

Functional Assessment of Non-coding Regulatory Variants in Familial Breast Cancer

Carol M. O'Brien

Thesis submitted to Trinity College Dublin
for the degree of Doctor of Philosophy

2020

Thesis Supervisor: Prof. Adrian P. Bracken

Cancer Epigenetics Laboratory,

Department of Genetics,

University of Dublin,

Trinity College, Dublin 2

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work, except where otherwise acknowledged.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Carol O'Brien

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Adrian Bracken for giving me the opportunity to pursue a PhD in his lab and for his patience, advice and guidance over the past few years. I would also like to thank Mr. James Geraghty for his constant support and encouragement. I am very grateful to you both for entrusting me with this project.

I am incredibly grateful to the St. Vincent's Foundation and the Irish Research Council for providing funding to carry out this project.

I would also like to thank Karsten Hokamp and Mark Doherty for their excellent bioinformatics advice, and Prof. Kevin Mitchell and Prof. Dan Bradley for their helpful comments during my transfer.

My sincerest thanks to all members of the Bracken Laboratory, past and present, for being such fantastic lab mates: Darren Fitzpatrick, Eleanor Glancy, Rachel McCole, Orla Deevy, Marlena Mucha, Gerry Brien, Cheng Wang, Hannah Neikes, James Nolan, Dáire Gannon, Francesca Matra and Amaranta Armesto Jiménez for helping me over the last hurdle.

I am particularly grateful to Evan Healy and Eric Conway for their company in the early years and their constant advice and encouragement. Thanks also to Fatima Aloraifi for her hard work in setting up this project, and to Siobhan Purcell, Jane O'Sullivan, Mary-Kate Corbally, Sarah Hill and Tamara Alrawahneh for their assistance in the lab.

Finally, I would like to express my deepest thanks to my mother and father, Eleanor and Eugene, my siblings Dáithi and Willow, and to my wife Fiona. Thank you all for your constant love, support and encouragement.

Table of Contents

List of Figures.....	11
List of Tables	13
Summary	14
Abbreviations	15
1. Introduction	17
1.1. Breast cancer epidemiology and risk factors.....	18
1.2. Breast cancer susceptibility genes	19
1.3. The non-coding genome and its role in regulation of transcription.....	21
1.4. Relevance of the non-coding genome to cancer.....	25
1.4.1. Variants in promoters implicated in cancer	26
1.4.2. Variants in enhancers implicated in cancer	27
1.4.3. Genome-wide association studies	28
1.5. Identification of rare exonic variants in an Irish cohort of BRCAx familial breast cancer patients	30
1.6. Aim of this project.....	33
2. Materials and methods	35
2.1. Targeted sequencing of cases and controls.....	36
2.1.1. Selection of study subjects	36
2.1.2. Selection and sequencing of target regions.....	37
2.1.3. Bioinformatic analysis of sequencing data.....	38
2.2. Cell culture	39
2.3. Luciferase Reporter Assay	40
2.3.1. Cloning of DNA fragments containing SNV into pCR8 entry vector	40

2.3.2. Transformation and purification of pCR8 entry vectors from competent bacteria	41
2.3.3. Restriction digest and sequencing of pCR8 entry vector clones	41
2.3.4. Subcloning of SNV fragments from pCR8 entry vector to pGL3/pGL4.23 destination vector.....	41
2.3.5. Transformation and purification of pGL3/pGL4.23 entry vectors from competent bacteria	42
2.3.6. Restriction digest and sequencing of pGL3/pGL4.23 destination vector clones	42
2.3.7. Transfection of HEK293T cells with calcium phosphate.....	42
2.3.8. Transfection of MCF7 cells with EugeneHD	43
2.3.9. Measuring luminescence	43
2.4. Transcription factor motif analysis.....	44
2.5. CRISPR/Cas9-mediated genomic deletions	44
2.5.1. Generation of an MCF10A cell line expressing Cas9 via lentiviral transduction	44
2.5.2. Cloning and lentiviral transduction of gRNAs into MCF10 cells.....	45
2.5.3. Screening for edited loci	45
2.6. RNA preparation and real-time quantitative PCR (RT-qPCR).....	46
2.7. Compilation and annotation of the GWAS-associated SNPs	46
2.8. Enrichment of transcription factors at SNPs.....	48
2.9. Primers and Oligonucleotides	50
3. Identification of potentially functional rare variants in promoters and enhancers located near breast cancer-associated genes	55
3.1. Introduction.....	56
3.2. Results	58
3.2.1. Study design for the identification of rare germline SNVs in promoters and enhancers located near cancer genes.....	58
3.2.2. Prioritisation of case-enriched SNVs for functional evaluation	62

3.2.3. Overview of clinical and biological characteristics of 28 potentially functional non-coding regulatory SNVs.....	66
3.2.4. Familial breast cancer-associated variants modify transcription factor binding motifs	68
3.2.5. Familial breast cancer-associated variants disrupt transcription in an <i>in-vitro</i> reporter assay	70
3.3. Discussion	75
4. Functional investigation of variants and characterisation of their mechanism of action.....	81
4.1. Introduction.....	82
4.2. Results	83
4.2.1. Identification of probable target genes of each variant locus.....	83
4.2.2. Strategy for the CRISPR-mediated deletion of variant loci.....	86
4.2.3. Strategy to screen edited pools of MCF10A cells for CRISPR deletions.....	87
4.2.4. Deletion of regulatory loci and the effect on expression of breast cancer-related target genes.....	90
4.3. Discussion	92
5. Integrative functional analyses of common breast cancer GWAS SNPs at 172 loci	95
5.1. Introduction.....	96
5.2. Results	98
5.2.1. Recurrent breast cancer GWAS-associated SNPs in an Irish cohort of familial breast cancer patients	98
5.2.2. Annotation and functional analyses of 172 breast cancer GWAS loci.....	100
5.2.3. GWAS-linked SNPs disrupt CTCF sites and transcription factor binding sites in super-enhancers	105
5.2.4. Putative target genes of each candidate causal SNP.....	109
5.3. Discussion	111

6. General Discussion	115
6.1. Discussion	116
6.1.1. The function of non-coding variants in breast cancer	116
6.1.2. Technological and other limitations of the current study.....	117
6.1.3. Future work.....	119
6.1.4. Challenges remaining in the field of non-coding regulatory genomics	120
6.1.5. Non-coding regulatory regions as targets for therapy.....	121
6.2. Conclusions.....	124
7. References	125
8. Appendix	141
8.1. 255 non-coding variants identified in BRCAx cases.....	142

List of Figures

Figure 1.1: Breast cancer causes and susceptibility genes.....	21
Figure 1.2: Distribution of regulatory elements in a typical eukaryotic gene.....	24
Figure 1.3: A generic model of enhancer activation of its target gene.....	25
Figure 1.4: Model depicting how a SNV in an enhancer could affect transcription.....	27
Figure 1.5: The relationship between the allele frequency of a variant and the strength of its effect.	29
Figure 1.6: Study design to identify rare hereditary breast cancer pathogenic variants in the exonic regions of 312 cancer-associated genes.....	32
Figure 1.7: Overview of all potentially pathogenic rare exonic variants identified the Irish BRCAx cohort.	33
Figure 3.1: Study design for the identification of germline SNVs in promoters and enhancers located near breast cancer genes.....	61
Figure 3.2: Genomic distribution of 255 non-coding variants enriched in familial breast cancer cases.....	63
Figure 3.3: SNVs near breast cancer-related genes were identified in a familial breast cancer cohort.....	65
Figure 3.4: Overview of clinical and biological characteristics of 28 potentially functional non-coding regulatory SNVs.....	67
Figure 3.5: Familial breast cancer-associated SNVs at promoters and enhancers which result in an increase in transcription.....	72
Figure 3.6: Familial breast cancer-associated SNVs at promoters and enhancers which result in a decrease in transcription.....	73
Figure 3.7: Familial breast cancer-associated SNVs at promoters and enhancers which result in minimal disruption to transcription.....	74
Figure 4.1 Strategy for the CRISPR-mediated deletion of variant loci.....	87

Figure 4.2: Screening edited pools of MCF10A cells for CRISPR deletions.....	89
Figure 4.3: RT-qPCR of potential target genes after deletion of the SNV regulatory regions using CRISPR.....	91
Figure 5.1: Localisation of SNPs in 172 breast cancer GWAS loci.....	102
Figure 5.2: Enrichment of transcription factors binding at GWAS-linked SNPs in open chromatin.....	104
Figure 5.3: Overview of analysis for prioritising candidate causal GWAS SNPs.....	106
Figure 5.4: rs1744959 is located in a TAD boundary and disrupts a CTCF motif.....	107
Figure 5.5: rs1011623 is located in a TAD boundary and creates a CTCF motif.	108

List of Tables

Table 2.1: Publicly available files used to annotate GWAS-associated SNPs.....	48
Table 2.2: PCR Primers.....	50
Table 2.3: gRNA sequences.....	51
Table 2.4: RT-qPCR mRNA Primers.....	53
Table 3.1: Prioritised potentially pathogenic variants.....	64
Table 3.2: Familial breast cancer-associated variants modify transcription factor binding motifs.....	69
Table 3.3: Familial breast cancer-associated variants disrupt transcription in an <i>in-vitro</i> reporter assay.....	71
Table 4.1: Potential target genes for each SNV locus.....	85
Table 5.1: Common SNPs are enriched in cases versus controls at two breast cancer GWAS loci.....	99
Table 5.2: Putative target genes of each candidate causal SNP.....	110

Summary

Breast cancer is the most common cancer in Irish women, with almost 3,000 cases diagnosed every year. It often runs in families, and some women inherit a greater risk of developing the disease. Landmark studies on these families led to the discovery of breast cancer susceptibility genes such as *BRCA1* and *BRCA2*. However, mutations in *BRCA1*, *BRCA2* and other susceptibility genes are only detected in ~25% of women with familial breast cancer. Despite extensive efforts to uncover the genetic susceptibility to breast cancer, much of the heritability remains unexplained.

To address this issue, we hypothesised that pathogenic germline mutations may be located in the non-protein-coding, regulatory regions of the genome. We carried out targeted sequencing of promoter and candidate enhancer regions of 143 cancer-associated genes in 104 high-risk non-*BRCA1/2* Irish familial breast cancer cases and 101 geographically-matched controls.

We found that case-enriched non-coding variants were capable of disrupting transcription in a luciferase reporter assay and that they are predicted to disrupt transcription factor binding sites. Furthermore, we show that deletion of an enhancer region containing a rare case-enriched single nucleotide variant (SNV), which is in proximity to *PIK3CA*, appears to result in a decrease in *ZMAT3* expression.

Our results demonstrate that interrogation of the non-protein-coding region in proximity to cancer-associated genes is a worthwhile endeavour. We anticipate this study to be a starting point for yet more sophisticated modelling of the effect of the SNV through the generation of isogenic cell lines containing just the single nucleotide mutated in patients. We hope that this project will make an important contribution to both our knowledge about the genetic causes of breast cancer, as well as adding to the discussion surrounding the functional validation of non-coding variants.

Abbreviations

3C	Chromosome Conformation Capture
4C	Circularized Chromosome Conformation Capture
5C	Chromosome Conformation Capture Carbon Copy
BCAC	Breast Cancer Association Consortium
BRCAX	Familial breast cancer case without a mutation in <i>BRCA1/2</i>
CCLE	Cancer cell line encyclopedia
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
COGS	Collaborative Oncological Gene-environment Study
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRV	Credible risk variant
CTCF	CCCTC binding factor
DHS	DNase hypersensitive site
DRIVE	Discovery, Biology and Risk of Inherited Variants in Breast Cancer Consortium
EMT	Epithelial-mesenchymal transition
ENCODE	Encyclopedia of DNA Elements
Enh	Enhancer
eQTL	Expression quantitative trait loci
eRNA	Enhancer RNA
GATK	Genome Analysis Toolkit
GMAF	Global Minor Allele Frequency
GTEx	Genotype Tissue Expression
GWAS	Genome-wide association study
H1-ESC	Human embryonic stem cells
H3K27ac	Histone 3 lysine 27 acetylation

H3K4me1	Histone 3 lysine 4 mono-methylation
H3K4me3	Histone 3 lysine 4 tri-methylation
HEK293T	Human Embryonic Kidney cells (with SV40 T antigen)
HMEC	Human Mammary Epithelial Cells
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
mRNA	Messenger RNA
PCA	Principal Component Analysis
PCR	Polymerase chain reaction
Pro	Promoter
PRS	Polygenic risk score
RT-qPCR	Reverse transcription quantitative PCR
SE	Super-enhancer
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TAD	Topological associating chromatin domain
TCGA	The Cancer Genome Atlas
TFBS	Transcription factor binding site
TF	Transcription Factor
TPM	Transcripts per kilobase million
TSS	Transcriptional start site
UTR	Untranslated region

1. Introduction

1.1. Breast cancer epidemiology and risk factors

Breast cancer is the most common cancer in women and the second most common cancer overall, both in Ireland and worldwide (Bray et al., 2018). In 2018, there were approximately 2.1 million new cases of breast cancer diagnosed worldwide; 3,334 of which were in Ireland (Bray et al., 2018). There was an average of 2,922 invasive cases of breast cancer diagnosed in women each year between 2012-2014 in Ireland, making up 30.1% of all invasive cancers in females (National Cancer Registry, 2017). As of 2018, Ireland had the 8th highest breast cancer incidence among women in Europe, and the 7th highest mortality rate (Bray et al., 2018). Breast cancer survival is high, with an 82.9% 5-year survival rate for females in Ireland, yet it was still the second leading cause of cancer deaths in women between 2012-2014, with an average of 709 deaths per year (National Cancer Registry, 2017).

There several known risk factors for breast cancer development which span genetic, environmental and lifestyle factors. Breast cancer becomes more likely with age; the incidence doubles approximately every 10 years until menopause when the increase slows (McPherson et al., 2000). Apart from age, the most important risk factor is family history (Amir et al., 2010), which suggests that searching for germline genetic risk factors is critical, as will be discussed in detail below.

There are also known hormonal and reproductive risk factors for developing breast cancer. These include exposure to endogenous estrogens from early menarche (age <12 years) and/or late menopause (age >55 years), exposure to exogenous estrogens from use of the oral combined contraceptive pill and hormone replacement therapy, as well as reproductive risk factors such as nulliparity and older age at first pregnancy (Amir et al., 2010). There are also a number of other environmental and lifestyle risk factors, such as diet, exercise, obesity, alcohol and smoking (McPherson et al., 2000).

1.2. Breast cancer susceptibility genes

The presence of a significant family history of breast cancer is an important risk factor. The risk of developing breast cancer increases with increasing numbers of affected first-degree relatives (Amir et al., 2010). One study estimated that the probability of developing breast cancer for women between the ages 20 and 80 in more developed countries with no affected relatives was 7.8%; this rises to 13.3% and 21.1% for women with one and two affected relatives respectively (Collaborative Group on Hormonal Factors in Breast Cancer, 2001).

In 1990, the first high-risk breast cancer gene *BRCA1*, was discovered through linkage analysis (Hall et al., 1990). Four years later, a second high-risk breast cancer gene was discovered and named *BRCA2*. It was discovered through linkage analysis of 15 high-risk breast cancer families that were not linked to *BRCA1* (Wooster et al., 1994). *BRCA1* and *BRCA2* are both highly penetrant; 79.5% and 81% respectively (Evans et al., 2008).

Familial breast cancer accounts for ~5-15% of all breast cancer cases (Shiovitz & Korde, 2015). Out of these familial cases, variants in *BRCA1* and *BRCA2* account for only about 15-20% (Anglian Breast Cancer Study Group, 2000; Peto et al., 1999). Additionally, there are other known high-penetrance susceptibility genes such as *TP53*, *PTEN*, *CDH1*, *STK11*. Inherited mutations in *TP53* and *PTEN* are rare, but are associated with familial Li-Fraumeni syndrome and Cowden's syndrome respectively, which both include a high risk of breast cancer, among other cancers (Shiovitz & Korde, 2015). Together, these known high-penetrance genes are thought to account for no more than 25% of cases (Antoniou & Easton, 2006). Moderate- or intermediate-penetrance susceptibility genes include *ATM*, *CHEK2*, *PALB2*, *BRIP1*, *RAD50*, *NBS1* and *MRE11A*, and are thought to account for a further 3-5% (Wendt & Margolin, 2019). In addition, a number of low-penetrance common variants from genome-wide association studies (GWASs) have been identified, and are estimated to account for a further 18% of the familial risk (Michailidou et

al., 2017). Despite these discoveries, <30% of all familial breast cancer cases have an identified causal variant (Shiovitz & Korde, 2015), and most of the familial risk of breast cancer remains unexplained.

The remaining cases of breast cancer may be caused by variants with effect sizes too small for GWASs to detect, or be rare variants that are too rare to be tagged by common SNPs included in GWAS arrays. They could also be hidden in a polygenic model of breast cancer, where a large number of loci each exert a small effect. Or, as we will hypothesise in this project, they may lie in the relatively unexplored non-protein-coding regions of the genome.

In Ireland, the Department of Clinical Genetics at Our Lady's Children's Hospital, Crumlin is the body responsible for providing diagnostic services for a range of inherited disorders, including familial breast cancer. They provide genetic testing of the breast cancer-related genes, *BRCA1* and *BRCA2*, and subsequent counselling. Despite these efforts however, the majority of women with familial breast cancer will test negative for the *BRCA1* and *BRCA2* genes. These women (hereafter referred to as BRCAx) are the focus of this project.

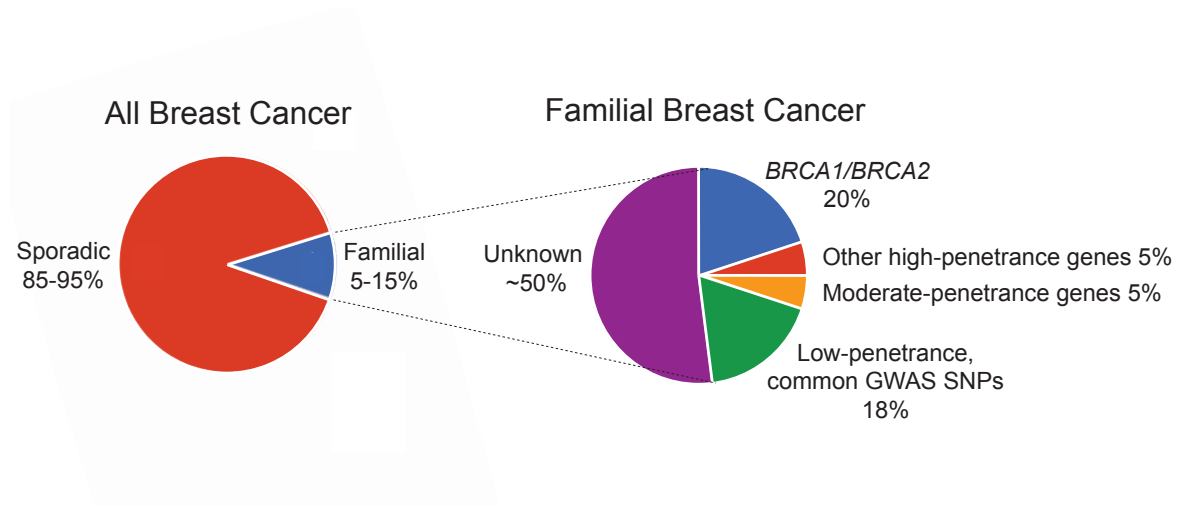


Figure 1.1: Breast cancer causes and susceptibility genes.

Familial breast cancer constitutes about 5–15% of total breast cancer. Variants in the *BRCA1* and *BRCA2* genes only account for about 20% of the familial risk. Other high-penetrance genes such as *PTEN*, *TP53*, *CDH1* and *STK11* account for a further 5%. Moderate-penetrance genes such as *CHEK2*, *BRIP1*, *ATM* and *PALB2* also account for approximately 5%. The 172 GWAS loci identified to date are estimated to account for a further 18% of relative familial breast cancer heritability.

1.3. The non-coding genome and its role in regulation of transcription

The human genome is made up of approximately 3 billion DNA base pairs. There are approximately 20,000 to 25,000 protein-coding genes, which together, make up ~1-2% of the genome (International Human Genome Sequencing Consortium, 2004).

Approximately 50% of the genome is made up of repetitive elements, which can be broadly divided into two classes: interspersed transposable elements (~45%, examples include the short interspersed nuclear element Alu and the long interspersed nuclear element L1) and short tandem repeats (such as microsatellites and telomeres). The remaining non-coding DNA consists of pseudogenes, non-coding RNA genes (such as

tRNA, rRNA and lncRNA), introns (~26%) and regulatory elements, in addition to many regions of no or unknown function (International Human Genome Sequencing Consortium, 2001). Regulatory elements can be further subdivided into promoters, enhancers, silencers and insulators.

Promoters are the stretches of DNA directly upstream of genes that define where transcription of that gene begins. Transcription initiation is mediated at the promoter by RNA polymerase and other transcription factors (Haberle & Stark, 2018). Active promoters are marked by H3K27ac and H3K4me3 (Heintzman et al., 2009), while repressed promoters are marked by H3K27me3 (Di Croce & Helin, 2013).

Enhancers are segments of non-coding DNA that are involved in regulating transcription of their target gene (Shlyueva et al., 2014). They allow genes to be more precisely regulated in specific cell types and at different stages of development. Enhancers were first identified in 1981 when it was found that a repeating sequence of 72 bases cloned from the small DNA virus SV40, was capable of stimulating the β -globin gene to produce 200 times more transcripts (Banerji et al., 1981). This 'enhancer' could function even when it was located thousands of bases away from the β -globin gene, and regardless of whether its sequence was inserted forwards or backwards - characteristics which we now know are typical of all enhancers. It was subsequently discovered that this 72bp region contained binding sites for two transcription factors, AP-1 and Sp1 (Lee et al., 1987). This result helped establish that enhancers are composed of clusters of sequence-specific transcription factor binding sites (Spitz & Furlong, 2012).

The ENCODE consortium has identified ~399,000 enhancer-like regions in the human genome (The ENCODE Project Consortium, 2012). This far outnumbers coding genes and suggests that one gene may be regulated by many enhancers. Identifying which gene an enhancer regulates is a challenging problem. Enhancers do not occur at a defined distance from their target gene, and target genes are often not the closest linear gene as

enhancers can be located at large distances upstream or downstream of their target (Shlyueva et al., 2014). One gene may have many enhancers, which may be active at different times, in different locations and in different cell types. Alternatively, one enhancer may regulate many genes. They can act on gene promoters on the same chromosome (in *cis*) or on a different chromosome (in *trans*) (Shlyueva et al., 2014).

Active enhancers are located in regions of open chromatin that are depleted of nucleosomes. These regions are sensitive to digestion by DNase, and so can be identified by a DNase hypersensitivity assay (Elgin, 1984; Gross & Garrard, 1988). These enhancer open regions are flanked by nucleosomes marked by H3K4me1 and H3K27ac: these enrichments are cell type-specific (Heintzman et al., 2009). Enhancers are often also bound by transcription factors (TFs), such as p300, Mediator and cohesin. Active enhancers have also been reported to be transcribed bidirectionally, producing enhancer RNA (eRNA) (Kim et al., 2010), though the function of these transcripts is unclear. Repressed enhancers are marked with H3K27me3 (Di Croce & Helin, 2013) and enhancers in silent heterochromatic regions are marked with H3K9me3 (Shlyueva et al., 2014).

Super-enhancers (SEs) are clusters of enhancers that are enriched for enhancer-associated chromatin marks, such as H3K27ac, Mediator and master transcription factors (Whyte et al., 2013). They are associated with key cell type-specific genes and thus are thought to play critical roles in the development and maintenance of cellular identity expression (Whyte et al., 2013). One study found that a neuroblastoma GWAS-associated SNP in a super-enhancer of LMO1 disrupts a GATA3 motif and is associated with lower LMO1 expression (Oldridge et al., 2015).

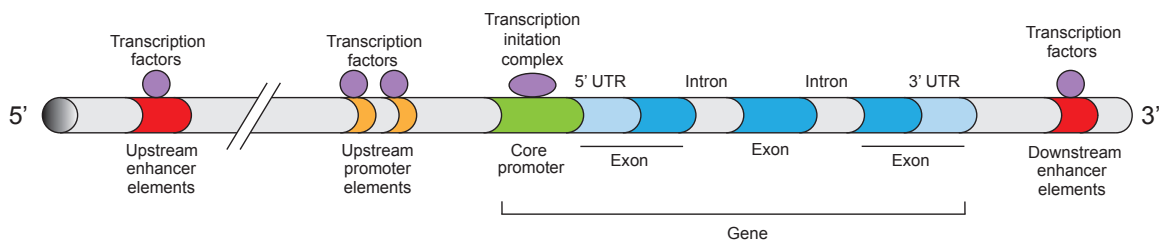


Figure 1.2: Distribution of regulatory elements in a typical eukaryotic gene.

A typical eukaryotic gene consists of exons, introns, a core promoter, upstream promoter elements and enhancers which can be located up- or downstream of a gene. The core promoter is the minimal sequence required to initiate transcription and is the point of assembly of the transcription initiation complex. Both upstream promoter elements and enhancers often contain binding sites for transcription factors and are capable of increasing the rate of transcription.

Enhancers are thought to physically interact with the promoters of their target genes through the three-dimensional organisation of chromatin (Sanyal et al., 2012). The most prominent model for this interaction is the 'looping' model, where enhancers are brought into the proximity of their target promoter by 'looping' around to the promoter (Dorsett, 1999) (Figure 1.3). Alternative models include 'tracking' of the enhancer across the intervening DNA towards the promoter and 'linking' of the enhancer and promoter through the spread of chromatin modifications (Bulger & Groudine, 1999). In reality, much of the exact details of how an enhancer contacts its promoter remains unknown. Interestingly, a new paradigm seems to be emerging – phase separation has now been touted as a possible new mechanism, where large numbers of transcription factors and coactivators compartmentalize and condense, forming separate liquid droplets – i.e. phase separation. Such condensates would incorporate both enhancers and promoters (Cho et al., 2018).

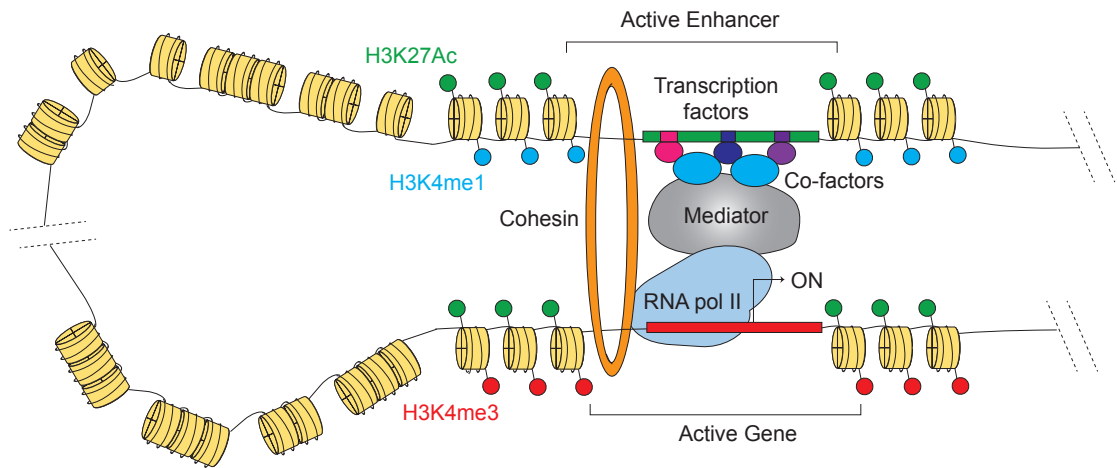


Figure 1.3: A generic model of enhancer activation of its target gene.

Enhancers contain binding sites for transcription factors and upregulate expression of their associated genes. Recruitment of the Mediator complex to the enhancer results in chromatin looping which brings the enhancer into the proximity of its target promoter, thus driving expression. Recruitment of cohesin is thought to stabilise the DNA loop. Flanking nucleosomes are marked by specific histone modifications: H3K27ac at active chromatin, H3K4me1 at enhancers, and H3K4me3 at promoters.

1.4. Relevance of the non-coding genome to cancer

Understanding how variants in the non-coding genome function is an ongoing challenge. In protein-coding regions, we understand the grammar of transcription and translation. No such guide exists for non-coding regions. Instead, the best we can do is to use biochemical and epigenomic annotation to guide us. There has been great progress

recently in developing technologies for functional characterisation of the whole genome. The ENCODE (Encyclopedia of DNA Elements) Consortium published their landmark paper in September 2012 claiming that 80% of the genome was functional (The ENCODE Project Consortium, 2012). It proved to be a very controversial paper (Graur et al., 2013) and has come under much scrutiny for two main reasons. Firstly, ENCODE chose to adopt a broad definition of the term “functional”. For their purposes, it was any region that showed a biochemical activity signal from an assay, such as a transcription factor binding site, transcription, had a histone modification present, DNaseI hypersensitive sites, or was an exon. A second criticism is that many of the associations found may merely be noise, or chance associations. A major caveat of the ENCODE assays is that these marks are often merely correlative. There may be much transient, non-functional contact between proteins such as transcription factors and DNA. Even after annotation with ENCODE data, extensive further experiments are required to elucidate the mechanisms of action of regulatory non-coding regions.

1.4.1. Variants in promoters implicated in cancer

Non-coding germline variants located in promoters may be tumorigenic, as exemplified by the high-penetrance *TERT* promoter mutations found in melanoma. These are single nucleotide variants (SNVs) which result in the creation of a binding site motif for the ETS family of transcription factors, which in turn results in increased expression of the *TERT* oncogene (Horn et al., 2013; Huang et al., 2013). In an earlier study, a T-to-G SNP in the promoter of *MDM2* was found to enhance binding of Sp1, a transcriptional activator, to the promoter of *MDM2* resulting in increased levels of expression. As MDM2 is a negative regulator of p53, this SNP can attenuate p53 tumour suppression and result in accelerated tumour formation (Bond et al., 2004).

1.4.2. Variants in enhancers implicated in cancer

Alterations in enhancers have been shown to contribute to phenotypic traits such as blonde hair (Guenther et al., 2014) and blue eyes (Sturm et al., 2008), as well as diseases such as β -thalassemia (Kioussis et al., 1983), pancreatic agenesis (Weedon et al., 2014) and cancer. A putative enhancer located upstream of the *MYC* oncogene contains a colon-cancer risk SNP. Upon deletion of the enhancer containing the risk SNP, mice were resistant to intestinal tumour formation (Sur et al., 2012). Another example can be found in T-cell acute lymphoblastic leukaemia (T-ALL), where somatic insertions in an enhancer upstream of the *TAL1* oncogene have been shown to lead to a *de novo* MYB binding site, in turn leading to aberrant expression of TAL1 (Mansour et al., 2014).

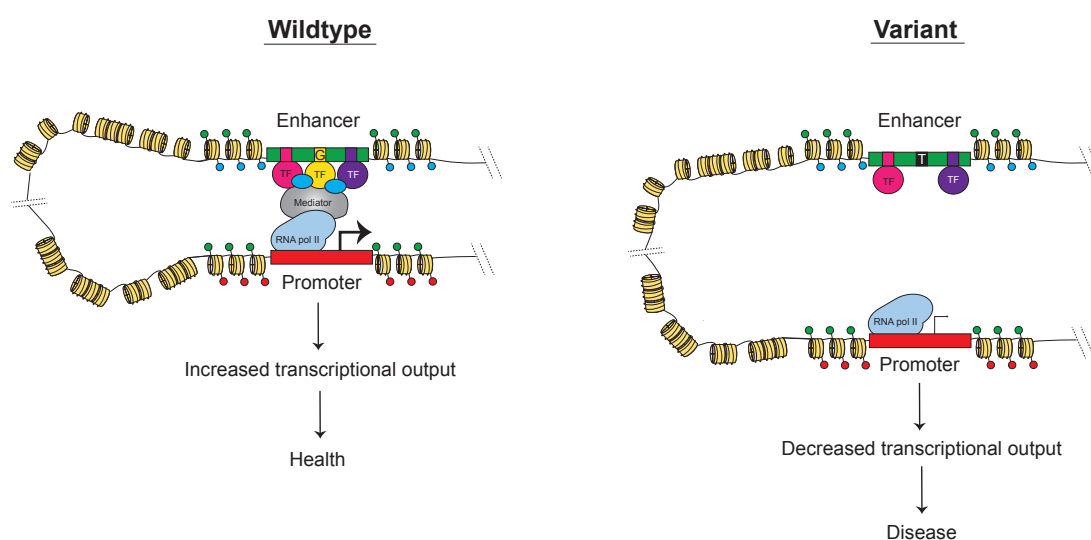


Figure 1.4: Model depicting how a SNV in an enhancer could affect transcription.

Within the enhancer, the reference G allele of the SNV binds a transcription factor (yellow circle) and activates the gene, resulting in normal, healthy levels of gene expression. However, the alternative T allele is no longer capable of binding the transcription factor, resulting in a less efficient enhancer with decreased transcription of the gene, and an increase in disease risk. Figure adapted from Palstra & Grosveld (2012).

1.4.3. Genome-wide association studies

Genome-wide association studies (GWASs) represented an important step forward from family-based linkage studies, which have limited power in complex diseases, and candidate gene studies, which are limited by our prior understanding of the underlying biology. GWASs are capable of detecting common variants, which are variants with a minor allele frequency (MAF) of greater than 5%. GWASs are based on the 'Common Disease – Common Variant' hypothesis, which posits that common variants may contribute significantly to the risk of common diseases such as cancer and diabetes (Lander, 1996).

Whereas common variants appear to have a modest effect size, in contrast, rare variants may have large effect sizes and could contribute significantly to missing heritability in disease susceptibility (Figure 1.5) (Manolio et al., 2009). Variants with large effect sizes are thought to reduce reproductive fitness, thus keeping them rare within the population (Manolio et al., 2009). However, exceptions do exist, the best example being the *APOE4* allele in which a common variant with a large effect size influences the risk of developing Alzheimer's disease (Saunders et al., 1993). Rare variants are not likely to be found by GWASs due to their low allele frequency; they are normally detected by next-generation sequencing methods.

Despite the success of GWAS in identifying SNPs which are statistically associated with many traits and diseases, there are criticisms and limitations of GWASs. In particular, there has been much debate around the limits of penetrance for common variants, and the fact GWASs have been unable to account for much of the heritability of traits and diseases (Manolio et al., 2009). Heritability is the proportion of variability that is due to genetic factors. For example, GWASs have been able to explain just 18% of the heritability of breast cancer and just 5% of heritability for height (Visscher, 2008). One recent pre-print used SNPs obtained from whole-genome sequencing on over 20,000

individuals to estimate the heritability for height and body mass index (Wainschtein et al., 2019). The authors concluded that for complex diseases, missing heritability is accounted for by rare variants, specifically rare variants occurring in protein-coding regions of low LD. Rare variants ($MAF < 0.1$) explained 0.54 of heritability for height, and 0.51 for BMI. Additionally, approximately 95% of SNPs identified by GWAS are located in non-coding regions with approximately 40% located in introns (Freedman et al., 2011; Maurano et al., 2012). This adds significantly to the difficulty in attempting to decipher the functionality of identified SNPs.

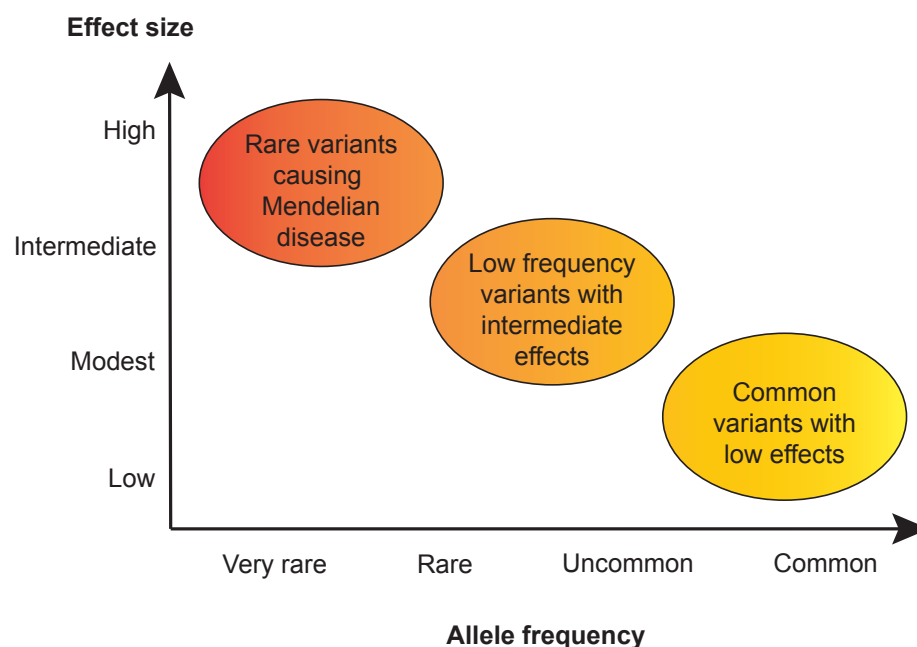


Figure 1.5: The relationship between the allele frequency of a variant and the strength of its effect.

Common variants, which appear to have modest effect sizes, can be detected through GWAS. In contrast, rare variants may have large effect sizes, and could contribute significantly to missing heritability in disease susceptibility. Figure adapted from Manolio et al. (2009).

GWASs have been carried out for most types of common cancers. The first breast cancer GWAS was published in 2007 and involved over 4,000 cases and 4,000 controls (Easton et al., 2007). This pioneering study identified five risk loci, four of which contained the genes *FGFR2*, *TNRC9*, *MAP3K1* and *LSP1*. Various consortia were founded to pool resources and carry out more extensive studies with more power, including the Breast Cancer Association Consortium (BCAC); Discovery, Biology and Risk of Inherited Variants in Breast Cancer Consortium (DRIVE); and the Collaborative Oncological Gene-environment Study (COGS), which even developed their own custom genotyping array. The most recent breast cancer GWAS was published in 2017 and was a large study involving over 120,000 cases and 100,000 controls of European ancestry (Michailidou et al., 2017). This study combined new cases and controls with those from previous GWASs and identified 65 new loci in addition to replicating 49 previously identified ones. To date, 172 loci have been identified from breast cancer GWASs, and are estimated to account for approximately 18% of the familial relative risk (Michailidou et al., 2017).

1.5. Identification of rare exonic variants in an Irish cohort of BRCAx familial breast cancer patients

Previous work in our laboratory sought to identify the missing heritability in familial breast cancer (Aloraifi et al., 2015). Our objective was to systematically identify rare, probably pathogenic variants in familial cases of breast cancer without *BRCA1/2* mutations (BRCAx). To do this, we sequenced the exons of 312 cancer genes in 104 high-risk Irish BRCAx cases and 101 geographically matched unaffected controls (Figure 1.6). The 312 genes included 18 high- and moderate- susceptibility breast cancer genes, 88 genes identified through breast cancer GWASs, 33 genes somatically mutated in breast cancer, 48 genes silenced through hypermethylation in breast cancer, genes implicated in a whole-exome study on hereditary breast cancer cases, genes reported to vary in copy

number in breast tumours, and genes encoding chromatin regulators found to be frequently mutated in many cancer types.

This approach identified rare, loss of function mutations in several well known high- and moderate-susceptibility genes, including *ATM* (~ 5%), *RAD50* (~ 3%), *CHEK2* (~ 2%), *TP53* (~ 1%), *PALB2* (~ 1%), and *MRE11A* (~ 1%). Additionally, we also identified mutations in several GWAS-associated genes, including *MAP3K1* (~1%), *CASP8* (~1%), and *RAD51B* (~1%), as well as in other candidate genes (e.g. *ERBB2* and *ZNF217*) (Figure 1.7). Despite the identification of these mutations, there still remains a large proportion (~70%) of BRCAx patients with no identifiable potentially causal mutation (Figure 1.7D). One possible reason is that the causal variants lie in non-protein-coding, regulatory regions, something that was not explored in this initial study. This hypothesis now forms the basis for this thesis.

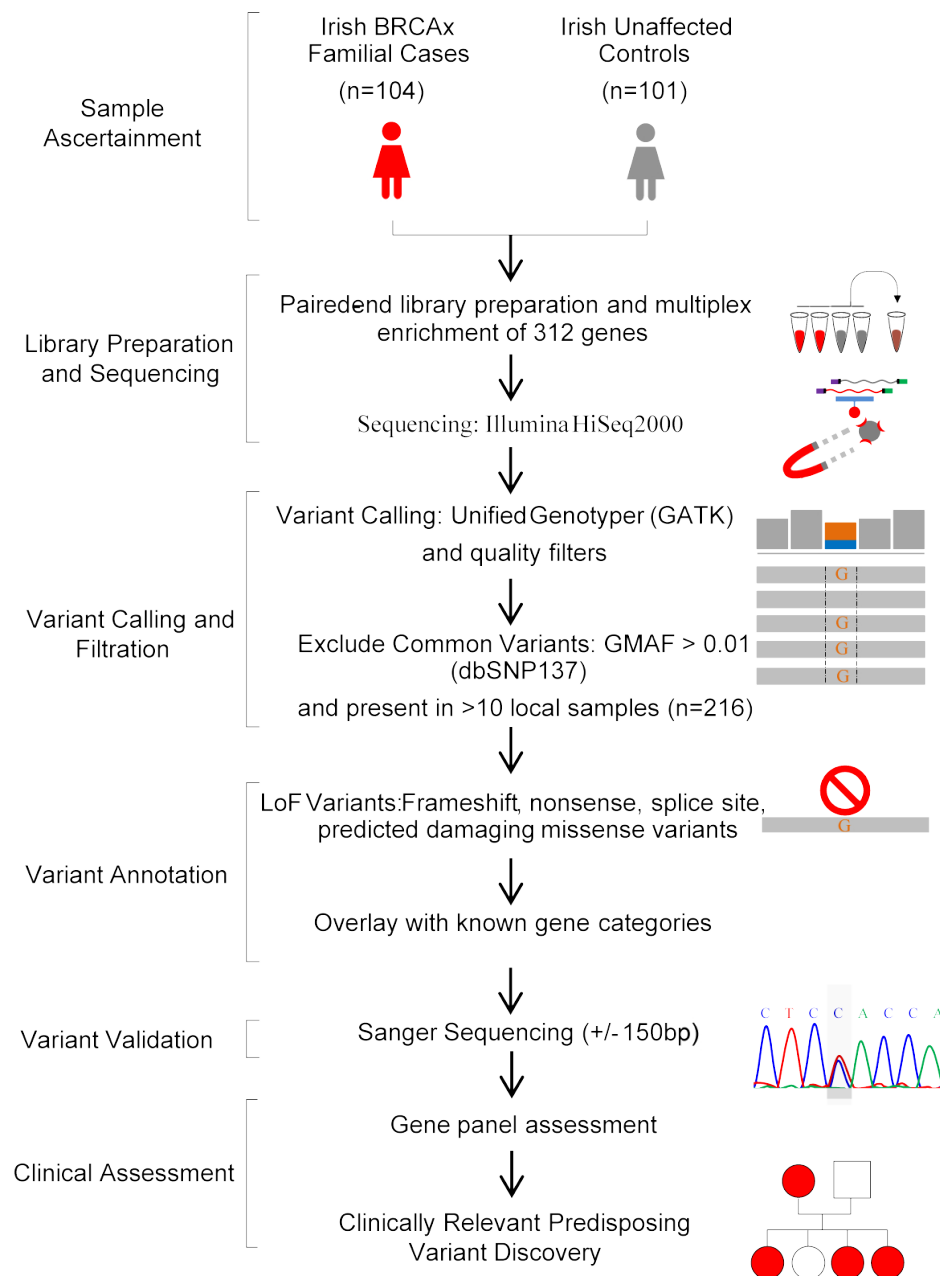


Figure 1.6: Study design to identify rare hereditary breast cancer pathogenic variants in the exonic regions of 312 cancer-associated genes.

Illumina's TruSeq enrichment was used to capture the exons of 312 genes with 104 BRCAx Cases and 101 geographically matched controls. Rare loss of function variants were retained for analysis. Figure adapted from Aloraifi et al. (2015).

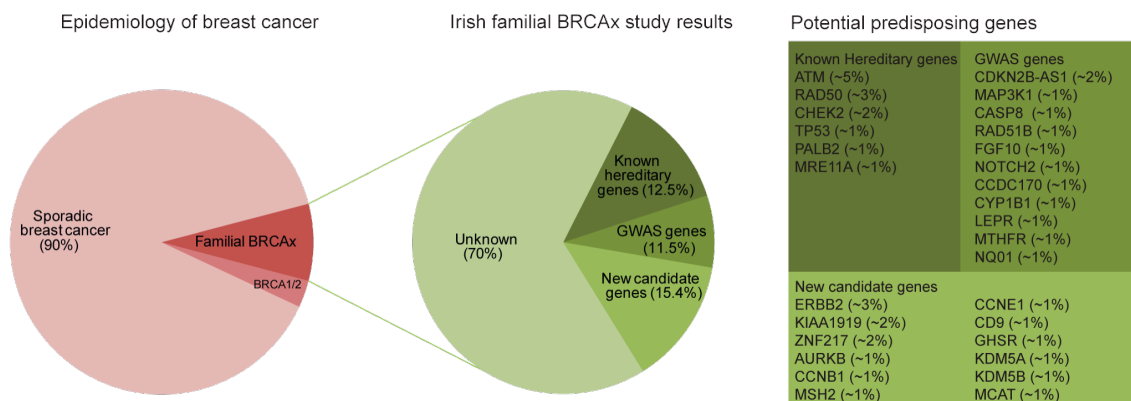


Figure 1.7: Overview of all potentially pathogenic rare exonic variants identified the Irish BRCAx cohort.

Venn diagram summary of familial breast cancer in Ireland based on the results of this study. In total, 11% of all breast cancers in Ireland are thought to be attributable to a hereditary predisposition, and 25% of those with a hereditary predisposition can be attributed to *BRCA1* or *BRCA2* mutations. In green are the proportions of the Irish BRCAx samples with rare potentially pathogenic mutations detected in this study. Figure from Aloraifi et al. (2015).

1.6. Aim of this project

The aim of this project is to investigate the missing heritability of familial breast cancer, by identifying SNVs and small indels in non-coding regions of the genome that are potentially pathogenic in familial breast cancer. We hypothesized that SNVs within promoter and enhancer regions of cancer-associated genes are capable of altering transcription of their target genes, leading to an increased risk of developing breast cancer. This project is a continuation of the work described in Section 1.5. Whereas previously, we had focused on rare exonic variants, we now expanded our analysis to include both rare and common non-coding variants.

In chapter 3, I prioritised the SNVs detected in our sequencing study and assessed their possible effects on transcription. In chapter 4, I deleted the SNV locus and assessed the impact on potential target genes. Finally, in chapter 5, I focused on common variants identified both in our sequencing study and through GWAS, and identified some likely functional variants.

2. Materials and methods

2.1. Targeted sequencing of cases and controls

The work in section 2.1 was carried out by Fatima Aloraifi as part of her PhD project. This work is described in detail in her PhD thesis (Aloraifi, 2014) and resulting paper (Aloraifi et al., 2015) but is repeated briefly here for clarity.

2.1.1. Selection of study subjects

104 unrelated women with a family history of breast cancer were recruited to this study. The inclusion criteria were: 1) a personal history of breast cancer, 2) an age of onset less than 60 years old, and 3) having either at least 2 affected first degree relatives with breast cancer or for probands with bilateral breast cancer, one affected relative. All 104 probands were confirmed to be BRCAx by screening for truncating variants of *BRCA1* or *BRCA2* using bidirectional sequencing and multiplex ligation-dependent probe amplification.

Additionally, 101 anonymous age-, sex-, and geographically-matched controls with no known history of disease were specifically selected as negative controls for this study and obtained from blood bank donors provided by the Trinity College Dublin Biobank. We also included 6 *BRCA1* mutation-positive and 5 *BRCA2* mutation-positive cases. These positive controls were provided by the National Centre for Medical Genetics, and were not included in the discovery cohort.

All subjects provided written informed consent and ethical approval was granted by the Research and Ethics Committee at Our Lady's Children's Hospital, Crumlin, Dublin, Republic of Ireland.

2.1.2. Selection and sequencing of target regions

Firstly, a literature search was performed to compile a list of candidate genes for sequencing. This list of genes included:

- known hereditary breast cancer genes (Walsh et al., 2010)
- genes that are reported to be epigenetically silenced in more than 20% of breast tumour samples according to the Pubmeth database (Ongenaert et al., 2008)
- genes implicated from genome-wide association studies of breast cancer (Peng et al., 2011; Zhang et al., 2011)
- genes reported to be somatically mutated in breast tumours (Forbes et al., 2011)
- nine other potentially interesting genes.

To identify potential promoter regions of these genes, H3K4me3 positive regions, DNase hypersensitive sites, and “active promoter” regions from ENCODE’s chromatin segmentation track were identified in human mammary epithelial cells (HMEC) and human embryonic stem cells (H1-ESC). To identify potential active enhancer regions of these genes, H3K27ac positive regions were identified in the HMEC, ES, MCF7 and T47D cell lines, and restricted to a region of 200 kb upstream and downstream of the gene’s transcriptional start in HMEC. Haplotype blocks of breast cancer GWAS SNPs were also identified, and any of these regions falling on “strong enhancer” or “weak enhancer” chromatin segmentation annotations were included. Finally, repeatmasker (Smit et al., 2010) was used to identify all repetitive DNA sections which were then removed as they are known to cause ambiguities in read alignment (Smit, 1996; Treangen & Salzberg, 2012).

DNA probes for the above-mentioned candidate gene exons, and potential promoter and enhancer regions were designed using DesignStudio aligning to the hg19 human reference genome. Genomic DNA was extracted from blood leucocytes. DNA libraries

were prepared using TruSeq v2 Sample Preparation kits and enriched using TruSeq Custom Enrichment kit, according to the manufacturer's protocols.

Initially, a pilot study of 24 samples (12 controls, 1 case and 11 positive controls) was sequenced with single-end 70-bp reads on the Illumina GenomeAnalyzer II with TrinSeq (St. James's Hospital). The remaining 192 samples were subsequently sequenced with paired-end 50-bp reads on the Illumina HiSeq 2000 platform with BGI Genomics (Hong Kong).

2.1.3. Bioinformatic analysis of sequencing data

Adapters were removed from the raw sequence reads and trimmed reads were aligned to the human reference genome (hg19) using Burrows-Wheeler Aligner (BWA) version 0.7.022. Aligned sequence reads were sorted, filtered and compressed using SAMtools version 1.4, and artefact PCR duplicates were removed using the MarkDuplicates tool in Picard. Local realignments, indel realignment, and base recalibration were performed using the Genome Analysis Toolkit (GATK, version 2.7-1). Poor quality alignments with a MAPQ of less than 20 were removed.

Unified Genotyper (GATK) was used to call variants in samples that overlapped the target region by +/- 10 bp (to detect splice-site variants). Variants were annotated with quality metrics and known variants from dbSNP137. Variant quality score recalibration (GATK) was used to build an adaptive error model using a set of likely true variant sites (HapMap, 1000 Genomes, dbSNP137, Mills and 1000 Genomes gold standard for Indels) to estimate the likelihood of a true versus a false positive variant.

Variants were then split into SNPs and indels. The following filters were used, as recommended by the Broad Institute's best-practice guidelines for smaller datasets: for

SNPs: QD < 2.0, MQ < 40, FS > 60.0, HaplotypeScore > 13.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0; for Indels: QD < 2.0, ReadPosRankSum < -20.0, Inbreeding Coeff < -0.8, and FS > 200.0. In addition, variants falling on repeat elements (www.repeatmasker.org) and having a depth of coverage (DP) < 10x were filtered for and removed.

In order to determine the extent of population structure in our data, we merged called variants with genotype data from selected African, European and Asian samples (n = 1105, ~2.4 M SNPs) from the 1000 Genomes project. This resulted in a dataset with ~1000 SNPs on which we performed principal component analysis (PCA) with SMARTPCA. European ancestry was confirmed except for one case with Asian ancestry which was removed from further analysis. Common known variants with a global minor allele frequency (GMAF) of greater than 1% were filtered out. Variants found to be common within the 216 samples sequenced (>10 samples) were also filtered out, as those were seen as population-specific common variants. Variants were then separated into protein-coding and non-coding regions; those in protein-coding exonic regions were further described in Aloraifi et al., 2015, while variants located in the non-coding potential promoter and enhancer regions were selected for analysis in this project.

2.2. Cell culture

Human embryonic kidney HEK293T cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, Sigma) supplemented with 10% (v/v) fetal bovine serum (FBS, Gibco), 100 U/mL penicillin (Gibco) and 100 U/mL streptomycin (Gibco).

Breast cancer MCF7 cells were cultured in Dulbecco's Modified Eagle Medium DMEM/F12 media (Sigma), supplemented with 10% FBS, 100 U/mL penicillin (Gibco) and 100 U/mL streptomycin (Gibco).

Mammary epithelial MCF10A cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, Sigma) supplemented with 5% (v/v) horse serum (Sigma), 10 µg/ml insulin (Sigma), 0.5 µg/ml hydrocortisone, 10 ng/ml cholera toxin, 10 ng/ml epidermal growth factor (Peprotech), 100 U/mL penicillin (Gibco) and 100 U/mL streptomycin (Gibco).

2.3. Luciferase Reporter Assay

2.3.1. Cloning of DNA fragments containing SNV into pCR8 entry vector

The DNA fragments surrounding the SNVs were PCR amplified from 100 ng of patient DNA using the Platinum PCR SuperMix High Fidelity (Invitrogen) and a 10 uM solution of primer pairs as listed in Table 2.2. Triplicate reactions were performed with 25 cycles of PCR amplification. Triplicate reactions were pooled and 20 ul of the amplified DNA was run on a 0.8% agarose gel to confirm correct amplicon size. The rest of the PCR product was purified using a Nal-based silica bead DNA extraction method. 3' A overhangs were added to the PCR product by incubating at 72°C for 30 min with 10 mM dATP, 10X PCR Buffer with Mg⁺ and Tag Polymerase (Roche). Amplified fragments were cloned into a pCR8 entry vector using the pCR[™]8/GW/TOPO® TA Cloning Kit by incubating at 22°C for 30 min with salt solution and the pCR8 vector.

2.3.2. Transformation and purification of pCR8 entry vectors from competent bacteria

Plasmids were transformed into One Shot Top10 Chemically Competent E. coli cells by heat shock at 42°C for 30 s. Cells were then left to recover in S.O.C. media shaking for 1 h at 37°C, before being plated on 100 ug/mL spectinomycin agar plates and incubated for 16 h at 37°C.

Multiple single colonies from each plate were picked and cultured overnight in 3 mL LB both with 100 ug/mL spectinomycin in a 37°C shaker at 220 RPM. Plasmid DNA was isolated using the Thermo Scientific GeneJET MiniPrep Kit according to the manufacturer's instructions.

2.3.3. Restriction digest and sequencing of pCR8 entry vector clones

To identify positive clones containing the insert, restriction digests with Eco RI were performed by digesting 400 ng of DNA for 2 h at 37°C, followed by 20 min at 68°C to denature the enzyme. Digested DNA was then run on a 0.8% agarose gel, and positive clones were then sent for sanger sequencing to determine fragment orientation and their wild-type or mutant status. A phenol-chloroform extraction was carried out on selected WT and mutant plasmids to remove impurities remaining from the MiniPrep.

2.3.4. Subcloning of SNV fragments from pCR8 entry vector to pGL3/pGL4.23 destination vector

Fragments contain the WT and mutant SNVs were subcloned from the pCR8 entry vector to the pGL3/pGL4.23 destination vector using the Gateway LR Clonase II Enzyme mix (Invitrogen) according to the manufacturer's instructions.

2.3.5. Transformation and purification of pGL3/pGL4.23 entry vectors from competent bacteria

As above, plasmids were transformed into One Shot Top10 Chemically Competent E. coli cells by heat shock at 42°C for 30 s. Cells were then left to recover in S.O.C. media shaking for 1 h at 37°C, before being plated on 100 ug/mL ampicillin agar plates and incubated for 16 h at 37°C.

Multiple single colonies from each plate were picked and cultured overnight in 9 mL LB broth with 100 ug/mL ampicillin in a 37°C shaker at 220 RPM. Plasmid DNA was isolated by performing a miniprep with LiCl precipitation, following the standard protocol.

2.3.6. Restriction digest and sequencing of pGL3/pGL4.23 destination vector clones

As above, to confirm that clones contained the insert, restriction digests with Eco RI were performed by digesting 400 ng of DNA for 2 h at 37°C, followed by 20 min at 68°C to denature the enzyme. Digested DNA was then run on a 0.8% agarose gel, and positive clones were then sent for Sanger sequencing with both forward and reverse primers to confirm fragment orientation, their wild-type or mutant status, and that sequences were error-free.

2.3.7. Transfection of HEK293T cells with calcium phosphate

24 h prior to transfection, cells were seeded in 24-well plates at a density of 4.0×10^4 cells/well. As a result, cells were approximately 40% confluent at the time of transfection. For each triplicate transfection, 200 µl of transfection mix was made up. 100 µl of 400 mM CaCl₂ solution containing 1091 ng pGL3/4 plasmid DNA and 109.1 ng pRL-TK plasmid

was made up. The DNA-CaCl₂ mix was then added dropwise to 100 µL sterile 2X HEPES buffered saline (HBS = 140 mM NaCl, 1.5 mM Na₂HPO₄ dibasic, 50 mM HEPES, pH7.0) while bubbling for a total of two min.

The transfection mix was left for 34 min to allow complete DNA-CaPO₄ complex formation. Chloroquine was added to the HEK293T plates to a final concentration of 25 µM. 50 µl of DNA-CaPO₄ mixture was added drop-wise directly onto the media of the HEK293T cells of each triplicate. The media was changed after 8-16 h and cells were harvested after 48 h.

2.3.8. Transfection of MCF7 cells with EugeneHD

24 h prior to transfection, cells were seeded in 24-well plates at a density of 5.0×10^4 cells/well. As a result, cells were approximately 50% confluent at the time of transfection. On the day of transfection, media was removed, and 475 µl of DMEM/F12 media (Sigma) supplemented with 10% FBS was added to each well. For each triplicate transfection, 730 ng of pGL3/4 plasmid DNA and 73 ng of pRL-TK plasmid was added to Opti-MEM for a total volume of 97.6 µl. 2.4 µl of EugeneHD (Promega) was then added and incubated for 10 min. 25 µl of the DNA-Eugene mix was added to each well of cells. Cells were harvested 48 h post-transfection.

2.3.9. Measuring luminescence

Cells were harvested 48 h post-transfection by adding 100 µl of 1X Passive Lysis Buffer (supplied with the Dual Luciferase® Reporter Assay System) and placing plates on a rotator for 15 min. Lysed cells were resuspended and centrifuged for 10 min at 14000 RPM. 20 µl of each cleared lysate was loaded into a 96-well luminometer plate and placed into a Plate-Reading GloMax Multi Luminometer (Promega). 40 µl of Luciferase Assay Reagent II (LAR II) was injected to each well and Firefly luminescence was measured.

40 µl of Stop & Glo® Reagent was subsequently injected to each well and Renilla luminescence was again measured. Firefly luminescence readings were then normalised to Renilla luminescence. Readings from the mutant vector were normalized to the wild-type vector. All experiments were performed in triplicate and repeated at least three times. P-values were calculated using an unequal variance two-tailed t test and then corrected for multiple testing using the Bonferroni correction.

2.4. Transcription factor motif analysis

The effect of SNVs on transcription factor binding sites was assessed using the motifbreakR package (Coetzee et al., 2015) which was run on RStudio Version 1.1.383. TF motifs were obtained from the 'motifbreakRmotif' library, which incorporates 2816 human motifs from HOCOMOCO, Homer, FactorBook and ENCODE. Scores were calculated using the default weighted sum method, which is the difference in probabilities for the two letters of the variant, weighted by the importance of the variant with the motif sequence. I used a p-value cutoff of $1e^{-4}$, and only retained only motifs which were determined by motifbreakR to have a “strong” effect. motifbreakR was not capable of scoring the SNVs and SNPs which were small indels. I then filtered these motifs for only TFs that were expressed in MCF7 (chapter3) or MCF10A (chapter 5).

2.5. CRISPR/Cas9-mediated genomic deletions

2.5.1. Generation of an MCF10A cell line expressing Cas9 via lentiviral transduction

MCF10A cells were engineered to stable expression of Cas9 using pLENTI FLAG Cas9 plasmid. To produce lentiviral particles containing FLAG-Cas9, HEK293T cells were

seeded at 2.1×10^6 cells per T75 flask 24 h prior to transfection. Using PEI, the cells were then transfected with 3 μ g of viral packaging vector (pPAX8) and 2 μ g of viral envelope vector (pVSVG) DNA and 4 μ g of pLENTI FLAG Cas9. Media containing the viral particles was collected 24 and 48 h after transfection and filtered through a 0.45 μ m filter. Viral particles were then used to infect MF10A cells for 18 h, in the presence of 5 μ g/ml polybrene. After infection, cells were then washed and selected in media containing 0.04 μ g/ml blasticidin for 2 weeks.

2.5.2. Cloning and lentiviral transduction of gRNAs into MCF10 cells

Guide RNAs #1 and #2 were cloned into the RFP-containing ipUSEPR lentiviral vector, and guide RNAs #3 and #4 were cloned into the GFP-containing LRG lentiviral vector. The gRNA sequences are listed in Table 2.3.

Lentivirus particles containing the gRNA plasmids were produced as described in Section 2.5.1, where 4 μ g of each gRNA was transfected into HEK293T cells along with pPAX8 and pVSVG. MCF10A cells stably expressing Cas9 were seeded at 0.3×10^6 cells per well of a 6-well plate 24 h prior to transfection. They were then infected with lentivirus particles containing a gRNAs for 16 h, in the presence of 5 μ g/ml polybrene. The same vector without a gRNA sequence was used as an empty vector negative control. The cells were then harvested and sorted for GFP+ and RFP+ cells in a FACSAria flow cytometer (BD Biosciences). Pools of double-positive cells were seeded in 24-well plates and expanded.

2.5.3. Screening for edited loci

CRISPR deletions were detected by extracting gDNA and performing a PCR across the SNV locus using the Platinum PCR SuperMix High Fidelity (Invitrogen) and a 10 μ M solution of primer pairs as listed in Table 2.2. Triplicate reactions were performed with 25

cycles of PCR amplification. The amplified DNA was run on a 0.8% agarose gel to confirm the presence of a deletion.

2.6. RNA preparation and real-time quantitative PCR (RT-qPCR)

Total RNA was extracted using the RNeasy kit (Qiagen) and cDNA was generated by reverse transcription PCR using the TaqMan Reverse Transcription kit (Applied Biosystems). Relative mRNA expression levels were determined by the SYBR Green I detection chemistry (Applied Biosystems) on the QuantStudio 3 Real-Time PCR System (Applied Biosystems). The levels of *RPLPO* and/or *GAPDH* were included as controls. The error bars indicate standard deviation of triplicate qPCR data. The qPCR primers used are listed in Table 2.4.

2.7. Compilation and annotation of the GWAS-associated SNPs

Breast cancer GWAS SNPs were obtained from Lilyquist et al., 2018. For completeness, the SNPs in LD with each lead SNP were calculated using two tools, LDlink's LDproxy tool (<https://ldlink.nci.nih.gov/?tab=ldproxy/>) and rAggr (<http://raggr.usc.edu>). LD SNPs were derived based on frequencies in all 5 European populations (CEU+FIN+GBR+IBS+TSI) from the 1000 Genomes Project Phase 3, and are $\pm$ 500 Kb of the query variant with an R^2 (measure of linkage disequilibrium) value ≥ 0.8 .

Additionally, I added a set of credible risk variants (CRVs) at 142 of the loci (Michailidou et al., 2017). CRVs are defined as variants which were located within 500 kb of the most significant SNP in each region and with P-values within two orders of magnitude of the most significant SNP in each region. The union of these three sets was then created using the bedtools merge tool. This total set of GWAS-associated variants consisted of 6,270 SNPs which were taken forward for further analysis.

SNPs were annotated using Ensembl's VEP tool (http://grch37.ensembl.org/Homo_sapiens/Tools/VEP), selecting the most severe consequence for each variant. Next, SNPs were annotated with regions of open chromatin, as identified by ATAC-seq in MCF10A (Liu et al., 2017) and breast tumours (Corces et al., 2018).

For the breast tumour data, the authors had aligned to the hg38 version of the human genome and identified 215,920 peaks. They then used LiftOver to convert to the hg19 version, and 215,310 peaks remained (99.7%). For the MCF10A data, the authors had also aligned to hg38, and I used Crossmap to convert to hg19, retaining 99.2% of the peaks. I then intersected the SNPs and peaks using bedtools.

SNPs were then annotated using publically available ChIP-seq data. CTCF ChIP-seq and HiC datasets were used to determine if GWAS SNPs potentially affect CTCF binding at TAD boundaries. Firstly, I overlapped the GWAS SNPs in open chromatin with CTCF peaks in 11 different cell lines relevant to breast cancer. I then annotated with ENCODE uniform CTCF sites from HMEC, HMF, T47d, and MCF7 in 6 different conditions (estrogen-treated, vehicle-treated and two no-treatment), as well as MCF10AT1, MCF10CA1, MCF10A from Fritz et al. (2018) (Table 2.1).

TAD boundaries were obtained from Rao et al. (2014). Super-enhancers were obtained from Ji et al. (2018).

Cell Line	Treatment	Source	File Name
HMEC	N/A	ENCODE Uniform TFBS	wgEncodeAvgTfbsBroadHmecCtcfUniPk.narrowPeak
HMEC	N/A	ENCODE Uniform TFBS	wgEncodeAvgTfbsUwHmecCtcfUniPk.narrowPeak
HMF	N/A	ENCODE Uniform TFBS	wgEncodeAvgTfbsUwHmfCtcfUniPk.narrowPeak
T47D	N/A	ENCODE Uniform TFBS	wgEncodeAvgTfbsHaibT47dCtcfsc5916V0416102Dm002p1hUniPk.narrowPeak
MCF7	Estrogen	ENCODE Uniform TFBS	wgEncodeAvgTfbsUtaMcf7CtcfEstroUniPk.narrowPeak
MCF7	Vehicle	ENCODE Uniform TFBS	wgEncodeAvgTfbsUtaMcf7CtcfVehUniPk.narrowPeak
MCF7	N/A	ENCODE Uniform TFBS	wgEncodeAvgTfbsUwMcf7CtcfUniPk.narrowPeak
MCF7	N/A	ENCODE Uniform TFBS	wgEncodeAvgTfbsUtaMcf7CtcfUniPk.narrowPeak
MCF10AT1	N/A	Fritz et al., 2018; GEO: GSE98551; Converted to hg19	GSE98551_MCF10AT1_CTCF_pooled_peaks_passIDR.05.narrowPeak
MCF10A	N/A	Fritz et al., 2018; GEO: GSE98551; Converted to hg19	GSE98551_MCF10A_CTCF_pooled_peaks_passIDR.05.narrowPeak
MCF10CA	N/A	Fritz et al., 2018; GEO: GSE98551; Converted to hg19	GSE98551_MCF10CA1_CTCF_pooled_peaks_passIDR.05.narrowPeak
HMEC	N/A	Rao et al., 2014; GEO: GSE63525	GSE63525_HMEC_HiCCUPS_looplist_with_motifs_peak_boundaries.bed
MCF-10A-ER-Src	+/- Tamoxifen	Ji et al., 2018; Supplementary Data 1	N/A

Table 2.1: Publicly available files used to annotate GWAS-associated SNPs.

2.8. Enrichment of transcription factors at SNPs

I wished to examine whether there was an enrichment of transcription factors binding at our breast cancer GWAS-associated SNPs. To do this, I downloaded all ChIP-seq datasets in breast or breast cancer cell lines from the ENCODE project. To make a background null matched random set for the 716 GWAS-linked SNPs in open chromatin, I generated 716 random SNVs located in the ATAC-seq peaks called in the MCF10A cell line and breast tumours. I then intersected both the random set of SNVs and the GWAS-linked set of SNPs with the peak regions of a TF of interest. I then repeated this 1000 times for each TF, and tabulated how many times the expected number of intersections (the random SNVs) was greater than the observed number of intersections (GWAS-associated SNPs).

The p-value was then defined as: $p\text{-value} = (1 + \sum(\text{expected intersections} \geq \text{observed intersections})) / (1001)$. I then took the number of observed intersections and divided this by the median of the 1000 iterated intersections of the null set to calculate an enrichment fold change, where a positive log₂ fold change indicates enriched SNP-TF overlap compared to random background, while negative indicates depleted. No adjustments were made to correct for multiple testing.

2.9. Primers and Oligonucleotides

ID	FWD Primer	REV Primer
rs576007456:PIK3CA #2 (-334) Pro.	ACTGACAGCGGCGGTTAG	CTCTACTCACGCTCGGGG
rs78248633:PIK3CA #3 (+1,028) Pro.	TTACTGGGTTTAAATCGCCCG	CCTCAGTGCTTCAGTTAAAAGGA
rs146585339:PIK3R1 (+5,520) Enh.	GGTGCTCTGGGTTGAATGTG	GCCCACAGCCCATATCTCTC
rs116906793:MRE11A (-184,861) Enh.	AGGCTTTCCTCTATCGTGAT	AATTCTGAGCCTTTTCTTTCT
rs141340086:PIK3CA (-42,899) Enh.	CTGACAGCGGCGGTTAGCCA	ATT CCC ACC GCA CCC GCT A
rs190491586:ATM (-1,134) Pro.	TGGATGGTGAACCATTTCC	TGTAGAGAGACCTTCCTGGCG
rs574133962:KRAS (+1,174) Pro	CCTGGGCAACAGAGCAGATG	GGGCTGTGACATTGCTGTGG
rs533075364:FHIT (-47) Pro.	GTGCCTCAGTTTCCCAATGTAT	AGGGCTGAGCGTAGTCGTAAATG
rs72682900:GADD45A (+4,385) Enh.	CAAGGTTTTTGTGCCCCCTAGT	CACCTACTGCTTGCTGGCA
rs372644648:GATA3 (+16,144) Enh.	TGTTAGCAGAGCGGTAATTGCC	ACTTCCCAGGAAAGCCTCC
rs181313855:TP53 (-2,242) Enh.	AACCTGAGAAGCTTCTGAAAG	AATCAATAAATAGCTATAAGCCAG
rs184741182:EP300 (-934) Enh.	TGAATTCGACCGGGGTTTTG	CGCGGGGTTATGTAATGGTC
rs73131819:EGFR (+2,415) Pro.	GTCTCCGGAGATGTTACTTGCC	GGAGAGTGTGCACCTCTTATGC
rs535822015:EGFR (+46,661) Enh.	GCACTTTCAAAGCCTCTAA	GCAGCTTTTTGGAAAAGTTT
rs145717085:GATA3 #1 (-7,784) Pro.	GGCTTCGTTCCCATATTTCC	TGGCTTAGGCACCAAGAATG
rs553237095:GATA3 #2 (-1,882) Pro.	GTCAGCATTTTCAAGACCTGGC	GAAGCCGTCCTGAAAGTAAG
rs61917081:CD9 (+29,486) Enh.	GAC ACG CGA CAC TTG G	GAGGAAGATTTTAGGATCACG
rs74951121:BRCA2 (-36,832) Enh.	TGGAAGATTTCTAGCCATATAT	ATT AAT GTC TGT CTC GCT G
rs935143708:BRCA2 (-436) Pro.	TGAGGTCGGGAGGTTGAGAC	GCTGACCCTTAAGAGTCCCAGG
rs375724242:CDH3 (+14,225) Enh.	CCTAAGAGTGCATGCGCATCTTTC	TTAATTTGAAATCCCAACCAAC
rs766631446:ERBB2 (+4,587) Enh.	TGTTCCGAGACTCGTG	AAAGAAGACAAGTGGCTGGCAG
rs529801433:PIK3CA (-350) Pro.	TATCTCTACCCAGCTCGCCTG	AACCGCATCGGGTCCCAG
rs369303269:TERT (-1,566) Enh.	GCTGAGGTCGGCAAACACTG	CTCCGCCTCCCATTGG
rs1030284040:TERT #2 (-20,608) Enh.	GAGCCCATCCTGCCAGGTAC	TGCATCAGCAGCCAAAGCCG
rs369388483:PIK3R1 (+1,201) Pro.	GGCCCTAACTCAGCGTTTATTG	AATGAACCTGTCTGCCCACTC
rs561575495:BRAF (+1,469) Pro.	AGAAGCTGGGGGAAAGGAACAG	CAAATCCTCCTGGACACTTTGG
rs551957572:EGFR #2 (-57,620) Enh.	AGCCATTGTCAGAGTGGATGG	GAGGGCAGTGTTGTGATCTGG
rs141940507:NOTCH1 (-79,891) Enh.	GGGCTCTGGGGAAGGGAACAC	AGAGTGGCGGGAGGCGCTC

Table 2.2: PCR Primers.

Forward gRNA name	Forward gRNA	Reverse gRNA name	Reverse gRNA
hPIK3CA.3-pro-gRNA-1Fwd	CACCGGTAAAGGGAGCGCAACAAGA	hPIK3CA.3-pro-gRNA-1Rev	AAACTCTTGTTGCGCTCCCTTTACC
hPIK3CA.3-pro-gRNA-2Fwd	CACCGAAAATAAACCCCTCGAGCCG	hPIK3CA.3-pro-gRNA-2Rev	AAACCGGCTCGAGGGGTTTATTTTC
hPIK3CA.3-pro-gRNA-3Fwd	CACCGAAACATCGTTCCGAAAACCC	hPIK3CA.3-pro-gRNA-3Rev	AAACGGGTTTTTCGGAACGATGTTTC
hPIK3CA.3-pro-gRNA-4Fwd	CACCGCAGGTCCTGCATGTGCACAA	hPIK3CA.3-pro-gRNA-4Rev	AAACTTGTGCACATGCAGGACCTGC
hPIK3R1-enh-gRNA-1Fwd	CACCGCATCTTATGTCTCTTGCAA	hPIK3R1-enh-gRNA-1Rev	AAACTTGCAAGAGGACATAAGATGC
hPIK3R1-enh-gRNA-2Fwd	CACCGAGAAGTATATTTGGTTCTGG	hPIK3R1-enh-gRNA-2Rev	AAACCCAGAACCAAATATACTTCTC
hPIK3R1-enh-gRNA-3Fwd	CACCGAGACTGTATGAATTTAGGTG	hPIK3R1-enh-gRNA-3Rev	AAACCACCTAAATTCATACAGTCTC
hPIK3R1-enh-gRNA-4Fwd	CACCGCAAGGCAGGAGGGTATGGAA	hPIK3R1-enh-gRNA-4Rev	AAACTTCCATACCCTCCTGCCTTGC
hMRE11A-enh-gRNA-1Fwd	CACCGAAAGTAATATAATTGACGAG	hMRE11A-enh-gRNA-1Rev	AAACCTCGTCAATTATATTACTTTC
hMRE11A-enh-gRNA-2Fwd	CACCGATAAAAAGAGACAGGCTCAG	hMRE11A-enh-gRNA-2Rev	AAACCTGAGCCTGTCTCTTTTTATC
hMRE11A-enh-gRNA-3Fwd	CACCGACAAACAAATGTGCATAAGG	hMRE11A-enh-gRNA-3Rev	AAACCCTTATGCACATTTGTTTGTC
hMRE11A-enh-gRNA-4Fwd	CACCGGAGGAGAGCAATCATTACCT	hMRE11A-enh-gRNA-4Rev	AAACAGGTAATGATTGCTCTCCTCC
hKRAS-pro-gRNA-1Fwd	CACCGATGTTTAAGTCTATTTTCAGG	hKRAS-pro-gRNA-1Rev	AAACCCTGAAATAGACTTAAACATC
hKRAS-pro-gRNA-2Fwd	CACCGTATTTTCAGGCGGTTAGCTGG	hKRAS-pro-gRNA-2Rev	AAACCCAGCTAACCGCCTGAAATAC
hKRAS-pro-gRNA-3Fwd	CACCGCTGTCTACACTCAACTAGCA	hKRAS-pro-gRNA-3Rev	AAACTGCTAGTTGAGTGTAGACAGC
hKRAS-pro-gRNA-4Fwd	CACCGATACATAAGAAATAGGGGAA	hKRAS-pro-gRNA-4Rev	AAACTTCCCCTATTTCTTATGTATC
hGADD45A-enh-gRNA-1Fwd	CACCGGATTGTATGTAGTCATGACT	hGADD45A-enh-gRNA-1Rev	AAACAGTCATGACTACATACAATCC
hGADD45A-enh-gRNA-2Fwd	CACCGCCCACAATTCTAGCAGATGT	hGADD45A-enh-gRNA-2Rev	AAACACATCTGCTAGAATTGTGGGC
hGADD45A-enh-gRNA-3Fwd	CACCGAGACACACAGTGATAAAGC	hGADD45A-enh-gRNA-3Rev	AAACGCTTTATCCACTGTGTGTCTC
hGADD45A-enh-gRNA-4Fwd	CACCGAGGTTGCAGAGTTGTAGGAG	hGADD45A-enh-gRNA-4Rev	AAACCTCCTACAACCTCTGCAACCTC
hGATA3-enh-gRNA-1Fwd	CACCGTAGTTTCATTCACTGCACCA	hGATA3-enh-gRNA-1Rev	AAACTGGTGCAGTGAATGAACTAC
hGATA3-enh-gRNA-2Fwd	CACCGCAGGGCCACAGCGAGAGAGT	hGATA3-enh-gRNA-2Rev	AAACACTCTCTCGCTGTGGCCCTGC
hGATA3-enh-gRNA-3Fwd	CACCGTGTATATAGACCACAGTCCC	hGATA3-enh-gRNA-3Rev	AAACGGGACTGTGGTCTATATACAC
hGATA3-enh-gRNA-4Fwd	CACCGATTTCGTGATTAAATCTGGA	hGATA3-enh-gRNA-4Rev	AAACTCCAGATTTTAATCACGAATC

Table 2.3: gRNA sequences.

Forward gRNA name	Forward gRNA	Reverse gRNA name	Reverse gRNA
hTP53-enh-gRNA-1Fwd	CACCGAATGTGCAGCATTGATGAAG	hTP53-enh-gRNA-1Rev	AAACCTTCATCAATGCTGCACATTC
hTP53-enh-gRNA-2Fwd	CACCGCACTTTTTCCATCTCCACGA	hTP53-enh-gRNA-2Rev	AAACTCGTGGAGATGGAAAAAGTGC
hTP53-enh-gRNA-3Fwd	CACCGGGGGCAGAAAACGCTACATG	hTP53-enh-gRNA-3Rev	AAACCATGTAGCGTTTTCTGCCCCC
hTP53-enh-gRNA-4Fwd	CACCGCTAAATAACTTCAGAATTCG	hTP53-enh-gRNA-4Rev	AAACCGAATTCTGAAGTTATTTAGC
hGATA3.2-pro-gRNA-1Fwd	CACCGTATCCTTAAATCTTAATCCC	hGATA3.2-pro-gRNA-1Rev	AAACGGGATTAAGATTTAAGGATAC
hGATA3.2-pro-gRNA-2Fwd	CACCGGTTTAGAGAGCGCTGTGAGC	hGATA3.2-pro-gRNA-2Rev	AAACGCTCACAGCGCTCTCTAAACC
hGATA3.2-pro-gRNA-3Fwd	CACCGCTTGGTCTCCCGTTATTAGG	hGATA3.2-pro-gRNA-3Rev	AAACCCTAATAACGGGAGACCAAGC
hGATA3.2-pro-gRNA-4Fwd	CACCGCTAATAACGGGAGACCAAGT	hGATA3.2-pro-gRNA-4Rev	AAACACTTGGTCTCCCGTTATTAGC
hERBB2-enh-gRNA-1Fwd	CACCGTACCCCCCAGGAAATCAACG	hERBB2-enh-gRNA-1Rev	AAACCGTTGATTTCTGGGGGGTAC
hERBB2-enh-gRNA-2Fwd	CACCGGGTAGGTCTTAATTCTACCC	hERBB2-enh-gRNA-2Rev	AAACGGGTAGAATTAGGACCTACCC
hERBB2-enh-gRNA-3Fwd	CACCGTACCCTGAGAACGGTTCCCA	hERBB2-enh-gRNA-3Rev	AAACTGGGAACCGTTCTCAGGGTAC
hERBB2-enh-gRNA-4Fwd	CACCGTCCCTGGGAACCGTTCTCA	hERBB2-enh-gRNA-4Rev	AAACTGAGAACGGTTCCCAGGGAAC
hPIK3R1-pro-gRNA-1Fwd	CACCGATGTGTGTCAACTGGGTTTG	hPIK3R1-pro-gRNA-1Rev	AAACCAAACCCAGTTGACACACATC
hPIK3R1-pro-gRNA-2Fwd	CACCGATGCAGAATGTGTGTCAACT	hPIK3R1-pro-gRNA-2Rev	AAACAGTTGACACACATTCTGCATC
hPIK3R1-pro-gRNA-3Fwd	CACCGGCAGCTCAAGGGATTAACGG	hPIK3R1-pro-gRNA-3Rev	AAACCCGTTAATCCCTTGAGCTGCC
hPIK3R1-pro-gRNA-4Fwd	CACCGTCCGATATTAATAGCGCCCC	hPIK3R1-pro-gRNA-4Rev	AAACGGGGCGCTATTAATATCGGAC
hBRAF-pro-gRNA-1Fwd	CACCGATAAATAAGAAGGAACGTGG	hBRAF-pro-gRNA-1Rev	AAACCCACGTTCTTCTTATTTATC
hBRAF-pro-gRNA-2Fwd	CACCGAAAGTTAAGTACCAAAAGCC	hBRAF-pro-gRNA-2Rev	AAACGGCTTTTGGTACTTAACTTTC
hBRAF-pro-gRNA-3Fwd	CACCGTCAGACCTATAAAAATGGGT	hBRAF-pro-gRNA-3Rev	AAACACCCATTTTTATAGGTCTGAC
hBRAF-pro-gRNA-4Fwd	CACCGCAGACAACAATGTTAGGGAG	hBRAF-pro-gRNA-4Rev	AAACCTCCCTAACATTGTTGTCTGC
hNOTCH1-enh-gRNA-1Fwd	CACCGTTAATCTGTAGCTCGGATCA	hNOTCH1-enh-gRNA-1Rev	AAACTGATCCGAGCTACAGATTAAC
hNOTCH1-enh-gRNA-2Fwd	CACCGTCAAGACTTAATCTGTAGCT	hNOTCH1-enh-gRNA-2Rev	AAACAGCTACAGATTAAGTCTTGAC
hNOTCH1-enh-gRNA-3Fwd	CACCGCAGCCGCAGACTAACCCTG	hNOTCH1-enh-gRNA-3Rev	AAACCAGGGGTTAGTCTGCGGCTGC
hNOTCH1-enh-gRNA-4Fwd	CACCGGTGAGGGGTGACCTATAGG	hNOTCH1-enh-gRNA-4Rev	AAACCTATAGGTGACCCCTCACC
hATM-pro-gRNA-1Fwd	CACCGATGGCTGTGAATACTATGGG	hATM-pro-gRNA-1Rev	AAACCCCATAGTATTCACAGCCATC
hATM-pro-gRNA-2Fwd	CACCGCAGCGCTGGGCATTTAAAAG	hATM-pro-gRNA-2Rev	AAACCTTTTAAATGCCAGCGCTGC

Table 2.3 (continued): gRNA sequences.

Forward gRNA name	Forward gRNA	Reverse gRNA name	Reverse gRNA
hATM-pro-gRNA-3Fwd	CACCGTAAAGAGGATTGTACTACAG	hATM-pro-gRNA-3Rev	AAACCTGTAGTACAATCCTCTTTAC
hATM-pro-gRNA-4Fwd	CACCGTGTATTGTCTGTGTGAGAGA	hATM-pro-gRNA-4Rev	AAACTCTCTCACACAGACAATACAC
hBRCA2-enh-gRNA-1Fwd	CACCGTTTGGTCTACCCAAGAAGCC	hBRCA2-enh-gRNA-1Rev	AAACGGCTTCTTGGGTAGACCAAAC
hBRCA2-enh-gRNA-2Fwd	CACCGTAATAAAAGCCTGTTAGAAT	hBRCA2-enh-gRNA-2Rev	AAACATTCTAACAGGCTTTTATTAC
hBRCA2-enh-gRNA-3Fwd	CACCGACACTCATGAGCAACAGACA	hBRCA2-enh-gRNA-3Rev	AAACTGTCTGTTGCTCATGAGTGTC
hBRCA2-enh-gRNA-4Fwd	CACCGAACTGCAACAAGGACCATCA	hBRCA2-enh-gRNA-4Rev	AAACTGATGGTCCTTGTGTCAGTTC
hEP300-enh-gRNA-1Fwd	CACCGCCGAGCCGCTCACACAATGG	hEP300-enh-gRNA-1Rev	AAACCCATTGTGTGAGCGGCTCGGC
hEP300-enh-gRNA-2Fwd	CACCGCGGAAGCGCTAACCGCGTTG	hEP300-enh-gRNA-2Rev	AAACCAACGCGGTTAGCGCTTCCGC
hEP300-enh-gRNA-3Fwd	CACCGTTACAGGCTAGCGGTAGCGG	hEP300-enh-gRNA-3Rev	AAACCCGCTACCGCTAGCCTGTAAC
hEP300-enh-gRNA-4Fwd	CACCGGGCGGCGCGGAAATGAGCCAA	hEP300-enh-gRNA-4Rev	AAACTTGGCTCATTTCCGGCCGCC

Table 2.3 (continued): gRNA sequences.

Gene	Forward	Reverse
<i>hZMAT3</i> mRNA	CCTTTAAGCCAGGAGGCCGAG	TTCTTTCCTGGCCCGCCG
<i>hZNF639</i> mRNA	AAAAGAAAAAGGAAGACTCTAC	TGCAATTCTGCTTATTCC
<i>hNDUFB5</i> mRNA	CCATGAGTTTGTGCGGCGG	CCCCCAAATCCGAGGCCA
<i>hRABL6</i> mRNA	TCCGCCCTGAAGAAGCTGGTG	TGCACCCCCTTGCGGAAG
<i>hKIN</i> mRNA	AGCCACCGAGAGCACATCCAC	AGTTCCAGTTGCCGGCGG
<i>hTAF3</i> mRNA	GAAGAGCAGGTGCCCCACTGATG	AGCCGAGGCCGCTTCATG
<i>hRASSF8</i> mRNA	TCAAGGCCAGCAGAGTCTGAGG	TCAATGGGCTCCGCTGGC
<i>hRPLPO</i> mRNA	TTCATTGTGGGAGCAGAC	CAGCAGTTTCTCCAGAGC

Table 2.4: RT-qPCR mRNA Primers.

3. Identification of potentially functional rare variants in promoters and enhancers located near breast cancer- associated genes

3.1. Introduction

After next-generation sequencing studies, potentially pathogenic variants are normally identified by filtering for loss-of-function mutations. Variants which are located within the protein-coding region can be classified by their effect on the resulting protein. They can be classified into synonymous (the variant does not change the amino acid sequence), nonsense (introduction of a stop codon resulting in a truncated protein), missense (a change in amino acid) or splice-site (occurring at an exon-intron boundary) variants. Additionally, if the variant is an indel with a length of three, or any other number divisible by three, it can result in in-frame insertions or deletions. If the indel is any other length, it may result in a frame-shift mutation. For such exonic variants, there are two widely used *in-silico* bioinformatic tools, SIFT (Ng & Henikoff, 2003) and Polyphen (Adzhubei et al., 2010), available to classify variants by their effect on protein structure and function. Both tools provide the user with a 'score' of potential pathogenicity, which can subsequently be used to prioritise the variants for further analysis.

However, outside the exomes, this task of prioritising potentially pathogenic variants becomes significantly more difficult. Unlike the codon-to-amino acid translation, there is no simple way to interpret the effects of variants in non-coding regions. Therefore, this chapter aimed to prioritise the non-coding variants discovered during next-generation sequencing of Irish familial breast cancer cases, to select an appropriate number for further functional validation.

In order to do this, I firstly filtered the variants for rare case-enriched variants and classified them both by frequency of occurrence and by the function of the cancer gene they were located in proximity to. Secondly, I identified which transcription factor motifs might be disrupted or created by the variants, and highlighted if any of these transcription factors have been implicated in breast cancer previously.

Thirdly, I established a reporter assay to screen the variants for effects on transcription. Such a technique has been used in many previous studies to screen for transcriptional effects of non-coding mutations. Using a luciferase reporter assay, I was able to identify potentially pathogenic variants. From an initial ~4,000 variants identified after sequencing, 15 variants were taken forward for further investigation, as described in chapter 4.

3.2. Results

3.2.1. Study design for the identification of rare germline SNVs in promoters and enhancers located near cancer genes

To address the hypothesis that non-coding DNA elements may have relevance in familial breast cancer, we wished to identify germline SNVs in promoters and enhancers located near known cancer genes. We developed a workflow to do this (Figure 3.1B), by firstly sequencing the promoters and potential enhancer regions of candidate cancer-associated genes, and secondly by functionally evaluating the most promising variants. Selecting and sequencing the target regions as described in this section (Section 3.2.1) was carried out by Fatima Aloraifi as part of her PhD project in the Bracken Lab (Aloraifi, 2014).

Candidate genes included known hereditary breast cancer genes (Walsh et al., 2010), genes that are reported to be epigenetically silenced in more than 20% of breast tumour samples according to the Pubmeth database (Ongenaert et al., 2008), genes implicated from genome-wide association studies of breast cancer (Peng et al., 2011; Zhang et al., 2011), genes reported to be somatically mutated in breast tumours (Forbes et al., 2011) as well as nine other potentially interesting genes.

Histone modifications were used to select potentially functional regulatory regions of these candidate genes. As regulatory regions often show tissue-specificity, we restricted analysis of these histone modifications to tissues and cell lines relevant to breast cancer. Germline variants may be exerting their effects prior to the development of cancer, so the HMEC human mammary cell line and H1-ESC embryonic cell lines were included to represent the non-cancerous and embryonic contexts. The breast cancer cell lines T47D and MCF7 were also included to cover possible functions after the onset of cancer.

To identify potential promoter regions of our candidate genes, H3K4me3 positive regions, DNase hypersensitive sites, and “active promoter” regions from ENCODE’s chromatin segmentation track were identified in HMEC and H1-ESC cell lines. To identify potential active enhancer regions of these genes, H3K27ac positive regions were identified in the HMEC, ES, MCF7 and T47D cell lines. Haplotype blocks of breast cancer GWAS SNPs were also identified, and any of these regions falling on “strong enhancer” or “weak enhancer” chromatin segmentation annotations were included.

Targeted sequencing of these regulatory regions was then carried out on 104 unrelated BRCAx patients and 101 age-, sex- and geographically-matched controls. BRCAx patients were defined as those with a personal history of breast cancer, an age of onset less than 60 years old, and having either at least 2 affected first degree relatives with breast cancer or for patients with bilateral breast cancer, one affected relative.

The potential promoter regions consisted of 367,700 bp of DNA representing regions identified in proximity to 118 cancer-associated genes. These regions had an average coverage across all samples of 83x, with 69% of the target regions covered at 30x. On average, 79% of the bases targeted were covered by 10x, a depth still sufficient for variant calling. The potential enhancer regions consisted of 986,381 bp of DNA representing 968 regions identified in proximity to 128 cancer-associated genes. These regions had an average coverage across all samples of 87x with 78% of the target regions covered at greater than 30x. Between the potential promoter and enhancer regions, 143 cancer-associated genes in total are represented.

Variant calling was performed and variants were filtered for only rare variants (those with a global minor allele frequency (GMAF) of <1%). Additionally, variants found in >10 individuals were also filtered out as these were likely to be population-specific common variants.

Principal component analysis was carried out to assess population stratification and showed all of the sequenced individuals had the same European ancestral background, except for one of Asian ancestry who was excluded from further analysis (Aloraifi et al., 2015).

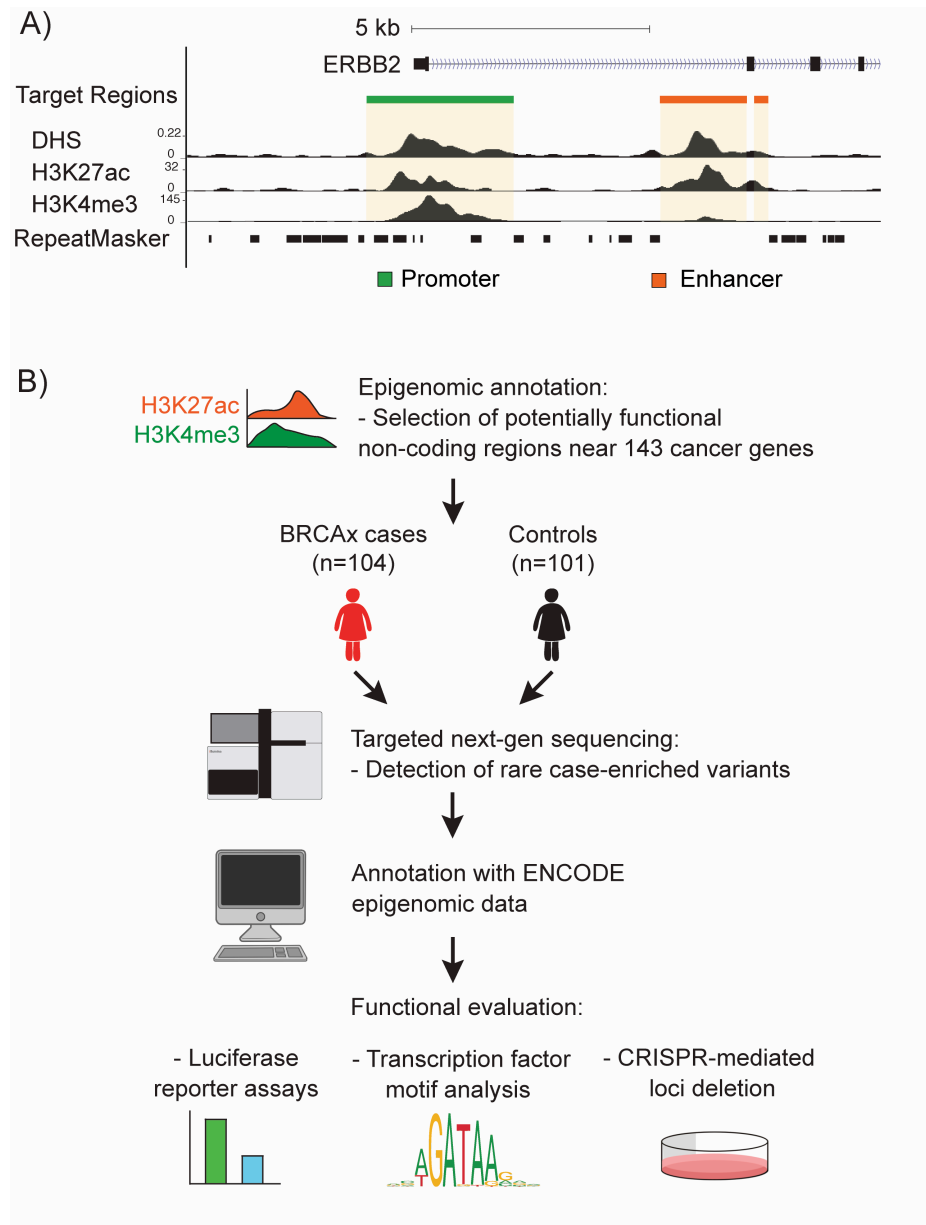


Figure 3.1: Study design for the identification of germline SNVs in promoters and enhancers located near breast cancer genes.

A) Selection of potentially functional non-coding regulatory regions located near 143 cancer genes for targeted sequencing. Promoters and candidate enhancer regions were selected based on the presence of H3K4me3 and H3K27ac respectively. Shown is the promoter of, and an enhancer within, the sporadic breast cancer-associated oncogene *ERBB2*, with histone modification data and DHSs from the HMEC cell line.

B) Workflow to identify familial breast cancer-associated pathogenic non-coding variants. Rare, case-enriched variants were identified from targeted sequencing. This was followed by functional evaluation through the use of luciferase assays, transcription factor motif analysis (chapter 3) and CRISPR-mediated deletion of the variant loci (chapter 4).

3.2.2. Prioritisation of case-enriched SNVs for functional evaluation

After filtering, variant calling identified 892 unique rare variants in the potential promoter regions, and 2822 unique rare variants in the potential enhancer regions. Variants were then filtered to retain only rare case-enriched variants. I defined 'case-enriched' by:

- removing all variants that occur more in controls than cases, or in an equal number of each
- removing all variants occurring in >2 controls
- only keeping the remaining variants occurring in controls if they are present in 3 or more cases than controls
- remove variants occurring just once in cases.

This reduced the number of rare case-enriched variants to 255 (Appendix 1). To predict the potential effects of these SNVs, I used Ensembl's Variant Effect Predictor to analyse where the variants were located. Most (56%) of these 255 SNVs were located in intronic regions, 8% were located in the 3' and 5' UTRs, and 2.7% in the exons of non-coding RNA genes (Figure 3.2).

From these 255 case-enriched variants, we selected 28 to be functionally investigated. These 28 variants were associated with genes from all four of the gene categories; hereditary, somatic, GWAS and hypermethylated. They also represented a selection of the frequency of occurrences in cases and controls (Figure 3.3; Table 3.1).

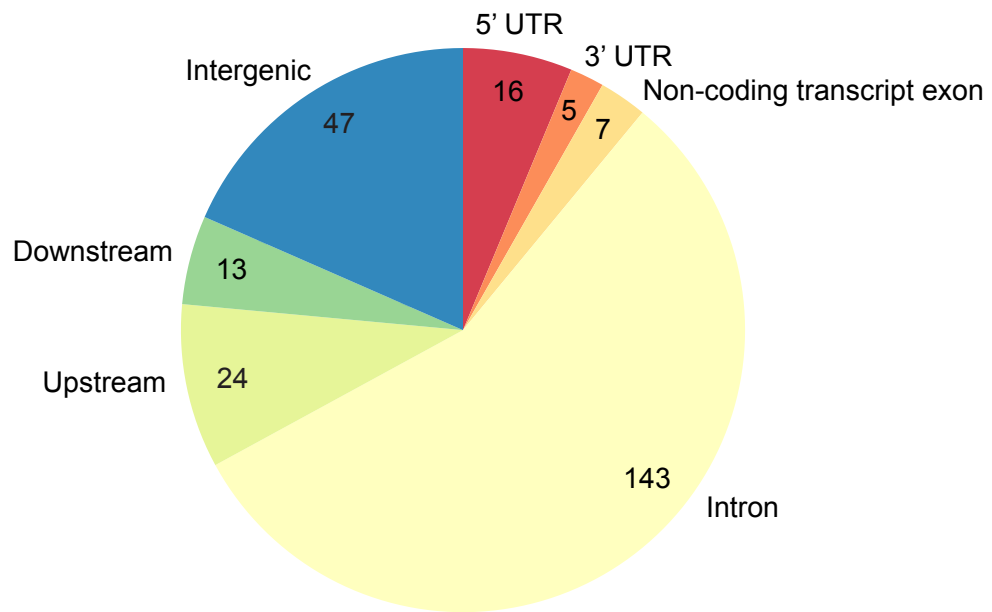


Figure 3.2: Genomic distribution of 255 non-coding variants enriched in familial breast cancer cases.

The consequence of each variant was determined using Ensembl's Variant Effect Predictor. Results are ordered clockwise from most severe to least severe (red to blue). Numbers within each slice indicate the numbers of variants in each category.

ID	Label	Location (hg19)	# Cases	# Controls	Ref	Alt	Gene Category	Relative to breast cancer gene Distance to TSS	Location	GMAF	Linear closest gene
rs576007456: <i>PIK3CA</i> #2 (-334)	Pro.	chr3:178865976	7	1	G	A	Somatic	-334	Upstream	0.0036	<i>LOC101928739 (PIK3CA)</i>
rs78248633: <i>PIK3CA</i> #3 (+1,028)	Pro.	chr3:178867338	7	2	T	C	Somatic	+1,028	Downstream	0.0038	<i>PIK3CA</i>
rs146585339: <i>PIK3R1</i> (+5,520)	Enh.	chr5:67517103	6	0	G	A	Somatic	+5,520	Downstream	0.0062	<i>PIK3R1</i>
rs116906793: <i>MRE11A</i> (-184,861)	Enh.	chr11:94411901	5	0	C	T	Hereditary	-184,861	Upstream	0.0016	<i>LOC105369438 (PIWIL4)</i>
rs141340086: <i>PIK3CA</i> (-42,899)	Enh.	chr3:178823411	5	0	A	G	Somatic	-42,899	Upstream	0.003	<i>LOC101928739 (ZMAT3)</i>
rs190491686: <i>ATM</i> (-1,134)	Pro.	chr11:108092424	4	0	G	A	Hereditary	-1,134	Upstream	0.0014	<i>NPAT</i>
rs533075364: <i>FHIT</i> (-47)	Pro.	chr3:61237180	4	0	T	C	Methylated	-47	Upstream	0.0004	<i>FHIT</i>
rs574133962: <i>KRAS</i> (+1,174)	Pro.	chr12:25402680	4	0	T	C	Somatic	+1,174	Downstream	0.0006	<i>KRAS</i>
rs181313855: <i>TP53</i> (-2,242)	Enh.	chr17:7593110	3	0	T	C	Somatic	-2,242	Upstream	0.0028	<i>WRAP53</i>
rs184741182: <i>EP300</i> (-934)	Enh.	chr22:41487679	3	0	G	A	Somatic	-934	Upstream	0.0006	<i>EP300</i>
rs372644648: <i>GATA3</i> (+16,144)	Enh.	chr10:8112810	3	0	C	T	Somatic	+16,144	Downstream	0.0008	<i>GATA3</i>
rs535822015: <i>EGFR</i> (+46,661)	Enh.	chr7:55133365	3	0	TG	T	Somatic	+46,661	Downstream	0.0006	<i>EGFR</i>
rs72682900: <i>GADD45A</i> (+4,385)	Enh.	chr1:68155244	3	0	C	T	Methylated	+4,385	Downstream	0.0018	<i>GADD45A</i>
rs73131819: <i>EGFR</i> (+2,415)	Pro.	chr7:55089139	3	0	T	C	Somatic	+2,415	Downstream	0.0048	<i>EGFR</i>
rs1030284040: <i>TERT</i> #2 (-20,608)	Enh.	chr5:1315770	2	0	AT	A	GWAS	-20,608	Upstream	0.0001	<i>CLPTM1L</i>
rs141940507: <i>NOTCH1</i> (-79,891)	Enh.	chr9:139520129	2	0	C	T	Somatic	-79,891	Upstream	0.0042	<i>LINC01451 (EGFL7)</i>
rs145717085: <i>GATA3</i> #1 (-7,784)	Pro.	chr10:8088862	2	0	G	T	Somatic	-7,784	Upstream	0.0034	<i>GATA3-AS1 (GATA3)</i>
rs3939303269: <i>TERT</i> (-1,566)	Enh.	chr5:1296728	2	0	G	A	GWAS	-1,566	Upstream	0.0001	<i>TERT</i>
rs3939388483: <i>PIK3R1</i> (+1,201)	Pro.	chr5:67512784	2	0	ATTC	A	Somatic	+1,201	Downstream	0.0018	<i>PIK3R1</i>
rs375724242: <i>CDH3</i> (+14,225)	Enh.	chr16:68692375	2	0	G	A	Methylated	+14,225	Downstream	0.0004	<i>CDH3</i>
rs529801433: <i>PIK3CA</i> (-350)	Pro.	chr3:178865960	2	0	G	T	Somatic	-350	Upstream	0.0004	<i>LOC101928739 (PIK3CA)</i>
rs551957572: <i>EGFR</i> #2 (-57,620)	Enh.	chr7:55029104	2	0	G	A	Somatic	-57,620	Upstream	0.0002	<i>EGFR</i>
rs553237095: <i>GATA3</i> #2 (-1,882)	Pro.	chr10:8094784	2	0	G	T	Somatic	-1,882	Upstream	0.0002	<i>GATA3-AS1 (GATA3)</i>
rs561575495: <i>BRAF</i> (+1,469)	Pro.	chr7:140623095	2	0	G	C	Somatic	+1,469	Downstream	0.0008	<i>BRAF</i>
rs61917081: <i>CD9</i> (+29,486)	Enh.	chr12:6338967	2	0	T	C	Methylated	+29,486	Downstream	0.0026	<i>CD9</i>
rs74961121: <i>BRCA2</i> (-36,832)	Enh.	chr13:32852784	2	0	G	A	Hereditary	-36,832	Upstream	0.0006	<i>FRY</i>
rs766631446: <i>ERBB2</i> (+4,587)	Enh.	chr17:37660840	2	0	A	G	Somatic	+4,587	Downstream	0.0007	<i>ERBB2</i>
rs935143708: <i>BRCA2</i> (-436)	Pro.	chr13:32889180	2	0	T	C	Hereditary	-436	Upstream	0.0007	<i>BRCA2</i>

Table 3.1: Prioritised potentially pathogenic variants.

Table summarising important additional information about the 28 SNVs selected for further functional evaluation. Shown are data about the location, rsID, frequency in cases and controls, reference and alternative alleles, and the global minor allele frequency. The relationship between the hypothesised target gene and breast cancer is shown in the 'Gene Category' column. Also shown is the location relative to, and distance between, the SNV and the transcription start site of the hypothesised target gene. The linear closest gene is also shown, but where this was an ncRNA gene, then the nearest protein-coding gene is also indicated in brackets.

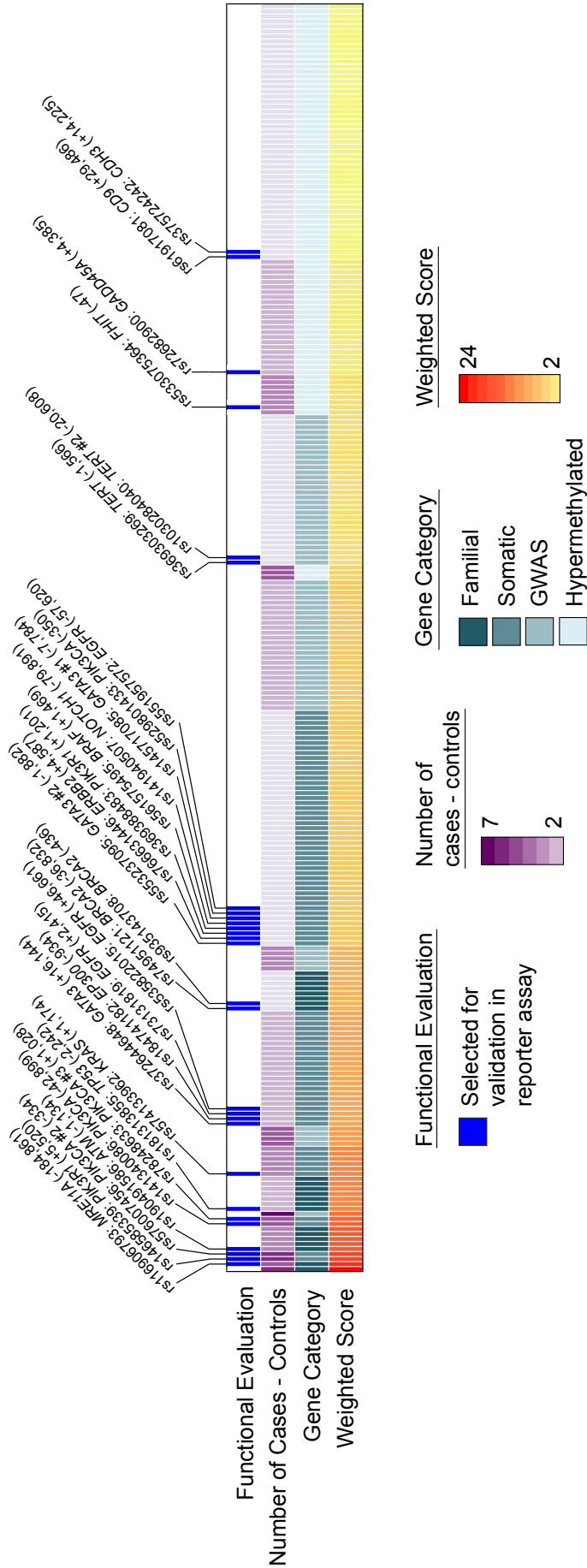


Figure 3.3: SNVs near breast cancer-related genes were identified in a familial breast cancer cohort.

The 255 case-enriched SNVs are ordered based on their weighted score. The weighted score is the frequency of the SNV in cases, minus the frequency in controls, weighted by the relevance of the implicated gene. A score of 4 was assigned to familial breast cancer genes, 3 to somatic breast cancer genes, 2 to genes implicated through GWAS and 1 to genes hypermethylated in breast cancer. A representative selection of SNVs was chosen for functional evaluation in a luciferase reporter assay (blue).

3.2.3. Overview of clinical and biological characteristics of 28 potentially functional non-coding regulatory SNVs

The selected 28 SNVs are present in a combined total of 52 patients. It is important to note that the vast majority of the patients with these prioritised non-coding variants do not have an exonic mutation that could explain their cancer (Figure 3.4). For example, if we take a look at the *BRCA2* promoter mutation rs935143708, we can see that both patients who have this mutation also have coding mutations – one has a frameshift mutation in *ATM* and the other a missense variant in *CHEK2*. Therefore we might be sceptical about this *BRCA2* promoter SNV being truly responsible for causing breast cancer in these two patients. In contrast, the two patients with the *BRCA2* enhancer mutation do not have any mutations in the coding regions of familial breast cancer genes, and so it remains plausible that this SNV may be pathogenic. In fact, just 7 out of 52 patients were found to have rare mutations in exons of cancer genes (Figure 3.4).

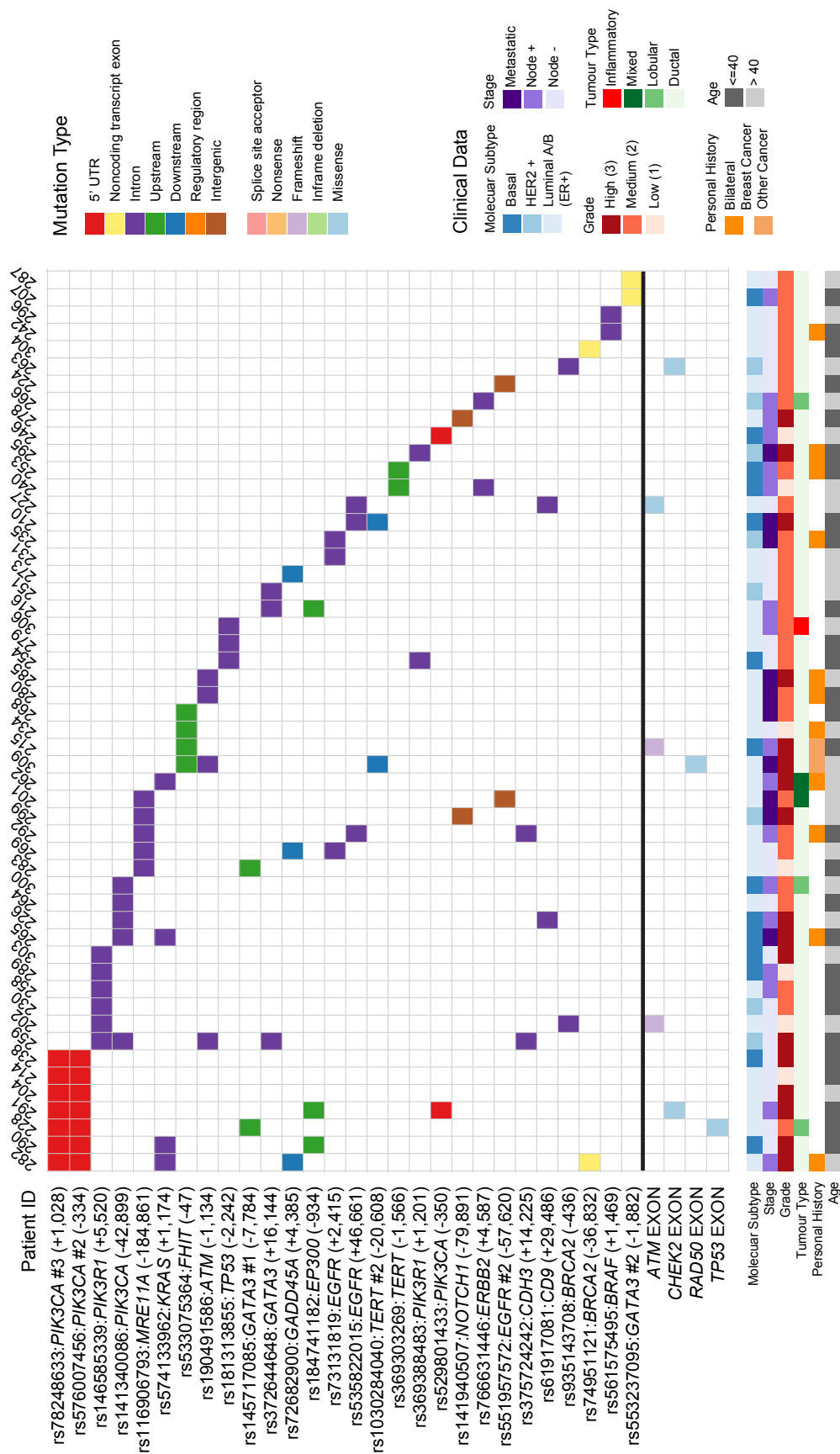


Figure 3.4: Overview of clinical and biological characteristics of 28 potentially functional non-coding regulatory SNVs.

Mutations in the coding regions of known familial breast cancer susceptibility genes are shown in the lower panel.

3.2.4. Familial breast cancer-associated variants modify transcription factor binding motifs

One mechanism for how SNVs can exert their effects is through the disruption or creation of transcription factor binding sites (TFBS). Variants causing a gain of a TFBS leading to increased expression of their target gene have been reported previously. In melanoma, *TERT* promoter mutations create an ETS family binding motif leading to increased expression of TERT (Horn et al., 2013; Huang et al., 2013). In leukaemia, *TAL1* enhancer mutations create MYB-binding motifs (Mansour et al., 2014). Therefore, I explored if the prioritised SNVs affected any TFBS. To do this, I used the program motifbreakR, to determine if any SNPs disrupt or create TF binding motifs. I then filtered to retain only TFs that were expressed in MCF7 cells. The most frequently affected TFs are REST, SP1 and STAT3, which all appear 3 times. To determine which of these TFs may be involved in breast cancer, I overlapped the list of TFs with a previously compiled list of 312 breast cancer-associated genes (Aloraifi et al., 2015) and a more recent list of TFs implicated in breast cancer (Zacksenhaus et al., 2017). Just 8 of the TFs affected in our study are associated with breast cancer (TFs in bold in Table 3.2).

ID	Label	Location (hg19)	TFBS DESTROYED	TFBS CREATED
rs576007456:PIK3CA #2 (-334)	Pro.	chr3:178865976	PLAG1, SP4, ZNF148, ZNF740, SP1, ZNF143, SP3, EGR2, ZNF281, KLF8, KLF7, FOXO6, KLF16, FOXO1 , FOXO3, FOXO4, KLF4, PATZ1, STAT3, TBP, EGR1	MZF1
rs78248633:PIK3CA #3 (+1,028)	Pro.	chr3:178867338	BCL6B	NFKB1, REL, RELA, RELB, STAT1, STAT3
rs146585339:PIK3R1 (+5,520)	Enh.	chr5:67517103	FOXJ3	SOX15
rs116906793:MRE11A (-184,861)	Enh.	chr11:94411901	ESRRA, REST	CDC5L
rs141340086:PIK3CA (-42,899)	Enh.	chr3:178823411	-	E2F1
rs190491586:ATM (-1,134)	Pro.	chr11:108092424	-	-
rs533075364:FHIT (-47)	Pro.	chr3:61237180	FOXO1 , FOXO6, REST, SMAD2, SMAD3, SMAD4	REST
rs574133962:KRAS (+1,174)	Pro.	chr12:25402680	-	-
rs181313855:TP53 (-2,242)	Enh.	chr17:7593110	FOXJ2, FOXJ3, EP300	SRF
rs184741182:EP300 (-934)	Enh.	chr22:41487679	MBD2, CTCF	-
rs372644648:GATA3 (+16,144)	Enh.	chr10:8112810	-	-
rs535822015:EGFR (+46,661)	Enh.	chr7:55133385	-	-
rs72682900:GADD45A (+4,385)	Enh.	chr1:68155244	-	FOXF2 , FOXJ3, FOXp1 , IRF3, IRF4, IRF5, IRF7, IRF8, IRF9
rs73131819:EGFR (+2,415)	Pro.	chr7:55089139	EN1	HDAC2
rs1030284040:TERT #2 (-20,608)	Enh.	chr5:1315770	-	-
rs141940507:NOTCH1 (-79,891)	Enh.	chr9:139520129	-	HSF1, SOX8, SOX9
rs145717085:GATA3 #1 (-7,784)	Pro.	chr10:8088882	-	-
rs369303269:TERT (-1,566)	Enh.	chr5:1296728	-	POU5F1
rs369388483:PIK3R1 (+1,201)	Pro.	chr5:67512784	-	-
rs375724242:CDH3 (+14,225)	Enh.	chr16:68692375	EGR1, EGR2, EGR3, RXRA	HIC1 , ASCL2, TAL1, TFAP4, TFCP2, TCF12, TCF3
rs529801433:PIK3CA (-350)	Pro.	chr3:178865960	SP4, MAZ, RAD21, ZNF281, ZNF148, SP1, SP3, MZF1, ZNF219, KLF7, CCNT2, KLF4, KLF5, PLAG1, REST, GABPA, IRF1, STAT3, TBP, SPI1	EGR2, EGR1
rs551957572:EGFR #2 (-57,620)	Enh.	chr7:55029104	-	-
rs553237095:GATA3 #2 (-1,882)	Pro.	chr10:8094784	-	GATA2, HDAC2, TEF, ZNF333, GATA3
rs561575495:BRAF (+1,469)	Pro.	chr7:140623095	-	-
rs61917081:CD9 (+29,486)	Enh.	chr12:6338967	-	-
rs74951121:BRCA2 (-36,832)	Enh.	chr13:32852784	ZBTB7A, CREB1	-
rs766631446:ERBB2 (+4,587)	Enh.	chr17:37860840	FOS, SP2, CEBPZ, RFX5, PBX3, NFY, NFYA, NFYB, SP1, IRF3, HDAC2	KLF12, GRHL1
rs935143708:BRCA2 (-436)	Pro.	chr13:32889180	-	POU2F2, ZBTB7C, ZNF219

Table 3.2: Familial breast cancer-associated variants modify transcription factor binding motifs.

Table of the effects of the SNVs identified in familial breast cancer on transcription factor binding motifs. Only those transcription factors expressed in MCF7 cells are listed.

Transcription factors previously associated with breast cancer are highlighted in bold.

3.2.5. Familial breast cancer-associated variants disrupt transcription in an *in-vitro* reporter assay

Next, I wished to functionally evaluate the potential consequences of SNVs on transcription. To do this, I used the luciferase reporter assay to screen if any of the SNVs were capable of altering transcription. The assays were carried out in a human embryonic kidney (HEK293T) cell line and a breast cancer (MCF7) cell line. In MCF7 cells, 17 variants were initially statistically significant; with 9 resulting in increased transcription and 8 resulting in decreased transcription. After correcting for multiple testing with the Bonferroni method, the statistically significant cut-off was $p < 0.0018$. 5 variants remained statistically significant with 2 resulting in increased transcription and 3 resulting in decreased transcription. SNVs which were capable of increasing transcription in the assay are shown in Figure 3.5, those capable of decreasing transcription are shown in Figure 3.6, and those with little effect are shown in Figure 3.7. Results of the assay in the MCF7 cell line is summarised in Table 3.3. It will be particularly interesting to follow up the SNVs located near oncogenes which increase transcription, and those located near tumour suppressor genes which decrease transcription.

ID	Label	Location (hg19)	MUT:WT ratio	Fold change	Direction	P value	CRISPR
rs116906793: <i>MRE11A</i> (-184,861)	Enh.	chr11:94411901	0.223	4.492	Decreasing	0.00002	Y
rs78248633: <i>PIK3CA</i> #3 (+1,028)	Pro.	chr3:178867338	2.092	2.092	Increasing	0.00006	Y
rs181313855: <i>TP53</i> (-2,242)	Enh.	chr17:7593110	0.595	1.681	Decreasing	0.00008	Y
rs372644648: <i>GATA3</i> (+16,144)	Enh.	chr10:8112810	0.638	1.567	Decreasing	0.00048	Y
rs146585339: <i>PIK3R1</i> (+5,520)	Enh.	chr5:67517103	5.738	5.738	Increasing	0.00093	Y
rs72682900: <i>GADD45A</i> (+4,385)	Enh.	chr1:68155244	0.358	2.796	Decreasing	0.00223	Y
rs141940507: <i>NOTCH1</i> (-79,891)	Enh.	chr9:139520129	0.724	1.382	Decreasing	0.00284	Y
rs553237095: <i>GATA3</i> #2 (-1,882)	Pro.	chr10:8094784	0.700	1.428	Decreasing	0.00375	Y
rs369388483: <i>PIK3R1</i> (+1,201)	Pro.	chr5:67512784	1.495	1.495	Increasing	0.00393	Y
rs561575495: <i>BRAF</i> (+1,469)	Pro.	chr7:140623095	1.704	1.704	Increasing	0.00886	Y
rs766631446: <i>ERBB2</i> (+4,587)	Enh.	chr17:37860840	0.765	1.307	Decreasing	0.01458	Y
rs74951121: <i>BRCA2</i> (-36,832)	Enh.	chr13:32852784	3.028	3.028	Increasing	0.02503	Y
rs141340086: <i>PIK3CA</i> (-42,899)	Enh.	chr3:178823411	0.662	1.511	Decreasing	0.02751	-
rs190491586: <i>ATM</i> (-1,134)	Pro.	chr11:108092424	2.384	2.384	Increasing	0.03137	Y
rs184741182: <i>EP300</i> (-934)	Enh.	chr22:41487679	3.291	3.291	Increasing	0.03200	Y
rs551957572: <i>EGFR</i> #2 (-57,620)	Enh.	chr7:55029104	1.207	1.207	Increasing	0.03633	-
rs369303269: <i>TERT</i> (-1,566)	Enh.	chr5:1296728	1.563	1.563	Increasing	0.04340	-
rs61917081: <i>CD9</i> (+29,486)	Enh.	chr12:6338967	1.269	1.269	Increasing	0.05740	-
rs576007456: <i>PIK3CA</i> #2 (-334)	Pro.	chr3:178865976	1.262	1.262	Increasing	0.07144	-
rs145717085: <i>GATA3</i> #1 (-7,784)	Pro.	chr10:8088882	0.836	1.196	Decreasing	0.20370	-
rs533075364: <i>FHIT</i> (-47)	Pro.	chr3:61237180	1.495	1.495	Increasing	0.21280	-
rs73131819: <i>EGFR</i> (+2,415)	Pro.	chr7:55089139	0.778	1.286	Decreasing	0.33030	-
rs574133962: <i>KRAS</i> (+1,174)	Pro.	chr12:25402680	1.754	1.754	Increasing	0.35510	Y
rs535822015: <i>EGFR</i> (+46,661)	Enh.	chr7:55133385	1.120	1.120	Increasing	0.36160	-
rs375724242: <i>CDH3</i> (+14,225)	Enh.	chr16:68692375	0.963	1.038	Decreasing	0.45050	-
rs935143708: <i>BRCA2</i> (-436)	Pro.	chr13:32889180	1.291	1.291	Increasing	0.50860	-
rs529801433: <i>PIK3CA</i> (-350)	Pro.	chr3:178865960	1.168	1.168	Increasing	0.56690	-
rs1030284040: <i>TERT</i> #2 (-20,608)	Enh.	chr5:1315770	1.014	1.014	Increasing	0.90280	-

Table 3.3: Familial breast cancer-associated variants disrupt transcription in an *in-vitro* reporter assay.

P-values were calculated using an unequal variance two-tailed t test and then corrected for multiple testing using the Bonferroni correction. Variants which are statistically significant before correction are indicated in bold, and significant after correction are indicated in red. Variants which were taken forward for further analysis are also indicated in the 'CRISPR' column.

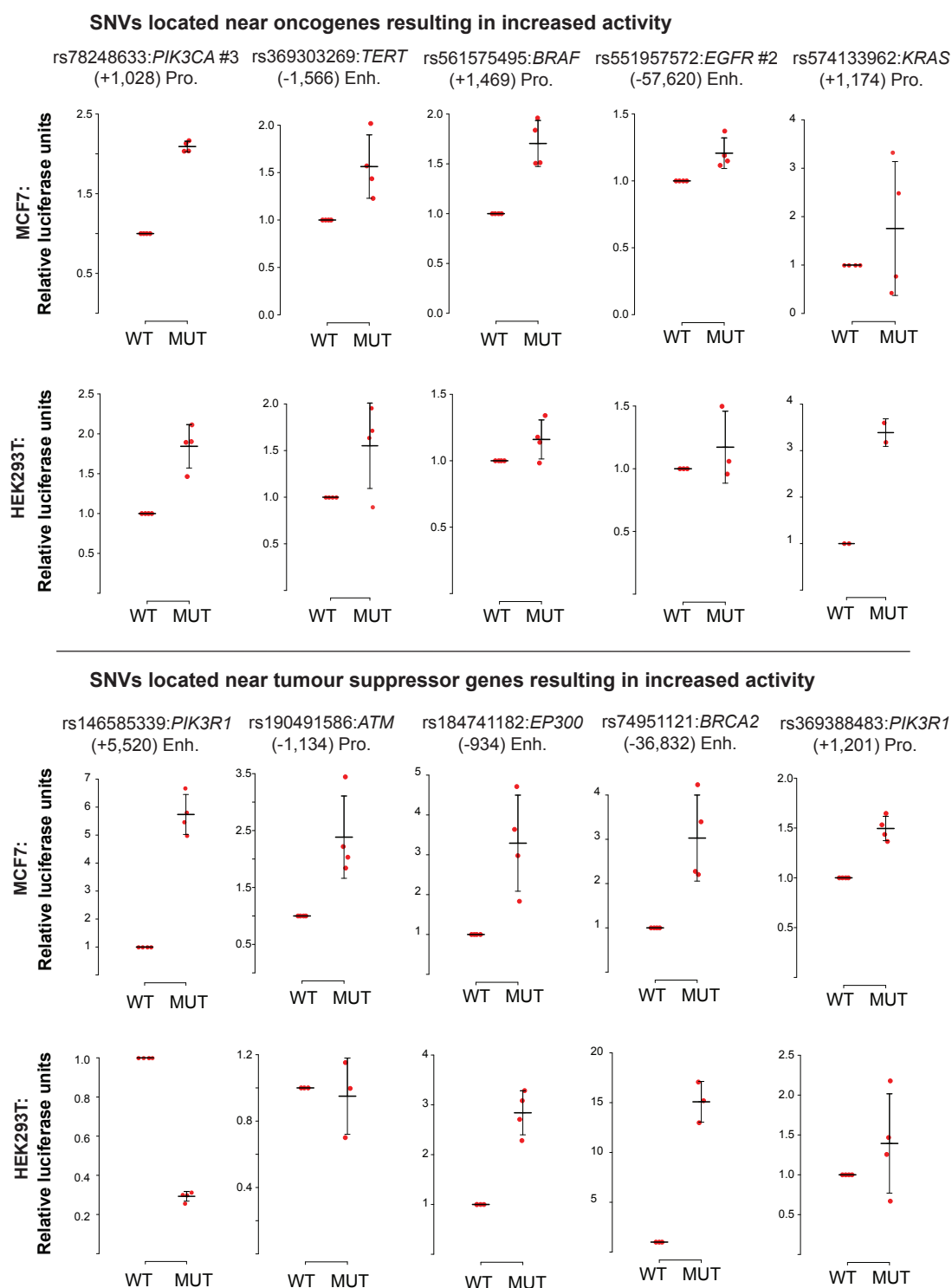


Figure 3.5: Familial breast cancer-associated SNVs at promoters and enhancers which result in an increase in transcription.

For each SNV, results of the assay are shown in MCF7 cells (top) and HEK293T cells (bottom).

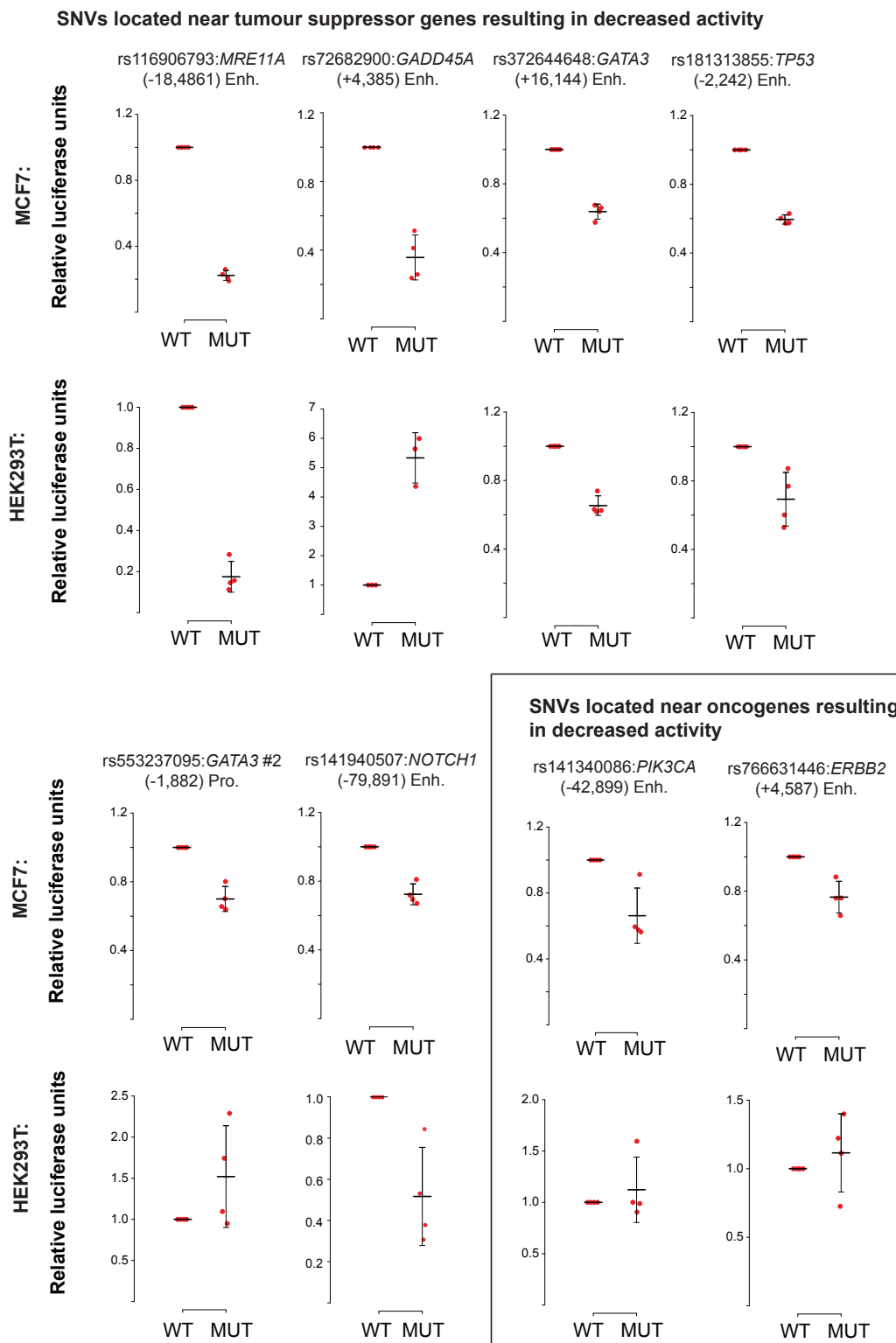


Figure 3.6: Familial breast cancer-associated SNVs at promoters and enhancers which result in a decrease in transcription.

For each SNV, results of the assay are shown in MCF7 cells (top) and HEK293T cells (bottom).

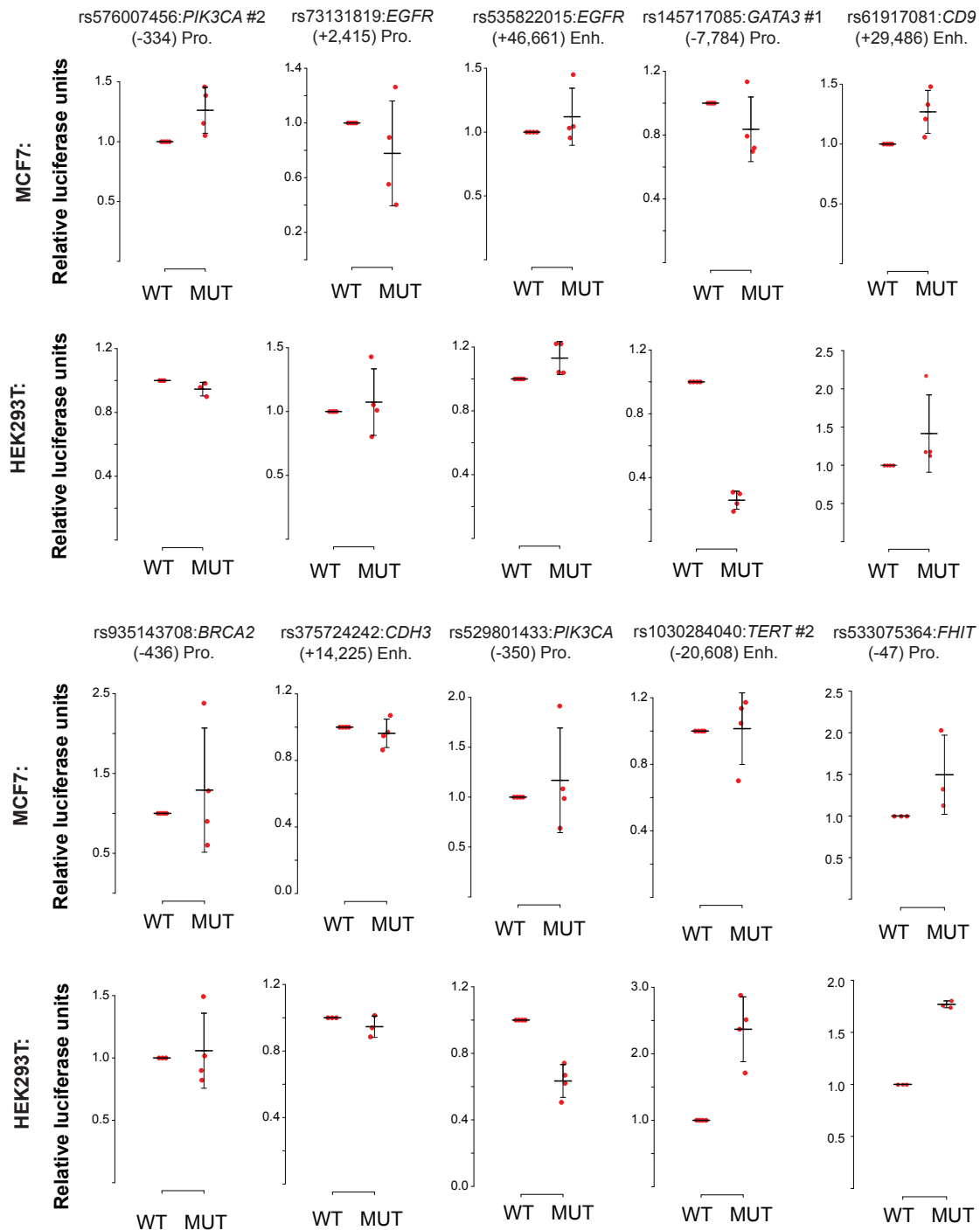


Figure 3.7: Familial breast cancer-associated SNVs at promoters and enhancers which result in minimal disruption to transcription.

For each SNV, results of the assay are shown in MCF7 cells (top) and HEK293T cells (bottom).

3.3. Discussion

This chapter aimed to prioritise the non-coding variants discovered during next-generation sequencing of Irish familial breast cancer cases. Ultimately, we wished to screen the variants for their ability to disrupt transcription so that we could assess which ones warranted further analysis. I began by selecting a range of rare SNVs located in promoter and enhancer regions located in proximity to breast cancer genes.

When selecting variants to prioritise, I chose to focus on the most recurrent variants in our cohort. However, an alternative method would have been to identify the most recurrently mutated regions, rather than the most recurrently mutated individual nucleotides. This method, burden testing, analyses whether variants in a particular set of genes, or in this case, set of gene regulatory regions, occur more often in cases than controls (Morgenthaler & Thilly, 2007). It involves collapsing all variants within a single region, for example, a gene, promoter or enhancer and examining the cumulative effects (Lee et al., 2014). Burden testing is a type of rare variant association test but has the advantage of increased power over single variant tests. Burden testing was carried out by Fatima Aloraifi (Aloraifi, 2014), but is not discussed here however, since this project is focused on the functional analysis of single variants, we chose to pursue the most recurrent variants instead of the most recurrently mutated regulatory regions.

As I focused on recurrent variants, I filtered out ultra-rare variants, defined as variants only present in one person or those absent from large genomic databases such as dbSNP. All of these 28 variants were present in the 1000K genomes project, so have been observed in other cohorts before. However, most of the variants identified in this study were ultra-rare: 517 out of 892 promoter variants and 1737 out of 2822 enhancer variants were present in one patient only. This is to be expected as the 1000K Genomes Project Consortium reported that the majority of variants in a population are rare (The 1000 Genomes Project Consortium, 2015).

Yet whilst these 28 variants were not ultra-rare they still had a MAF <1% in the general population as determined by 1000K Genomes Project MAFs. Adding in more recently available data from GnomAD (V2.1.1) (Karczewski et al., 2019), which was not available at the beginning of this project, the MAF of all variants remained <1%, except for rs146585339: *PIK3R1* (+5,520) enhancer and rs73131819: *EGFR* (+2,415) promoter which rose to 1.143% and 1.283% respectively, which is still not common. Interestingly, however, the frequency in our cohort was much elevated over the 1000K Genomes Project MAFs - these variants are enriched in our cases.

Once I had selected a shortlist of SNVs, I then carried out an analysis to assess whether they disrupted or created a TFBS motif. 18 out of the 28 SNVs did, but only 7 SNVs affected TF motifs previously associated with breast cancer:

- rs184741182:*EP300* (-934) enhancer SNV destroys a CTCF motif. CTCF is involved in the 3D organisation of the genome.
- rs145717085:*GATA3* #1 (-7,784) promoter SNV creates a GATA3 motif in a possible feedback loop. GATA3 is involved in the differentiation of many tissues including the mammary gland and is highly significantly mutated in breast cancer.
- rs375724242:*CDH3* (+14,225) enhancer SNV creates a site for HIC1, which is hypermethylated in breast cancer.
- rs576007456:*PIK3CA* #2 (-334) promoter SNV destroys a FOXO1 site.
- rs181313855:*TP53* (-2,242) enhancer SNV disrupts binding of EP300, a histone acetyltransferase.
- rs533075364:*FHIT* (-47) promoter SNV also disrupts two breast cancer transcription factors: FOXO1, and SMAD4 which acts as a tumour suppressor with an inhibitory effect on cell proliferation.

Having these predictions of TFBS disruptions and creations will guide our functional experiments later on. For example, if a SNV creates a *de novo* binding site for a particular transcriptional activator, we could repeat the luciferase assays with a knockdown of that

specific TF and test if the change mediated by the SNV is diminished. We could perform chromatin immunoprecipitations for the various TFs of interest, and see if they are present at the SNV loci. These TFs provide us with clues about the mechanism of action of the variant.

Next, I performed luciferase reporter assays to determine whether any of the SNVs were capable of disrupting transcription. I found that 17 of them were, with 5 showing a statistically significant difference in transcription between wild type and mutant even after correcting for multiple testing.

Of particular interest for follow up are those SNVs located near oncogenes which increase transcription (rs78248633:*PIK3CA* #3 (+1,028) promoter), and those located near tumour suppressor genes which decrease transcription (rs116906793:*MRE11A* (-184,861) enhancer, rs181313855:*TP53* (-2,242) enhancer, rs372644648:*GATA3* (+16,144) enhancer). Interestingly, 3 out of these top 4 SNVs also affected TF motifs which are of interest in breast cancer. All four of these SNVs were selected for further functional analysis. Additionally, I decided a further 11 SNVs also warranted further analysis. Therefore, a total of 15 SNVs were selected for functional analysis where the region surrounding the SNV was deleted using the CRISPR method.

When integrating the TFBS predictions with the luciferase results, for most of the variants, a clear mechanism of action does not emerge. There are however, a few interesting correlations.

Most interesting out of the 7 SNVs with breast cancer-associated TFs is the rs72682900:*GADD45A* (+4,385) enhancer SNV which creates a FOXP1 and FOXF2 motif, both TFs belonging to the forkhead box TF superfamily. Conflicting results about the function of FOXP1 and FOXF2 in breast cancer have been published, however, and context appears to be critical. FOXP1 is a repressor, possibly explaining the decreased

expression found in the luciferase assay. It has been thought to be a tumour suppressor as reduced expression in breast cancer is correlated with poor survival (Xiao et al., 2016) but also other studies show that cytoplasmic FoxP1 is associated with ER expression and reduced overall survival in invasive ductal carcinoma (Yu et al., 2018). Another small study showed that FOX1 expression is associated with improved survival in familial cancers, but also showed through immunohistochemical staining that expression of FOXP1 is higher in familial breast cancer as compared to sporadic breast cancer, that it correlates with Era and Erb expression in familial breast cancer (Rayoo et al., 2009). This rs72682900:*GADD45A* (+4,385) enhancer SNV also created a binding site for FOXF2. FOXF2 is has been found to negatively regulate TWIST1 (Wang et al., 2015) and FOXC2 (Cai et al., 2015) leading to inhibition of the epithelial-mesenchymal transition (EMT). In contrast, however, other studies have also shown FOXF2 to be capable of activating the EMT and thus function in an oncogenic manner in triple-negative tumours (Lo et al., 2016). Despite the unclear results from the literature, however, both FOXP1 and FOXF2 clearly have a role in breast cancer, and are interesting candidates for mediation of the effects of this SNV, and will require further study to validate.

The rs181313855:*TP53* (-2,242) enhancer SNV destroys an EP300 motif. The *EP300* gene encodes the protein p300, which is a transcriptional co-activator which functions as a histone acetyltransferase through chromatin remodelling. p300 is often found at enhancers (Visel et al., 2009). Therefore, loss of p300 at a region is likely to lead to decreased transcription, which is what we see in our luciferase assay. However, EP300 is known to be a non-sequence-specific TF. The source of the EP300 motif which was disrupted here is the ENCODE consortium, who state that such a motif likely represents a specific region which is bound by p300 in this context but would not capture all regions bound by p300 (Kheradpour & Kellis, 2014). Thus, it at least provides us with an indication that this region may be transcriptionally active.

The rs375724242: *CDH3* (+14,225) enhancer SNV creates a binding site for HIC1, which is a transcriptional repressor (Pinte et al., 2004). Therefore, creation of a binding site should decrease expression, as we see in the luciferase assay.

The rs78248633: *PIK3CA* #3 (+1,028) promoter SNV creates motifs for STAT1, STAT3 and NFkB1, which are all transcriptional activators. Creation of such motifs is consistent with the increase seen in our luciferase assay. STAT1 and STA3T play a role in cell growth, and NFkB1 is involved in a wide array of biological functions such as cell growth and apoptosis as well as inflammation and immunity.

Additionally, the rs116906793: *MRE11A* (-184,861) enhancer SNV destroyed a motif for ERRA. Although ERRA is related to the estrogen receptor, it is an orphan receptor, i.e. its ligand has not yet been identified. It can, however, bind to estrogen response elements (Ariazi & Jordan, 2006). This possible link to estrogen is a potentially exciting connection which warrants future analysis. Also of note, the rs116906793: *MRE11A* (-184,861) enhancer, had a 4.5-fold decrease in transcription over the wild type. This is a substantial decrease compared to other published studies which have used luciferase assays to investigate the consequences of SNVs.

It was difficult to make correlations for the other SNVs; for many of the transcription factors listed it is unclear whether they are activators or repressors and their functions could be context or cell-specific. As an example, NFkB1 can act as a transcriptional repressor or activator.

Luciferase assays were carried out in MCF7 cells and HEK293T cells. However, as these are germline variants it is important to know the function of the SNVs in non-cancerous context. Towards this end, we are currently repeating some of these luciferase assays in MCF10A cells.

One limitation of the luciferase assay is that the region of interest is cloned into a plasmid, and thus is not assayed in its native chromatin context, an aspect which is particularly crucial for regulatory regions. Additionally, it is important to remember that such an assay tells us nothing about the effect of the regulatory region on the hypothesised target gene – as that gene is not included in the plasmid. For these reasons, follow-up studies such as the ones presented in chapter 4 are critical.

4. Functional investigation of variants and characterisation of their mechanism of action

4.1. Introduction

Once I had established that some of the selected SNVs were capable of altering transcription, and thus potentially pathogenic, I selected 15 of them for further functional analysis. In this chapter, I aimed to delete the regions surrounding the variant, and assess the effect on gene expression of the flanking putative target genes.

To do this, I first needed to compile a list of potential target genes for our regulatory regions. This is not a trivial task, as enhancer-promoter connections can be complex. Enhancers can regulate more than one gene, and one gene can be regulated by more than one enhancer. Additionally, the distance between enhancers and their target genes can vary greatly. Annotating enhancers to their nearest promoter is a common approach to address this problem, and annotating them to the nearest expressed gene has been suggested as a further improvement to this approach. Yet both these methods likely result in many incorrect associations. A 2012 ENCODE paper which used 5C to assess interactions between promoters and distal elements, found that only 27% of distal elements had a looping interaction with their nearest TSS. When gene expression was taken into account this figure increased, yet still, just 47% of distal elements had a looping interaction with their nearest expressed TSS (Sanyal et al., 2012).

As even our promoter variants are located in proximal and distal promoters rather than core promoters, I included them in this analysis, in order to cover all possible target genes. Therefore the first aim of this chapter was to identify potential target genes of the regulatory loci containing our SNVs.

Secondly, I wished to delete the region surrounding the SNV and assess the effects of this on the expression of the potential target genes. I used the CRISPR/Cas9 genetic editing method to achieve this. Such a method has been used for similar aims in other studies.

4.2. Results

4.2.1. Identification of probable target genes of each variant locus

To carry out further functional analysis, I needed to know what the target gene of each non-coding region was likely to be, by linking enhancers and distal promoters to their target genes. To determine the possible target genes for each SNV, three complementary analyses were carried out.

Firstly, I obtained the nearest expressed gene to each SNV. To do this, I obtained RNA-seq data from the EBI expression atlas, a public database of RNA and protein expression experiments (Petryszak et al., 2016). Here I was able to obtain MCF10A RNA-seq data from an experiment entitled “RNA-seq of 675 commonly used human cancer cell lines” performed by Genentech and re-analysed by the EBI. I defined a gene as expressed if it had an FPKM >1.

Secondly, given the median size of a TAD has been found to be 185 kb with most under 2 Mb (Rao et al., 2014). I obtained a list of all genes 1 Mb up- and downstream of each SNV. Next, I wished to extract all genes suspected to be involved in cancer from this list. I did this by taking the 254 genes found to be significantly mutated in any cancer as according to the Tumor Portal analysis (Lawrence et al., 2014). I also used a list of 312 genes involved in breast cancer, which had been previously curated by our lab (Aloraifi et al., 2015). I also used the 723 genes in the COSMIC cancer gene census - an expert-curated list of genes which contain mutations that have been causally implicated in cancer (Sondka et al., 2018). This resulted in a master list of 1081 cancer-associated genes. I then cross-referenced my master list of 1081 cancer-associated genes with the genes falling within 1Mb of my SNVs. Importantly, I wished to restrict my analyses to only those genes which were expressed in MCF10A cells, so I again used the RNA-seq dataset

described above to filter out genes which were not expressed. This resulted in just 30 genes (Table 4.1).

Thirdly, I searched for functional data from chromatin conformation capture experiments to link the regulatory regions to any gene. To do this, I obtained ENCODE ChIA-PET data, which was available in MCF7 but not MCF10A cells. ENCODE had 4 RNA Polymerase II replicates, 2 CTCF replicates, and 3 ER-Alpha replicates. No SNVs were located in ER-Alpha interactions, two SNVs (rs553237095:*GATA3* #2 (-1,882) promoter and rs141940507:*NOTCH1* (-79,891) enhancer) were located in CTCF interactions, and all SNVs (except rs116906793:*MRE11A* (-184,861) enhancer) were located in the RNA Polymerase II interactions. Many of these interactions led to unannotated non-coding regions, but some linked the SNVs to genes. As before, only those genes expressed in MCF10A cells were retained. In total, 13 additional genes were added to our analyses. A summary of all potential target genes and their reason for inclusion can be seen in Table 4.1.

ID	Label	Potential	RNA-seq	Reason for inclusion
		Target Genes	TPM	
rs72682900: <i>GADD45A</i> (+4,385)	Enh.	<i>GADD45A</i>	94	Nearest expressed gene, Cancer gene within 1Mb
		<i>GNG12</i>	271	Link from ChIA-PET
rs116906793: <i>MRE11A</i> (-184,861)	Enh.	<i>MRE11A</i>	32	Cancer gene within 1Mb
		<i>AMOTL1</i>	81	Nearest expressed gene
rs766631446: <i>ERBB2</i> (+4,587)	Enh.	<i>ERBB2</i>	23	Nearest expressed gene, Cancer gene within 1Mb, Link from ChIA-PET
		<i>CDK12</i>	39	Cancer gene within 1Mb
		<i>LASP1</i>	184	Cancer gene within 1Mb
		<i>RARA</i>	8	Cancer gene within 1Mb
		<i>RPL23</i>	582	Cancer gene within 1Mb
		<i>SMARCE1</i>	26	Cancer gene within 1Mb
		<i>PGAP3</i>	8	Link from ChIA-PET
		<i>MIEN1</i>	50	Link from ChIA-PET
		<i>GRB7</i>	19	Link from ChIA-PET
rs369388483: <i>PIK3R1</i> (+1,201)	Pro.	<i>PIK3R1</i>	75	Nearest expressed gene, Cancer gene within 1Mb, Link from ChIA-PET
		<i>CCNB1</i>	446	Cancer gene within 1Mb
		<i>MCIDAS</i>	4	Link from ChIA-PET
rs146585339: <i>PIK3R1</i> (+5,520)	Enh.	<i>PIK3R1</i>	75	Nearest expressed gene, Cancer gene within 1Mb, Link from ChIA-PET
		<i>CCNB1</i>	446	Cancer gene within 1Mb
rs181313855: <i>TP53</i> (-2,242)	Enh.	<i>TP53</i>	101	Cancer gene within 1Mb, Link from ChIA-PET
		<i>WRAP53</i>	14	Nearest expressed gene, Cancer gene within 1Mb
		<i>AURKB</i>	80	Cancer gene within 1Mb
		<i>GPS2</i>	4	Cancer gene within 1Mb
		<i>KDM6B</i>	6	Cancer gene within 1Mb
		<i>PER1</i>	35	Cancer gene within 1Mb
		<i>PHF23</i>	58	Cancer gene within 1Mb
		<i>EFNB3</i>	4	Link from ChIA-PET
		<i>MPDU1</i>	74	Link from ChIA-PET
rs561575495: <i>BRAF</i> (+1,469)	Pro.	<i>BRAF</i>	10	Nearest expressed gene, Cancer gene within 1Mb
		<i>MRPS33</i>	39	Link from ChIA-PET
rs141940507: <i>NOTCH1</i> (-79,891)	Enh.	<i>NOTCH1</i>	19	Cancer gene within 1Mb, Link from ChIA-PET
		<i>EGFL7</i>	2	Nearest expressed gene, Link from ChIA-PET
		<i>SNHG7</i>	20	Link from ChIA-PET
		<i>SEC16A</i>	50	Link from ChIA-PET
		<i>RABL6</i>	45	
rs553237095: <i>GATA3</i> #2 (-1,882)	Pro.	<i>GATA3</i>	19	Nearest expressed gene, Cancer gene within 1Mb
		<i>TAF3</i>	14	Cancer gene within 1Mb, Link from ChIA-PET
		<i>TCF7L2</i>	27	Link from ChIA-PET
rs372644648: <i>GATA3</i> (+16,144)	Enh.	<i>GATA3</i>	19	Nearest expressed gene, Cancer gene within 1Mb
		<i>KIN</i>	7	
		<i>TAF3</i>	14	
rs78248633: <i>PIK3CA</i> #3 (+1,028)	Pro.	<i>PIK3CA</i>	25	Nearest expressed gene, Cancer gene within 1Mb
		<i>ZNF639</i>	24	Link from ChIA-PET
		<i>ZMAT3</i>	48	
		<i>NDUFB5</i>	41	
rs190491586: <i>ATM</i> (-1,134)	Pro.	<i>ATM</i>	14	Cancer gene within 1Mb
		<i>NPAT</i>	14	Nearest expressed gene
		<i>DDX10</i>	40	Cancer gene within 1Mb
rs74951121: <i>BRCA2</i> (-36,832)	Enh.	<i>BRCA2</i>	11	Cancer gene within 1Mb, Link from ChIA-PET
		<i>FRY</i>	2	Nearest expressed gene
rs184741182: <i>EP300</i> (-934)	Enh.	<i>EP300</i>	11	Cancer gene within 1Mb
		<i>TOB2</i>	35	Link from ChIA-PET
		<i>RBX1</i>	55	Link from ChIA-PET
		<i>RANGAP1</i>	167	Link from ChIA-PET
rs574133962: <i>KRAS</i> (+1,174)	Pro.	<i>KRAS</i>	33	Nearest expressed gene, Cancer gene within 1Mb
		<i>RASSF8</i>	24	

Table 4.1: Potential target genes for each SNV locus.

Potential target genes were identified by annotating the SNV to the nearest expressed gene, by selecting all expressed cancer-associated genes within 1 Mb and by interrogating CTCF, ER-Alpha and RNA Polymerase II ChIA-pet interactions in the MCF7 cell line. Also shown is the RNA-expression in the MCF10A cell line in transcripts per million (TPM) units.

4.2.2. Strategy for the CRISPR-mediated deletion of variant loci

To determine the function of the SNV locus, I developed a strategy to delete the region containing the SNV and measure the effect on the expression of potential target genes. This would allow us to confirm which genes might be affected by the SNV and hypothesise potential mechanisms of action. To delete the SNV locus, I used the CRISPR/Cas9 genetic editing system. This strategy is outlined in Figure 4.1. I designed two guide RNAs, one targeted to either side of the SNV. This would result in the Cas9 cutting the DNA at the locations of the guide RNAs, effectively resulting in a small deletion of approximately 200 bp around the SNV. For redundancy, I designed two pairs of guide RNAs per SNV.

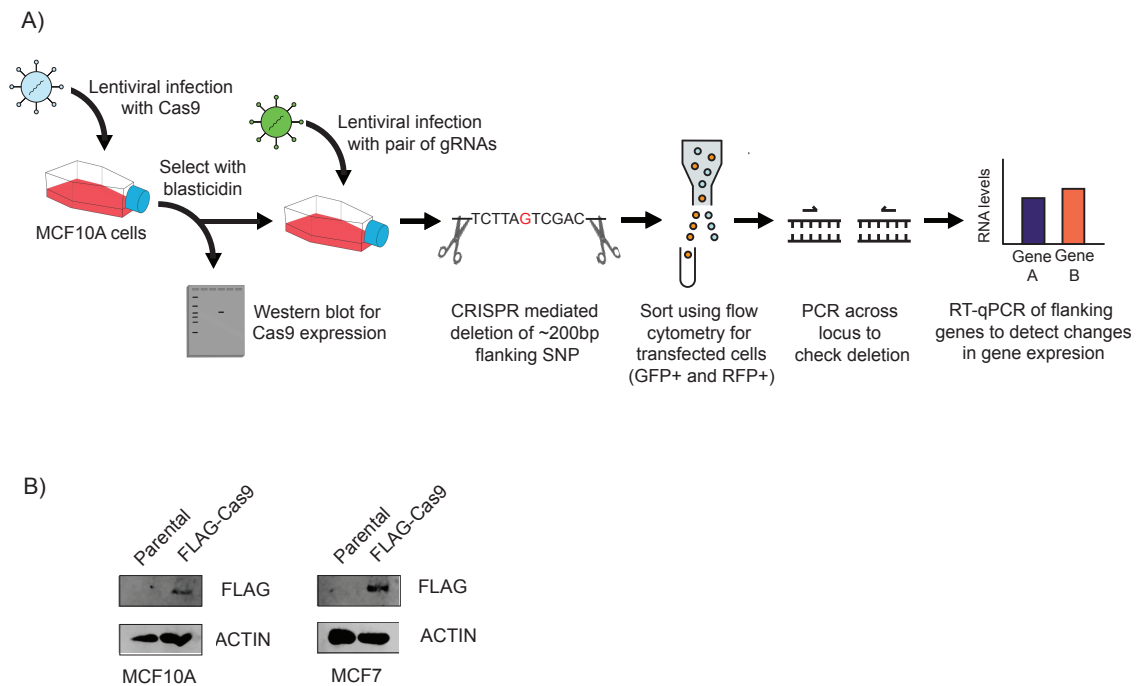


Figure 4.1 Strategy for the CRISPR-mediated deletion of variant loci.

A) MCF10A cells were infected with Cas9, so they would stably express Cas9. These cells were then infected with two guide RNAs, one targeted to either side of the SNV. Cells were sorted for GFP and RFP double-positive cells, then screened for deletions by PCR. qRT-PCRs were then carried out on the hypothesised target and flanking genes.

B) Western blot showing stable expression of Cas9 in the MCF10A cell line. Also shown is an MCF7 cell line expressing Cas9.

4.2.3. Strategy to screen edited pools of MCF10A cells for CRISPR deletions

To determine if the regulatory regions in the MCF10A cells were successfully deleted, I performed a PCR to amplify the SNV locus. By running the resulting amplicon on a gel I could identify the size of the DNA fragments – if the region was successfully deleted, then the DNA fragment would be ~200 bp shorter. As this is a mixed population of cells, we would expect some cells to contain the deletion and some to remain un-edited, thus resulting in fragments of two different sizes on the gel. Additionally, the efficiency of editing can be estimated by the relative intensities of the upper and lower bands on the gel.

I performed PCRs on the 15 SNV loci, and successfully obtained deletions in 10 of them (Figure 4.2). Unfortunately, the deletion efficiency was low for many of the SNVs, and because of this, the effect of the deletions on gene expression is likely to be underestimated by any qRT-qPCRs. There were four SNVs which appeared to have relatively good efficiencies: rs372644648:*GATA3* (+16,144) enhancer gRNA pair2, rs78248633:*PIK3CA* #3 (+1,028) promoter gRNA pair2, rs141940507:*NOTCH1* (-79,891) enhancer gRNA pair1 and rs574133962:*KRAS* (+1,174) promoter gRNA pair2 all appear to have large proportions of edited cells. These SNVs look to have approximately >30% of the cells with a deletion (Figure 4.2; top row).

4.2.4. Deletion of regulatory loci and the effect on expression of breast cancer-related target genes

Having obtained good deletion efficiencies with four SNVs, I then extracted RNA and carried out RT-qPCRs for expression of the putative target genes listed in Table 4.1. None of the deletions resulted in decreased expression of our selected candidate target gene (Figure 4.3). However, deletion of the rs78248633:*PIK3CA* (+1,028) promoter SNV showed decreased expression of *ZMAT3* and increased expression of *ZNF639*. Reassuringly, the expression of *GAPDH* did not change after CRISPR. It was introduced as a negative control as it is not within the same TAD as the possible target genes, and is on a different chromosome to them it should not be affected by deletion of the region surrounding the SNVs.

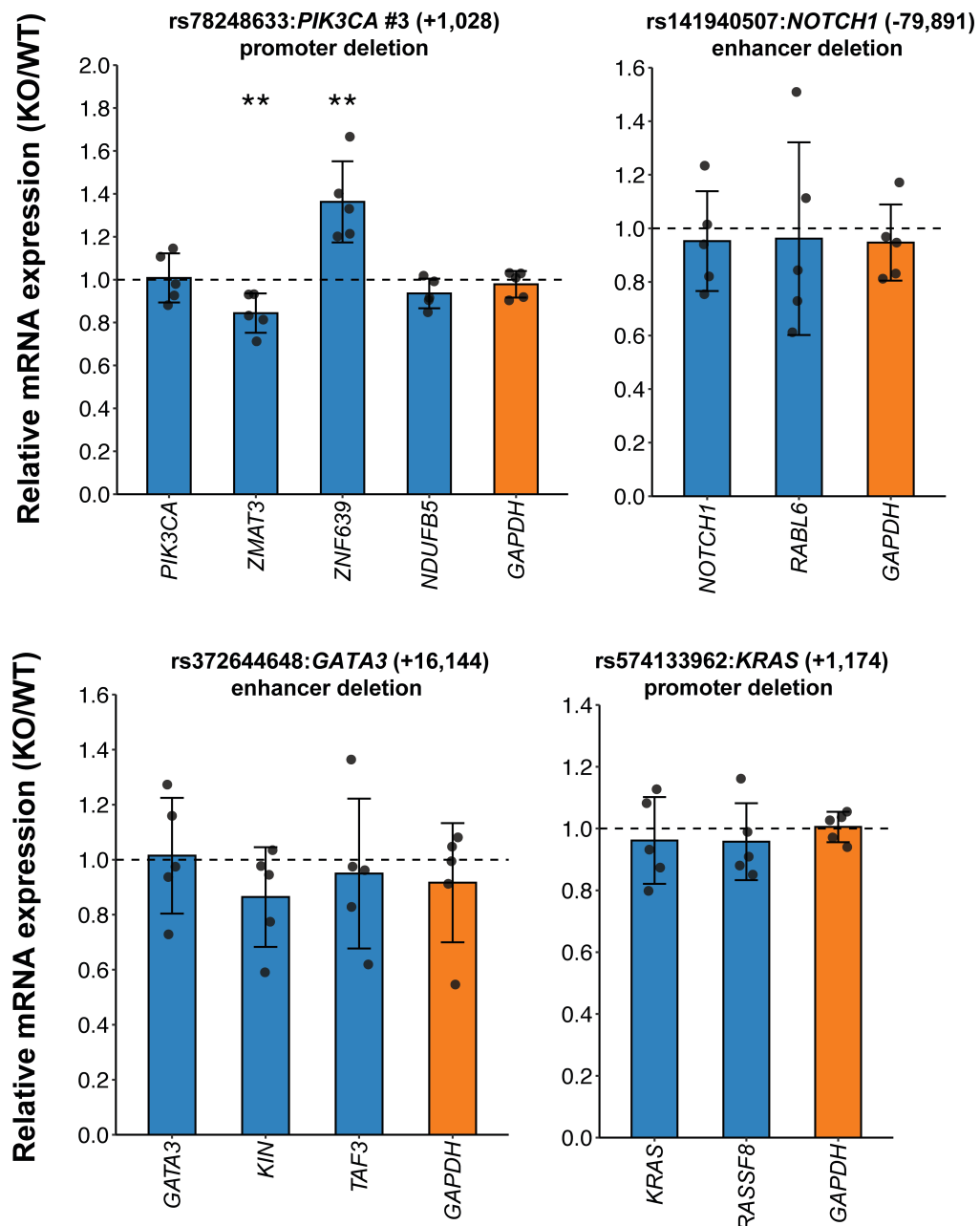


Figure 4.3: RT-qPCR of potential target genes after deletion of the SNV regulatory regions using CRISPR.

Quantitative RT-PCR analyses of the indicated mRNA transcripts in pools of cells containing the desired CRISPR regulatory region deletion. None of the deletions resulted in decreased expression of our hypothesised candidate target gene. However, deletion of the *PIK3CA* promoter SNV region showed decreased expression of *ZMAT3* and increased expression of *ZNF639*. Note that expression of the control, non-TAD gene *GAPDH* does not change after deletion. Results were firstly normalised to levels of the *RPLPO* housekeeping gene, then normalised to expression in WT cells. The error bars indicate standard deviation of five replicates.

4.3. Discussion

In this chapter, I aimed to delete the regulatory regions flanking our case-enriched SNVs and assess the effects on potential target gene expression. I selected 15 SNVs which had shown changes in transcription in the luciferase reporter assay. I used the CRISPR/Cas9 genetic engineering system with two pairs of gRNAs targeted to either side of the SNV. I then assessed the changes in gene expression changes after deletion of four of these variant regions: rs78248633:*PIK3CA* #3 (+1,028) promoter, rs141940507:*NOTCH1* (-79,891) enhancer, rs372644648:*GATA3* (+16,144) enhancer and rs574133962:*KRAS* (+1,174) promoter. Three of them showed no changes in expression of potential target genes while one did; deletion of the rs78248633:*PIK3CA* #3 (+1,028) promoter region showed decreased expression of *ZMAT3* and increased expression of *ZNF639*.

ZMAT3, also known as Wig1, is a target gene of p53, and can regulate cell cycle death and cell apoptosis (Bersani et al., 2014). Reduced expression shows that our SNV lies in a regulatory region capable of affecting *ZMAT3*. However, in the luciferase assay, this SNV resulted in an increase in transcription, meaning that it has the potential to increase expression of *ZMAT3*, although further experiments are needed to determine the oncogenic potential of *ZMAT3*. Deletion of this *PIK3CA* promoter region also resulted in an increase in *ZNF639*, also known as *ZASC1*, a zinc finger protein capable of binding DNA which has been found to be overexpressed in oesophageal squamous cell carcinomas (Imoto et al., 2003) and showed copy number increases in oral carcinoma (Chiang et al., 2011). Again, further experiments are needed, but this established link to cancer is promising.

When compiling target genes, I initially attempted to use HiC performed in HMEC cells to define TADs, and then assess the effect on expression of all these genes through RT-qPCR. However, TAD the resolution was very low, meaning the TAD boundaries are very large, and many genes fell into the TADs, too many to RT-qPCR (Rao et al., 2014).

Therefore, alternative methods were needed to define potential target promoters. In an ideal scenario, I would have preferred to use RNA-seq to determine the target genes of our enhancers. I would have done this by performing RNA-seq in the parental cell line, and again post-CRISPR, and compared the two to figure out what genes changed expression, and thus what ones are under the control of the deleted locus. However, RNA-seq is costly, especially as we are investigating 15 SNVs. Another excellent alternative would be to perform one of the chromosome conformation capture techniques (3C, 4C, 5C), but this was also not feasible. I have proposed a cost-effective alternative method which offers a good chance of finding the affected target genes.

Another limitation of the work performed in this chapter is that MCF10A cells do not grow to single-cell density. The RT-qPCRs were performed on a pool of cells with a mixed population, where some cells were edited successfully, and some were not. As a consequence of this, the gene expression changes seen in the RT-qPCRs are likely to be underestimated. To overcome this issue, I could have used the MCF7 cell line, which is a cancer cell line. These cells do grow at single-cell density, but they are cancer cells, and so are not as physiologically relevant as MCF10A cells and have abnormal karyotypes, which make it difficult to ensure that the CRISPR has edited every locus. Instead, I have successfully extracted homogenous cells by growing them at an extremely low density, so that the cells are grown in proximity to each other, but a single colony of homogenous cells can still be isolated. This work is ongoing, but we will repeat the RT-qPCRs in these clonal cell lines.

Our final aim is to choose the most promising variant and attempt to introduce just the single nucleotide change into the cell line (as opposed to generating cells carrying a ~200 bp deletion), repeat the RT-qPCR assays, and perform other biochemical characterisations such as growth and proliferation assays.

In addition to the ongoing work on the CRISPR deletions described in this chapter, we are also using the complementary techniques of CRISPR interference (CRISPRi) and CRISPR activation (CRISPRa). With these techniques, gRNAs can be used to recruit a catalytically inactive version of the Cas9 protein which lacks endonuclease activity (dCas9), to a particular region of interest. The dCas9 then blocks transcription in a process known as CRISPRi (Qi et al., 2013). Furthermore, this dCas9 can be fused to effector proteins such as the transcriptional repressor domain KRAB or the transcriptional activator VP64 to further repress or activate gene expression respectively (Gilbert et al., 2014).

For this project, we are using CRISPRi and CRISPRa with gRNAs targeted to our promoters and enhancers, and subsequently performing RT-qPCRs of potential target genes. This will allow us to determine if our SNVs are located in functional regulatory regions and what their target genes are. These techniques are useful tools for assessing the regulatory capabilities of non-coding regions and have the advantage over our CRISPR deletions that they do not alter the DNA sequence. This work is currently ongoing in the lab, but we expect the results will add a second layer of evidence for whether or not our variants are located in functional non-coding regions.

5. Integrative functional analyses of common breast cancer GWAS SNPs at 172 loci

5.1. Introduction

To date, 172 loci have been identified from breast cancer GWASs, and are estimated to account for approximately 18% of the familial relative risk (Michailidou et al., 2017). This means that common, low-penetrance variants, account for a small but significant proportion of heritability, and it is therefore worth investigating their mechanism of action. GWASs compliment other investigative methods such as sequencing studies, and if we are to understand the complete inherited risks of breast cancer, it is critically important to address the impact of both rare and common variants on breast cancer risk and development.

Despite the success of GWASs at identifying SNPs and loci associated with a disease or trait, there are many limitations of GWASs. Crucially, the GWAS-identified SNP merely acts as a proxy for all SNPs in linkage disequilibrium (LD) within the haplotype block. Therefore, GWAS does not address the central question of which variant in the locus is causal. It reflects a statistical association but offers no insight into biological mechanisms. Follow-up studies are required to subsequently uncover the causal variant and to elucidate the mechanism of action. Such studies are challenging, hampered by the fact that ~95% of variants identified by GWAS fall into non-coding regions of the genome (Maurano et al., 2012). There remains a significant gap in the field and a crucial need for these functional follow-up studies.

This chapter aims to address this problem by annotating breast cancer GWAS variants with publicly available biochemical data in an attempt to identify the likely causal SNP at known GWAS loci. To do this, I hypothesised that functional SNPs were likely to be located in open chromatin and prioritised GWAS SNPs in these open chromatin regions.

One mechanism of how SNPs in non-coding regions can affect function is through the disruption of CTCF sites. Therefore, I investigated whether any of these SNPs were

capable of disrupting CTCF binding. Additionally, another way that SNPs in non-coding regions can exert their effects is through the disruption of a transcription factor binding site. I prioritised SNPs in super-enhancers and those affecting transcription factors thought to be important in breast cancer: FOXA1, ESR1, GATA3, E2F1 and TCF7L2.

This chapter builds on the work of the many previous breast cancer GWASs, but adds a new angle to the discussion, taking us closer to understanding the underlying biology. I shortlist some promising variants that would be suitable for further functional follow-up studies and discuss their possible mechanisms of action.

5.2. Results

5.2.1. Recurrent breast cancer GWAS-associated SNPs in an Irish cohort of familial breast cancer patients

When designing the target regions for this sequencing study, we included 88 genes identified in two meta-analyses of GWASs of breast cancer (Peng et al., 2011; Zhang et al., 2011). Additionally, we also included the regulatory DNA elements within the haplotype blocks of the most promising SNPs that showed significant association with breast cancer predisposition. This allowed us to extend our investigation to common variants, expanding on the previous two chapters, which had focused on rare variants.

After analysing the sequencing data, I found 151 breast cancer-associated lead or LD SNPs present in our cases and controls. To test whether any of these SNPs are enriched in cases versus controls, I used SnpSift to calculate the enrichment p-value using different genetic models: dominant, recessive, allelic and co-dominant. Two interesting loci emerged (Table 5.1). In both cases, the breast cancer-associated lead SNP was not significantly enriched in our cases over controls, while other LD SNPs in the loci were - suggesting that it is not the lead SNP, but rather another SNP in LD that may be the causal one. What is interesting about this is that in the case of the first locus, rs3817198, the variants enriched in our cohort were ones identified as credible risk variants in a fine-mapping study by Michailidou and colleagues (Michailidou et al., 2017). Although the sample size in this study is very small for an analysis of common variants, this work highlights two potentially interesting loci to follow up.

						Cases genotypes			Controls genotypes			P-values					
Chr	Pos	rsID	Ref	Alt	Source	Homo.	Hetero.	Total	Homo.	Hetero.	Total	Cochran-Armitage trend model	Genotypic model	Allelic model	Dominant model	Recessive model	
chr11	1909006	rs3817198	T	C	Lead SNP	14	46	74	15	43	73	0.3108	0.5867	0.1793	0.2025	0.3338	
chr11	1898664	rs1973765	T	C	CRV	15	56	86	6	33	45	0.0006	0.0017	0.0005	0.0004	0.0653	
chr11	1896957	rs588321	G	C	CRV	13	56	82	4	22	30	0.0029	0.0065	0.0025	0.0014	0.1592	
chr11	1895708	rs4980386	C	A	CRV	13	55	81	6	33	45	0.0029	0.0076	0.0024	0.0016	0.1313	
chr8	128355618	rs13281615	A	G	Lead SNP	25	45	95	17	33	67	0.1206	0.2388	0.0577	0.0628	0.2459	
chr8	128352967	rs11328185	CA	C	LD SNP	31	41	103	37	37	111	0.0250	0.0704	0.0098	0.0660	0.0235	
chr8	128352234	rs4871782	G	A	LD SNP	26	43	95	15	33	63	0.0417	0.1184	0.0174	0.0365	0.1010	
chr8	128344402	rs688201	G	T	LD SNP	23	47	93	14	32	60	0.0475	0.1009	0.0222	0.0241	0.1849	
chr8	128351048	rs59351655	A	T	LD SNP	25	44	94	15	33	63	0.0500	0.1307	0.0219	0.0365	0.1314	
chr8	128344591	rs687324	T	C	LD SNP	24	46	94	15	33	63	0.0576	0.1299	0.0266	0.0326	0.1770	
chr8	128351442	rs56898669	A	C	LD SNP	24	46	94	15	33	63	0.0576	0.1299	0.0266	0.0326	0.1770	
chr8	128344410	rs595821	G	A	LD SNP	23	46	92	14	32	60	0.0593	0.1329	0.0277	0.0333	0.1849	
chr8	128344380	rs389143	C	T	LD SNP	24	46	94	15	30	60	0.0601	0.1089	0.0270	0.0254	0.2210	
chr8	128351773	rs10956363	G	A	LD SNP	23	47	93	15	33	63	0.0687	0.1354	0.0330	0.0326	0.2217	
chr8	128343879	rs420101	G	A	LD SNP	24	44	92	15	32	62	0.0708	0.1648	0.0316	0.0432	0.1770	
chr8	128344601	rs594869	T	C	LD SNP	24	45	93	15	33	63	0.0714	0.1674	0.0330	0.0445	0.1770	
chr8	128344602	rs594868	G	A	LD SNP	24	45	93	15	33	63	0.0714	0.1674	0.0330	0.0445	0.1770	
chr8	128352730	rs10087810	C	T	LD SNP	30	43	103	32	30	94	0.0749	0.1970	0.0332	0.0654	0.1256	
chr8	128351524	rs61599109	A	AG	LD SNP	24	44	92	15	33	63	0.0875	0.2104	0.0406	0.0594	0.1770	

Table 5.1: Common SNPs are enriched in cases versus controls at two breast cancer GWAS loci.

Note that the lead SNP in both of these loci is not enriched in cases, suggesting that the lead SNP may not be the causal SNP.

Statistical significance for enrichment was calculated using four different models, and significant p-values are in bold.

5.2.2. Annotation and functional analyses of 172 breast cancer GWAS loci

To date, 172 breast cancer risk loci have been identified through GWAS. I wished to identify which SNPs in these loci were candidate causal SNPs. Firstly, I identified all SNPs in LD with the 172 breast cancer-associated lead SNPs. Additionally, I added the set of 'credible risk variants' (CRVs) at 142 of the GWAS loci from a recent GWAS meta-analysis (Michailidou et al., 2017). The authors of this meta-analysis, pursued a statistical fine-mapping approach, in an attempt to reduce the number of candidate causal variants at these 142 loci. Such a fine-mapping approach requires genotyping all known variants within the original GWAS-identified locus, then retaining only the variants with the strongest association to the trait, as these are considered to be the best candidate causal variants (Udler et al., 2010). Instead of taking all SNPs in LD with a lead SNP, the authors included only their fine-mapped CRVs, where CRVs were defined as variants which were located within 500 kb of the most significant SNP in each region, and with p-values within two orders of magnitude of the most significant SNPs in each region, i.e. they excluded all variants with a likelihood of <100 worse than the strongest SNP in the locus, a threshold which was determined in Udler et al., 2010.

In order to maximise the set of variants for functional analysis in this chapter, I included all SNPs in LD with the lead SNP at 172 loci, as well as CRVs at 142 loci. Including only fine mapped CRVs would have necessitated restricting my analysis to 142 loci instead of 172. Additionally, CRVs were determined in a different dataset to the LD SNPs, and so I expected that there may be differences in identified SNPs, which indeed there was. Although the LD and CRV sets were largely overlapping, the CRVs contained some SNPs not included in my LD set, and inversely, the LD SNPs contained some not included in the CRV set, therefore I felt it was important to include all SNPs. This complete set of breast cancer CRV and GWAS-derived lead and linked SNPs consisted of 6,270 SNPs which were taken forward for further analysis.

Although the vast majority of breast cancer-associated lead SNPs are located in non-coding regions, I wished to investigate whether any SNPs in LD were located within exons, and thus could explain the locus pathogenicity. I determined the most severe consequences of all the linked SNPs and found that 46 loci contained SNPs which were located in exons or at exon-intron boundaries that could affect splicing, with just 8 of those 46 loci containing variants that could result in a truncated protein (Figure 5.1A). The remaining loci contained SNPs located only in non-coding regions, which are of interest to us in this project, and thus it was necessary to develop a system to determine the likely causal variants.

In order to identify the SNPs which are most likely to be functional, I extracted all SNPs located within open chromatin. To do this, I obtained ATAC-seq data in the MCF10A cell line (Fritz et al., 2018) as well as ATAC-seq from breast tumours (Corces et al., 2018). I then intersected 6,270 SNPs with the ATAC-seq peaks and found that 716 SNPs were located in open chromatin in either MCF10A cells or breast tumours, while 110 SNPs were located in open chromatin in both (Figure 5.1B).

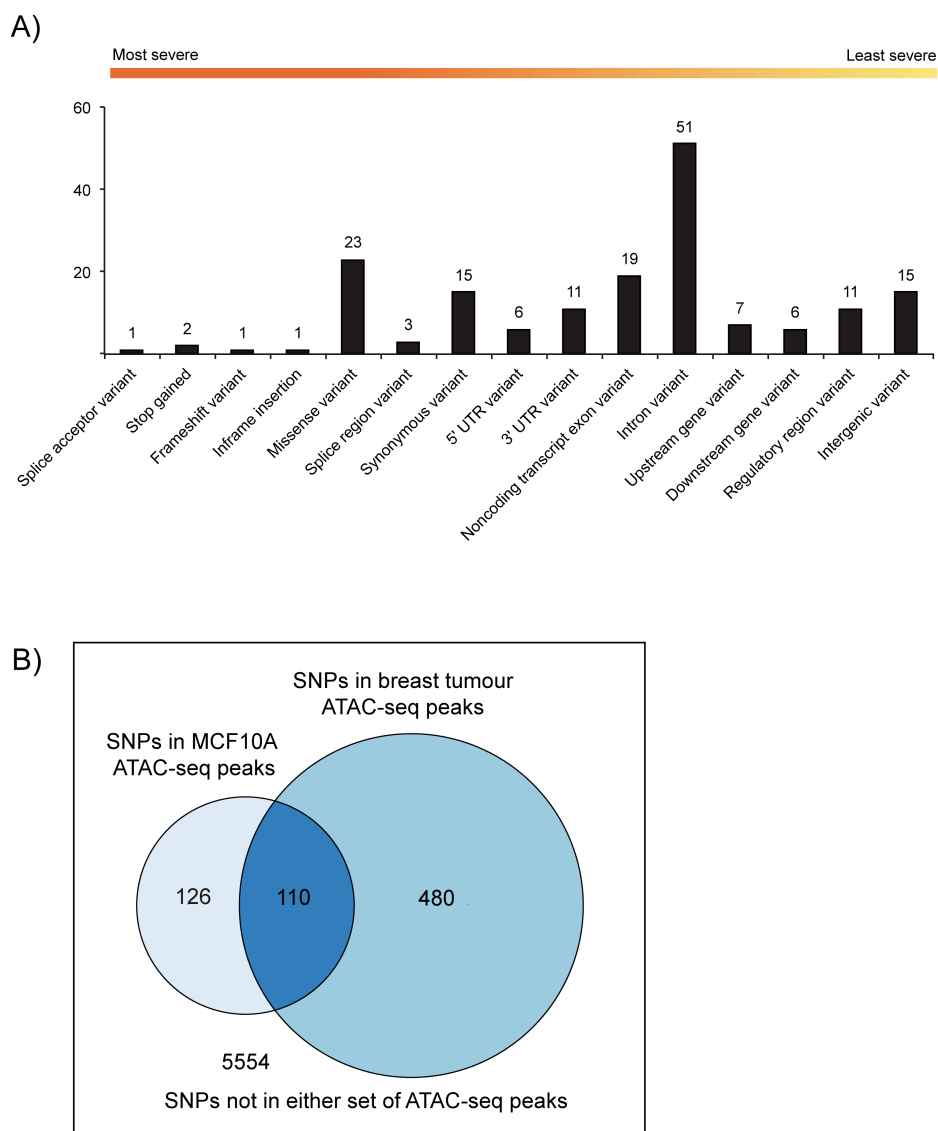


Figure 5.1: Localisation of SNPs in 172 breast cancer GWAS loci.

A) The most severe consequence of SNPs at each risk locus. Just 46 loci contained SNPs located in exons or at exon-intron boundaries that could affect splicing. The remaining loci only contained SNPs located in non-coding regions, complicating their interpretation.

B) 716 breast cancer-associated lead and linked SNPs were located in open chromatin, as defined by ATAC-seq. These were the SNPs which were prioritised for further analysis.

Once I had identified the SNPs located in open chromatin, I next wished to investigate the likely mechanisms of action of these SNPs. I wished to determine whether they affected the binding of any transcription factors. A previous study which focused on breast cancer-associated SNPs found that they are located in enhancer regions and co-localise with the transcription factors FOXA1 and ESR1 (Cowper-Salari et al., 2012). Additionally, the set of CRVs mentioned above (Michailidou et al., 2017) was enriched for transcription factor binding sites for FOXA1, ESR1, GATA3, E2F1 and TCF7L2. Therefore, these were the transcription factors I decided to focus on for this analysis. However, these two studies had carried out their analyses on a larger set of breast cancer GWAS-derived SNPs; as I had filtered my SNPs for only those which were located in open chromatin, I wished to see if the association with these transcription factors still held, or if there was evidence of any other transcription factors being involved. In order to do this, I examined whether our SNPs were co-localised with any particular transcription factors more than expected by chance.

I carried out this analysis in ENCODE ChIP-seq data sets for all proteins in any cell line relevant to breast cancer. The number of intersections of TF binding with a background set of 716 random SNVs in open chromatin was compared to the number of intersections of TF binding with the 716 GWAS-associated SNPs in open chromatin. There were just two TFs that had a p-value <0.05 and a fold-enrichment >2 : ESR1 and TCF7L2 (Figure 5.2). These are two transcription factors which were already implicated in breast cancer, therefore no new transcription factors were added to my subsequent analysis.

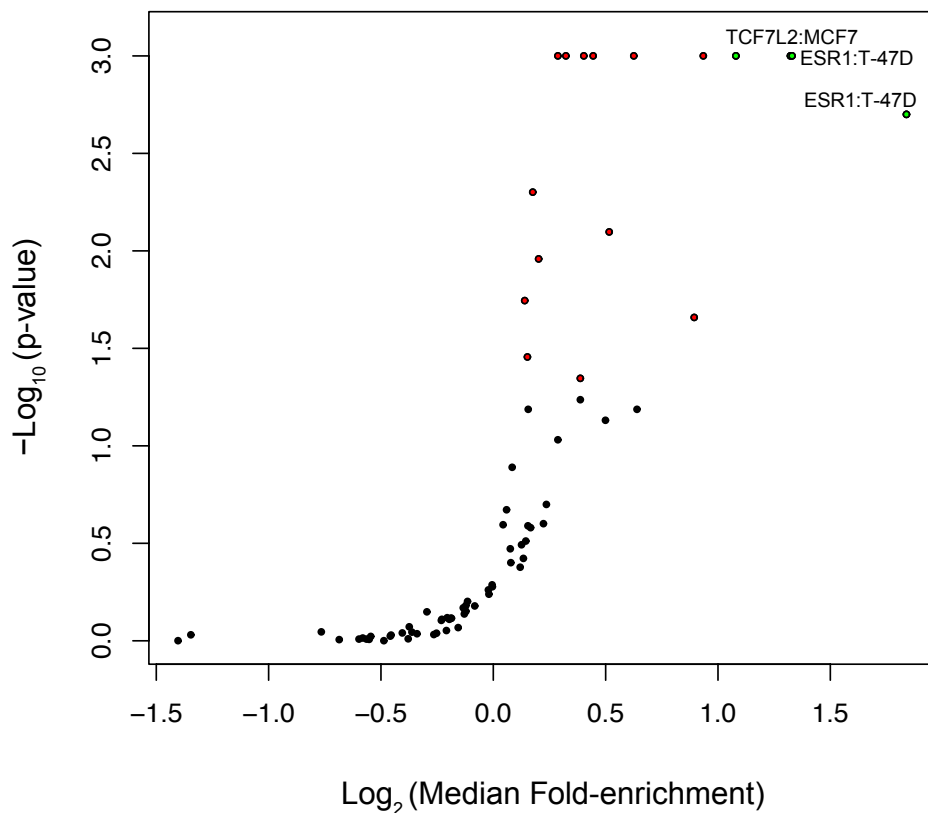


Figure 5.2: Enrichment of transcription factors binding at GWAS-linked SNPs in open chromatin.

TF binding at 716 breast cancer-associated lead and linked SNPs in open chromatin was compared to a background set of 716 random SNVs in open chromatin. For each TF in each cell line, I sampled 1000 sets of 716 random SNPs in open chromatin as defined by ATAC-seq in the MCF10A cell line and breast tumours. The number of intersections of TF binding with random SNVs (expected intersections) was compared to the number of intersections of TF binding with the GWAS-associated SNPs (observed intersections). Median fold enrichment was defined as the number of observed intersections divided by the median of the 1000 expected intersections of the background set. The p-value was then defined as: $p\text{-value} = (1 + \sum(\text{expected intersections} \geq \text{observed intersections})) / (1001)$. The x-axis represents the fold enrichment over background, while the y-axis represents significance. No adjustments were made to correct for multiple testing. TFs plotted in red have a p-value < 0.05, and those in green, with names, have a p-value < 0.05 and a fold enrichment of ≥ 2 .

5.2.3. GWAS-linked SNPs disrupt CTCF sites and transcription factor binding sites in super-enhancers

The next potential mechanism of action I wished to investigate was whether any of the SNPs disrupted CTCF sites. Firstly, I overlapped the 716 SNPs in open chromatin with 11 different CTCF ChIP-seq datasets in various breast and breast cancer cell lines including HMEC, HMF, T47D, MCF7, MCF10A, MCF10AT1 and MCF10CA. 66 SNPs were located at a CTCF binding site in at least one condition in one of these cell lines (Figure 5.3). Next, I wished to see whether any of these SNPs were also located in a TAD boundary. To do this, I obtained HiC data from HMEC cells (Rao et al., 2014) and extracted all the TAD boundary regions which contained a CTCF motif. 19 of the 66 SNPs were located within a HiC boundary. Finally, I wished to determine if any of these 19 SNPs disrupted or created a CTCF motif. To do this, I used the program motifbreakR to determine if any SNPs disrupt or create TF binding motifs. I found that just 2 of these 19 SNPs affected a CTCF motif: rs1744959 (Figure 5.4) and rs1011623 (Figure 5.5). The seven variants which were located in super-enhancers and disrupted a motif for FOXA1, ESR1, GATA3, E2F1 or TCF7L2 were also retained for further analysis (Figure 5.3).

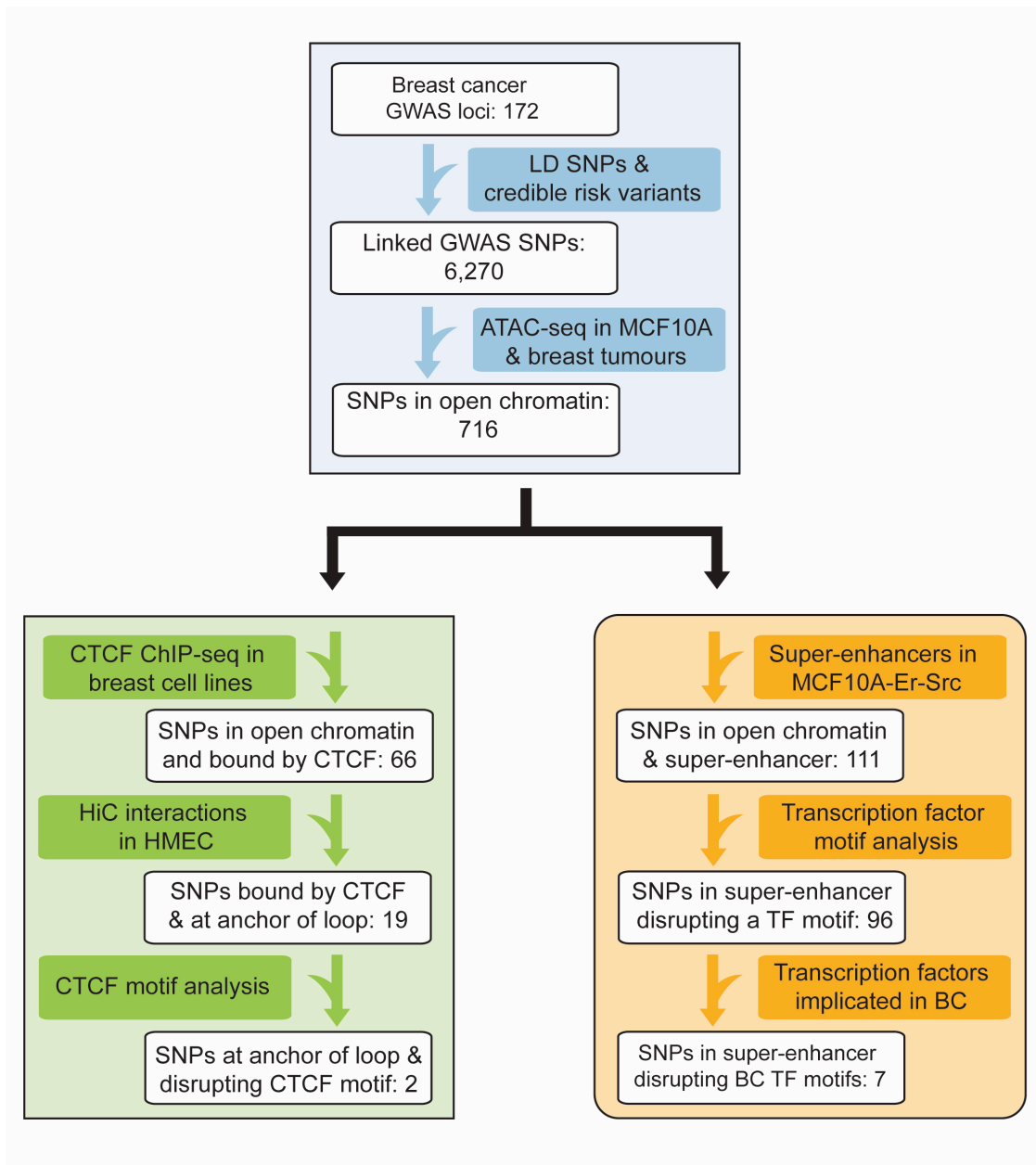


Figure 5.3: Overview of analysis for prioritising candidate causal GWAS SNPs.

Two SNPs were located in the anchor of a HiC loop in HMEC cells and disrupted or created a CTCF motif. Seven SNPs were located in super-enhancers and found to be capable of disrupting or creating a motif for a TF previously implicated in breast cancer.

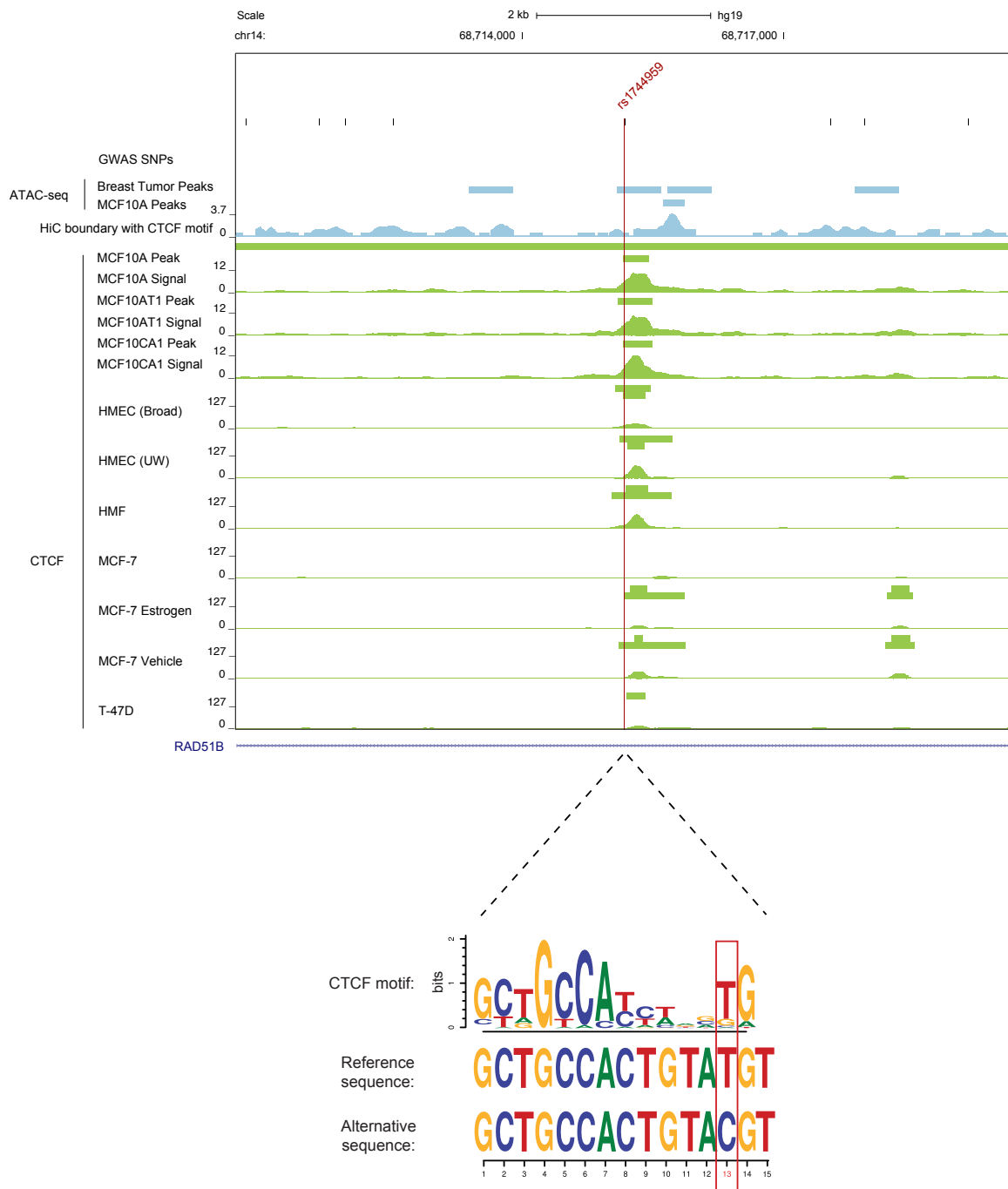


Figure 5.4: rs1744959 is located in a TAD boundary and disrupts a CTCF motif.

rs1744959 is located in the intron of RAD51B, a gene known to be essential for DNA repair.

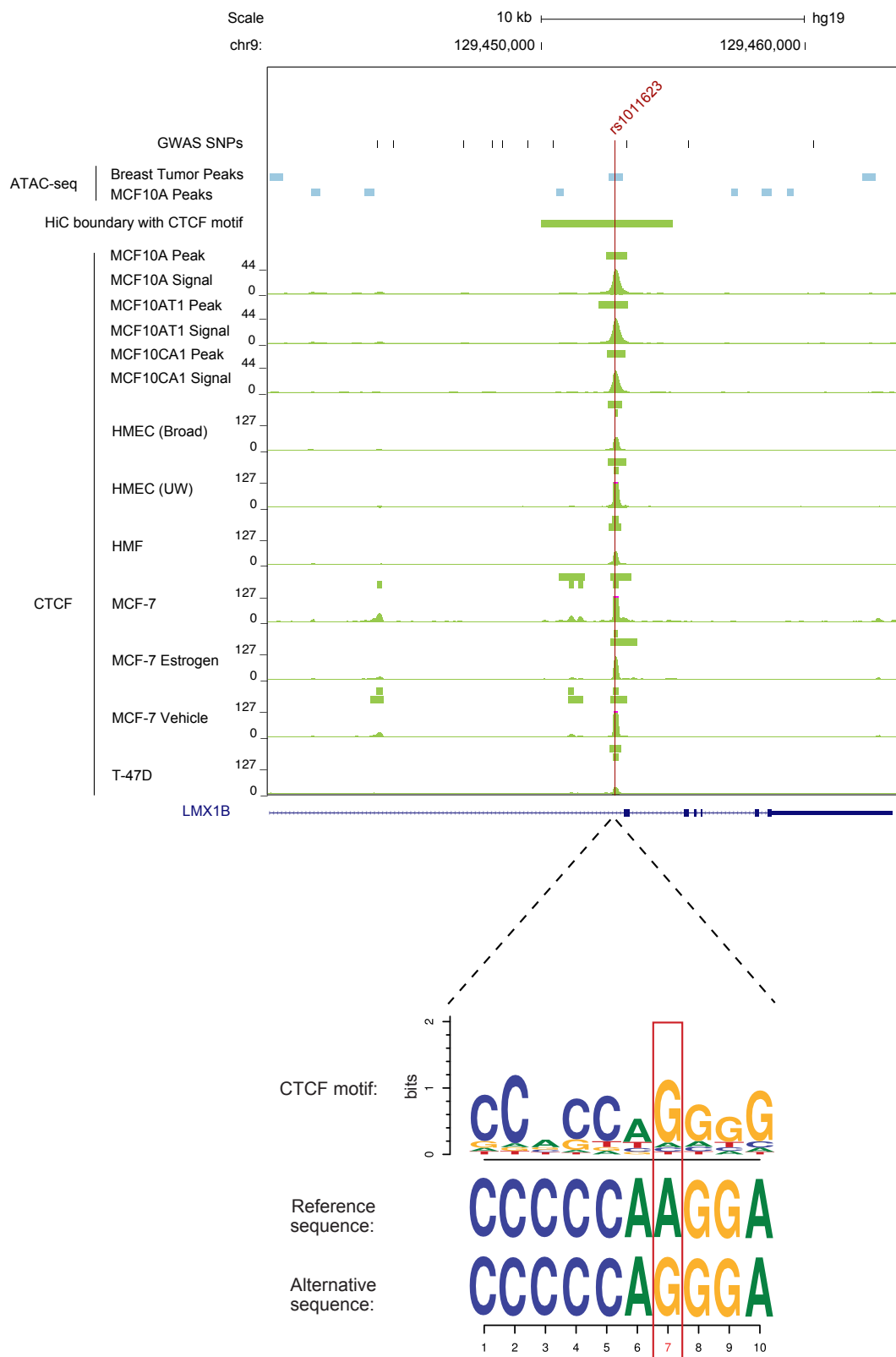


Figure 5.5: rs1011623 is located in a TAD boundary and creates a CTCF motif.

rs1011623 is located in an intron of LMX1B, which is a transcription factor involved in development.

5.2.4. Putative target genes of each candidate causal SNP

In order to determine the potential target gene of each SNP, I queried the Genotype-Tissue Expression (GTEx) project, a public database of expression quantitative trait loci (eQTLs). These eQTLs are derived from genotyping and analysing RNA expression in various human tissues. Just two SNPs, rs11669175 and rs11665924, which are at the same locus, were determined to be an eQTL for *ZNF404*, a zinc finger protein of undetermined significance. This locus has been studied before - another paper predicted these two SNPs, along with two others to be candidate causal SNPs (Liu et al., 2017). These authors also suggest that rs11669175 affects USF1 and ZBTB33 binding, while rs11665924 affects GRHL2 binding.

Secondly, I downloaded predicted enhancer-to-gene links (Corces et al., 2018). In this paper, the authors used a Pearson correlation between the accessibility of an ATAC-seq peak and a gene's expression to infer which peaks of open chromatin were likely to regulate which genes. For breast cancer, they identified 9,711 ATAC-seq peak-to-gene links, which I overlapped with our 14 SNPs. Again, just two SNPs showed an association, rs12609846 to *KCNN4*, a calcium-activated potassium channel, and rs11086307 which is associated with *CEBPB*, which is involved in immune and inflammation responses.

Thirdly, I obtained eQTL associations calculated using GWAS SNPs and the TCGA and METABRIC datasets (Michailidou et al., 2017). The authors of this paper used their heuristic scoring system, INQUISIT, to predict target genes for every GWAS locus, and they were able to make predictions for 128 out of 142 loci analysed. For the two SNPs of which Corces et al. (2018) could predict the target, Michailidou et al. (2017) predicted different genes. There was a lack of corroboration between the three methods, and no breast cancer-related genes were identified as potential targets for any of these SNPs (Table 5.2).

rsID	Ref	Alt	GWAS lead SNP rsID	Motif disrupted	Motif created	Gene in proximity	GTEX eQTL - mammary tissue	GTEX eQTL - any tissue	Corces et al. prediction	Michailidou et al. prediction
rs10116233	A	G	rs10760444	.	CTCF	LMX1B	.	.	.	GARNL3
rs1744959	T	C	rs2588809	CTCF	.	RAD51B	.	.	.	.
rs17755925	A	G	rs999737	FOXA1:AR	.	RAD51B	.	.	.	.
rs897786	T	C	rs2594714	.	E2F1	.	.	MRI1, ZSWIM4	.	ILVBL, NDUF7, ZNF136, STX10
rs12609846	C	T	rs3760982	.	E2F1	KCCN4/ LYPD5	.	AC115522.3, KCCN4, LYPD5, RP11-15A1.2, RP11-15A1.3, ZNF225, ZNF404	KCCN4	LYPD3, OPA3, DMPK, PVRL2, ZNF404, XRCC1, APOC1P1, ZNF155
rs11669175	A	G	rs3760982	GATA3	.	KCCN4/ LYPD5	ZNF404	AC115522.3, KCCN4, LYPD5, RP11-15A1.2, RP11-15A1.3, ZNF155, ZNF404	.	.
rs11665924	G	A	rs3760982	ESR1	.	KCCN4/ LYPD5	ZNF404	AC115522.3, KCCN4, LYPD5, RP11-15A1.2, RP11-15A1.3, ZNF155, ZNF404	.	.
rs73167096	T	C	rs6001930	.	FOXA1	MKL1	.	RP5-1042K10.10, MKL1, CTA- 223H9.9	.	.
rs72722756	T	C	rs11780156	FOXA1:AR	.	MYC	.	.	.	.

Table 5.2: Putative target genes of each candidate causal SNP.

Three different methods were used to predict target genes: GTEx eQTLs, correlation between ATAC-seq peaks and gene expression from Corces et al. (2018), and the INQUISIT scoring method from Michailidou et al. (2017).

5.3. Discussion

This chapter aimed to prioritise candidate causal SNPs at GWAS loci. It aimed to address the lack of GWAS follow-up studies and the lack of functional validation. I developed a pipeline for prioritisation of variants and addressed the potential mechanisms of action of some promising SNPs.

Towards further exploring the functional mechanisms of these nine SNPs, my next step is to download genotype information for cell lines and check for the presence of the lead SNP in the cancer cell line encyclopedia (CCLE) project (Barretina et al., 2012). If the SNPs are present in any cell line, I will then attempt to obtain ChIP-seq data for the implicated TF and assess if there is any allele-specific bias in the data. After that, an allele-specific ChIP-qPCR for the TF will be performed to confirm the bioinformatic observations, and a 4C assay carried out to confirm the target gene. Finally, the TFBS will be deleted using CRISPR, and any functional effects analysed.

Assessing whether a SNP disrupts a CTCF site has been used previously to functionally investigate GWAS SNPs. One recent paper identified prostate cancer GWAS loci involved in chromatin looping, and deletion of these CTCF sites resulted in disruption of gene expression within those loops (Guo et al., 2018). My study only identified two SNPs in CTCF loop anchors in breast cancer, in comparison to Guo et al. (2018) who, using a similar methodology, identified 85 such SNPs in prostate cancer. Although our study didn't identify many SNPs involved in looping, this paper suggests that the hypothesis is reasonable and that further exploration of the two identified variants may lead to important results.

Interestingly, a number of very recent papers have begun to question the function of CTCF sites, and whether disruption of a single CTCF site could change the local 3D architecture of the chromatin sufficiently to be pathogenic. CTCF sites are relatively

unchanging across cell types (Heintzman et al., 2009). A recent pre-print showed that deletion of CTCF sites at a TAD boundary did not have an impact on overall TAD structure, and suggested that there is high redundancy in CTCF sites which form TADs (Despang et al., 2019). Therefore, deletion of just one CTCF site may not cause significant deregulation of gene expression, as there may be other CTCF sites in close proximity which can adopt the function of the deleted site.

In this chapter, my work on identifying potential target genes did not provide conclusive results. The results lacked any breast cancer-related genes (Table 5.2), although it is important to note that these GWAS SNPs may not be acting through known breast cancer-associated pathways and so identification of a non-breast cancer gene could still be interesting. The lack of concordance between the methods as to the predicted target gene is not necessarily surprising: they all use different methods to determine the target gene. The first method was eQTLs from the GTEx project, which correlated genotype and tissue-specific gene expression levels to determine target genes. The second method correlated chromosome accessibility in breast tumours, as determined by ATAC-seq, and gene expression to predict which open regions might regulate which genes (Corces et al., 2018). None of the other methods incorporated ATAC-seq in their analyses. The third method incorporated numerous genomic datasets such as ChIA-PET, HiC, TF binding, RNA expression, TAD boundaries, and eQTLs into a scoring system (Michailidou et al., 2017). A better way to determine the target gene would have been through the use of high-throughput CRISPR perturbations, but in the absence of such data, these 3 methods offered an alternative.

Perhaps a better way of dealing with common variants in disease is not to attempt to investigate the effects of each individual SNP, but rather to use the combined effects of each SNP to assess risk. Recently, there has been much progress in developing polygenic risk scores (PRS). Although the effect size of each individual SNPs is small, PRSs use these GWAS-derived effect sizes to summarise the combined effects of

multiple SNPs on disease risk (Wray et al., 2007). PRSs represent an important progression in understanding the contribution of common variants to disease. PRSs have become more powerful as GWAS sample sizes have increased, and are set to become an important part of clinical risk management. Focusing first on those variants which have the greatest effect size, and thus greatest contribution to the PRS may be an alternative way to prioritise variants for functional interpretation. Perhaps only then, once a shortlist of potentially functional variants of the highest effect sizes have been identified, should biochemical means such as the ones used in this chapter be employed. For example, the most recent study of PRSs in breast cancer, incorporated just 313 SNPs, and found that women who were in the top centile of the PRS had a lifetime risk of breast cancer of 32.6%, while women in the top 1% of the distribution had a 4-fold larger predicted risk than the middle quintile (Mavaddat et al., 2019). Because the PRS attempts to explain the relative risk without knowing the underlying mechanisms, an interesting future approach might be to annotate these 313 SNPs with histone modifications, transcription factor binding sites, and conservation scores. As an aside, none of the 14 SNPs highlighted in this chapter were contained in the authors list of 313 SNPs.

In summary, this section described a pipeline for prioritising SNPs identified through GWAS for functional follow-up. In the case of breast cancer, from a starting list of >6,000 SNPs, I have highlighted nine potentially causal variants, possible target genes and possible TFs through which they mediate their effects. These nine variants are promising candidates for further functional studies, and could be investigated in a manner similar to that described in chapter 4 and chapter 5.

6. General Discussion

6.1. Discussion

6.1.1. The function of non-coding variants in breast cancer

The goal of the work presented in this thesis was to investigate the missing heritability of familial breast cancer, by identifying SNVs and small indels in non-coding regions of the genome that are potentially pathogenic in familial breast cancer. To my knowledge, this is the first study of non-coding variants in an Irish breast cancer cohort. We found that case-enriched non-coding variants were capable of disrupting transcription in a luciferase reporter assay and that they are predicted to disrupt transcription factor binding sites. Furthermore, we show that deletion of an enhancer region containing a rare case-enriched SNV, which is in proximity to *PIK3CA*, appears to result in a decrease in *ZMAT3* expression and an increase in *ZNF639* expression. The lack of gene expression changes in three of the CRISPR experiments demonstrates the complexity of assigning function to non-coding regions, and the urgent necessity of an accurate method of assigning enhancers to their target genes.

Significant work remains to uncover the mechanisms of action of these variants but we have demonstrated both proof of principle, that our approach is capable of finding functional non-coding variants, and that there are potentially pathogenic variants located near breast cancer genes in our cohort of patients.

Overall, our results demonstrate that interrogation of variants in the non-protein coding region in proximity to cancer-associated genes is a worthwhile endeavour. In summary, we hope that this project will make an important contribution to both our knowledge about the genetic causes of breast cancer, as well as adding to the discussion surrounding the functional validation of non-coding variants.

6.1.2. Technological and other limitations of the current study

There are a number of limitations to the work proposed here. Firstly, 104 cases and 101 controls are relatively small numbers for a sequencing study. There is always the possibility that a mutation classed as rare on the basis of this study would not be rare had more individuals been sequenced. One future aim is to perform genotyping for these variants in an independent cohort. However, this has proved difficult to organise as it would require us to obtain germline data which is subject to strict privacy regulations. It has also proven difficult to obtain an adequate dataset as most cancer sequencing studies are constrained to the coding genome, and a lot are focused on somatic variants as opposed to germline variants. Determining if these variants are unique to our Irish patients, or are found in other populations will be another interesting issue. We are currently in the process of attempting to set up a collaboration to clarify this point.

Throughout this project, I made use of publicly available data sets from the ENCODE consortium and others. While this has been an excellent resource, ENCODE is a limited dataset, and thus I was restricted in the scope of analyses to histone modifications, transcriptions factors and cell lines already chosen and analysed by ENCODE - these represent only a fraction of the total number of human TFs.

One reason why we might see no change in gene expression after deletion of our regulatory regions is that many enhancers are redundant, or that within an enhancer, there is redundancy within the TFBS affected. Additionally, the effects of deleting the regulatory region, or the effects of the mutation themselves, might only be seen under specific circumstances or stressors.

At the beginning of this project, making precise and specific edits to a single DNA nucleotide was technologically challenging. It was for this reason that we decided to make small deletions surrounding our potential regulatory regions of interest. Until recently, the

only way to introduce single nucleotide changes with CRISPR was through the use of CRISPR/Cas9 with a repair template containing the desired substitution. A major problem with this approach, however, is that it relies on the cell using its extremely low-efficiency homology-directed repair pathway, which is the one required to produce specific edits such as single nucleotide changes. The majority of the double-strand breaks will be repaired by the error-prone, indel-inducing, non-homologous end-joining repair pathway, and only a small fraction by the high-fidelity homology-directed repair pathway.

Recently, there have been significant improvements in CRISPR technology, with two methods of particular interest for modelling and correcting the variants described in this project. Firstly, the development of base-editors, which make use of Cas9 fused to a cytidine deaminase enzyme to convert cytidine to uridine allowing for a C>T or G>A substitution (Komor et al., 2016), or Cas9 fused to an adenosine deaminase allowing for A>G and T>C substitutions (Gaudelli et al., 2017). However, base editing can only perform these 4 specific base changes and cannot perform any of the other 8 possible substitutions such as G>C, T>A, or C>A.

A more recent development is that of the CRISPR prime editing technique, a more precise and efficient editing technique (Anzalone et al., 2019). This new method includes a modified gRNA, which contains both the sequence to be edited and the sequence to be inserted, a modified Cas9 which cuts at only one strand, and a reverse transcriptase enzyme which transcribes the DNA to be placed into the cut site. The inventors of this approach demonstrate its promise by, among others, correcting the point mutation which causes sickle cell anaemia in HEK293T cells (Anzalone et al., 2019). These increases in the specificity of CRISPR editing are of critical importance, considering that point mutations account for ~50% of known human pathogenic genetic variants (Anzalone et al., 2019).

6.1.3. Future work

We anticipate this study to be a starting point for yet more sophisticated modelling of the effect of the SNV. As described in Section 4.3, the CRISPR results presented in this thesis are from pooled populations of cells, where some of the cells contain the desired deletion, and some do not. There are limitations to this approach however, and I have been extracting a clonal population of cells by growing them at an extremely low density so that the cells are grown in proximity to each other, but a single colony of homogenous cells can still be isolated. This labour-intensive method has allowed me to isolate a clonal population where all the cells contain the deletion. We are currently in the process of repeating the RT-qPCRs on these populations of cells. This will allow us to obtain more robust results.

As also mentioned in Section 4.3, our next aim is to choose the single most promising variant and attempt to further characterise its mechanism of action. We will introduce just the single nucleotide change into the cell line (as opposed to generating cells carrying a ~200 bp deletion) and extract a clonal population of cells. After we have established these isogenic cell lines, we will perform functional assays to further characterise the mechanism of action. Such assays would include *in vitro* 3T3 growth assays to monitor alterations in the rate of cellular proliferation between wild-type and mutant cell lines, which would be expected to increase after the introduction of a potential oncogenic mutation, as well as monitoring the activation status of specific downstream components of relevant signalling pathways. This work also opens the possibility of treating these cell lines with small molecule inhibitors, or various targeted 'biological' drugs, for example in cases where a regulatory mutation results in overexpression of an oncogene.

Another issue we wish to address is whether these variants are unique to the Irish population, or if they are present in other populations. We are currently in the process of setting up a collaboration to genotype our variants in an independent cohort.

Finally, as mentioned in Section 4.3, we are also currently working on using CRISPRa and CRISPRi as complementary methods to determine if our SNVs are located in functional regulatory regions and what their target genes are.

6.1.4. Challenges remaining in the field of non-coding regulatory genomics

As the cost of next-generation sequencing has decreased in recent years, it has become possible to sequence larger cohort sizes, and sequence regions outside of the exomes. This has increased the numbers of variants identified, but the functional interpretation of these variants has lagged behind. Challenging areas include problems such as determining the function of non-coding variants as well as, for example, separating driver from passenger mutations in tumour sequencing data. If we are ever to functionally assess all possible functional non-coding variants, high-throughput methods will be key. Such methods include massively parallel reporter assays (Inoue & Ahituv, 2015) and genome-wide CRISPR/Cas9 screens to interrogate cis-regulatory elements (Gasperini et al., 2017).

Although there have been many efforts to replicate the success of the TERT promoter mutations (Horn et al., 2013; Huang et al., 2013), it has proven difficult. Many papers in the field of non-coding variants have turned up very little. For example, one recent paper made over 4,000 overlapping 1 kb and 2 kb deletions across the exonic and non-coding regions of *HRPT1*, a gene involved in Lesch-Nyhan syndrome, and did not find any associated non-coding region apart from the promoter (Gasperini et al., 2017). This study had a rigorous and in-depth experimental design yet still found no regulatory regions of *HRPT1*. However, this was just one gene, and we can't generalise these results to every gene.

An alternative method of prioritising variants is to use some of the many *in-silico* bioinformatic tools, such as Funseq2 (Fu et al., 2014) and CADD (Rentzsch et al., 2019), which claim to predict likely functional non-coding variants. However, after using many such tools throughout my project, I found there were discrepancies in their results (data not shown). Other studies have also shown discordant results, and suggest that the use of multiple tools in parallel might be a better approach (Nishizaki & Boyle, 2017). There is a lack of a 'gold standard' set of pathogenic non-coding SNVs, which would allow us to benchmark such bioinformatic methods. Having a standard accurate method of prioritising non-coding variants would revolutionise the field. This is an area that needs critical methodological refinement.

Cis-regulatory variation has been reported in one study to modify the penetrance of coding variants (Castel et al., 2018). The authors of this particular study took a somewhat unique approach in considering both rare and common variants together, which are often studied separately, but of course, exist *in vivo* alongside each other. The authors showed that common eQTL regulatory SNPs predicted to increase the expression of the pathogenic allele were enriched in a cohort of cancer patients. In future, looking at non-coding variants along with any coding variants in corresponding genes may prove important.

6.1.5. Non-coding regulatory regions as targets for therapy

There has been rapid progress over the last few years, and one interesting example of the opportunities exploring regulatory regions offers is the case of β -thalassemia and sickle cell anaemia which result in severe anaemia and red blood cell sickling respectively. Both are common inherited blood disorders, caused by mutations in the *HBB* gene, which produces the β chain of adult haemoglobin. Treatment usually involves an allogeneic hematopoietic stem cell transplantation (Ikawa et al., 2019). Gene therapy is a promising

new option, but interestingly, one approach is not to target the *HBB* gene itself, but rather an enhancer of *BCL11A*, a repressor of fetal haemoglobin.

This new approach is based around the concept that induction of endogenous fetal haemoglobin, which is not normally produced after birth, counteracts the effects of the defective adult haemoglobin and alleviates the symptoms. The method involves targeting an enhancer of *BCL11A*, which mediates the switch from fetal to adult haemoglobin. CRISPR/Cas9 cleavage at a GATA1 binding site within the erythroid-specific enhancer of *BCL11A* results in a reduction of BCL11A and an increase in fetal haemoglobin (Wu et al., 2019). In β -thalassemia, the fetal haemoglobin compensates for the deficiencies in β -globin, while in sickle cell anaemia the fetal γ -globin chains are incorporated into haemoglobin and they do not sickle (Ikawa et al., 2019).

Various pharmaceutical companies have taken notice of this line of research. CRISPR Therapeutics and Vertex Pharmaceuticals' have adopted a CRISPR/Cas9 approach to target the enhancer, and phase 1/2 clinical trials are underway for sickle cell disease (Vertex Pharmaceuticals Incorporated, 2018a) and β -thalassemia (Vertex Pharmaceuticals Incorporated, 2018b). In contrast, Sangamo Therapeutics is using zinc finger nucleases to disrupt the BCL11A enhancer (Sangamo Therapeutics, 2018). With the CRISPR strategy, the aim is to carry out *ex-vivo* editing of patients own hematopoietic stem cells. The first β -thalassemia patient has already been treated with the CRISPR approach, and as of July 2019 remained transfusion-independent four months after treatment (CRISPR Therapeutics, 2019).

Another interesting example comes from the use of CRISPRa to activate regulatory regions. Loss of function mutations in one copy of a gene, termed haploinsufficiency, decreases expression, and in the case of *SIM1* and *MC4R*, results in obesity. Activation of the promoters of the remaining functional copy of these genes in mice with CRISPRa rescued the obesity phenotype of *Sim1* and *Mc4r* heterozygous mice. In the case of *Sim1*,

a 270 kb distal hypothalamic specific enhancer was also targeted, allowing for upregulation of *Sim1* specifically in the hypothalamus, also sufficient to rescue the obesity phenotype (Matharu et al., 2019).

This CRISPRa approach offers a potential alternative to conventional gene therapy strategies and the associated issues with ectopic gene expression. This concept also opens up a CRISPRi approach, where oncogenes with known enhancers seem to be a logical next step for such an approach. For example, this approach could potentially target the *MYC* oncogene which has been shown to be dysregulated in cancer. It could also potentially be used to specifically target a *de novo* enhancer in the disease state, thus tailoring the therapy to only the diseased tissue.

These two examples of β -thalassemia and obesity demonstrate the latent potential of non-coding regions to be exploited for therapeutic gain and offer exciting paradigms for how it might be done. Critically, the fact that many regulatory elements are tissue-specific, allows for the therapeutic effect to be obtained specifically in the diseased tissue. This offers an exceptionally precise way to control gene expression and is an even more attractive option than conventional gene-targeted gene editing or drug treatment, whose effects may be present in healthy tissues. To relate these concepts to my work, where a SNVs results in activation of a regulatory region resulting in upregulation of an oncogene, the enhancer or promoter could potentially be targeted with CRISPRi. Similarly, for SNVs resulting in a decrease in expression of a tumour suppressor, gene expression could be increased by targeting with CRISPRa.

6.2. Conclusions

The data presented in this thesis support the following conclusions:

- Single nucleotide mutations in the promoters and candidate enhancers of cancer-associated genes are capable of affecting transcription in reporter assays.
- rs78248633 is a rare variant found in Irish BRCAx cases. It is located in the distal promoter region of *PIK3CA* and is capable of increasing transcription in a reporter assay, possibly through the creation of a binding site for STAT1 and/or STAT3 transcriptional activators.
- Deletion of the region surrounding rs78248633 causes decreased expression of *ZMAT3*, which is involved in cell death and the cell cycle, and increased expression of *ZNF639*.
- Integrative functional annotation of GWAS loci with publicly available data is a useful method to identify potential causal SNPs, and here I identified nine SNPs from 6,270 as good candidates for functional follow-up studies.
- Single nucleotide mutations in regulatory regions represent a promising mechanism for influencing disease risk but the lack of knowledge about the function of the non-coding genome still represents a large barrier in understanding their exact mechanisms of action.

7. References

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7, 248–249.
- Aloraifi, F. (2014). *The search for the missing heritability in familial breast cancer: Identification of rare variants in BRCAx families through targeted capture and next-generation sequencing*. (Doctoral thesis). Trinity College Dublin, Ireland.
- Aloraifi, F., McDevitt, T., Martiniano, R., McGreevy, J., McLaughlin, R., Egan, C. M., Cody, N., Meany, M., Kenny, E., Green, A. J., et al. (2015). Detection of novel germline mutations for breast cancer in non-BRCA1/2 families. *FEBS Journal*, 282, 3424–3437.
- Amir, E., Freedman, O. C., Seruga, B., & Evans, D. G. (2010). Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models. *JNCI/ Journal of the National Cancer Institute*, 102, 680–691.
- Anglian Breast Cancer Study Group. (2000). Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *British Journal of Cancer*, 83, 1301–1308.
- Antoniou, A. C., & Easton, D. F. (2006). Models of genetic susceptibility to breast cancer. *Oncogene*, 25, 5898–5905.
- Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, 576, 149–157.
- Ariazi, E. A., & Jordan, V. C. (2006). Estrogen-related receptors as emerging targets in cancer and metabolic disorders. *Current Topics in Medicinal Chemistry*, 6, 203–215.

- Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β -globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27, 299–308.
- Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483, 603–607.
- Bersani, C., Xu, L.-D., Vilborg, A., Lui, W.-O., & Wiman, K. G. (2014). Wig-1 regulates cell cycle arrest and cell death through the p53 targets FAS and 14-3-3 σ . *Oncogene*, 33, 4407–4417.
- Bond, G. L., Hu, W., Bond, E. E., Robins, H., Lutzker, S. G., Arva, N. C., Bargonetti, J., Bartel, F., Taubert, H., Wuerl, P., et al. (2004). A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell*, 119, 591–602.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68, 394–424.
- Bulger, M., & Groudine, M. (1999). Looping versus linking: toward a model for long-distance gene activation. *Genes & Development*, 13, 2465–2477.
- Cai, J., Tian, A.-X., Wang, Q.-S., Kong, P.-Z., Du, X., Li, X.-Q., & Feng, Y.-M. (2015). FOXF2 suppresses the FOXC2-mediated epithelial–mesenchymal transition and multidrug resistance of basal-like breast cancer. *Cancer Letters*, 367, 129–137.
- Castel, S. E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., & Lappalainen, T. (2018). Modified

penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nature Genetics*, 50, 1327–1334.

Chiang, W.-F., Hung, P.-S., Liu, S.-Y., Yuan, T.-C., Chang, K.-W., Chen, Y.-P., Liu, Y.-C., & Lin, S.-C. (2011). Increase of ZASC1 gene copy number in recurrent oral carcinoma. *Oral Diseases*, 17, 53–59.

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V., & Cisse, I. I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361, 412–415.

Coetzee, S. G., Coetzee, G. A., & Hazelett, D. J. (2015). motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics*, 31, 3847–3849.

Collaborative Group on Hormonal Factors in Breast Cancer. (2001). Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet*, 358, 1389–1399.

Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, S. W., et al. (2018). The chromatin accessibility landscape of primary human cancers. *Science*, 362, 6413.

Cowper-Salari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoutte, J., Moore, J. H., & Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics*, 44, 1191–1198.

CRISPR Therapeutics. (2019). CRISPR Therapeutics Provides Business Update and Reports Second Quarter 2019 Financial Results.

<http://ir.crisprtx.com/news-releases/news-release-details/crispr-therapeutics->

provides-business-update-and-reports-1.

- Despang, A., Schöpflin, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.-L., Timmermann, B., Wittler, L., Vingron, M., et al. (2019). Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nature Genetics*, 51, 1263–1271.
- Di Croce, L., & Helin, K. (2013). Transcriptional regulation by Polycomb group proteins. *Nature Structural & Molecular Biology*, 20, 1147–1155.
- Dorsett, D. (1999). Distant liaisons: long-range enhancer–promoter interactions in *Drosophila*. *Current Opinion in Genetics & Development*, 9, 505–514.
- Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struwing, J. P., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447, 1087–1093.
- Elgin, S. C. R. (1984). Molecular biology: Anatomy of hypersensitive sites. *Nature*, 309, 213–214.
- Evans, D. G., Shenton, A., Woodward, E., Laloo, F., Howell, A., & Maher, E. R. (2008). Penetrance estimates for BRCA1 and BRCA2 based on genetic testing in a Clinical Cancer Genetics service setting: risks of breast/ovarian cancer quoted should reflect the cancer burden in the family. *BMC Cancer*, 8, 155.
- Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research*, 39, D945-50.
- Freedman, M. L., Monteiro, A. N. A., Gayther, S. A., Coetzee, G. A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D., et al. (2011).

Principles for the post-GWAS functional characterization of cancer risk loci.

Nature Genetics, 43, 513–518.

Fritz, A. J., Ghule, P. N., Boyd, J. R., Tye, C. E., Page, N. A., Hong, D., Shirley, D. J., Weinheimer, A. S., Barutcu, A. R., Gerrard, D. L., et al. (2018). Intranuclear and higher-order chromatin organization of the major histone gene cluster in breast cancer. *Journal of Cellular Physiology*, 233, 1278–1290.

Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X. J., Yip, K. Y., Khurana, E., & Gerstein, M. (2014). FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*, 15, 480.

Gasperini, M., Findlay, G. M., McKenna, A., Milbank, J. H., Lee, C., Zhang, M. D., Cusanovich, D. A., & Shendure, J. (2017). CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *The American Journal of Human Genetics*, 101, 192–205.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., & Liu, D. R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*, 551, 464–471.

Gilbert, L. A., Horlbeck, M. A., Adamson, B., Villalta, J. E., Chen, Y., Whitehead, E. H., Guimaraes, C., Panning, B., Ploegh, H. L., Bassik, M. C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, 159, 647–661.

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., & Elhaik, E. (2013). On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*, 5, 578–590.

Gross, D. S., & Garrard, W. T. (1988). Nuclease Hypersensitive Sites in

- Chromatin. *Annual Review of Biochemistry*, 57, 159–197.
- Guenther, C. A., Tasic, B., Luo, L., Bedell, M. A., & Kingsley, D. M. (2014). A molecular basis for classic blond hair color in Europeans. *Nature Genetics*, 46, 748–752.
- Guo, Y., Perez, A. A., Hazelett, D. J., Coetzee, G. A., Rhie, S. K., & Farnham, P. J. (2018). CRISPR-mediated deletion of prostate cancer risk-associated CTCF loop anchors identifies repressive chromatin loops. *Genome Biology*, 19, 160.
- Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19, 621–637.
- Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., & King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250, 1684–1689.
- Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459, 108–112.
- Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science*, 339, 959–961.
- Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V, Chin, L., & Garraway, L. a. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science*, 339, 957–959.
- Ikawa, Y., Miccio, A., Magrin, E., Kwiatkowski, J.L. Rivella, S., Cavazzana, M. (2019). Gene therapy of hemoglobinopathies: progress and future challenges.

Human Molecular Genetics, 28, R24-30.

- Imoto, I., Yuki, Y., Sonoda, I., Ito, T., Shimada, Y., Imamura, M., & Inazawa, J. (2003). Identification of ZASC1 encoding a Krüppel-like zinc finger protein as a novel target for 3q26 amplification in esophageal squamous cell carcinomas. *Cancer Research*, 63, 5691–5696.
- Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, 106, 159–164.
- International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.
- International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431, 931–945.
- Ji, Z., He, L., Rotem, A., Janzer, A., Cheng, C. S., Regev, A., & Struhl, K. (2018). Genome-scale identification of transcription factors that mediate an inflammatory network during breast cellular transformation. *Nature Communications*, 9, 2068.
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, doi:10.1101/531210.
- Kheradpour, P., & Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Research*, 42, 2976–2987.
- Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*,

465, 182–187.

Kioussis, D., Vanin, E., DeLange, T., Flavell, R. A., & Grosveld, F. G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia.

Nature, 306, 662–666.

Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016).

Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, 533, 420–424.

Lander, E. S. (1996). The new genomics: global views of biology. *Science*, 274, 536–539.

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., & Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505, 495–501.

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *American Journal of Human Genetics*, 95, 5–23.

Lee, W., Haslinger, A., Karin, M., & Tjian, R. (1987). Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40. *Nature*, 325, 368–372.

Lilyquist, J., Ruddy, K. J., Vachon, C. M., & Couch, F. J. (2018). Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. *Cancer Epidemiology, Biomarkers & Prevention*, 27, 380–394.

Liu, Y., Walavalkar, N. M., Dozmorov, M. G., Rich, S. S., Civelek, M., & Guertin, M. J. (2017). Identification of breast cancer associated variants that modulate transcription factor binding. *PLOS Genetics*, 13, e1006761.

Lo, P.-K., Lee, J. S., Liang, X., & Sukumar, S. (2016). The dual role of FOXF2 in

regulation of DNA replication and the epithelial-mesenchymal transition in breast cancer progression. *Cellular Signalling*, 28, 1502–1519.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461, 747–753.

Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., Etchin, J., Lawton, L., Sallan, S. E., Silverman, L. B., et al. (2014). An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*, 346, 1373–1377.

Matharu, N., Rattanasopha, S., Tamura, S., Maliskova, L., Wang, Y., Bernard, A., Hardin, A., Eckalbar, W. L., Vaisse, C., & Ahituv, N. (2019). CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. *Science*, 363, 6424.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337, 1190–1195.

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., Tyrer, J. P., Chen, T.-H., Wang, Q., Bolla, M. K., et al. (2019). Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics*, 104, 21–34.

McPherson, K., Steel, C. M., & Dixon, J. M. (2000). ABC of breast diseases. Breast cancer-epidemiology, risk factors, and genetics. *BMJ*, 321, 624–628.

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al. (2017). Association analysis

- identifies 65 new breast cancer risk loci. *Nature*, 551, 92–94.
- Morgenthaler, S., & Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615, 28–56.
- National Cancer Registry. (2017). *Cancer in Ireland 1994-2015 with estimates for 2015-2017: Annual Report of the National Cancer Registry*. Cork, Ireland.
- Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31, 3812–3814.
- Nishizaki, S. S., & Boyle, A. P. (2017). Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. *Trends in Genetics*, 33, 34–45.
- Oldridge, D. A., Wood, A. C., Weichert-Leahey, N., Crimmins, I., Sussman, R., Winter, C., McDaniel, L. D., Diamond, M., Hart, L. S., Zhu, S., et al. (2015). Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature*, 528, 418–421.
- Ongenaert, M., Van Neste, L., De Meyer, T., Menschaert, G., Bekaert, S., & Van Criekinge, W. (2008). PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Research*, 36, D842-6.
- Palstra, R.-J., & Grosveld, F. (2012). Transcription factor binding at enhancers: shaping a genomic regulatory landscape in flux. *Frontiers In Genetics*, 3, 195.
- Peng, S., Lü, B., Ruan, W., Zhu, Y., Sheng, H., & Lai, M. (2011). Genetic polymorphisms and breast cancer risk: evidence from meta-analyses, pooled analyses, and genome-wide association studies. *Breast Cancer Research and Treatment*, 127, 309–324.
- Peto, J., Collins, N., Barfoot, R., Seal, S., Warren, W., Rahman, N., Easton, D. F., Evans, C., Deacon, J., & Stratton, M. R. (1999). Prevalence of BRCA1 and

BRCA2 Gene Mutations in Patients With Early-Onset Breast Cancer. *JNCI/ Journal of the National Cancer Institute*, 91, 943–949.

Petryszak, R., Keays, M., Tang, Y. A., Fonseca, N. A., Barrera, E., Burdett, T.,

Füllgrabe, A., Fuentes, A. M.-P., Jupp, S., Koskinen, S., et al. (2016).

Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*, 44, D746–D752.

Pinte, S., Stankovic-Valentin, N., Deltour, S., Rood, B. R., Guérardel, C., &

Leprince, D. (2004). The tumor suppressor gene HIC1 (hypermethylated in cancer 1) is a sequence-specific transcriptional repressor: definition of its consensus binding sequence and analysis of its DNA binding and repressive properties. *The Journal of Biological Chemistry*, 279, 38313–38324.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P.,

& Lim, W. A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, 152, 1173–1183.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D.,

Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159, 1665–1680.

Rayoo, M., Yan, M., Takano, E. A., Bates, G. J., Brown, P. J., Banham, A. H., &

Fox, S. B. (2009). Expression of the forkhead box transcription factor FOXP1 is associated with oestrogen receptor alpha, oestrogen receptor beta and improved survival in familial breast cancers. *Journal of Clinical Pathology*, 62, 896–902.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019).

CADD: predicting the deleteriousness of variants throughout the human

- genome. *Nucleic Acids Research*, 47, D886–D894.
- Sangamo Therapeutics. (2018). A Study to Assess the Safety, Tolerability, and Efficacy of ST-400 for Treatment of Transfusion-Dependent Beta-thalassemia (TDT). <https://clinicaltrials.gov/ct2/show/NCT03432364>.
- Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489, 109–113.
- Saunders, A. M., Strittmatter, W. J., Schmechel, D., George-Hyslop, P. H., Pericak-Vance, M. A., Joo, S. H., Rosi, B. L., Gusella, J. F., Crapper-MacLachlan, D. R., & Alberts, M. J. (1993). Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, 43, 1467–1472.
- Shiovitz, S., & Korde, L. A. (2015). Genetics of breast cancer: a topic in evolution. *Annals of Oncology*, 26, 1291–1299.
- Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, 15, 272–286.
- Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Current Opinion in Genetics & Development*, 6, 743–748.
- Smit, A., Hubley, R., & Green, P. (2010). <http://www.repeatmasker.org>.
- Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18, 696–705.
- Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13, 613–626.
- Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K., Martin, N. G., & Montgomery, G. W. (2008). A Single SNP in an Evolutionary

Conserved Region within Intron 86 of the HERC2 Gene Determines Human Blue-Brown Eye Color. *The American Journal of Human Genetics*, 82, 424–431.

Sur, I. K., Hallikas, O., Vaharautio, A., Yan, J., Turunen, M., Enge, M., Taipale, M., Karhu, A., Aaltonen, L. A., & Taipale, J. (2012). Mice Lacking a Myc Enhancer That Includes Human SNP rs6983267 Are Resistant to Intestinal Tumors. *Science*, 338, 1360–1363.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526, 68.

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

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13, 36–46.

Udler, M. S., Tyrer, J., & Easton, D. F. (2010). Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genetic Epidemiology*, 34, 463–468.

Vertex Pharmaceuticals Incorporated. (2018a). A Safety and Efficacy Study Evaluating CTX001 in Subjects With Severe Sickle Cell Disease. <https://clinicaltrials.gov/ct2/show/NCT03745287>.

Vertex Pharmaceuticals Incorporated. (2018b). A Safety and Efficacy Study Evaluating CTX001 in Subjects With Transfusion-Dependent β -Thalassemia. <https://clinicaltrials.gov/ct2/show/NCT03655678>.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457, 854–858.

- Visscher, P. M. (2008). Sizing up human height variation. *Nature Genetics*, *40*, 489.
- Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z., Group, Topm. A. W., Consortium, T.-O. for P. M., Cupples, L. A., Shadyab, A. H., McKnight, B., Shoemaker, B. M., et al. (2019). Recovery of trait heritability from whole genome sequence data. *BioRxiv*, doi:10.1101/588020.
- Walsh, T., Lee, M. K., Casadei, S., Thornton, A. M., Stray, S. M., Pennil, C., Nord, A. S., Mandell, J. B., Swisher, E. M., & King, M.-C. (2010). Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *PNAS*, *107*, 12629–12633.
- Wang, Q.-S., Kong, P.-Z., Li, X.-Q., Yang, F., & Feng, Y.-M. (2015). FOXF2 deficiency promotes epithelial-mesenchymal transition and metastasis of basal-like breast cancer. *Breast Cancer Research*, *17*, 30.
- Weedon, M. N., Cebola, I., Patch, A.-M., Flanagan, S. E., De Franco, E., Caswell, R., Rodríguez-Seguí, S. A., Shaw-Smith, C., Cho, C. H.-H., Allen, H. L., et al. (2014). Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nature Genetics*, *46*, 61–64.
- Wendt, C., & Margolin, S. (2019). Identifying breast cancer susceptibility genes – a review of the genetic background in familial breast cancer. *Acta Oncologica*, *58*, 135–146.
- Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., & Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*, 307–319.
- Wooster, R., Neuhausen, S. L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., & Averill, D. (1994). Localization of a breast

cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*, 265, 2088–2090.

Wray, N. R., Goddard, M. E., & Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Research*, 17, 1520–1528.

Xiao, J., He, B., Zou, Y., Chen, X., Lu, X., Xie, M., Li, W., He, S., You, S., & Chen, Q. (2016). Prognostic value of decreased FOXP1 protein expression in various tumors: a systematic review and meta-analysis. *Scientific Reports*, 6, 30437.

Yu, B.-H., Li, B.-Z., Zhou, X.-Y., Shi, D.-R., & Yang, W.-T. (2018). Cytoplasmic FOXP1 expression is correlated with ER and calpain II expression and predicts a poor outcome in breast cancer. *Diagnostic Pathology*, 13, 36.

Zacksenhaus, E., Liu, J. C., Jiang, Z., Yao, Y., Xia, L., Shrestha, M., & Ben-David, Y. (2017). Transcription Factors in Breast Cancer—Lessons From Recent Genomic Analyses and Therapeutic Implications. *Advances in Protein Chemistry and Structural Biology*, 107, 223–273.

Zhang, B., Beeghly-Fadiel, A., Long, J., & Zheng, W. (2011). Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *The Lancet Oncology*, 12, 477–488.

8. Appendix

8.1. 255 non-coding variants identified in BRCAX cases

Chr	Position	rsID	REF	ALT	CASES_TOTAL	CONTROLS_TOTAL
chr11	69306634	rs77797605	G	A	8	1
chr3	178865976	rs576007456	G	A	7	1
chr3	178867338	rs78248633	T	C	7	2
chr5	1296727	rs33977403	C	T	7	2
chr6	127589197	-	T	TAGAGAGAAA	7	2
chr11	1896319	rs117248447	C	T	7	2
chr5	67517103	rs146585339	G	A	6	0
chr19	1172229	rs767671954	C	G	6	0
chr7	27193732	rs118016485	T	C	6	2
chr10	50762378	rs55825410	T	C	6	2
chr16	82688068	rs150085830	T	G	6	2
chr16	82692647	rs117805694	A	G	6	2
chr3	24548891	rs141542764	G	T	5	0
chr3	30388881	rs147325121	C	T	5	0
chr3	178823411	rs141340086	A	G	5	0
chr7	116082888	-	G	A	5	0
chr9	22159250	-	T	G	5	0
chr11	94411901	rs116906793	C	T	5	0
chr5	56110339	rs80096098	C	A	5	1
chr10	8108022	rs200042710	G	GT	5	1
chr11	2002498	rs113602176	C	T	5	1
chr14	64804831	rs142119488	T	C	5	1
chr17	41445483	rs559003818	C	G	5	1
chr17	56709586	rs117013833	G	A	5	1
chr3	41158906	rs62260281	C	G	5	2
chr5	1328747	rs116555409	C	T	5	2
chr5	55911754	rs148631788	G	A	5	2
chr5	56112594	rs145021294	G	A	5	2
chr5	67600218	rs148153519	T	C	5	2
chr6	1610389	-	G	T	5	2
chr7	27200256	rs150960231	G	C	5	2
chr8	128807120	rs571724121	C	T	5	2
chr9	21975319	rs36228834	T	A	5	2
chr11	2004159	rs141797563	G	A	5	2
chr11	69341636	rs185085218	C	T	5	2
chr11	69450976	rs201771970	G	GA	5	2
chr13	48706391	rs117116927	A	G	5	2
chr14	32694563	-	A	AT	5	2
chr14	64805696	rs186918491	G	A	5	2
chr16	52579900	rs112672919	C	T	5	2
chr17	41277944	rs8176072	A	T	5	2

Chr	Position	rsID	REF	ALT	CASES_TOTAL	CONTROLS_TOTAL
chr3	58037101	rs115789493	C	T	4	0
chr3	61237180	rs533075364	T	C	4	0
chr6	19837556	rs570152569	C	T	4	0
chr7	91571432	rs556758385	A	G	4	0
chr7	116082881	-	A	T	4	0
chr9	139493442	rs184109994	C	T	4	0
chr11	2017762	rs377448027	C	T	4	0
chr11	69302067	rs553538570	G	GA	4	0
chr11	108092424	rs190491586	G	A	4	0
chr12	6324825	rs192478766	G	A	4	0
chr12	25402680	rs574133962	T	C	4	0
chr13	32829909	-	A	AAG	4	0
chr13	32829911	-	ACT	A	4	0
chr16	82667341	rs8053232	T	A	4	0
chr2	217925670	rs17835248	G	A	4	1
chr5	39394660	rs189962433	T	C	4	1
chr5	45284775	rs77436149	G	A	4	1
chr5	112293944	rs79844138	T	A	4	1
chr7	54907530	rs145835677	T	G	4	1
chr8	41179645	rs78858487	G	A	4	1
chr11	69353192	rs182883964	A	G	4	1
chr11	108093077	rs4987879	C	T	4	1
chr11	108094538	rs3092847	G	A	4	1
chr11	115489329	rs553593294	C	G	4	1
chr14	69016002	rs113671762	A	G	4	1
chr15	32951756	rs143319100	C	T	4	1
chr15	32999109	rs80018808	T	C	4	1
chr16	22636190	rs553008968	C	T	4	1
chr16	31213181	rs186112291	C	T	4	1
chr16	52586968	rs189388211	G	A	4	1
chr16	82694959	rs62040567	G	T	4	1
chr17	7651739	rs190219584	T	C	4	1
chr17	52978355	rs774191752	T	C	4	1
chr20	39716104	rs183191678	G	A	4	1
chr1	68155244	rs72682900	C	T	3	0
chr1	206945387	rs151283261	G	A	3	0
chr3	24537012	rs180941378	C	G	3	0
chr3	30535505	rs189344359	C	T	3	0
chr3	30544773	rs114701863	G	A	3	0
chr3	57994597	rs200721532	C	G	3	0
chr3	172167234	rs115459563	G	A	3	0
chr5	56060002	rs146045545	C	T	3	0
chr5	67238911	rs79920499	C	T	3	0
chr5	67519897	rs138117821	G	T	3	0
chr5	67523867	rs537192367	C	T	3	0

Chr	Position	rsID	REF	ALT	CASES_TOTAL	CONTROLS_TOTAL
chr7	27195307	rs552697857	C	T	3	0
chr7	55089139	rs73131819	T	C	3	0
chr7	55133385	rs535822015	TG	T	3	0
chr7	55133915	rs112156795	C	T	3	0
chr7	55167774	rs182274005	A	G	3	0
chr7	55179566	rs192274478	A	G	3	0
chr7	91576325	rs531278019	A	G	3	0
chr8	91264291	rs1235389	T	A	3	0
chr9	21833095	rs148745437	C	T	3	0
chr9	22210607	rs181743757	T	G	3	0
chr10	5902830	rs557836805	TAAC	T	3	0
chr10	8112810	rs372644648	C	T	3	0
chr10	123349634	rs181229508	C	G	3	0
chr10	131266247	rs140393798	T	C	3	0
chr11	519012	rs186960061	G	C	3	0
chr11	1897836	rs80005225	G	A	3	0
chr11	1990805	rs766195279	G	A	3	0
chr11	2024090	rs117240203	C	T	3	0
chr11	2934284	rs76436178	C	T	3	0
chr11	94226524	rs542694969	C	A	3	0
chr12	65515117	rs117397017	G	A	3	0
chr14	32943700	rs754015635	T	C	3	0
chr14	69064804	rs138607586	G	C	3	0
chr16	22652330	rs557157593	C	T	3	0
chr16	82661466	rs574651820	C	T	3	0
chr17	1957912	rs145884851	T	C	3	0
chr17	7593110	rs181313855	T	C	3	0
chr19	17416570	rs142838944	A	G	3	0
chr20	39765429	rs539920793	G	C	3	0
chr22	33196191	rs575217423	CCT	C	3	0
chr22	41487679	rs184741182	G	A	3	0
chr1	14085253	rs544967575	T	C	2	0
chr1	45814707	rs550806032	AC	A	2	0
chr1	68149230	-	GT	G	2	0
chr1	68149233	-	T	A	2	0
chr1	68149239	-	A	AC	2	0
chr1	68152294	rs201162918	C	T	2	0
chr1	121261022	rs587742772	T	C	2	0
chr1	206940810	rs185549928	C	G	2	0
chr2	19258962	rs562086590	T	TG	2	0
chr2	19304233	rs955824958	C	T	2	0
chr2	48010255	rs556432240	G	A	2	0
chr2	113570814	rs140444889	G	A	2	0
chr2	113593225	rs200183218	A	G	2	0
chr2	190075716	rs75118903	G	A	2	0

Chr	Position	rsID	REF	ALT	CASES_TOTAL	CONTROLS_TOTAL
chr3	24550704	-	A	G	2	0
chr3	24640398	rs140830129	C	T	2	0
chr3	24640945	rs186255448	G	A	2	0
chr3	27411006	rs77487712	G	A	2	0
chr3	27492600	rs146215419	A	T	2	0
chr3	37031991	rs187318406	T	C	2	0
chr3	37036112	rs560647889	G	GT	2	0
chr3	41159228	-	G	A	2	0
chr3	41242951	rs1000012468	C	T	2	0
chr3	50371791	rs145333395	C	G	2	0
chr3	59640759	rs150092578	A	G	2	0
chr3	61237015	rs114542366	G	T	2	0
chr3	178865960	rs529801433	G	T	2	0
chr3	178876932	rs185729982	G	A	2	0
chr5	1296728	rs369303269	G	A	2	0
chr5	1315770	rs1030284040	AT	A	2	0
chr5	52385354	rs112722993	A	G	2	0
chr5	52466446	rs766030569	T	C	2	0
chr5	55933349	-	C	T	2	0
chr5	55934859	rs142847580	C	T	2	0
chr5	56114886	rs770208721	TTTAG	T	2	0
chr5	67159075	rs891349929	T	TA	2	0
chr5	67511340	rs251396	G	C	2	0
chr5	67512784	rs369388483	ATTC	A	2	0
chr5	67514648	-	G	A	2	0
chr5	67518950	-	CA	C	2	0
chr5	67520306	rs706712	C	G	2	0
chr5	67600733	rs148927952	G	A	2	0
chr5	111811544	rs184367537	T	C	2	0
chr5	111815133	-	C	T	2	0
chr5	112043706	-	C	T	2	0
chr5	146548046	-	C	T	2	0
chr6	1615264	-	C	A	2	0
chr7	6055848	-	A	C	2	0
chr7	6056139	rs149880598	C	T	2	0
chr7	19157053	-	C	G	2	0
chr7	46016847	rs73108561	G	A	2	0
chr7	54965988	rs146489208	C	T	2	0
chr7	55029104	rs551957572	G	A	2	0
chr7	55112171	rs888203607	C	T	2	0
chr7	55130989	rs146353504	G	C	2	0
chr7	55154714	rs113242171	T	C	2	0
chr7	55169255	rs111397519	G	A	2	0
chr7	55195503	rs539048920	A	G	2	0
chr7	55197485	rs918174293	C	A	2	0

Chr	Position	rsID	REF	ALT	CASES_TOTAL	CONTROLS_TOTAL
chr7	55198378	rs113759509	C	T	2	0
chr7	55199959	rs561917236	A	G	2	0
chr7	91686009	rs781548036	A	G	2	0
chr7	116152880	rs537960842	T	C	2	0
chr7	140623095	rs561575495	G	C	2	0
chr8	41199900	rs540193258	TAAG	T	2	0
chr8	128403366	rs529551471	A	T	2	0
chr9	21831628	rs906510540	C	T	2	0
chr9	21987155	rs3731208	A	G	2	0
chr9	22030517	-	C	A	2	0
chr9	22044804	rs13290048	C	T	2	0
chr9	22170708	-	A	G	2	0
chr9	22209701	rs191272696	A	C	2	0
chr9	89998310	rs185022702	G	A	2	0
chr9	139471247	rs750567800	G	A	2	0
chr9	139520129	rs141940507	C	T	2	0
chr10	7914370	rs745616341	A	G	2	0
chr10	7998854	rs745430142	A	T	2	0
chr10	8088882	rs145717085	G	T	2	0
chr10	8094784	rs553237095	G	T	2	0
chr10	8096386	rs907322755	C	G	2	0
chr10	8101921	rs78610022	C	A	2	0
chr10	8112139	rs147357082	C	A	2	0
chr10	8113726	rs191488644	G	A	2	0
chr10	8116650	rs192494656	C	T	2	0
chr10	50743360	rs75448943	G	A	2	0
chr10	131265898	rs773078400	C	T	2	0
chr11	2006266	rs545678735	G	A	2	0
chr11	2008147	rs147601431	A	G	2	0
chr11	69312202	rs183236230	C	T	2	0
chr11	69352126	rs755977252	G	A	2	0
chr11	69353155	rs546015627	C	T	2	0
chr11	69408470	rs186153098	G	A	2	0
chr11	69454136	rs36225069	A	T	2	0
chr11	94439740	rs575658665	G	A	2	0
chr11	115373606	rs973509490	C	A	2	0
chr11	115524945	rs536910855	G	A	2	0
chr12	6310778	rs975436941	T	C	2	0
chr12	6322750	rs190055280	A	C	2	0
chr12	6323140	rs117761448	G	A	2	0
chr12	6330600	rs11568229	G	T	2	0
chr12	6338967	rs61917081	T	C	2	0
chr12	6341502	rs7966772	G	A	2	0
chr12	25649109	rs185874675	T	C	2	0
chr12	65515425	rs182995179	T	C	2	0

Chr	Position	rsID	REF	ALT	CASES_TOTAL	CONTROLS_TOTAL
chr12	65515469	rs188036546	G	A	2	0
chr13	21591482	-	G	T	2	0
chr13	21649277	rs151001777	A	G	2	0
chr13	21653975	rs190358351	C	A	2	0
chr13	32852784	rs74951121	G	A	2	0
chr13	32889180	rs935143708	T	C	2	0
chr14	32942055	rs147350664	C	T	2	0
chr14	32943747	rs188037873	G	C	2	0
chr14	64699316	rs146413257	T	C	2	0
chr14	69003273	rs78231121	G	A	2	0
chr16	18936098	-	T	C	2	0
chr16	18970844	rs184841913	T	A	2	0
chr16	68679247	rs373351410	C	T	2	0
chr16	68680390	rs573336656	C	T	2	0
chr16	68692375	rs375724242	G	A	2	0
chr16	68788518	rs201066116	A	G	2	0
chr16	82663807	rs184867010	A	G	2	0
chr16	82672919	rs138964780	C	T	2	0
chr17	11852214	rs181341817	G	C	2	0
chr17	37852066	rs551511055	AGGCCGCT	A	2	0
chr17	37860840	rs766631446	A	G	2	0
chr17	41393325	rs780031517	C	T	2	0
chr17	41445732	rs560589736	A	G	2	0
chr17	52980892	rs542297830	G	C	2	0
chr18	21301768	rs144129809	A	C	2	0
chr18	21406062	rs550292630	G	C	2	0
chr18	21409330	rs569281838	C	T	2	0
chr18	21454639	rs145260044	C	T	2	0
chr18	28626103	rs757687826	C	T	2	0
chr18	28628804	rs78952600	A	T	2	0
chr19	1170216	rs112654501	C	T	2	0
chr19	1206602	rs149756065	C	T	2	0
chr19	17246688	rs556004011	G	A	2	0
chr22	33199303	rs554665607	G	C	2	0