the body, but metastases often occur in bones, brain, liver and lungs.

Most breast cancers can be sensitive to oestrogen because they express the oestrogen receptor on the surface of their cells, these are called oestrogen receptor-positive (ER+) cancers. This indicates that oestrogen lead to growth of breast tumour (Yerushalmi et al., 2009).

Germline mutations may occur *de novo* or be inherited from parent's germ cells – eggs or sperm – and can be transmitted to the next generation. Somatic mutations can occur in any cells in the body and the genes involved are usually located in autosomal chromosomes. These mutations are not passed along to the next generation. There are many genes whose germline mutations can contribute to breast cancer susceptibility (Weinberg, 2007). These include: *BRCA1*, *BRCA2*, *ATM*, *TP53*, *CHECK2*, *PTEN*, *CDH1*, *STK11*, among others. Rare germline mutations in *BRCA1* and *BRCA2* are associated with very high risk and are responsible for approximately 20% of familial cases. Germline mutations in *TP53* are even more rare but have the highest risk (approximately 20-fold). Somatic mutations in *TP53* are also found in breast cancer in about 15-20% of tumours (Hindorff et al., 2011).

1.2 *TP53* gene

TP53 is a tumour suppressor gene that encodes a protein called p53. The p53 protein responds to diverse cellular stresses to regulate the expression of target genes. This protein induces cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. The activity of p53 is regulated by post-transcriptional, multiple transcriptional and translational mechanisms in response to a broad range of biological and physical stresses. This protein has consequently an important role in preventing DNA replication and cell division in conditions that damage genetic integrity (Liang et al., 2013).

The *TP53* gene is one of the most common mutated genes in human cancers. This gene is an important homeostatic regulator, acting over the DNA repair, negative growth control, multiple pathways of cell cycle regulation and apoptosis. *TP53* has multiple transcript variants and isoforms that result from alternative splicing and the use of alternate translation initiation codon and/or promoters. This gene is located on the short (p) arm of chromosome 17 at position 13.1 and consist of 11 exons (Smith et al., 2011) (Figure 1.2).

Germline mutations in *TP53* occur in familial cases with Li-Fraumeni syndrome (LFS), which confer an increased risk of developing various cancers comprising mostly breast cancer, sarcoma, leukaemia and brain tumours. Breast cancer accounts for about 25% of all tumours in LFS patients and this syndrome can occur at any point in an individual's lifetime, including childhood (Costa et al., 2008; Walerych et al., 2012).

TP53 somatic mutations are present in approximately 23% of breast cancers and this is second most mutated gene, where the *PI3KCA* proto-oncogene is the first most mutated gene. Usually somatic mutations in *TP53* are accompanied by Loss of Heterozygosity (LOH) indicating that the remaining *wild-type* allele is also lost in the tumours (Strachan et al, 2010).

In breast cancer, the most investigated polymorphisms in *TP53* are the Arg72Pro (a G to C transversion in codon 72 of exon 4), and the PIN3 Ins16bp (a 16 base pair duplication in intron 3). The *TP53* Arg72Pro (rs1042522) results in an amino acid change from arginine to proline and is located in a proline-rich region of the protein, which has been known to be important for growth suppression and apoptotic functions. This polymorphism results in a structural change in the protein, giving rise to two isoforms of p53 that differ in biochemical and biological properties. However, several studies have indicated different functions for the two isoforms and these contradictory results have demonstrated that both *TP53* Arg72Pro variants can differently regulate specific cellular functions (Dumont et al., 2003; Pim et al., 2004; Ohayon et al., 2005). The PIN3 Ins16bp (rs17878362) has been reported to affect mRNA splicing, altering the coding regions and is implicated in the regulation of gene expression and DNA-protein interactions, which can result in a defective protein. Nevertheless, the biological effects of this polymorphism are not yet well understood. The *TP53* Arg72Pro and PIN3 Ins16bp polymorphisms increase breast cancer in familial and sporadic cases (Costa et al., 2008; Guleria et al., 2012).

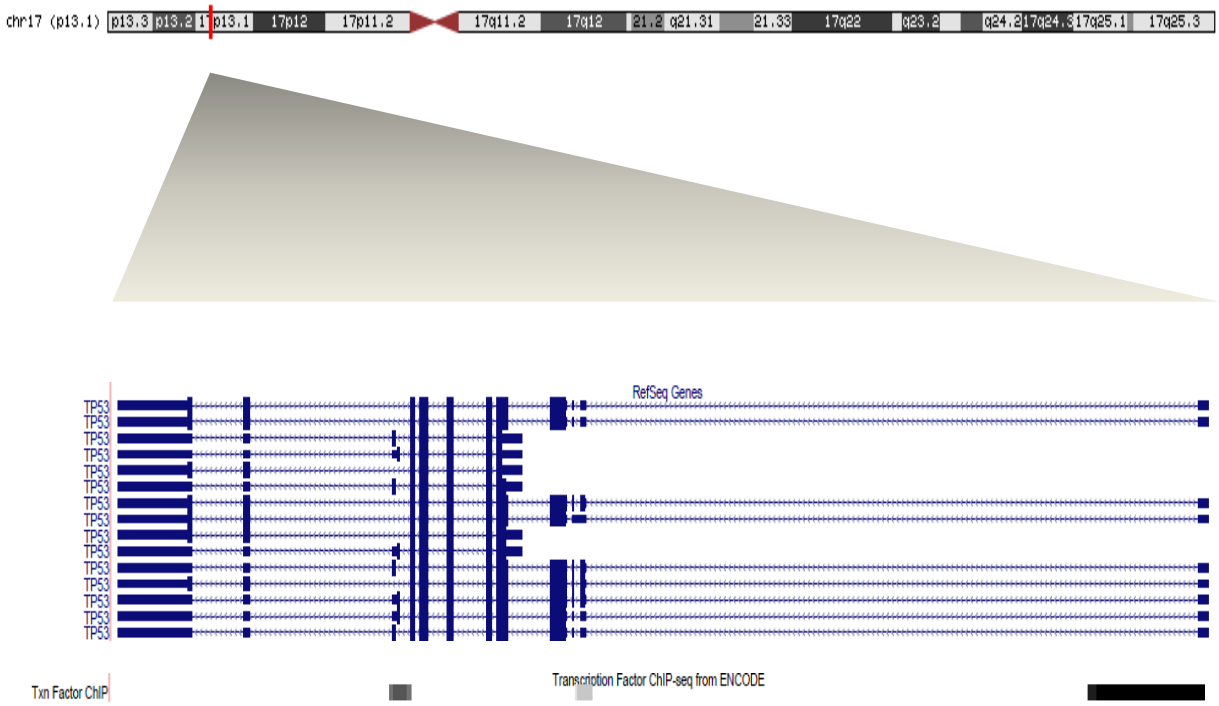

Figure 1.2 - *TP53* gene structure and the different transcript variants. RefSeq genes mapped to the area of interest around the gene *TP53* and Transcription Factor ChIP-seq according to the Genome Browser. In Transcription Factor ChIP-seq data, the black band means that the binding is stronger and more occupancy by transcription factors and the grey band indicate that binding is weaker and less occupancy by transcription factor.

1.3 Cis-regulatory SNPs

Millions of genetic variants occur in the human genome and are designated as polymorphisms. These genetic variants represent approximately 1% of the genome. There are some variants that cause numerous phenotypic differences and are responsible for intra-species variability. Some genetic variants have little impact on human health but the majority of these variants can have a strongly impact on disease susceptibility (Wilkins et al., 2007; Vernot et al., 2012). Common polymorphisms include deletions, insertions and/or duplications of segments (copy number variations), sets of repeated segments (mini and/or microsatellites) and single nucleotide polymorphisms (SNPs). SNPs are very frequent polymorphisms and constitute about 90% of all known sequence variations.

SNPs located within the coding region of a gene may cause amino acid alterations (non-synonymous variants). These can induce protein polarity shift, misfolding, unsuitable phosphorylation and other functional consequences. SNPs located in non-coding regions are mostly considered as non-functional, nevertheless, this type of alterations can affect

regulatory elements of genes like enhancers, promoters or silencers, which may lie close to the gene, but can also be found hundreds of kilobases away from the gene that they regulate. These regulatory elements can collectively modulate timing, magnitude and cell-specificity of gene expression (Pastinen et al., 2006).

Cis-regulatory SNPs are the genetic variants that affect the gene expression. Cis-regulatory SNPs can act through different mechanisms by affecting transcription factor binding sites (TFBS), histone modification and DNA methylation, for example. Probably the most common mechanism is the modification of TFBS. A SNP in a TFBS can have three main consequences: (1) a SNP may decrease or increase binding of a transcription factor (TF) and cause allele-specific gene expression alteration; (2) a SNP may delete an existing binding site or create a novel one; and (3) a SNP may not have any effect in the binding of the TF and consequently not change the expression, since a TF can recognize numerous binding sites. Therefore, a SNP in a TFBS can affect disease susceptibility because it can lead to differences in gene expression and phenotypes (Rockman et al., 2002; Serre et al., 2008; Maia et al., 2012).

1.4 Differential Allelic Expression

An important area of medical research is the understanding of the role of cis-regulatory SNPs on disease. Because cis-regulatory SNPs generate imbalances in the expression of the alleles of a gene, differential allelic expression (DAE) analysis is a powerful method for studying them. To study DAE is to compare the relative expression of the two alleles in one heterozygous individual (Figure 1.3). In other words, the effect of cis-regulatory SNPs in a target gene can be detected by measuring the relative expression of the two alleles of that gene, by using transcribed SNPs in heterozygous individuals as allelic markers (Wilkins et al., 2007; Maia et al., 2009).

DAE is largely responsible for intra-species variability and can lead to the appearance of susceptibility to complex disease (Cheung et al., 2010). Variation in cis-regulatory elements is common in the human genome and DAE has been estimated to affect 20-50% of genes, depending on the method used. Some studies have reported that genes demonstrating DAE are tissue specific, with differences in allelic expression referred to liver, brain, kidney and spleen. To identify DAE that is likely to play an important biological role, recent genome-

wide association studies (GWAS) have been focused on the interpretation of coding or other SNPs in transcribed regions (Wilkins et al., 2007; Gagneur et al., 2009).

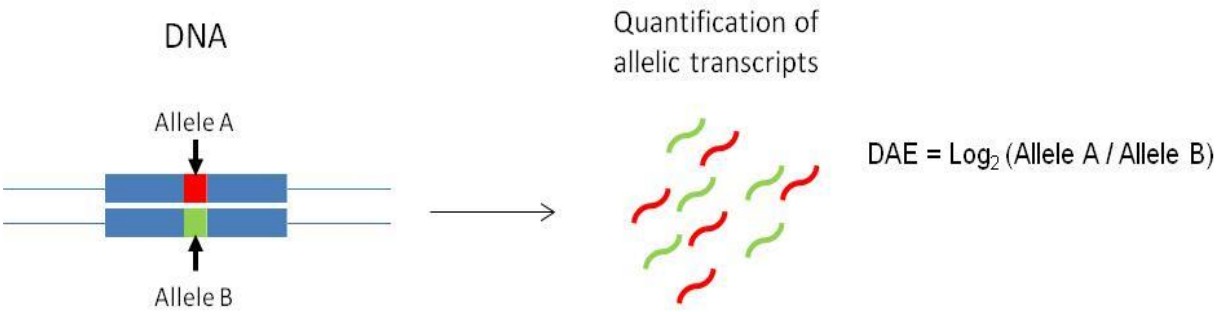

Figure 1.3 - Differential allelic expression analysis.

Our group is mainly interested in studying DAE in breast cancer susceptibility. Previously, it has been observed that *TP53* displays DAE and all heterozygous individuals express more of the same allele, both in blood and breast tissue (Figure 1.4). Because the allelic expression ratio is in log scale in Figure 1.4, all of the individuals near 0 express the two alleles equally but the individuals that are distant from 0 express more of one allele, i.e. have DAE. In the case of *TP53,* all of the heterozygous individuals express more of the same allele, indicating that the marker SNP is in complete linkage disequilibrium (LD) with the cis-regulatory variant (or variants). As normal blood and normal breast tissue express the same allele, we can conclude that *TP53* have cis-regulatory variants.

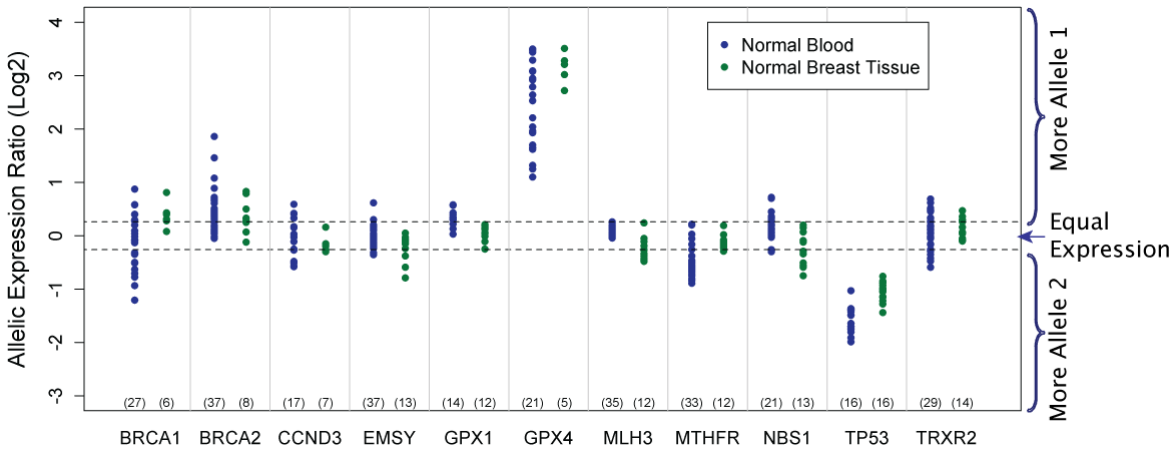

Figure 1.4 - Comparison of DAE in blood and breast tissue in different genes. Heterozygous individuals are represented as dots. The numbers in parentheses are the quantity of individuals studied for each sample (adapted from Maia et al., 2009).

Preliminary data from our group also indicates that tumours have a completely different DAE pattern from control blood and normal breast tissue (Figure 1.5). The control and normal breast samples exhibit a unilateral allelic expression. And the tumours samples show a wide bidirectional variation in the allelic expression, with the allelic expression in these samples showing a large discrepancy. Consequently, this can have consequences on the characteristics of tumours.

Therefore, we are interested in studying whether cis-regulatory polymorphisms of *TP53* are involved in the susceptibility to breast cancer as well as in determining the clinical characteristics of tumours.

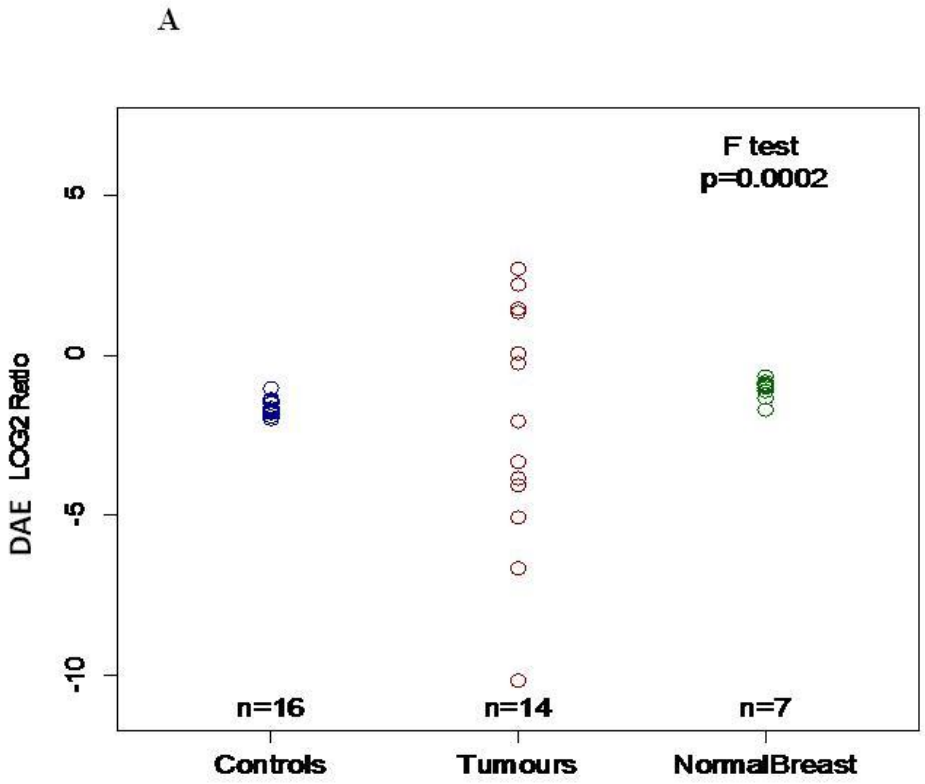

Figure 1.5 - DAE analysis of *TP53* in control, normal breast and tumours. "ControlS" corresponds to control blood, the "Tumours" correspond to breast tumours and the "Normal Breast" corresponds to normal breast tissue. Each circle represents a heterozygous individual and "n=" represent the number of samples. The p-value corresponds to an F test to compare the variance of the three groups. (Maia AT personal communication).

1.5 Aims

The main goal of the present work was to map the cis-regulatory variants of *TP53* in breast cancer. The specific aims of this work were:

1. Identification of candidate cis-regulatory SNPs in the gene *TP53*;
2. prediction of transcription factor binding sites in this locus through *in silico* analysis;
3. study DNA-protein interactions through *in vitro* and *in vivo* analysis.

2 MATERIALS AND METHODS

2.1 Cell lines and cell culture

In this study, we used different cell lines: breast cancer (MCF-7, T47D, HCC1954 and MDA-MB-436) and lymphoblastoid (GM12878 and GM06991) for nucleic acid and nuclear protein extraction.

Breast cancer cell lines MCF-7 and T47D are positive for oestrogen receptor and derived from invasive ductal carcinoma (Keydar et al., 1979; Soule et al., 1990), HCC1954 is negative for oestrogen receptor and derived from a ductal carcinoma (Gazdar et al., 1998) and MDA-MB-436 (Cailleau et al., 1978) is negative for oestrogen receptor and derived from an invasive ductal carcinoma (Di Leva et al., 2013). These cell lines were kindly provided by Dr. Suet-Feung Chin of the University from Cambridge.

Lymphoblastoid cell lines GM12878 and GM06991 were obtained from the Centre d'Étude du Polymorphisme Humain (CEPH) collection, which is an international resource of cultured lymphoblastoid cell lines (LCL) from 1050 individuals in 52 world populations.

MCF-7 and T47D cell lines were cultured in DMEM medium (Dulbecco's Modified Eagle Medium), HCC1954, GM12878 and GM06991 cell lines were cultured in RPMI1640 medium (Roswell Park Memorial Institute) and MDA-MB-436 cell line was cultured in Leibovitz's L-15 medium. All culture media were supplemented with 10% foetal bovine serum (FBS) and 1% penicillin/streptomycin. All media and supplements were obtained from Invitrogen. These cell lines were incubated at 37^{O}C with 5% CO_2.

For sub-culturing cells, the culture medium was removed from the T-flask and the cells were washed with PBS 1x (Sigma), to withdraw all residues and dead cells. Then, Tripsin-EDTA solution was added and the cells were incubated at 37^{O}C for 2 minutes to facilitate the dispersion of the cell layer. The cells were observed under an inverted microscope to ensure that the cell layer was dispersed throughout. Complete growth medium was then added and the cells were resuspended and transferred to a centrifuge tube. The cell suspension was centrifuged at approximately 300rpm for 5 minutes, and the supernatant was discarded. The cell pellet was resuspended in fresh growth medium and divided into new T-flasks.

For freezing cells, the procedure described above was performed with the exception that complete growth medium was supplemented with 5% DMSO (Sigma). The cell suspension was aliquoted in vials and these were placed in a freezing container (Nalgene) at -80^{o}C. The freezing container has isopropyl alcohol that allows a rate of cooling approximately of 1^{o}C/minute, which does not compromise cell viability.

For thawing cells, the cell vial was placed in a 37^{o}C water bath with gentle agitation. The vial was removed from the water bath as soon as the contents were thawed. The cell suspension was removed immediately to a centrifuge tube that was prepared previously with 10ml of culture medium. This was centrifuged at approximately 300rpm during 5 minutes. The supernatant was discarded and the cell pellet was resuspended in 5ml of fresh growth medium. The cells were transferred to a 25cm^{2} T-flask and were observed under an inverted microscope and incubated as described above.

2.2 Nucleic acid extraction and processing

Genomic DNA was extracted from the cell lines with the QIAamp DNA Mini kit (QIAGEN) following the manufacturer's instructions. Initially, 5 x 10^{6} cells were centrifuged at 300xg for 5 minutes. The pellet was resuspended in 200µl PBS and 20µl of proteinase K were added. 200µl Buffer AL were then added and mixed thoroughly, followed by an incubation at 56^{o}C for 10 min. 200µl ethanol (100%) were added to the sample and mixed by vortexing. The mixture was pipetted into the DNeasy Mini spin column and centrifuged at 6000xg for 1 minute. Next, 500µl Buffer AW1 were added, followed by a centrifugation at 6000xg for 1 minute. 500µl Buffer AW2 was added in column and centrifuged at 20,000xg for 3 minutes to dry the DNeasy membrane. The column was placed in a clean tube and 200µl Buffer AE were pipetted directly onto the DNeasy membrane. This column was incubated at room temperature for 1 minute and then was centrifuged at 6000xg for 1 minute to elute.

Total RNA was extracted from the cell lines using TRI Reagent (Sigma-Aldrich). The TRI Reagent is a monophasic solution of phenol and guanidine thiocyanate that at the same time dissolves DNA, RNA, and protein on homogenization or lysis of tissue sample. For monolayer cells (MCF-7, T47D, HCC1954 and MDA-MB-436 cell lines), 1ml of the TRI Reagent was added per flask. Then, the cell lysate was passed several times through a pipette to form a homogenous lysate. For suspension cells (GM12878 and GM06991 cell lines) 5-10x10^{6} cells

were centrifuged and then lysed in 1ml of TRI Reagent, with repeated pipetting. To ensure complete dissociation of nucleoprotein complexes, the samples were left for 5 minutes at room temperature and then 0.2ml of chloroform (Merck) per ml of TRI Reagent were added. The chloroform is used to separate DNA, RNA and proteins in different phases. The samples were mixed vigorously for 15 seconds and incubated at room temperature, during 15 minutes. The mixture was centrifuged at 12,000g for 15 minutes, at 4^{o}C. Centrifugation separated the mixture into 3 phases: a red organic phase (containing protein), an interphase (containing DNA) and a colourless upper aqueous phase (containing RNA). The aqueous phase was transferred to a fresh tube and added 0.5ml of isopropanol (Sigma) per ml of TRI Reagent. The isopropanol serves to precipitate the RNA, originating a pellet after centrifugation. The sample stood 5-10 minutes at room temperature and then was centrifuged at 12,000g for 10 minutes at 4^{o}C. The supernatant was removed and the RNA pellet was washed by adding 1ml of 75% ethanol per ml of TRI Reagent. 75% ethanol is used as a wash solution because RNA is a precipitate in this percentage of ethanol, while most proteins and salts remain in solution. The sample was vortexed and then centrifuged at 7,500g for 5 minutes, at 4^{o}C. The RNA pellet was air-dried for 5-10 minutes and then an appropriated volume of water was added and mixed by repeated pipetting. The RNA was subsequently treated with DNase. For DNase treatment, the reactions were performed using: 2U of RNase-Free DNase, 1µl of RNase-Free DNase 10x Reaction Buffer, 1µg of RNA and Nuclease-free water to a final volume of 10µl. The samples were incubated at 37^{o}C for 30 minutes. Then, the reaction was terminated with the addition of 1µl of DNase Stop Solution and incubated at 65^{o}C, for 10 minutes. All the reagents used for DNase treatment were purchased from Invitrogen.

cDNA was prepared from total RNA with Super-Script III First Strand SuperMix (Invitrogen) using oligo-dT primer according to manufacturer's instructions. For First Strand cDNA Synthesis, the reactions were performed using: 5µg of total RNA, 50µM oligo(dT), 1µl of annealing buffer and RNase/DNase-free water in a final volume of 10µl. The PCR tubes were incubated at 65^{o}C for 5 minutes, immediately put on ice for at least 1 minute and then 10µl of First-Strand Reaction Mix and 2µl of SuperScript III/RNaseOUT Enzyme Mix were added. The samples were briefly vortexed and collected by centrifugation, and then incubated at 50^{o}C, for 5 minutes. To terminate the reaction, the samples were incubated at

85^{o}C, for 5 minutes and then placed on ice. The cDNA synthesis reactions were stored at -20^{o}C.

2.3 Quantification of DNA and RNA

DNA and RNA concentrations were assessed with the Thermo Scientific NanoDrop 2000c Spectrophotometer. Spectrophotometric analysis is one of the more commonly used as a measure of quantity and purity of nucleic acid, as this absorbs ultraviolet light in a specific pattern. DNA and RNA samples are exposed to ultraviolet light at 260 nanometres (nm) and a photodetector measures the ultraviolet light which passes through the samples. The ratio of the absorbance at 260 and 280nm is used to measure the purity of nucleic acid. For DNA, a pure sample will yield a ratio of approximately 1.8. A pure RNA sample will yield a ratio of approximately 2.0. The ratio is frequently used to assess the contamination; it may indicate the presence of protein, phenol or other contamination that absorb light at 280nm.

2.4 SNPs and haplotype analyses

HapMap is a catalogue of common human genetic variants, resulting from an international effort. It contains information on location and population distribution of these genetic variants. HapMap release #27 (http://www.hapmap.ncbi.nlm.nih.gov/) was accessed to download the genotyping data of the area of interest around the gene *TP53*. Haploview software (http://www.broadinstitute.org/haploview) was used for the analysis of haplotype and linkage disequilibrium. Genome Browser (http://genome.ucsc.edu/) was accessed to analyse the region selected. This site contains the reference sequence of the human genome, amongst others, and data from the ENCODE project, on regulatory elements of our genome.

2.5 *in silico* prediction of transcription factor binding sites

The prediction of transcription factor binding sites (TFBS) started with the analysis of our region of interest with different TF search engines such as the Jaspar database, the

Genomatix software and the ALGGEN website. Jaspar database, version 6.0_ALPHA (http://jaspar.binf.ku.dk/) is a collection of smaller databases (TRANSFAC, etc). Genomatix software suite (http://www.genomatix.de, Genomatix Software GmbH, Munich, Germany)(Quandtl et al., 1995) uses the MatInspector release 8.0.6 (Cartharius et al., 2005). ALGGEN website (http://alggen.lsi.upc.es/) uses PROMO with version 8.3 of TRANSFAC. We analysed the region surrounding all the SNPs, as well as used the sequences with the different alleles for each SNP. We used a cut-off of 0.9 for the matrix and the core similarity score that correspond to the quality of a match between DNA sequence and TF binding matrix.

2.6 Mobility shift DNA – binding assay

2.6.1 General description of the assay

Electrophoretic Mobility Shift Assay (EMSA) is a powerful technique that allows *in vitro* detection of the interaction between a protein and a labelled DNA probe. In the assay a labelled double-stranded oligonucleotide containing a known sequence with a putative TFBS is added to the nuclear extract, thereby allowing DNA-protein complexes to form. The DNA-protein probe mixture is loaded on a polyacrylamide gel and electrophoresed. If proteins are bound to the labelled DNA, the protein-DNA complex will migrate more slowly through the gel, creating a shift relative to the unbound oligonucleotide.

Binding affinity of a particular protein for a target binding sequence is distinguished with competitive EMSA by the use of competing non-labelled sequences in the binding reaction. Additionally, antibodies that recognize epitopes of a bound protein generate a decrease mobile protein-DNA complex, resulting in a supershift on the gel which indicates specificity of DNA-protein binding (Jiang et al., 2010; Pagano et al., 2011).

2.6.2 Preparation of Nuclear Extracts

2-6x10^7 cells were harvested with the addition of Tripsin-EDTA solution to facilitate the dispersion of the cell layer, and centrifuged at 1500rpm for 5 minutes to pellet; subsequent steps were performed at 4^{o}C. Pelleted cells were re-suspended in 1ml of PBS 1x to withdraw all residues and dead cells and then collected by centrifugation at full speed for 25 seconds.

The cells were resuspended in 250µl of HB buffer (10mM Tris pH 7.4, 10mM KCl (Sigma), 1.5mM MgCl (USB), 1mM DTT and 1x Protease Inhibitor (PI) (Roche)) and centrifuged at full speed for 25 seconds. The cells were resuspended in 300µl of HB buffer with 0.4% of NP40 (Fluka) and left on ice for 5 minutes. Then, the cells were centrifuged at full speed for 5 minutes, frozen in liquid N_2 and storage at -80^{o}C. The pellet was resuspended in 100µl of buffer C (20Mm Hepes (Sigma) pH 7.9, 0.4M NaCl (Merck), 1mM EDTA, 20% Glycerol (VWR), 1mM DTT and 1x PI)/ 2x10^{7} cells. The tubes were vigorously rocked at max speed for 30 minutes and then centrifuged at full speed for 5 minutes. The supernatant, designated as the nuclear extract, was used immediately.

2.6.3 Determination of total protein concentration

Nuclear extract concentration was assessed with the Qubit 2.0 Fluorometer (Life Technologies) according to the manufacturer's instructions. A working solution was made by diluting the Qubit Reagent in Qubit Buffer (1:200) in appropriate PCR tubes. This assay requires three standards tubes with 190µl of working solution and 10µl of Qubit standards. The same process was made for the nuclear extract samples. PCR tubes were mixed by vortexing and incubated at room temperature for 15 minutes. Then, the standards were measured first to calibrate the Qubit 2.0 Fluorometer followed by the samples.

This method uses a fluorescent dye (Qubit dye) that intercalates nucleic acid or protein to determine its concentration. The amount of fluorescence signal from the mixture is directly proportional to the concentration of nucleic acid or protein in the sample solution. The Qubit fluorometer reads this fluorescence signal and extrapolates the sample concentration by using the Qubit standards of known concentration.

2.6.4 Preparation of DNA probes

Primers were designed in the regions that contain the selected SNPs and are summarized in Table 2.2. Single-stranded DNA probes (Invitrogen) and unlabelled control were labelled with Biotin 3' End DNA Labelling Kit (Thermo Scientific). For the labelling reaction, in a total volume of 50µl, 25µl of ultrapure water were mixed with 1x TdT Reaction Buffer, 100nM Unlabeled Oligos, 0.5µM Biotin-11-UTP and 10U Diluted TdT. The reaction mixture was incubated at 37^{o}C for 30 minutes, and then 2.5µl of 0.2 M EDTA were added to stop the reaction. 50µl of chloroform:isoamyl alcohol (24:1) (Sigma) were added to the

mixture to extract the TdT, and the phases were separated through centrifugation at high speed for 2 minutes. The complementary oligos that were end-labelled separately were then annealed by mixing together equal amounts of labelled complementary oligos and incubated at 80^{o}C for 10 minutes and then at room temperature for at least 3 hours.

Table 2.1 – Primers designed for EMSA.

SNPs	Alleles	Strand	Sequence*	Primer
rs1042522	G	Forward	GCAGGGGCCACGGGGGGAGCAGCCT	ATM1
		Reverse	AGGCTGCTCCCCCCGTGGCCCCTGC	ATM2
	C	Forward	GCAGGGGCCACGCGGGGAGCAGCCT	ATM3
		Reverse	AGGCTGCTCCCCGCGTGGCCCCTGC	ATM4
rs17878362	CCCCAGCCCTCCAGGT	Forward	TCAGCCCCCCAGCCCCCCAGCCCTCCAGGTCCCCAGCCC TCCAGGTCCCCAGCCCAACCC	ATM13
		Reverse	GGGTTGGGCTGGGGACCTGGAGGGCTGGGGACCTGGA GGGCTGGGGGGCTGGGGGGCTGA	ATM14
	-	Forward	CCAGGTCCTCAGCCCCCCAGCC-CCCCAGCCCTCCAGGTCCCCAGCCCAACCCTTGTCCTT	ATM15
		Reverse	AAGGACAAGGGTTGGGCTGGGGACCTGGAGGGCTGGG G-GGCTGGGGGGCTGAGGACCTGG	ATM16
rs2307496	G	Forward	GTCAGTCCCATGGAATTTTCGCTTC	ATM17
		Reverse	GAAGCGAAAATTCCATGGGACTGAC	ATM18
	-	Forward	GTCAGTCCCATG-AATTTTCGCTTTC	ATM19
		Reverse	GAAAGCGAAAATT-CATGGGACTGAC	ATM20
rs2981578	G	Forward	CTCTATGCAAATATGCGGTTTGGAGCAGGG	FGFR2-13
		Reverse	CCCTGCTCCAAACCGCATATTTGCATAGAG	FGFR2-13

*The underlined nucleotides indicate the SNPs; - : indicates the deletions

The labelling efficiency was determined by dot blot using hand spotting following the manufacturer's instructions. The samples and standards were put onto the nylon membrane (Thermo Scientific) that was previously hydrated in TE buffer (10mM Tris pH 8.0 and 1mM EDTA) and the crosslink of the membrane was performed in a transilluminator with the membrane face down, for 15 minutes.

For the detection of the spotted standards and samples, the Streptavidin-Horseradish Peroxidase Conjugate was used together with the Chemiluminescent Substrate from the Chemiluminescent Nucleic Acid Detection Module (Thermo Scientific). To determine the labelling efficiency the spot intensities of the samples were compared to those of the Biotin-EBNA Control DNA.

2.6.5 Identification of DNA-protein binding reaction

EMSA were performed with LightShift Chemiluminescent EMSA Kit (Thermo Scientific) according to manufacturer's instructions. A 6% polyacrylamide gel was prepared as follows: 7333µl of H_2O, 125µl of 5xTBE (450mM Tris, 450mM boric acid (Sigma), 10mM EDTA), 1000µl of Glycerol, 1500µl of 30% acrylamide (Ambion), 30µl of 20% APS (Sigma) and 12µl of TEMED (Nzytech). The gel was pre-run for 30 minutes at 100V. A complete set of three reactions were performed for each SNP and for Biotin-EBNA control DNA (summarized in Table 2.4) and the binding reactions were incubated at room temperature, for 20 minutes. To each 20µl binding reaction, 5µl of 5X Loading Buffer was added and then 20µl of each sample was loaded onto the wells of the polyacrylamide gel.

After the electrophoresis, the binding reactions on the polyacrylamide gel were transferred onto a nylon membrane and were sandwiched between three chromatography paper sheets on each side soaked in transfer buffer (0.5X TBE). The transfer was carried out by a Trans-Blot SD Smi-Dry Transfer Cell (Bio-Rad) at 20V, for 25 minutes. Crosslink and detection were performed as described in Section 2.7.4.

Table 2.2- Binding reaction calculations used for each SNP

Component	**Final Amount**	**Reaction**		
		1	2	3
Ultrapure water	-	12µl	µl	µl
10X Binding Buffer	1X	2µl	2µl	2µl
1µg/µl Poly (dI·dC)	50ng/µl	1µl	1µl	1µl
50% Glycerol	2.5%	1µl	1µl	1µl
1% NP-40	0.05%	1µl	1µl	1µl
100Mm $MgCl_2$	5mM	1µl	1µl	1µl
Unlabelled Target DNA	4pmol	-	-	
Protein Extract		-		
Biotin End-Labelled Target DNA	20fmol			
Total volume	-	20µl	20µl	20µl

2.7 Genotyping

Genomic DNA from cell lines, as described in Section 2.2, was amplified by the Polymerase Chain Reaction (PCR) in a total volume of 50µl. The reactions were performed with 50ng of genomic DNA, 1x PCR buffer, 2,5mM of $MgCl_2$, 0.2mM of deoxynucleotidyl triphosphates (dNTPs), 0.125µM of each primer (Table 2.1) and 5U of Taq DNA Polymerase (all from Invitrogen). The reaction mixture was initially incubated for 2 minutes at 95ºC and then submitted to 35 cycles including in each one denaturing (30 seconds at 95ºC), annealing (30 seconds at 60ºC) and extension (45 seconds at 72ºC) steps, and a final extension step of 7 minutes at 72ºC.

cDNA samples obtained from cell lines was used in a PCR reaction with the same conditions.

Specific forward and reverse primers for the region around the selected SNPs were designed with Primer3 program (available at http://frodo.wi.mit.edu/), and are summarized in Table 2.3.

Table 2.3- Primers for the region around the selected SNPs.

	Strand	Oligonucleotides
DNA	Forward	GCCAGGCATTGAAGTCTCAT
	Reverse	TGGAAGTGTCTCATGCTGGA
cDNA	Forward	CCCCTCTGAGTCAGGAAACA
	Reverse	AGAATGCAAGAAGCCCAGAC

The resulting PCR products were submitted to electrophoresis in 1.5% agarose gel (UltrapureTM Agarose, Invitrogen) prepared with 1x TAE buffer (40mM Tris (Sigma), 20mM acetic acid (Merck), 1mM EDTA (Sigma)) and 2.5µl/100 ml GreenSafe (Nzytech). A Gene Ruler 1kb (Thermo Scientific) was used as a DNA size marker. The gels were visualized using UV transilluminator (Gel Image Analyser, GelDoc 2000, BioRad).

The DNA fragments of interest were excised from the agarose gel and extracted using QIAquick PCR Purification Kit (QIAGEN), according to the manufacturer's instructions. The final product was diluted in 30 µl of ultrapure water (Sigma).

The DNA sequencing was performed in the Centre of Marine Sciences (CCMAR; Serviço de Biologia Molecular). CCMAR used Sanger sequencing with BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems).

3 RESULTS

3.1 *in silico* analysis

The genotyping data of area of the interest around *TP53* was collected from HapMap release #27. HapMap is a database with genotyping information of a large proportion of SNPs, obtained from normal controls. Then this genotyping data was inserted in the Haploview software. In this software the haplotype structure and linkage disequilibrium (LD) of this region were analysed. LD is the higher probability of two alleles of two SNPs being inherited together, than it would be expected by chance alone. In Haploview software, we identified two main haplotype blocks (Figure 3.1) corresponding in block 1 to two haplotypes and in block 2 to four haplotypes in the European population (CEU samples in HapMap) (Figure 3.2). SNP rs1042522 was used by Maia et al (2009) as a marker SNP in a breast tissue DAE study, and we searched for SNPs in complete LD with it, as all heterozygotes identified in that study had the same allele being preferentially expressed. rs1042522 is included in block 2, in which haplotype 1 is the most common in the European population, with an approximate frequency of 72%. Our results showed that the SNP rs1642785 is in complete LD with rs1042522 (Figure 3.2). Both allele C in rs1042522 and allele C in rs1642785 are the most frequent in the population. They are in complete LD so when one SNP has allele C the other SNP has allele C, as we can see in Figure 3.2. The allele C of rs1042522 has a frequency of 74% that corresponds to the haplotype 1 and haplotype 4.

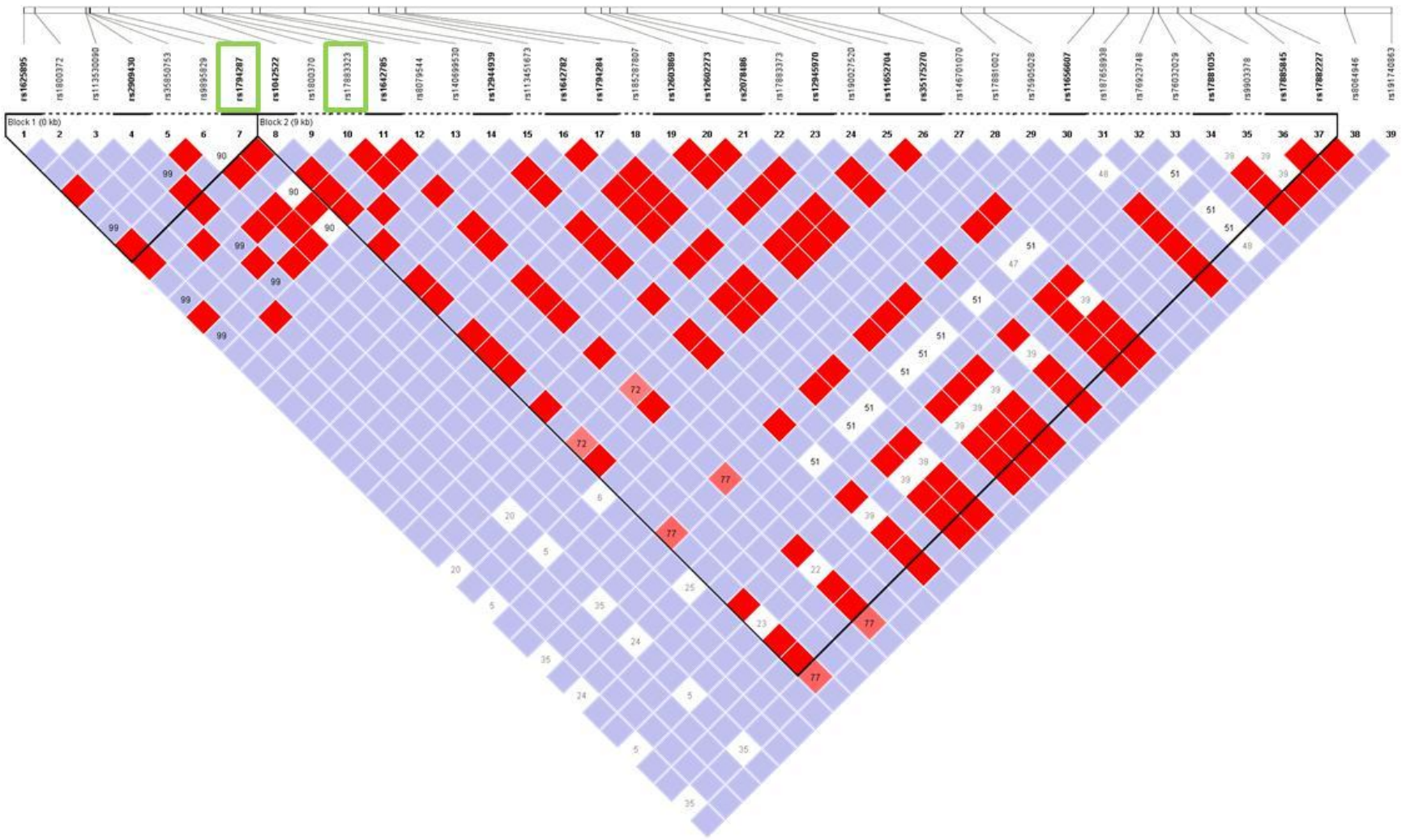

Figure 3.1 - LD plot for the region around *TP53*. Two main haplotype blocks (black triangles) were identified. The red squares denote the pairs of SNPs with stronger LD and the light blue squares are those with weaker LD. The two SNPs that were selected for this study are shown in green (http://www.broadinstitute.org/haploview).

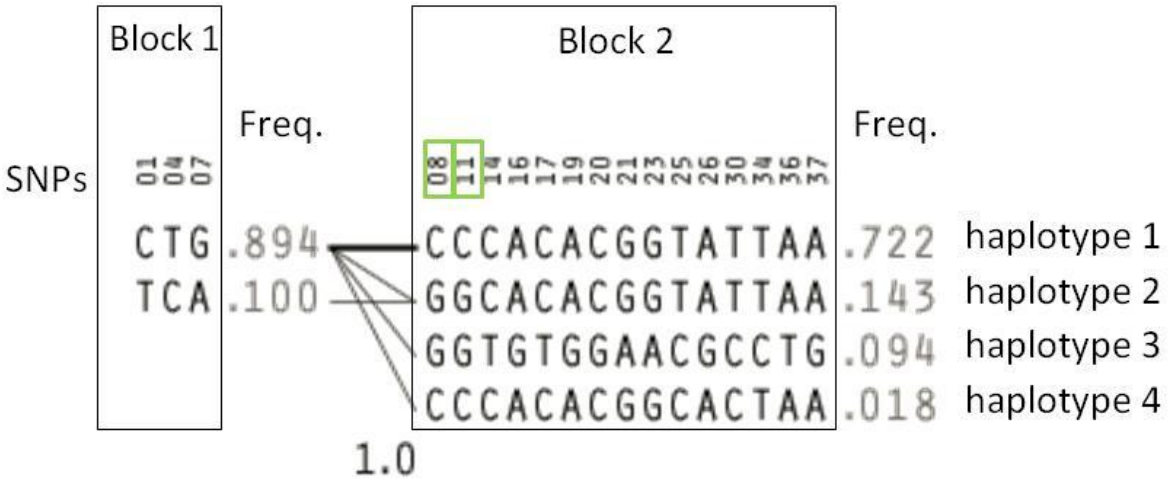

Figure 3.2 - Haplotype blocks and haplotype frequencies in *TP53*. In green rectangle are the two SNPs selected and each letter corresponds to one allele of each SNP. The SNPs 8 and 11 are, respectively rs1042522 and rs1642785 (http://www.broadinstitute.org/haploview).

In Haploview analysis we only obtained two SNPs, then we analysed the region between this two SNPs using the UCSC Genome Browser. This region has the possible existence of regulators elements where there can be cis-regulatory SNPs. The Genome Browser contains the reference sequence for a large collection of genomes, plus information on histone modifications and DNAse hypersensitivity regions. The histone modifications indicate the possibility of existence of enhancers, promoters and other regulatory elements. The DNAse hypersensitivity regions or clusters are areas of open chromatin that indicate the presence of active chromatin (see in Genome Browser according with The ENCODE Project Consortium, 2011). We used this information to investigate whether our region of interest had potential regulatory elements. In the Genome Browser there is the indication that this region may contain binding sites for several transcription factors. In the version Human Feb. 2009 (GRCh/hg19) Assembly from Genome Browser, we observed the possible existence of SNPs in the region between two SNPs selected in Haploview. And here we found four SNPs that are in overlap with regions containing possible regulatory elements (Figure 3.3). Two of these SNPs (rs1042522 and rs1800370) are in the coding region (exon 4) and the other four (rs17883323, rs1787832, rs2307496 and rs1642785) are in non-coding regions (first two are found in intron 3 and the others in intron 2), as shown the Figure 3.3.

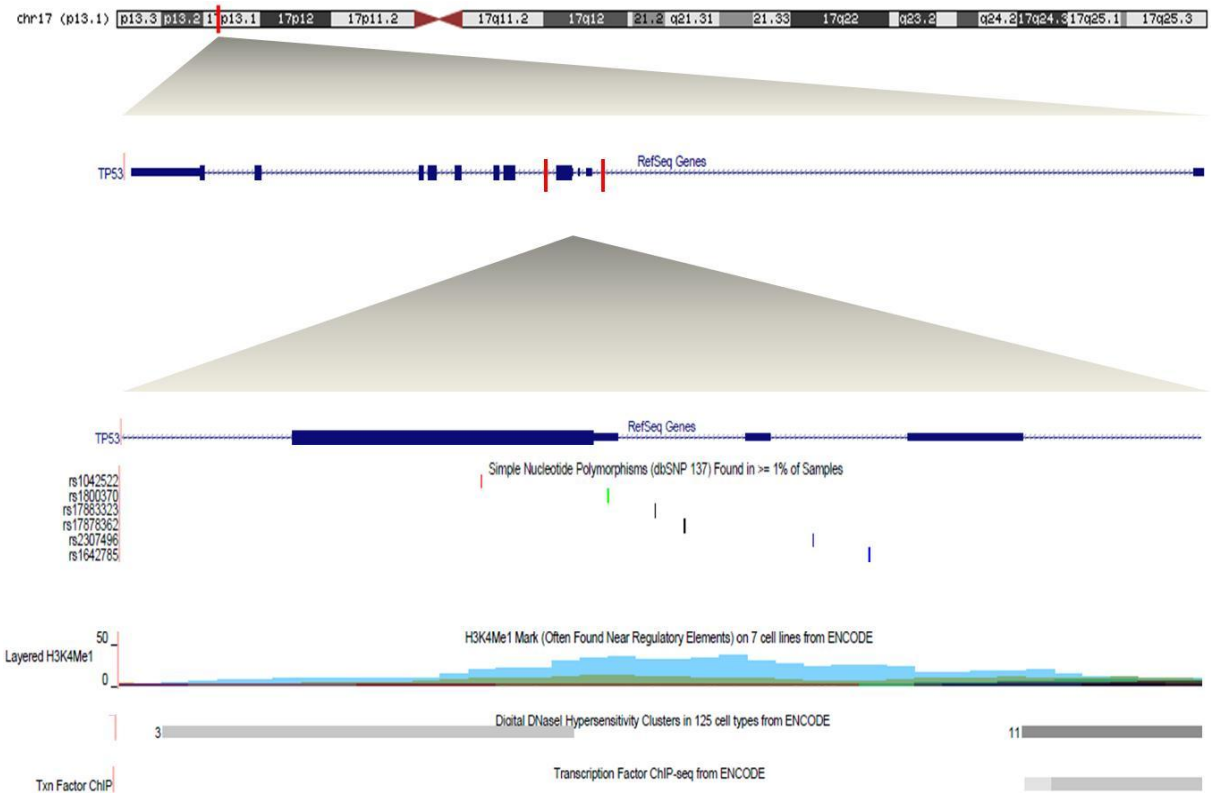

Figure 3.3 - Diagram of the *TP53* gene. RefSeq genes mapped to the area of interest around the gene *TP53*, position of SNPs and regions where there are histone modifications, DNAse clusters and transcription factors according to the Genome Browser (http://genome.ucsc.edu/).

3.2 Transcription factor binding sites results

Initially, we analysed the putative transcription factor binding sites (TFBS) in three different websites that are: Jaspar database, Genomatix software and ALGGEN. In this analysis we included the sequence around all SNPs with the two different alleles, to search for allelic differences in TF binding. Different TFBS were observed for each allele but only to three SNPs the predicted TF is known to be expressed in breast tissue (Table 3.2). Although the two alleles of rs17878362 have the same TF predicted we selected them because this SNP has a large allele (sequence with 16 alleles) that can have consequences on the number of possible TFs binding. SNP rs2307496 in G allele has one possible transcription factor and in the allele with a deletion there are two possible transcription factors. In rs1042522 there are two possible TFs binding in the presence of allele G that are not in present for allele C.

Table 3.1 - SNPs selected for analysis of TFBS.

SNPs	Alleles	Transcription Factor
rs17878362	CCCCAGCCCTCCAGGT	SP1, CTCF
	-	
rs2307496	G	HMGY
	-	HMGY, ETS1
rs1642785	G	-
	C	-
rs1042522	G	NMYC, HIF
	C	-

3.3 *in vitro* analysis

3.3.1 Identification of DNA-protein binding

The EMSA is used for studying DNA-protein interactions. For EMSA, we prepared protein nuclear extract from MCF-7, T47D, HCC1954 and GM12878 cell lines. The primers containing the SNPs of interest (Table 2.1) and the EMSA control primer (FGFR2-13) were labelled with biotin. The FGFR2-13 has been reported to show binding in EMSAs using nuclear extracts from these cell lines (Meyer et al., 2008). The level of labelling of all primers was between 75-100% of the Biotin-EBNA Control DNA, as shown in Figure 3.4.

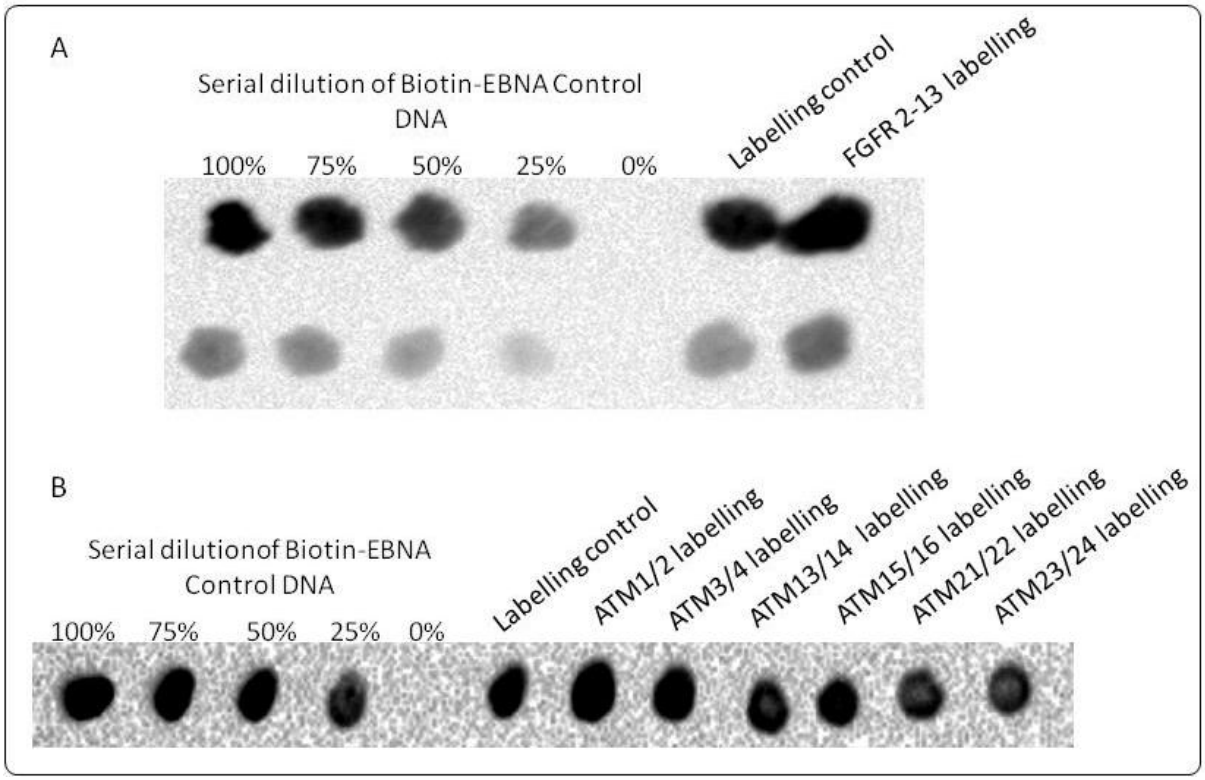

Figure 3.4 - Level of labelling from all primers and Biotin-EBNA Control DNA. (A) Detection of labelling from FGFR2-13 and Biotin-EBNA Control DNA. (B) Detection of labelling from all SNPs.

The Biotin-EBNA Control DNA was tested and EMSA was performed for nuclear extract of MCF-7, T47D, HCC1954 and GM12878 cell lines with primers detailed in Table 2.1. Initially, we tested the SNPs for all cell lines and saw that the shift did not occur (Figure3.5A). However, when we tested a control primer (FGFR2-13) for the nuclear extracts, this did not produce a shift as well (Figure 3.5B). This primer has previously been shown to bind to Oct1 and Runx2 (Meyer et al. 2008). This result suggests that our nuclear extracts were not working. Amongst other possibilities, we believe that the temperature of the laboratories at the time of preparation of the nuclear extracts was too high and may have interfered with the efficiency of the preparation. Therefore, we propose to re-extract the nuclear proteins this time using a different protocol, the NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific), and keeping strict temperature control of our experiments.

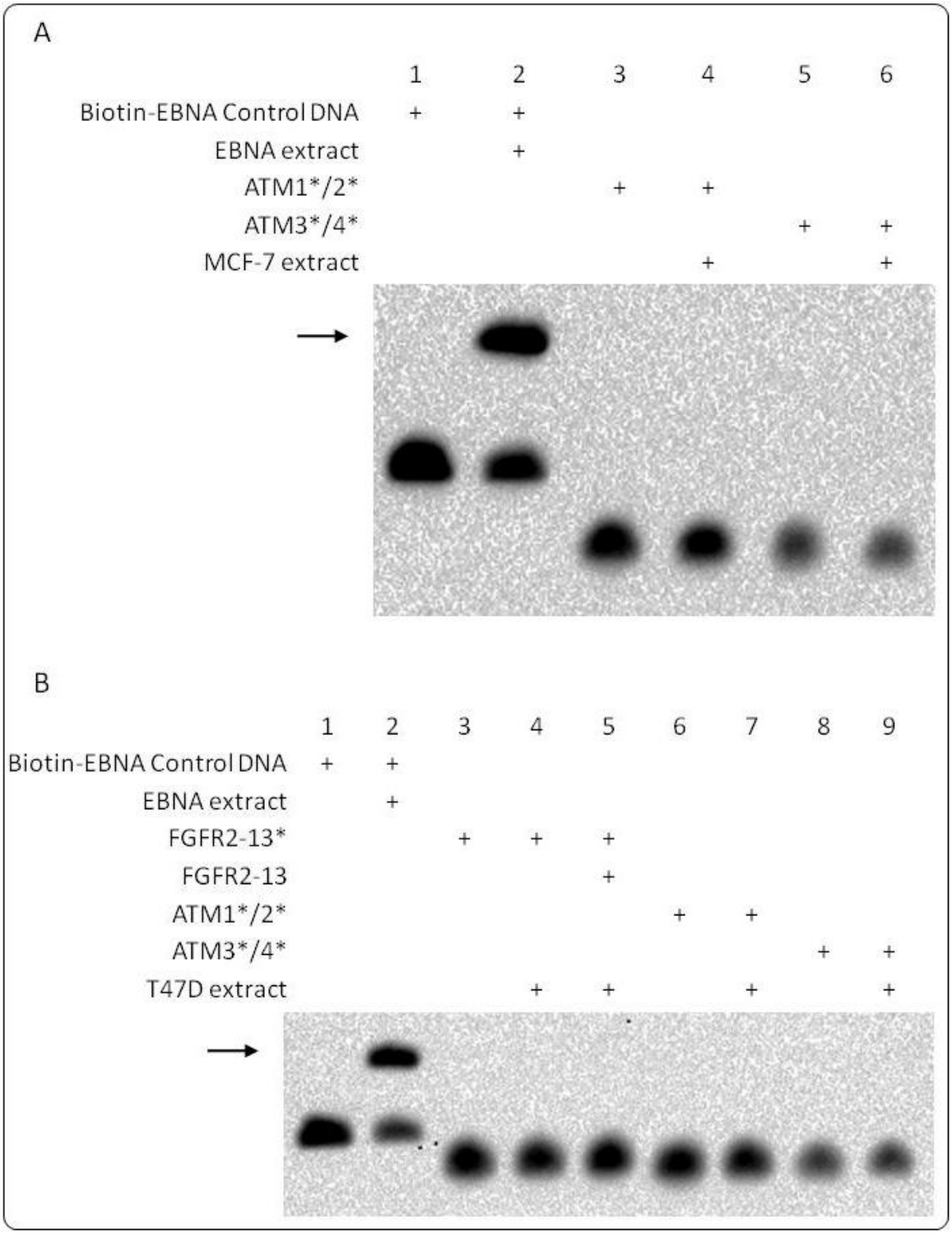

Figure 3.5 – *in vitro* DNA-protein binding studies. (A) Analysis of rs1042522, using MCF-7 cell extract. (B) Analysis of FGFR2-13 (control positive) and rs1042522, using T47D cell extract. In both analysis was tested Biotin-EBNA Control DNA. Arrows indicate specific bands for DNA-protein interactions of Biotin-EBNA Control DNA. *: represent the primers labelled.

3.4 Genotyping analysis

In HapMap there is genotyping information for SNPs rs1642785 and rs1042522. Therefore, all other SNPs were sequenced to determine the genotype in all cell lines. The aim of genotyping is to define which cell lines should be used for Chromatin Immunoprecipitation (ChIP). In Genome Browser there is no information from ENCODE on the breast cancer cell lines we are using but there is information on the lymphoblastoid cell lines (GM12878 and GM06991). Therefore, lymphoblastoid cell lines are the positive controls for the ChIP analysis. Genotyping of the breast cancer and lymphoblastoid cell lines was performed by PCR amplification of the region of interest followed by direct sequencing.

The genotype is heterozygous only for three SNPs in GM12878 cell lines (Figure 3.6), all other SNPs are homozygous for all cell lines (Table 3.2).

Table 3.2 - Analysis of genotype from different cell lines.

SNPs	MAF	Region	Cell lines					
			MCF-7	T47D	HCC1954	MDA-MB436	GM12878	GM06991
rs1642785	36.4	Non-Coding	CC	CC	ND	ND	CG	ND
rs2307496	NA	Non-Coding	+/+	+/+	ND	ND	+/+	ND
rs17878362	NA	Non-Coding	-/-	+/+	ND	ND	+/-	ND
rs17883323	7.3	Non-Coding	AA	CC	ND	ND	CC	ND
rs1800370	1.2	Coding	GG	GG	CC	CC	GG	CC
rs1042522	39.8	Coding	CC	CC	GG	CC	CG	CC

MAF: Minor Allele Frequency; ND: Not Determined; NA: Not Available.

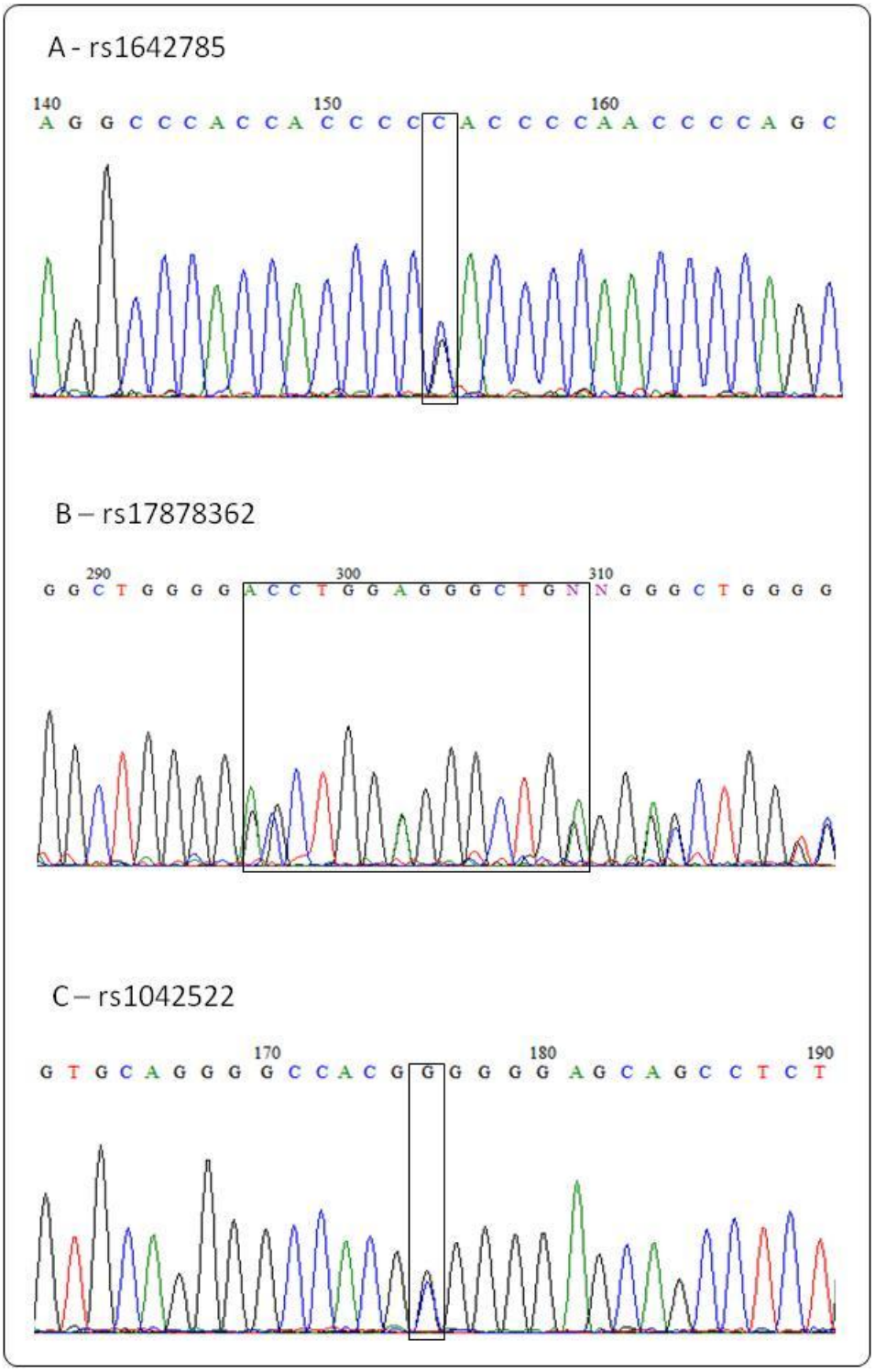

Figure 3.6 - Results of sequencing from heterozygous SNPs in GM12878 cell line. (A) rs1642785 is in negative strand and has the allele C and allele G. (B) rs17878362 is in negative strand, this SNP is a deletion of the sequence ACCTGGAGGGCTGGGG. (C) rs1042522 is in positive strand and has the allele C and allele G. Black rectangles indicate the position of the SNPs.

DISCUSSION

Our aim in this study was to investigate whether cis-regulatory SNPs of *TP53* have any effect in breast cancer. For this, we proposed to map the cis-regulatory SNPs affecting the gene expression of *TP53* and the predicted TFBSs in this region.

Initially, we chose the SNP (rs1042522) that showed previously differential allelic expression in normal breast tissue and blood (Maia et al., 2009). Our *in silico* analysis showed that rs1042522 is in complete LD with another SNP (rs1642785). This led us to decide to investigate the region between these two SNPs for histone modifications, DNase clusters and transcription factors binding sites. We found that in this interval there is a very active region, with the possibility of existence of regulatory elements. Four other SNPs overlap these elements and are candidates to be cis-regulatory SNPs. So, we analysed the TFBSs for six cis-regulatory SNPs in total. Each of these SNPs has two different alleles, which were analysed individually for differential TF binding. Three were identified to have possible TFBSs that are expressed in breast tissue, with differences for the two corresponding alleles. SNP rs17878362 has the same TFBS predicted for both alleles (for SP1 and CTCF) and we selected it because we think that the large deletion allele can have consequences in terms of numbers of TFs binding (SP1 has different number of binding sites for each allele). The G allele of SNP rs2307496 has one possible TFBS (for HMGA2). The alternative allele, with a deletion in that position, hast two possible TFBS (for HMGA2 and ETS1). In rs1042622 there are two possible TFBS (for NMYC and HIF1) in presence of G allele that do not appear in the presence of the C allele.

All of these transcription factors have been reported to be involved in cancer. Specific protein 1 (SP1) is a transcription factor that regulates the expression of many genes involved in processes like differentiation, cell growth, apoptosis, among others (Zhang et al., 2013). CCCT-binding Factor (CTCF) is a transcription factor and an epigenetic regulator that is associated with Wilms tumours, breast cancer and prostate cancer (Ross-Innes et al., 2011). High mobility group AT-HOOK 2 (HMGA2) participates in a wide variety of cellular processes like induction of neoplastic transformation and promotion of metastatic progression of cancer cells (Peluso et al., 2010). Protein C-ets-1 (ETS1) is a transcription factor that may control the differentiation, survival and proliferation of cells and may also regulate

angiogenesis (Kalet et al., 2012). N-myc proto-oncogene protein (NMYC) is transcription factor found in cancer cells and for efficient DNA binding requires dimerization with another protein (MAX) (Myzukami at al., 1995). Hypoxia-inducible factor 1 (HIF1) is expressed in most tissues and overexpressed in the majority of common cancers (Francesco et al., 2013).

To investigate whether the sequence surrounding these candidates cis-regulatory SNPs can indeed bind TF, we analysed protein-DNA interactions through EMSA, using oligonucleotides containing the SNPs of interest. We performed several experiments, but did not detect any shift, corresponding to a positive protein-DNA interaction. However, we believe that our nuclear extracts might not have been optimal, as the positive control (primer FGFR2-13) used to evaluate the quality of the extracts did not show a shift as expected. We have two possible explanations: (1) there is indeed no interaction between TFs and our oligonucleotides, therefore these SNPs are not cis-regulatory; and (2) the TF can bind to our oligonucleotides but the extracts are suboptimal. We will carry out further experiments to resolve this issue. If the second scenario is confirmed, we will study whether this binding is confirmed by *in vivo* analysis through Chromatin Immunoprecipitation (ChIP).

In the Genome Browser there is evidence of the existence of enhancers and active chromatin in our region of interest. We propose to extract the nuclear proteins of breast cancer cell lines using a different protocol (the NE-PER Nuclear and Cytoplasmic Extraction Reagents). If the protein-DNA interaction will be confirmed, the Chromatin Immunoprecipitation (ChIP) method will be used to study this interaction *in vivo*. In ChIP experiments, protein complexes that contact with DNA are crosslinked to their binding sites, by treatment with formaldehydde. The chromatin is then sheared in short fragments, and the specific DNA fraction that interacts with the protein of interest is isolated by immunoprecipitation and is then sequenced (Schmidt et al., 2009).

The central goal of this work was to study the cis-regulatory variants of *TP53* in breast cancer. In this work, we did not confirm whether any of the candidate SNPs in this region may be causing DAE observed in *TP53* gene. Therefore, we need to continue this work in order to obtain conclusive results. To investigate a link between cis-regulatory SNPs and breast cancer, we will have to compare their effect in normal and tumour samples in future work.

REFERENCES

Cailleau, R., Olive, M., Cruciger, Q. V. J. (1978). Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *14*(11), 3–7.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, *21*(13), 2933–42.

Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y. et al (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Research*, 22:1658-1667

Cheung, V. G., Nayak, R. R., Wang, I. X., Elwyn, S., Cousins, S. M., Morley, M., Spielman, R. S. (2010). Polymorphic cis- and trans-regulation of human gene expression. *PLoS biology*, *8*(9).

Cheung, V. G., Spielman, R. S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nature Reviews*, 10(9):595-604

Costa, S., Pinto, D., Pereira, D., Rodrigues, H., Cameselle-Teijeiro, J., Medeiros, R., Schmitt, F. (2008). Importance of TP53 codon 72 and intron 3 duplication 16bp polymorphisms in prediction of susceptibility on breast cancer. *BMC cancer*, *8*, 32.

Di Leva, G., Piovan, C., Gasparini, P., Ngankeu, A., Taccioli, C., Briskin, D., Cheung, D.G., Bolon, B., Anderlucci, L., Alder, H. et al. (2013). Estrogen mediated-activation of miR-191/425 cluster modulates tumorigenicity of breast cancer cells depending on estrogen receptor status. *PLoS genetics*, *9*(3), e1003311.

Dumont, P., Leu, J. I.-J., Della Pietra, A. C., George, D. L., Murphy, M. (2003). The codon 72 polymorphic variants of p53 have markedly different apoptotic potential. *Nature genetics*, *33*(3), 357–65.

Francesco, E., Lappano, R., Francesca, M., Marsico, S., Caruso, A., Maggiolini, M. (2013). HIF-1alpha/GPER signaling mediates the expression of VEGF induced by hypoxia in breast câncer associated fibroblasts (CAFs). *Breast Cancer Research*, 15:R64

Freedman, M., Monteiro, A. N. A., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C., Casey, G., Biasi, M. D., Carlson, C., Duggan, D. et al (2011). Principles for the post-GWAS functional characterization of cancer risk loci. *Nature Genetics*, 43(6): 513-518.

Gagneur, J., Sinha, H., Perocchi, F., Bourgon, R., Huber, W., Steinmetz, L. M. (2009). Genome-wide allele- and strand-specific expression profiling. *Molecular Systems Biology*, *5*(274), 274.

Gazdar, A. F., Kurvari, V., Virmani, A., Gollahon, L., Sakaguchi, M., Westerfield, M., Kodagoda, D., Stasny, V., Cunningham, H. T., Wistuba, I. I. et al. (1998). Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *International journal of cancer*, *78*(6), 766–74.

Genomatix Software. [http://www.genomatix.de].

Genome Browser. [http://genome.ucsc.edu/].

Guleria, K., Sharma, S., Manjari, M., Uppal, M. S., Singh, N. R., Sambyal, V. (2012). p.R72P, PIN3 Ins16bp polymorphisms of TP53 and CCR5Δ32 in north Indian breast cancer patients. *Asian Pacific journal of cancer prevention*, *13*(7), 3305–11.

Hanahan, D., Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674.

Haploview. [http://www.broadinstitute.org/haploview].

Hindorff, L. A., Gillanders, E. M., Manolio, T. A. (2011). Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis*, *32*(7), 945–54.

Jemal, A., Center, M. M., DeSantis, C., Ward, E. M. (2010). Global patterns of cancer incidence and mortality rates and trends. *Cancer epidemiology, biomarkers & prevention*, *19*(8), 1893–907.

Jiang, D., Jarret, H. W., Haskins, W. E. (2010). Methods for proteomic analysis of transcription factors. *Elsevier*, *1216*(41), 6881–6889.

JASPAR database. [http://jaspar.binf.ku.dk/].

Kalet, B.T., Anglin, S. R., Handschy, A., O'Donoghue, L. E., Halsey, C., Chubb, L., Korch, C., Duval, D.L. (2013). Transcription factor Ets1 cooperates with estrogen receptor α to stimulate estradiol-dependent growth in breast cancer cells and tumors. *PLoS ONE* 8(7): e68815

Kelsey, J. L., Berkowitz, G. S. (1988). Breast cancer epidemiology. *Cancer research*, (48), 5615–5623.

Keydar, I., Chen, L., Karby, S., Weiss, F. R., Delarea, J., Radu, M., Chaitcik, S., Brenner, H. J. (1979). Establishment and characterization of a cell line of human breast carcinoma origin. *European journal of cancer*, *15*(5), 659–70.

Levin, I. (1913). The mechanisms of metastasis formation in experimental cancer. *J Exp Med*, 18(4):397-405.

Liang, Y., Liu, J., Feng, Z. (2013). The regulation of cellular metabolism by tumor suppressor p53. *Cell & bioscience*, *3*(1), 9.

Maia, A.-T., Spiteri, I., Lee, A. J. X., O'Reilly, M., Jones, L., Caldas, C., Ponder, B. A J. (2009). Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. *Breast cancer research*, *11*(6), R88.

Maia, A.-T., Antoniou, A. C., O'Reilly, M., Samarajiwa, S., Dunning, M., Kartsonaki, C., Chin, S.-F., Curtis, C. N., McGuffog, L., Domchek, S. M. et al. (2012). Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. *Breast cancer research, 14*(2), R63.

Meyer, K. B., Maia, A-T., O'Reilly, M., Teschendorff, A. E., Chin, S.-F., Caldas, C., Ponder, B. A. J. (2008). Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. *PLoS Biology*, 6(5): e108.

Mizukami, Y., Nonomura, A., Takizawa, T., Noguchi, M., Nakamura, S., Ishizaki, T. (1995). N-myc protein expression in human breast carcinoma: prognostic implications. *Anticancer Res*, 15(6b),2899-905

Ohayon, T., Gershoni-Baruch, R., Papa, M. Z., Distelman Menachem, T., Eisenberg Barzilai, S., Friedman, E. (2005). The R72P P53 mutation is associated with familial breast cancer in Jewish women. *British journal of cancer*, *92*(6), 1144–8.

Pagano, J. M., Clingman, C. C., Ryder, S. P. (2011). Quantitative approaches to monitor protein – nucleic acid interactions using fluorescent probes. *Cold Spring Harbor Laboratory Press*, 14–20.

Pastinen, T., Ge, B., Hudson, T. J. (2006). Influence of human genome polymorphism on gene expression. *Human molecular genetics*, *15 Spec No*(1), R9–16.

Pelusco, S., Chiappetta, G. (2010). High-Mobility Group A (HMGA) proteins and breast cancer. Breast Care, 5:81-85

Pim, D., Banks, L. (2004). P53 polymorphic variants at codon 72 exert different effects on cell cycle progression. *International journal of cancer*, *108*(2), 196–9.

Primer3. [http://frodo.wi.mit.edu/].

Quandtl, K., Werner, T., Karas, H., Wingender, E., Neuherberg, D., Biotechnologische, G. (1995). Matind and Matinspector : new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *23*(23), 4878–4884.

Rockman, M. V, & Wray, G. a. (2002). Abundant raw material for cis-regulatory evolution in humans. *Molecular biology and evolution*, *19*(11), 1991–2004.

Ross-Innes, C. S., Brown, G. D., Carrol, J. S. (2011). A co-ordinated interaction between CTCF and ER in breast cancer cells. BMC Genomics, 12:593

Russnes, H. G., Navin, N., Hicks, J. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. *The Journal of Clinical Investigation*, *121*(10).

Scully, O. J., Bay, B., Yip, G., Yu, Y. (2012). Breast cancer metastasis. *Cancer genomics & proteomics*, *320*, 311–320.

Serre, D., Gurd, S., Ge, B., Sladek, R., Sinnett, D., Harmsen, E., Bibikova, M., Chudin, E., Barker, D. L., Dickinson, T., Fan, J.-B., Hudson, T. J. (2008). Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genetics*, *4*(2), 16.

Smith, T. R., Liu-Mares, W., Van Emburgh, B. O., Levine, E. a, Allen, G. O., Hill, J. W., Reis, I. M., Kresty, L. A., Pegram, M. D., Miller, M. S., Hu, J. J. (2011). Genetic polymorphisms of multiple DNA repair pathways impact age at diagnosis and TP53 mutations in breast cancer. *Carcinogenesis*, *32*(9), 1354–60.

Soule, H. D., Maloney, T. M., Wolman, S. R., Peterson, W. D., Brenz, R., Mcgrath, C. M., Russo, J., Pauley,R. J., Jones, R. F., Brooks, S. C. (1990). Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer research*, 6075–6086.

Strachan, T., Read, A. P. (2010). Human Molecular Genetics. New York: Garland Science.

The Algorithmics and Genetics Group (ALGGEN). [http://alggen.lsi.upc.es/].

The International Agency for Research on Cancer. [http://www.eu-cancer.iarc.fr/EUCAN/].

The International HapMap Project. [http://www.hapmap.ncbi.nlm.nih.gov/].

The ENCODE Project Consortium (2011). A user's guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology* 9(4):e1001046.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57-74.

Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., Stamatoyannopoulos, J. A., Akey, J. M. (2012). Personal and population genomics of human regulatory variation. *Genome Research*, 22:1689–1697.

Walerych, D., Napoli, M., Collavin, L., Del Sal, G. (2012). The rebel angel: mutant p53 as the driving oncogene in breast cancer. *Carcinogenesis*, *33*(11), 2007–17.

Weinberg, R. A. (2007). The Biology of Cancer. New York: Garland Science

Wilkins, J. M., Southam, L., Price, A. J., Mustafa, Z., Carr, A., Loughlin, J. (2007). Extreme context specificity in differential allelic expression. *Human molecular genetics*, *16*(5), 537–46.

Yerushalmi, R., Hayes, M. M., Gelmon, K. a. (2009). Breast carcinoma--rare types: review of the literature. *Annals of oncology*, *20*(11), 1763–70.

Zhang, Y., Zhao, Y., Shen, Y., Cai, X., Zhang, X., Ye, L. (2013). The oncoprotein HBXIP upregulates PDGFB via activating transcription factor Sp1 to promote the proliferation of breast cancer cells. *Biochem Biophys*, 434(2):305-10