Molecular genetics of lobular breast cancer ductal carcinoma in situ

Petridis, Christos

Awarding institution:
King's College London

Molecular genetics of lobular breast cancer and ductal carcinoma *in situ*

Christos Petridis

K1208048

Department of Medical and Molecular Genetics

November 2016

Thesis submitted to King's College London in fulfilment of the degree of

Doctor of Philosophy

Declaration

I hereby declare that the work presented in this PhD thesis is my own with the exception of collaborators' contributions stated here. Dr Mark Brook performed the meta-analysis of the GLACIER and ICICLE data sets with the corresponding ones from BCAC studies under the supervision of Professor Montserrat Garcia-Closas. Nik Ruggles curated the GLACIER and ICICLE databases and updated the clinical information on individuals, and Iteeka Arora assisted in sample handling for the library preparation of the targeted sequencing. Finally, part of the work presented in Chapters 2,3, and 5 of my thesis is part of our previous publications and the result of a big collaborative project with many researchers including myself, contributing towards the data collection/processing/analysis and manuscript preparation.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisors, Prof Michael Simpson and Dr Elinor Sawyer for the continuous support of my PhD study and related research, for their patience, motivation, and immense knowledge. Their guidance helped develop all the skills that I acquired over the last years. Both were very supportive during a very stressful period during the writing of this thesis. I could not have imagined having better supervisors for my PhD. Both Elly and Michael have been role models over the last years that I've been working under their supervision. It was a great pleasure to meet them and have them in my life.

I enjoyed every single second of working in both the Medical and Molecular Genetics Department and the Research Oncology Department. It was an extremely friendly and motivating environment. Throughout my course I was looking forward to go to work and participate in stimulating conversations. I would like to thank all my lab mates and in particular Inti for his extreme motivation for knowledge sharing and his passion for data analysis.

I would also like to thank my I-Cell Therapeutics comrades who made my time in the campus very pleasurable. Bauke, Maarten, and Thivyan were always there. The four of us had several occasions of knowledge sharing, entrepreneurial thinking, and of course a lot of fun.

All the members of the Guy's club who made (almost) every Friday special after the long working weeks. Special thanks to Chronis, Dorita, Ines, Matt, Jake, Diego and Sebas.

Additionally, I would like to thank my good friends Christos, and Salena for all the valuable conversations, advice, suggestions and quality time.

During my PhD, apart from friends, I also met a family member, Seth Seegobin. I deeply thank my two non-blood brothers, Seth and Kostas. Both of them were always there for me in good and bad days and we have been through many different situations together that made us stronger and I really appreciate their help and presence in my life… "you are allowed to fall, you are obliged to stand back up"

I don't have words to describe the gratitude I feel for the three people I share half of my genome with. My parents and my brother mean everything to me and I would be literally nowhere at least without my parents… I have all three of them on my side for all my life and I would like to thank them from the bottom of my heart for making me who I am and shaping my personality.

Finally, I would like to deeply thank and devote my thesis to the person I devoted my life; my wife, Ellie, for her endless love and support over the last 12 years and for many more to come.

Abstract

Ductal carcinoma *in situ* (DCIS) and lobular carcinoma *in situ* (LCIS) are clinically undetectable forms of non-invasive breast cancer. DCIS is considered a non-obligate precursor of invasive ductal carcinoma (IDC). LCIS shares many of the same genetic aberrations as invasive lobular breast cancer (ILC), which accounts for 10-15% of all invasive breast cancer. With the advent of screening mammography, the diagnosis of pure DCIS (with no invasive component) and LCIS has become more common, and approximately 20% of screen detected tumours are pure DCIS.

The aim of this project is to test the hypothesis that breast cancer is a heterogeneous disease and that by focusing on specific histological subtypes we can increase the power to detect genetic variants that predispose to DCIS/LCIS/ILC. We also exploited the extreme phenotype hypothesis, having focused on cases with a severe phenotype such as early-onset or bilateral disease.

During this PhD we assessed the role of rare coding variants using next generation sequencing approaches. We also interrogated data on 211,000 SNPs, genotyped on the iCOGS platform in 3,000 DCIS cases 2500 LCIS/ILC and 5000 controls, to evaluate common variants that predispose to these subtypes of breast cancer.

Some of the key findings include the excess of *CDH1* protein truncating variants in cases with bilateral lobular lesions (8%), and the identification of a novel lobular specific locus on 7q34. We were also able to estimate the prevalence of rare variants predisposing to breast cancer in the context of sporadic cases with DCIS/LCIS/ILC. Further analyses and validation is required in order to assess any of the novel putative genes can be linked with ILC development.

Once such variants have been validated they can be used to predict which women are at high risk of developing DCIS/LCIS/ILC and such women can be offered intensive screening or chemoprevention.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

Abbreviation	Meaning
BBCS	British Breast Cancer Study
BCAC	Breast cancer association consortium
BIC	Breast Cancer Information Core
BMI	Body mass index
BOADICEA	Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm
CADD	Combined annotation–dependent depletion
CAF	Combined allele frequency
CI	Confidence interval
COGS	Collaborative oncological gene-environment study
DCIS	Ductal carcinoma in situ
DP	Read depth
EPACTS	Efficient and parallelizable association container toolbox
eQTL	Expression quantitative trait locus
ER	Estrogen receptor
ESP	Exome sequencing project
EVS	Exome variant server
FRR	Familial relative risk
GATK	Genome analysis toolkit
GLACIER	Genetics of LobulAr Carcinoma In situ in EuRope
GQ	Genotyping quality
GWAS	Genome wide association study
GxE	Gene-environment
HDGC	Hereditary diffuse gastric cancer
HR	Hazard ratio
HRT	Hormone replacement therapy
HWE	Hardy Weinberg equilibrium
ICICLE	Investigation of the genetiCs of In situ Carcinoma of the ductaL subtypE
IDC	Invasive ductal carcinoma
ILC	Invasive lobular carcinoma
KASP	Kompetitive allele specific PCR genotyping system
KC	Kinship coefficient
LCIS	Lobular carcinoma in situ
LoF	Loss of function
LOH	Loss of heterozygosity
LRT	Likelihood ratio test
MAC	Minor allele count
MAF	Minor allele frequency
MLPA	Multiplex ligation-dependent probe amplification
MRI	Magnetic resonance imaging
NGS	Next generation sequencing
NICE	The National Institute for Health and Care Excellence
NST	No special type
OC	Oral contraceptive
OCCR	ovarian cancer cluster region
OR	Odds ratio

PCA	Principal component analysis
PCR	Polymerase chain reaction
pLI	Probability of loss of function intolerance
PolyPhen2	Polymorphism Phenotyping v2
PR	Progesteron receptor
PRACTICAL	Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome
PROVEAN	Protein variation effect analyser
Q	quality base call
QC	Quality controls score
qq	Quantile-quantile
RR	Relative risk
RVAS	Rare variant association study
SBCS	Sheffield Breast Cancer Study
SBS	Sequencing by synthesis
SEARCH	Study of Epidemiology & Risk Factors in Cancer Heredity
SIFT	Sorting intolerant from tolerant
SNP	Single nucleotide polymorphism
SVM	Support vector machine
TCGA	The cancer genome atlas
TNBC	Triple negative breast cancer
TNM	Tumour node metastasis
UCSC	University of California, Santa Cruz
UKBGS	UK breakthrough generations study
VEP	Variant effect predictor
WES	Whole exome sequencing

Chapter 1 General introduction

1.1 Breast cancer

Breast cancer is a malignant neoplasm arising from the lactiferous ducts or the milk-producing breast lobules. There are several molecular features which differentiate breast tumours, while also determining prognosis and treatment.

1.1.1 Epidemiology

Breast cancer is the most common malignancy amongst women, accounting for 23% of all cancers. More than 1.4 million women develop breast cancer every year worldwide, 50,000 of which are in the UK [1]. Breast cancer is the second most common cause of cancer death in women, and is estimated that approximately 12,000 women die of breast cancer every year in the UK [2]. The prevalence of the disease is about 1/1000 in the European population and the life-time risk of developing breast cancer for a woman is about 1/8 (12.5%).

There was a steady increase in breast cancer incidence rates between 1975 and 2002 which can be attributed to several factors such as decreased and later-onset parity, increased obesity, excessive use of postmenopausal hormone replacement therapy (HRT), and more importantly likely represents an increased detection rate as a result of the dramatic increase of screening mammography [3]. This pattern of increasing incidence reversed in 2002 due to the results from Women's Health Initiative Trial, where they associated the use of HRT with increased risk of breast cancer and coronary heart disease [4]. It is estimated that there was an annual decline of 8% on breast cancer incidence rates between 2001 and 2004 [5]. Over the last decade there has been a significant increase in breast cancer incidence, reaching almost 20% and this has been attributed to lifestyle changes. A synchronous reduction of mortality for about 15% has also been seen, thought to be due to advances both in treatment modalities and prevention with early diagnosis [6]. *In situ* breast cancer incidence rates have also dramatically increased while use of mammographic screening has been universally adopted, and their detection has become more frequent.

1.1.2 Risk factors

Breast cancer is a complex, multifactorial disease and there are several risk factors contributing to disease development with the major ones being age and gender. Several other risk factors that confer increased risk towards breast cancer development have been identified and can be grouped into four main categories; (i) Genetic factors: family history of disease, mutations in known breast cancer predisposition genes, common polymorphisms associated with the disease. (ii) Breast features: High density of breast tissue, personal history of invasive breast cancer, *in situ* disease, atypical proliferation. (iii) Reproductive and menstrual factors: Number of menstrual cycles, late or no parity, use of HRT, use of oral contraceptives. (iv) Lifestyle factors: Obesity, alcohol consumption, and exposure to radioactive compounds/chemicals. There are also factors that are associated with reduced risk of developing breast cancer and these include low-fat diet, regular exercise and minimal exposure to exogenous hormones such as estrogen [7-9].

1.1.2.1 **Genetic factors**

Family history of the disease is a major risk factor, with women having a first degree relative with breast cancer being approximately two times more likely to develop the disease. However altogether, less than 20% of the women who have breast cancer have a first degree relative with the disease. The inherited genetic factors contributing towards breast cancer development are discussed in section 1.3.

1.1.2.2 **Breast features**

Breast density has been associated with increased risk of breast cancer. It is estimated that about 16% of all breast cancers are attributed to mammographic density. However, breast density is modifiable by several factors such as hormones, parity, body mass index (BMI), and age [10]. Breast density is nowadays routinely reported, and classifies individuals in four categories for each density quartile. The relative risk (RR) for each quartile compared to the least dense quartile (D1) is RR=2.04 for D2, RR=2.81 for D3, and RR=4.08 for D4 [11]. Other breast features that have been identified as breast cancer risk factors are non-invasive lesions such as ductal or lobular carcinoma *in situ* which will be discussed in detail in section 1.2.1.2 and section 1.2.1.4 respectively.

1.1.2.3 **Reproductive and menstrual factors**

One of the risk factors with a significant impact on breast cancer risk is combined estrogen and progesterone hormone replacement therapy (HRT). The relative risk for current users versus never users is 1.66 (95%CI 1.58, 1.75) but can vary depending on duration, and type of hormone therapy but also on the histological subtype of the breast cancer [12, 13]. Investigating a potential association between HRT and breast cancer goes back to the 70s. Since then, several studies have been published reporting association between the use of HRT and breast cancer development, irrespective of whether HRT was estrogen only or combined estrogen and progesterone. In the late 90s, a large meta-analysis confined the risk to current users of HRT [14]. The risk increased along with duration of HRT and reverted back to normal after the end of the treatment course. Five years later, the UK Million women study revealed that the risk of combined estrogen and progesterone HRT had a significantly stronger effect on breast cancer risk both with regards to incidence and death rates [15]. Different morphological subtypes of breast cancer also appear to have distinct aetiological associations with hormonal risk factors. Comparison of invasive lobular and ductal cancers in the UK Million women study has shown that current use of HRT has a stronger association with lobular than ductal cancer [13], and that this risk is higher for those who have used HRT for longer and for those using combined estrogen-progesterone therapy [13, 16].

There is a minor increase in breast cancer risk for women using oral contraceptives (OC) for up to 10 years after cessation. There is no increased risk after that period of 10 years. However, tumours in women taking OC are more likely to be less advanced and there is a reduced relative risk of developing breast cancer compared to non-users (RR=0.88 (95% confidence interval (CI) 0.81, 0.95)) [2]. A more recent study investigating association of reproductive risk factors with ER positive or ER negative breast cancer failed to identify any association of OC with either subtype [17]. However, a study from Iceland found an increased breast cancer risk for ever OC use (HR = 1.32, 95% CI 1.02–1.70) incorporating data from 16,928 individuals [18]. OC usage has been shown to be associated with a 2.5-fold increased risk for triple-negative breast cancer (95%CI, 1.4, 4.3) and no significantly increased risk for non-triple-negative breast cancer (*P*-heterogeneity = 0.008) in women under 40 years of age [19].

Two risk factors that can be summarised by the number of menstrual cycles are the age of menarche and the age at menopause. Having an early menopause is a risk factor for developing breast cancer. Females having their first period before 12 years of age are more likely to develop breast cancer than females having their first period after 12 [20]. Additionally, it has been shown that breast cancer risk is increased by 1.05 for every year younger at menarche and by 1.03 for every year older at menopause. Age-adjusted analysis for individuals between 45 and 54 years of age showed that pre-menopausal women were at increased risk compared to post-menopausal women with a relative risk of 1.43.

A really important and concurrently interesting risk factor for breast cancer is parity. Parity has a dual role on conferring risk for breast cancer, being both protective and harmful. It increases the risk of breast cancer for the first 15 years, while that risk is reduced later on, and switches to a protective factor. It has also been shown that uniparous women with late pregnancies, later than 35 years of age, are at increased risk of breast cancer 5 years after their pregnancy with an odds ratio (OR) of 1.26, and 95% CI of 1.10-1.44 [21]. The risk of developing breast cancer increases by 3% for every year a woman ages after she gives birth for the first time [22]. Nulliparity is most strongly associated with risk of ER positive breast cancer (hazard ratio (HR) = 1.31, 95%CI,:1.23-1.39); whereas late age at first birth is most strongly associated with risk of Estrogen receptor (ER) negative, Progesterone receptor (PR) negative, and HER-2 positive disease (HR = 1.83, 95% CI: 1.31, 2.56) [23].

The effect of breastfeeding was unclear until the collaborative group on hormonal factors in breast cancer meta-analysed data from 47 studies and identified breastfeeding as a protective factor for breast cancer. The relative risk of breast cancer is decreasing by 4% for every year of breastfeeding irrespective of the parity effect [22]. Another important finding of this study conducted in 2002, is that breast cancer incidence could be reduced to half in developed countries if women followed the same average births and breastfeeding patterns as they used to in the early 90s.

1.1.2.4 **Lifestyle factors**

Obesity is associated with increased breast cancer risk in postmenopausal women. A study including approximately 200,000 postmenopausal women, estimated that the risk of developing breast cancer was 20%-40% greater for obese women (BMI ≥ 30) compared to those with a normal weight composition (BMI of 18.5–24.9). It has also been shown that the risk is increased with age. The hazard ratio for BMI ≥35 vs 18.5-25 was 1.24 (95% CI 0.97-1.58) for 50-59 year olds, 1.39 for 60-69 (95% CI 1.24-1.57) and 1.46 (95% CI 1.26-1.70) for ≥ 70 years [24]. Other studies also support this finding [25-27]. Several studies have identified an increase of weight during adulthood being associated with an increased risk of breast cancer [28, 29].

A very large trial evaluating the effect of fatty dietary restrictions failed to identify an advantageous effect [30]. Approximately 50,000 women were allocated to either of the two branches of the trial, which were either a diet with 20% reduced fat or a regular diet. There was no significant value on this trial which was discontinued, even though it is very likely that there was a subset of obese women who had benefited from the trial.

There is an association between alcohol consumption and breast cancer risk, where regular daily drinking can increase the risk by 4% and up to 40-50% if it becomes excessive [31]. The exact mechanisms and pathways through which alcohol can increase breast cancer risk are not widely explored, but since alcohol increases estrogen levels in the blood stream, this is a possible mode of action. This hypothesis can be supported by the fact that the association of alcohol usage is stronger with ER positive than ER negative breast cancers. However, some alcohol metabolites are also known to be carcinogens. Alcohol use in early adulthood has a stronger effect on breast cancer risk. The most important measure of alcohol intake in terms of risk estimation is likely to be the cumulative alcohol consumption [32]. There has been no evidence of differential association depending on the type of alcohol that is being consumed.

A recent study identified working long hours as a risk factor for developing breast cancer. The OR conferred by working more than 55 hours per week was estimated to be 1.6 (95% CI 1.12-2.29) [33]. However, the association is not clear since it could be influenced by parity and therefore more studies would be required to replicate and validate this finding.

High levels of physical activity have been found to reduce the risk of developing breast cancer especially during adolescence or early adulthood [34]. According to estimates from a large cohort study, the equivalent of 10 hours walking exercise per week resulted in 21% reduction in breast cancer risk for all women and 38% for premenopausal women [35]. An interesting finding coming from a study using data from more than 200,000 females indicated that specifically house-holding activities, as opposed to occupational or recreational activities can have a stronger protective effect towards breast cancer [36].

1.1.2.5 **Risk prediction tools**

The major risk factor upon which risk predictions can be made is family history of the disease. Apart from family history, all the aforementioned risk factors could be added in a statistical model and depending on their estimated effect size, predict one's risk of developing the disease. There are several risk prediction tools that have been developed over the last years, the majority of which can be accessed online. The major advantage of these prediction tools is that the clinician can get a single value as an outcome and decide on the appropriate intervention. As expected, there are many discrepancies between different tools mainly due to the different underlying statistical models used but also due to the different risk factors taken into account. Some of those tools, such as BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm), were initially developed to estimate one's risk of being a BRCA mutation carrier, but can also be used to estimate the risk of developing breast cancer [37]. It has been shown to outperform other tools in terms of predicting accurately the likelihood of carrying a *BRCA1* or *BRCA2* mutation [38]. Some tools, such as the Claus model or the BRCAPRO include only family history in their prediction, whereas other models such as the Gail model or the Disease Risk Index emphasise less on family history and focus on other risk factors. The model that includes more risk factors than any other is the IBIS model (International Breast Cancer Intervention Studies) which was the most accurate of all other models at the time its algorithm was published [39].

1.2 Breast cancer classifications

Several different features can be used to classify breast carcinomas. These features can be used in order to stratify individuals in terms of disease aggressiveness, response to treatment, overall prognosis, disease-free survival and other important clinical factors. These features include cyto-nuclear grade of the disease, stage of the disease, and molecular characteristics such as hormone receptor status, Herceptin (HER-2) expression and gene expression patterns of genes associated with proliferation or genes that could be implicated in cancer development in general. Breast carcinomas can also be classified based on their histological and morphological features.

1.2.1 Histological subtypes

There are several different histological subtypes of breast cancer. The most common breast cancer subtype is invasive ductal carcinoma (IDC) with invasive lobular carcinoma (ILC) being second with respect to prevalence. Due to the fact that IDC is a broad term, since most breast cancers arise from the breast ducts, it is more common to refer to it as invasive breast cancer of no special type (NST). According to the WHO classification there are 17 different histological subtypes of breast cancer, some of which are so rare that their combined prevalence accounts for less than 1% of all breast cancers. Some relatively common subtypes of breast cancer are the tubular breast cancers, which account for 2-4% of all breast cancers and are characterised by multiple tubules formed by layers of cancer cells, the mucinous (2%), which form cancer cell clusters at the extracellular mucin, invasive papillary, neuroendocrine, invasive apocrine, metaplastic, invasive micropapillary, medullary, and invasive cribriform carcinomas.

1.2.1.1 **Invasive ductal cancer**

The most common histological subtype of breast cancer is the invasive ductal cancer of no special type (IDC or NST). IDCs account for 70%–80% of all breast cancers. This group of breast tumours comprises all breast tumours that their features do not characterise any of the special subtypes. Therefore, diagnosis of NST ductal invasive carcinomas is based on excluding characteristics associated with specific types of breast cancer [40]. The prognosis of IDCs depends on several factors such as grade, stage, hormone receptor status, HER-2 expression, presence of lympho-vascular invasion and others.

One way of classifying tumours is the TNM (Tumour Node Metastasis) method and is the one that has been established as the standard. The American Joint Committee on Cancer (AJCC), was formed to define the factors that would be included in the classification [41]. More recently, the AJCC has joined powers with the Union for International Cancer Control (UICC) to generate a globally used and standardised system. Depending on the stage of the disease, there can be alternative treatment offered to individuals. Apart from alternative therapeutic options that are offered, prognosis is also very different between different stages of the disease. The classifications have been standardised by the AJCC and UICC committees and range from stage 0 to stage 4 depending on three main features; size of the tumour, lymph node involvement and presence of metastasis.

Grading of tumours is a well-established method, firstly described by Patey and Scharff in 1928. The principals upon which they established their method were firstly reported by Greenough (1925). There are three main features that are used for grading. The first feature is the differentiation status of the cells that can be assessed by the presence of the tubular arrangement of the cells. The second feature is referring to the nuclei of the cells and can be assessed by the variation in shape, staining and size of the nuclei. Finally, the last feature is the frequency or proportion of mitotic cells as described by Bloom and Richardson [42]. Each feature is assessed with a score of 1-3 and all three scores are summed to lead to the final grade points. Tumours with 3-5 grade points are grade I, whereas tumours with 6-7 points are grade II. Finally, tumours with 8-9 grading points are characterised as grade III. This approach has been proven to be very helpful and efficient in categorising patients since the survival rates are different amongst different groups. However, one needs to keep in mind that this scoring system is based on arbitrary threshold selections. This is a continuous scale that has been categorised in order to accommodate an easier way to communicate this information. Low grade tumours have the best prognosis amongst different grades and high grade tumours show the worst outcomes.

A large proportion of IDCs (45–78%) is also associated with ductal carcinoma *in situ* (DCIS) which is a pre-invasive lesion of the breast [43, 44].

1.2.1.2 **Ductal carcinoma *in situ***

DCIS is a non-obligate precursor of invasive breast cancer including IDC. Since the introduction of screening mammography there has been a 7-fold increase in reported DCIS incidence in the USA, primarily in postmenopausal women [45], with about 20% of screen-detected tumours being DCIS [46, 47]. Approximately 55% of all invasive breast cancers are associated with DCIS [43, 44]. For the majority of these cases it is hypothesised that the invasive component has arisen from the DCIS as they generally share somatic genetic changes [48, 49]. About 5% of DCIS cases are bilateral [50]. DCIS subtypes can also be subdivided by morphological features of the tumour [51].

As most DCIS cases are treated surgically, the natural progression of untreated DCIS is not known. Currently there are no accurate methods for predicting the behaviour of DCIS [52]. Although grade has not been shown to be a good predictor of recurrence many clinicians use this classification to determine the use of radiotherapy following breast-conserving surgery. There is a strong correlation between the grade of the *in situ* and co-existing invasive components in IDC, suggesting that DCIS does not progress from low through to high grade before becoming invasive [53, 54].

1.2.1.3 **Invasive lobular cancer**

Invasive lobular breast cancer (ILC) accounts for about 10-15% of all invasive breast cancers and is the second most common subtype of invasive breast cancer after the ductal subtype. Its prevalence has increased over the past years, possibly due to the increase in HRT usage in post-menopausal women [55]. The use of HRT is more strongly associated with ILC compared to IDC (RR=3.1 (95% CI 2.41-4.05) for ILC and RR=1.7 (95% CI 1.57-1.95) for IDC) [7]. The great majority of lobular tumours are characterised as ER and PR positive. Lobular carcinomas arise in the breast lobule as opposed to the ducts, where the more common ductal carcinomas arise (Figure 1.1) and it has been demonstrated that there is a distinct molecular aetiology as well as clinical and biological characteristics [56]. Due to their nature of infiltrating the cell stroma in single file sheets, it is difficult to detect ILCs with mammography. They are also often detected at a more advanced stage than IDC. The majority of ILCs are grade 2. ILCs are less sensitive to chemotherapy compared to IDCs and the 10-year survival rate of women with ILC is lower than that of ER positive IDCs [57, 58]. A recent study that incorporated genomic, transcriptomic and proteomic data in a cohort of ILCs revealed that there are two distinct molecular subtypes of ILC. The two subtypes had distinct gene expression signatures with one of them being immune related whereas the other being hormone related [59]. The vast majority of ILCs are characterised by loss of E-Cadherin expression which is an adhesion molecule encoded by the *CDH1* gene. E-cadherin has been implicated in the development of several different cancer types including breast cancer. ILC is also often associated with a pre-invasive form of lobular carcinoma, known as lobular carcinoma *in situ* (LCIS).

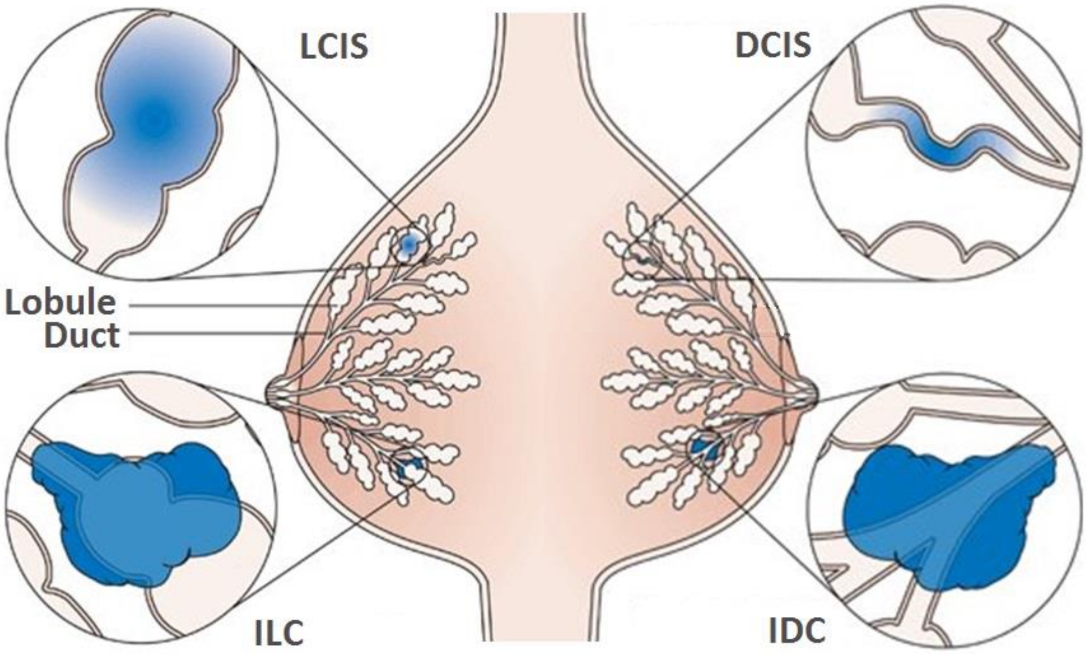


Figure 1.1. Representation of the breast morphology with a focus on the two most common histological subtypes of breast cancer. *In situ* lesions are shown on top and invasive on the bottom part of the figure while the lobular subtype is indicated on the left side with the ductal on the right side of the figure. This figure has been adapted and edited from CancerHelp UK CRUK.org.

1.2.1.4 **Lobular carcinoma *in situ***

LCIS is a non-invasive breast lesion that is typically found incidentally on biopsy. The increased breast biopsy rate associated with screening mammography has led to an increase in the diagnosis of LCIS in post-menopausal women [45]. LCIS is often associated with ILC, and shares many of the same genetic aberrations as ILC including E-cadherin loss, suggesting that it is a precursor lesion in an analogous manner to DCIS preceding IDC [60]. Women who have had LCIS are 4 times (95% CI 2·1–7·5) more likely to develop invasive breast cancer compared to the general population, with an excess of ILC (23-80% of cases) [61-63]. LCIS increases the risk of developing breast cancer and especially ILC either in the ipsilateral or contralateral breast. The cumulative risk of developing invasive breast cancer at 15 years after LCIS diagnosis is 26% [64]. There is a debate of whether it is a precursor lesion or simply a risk factor for ILC. However the invasive cancers associated with LCIS are not exclusively ILC and can often be IDC, tubular cancers or mixed ductal-lobular cancers. In addition, unlike DCIS, LCIS is also a risk factor for developing invasive cancer in the contralateral breast [62].

1.2.2 Molecular subtypes

There are broadly five different molecular signatures identified over the last years that have been widely used to characterise and classify breast cancer. These five patterns have been identified by gene expression profiling. The expression of three main markers; ER, PR, and HER-2, can also classify breast cancers in a relatively robust manner. Along with those three markers, there are specific sets of genes (usually related to proliferation) that are generally switched on or off depending on the breast cancer classification. The five molecular subtypes of breast cancer are generally classified as *luminal-A*, *luminal-B*, *Basal*, *normal-like*, and *HER-2 positive*. Apart from staining for ER, PR and HER-2, that can classify tumours into different molecular subtypes, several different microarray gene expression panels have also been used as described in previous publications [65, 66]. More recently, gene expression panels based on microarray technology or qPCR methods, such as the PAM50, and the MammaPrint have been designed for molecular subtype classification based on the intrinsic subtypes [67-69]. Additionally, a recent study showed that genotyping specific genetic markers across the genome and using machine learning approaches such as support vector machine (SVM) can successfully distinguish tumours based on their ER status [70]. There are different prognostic and therapeutic implications for each different biological or molecular intrinsic subtype [71].

Some limitations of this classification is that gene expression patterns can separate specific subtypes better than others and there is no consistency amongst different methods or gene-sets especially on classifying luminal subtypes. Moreover, it has been shown that normal tissue contamination can significantly alter the gene expression profile and misclassify tumours into less aggressive categories. The molecular classification of breast tumours based on the well-defined immuno-histochemical markers remains a more practical method compared to more complicated gene expression profiles. Until novel molecular targets that can drive individual tumours to grow are identified and can be targeted either for diagnosis or for treatment, it is likely that immunohistochemistry will play a major role in defining breast cancer clinical practice [71].

1.2.2.1 **Luminal A**

Luminal-A breast cancers are characterised by high levels of ER and low levels of proliferation promoting genes. These tumours account for approximately 50-60% of all breast cancers and constitute the most common breast cancer molecular subtype [71]. They are generally characterised by low grade and can include several different histological subtypes. The prognosis for *luminal-A* breast cancers is good compared to other subtypes. Ki67 is a cell proliferation nuclear antigen, the expression of which has been broadly used as a factor to discriminate the luminal subtypes in clinical practice. Its expression follows a continuous pattern but Cheang and colleagues decided on a cut-off of 14%. Luminal breast cancers with Ki67 low (<14%) are classified as *luminal-A* and tumours with high Ki67 (≥ 14%) as *luminal-B* [23].

1.2.2.2 **Luminal B**

Luminal-B breast cancers are usually more aggressive than *luminal-A* and are characterised by higher nuclear grade. They are also associated with a worse prognosis, both in terms of relapse and survival rate [72, 73]. They account of for approximately 15-20% of all breast cancers.

1.2.2.3 **Basal**

Basal breast cancers have the worst prognosis. The vast majority of these tumours are characterised by loss of expression of ER, PR, and HER-2 and are therefore usually being referred to as triple negative breast cancer (TNBC). The vast majority of triple negative breast cancers (TNBC) are high grade invasive ductal carcinomas [74]. Basal and TNBC are not synonyms since there is about 20% discrepancy between the two [75]. They account for 10-35% of all breast cancers depending on the proportion of grade 3 cases included in the studies [76], and they are usually characterised by presence of necrotic zones, and poor tubule formation [71]. Approximately 50% of TNBC cancers respond to chemotherapy, whereas there is a large group that is chemotherapy resistant. This, along with the fact that these tumours do not express the ER, PR, and HER-2, and therefore cannot be treated with endocrine therapy or HER-2 targeted therapy, are some of the reasons why these tumours are more difficult to treat and have the worst prognosis amongst all molecular subtypes.

1.2.2.4 **Normal breast-like (unclassified)**

One intrinsic molecular subtype has been characterised as normal breast-like since the gene expression patterns of these tumours occasionally show similarities with normal breast tissue expression profiles. Between 5 and 10% of breast tumours are characterised as normal breast-like and their prognosis is intermediate, between luminal and basal subtypes. Since there are many similarities with normal breast tissue, a hypothesis that this subtype is the result of poor sample extraction with normal tissue contamination has risen [71].

1.2.2.5 **HER2 overexpression**

Human epidermal growth factor receptor-2 belongs to the membrane tyrosine kinase family. It is encoded by the *HER2* gene located on chromosome 17q21. Her2 operates in homodimers, transducing growth signals within cells. HER-2 overexpressing cells are therefore prone to tumour growth [77]. HER-2 overexpressing tumours are also characterised by overexpression of several HER-2 amplicon associated genes and low expression of ER, PR, and their associated genes. These tumours account for 15-20% of all breast cancers. With regard to the clinical features, HER-2 tumours are more likely to be of high grade and spread to lymph nodes. Initially, HER-2 overexpressing tumours and TNBC had a similar poor prognosis, but with the development of anti-HER-2 targeted therapies, the prognosis for HER-2-overexpressing tumours improved significantly. In general they still have a relatively poor prognosis but they respond to treatment such as trastuzumab (Herceptin) and anthracycline-based chemotherapy [78]. However, Staaf *et al* identified three different subtypes with distinct gene expression profiles, one of which was associated with dramatically poorer prognosis (12% vs 50-55% 10 year survival) compared to the other two [79]. There are several ways to assess the HER-2 status, such as immunohistochemistry, chromogenic *in situ* hybridisation, or fluorescent *in situ* hybridisation. Two studies investigating the correlation between the results of those three different techniques have shown that there is a high concordance between methods when they standard operating procedures are followed [80, 81].

1.3 Genetic predisposition to breast cancer

Breast cancer, like other common complex disorders has a significant inherited genetic component contributing towards disease development, with an estimated heritability of 25% [82]. The heritability of a trait is defined as the phenotypic differences observed that are attributed to genetic variation. The familial relative risk (FRR) is a measure of the genetic component of a trait which corresponds to the familial clustering. FRR is the ratio of the risk of developing breast cancer having a first degree relative with the disease compared to the general population. The FRR for breast cancer can vary based on several factors and can range from 1.4 for someone diagnosed over the age of 60 with a relative diagnosed over 60, to more than 5 for an index case diagnosed before 40 with a relative diagnosed before 40 [83]. However, meta-analyses of several epidemiological studies investigating the familial clustering of breast cancer estimated an overall two fold increased risk of breast cancer in women with an affected first degree relative, RR= 2.1 (95% CI= 2.0-2.2) [84]. The excessive disease correlation amongst monozygotic twins compared to dizygotic demonstrates that the familial risk is predominantly due to genetic factors. Simulation studies have shown that the effect of environmental risk factors would have to increase breast cancer risk on the magnitude of 10-fold to account for and explain part of the familial relative risk and such risk factors have not been identified in the context of breast cancer. Therefore it is hypothesised that breast cancer FFR could be explained by genetic factors [85].

Linkage studies have been used to map disease associated loci by interrogating the co-segregation of genetic markers with affected family members of large families. Having multiple affected and unaffected individuals from the same family can assist to underpin the associated region. The region can then be screened by positional cloning to identify the causative gene. However, using this study design, highly penetrant alleles are required since moderate or low penetrant alleles might not provide enough power to generate a strong linkage signal.

Candidate gene approaches usually include sequencing to identify mutations that are found disproportionally more frequently in affected individuals compared to healthy controls. Biologically plausible candidate genes are sequenced to identify novel breast cancer predisposition genes. Such candidates can include genes that interact with *BRCA1* and *BRCA2*, as well as genes that are involved in similar DNA repair pathways. Other groups of genes that could be implicated with breast cancer include genes involved in cell cycle, checkpoint control, apoptosis, and genes involved in hormone metabolism.

Association studies have been broadly used during the last decade to identify common low penetrance variants associated with traits. These studies can utilise up to more than a million common genetic markers across the genome and investigate potential association across the genome. In a case control association study, genetic markers are investigated to assess their frequency in the case and control populations and can lead to the identification of loci that are significantly associated with the trait of interest, by observing differences in the frequency of the variant between the study populations that are not likely to occur by chance. The vast majority of association studies in breast cancer are genome-wide association studies (GWAS). Initial attempts were underpowered and several type I errors were reported and failed to replicate. Since then, several consortia have been formed establishing large enough data sets to ensure statistical power to detect associations. A study design that minimises type I errors, but also dramatically reduces the experimental costs is the separation of the study into different phases where the top candidates of a phase I study are followed up in a second replication phase II cohort, reducing the cost of the experiment as well as the false positive rates.

The genetic architecture of breast cancer is not completely understood despite the huge efforts. This is possibly due to the complexity and the genetic heterogeneity of the disease. Therefore, even after fine-mapping and identifying associated regions with moderate and small effect size, there is still about 50% unexplained heritability and the aim of this project is to identify a part of it which could be hidden in either common or rare variants (Figure 1.2). It is broadly known that different subtypes of breast cancer have different characteristics and possibly different aetiology [86]. It is now evident that low-risk susceptibility loci are associated with the pathological subtype of breast cancer and support the hypothesis that breast tumours arise through distinct aetiological pathways [87]. The mechanism through which these susceptibility loci contribute to disease development is in most of the cases unknown and is one of the big challenges that researchers are facing.

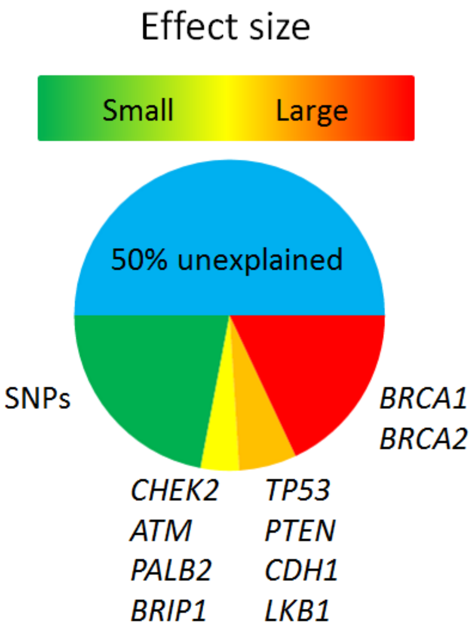


Figure 1.2: Pie chart showing breast cancer heritability attributed to mutations in several genes as well as risk-conferring SNPs. However, there is still approximately 50% missing heritability that remains to be revealed.

1.3.1 High penetrance genes

Two genes, *BRCA1* and *BRCA2* account for about 20% of familial breast cancers and have been identified using linkage family studies and positional cloning at the early 90s [88, 89]. Variants in those two genes have a detrimental effect on the proteins' function leading to a dramatically increased risk (OR>10) of developing breast cancer. There are diagnostic tests developed based on these genes, and women with *BRCA1* or *BRCA2* mutations can be offered intensive surveillance or a risk reducing surgery [90]. Mutations in *TP53* confer high risk towards breast cancer, even though they are uncommon in non-Li-Fraumeni syndrome families, and therefore only account for a small fraction of the familial risk for breast cancer. Another gene harbouring highly penetrant variants is the *CDH1* gene which specifically predisposes to the lobular histology.

1.3.1.1 ***BRCA1***

In 1990, an association between early onset breast cancer and a locus on chromosome 17q21 was found utilising linkage analysis [91]. This led to the identification of *BRCA1* gene four years later using positional cloning [88]. The population frequency of *BRCA1* mutations is estimated to be approximately 1/1000 [92]. BRCA1 has a major role in DNA damage response, DNA repair. It also has a function in regulating chromatin formation and cell cycle. Loss of *BRCA1* can lead to malfunctioning DNA repair that is error prone. Therefore, loss of function (LoF) mutations in

BRCA1 result in genomic instability that confers increased risk of tumorigenesis [93]. Breast cancers arising from a *BRCA1* mutation show a triple negative phenotype and are generally characterized by high grade, with frequent similarities with medullary tumours characteristics [94, 95].

The cumulative breast cancer risk for *BRCA1* mutation carriers at the age of 70 is 57% (95% CI, 47% to 66%) [96]. *BRCA1* mutations confer a lifetime risk of developing breast cancer up to 85%. The age-corrected relative risk of breast cancer is higher in individuals under the age of 40. Pathogenic mutations in *BRCA1* also increase the risk of developing ovarian cancer. *BRCA1* is a large gene comprising 24 exons. *BRCA1* mutations are distributed across all exons of the gene. Most mutations are protein truncating and include nonsense variants, frameshift insertions or deletions (indels) and variants that alter the wild type splice sites. There are some founder mutations that are relatively frequent in certain populations, such as the Ashkenazi Jewish population and the Polish population. However, the vast majority of *BRCA1* mutations are very rare individually and some have been reported only once. Although there is a large number of more than 1,300 pathogenic *BRCA1* variants according to ClinVar database, one of the main issues that needs to be overcome is the classification of variants currently of unknown significance (VUS).

1.3.1.2 ***BRCA2***

Following the identification of *BRCA1, BRCA2* on chromosome 13q12-13 [89] was linked with breast cancer, and cloned one year later [97]. BRCA2 is a key role player in homologous recombination, both during meiosis and double-strand break DNA repair [98]. *BRCA2* mutations are rare with population frequency estimates ranging from 1/600 to 1/800, conferring a cumulative risk of breast cancer at age 70 years of 49% (95% CI, 40% to 57%). The risk of ovarian cancer is 18% (95% CI, 13% to 23%), which is relatively lower than *BRCA1* carriers [96]. Male *BRCA2* mutation carriers have also elevated risk of 80-100 folds towards breast cancer development, with a life time risk of 10% which is very similar to that of a non-*BRCA2* mutation carrier woman. Prostate cancers are also frequent amongst male *BRCA2* carriers [92].

BRCA2 is also a large gene with 27 exons. Mutations are scattered across the coding portions of the gene and the majority of them are frameshift indels. Several missense variants also exist, the pathogenicity of many of which remains unclear. Large gene rearrangements also exist but are relatively rare. The ovarian cancer cluster region (OCCR) exists in the central part of the gene, on exon 11. Mutations in this region are characterised by an increased risk of ovarian cancers, although the underlying molecular mechanisms remain unknown [99]. Bi-allelic mutations can cause Fanconi anaemia, subtype D1. This rare condition is characterised by skeletal abnormalities, predisposition to several tumour types. There are common founder mutations in *BRCA2* that have been reported in the Ashkenazi population and the Icelanders. Morphologically *BRCA2* mutated breast cancers are heterogeneous and they show similar patterns to sporadic breast tumours as opposed to *BRCA1* related tumours which usually show a distinct morphology. Breast cancers in *BRCA2* mutation carriers are usually ER positive, HER2 negative, and of luminal molecular subtype [94, 95]. They are usually higher grade than sporadic tumours and are less pleomorphic with lower mitotic rates in comparison to *BRCA1* related tumours. Unlike *BRCA1* mutated tumours, they are not characterised by p53 abnormalities [100].

1.3.1.3 ***TP53***

TP53 gene is encoding for the p53 protein which has been implicated in several different cancer types. The gene is located on chromosome 17p13.1 and the encoded nuclear phosphoprotein is a transcription factor that controls the cell cycle progression, DNA damage repair, genomic stability and also apoptosis [101]. Germline mutations in the *TP53* gene can cause Li-Fraumeni syndrome. In 1969, Li and Fraumeni firstly described this syndrome after reviewing medical records and death certificates of 648 childhood rhabdomyosarcoma cases. They identified four families having members affected with childhood sarcoma [102]. Those families also had a high incidence of other malignancies including breast cancer. Therefore, *TP53* has been recognised as a gene that highly penetrant alleles can cause breast cancer. However, the prevalence of germline mutations amongst breast cancer cases is very low [103, 104]. The majority of *TP53* mutations are single nucleotide substitutions that lead to defective DNA binding and activation of other downstream genes [105, 106]. There is a large number of missense variants that have been implicated with cancer development. The *TP53* gene is usually somatically mutated in *BRCA1* related breast cancer [107]. In ClinVar database there are 46 variants that are classified as likely pathogenic and 95 variants classified as pathogenic that could predispose to breast cancer.

1.3.1.4 ***CDH1***

Cadherin-1 or E-cadherin is a classical and very well studied protein of the cadherin superfamily. The encoded product is a cell-cell adhesion glycoprotein consisting of five extracellular cadherin repeats, a transmembrane region, and a cytoplasmic tail that is highly conserved and phosphorylated. Germline *CDH1* mutations were initially reported in gastric cancer patients with a syndrome called hereditary diffuse gastric cancer (HDGC) [108]. Both *in situ* and invasive lobular carcinomas are generally characterised by loss of E-cadherin expression. About 30% of HDGC families where the cause is a *CDH1* germline mutation have at least one individual with ILC [109-112]. However, germline *CDH1* mutations in women with ILC and no family history of HDGC do not seem to be a common event. Out of the 343 women with ILC and no personal or family history of HDGC that have been screened for *CDH1* mutations and reported before, only 3 germline mutations have been identified leading to a CAF<1% (Combined Allele Frequency) [113-116]. The cases in these studies were selected mainly on the basis of early onset disease or family history of ILC.

1.3.2 Moderate risk

Several genes, most of them identified using a candidate gene approach, have alleles that confer moderate to high risk (OR>2) such as *CHEK2*, *PALB2, PTEN, LKB1, BRIP,* and *ATM* [117]. These genes are involved in similar pathways and molecular processes to the two BRCA genes, which made them good candidates to study. The advent of next generation sequencing and the capacity to sequence thousands of individuals at a relatively low cost, either in a whole genome/ exome context or in a targeted gene panel approach, allowed the identification of several other genes that are or might be implicated in breast cancer risk.

1.3.2.1 ***CHEK2***

CHEK2 gene, located on chromosome 22q12.1 encodes for the checkpoint kinase 2, which is a mediator of cellular response to DNA damage. CHEK2 phosphorylates both p53 and BRCA1 and acts as a tumour suppressor [118]. The first mutation reported on the *CHEK2* gene was the 1100delC, having been identified in a study investigating families with Li-Fraumeni syndrome [119]. However, its overall relatively high frequency in the general population classifies it as an intermediate penetrance mutation. It is also the most common *CHEK2* mutation. The frequency of 1100delC varies between populations, normally around 1%. Its frequency is increased in individuals with breast cancer. Particularly in a study interrogating individuals with family history and *BRCA1*/*BRCA2* negative genetic testing the frequency of the 1100delC mutation was almost 5% [120]. The CHEK2 Breast Cancer Case–Control Consortium (2004) investigated this mutation in 10,860 breast cancer cases and 9,065 controls with a frequency of 1.9% for cases and 0.7% for controls (OR 2.34; 95%CI 1.72, 3.20) [121]. Researchers from the breast cancer association consortium (BCAC) found evidence of association for rare missense variants in a cohort of 1,303 breast cancer cases and 1,109 heathy controls using a previously developed algorithm that grades missense variants based on their likelihood of being evolutionary tolerant. This study concluded that there are several *CHEK2* rare variants, some of which are missense, that confer breast cancer risk [122]. One example is the missense mutations p.I157T. This variant (c.T470C) has been associated with breast cancer but confers a lower risk than the 1100delC allele. A meta-analysis of the p.I157T variant, including 19,621 cases and 27,001 controls, estimated an OR=1.48 (95%CI 1.31, 1.66) for unselected breast cancer as well as an increased risk of the lobular subtype with OR=4.17 (95%CI = 2.89, 6.03) [123].

1.3.2.2 ***PALB2***

The *PALB2* gene (partner and localizer of BRCA2), which is located at 16p12.2, encodes for a protein that is a nuclear partner of BRCA2 that facilitates BRCA2 functions in DNA repair [124]. Furthermore, PALB2 can also bind to BRCA1 [125]. It has been characterised as a component of the BRCA complex and can link BRCA1 with BRCA2 [126]. Even though it has been shown that bi-allelic mutations in *PALB2* are associated with Fanconi anemia [127, 128], mono-allelic truncating *PALB2* variants have been identified in cases with breast cancer and were estimated to confer a 2 to 6 fold increased risk of developing breast cancer [129, 130]. One frameshift deletion was identified in one individual in a cohort screening breast cancer families that tested negative during BRCA screening in Spain [131]. Another single frameshift deletion, c.1592delT, accounts for 1% of breast cancer incidence in Finland [130]. This variant has been identified after screening more than 4,500 individuals, including familial cases, sporadic cases, and healthy controls [132]. Moreover, a stop-gain truncating variant (Gln775X) has been found in the French Canadian population [133]. There are 89 pathogenic and 16 likely pathogenic variants associated with breast cancer in the ClinVar database.

1.3.2.3 ***BRIP1***

BRIP1 (or *BACH1*) is a gene located at 17q22.2, encoding a helicase that interacts with BRCA1. This complex has a role in double-strand break repair but also in checkpoint control

[134]. It has been shown that bi-allelic mutations in the *BRIP1* gene cause Fanconi anemia [135]. Since it was known that LoF mutations in *BRCA1* gene predispose to breast cancer, and BRCA1 interacts with BRIP1, it was plausible that there would be germline mutations in the *BRIP1* gene, that predispose to breast cancer in a similar manner to *BRCA1* mutations. In a study investigating more than 1,000 index cases with familial breast cancer that were BRCA negative and more than 2,000 controls, there was a significant enrichment of truncating variants in cases over controls, assigning BRIP1 as a low-moderate breast cancer predisposition gene with an estimated relative risk of 2 [136]. However, very recently a study that included 64,033 individuals with breast cancer and 51,538 healthy controls concluded that there is no association between truncating variants in *BRIP* gene and risk of developing breast cancer [137].

1.3.2.4 ***PTEN***

PTEN (phosphatase and tensin homolog), is another tumour suppressor gene which is located on chromosome 10q23.3. It encodes for the phosphatidylinositol phosphate phosphatase which plays a role in cellular regulation [138, 139]. Germline mutations can cause Cowden disease, a rare autosomal dominant inherited cancer syndrome. Cowden syndrome is characterised by increased risk of developing several types of cancer including breast cancer [140]. Since breast cancer is a characteristic phenotype in Cowden syndrome patients, PTEN is also considered a breast cancer predisposition gene. Due to the fact that most studies estimate the risk of breast cancer conferred by *PTEN* germline mutations using selected patients, they might overestimate the actual effect [141]. According to the ClinVar database, there are 8 variants classified as pathogenic and 1 variant classified as likely pathogenic.

1.3.2.5 ***ATM***

ATM gene (ataxia-telangiectasia mutated), located at chromosome 11q22.3, encodes for ATM which is a member of the PI3K-related protein kinases. ATM is an active serine/threonine kinase. This kinase that has a role in cellular response to DNA double strand breaks by phosphorylating the protein products of genes involved in breast cancer such as *BRCA1*, *TP53*, and *CHEK2* [142]. *ATM* gene was initially discovered as being responsible for the cause of an autosomal recessive condition called ataxia-telangiectasia. It was not long until researchers identified a link between ataxia-telangiectasia and increased risk of breast cancer having studied families with ataxia-telangiectasia [143]. Soon, it became apparent that mutations at the

ATM gene can also predispose to breast cancer. Several studies investigated the link between *ATM* mutations and breast cancer, with contradicting results, either due to different study designs or sample size. A study that screened 443 familial breast cancer cases and 521 healthy controls identified that the same mutations that can cause ataxia-telangiectasia in bi-allelic carriers, can also predispose to breast cancer in mono-allelic carriers [144]. In a study investigating mutations in 2,531 breast cancer cases and 2,245 controls, they found evidence that protein truncating, as well as a subset of missense variants, contribute to breast cancer risk. They hypothesised that missense variants around the FRAP-ATM-TRRAP (FAT) domain, FAT C-terminal (FATC), and kinase domains can have a larger effect on the function of ATM than protein truncating variants [145].

A recent meta-analysis of the previously published studies revealed that the increased risk that carriers have at 50 years of age is estimated to be 6.02% and 32.83% for carriers at the age of 80 [142]. Data from epidemiological studies and segregation analysis of familial breast cancer cases lead to estimates of a two-fold relative risk of breast cancer for *ATM* mutation carriers [144, 146]. Further research needs to reveal whether *ATM* mutation carriers could benefit from alternative treatment or have different response, and what will the clinical utility of *ATM* genetic testing be [147]. To conclude, *ATM* is another DNA repair gene, which along with *CHEK2* confers moderate risk towards breast cancer.

1.3.2.6 ***STK11 / LKB1***

STK11 (or *LKB1*) is a tumour suppressor gene, located on chromosome 19p13.3, and encodes for a member of the serine/threonine kinase family. This kinase acts as a cellular proliferation inhibitor and controller of cell polarity. It has also been shown that it is involved in the mTOR pathway. LoF mutations are the major cause of Peutz-Jeghers syndrome, which follows the autosomal dominant model of inheritance. It is a syndrome characterised by gastrointestinal polyposis and various cancers on different organs including breast cancer [148]. Different studies have estimated the lifetime relative risk of developing breast cancer to be approximately 32%–55% for *STK11* mutation carriers [149-151]. In a study investigating individuals with loss of heterozygosity (LOH) in the region of 19p13 from 14 families with breast cancer there was no mutation found in *STK11* in any of the families, rejecting the hypothesis that this LOH and *STK11* mutations are associated events that lead to breast cancer [152].

1.3.3 Low risk variants

In addition, more common susceptibility loci with low risk alleles have been identified. To date, approximately 100 independent breast cancer susceptibility loci have been identified in GWAS or large scale genotyping studies [153-174]. These associated loci are usually identified by single nucleotide polymorphisms (SNPs) and the vast majority of them are intergenic. The SNP-disease association field exploded about a decade ago with the development of the GWAS. Just over a decade ago, microarrays were developed that can explore known genetic variation across the genome in a cost efficient way. By utilising this technology, there was a vast expansion on case control studies exploring the genetic aetiology of several complex diseases, including breast cancer. More recently, fine mapping is being performed by utilising imputation methodology but also denser genotyping platforms with the aim to capture specific regions important to particular diseases or pathways and identify the functional variants that are tagged by the associated SNPs.

Initially, association studies were focused on variants with some prior evidence of implication in breast cancer such as variants in genes involved in DNA repair, cell cycle regulation, checkpoint control, apoptosis, and hormone metabolism. A BCAC study in 2006 identified borderline associations of 5 SNPs having interrogated 16 SNPs [175]. One of these SNPs was validated the following year as a susceptibility locus to breast cancer by the same consortium. This variant is a common missense variant (p.Asp302His, rs1045485) in *CASP8* [176]. It has a minor allele frequency (MAF) of 10% in the European population and reduces the risk of developing breast cancer (OR=0.88 (95%CI 0.84, 0.92) P= 1.1x10^{-7}). This variant was prioritised due to the fact that *CASP8* is involved in apoptosis and therefore constitutes a good candidate gene. However, this variant failed to replicate in subsequent analyses [154].

The first breast cancer GWAS was a study including almost 45,000 individuals. It was a study separated in three stages starting with screening 266,732 SNPs during stage 1 and finally validating a set of 30 SNPs in stage 3 [153]. An intronic *FGFR2* variant (rs2981582) was associated with familial breast cancer with P=2x10^{-76}, (OR=1.26, 95% CI= 1.23-1.30). One of the most significant and well established associations is the one at the *FGFR2* locus. The *FGFR2* gene is located on chromosome 10q26.13 and encodes for the fibroblast growth factor receptor 2. Along with the *FGFR2* variant, a further 4 common variants were identified as being associated with breast cancer, Table 1.1. Since then, several GWAS have been performed and up to 2013, 37 independent loci have been identified.

A few years ago, different cancer consortia (breast, ovarian, and prostate) combined their powers to establish the Collaborative Oncology Gene-environment Study (COGS). In collaboration with Illumina, they designed a custom cancer platform called iCOGS, which includes more than 210,000 markers across the genome selected from previous GWAS as suggestive cancer susceptibility loci or for fine-mapping in known cancer associated genomic regions. There were four consortia involved in the iCOGS development and the variant selection, one of which is the BCAC.

In 2013, a large scale genotyping study using the iCOGS platform on approximately 50,000 cases and 50,000 controls identified 41 novel loci associated with breast cancer bringing the total to 78 [154]. A year later, two more studies identified 5 further alleles associated with breast cancer [177, 178]. In a more recent study of the same consortium, using the iCOGS genotyping platform and utilising imputation methodologies, incorporating data from more than 120,000 individuals, 15 further loci were identified [155]. A complete up to date list of independent loci that have been associated with breast cancer is reported on Table 1.1. Some of those variants show a stronger association with specific subtypes of breast cancer and that is also indicated in the same table. 15 loci show a stronger association with ER positive disease, whereas 10 are more significantly associated with ER negative breast cancer.

Table 1.1: Common known breast cancer predisposition loci.

Study	Year	SNP	Locus	Gene	Subtype	OR	*P* value
Easton *et al*. [153]	2007	rs889312	5q11.2	intergenic	BC	1.13	7x10^{-12}
		rs13281615	8q24.21	*CASC8*	BC	1.08	5x10^{-12}
		rs2981582	10q26.13	*FGFR2*	BC	1.26	2x10^{-76}
		rs3817198	11p15.5	*LSP1*	BC	1.07	3x10^{-9}
		rs3803662	16q12.1	*CASC16*	BC	1.2	1x10^{-36}
Stacey *et al*. [159]	2007	rs13387042	2q35	intergenic	BC	1.2	1.3x10^{-13}
Stacey *et al*. [160]	2008	rs10941679	5p12	intergenic	ER+	1.27	2.5x10^{-12}
Zheng *et al*. [162]	2009	rs2046210	6q25.1	intergenic	BC	1.29	2x10^{-15}
Ahmed *et al*. [161]	2009	rs4973768	3p24.1	*SLC4A7*	BC	1.11	4.1x10^{-23}
		rs6504950	17q22	*STXBP4*	BC	0.95	1.4x10^{-8}
Thomas *et al*. [163]	2009	rs11249433	1p11.2	*EMBP1*	BC	1.16	1.2x10^{-18}
		rs999737	14q24.1	*RAD51B*	BC	0.94	1.7x10^{-7}
Turnbull *et al*. [164]	2010	rs3757318	6q25.1	*CCDC170*	BC	1.3	2.9x10^{-6}
		rs1562430	8q24.21	*CASC8*	BC	1.17	5.8x10^{-7}
		rs1011970	9p21.3	*CDKN2B*	BC	1.09	2.5x10^{-8}
		rs2380205	10p15.1	intergenic	BC	0.94	4.6x10^{-7}
		rs10995190	10q21.2	*ZNF365*	BC	0.86	5.1x10^{-15}
		rs704010	10q22.3	*ZMIZ1*	BC	1.07	3.7x10^{-9}
		rs909116	11p15.5	*TNNT3*	BC	1.17	7.3x10^{-7}
		rs614367	11q13.3	intergenic	BC	1.15	3.2x10^{-15}
Antoniou *et al*. [165]	2010	rs8170	19p13.11	*BABAM1*	TNBC	1.28	1.2x10^{-6}
		rs2363956	19p13.11	*ANKLE1*	TNBC	0.80	1.1x10^{-7}
Haiman *et al*. [167]	2011	rs10069690	5p15.33	*TERT*	ER-	1.18	1x10^{-10}
Fletcher *et al*. [166]	2011	rs9383938	6q25.1	*ESR1*	BC	1.18	1.4x10^{-7}
		rs865686	9q31.2	intergenic	BC	0.89	1.7x10^{-10}
Cai *et al*. [169]	2011	rs10822013	10q21.2	*ZNF365*	BC	1.12	5.9x10^{-9}
Ghoussaini *et al*. [174]	2012	rs10771399	12p11.22	intergenic		0.85	2.7x10^{-35}
		rs1292011	12q24.21	intergenic	ER+	0.90	2x10^{-15}
		rs2823093	21q21.1	intergenic	ER+	0.93	4.6x10^{-8}
Siddiq *et al*. [168]	2012	rs17530068	6q14.1	intergenic	BC	1.12	1.1x10^{-9}
		rs2284378	20q11.22	*RALY*	ER-	1.16	1.1x10^{-8}
Long *et al*. [170]	2012	rs9485372	6q25.1	*TAB2*	BC	0.9	3.9x10^{-12}
Kim *et al*. [171]	2012	rs13393577	2q34	*ERBB4*	BC	1.53	8.8x10^{-14}
Couch *et al*. [172]	2013	rs2290854	1q32.1	*MDM4*	ER-	1.16	1.3x10^{-7}
Garcia-Closas *et al*. [157]	2013	rs6678914	1q32.1	*LGR6*	ER-	1.10	1.4x10^{-8}
		rs12710696	2p24.1	intergenic	ER-	1.10	4.6x10^{-8}
		rs11075995	16q12.2	*FTO*	ER-	1.11	4x10^{-8}
Michailidou *et al*. [154]	2013	rs616488	1p36.22	*PEX14*	BC	0.94	2x10^{-10}
		rs11552449	1p13.2	*DCLRE1B*	BC	1.07	1.8x10^{-8}
		rs4849887	2q14.2	intergenic	BC	0.91	3.7x10^{-11}
		rs2016394	2q31.1	intergenic	ER+	0.94	1.1x10^{-8}
		rs1550623	2q31.1	intergenic	BC	0.94	3x10^{-8}
		rs16857609	2q35	*DIRC3*	BC	1.08	1.1x10^{-15}
		rs6762644	3p26.1	*ITPR1*	ER+	1.07	1.4x10^{-8}
		rs12493607	3p24.1	*TGFBR2*	ER+	1.07	1x10^{-7}
		rs9790517	4q24	*TET2*	BC	1.05	4.2x10^{-8}
		rs6828523	4q34.1	*ADAM29*	ER+	0.87	2.9x10^{-14}
		rs10472076	5q11.2	intergenic	BC	1.05	2.9x10^{-8}
		rs1353747	5q11.2	*PDE4D*	BC	0.92	2.5x10^{-8}
		rs1432679	5q33.3	*EBF1*	BC	1.07	2x10^{-14}
		rs11242675	6p25.3	intergenic	BC	0.94	7.1x10^{-9}
		rs204247	6p23	intergenic	ER+	1.06	9x10^{-8}
		rs720475	7q35	*ARHGEF5*	ER+	0.93	2.9x10^{-8}
		rs9693444	8p12	intergenic	BC	1.07	9.2x10^{-14}
		rs6472903	8q21.11	*CASC9*	BC	0.91	1.7x10^{-17}
		rs2943559	8q21.11	*HNF4G*	BC	1.13	5.7x10^{-15}
		rs11780156	8q24.21	intergenic	BC	1.07	3.4x10^{-11}
		rs10759243	9q31.2	intergenic	BC	1.06	1.2x10^{-8}
		rs7072776	10p12.31	intergenic	ER+	1.09	2.5x10^{-11}
		rs11814448	10p12.31	intergenic	BC	1.26	9.3x10^{-16}
		rs7904519	10q25.2	*TCF7L2*	BC	1.06	3.1x10^{-8}
		rs11199914	10q26.12	intergenic	ER+	0.94	9.1x10^{-8}
		rs3903072	11q13.1	intergenic	BC	0.95	8.6x10^{-12}
		rs11820646	11q24.3	intergenic	BC	0.95	1.1x10^{-9}
		rs12422552	12p13.1	intergenic	BC	1.05	3.7x10^{-8}
		rs17356907	12q22	intergenic	BC	0.91	1.8x10^{-22}

Study	Year	SNP	Locus	Gene	Subtype	OR	*P* value
		rs11571833	13q13.1	*BRCA2*	BC	1.26	4.9x10^{-8}
		rs2236007	14q13.3	*PAX9*	ER+	0.91	1.9x10^{-10}
		rs2588809	14q24.1	*RAD51B*	ER+	1.10	5.7x10^{-9}
		rs941764	14q32.11	*CCDC88C*	BC	1.06	3.7x10^{-10}
		rs17817449	16q12.2	*FTO*	BC	0.93	6.4x10^{-14}
		rs13329835	16q23.2	*CDYL2*	ER+	1.09	3.4x10^{-10}
		rs527616	18q11.2	intergenic	BC	0.95	1.6x10^{-10}
		rs1436904	18q11.2	*CHST9*	ER+	0.93	7.3x10^{-8}
		rs4808801	19p13.11	*ELL*	BC	0.93	4.6x10^{-15}
		rs3760982	19q13.31	intergenic	BC	1.06	2.1x10^{-10}
		rs132390	22q12.2	*EMID1*	BC	1.12	3.1x10^{-9}
		rs6001930	22q13.1	*MKL1*	BC	1.12	8.8x10^{-19}
Sawyer *et al*. [173]	2014	rs11977670	7q34	intergenic	ILC	1.13	6x10^{-10}
Cai *et al*. [177]	2014	rs4951011	1q32.1b	*ZC3H11A*	BC	1.09	8.8x10^{-9}
		rs10474352	5q14.3	intergenic	BC	1.09	1.7x10^{-9}
		rs2290203	15q26.1	*PRC1*	BC	1.08	4.2x10^{-8}
Milne *et al*. [178]	2014	rs1053338	3p14.1	*ATXN7*	BC	1.07	1x10^{-8}
		rs6964587	7q21.2	*AKAP9*	BC	1.05	2x10^{-10}
Michailidou *et al*. [155]	2015	rs12405132	1q21.1	*RNF115*	BC	0.95	7.9x10^{-9}
		rs12048493	1q21.2	*OTUD7B*	BC	1.07	1.1x10^{-9}
		rs72755295	1q43	*EXO1*	BC	1.15	1.8x10^{-8}
		rs6796502	3p21.3	intergenic	BC	0.92	1.8x10^{-8}
		rs13162653	5p15.1	intergenic	BC	0.95	1.1x10^{-10}
		rs2012709	5p13.3	*SUB1*	BC	1.05	6.4x10^{-9}
		rs7707921	5q14	*ATG10*	BC	0.93	5x10^{-11}
		rs9257408	6p22.1	intergenic	BC	1.05	4.8x10^{-8}
		rs4593472	7q32.3	*LINC-PINT*	BC	0.95	1.8x10^{-9}
		rs13365225	8p11.23	intergenic	BC	0.95	1.1x10^{-8}
		rs13267382	8q23.3	*LINC00536*	BC	1.05	1.7x10^{-8}
		rs11627032	14q32.12	*RIN3*	BC	0.94	4.5x10^{-9}
		rs146699004	17q11.2	*TEFM*	BC	0.93	3.3x10^{-8}
		rs745570	17q25.3	intergenic	BC	0.95	1.4x10^{-9}
		rs6507583	18q12.3	*SETBP1*	BC	0.91	3.2x10^{-8}
Couch *et al*. [156]	2016	rs67073037	2p23.2	*WDR43*	ER-	0.92	4.8x10^{-9}
		rs6562760	13q22	intergenic	ER-	0.92	5x10^{-10}
		rs17181761	13q22	intergenic	ER-	1.09	4.2x10^{-8}
		rs188686860	2q33	intergenic	ER-	1.36	8.3x10^{-8}

The underlying mechanisms of action for the vast majority of the common alleles that are associated with breast cancer are unknown. Fine-mapping and functional studies are required to elucidate the role and contribution of these loci for breast cancer. Several studies have been conducted in order to identify the functional role of GWAS hits [179]. However it has been proved difficult to assess the role of these loci.

Identifying a functional role of signals that lie within gene deserts with the nearest gene being several kilo-bases away, such as the 2q and 8q loci, is even more challenging. The opposite phenomenon is also not trivial, where several genes, some of which might be good biological candidates lie within the associated linkage disequilibrium block. It is likely that the true functional risk modifying variants that are tagged by the association signal will be identified by large-scale sequencing projects such as the 100,000 genomes.

Eight of the associated loci have been investigated thoroughly with fine-mapping studies to identify the true functional variant. Fine-mapping at the 5p15 locus identified several variants that lie in the *TERT* gene and have an effect on the encoded protein's activity [180]. Two variants near the *FGFR2* gene have been found to alter important binding domains of the E2F1 and FOXA1/Era. The target gene of these variants has been found to be *FGFR2* [181]. In a similar manner three variants at the 11q1 locus regulate the *CCND1* gene, which encodes for the Cyclin D1 by altering transcription factor domains [182]. A variant (rs4442975) at the 2q35 locus has been found to confer risk towards breast cancer by regulating the expression of *IGFBP5* gene which encodes for the insulin-like growth factor binding protein 5 [183].

Other approaches that could increase the power to detect associations are to focus on other populations, and especially population isolates, where associations might be easier to detect due to possible bottleneck effects. There is evidence that associations between populations are different and by focusing on different populations we can underpin more novel associations [184]. Specific phenotypic features can also be used as proxies for breast cancer and therefore can also be used to identify common alleles associated with the disease. Such features can include mammographic density or molecular profiling of the individuals [185]. Another approach is to enrich the data set for genetic predisposition by selecting cases with early onset of disease, bilateral disease or family history [90].

Most of the studies have been focusing on breast cancer as one entity apart from separating cases based on their ER expression. Several studies have stratified individuals based on the ER status and identified loci that are either differentially associated between ER positive and ER negative breast cancer or specifically associated with ER positive or negative breast cancer [156, 157]. Two different studies from BCAC identified a total of 8 novel breast cancer predisposition loci being specifically associated with ER negative disease. There is clear evidence that the strength of associations varies between ER positive and ER negative disease. Several studies have interrogated the association of breast cancer predisposition SNPs with the five molecular breast cancer subtypes [87, 186, 187]. One consistent finding is the stronger association of the variant rs2981582 with the *Luminal-A* subtype. Simulation methods predicted that more than another 1,000 independent loci are likely to be associated with breast cancer but with the current data sets and technology we are unable to detect them at the genome-wide significance level [154]. The majority of these loci are expected to have a very small effect size ($1.02<OR<1.05$). Given the fact that the smallest effect size being detected via GWAS to date is approximately 1.05, it becomes apparent that larger data sets and possibly different analytical approaches might be required to identify the excess of unrevealed loci that could increase the explained heritability.

1.3.4 Missing heritability

Even though a huge leap forward has been made in defining the molecular genetic contribution over the last couple of decades, with the advancements in technology and the identification of several genetic loci associated with traits, the problem of missing heritability persists almost for all complex traits. The heritability of a disease is defined as the proportion of the phenotypic difference that is attributable to the genotypes [188]. There are several potential roots for this problem and some arise from the tools that are being used, but also from the way data is interpreted. Some of the possible reasons for the problem of high levels of unexplained heritability in any trait are the following [189].

Firstly, the most significantly associated variant in a region might be tagging an actual functional variant that has not been identified yet. This leads to underestimation of the actual contribution of a genomic locus to a trait.

Another explanation for a fraction of the missing heritability could be that one might need to consider facts outside the standard pre-set models, such as those of rare variants with large effect and common variants with small effect. There might be rare variants with small effect, which would translate to the necessity of millions of samples in order to identify and quantify their effect in a phenotype.

Similarly, there could be rare variants acting as expression Quantitative Trait Loci (eQTLs) for target genes and therefore explain a fraction of the phenotypic variation. Since the majority of loci associated with traits do not lie within exonic portions of genes, it becomes apparent that finding their functional role and mechanism of action towards a phenotype is challenging. Part of these associations can be explained by eQTLs where a variant in a position can regulate the expression of gene and predispose to a phenotype through this mechanism. An eQTL study in breast cancer that was published in 2013 where the authors used publicly available data from TCGA (The Cancer Genome Atlas) and identified 3 cis and 3 trans-acting eQTLs out of the 15 breast cancer loci that they investigated [190].

Furthermore, the statistical models that are currently used are not optimal to assess the effect of common variants with minimal effects. Since the P<5x10^{-8} threshold is used in every GWAS or large scale genomic study, it is limiting the pool of variants that reach that level of significance.

Implementing a different approach that could use the information of the variants that could be hidden below the pre-set level of genome-wide significance could help explaining a part of heritability.

There are methods being developed and used in order to estimate the total amount of the heritability explained by a study, using the information from the whole data set rather than from the "significant hits" only [191].

On the other hand, diseases such as breast cancer are very heterogeneous and by treating them as one entity might not reveal the genetic architecture underlying the disease development. The majority of breast genetic studies focus on breast cancer as one disease, with the exception of ER status stratification [156, 157]. There are studies in psychiatric and quantitative genetics illustrating that dissecting a case cohort into distinct pathological subtypes can increase the power and improve the ability to detect association with a trait when there is evidence of distinct inheritance patterns [192].

The opposite approach has also been shown to be successful. Instead of phenotypic stratification, joining phenotypes can also increase the likelihood of identifying novel genetic variants predisposing to disease since the sample size can dramatically increase, without having a huge impact on the genetic heterogeneity of the samples. A recent study meta-analysed GWAS data from breast ovarian and prostate cancer consortia and identified seven novel loci that are associated with at least two of the three different cancer types all of which included breast cancer (rs17041869, rs7937840, rs1469713, rs200182588, rs8037137, rs5013329, rs9375701) [193].

1.4 Aim of the study

The aim of this study is to investigate the genetic predisposition to distinct morphological subtypes of breast cancer and expand our understanding of breast cancer pathogenesis. Our efforts were focused on DCIS, ILC, and LCIS, with the objective to understand the similarities and differences between different histological subtypes in the context of both rare and common inherited genetic variation.

1.4.1 Distinct subtypes to increase genetic homogeneity

Despite its complexity and heterogeneity, breast cancer has been, in general, treated as one disease in most genetic studies up to date. A few exceptions are studies that looked into ER specific cases and identified ER specific loci [156, 157]. The main hypothesis of this project is based on the fact that breast cancer is a heterogeneous complex disease or even a heterogeneous complex of diseases. This heterogeneity could also be translated to genetic heterogeneity. We focused our efforts on specific histological subtypes of breast cancer based on the hypothesis that by carefully defining the phenotype and focusing on specific morphologies, one is more likely to increase the genetic homogeneity of the sample and therefore increase statistical power to detect association.

1.4.2 Extreme phenotype selection

One of the main hypotheses applied during the course of this PhD was the extreme phenotype hypothesis. There are several studies that have shown that cases with more severe symptoms are more likely to carry a highly penetrant variant contributing towards disease development. Following this principal, we hypothesise that cases with early onset of disease, bilateral disease, or family history are more likely to carry a rare highly penetrant variant predisposing to breast cancer. Individuals that fulfilled certain criteria were prioritised for analysis.

Chapter 2 Rare genetic variation predisposing to ILC and LCIS

2.1 Introduction

Lobular breast cancer is the second most common subtype of breast cancer with an increasing incidence among the female population over the last years. It has been previously shown that both ILC and LCIS have higher familial risks than their ductal counterparts, with the presence of stronger family history and bilateral lesions [50, 194, 195]. These suggest the presence of rare penetrant variants that contribute to breast cancer development.

Lobular carcinomas are characterised by loss of E-cadherin expression and somatic mutations are a common cause of this loss. It was therefore not unreasonable to hypothesise that germline *CDH1* variants may contribute to the familial risk of lobular cancer. Cadherin-1 or E-cadherin is a classical and very well studied protein of the cadherin superfamily. The encoded product is a cell-cell adhesion glycoprotein comprised of five extracellular cadherin repeats, a transmembrane region, and a cytoplasmic tail that is highly conserved and phosphorylated. Loss of 16q and acquired events in the *CDH1* gene, which is known for its tumour suppression function, have been implicated with several different carcinomas including gastric, breast, colorectal, thyroid, and ovarian cancers. E-cadherin's loss of function is thought to contribute to cancer development by promoting proliferation, invasion, and metastasis [196].

Rare missense variants in the *CDH1* gene have not been thoroughly investigated due to their frequency and their expected lower penetrance than protein truncating variants. However, several studies investigating predisposition to breast cancer highlighted a possible association of rs35187787, with breast cancer [197]. This variant is a missense variant (p.A592T) located in exon 12 of *CDH1*.

There is very little data on the prevalence of the other known moderate and high penetrance breast cancer predisposition genes in lobular breast cancer. Lobular cancers have been shown to be more frequent among *BRCA2* carriers than *BRCA1* carriers [198] and there is anecdotal evidence that *CHEK2* and *PALB2* mutations may be associated with ILC [199, 200].

2.2 Aims and strategy

With the aim of understanding genetic predisposition to lobular breast cancer 2,539 individuals recruited through the GLACIER study (Genetics of LobulAr Carcinoma In situ in EuRope) were investigated to study the genetics of both ILC and LCIS. The numbers of cases with different phenotypic characteristics within the GLACIER study are indicated in Figure 7.1, page 157. One of the aims of this project was to assess the frequency of germline mutations in *CDH1* in bilateral LCIS/ILC testing the hypothesis that cases of bilateral LCIS/ILC are more likely to be underpinned by an inherited component. This hypothesis is supported by a recent retrospective study from France of all index cases (165 in total) who had undergone *CDH1* mutation screening in their region from 2006-2012. A second aim of our study was to identify a possible association of rs35187787 with any form of lobular disease, since there was some prior evidence of association with breast and other types of cancer. A third aim of this project was to assess the prevalence of the known rare variants in sporadic lobular breast cancer and also to identify novel rare variants that predispose to ILC. Currently women with lobular breast cancer are only eligible for *CDH1* screening if there is a family history of diffuse gastric cancer. Although *CDH1*, *BRCA2*, *CHEK2* and *PALB2* mutations have been described in ILC [201] the prevalence of mutations in these genes in sporadic lobular breast cancer is unknown. Mutations in *BRCA1* and *TP53* are not well described in ILC, due to their rarity and in our study we aim to assess their contribution towards this histological subtype of breast cancer.

We have employed different approaches to identify variants in either known breast cancer predisposition loci or novel putative genes. In an initial attempt to identify rare variants that predispose to ILC, we utilised WES technology to interrogate the exonic portions along with the splicing flanking regions of all coding genes.

2.3 Pilot whole exome sequencing of seven individuals

As a pilot study we selected 7 individuals with a severe phenotype, including early onset of disease, or bilateral lobular carcinoma. The list of individuals that were prioritised for whole exome sequencing is indicated in Table 2.1. Amongst those individuals, a *CDH1* protein truncating variant was identified, along with a *BRCA2* protein truncating variant. Both variants have been previously described as pathogenic. The *BRCA2* variant was present in an individual diagnosed with unilateral ILC and LCIS at the age of 38. This case had two first degree relatives affected with breast cancer*.* The *CDH1* variant was found in an individual with bilateral LCIS and unilateral ILC at the age of 36 that has been previously tested negative in BRCA genetic screening. This finding led to an extension of this study to investigate the prevalence of *CDH1* truncating variants in a cohort of cases with bilateral lobular lesions.

Table 2.1: Characteristics of 7 individuals selected for the pilot whole exome sequencing project.

Study ID	FHx of Breast Cancer	Age at Dx	Details	Bilateral	FHx of Breast Cancer
TG00978	yes	35	ILC + extensive LCIS	no	yes
TG00784	no	36	L ILC + LCIS; R pure extensive LCIS	yes	no
TG00276	yes	38	ILC + LCIS	no	yes
TG00675	yes	36	ILC + extensive LCIS	no	yes
TG00211	no	37	ILC + extensive LCIS	yes	no
TG00384	no	39	ILC + extensive LCIS	no	no
TG00115	no	40	ILC + extensive LCIS	no	no

2.4 *CDH1* mutations are frequent in bilateral LCIS/ILC

2.4.1 Methods

We screened the GLACIER database for individuals with bilateral lobular lesions, either invasive or *in situ*. At the time of this analysis, DNA from 2,210 cases of LCIS/ILC that have been recruited from 97 UK hospitals was available. Diagnosis from local pathology reports had been confirmed in 1,960 cases. We identified 50 individuals of European ethnicity, having bilateral LCIS/ILC. Cases were considered bilateral if they had evidence of bilateral LCIS with or without invasive carcinoma, or pure ILC either synchronously or sequentially. All cases with FFPE tissue blocks (N=26) underwent histological review to confirm the diagnosis by specialised pathologists. None of the 26 samples that were reviewed had the histological diagnosis changed following review. A proportion of those 50 samples (N=29) were prioritised for whole exome sequencing, whereas the rest were PCR amplified and Sanger sequenced using protocols described in section 7.2. The entire coding sequence and associated splice sites of the *CDH1* gene were screened. Additionally, MLPA was performed in order to investigate the presence of large deletions or duplication including whole exons of the *CDH1* gene.

2.4.2 Results

The characteristics of each bilateral case are summarised in Table 2.2. The majority of patients had bilateral LCIS with or without invasive disease. One case had bilateral ILC with no LCIS and four had bilateral ILC with unilateral LCIS. The median age of diagnosis was 51 (range 36-60, IQR=6). For pure bilateral LCIS the median age was 49 (IQR=5,5), and for LCIS with ILC the median was 51 (IQR=6). The median age of diagnosis for all 50 bilateral cases was 51 years, which was similar to the median age of the 1,791 unilateral cases (51 years) of LCIS/ILC (all non-European cases excluded) collected through GLACIER (*P*=0.89, Mann-Whitney U test).

Family history of breast cancer (not confined to first degree relative) was more frequent in bilateral cases (56%) than unilateral (41%) (*P*=0.042, Fisher's exact test), section 7.3.2. There was no significant excess of family history of gastric cancer in patients with bilateral disease (10%) compared to unilateral (8%) (*P*=0.61 Fisher's exact test).

Four germline mutations were identified in four individuals: one donor splice site mutation (c.48+1G>A), two frame-shift mutations (c.1465insC, c.2398delC), and a nonsense substitution (c.1942G>T), Table 2.3, Figure 2.1, Appendix 3.

Table 2.2: Characteristics of 50 bilateral cases from the GLACIER study. The four carriers are indicated in bold.

Sample ID	Age of first LCIS/ILC	Age of second LCIS/ILC	Pathology left breast	Pathology right breast	Family History of breast cancer	Family History of gastric cancer
TG00092	52		LCIS	LCIS/Mixed invasive	-	
TG00107	49		LCIS	LCIS	Aunt (50) Mother (Paget's)	
TG00138	51		LCIS	LCIS/ILC	-	
TG00144	46		LCIS/ILC	LCIS	-	Grandmother Uncle
TG00162	46		LCIS	LCIS	Mother(43)	
TG00170	54		ILC	LCIS/IDC, ILC	-	Grandfather, Uncle
TG00197	55		LCIS/ILC	LCIS	-	
TG00325	57		LCIS/IDC	LCIS/IDC	-	
TG00413	44		LCIS/ILC	LCIS/IDC	Aunt	
TG00457	53		LCIS/Mixed invasive	LCIS/Mixed invasive	Sister(52), Aunt(60)	Cousin (54)
TG00541	46		LCIS/ILC	LCIS/ILC	Grandmother (70), Aunt (70), Cousin (45)	
TG00544	52		LCIS/IDC	LCIS	-	
TG00598	51		LCIS	LCIS/IDC	-	
TG00617	51		LCIS/ILC	LCIS	-	
TG00632	58		LCIS/ILC	LCIS	-	
TG00645	51		LCIS	LCIS	-	
TG00669	45		LCIS/ILC	LCIS	Sister, Grandmother	
TG00705	44		LCIS	LCIS	-	
TG00762	53	60	LCIS	LCIS/Mixed invasive	Great-grandmother	
TG00777	56		LCIS	LCIS	Grandmother, Aunt, Aunt, Aunt	
TG00784	36	37	LCIS/ILC	LCIS	-	
TG00800	50		LCIS	LCIS/ILC	-	
TG00852	50		LCIS/ILC	LCIS	Mother (50)	
TG00857	54	55	LCIS/ILC	LCIS	Mother (54)	
TG00945	57	59	LCIS/ILC	LCIS	Mother (40), Sister (35)	
TG00985	58		LCIS/ILC	LCIS	Sister	
TG01026	52		ILC	LCIS/ILC	-	
TG01038	52		LCIS	LCIS/Mixed invasive	Great-grandmother (70)	
TG01154	50		LCIS/ILC	LCIS	Sister (57), Cousin	Aunt (65)
TG01192	56		LCIS/IDC	LCIS/ILC	Grandmother, Aunt	
TG01212	60		LCIS	LCIS/ILC	Aunt(55)	
TG01295	48	50	LCIS	LCIS/ILC	Aunt (70)	
TG01366	57		ILC	LCIS/ILC	Mother(55), Cousin(59)	
TG01420	52		LCIS/ILC	LCIS/ILC	-	
TG01457	52		LCIS/ILC	LCIS/ILC	Sister(40), Sister(56)	

Sample ID	Age of first LCIS/ILC	Age of second LCIS/ILC	Pathology left breast	Pathology right breast	Family History of breast cancer	Family History of gastric cancer
TG01483	46		LCIS	LCIS	Sister(46)	
TG01490	60		LCIS/Tubular	LCIS/IDC	Cousin (50)	
TG01513	48		LCIS	LCIS/ILC	Mother(39), Sister(38)	
TG01514	40		LCIS	LCIS/ILC	-	
TG01528	56		LCIS/ILC	LCIS	-	
TG01534	43		LCIS/ILC, Tubular	LCIS	Mother (74)	
TG01648	50		ILC	ILC	-	
TG01666	52	56	LCIS	LCIS/ILC	Grandmother	
TG01672	51		LCIS/ILC	LCIS/ILC	Cousin (55)	
TG01705	52	53	LCIS	LCIS	Grandmother (67)	
TG01777	47		LCIS	LCIS/IDC	Aunt (65)	
TG01824	42	44	LCIS/ILC	LCIS/ILC	-	
TG01844	49		LCIS	LCIS	-	
TG02040	50		LCIS/ILC	ILC	Mother (60), Sister (51), Grandmother (56), Grandmother (60)	
TG02417	49		LCIS/ILC	LCIS/ILC, Tubular	-	

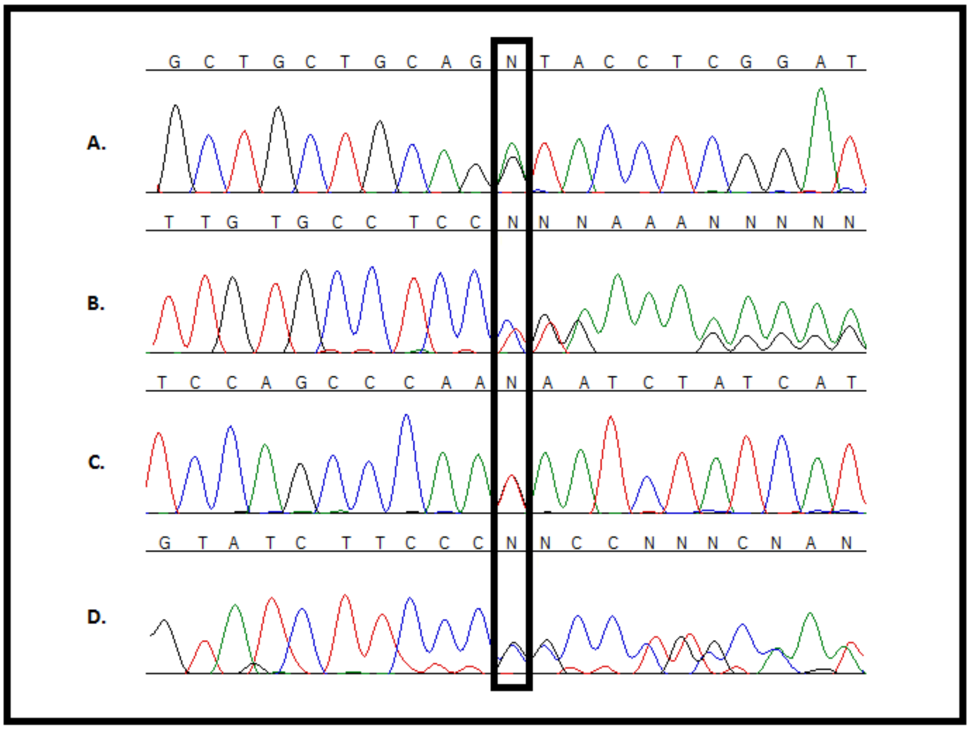


Figure 2.1: Chromatograms of the four protein truncating *CDH1* variants identified in cases with bilateral lobular disease, (A) TG01672:c.48+1G>A, (B) TG00162:c.1465insC, (C) TG01514:c.1942G>T and (D) TG00784:c.2398delC..

The c.48+1G>A mutation is novel, affecting the donor splice site of intron 1 and occurred in an individual with bilateral LCIS and ILC at the age of 51 with a family history of breast cancer. The other three mutations introduce a premature stop codon in exon 10, 13 and 15 respectively, leading to a protein lacking all or part of the intracellular domain. These three mutations have been previously described: c.1942G>T, a somatic mutation in colon cancer [202]; c.1465insC, a germline mutation in diffuse gastric cancer [203]; and c.2398delC, a founder mutation in four Newfoundland families with diffuse gastric and ILC [110]. The c.1465insC mutation occurred in an individual with bilateral pure LCIS at the age of 46 whose mother had breast cancer at the age of 43. The remaining two carriers had bilateral LCIS and unilateral concurrent ILC with no family history of breast cancer. None of the four cases had a family history of gastric cancer. There was no evidence of exonic deletions/duplications in any of the remaining 46 cases without a germline mutation as revealed by MLPA. Using a control data set from the King's College London exome sequencing database, we found no truncating or splice site mutations in *CDH1* in 536 ethnicity matched female individuals (*P*<0.0001, Fisher's exact test).

Table 2.3: Variant information along with pathological features of individuals with *CDH1* protein truncating variants. None of these variants was present in the ExAC population.

Sample ID	Exon	Nucleotide substitution	Amino-acid substitution	Age of diagnosis	Pathology
TG01672	1	c.48+1G>A	Donor splice site	51	Bilateral LCIS, Bilateral ILC
TG00162	10	c.1465insC	p.P489fs	46	Bilateral LCIS, No ILC
TG01514	13	c.1942G>T	p.E648X	40	Bilateral LCIS, Unilateral ILC
TG00784	15	c.2398delC	p.P799fs	37	Bilateral LCIS, Unilateral ILC

Four (8%) of the bilateral cases in our cohort of LCIS/ILC were found to have a germline mutation in *CDH1*. All are predicted to be loss of function, with one being a splicing mutation and the remaining three being truncating mutations. Two have previously been shown to be pathogenic.

Germline *CDH1* mutations were initially reported in patients with hereditary diffuse gastric cancer (HDGC) [108]. Approximately 30% of families with HDGC due to *CDH1* germline mutations also include individuals with ILC [109-112]. However germline *CDH1* mutations in women with ILC that present without a family history of HDGC appear to be rare. Of the 408 cases of LCIS/ILC with no family history of HDGC screened for *CDH1* mutations and reported in the literature, only three germline mutations have been described, all in cases of ILC [113-116], Table 2.4. The cases in these studies were selected mainly on the basis of early onset disease or family history of ILC. In a study where 165 were screened for *CDH1* mutations, they identified 18 individuals with *CDH1* mutations, three of which had bilateral ILC before the age of 50 prior to developing gastric cancer [204].

Table 2.4: Published breast cancer studies that screened the *CDH1* gene for germline mutations in cases with no family history of gastric cancer.

Studies	Phenotypes	Mutation carriers	Total cases
Current study	**Bilateral LCIS/ILC**	**4**	50
	FH of breast cancer[1]	2	27
	Early onset (<45 years)	2	7
	Bilateral LCIS/ILC	4	50
Rahman *et al*. 2000	**LCIS**	**0**	65
	FH of breast cancer	0	20
	Early onset (<45 years)	0	Unknown
	Bilateral LCIS	0	17
Masciari *et al*. 2007	**9 ILC/ 14 mixed pathology**	**1**	23
	FH of breast cancer	1	19
	Early onset (<45 years)	1	4
	Bilateral ILC	Unknown	Unknown
Schrader *et al*. 2011	**ILC**	**0**	318
	FH of breast cancer	0	104
	Early onset (<45 years)	0	214
	Bilateral ILC	0	Unknown
Xie *et al*. 2011	**Familial ILC[2]**	**2**	2
	FH of breast cancer	2	2
	Early onset (<45 years)	1	1
	Bilateral ILC	2	2

The frequency of *CDH1* mutations is much higher than previous studies of LCIS/ILC without a personal or family history of gastric cancer where only 0.7% of the sporadic or familial cases of LCIS/ILC without HDGC carry *CDH1* mutations (*P*=0.003, Fisher's exact test, comparison with previous literature). The median age of the mutation carriers at presentation was eight years lower than that of the 46 *CDH1* negative bilateral cases (43 years versus 51 years respectively). Interestingly only two cases had a family history of breast cancer with one having a first degree relative with the disease (subtype unknown) and none of them having any family history of gastric cancer. This is in accordance with the findings of Claus *et al* [50], who showed that few cases of bilateral LCIS had a family history of breast cancer despite the fact that both bilaterality and family history of breast cancer are reported to be more frequent in LCIS than other breast pathologies [50, 194, 195].

2.5 *CDH1* rare variant c.G1774A:p.A592T (rs35187787) and association with lobular breast cancer

During the pilot WES screening of 7 individuals, we incidentally identified a missense variant (c.G1774A : p.A592T: rs35187787) in one case. Previous studies suggested an association of rs35187787 with cancer having found co-segregation of this variant in families with multiple affected members. Variant rs35187787 has been initially implicated with cancer in a study

[1] Not confined to first-degree relative

[2] Index case only described, Family A, 5 ILC cases, 2 bilateral; Family B, 1 ILC

investigating the genetics of colon cancer. This variant was found in two unrelated individuals with colon cancer one of which had a family history of colon and gastric cancer where the variant co-segregated with both diseases [205]. The same researchers proceeded with a follow up screening of around 1,800 individuals in order to explore their findings further, and to assess the frequency of this variant in breast cancer and a healthy control group. They screened over 1,300 cases with either sporadic or familial breast cancer and approximately 500 controls. The frequency for this variant across different groups (familial cases, sporadic cases, BRCA carriers, early onset cases, controls) varied between 0.56% and 0.83% [206]. Several studies have reported this variant while screening the *CDH1* gene for pathogenic variants. However, due to its prevalence in healthy individuals, most studies are either inconclusive or treat this variant as benign [207].

Since there is a prior knowledge of *CDH1* being associated with breast cancer of lobular subtype, and other researchers have looked at the association of rs35187787 with breast cancer we followed a phenotypic stratification approach where we investigated a possible association of rs35187787 in the context of lobular disease. Rather than including all breast cancer cases in our analysis, we included individuals with lobular breast cancer since there is a prior knowledge of association between *CDH1* and ILC.

2.5.1 Methods

We genotyped 2,630 cases with either LCIS or ILC and 1,471 matched controls from the GLACIER study. Samples were quantified and plated into 96 well plates. Genotyping was outsourced to LGC for this project. The genotyping technology used was KASP and genotypes were analysed in plink. Samples were excluded based on self-reported non-European ethnicity. The final data set comprised 2,440 lobular cases and 1,349 controls. A Fisher’s exact test was used in order to test for association of the variant with lobular breast cancer in general, or any specific subgroup based on onset of disease, bilaterality, and family history. Different prediction tools have been used to assess the deleteriousness of this variant *in silico*. In order to assess potential causality of rs35187787, which has a population frequency of <0.01, we used PROVEAN (Protein Variation Effect Analyser), SIFT (Sorting intolerant from tolerant variants), Polyphen2 (Polymorphism Phenotyping v2), CADD (Combined annotation–dependent depletion), and DANN which are bioinformatics tools that predict a potential functional role of variants including non-synonymous variants. In the context of this thesis, non-synonymous variants are defined as single nucleotide substitutions that alter the amino-acid sequence of a protein coding gene (missense variants).

2.5.2 Results

To assess the frequency and the potential causality of this variant, we used genotypic data obtained from 2,440 lobular cases along with 1,349 matched healthy controls with no personal or family history of breast cancer. No homozygotes were identified. This variant was present in a heterozygous state in 33 cases and 22 controls. No evidence of association was found. There was no underlying association in subgroup analyses of individuals with family history, bilateral lesions or early onset of disease, Table 2.5. This large study of approximately 4,000 individuals provides enough evidence that rs35187787 is not associated with any form of lobular breast cancer. In order to attempt to assess a potential functional effect, we used several bioinformatics tools including PROVEAN (Protein Variation Effect Analyser), SIFT (Sorting intolerant from tolerant variants), and Polyphen2 (Polymorphism Phenotyping v2). All tools suggest that this variant has a significant and damaging effect on the predicted protein's structure and function. Table 2.6 shows the *in silico* predictions of functional consequences for rs35187787. Figure 2.2 indicates that rs35187787 is a highly evolutionary conserved locus.

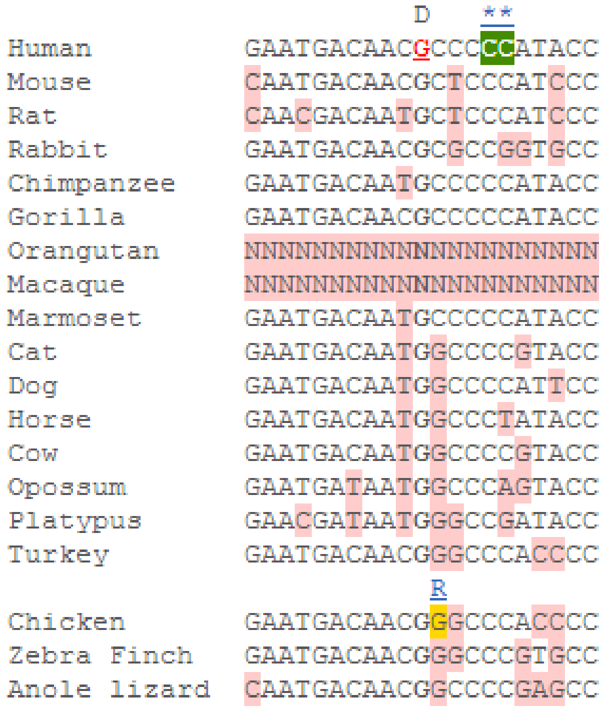
Table 2.5: Summary data on rs35187787 for different GLACIER study populations including a reference European ExAC population group. *P* values are calculated using a Fisher's exact test.

Group	**Carriers**	**Total N**	**MAF**	***P* value**
ExAC	292	66740	0.0043	-
Controls	22	1349	0.0082	Ref
All cases	33	2440	0.0068	0.8
ILC	20	1419	0.0070	0.73
Age<45 or FH any	17	968	0.0088	0.47
Age<45 or FH ILC	9	618	0.0073	0.68
FH ILC	5	432	0.0058	0.82
<45 ILC	7	388	0.0090	0.48
Age<45 and FH ILC	1	64	0.0078	0.66
Bilateral	1	63	0.0079	0.58

However, our data show no significant association using a Fisher's exact test (*P*=0.8) (Table 2.5). Therefore, it is extremely unlikely that this variant is associated with lobular breast cancer. Its frequency in the European population is 0.4% (ExAC).

Table 2.6: *In silico* predictions for rs35187787 using different prediction tools.

Tool	PROVEAN	CADD	DANN	SIFT	PolyPhen2
Score	-2.75	15.26	0.995	0	0.266/ 0.492
Prediction	Damaging	Possibly damaging	Damaging	Deleterious	Benign/ Possibly damaging



```
                                D    **
Human           GAATGACAACGCCCCCATACC
Mouse           CAATGACAACGCTCCCATCCC
Rat             CAACGACAATGCTCCCATCCC
Rabbit          GAATGACAACGCGCCGGTGCC
Chimpanzee      GAATGACAATGCCCCCATACC
Gorilla         GAATGACAACGCCCCCATACC
Orangutan       NNNNNNNNNNNNNNNNNNNNN
Macaque         NNNNNNNNNNNNNNNNNNNNN
Marmoset        GAATGACAATGCCCCCATACC
Cat             GAATGACAATGGCCCCGTACC
Dog             GAATGACAATGGCCCCATTCC
Horse           GAATGACAATGGCCCTATACC
Cow             GAATGACAATGGCCCCGTACC
Opossum         GAATGATAATGGCCCAGTACC
Platypus        GAACGATAATGGGCCGATACC
Turkey          GAATGACAACGGGCCCACCCC
                                R
Chicken         GAATGACAACGGGCCCACCCC
Zebra Finch     GAATGACAACGGGCCCGTGCC
Anole lizard    CAATGACAACGGCCCCGAGCC
```

Figure 2.2: Conservation status of the rs35187787 variant across 19 species.

2.6 Exome sequencing of extreme phenotypes

In an attempt to identify novel breast cancer predisposition genes we exploited the extreme phenotype hypothesis. The underlying basis of this hypothesis is that cases that have a strong family history, or a more severe phenotype such as bilateral disease or early onset of disease, are more likely to carry a rare variant that will confer high risk towards disease development. We performed whole exome sequencing on 51 individuals. The characteristics of the 51 cases with lobular disease that underwent exome sequencing are shown in Table 2.7. In addition, we downloaded data for 110 individuals with ILC from TCGA. Individuals were initially screened for rare variants in known breast cancer predisposition genes. For those with no obvious pathogenic variants identified, an exome-wide gene burden test was performed using 536 female Europeans as controls. The controls have been sequenced in-house for other projects investigating the genetic predisposition to other non-cancer diseases. However, we cannot exclude the possibility that a portion of them might have a history of breast or ovarian cancer even though they were screened for a different trait. All individuals that remained in the final data set, irrespective of their status were of European ancestry, confirmed by principal component analysis (PCA). We identified eight genes not known to be associated with breast cancer that contained rare protein truncating or non-synonymous variants that were predicted to be deleterious in cases of ILC and were either absent or present at low frequency in the controls. The number of variants identified during this study along with some key characteristics of the genes are highlighted in Table 2.12, page 71. The aim of this study was to identify good candidates that would be validated in a larger targeted sequencing phase II study.

Table 2.7: Individuals selected for exome sequencing. Age of diagnosis, pathological phenotype, laterality, family history of breast cancer along with number of affected family members are indicated on the table.

Sample ID	Age	Pathology	Bilateral	FH of BC	N FH 1st d	N FH 2nd d
TG00475	29	ILC/LCIS				
TG00386	34	ILC/LCIS				
TG00396	34	ILC/LCIS				
TG00978	35	ILC/LCIS		x	2	
TG00484	35	ILC/LCIS		x		1
TG01161	35	ILC/LCIS				
TG01238	35	ILC/LCIS				
TG00675	36	ILC/LCIS		x	1	1
TG01543	36	ILC/LCIS		x	1	
TG00784	36	ILC/LCIS	x			
TG00211	37	ILC/LCIS	x			
TG00276	38	ILC/LCIS		x	2	
TG00384	39	ILC/LCIS				
TG01514	40	ILC/LCIS	x			
TG00115	40	ILC/LCIS				
LI_2	40	ILC	?		2	
TG01824	42	ILC/LCIS	x			
TG02438	42	ILC/LCIS	x			
TG01534	43	ILC/LCIS	x	x	1	
TG01901	44	ILC	x	x		3
TG00669	45	ILC/LCIS	x			
TG00541	46	ILC/LCIS	x	x		1
TG00144	46	ILC/LCIS	x			
TG01483	46	LCIS	x			
TG01513	48	ILC/LCIS	x	x	2	
TG01295	48	ILC/LCIS	x	x	1	
TG00107	49	LCIS	x	x	1	1
TG02417	49	ILC/LCIS	x			
TG02525	49	ILC/LCIS	x			
TG01844	49	LCIS	x			
LI_1	49	ILC	?	x	2	
TG01154	50	ILC/LCIS	x	x	1	1
TG02040	50	ILC/LCIS	x	x	2	2
ES_Sister1	50	ILC/LCIS		x	1	
TG01648	50	ILC	x			
TG00852	50	ILC/LCIS	x			
TG01672	51	ILC/LCIS	x	x		1
TG00053	51	ILC/LCIS		x	1	
TG00612	51	ILC/LCIS	x			
TG02394	51	ILC/LCIS	x			
TG01457	52	ILC/LCIS	x	x	2	
TG00088	52	ILC/LCIS				
TG02504	53	ILC	x			
TG02095	56	ILC/LCIS	x	x	1	
101190-INV	56	ILC/LCIS	x			
TG01473	57	ILC	x			
101185-INV	57	ILC/LCIS				
101181-INV	75	ILC/LCIS				
101194-INV	81	ILC/LCIS				
GLC_Sister1	48	ILC/LCIS	?	x	2	
GLC_Sister2	55	IDC/LCIS	?	x	2	

2.6.1 Findings in known breast cancer predisposition genes

In addition to the *CDH1* findings, we investigated the existence of high penetrant variants that have been previously associated with breast cancer. Preliminary exome sequencing analysis revealed the presence of 5 variants in 6 individuals (Table 2.8). Three truncating *BRCA2* mutations have been identified in three individuals with ILC diagnosed at 38, 57 and 65 respectively. Two of the variants are only 10 base pairs apart and lead to a stop codon at exon 11 (c.C5645A, and c.C5655A). The third *BRCA2* variant is a dinucleotide (AG) deletion on exon 20 that leads to a frame-shift (c.8537_8538del). These findings suggest that the prevalence of germline *BRCA2* mutations is approximately 2% for ILC based on our phase I study incorporating data from 110 TCGA samples and 51 samples sequenced in house. This is consistent with previous literature on prevalence of *BRCA2* mutations which is estimated to be between 1% and 5% [208]. Two individuals from the TCGA with ILC being diagnosed at the age of 63 and 80 respectively, carry the *CHEK2* variant c.T470C. This is a non-synonymous missense variant (p.I157T) conferring an overall risk of developing breast cancer of OR=1.5 (95% CI = 1.3-1.7). However, according to a meta-analysis study investigating the prevalence of that variant, the risk of developing lobular breast cancer is significantly higher (OR=4.2, 95%CI CI = 2.9-6) [123]. This provides more evidence for the shared but distinct aetiology of different morphological subtypes of breast cancer. Finally, the last rare penetrant variant that we found in the phase I analysis is located on exon 40 of the *ATM* gene. The individual, age of 72 when diagnosed with ILC, carried the nonsense c.G5932T variant (p.E1978X). This variant has been previously shown to be associated with breast cancer across different populations with an overall OR=5.6 (95% CI 1.3-21.4) [209].

Table 2.8: Established mutations in known breast cancer predisposition genes apart from *CDH1*.

Sample ID	Age	Gene	Exon	AA change	Nt change
TCGA_BC026	72	*ATM*	40	p.E1978X	c.G5932T
S0072	38	*BRCA2*	11	p.S1882X	c.C5645A
S0251	57	*BRCA2*	11	p.C1885X	c.C5655A
TCGA_BC018	65	*BRCA2*	20	p.2846fs	c.8537_8538del
TCGA_BC007	80	*CHEK2*	4	p.I157T	c.T470C
TCGA_BC021	63	*CHEK2*	4	p.I157T	c.T470C

2.6.2 Familial cases

Exome sequencing two individuals from the same family can dramatically increase the power to identify causative variants. Since the expectation is that the same variant is predisposing all family members to breast cancer, having more affected individuals reduces the number of shared variants across all of them and therefore can indicate the true “pathogenic” variant. Two members of the same family have been exome sequenced in two occasions to follow a family based approach in order to enrich the analysis for true pathogenic variants.

In order to stratify and filter variants according to deleteriousness, we applied strict filtering criteria with CADD>30 for non-synonymous variants. An exploratory analysis including known pathogenic and non-pathogenic variants in BRCA1 and BRCA2 indicated that a CADD>30 is a reasonable cut-off. The false positive rate amongst 1,500 variants was < 1%. Protein truncating variants were also included in the analysis. Finally, a MAF<1% cut-off was applied using the 1000 Genomes project and the Exome sequencing project (ESP).

A family pedigree is shown in Figure 2.3, where two siblings were exome sequenced. A list of 6 variants that correspond to 6 genes was the output of the analysis. The genes are; *ATRIP*, *STK38L*, *ULK2*, *OR2AP1*, *RREB1*, and *BAIAP3*, Table 2.9.

Table 2.9: Best candidate genes from Family 1. The frequency of those variants was interrogated in our set of 536 matched controls and in the ExAC population.

Gene	Controls N=536	ExAC frequency	Variant description
ATRIP	0	-	ATRIP:NM_130384:exon12:c.2263delG:p.V755fs
STK38L	0	-	NM_015000:exon4:c.G211A:p.A71T
ULK2	0	-	NM_014683:exon26:c.C2971T:p.R991C
OR2AP1	1	-	NM_001258285:exon1:c.93delT:p.L31fs
RREB1	2	0.00039	NM_001003700:exon10:c.C3191T:p.P1064L
BAIAP3	2	-	NM_003933:exon17:c.1617-1G>T

The *ATRIP* gene was the most plausible candidate both due to the absence of similar variants to a group of controls but also due to its function as a transcription factor. This variant is a novel frameshift deletion in exon 12 (c.2263delG:p.V755fs). DNA from a third affected member of the family was obtained, and the presence of the *ATRIP* protein truncating variant was confirmed using Sanger Sequencing. Two siblings diagnosed at 40 and 49 respectively were selected for exome sequencing whereas germline DNA from their mother was used for Sanger sequencing validation.

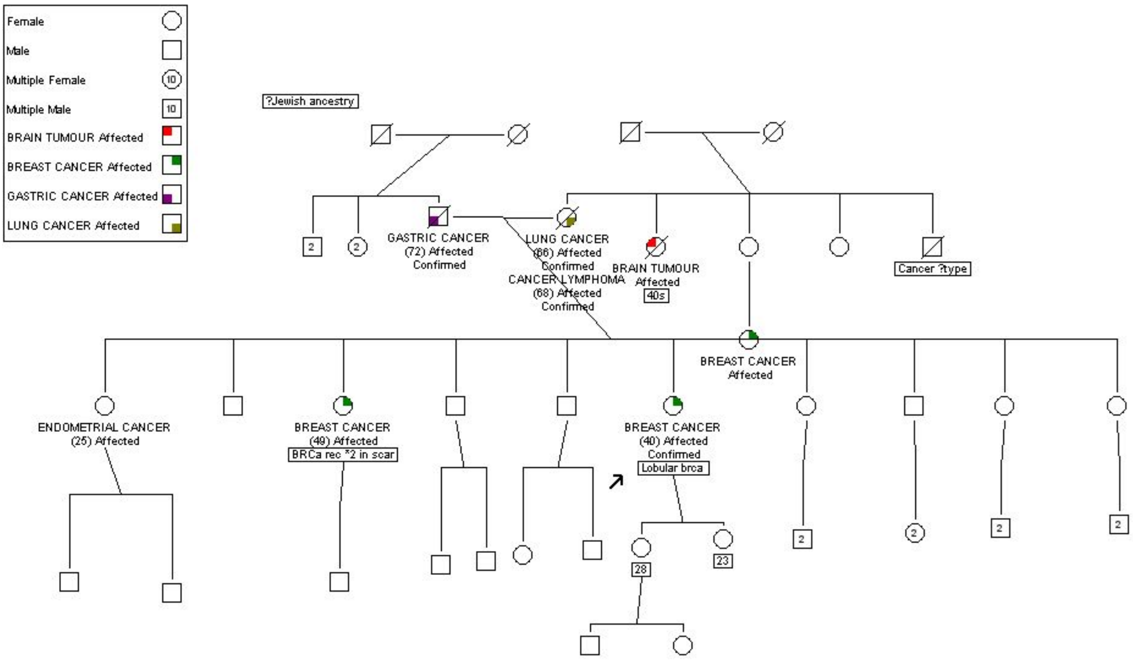


Figure 2.3: Pedigree of a family with three affected individuals, two of which were confirmed to be of lobular histology. Both of those individuals were exome sequenced.

For the second family with three siblings affected with different histological subtypes of breast cancer, a similar approach was followed. A representative pedigree is shown in Figure 2.4. The top candidate genes are reported on Table 2.10. The two genes that seemed to be the best candidates to pursue at the validation stage were *ESR2* and *FMO2*. The *ESR2* variant is a stop-gain protein truncating variant (c.C335A: p.S112X, rs141516067). Its frequency in the general population is less than 0.02%.

ESR2 encodes for estrogen receptor β. The specific role of ER-β, in breast cancer development is currently unknown. However, due to the key role of estrogen in the development of breast cancer [210], and the involvement of ER in the estrogen signal transduction, it has been speculated that variants at the *ESR2* gene could be associated with breast cancer.

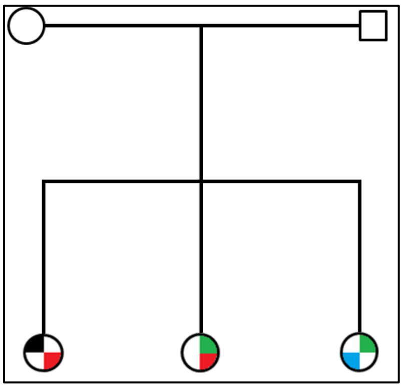


Figure 2.4: Three siblings with different histological subtypes of breast cancer. IDC black), ILC (green), DCIS (blue), and LCIS (red) are shown in the three siblings.

The *FMO2* variant is a novel frameshift insertion in exon 7 (c.893_894insA:p.K298fs). Due to the limitation on the number of amplicons that could be included, the second candidate, *FMO2*, was partially screened with only the exon where the mutation was found being included in the gene panel. The presence of both variants was confirmed on the third sibling using Sanger Sequencing.

Table 2.10: Best candidates from Family 2 with three affected siblings. The frequency of those variants was interrogated in our set of 536 matched controls and in the ExAC population.

Gene	**Controls N=536**	ExAC frequency	Variant description
ESR2	2	0.000092	NM_001040275:exon2:c.C335A:p.S112X
TRPC6	2	0.000033	NM_004621:exon4:c.G1196A:p.R399Q
AHCTF1	1	0.0011	NM_015446:exon27:c.3374+1G>T
BUB1B	1	-	NM_001211:exon2:c.G61A:p.E21K
FMO2	1	-	NM_001460:exon7:c.893_894insA:p.K298fs
IL17RE	1	0.00049	NM_001193380:exon14:c.1297-2A>T
PCSK9	1	0.000039	NM_174936:exon11:c.1863+1G>A
RNF123	1	-	NM_022064:exon13:c.1110+1G>C
ZNF594	1	0.000066	NM_032530:exon2:c.2173_2174insT:p.K725_H726delinsX
DPH6	0	0.00059	NM_080650:exon3:c.G136A:p.D46N
FMNL3	0	-	NM_175736:exon18:c.G2048A:p.R683H
LMNTD1	0	0.0053	NM_001145728:exon3:c.89+1G>A
METTL1	0	0.00012	NM_005371:exon3:c.G326A:p.R109Q
PLEKHN1	0	0.000018	NM_032129:exon14:c.1632delG:p.Q544fs
RILPL1	0	-	NM_178314:exon1:c.G46T:p.E16X
RILPL2	0	-	NM_145058:exon4:c.610delT:p.F204fs
SLC17A6	0	-	NM_020346:exon8:c.1041+1G>A
TM7SF2	0	-	NM_001277233:exon7:c.810_811insC:p.T270fs
UNC5D	0	-	NM_080872:exon9:c.1271_1272insT:p.F424fs
ZNF221	0	-	NM_013359:exon6:c.1294delT:p.Y432fs

2.7 Gene based case control rare variant association study

2.7.1 Design

The gene collapsing method has been followed in order to identify rare, non-synonymous or truncating variants that predispose to lobular breast cancer. We therefore conducted a gene based case control study based on rare variants that are likely to cause a phenotype. This method merges all variants of a certain class, or variants that fulfil certain criteria and measures the excess of their frequency in one group (cases) over the other (controls). Since the prevalence of the disease is approximately 1/10,000 and there is high expected genetic heterogeneity, the expectation for that project would be to identify two or three variants per gene in the cases and that would correspond to an excess compared to the control population where we would expect to observe none or one variant on that same gene.

2.7.2 Methods

Our attempts were focused on 42 cases with severe phenotype that have been exome sequenced in-house and were not carriers of known germline mutations that predispose to breast cancer and 110 ILC cases with ILC that have been downloaded from TCGA. The samples downloaded from the TCGA were of unknown or European ethnic background. We focused our analysis on individuals of European ancestry, and therefore we excluded 1 case based on non-European ancestry after PCA (Figure 2.5).

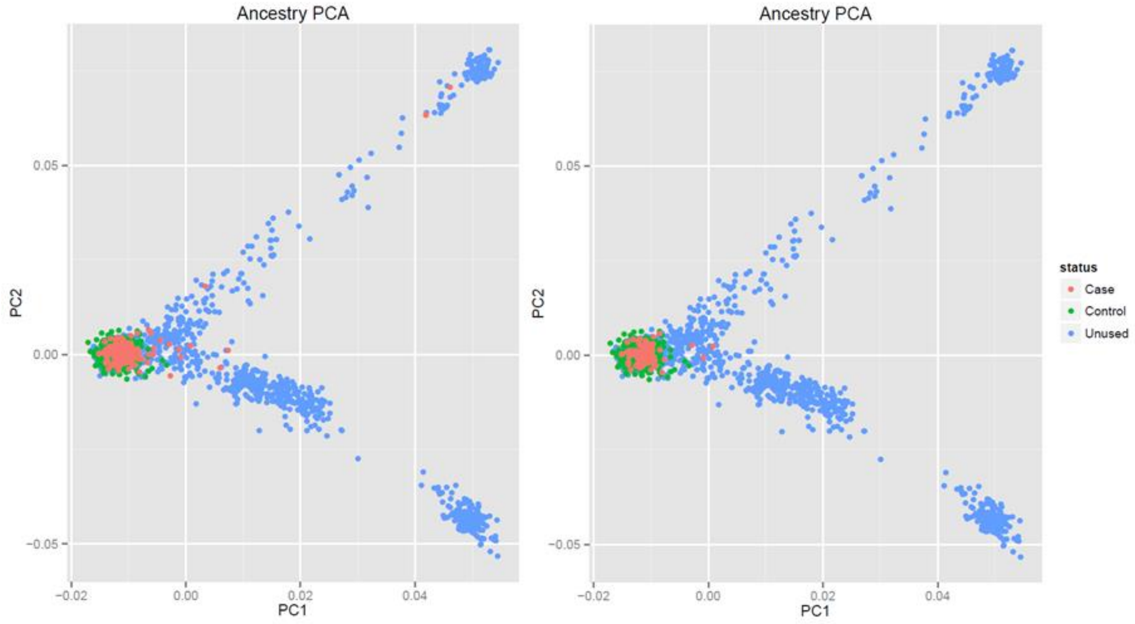


Figure 2.5: Principal component analysis using 9568 common variants (MAF>5%) corresponding to ethnic differences. PC1 vs PC2 are plotted before (left) and after (right) removal of outliers. Red dots correspond to cases (GLACIER and TCGA), green to controls, and blue to unused samples.

Possible sample contamination was suspected in 3 TCGA cases due to high levels of heterozygosity, Figure 2.6, and high relatedness amongst them, and therefore these individuals were also excluded from downstream analyses.

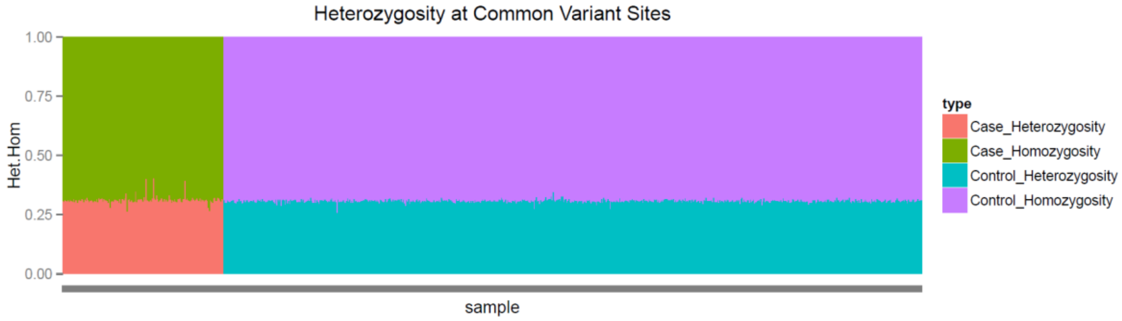


Figure 2.6: Heterozygosity plot from cases and controls included in analysis. Three TCGA cases are contaminated and have significantly higher heterozygosity values.

A total of 41 individuals out of 102 that remained in the final TCGA data set have been diagnosed before the age of 60. The age of diagnosis ranged from 40-90 with a mean of 62 for the TCGA data set. A density plot indicating the age of diagnosis of the GLACIER and the TCGA cases is shown in Figure 2.7. 536 European females that have been exome sequenced in house to identify rare variants predisposing to rare conditions not associated with cancer were used as controls for this study.

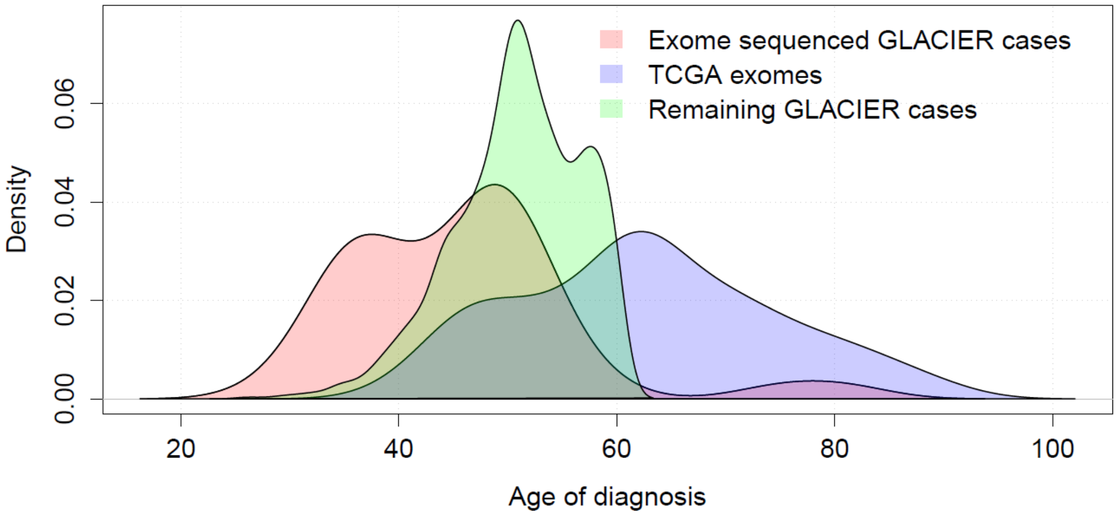


Figure 2.7: Density plot of different case groups. Pink indicates the cases that have been sequenced in-house, purple corresponds to TCGA cases, and green to GLACIER cases that have not been exome sequenced.

The exome sequencing pipeline is described in section 7.4.3. Samples were analysed using gene collapsing tests on EPACTS software as well as manually. Using these approaches we separated the different classes of variants and collapsed all the variants of the same class in a gene. Subsequently, an excess of variants of a specific class on each gene was measured and quantified, which gave an indication of an overall gene burden.

A major filtering criterion is the MAF<1%. Additionally, a variant class filter was applied where all protein truncating variants were kept in the analysis, along with missense variants with a high deleteriousness prediction (CADD>30). The average number of variants post-filtering per individual was 29. A representative quantile-quantile plot is shown in Figure 2.8.

The frequency threshold that has been used in order for variants to remain in the analyses was MAF<1%. However, taking into account the prevalence of the disease (1/10,000) along with the expected genetic heterogeneity and burden of the variants, we can enrich for potentially more damaging variants by altering the threshold and include only very rare variants with MAF<0.005 or MAF<0.001. Indeed, the vast majority of the variants that remained in the analysis were rare with MAF<0.001%. Due to the nature of the phenotype and the prevalence of the disease, we expect very rare events to have a large effect on the phenotype. This means that it is less likely for a single variant to explain a big proportion of the phenotypic variation. We can therefore de-prioritise genes where the "burden signal" comes from a single variant or most individuals carry the same variant/s. On the contrary, if a gene has several unique variants (variants that are found in only one individual), enriched in the cases against the controls, it is more likely that there is some biological significance underlying the statistical differences observed. Moreover, since our hypothesis is to investigate rare "pathogenic" variants, we can focus our research on genes where the number of putative pathogenic variants in the control data set is either none or very close to none. Due to our strict filtering steps, we expect the variants that remained in the analysis to have a substantial effect on the gene's function. Identifying several "highly penetrant variants" in the control group contradicts the initial hypothesis and therefore genes that show this pattern get de-prioritised irrespective of any excess of variants in cases over controls and the significance *P* value.

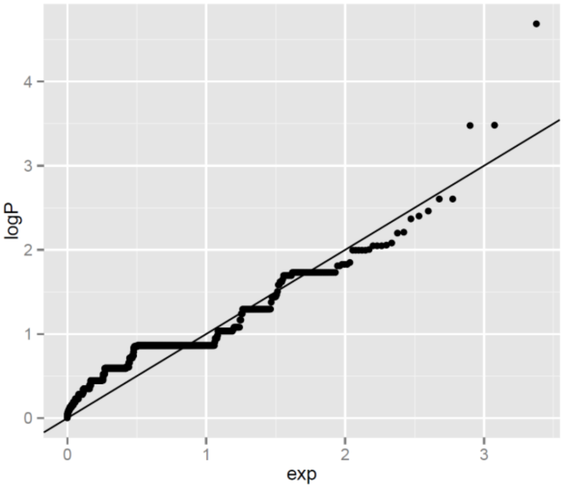


Figure 2.8: Representation of *P* value distribution in a QQ-plot format. The observed p values are plotted against the expected.

2.7.3 Results

The criteria followed for a gene to be ranked in the final list, were either to have at least 2 variants (post-filtered) in GLACIER cases, and no variants in controls, or 2 or more variants in GLACIER and TCGA cases and maximum of 1 variant in controls. Genes that fulfilled these criteria are shown in Table 2.11. We identified twelve genes not known to be associated with breast cancer that contained rare, likely deleterious non-synonymous or protein truncating variants in cases of ILC and were either absent or present only once in the controls. Variants found in those genes underwent manual inspection in Integrative Genomics Viewer (IGV) to ensure that they are efficiently covered and do not result from spurious amplification/ sequencing/ variant calling errors. After manual inspection of these variants in IGV, we selected 8 genes to follow up on a phase II large scale case control study. The selected genes are indicated in bold on Table 2.11. During this analysis, we identified the same non-synonymous variant (rs35610885) in 5 GLACIER cases, 4 TCGA cases, and 7 Controls. This variant is predicted to be deleterious based on its high CADD score (>30) and therefore warrants further investigation. The frequency of the variant in the cases was extremely higher than the controls ($MAF_{GLACIER}$=6%, MAF_{TCGA}=2.5%, $MAF_{Controls}$=0.6%). That variant lies within exon 2 of the *SRA1* gene. Due to the role of *SRA1* gene in hormone metabolism and its usage as a prognostic factor in ER positive breast cancers, we included the exonic portions of that gene in our gene panel of the follow up study using a targeted sequencing approach.

Table 2.11: Top genes from the phase I rare variant case control analysis including 42 GLACIER cases, 107 TCGA cases, and 536 controls. The genes highlighted in bold are the ones that were followed up in phase 2.

Gene	GLACIER	TCGA	All cases	Controls	Bilateral
WDR17	2	1	3	0	1
GOLGB1	2	1	3	0	0
IDE	3	0	3	0	2
MME	2	2	4	0	0
PABPN1L	2	3	5	1	2
SLC15A2	2	1	3	1	1
CYP4F11	2	0	2	0	2
UBTFL1	2	0	2	0	1
NEURL2	2	0	2	0	0
LIMCH1	2	0	2	0	1
DCLRE1B	2	0	2	0	1
ATG2B	2	0	2	0	1

We have therefore selected a set of novel genes, from our exome sequencing case control study of 144 germline samples from women with ILC, that contain rare variants showing a possible association with ILC, Table 2.12. The 144 exomes consisted of 42 cases sequenced in house, selected for their young age of onset (<40 years) or presence of bilateral disease, along with 102 germline ILC exomes downloaded from the Cancer Genome Atlas (https://tcga-data.nci.nih.gov/tcga/) (Accessed 03/03/2014). Carriers of known mutations predisposing to breast cancer were removed prior to the analysis.

Table 2.12: Number of variants in cases and controls along with size of gene for the 8 putative novel loci identified during the phase I case control analysis. Column "Variants" corresponds to the carrier number of GLACIER, TCGA, and Controls.

Gene	Length	Total amplicons in gene	Exons	Variants	Amplicons selected	Coverage
MME	2253	40	30	1.2.0	40	Full
IDE	3060	48	28	3.0.0	48	Full
DCLRE1B	1599	17	4	2.0.0	17	Full
PABPN1L	837	12	8	2.3.1	12	Full
SLC15A2	2190	29	23	2.1.1	4	Exons
GOLGB1	9789	109	26	2.1.0	6	Exons
WDR17	3969	62	33	2.1.0	7	Exons
ATG2B	6237	85	45	2.0.0	6	Exons

2.8 Targeted sequencing of known and putative novel breast cancer predisposition genes

As a phase II study we incorporated findings from phase I study as well as genes with prior evidence of association with breast cancer, and designed a custom targeted sequencing panel comprising 20 genes, Table 2.13.

Six of these genes are known breast cancer susceptibility genes *CDH1*, *BRCA2*, *BRCA1*, *TP53*, *CHEK2*, and *PALB2*. Although *CDH1*, *BRCA2*, *CHEK2* and *PALB2* mutations have been described in ILC [16] the prevalence of mutations in these genes in sporadic lobular breast cancer is unknown. Mutations in *BRCA1* and *TP53* are not well described in ILC, due to their rarity and this study will assess whether they have any association with ILC.

Apart from these 6 known breast cancer predisposition genes, we included 14 putative novel breast cancer genes. The majority of these stem from the phase I exome sequencing rare variant case control study. As shown in Table 2.12, 8 genes were direct candidates from the case control study including 42 ILC cases from the GLACIER study, 102 ILC cases downloaded from the TCGA, and 536 controls. This list is comprised of *MME*, *IDE*, *DCLRE1B*, *PABPN1L*, *SLC15A2*, *GOLGB1*, *WDR17*, and *ATG2B*. Three genes were selected based on individual family analysis after intersecting the results with publicly known databases and the 536 controls that have been used in the phase I study. These include *ESR2*, *FMO2*, and *ATRIP*. A single variant at the *SRA1* gene was found to be significantly enriched in the cases over the controls during the phase I study and due to the function of the gene it constitutes a potential interesting candidate. Therefore, we included all exonic portions and flanking splicing junctions of the gene in our panel. Finally, two genes that have been recently implicated with cancer syndromes, *FAM175A*, and *CTNNA1*, were included in our targeted resequencing gene panel. With regards to the *CTNNA1* gene, there was one individual with ILC in our study with a non-synonymous variant with CADD>30, c.G670A:p.A224T, whereas there was no genetic evidence in our study for *FAM175A*. CTNNA1 is a main interactor of CDH1 in the catenin-cadherin pathway and mutations in that gene could cause a similar phenotype as mutations in the *CDH1* gene, affecting cell adhesion [211]. A recent study identified two protein truncating variants in two families with HDGC [212]. Immunohistochemistry of tumours from carriers revealed loss of α-catenin expression, suggesting a somatic event at the *CTNNA1* gene. They concluded that mutations can cause HDGC and therefore *CTNNA1* screening should be included in genetic testing of prospective families. In addition, *FAM175A* has been recently implicated with cancer predisposition and was deemed a good candidate to include in the targeted sequencing panel. A study from Finland identified an excess of a non-synonymous variant in a cohort of 126 breast cancer families and 868 controls [213]. Additionally, two separate protein truncating mutations have been identified as pathogenic in two different studies investigating hereditary cancer syndromes. The first one is a frameshift insertion, c.1106_1107insG and has been reported in two individuals with ovarian cancer [214]. The second one is another frameshift insertion, c.1032dupT, identified by GeneDX during clinical diagnostics screening. These two variants lie within exon 9. Due to the limitation in amplicons, we included only exon 9 of the gene where the pathogenic variants have been previously identified.

Samples were amplified and sequenced according to the protocols mentioned in section 7.2. The analysis pipeline that was used for data analysis is described in section 7.4. In brief, sequences were aligned to the reference genome (hg19) using Novoalign, primers were stripped using Btrim, and variants were called using GATK’s Haplotype caller and Samtools.

The intersection of those two variant calling algorithms was used as the final data set. Subsequently, Annovar was used to annotate variants according to gene content, population frequency and potential pathogenicity. The databases used are described in section 7.4.

Table 2.13: Information on 20 genes included in the targeted sequencing panel. The transcripts mentioned in this table were used to obtain the HGVS nomenclature for the position of each variant.

Gene	Transcript	Length in bp	Total amplicons in gene	Exons	Amplicons	Coverage	Justification
CDH1	NM_004360	2,649	36	16	36	Full	Lobular breast cancer predisposition
BRCA2	NM_000059	10,257	126	28	126	Full	Lobular breast cancer predisposition
BRCA1	NM_007300	5,552	68	24	68	Full	Breast cancer predisposition
CHEK2	NM_007194	1,761	28	22	28	Full	Breast cancer predisposition
PALB2	NM_024675	3,561	48	15	48	Full	Breast cancer predisposition
FAM175A	NM_139076	1,230	16	10	5	Exons	Suggestive literature evidence
TP53	NM_000546.4	1,146	15	10	15	Full	Breast cancer predisposition
CTNNA1	NM_001903	2,721	32	22	38	Full	Suggestive literature evidence + 1 carrier
ESR2	NM_001291712	1,747	22	20	23	Full	Best candidate from family 1
ATRIP	NM_130384	2,344	28	15	29	Full	Best candidate from family 2
FMO2	NM_018881	1,416	17	9	5	Exons	Second candidate from family 1
SRA1	NM_001035235	891	9	5	12	Full	Excess of a single variant in phase I
MME	NM_007287	2,253	37	30	40	Full	Candidate from phase I study
IDE	NM_004969	3,060	43	28	48	Full	Candidate from phase I study
DCLRE1B	NM_022836	1,599	16	4	17	Full	Candidate from phase I study
PABPN1L	NM_001080487	837	11	8	12	Full	Candidate from phase I study
SLC15A2	NM_021082	2,190	29	23	4	Exons	Candidate from phase I study
GOLGB1	NM_001256486	9,789	109	26	6	Exons	Candidate from phase I study
WDR17	NM_170710	3,969	62	33	7	Exons	Candidate from phase I study
ATG2B	NM_018036	6,237	85	45	6	Exons	Candidate from phase I study

In order to evaluate the role of rare protein coding variation in each gene, we undertook a gene burden analysis. All variants of a certain class (missense, nonsense, etc.) were pooled together to investigate the relative risk conferred by variants of a particular class within a gene rather than each variant individually. This approach allows for meaningful statistical analysis and provide sufficient statistical power to observe the effect sizes of the magnitude that have been previously described for rare variation in cancer predisposition. Our hypothesis is underpinned by an expectation that rare alleles will contribute to the disease. We therefore employed a one-tailed test. The corrected level of significance was set at *P*<0.0036 to correct for multiple testing using the Bonferroni correction (α/N_{genes}) for 14 genes.

The different age groups of individuals across different lobular subgroups groups are indicated in Table 2.14.

Table 2.14: Individuals included in analysis separated by age of diagnosis (cases) and recruitment (controls).

Age groups	**Controls**	**All lobular**	**ILC**	**LCIS**	**LCIS & Non ILC invasive**
≤ 40	0	118	72	24	22
≤ 50	685	958	616	163	179
≤ 60	1,277	2,215	1,443	366	406
All	1,611	2,215	1,443	366	406

Variants that were called by both variant callers (GATK, Samtools) were used for this analysis in order to minimise the number of false positive calls. Variants were further filtered based on read depth (DP), quality control score (QC), and genotypic quality (GQ). Having removed all variants that failed the QC metrics used, we ended up with 1,210 variants in the 6 known breast cancer predisposition genes and 956 variants in the 14 genes under investigation for novel association with lobular breast cancer.

2.8.1 Assess prevalence of known breast cancer predisposition genes

During the phase II study we screened 2,215 individuals diagnosed ≤ 60 with any form of lobular disease out of which 1,443 were ILCs. The exact numbers and phenotypic characteristics of all screened samples are reported in Table 2.14. For variants in known breast cancer predisposition genes, we assigned labels of benign, VUS, and pathogenic, based on the publicly available annotations from ClinVar. With regards to *BRCA1* and *BRCA2*, we also utilised the use of the Breast Cancer Information Core (BIC) BRCA database that incorporates findings from several screening studies including the Myriad database. We therefore identified 46 pathogenic, and likely pathogenic variants across all 6 genes under investigation. A total of 179 VUS were identified across those 6 genes.

2.8.1.1 **BRCA1**

During *BRCA1* screening, 90 non-synonymous/ missense variants were identified in 1,443 cases and 85 in 1,611 controls. In addition, one previously described as pathogenic variants was identified. This variant is non-synonymous and found in an individual diagnosed with ILC at the age of 54. It lies in exon 10 of the gene, c.G1789A:p.E597K, rs55650082. The same variant was found in one of the 366 LCIS cases that were screened as part of that project. One novel finding is the presence of a significant excess of VUS in ILC cases over controls across different age groups. A total of 40 VUS have been identified in controls and 55 in 1,443 ILC cases diagnosed ≤ 60. This leads to a significant enrichment in VUS in ILC compared to controls with OR_{ILC}=1.56 (95% CI 1.03-2.35). This enrichment remains when we investigate specific subgroups of early onset ILC with $OR_{ILC\ Age\leq50}$=1.8 (95% CI 1.09-2.56) and $OR_{ILC\ Age\leq60}$= 2.93 (95% CI 1.12-7.66), Table 2.15, Figure 2.9. *BRCA1* shows an excess of VUS in ILC cases and that is observed across all different age groups.

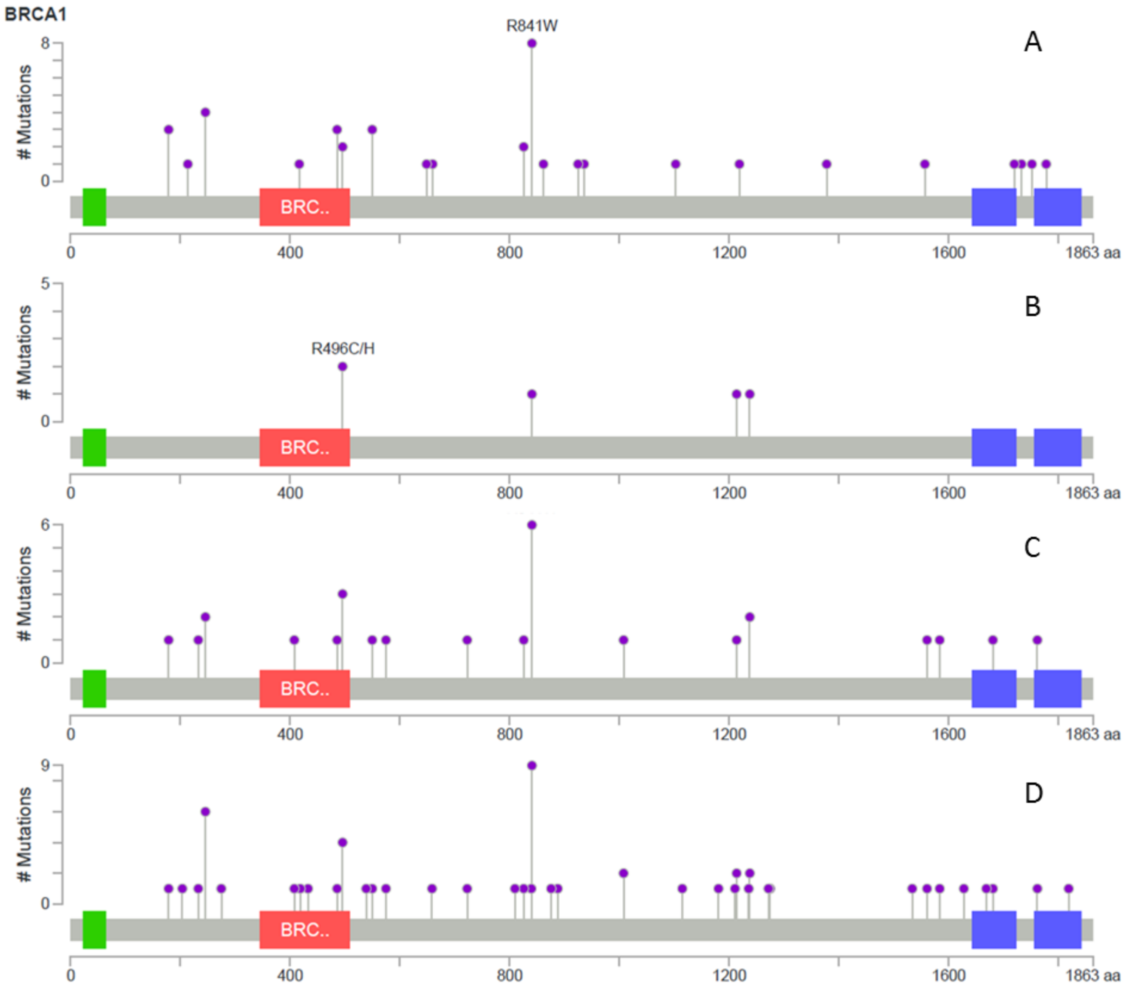


Figure 2.9: Distribution of *BRCA1* VUS across a) controls b) cases with AoD ≤ 40 c) cases with AoD ≤ 50 and d) cases with AoD ≤ 60.

Table 2.15: Prevalence of *BRCA1* VUS in ILC cases and healthy controls. OR and *P* correspond to a Fisher's exact test. ILC are separated into age of onset groups.

Age of diagnosis	Carriers controls N (%)	Carriers cases N (%)	OR	95% CI	*P*
≤ 40	40 (2,5%)	5 (7%)	2.93	1.12, 7.66	0.042
≤ 50	40 (2,5%)	27 (4.3%)	1.8	1.09, 2.96	0.026
≤ 60	40 (2,5%)	55 (3.8%)	1.56	1.03, 2.35	0.037

2.8.1.2 **BRCA2**

BRCA2 is the most enriched gene amongst the 6 that were tested in our targeted sequencing experiment, in terms of pathogenic variants. There are 22 protein truncating variants in ILC cases and only 3 in the control population. This leads to an 8-fold increase of protein truncating variants in the ILC group compared to the controls (OR=8.30, 95%CI 2.5-27.8, *P*=2.7x10^{-5}). Three of those variants have not been described before to our knowledge, indicated with green in Figure 2.10. The first one lies in exon 11 and is a frameshift deletion: c.6068delA: p.D2023fs, found in an individual diagnosed with ILC at the age of 44 with a very high incidence of breast cancer in her family (4 affected relatives). The second one is a deletion in exon 22, c.8942_8943del:p.E2981fs, found in a case diagnosed with ILC at the age of 49. Her mother has also developed breast cancer at the age of 60. Finally, the last novel variant is located in exon 27 and is also a frameshift deletion, c.9719delT:p.V3240fs. This individual was diagnosed at the age of 48 and had no family history of breast cancer.

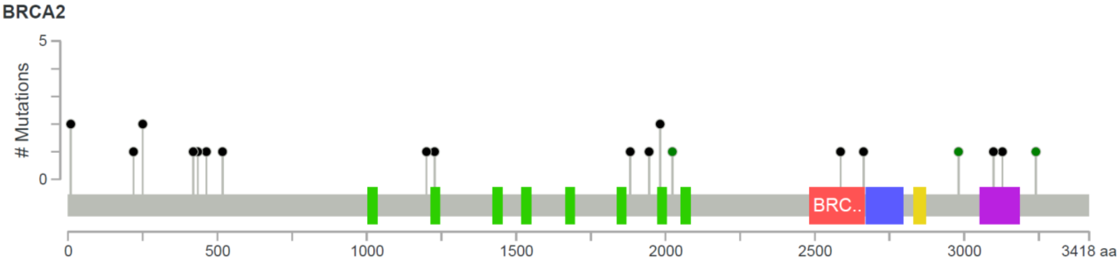


Figure 2.10: Distribution of protein truncating variants identified in ILC cases from our cohort. Black lollipops correspond to previously known pathogenic variants whereas green correspond to variants that have not been described before.

This enrichment gets stronger as we interrogate individuals with early onset of disease. A breakdown of the prevalence of *BRCA2* protein truncating variants amongst different age groups is indicated on Table 2.16. A complete list of *BRCA2* pathogenic variants along with three novel protein truncating variants that were identified in our study is reported in Table 2.17.

Table 2.16: Case control analysis including known breast cancer predisposition variants and novel protein truncating in *BRCA2* in 1,443 ILC cases and 1,611 controls.

Age group	*BRCA2* carriers	Frequency	OR	95% CI	*P*
≤ 40	4	5.5%	31.53	6.92, 143.66	1x10^{-4}
≤ 50	16	2.6%	14.29	4.15, 49.23	4x10^{-7}
≤ 60	22	1.5%	8.30	2.48, 27.78	3x10^{-5}

Table 2.17: Details of *BRCA2* pathogenic variants identified in our cohort.

Sample ID	Nt change	AA change	Class	Pathology	Age
CG00386	c.5946delT	p.S1982fs	Frameshift	Control	41
CG00552	c.5098delG	p.G1700fs	Frameshift	Control	41
CG01138	c.C2612A	p.S871X	Stop-gain	Control	43
TG00276	c.C5645A	p.S1882X	Stop-gain	ILC	38
TG01228	c.750_753del	p.V250fs	Frameshift	ILC	39
150349	c.A7988T	p.E2663V	Non-synonymous	ILC	39
TG01442	c.1257delT	p.C419fs	Frameshift	ILC	40
TG01783	c.517-2A>G		Splicing	ILC	42
TG01787	c.750_753del	p.V250fs	Frameshift	ILC	42
TG01686	c.1301_1304del	p.K434fs	Frameshift	ILC	42
TG01352	c.C9294G	p.Y3098X	Stop-gain	ILC	43
TG01378	c.6068delA	p.D2023fs	Frameshift	ILC	44
150368	c.C9382T	p.R3128X	Stop-gain	ILC	44
TG00783	c.1389_1390del	p.T463fs	Frameshift	ILC	47
TG01338	c.9719delT	p.V3240fs	Frameshift	ILC	48
150411	c.657_658del	p.T219fs	Frameshift	ILC	49
TG00870	c.5946delT	p.S1982fs	Frameshift	ILC	49
TG01141	c.G7757A	p.W2586X	Stop-gain	ILC	49
TG01606	c.8942_8943del	p.E2981fs	Frameshift	ILC	49
TG02084	c.25delC	p.P9fs	Frameshift	ILC	53
TG02115	c.25delC	p.P9fs	Frameshift	ILC	53
150476	c.3598_3599del	p.C1200fs	Frameshift	ILC	55
TG02408	c.3680_3681del	p.L1227fs	Frameshift	ILC	55
TG01934	c.5835dupA	p.I1945fs	Frameshift	ILC	57
TG00467	c.5946delT	p.S1982fs	Frameshift	ILC	60
TG00306	c.517-2A>G		Frameshift	LCIS/IDC	55
TG01559	c.C9294G	p.Y3098X	Frameshift	LCIS	43
TG01506	c.517-2A>G		Splicing	LCIS/ Mixed Duct/Lob	49
TG00255	c.C5682G	p.Y1894X	Stop-gain	LCIS/Mixed Duct/Lob	35
TG00463	c.5835dupA	p.I1945fs	Frameshift	LCIS/Mixed Duct/Lob	58
TG00286	c.8067delT	p.C2689fs	Frameshift	LCIS/Mixed Duct/Lob	33
TG01400	c.9117+1G>A	c.9117+1G>A	Splicing	Mixed Duct/Lob	44

The prevalence of *BRCA2* mutations is relatively high, with 1.4% of all cases with lobular features and 1.5% of ILC cases being carriers of a pathogenic variant. The frequency of pathogenic variants in *BRCA2* increases in groups of early onset disease. As indicated in Table 2.16, the frequency of carriers amongst ILC cases diagnosed ≤ 40 is 5.5%, whereas the frequency of *BRCA2* pathogenic variants is 2.6% for cases diagnosed ≤ 50, dropping down to 1.5% for cases diagnosed ≤ 60. A significant enrichment in VUS with CADD>20 was also observed in *BRCA2*, with 25 ILC cases and 14 controls being carriers (OR=2.01, 95%CI 1.04-3.8, *P*=0.036).

2.8.1.3 **Frequency of protein truncating mutations in the *CDH1* gene**

CDH1 has 6 truncating variants In ILC and none in controls. Four additional protein truncating variants have been identified in individuals with either pure LCIS or LCIS concurrent with non-lobular invasive breast cancer. The location of the six variants found in ILC cases across the *CDH1* gene is indicated in Figure 2.11. A detailed list of all 10 individuals that are carriers, along with some key phenotypic characteristics and variant information, is shown in Table 2.18. Those 10 individuals include the four that have been screened in the cohort of 50 cases with bilateral lobular disease as described in section 2.4. Two of the remaining five variants were novel whereas the other three have been previously described. One variant, rs121964875, has been described in a family with three individuals with gastric cancer [215]. The second variant, rs587781919, has been identified in an individual diagnosed with breast cancer at the age of 60, in a study evaluating the use of targeted diagnostic sequencing [216]. Finally, the last variant, c.1487_1493del, a deletion of seven bases, has been described in an index case from a HGC family, diagnosed at the age of 30 [217]. The three novel variants are frameshift deletions. The first one, c.3delG alters the start codon and is expected to have a detrimental effect to the gene’s function. The second one is located in exon 7, 933delC, whereas the latter lies within the last exon of the gene and even though it is a protein truncating variant, its exact effect cannot be easily estimated. The location of the variant is 18 amino-acids before the stop-codon. The frameshift causes the stop codon to be skipped and the next stop codon is downstream of 46 additional amino-acids. There was no protein truncating variant identified in any of the 1,611 healthy controls. A non-significant excess of rare VUS with CADD>20 was also observed in lobular cases (OR=2.61, *P*=0.21) compared to controls, with 7 ILC carriers and 3 out of 1,611 controls. The coverage of the first exon was suboptimal, probably due to higher GC content, with 31% of the samples having less than 10 reads which was set as the minimum required coverage for variants to be included. The same suboptimal pattern of amplification was observed for exon 12. Therefore, it is possible that a portion of the variation in *CDH1* gene, and specifically in exons 1 and 12, is missed due to technological limitations.

Table 2.18: All *CDH1* protein truncating variants found in our cohort are reported along with characteristics of carriers. Bold indicates variants that have not been previously described.

ID	Status	AoD	Breast cancer FH	Gastric cancer FH	Exon	Amino-Acid change	Nt change
TG02059	**LCIS/IDC**	**43**	**x**	**x**	**1**	**p.M1fs**	**c.3delG**
TG01672	ILC	51	Cousins	x	1		c.48+1G>A
TG01223	LCIS	49	x	x	2	p.W20X	c.G59A
TG50400	ILC	47	x	x	3		c.387+1G>A
TG00589	**LCIS/Mixed**	**51**	**x**	**Father (70)**	**7**	**p.L311fs**	**c.933delC**
TG00162	ILC	46	Mother	x	10	p.P488fs	c.1465dupC
TG00323	LCIS/IDC	58	Mother/sister	Grandfather	10	p.S496fs	c.1487_1493del
TG01514	ILC	40	x	x	13	p.E648X	c.G1942T
TG00784	ILC	36	x	x	15	p.P799fs	c.2398delC
TG01112	**ILC**	**55**	**Mother**	**x**	**16**	**p.W865fs**	**c.2594delG**

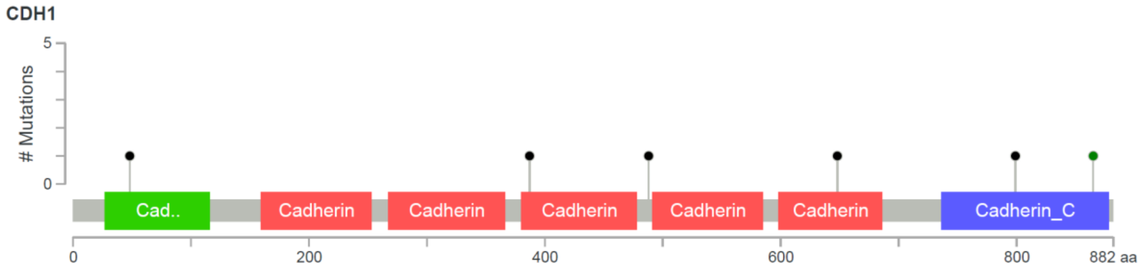


Figure 2.11: Distribution of protein truncating variants identified in ILC cases. Black lollipops correspond to known pathogenic variants whereas green indicates the novel variant. Pfam domains are also indicated across the *CDH1* gene.

2.8.1.4 **TP53**

Having screened 1,443 ILC cases and 1,611 controls for the *TP53* gene, we identified 9 rare non-synonymous variants along with 1 protein truncating variant. Protein truncating variants are rare in *TP53* gene. Missense variants are usually linked with Li-Fraumeni syndrome. *TP53* has a very high pLI score (pLI=0.91) which constitutes an indicator for low tolerance in LoF mutations. A variant that has been previously reported as pathogenic is c.G818A: p.R273H: rs28934576 was found in an individual diagnosed with ILC at the age of 55, with an affected sister with thyroid cancer. It is a non-synonymous variant in exon 8. Intersecting our data with ClinVar, we identified 3 VUS in cases and 5 in controls. No significant enrichment of VUS has been observed. This is the first study to date to estimate the prevalence of *TP53* germline mutations in a series of unselected lobular breast cancer cases, which as expected is very low (0.07%).

2.8.1.5 **CHEK2**

CHEK2 analysis indicates a non-significant enrichment in both VUS and pathogenic variants in cases but we need larger sample size since some of the variants are relatively common (0.01% < MAF < 1%) and present in the control population. There are seven protein truncating variants in individuals with any form of lobular disease, out of which 3 in ILC cases whereas there was only one in a healthy control, Table 2.19.

Even though previous literature suggested that *CHEK2* variant, I157T, might be more strongly associated with development of lobular breast cancer compared to IDC, our data do not support this hypothesis. This variant was present in 3 controls and 2 cases (*P*=0.99). A non-significant enrichment of rare variants has been observed but no conclusions can be drawn from our analytical approach on *CHEK2* predisposition to breast cancer. This is due to the fact that the effect size differs amongst different *CHEK2* variants.

Suboptimal amplification was also observed in two amplicons at the *CHEK2* gene, Table 7.6, page 179. With regards to the c.1100C deletion variant that has been previously associated with increased risk of developing breast cancer, there were three individuals with ILC that were carriers of this variant. However, Samtools algorithm failed to identify them and call them. Therefore, this variant was not included in the final analysis.

Table 2.19: *CHEK2* variants that have been previously described to be associated with breast cancer development and were present in our cohort.

Status	Sample ID	Nt change	AA change	Variant class
Control	CG00303	c.T470C	p.I157T	Non-synonymous
Control	CG00741	c.T470C	p.I157T	Non-synonymous
Control	CG01482	c.T470C	p.I157T	Non-synonymous
Control	CG00400	c.C1196T	p.S399F	Non-synonymous
Control	CG00417	c.C1196T	p.S399F	Non-synonymous
Control	CG01651	c.1375-2A>G		Splicing
ILC	TG00138	c.A349G	p.R117G	Non-synonymous
ILC	TG01704	c.A349G	p.R117G	Non-synonymous
ILC	TG01829	c.T470C	p.I157T	Non-synonymous
ILC	TG02526	c.T470C	p.I157T	Non-synonymous
ILC	TG01819	c.G715A	p.E239K	Non-synonymous
ILC	TG00217	c.1176delT	p.L392fs	Frameshift
ILC	TG00735	c.1176delT	p.L392fs	Frameshift
ILC	TG01163	c.C1468T	p.R490X	Stop-gain
LCIS/IDC	TG00892	c.A529T	p.K177X	Stop-gain
LCIS/IDC	TG01579	c.1176delT	p.L392fs	Frameshift
LCIS	TG00511	c.G697T	p.E233X	Stop-gain
LCIS	TG00061	c.188_189insC	p.L63fs	Frameshift

2.8.1.6 **PALB2**

In addition to *BRCA2*, *PALB2* also shows a significant enrichment in terms of known pathogenic/ likely pathogenic variants, with 8 being present in ILC cases and 1 in a control sample. A case control analysis that included only pathogenic and likely pathogenic variants, irrespectively of whether they have been described before, indicates a strong effect size for *PALB2* in the context of ILC. There is a 10-fold increase in the prevalence of *PALB2* protein truncating variants in ILC cases compared to controls, Table 2.20.

Table 2.20: Pathogenic variants in ILC cases and controls; OR and *P* correspond to a Fisher's exact test using 1443 cases and 1611 controls.

Carrier controls N (%)	Carrier ILC cases N (%)	OR	95% CI	*P*
1 (0.06%)	8 (0.6%)	8.98	1.12, 71.85	0.016

The list of individuals with a *PALB2* protein truncating variant is shown in Table 2.21. All those 9 *PALB2* variants are protein truncating and 4 out of the 8 in the cases are novel in the context of publically available databases and are highlighted in bold. Apart from 8 ILC cases with protein truncating variants, there were three cases with LCIS (two of which with concurrent IDC) that also carried a known pathogenic protein truncating variant. The list of individuals with a *PALB2* protein truncating variant is shown in Table 2.21. Out of 1,443 individuals with ILC there were 8 carriers. Another 3 carriers with any form of lobular disease were also identified. The prevalence of these variants is 0.6% in ILC. This can translate to an enrichment of the magnitude of 10 fold compared to controls.

Table 2.21: Details of *PALB2* carriers from the GLACIER study. Novel variants are highlighted in bold.

Sample ID	Nt change	AA change	Class	Pathology	Age
CG01848	c.2052delC	p.P684fs	Frameshift	Control	41
150378	c.C3256T	p.R1086X	Stop-gain	ILC	53
TG01120	c.C3256T	p.R1086X	Stop-gain	ILC	50
TG00941	**c.2748+1G>A**		**Splicing**	**ILC**	**54**
TG01021	c.G2718A	p.W906X	Stop-gain	ILC	59
150477	**c.2488delG**	**p.E830fs**	**Frameshift**	**ILC**	**51**
TG00571	c.1317delG	p.G439fs	Frameshift	ILC	41
TG02368	**c.1172delC**	**p.A391fs**	**Frameshift**	**ILC**	**50**
TG01249	**c.G412T**	**p.E138X**	**Stop-gain**	**ILC**	**54**
TG00239	c.G3113A	p.W1038X	Stop-gain	LCIS/IDC	47
TG00565	c.G2386T	p.G796X,	Stop-gain	LCIS/IDC	60
TG01329	c.G3113A	p.W1038X	Stop-gain	LCIS	55

2.8.2 Identification of putative novel breast cancer predisposition genes

In an attempt to identify novel breast cancer predisposition genes that were either candidates from the phase I exome sequencing study or suggested by literature, we included exons and splicing junctions from 14 genes that showed some suggestive evidence of association.

Due to experimental limitations on the number of amplicons that could be included, four genes were not fully screened and only a small number of selected exons were captured, Table 2.13, page 73.

We conducted gene burden tests to test for enrichment of rare likely pathogenic variants cases over controls. Two different analyses were conducted; the first one included cases with any form of lobular disease whereas the second one included only cases with ILC. The final data set comprised of 1,611 healthy controls and 2,215 cases with lobular disease out of which 1,443 had ILC. The remaining 772 cases included 366 cases of pure LCIS and 406 cases with LCIS and non ILC invasive disease (either IDC or mixed ductal lobular).

The total number of truncating variants across all 14 putative novel genes was 20 for lobular (17 ILC) cases and 17 for controls. The complete list of protein truncating variants is reported in Table 2.22.

Table 2.22: Details of protein truncating identified in ILC cases and controls in the 14 putative novel genes under investigation.

Sample	Age	Pathology	Gene	Class	Exon	Nt change	AA change
CG01957	67	Control	*CTNNA1*	Stop-gain	10	c.C1351T	p.R451X
CG00990	54	Control	*ESR2*	Stop-gain	7	c.C335A	p.S112X
CG00299	92	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG00393	50	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG00730	56	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG00814	53	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG01160	63	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG01340	60	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG01345	51	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG01423	63	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG01454	53	Control	*IDE*	Stop-gain	1	c.13delC	p.L5X
CG00850	50	Control	*MME*	Splicing	5	c.439+1G>A	x
CG00940	46	Control	*MME*	Frameshift	6	c.466delC	p.P156fs
CG01712	56	Control	*MME*	Frameshift	6	c.466delC	p.P156fs
CG00552	41	Control	*MME*	Frameshift	12	c.1186delA	p.K396fs
CG00695	47	Control	*SRA1*	Splicing	5	c.391-2A>G	x
CG00912	55	Control	*SRA1*	Frameshift	3	c.236dupC	p.P79fs
TG02438	43	ILC	*DCLRE1B*	Stop-gain	3	c.C508T	p.R170X
TG00025	48	ILC	*ESR2*	Stop-gain	7	c.C335A	p.S112X
TG02470	47	ILC	*ESR2*	Stop-gain	7	c.C335A	p.S112X
TG01952	52	ILC	*ESR2*	Splicing	8	c.362+2T>C	x
TG00978	35	ILC	*GOLGB1*	Frameshift	20	c.9476_9477insC	p.E3159fs
TG01172	58	ILC	*GOLGB1*	Frameshift	18	c.9221dupA	p.Q3074fs
TG01819	52	ILC	*IDE*	Stop-gain	2	c.C184T	p.R62X
TG00140	48	ILC	*IDE*	frameshift	1	c.13delC	p.L5X
TG00386	34	ILC	*IDE*	frameshift	1	c.13delC	p.L5X
TG01824	42	ILC	*IDE*	frameshift	1	c.13delC	p.L5X
TG02215	53	ILC	*MME*	Stop-gain	2	c.C11G	p.S4X
TG02012	56	ILC	*MME*	Frameshift	6	c.466delC	p.P156fs
TG00591	50	ILC	*MME*	Frameshift	9	c.763dupT	p.R254fs
TG01147	51	ILC	*SRA1*	Splicing	5	c.391-2A>G	x
TG01499	46	ILC	*SRA1*	Frameshift	5	c.598dupA	p.R200fs
TG01457	52	ILC	*WDR17*	Frameshift	12	c.1592delA	p.Q531fs
TG01648	50	ILC	*WDR17*	Frameshift	14	c.1892_1896del	p.D631fs
TG01561	57	LCIS/IDC	SRA1	Splicing	5	c.391-2A>G	x
TG01691	37	LCIS	ATRIP	Frameshift	11	c.1561delT	p.L521fs
TG02219	46	LCIS/Mixed Duct/Lob	IDE	Stop-gain	1	c.13delC	p.L5X

An additional 368 non-synonymous variants in controls and 322 in ILC cases were identified. From those, 181 had CADD>20 for controls and 171 for cases, while 15 controls and 22 ILC cases had variants with CADD>30.

With regards to non-frameshift variants, the same *FAM175A* variant was identified in 4 controls and 1 ILC case. Another *FAM175A* variant was identified in one control along with a variant in *PABPN1L*. The same *ESR2* variant was also identified in two individuals with ILC. The complete list of all rare non-frameshift variants identified is reported in Table 2.23.

Table 2.23: Non-frameshift variants identified in the putative novel genes.

Sample	Age	Pathology	Gene	Exon	Nt Change	AA change
CG01524	46	Control	*FAM175A*	9	c.1102_1104del	p.368_368del
CG02086	51	Control	*PABPN1L*	3	c.412_414del	p.138_138del
CG00542	56	Control	*FAM175A*	9	c.826_828del	p.276_276del
CG00603	54	Control	*FAM175A*	9	c.826_828del	p.276_276del
CG00853	48	Control	*FAM175A*	9	c.826_828del	p.276_276del
CG01370	41	Control	*FAM175A*	9	c.826_828del	p.276_276del
TG00022	53	ILC	*ESR2*	9	c.541_543del	p.181_181del
TG01663	47	ILC	*ESR2*	9	c.541_543del	p.181_181del
TG02106	53	ILC	*FAM175A*	9	c.826_828del	p.276_276del

None of the genes under investigation reached the Bonferroni corrected threshold when included only protein truncating variants or when combining protein truncating variants along with non-synonymous variants with CADD>30 or with CADD>20.

One gene that shows some suggestive evidence of association is *DCLRE1B*. There is a borderline significance of rare non-synonymous variants in *DCLRE1B* with 16 variants in lobular cases and only 3 in controls (*P*=0.02, OR=3.9 95%CI 1.1-13.4). This analysis included non-synonymous variants with CADD>20 along with protein truncating variants. However, further screening projects including larger sample size are required in order to validate or reject our initial suggestive association.

Three tables show the summary results of the gene based analysis for the 14 putative novel breast cancer predisposition genes. The analysis was conducted using either i) purely protein truncating variants (Table 2.24), ii) including truncating and non-synonymous with CADD>30 (Table 2.25), and finally iii) truncating and non-synonymous with CADD>20 (Table 2.26). Two different case sets were used for these analyses. The first group includes 1,443 cases diagnosed with ILC ≤ 60 and the second one 2,215 cases diagnosed ≤ 60 with any form of lobular disease, including LCIS with our without any form of invasive breast cancer. The only suggestive evidence of association comes from the analysis including variants that are either protein truncating or non-synonymous with CADD > 20.

DCLRE1B, a gene involved in DNA repair, shows a nominal association with lobular breast cancer. Post-filtering, there were 16 variants in the cases group and 3 in the controls. This leads to a non-significant after multiple testing correction OR=3.9 (95% CI 1.13-13.41) and *P*=0.02, Table 2.26.

Another good biological candidate was the *ESR2* gene. During the phase II screening of 2,215 lobular cases and 1,611 controls we identified three individuals diagnosed with ILC (age of diagnoses 47, 48, 52) and one healthy control (age = 54) with a protein truncating variant. Due to its biological significance in hormone regulation, the *ESR2* gene was more thoroughly investigated. In a restricted analysis, investigating rare variants with MAF<0.001, we identified 20 individuals with any form of lobular disease being carriers and only 5 control carriers ($OR_{Any\ lobular}$=2.92, *P*=0.025. However, this analysis does not generate enough evidence for association and therefore further analyses incorporating larger data sets should be conducted to assess whether rare variants at the *ESR2* gene can predispose to breast cancer.

Table 2.24: Case control analysis including protein truncating variants.

Gene	Control carriers	ILC carriers	OR	95% CI	*P*	All lobular	OR	95% CI	*P*
ATG2B	0	0	nan	nan	0.999	0	nan	nan	0.999
ATRIP	0	0	nan	nan	0.999	2	inf	-	0.512
CTNNA1	1	0	0	-	0.999	0	0	-	0.421
DCLRE1B	0	1	inf	-	0.472	1	inf	-	0.999
ESR2	1	3	3.35	0.35, 32.28	0.35	4	2.91	0.33, 26.09	0.405
FAM175A	0	0	nan	nan	0.999	0	nan	nan	0.999
FMO2	0	0	nan	nan	0.999	0	nan	nan	0.999
GOLGB1	0	2	inf	-	0.223	2	inf	-	0.512
IDE	9	4	0.49	0.15, 1.61	0.275	7	0.56	0.21, 1.52	0.312
MME	4	3	0.84	0.19, 3.75	0.999	4	0.73	0.18, 2.91	0.728
PABPN1L	0	0	nan	nan	0.999	0	nan	nan	0.999
SLC15A2	0	0	nan	nan	0.999	0	nan	nan	0.999
SRA1	2	2	1.12	0.16, 7.94	0.999	3	1.09	0.18, 6.54	0.999
WDR17	0	2	inf	-	0.223	3	inf	-	0.269

Table 2.25: Case control analysis including protein truncating and non-synonymous (CADD>30) variants.

Gene	Control carriers	ILC carriers	OR	95% CI	*P*	All lobular	OR	95% CI	*P*
ATG2B	2	5	2.8	0.54, 14.44	0.266	6	2.19	0.44, 10.84	0.480
ATRIP	0	0	nan	nan	0.999	1	inf	-	0.999
CTNNA1	3	0	0	-	0.252	3	0.73	0.15, 3.61	0.701
DCLRE1B	0	3	inf	-	0.105	4	inf	-	0.144
ESR2	1	3	3.35	0.35, 32.28	0.35	4	2.91	0.33, 26.09	0.405
FAM175A	0	1	inf	-	0.472	1	inf	-	0.999
FMO2	0	0	nan	nan	0.999	0	nan	nan	0.999
GOLGB1	0	2	inf	-	0.223	2	inf	-	0.512
IDE	11	7	0.71	0.27, 1.83	0.637	9	0.59	0.25, 1.44	0.262
MME	7	8	1.28	0.46, 3.53	0.797	11	1.14	0.44, 2.96	0.817
PABPN1L	1	3	3.35	0.35, 32.28	0.35	3	2.18	0.23, 21.01	0.643
SLC15A2	0	1	inf	-	0.472	1	inf	-	0.999
SRA1	2	2	1.12	0.16, 7.94	0.999	3	1.09	0.18, 6.54	0.999
WDR17	5	4	0.89	0.24, 3.33	0.999	6	0.87	0.27, 2.86	0.999

Table 2.26: Case control analysis including protein truncating and non-synonymous (CADD>20) variants.

Gene	Control carriers	ILC carriers	OR	95% CI	*P*	All lobular	OR	95% CI	*P*
ATG2B	3	6	2.24	0.56, 8.97	0.322	7	1.7	0.44, 6.58	0.534
ATRIP	37	30	0.9	0.56, 1.47	0.712	45	0.88	0.57, 1.37	0.574
CTNNA1	35	28	0.89	0.54, 1.47	0.703	44	0.91	0.58, 1.43	0.730
DCLRE1B	3	8	2.99	0.79, 11.29	0.129	16	3.9	1.13, 13.41	**0.020**
ESR2	22	22	1.12	0.62, 2.03	0.762	32	1.06	0.61, 1.83	0.890
FAM175A	0	2	inf	-	0.223	2	inf	-	0.512
FMO2	0	0	nan	nan	0.999	0	nan	nan	0.999
GOLGB1	5	6	1.34	0.41, 4.40	0.765	10	1.46	0.50, 4.27	0.605
IDE	19	11	0.64	0.31, 1.36	0.274	18	0.69	0.36, 1.31	0.315
MME	27	24	0.99	0.57, 1.73	0.999	37	1	0.60, 1.64	0.999
PABPN1L	14	15	1.2	0.58, 2.49	0.71	21	1.09	0.55, 2.15	0.865
SLC15A2	0	3	inf	-	0.105	5	inf	-	0.078
SRA1	21	19	1.01	0.54, 1.89	0.999	33	1.15	0.66, 1.99	0.679
WDR17	12	14	1.31	0.60, 2.83	0.557	21	1.28	0.63, 2.60	0.597

2.9 Discussion

We have shown for the first time that *CDH1* mutations predispose to LCIS, with 12.5% (1/8 cases) of pure LCIS (no invasion) and 9% (4/45) of bilateral LCIS with or without invasive disease of any subtype having *CDH1* mutations. Interestingly, none of the cases with bilateral LCIS and non-lobular invasive disease (11 cases) had *CDH1* mutations, suggesting that the presence of a germline *CDH1* mutation in bilateral LCIS predisposes to the development of ILC rather than IDC. In the study by Rahman *et al*. only 17 cases (26%) of the 65 cases of LCIS screened had bilateral disease, Table 2.4, page 57, which may explain why no *CDH1* mutations were detected.

Our findings suggest that *CDH1* testing should be offered to individuals with bilateral lobular lesions under the age of 50, enabling us to identify *CDH1* mutation carriers to whom MRI screening and endoscopic surveillance for diffuse gastric cancer will be beneficial. To conclude, there is evidence that *CDH1* mutations are associated with early-onset (<50 years) bilateral disease which however is not observed in the remaining mutations found on the other genes where there seems to be a broader distribution of age. One limitation of this analysis is that we do not have information on family history and existence of bilateral disease on the TCGA samples.

One of the limitations of the *CDH1* study is that family history is self-reported by the index case and we are therefore unable to ascertain what subtype of breast or gastric cancer family members suffered from. It does not appear that there is an excess of personal or family history of gastric cancer in the GLACIER cohort comparing it with the UK statistics produced by the Cancer Research UK, but we cannot be certain that diffuse gastric cancer is not overrepresented in our bilateral cases.

The majority of breast cancer predisposition genes confer high risk to disease showing an excess of truncating variants that will lead to loss of function. The main exception in that rule is the *TP53* gene, where several missense variants also confer high risk. With the exception of *TP53,* researchers struggle to refine the effect of missense variants and the vast majority of them are still considered as VUS. There are certain missense variants in *BRCA1* and *BRCA2* that confer high risk of breast cancer, but the vast majority do not [218, 219]. Our data support this statement. We have identified a significant excess of VUS in *BRCA1* and pathogenic variants *BRCA2* for individuals with ILC. For these genes, algorithms based on conservation, pedigree data, and analysis of tumour subtype can be used to predict the pathogenicity of some variants [220, 221]. A similar approach can be followed on other breast cancer susceptibility genes such as *PALB2, ATM,* and *CHEK2*. Missense variants that lie in important functional domains of genes or are evolutionary conserved are more likely to confer risk towards disease [122]. However, accurate risk estimations associated with the majority of the missense variants still remain to be obtained, even for very well studied and thoroughly sequenced genes such as *BRCA1* and *BRCA2*. In the future, and with advances in sequencing technology, it will be possible to identify and stratify the effect of individual variants and confine individualised risks per variant. One such occasion could be the rs35187787 variant, a non-synonymous *CDH1* variant for which we failed to identify any correlation with the lobular phenotype. Our study concluded that rs35187787 is not associated with ILC. This missense *CDH1* variant has been previously described as pathogenic in a family of diffuse gastric and colorectal cancer [45].

However, there is now evidence that it is unlikely to be associated with breast cancer development. Its frequency in the European population is relatively common for a variant with moderate-high penetrance. If this variant conferred susceptibility to cancer, its penetrance would be low since its frequency in the general population is relatively high, reaching almost 1% in the Europeans. In our pilot study of 7 individuals, we identified one carrier. However, we were unable to identify an association when we screened a larger sample size, where the frequency of the variant was very similar between the cases and control groups (ranging between 0.58%-0.9%). Mutations in the *CDH1* gene can lead to non-malignant carcinogenesis such as LCIS and lobular neoplasia. These pathological features are randomly detected in core biopsies and they might exist in a higher proportion of the population than expected. Identifying genes and variants that predispose to these forms of disease can be challenging because a proportion of the control population can actually be pre-symptomatic or at high risk. This could partially explain the fact that rs35187787 exists in controls at similar proportions as in breast cancer cases in general [206].

In the context of the putative novel gene identification, there are some limitations in the study design. The absence of family history or age of diagnosis of their other disorder for the controls constitutes another limitation of this study. Furthermore, the presence of common susceptibility genes between different diseases could impact our ability to detect a true association or enrichment in the cases versus the controls, since a portion of the controls could also be carriers of certain mutations that predispose both to breast cancer and another syndrome or rare disorder.

During the phase I exome sequencing study, we identified eight genes not known to be associated with breast cancer that contained rare protein truncating or non-synonymous variants that were predicted to be deleterious in cases of ILC and were either absent or present at low frequency in the controls. The number of variants identified during this study along with some key characteristics of the genes, are highlighted in Table 2.12. We therefore designed a phase II study where we investigated these genes in a larger cohort. Along with those putative novel genes, we included known breast cancer predisposition genes in our custom panel comprising 20 genes.

Our targeted sequencing project revealed that the prevalence of *BRCA2* mutations is relatively high in cases with lobular carcinoma and in particular ILC cases. More than 3% of ILC cases that were diagnosed with the disease before the age of 40 are carriers of a *BRCA2* pathogenic variant.

We have also identified an enrichment of VUS in the *BRCA1* gene in ILC cases compared to the control population. This enrichment is more profound in individuals with early onset ILC. This is the first study to date that shows an association of *BRCA1* variants and lobular breast cancer. Further breakdown and classification of these variants using either larger screening studies, *in silico*, or *in vitro* studies could elucidate their exact role in breast cancer development, and accurately assess their effect size. The clinical utility of diagnostic genetic tests can increase by decreasing the number of VUS and by being able to assess risks conferred by specific variants with higher precision.

Significant differences were observed for 4 out of 6 genes with *TP53* not reaching significance due to the very small number of carriers, and *CHEK2* due to the presence of pathogenic variants in the control group. However, with regards to *CHEK2*, it has been observed that different variants can have different effect size towards breast cancer development, and therefore an alternative approach where variants could be stratified further according to their class or domain might be more appropriate. Previous literature suggests a stronger link between *CHEK2* and lobular disease in comparison to the more common ductal breast cancer. However, our data does not support this statement with an overall OR=1.5 (95% CI 0.52-4.31) for all pathogenic variants across the *CHEK2* gene. However, the gene based approach might not be the most appropriate for *CHEK2* since there is evidence for differential effect size among different variants with non-synonymous variants usually having a smaller effect size. The proportions of non-synonymous and truncating variants differ between cases and controls. The ratio (R) of non-synonymous to truncating variants R=6 for controls and R=1.66 for cases which indicates an enrichment in more penetrant variants in cases. Nevertheless, these results are not statistically significant. A larger sample size would be required to be able to identify further variants and stratify them based on their properties to draw robust conclusions on their estimated effect.

One individual was a carrier of a *BRCA1* mutation. A further missense pathogenic *TP53* variant has been identified in an individual with ILC. With regards to *BRCA1*, we identified an excess of variants previously classified as VUS in ILC cases over controls. This observation was more profound in early onset cases with approximately 7% of cases carrying a VUS. This finding is novel and requires further validation in larger cohorts to establish the underpinning mechanisms of action of these variants. The fact that the effect size gets stronger with younger age of diagnosis is another finding that suggests that there is a true signal in the observed enrichment in VUS.

A group in Poland conducted a *CHEK2* screening study in a Polish population including 211 ILC cases and found 29 (13.7%) pathogenic variants out of which 4 (1.9%) were protein truncating [222]. The same group, a year later published another analysis on *CHEK2* where they found 111 (11.5%) mutations in ILCs out of which 23 (2.4%) were protein truncating in a total of 960 ILC cases [223]. Researchers from the same consortium, identified that *PALB2* mutations were present in 0.5% of their ILC Polish cases (7/1306) [224]. This comes in concordance with our findings where 0.5% of our unselected population with ILC are carriers of a protein truncating *PALB2* variant.

A recent study reported a borderline association of germline *PALB2* mutations with LCIS, having identified an enrichment of LCIS in their population of *PALB2* positive samples. However, we identified only one individual with pure LCIS carrying a *PALB2* protein truncating variant, rs180177132, in our cohort of 366 LCIS cases [225].

One limitation of the targeted sequencing study with regards to known genes is the suboptimal coverage of specific exons in *CDH1* and *CHEK2* genes. It is likely that pathogenic mutations exist in these genes and we do not have the means to identify them.

In an attempt to identify novel genes predisposing to lobular breast cancer, we conducted a two-phased gene based rare variant association study. Due to the expected genetic heterogeneity and the small sample size of the phase I study, we failed to identify strong candidates during the initial screening. However, having found some evidence of enrichment in 8 genes, we incorporated them in a phase II targeted sequencing project along with another 6 candidates and 6 known breast cancer predisposition genes.

After having screened the whole GLACIER cohort of 2,215 lobular cases and 1,611 controls and conducted a case control study for these genes, there were two genes that warranted further investigation; *DCLRE1B* and *ESR2.* A common non-synonymous variant in *DCLRE1B* has been previously associated with breast cancer [154]. This gene encodes for DNA cross-link repair 1B protein. *DCLRE1B* is an evolutionarily conserved gene involved in repair of inter-strand cross-links. Its role in genome stability constitutes that gene as a good biological candidate [226]. Having found a nominally significant excess of rare variants in the gene might elucidate its role on breast cancer development. Further studies should validate whether there is a true signal in rare protein truncating variants or missense variants with high likelihood of pathogenicity. Due to its function, alterations caused by rare variants could be related to breast cancer development in a similar manner to *BRCA1* and *BRCA2*. The second gene with some suggestive evidence, *ESR2*, is also biologically relevant. Several association studies have been conducted to identify the relationship between common SNPs at the *ESR2* locus and the risk of developing breast cancer [227]. Overall, results have been inconclusive with some studies showing nominal associations in specific populations that fail to replicate. However, rare variants at the *ESR2* gene could be implicated with specific subtypes of breast cancer in certain populations [228]. In our study of more than 2,200 lobular cases and 1,600 controls, we have shown some evidence of enrichment in rare variants but this enrichment is not significant after correcting form multiple testing. Further studies should validate whether there is a contribution of rare *ESR2* variants towards lobular breast cancer development.

We are in no position to draw conclusions for any of the two putative novel breast cancer predisposition genes. If these genes are validated as breast cancer predisposition genes, they could be included in targeted sequencing panels that are broadly used in diagnostic laboratories in clinical practise.

The outcome of the novel gene discovery study was not as fruitful as initially expected and there are a few possible reasons on why that might be. A key consideration is whether there would be that many variants in the selected genes to reach a CAF of 0.1% or 0.5%. Taking into account the genetic heterogeneity of breast cancer and the extremely low frequency of pathogenic variants that have already been associated with breast cancer, it becomes apparent that we were underpowered for the phase 1 to identify genes that can confer susceptibility to breast cancer with this case control study. The number of cases (144) is very low for such a study. Investigating the proportions of rare truncating variants in the putative novel genes, highlights the fact that there is no significant contribution of any of these genes towards lobular breast cancer development since there is no gene with more than 4 protein truncating variants and no gene with more than 9 combined protein truncating and non-synonymous with CADD>30 in ILC cases. A total of 9 variants in cases would indicate a possible signal but the presence of similar proportions of variants in the control group, diminishes this possibility.

To conclude, we have shown evidence for involvement with ILC of all 6 known breast cancer predisposition genes under investigation in our study. The prevalence of *BRCA2* pathogenic variants is higher than *CDH1*. The latter is found mutated in a similar portion as the *PALB2*.

However, *CDH1* mutations are more prevalent amongst bilateral cases. We have shown a significant excess of *BRCA1* VUS in ILC compared to controls, as well as the presence of 1 pathogenic *TP53* variants in an individual with ILC. *CHEK2* protein truncating variants are not as common in our study. We have also shown some evidence of involvement of two novel genes but these findings require further validation.

Chapter 3 Common variants predisposing to lobular carcinoma

3.1 Introduction

Several studies have been conducted over the last ten years in order to identify genetic markers associated with breast cancer [153-174]. However, most studies have been treating breast cancer as one disease with the exception of ER stratification. It has been shown that breast cancer is extremely heterogeneous and therefore we hypothesise that by focusing on specific morphological breast cancer subtypes we are more likely to increase power to detect association by increasing the genetic homogeneity of the sample set.

GWAS in breast cancer have identified loci that predispose to invasive breast cancer in general, or specifically to ER positive or ER negative disease GWAS [153-174]. However, no previous study has focused specifically on lobular carcinomas. Only one common single nucleotide polymorphism (SNP; rs11249433 at 1p11.2) has been shown to be more strongly associated with lobular than ductal histology [229]. This locus has been recently investigated thoroughly in a fine-mapping study from BCAC incorporating data from more than 90,000 individuals [230]. After imputation the same variant still showed the strongest association, and no significant eQTLs were observed. Using an *in silico* analysis, utilising both UCSC Genome Browser and HaploReg v3 to determine altered regulatory motifs using ENCODE data, Horne *et al* observed that the variant is located in an enhancer/promoter region. For the remaining SNPs predisposing to ER positive tumours, it is unclear whether the studies have lacked statistical power to identify differential associations by histology, or whether associations tend to be non-differential by morphology after accounting for ER status. We therefore screened individuals with ILC and or LCIS on the iCOGS genotyping platform and conducted a case control study using 5,000 controls coming from 4 different studies (SEARCH, BBC, SBCS, UKBGS).

Exogenous hormone use, reproductive behaviour, early menarche and late menopause are well established risk factors for invasive breast cancer [231]. However many of the initial risk factor studies considered breast cancer as one disease and did not study the different subtypes as defined by ER and Her2 expression or by morphological. They were therefore biased towards the more common ductal ER positive cancers.

More recent studies have started to look at risk factors by breast cancer subtype and found some differences. For example, nulliparity is most strongly associated with risk of ER positive breast cancer (hazard ratio, HR=1.31, 95% confidence interval,CI,:1.23-1.39); whereas late age at first birth is most strongly associated with risk of ER-/PR-/HER2+ disease (HR = 1.83, 95% CI: 1.31-2.56) [232]. Oral contraceptive use has been shown to be associated with a 2.5-fold increased risk for triple-negative breast cancer (95% confidence interval, 1.4-4.3) and no significantly increased risk for non-triple-negative breast cancer (*P*-het=0.008) in women under 40 years of age [19]. Lobular cancers show stronger associations with the use of hormone replacement therapy (HRT) than IDC, [233] and their incidence follows a similar temporal pattern as the use of combined HRT [7]. A population-based study by Flesch-Janys *et al* conducted in Germany, observed a more than 2-fold higher risk for lobular than for ductal cancer for current HRT users [234].

In situ disease has been shown to be associated with some of the risk factors common to all breast cancers [235, 236]. Claus *et al* demonstrated that reproductive risk factors, including age at menarche, age at first birth, parity and age at menopause showed very similar associations with DCIS and invasive ductal cancer suggesting that most risk factors affect the risk of invasive ductal cancer primarily through their effects on the risk of DCIS. Reeves *et al*, 2011 [237], showed that combined HRT was associated more strongly with IDC then DCIS and the opposite was found in the estrogen only type. There is no evidence that reproductive history contributes to progression of *in situ* to invasive disease [238].

There is relatively little reported on risk factors associated with LCIS. This is mainly due to the difficulty of getting a large enough study population of LCIS, as it is commonly only discovered as an incidental finding. Reeves *et al* [13] reported on 86 cases of LCIS and showed that compared to never users of hormone therapy the relative risk for LCIS was higher than that for DCIS (2.82. 95CI 1.72-4.63 and 1.56, 95 CI 1.38-1.75 respectively). Claus *et al* [235] showed that in 123 cases of LCIS the risk factors associated with LCIS and DCIS were similar to those associated with invasive breast cancer. In this study we intended to explore the hormonal and reproductive risk factors associated with LCIS and ILC. Since we already have genetic data from a large scale genotyping project for the same individuals, we also assessed potential gene-environment (GxE) interactions between HRT use and loci that have been previously shown to be associated with lobular breast cancer.

3.2 Case control association study

The aim of this study was to identify new breast cancer susceptibility loci specific to lobular carcinoma, and to evaluate the heterogeneity of associations of known loci by morphology. This involved pooling genotyping data from over 6,000 cases of lobular carcinoma (ILC and/or LCIS) and over 34,000 controls genotyped using the iCOGS chip, a custom SNP array that comprises 211,155 SNPs enriched at predisposition loci for breast and other cancers. The analyses that were conducted during this PhD included a total of 2,527 lobular cases from the GLACIER study along with 5,000 controls.

3.2.1 Methods

Cases and controls originate from GLACIER and 34 studies forming part of the BCAC included in the COGS Project [159]. The GLACIER study recruited a total of 2,539 cases: 2,167 were identified from local pathology reports in 97 UK hospitals, 346 cases were identified through the British Breast Cancer Study (BBCS) using UK Cancer Registry data and 26 cases from the Royal Marsden Breast Tissue Bank. BCAC studies recruited all types of breast cancer.

All these cases were genotyped with the iCOGS chip and compared to 5,000 UK controls selected from four UK studies participating in BCAC and already typed on the iCOGS chip. Controls were randomly selected prior to analysis so that each of these UK studies, including GLACIER, had a case-control ratio of at least 1:2, Table 3.1.

Table 3.1: Dissemination of control samples allocated to GLACIER or BCAC analysis.

UK study	Nº of lobular cases	Total Nº of controls	Source of Controls	Nº of controls selected for BCAC analysis	Nº of controls selected for GLACIER analysis
BBCS British Breast Cancer Study	83	1,397	A friend or non-blood relative of cases, recruited from throughout UK	166	**1,231**
SBCS Sheffield Breast Cancer Study	72	848	Unselected women attending Sheffield Mammography Screening Service with no evidence of a breast lesion	144	**704**
UKBGS Breakthrough Generations Study	50	470	Women from throughout the UK who had not had breast cancer or in situ disease before entry into the cohort study	100	**370**
SEARCH Study of Epidemiology & Risk Factors in Cancer Heredity	1,234	8,069	(a) from the EPIC-Norfolk cohort study, (b) women attending GP practices, matched to cases by age and geographic region (East Anglia)	5,374	**2,695**
Total	**1,439**	**10,784**		**5,784**	**5,000**

Pathological information in BCAC was collected by the studies individually but combined and checked through standardized data control in a central database. A total of 4,152 ILC and 89 LCIS cases were identified by the central BCAC pathology database.

This study includes only cases of pure LCIS or ILC with or without LCIS. Cases of LCIS with IDC or mixed lobular and ductal carcinoma in GLACIER were excluded in order to perform meta-analyses with the BCAC studies which do not have information on the presence or absence of LCIS associated with an invasive cancer. Germline DNA extracted from peripheral blood was used for this project. All GLACIER samples were genotyped on the iCOGS custom Illumina iSelect platform as part of my PhD, section 7.2.2. The remaining cases and controls were genotyped as part of the COGS project by collaborators, as described in detail elsewhere [154]. The GLACIER cases were analysed using the same QC criteria as the COGS project. Briefly, genotypes were called using Illuminas proprietary GenCall algorithm and 10,000 SNPs were manually inspected to verify the algorithm calling. Individuals were excluded if genotypically not female, had overall call rate <95% or were ethnic outliers (248 cases) as identified by PCA, combining the genotyping data with the three Hapmap2 populations. SNPs with a Gencall rate of < 0.25, call rate <95% (call rate <99% if MAF <0.1) and P-$_{\mathrm{HWE}}$<10^{-7} or evidence of poor clustering on inspection of cluster plots were excluded. All SNPs with MAF <0.01 were excluded from this analysis.

A cryptic relatedness analysis of the GLACIER case control data set was performed using 46,918 uncorrelated SNPs and there was no evidence of any duplicates. The same analysis showed no evidence of overlap between the GLACIER samples and the BCAC samples. For each SNP, a per-allele OR was calculated by logistic regression, including the first five principal components (PCs) as covariates, using plink (http://pngu.mgh.harvard.edu/purcell/plink/), section 7.4.1. Genotyping and analysis of BCAC studies is described in detail elsewhere [154], All analyses were performed in subjects of European ancestry (determined by PCA). The meta-analysis was performed by other members of the BCAC. In brief, case-control OR for ILC or LCIS cases vs controls from BCAC and GLACIER were combined using inverse variance-weighted fixed-effects meta-analysis, as implemented in METAL [239]. Case-only analyses were also carried out to compare genotype frequencies for ILC vs LCIS and potentially identify invasive or *in situ* specific associations.

For GLACIER cases and controls, PCA was carried out on a subset of 46,918 uncorrelated SNPs and used to exclude individuals or groups distinct from the main cluster using the first PCs, Figure 3.1. Following removal of outliers (166 cases and 245 controls), the PCA was repeated and the first five PCs included as covariates in the analysis, Figure 3.2.

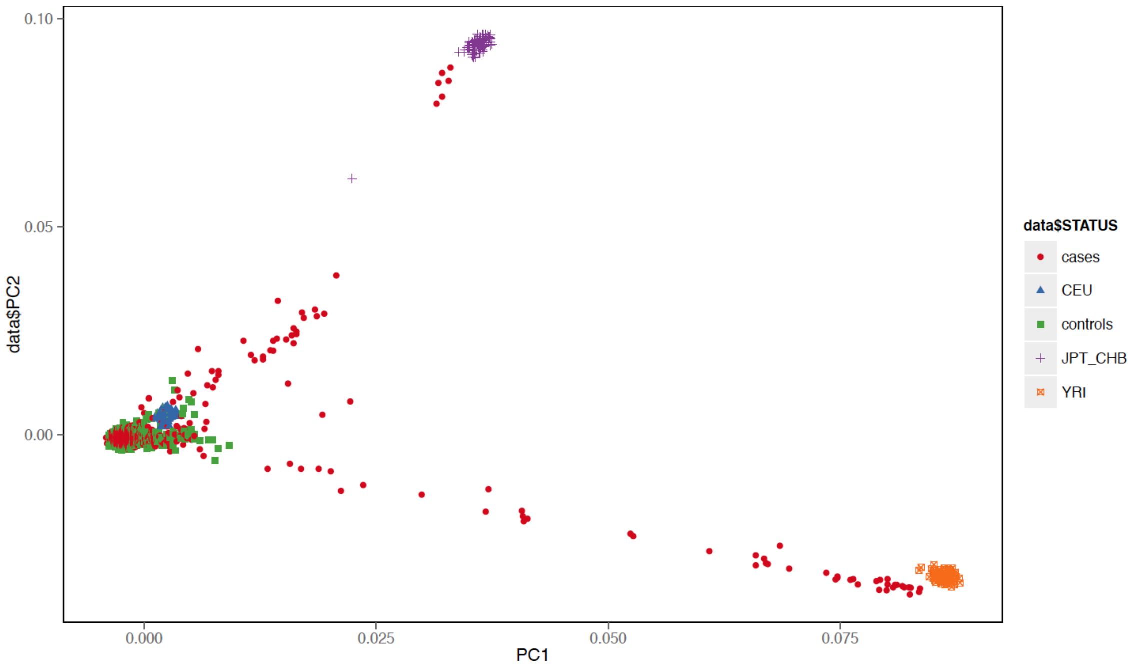


Figure 3.1: Ethnic differences on samples genotyped on iCOGS platform defined by PCA.

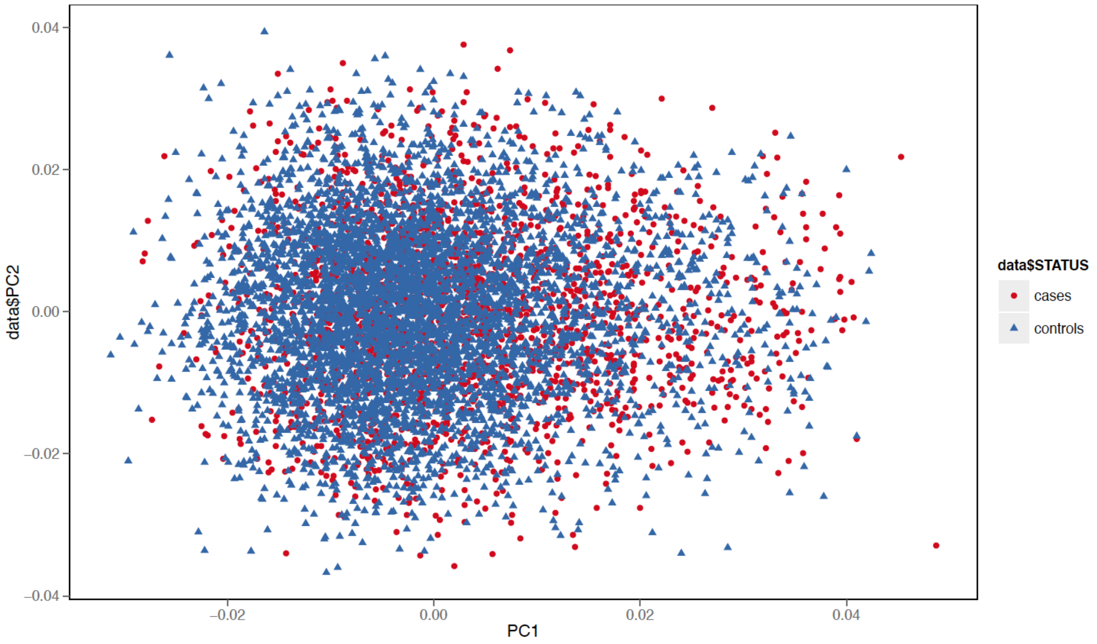


Figure 3.2: Principal component 1 vs Principal component 2 for samples genotyped on iCOGS after removal of ethnic outliers.

The adequacy of the case-control matching was evaluated using quantile-quantile plots of test statistics and the inflation factor (λ) calculated using only 37,544 uncorrelated SNPs that were not selected by BCAC and were not within one of the four common fine-mapping regions, to minimize selection for SNPs associated with breast cancer, Figure 3.3. As the majority of the SNPs on the iCOGs array are associated with breast, ovarian or prostate cancer, the SNPs selected for this analysis were taken from the set of prostate cancer SNPs, with the assumption that these SNPs are more likely to be representative of common SNPs in terms of population structure in our study. After excluding individuals based on genotyping quality and non-European ancestry, data for the GLACIER study available for analyses included 1,782 cases (1,470 ILC (with or without LCIS), 312 pure LCIS) and 4,755 controls.

A further 518 cases (482 ILC, 36 LCIS) and 1,465 controls were analysed as part of a phase II study. Controls were recruited through the GLACIER study, but were not genotyped in phase I on the iCOGS chip to reduce costs. Cases came from the following studies: 232 cases from GLACIER, 176 from BBCS, 71 from DietCompLyf [239], 39 from Kings Health Partners Cancer Biobank (KHP-CB). All cases were white West European, apart from the 39 samples from the KHP-CB where there were no associated ethnicity data. These samples were genotyped at LGC Genomics.

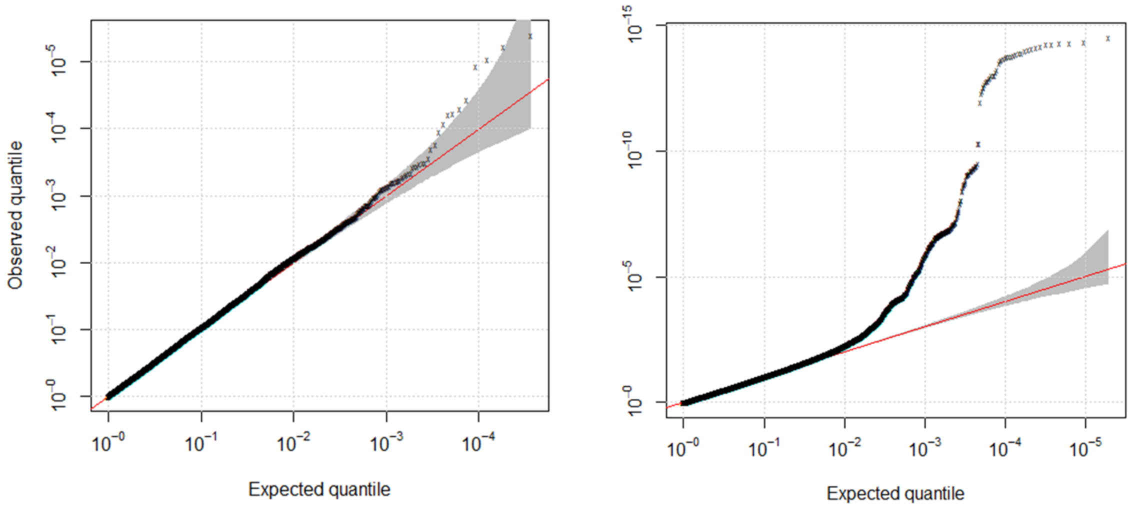


Figure 3.3: QQ-plots for GLACIER data set based on 37,544 uncorrelated SNPs not selected on the basis of breast cancer (left), and all SNPs in data set (right).

In a phase I analysis, we evaluated associations between SNPs on the iCOGS chip and risk of ILC and LCIS using 1,782 lobular cases (1,470 ILC with or without LCIS, 312 pure LCIS) from GLACIER, and 4,755 UK controls from BCAC. There was little evidence for systematic inflation of the test statistics, based on 37,544 uncorrelated SNPs that had not been selected on the basis of breast cancer risk (λ=1.04, Figure 3.3). Data were combined by meta-analysis with a further 4,241 cases (4,152 ILC, 89 LCIS) and 29,519 controls of European ancestry, derived from 34 studies in BCAC, and previously typed on the iCOGS chip. The final meta-analysis, conducted by BCAC collaborators, incorporated data from a total of 6,023 cases (5,622 ILC, 401 LCIS) and 34,271 controls with genotypes on 199,961 iCOGS SNPs (after quality control exclusions and MAF >0.01).

3.2.2 Novel breast cancer predisposition loci

All SNPs reaching genome-wide significance (*P*<5x10^{-8}) in the meta-analysis were correlated with one of the known breast cancer predisposition loci (Figure 3.4). In an attempt to identify novel loci predisposing to lobular carcinoma, we selected 6 SNPs (rs11977670, rs2121783, rs2747652, rs3909680, rs9948182, rs7034265) that were not correlated (r^{2}<0.25) with known loci and that showed the best evidence of association (*P* between 5x10^{-8} and 5x10^{-5}) in the overall lobular case-control analysis (ILC and LCIS). This group of SNPs was genotyped in a phase II study comprised of 516 European cases (481 ILC, 35 LCIS) and 1,467 ethnically matched controls. The power we had to detect genome wide significance of variants with MAF=0.2 and an effect size of OR=1.2 was 97%.

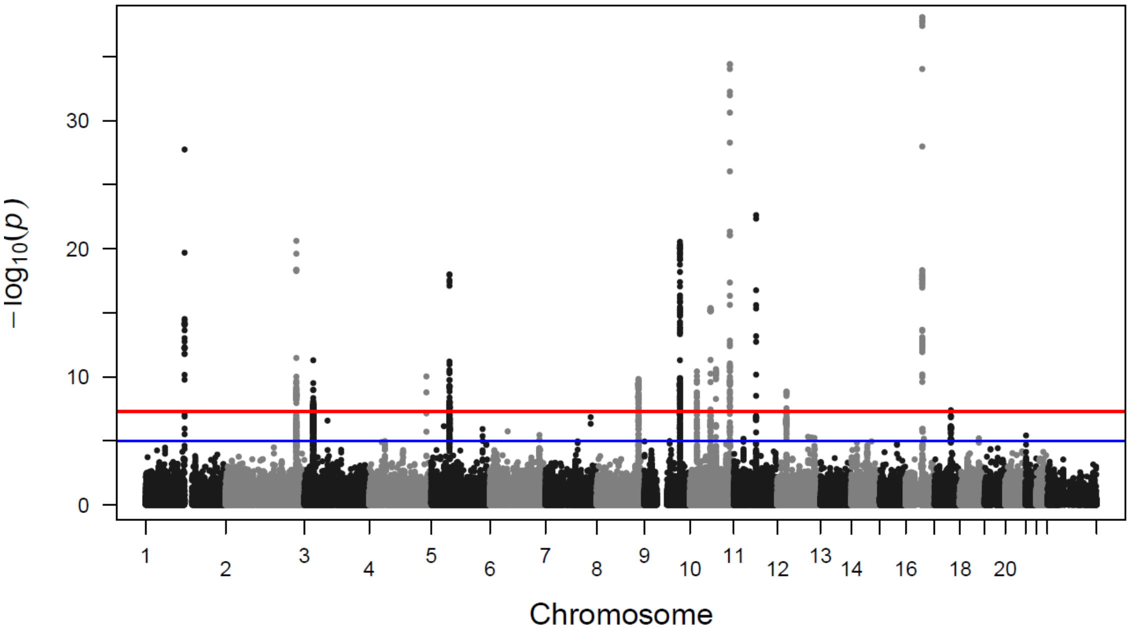


Figure 3.4: Manhattan plot showing results from meta-analysis of 5,622 ILC cases and 34,243 controls.

One of the six SNPs, rs11977670 at 7q34, reached genome-wide significance in a pooled analysis of phase I and II ILC cases and controls (OR=1.13, 95%CI=1.09-1.18, P=6.0x10^{-10}), Table 3.2, Figure 3.6, Appendix 4. rs11977670 showed a similar association with LCIS (*P*-het for ILC vs LCIS=0.198), and no association with IDC (OR=1.02, 95%CI=1.00-1.05, *P*=0.07; *P*-het $_{\text{ILC vs IDC}}$ =1.3 x10^{-5}), indicating that this is a lobular specific predisposition locus. The risk allele appeared to act in a dominant rather than additive manner: OR$_{\text{AG}}$=1.21, 95%CI=1.14-1.30; OR$_{\text{AA}}$=1.27, 95%CI=1.17-1.38 (GLACIER only analyses: *P*$_{\text{Add}}$=9.8x10^{-5}, *P*$_{\text{Dom}}$=1.2x10^{-5}, *P*$_{\text{Rec}}$=0.07). None of the other 5 SNPs genotyped were associated with lobular breast cancer at a genome-wide significance level.

Table 3.2: Common variants reaching a suggestive significance threshold of P<5x10^{-5} that were screened on phase II.

SNP	RAF	Phase I OR (95% CI)	Phase I *P*	Phase II OR (95% CI)	Phase II *P*	Phase I + II OR (95% CI)	Phase I + II *P*
rs2121783	*0.41*	*1.11 (1.07, 1.16)*	*7.1x10^{-7}*	*1.08 (0.93, 1.25)*	*0.31*	*1.11 (1.07, 1.15)*	*4.5x10^{-7}*
rs2747652	*0.47*	*0.91 (0.87, 0.95)*	*1.4x10^{-5}*	*0.98 (0.85, 1.14)*	*0.83*	*0.92 (0.88, 0.95)*	*2.2x10^{-5}*
rs11977670	*0.43*	*1.12 (1.07, 1.16)*	*1.4x10^{-7}*	*1.38 (1.19, 1.60)*	*2.9x10^{-5}*	*1.13 (1.09, 1.18)*	*6.1x10^{-10}*
rs3909680	*0.43*	*1.10 (1.05, 1.14)*	*1.0x10^{-5}*	*1.05 (0.91, 1.22)*	*0.51*	*1.09 (1.05, 1.14)*	*9.9x10^{-6}*
rs9948182	*0.34*	*0.90 (0.86, 0.94)*	*3.7x10^{-6}*	*1.02 (0.88, 1.18)*	*0.83*	*0.91 (0.87, 0.95)*	*1.2x10^{-5}*
rs7034265	*0.19*	*0.90 (0.85, 0.95)*	*7.2x10^{-5}*	*0.93 (0.77, 1.12)*	*0.44*	*0.90 (0.85, 0.95)*	*5.6x10^{-5}*

rs11977670 at 7q34 (position:139942304, GRCh Build 37) is intergenic, Figure 3.5, 65 kb from the nearest gene, *JHDM1D*, a histone demethylase and 500 kb from *BRAF*, a gene frequently somatically mutated in melanoma.

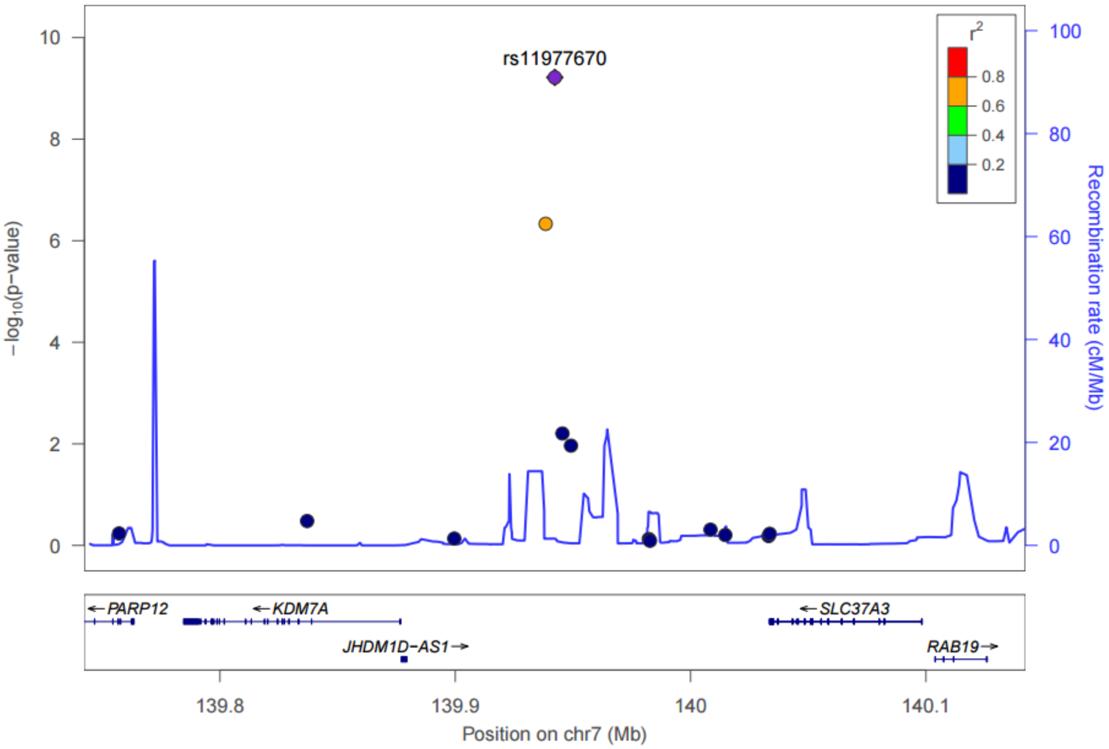


Figure 3.5: Plot of recombination, gene location and other SNPs typed in iCOGs in a 200KB region eitherside of the novel lobular specific locus.

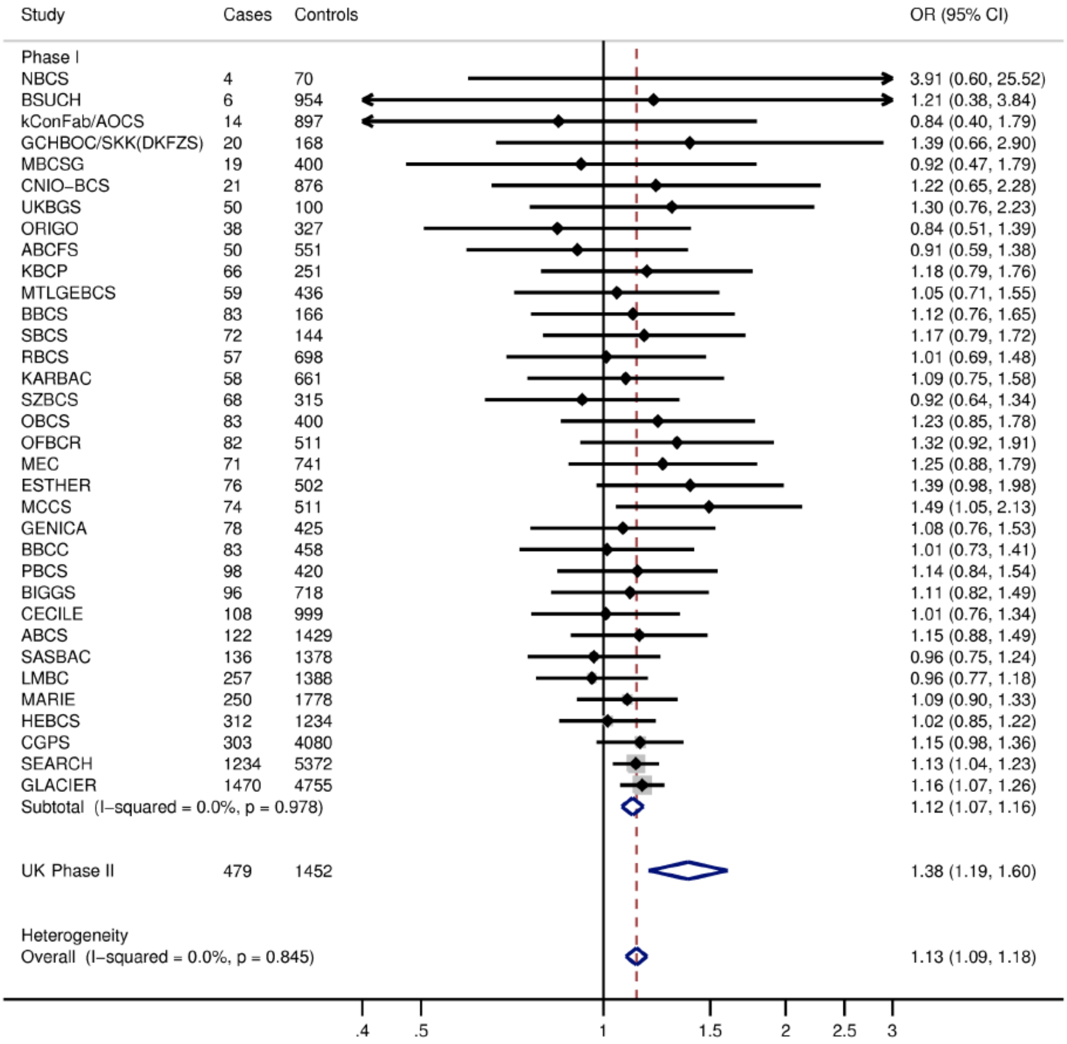


Figure 3.6: Forest plot for rs11977670 indicating effect size of different BCAC studies along with GLACIER and phase II study.

3.2.3 Known breast cancer predisposition loci for ILC

The majority (56) of the 75 known (at the time of analysis) common breast cancer susceptibility loci were associated with ILC at *P*<0.05 with the effect in the same direction as previously reported (31 were significant at the Bonferroni corrected *P*<0.00066), and 14 of these reached genome-wide significance (*P*<5x10^{-8}, Table 3.3). The strongest associations were with SNPs close to *FGFR2* (rs2981579, OR=1.38, P=5.1x10^{-52}), *TOX3* (rs3803662, OR=1.33, *P*=1.1x10^{-35}), at 1p11.2 (rs11249433, OR=1.25, *P*=2.7x10^{-25}) and 11q13.3 (rs554219, OR=1.33, *P*=1.6x10^{-22}). All 14 loci had previously been shown to be associated with ER positive breast cancer and one locus, rs11249433 (1p11.2), with lobular histology in subgroup analysis. Of the remaining 19 SNPs with *P*>0.05, 18 had ORs in the same direction as previously reported for overall breast cancer. Only one of the seven ER negative specific loci on the iCOGS array showed a borderline significant association with ILC (rs12710696, *P*=0.037), but none reached the Bonferroni corrected *P*<0.00066.

Table 3.3: Previously reported SNPs that are associated with ILC (P<5x10^{-8}) in our meta-analysis.

Cytoband	SNP	MAF Controls	OR (95% CI)	*P*
10q26.13	rs2981579	0.4	1.38 (1.32, 1.44)	**5.1x10^{-52}**
16q12.1	rs3803662	0.26	1.33 (1.27, 1.39)	**1.1x10^{-35}**
1p11.2	rs11249433	0.4	1.25 (1.20, 1.30)	2.7x10^{-25}
11q13.3	rs554219	0.12	1.33 (1.26, 1.41)	1.6x10^{-22}
9q31.2	rs865686	0.38	0.83 (0.79, 0.86)	1.0x10^{-17}
2q35	rs13387042	0.49	0.84 (0.80, 0.87)	5.7x10^{-17}
11q13.3	rs75915166	0.06	1.40 (1.29, 1.51	1.2x10^{-16}
11q13.3	rs614367	0.14	1.24 (1.18, 1.31)	7.2x10^{-15}
10q21.2	rs10995190	0.16	0.80 (0.75, 0.85)	1.7x10^{-13}
5q11.2	rs889312	0.28	1.18 (1.13, 1.23)	9.1x10^{-13}
10q22.3	rs704010	0.38	1.14 (1.10, 1.19)	3.7x10^{-10}
10p12.31	rs1243182	0.32	1.14 (1.09, 1.19)	6.1x10^{-9}
4q34.1	rs6828523	0.12	0.82 (0.77, 0.88)	1.6x10^{-8}
8q24.21	rs13281615	0.41	1.13 (1.08, 1.18)	2.1x10^{-8}

A case only analysis including 3,201 ILC cases from GLACIER and BCAC, 15,023 ER positive IDC cases from BCAC, and 29,273 controls from BCAC has yielded some loci that seem to be more strongly associated with one histology over the other, Table 3.4. We have shown evidence for three loci to be more strongly associated with lobular histology (rs11249433, rs2981579, rs10995190).

Table 3.4: Loci that show differential association between ILC and IDC.

Cytoband	SNP	MAF	IDC OR (95%CI)	IDC *P*	ILC OR (95%CI)	ILC *P*	*P*-het
1p11.2	rs11249433	0.4	1.09 (1.06, 1.13)	2.3x10^{-9}	1.28 (1.22, 1.35)	7.2x10^{-20}	2.8x10^{-8}
5p12	rs10941679	0.25	1.17 (1.13, 1.21)	4.1x10^{-21}	1.03 (0.97, 1.10)	0.32	1.6x10^{-4}
10q26.13	rs2981579	0.40	1.31 (1.27, 1.35)	1.3x10^{-71}	1.42 (1.35, 1.50)	3.5x10^{-38}	5.4x10^{-3}
14q24.1	rs2588809	0.16	1.12 (1.08, 1.17)	8.7x10^{-9}	0.99 (0.92, 1.07)	0.87	0.001
10q21.2	rs10995190	0.16	0.87 (0.84, 0.91)	4.2x10^{-11}	0.76 (0.71, 0.83)	1.3x10^{-11}	0.002
8q21.11	rs6472903	0.18	0.89 (0.85, 0.92)	3x10^{-9}	1 (0.93, 1.07)	0.89	0.004
2q31.1	rs1550623	0.16	0.93 (0.89, 0.96)	2.2x10^{-4}	1.01 (0.94, 1.08)	0.84	0.031

As mentioned in Chapter 2, *CDH1* has a key role in lobular breast cancer development and therefore it was not unreasonable to hypothesise that common variants tagging the gene might be associated with the disease. We therefore investigated possible association of 56 variants with MAF>1% that were present in iCOGS platform and spanning a region 184KB surrounding and tagging the *CDH1* gene. None of these 56 variants reached even a nominal association level of P<0.05. We therefore concluded that common variants at the *CDH1* locus are unlikely to be associated with ILC, even though the genotyping platform used did not include a GWAS backbone which could implicate that there might still be variants that are not optimally tagged.

3.2.4 Known breast cancer predisposition loci for LCIS

For the 75 known breast cancer susceptibility loci, case-control analysis for the 401 cases of pure LCIS (without invasive disease) and 24,045 controls, revealed 15 out of 75 SNPs associated with LCIS at P<0.05 (Table 3.5). The strongest associations were for rs865686 (9q31.2, P=2.2x10^{-5}); rs3803662 (*TOX3*, P=1.2x10^{-4}), rs75915166 (11q13.3, P=7.8x10^{-4}) and rs1243482 (*MLLT10*, 10p12.31, P=7.8x10^{-4}) that is partially correlated (r^{2}=0.69) with rs7072776, a recently identified ER positive breast cancer predisposition locus that showed a weaker association with LCIS (OR=1.17, 95%CI =1.00-1.36, P=0.05). Forty-seven of the remaining sixty SNPs at P>0.05 had ORs in the same direction as for ILC. This is greater than what expected by chance (Sign test conducted by BCAC collaborators, P=1.2x10^{-5}) suggesting many of these SNPs predispose to LCIS, but our study did not have enough power to detect these associations with the small sample size. They also conducted a global test in case-only analysis (ILC vs LCIS), which indicated no significant differences in associations of the 75 SNPs between LCIS and ILC (LRT=0.438). However, individual SNP analyses suggested some differences. Two loci showed stronger associations with ILC than pure LCIS: rs2981579, *FGFR2* (P-het=0.02); and rs889312, 5q11.2 (P-het=0.03). Case-only analysis also suggested that two ER negative specific SNPs [157, 168] were more strongly associated with LCIS than ILC: rs6678914, 1q32.1 (P-het=0.0007) and rs17529111, 6q14.1 (P-het=0.04), even though they did not reach the Bonferroni corrected threshold of significance. The remaining SNPs showed no significant heterogeneity between ILC and LCIS, Table 3.5.

Table 3.5: Previously reported SNPs that are associated with LCIS (*P*<0.05) in a pooled meta-analysis including 401 LCIS cases and 24,045 controls.

Cytoband	**SNP**	**MAF**	**OR (95% CI)**	***P***	***P*-het ILC vs LCIS**
1q32.1	rs6678914	0.41	0.77 (0.67, 0.90)	$8.0\text{x}10^{-4}$	**0.0007**
10q26.13	rs2981579	0.4	1.19 (1.03, 1.37)	0.019	**0.04**
6q14.1	rs17529111	0.22	1.25 (1.06, 1.48)	0.009	**0.04**
10q21.2	rs10995190	0.16	0.69 (0.55, 0.87)	0.002	0.1
2p24.1	rs12710696	0.36	1.17 (1.01, 1.35	0.034	0.1
2q14.2	rs4849887	0.1	0.71 (0.54, 0.93)	0.012	0.11
8q21.11	rs6472903	0.18	0.81 (0.66, 0.99)	0.036	0.11
9q31.2	rs865686	0.38	0.72 (0.61, 0.84)	$2.2\text{x}10^{-5}$	0.12
5p15.33	rs10069690	0.26	1.18 (1.01, 1.38)	0.04	0.19
5p15.33	rs7726159	0.34	1.22 (1.05, 1.42)	0.008	0.254
11q13.3	rs614367	0.14	1.32 (1.10, 1.58)	0.003	0.46
10p12.31	rs1243182	0.32	1.29 (1.11, 1.49)	$7.8\text{x}10^{-4}$	0.49
11q13.3	rs75915166	0.06	1.55 (1.20, 2.01)	$7.8\text{x}10^{-4}$	0.54
2q35	rs16857609	0.26	1.25 (1.07, 1.46)	0.006	0.625
11q13.3	rs554219	0.12	1.31 (1.08, 1.60)	0.007	0.8
16q12.1	rs3803662	0.26	1.35 (1.16, 1.57)	$1.2\text{x}10^{-4}$	0.99

3.3 Environmental risk factors and interactions

Several studies have identified potential GxE interactions in the context of breast cancer. Some studies have found evidence of interactions in the *FGFR2* locus [240, 241] as well as other loci [242]. Recent joined efforts from consortia revealed that it is unlikely that there will be a large enough interaction to be clinically useful [243]. However, this study looked at age at menarche, parity, age at first birth and BMI as environmental exposures. Findings from the UK Million Women Study also support this statement with no major interaction being identified after investigating 12 SNPs and 10 environmental risk factors in a cohort of 7,610 breast cancer cases and 10,196 controls [244]. Additionally, a collaboration study recently failed to identify any interaction with HRT after correcting for multiple testing. In this two-phased study, investigators tested the hypothesis of lobular specific interactions with no significant findings [245]. However, the number of lobular cases was very small. Another BCAC study investigating interactions between SNPs and HRT in the context of lobular breast cancer identified some potential signals that warrant further investigation [246].

We therefore followed a stratified phenotype approach, focusing on the lobular histology to interrogate environmental risk factors as well as investigate possible interactions between HRT and known SNPs that have been associated with lobular breast cancer.

3.3.1 Methods

Genotypic data for 12 single nucleotide polymorphisms (SNPs) that have been previously found to be associated with lobular breast cancer were obtained using two different genotyping approaches including the iCOGS custom platform from Illumina [154] and KASP genotyping technology. Individual effect sizes as well as the presence of GxE interactions were investigated. Genotypes were converted in dosage values (0, 1, 2) based on minor allele frequency (MAF) and minor allele count (MAC) was used as the genotypic value.

Controls aged <35 years or >60 years were excluded in order to match the age range of the cases. As all controls were required by definition not to have a family history of breast cancer, all cases with a family history of breast cancer were excluded as this might influence the use of exogenous hormones. For the same reason only cases and controls born between 1948 and 1971 were included in this analysis. Age was defined as the age of diagnosis (age on day of pathology report) for the cases and for controls as their age on the day questionnaire was received.

Women who reported a natural cessation of periods or a bilateral oophorectomy were classified as postmenopausal. Since the use of HRT whilst peri-menopausal and surgical interventions such as a hysterectomy without a bilateral oophorectomy, may mask the natural cessation of periods, anyone who fell in to these categories but was aged 56 and over, were assumed to be postmenopausal, and those younger were scored as having unknown menopausal status.

Parity was defined as the number of live births, excluding still births.

Risk factor data were analysed using SAS/STAT® software. All subjects were analysed together and then separate analysis was performed for post-menopausal women. Using multivariate logistic regression accounting for age in the model, odds ratios (ORs) and 95% confidence intervals (95%CI) for cases and controls were computed for age of menopause, age of menarche, parity (never, parous, 1-2, >2), age of first birth (≤ 25, > 25), ever breastfed in parous women, ever use of OC (combined estrogen-progesterone, progesterone only, combined and progesterone only), years of OC use, ever use of levonorgestrel-releasing intrauterine system (Mirena), ever use of HRT, years of HRT use (never, 0-4, 5-9, ≥ 10), type of HRT used (estrogen only, combined), and ever use of exogenous hormones (either HRT or OC). A case-only comparison of pure LCIS (ref) *vs* ILC +- LCIS was also performed for HRT.

The final data set included 1,095 cases consisted of 658 ILC + LCIS, 191 pure ILC, 39 mixed ILC with unknown LCIS status and 207 pure LCIS. Out of those 1,095 individuals, we obtained genotypic data for 1,050 of them. Out of the 1,326 controls that have been used in this study, we obtained genotypic data for 1,224 of them.

3.3.2 Environmental risk factors

There is a non-significant excess of HRT usage amongst cases (OR: 1.23; 95%CI: 1.00-1.48, *P*=0.056), Table 3.6. This association becomes significant when looking in postmenopausal women (OR: 1.37; 95%CI: 1.04-1.80, *P*=0.025). Also, women who had taken HRT for >10 years had a significantly increased risk (OR: 3.07; 95%CI: 1.87-5.03, *P*<0.0001). Breast feeding was protective (OR: 0.67; 95%CI: 0.54-0.83, *P*=0.0003). Age of first birth (≥ 26) was associated with increased risk of lobular disease (OR: 1.23; 95%CI: 1.03-1.47, *P*=0.0258) amongst parous women. There was a non-significant trend towards an association with combined HRT.

Table 3.6: Demographic and association data in women aged 35-60 born between 1948 and 1971.

	Controls N=1326	**Cases** N=1095	
	N (%)	N (%)	OR (95% CI) *P*
	Age		
Age (mean±SD)	49.55 (± 6.00)	50.75 (± 5.5)	**1.04 (1.02, 1.05), p <.0001**
	Age at menopause		
Mean age at menopause	48.79 (±5.32)	48.65 (±5.08)	0.99 (0.96, 1.02), p=0.44
	Age at menarche		
Mean age at menarche	12.82 (± 1.56)	12.85 (± 1.6)	1.01 (0.96, 1.06), p=0.72
	Parity		
Nulliparous	261 (19.68)	182 (16.62)	1 (ref)
Parous	1065 (80.32)	913 (83.38)	1.18 (0.96, 1.46), p= 0.12
	Age of First Birth		
≤ 25	749 (56.49)	557 (50.87)	1 (ref)
> 25	577 (43.51)	538 (49.13)	**1.23 (1.03, 1.47), p= 0.026**
	Breastfeeding		
Never	198 (18.66)	233 (25.55)	1 (ref)
Breastfed	863 (81.34)	679 (74.45)	**0.67 (0.54, 0.83), p= 0.0003**
	Oral Contraceptive (OC) use and type		
Never	172 (13.14)	151 (13.97)	1 (ref)
OC used	1137 (86.86)	930 (86.03)	0.95 (0.75, 1.20), p= 0.65
	Mirena Coil (MC) use		
Never	149 (58.66)	143 (71.50)	1 (ref)
MC used	105 (41.34)	57 (28.50)	**0.60 (0.40, 0.89), p= 0.011**
	Years of Contraceptive use		
Mean years of use	9.58 (± 7.38)	9.73 (± 7.2)	1.01 (0.99, 1.02) p=0.50
	Hormonal Replacement Therapy (HRT) use		
Never	1047 (80.35)	789 (74.02)	1 (ref)
HRT used	256 (19.65)	277 (25.98)	1.23 (1.00, 1.51), p= 0.056
	Years of HRT use		
Never	1047 (84.44)	789 (76.31)	1 (ref)
0 < x < 5	118 (9.52)	124 (11.99)	1.25 (0.94, 1.64), p= 0.12
5 ≤ x < 10	52 (4.19)	59 (5.71)	1.28 (0.87, 1.91), p= 0.21
≥ 10	23 (1.85)	62 (6.00)	**3.07 (1.87, 5.03), p <.0001**
	Type of HRT used		
None used	1047 (87.62)	789 (85.67)	-
estrogen only	70 (5.86)	51 (5.54)	1 (ref)
Combined	78 (6.53)	81 (8.79)	1.41 (0.87, 2.27), p= 0.16
	Any exogenous hormone use (either OC or HRT)		
Never	152 (6.35)	116 (4.85)	1 (ref)
Ever	1160 (88.41)	964 (89.26)	1.08 (0.84, 1.40)

The use of mirena coil has a protective effect since a smaller proportion of cases were using this method of contraception (OR=0.60 (95%CI 0.40, 0.89), *P*= 0.011), Figure 3.7.

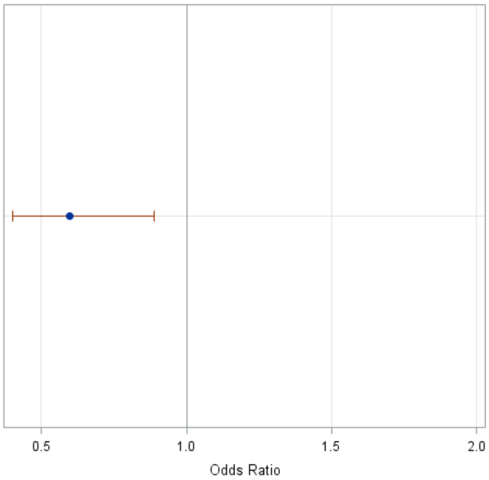


Figure 3.7: Effect size of mirena coil in lobular breast cancer development.

No association was found with use of OC (OR=0.95 (95%CI 0.75, 1.20), *P*=0.65) or use of any exogenous hormones (OR=1.08 95%CI 0.84, 1.40, *P*=0.54).

3.3.3 Genetic risk factors

12 SNPs that have been previously found to be associated with lobular breast cancer at a genome-wide level have been assessed in the context of this study. Table 3.7 shows the association of these loci using data from 1050 cases and 1224 controls. No covariates have been added in the logistic regression. The significance threshold has been Bonferroni corrected (p<0.0042) for 12 tests. Significant association was found with 7 SNPs in this data set. The associated loci were then followed up for a GxE interaction investigation with HRT usage as a binary variable (never versus ever users).

Table 3.7:Association of 12 loci previously associated with ILC in our data-set of 1050 lobular cases and 1224 controls.

SNP	Estimate	St Error	Chi-Square	P
rs75915166	0.3178	0.125	6.4642	0.011
rs10995190	-0.1262	0.0905	1.9421	0.1634
rs11249433	0.184	0.0619	8.8213	**0.003**
rs11977670	0.3136	0.0622	25.4186	**<.0001**
rs13387042	-0.053	0.0608	0.7602	0.3833
rs2981582	0.3667	0.0637	33.0933	**<.0001**
rs3803662	0.3149	0.0678	21.5585	**<.0001**
rs554219	0.3529	0.0902	15.3176	**<.0001**
rs6678914	-0.076	0.0629	1.4606	0.2268
rs704010	0.2277	0.061	13.9108	**0.0002**
rs865686	-0.2656	0.0641	17.1944	**<.0001**
rs889312	0.1214	0.0669	3.2943	0.0695

3.3.4 Gene-environment interactions

In order to look for the combined effects of HRT and genotypes, we selected only postmenopausal women from both cases and controls. This corresponds to 406 cases and 481 controls. The proportion of definitely postmenopausal women was equal amongst the two groups (37%). Table 3.8 shows the number of individuals per group of genotype and HRT status for the 7 SNPs that reached the corrected level of significance for this study. The corresponding effect sizes of HRT per genotype along with the significance *P* values across the 7 SNPs under investigation are indicated in Table 3.9. Evidence for three interactions has been found, having investigated 406 postmenopausal cases and 481 matched controls,

Table 3.10. Data is represented in two different ways plotting the 6 data points (genotypes 3x2 HRT status), Figure 3.8, Figure 3.9, Figure 3.10. Three different genetic models have been investigated, firstly the additive model where the exact allele count is considered, but also the dominant and recessive models where the heterozygotes are merged in one group with the homozygotes of the rare or the common allele respectively. For rs704010 and rs865686, the interaction seems to act in a recessive manner, where two copies of the risk allele are required to observe the effect from HRT. On the contrary, the interaction between rs2981582 (*FGFR2*) and HRT seems to follow the additive model. However, due to the relatively small sample size, these possible interactions need to be validated in additional samples for any conclusions to be drawn.

Table 3.8: Number of individuals stratified across different genotypic groups and ever vs never HRT users.

N of minor alleles	**0**				**1**				**2**			
	Controls		**Cases**		**Controls**		**Cases**		**Controls**		**Cases**	
SNP	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**
rs11249433	98	55	64	45	139	77	109	92	52	22	41	33
rs11977670	94	58	55	50	153	74	108	90	43	21	51	32
rs2981582	115	50	56	52	145	83	112	88	28	22	46	30
rs3803662	167	88	104	74	104	59	87	70	19	8	23	26
rs554219	229	118	160	119	59	31	49	44	2	6	5	5
rs704010	116	67	67	50	124	72	104	68	48	16	43	50
rs865686	125	55	95	81	130	76	93	82	35	24	26	8

Table 3.9: Interaction between ever vs never use of HRT with 7 SNPs that were significantly associated with lobular breast cancer.

N of minor alleles	**0**		**1**		**2**	
SNP	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***
rs11249433	1.11 (0.77, 1.61)	0.578	1.27 (0.93, 1.73)	0.127	1.54 (0.92, 2.60)	0.102
rs11977670	1.24 (0.85, 1.83)	0.267	1.32 (0.97, 1.79)	0.078	1.26 (0.76, 2.11)	0.374
rs2981582	1.82 (1.25, 2.66)	0.002	1.04 (0.77, 1.40)	0.816	0.86 (0.50, 1.48)	0.587
rs3803662	1.29 (0.95, 1.75)	0.104	1.08 (0.77, 1.52)	0.654	2.01 (0.93, 4.33)	0.077
rs554219	1.26 (0.98, 1.61)	0.078	1.09 (0.70, 1.70)	0.698	0.92 (0.22, 3.78)	0.905
rs704010	1.16 (0.81, 1.66)	0.415	1.13 (0.82, 1.56)	0.454	1.82 (1.06, 3.11)	0.029
rs865686	1.31 (0.93, 1.85)	0.119	1.35 (0.98, 1.85)	0.066	0.94 (0.52, 1.72)	0.849

Table 3.10: Interaction *P* values based on 3 different genetic models (additive, dominant and recessive) between ever vs never use of HRT vs 7 SNPs that were significantly associated with lobular breast cancer.

SNP	Locus	MAF	Additive interaction *P*	Dominant interaction *P*	Recessive interaction *P*
rs11249433	1p11.2	0.43	0.3237	0.4562	0.4097
rs11977670	7q34	0.43	0.8524	0.8142	0.5616
rs2981582	*FGFR2*	0.42	0.0309	0.0702	0.0945
rs3803662	*TOX3*	0.29	0.2868	0.5204	0.1842
rs554219	*CCND1*	0.12	0.7396	0.8538	0.1543
rs704010	*ZMIZ1*	0.41	0.0581	0.5072	0.0065
rs865686	9q31	0.37	0.0199	0.1496	0.0099

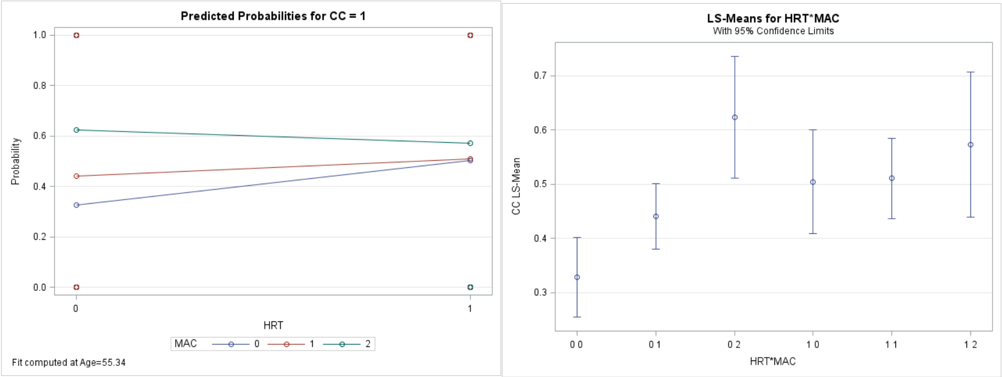


Figure 3.8: Interaction plots for rs2981582 (*FGFR2*). Interaction *P*= 0.030, lobular association *P*=5x10^{-52}.

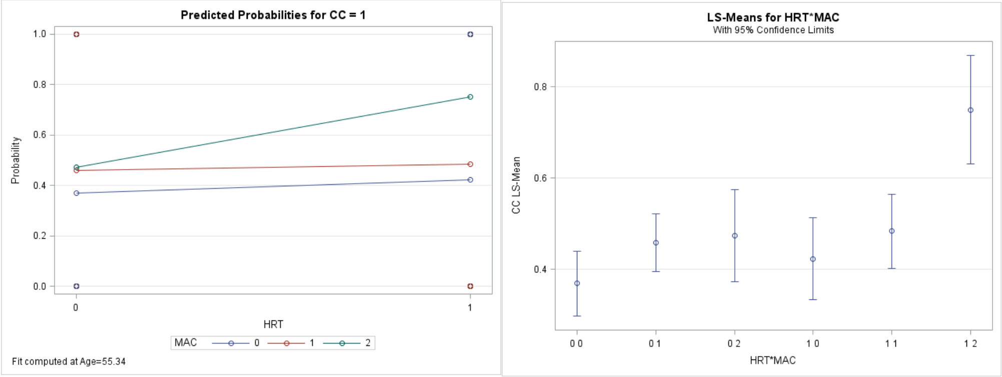


Figure 3.9: Interaction plots for rs704010 (*ZMIZ1*). Interaction under recessive model *P*= 0.0065, lobular association *P*=3x10^{-10}.

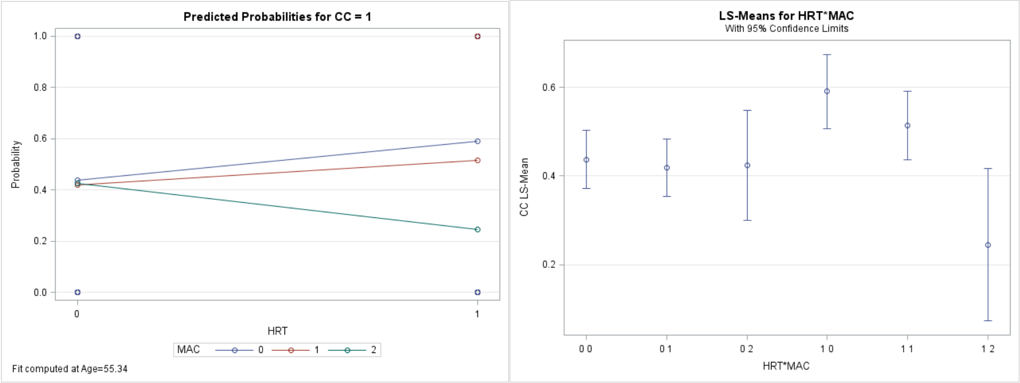


Figure 3.10: Interaction plots for rs865686 (9q31.2). Interaction *P*= 0.019, lobular association P=1x10^{-17}

3.4 Discussion

In this study, where we focused on lobular breast cancer, we identified a novel variant that appears to be specific to this morphological subtype. We also ascertained which of the known variants predispose specifically to lobular breast cancer and shown for the first time that some of these loci are also associated with LCIS. Our study showed that the genetic pathways of invasive lobular cancer and ER positive IDC mainly overlap, but with key differences.

Analysing of a total of 6,539 lobular breast cancer cases (including 436 cases of pure LCIS) and 35,710 controls led to the first identification of a lobular-specific SNP, rs11977670 (*JHDM1D*; OR=1.13 P=4.2x10^{-10}, that showed little or no evidence of association with IDC (P=0.064) or DCIS (P= 0.44). Fine-mapping of the region is required, followed by functional assays to determine whether the associated SNPs regulate the function of certain genes. Preliminary *in silico* functional analysis suggests that SNPs in this region may be influencing expression of *JHDM1D* (histone demethylase) and *SLC37A3* (sugar-phosphate exchanger). An eQTL analysis using a surrogate SNP, rs13225058, revealed that the risk allele is associated with increased expression of both *JHDM1D* and *SLC37A3*. This analysis included 335 ER positive primary tumours where both genotyping and expression data were available and were downloaded from TCGA. There are little data on the role of these genes in cancer. There is some evidence that increased expression of *JHDM1D* can suppress tumour growth by regulating angiogenesis [247] and decreased expression promotes invasiveness, which is contrary to what one would expect from the risk data. Studies of syndecan-1-deficient breast cancer cells, which show increased cell motility and invasiveness, demonstrate decreased expression of both *JHDM1D* and E-cadherin [248], suggesting the two genes may interact.

Somatic mutations in *CDH1* are frequent in ILC and rare germline frameshift mutations in *CDH1* have been described in ILC, particularly in families with hereditary diffuse gastric cancer (HDGC), but also in cases of familial ILC with no HDGC [115, 249, 250]. However, none of the *CDH1* tagging SNPs that were typed on the iCOGs chip showed any association with lobular cancer at P<0.05.

A total of 75 of the known common breast cancer susceptibility loci were assessed for association with ILC and LCIS. As cases of ILC are generally ER positive with the majority of ILCs classified as luminal tumours [251], it does not come as a surprise that the majority of SNPs that we found to be associated with ILC were previously known to predispose to ER positive breast cancer. However, some loci were only associated with ER positive IDC and not with ILC, particularly rs10941679 at 5p12, previously shown to more strongly predispose to ER positive, lower-grade cancers [252], *P*-het=2.7x10^{-8}. Another study showed a much stronger association with ILC than IDC, particularly rs11249433 at 1p11.2 [229]. These data show evidence for specific aetiological pathways in the development of different histological subtypes of breast cancer, in addition to common pathways that predispose to multiple tumour subtypes.

Our analyses have shown for the first time that many of the SNPs that predispose to ILC also predispose to LCIS, even though the number of pure LCIS cases is small. Although only 15 of the known breast cancer SNPs were associated with LCIS risk at *P*<0.05, 47 of the remaining 60 SNPs at *P*>0.05 had ORs in the same direction as for ILC (Sign Test *P*=1.2x10^{-5}) suggesting that many more SNPs are likely to be associated with pure LCIS but did not reach statistical significance because of the relatively few LCIS cases without associated ILC in our study. This is not unexpected if LCIS is an intermediate phenotype for ILC. However, a small number of SNPs had differential effects on LCIS or ILC risk. Specifically, rs6678914 at 1q32.1 (*LGR6*), known to be an ER negative specific SNP [157], that appeared to be associated with LCIS but not ILC (*P*-het=0.0007), and rs17529111 at 6q14 preferentially associated with ER negative tumours [168] that had a stronger association with LCIS than ILC (*P*-het=0.04). We also identified SNPs in *FGFR2* and at 5q11.2 (*MAP3K1*) that appear only to predispose to ILC. These findings are surprising but based on small numbers and therefore need confirmation in future studies.

Some of the SNPs associated with both ILC and LCIS showed a stronger effect size in LCIS compared to ILC (for example SNPs at *TOX3*, 9q31.2, 11q13.3, *ZNF365* and *MLLT10*). It is possible that the SNPs that showed an association with both LCIS and ILC predispose to the development of LCIS rather than ILC, and that the effect size is smaller in ILC as not all cases of LCIS will become invasive cancer. SNPs that predispose strongly to LCIS were also associated with ER positive IDCs but again with stronger effect sizes in LCIS, consistent with the fact that 30-40% of invasive tumours associated with LCIS will not be ILC but will be IDC, mixed ductal-lobular or other morphology.

With regards to environmental risk factors, we do not observe any deviation from what is already known, with late age of first birth and long periods of HRT usage increasing the risk, while breastfeeding having a protective effect.

Enough power was obtained to validate 7 out of 12 loci that have previously been found to be associated with lobular breast cancer on the same subset of our data set that has been used to investigate the effect of environmental risk factors. Using the seven SNPs that reached the corrected significance threshold, we investigated potential GxE interactions using genotypes from those seven SNPs and use of HRT in postmenopausal women.

Previous studies have highlighted the potential interaction between HRT and SNPs located in the *FGFR2* locus. We have also shown evidence of an interaction on this locus. The log-additive increased risk that is conferred by the risk allele is observed in the group of individuals who have not used HRT. This finding suggests that use of HRT can mask the effect of a SNP since individuals using HRT are at equally higher risk than no HRT subjects irrespective of their genotype.

In addition to that, a common polymorphism in the *ZMIZ1* locus shows a possible interaction with use of HRT. As shown in Figure 3.9, the effect of the SNP and HRT is higher in HRT users that are carriers of two risk alleles. The significance of the recessive model (where heterozygotes and homozygotes of the common allele are treated as one group) is higher with *P*=0.0065 compared to the additive model where the number of alleles per individual is counted and taken into consideration. Larger sample size would be required to confirm this finding.

Finally, the last possible interaction is found with rs865686 (9q31.2) where we observe the same effect with users of HRT and carriers of two risk alleles.

With the sample size of the current study, and an interaction that is not expected to be very strong in terms of increased risk, it is evident that we are underpowered and more samples would be required in order to dissect and accurately estimate possible interactions between the use of HRT and SNPs that predispose to ILC.

In conclusion, we have identified a novel lobular-specific predisposition SNP at 7q34 close to *JHDM1D* that does not appear to be associated with IDC. Most known breast cancer predisposition SNPs also predispose to ILC, with some differential effects between ILC and IDC. In addition, many SNPs predisposing to invasive cancer are also likely to increase the risk for LCIS. We have also shown for the first time that common breast cancer polymorphisms predispose to LCIS. Furthermore, we have shown that many of the ER positive breast cancer predisposition loci also predispose to ILC, although there is some heterogeneity between ER positive lobular and ER positive IDC tumours. Overall, our analyses show that genetic predisposition to IDC and ILC overlap to a large extent, but there are important differences that are likely to prove insightful. No GxE interaction for postmenopausal women was deemed significantly associated with ILC after correcting for multiple testing but we have shown suggestive evidence of three possible interactions that qualify for further investigation.

Chapter 4 Rare variants in known predisposition genes contributing to DCIS development

4.1 Introduction

Most non-genetic risk factors for breast cancer have similar associations with DCIS and IDC, supporting the notion that DCIS is a precursor of invasive cancer [235, 237]. There is also evidence from epidemiological studies that there is an inherited predisposition to DCIS. Claus *et al* showed that women with DCIS are more likely to have a first-degree relative with breast cancer than controls with an odds ratio of 1.6 (95% CI 1.3-2.1). They have also demonstrated that DCIS cases are 2.4-times (non -significant) (95%CI=0.8-7.2) more likely to have an affected mother and sister with breast cancer than controls [253]. Furthermore, there is evidence from a study of almost 40,000 women that the familial relative risk of DCIS is greater than that of invasive breast cancer. For women with a family history of breast cancer aged 30-49 the OR for developing DCIS was 2.4 (95%CI=1.1, 4.9) compared to an OR of 1.7 (95%CI=0.9, 3.4) for invasive cancer. For women aged 50 and above the risks were slightly reduced, but still higher for DCIS (OR=2.2, 95%CI=1.0, 4.2) than invasive disease (OR=1.5, 95%CI=1.0, 2.2) [254]. However, this was not confirmed in the UK Million Women Study, which showed a similar association with family history for DCIS and IDC [237].

A part of this inherited predisposition is explained by *BRCA1* and *BRCA2* mutations, as mutations in these genes are found in a similar proportion of DCIS and invasive breast cancer cases [255]. In a study that screened a total of 7,295 with *in situ* breast cancer (with or without presence of personal or family history of invasive breast cancer) they estimated the prevalence of *BRCA1* and *BRCA2* mutations to be 5.9%. In a more restricted analysis investigating individuals with no personal or family history of invasive breast or ovarian cancer, they identified 17 carriers out of 738 (Prevalence =2.3%) [256]. This is in agreement with earlier findings from Claus *et al.* where they found the prevalence of BRCA1/2 mutations to be 3.2% amongst 369 DCIS cases that they screened [255].

In order to assess the prevalence of germline mutations in known breast cancer predisposition genes including *BRCA1* and *BRCA2* in the context of an unselected population of DCIS, we screened individuals that were diagnosed with the disease before the age of 50.

4.2 Rare variants in known breast cancer predisposition genes

4.2.1 Methods

Samples were selected from the ICICLE study (Investigation of the genetiCs of In situ Carcinoma of the ductaL subtypE), and after a review of the pathology reports, we identified 680 individuals that were eligible for this project. All individuals were under 50 years of age when diagnosed and had no invasive disease. The final data set includes 657 individuals with DCIS diagnosed ≤ 50. A total of 124 cases were diagnosed ≤ 40. Along with the DCIS cases, we additionally screened 1,611 healthy European females that were not diagnosed with any form of breast cancer by at least the age of 40. The age for controls ranged between 40 and 92, with a median of 52. The numbers of individuals included in the final analysis stratified by ER status and nuclear grade are indicated in Table 4.1. For 94 cases, immunohistochemistry was performed by our research group members to assess the ER status of the lesion since the pathology reports were incomplete. ER staining was reviewed by our study pathologist.

Table 4.1: ER status and nuclear grade information for DCIS cases included in the final analysis.

	ER positive	ER Negative	Missing	All
High grade	229	74	94	397
Intermediate grade	118	3	46	167
Low grade	30	1	23	54
Missing	2	0	37	39
All	379	78	200	657

In our cohort of 657 unselected DCIS cases, we have information on family history for 633 of them. A total of 168 (25.6%) cases had a first degree relative with breast cancer. From those 168 cases, 31 (5%) were diagnosed ≤ 40. 465 (70.7%) cases had no affected first degree relatives, and the remaining 24 (3.7%) cases had missing data.

In order to assess the prevalence of rare variants in known breast cancer predisposition genes we utilised a targeted sequencing method. For this project we interrogated the presence of rare variants in 6 known breast cancer predisposition genes in the context of early onset DCIS. Apart from *BRCA1* and *BRCA2*, we screened *TP53*, *CDH1*, *CHEK2*, and *PALB2*.

The Fluidigm Access Array technology has been used to amplify germline DNA from those individuals. Samples were pooled in a 960 multiplexing format and sequenced on a HiSeq2500 lane. The cluster density for the lane that included all DCIS samples was 713 +/- 39 K/mm^2.

The analysis pipeline is described in section 7.4.3. A total of 230 variant sites passed the QC metrics, including a MAF<1% cut-off. Out of those variants 177 were missense, 2 were non-frameshift deletions, and the remaining 51 were protein truncating variants.

Variants were classified as benign, VUS, and pathogenic according to ClinVar. For *BRCA1* and *BRCA2* in particular, a more accurate database (BIC) was used that incorporates findings from several validated studies. Variants that are reported in ClinVar as pathogenic or likely pathogenic were merged in the pathogenic category along with protein truncating variants that were not present in ClinVar. Variants that were previously classified as benign or likely benign were merged into the benign category for the purpose of our analyses. The remaining variants were considered as VUS.

There is no enrichment with regards to benign variants compared to controls in any of the genes or overall. This observation also serves as an internal quality control and as a metric of calling variants across cases and controls with no bias. The prevalence of benign variants in the six genes under investigation is 17.5% across both DCIS case and control populations.

Table 4.2 shows the prevalence of pathogenic mutations in our panel of six genes stratified by family history of breast cancer and age of diagnosis. Family history is defined as having at least one first degree relative with breast cancer.

With regards to pathogenic variants, there is an overall significant enrichment in DCIS cases compared to controls for all DCIS cases ≤ 50 as well as DCIS cases diagnosed ≤ 40 (Fisher's exact test, section 7.3.2. The prevalence of pathogenic variants in DCIS cases diagnosed ≤ 50 is 5% and reaches almost 10% for cases diagnosed ≤ 40. The distributions of mutations across all six genes under investigation for cases diagnosed ≤ 50 and ≤ 40 are indicated in Figure 4.1.

Table 4.2: Prevalence of germline mutations identified in our cohort of DCIS stratified by family history and age of onset. FH corresponds to having a first degree relative with breast cancer. OR and *P* are generated using a Fishers exact test comparing the cases to our set of 1,611 controls. Cases with missing information on family history were excluded from family history related analyses.

Group	Carriers	Frequency	OR	95% CI	*P*
Controls	10	0.6%	Ref	Ref	Ref
Age ≤ 40 +FH	6	19.4%	38.42	13, 113.90	$1.9\text{x}10^{-7}$
Age ≤ 40 -FH	5	5.9%	10.13	3.38, 30.35	0.00054
Age ≤ 40 all	12	9.7%	17.15	7.25, 40.57	$3.7\text{x}10^{-9}$
Age ≤ 50 +FH	17	10.1%	18.02	8.11, 40.06	$6.4\text{x}10^{-12}$
Age ≤ 50 -FH	15	2.7%	5.34	2.38, 11.96	$5\text{x}10^{-5}$
Age ≤ 50 all	33	5%	8.46	4.14, 17.28	$7.9\text{x}10^{-11}$

The presence of family history is more profound amongst carriers when compared to non-carriers. As indicated in Table 4.3, significant differences are observed for cases diagnosed before 50 and the same trend is observed for cases diagnosed before 40 even though it does not reach significance, probably due to small sample size. Carriers are more likely to have a first degree relative with breast cancer compared to non-carriers.

Table 4.3: Presence of first degree relative with breast cancer is more frequent amongst germline mutation carriers compared to non-carrier DCIS cases.

Group	N (%) cases with 1st degree relative affected	N (%) carriers without 1st degree relative affected	OR	95% CI	*P*
Age ≤ 40	6 (19.4%)	5 (5.9%)	3.79	1.06, 13.49	0.06
Age ≤ 50	17 (10.1%)	15 (3.2%)	3.38	1.65, 6.93	0.0014

The same analysis was conducted for the two main breast cancer predisposition genes and the results are shown in Table 4.4.

Table 4.4: Prevalence of pathogenic variants in *BRCA1* and *BRCA2* amongst different groups of DCIS stratified by age of diagnosis and having a first degree relative with breast cancer.

Groups	Gene	N Carriers	Frequency	OR	95% CI	*P*
All DCIS cases diagnosed ≤ 50 all (N=657)	*BRCA1*	4	0.61%	inf	-	0.007
	BRCA2	18	2.74%	15.10	4.43, 51.43	$9.1x10^{-8}$
	Combined	21	3.2%	17.70	5.26, 59.54	$3.1x10^{-9}$
Age ≤ 50 with affected 1st degree relative (N=168)	*BRCA1*	3	1.79%	inf	-	0.00083
	BRCA2	9	5.36%	30.34	8.13, 113.21	$8.4x10^{-8}$
	Combined	11	6.55%	37.55	10.37, 136.03	$1.1x10^{-9}$
Age ≤ 50 with no affected 1st degree relative (N=465)	*BRCA1*	1	0.22%	inf	-	0.22
	BRCA2	8	1.72%	9.38	2.48, 35.51	0.00052
	Combined	9	1.94%	10.58	2.85, 39.23	0.00015
All DCIS cases diagnosed ≤ 40 (N=124)	*BRCA1*	2	1.61%	inf	-	0.0051
	BRCA2	8	6.45%	36.97	9.68, 141.21	$7.5x10^{-8}$
	Combined	9	7.26%	41.95	11.20, 157.07	$6.7x10^{-9}$
Age ≤ 40 with affected 1st degree relative FH (N=31)	*BRCA1*	1	3.23%	inf	-	0.019
	BRCA2	4	12.90%	79.41	16.95, 372.09	$3.5x10^{-6}$
	Combined	5	12.90%	103.08	23.40, 454.13	$9.2x10^{-8}$
Age ≤ 40 with no affected 1st degree relative (N=84)	*BRCA1*	1	1.19%	inf	-	0.049
	BRCA2	3	3.57%	19.85	3.95, 99.90	0.0021
	Combined	4	4.76%	26.80	5.90, 121.77	0.00018
Controls (N=1,611)	*BRCA1*	0	0%	Ref	Ref	Ref
	BRCA2	3	0.18%	Ref	Ref	Ref
	Combined	3	0.18%	Ref	Ref	Ref

Almost 20% of individuals under the age of 40 with a first degree relative with breast cancer are carriers of a pathogenic mutation in one of the 6 genes that were screened in our study. The prevalence of mutations was 10.1% for individuals with an affected first degree relative and their age of diagnosis ≤ 50.

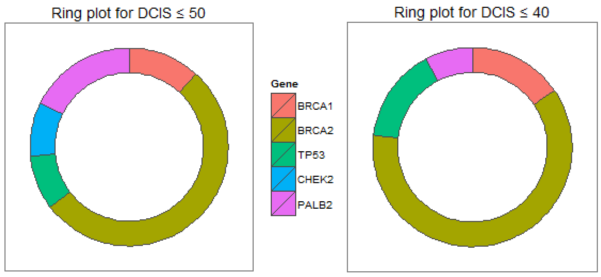


Figure 4.1: Distribution on pathogenic mutations in screened genes across individuals diagnosed with DCIS ≤ 50 years (left), and ≤ 40 years (right).

4.2.2 Prevalence of *BRCA1* rare variants

A total of 4 *BRCA1* truncating variants were identified in cases with DCIS. All of those variants have been previously assessed as pathogenic. These variants are indicated in Figure 4.2 and Table 4.5. The frequency of *BRCA1* pathogenic variants is 0.6% in our cohort, Table 4.6.

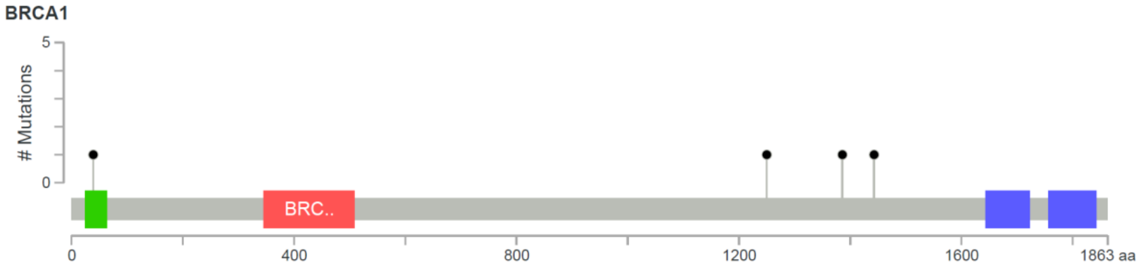


Figure 4.2: Distribution of *BRCA1* protein truncating variants identified in 4 individuals with DCIS diagnosed ≤ 50.

Table 4.5: Details of *BRCA1* protein truncating variants identified in DCIS cases.

Sample ID	Nt change	AA change	Class	Age	Family history	Grade	ER status
IT00761	c.C4327T	p.R1443X	Stop-gain	49	Sister, breast, 48	High	Negative
IT00811	c.3750delG	p.E1250fs	Frameshift	38	Mother, cervical, 37	High	Positive
IT01459	c.117_118del	p.C39fs	Frameshift	49	Mother-aunt-sister, breast-breast-ovarian, 38-40-32	Low	Negative
IT02701	c.4158_4162 del	p.S1386fs	Frameshift	40	Sister-sister, breast-breast, 30s-30s	High	Negative

One individual diagnosed with high grade ER positive DCIS at the age of 38, was a carrier of a pathogenic variant in both *BRCA1* and *BRCA2*. Both variants are frameshift deletions; *BRCA1*: c.3750delG, p.E1250fs, and *BRCA2*: c.4445delA, p.E1482fs. This was the only carrier with ER positive DCIS whereas the remaining three had ER negative DCIS.

Table 4.6: Rare variants identified in controls and DCIS cases at the *BRCA1* gene.

Class	N (%) Controls	N (%) DCIS	OR	95% CI	*P*
Benign	44 (2.73)	14 (2.13)	0.76	0.41, 1.39	0.46
VUS	40 (2.48)	18 (2.74)	1.11	0.63, 1.94	0.77
Pathogenic	0 (0)	4 (0.61)	inf	-	0.007

4.2.3 Prevalence of *BRCA2* rare variants

A total of 18 protein truncating variants were identified in our cohort of 657 DCIS cases, Figure 4.3. The prevalence of *BRCA2* pathogenic variants in our unselected population of DCIS cases diagnosed before the age of 50 is 2.7%, Table 4.7. All of these variants have been previously described as pathogenic. Some key characteristics of the carriers are indicated in Table 4.8. It is of note that there was no ER negative DCIS amongst the DCIS *BRCA2* carriers.

Table 4.7: Rare variants identified in controls and DCIS cases at the *BRCA2* gene.

Class	N (%) Controls	N (%) DCIS	OR	95% CI	*P*
VUS	177 (11)	77 (11.72)	1.08	0.81, 1.43	0.61
Benign	47 (2.92)	29 (4.41)	1.54	0.96, 2.46	0.093
Pathogenic	3 (0.19)	18 (2.74)	15.1	4.43, 51.4	$9x10^{-8}$

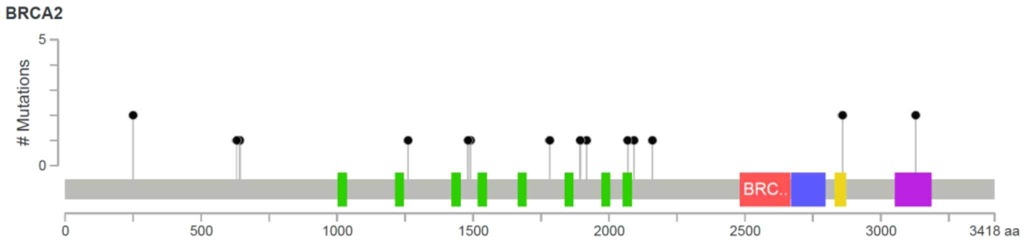


Figure 4.3: Distribution of *BRCA2* protein truncating variants identified in individuals with DCIS diagnosed ≤ 50.

Table 4.8: Details of *BRCA2* protein truncating variants identified in DCIS cases and 3 controls.

Sample ID	Nt change	AA change	Class	Status	Age	Grade	ER status
CG00386	c.5946delT	p.S1982fs	Frameshift	Control	41	-	-
CG00552	c.5098delG	p.G1700fs	Frameshift	Control	41	-	-
CG01138	c.C2612A	p.S871X	Stop-gain	Control	43	-	-
IT02381	c.631+2T>G		Splicing	DCIS	46	Low	Positive
IT01335	c.750_753del	p.V250fs	Frameshift	DCIS	48	High	Positive
IT03243	c.750_753del	p.V250fs	Frameshift	DCIS	39	High	Missing
IT03343	c.1929delG	p.V643fs	Frameshift	DCIS	41	High	Missing
IT01733	c.C3785G	p.S1262X	Stop-gain	DCIS	37	Intermediate	Positive
IT00811	c.4445delA	p.E1482fs	Frameshift	DCIS	38	High	Positive
IT01134	c.4473_4476del	p.L1491fs	Frameshift	DCIS	44	Intermediate	Positive
IT00953	c.5345_5346del	p.Q1782fs	Frameshift	DCIS	41	High	Missing
IT02863	c.C5682G	p.Y1894X	Stop-gain	DCIS	30	High	Positive
IT02864	c.5754dupT	p.H1918fs	Frameshift	DCIS	47	Intermediate	Missing
IT02922	c.T6206G	p.L2069X	Stop-gain	DCIS	47	Intermediate	Missing
150452	c.6275_6276del	p.L2092fs	Frameshift	DCIS	38	Missing	Missing
IT00562	c.6482_6485del	p.D2161fs	Frameshift	DCIS	47	High	Positive
IT01747	c.7977-1G>C		Splicing	DCIS	34	Intermediate	Positive
IT02197	c.8575delC	p.Q2859fs	Frameshift	DCIS	44	Intermediate	Positive
IT02811	c.8575delC	p.Q2859fs	Frameshift	DCIS	40	High	Positive
IT00033	c.C9382T	p.R3128X	Stop-gain	DCIS	42	High	Positive
IT00099	c.C9382T	p.R3128X	Stop-gain	DCIS	34	High	Positive

4.2.4 No protein truncating variants identified in the *CDH1* gene

CDH1 was selected for its prior association with the lobular histology and our results validate the previous notion that it is a lobular specific locus. No protein truncating variant was identified in individuals with DCIS, Table 4.9. Additionally there was no enrichment of missense variants in a case control manner CAF_{DCIS}=1.3% and $CAF_{Controls}$=1.5%, *P*=0.66. The coverage of exon 1 and exon 12 was suboptimal, but this it is not expected to dramatically change the analysis in terms of identifying pathogenic variants.

Table 4.9: Rare variants identified in controls and DCIS cases at the *CDH1* gene.

Class	N (%) Controls	N (%) DCIS	OR	95% CI	*P*
Benign	29 (1.8)	11 (1.67)	0.93	0.46, 1.87	0.99
VUS	10 (0.62)	7 (1.07)	1.72	0.65, 4.55	0.29
Pathogenic	0 (0)	0 (0)	nan	nan	nan

4.2.5 Three pathogenic *TP53* variants identified in DCIS cases

Another gene with highly penetrant variants that has been linked to breast cancer is *TP53*. However, germline mutations in that gene are rare since they usually exist in families with Li-Fraumeni syndrome. Two variants that have been previously found to be pathogenic were identified (c.C916T: p.R306X: rs121913344, and c.G542A: p.R181H: rs397514495). A further protein truncating variant was also identified and is very likely to be causal. It has not been described before as a germline variant but it has been found as a somatic alteration in a colon carcinoma specimen [257], Table 4.10, Figure 4.4. The prevalence of *TP53* germline mutations in our cohort is approximately 0.5%, Table 4.11.

Table 4.10: Details of three *TP53* variants that have been previously found to be pathogenic and were identified in three individuals with DCIS.

Sample ID	Nt change	AA change	Class	Age	Family history (relative, type, age)	Grade	ER status
IT03363	c.C916T	p.R306X	Stop-gain	40	Father unknown 50	High	Negative
IT01091	c.G542A	p.R181H	Non-synonymous	45	Grandmother-cousin, breast-breast, 60-60	Intermediate	Positive
IT00768	c.G272A	p.W91X	Stop-gain	35	Mother, breast, 39	High	Positive/Negative

Table 4.11: Rare variants identified in controls and DCIS cases at the *TP53* gene.

Class	N (%) Controls	N (%) DCIS	OR	95% CI	*P*
Benign	2 (0.12)	2 (0.3)	2.46	0.35, 17.48	0.33
VUS	5 (0.31)	2 (0.3)	0.98	0.19, 5.07	0.99
Pathogenic	0 (0)	3 (0.46)	inf	-	0.024

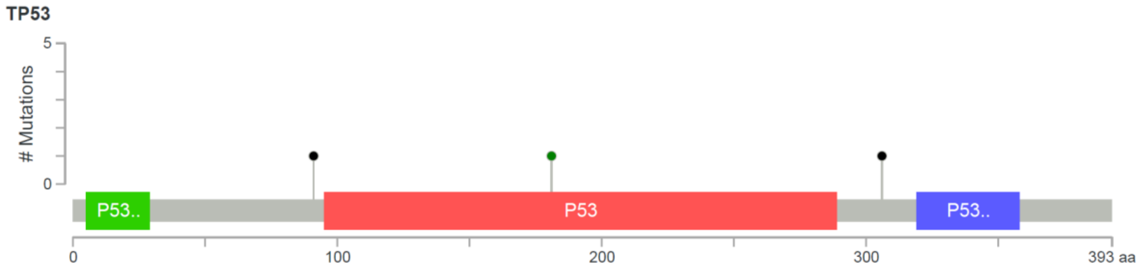


Figure 4.4: Distribution of *TP53* protein truncating (black lollipop) and one non-synonymous (green lollipop) variant identified in 3 individuals with DCIS diagnosed ≤ 50.

4.2.6 CHEK2

Three protein truncating variants were identified amongst the 657 DCIS cases, Figure 4.5. These are p.L200fs, p.D134fs, p.E236fs and the age of diagnosis of the carriers were 49, 42, and 41 respectively, Table 4.12. The case control comparison in the context of pathogenic variants includes all variants that have been reported as pathogenic in ClinVar, and therefore includes two non-synonymous variants that confer moderate risk towards breast cancer and are present in 5 controls. However, due to the fact that the penetrance of *CHEK2* variants varies, no conclusions can be drawn for its contribution towards DCIS. An additional analysis was performed, including only frameshift mutations which are very likely to disrupt the genes function, Table 4.13.

Table 4.12: Details of *CHEK2* variants that have been previously found to be pathogenic identified in 3 DCIS cases and 6 controls.

Sample ID	Nt change	AA change	Class	Status	Age	Grade	ER status
CG00303	c.T470C	p.I157T	Non-synonymous	Control	47	-	-
CG00741	c.T470C	p.I157T	Non-synonymous	Control	60	-	-
CG01482	c.T470C	p.I157T	Non-synonymous	Control	40	-	-
CG00400	c.C1196T	p.S399F	Non-synonymous	Control	75	-	-
CG00417	c.C1196T	p.S399F	Non-synonymous	Control	41	-	-
CG01651	c.1375-2A>G		Splicing	Control	48	-	-
IT01888	c.1176delT	p.L392fs	Frameshift	DCIS	49	High	Positive
IT01444	c.402_403del	p.D134fs	Frameshift	DCIS	42	High	Positive
IT01057	c.1281dupA	p.E428fs	Frameshift	DCIS	41	High	Negative

Table 4.13: Rare variants identified in controls and DCIS cases at the *CHEK2* gene.

Class	N (%) Controls	N (%) DCIS	OR	95% CI	*P*
Benign	1 (0.06)	0 (0)	0	-	0.99
VUS	18 (1.12)	16 (2.44)	2.21	1.12, 4.36	0.034
Pathogenic	6 (0.37)	3 (0.46)	1.23	0.31, 4.92	0.72
Frameshift indels	0 (0)	3 (0.46)	inf	-	0.024

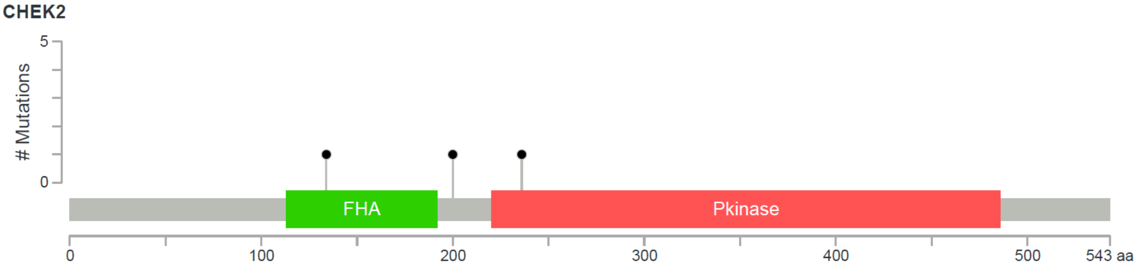


Figure 4.5: Distribution of the three protein truncating variants identified in DCIS cases across the *CHEK2* gene.

4.2.7 PALB2

A significant enrichment of *PALB2* protein truncating and highly likely pathogenic variants has been observed in our cohort, Table 4.14. A total of 6 protein truncating variants have been identified in 6 individuals with DCIS with their age of diagnosis ranging from 39 to 49. The presence of *PALB2* variants at this frequency in DCIS is a novel finding that could have clinical implications. A total of six protein truncating variants were identified in *PALB2*, Figure 4.6. Variant p.W1038X was identified in two individuals with age of diagnosis at 43 and 49 respectively, Table 4.15. This variant was initially reported by Rahman *et al.* when *PALB2* was identified as a breast cancer predisposition gene [129]. A second variant that was initially reported at the same study was identified in an additional individual from our study. Case IT03307 is a carrier of a stop-gain variant, p.Y1183X and was diagnosed with DCIS at the age of 46. Another protein truncating variant, p.F776fs was present in a case diagnosed with DCIS at the age of 39. This individual had both her mother and her grandmother affected with breast cancer at the age of 59 and 65 respectively. The same variant has been recently described in a study investigating the prevalence of *PALB2* germline mutations in familial cases of breast cancer [258]. They identified one carrier (diagnosed with breast cancer at 39) who had breast cancer, with four maternal and one paternal family member also having breast cancer. Another variant identified in our study, p.R170fs in a DCIS case diagnosed at 44, has also been previously described as a mutation in a cohort of breast and ovarian cancer in the Polish population [259]. Finally, a protein truncating variant, that to our knowledge has not been previously described, was identified in an individual diagnosed with DCIS at the age of 41. This case had her mother and her grandmother (maternal side) affected with breast cancer at the age of 52 and 54 respectively. The variant is a frameshift deletion in exon 4 (c.833delT:p.L278fs).

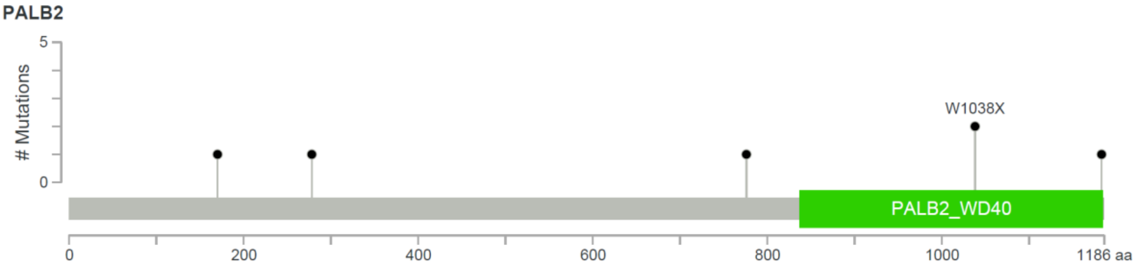


Figure 4.6: Distribution of *PALB2* protein truncating variants identified in 6 individuals with DCIS diagnosed ≤ 50.

Table 4.14: Rare variants identified in controls and DCIS cases at the *PALB2* gene.

Class	N (%) Controls	N (%) DCIS	OR	95% CI	*P*
Benign	30 (1.8)	11 (1.6)	0.9	0.45, 1.80	0.86
VUS	37 (2.9)	6 (1.1)	0.39	0.16, 0.93	0.027
Pathogenic	1 (0.06)	6 (1.1)	14.84	1.78, 123.5	0.0031

Table 4.15: Details of *PALB2* protein truncating variants identified in DCIS cases and 1 control.

Sample ID	Nt change	AA change	Class	Status	Age	Grade	ER status
CG01848	c.2052delC	p.P684fs	Frameshift	Control	41	-	-
IT03307	c.C3549G	p.Y1183X	Stop-gain	DCIS	46	High	Positive
IT00229	c.G3113A	p.W1038X	Stop-gain	DCIS	43	Intermediate	Positive
IT02690	c.G3113A	p.W1038X	Stop-gain	DCIS	49	High	Positive
IT03010	c.833delT	p.L278fs	Frameshift	DCIS	41	High	Missing
IT01887	c.2325dupA	p.F776fs	Frameshift	DCIS	39	High	Positive
IT03240	c.509_510del	p.R170fs	Frameshift	DCIS	44	High	Positive

4.2.8 *In situ* comparison of mutations in known breast cancer predisposition genes

In order to investigate the differential association between two different histological subtypes of *in situ* breast carcinoma, we conducted a case only analysis including 657 DCIS cases and 163 LCIS cases. All individuals included in the analysis were diagnosed ≤ 50 and had pure *in situ* lesions with no presence of invasive disease.

A total of 2 LCIS cases were carriers leading to a prevalence of 1.2% for LCIS. On the contrary, 33 DCIS carriers lead to a prevalence of 5% for DCIS. This analysis shows that rare pathogenic variants in known breast cancer predisposition genes are more common amongst DCIS cases compared its lobular counterpart, LCIS, (OR=4.26, 95% CI 1.01-17.93, *P*=0.03).

4.3 Discussion

Previous literature in the prevalence of pathogenic germline mutations in individuals with DCIS is limited. A study that screened *BRCA1* and *BRCA2* in individuals with *in situ* disease identified that the prevalence of mutations varies between 2.3%, for cases with no family history of breast or ovarian cancer and 7.7% for cases with family history of both breast and ovarian cancer [255]. Our study confirms the previous literature having identified similar proportions of germline mutations in our study of DCIS.

Initially Claus *et al*. reported the combined prevalence of 3.3% for pathogenic variants in *BRCA1* and *BRCA2* in a population study of DCIS. Our findings come in concordance with these data. In our cohort of 657 cases of DCIS diagnosed ≤ 50 we identified 22 pathogenic variants in *BRCA1* and *BRCA2* in 21 individuals which leads to a carrier frequency of 3.2%.

There was no large study investigating the prevalence of mutations of other breast cancer predisposition genes such as *CDH1*, *TP53*, *PALB2*, and *CHEK2*. Even though the prevalence of mutations in these genes is unknown for DCIS, they have been described before.

As expected, *CDH1* does not contribute to DCIS pathogenesis. Our study confirmed the previous notion that *CDH1* is a predisposition gene exclusively to the lobular histological subtype of breast cancer.

A study interrogating morphological and molecular features of 39 *TP53* germline mutation carriers identified that 11 of them were DCIS [260]. Another study with the aim of characterising the features of breast cancers arising from germline *TP53* mutations, identified 7 DCIS cases out of the 29 of their whole sample set [261]. However, these two studies were focused on germline mutation carriers and cannot be compared to our study. In our study we can provide an estimate of the prevalence of *TP53* mutations in an unselected cohort of DCIS.

In a cohort investigating the prevalence of *CHEK2* variants across different subtypes of breast cancer identified 13 protein truncating *CHEK2* variants out of 203 DCIS cases screened. However, those cases had concurrent micro-invasion lesions which might represent a different histological population compared to our study [223].

A recent case study investigating the primary tumour of a 68 year old affected with bilateral ovarian cancer and grade 2 IDC identified a novel germline *PALB2* mutation. Histological review of the breast lesion revealed presence of DCIS. The index case had a mother diagnosed with breast cancer at 42 and her sister diagnosed with DCIS at the age of 54. The *PALB2* protein truncating variant was also present in the sister [262]. Here we show strong evidence for involvement of *PALB2* protein truncating variants in the development of DCIS with the prevalence of its pathogenic germline variants reaching 1% in our unselected population of DCIS.

This is the first study that investigates the prevalence of rare variants in the context of young onset DCIS (diagnosed ≤ 50). Intersection with databases on known breast cancer predisposition variants showed that the vast majority of protein truncating variants identified in our study has been previously described as pathogenic. A major exception to this rule was a relatively common *BRCA2* variant that leads to a stop codon in exon 27, c.A9976T:p.K3326X, rs11571833. This variant was present in 29 controls and 17 cases ($MAF_{Controls}$=0.8%, MAF_{DCIS}=1.3%) and it has previously been described as benign due to the fact that it lies in the last exon of the gene and due to its high frequency in healthy individuals .This variant has been recently found to be a low penetrance variant in a cohort of familial breast cancers [263].

To conclude, we have identified enrichment in pathogenic variants in *BRCA1*, *BRCA2*, *TP53*, and *PALB2*. *CHEK2* requires a larger sample size to be able to conclude on the effect size and its importance towards DCIS development. Our data shows that if we focus on early onset individuals with one first degree relative with breast cancer we can enrich for pathogenic variants and yield a positive result for more than 20% of the cases. This finding has implications for clinical practise. Early onset DCIS cases with other affected family members could be offered genetic testing and intensive surveillance if considered necessary.

We have also shown that the patterns and prevalence of pathogenic mutations differ between the two different histological subtypes of *in situ* breast lesions. Germline mutations were found to be more common in DCIS compared to LCIS even though previous literature suggest that there is a stronger familial component or risk in LCIS compared to DCIS. Our findings do not support this notion but we have a limitation of a relatively small sample size. Additionally, the familial risk conferred in LCIS might be attributed to other genetic factors that are yet to be explored. Therefore, we cannot draw any definite conclusions towards the familiality of DCIS and LCIS even though the prevalence of pathogenic germline mutations is higher in the DCIS study.

Chapter 5 Common variation predisposing to DCIS

5.1 Introduction

Several lines of evidence from epidemiological studies point towards an inherited predisposition to DCIS. For low risk common breast cancer predisposition alleles most of the initial breast cancer association studies were not powered to identify associations with DCIS, so it is not clear whether all the low-risk susceptibility loci that have been identified are associated with DCIS and what the strength of any associations are. There are no large scale genetic/genomic studies to date, that focus on the inherited component of DCIS. However, the latest BCAC study which identified 41 loci associated with breast cancer, also looked at 2355 DCIS cases, but no novel findings that are DCIS specific were reported [154]. Nevertheless, the DCIS samples show comparable ORs to the invasive samples for the loci that were identified in this study [154]. A recent study investigating the association between 39 of the known breast cancer predisposition loci, identified rs1011970 (9p21.3, *CDKN2BAS*) to be more strongly associated with *in situ* breast cancer compared to invasive disease ($P\text{-Het}_{\text{BCIS/BC}}$ = 0.0065) [264]. This trend remained in a DCIS vs BC analysis ($P\text{-Het}_{\text{DCIS/BC}}$ = 0.021).

It is now evident that some low-risk susceptibility loci are associated with different pathological subtypes of breast cancer and support the hypothesis that breast tumour subtypes arise through distinct molecular pathways [157, 173, 265]. In order to identify further low-risk susceptibility loci, it will be necessary to look at specific morphological subtypes including DCIS as well as the cytonuclear grade and ER status of the disease. In this study we analysed 3,078 cases of pure DCIS collected through the ICICLE study and performed a meta-analysis with 2,352 *in situ* cases collected through the BCAC. Our aims were to assess whether any of the known low risk breast susceptibility alleles have different associations for DCIS and IDC, and to identify if there are any DCIS-specific low risk alleles.

It is evident that DCIS and IDC represent various phases of the same disease process, however, in this study we intend to identify potential links and differences between these stages of the disease and pinpoint potential reasons as to why not all DCIS cases progress to invasive disease.

It still remains unclear why not all DCIS progress to invasive lesions. In particular, there are concerns regarding over-diagnosis and over-treatment of DCIS through screening programmes [266]. Current methods for accurately predicting the behaviour of an individual DCIS lesion are poor, with many researchers attempting to identify molecular biomarkers that can be used to distinguish between aggressive and non-aggressive DCIS with, little success to date. The rationale for our study was to determine whether a subgroup of non-aggressive DCIS could be identified by examining low-risk genetic susceptibility loci.

In our study we examined the extent to which DCIS without associated invasive disease (5,067 cases) and IDC (24,584 cases) share low-risk susceptibility loci and whether there were any differences in the strength of the associations. We conducted subgroup case only analyses including grade, ER, and age of diagnosis stratification. We also interrogated the differences and similarities between DCIS and LCIS.

We also examined whether there are any putative novel loci associated with early stage breast cancer that would be easier to identify using a data set of DCIS rather than invasive breast cancer.

5.2 Methods

5.2.1 Clinical resource

Cases derived from the ICICLE study, a UK case-only study of DCIS. A total of 3,078 cases were genotyped with the iCOGS platform and compared to 5,000 UK controls selected from four UK studies (BBCS -1,231 controls, SBCS - 704 controls, UKBGS - 370 controls, SEARCH - 2,695 controls) participating in BCAC and already typed on the iCOGS genotyping platform. Controls were randomly selected prior to analysis, and were excluded from case-control comparisons with BCAC cases from the originating study. After excluding individuals based on genotyping quality and non-European ancestry, data for the ICICLE study available for analyses included 2,715 DCIS cases and 4,813 controls.

In a meta-analysis, the study was combined with samples from 29 studies with DCIS cases forming part of the BCAC included in the COGS Project [154].

BCAC studies recruited all types of breast cancer. Pathological information in BCAC was collected by the studies individually but combined and checked through standardized data control in a central database. A total of 2,352 cases with DCIS were identified by the central BCAC pathology database. Controls came from the 29 BCAC studies (37,654 in total).

5.2.2 Genotyping and Analysis

A detailed description of the methodology used is reported in sections 7.2 and 7.4. Genotypes were called using Illuminas proprietary GenCall algorithm and 10,000 SNPs were manually inspected to verify the algorithm calling. Individuals were excluded if genotypically non-European or not female, or had overall call rate <95% (248 cases). SNPs with a Gen-Train score of < 0.4, call rate <95% (call rate <99% if MAF <0.1) and $P_{\text{-HWE}}$<10^{-7} or evidence of poor clustering on inspection of cluster plots were excluded. All SNPs with MAF <0.01 were excluded. A cryptic relatedness analysis of the whole data set was performed using 46,789 uncorrelated SNPs and led to the exclusion of 28 cases and 18 controls due to relatedness between the ICICLE and BCAC samples (PIHAT>0.1875).

For ICICLE cases and controls, PCA was carried out on a subset of 46,789 uncorrelated SNPs and individuals or groups distinct from the main cluster (327 cases and 164 controls) were excluded using the first five PCs, Figure 5.1. Following removal of outliers, the PCA was repeated and the first five PCs were included as covariates in the analysis.

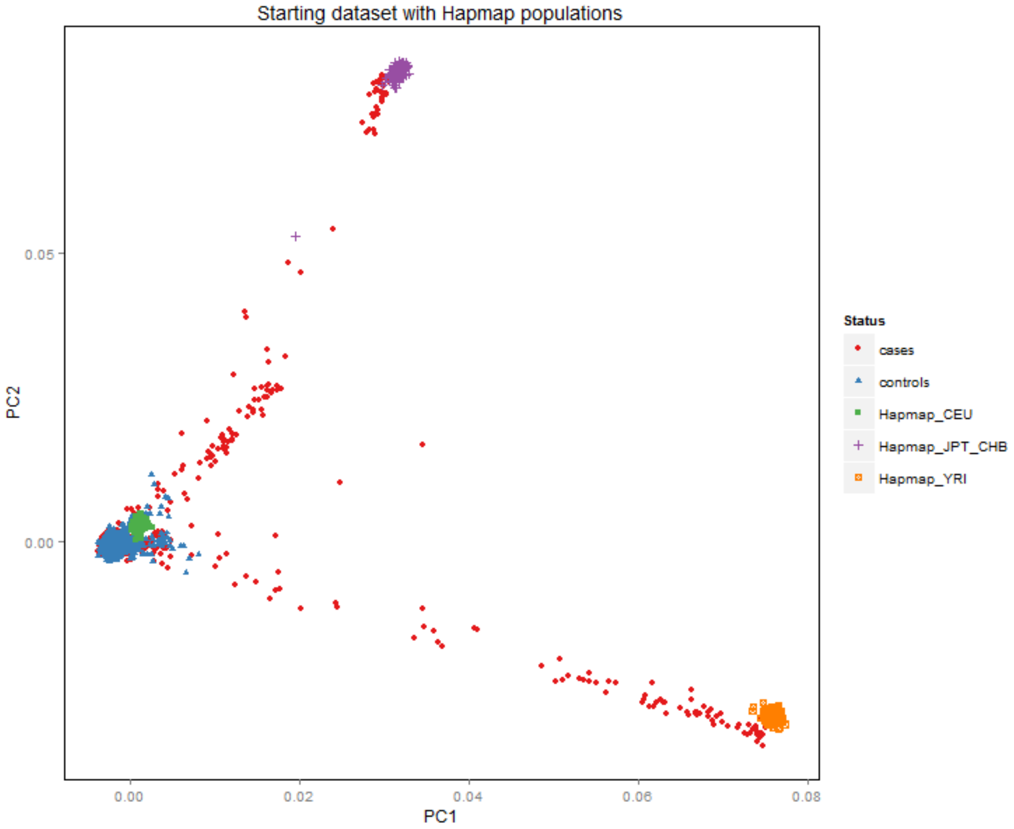


Figure 5.1: Principal component analysis indicating the genetic ancestry of the samples compared to the three Hapmap2 populations. Individuals not clustering within the European population were removed from further analyses.

The adequacy of the case-control matching was evaluated using quantile-quantile plots of test statistics and the inflation factor (λ) calculated using 37,289 uncorrelated SNPs that were not selected by BCAC and were not within one of the four common fine-mapping regions, to minimize selection for SNPs associated with breast cancer, Figure 5.2. As the majority of the SNPs on the iCOGS array are associated with breast, ovarian or prostate cancer, Figure 5.3, the SNPs selected for this analysis were taken from the set of prostate cancer SNPs, with the assumption that these SNPs were more likely to be representative of common SNPs in terms of population structure in our study, Figure 5.2. The inflation factor was λ=1.06 using this subset of SNPs whereas it increased to λ=1.09 when included the whole SNP data set.

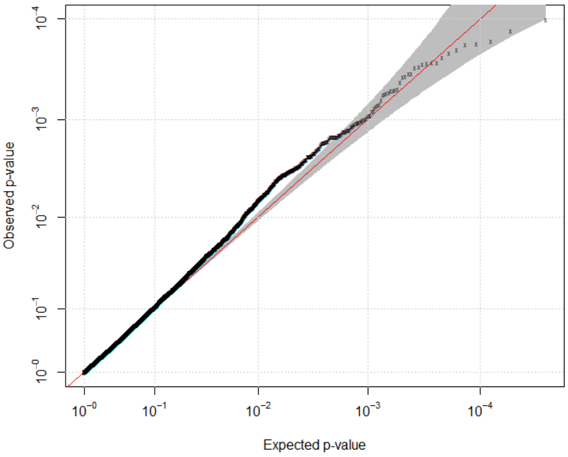


Figure 5.2: QQ-plot using 37,289 SNPs that have been selected on the basis of association with prostate cancer and have no prior evidence of association with breast cancer (**λ**=1.065).

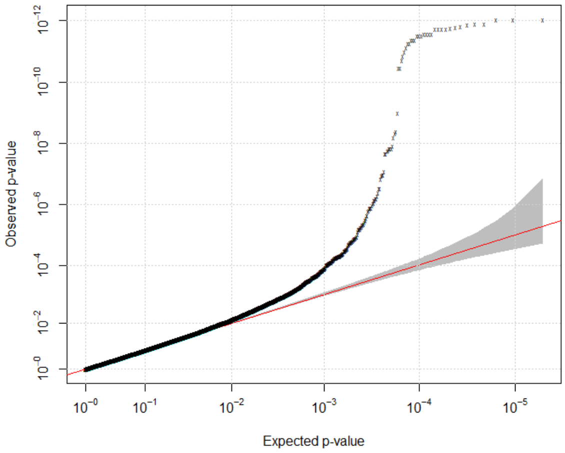


Figure 5.3: QQ-plot showing all post-qc SNPs in iCOGS platform. Inflation is due to the enrichment on breast cancer predisposition loci on the iCOGS platform (**λ**=1.092).

For each SNP, we estimated a per-allele OR and reported corresponding 95% confidence intervals by logistic regression, including the five PCs as covariates, using PLINK v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/).

Genotyping and analysis of BCAC studies have been described in detail elsewhere [154]. In brief, data were analysed using the Genotype Library and Utilities (GLU) package to estimate per-allele ORs for each SNP using unconditional logistic regression. All analyses were performed in subjects of European ancestry (determined by PC analyses) and adjusted for study and seven PCs.

All the meta-analyses were conducted by collaborators from BCAC and the summary statistics were sent back to us for analysis. Case-control ORs for DCIS cases vs controls from BCAC and ICICLE were combined using inverse variance-weighted fixed-effects meta-analysis, as implemented in METAL [267]. This was done by collaborators from BCAC. Case-only analyses were also carried out to compare genotype frequencies for (i) ER positive vs ER negative DCIS, (ii) high grade DCIS vs low and intermediate grade DCIS, and (iii) DCIS vs IDC, and were used as a test for heterogeneity of ORs by tumor subtype. Only studies with data on both subtypes contributed to case-only analysis comparing these subtypes. Similar case-only analyses were performed for the IDC cases in these studies to assess whether any heterogeneity evident in DCIS also occurred in IDC.

For the known breast cancer predisposition loci $P<0.00066$ was considered statistically significant (Bonferroni correction for multiple testing on 76 known loci). All of the known breast cancer susceptibility loci at the time of analysis were included in the iCOGS chip with the exception of rs2284378 (20q11) that was identified as an ER negative breast cancer predisposition SNP after the iCOGS chip was developed [168].

5.3 Known variants predisposing to DCIS

Fifty-one of the 76 known common breast cancer susceptibility loci genotyped on the iCOGS array showed an association at $P<0.05$ with DCIS, with effect in the same direction as previously reported in IDC (Figure 5.4, Figure 5.5). Sixteen SNPs showed a significant association with DCIS at $P<0.00066$ with three reaching genome-wide significance ($P<5\text{x}10^{-8}$), Table 5.1. The strongest associations were with loci in *FGFR2* (rs2981579, OR=1.29, 95%CI=1.24-1.35, $P=9.0\text{x}10^{-30}$) and *TOX3* (rs3803662, OR=1.15, 95%CI=1.1-1.21, $P=1.7\text{x}10^{-8}$). For the majority of known loci (N=46) the risk allele for invasive breast cancer is the minor allele. For the ORs presented on the forest plot (Figure 5.5) the reference allele was set as the non-risk allele in order to determine whether the association with DCIS was in the same direction as previously published for invasive breast cancer. Thus ORs for DCIS will be >1 if in the same direction as invasive disease and <1 if in the opposite direction, Appendix 5.

Table 5.1: Three out of the 76 previously published loci are associated with DCIS at $P<5\text{x}10^{-8}$.

Chr	SNP	Locus	MAF	OR (95%CI)	*P*
10	rs2981579	*FGFR2*	0.4	1.29 (1.24-1.35)	$1.16\text{x}10^{-29}$
10	rs2981582	*FGFR2*	0.38	1.28 (1.22-1.34)	$2.66\text{x}10^{-27}$
16	rs3803662	*TOX3*	0.26	1.15 (1.1-1.21)	$1.91\text{x}10^{-08}$

The case-only analysis (DCIS vs IDC) confirmed the shared genetic susceptibility between DCIS and IDC as none of the heterogeneity *P* values were significant after a Bonferroni adjustment for 76 SNPs. Five loci that show a *P*-Het<0.05 are indicated with red in Figure 5.4, Table 5.2. One of those 5 variants, rs4245739 (*MDM4*), with a MAF=0.26 shows a stronger association with DCIS OR=1.09 (95% CI 1.04, 1.14), *P*=0.00073 as opposed to IDC with OR=1.02 (95%CI 1.00, 1.05), *P*=0.087, without however reaching statistical significance after multiple testing, *P*-het=0.017.

Table 5.2: Five loci showing a borderline differential association between DCIS and IDC. None of these loci reached a Bonferroni corrected *P*<0.00066.

Chr	**SNP**	**Locus**	**RAF**	**DCIS OR (95% CI)**	**DCIS *P***	**IDC OR (95% CI)**	**IDC *P***	***P-Het***
1	rs11249433	1p11.2	0.41	0.98 (0.94, 1.02)	0.36	1.08 (1.05, 1.11)	$4.4x10^{-9}$	0.011
1	rs4245739	*MDM4*	0.26	1.09 (1.04, 1.14)	0.00073	1.02 (1.00, 1.05)	0.086	0.017
10	rs7072776	*DNAJC1*	0.28	1.00 (0.95, 1.05)	0.96	1.06 (1.03, 1.09)	$3.5x10^{-5}$	0.036
9	rs10759243	9q31.2	0.28	1.03 (0.98, 1.08)	0.23	1.06 (1.03, 1.09)	$7.4x10^{-5}$	0.041
3	rs12493607	*TGFBR2*	0.34	0.99 (0.95, 1.04)	0.80	1.07 (1.04, 1.09)	$1.4x10^{-6}$	0.044

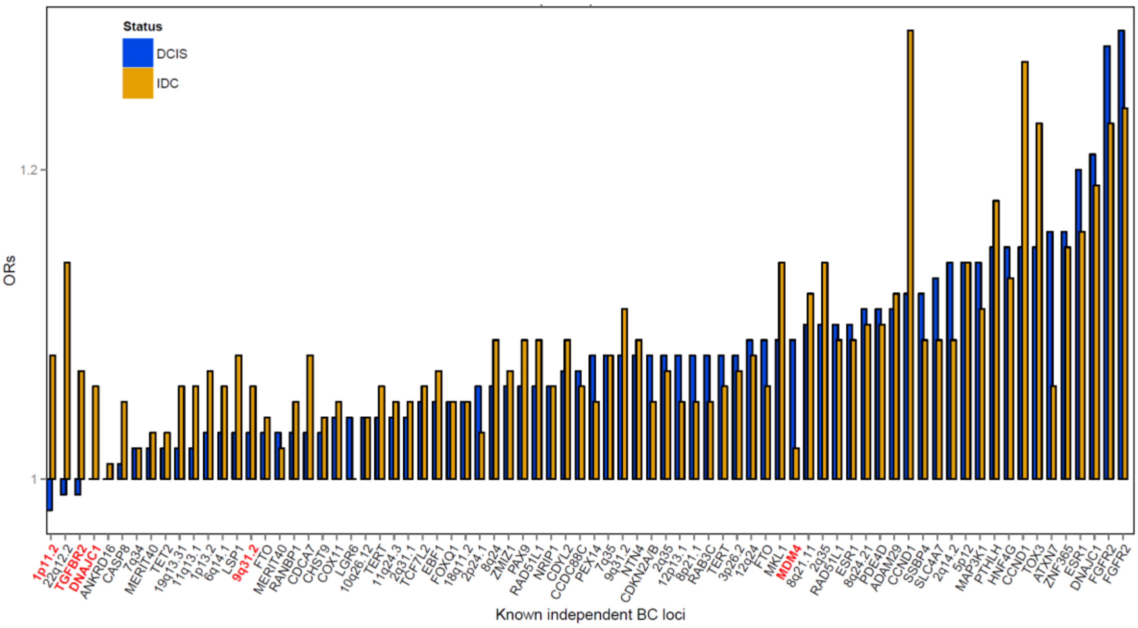


Figure 5.4: Effect size in form of OR of known breast cancer predisposition loci for DCIS (blue) and IDC (yellow). Highlighted in red are the five loci that show a borderline differential association between the two groups.

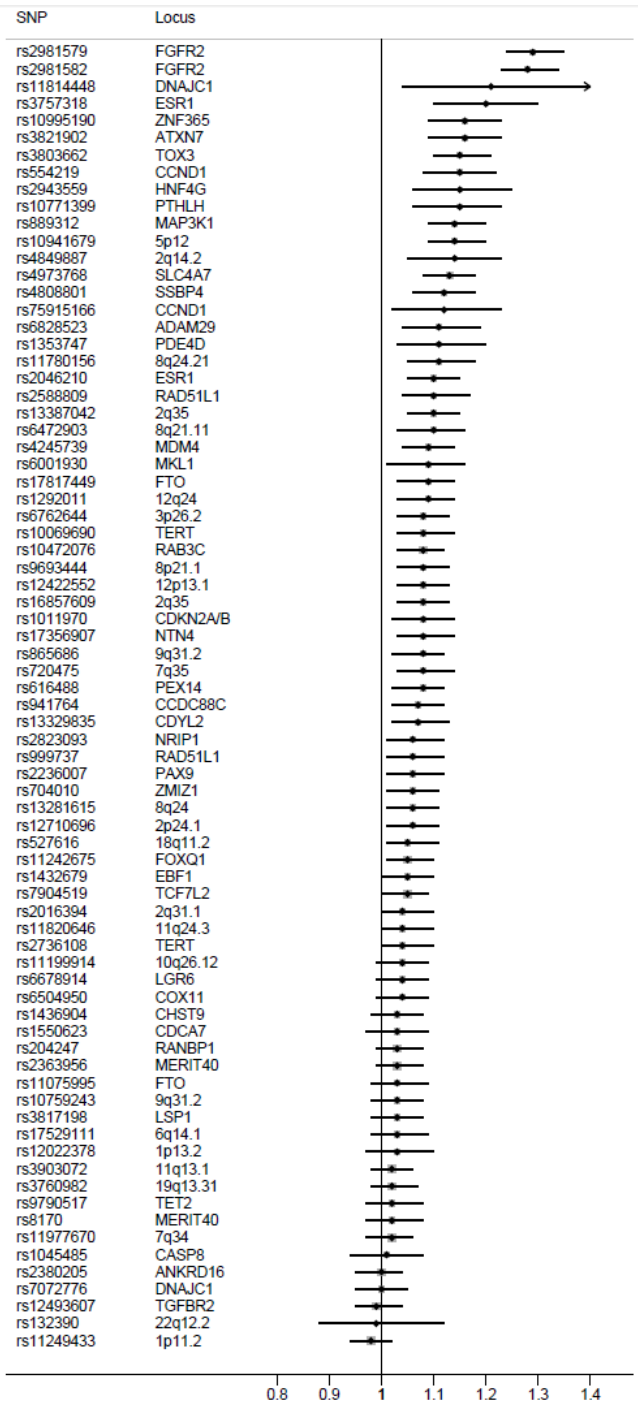


Figure 5.5: Forest plot showing ORs and 95% CI for DCIS on 76 known breast cancer predisposition SNPs.

5.4 Novel variants predisposing to DCIS

Novel SNPs showing the strongest evidence for association with DCIS (*P*<6x10^{-6}) in the meta-analysis (after excluding previously reported loci) were genotyped in a Phase II analysis at LGC Genomics. The Phase II samples consisted of 653 DCIS cases from the ICICLE and Breakthrough Generation Studies and 1,882 controls from the ICICLE study not previously genotyped on the iCOGS chip. All individuals included in the analysis were of European ancestry (self-reported).

All SNPs reaching genome-wide significance (P<5x10^{-8}) in the meta-analysis were correlated with one of the known breast cancer predisposition loci. In particular 5 previously published regions reached genome-wide significance (2 independent *FGFR2* loci, the *MAP3K1* locus, the *NEK10* locus, and the *TOX3* locus). There were three SNPs that were not correlated with known loci at P<6x10^{-6}, Figure 5.6, all with very little evidence of an association with IDC, Table 5.3.

Table 5.3: Three loci showing a suggestive association with DCIS.

SNP	**rs12631593**	**rs13236351**	**rs73179023**
Chromosome	3	7	22
Position	60701884	97772513	43424477
Nearest genes	*FHIT*	*LMTK2*	*PACSIN2:TTLL1*
MAF	0.11	0.032	0.13
	ICICLE DCIS phase I		
OR (95% CI)	1.15 (1.04, 1.28)	1.31 (1.10, 1.56)	0.83 (0.75, 0.91)
P	0.0088	0.0029	0.00020
	BCAC DCIS		
OR (95% CI)	1.25 (1.14, 1.36)	1.3 (1.12, 1.51)	0.86 (0.79, 0.94)
P	1.0x10^{-6}	0.00060	0.0012
	Meta-analysis phase I		
OR (95% CI)	1.21 (1.13, 1.29)	1.3 (1.16, 1.46)	0.85 (0.79, 0.90)
P	5.5x10^{-8}	5.7x10^{-6}	1.1x10^{-6}
	Phase II DCIS		
OR (95% CI)	0.93 (0.76, 1.14)	0.91 (0.63, 1.31)	0.95 (0.78, 1.15)
P	0.49	0.61	0.57
	Meta-analysis phase II		
OR (95% CI)	1.18 (1.10, 1.25)	1.26 (1.13, 1.41)	0.86 (0.80, 0.91)
P	7.8x10^{-7}	2.9x10^{-5}	1.7x10^{-6}
	BCAC IDC		
OR (95% CI)	1.01 (0.97, 1.05)	1.05 (0.99, 1.13)	0.97 (0.93, 1.00)
P	0.54	0.13	0.060
DCIS vs IDC *P*-Het	0.0048	0.17	0.0099

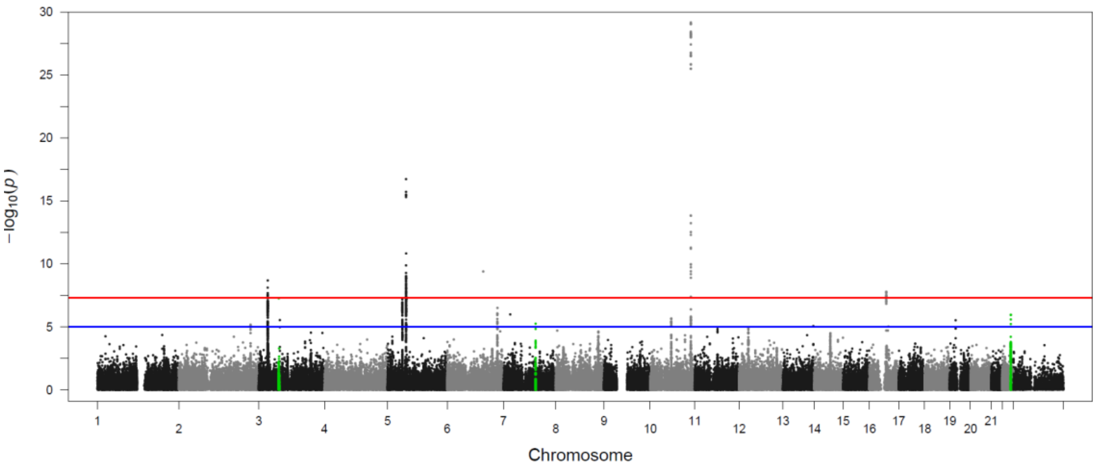


Figure 5.6: Manhattan plot for DCIS vs controls analysis. Three suggestive novel regions are highlighted in green.

Of these novel SNPs, rs12631593, 3p14.2, (an intronic variant in *FHIT*, chr3: 60726844) showed the strongest association with DCIS (OR=1.21, 95%CI=1.13-1.29, P=5.5x10^{-8}). This

SNP showed little association with IDC (OR=1.01, 95%CI=0.97-1.05, *P*=0.54) and this was supported by the case-only analysis (*P*-Het$_{DCIS/IDC}$=0.0048).

The second locus was on 22q13.2, rs73179023 (DCIS only: OR=0.85, 95%CI=0.79-0.90, *P*=1.11 x10^{-6}; IDC only: OR=0.97, 95%CI=0.93-1.00, *P*=0.06, *P*-Het$_{DCIS/IDC}$=0.0099).

Finally, the third locus was on and 7q21.3, rs13236351 (DCIS only: OR=1.30, 95%CI=1.16-1.46, *P*=5.71x10^{-6}; IDC only: OR=1.05, 95%CI=0.99-1.13, *P*=0.13, *P*-Het$_{DCIS/IDC}$=0.172).

These SNPs were genotyped in a validation study including further 653 DCIS cases and 1,882 controls, however for all three loci there was no evidence of an association (rs12631593, rs13236351, and rs73179023 *P*=0.49, 0.61, and 0.57 respectively) and none reached genome-wide significance following a meta-analysis of all data (*P*=7.8x10^{-7}, 2.9x10^{-5}, and 1.7x10^{-6} respectively), Table 5.3. The power we had to detect genome wide significance of variants with MAF=0.2 and an effect size of OR=1.2 was 96%.

5.4.1 Imputation to fine map novel loci

Using data from 2,715 DCIS cases from ICICLE along with 4,813 controls from BBCS, SBCS, UKBGS, and SEARCH through BCAC, and utilising 186,038 variants from iCOGS we imputed 38,016,337 variants using impute2, section 7.4.2. The genotypes were imputed based on haplotype information based on the version 3 of 1000 genomes phase I data set. From those variants 11,301,527 were imputed with an info score >0.5. These variants were then filtered based on their *P* value and 105,231 variants had *P*<0.01. A total of 2,651 variants that were genotyped on iCOGS were on that list whereas the remaining were imputed with an info score >0.5. The three suggestive novel DCIS predisposition loci were interrogated in detail. The three regions of interest are plotted using LocusZoom, Figure 5.7, Figure 5.8, Figure 5.9 [268].

The region surrounding rs12631593 in chromosome 3 is not very well captured in iCOGS and there were only 17 SNPs genotyped in the region of 200KB, Figure 5.7. From the imputed SNPs, the one with the lowest *P* value was rs73836108, OR= 1.27 (95%CI 1.07, 1.51), *P*=0.00034, Table 5.4.

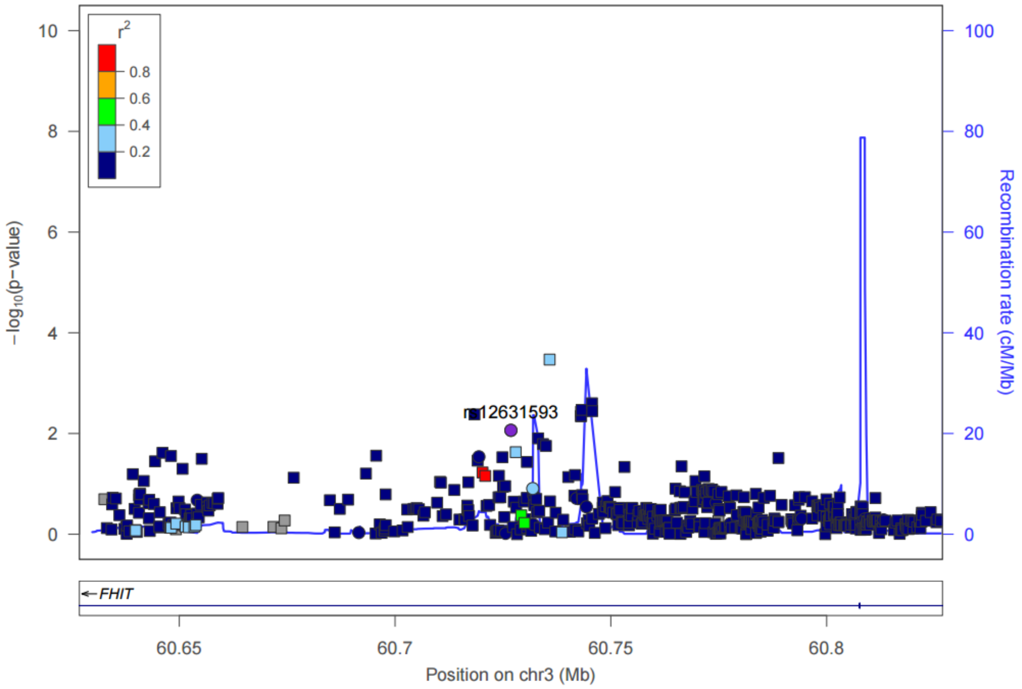


Figure 5.7: Fine-mapping of putative novel DCIS specific region on chromosome 3. Circles correspond to genotyped SNPs whereas squares correspond to imputed SNPs. The SNP showing the strongest association in the genotyping-based meta-analysis is highlighted in purple.

In the region 100KB upstream, and downstream of rs13236351 in chromosome 7, there were 138 genotyped variants and another 531 imputed with info score >0.5, Figure 5.8. The SNP with the lowest *P* value was rs12704976 as indicated in Table 5.4.

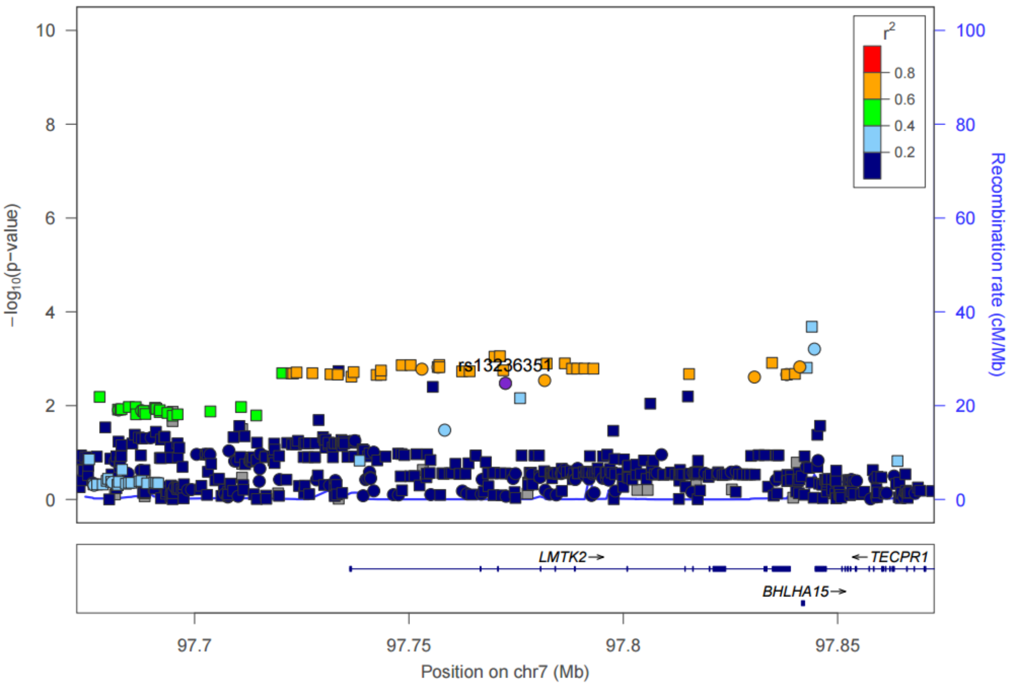


Figure 5.8: Fine-mapping of putative novel DCIS specific region on chromosome 7. Circles correspond to genotyped SNPs whereas squares correspond to imputed SNPs. The SNP showing the strongest association in the genotyping-based meta-analysis is highlighted in purple.

Finally, a total of 637 variants were imputed with info score>0.5 in the region spanning 100 KB each side of rs73179023 in chromosome 22. The imputed SNP with the strongest association was rs9611991 which is an intronic variant on the *TTLL1* gene (OR=0.83 95%CI 0.77, 0.89, *P*=2.67x10^{-7}), Figure 5.9, Table 5.4.

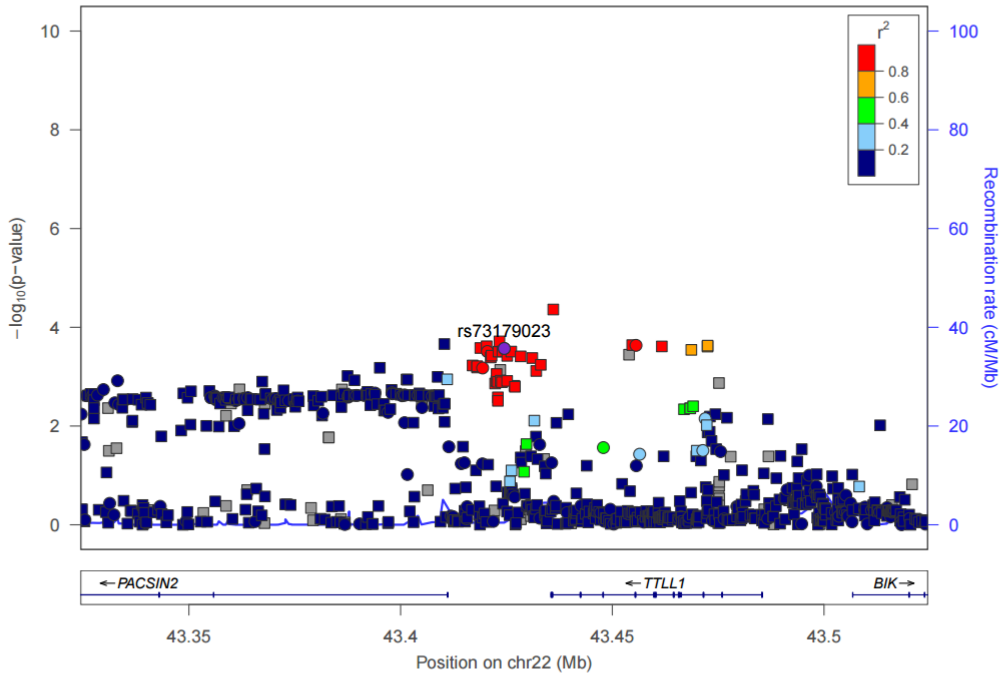


Figure 5.9: Fine-mapping of putative novel DCIS specific region on chromosome 22. Circles correspond to genotyped SNPs whereas squares correspond to imputed SNPs. The SNP showing the strongest association in the genotyping-based meta-analysis is highlighted in purple.

Overall, using imputation method we identified a stronger association in all three regions under investigation, Table 5.4. However, this is a preliminary analysis, including only data from the ICICLE study. Since those three suggestive loci resulted from a meta-analysis including BCAC samples, it would be reasonable to conduct this meta-analysis using the imputed data and investigate whether any of the associated loci reaches genome wide significance. If there are any significant findings, they would need to be replicated in further studies.

Table 5.4: Genotyped and imputed SNPs in the three putative novel DCIS predisposition loci. Imputed SNPs are highlighted in bold. Info score measures the quality of the imputation and r^2 is a measure of LD between the genotyped and imputed SNPs at each locus.

SNP	Chr	Distance from genotyped SNP	r^2	MAF	Info score	ICICLE OR	*P*
rs12631593	3	-		0.11	-	1.15 (1.04, 1.28)	0.0088
rs73836108	3	8,968	0.40	0.04	0.65	1.27 (1.07, 1.51)	0.00034
rs13236351	7	-		0.04	-	1.31 (1.10, 1.56)	0.0029
rs12704976	7	71,466	0.41	0.07	0.95	1.28 (1.13, 1.46)	0.00021
rs73179023	22	-		0.13	-	0.83 (0.75, 0.91)	0.0002
rs9611991	22	11,578	0.43	0.12	0.95	0.81 (0.73, 0.90)	4.3x10^{-5}

5.5 Grade specific analysis

Grade data were available for 95% of ICICLE DCIS cases; 1,635 (60%) were high cytonuclear grade and 943 (35%) were low/intermediate grade. The grade data on the BCAC DCIS were less complete with data only available on 35% of cases: 306 (13%) high grade and 522 (22%) low/intermediate grade cases, Table 5.5. A case-control analysis was performed on the low/intermediate and high grade subsets separately and a case-only analysis of low/intermediate grade vs high grade DCIS was performed in order to assess whether any of these loci are grade specific after a Bonferroni correction (*P*<0.00066).

Table 5.5: DCIS cases from the ICICLE and BCAC studies stratified by grade.

Grade	N of ICICLE cases	N of BCAC cases
High	1635	306
Intermediate	693	247
Low	250	275
Missing	137	1524
Total	**2715**	**2352**

Analysis of DCIS by grade revealed that although the majority of SNPs predispose to all grades of DCIS some are grade specific, Table 5.6. The two SNPs close to *CCND1* showed a strong association with low/intermediate grade DCIS (rs75915166, OR=1.36, 95%CI=1.17-1.59, P=7.2x10^{-5}; rs554219, OR=1.32, 95%CI=1.18-1.48, P=8.2x10^{-7}) and no association with high grade DCIS. Case-only analysis confirmed that these loci were low/intermediate grade specific (rs75915166, P-Het$_{low/highgrade}$=0.00014; rs554219, P-Het$_{low/highgrade}$=0.00013) and this was independent of ER status (adjusted for ER status rs75915166, *P*=0.0050; rs554219, *P*=0.019).

Table 5.6: Grade specific associations for two *CCND1* variants reaching the Bonferroni corrected *P*<0.00066 and a further 3 variants reaching nominal significance of *P*<0.05.

			Low/intermediate grade		High grade		low/ inter vs high grade
SNP	**Chr**	**Locus**	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***	***P*-Het**
rs554219	11	*CCND1*	1.32 (1.18, 1.48)	8.2x10^{-07}	1.02 (0.91, 1.14)	0.75	0.00013
rs75915166	11	*CCND1*	1.36 (1.17, 1.59)	7.2x10^{-05}	0.92 (0.79, 1.08)	0.31	0.00014
rs7072776	10	*DNAJC1*	1.09 (1.00, 1.18)	0.051	0.95 (0.87, 1.03)	0.18	0.0019
rs10941679	5	5p12	1.26 (1.15, 1.37)	2.1x10^{-07}	1.09 (1.01, 1.19)	0.030	0.0033
rs10995190	10	*ZNF365*	1.05 (0.94, 1.18)	0.34	1.30 (1.16, 1.43)	2.1x10^{-06}	0.017

A similar-case-only analysis of IDC by grade that was conducted by BCAC collaborators confirmed that the two SNPs on 11q13.3 close to *CCND1* were also invasive grade 1/2 specific in IDC (rs75915166, OR=1.42, P=1.7x10^{-30}, P-Het=2.8x10^{-10}; rs554219, OR=1.39, P=4.7x10^{-49}, *P*-Het=$1.3\text{x}10^{-17}$) and again were independent of ER status (*P*=$1.3\text{X}10^{-6}$, *P*=$1.6\text{x}10^{-6}$, respectively).

rs10941679, 5p12 showed a borderline association with low/intermediate grade DCIS (OR=1.26, *P*=$2.1\text{x}10^{-7}$, *P*-Het$_{\text{low/highgrade}}$=0.0033). This locus has previously been shown to be associated with low grade PR positive IDC [252]. A variant at the *ZNF365* locus shows evidence of a stronger association with high grade DCIS. This has already been observed in overall breast cancer (BCAC Website).

5.5.1 Novel variants with evidence of association with high grade DCIS

In an attempt to identify novel common variants associated with specific sub-phenotypes of DCIS, we conducted a case control analysis including only cases with high grade DCIS. This analysis including 1,635 cases and 4,813 controls yielded a possible novel locus at chromosome 17, Figure 5.10. One of the variants in the association linkage disequilibrium (LD) block (rs9302935, *P*=$2.4\text{x}10^{-5}$) was selected to be genotyped on a phase II study where 172 additional high grade DCIS cases were screened along with 1,882 controls. Additionally, data from BCAC was added to the meta-analysis. The final data set that was used for the meta-analysis is reported in Table 5.7. Genome wide significance was not reached. Due to the fact that the variant is relatively rare MAF<5%, additional samples could be used in order to establish whether the observed nominal association stands true, Table 5.8. The observed association is driven primarily by the ICICLE study, even though all studies have the effect on the same direction. Screening additional cohorts of high grade DCIS could elucidate whether this association stands true.

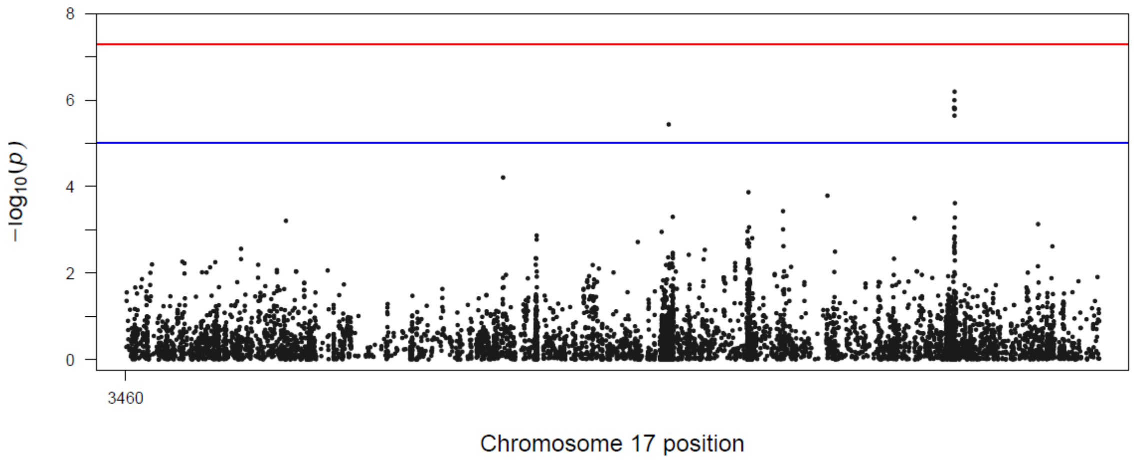


Figure 5.10: Association plot on chromosome 17 for high grade DCIS. Rare variants in region near the *CASC17* gene. Shows a suggestive signal with high grade DCIS.

Table 5.7: Number of individuals used for the single SNP (rs9302935) analysis in the putative novel locus on chromosome 17.

Study	Controls	Cases	Total
ICICLE	4,813	1,635	6,448
BCAC	27,389	306	27,695
ICICLE phase II	1,882	55	1,937
Total	34,084	1,996	36,080

Table 5.8: Meta-analysis results for rs9302935, a putative locus for high grade DCIS.

Study	MAF	OR	95% CI	*P* value
ICICLE	0.032	1.54	(1.26, 1.89)	2.4x10^{-5}
BCAC	0.038	1.46	(0.96, 2.22)	0.08
Phase II	0.039	1.17	(0.46, 2.94)	0.75
Meta-analysis	0.034	1.51	(1.26, 1.81)	5.5x10^{-6}

5.6 ER specific analysis

Following immunohistochemistry for ER of the ICICLE study samples, 1,484 (54%) cases were classified as ER positive and 383 (14%) as ER negative. The ER data on BCAC DCIS were less complete with 664 (28%) ER positive, 301 (13%) ER negative and 1,387 (59%) ER unknown cases, Table 5.9.

Table 5.9: DCIS cases from ICICLE and BCAC studies stratified by ER status.

ER status	N of ICICLE cases	N of BCAC cases
Positive	1484	664
Negative	383	301
Missing	848	1387
Total	**2715**	**2352**

One SNP, rs527616, on chromosome 18q11.2 reached the Bonferroni corrected significance threshold (76 tests) with a *P*-Het=0.00036 for differential association between ER positive and ER negative DCIS, Table 5.10. This is one of the known ER positive loci in the context of invasive breast cancer. Our results indicate that DCIS behaves in a similar manner to IDC in terms of genetic predisposition stratified by ER status.

Table 5.10: Seven known predisposition loci showing evidence for differential association between ER positive and ER negative DCIS at *P*<0.05.

			ER positive DCIS		**ER negative DCIS**		**ER+ vs ER-**
SNP	**Chr**	**Locus**	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***	***P*-Het**
rs527616	18	18q11.2	1.14 (1.06, 1.22)	0.00026	0.88 (0.79, 0.98)	0.023	0.00036
rs554219	11	*CCND1*	1.24 (1.12, 1.36)	$1.8x10^{-05}$	0.94 (0.80, 1.12)	0.51	0.0078
rs11977670	7	7q34	1.06 (0.99, 1.13)	0.087	0.93 (0.84, 1.04)	0.22	0.021
rs6678914	1	*LGR6*	1.01 (0.95, 1.09)	0.67	1.14 (1.02, 1.28)	0.020	0.029
rs9693444	8	8p21.1	1.08 (1.01, 1.16)	0.031	1.22 (1.08, 1.36)	0.00084	0.039
rs10771399	12	*PTHLH*	1.23 (1.11, 1.37)	0.00014	1.04 (0.88, 1.23)	0.65	0.049
rs10995190	10	*ZNF365*	1.16 (1.06, 1.28)	0.0016	1.37 (1.16, 1.61)	0.00013	0.049

5.7 Age specific analysis

In order to interrogate whether any of the known breast cancer predisposition loci is associated with the onset of DCIS, we conducted a case-only analysis separating individuals with at the cut-off of 50 years of age. There were 573 DCIS cases from ICICLE with age of diagnosis <50, and 2,003 diagnosed at the age of 50 or over. This data set was combined with 410 individuals with age of diagnosis <50 and 1,648 with age of diagnosis ≥ 50, Table 5.11. Six out of the 76 known breast cancer predisposition loci showed a borderline differential association between early (< 50) and late (≥ 50) onset of DCIS, Table 5.12. One of the variants, rs527616-18q11.2, reached the Bonferroni corrected *P* value with a *P*-Het=0.0003.

Table 5.11: DCIS cases from ICICLE and BCAC studies stratified by onset of disease.

Age of diagnosis	**N of ICICLE cases**	**N of BCAC cases**
< 50	573	410
≥ 50	2003	1648
Missing	139	294
Total	**2715**	**2352**

Table 5.12: Six known breast cancer predisposition loci that show evidence for differential association between early and late onset DCIS. ICICLE and BCAC ORs are indicated separately and *P*-het corresponds to combined ICICLE and BCAC age stratified analysis of age of diagnosis ≥ 50 versus < 50.

Chr	**SNP**	**RAF**	**50 ≥ BCAC**	**50 ≥ ICICLE**	**< 50 BCAC**	**< 50 ICICLE**	***P*-Het**
18	rs527616	0.63	1.00 (0.93, 1.06)	0.95 (0.88, 1.02)	0.91 (0.79, 1.05)	0.75 (0.66, 0.86)	0.0003
16	rs3803662	0.26	1.13 (1.05, 1.21)	1.14 (1.05, 1.24)	1.35 (1.15, 1.58)	1.28 (1.12, 1.47)	0.0041
2	rs13387042	0.51	0.86 (0.80, 0.92)	0.91 (0.85, 0.98)	0.87 (0.75, 1.00)	1.08 (0.96, 1.22)	0.015
14	rs2236007	0.79	0.95 (0.88, 1.03)	0.98 (0.90, 1.08)	0.91 (0.76, 1.08)	0.82 (0.70, 0.96)	0.016
7	rs11977670	0.43	1.02 (0.96, 1.09)	0.96 (0.89, 1.03)	1.10 (0.96, 1.27)	1.12 (0.99, 1.27)	0.030
3	rs6762644	0.4	1.09 (1.02, 1.17)	1.03 (0.95, 1.11)	0.99 (0.86, 1.15)	1.26 (1.11, 1.43)	0.039

5.8 *In situ* comparison

In an attempt to distinguish the signals from known breast cancer predisposition loci we conducted an *in situ* case only analysis, comparing cases with DCIS with cases with LCIS. This analysis included 1,484 ER positive DCIS cases along with 231 LCIS cases. Figure 5.11 shows the difference in effect size for LCIS and ER positive DCIS cases in 76 known breast cancer predisposition loci.

Table 5.13 shows the SNPs that had a *P*-het<0.05 between ER positive DCIS and LCIS. Four of these loci rs865686, rs12710696, rs17529111, and rs11249433 seem to be associated with LCIS whereas the remaining one, rs4973768 at the *SLC4A7* locus seems to be associated with ER positive DCIS. In the case control analyses there were 1,484 ER positive DCIS cases and 4,813 controls, and 231 LCIS cases and 4,755 controls. The only SNP that passes the Bonferroni correction of *P*<0.00066 is rs865686 at chromosome 9q31.2. This variant seems to be strongly associated with LCIS (OR=1.57 (95%CI 1.28, 1.92) *P*=1.8x10^{-5}, but not with ER positive DCIS.

Table 5.13: Five SNPs showing evidence for differential association between ER positive DCIS and LCIS.

			ER+ DCIS		LCIS		
SNP	**Chr**	**Locus**	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***	***P*-Het**
rs865686	9	9q31.2	1.07 (0.98, 1.16)	0.14	1.57 (1.28, 1.92)	1.8x10^{-5}	0.00065
rs12710696	2	2p24.1	0.99 (0.90, 1.08)	0.76	1.29 (1.06, 1.56)	0.0096	0.0066
rs4973768	3	SLC4A7	1.19 (1.09, 1.29)	5.4x10^{-5}	0.94 (0.78, 1.14)	0.53	0.018
rs17529111	6	6q14.1	1.01 (0.91, 1.12)	0.79	1.33 (1.07, 1.65)	0.011	0.022
rs11249433	1	1p11.2	0.99 (0.91, 1.08)	0.90	1.22 (1.01, 1.47)	0.038	0.049

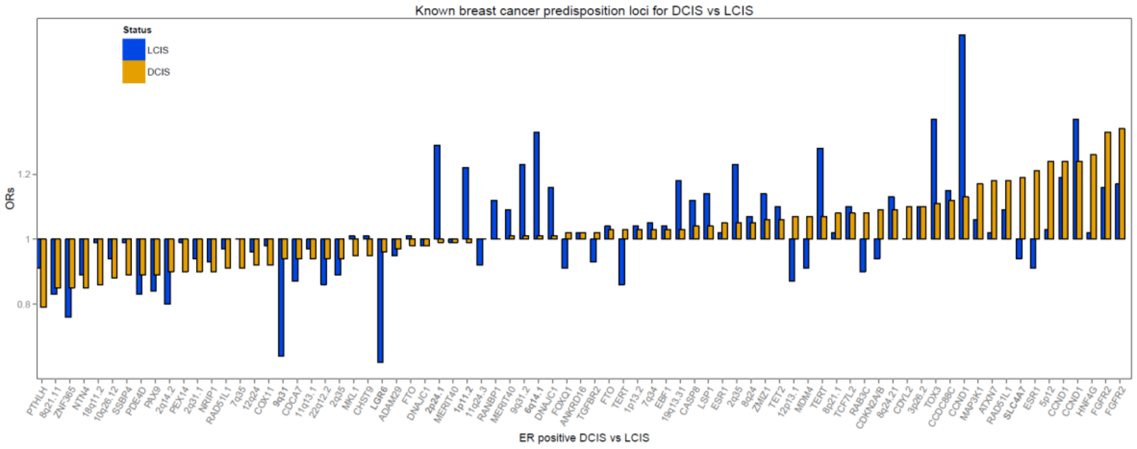


Figure 5.11: Effect size for ER positive DCIS versus LCIS based on 76 known breast cancer predisposition loci.

5.9 Discussion

This study provides the strongest evidence to date for a shared genetic susceptibility between DCIS and IDC, based on 5,067 cases with pure DCIS (no invasive disease) and 24,670 cases with IDC. It differs from previous BCAC analyses of DCIS, as it has included an aditional 3,078 DCIS cases, excluded all cases of pure LCIS and has also compared DCIS to IDC rather than all invasive disease. We found no significant differences, although there were some DCIS-specific loci that did not replicate or reach genome-wide significance. Therefore, our data largely support the theory that DCIS and IDC are a continuum of the same disease, without excluding the possibility that there may be some low-risk susceptibility loci that have a strong association with stage 0 (DCIS) and a weaker association with other stages of breast cancer. Identifying any such loci is important as it would identify a subset of DCIS that has a low risk of progression.

An important finding of this study is the lack of DCIS / IDC specific loci among the known breast cancer predisposition loci. Of the five breast cancer predisposition alleles originally reported by Easton *et al* [153], three were shown to be associated with *in situ* (998 cases of DCIS and LCIS) disease (rs2981582-*FGFR2*, rs3803662-*TOX3*, rs889312-*MAP3K1*) with rs889312 showing a stronger association with DCIS (*P*-trend 0.007, per allele OR 1.30 for DCIS, per allele OR 1.13 for invasive disease). However this finding of potential DCIS specific loci was not confirmed in the UK Million Women Study which found no differential association with DCIS *vs* IDC for twelve breast cancer susceptibility loci, including rs889312, although their sample size was smaller (873 DCIS and 4,959 IDC) [237]. In the recent BCAC COGS analysis all 41 novel SNPs identified on the iCOGS chip showed comparable ORs for invasive and *in situ* disease (based on data from 2,335 *in situ*, and 42,118 invasive cases), with the exceptions of rs12493607 (*TGFBR2*), and rs3903072 (11q13.1), for which associations seemed to be restricted to invasive disease [154]; however we found no evidence of an IDC specific association for these loci after correcting for multiple testing. We have also shown for the first time that seven of the known invasive breast cancer predisposition loci not previously shown to have an association with DCIS show comparable ORs for IDC and DCIS: rs4973768 (*SLC4A7),* rs3821902 (*ATXN7*) [178], rs10995190 (*ZNF365*), rs554219 (*CCND1*), rs3757318 and rs2046210 (*ESR1*).

This lack of DCIS / IDC specific loci is in contrast to our previous study of lobular cancer where we showed that there were loci that were specific to invasive lobular cancer (ILC), showing no association with LCIS and there was also the suggestion of the LCIS specific loci [173]. When we compare the DCIS data presented here to our previous LCIS analyses it reveals that there is some overlap between loci that are associated with ER positive DCIS and LCIS. However there are also some differences: rs6678914, *LGR6* and rs865686, 9q31.2 show a strong association with LCIS but little evidence of an association with ER positive DCIS (*P*-Het$_{\text{DCIS/LCIS}}$ = 7.4x10^{-5} and 6.6x10^{-4} respectively) Table 5.13, Figure 5.11. It has been previously shown that rs865686 is associated with ER positive IDC, however we do not observe this association on DCIS [269].

We have also previously shown that rs11249433, 1p11.2 and rs11977670, 7q34 have a stronger association with invasive lobular cancer than IDC [173]. These loci showed only a weak association with LCIS and no association with ER positive DCIS in this analysis.

Most association studies of invasive breast cancer perform subgroup analyses based on ER status. In contrast to invasive breast cancer, ER status is not routinely assessed in DCIS in all centres, despite the NSABP B-24 trial showing a benefit from endocrine therapy in ER positive DCIS [270]. A national audit of DCIS in the UK revealed that only 50% of DCIS cases had ER status assessed and ER positivity in low and intermediate grade DCIS was significantly more common than in high grade DCIS ($P< 0.001$) (ER positive high grade 69%, intermediate grade 94%, low grade 99%) [271]. In order to overcome this issue we performed ER immunohistochemistry on the samples from ICICLE without ER status. However there was still a large amount of data on ER status missing from the BCAC cases, resulting in only 684 ER negative DCIS cases being available for analyses making it difficult to draw definitive conclusions about ER negative DCIS. In essence, the findings are similar to invasive breast cancer, with ER negative and ER positive DCIS having different genetic susceptibility profiles and ER positive DCIS having a very similar profile to ER positive IDC.

Cytonuclear grade of DCIS is used by many clinicians to select those cases most likely to benefit from radiotherapy despite the fact that grade has not been shown to be a good predictor of recurrence. In the UK audit of DCIS, grade data were available for 99% of DCIS with 59% classified as high grade, 29% as intermediate and 11% as low grade [271]. Similarly, in our study, data on grade were available for 95% cases in ICICLE. In invasive disease only a minority of predisposition loci have been shown to be grade specific; rs2981582 (*FGFR2*) and rs13281615 (8q24) [272, 273] and rs10941679 (5p12) [252]. We have shown that analysis of DCIS by grade revealed other known loci that are grade specific. The loci with the strongest association with grade were SNPs on 11q13, showing a stronger association with low/intermediate grade DCIS and IDC, than high grade lesions. The finding of a strong association with low and intermediate grade ductal carcinomas that is independent of ER status in both DCIS and IDC for these loci is novel. Variant rs614367 was the first locus on 11q13 shown to be associated with invasive breast cancer [274]. Fine mapping of the region subsequently showed two independent signals (rs554219 and rs78540526, $r^2= 0.38$), which are the loci reported in this analysis. Functional analyses demonstrated that the risk variants modify enhancer and silencer elements with the likely target gene being *CCND1* [182]. Both SNPs map to a transcriptional enhancer element and the risk alleles apart from increasing the risk of developing breast cancer, also reduce the binding of ELK4 transcription factor and luciferase activity in reporter assays, and can therefore be associated with low cyclin D1 protein levels in tumours.

A study of 150 cases of subsequent breast cancer (invasive and *in situ*) after DCIS showed a significant association for both grade and ER status between the index DCIS and the subsequent breast cancer (whether ipsilateral or contralateral), suggesting that women with DCIS are at risk of developing subsequent breast cancers of a similar phenotype [275]. This finding supports the genetic predisposition data presented here, with ER and grade specific loci in DCIS showing similar specificity in IDC.

Although we did not identify any novel loci that reached genome-wide significance, we did identify three potential novel DCIS predisposition loci, two of which were DCIS specific (rs12631593, rs73179023), and therefore need further investigation in other cohorts of DCIS. As at least 55% of IDC have associated DCIS present at diagnosis consistent with direct precursor behaviour, it may seem biologically implausible that a SNP predisposes to DCIS but shows no association with IDC. However it is possible that there is a subset of DCIS with very low probability of progression. If the finding of DCIS specific predisposition loci was confirmed in other studies, identifying such a subset of low risk DCIS would be clinically valuable.

In conclusion, this is the largest study to assess genetic predisposition in DCIS and shows that the majority of invasive breast cancer predisposition loci also predispose to DCIS. It highlights that, as for invasive disease, different SNPs predispose to ER positive and ER negative DCIS. In addition it shows the importance of grade in both DCIS and IDC.

Chapter 6 General discussion

In this thesis I have analysed rare and common variants in two different subtypes of breast cancer; lobular (ILC and LCIS) and DCIS. The rationale for selecting these two subtypes as the focus of my research project was because did form a large proportion of breast cancer case control studies and were therefore relatively understudied compared to IDC, but are important subtypes as their incidence rate is increasing.

As ILC can be difficult to detect by screening mammogram, the ability to identify those women at high risk of ILC with genetic screening would be highly beneficial for the population. Similarly, DCIS can be a precursor to IDC and therefore identifying those at risk of developing DCIS would enable us to offer prevention prior to the development of invasive disease.

There was already evidence of a distinct genetic aetiology between the different breast cancer subtypes and we therefore followed a phenotypic stratification approach to identify loci predisposing to specific histological subtypes.

The work outlined in this thesis has added to our current knowledge of genetic predisposition to these two subtypes by identifying which of the known rare variants predisposing to each subtype as well as identifying novel loci. Overall, we have shown that there is a significant overlap in the genetic predisposition of different histological subtypes of breast cancer but there are also distinct associations that could help us understand the aetiology of these subtypes, which in turn may allow us to identify subtype specific therapies.

6.1 Clinical significance of findings

6.1.1 Lobular breast cancer

CDH1 protein truncating mutations are more common than previously thought [250] and these seem to be more in bilateral cases with 4/50 (8%) of individuals with bilateral lobular disease in our initial series having a germline mutation. None of our four mutation carriers satisfy the current criteria for *CDH1* testing given the necessity to have a family history of diffuse gastric cancer in order to be eligible for screening. Although they gave no family history of gastric cancer it is possible that they may develop diffuse gastric cancer in the future, as Pharoah *et al* (2001) showed that the estimated cumulative risk of gastric cancer is higher (83%) than that for breast cancer (39%) in women with *CDH1* mutations. However this was calculated using families with at least three cases of DGC and, as discussed by Pharoah *et al*, may not apply to individuals with a minimal family history, in whom the risks are likely to be lower [111].

Benusiglio *et al*. suggested that women with a personal or family history of at least two ILC before the age of 50 should be offered *CDH1* screening. However, none of the four carrier cases identified in our study would fulfil these criteria as they do not take into account bilateral LCIS. On the basis of our study we recommend that this should be extended to include women with bilateral LCIS.

Two studies, including our own, have shown that *CDH1* mutations in bilateral LCIS or ILC are more common than previously thought [204]. If further studies confirm these findings then, *CDH1* testing could be offered to individuals with bilateral LCIS/ILC under the age of 50, enabling us to identify patients with *CDH1* mutations who may benefit from regular breast MRI screening and endoscopic surveillance for diffuse gastric cancer.

Apart from the bilateral analysis, we screened 2,215 individuals with lobular breast cancer out of which 1,443 had ILC. We identified 6 further protein truncating variants. Out of the 10 protein truncating variants identified in our study, we identified four novel *CDH1* protein truncating variants including the one that has been published already [250].

The prevalence of *CDH1* protein truncating mutations is high amongst individuals with bilateral lobular lesions (8%) but not in individuals with unilateral lesions (<0.5%).

This is the first study at this scale to assess the prevalence of pathogenic variants in the context of unselected individuals with lobular breast cancer. The prevalence of *CDH1* mutations is relatively low, <1%. However, *CDH1* is relatively intolerant to LoF variants since it has a high pLI which is an indicator of how tolerant a gene is to LoF variants.

With regards to the two main breast cancer predisposition genes, *BRCA1*, and *BRCA2*, the clinical utility of genetic testing in the context of lobular carcinoma is not well defined. In particular, *BRCA1* mutations are less frequent amongst individuals with lobular carcinomas whereas the clinical characteristics of *BRCA2* carriers are more heterogeneous compared to *BRCA1* carriers that are usually TNBC.

BRCA2, with a low pLI which indicates that it is prone to LoF variants, was mutated in almost 2% of our ILC cases, and in more than 5% of the cases when restricting analysis to early-onset ILC (≤ 40). The analysis of known predisposition genes revealed that *BRCA2* has a significant contribution to lobular breast cancer. However, these findings are not strong enough to recommend individuals with ILC for *BRCA2* screening.

Our targeted sequencing project allowed us to ascertain the prevalence of rare breast cancer predisposition variants in ILC and generated data that can influence the recommendations for genetic testing in breast cancer. It is particularly important to identify women at high or moderate risk of ILC as this type of breast cancer is often undetectable by mammogram, so women at risk can be offered other alternatives such as MRI screening and / or chemoprevention. We identified a known pathogenic mutation or a novel protein truncating mutation in 3.25% of our unselected ILC patients diagnosed before the age of 60.

I have also shown preliminary evidence for three GxE interactions which will need validation in larger cohorts. These interactions are between HRT usage and three SNPs that have been previously shown to be associated with ILC (rs2981582, rs704010, and rs865686). Nevertheless, these preliminary findings are interesting and may have clinical utility if validated, for example by assessing breast cancer risk by genotype before offering HRT. This is particularly pertinent to lobular breast cancer which has a stronger association with HRT than IDC.

6.1.2 DCIS

The current NICE guidelines for *BRCA1* and *BRCA2* screening are based on a minimum combined probability of a mutation identification of at least 10%. There are methods such as the BOADICEA (see section 1.1.2.5) and the Manchester scoring system that can provide carrier probabilities based on family history and other clinical characteristics. However, these do not include the presence of DCIS lesions as a risk factor in their calculations.

Clinical guidelines are focused on TNBC which is associated with *BRCA1* germline mutations and all women with TNBC under the age of 40 are now offered *BRCA1* screening. *BRCA2* mutations are more common in ER positive carcinomas, and the majority or our DCIS population has an ER positive phenotype.

In our small subset of 31 cases diagnosed with DCIS ≤ 40 and having a first degree relative affected with breast cancer, the frequency of *BRCA2* pathogenic variants is 13%, suggesting that women with young onset DCIS should be referred for genetic screening, in the same way that invasive cases are.

In the group of individuals diagnosed with DCIS ≤ 40 with no first degree relative of breast cancer, the frequency of *BRCA1* and *BRCA2* pathogenic variants is 4.76% which highlights the potential benefit of genetic testing in cases with young onset DCIS, even in the absence of family history. However, our data-set is relatively small and these findings need to be validated in order for them to be translated into clinical practise.

A recent study used the BRCAPRO algorithm to assess the scores of DCIS cases with or without a *BRCA1* or *BRCA2* mutation [276]. Carriers had overall higher score than non-carriers but there was no significant discrimination between the two groups irrespectively of whether DCIS was treated as breast cancer or as not cancer.

We have also shown evidence for DCIS predisposition in other breast cancer susceptibility genes such as the *PALB2* gene, with the frequency of pathogenic mutations being higher than 1% in our DCIS cohort. Our findings illustrate that the prevalence of known breast cancer predisposition genes is relatively high (5%) amongst our unselected cohort of DCIS and increases dramatically when including certain criteria such as early onset of disease or family history.

6.2 Novel predisposition loci

6.2.1 Rare and common variants in lobular breast cancer

One of the best biological candidates from our WES study of extreme cases with ILC was the *ESR2* gene. A protein truncating variant was identified in a family with three affected sisters. Overall we identified a further three protein truncating variants in three individuals with ILC. However, due to the fact that the same initially identified stop-gain variant was present in a healthy control, there is no statistical significant enrichment. Since *ESR2* encodes for estrogen receptor β, we explored the possibility that these variants are likely to express the phenotype in the presence of an environmental trigger, with the most plausible risk factor being use of HRT. However, only one of the three ILC carriers in the replication cohort had taken HRT. Therefore, further studies screening the coding portions of the *ESR2* gene should be conducted in order to validate whether there is a significant contribution of rare variants predisposing to breast cancer.

We have provided evidence suggestive of association between another putative novel gene, *DCLRE1B* with lobular breast cancer, using a rare variant gene-based approach. However, due to the low frequency of events we are unable to draw conclusions on whether rare variants in *DCLRE1B* predispose to breast cancer. Other approaches including more samples or strict phenotypic criteria should be used in the future in order to validate any of the candidate genes from our study. Nevertheless, this gene constitutes a very good candidate since it is involved in inter-strand cross-link DNA repair and warrants further investigation. Additional evidence of involvement of that gene with breast cancer development stems from the fact that a common nonsynonymous variant in that gene, rs11552449, has been previously shown to be associated with breast cancer.

Our study has also demonstrated the existence of distinct molecular pathways underpinning different histological subtypes of breast cancer by identifying a novel locus, rs11977670, at chromosome 7q34 that is specifically associated with invasive lobular breast cancer. This was the first time that unselected individuals with lobular carcinomas were screened at this level to identify novel genetic risk factors. Further functional work is required and currently being undertaken in our research group in order to identify a possible role and mechanism of action for this locus. A histone demethylase, JHDM1D, is 65KB upstream and could possibly explain this functional role. Furthermore, fine-mapping approaches have been followed but failed to identify further neighbouring variants that would show a higher association.

6.2.2 Novel findings DCIS predisposition

None of the SNPs that were genotyped in iCOGS, and were not in LD with any of the previously known breast cancer susceptibility loci, reached genome-wide significance. Three loci that were found to be associated were genotyped in a phase II stage but none achieved replication or overall significance. The locus in chromosome 22 (rs73179023), showed comparable effect sizes across all different studies (ICICLE, BCAC, and phase II) and therefore constitutes the best candidate out of three. In support of this, the risk for overall breast cancer is on the same direction as our DCIS cohort with a suggestive level of significance. A larger replication study could potentially validate this finding.

6.3 Evidence for distinct genetic aetiology between different histological subtypes of breast cancer

While conducting *in situ* breast carcinoma comparisons, we identified a significant excess of rare variants in 6 known breast cancer predisposition genes under investigation in DCIS over LCIS. However, previous studies have shown that there is a higher familial relative risk amongst lobular carcinomas compared to the ductal subtype. Our findings do not necessarily reject the previous notion that there is strong genetic contribution towards lobular carcinoma development. The case could be that there is strong genetic basis for LCIS and ILC development, which we were unable to identify. One possible explanation is that due to the underlying breast cancer heterogeneity, a group of different genes that is yet to be identified can predispose to the lobular histology. Another possible explanation is that the genetic aetiology of lobular cancers is more complex: interactions either in the form of epistasis, genetic modifiers, or GxE interactions play a crucial role. As it has been previously shown that HRT has a stronger association with ILC compared to IDC, we cannot exclude the possibility that HRT might trigger ILC in the presence of specific genetic variation.

Our common-variant analysis of lobular breast cancer has validated the hypothesis of a shared genetic aetiology of different breast cancer subtypes with some key differences and distinct associations. A key finding is that the strength of association can differ between different histological subtypes. We have identified the first lobular specific locus and have shown evidence for three SNPs that seem to be more strongly associated with lobular histology when comparing ILCs to IDCs [173]. The SNP with the most significant differences was rs11249433 which has been previously described as being strongly associated with ER positive breast cancer and in particular of lobular histology. The major finding from our DCIS common variant analysis was that DCIS and IDC share most of their common susceptibility loci [277]. The finding by Campa *et al* that rs1011970 is more strongly associated with *in situ* disease was not validated in our study which has a larger sample size [264]. This leads to the conclusion that these two distinct lesions are two different phases of the same disease. This does not come as a surprise since it is broadly hypothesised that DCIS is a precursor lesion to IDC.

It has been illustrated that polymorphisms at the *CCND1* locus are associated with breast cancer and that genomic aberrations in this locus are frequent in ER positive carcinomas. CCND1 alterations belong to early events in tumour development and appear to be present in LCIS and DCIS lesions. We have shown that variants in the *CCND1* locus are also associated with LCIS and DCIS. While conducting grade and ER stratified DCIS case-only conditional analyses, we have shown that two independent loci near the *CCND1* gene are associated with low/ intermediate grade and ER positive DCIS in an independent manner. This finding was also validated in individuals with IDC from the BCAC studies.

6.4 Excess familial risk and unexplained heritability

The genetic architecture of breast cancer is highly complex with several loci having been identified to be associated with some form of the disease. As previously described, there are genes with highly penetrant variants increasing the risk of developing the disease drastically. There are several other genes with rare variants conferring a moderate risk towards breast cancer development. Also, there are more than 100 independent signals attributed to SNPs across the genome associated with the disease. Those loci increase the risk of developing the disease by a small fraction, usually with OR <1.5. Several research groups have tried to combine data from multiple low risk loci and introduce them as one single value of polygenic risk score (PRS). Researchers from BCAC have utilised information from 77 SNPs and introduced a PRS for breast cancer utilising their previously estimated individual risks conferred by each of these variants. Even though the results are promising and may be clinically useful for the top 1% at the PRS scale, there is still a lot room for improvement and discrimination of cases and controls [278]. They stratified their groups into percentiles and assessed the risk for each percentile group. The average lifetime risk for breast cancer is 12%. It was found that the absolute risk of breast cancer by the age of 80 for the top 99 percentile according to the PRS is 29%. It was also estimated that the risk for the highest quintile is 17.2% (95% CI 16.1%-18.1%). Since the NICE guidelines recommend secondary care to individuals with a lifetime risk higher than 17%, individuals at the PRS scale could benefit from intensive screening. Another interesting but expected finding was that a stronger FRR was observed for women at the lowest percentile of the PRS. This suggests that rare highly penetrant variants may have a stronger impact on familial breast cancer. Collaborators from BCAC also conducted similar analyses in the context of ILC and found more significant differences (unpublished data). The separation of different percentiles is more profound for ILC cases. However, the lobular analysis is based on 72 SNPs whereas the overall breast cancer analysis is based on 77.

Our approach of phenotypic stratification has not produced strong evidence of novel loci associated with breast cancer. We have used a conventional approach by focusing on phenotypic similarities amongst the samples used in order to potentially enrich the genetic homogeneity of the data set and therefore increase the power to identify true associations. However, as it is well understood, the genetic architecture and aetiology of breast cancer is heterogeneous and interactions can occur across the genome as well as between genetic and environmental risk factors. Millions of individuals would be required to be screened in order to identify a portion of these interactions. It has also been shown that regulatory elements across the genome also play a big role in disease-causing and in particular in breast cancer. Larger projects such as the 100,000 Genomes Project are under way with the aim of whole genome sequencing individuals to identify predisposition variants that not necessarily lie within coding regions. By furthering this approach, and introducing several other large scale sequencing projects in the future, we would be able to use this information to identify variants that affect regulatory elements and link them with disease development.

One successful strategy to identify rare variants predisposing to breast cancer would be to focus on individuals with a severe phenotype or individuals that are more likely to carry a rare variant contributing to disease development due to strong family history. However, evidence of merging samples even at a very high level, that one would expect to increase the genetic heterogeneity drastically, has also been shown to be successful. One approach for lobular carcinoma would be to identify other types of cancer with similar patterns of phenotypic or molecular characteristics such as diffuse gastric cancer and merge those data sets into a larger cohort to increase power. This could work both in terms of common and rare variant identification.

6.5 Contribute to database annotation - VUS

Besides the direct clinical utility of studies such as ours, by leading to the incorporation of genes into diagnostic panels or risk prediction tools, there are further contributory elements in addition to scientific knowledge expansion. Several databases exist where genes or variants are annotated with different features and can be used by research studies investigating genetics of specific diseases. Identifying novel variants or genes associated with disease can increase the pool of variants, genes, or phenotypes that exist in databases and can lead to less resource wasting and better science communication. The finding of excess of VUS in ILC cases compared to controls highlights the potential benefit of screening larger cohorts to identify which of these variants can be clinically useful.

6.6 Limitations

Even though during this project we followed a phenotypic stratification approach by focusing on specific histological subtypes, there is still a high expected heterogeneity within our data set. For the lobular subtype study, our cohort is comprised of cases with any features of lobular disease, including cases with pure LCIS, pure ILC, concurrent ILC and LCIS, and LCIS concurrent with non-lobular invasive disease. Depending on the hypothesis, different sample-sets were used, and expected heterogeneity had its toll towards identifying novel signals.

It should be noted that our common variant case control studies are not true GWAS. The genotyping platform used was designed by four consortia for tagging or fine-mapping of loci with prior suggestive evidence of association with cancer, without an actual GWAS backbone and therefore our studies cannot be considered GWAS. This has to leave a window of uncertainty in terms of any possible other loci that could be associated with the phenotypes of our interest but we were unable to capture due to the technology used. SNPs on the iCOGs chips were chosen on the basis of some prior evidence of association with breast cancer as a whole. Although ILC would have been a small proportion of the samples in the discovery sets for these SNPs it is possible that other lobular specific loci exist that have not been included on the iCOGs chip. This is particularly true for LCIS, which would only have been included in the discovery set as a parallel phenotype when associated with invasive disease.

One limitation of the targeted sequencing project is that, due to cost constraints, we could only select a small number of genes to be tested in the targeted sequencing. Some of those genes were not screened in full. Alternative technologies that allow for larger capture regions could be used that would allow a more thorough investigation of loci suspected to be associated with breast cancer and lobular disease in particular.

Finally, we identified several limitations with regards to WES case control analysis. The controls that have been used for this analysis do not have family history or age of diagnosis of their other disorder. We can also not exclude the possibility that they may develop breast cancer. Another possible limitation is that there might be common susceptibility genes between different diseases, and therefore we would not be able to detect a true association or enrichment in the cases over the controls since a portion of the controls are also carriers of certain mutations that predispose both to breast cancer and another syndrome or rare disorder.

One major limitation of the phase I WES rare variant case control study is the lack of power due to small sample size. Taking into account the genetic heterogeneity of the disease along with the phenotypic diversity of the samples, we did not have enough power to detect associations with statistical significance. Therefore, we followed a rare event approach where genes were prioritised in the absence of variants in the control population. This approach might not be ideal since there might be false positive results that can reach the final gene selection, but at the time of this analysis and with the data available, this was the most cost effective approach to investigate putative novel genes that could predispose to lobular breast cancer. A larger phase I study could potentially yield a different and more accurate gene list but due to time and cost constraints we were forced to select our candidate genes based on a smaller phase I exome-wide gene based study. Future analyses, incorporating more samples could yield a different gene list that could be later on tested in a targeted sequencing project.

With advances in sequencing technologies, screening a large number of genes is becoming more cost effective and we are now able to screen more than 100x the region of our targeted sequencing panel, for the same reagent cost. If that technology existed while we were designing the phase II study, we could have incorporated a larger set of genes identified as possible candidates from phase I as well as a larger set of genes that have been implicated with breast cancer development in the past. Finally, genes near common variant association loci such as the region in chromosome 7q34, which is lobular specific could also be screened. This would allow a thorough investigation of these regions and potentially could lead to the identification of rare variants in the region that are functional and are tagged by common tagging SNPs. We would then be able to explain a larger proportion of the genetic aetiology of lobular breast cancer.

6.7 Summary of findings

To summarise the key findings of our study, we have identified a germline protein truncating variant in 8% (95% CI 5%-15.5%) of individuals with any bilateral form of lobular disease. This is significantly higher compared to unilateral cases and healthy individuals.

We have estimated the prevalence of *BRCA2* mutations to be higher than *CDH1* mutations in the context of lobular cancer. A total of 1.4% (95% CI 0.8%-2%) and 1.5% (95% CI 1%-2%) of ILC cases or cases with any form of lobular disease are *BRCA2* carriers whereas less than 0.5% (ILC 95% CI 0.1%-0.9%) (Any lobular 95% CI 0.2%-0.8%) are *CDH1* carriers in both of the aforementioned groups. We have provided evidence of an excess of rare *BRCA1* VUS in ILC compared to controls. We have also shown that the prevalence of *PALB2* mutations is relatively high 0.6% (95% CI 0.2%-1%) for ILC and 0.5% (95% CI 0.2%-0.8%) for any lobular carcinoma. A total of 3.2% (95% CI 2.3%-4.1%) of ILC cases screened carry a germline pathogenic variant in one of the 6 breast cancer predisposition genes that we screened. We have identified a common novel breast cancer predisposition locus that is specific to the lobular histology. This locus is currently being interrogated in depth in order to identify the functional role of this region. Further analyses are required to validate any of the putative novel genes assessed in our study.

We have also assessed the prevalence of pathogenic mutations in known breast cancer predisposition genes in a cohort of DCIS cases. The combined prevalence of *BRCA1* and *BRCA2* germline mutations is 3.2% (95% CI 1.9%-4.5%) in our population. It is increased to

12.9% (95% CI 1.1%-24.7%) when taking into account only individuals with a first degree relative with breast cancer and age of DCIS diagnosis ≤ 40. We were able to identify a pathogenic mutation in 19.4% (95% CI 5.5%-33.3%) of individuals having a first degree relative with breast cancer and were diagnosed with DCIS ≤ 40. Finally, we have shown evidence for shared genetic susceptibility between DCIS and IDC with regards to common variants.

Overall, we have contributed to current knowledge on breast cancer predisposition by focusing on specific histological subtypes that have previously been understudied. More studies, incorporating larger data-sets can underpin those similarities and differences and assist in understanding the complexity of the genetic architecture underlying breast cancer predisposition.

Chapter 7 Materials and methods

7.1 Clinical resource

7.1.1 GLACIER study

The GLACIER study has recruited 2,539 patients from throughout the UK with the aim of understanding genetic predisposition to LCIS and ILC (MREC 06/Q1702/64). Women who have or had LCIS (with or without invasive disease of any morphological subtype) or pure invasive lobular carcinoma before the age of 60 at the time of diagnosis were eligible for enrolment. Figure 7.1 shows the different pathological features of the cases enrolled in GLACIER. Cases in GLACIER study are split into 6 morphological categories from left to right: Pure LCIS (388), Pure ILC (382), ILC with LCIS (1152), IDC with LCIS (277), invasive of mixed morphology with LCIS (156), missing pathology (143), and ineligible based on predefined criteria (37).

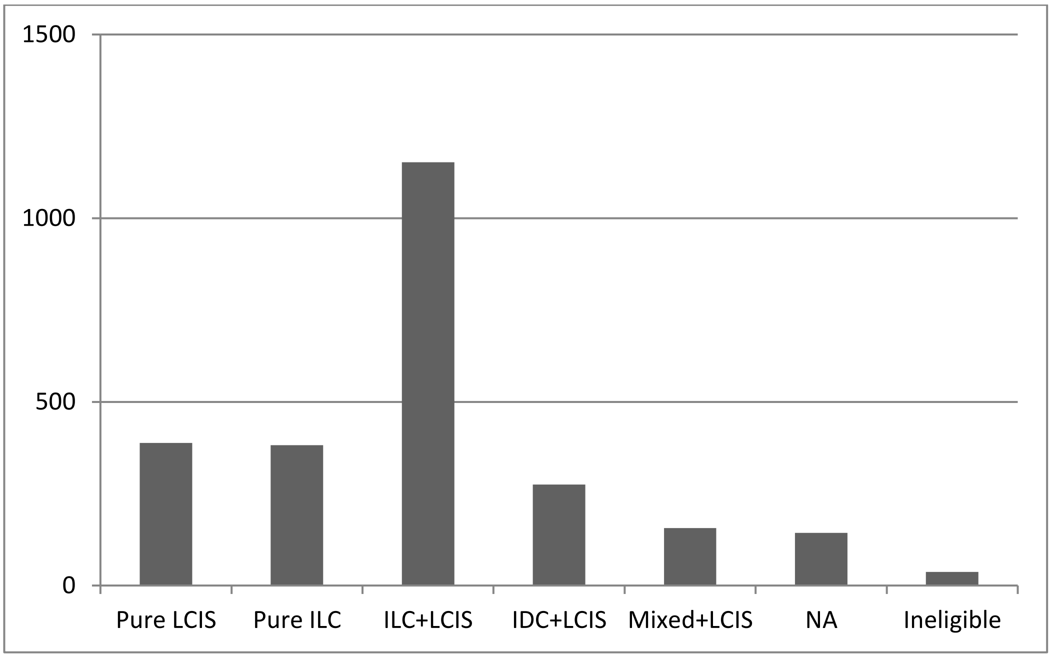


Figure 7.1 Proportion of different morphological subtypes in the GLACIER study.

Pathology reports were requested and were filed for more than 95% of the participants. Peripheral blood samples, and formalin fixed paraffin embedded (FFPE) tissue blocks were collected from participants alongside family history data and other risk factor information. Individuals were asked to complete a questionnaire reporting demographic, ethnic, reproductive and family history information.

7.1.2 ICICLE study

The ICICLE study ascertained 3,371 DCIS cases across the UK (MREC 08/H0502/4). The aim of this study is to identify genetic predisposition to DCIS. Participants were selected based on the following eligibility criteria 1) cases with any grade of DCIS and no invasion, 2) age of diagnosis, with participants having to be younger than 60 years of age when diagnosed with DCIS. Recruitment was followed by collection of information on the grade and size of the tumours by pathologists. For cases where an FFPE tissue was available, sections were stained and reviewed to assess the grade and ER status. The concordance between these scores and the pathology reports was assessed and more than 95% correlation was observed. The distribution of the age of diagnosis is provided in Figure 7.2.

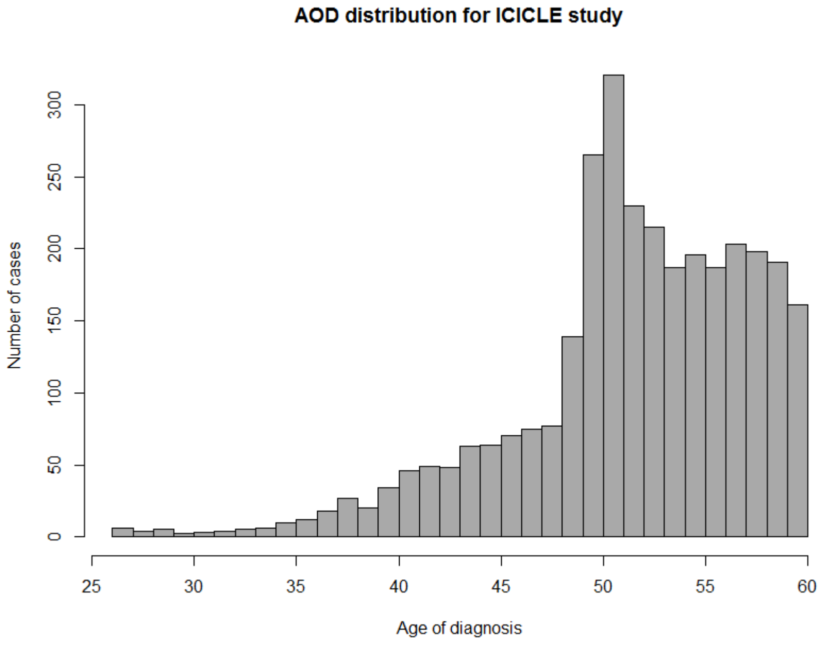


Figure 7.2 Distribution of the age of diagnosis of DCIS for all cases recruited in the ICICLE study.

7.1.3 Samples from the KHP Cancer Biobank

Although recruitment for the GLACIER and ICICLE studies was completed in 2012, individuals with ILC or DCIS were recruited via Guys Hospital breast tissue bank (NHS REC ref. 12-EE-0493) as a continuation research initiative. Since its start in 2015 this study has recruited more than 200 individuals with either ILC or DCIS.

7.1.4 GLACIER and ICICLE healthy controls

As part of the GLACIER and ICICLE studies, healthy volunteers were also recruited using two methods: 1) by requesting the recruited individuals to approach their female peers with no personal or family (up to 2nd degree) history of breast cancer, LCIS, DCIS, or benign breast disease, 2) via posters requesting healthy volunteers with no personal or family (up to 2nd degree) history of breast cancer, LCIS, DCIS, or benign breast disease) within the recruiting hospitals. All participants donated a blood sample and were asked to complete a self-administered paper-based questionnaire on their family history and reproductive and hormonal risk factors after giving full consent for the study. In total, 2,121 healthy volunteers have been recruited.

7.1.5 Additional samples

In addition to the individuals recruited through the GLACIER and ICICLE studies, there were other DNA samples used for particular analyses. In more detail, 200 ILC/LCIS samples from University of Westminster and 300 ILC samples from ICR were used for the iCOGS genotyping studies. These samples were processed as part of the GLACIER study at the BRC genomics facility with respect to the genotyping project.

7.1.5.1 **TCGA**

An additional data set used in this project was downloaded from TCGA, which constitutes a repository of different data sets, including whole exome sequencing, on different types of cancer. A user key was obtained by our research group from the Cancer Genomics Hub (CGHub) in order to obtain access to downloading data. Using the online interface, a manifest of samples list was generated and was fed to the Gene-Torrent downloading software in order to download the .bam (mapped to the genome) files from the Santa Cruz server where the TCGA data is stored (https://tcga-data.nci.nih.gov/tcga/). It is likely that the .fastq (raw sequence) files from TCGA have been processed in a different way in order to produce the .bam files and the .vcf (variant calling) files, and therefore it was prudent to reformat the data to its initial form and apply the same filtering criteria as per exome samples sequenced in-house. This was done to ensure data was aligned in the same way amongst the TCGA samples and our in-house samples. Following an interrogation of clinical data from more than 1,000 individuals with breast cancer, we identified 110 germline exomes from individuals with ILC. Files that were already mapped to the reference genome were downloaded (.bam files). Some clinical characteristics of the TCGA individuals such as age of diagnosis of ILC along with menopause status and ER status are reported in Appendix 1.

7.1.5.2 **Controls for exome sequencing case control analysis**

More than 6000 samples have been exome sequenced in the BRC genomics facility over the last 5 years. The vast majority of these samples were sequenced to identify rare variants predisposing to rare disorders and syndromes not associated with cancer development. A systematic approach was followed to select control samples from this sample set. A total of 536 European unaffected females with no personal history of breast cancer were used as controls for this study. Cryptic relatedness analysis was conducted, as well as PCA that led to the identification of a cluster of individuals with European ancestry. A gender identification script was performed to ensure that individuals used in the final analysis were all females. The disorders that the control samples were selected for are mentioned in Table 7.1.

Table 7.1: Disorders that controls samples were sequenced for. All projects including 10 or more individuals are mentioned.

Project	Samples
AGEP	44
Renal Disease	42
EB	35
SIDS	33
SRNS	33
Lymphoedema	22
Hidradenitis Suppurativa	22
Renal	16
SRS	13
Epilepsy	12
Wiedemann Steiner	10
Paradoxical Psoriasis	10
Other	244

7.2 Laboratory experiments

7.2.1 DNA quantification

All samples were initially quantified using PicoGreen in the outsourced centre (Tepnel, Manchester, UK) following DNA extraction. Subsequently, several different DNA quantification methods have been used for the germline DNA that we obtained from the GLACIER and ICICLE studies, these included Nanodrop, Agilents Bioanalyzer and Tapestation, as well as Qubit.

7.2.1.1 **Nanodrop**

All samples that were processed in the targeted sequencing experiment underwent Nanodrop quantification and quality control. The Nanodrop 8000 (Thermo Fisher Scientific) spectrophotometer was used to measure the quantity and quality of all samples. DNA quantification was performed using 1 μl of sample and DNA quality and purity was tested by spectral scan observation and considered when a single prominent A260 peak and an ~1.8 A260/A280 ratio was found.

7.2.1.2 Qubit

The Qubit Fluorometer (Q32857, Thermo Fisher Scientific) has been used to quantify samples for all samples that have been selected for exome sequencing have been quantified using the Qubit dsDNA Broad Range Assay Kit (Q32853, Thermo Fisher Scientific) whereas the pooled libraries were quantified using Qubit dsDNA High Sensitivity Assay Kit (Q32851, Thermo Fisher Scientific) according to manufacturers instructions. Libraries were quantified and aliquoted appropriately to load on the sequencing machine. Finally, 5% of the samples used on the targeted sequencing experiment were also quantified using Qubit since the Nanodrop measures usually overestimates the amount of DNA present due to the fact that Qubit only quantifies dsDNA. This was performed in order get an estimate of the variation between Nanodrop and Qubit for which the average Qubit/Nanodrop measure ratio was of 0.53x. Therefore, the Nanodrop measurements were multiplied by 0.53 while calculating the optimal input DNA for the targeted sequencing experiment.

7.2.1.3 Bioanalyser

Samples selected for exome sequencing were run in an Agilent 2100 Bioanalyzer system which provides sizing, quantitation and quality control of DNA. This was performed to ensure that the DNA integrity was optimal across different stages of the library preparation step before the samples were ready for sequencing.

This method includes an in-chip electrophoresis where DNA molecules migrate through their wells and are separated based on their size. Fragments are detected using fluorescence.

The peaks are analysed in the incorporated software (2100 expert software), where the fluorescence intensity is plotted against the migration time to produce electrophoregrams. The quantification of the samples occurs by comparing the sample peaks to the ones of the upper standard which is of known concentration. The two different kits used are indicated in Table 7.2

Table 7.2: Bioanalyzer kits used for different stages of the WES library preparation quantification.

Kit	High Sensitivity	DNA 1000
Quantitative range	5-500 pg/µL	0.5-50 ng/µL
Size range	50-7000	25–1000
Peaks for ladder	15	13
Lower/Upper marker	35/10380 bp	15/1500 bp
Lot-Number	5067-4626	5067-1504

7.2.1.4 **Tapestation**

For the Fluidigm amplified libraries, a selection of 8 samples per plate, or 4 samples per array, which included harvested and barcoded products, were randomly selected for screening using the Tapestation 2200 and the 1000D screen tapes. Additionally, final pooled libraries were run in order to visualise the purity of the libraries and to estimate the variance in fragment length and the average length of the fragment size. A representative example is shown in Figure 7.3. The average length of the fragment size was used to calculate the appropriate amount of library that needs to be loaded on the sequencer for sequencing (see section 7.2.10). To prepare libraries at the optimal input that the Genomics facility requires them, 4μM, we can calculate the molar concentration of the libraries using the formula: $Molar\ Concentration\ (in\ nM) = Concentration\ (in\ ng) x 10^{6} x (\frac{1}{649}) x (\frac{1}{average\ size\ (in\ bp)})$

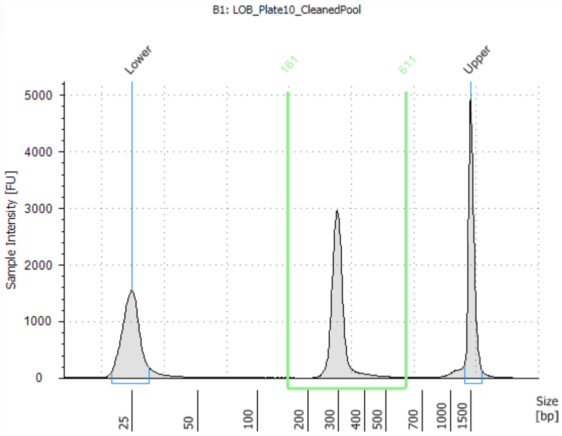


Figure 7.3: Targeted sequencing pooled library quantification. The concentration along with the average size of the fragments is calculated by selecting the area of interest.

7.2.2 iCOGS genotyping

A total of 200ng per sample of germline DNA from both the GLACIER and ICICLE studies was used for genotyping in the iCOGS platform. Samples genotyped included 3,160 ICICLE cases, 2,210 GLACIER cases together with 5,000 ethnicity-matched controls. The cases were genotyped at the BRC Genomics core facility at Guys Hospital. Regarding the 5,000 matched controls used, these came from 4 different studies across the UK (SEARCH, BBCS, SBCS, UKBGS).

The iCOGS array platform has been designed by four different consortia including the BCAC, with the aim of identifying individuals at higher risk of developing certain types of cancer. The platform utilises the Infinium array technology from Illumina where hundreds of bi-allelic markers can be assessed. The iCOGS custom Illumina iSelect, contains 211,155 variants, most of which are single nucleotide polymorphisms (SNPs). In brief, the first step of the protocol involves whole genome DNA amplification for approximately 20 hours and its followed by DNA fragmentation, precipitation and resuspension. The resuspended DNA is then hybridised onto the BeadChip for approximately 17-20 hours. The following day the BeadChips are washed and stained, imaging is performed at least one hour after the staining to ensure that BeadChips have dried properly. Suboptimal imaging was observed initially in the bottom two rows of the BeadChips. A problem during the staining process was identified after liaising with Illumina and kits were replaced. Samples that failed were re-run in a new batch of replacement BeadChips.

7.2.3 Primer design

In order to capture the exonic portions of the *CDH1* gene, exon-flanking intronic primers were designed using Primer3 (http://frodo.wi.mit.edu/).

For the targeted sequencing project, primers were designed using the D3 tool that is incorporated on an online version of the Fluidigm website. D3s underlying algorithm is based on Primer3. Since this is a complex high-throughput multiplex PCR based experiment, several thresholds have to be optimised in order to minimize non-specific binding and primer-dimer generation. In this regard, repeat regions, high GC content regions, as well as regions with common variants were avoided when possible from primer sites. Primers were designed within the flanking regions of the target fragment (exons +/- 10 bp) so the product size would range 150-200bp, including the primers, Appendix 2. A second stage of primer selection occurred where we excluded all common SNPs (5% MAF) based on the latest, at the time of analysis, dbSNP version (dbSNP_142). Universal sequence tags were added to each primer to ensure that a second step PCR can also occur.

7.2.4 PCR

Samples with suspected mutations or variants of interest were PCR-amplified for subsequent Sanger sequencing in order to validate or reject the initial hypothesis supporting the existence of a rare variant. A total of 10-100 ng of germline DNA were amplified using standard conditions. Three minutes of hot start at 94°C were followed by 30 cycles of 45 sec at 94°C, 30 sec at 55°C and 90 sec at 72°C. The last step of the PCR is 10 min at 72°C and the products were later on stored at -20°C. For fragments failing to amplify using the standard protocol adjustments on the annealing temperature were made to ensure optimal fragment amplification.

7.2.5 Gel electrophoresis

To ensure proper amplification before Sanger Sequencing, PCR products were electrophoresed on 1% TBE agarose gels stained with GelRed (Biotium) and visualised under a UV transilluminator.

7.2.6 Sanger sequencing

PCR products were initially purified using ExoSAP (New England Biolabs). Subsequently, sequencing reaction was performed using 3.5µl purified PCR-amplified product, 1X sequencing buffer (Applied Biosystems), 5pM primer, and 0.25ul of BigDye terminator v1.3 Cycle sequencing kit (Applied Biosystems) for a final volume of 5.25µL. Samples were then purified via ethanol precipitation and purified DNA products were sequenced on the ABI 3730 Genetic Analyser (Applied Biosystems). Sequencing data were analysed with the Sequencher software V4.9 (Gene Codes). The conditions that were used for the sequencing reaction were 30 cycles of 30 sec at 96°C, 15 sec at 50°C, and 60 sec at 60°C.

Sanger sequencing was performed in 32 samples of individuals with bilateral lobular breast cancer, where all exonic portions and flanking splicing junctions of the *CDH1* gene were amplified. Primers used to amplify the coding portions of the *CDH1* gene are reported on Table 7.3. Additionally, this sequencing method was used to validate variants identified through next generation sequencing, either WES or the targeted sequencing experiment.

Table 7.3: Primer pairs for *CDH1* screening.

Primer ID	Primer Sequence	Fragment length
ECADX1F	TAGAGGGTCACCGCGTCTAT	378
ECADX1R	AATGCGTCCCTCGCAAG	
ECADX2.F	TCACCCGGTTCCATCTAC	198
ECADX2.R	TTCCAACCCCTCCCTACT	
ECADX3F	TGTCCAATTTCCTAATCTCTGTGA	300
ECADX3R	AAAACAACAGCGAACTTCTCA	
ECADX4.F	CCTGAAGTATCCGTCTTGAATTG	235
ECADX4.R	TCCCTCCCAGAGAAACAGAG	
ECADX5.F	GTTGGGATCCTTCTTTACTA	296
ECADX5.R	AAATCCTGGGTGGATGTTAC	
ECADX6F	TTCCTCATCAGAGCTCAAGTCA	248
ECADX6R	TTTGGGGTCCAAAGAACCTA	
ECADX7F	GCAGCTTGTCTAAACCTTCATC	250
ECADX7R	TCCTCCACACCCTCTGGAT	
ECADX8F	GTTCCTGGTCCTGACTTGGT	247
ECADX8R	CCATGAGCAGTGGTGACACTT	
ECADX9F	AATCCTTTAGCCCCCTGAGA	382
ECADX9R	TCTGGGAAAGTCACCCTGTC	
ECADX10F	TTTTTAACTTCATTGTTTCTGCTCTC	299
ECADX10R	TCAGTTGAAAAATCCTCACACTT	
ECADX11F	ACATGTTGTTTGCTGGTCCT	229
ECADX11R	AGGCAGCAAAGGCTCAGAT	
ECADX12F	CAAGCTGCCACATTTTCTGT	296
ECADX12R	TGGAGCAAAGTTGCCAAA	
ECADX13F	TCCCCTGGTCTCATCATTTC	300
ECADX13R	TCAAAGGCTGAGTCACTTGC	
ECADX14.F	CTCTCAACACTTGCTCTGTC	206
ECADX14.R	AGAGATCACCACTGAGCTAC	
ECADX15.F	TCCAACCATAATCTATAAACTGAACA	298
ECADX15.R	TGACACAACTCCTCCTGAGC	
ECADX16.1F	AAGATGCTTTTGTCCCTTCTTC	493
ECADX16.1R	TCTTTTGGACATCACCACCA	
ECADX16.2F	CAGCTCCCTTCCCTTGAGAT	396
ECADX16.2R	AAAAAGGCAGAGGGACACAC	
ECADX16.3F	CCAGCACCTTGCAGATTTTC	400
ECADX16.3R	CCAAGATGGGAGGATCACTT	

7.2.7 MLPA

Multiplex Ligation-dependent Probe Amplification (MLPA) was performed in order to investigate the presence of abnormal copy number of exons in the *CDH1* gene. MLPA is a multiplex PCR method using a probe-mix which is designed to detect deletions and duplications of one or more sequences in a particular gene of a DNA sample. MLPA was performed using the MRC Holland kit. In brief, DNA samples were amplified and PCR products were analysed on the ABI 3730 Genetic Analyser (Applied Biosystems) using TAMRA 500 according to manufacturer's instructions. Data analysis was conducted using Coffalyser.Net (MRC Holland, Amsterdam, Netherlands), which allows for MLPA analysis. Data from the sequenced fragments is standardised and visualised using this tool to call potential exonic deletions/rearrangements.

7.2.8 Agilent Sure-select exon capture

Samples were prepared so a total of 3µg of germline DNA was used as input for WES. The library preparation step was performed using the target enrichment capture SureSelectXT2 Human All Exon V4 from Agilent. One library was prepared for each individual sample that was exome sequenced. The Qubit system was used to quantify genomic DNA before library preparation. The initial DNA fragmentation or shearing step, was performed using sonication with Covaris E220. The Covaris instrument was degassed for least 30 minutes before use, and the chiller temperature was set between 2°C to 5°C to ensure that the temperature reading in the water bath displays 5°C. A tapered pipette tip was used to slowly transfer 130µl of DNA sample through the pre-split septa of the Covaris microTUBE. The target DNA fragment size after shearing is 150 to 200 bp. Once the shearing was completed, DNA was removed while keeping the snap-cap on, and inserting a pipette tip through the pre-split septa and slowly. Subsequently, the sheared DNA was purified using AMPure XP beads. The quality of the DNA was assessed using the 2100 Bioanalyzer and the DNA 1000 assay. The next step included repair of the ends of the DNA fragments, using the SureSelect XT Library Prep Kit ILM, followed by another DNA purification using AMPure XP beads. Subsequently, the 3 end of the DNA fragments were adenylated and libraries were once again purified using the AMPure XP beads. This purification was followed by the ligation of the paired-end adaptor. This step can produce varying results, based on the quality and quantity of the input DNA. Usually five cycles produced adequate DNA yield for the subsequent capture without introducing bias. This specific amplification step is of great importance as several different genomic attributes, such as GC content or repetitive sequences increase the difficulty of region amplification, which may therefore be underrepresented in the final sequencing data. This constitutes a common feature of high throughput experiments where amplification of these regions tends to fail or amplify sub-optimally for the majority of the samples used under the same experimental conditions. The libraries were finally purified with AMPure XP beads. In order to assess the quality and quantity of the libraries, the 2100 Bioanalyzer instrument and DNA 1000 assay were used. The electropherograms were inspected and showed a distribution of DNA fragment size peak of approximately 225 to 275 bp. The concentration of the library DNA was determined by integrating under the peak.

7.2.9 Fluidigm Access Array

The Access Array technology from Fluidigm has been selected for the targeted sequencing project. This method allows for simultaneous amplification of all different fragments of interest utilising microfluidics technology. The objective of this protocol was to create an amplicon library of the regions of interest, making them suitable for next-generation sequencing as it allows the interrogation and identification of single nucleotide substitution variants as well as small indels. This is an amplicon-based method where different primers are designed to capture the target of interest and regions are PCR-amplified in micro-chambers.

The principle of Fluidigm Access Array is based on a two-step PCR amplification process. Approximately 100ng (range between 50ng-250ng) of high quality germline DNA has been used to amplify samples from the GLACIER and the ICICLE studies on a custom made targeted sequencing panel from Fluidigm. This comprised 573 amplicons translating to 218 exons of 20 genes being amplified in a core of an array for 48 samples each and for a total of 97 arrays, which translates to 27,504 different PCR reactions being performed concurrently.

The sample mix for the initial PCR amplification includes 3µl of diluted DNA sample (ranging between 50 and 250ng), 0.5µl of 10X FastStart High Fidelity Reaction buffer (without $MgCl_2$), 0.9µl of 25mM $MgCl_2$, 0.25µl DMSO, 0.1µl of 10mM PCR Grade Nucleotide Mix, 0.05µl 5U/ul FastStart High Fidelity Enzyme Blend, and 0.25µl 20X Access Array Loading Reagent. Note that all samples, primers, and mixes were vortexed and centrifuged appropriately to allow for homogeneous delivery of the reagents into the micro-chambers since there is no active mixing in the array.

The Fluidigm designed primers were supplied in single-plex with forward and reverse primers combined, these are multiplexed according to suppliers instruction to achieve optimal efficiency, a total of 48 sets of up to 13 different pairs of primers were used in this project. Target-specific primers incorporate common sequence tags, Figure 7.4 attached to the 5 and 3 ends, Appendix 2. These sequences allow the second PCR to occur as they provide a binding site for sample-specific primers, Figure 7.4. The latter primers not only contain the complementary sequence of the common sequences but also sample specific barcodes and paired-end Illumina sequencing primer-annealing sites.

A 5µM multiplexed primer stock plate is prepared according to manufacturer’s instructions and is followed by the preparation of the 20x primer plate by adding 20µl of the previously diluted 5µM Multiplexed primer, 5µl of 20x Access Array Loading Reagent, and 75µl of 1xTE. This dilution results in a primer mix with a final primer concentration of 1µM from which 4µl were loaded into the primer inlets on the right hand side of the arrays. Once the primer mix and the sample mix were loaded into the array, the latter was loaded into the IFC Controller AX for the “Sample load” process which distributed all mixes to the microchambers for the amplification step to occur.

The described sample loading process was followed by the transference of the arrays into the FC1 Thermal Cycler (Fluidigm) onto the vacuum controlled surface. Application of vacuum on the array is crucial as it allows the maintenance of homogeneous conditions across the surface of the different reaction chambers along the 2 hours and 20 minutes of the amplification step.

Upon completion of the thermal cycling program the “harvest” procedure was performed on a post-PCR IFC Controller AX, where all amplified samples from the chamber were driven back to the sample inlets and transferred to a 96 well plate. The final volume of the harvested product was approximately 10µl.

A fraction of these samples corresponding to different arrays underwent an amplification check using the D1000 screentapes on the Tapestation. In this regard, harvested products were diluted 1:100 in 1x TE, this step is of great importance since excessive amount of DNA in the following step could inhibit the reaction or cause biased amplification towards specific fragments. Consequently, this could lead to a non-uniform representation of the fragments during sequencing.

During the barcoding step and using the common sequence as a bridge, the sequencing primer sites as well as the sample specific barcodes for multiplexing during sequencing were introduced. The barcodes used are shown in the description of Appendix 2. The components used in the barcoding step include 1µl of the diluted harvested libraries, 2µl 10X FastStart High Fidelity Reaction buffer (without $MgCl_2$) (Roche), 3.6µl 25mM $MgCl_2$, 1µl DMSO, 0.4µl 10mM PCR Grade Nucleotide Mix, 0.2µl 5U/ul FastStart High Fidelity Enzyme Blend, 7.8µl PCR certified water, and 4µl of the barcode library primer, leading to a total volume of 15µl.

Amplified products were inspected in Tapestation to verify that the fragments were of the expected range, with no primer dimers or artefacts of unwanted size. A shift of approximately 60 bp was observed in all samples when comparing the harvest of the first amplification and the product of the barcoding PCR. This corresponds to the addition of the sequencing primers and barcodes.

The final step of this protocol included a magnetic bead clean-up step where 2µl of each sample from each 96 well plate were pooled together into one Eppendorf tube. A total of 12µl of the pooled library was then mixed with 24µl of 1xTE buffer, and 36µl of Ampure XP beads (Beckman Coulter). The clean-up protocol involved two washes with freshly made 70% Ethanol, and a final elution of 40µl in 1xTE buffer. This was followed by quantification of the purified libraries using Qubit High Sensitivity Assay Kit and their average length size was measured in Tapestation using the D1000 screentape. The two values obtained in each of the readings were then used to calculate the final Molar concentration in order to prepare each library at 4nM. The next step involved the pooling of 10 pooled libraries, note that at this stage each library corresponds to 96 samples, which leads to a multiplexing of 960 individual samples per final quantified library.

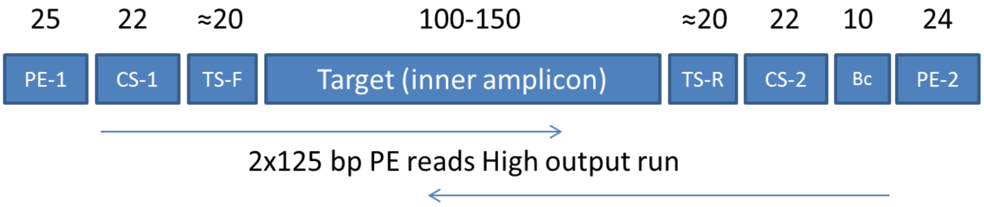


Figure 7.4: Representation of the two step PCR reaction with sizes of different amplicon components. PE-1 and PE-2 correspond to sequences that will bind on the sequencing flow cell, CS-1 and CS-2 are the common sequences that are used as bridges to allow for the initial PCR to be used as a template using universal primers, TS-F and TS-R correspond to the target specific primers, and the target corresponds to the region of interest. Numbers on top of each fragment correspond to the base-pair length of each component.

A major key point to be taken into consideration while performing and planning this protocol is that sample input should be carefully selected since too little DNA can lead to clonal amplification whereas too much DNA can lead to PCR inhibition. Additionally, pipetting errors at any stage could lead to preferential amplification of specific primer sets or inconsistent amplification in general. Ensuring there are no bubbles in all of the array inlets while loading the sample mix and the primer mix on the arrays is also crucial. If any bubbles are observed, they can be removed using a microlance. Another key consideration is that there is no active mixing within the array, therefore it is important to ensure that all reagents and mixes have been adequately vortexed and briefly centrifuged prior to pipetting to the array. Finally, an incorrect proportion of beads during purification may lead to loss of product, or failure in primer dimer removal. Furthermore, it is critical to remove the Ampure XP beads from the fridge to equilibrate at room temperature for at least 30mins and also prepare a fresh stock of 70% Ethanol whenever the clean-up step is performed.

7.2.10 Sequencing by synthesis

Illuminas platforms (i) MiSeq and HiSeq2500 were selected for sequencing the pooled libraries of the targeted sequencing and (ii) GAIIx and HiSeq2000 for the samples that underwent whole exome sequencing. All platforms are based on the same principal, sequencing by synthesis chemistry. In brief, once the libraries are amplified and bound to the flow cells clusters are generated. Subsequently, several different rounds of single nucleotide extension occur, each of which is accompanied by laser scan to identify each incorporated base.

There are two different sequencing methods, namely the single read and the paired end read. Here, we performed the paired end read protocol where there is higher confidence on the reads since each fragment is being scanned twice, once from each end. Therefore the outcome reads has their pair assigned which gives higher confidence to a read being mapped appropriately or highlight errors if paired reads do not map on the same chromosomal positions. This sequencing method is based on reversible dye-terminators that enable the identification of single bases as they are introduced into DNA strands. This technique was used both for our whole exome sequencing project and the targeted cancer panel amplicon based sequencing.

There are four main elements or main procedures in Illumina sequencing by synthesis technology. The first one is the library preparation step where the regions of interest are amplified and attached to special adaptors on both 5 and 3 ends in order for them to be compatible with sequencing. Library preparation may follow one of two different methods with distinct principals that lead to amplification of specific target regions. One of the methods is called target enrichment whereas the other is amplicon based. Both of these approaches have been used in this project, the target enrichment for whole exome sequencing and the amplicon based for the targeted cancer panel sequencing.

Library preparation and purification is then followed by cluster generation. In a brief description, the DNA molecules bind on complementary sequences on the flow cell, the molecules bend to form a bridge on the flow cell and attach on a second complimentary oligo, the reverse strand is synthesised using a polymerase, the two strands are then straightened and ready for another round of amplification. This procedure occurs in close proximity on the flow cell and therefore leads to clusters of thousands of amplified clones from the same molecules. Once the bridge amplification is finished all the reverse strands are washed off and removed from the flow cells surface. The next step includes the incorporation of single nucleotides per sequencing read and the record their fluorescent emissions. Each of the four different nucleotides have their own unique fluorescent label and when excited they emit a specific signal that is captured by a laser camera. These emissions are captured and converted to base calls that are later combined to form the sequencing reads.

One of the advantages of SBS is that the base calls are accompanied by intensities from the excitation and therefore can be quantified. It has been shown that this method delivers the highest proportion of high quality base calls (Q>30). Depending on the reagents and equipment used, the read length can vary between 100 up to 250 bases. For exome sequencing the read length selected was 2x100 bp. The 2x corresponds to the fact that the sequencing is paired end and therefore each fragment is sequenced from both ends irrespective of how long it is. For the targeted sequencing project, the read length was 2x150 for the MiSeq and 2x125 for the HiSeq2500.

It is important to calculate the read length appropriately and ensure that all regions in the fragments are captured by at least one read. Therefore, all fragments designed for the targeted sequencing project were ≤ 200 bp and the sequencing length was of 125 to ensure an optimal coverage across all regions by at least read 1 or read 2 if not from both. Additionally, the nucleotides, apart from the fluorescent tags, have reversible 3 blockers that will not allow the incorporation of more than one base per read. Each sequencing cycle consists in the incorporation of a base and a scan of the flow cell that will convert the emissions to base calls and all non-incorporated nucleotides are washed off. A step where the blockers are removed is also included to allow for the next read or cycle to occur. This process is repeated as many times as the reagents allow, or until the DNA molecule is completely sequenced. Once read 1 is finished, a similar bridge amplification occurs but this time the forward strands get washed off, leaving the reverse strands to be sequenced. Before the initiation of read 2, the index read is performed. This is a step that allows for demultiplexing samples that have been pooled on the same lane of the sequencing flow cell.

7.2.10.1 **MiSeq**

As a pilot sequencing experiment, the first 94 samples were sequenced on a MiSeq to assess the uniformity of coverage across different samples and amplicons. In this regard, 6pM of 94 samples pooled library, plus a positive control and a negative control, were run on MiSeq as part of a Pilot study. Samples were run at 150 bases paired-end run on an Illumina MiSeq system (v2 reagents). The optimal cluster density for MiSeq system, i.e. the number of clusters per square millimetre for the run on a flow cell, is approximately 800K/mm^2, and the cluster density achieved for this pilot experiment was 630 ±14K/mm^2. The successful outcome of this experiment, Figure 7.5, allowed us to proceed with the experiment and the remaining samples were processed and sequenced on a HiSeq2500.

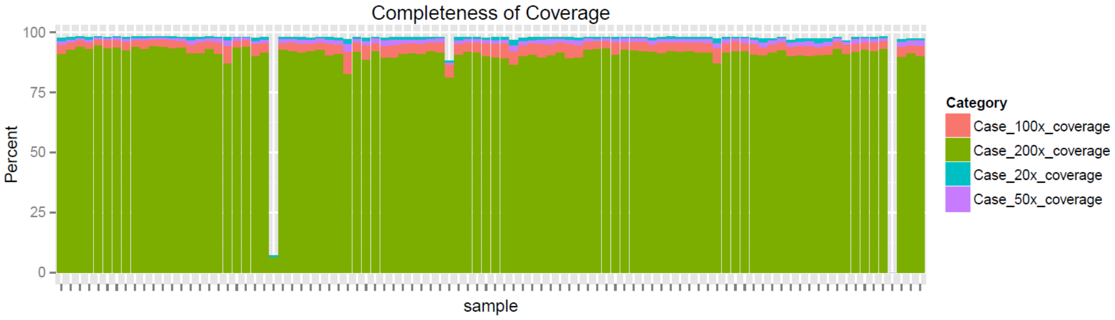


Figure 7.5: Coverage graph from MiSeq pilot experiment. For each sample (x axis) we achieved high coverage (at least 200x for 80% of the target capture, green). There was one sample that failed to amplify apart from the negative control.

7.2.10.2 **HiSeq 2500**

All samples were sequenced on HiSeq2500 as 125 bases paired-end run (v4 reagents) using 9pM of pooled library. The expected average coverage was of 800X. The optimal cluster density for a HiSeq2500 run is estimated to be 950 K/mm2. Table 7.4 shows the cluster density that was achieved for each of the 5 pools that were sequenced on the HiSeq2500. The proportion of high quality bases is also indicated on the same table with more than 95% of the sequenced data having Q>30. The "Q" score is a Phred scaled score that indicates the probability that a given base is incorrectly called during sequencing. A score of Q=30 corresponds to a probability of incorrect call for a base of 0.001. Some basic sequencing metrics are indicated in Table 7.4. Information on cluster density and total number of reads and high quality reads per lane is reported in Table 7.5.

Table 7.4: Summary statistics on basic quality control metrics for targeted sequencing experiment.

	Reads	Raw clusters per lane	Perfect index reads	% Q30 Bases	Mean Quality Score
1st IQR	379,659	0.09	98.21	94.9	35.82
Median	433,992	0.1	98.41	95.33	35.9
Mean	457,607	0.1	98.16	95.36	35.91
3rd IQR	502,337	0.11	98.57	96.09	36.06

Table 7.5: Information on cluster density and reads for the 5 HiSeq2500 lanes that were used for sequencing the amplified libraries of the targeted sequencing experiment.

Lane	Tiles	Density (K/mm^2)	Clusters PF (%)	Reads (M)	Reads PF (M)	% ≥ Q30
5	96	731 +/- 38	97.89 +/- 0.72	203.01	198.71	98.37
6	96	916 +/- 46	97.05 +/- 0.82	254.52	247	97.66
7	96	932 +/- 46	96.87 +/- 0.83	258.99	250.88	97.58
8	96	713 +/- 39	97.90 +/- 0.73	198.08	193.91	98.38
8.2	96	853 +/- 44	97.0 +/- 0.8	237.06	229.92	96.6

7.3 Statistical methods

7.3.1 Logistic regression

Logistic regression models have been used to assess possible associations between genetic loci and specific subtypes of breast cancer. For the genotyping data that was on the iCOGS platform we performed the analysis using plink. For imputed data the logistic regression models were built in snptest.

We also investigated potential GxE interactions introducing multivariate models taking into account genetic and environmental factors, the interaction term, and age as an additional covariate. These models were run on SAS version 9.4.

Logistic regression models were deemed more appropriate to use since they allow the inclusion of more than one explanatory variable (dependent variable) and those can either be dichotomous, ordinal, or continuous. This model also provides a quantified value for the strength of the association adjusting for other variables (removes confounding effects). The exponential of coefficients correspond to odd ratios for the given factor. The major advantage of this model as opposed to a Cochran-Armitage (CA) association test is that it is suitable to construct odds ratios and confidence intervals since it can directly measure the effect size of genotypes to phenotypes. On the contrary, CA test measures the difference in counts between the cases and the controls with respect to the average of risk alleles in each group [279].

7.3.2 Fishers exact test

Fishers exact test has been used in a case-control manner for gene based rare variant analysis. One sided test was selected since the expectation was enrichment rather than deficit of variants in cases over controls. The Fishers exact test seems more appropriate in rare variant association studies (RVAS) since it is more accurate than other tests when the expected numbers are relatively small which is true in the instance of rare variants.

7.3.3 Mann-Whitney U-test

Mann Whitney U-test has been used when comparing continuous data such as age of diagnosis. This test was preferentially selected over the traditional students t-test since it is non parametric and does not rely on assumptions such as normal distribution of the data.

7.3.4 Multiple testing corrections

Bonferroni correction has been used in several different analytical processes during the course of this project. In GWAS the gold standard is to use the genome wide significance cut-off of $P<5x10^{-8}$. It has been estimated that there are approximately one million independent loci across the genome. A Bonferroni correction for 1 one million markers has led to the usage of the genome wide significance by almost all scientific genetic community. The Bonferroni correction corresponds to the division of the α by the number of tests. For α=0.05 and 5 different tests, we can correct for 5 tests by dividing α by 5 and the new adjusted significance threshold will be 0.01.

7.3.5 Power calculations

In order to estimate the appropriate sample size to allow the identification of rare variant association during the phase II targeted sequencing project, a power calculation was conducted. In a power calculation several factors are taken into account to estimate the power of identifying true signals. The power of a study is defined as the probability of identifying a true association and corresponds to one minus the probability of falsely accepting the null hypothesis of no association. This can be translated to one minus type II error where type II error = b and $Power = 1 - b$. The factors included in the model are the minor allele frequency (MAF) of the variant of interest, the expected effect size (OR), the prevalence of the disease, as well as the significance threshold α. Due to the nature of the project, aiming the investigation of rare variants hypothesising that variants on the same gene with similar features might have the same effect, we can replace the MAF on the model with the combined allele frequency (CAF) of aggregated rare variants with similar characteristics, protein truncating or predicted to be deleterious, and estimate the power to detect gene based association.

Using the Bonferroni corrected threshold of significance (*P*=0.0036), and the sample size that we have (approximately 2,300 cases and 1,600 controls), we obtain enough power (>97%) to detect any association, for highly penetrant with OR=10 and combined allele frequency of 0.01%. In the occasion of more common variants with CAF=0.5% and moderate penetrance with OR=3, we have 87% power to detect association [280].

7.4 Analysis/ Bioinformatics tools

7.4.1 Genotypic data analysis

Several different tools and methods have been used to analyse the genotyping data set that was generated from the iCOGS custom Illumina arrays. The initial scanned fluorescent signals are converted into genotypes and undergo quality control. Furthermore, statistical analysis allows the evaluation of possible association of candidate markers with specific breast cancer subtypes. The process that was followed will be thoroughly described in this chapter.

7.4.1.1 **Genome Studio**

Initial quality control of the data was performed using genome studio (Illumina). The raw scanning files (.idat) were loaded into genome studio where all quality control metrics were investigated to ensure these were within the acceptable thresholds. Once the whole data set is loaded, genotypes were called using Illuminas GenCall algorithm (embedded on Genome Studio). Moreover, Genotypic statistics such as call frequency and call rates were calculated.

Sample and variant quality control can occur both in Genome Studio and after the genotypic data has been extracted in a .ped and .map format. Initial QC occurred in Genome Studio with four major filtering criteria: (i) Gen-train score >0.4, which corresponds to how well the three genotypic clusters are separated from each other, with values ranging from 0-1, (ii) call frequency > 95%, which corresponds to the proportion of samples that have been called for each individual variant, (iii) call rate > 95%, which corresponds to the proportion of variants that have been successfully called for each individual, and (iv) the Gencall rate > 0.25, which relates to individual genotypes. Gencall rate ranges from 0-1 and corresponds to how far a genotypic point is from the centre of the genotypic cluster that it belongs.

7.4.1.2 **Plink**

Plink (v1.07) was used for additional quality control of the genotyping data [281]. In more detail, the following criteria were used in order to exclude variants, MAF, call frequency and deviation from Hardy-Weinberg equilibrium (HWE). In addition, individuals were excluded based on low call rates.

Recoding of the genotyping data into dosage was also conducted in Plink using the –recode-A command, additionally this toolset was used to conduct the association study using the final individual data sets and variants that passed all filtering criteria.

The logistic model used included the disease phenotype as a response variable (case control), individual variants as predictor variables, and the first five PCs, after excluding ethnic outliers, as covariates in the model to correct for any underlying residual noise.

Genomic inflation factor was calculated to ensure it remained under the acceptable threshold of $\lambda<1.10$. The platform used was a custom chip designed by cancer consortia, including BCAC, and is enriched for variants that already predispose to different types of cancer. In this regard, in order to assess inflation we selected a subset of SNPs that have been previously selected by the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) in an attempt to exclude regions that have already been associated with breast cancer and have been selected for fine-mapping.

QQ-plots were also generated in order to visualise the distribution of association and inspect possible deviation of the observed associations compared to the expected. By comparing the QQ-plots that have been generated using the final data set of variants and the one that has been generated using only SNPs selected by PRACTICAL, dramatic genotyping platform enrichment for variants that are associated with breast cancer becomes apparent.

In order to identify individuals that are related we conducted a cryptic related analysis and excluded one individual for every pair that had a PIHAT >0.185. That led to the inclusion of individuals with relatedness more distant than second degree in the analysis.

7.4.1.3 **EIGENSTRAT**

PCA has been utilised in order to identify residual factors or vectors that might influence the case control analysis. This method is based on reducing dimensionality of the data and identifying and quantifying the key vectors that appear to influence the data. The software used for this analysis was EIGENSOFT from EIGENSTRAT (v3.0). This method was applied for the common variant genotyping analysis as well as for the rare variant case control analysis based on exome sequencing. For both studies, a subset of uncorrelated common variants ($r^2<0.5$, MAF<5%) were selected. The most significant vectors generated by the initial PCA correspond to ethnic differences. Samples were excluded if they failed to lie within the European cluster. In terms of the common variant genotyping project, genotypic data on three HapMap2 populations (CEU, JPT_CHB, and YRI) were used as a reference to ensure samples were clustering within the expected population. The first five PCs or Eigenvectors were used as covariates in the logistic regression model that was used to identify possible associations of common variants with different breast cancer subtypes.

7.4.2 Imputation

An imputation step was also undertaken to fine-map potential loci predisposing to DCIS. In brief, imputation exploits linkage disequilibrium properties of variants that are more likely to be inherited together and with this information it estimates the genotypes of variant sites that are not captured by the genotyping platform. As previously mentioned, iCOGS custom array constituted the genotyping platform used and it comprises more than 211,000 variants. By employing the described imputation methodology, we were able to impute genotypes of more than 2,000,000 variants across the genome using the phase I of the 1000 genomes version 3 as a reference study population.

7.4.2.1 **SHAPEIT**

Shapeit (v2.r644) was used to reformat the genotypes and generate haplotypes that can be compared to the reference haplotypes downstream [282]. Using Shapeit allows not only phasing of whole chromosomes, including chromosome X but also phasing of individuals with any level of relatedness, preparing the input for the imputation process. The missingness in regards to individuals and variants was set to 5%.

7.4.2.2 **Impute2**

The imputation step was conducted utilising the Impute2 tool [283]. Once all samples had been phased using Shapeit genotypes were imputed based on linkage disequilibrium properties. Including the -use_prephased_g option allows for imputation using pre-phased haplotypes. Regions were split into 578 chunks of 5Mb for computational reasons. Imputed genotypes were subsequently merged per chromosome.

7.4.2.3 **SNPtest**

The output of the imputation is probabilistic, meaning each variant will not get called as homozygous or heterozygous but will instead get three probability values for each of the three possible genotypes (for bi-allelic markers). A genetic tool that has been designed to handle this format of input data in order to conduct association studies is snptest (v2.4.1) [284]. The model used is very similar to the one used for the genotyped variant association studies, where a logistic regression model is built and each marker is tested individually taking into account the same five PCs as covariates. Due to the large number of tests, each chromosome was individually tested and output files were stored separately. Variants with $P<0.05$ were then combined in a final output file.

7.4.3 Sequencing data analysis

Our attempts at identifying rare variants predisposing to breast cancer were focused on next generation sequencing and, more specifically, WES. The top candidate genes were also followed up in a phase II study where they were thoroughly interrogated using a custom made gene panel, using a targeted sequencing method. The sequencing method that has been used for both phases of the study was the sequencing by synthesis (SBS) provided by Illumina. The data processing, quality control and analysis, along with the different tools used, will be described in depth along this chapter.

7.4.3.1 **Btrim**

Btrim was the tool chosen to trim the sequences from their 5 end, which constitutes a very important procedure in amplicon-based experiments [285]. Since there is a PCR amplification step, the primer sites do not correspond to the genomes of the individuals but to the predefined primers. Therefore it is crucial to remove those bases from the sequencing reads since they influence the final calls. To overcome this Btim can take a set of sequences that correspond to the primers and remove them from each 5 end of the sequencing reads. Read 1 and read 2 are getting treated separately, and therefore are getting desynchronised, so a second Btrim command should be used to synchronise the reads again in order for these to be paired.

7.4.3.2 **Novoalign**

Once the primer sites had been trimmed from the 5 ends of the reads, each read was aligned to the reference genome (http://www.novocraft.com). The GRCh37 version of the genome has been used as a reference. The tool to perform the alignment was Novoalign. Gap opening penalty=65 and gap extension penalty=7 thresholds were applied.

7.4.3.3 **Picard tools**

The AddOrReplaceReadGroups command from Picard tools (v1.74) was used to add the meta-information in each .bam file (https://github.com/broadinstitute/picard). Additionally, Picard tools have been used to convert the mapped (.bam) files downloaded from the TCGA to raw sequencing format (.fastq) using the SamToFASTQ command, this step is crucial in terms of data homogeneity. Moreover, the downloaded .bam files from TCGA were converted back to .fastq files, this step allows us to perform the alignment as previously described in Section 7.4.3.2.

7.4.3.4 **Bedtools**

In order to assess the proportion of the reads that have been mapped to the region of our interest, Bedtools (v2.17.0) has been used [286]. The coverageBed command from Bedtools also provides information on the exact depth of coverage for each specific region, and therefore apart from average coverage, and percentage of target region covered by a certain depth, we can identify the proportion of a gene that has been captured efficiently at a certain cut-off. Using Rs library ggplot2, we can visualise the coverage data in the form of a histogram per target region incorporating data across all individuals. We achieved at least 20x for 90 per cent of the capture region for 4536 out of 4599 samples (98.6%), Figure 7.6.

The coverage achieved for the regions of interest was satisfactory, with more than 97% of the target region being captured by at least 20 reads on average per sample. The mean coverage of our target region was 803 reads on average across all samples. There were specific amplicons that failed to amplify, and that pattern was consistent across the majority of the samples. A list of amplicons that failed to amplify is reported in Table 7.6.

Table 7.6: Amplicons failed to amplify during the targeted sequencing project.

Amplicon ID	Exonic region	Amplicon size	% GC	Comment
ATRIP_t1_1	chr3:48488240-48488506	176	77	High GC content
ATRIP_t1_2	chr3:48491465-48491586	178	78	High GC content
BRCA1_t6_5	chr17:41251782-41251907	195	44	Designed without SNP and repeat annotation
CDH1_t1_1	chr16:68771309-68771376	200	71	High GC content
CDH1_t12_1	chr16:68855894-68856138	199	47	
CHEK2_t1_1	chr22:29130381-29130719	191	47	Primers designed within a repeat region
CHEK2_t1_3	chr22:29121221-29121365	185	58	
IDE_t16_1	chr10:94235631-94235761	192	26	Primer designed within a repeat region
IDE_t24_3	chr10:94215323-94215410	195	35	
MME_t4_2	chr3:154834258-154834358	182	33	
MME_t5_1	chr3:154834443-154834558	198	30	
PALB2_t13_3	chr16:23614770-23615000	200	41	
PALB2_t4_21	chr16:23646173-23647665	180	21	
SRA1_t1_e2_3	chr5:139936722-139937047	191	65	

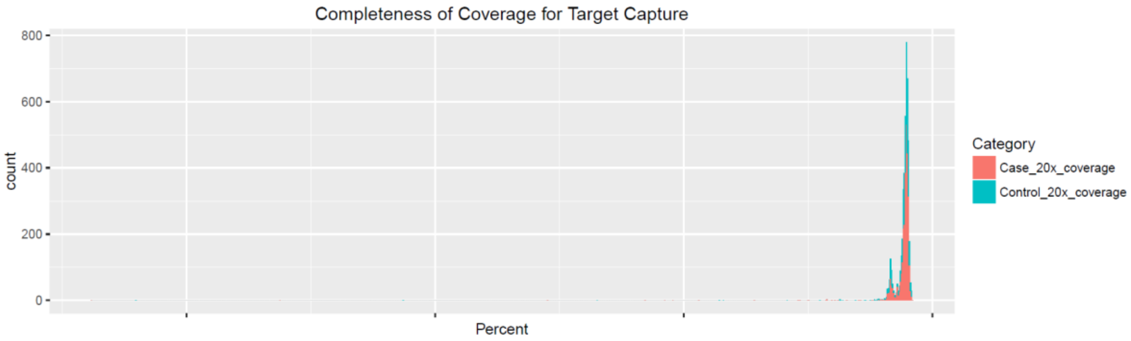


Figure 7.6: Completeness of coverage across all samples for the whole target region incorporating genomic locations across 20 genes.

7.4.3.5 **Samtools**

Using Samtools (v0.1.18) -view command, the .sam files which are the output of the alignment are converted to the binary format .bam for computational efficiency [287]. There are two main algorithms used in this project for variant calling. One of them is the Samtools algorithm with the -mpileup command where each sample is getting called individually and the final data is later on combined for the statistical analysis (see commands). Samtools has been used for variant calling and in particular the mpileup command. Several quality metrics were taken into account when calculating variants that pass the filtering criteria. Variants were filtered using the vcfutils.pl varFilter command and some of the default settings altered to accommodate the particularities of the targeted sequencing project which is amplicon based. The minimum read depth was increased to 10 from the default value which was 2. We reduced the threshold of the minimum number for a single nucleotide substitution to be within a certain number of bp around a gap to be filtered. The default was 3 but this was altered to 0. The strand bias filter was also removed since we expect strand bias in our sequencing experiment. The expected strand bias is due to the fact that not all regions are covered by both the forward and the reverse primers (which correspond to read 1 and read 2), due to the fact that the amplicons are up to 200 bp whereas the sequencing reads are 125 bp.

7.4.3.6 **GATK**

The second variant calling approach followed is using GATKs algorithms (v3.2-2) where each sample is initially called individually but later on combined using a multi-sample calling algorithm [288]. It employs the haplotype caller from GATK in conjunction with the joint genotype caller that allows for multi-sample calling. This approach could be more appropriate for statistical evaluation of variants where each variant position is compared against all other samples. The statistical outcome of this approach is more accurate for relatively common variants. Apart from calling a variant or not at a certain position, this algorithm gives information on whether that position is captured for each sample. Therefore, the statistical analysis can include only individuals that each position has been appropriately covered. This method has the advantage of investigating variant and non-variant positions across all samples in a data set which allows for more accurate genotype calls [289].

7.4.3.7 **VCFtools**

A filtering step of the vcf files can occur using vcftools (v0.1.14) where we can redefine the target region as well as specific quality metrics such as the minimum base quality allowed for a variant to remain in the final called variant list [290]. The minimum genotyping quality was set to 20 and minimum read depth to 10.

7.4.3.8 **Annovar**

Variants that have fulfilled all filtering criteria in terms of quality control metrics were annotated using the Annovar tool (Apr_2015) [291]. Annovar has several different built in databases incorporated and allows for different annotations. These annotations can be separated by their type and can include genomic reference, gene based annotations, variant type annotations, population frequencies, but also pathogenicity prediction scores. In brief, the annotations used were gene based; variant type and class was assessed including the nucleotide and amino-acid change. Variant frequency on the European population was also annotated using three different sources (1000 genomes, ESP, ExAC) that will be discussed downstream on section 7.4.6. Finally, variant deleteriousness was assessed using prediction scores including CADD, DANN, SIFT, and two different versions of PolyPhen2. There are other alternative methods of annotating the variants such as the Variant Effect Predictor (VEP) from ENSEMBL and the SnpEff from the GATK but the Annovar method has already been established in our research group and since it contains all the information that was necessary to conduct these studies and annotate the variants, it was used in preference compared to the other methods.

7.4.3.9 **Variant filtering**

Variants that were called by both variant callers (GATK, Samtools) were used for the targeted sequencing data analysis in order to minimise the number of false positive calls. Variants were further filtered based on read depth (DP), quality control score (QC), and genotypic quality (GQ). All variants with DP<10, QC<20, or GQ<20 were excluded from the analysis. Furthermore, variants were filtered according to their class. Synonymous variants were excluded, along with intronic variants that are more than two base-pairs apart from the splice junction. Frameshift indels, stop-gain, stop-loss and splicing variants were considered protein truncating whereas non-synonymous variants were considered protein altering as per Table 7.7.

Table 7.7. Variant type definitions and descriptions.

Variant class	Variant type	Description
Missense/ non-synonymous	Protein altering	A single nucleotide substitution that leads to an amino-acid substitution
Stop-gain	Protein truncating	A single nucleotide substitution that leads to the introduction of a premature stop codon
Stop-loss	Protein truncating	A single nucleotide substitution that leads to the loss of the wild type stop codon
Frameshift indel	Protein truncating	An insertion or deletion of a number of nucleotides that leads to a frame-shift of the amino-acid sequence
Non-frameshift indel	Protein altering	An insertion or deletion of a number of nucleotides that leads to the addition or deletion of a number of amino-acids
Splicing	Protein truncating	A single nucleotide substitution in the essential splice site 1 or 2 nucleotides adjacent to the splice site
Synonymous	Silent	A single nucleotide substitution that leads to the same amino-acid being encoded

7.4.3.10 **KING**

In terms of the exome sequencing data set, cryptic relatedness analysis was conducted and a kinship matrix was generated. The pairwise relatedness matrix was constructed using the KING tool [292]. This method generates pairwise relationship values amongst every pair of samples in the data set. The outcome kinship coefficient (KC) values correspond to duplicates/monozygotic twins (KC>0.354), first degree relatives (0.177<KC<0.354), second degree relatives (0.0884<KC<0.177), third degree relatives (0.0442<KC<0.0884) and less than third degree (KC<0.0442). For every pair of samples with a KC>0.0884 one sample was randomly removed. A set of 9,569 common uncorrelated variants was extracted from the WES data and was reformatted into .map and .ped format to be used as an input for the relatedness analysis. The same set of variants was used for the PCA with regards to WES data.

7.4.3.11 **EPACTS**

Efficient and Parallelizable Association Container Toolbox (EPACTS) is a framework that enables gene based or single variant association using next generation sequencing (NGS) data. Gene burden methods test for association of a gene with a disease by examining whether a particular class of alleles in a gene is enriched or depleted in cases versus random controls from the general population. Using EPACTS (v3.2.3) and applying a Fishers exact test we conducted gene burden tests under the dominant model of inheritance. In order to test for association of a particular class of variants in a gene with the phenotype of our interest, a threshold cut-off of MAF<0.01 has been used. Due to the nature of breast cancer, we expect enrichment and not deficit of damaging alleles in our cases versus the controls, which allows using a one-tailed chi square test. The Fishers exact test seems more appropriate test in RVAS since it is more accurate than other tests when the expected numbers are relatively small such as in the instance of rare variants.

7.4.4 Prediction tools

One of the key challenges in variant identification and prioritisation in clinical genetics is the separation of variants with a suggestive effect on the genes function from those that are more likely to be neutral and have no significant effect. Several prediction tools have been developed over the last year in order to prioritise variants and to assess a score of potential deleteriousness on each variant. Different algorithms use different features such as evolutionary tolerance, structural similarities in terms of amino acid changes, and others. The prediction tools that have been used in this project will be briefly described separately.

7.4.4.1 **SIFT**

SIFT is an online tool that has been developed over a decade ago and is used in genetics to assess potential deleteriousness for non-synonymous variants [293]. Their algorithm is based on how evolutionary tolerant a variant is. Variants in regions that are highly conserved amongst different species are more likely to have a larger effect on the genes function. A variant with a SIFT score of <0.05 is predicted to be deleterious.

7.4.4.2 **Polyphen2**

Another tool that has been used is Polyphen2 [294]. This tool also performs predictions for non-synonymous variants. This algorithm is heavily dependent on structural changes that a variant might cause to the protein. It takes into account the amino-acid similarities and differences and scores variants according to the expected impact of the amino-acid change. Two different databases of Polyphen2 were used, with one of them focusing on rare “disease causing” variants, and the second one scoring variants for an “effect on genes function”. The thresholds for deleteriousness are slightly different between the two databases even though both databases use a spectrum of scores that range between 0 and 1.

HVAR is used for diagnostics of Mendelian diseases and variants with a score higher than 0.909 are considered probably damaging. The HDIV database is used for evaluating rare variants at loci involved in complex diseases and variants with a score higher than 0.957 are considered probably damaging.

7.4.4.3 **Combined Annotation–Dependent Depletion (CADD)**

CADD is a recently developed tool that predicts the deleteriousness of all possible single nucleotide variants as well as small indels that could occur in the human genome [295]. The algorithm is based on a support vector machine (SVM). It has two main advantages over the majority of the other commonly used tools. The first one is the fact that it can score all possible substitutions and small indels rather than just non-synonymous variants, and secondly the predictions are made by using a matrix of several different features including conservation, structural effect, gene based location, co-localisation with regulatory elements, effect on splicing, and others. CADD is based on integrating many diverse annotations to provide a single Phred like score for each variant. The CADD score correlates with allelic diversity, annotations of functionality, pathogenicity, disease severity, experimentally measured regulatory effects and complex trait associations, highly ranks known pathogenic variants. Since CADD scores are based on Phred probabilities, CADD score of 10 corresponds to a variant predicted to be on the top 10% of variants in terms of deleteriousness. CADD score of 20 corresponds to the top 1% and CADD score of 30 corresponds to the top 0.1%. The authors of the tool suggest a cut-off of 15 for suggestive deleterious variants. After correlating CADD scores with known variants with large effect size that predispose to breast cancer we concluded that a cut-off of 30 will be more appropriate for our analyses.

7.4.4.4 **DANN**

DANN is a tool developed soon after the CADD tool. It utilises the same information as the CADD tool but differs in terms of calculating deleteriousness since it is based on deep neural network machine learning as opposed to SVM which is the method that the CADD developers preferred [296]. They have shown that it outperforms CADD in terms of identifying truly pathogenic variants. This algorithm can take into account non-linear relationships between different factors. Another difference between CADD and DANN is that CADDs output is the C score which is in the Phred scale whereas DANN output values range between 0-1. Their developers set a cut-off of 0.995 to distinguish likely deleterious variants from likely benign variants with variants having a DANN score >0.995 being more likely to have a detrimental effect on the genes product function.

7.4.5 Genome browsers

7.4.5.1 UCSC

The University of California Santa Cruz (UCSC) (http://genome.ucsc.edu/) genome browser has been used for several tasks throughout the course of this PhD [297]. During primer designing, the BLAST tool has been used to ensure that primers designed were not annealing at multiple locations on the genome and were specific to the target region. The primer sites were also intersected with known polymorphisms from dbSNP (v142) to ensure that no common variant lie within the primer sites that could lead to preferential allelic amplification. This process was performed using UCSCs variant annotation integrator. The final usage of UCSC was the lift-over tool to convert coordinates from different genome assemblies. This tool was particularly useful when the iCOGS genotyping data which was in hg18 was required to be converted to hg19.

7.4.5.2 ENSEMBL

Ensembl (http://www.ensembl.org/) is a joint project between EMBL-EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes including human [298]. It has been used to investigate particular variants of interest. There are several features available that allow exploring different features of variants such as whether there is a phenotype associated with them, and what is their frequency in different populations. It also provides information on linkage disequilibrium (LD) variants and evolutionary information in terms of phylogenetic context. A phylogeny analysis a rare *CDH1* variant and its evolution conservation (how evolutionary conserved a variant is across species) was also conducted using the Ensembl browser. Moreover, Ensembl allows access to the nucleotide sequence of genes of interest, including or excluding introns. This can be used as a template in occasions of primer designing. It also allows to interrogate gene sequences and visualise whether known variants exist or whether a variant identified during Sanger sequencing has been previously reported.

7.4.6 Online databases

The following databases have been used to filter the variants that were the output of sequencing experiments. A general cut-off of minor allele frequency (MAF) <1% was used throughout the analyses unless otherwise specified. Any variant with MAF>1% in any of the following databases was discarded.

7.4.6.1 **ClinVar**

ClinVar is a database where relationships between variants and human disease are archived and aggregated (https://www.ncbi.nlm.nih.gov/clinvar/). A clinically useful website that has been established and serves as a variant repository is ClinVar. Different research groups, consortia or diagnostic laboratories can upload information on the clinical utility of variants identified and help to build an accurate database on genetic variation across all genes implicated with disease. Variants can be classified as benign, likely benign, unknown significance, likely pathogenic and pathogenic depending on different genomic features as well as ascertainment criteria.

7.4.6.2 **1000 Genomes**

The 1000 Genomes project is a deep catalogue of Human Genetic variation. It includes low coverage (4x) sequencing data on 2,504 individuals from European, Asiatic, African and American populations.

7.4.6.3 **ExAC**

The Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org) is a coalition of researchers that collaborate in order to aggregate exome sequencing data from different large-scale sequencing projects, enabling data access to the scientific community. The data set provided includes 60,706 unrelated individuals undergone WES as part of various genetic studies. Individuals affected by severe paediatric conditions, have been removed in order for the data set to be used as a reference for allele frequencies.

7.4.6.4 **ESP**

The aim of the NHLBI GO Exome Sequencing Project (ESP) is to understand the contribution of rare genetic variation to heart, lung and blood disorders by utilising WES technology on deeply phenotyped populations. The current data set includes variants on 4,300 European-Americans unrelated individuals.

References

1. Kohler BA, Sherman RL, Howlader N, Jemal A, Ryerson AB, Henry KA, Boscoe FP, Cronin KA, Lake A, Noone AM *et al*: **Annual Report to the Nation on the Status of Cancer, 1975-2011, Featuring Incidence of Breast Cancer Subtypes by Race/Ethnicity, Poverty, and State (vol 107, djv048, 2015)**. *Jnci-J Natl Cancer I* 2015, **107**(7).
2. Dixon JM: **ABC of breast diseases**, 4th edn. Chichester, West Sussex: Blackwell Pub.; 2012.
3. DeSantis C, Ma J, Bryan L, Jemal A: **Breast cancer statistics, 2013**. *CA: a cancer journal for clinicians* 2014, **64**(1):52-62.
4. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML, Jackson RD, Beresford SA, Howard BV, Johnson KC *et al*: **Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Womens Health Initiative randomized controlled trial**. *Jama* 2002, **288**(3):321-333.
5. Ravdin PM, Cronin KA, Howlader N, Berg CD, Chlebowski RT, Feuer EJ, Edwards BK, Berry DA: **The decrease in breast-cancer incidence in 2003 in the United States**. *The New England journal of medicine* 2007, **356**(16):1670-1674.
6. Henderson IC: **Breast cancer : fundamentals of evidence-based disease management**.
7. Kotsopoulos J, Chen WY, Gates MA, Tworoger SS, Hankinson SE, Rosner BA: **Risk factors for ductal and lobular breast cancer: results from the nurses health study**. *Breast Cancer Research* 2010, **12**(6).
8. Clavel-Chapelon F, Group EN: **Cumulative number of menstrual cycles and breast cancer risk: results from the E3N cohort study of French women**. *Cancer causes & control : CCC* 2002, **13**(9):831-838.
9. Anothaisintawee T, Wiratkapun C, Lerdsitthichai P, Kasamesup V, Wongwaisayawan S, Srinakarin J, Hirunpat S, Woodtichartpreecha P, Boonlikit S, Teerawattananon Y *et al*: **Risk factors of breast cancer: a systematic review and meta-analysis**. *Asia-Pacific journal of public health / Asia-Pacific Academic Consortium for Public Health* 2013, **25**(5):368-387.
10. Assi V, Warwick J, Cuzick J, Duffy SW: **Clinical and epidemiological issues in mammographic density**. *Nature reviews Clinical oncology* 2012, **9**(1):33-40.
11. McCormack VA, Silva IDS: **Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis**. *Cancer Epidem Biomar* 2006, **15**(6):1159-1169.
12. Banks E, Beral V, Bull D, Reeves G, Austoker J, English R, Patnick J, Peto R, Vessey M, Wallis M *et al*: **Breast cancer and hormone-replacement therapy in the Million Women Study**. *Lancet* 2003, **362**(9382):419-427.
13. Reeves GK, Beral V, Green J, Gathani T, Bull D, Million Women Study C: **Hormonal therapy for menopause and breast-cancer risk by histological type: a cohort study and meta-analysis**. *The Lancet Oncology* 2006, **7**(11):910-918.
14. **Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. Collaborative Group on Hormonal Factors in Breast Cancer**. *Lancet* 1997, **350**(9084):1047-1059.
15. Beral V, Million Women Study C: **Breast cancer and hormone-replacement therapy in the Million Women Study**. *Lancet* 2003, **362**(9382):419-427.
16. Reeves GK, Pirie K, Green J, Bull D, Beral V, Million Women Study C: **Reproductive factors and specific histological types of breast cancer: prospective study and meta-analysis**. *British journal of cancer* 2009, **100**(3):538-544.
17. Ritte R, Tikk K, Lukanova A, Tjonneland A, Olsen A, Overvad K, Dossus L, Fournier A, Clavel-Chapelon F, Grote V *et al*: **Reproductive factors and risk of hormone receptor positive and negative breast cancer: a cohort study**. *BMC cancer* 2013, **13**.
18. Thorbjarnardottir T, Olafsdottir EJ, Valdimarsdottir UA, Olafsson O, Tryggvadottir L: **Oral contraceptives, hormone replacement therapy and breast cancer risk: A cohort study of 16 928 women 48 years and older**. *Acta Oncol* 2014, **53**(6):752-758.
19. Dolle JM, Daling JR, White E, Brinton LA, Doody DR, Porter PL, Malone KE: **Risk factors for triple-negative breast cancer in women under the age of 45 years**. *Cancer epidemiology, biomarkers & prevention : a publication of the American*

Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 2009, **18**(4):1157-1166.

20. Collaborative Group on Hormonal Factors in Breast C: **Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies**. *The Lancet Oncology* 2012, **13**(11):1141-1151.
21. Lambe M, Hsieh CC, Trichopoulos D, Ekbom A, Pavia M, Adami HO: **Transient Increase in the Risk of Breast-Cancer after Giving Birth**. *New Engl J Med* 1994, **331**(1):5-9.
22. Collaborative Group on Hormonal Factors in Breast C: **Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease**. *Lancet* 2002, **360**(9328):187-195.
23. Cheang MC, Chia SK, Voduc D, Gao D, Leung S, Snider J, Watson M, Davies S, Bernard PS, Parker JS *et al*: **Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer**. *J Natl Cancer Inst* 2009, **101**(10):736-750.
24. Brinton LA, Smith L, Gierach GL, Pfeiffer RM, Nyante SJ, Sherman ME, Park Y, Hollenbeck AR, Dallal CM: **Breast cancer risk in older women: results from the NIH-AARP Diet and Health Study**. *Cancer causes & control : CCC* 2014, **25**(7):843-857.
25. Franceschi S, Favero A, LaVecchia C, Baron AE, Negri E, dalMaso L, Giacosa A, Montella M, Conti E, Amadori R: **Body size indices and breast cancer risk before and after menopause**. *International journal of cancer* 1996, **67**(2):181-186.
26. Lubin F, Ruder AM, Wax Y, Modan B: **Overweight and Changes in Weight Throughout Adult Life in Breast-Cancer Etiology - a Case-Control Study**. *American journal of epidemiology* 1985, **122**(4):579-588.
27. vandenBrandt PA, Dirx MJM, Ronckers CM, vandenHoogen P: **Height, weight, weight change, and postmenopausal breast cancer risk: The Netherlands Cohort Study**. *Cancer Cause Control* 1997, **8**(1):39-47.
28. Ahn J, Schatzkin A, Lacey JV, Albanes D, Ballard-Barbash R, Adams KF, Kipnis V, Mouw T, Hollenbeck AR, Leitzmann MF: **Adiposity, adult weight change, and postmenopausal breast cancer risk**. *Arch Intern Med* 2007, **167**(19):2091-2102.
29. Krishnan K, Bassett JK, MacInnis RJ, English DR, Hopper JL, McLean C, Giles GG, Baglietto L: **Associations between Weight in Early Adulthood, Change in Weight, and Breast Cancer Risk in Postmenopausal Women**. *Cancer Epidem Biomar* 2013, **22**(8):1409-1416.
30. Thomson CA, Van Horn L, Caan BJ, Aragaki AK, Chlebowski RT, Manson JE, Rohan TE, Tinker LF, Kuller LH, Hou LF *et al*: **Cancer Incidence and Mortality during the Intervention and Postintervention Periods of the Womens Health Initiative Dietary Modification Trial**. *Cancer Epidem Biomar* 2014, **23**(12):2924-2935.
31. Seitz HK, Pelucchi C, Bagnardi V, La Vecchia C: **Epidemiology and Pathophysiology of Alcohol and Breast Cancer: Update 2012**. *Alcohol Alcoholism* 2012, **47**(3):204-212.
32. Chen WY, Rosner B, Hankinson SE, Colditz GA, Willett WC: **Moderate alcohol consumption during adult life, drinking patterns, and breast cancer risk**. *Jama* 2011, **306**(17):1884-1890.
33. Heikkila K, Nyberg ST, Madsen IE, de Vroome E, Alfredsson L, Bjorner JJ, Borritz M, Burr H, Erbel R, Ferrie JE *et al*: **Long working hours and cancer risk: a multi-cohort study**. *British journal of cancer* 2016, **114**(7):813-818.
34. Lagerros YT, Hsieh SF, Hsieh CC: **Physical activity in adolescence and young adulthood and breast cancer risk: a quantitative review**. *Eur J Cancer Prev* 2004, **13**(1):5-12.
35. Catsburg C, Kirsh VA, Soskolne CL, Kreiger N, Bruce E, Ho T, Leatherdale ST, Rohan TE: **Associations between anthropometric characteristics, physical activity, and breast cancer risk in a Canadian cohort**. *Breast Cancer Res Tr* 2014, **145**(2):545-552.
36. Lahmann PH, Friedenreich C, Schuit AJ, Salvini S, Allen NE, Key TJ, Khaw KT, Bingham S, Peeters PHM, Monninkhof E *et al*: **Physical activity and breast cancer risk: The European prospective investigation into cancer and nutrition**. *Cancer Epidem Biomar* 2007, **16**(1):36-42.
37. Antoniou AC, Pharoah PD, McMullan G, Day NE, Stratton MR, Peto J, Ponder BJ, Easton DF: **A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes**. *British journal of cancer* 2002, **86**(1):76-83.

38. Antoniou AC, Hardy R, Walker L, Evans DG, Shenton A, Eeles R, Shanley S, Pichert G, Izatt L, Rose S *et al*: **Predicting the likelihood of carrying a BRCA1 or BRCA2 mutation: validation of BOADICEA, BRCAPRO, IBIS, Myriad and the Manchester scoring system using data from UK genetics clinics**. *Journal of medical genetics* 2008, **45**(7):425-431.
39. Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, Wilson M, Howell A: **Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme**. *Journal of medical genetics* 2003, **40**(11):807-814.
40. Sinn HP, Kreipe H: **A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition**. *Breast care* 2013, **8**(2):149-154.
41. Edge SB, Compton CC: **The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM**. *Annals of surgical oncology* 2010, **17**(6):1471-1474.
42. Bloom HJ, Richardson WW: **Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years**. *British journal of cancer* 1957, **11**(3):359-377.
43. Wong H, Lau S, Yau T, Cheung P, Epstein RJ: **Presence of an in situ component is associated with reduced biological aggressiveness of size-matched invasive breast cancer**. *British journal of cancer* 2010, **102**(9):1391-1396.
44. Ruszczyk M, Zirpoli G, Kumar S, Bandera EV, Bovbjerg DH, Jandorf L, Khoury T, Hwang H, Ciupak G, Pawlish K *et al*: **Breast cancer risk factor associations differ for pure versus invasive carcinoma with an in situ component in case-control and case-case analyses**. *Cancer Cause Control* 2016, **27**(2):183-198.
45. Li CI, Daling JR, Malone KE: **Age-specific incidence rates of in situ breast carcinomas by histologic type, 1980 to 2001**. *Cancer Epidem Biomar* 2005, **14**(4):1008-1011.
46. Leonard GD, Swain SM: **Ductal carcinoma in situ, complexities and challenges**. *Journal of the National Cancer Institute* 2004, **96**(12):906-920.
47. Kerlikowske K: **Epidemiology of ductal carcinoma in situ**. *Journal of the National Cancer Institute Monographs* 2010, **2010**(41):139-141.
48. Kim SY, Jung SH, Kim MS, Baek IP, Lee SH, Kim TM, Chung YJ, Lee SH: **Genomic differences between pure ductal carcinoma in situ and synchronous ductal carcinoma in situ with invasive breast cancer**. *Oncotarget* 2015, **6**(10):7597-7607.
49. OConnell P, Pekkel V, Fuqua SAW, Osborne CK, Clark GM, Allred DC: **Analysis of loss of heterozygosity in 399 premalignant breast lesions at 15 genetic loci**. *J Natl Cancer I* 1998, **90**(9):697-703.
50. Claus EB, Risch N, Thompson WD, Carter D: **Relationship between Breast Histopathology and Family History of Breast-Cancer**. *Cancer* 1993, **71**(1):147-153.
51. Ross DS, Wen YH, Brogi E: **Ductal carcinoma in situ: morphology-based knowledge and molecular advances**. *Advances in anatomic pathology* 2013, **20**(4):205-216.
52. Pinder SE: **Ductal carcinoma in situ (DCIS): pathological features, differential diagnosis, prognostic factors and specimen evaluation**. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2010, **23 Suppl 2**:S8-13.
53. Aguiar FN, Mendes HN, Bacchi CE, Carvalho FM: **Comparison of nuclear grade and immunohistochemical features in situ and invasive components of ductal carcinoma of breast**. *Revista brasileira de ginecologia e obstetricia : revista da Federacao Brasileira das Sociedades de Ginecologia e Obstetricia* 2013, **35**(3):97-102.
54. Leong AS, Sormunen RT, Vinyuvat S, Hamdani RW, Suthipintawong C: **Biologic markers in ductal carcinoma in situ and concurrent infiltrating carcinoma. A comparison of eight contemporary grading systems**. *American journal of clinical pathology* 2001, **115**(5):709-718.
55. Biglia N, Mariani L, Sgro L, Mininanni P, Moggio G, Sismondi P: **Increased incidence of lobular breast cancer in women treated with hormone replacement therapy: implications for diagnosis, surgical and medical treatment**. *Endocrine-related cancer* 2007, **14**(3):549-567.
56. Yoder BJ, Wilkinson EJ, Massoll NA: **Molecular and morphologic distinctions between infiltrating ductal and lobular carcinoma of the breast**. *Breast J* 2007, **13**(2):172-179.

57. Cristofanilli M, Gonzalez-Angulo A, Sneige N, Kau SW, Broglio K, Theriault RL, Valero V, Buzdar AU, Kuerer H, Buchholz TA *et al*: **Invasive lobular carcinoma classic type: response to primary chemotherapy and survival outcomes**. *J Clin Oncol* 2005, **23**(1):41-48.

58. Tubiana-Hulin M, Stevens D, Lasry S, Guinebretiere JM, Bouita L, Cohen-Solal C, Cherel P, Rouesse J: **Response to neoadjuvant chemotherapy in lobular and ductal breast carcinomas: a retrospective study on 860 patients from one institution**. *Annals of oncology : official journal of the European Society for Medical Oncology / ESMO* 2006, **17**(8):1228-1233.

59. Michaut M, Chin SF, Majewski I, Severson TM, Bismeijer T, de Koning L, Peeters JK, Schouten PC, Rueda OM, Bosma AJ *et al*: **Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer**. *Scientific reports* 2016, **6**:18517.

60. Hwang ES, Nyante SJ, Chen YY, Moore D, DeVries S, Korkola JE, Esserman LJ, Waldman FM: **Clonality of lobular carcinoma in situ and synchronous invasive lobular carcinoma**. *Cancer* 2004, **100**(12):2562-2572.

61. Warnberg F, Yuen J, Holmberg L: **Risk of subsequent invasive breast cancer after breast carcinoma in situ**. *Lancet* 2000, **355**(9205):724-725.

62. Chuba PJ, Hamre MR, Yap J, Severson RK, Lucas D, Shamsa F, Aref A: **Bilateral risk for subsequent breast cancer after lobular carcinoma-in-situ: analysis of surveillance, epidemiology, and end results data**. *J Clin Oncol* 2005, **23**(24):5534-5541.

63. Fisher ER, Land SR, Fisher B, Mamounas E, Gilarski L, Wolmark N: **Pathologic findings from the national surgical adjuvant breast and bowel project - Twelve-year observations concerning lobular carcinoma in situ**. *Cancer* 2004, **100**(2):238-244.

64. King TA, Pilewskie M, Muhsen S, Patil S, Mautner SK, Park A, Oskar S, Guerini-Rocco E, Boafo C, Gooch JC *et al*: **Lobular Carcinoma in Situ: A 29-Year Longitudinal Experience Evaluating Clinicopathologic Features and Breast Cancer Risk**. *J Clin Oncol* 2015, **33**(33):3945-+.

65. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA *et al*: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747-752.

66. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *P Natl Acad Sci USA* 2001, **98**(19):10869-10874.

67. Prat A, Parker JS, Fan C, Perou CM: **PAM50 assay and the three-gene model for identifying the major and clinically relevant molecular subtypes of breast cancer**. *Breast Cancer Res Tr* 2012, **135**(1):301-306.

68. vant Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.

69. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He XP, Hu ZY *et al*: **Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes**. *J Clin Oncol* 2009, **27**(8):1160-1167.

70. Upstill-Goddard R, Eccles D, Ennis S, Rafiq S, Tapper W, Fliege J, Collins A: **Support Vector Machine classifier for estrogen receptor positive and negative early-onset breast cancer**. *PloS one* 2013, **8**(7):e68606.

71. Yersal O, Barutca S: **Biological subtypes of breast cancer: Prognostic and therapeutic implications**. *World journal of clinical oncology* 2014, **5**(3):412-424.

72. Creighton CJ: **The molecular profile of luminal B breast cancer**. *Biologics : targets & therapy* 2012, **6**:289-297.

73. Ellis MJ, Tao Y, Luo J, AHern R, Evans DB, Bhatnagar AS, Chaudri Ross HA, von Kameke A, Miller WR, Smith I *et al*: **Outcome prediction for estrogen receptor-positive breast cancer based on postneoadjuvant endocrine therapy tumor characteristics**. *J Natl Cancer Inst* 2008, **100**(19):1380-1388.

74. Lehmann BD, Pietenpol JA: **Clinical implications of molecular heterogeneity in triple negative breast cancer**. *Breast* 2015, **24 Suppl 2**:S36-40.

75. Kreike B, van Kouwenhove M, Horlings H, Weigelt B, Peterse H, Bartelink H, van de Vijver MJ: **Gene expression profiling and histopathological characterization of triple-negative/basal-like breast carcinomas**. *Breast cancer research : BCR* 2007, **9**(5):R65.

76. Rakha EA, Elsheikh SE, Aleskandarany MA, Habashi HO, Green AR, Powe DG, El-Sayed ME, Benhasouna A, Brunet JS, Akslen LA *et al*: **Triple-Negative Breast Cancer: Distinguishing between Basal and Nonbasal Subtypes**. *Clin Cancer Res* 2009, **15**(7):2302-2310.
77. Leong ASY, Zhuang ZP: **The Changing Role of Pathology in Breast Cancer Diagnosis and Treatment**. *Pathobiology* 2011, **78**(2):99-114.
78. Schnitt SJ: **Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy**. *Modern Pathol* 2010, **23**:S60-S64.
79. Staaf J, Ringner M, Vallon-Christersson J, Jonsson G, Bendahl PO, Holm K, Arason A, Gunnarsson H, Hegardt C, Agnarsson BA *et al*: **Identification of Subtypes in Human Epidermal Growth Factor Receptor 2-Positive Breast Cancer Reveals a Gene Signature Prognostic of Outcome**. *J Clin Oncol* 2010, **28**(11):1813-1820.
80. Bilous M, Morey A, Armes J, Cummings M, Francis G: **Chromogenic in situ hybridisation testing for HER2 gene amplification in breast cancer produces highly reproducible results concordant with fluorescence in situ hybridisation and immunohistochemistry**. *Pathology* 2006, **38**(2):120-124.
81. Pothos A, Plastira K, Plastiras A, Vlachodimitropoulos D, Goutas N, Angelopoulou R: **Comparison of chromogenic in situ hybridisation with fluorescence in situ hybridisation and immunohistochemistry for the assessment of Her-2/neu oncogene in archival material of breast carcinoma**. *Acta Histochem Cytoc* 2008, **41**(3):59-64.
82. Czene K, Lichtenstein P, Hemminki K: **Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database**. *International journal of cancer* 2002, **99**(2):260-266.
83. Easton DF: **Familial risks of breast cancer**. *Breast Cancer Research* 2002, **4**(5):179-181.
84. Pharoah PDP, Day NE, Duffy S, Easton DF, Ponder BAJ: **Family history and the risk of breast cancer: A systematic review and meta-analysis**. *International journal of cancer* 1997, **71**(5):800-809.
85. Hopper JL, Carlin JB: **Familial Aggregation of a Disease Consequent Upon Correlation between Relatives in a Risk Factor Measured on a Continuous Scale**. *American journal of epidemiology* 1992, **136**(9):1138-1147.
86. Garcia-Closas M, Chanock S: **Genetic susceptibility loci for breast cancer by estrogen receptor status**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2008, **14**(24):8000-8009.
87. Broeks A, Schmidt MK, Sherman ME, Couch FJ, Hopper JL, Dite GS, Apicella C, Smith LD, Hammet F, Southey MC *et al*: **Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium**. *Hum Mol Genet* 2011, **20**(16):3289-3303.
88. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W *et al*: **A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1**. *Science* 1994, **266**(5182):66-71.
89. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D *et al*: **Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13**. *Science* 1994, **265**(5181):2088-2090.
90. Turnbull C, Rahman N: **Genetic predisposition to breast cancer: past, present, and future**. *Annual review of genomics and human genetics* 2008, **9**:321-345.
91. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC: **Linkage of early-onset familial breast cancer to chromosome 17q21**. *Science* 1990, **250**(4988):1684-1689.
92. Lalloo F, Evans DG: **Familial Breast Cancer**. *Clin Genet* 2012, **82**(2):105-114.
93. Narod SA, Foulkes WD: **BRCA1 and BRCA2: 1994 and beyond**. *Nat Rev Cancer* 2004, **4**(9):665-676.
94. Lakhani SR, Jacquemier J, Sloane JP, Gusterson BA, Anderson TJ, van de Vijver MJ, Farid LM, Venter D, Antoniou A, Storfer-Isser A *et al*: **Multifactorial analysis of differences between sporadic breast cancers and cancers involving BRCA1 and BRCA2 mutations**. *J Natl Cancer I* 1998, **90**(15):1138-1145.
95. Vargas AC, Reis JS, Lakhani SR: **Phenotype-Genotype Correlation in Familial Breast Cancer**. *J Mammary Gland Biol* 2011, **16**(1):27-40.
96. Chen SN, Parmigiani G: **Meta-analysis of BRCA1 and BRCA2 penetrance**. *J Clin Oncol* 2007, **25**(11):1329-1333.

97. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G: **Identification of the breast cancer susceptibility gene BRCA2**. *Nature* 1995, **378**(6559):789-792.
98. Thorslund T, Esashi F, West SC: **Interactions between human BRCA2 protein and the meiosis-specific recombinase DMC1**. *Embo J* 2007, **26**(12):2915-2922.
99. Thompson D, Easton D, Consortium BCL: **Variation in cancer risks, by mutation position, in BRCA2 mutation carriers**. *American journal of human genetics* 2001, **68**(2):410-419.
100. Larsen MJ, Thomassen M, Gerdes AM, Kruse TA: **Hereditary breast cancer: clinical, pathological and molecular characteristics**. *Breast cancer : basic and clinical research* 2014, **8**:145-155.
101. Vogelstein B, Lane D, Levine AJ: **Surfing the p53 network**. *Nature* 2000, **408**(6810):307-310.
102. Li FP, Fraumeni JF, Jr.: **Soft-tissue sarcomas, breast cancer, and other neoplasms. A familial syndrome?** *Annals of internal medicine* 1969, **71**(4):747-752.
103. Rath MG, Masciari S, Gelman R, Miron A, Miron P, Foley K, Richardson AL, Krop IE, Verselis SJ, Dillon DA *et al*: **Prevalence of germline TP53 mutations in HER2+breast cancer patients**. *Breast Cancer Res Tr* 2013, **139**(1):193-198.
104. Castera L, Krieger S, Rousselin A, Legros A, Baumann JJ, Bruet O, Brault B, Fouillet R, Goardon N, Letac O *et al*: **Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes**. *Eur J Hum Genet* 2014, **22**(11):1305-1313.
105. Ko LJ, Prives C: **p53: Puzzle and paradigm**. *Gene Dev* 1996, **10**(9):1054-1072.
106. Sigal A, Rotter V: **Oncogenic mutations of the p53 tumor suppressor: The demons of the guardian of the genome**. *Cancer Res* 2000, **60**(24):6788-6793.
107. Holstege H, Joosse SA, van Oostrom CTM, Nederlof PM, de Vries A, Jonkers J: **High Incidence of Protein-Truncating TP53 Mutations in BRCA1-Related Breast Cancer**. *Cancer Res* 2009, **69**(8):3625-3633.
108. Guilford P, Hopkins J, Harraway J, McLeod M, McLeod N, Harawira P, Taite H, Scoular R, Miller A, Reeve AE: **E-cadherin germline mutations in familial gastric cancer**. *Nature* 1998, **392**(6674):402-405.
109. Brooks-Wilson AR, Kaurah P, Suriano G, Leach S, Senz J, Grehan N, Butterfield YS, Jeyes J, Schinas J, Bacani J *et al*: **Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria**. *Journal of medical genetics* 2004, **41**(7):508-517.
110. Kaurah P, MacMillan A, Boyd N, Senz J, De Luca A, Chun N, Suriano G, Zaor S, Van Manen L, Gilpin C *et al*: **Founder and recurrent CDH1 mutations in families with hereditary diffuse gastric cancer**. *JAMA : the journal of the American Medical Association* 2007, **297**(21):2360-2372.
111. Pharoah PD, Guilford P, Caldas C, International Gastric Cancer Linkage C: **Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families**. *Gastroenterology* 2001, **121**(6):1348-1353.
112. Suriano G, Yew S, Ferreira P, Senz J, Kaurah P, Ford JM, Longacre TA, Norton JA, Chun N, Young S *et al*: **Characterization of a recurrent germ line mutation of the E-cadherin gene: implications for genetic testing and clinical management**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2005, **11**(15):5401-5409.
113. Masciari S, Larsson N, Senz J, Boyd N, Kaurah P, Kandel MJ, Harris LN, Pinheiro HC, Troussard A, Miron P *et al*: **Germline E-cadherin mutations in familial lobular breast cancer**. *Journal of medical genetics* 2007, **44**(11):726-731.
114. Schrader KA, Masciari S, Boyd N, Salamanca C, Senz J, Saunders DN, Yorida E, Maines-Bandiera S, Kaurah P, Tung N *et al*: **Germline mutations in CDH1 are infrequent in women with early-onset or familial lobular breast cancers**. *Journal of medical genetics* 2011, **48**(1):64-68.
115. Xie ZM, Li LS, Laquet C, Penault-Llorca F, Uhrhammer N, Xie XM, Bignon YJ: **Germline Mutations of the E-Cadherin Gene in Families With Inherited Invasive Lobular Breast Carcinoma But No Diffuse Gastric Cancer**. *Cancer* 2011, **117**(14):3112-3117.
116. Rahman N, Stone JG, Coleman G, Gusterson B, Seal S, Marossy A, Lakhani SR, Ward A, Nash A, McKinna A *et al*: **Lobular carcinoma in situ of the breast is not caused by constitutional mutations in the E-cadherin gene**. *British journal of cancer* 2000, **82**(3):568-570.

117. Antoniou AC, Easton DF: **Models of genetic susceptibility to breast cancer**. *Oncogene* 2006, **25**(43):5898-5905.

118. Wu X, Webster SR, Chen J: **Characterization of tumor-associated Chk2 mutations**. *The Journal of biological chemistry* 2001, **276**(4):2971-2974.

119. Bell DW, Varley JM, Szydlo TE, Kang DH, Wahrer DC, Shannon KE, Lubratovich M, Verselis SJ, Isselbacher KJ, Fraumeni JF *et al*: **Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome**. *Science* 1999, **286**(5449):2528-2531.

120. Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M *et al*: **Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations**. *Nature genetics* 2002, **31**(1):55-59.

121. Consortium CBCC-C: **CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies**. *American journal of human genetics* 2004, **74**(6):1175-1182.

122. Le Calvez-Kelm F, Lesueur F, Damiola F, Vallee M, Voegele C, Babikyan D, Durand G, Forey N, McKay-Chopin S, Robinot N *et al*: **Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study**. *Breast cancer research : BCR* 2011, **13**(1):R6.

123. Liu C, Wang Y, Wang QS, Wang YJ: **The CHEK2 I157T variant and breast cancer susceptibility: a systematic review and meta-analysis**. *Asian Pacific journal of cancer prevention : APJCP* 2012, **13**(4):1355-1360.

124. Xia B, Sheng Q, Nakanishi K, Ohashi A, Wu J, Christ N, Liu X, Jasin M, Couch FJ, Livingston DM: **Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2**. *Molecular cell* 2006, **22**(6):719-729.

125. Zhang F, Ma JL, Wu JX, Ye L, Cai H, Xia B, Yu XC: **PALB2 Links BRCA1 and BRCA2 in the DNA-Damage Response**. *Curr Biol* 2009, **19**(6):524-529.

126. Sy SMH, Huen MSY, Chen JJ: **PALB2 is an integral component of the BRCA complex required for homologous recombination repair**. *P Natl Acad Sci USA* 2009, **106**(17):7155-7160.

127. Xia B, Dorsman JC, Ameziane N, de Vries Y, Rooimans MA, Sheng Q, Pals G, Errami A, Gluckman E, Llera J *et al*: **Fanconi anemia is associated with a defect in the BRCA2 partner PALB2**. *Nature genetics* 2007, **39**(2):159-161.

128. Reid S, Schindler D, Hanenberg H, Barker K, Hanks S, Kalb R, Neveling K, Kelly P, Seal S, Freund M *et al*: **Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer**. *Nature genetics* 2007, **39**(2):162-164.

129. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T *et al*: **PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene**. *Nature genetics* 2007, **39**(2):165-167.

130. Erkko H, Dowty JG, Nikkila J, Syrjakoski K, Mannermaa A, Pylkas K, Southey MC, Holli K, Kallioniemi A, Jukkola-Vuorinen A *et al*: **Penetrance analysis of the PALB2 c.1592delT founder mutation**. *Clin Cancer Res* 2008, **14**(14):4667-4671.

131. Garcia MJ, Fernandez V, Osorio A, Barroso A, Llort G, Lazaro C, Blanco I, Caldes T, de la Hoya M, Ramon YCT *et al*: **Analysis of FANCB and FANCN/PALB2 fanconi anemia genes in BRCA1/2-negative Spanish breast cancer families**. *Breast Cancer Res Treat* 2009, **113**(3):545-551.

132. Erkko H, Xia B, Nikkila J, Schleutker J, Syrjakoski K, Mannermaa A, Kallioniemi A, Pylkas K, Karppinen SM, Rapakko K *et al*: **A recurrent mutation in PALB2 in Finnish cancer families**. *Nature* 2007, **446**(7133):316-319.

133. Foulkes WD, Ghadirian P, Akbari MR, Hamel N, Giroux S, Sabbaghian N, Darnel A, Royer R, Poll A, Fafard E *et al*: **Identification of a novel truncating PALB2 mutation and analysis of its contribution to early-onset breast cancer in French-Canadian women**. *Breast Cancer Research* 2007, **9**(6).

134. Cantor SB, Bell DW, Ganesan S, Kass EM, Drapkin R, Grossman S, Wahrer DCR, Sgroi DC, Lane WS, Haber DA *et al*: **BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function**. *Cell* 2001, **105**(1):149-160.

135. Levran O, Attwooll C, Henry RT, Milton KL, Neveling K, Rio P, Batish SD, Kalb R, Velleuer E, Barral S *et al*: **The BRCA1-interacting helicase BRIP1 is deficient in Fanconi anemia**. *Nature genetics* 2005, **37**(9):931-933.

136. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K *et al*: **Truncating mutations in the Fanconi anemia J gene**

BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nature genetics* 2006, **38**(11):1239-1241.

137. Easton DF, Lesueur F, Decker B, Michailidou K, Li J, Allen J, Luccarini C, Pooley KA, Shah M, Bolla MK *et al*: **No evidence that protein truncating variants in BRIP1 are associated with breast cancer risk: implications for gene panel testing**. *Journal of medical genetics* 2016, **53**(5):298-309.

138. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, Puc J, Miliaresis C, Rodgers L, McCombie R *et al*: **PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer**. *Science* 1997, **275**(5308):1943-1947.

139. Steck PA, Pershouse MA, Jasser SA, Yung WKA, Lin H, Ligon AH, Langford LA, Baumgard ML, Hattier T, Davis T *et al*: **Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23.3 that is mutated in multiple advanced cancers**. *Nature genetics* 1997, **15**(4):356-362.

140. Nelen MR, Padberg GW, Peeters EAJ, Lin AY, vandenHelm B, Frants RR, Coulon V, Goldstein AM, vanReen MMM, Easton DF *et al*: **Localization of the gene for Cowden disease to chromosome 10q22-23**. *Nature genetics* 1996, **13**(1):114-116.

141. Easton DF, Pharoah PDP, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M *et al*: **Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk**. *New Engl J Med* 2015, **372**(23):2243-2257.

142. Marabelli M, Cheng SC, Parmigiani G: **Penetrance of ATM Gene Mutations in Breast Cancer: A Meta-Analysis of Different Measures of Risk**. *Genetic epidemiology* 2016.

143. Swift M, Reitnauer PJ, Morrell D, Chase CL: **Breast and Other Cancers in Families with Ataxia-Telangiectasia**. *New Engl J Med* 1987, **316**(21):1289-1294.

144. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K *et al*: **ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles**. *Nature genetics* 2006, **38**(8):873-875.

145. Tavtigian SV, Oefner PJ, Babikyan D, Hartmann A, Healey S, Le Calvez-Kelm F, Lesueur F, Byrnes GB, Chuang SC, Forey N *et al*: **Rare, Evolutionarily Unlikely Missense Substitutions in ATM Confer Increased Risk of Breast Cancer**. *American journal of human genetics* 2009, **85**(4):427-446.

146. Thompson D, Duedal S, Kirner JFR, McGuffog L, Last J, Reiman A, Byrd P, Taylor M, Easton DF: **Cancer risks and mortality in heterozygous ATM mutation carriers**. *J Natl Cancer I* 2005, **97**(11):813-822.

147. Ahmed M, Rahman N: **ATM and breast cancer susceptibility**. *Oncogene* 2006, **25**(43):5906-5911.

148. Shen Z, Wen XF, Lan F, Shen ZZ, Shao ZM: **The tumor suppressor gene LKB1 is associated with prognosis in human breast carcinoma**. *Clin Cancer Res* 2002, **8**(7):2085-2090.

149. Hearle N, Schumacher V, Menko FH, Olschwang S, Boardman LA, Gille JJ, Keller JJ, Westerman AM, Scott RJ, Lim W *et al*: **Frequency and spectrum of cancers in the Peutz-Jeghers syndrome**. *Clin Cancer Res* 2006, **12**(10):3209-3215.

150. Giardiello FM, Brensinger JD, Tersmette AC, Goodman SN, Petersen GM, Booker SV, Cruz-Correa M, Offerhaus JA: **Very high risk of cancer in familial Peutz-Jeghers syndrome**. *Gastroenterology* 2000, **119**(6):1447-1453.

151. Lim W, Olschwang S, Keller JJ, Westerman AM, Menko FH, Boardman LA, Scott RJ, Trimbath J, Giardiello FM, Gruber SB *et al*: **Relative frequency and morphology of cancers in STK11 mutation carriers**. *Gastroenterology* 2004, **126**(7):1788-1794.

152. Chen JD, Lindblom A: **Germline mutation screening of the STK11/LKB1 gene in familial breast cancer with LOH on 19p**. *Clin Genet* 2000, **57**(5):394-397.

153. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci**. *Nature* 2007, **447**(7148):1087-1093.

154. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK *et al*: **Large-scale genotyping identifies 41 new loci associated with breast cancer risk**. *Nature genetics* 2013, **45**(4):353-361, 361e351-352.

155. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, Maranian MJ, Bolla MK, Wang Q, Shah M *et al*: **Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer**. *Nature genetics* 2015, **47**(4):373-U127.

156. Couch FJ, Kuchenbaecker KB, Michailidou K, Mendoza-Fandino GA, Nord S, Lilyquist J, Olswold C, Hallberg E, Agata S, Ahsan H *et al*: **Identification of four novel**

susceptibility loci for oestrogen receptor negative breast cancer. *Nature communications* 2016, **7**:11375.

157. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, Orr N, Rhie SK, Riboli E, Feigelson HS *et al*: **Genome-wide association studies identify four ER negative-specific breast cancer risk loci**. *Nature genetics* 2013, **45**(4):392-398, 398e391-392.

158. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang ZM, Welch R, Hutchinson A *et al*: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer**. *Nature genetics* 2007, **39**(7):870-874.

159. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A *et al*: **Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer**. *Nature genetics* 2007, **39**(7):865-869.

160. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK *et al*: **Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer**. *Nature genetics* 2008, **40**(6):703-706.

161. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R *et al*: **Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2**. *Nature genetics* 2009, **41**(5):585-590.

162. Zheng W, Long JR, Gao YT, Li C, Zheng Y, Xiang YB, Wen WQ, Levy S, Deming SL, Haines JL *et al*: **Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1**. *Nature genetics* 2009, **41**(3):324-328.

163. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K *et al*: **A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1)**. *Nature genetics* 2009, **41**(5):579-584.

164. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS *et al*: **Genome-wide association study identifies five new breast cancer susceptibility loci**. *Nature genetics* 2010, **42**(6):504-U547.

165. Antoniou AC, Wang XS, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, Healey S, Morrison J, Kartsonaki C, Lesnick T *et al*: **A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population**. *Nature genetics* 2010, **42**(10):885-+.

166. Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, Zelenika D, Gut I, Heath S, Palles C *et al*: **Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study**. *J Natl Cancer I* 2011, **103**(5):425-435.

167. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, Wang XS, Ademuyiwa F, Ahmed S, Ambrosone CB *et al*: **A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer**. *Nature genetics* 2011, **43**(12):1210-U1261.

168. Siddiq A, Couch FJ, Chen GK, Lindstrom S, Eccles D, Millikan RC, Michailidou K, Stram DO, Beckmann L, Rhie SK *et al*: **A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11**. *Hum Mol Genet* 2012, **21**(24):5373-5384.

169. Cai QY, Long JR, Lu W, Qu SMA, Wen WQ, Kang D, Lee JY, Chen KX, Shen HB, Shen CY *et al*: **Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium**. *Hum Mol Genet* 2011, **20**(24):4991-4999.

170. Long J, Cai Q, Sung H, Shi J, Zhang B, Choi JY, Wen W, Delahanty RJ, Lu W, Gao YT *et al*: **Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer**. *PLoS genetics* 2012, **8**(2):e1002532.

171. Kim HC, Lee JY, Sung H, Choi JY, Park SK, Lee KM, Kim YJ, Go MJ, Li L, Cho YS *et al*: **A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study**. *Breast Cancer Research* 2012, **14**(2).

172. Couch FJ, Wang XS, McGuffog L, Lee A, Olswold C, Kuchenbaecker KB, Soucy P, Fredericksen Z, Barrowdale D, Dennis J *et al*: **Genome-Wide Association Study in BRCA1 Mutation Carriers Identifies Novel Loci Associated with Breast and Ovarian Cancer Risk**. *PLoS genetics* 2013, **9**(3).

173. Sawyer E, Roylance R, Petridis C, Brook MN, Nowinski S, Papouli E, Fletcher O, Pinder S, Hanby A, Kohut K *et al*: **Genetic Predisposition to In Situ and Invasive Lobular Carcinoma of the Breast**. *PLoS genetics* 2014, **10**(4).

174. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, Dennis J, Wang Q, Humphreys MK, Luccarini C *et al*: **Genome-wide association analysis identifies three new breast cancer susceptibility loci**. *Nature genetics* 2012, **44**(3):312-318.

175. Pharoah P, Consortium BCA: **Commonly studied single-nucleotide polymorphisms and breast cancer: Results from the Breast Cancer Association Consortium**. *Jnci-J Natl Cancer I* 2006, **98**(19):1382-1396.

176. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MWR, Pooley KA, Scollen S, Baynes C, Ponder BAJ, Chanock S *et al*: **A common coding variant in CASP8 is associated with breast cancer risk**. *Nature genetics* 2007, **39**(3):352-358.

177. Cai QY, Zhang B, Sung H, Low SK, Kweon SS, Lu W, Shi JJ, Long JR, Wen WQ, Choi JY *et al*: **Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1**. *Nature genetics* 2014, **46**(8):886-890.

178. Milne RL, Burwinkel B, Michailidou K, Arias-Perez JI, Zamora MP, Menendez-Rodriguez P, Hardisson D, Mendiola M, Gonzalez-Neira A, Pita G *et al*: **Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium**. *Hum Mol Genet* 2014, **23**(22):6096-6111.

179. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, Kozarewa L *et al*: **Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C**. *Genome research* 2014, **24**(11):1854-1868.

180. Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, Edwards SL, Pickett HA, Shen HC, Smart CE *et al*: **Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer**. *Nature genetics* 2013, **45**(4):371-384.

181. Meyer KB, OReilly M, Michailidou K, Carlebur S, Edwards SL, French JD, Prathalingham R, Dennis J, Bolla MK, Wang Q *et al*: **Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1**. *American journal of human genetics* 2013, **93**(6):1046-1060.

182. French JD, Ghoussaini M, Edwards SL, Meyer KB, Michailidou K, Ahmed S, Khan S, Maranian MJ, OReilly M, Hillman KM *et al*: **Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers**. *American journal of human genetics* 2013, **92**(4):489-503.

183. Ghoussaini M, Edwards SL, Michailidou K, Nord S, Lari RCS, Desai K, Kar S, Hillman KM, Kaufmann S, Glubb DM *et al*: **Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation**. *Nature communications* 2014, **5**.

184. Hein R, Maranian M, Hopper JL, Kapuscinski MK, Southey MC, Park DJ, Schmidt MK, Broeks A, Hogervorst FBL, Bueno-de-Mesquit HB *et al*: **Comparison of 6q25 Breast Cancer Hits from Asian and European Genome Wide Association Studies in the Breast Cancer Association Consortium (BCAC)**. *PloS one* 2012, **7**(8).

185. Lindstrom S, Thompson DJ, Paterson AD, Li JM, Gierach GL, Scott C, Stone J, Douglas JA, Dos-Santos-Silva I, Fernandez-Navarro P *et al*: **Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk**. *Nature communications* 2014, **5**.

186. OBrien KM, Cole SR, Engel LS, Bensen JT, Poole C, Herring AH, Millikan RC: **Breast Cancer Subtypes and Previously Established Genetic Risk Factors: A Bayesian Approach**. *Cancer Epidem Biomar* 2014, **23**(1):84-97.

187. Han W, Woo JH, Yu JH, Lee MJ, Moon HG, Kang D, Noh DY: **Common Genetic Variants Associated with Breast Cancer in Korean Women and Differential Susceptibility According to Intrinsic Subtype**. *Cancer Epidem Biomar* 2011, **20**(5):793-798.

188. Visscher PM, Hill WG, Wray NR: **Heritability in the genomics era--concepts and misconceptions**. *Nature reviews Genetics* 2008, **9**(4):255-266.

189. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A *et al*: **Finding the missing heritability of complex diseases**. *Nature* 2009, **461**(7265):747-753.

190. Li Q, Seo JH, Stranger B, McKenna A, Peer I, Laframboise T, Brown M, Tyekucheva S, Freedman ML: **Integrative eQTL-based analyses reveal the biology of breast cancer risk loci**. *Cell* 2013, **152**(3):633-641.

191. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW *et al*: **Common SNPs explain a large proportion of the heritability for human height**. *Nature genetics* 2010, **42**(7):565-569.

192. Hu VW, Addington A, Hyman A: **Novel autism subtype-dependent genetic variants are revealed by quantitative trait and subphenotype association analyses of published GWAS data**. *PloS one* 2011, **6**(4):e19067.

193. Kar SP, Beesley J, Al Olama AA, Michailidou K, Tyrer J, Kote-Jarai ZA, Lawrenson K, Lindstrom S, Ramus SJ, Thompson DJ *et al*: **Genome-Wide Meta-Analyses of Breast, Ovarian, and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by at Least Two Cancer Types**. *Cancer Discov* 2016, **6**(9):1052-1067.

194. Cannonalbright LA, Thomas A, Goldgar DE, Gholami K, Rowe K, Jacobsen M, Mcwhorter WP, Skolnick MH: **Familiality of Cancer in Utah**. *Cancer Res* 1994, **54**(9):2378-2385.

195. Claus EB, Risch NJ, Thompson WD: **Using Age of Onset to Distinguish between Subforms of Breast-Cancer**. *Ann Hum Genet* 1990, **54**:169-177.

196. Droufakou S, Deshmane V, Roylance R, Hanby A, Tomlinson I, Hart IR: **Multiple ways of silencing E-cadherin gene expression in lobular carcinoma of the breast**. *International journal of cancer Journal international du cancer* 2001, **92**(3):404-408.

197. Salahshor S, Lei HX, Huo HG, Kristensen VN, Loman N, Sjoberg-Margolin S, Borg A, Borresen-Dale AL, Vorechovsky I, Lindblom A: **Low frequency of E-cadherin alterations in familial breast cancer**. *Breast Cancer Research* 2001, **3**(3):199-207.

198. Mavaddat N, Barrowdale D, Andrulis IL, Domchek SM, Eccles D, Nevanlinna H, Ramus SJ, Spurdle A, Robson M, Sherman M *et al*: **Pathology of Breast and Ovarian Cancers among BRCA1 and BRCA2 Mutation Carriers: Results from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA)**. *Cancer Epidem Biomar* 2012, **21**(1):134-147.

199. Cybulski C, Gorski B, Huzarski T, Byrski T, Gronwald J, Debniak T, Wokolorczyk D, Jakubowska A, Kowalska E, Oszurek O *et al*: **CHEK2-positive breast cancers in young Polish women**. *Clin Cancer Res* 2006, **12**(16):4832-4835.

200. Teo ZL, Park DJ, Provenzano E, Chatfield CA, Odefrey FA, Tu ND, Dowty JG, Hopper JL, Winship I, Goldgar DE *et al*: **Prevalence of PALB2 mutations in Australasian multiple-case breast cancer families**. *Breast Cancer Research* 2013, **15**(1).

201. van der Groep P, van der Wall E, van Diest PJ: **Pathology of hereditary breast cancer**. *Cellular oncology* 2011, **34**(2):71-88.

202. Cancer Genome Atlas N: **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**(7407):330-337.

203. Barber M, Murrell A, Ito Y, Maia AT, Hyland S, Oliveira C, Save V, Carneiro F, Paterson AL, Grehan N *et al*: **Mechanisms and sequelae of E-cadherin silencing in hereditary diffuse gastric cancer**. *The Journal of pathology* 2008, **216**(3):295-306.

204. Benusiglio PR, Malka D, Rouleau E, De Pauw A, Buecher B, Nogues C, Fourme E, Colas C, Coulet F, Warcoin M *et al*: **CDH1 germline mutations and the hereditary diffuse gastric and lobular breast cancer syndrome: a multicentre study**. *Journal of medical genetics* 2013, **50**(7):486-489.

205. Salahshor S, Hou H, Diep CB, Loukola A, Zhang H, Liu T, Chen J, Iselius L, Rubio C, Lothe RA *et al*: **A germline E-cadherin mutation in a family with gastric and colon cancer**. *International journal of molecular medicine* 2001, **8**(4):439-443.

206. Salahshor S, Haixin L, Huo H, Kristensen VN, Loman N, Sjoberg-Margolin S, Borg A, Borresen-Dale AL, Vorechovsky I, Lindblom A: **Low frequency of E-cadherin alterations in familial breast cancer**. *Breast cancer research : BCR* 2001, **3**(3):199-207.

207. Valente AL, Rummel S, Shriver CD, Ellsworth RE: **Sequence-based detection of mutations in cadherin 1 to determine the prevalence of germline mutations in patients with invasive lobular carcinoma of the breast**. *Hereditary cancer in clinical practice* 2014, **12**(1):17.

208. de Sanjose S, Leone M, Berez V, Izquierdo A, Font R, Brunet JM, Louat T, Vilardell L, Borras J, Viladiu P *et al*: **Prevalence of BRCA1 and BRCA2 germline mutations in young breast cancer patients: a population-based study**. *International journal of cancer Journal international du cancer* 2003, **106**(4):588-593.

209. Bogdanova N, Cybulski C, Bermisheva M, Datsyuk I, Yamini P, Hillemanns P, Antonenkova NN, Khusnutdinova E, Lubinski J, Dork T: **A nonsense mutation (E1978X) in the ATM gene is associated with breast cancer**. *Breast cancer research and treatment* 2009, **118**(1):207-211.
210. Yager JD, Davidson NE: **Mechanisms of disease: Estrogen carcinogenesis in breast cancer**. *New Engl J Med* 2006, **354**(3):270-282.
211. Sun YT, Zhang JS, Ma L: **alpha-catenin A tumor suppressor beyond adherens junctions**. *Cell Cycle* 2014, **13**(15):2334-2339.
212. Hansford S, Kaurah P, Li-Chang H, Woo M, Senz J, Pinheiro H, Schrader KA, Schaeffer DF, Shumansky K, Zogopoulos G *et al*: **Hereditary Diffuse Gastric Cancer Syndrome CDH1 Mutations and Beyond**. *Jama Oncol* 2015, **1**(1):23-32.
213. Solyom S, Aressy B, Pylkas K, Patterson-Fortin J, Hartikainen JM, Kallioniemi A, Kauppila S, Nikkila J, Kosma VM, Mannermaa A *et al*: **Breast cancer-associated Abraxas mutation disrupts nuclear localization and DNA damage response functions**. *Science translational medicine* 2012, **4**(122):122ra123.
214. Pennington KP, Walsh T, Harrell MI, Lee MK, Pennil CC, Rendi MH, Thornton A, Norquist BM, Casadei S, Nord AS *et al*: **Germline and somatic mutations in homologous recombination genes predict platinum response and survival in ovarian, fallopian tube, and peritoneal carcinomas**. *Clin Cancer Res* 2014, **20**(3):764-775.
215. Richards FM, McKee SA, Rajpar MH, Cole TRP, Evans DGR, Jankowski JA, McKeown C, Sanders DSA, Maher ER: **Germline E-cadherin gene (CDH1) mutations predispose to familial gastric cancer and colorectal cancer**. *Hum Mol Genet* 1999, **8**(4):607-610.
216. Shirts BH, Casadei S, Jacobson AL, Lee MK, Gulsuner S, Bennett RL, Miller M, Hall SA, Hampel H, Hisama FM *et al*: **Improving performance of multigene panels for genomic analysis of cancer predisposition**. *Genetics in medicine : official journal of the American College of Medical Genetics* 2016, **18**(10):974-981.
217. Wang HD, Ren J, Zhang L: **CDH1 germline mutation in hereditary gastric carcinoma**. *World journal of gastroenterology* 2004, **10**(21):3088-3093.
218. Lindor NM, Guidugli L, Wang X, Vallee MP, Monteiro AN, Tavtigian S, Goldgar DE, Couch FJ: **A review of a multifactorial probability-based model for classification of BRCA1 and BRCA2 variants of uncertain significance (VUS)**. *Human mutation* 2012, **33**(1):8-21.
219. Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, Tavtigian SV, Monteiro AN, Iversen ES, Couch FJ *et al*: **A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes**. *American journal of human genetics* 2007, **81**(5):873-883.
220. Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, Couch FJ, Breast Cancer Information Core Steering C: **Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2**. *American journal of human genetics* 2004, **75**(4):535-544.
221. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV *et al*: **Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results**. *Human mutation* 2008, **29**(11):1282-1291.
222. Domagala P, Wokolorczyk D, Cybulski C, Huzarski T, Lubinski J, Domagala W: **Different CHEK2 germline mutations are associated with distinct immunophenotypic molecular subtypes of breast cancer**. *Breast Cancer Res Tr* 2012, **132**(3):937-945.
223. Cybulski C, Wokolorczyk D, Jakubowska A, Huzarski T, Byrski T, Gronwald J, Masojc B, Debniak T, Gorski B, Blecharz P *et al*: **Risk of Breast Cancer in Women With a CHEK2 Mutation With and Without a Family History of Breast Cancer**. *J Clin Oncol* 2011, **29**(28):3747-3752.
224. Cybulski C, Kluzniak W, Huzarski T, Wokolorczyk D, Kashyap A, Jakubowska A, Szwiec M, Byrski T, Debniak T, Gorski B *et al*: **Clinical outcomes in women with breast cancer and a PALB2 mutation: a prospective cohort analysis**. *Lancet Oncology* 2015, **16**(6):638-644.
225. Teo ZL, Provenzano E, Dite GS, Park DJ, Apicella C, Sawyer SD, James PA, Mitchell G, Trainer AH, Lindeman GJ *et al*: **Tumour morphology predicts PALB2 germline mutation status**. *British journal of cancer* 2013, **109**(1):154-163.

226. Hemphill AW, Bruun D, Thrun L, Akkari Y, Torimaru Y, Hejna K, Jakobs PM, Hejna J, Jones S, Olson SB *et al*: **Mammalian SNM1 is required for genome stability**. *Mol Genet Metab* 2008, **94**(1):38-45.

227. Yu KD, Rao NY, Chen AX, Fan L, Yang C, Shao ZM: **A systematic review of the relationship between polymorphic sites in the estrogen receptor-beta (ESR2) gene and breast cancer risk**. *Breast Cancer Res Tr* 2011, **126**(1):37-45.

228. Haldosen LA, Zhao CY, Dahlman-Wright K: **Estrogen receptor beta in breast cancer**. *Mol Cell Endocrinol* 2014, **382**(1):665-672.

229. Figueroa JD, Garcia-Closas M, Humphreys M, Platte R, Hopper JL, Southey MC, Apicella C, Hammet F, Schmidt MK, Broeks A *et al*: **Associations of common variants at 1p11.2 and 14q24.1 (RAD51L1) with breast cancer risk and heterogeneity by tumor subtype: findings from the Breast Cancer Association Consortium**. *Hum Mol Genet* 2011, **20**(23):4693-4706.

230. Horne HN, Chung CC, Zhang H, Yu K, Prokunina-Olsson L, Michailidou K, Bolla MK, Wang Q, Dennis J, Hopper JL *et al*: **Fine-Mapping of the 1p11.2 Breast Cancer Susceptibility Locus**. *PloS one* 2016, **11**(8).

231. Key TJ, Verkasalo PK, Banks E: **Epidemiology of breast cancer**. *The Lancet Oncology* 2001, **2**(3):133-140.

232. Phipps AI, Buist DS, Malone KE, Barlow WE, Porter PL, Kerlikowske K, Li CI: **Reproductive history and risk of three breast cancer subtypes defined by three biomarkers**. *Cancer causes & control : CCC* 2011, **22**(3):399-405.

233. Reeves GK, Beral V, Green J, Gathani T, Bull D: **Hormonal therapy for menopause and breast-cancer risk by histological type: a cohort study and meta-analysis**. *Lancet Oncology* 2006, **7**(11):910-918.

234. Fiesch-Janys D, Slanger T, Mutschelknauss E, Kropp S, Obi N, Vettorazzi E, Braendle W, Bastert G, Hentschel S, Berger J *et al*: **Risk of different histological types of postmenopausal breast cancer by type and regimen of menopausal hormone therapy**. *International journal of cancer* 2008, **123**(4):933-941.

235. Claus EB, Stowe M, Carter D: **Breast carcinoma in situ: risk factors and screening patterns**. *Journal of the National Cancer Institute* 2001, **93**(23):1811-1817.

236. Trentham-Dietz A, Newcomb PA, Storer BE, Remington PL: **Risk factors for carcinoma in situ of the breast**. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2000, **9**(7):697-703.

237. Reeves GK, Pirie K, Green J, Bull D, Beral V, Million Women Study C: **Comparison of the effects of genetic and environmental risk factors on in situ and invasive ductal breast cancer**. *International journal of cancer* 2012, **131**(4):930-937.

238. Wohlfahrt J, Mouridsen H, Andersen PK, Melbye M: **Reproductive risk factors for breast cancer by receptor status, histology, laterality and location**. *International journal of cancer* 1999, **81**(1):49-55.

239. Swann R, Perkins KA, Velentzis LS, Ciria C, Dutton SJ, Mulligan AA, Woodside JV, Cantwell MM, Leathem AJ, Robertson CE *et al*: **The DietCompLyf study: A prospective cohort study of breast cancer survival and phytoestrogen consumption**. *Maturitas* 2013, **75**(3):232-240.

240. Prentice RL, Huang Y, Hinds DA, Peters U, Pettinger M, Cox DR, Beilharz E, Chlebowski RT, Rossouw JE, Caan B *et al*: **Variation in the FGFR2 Gene and the Effects of Postmenopausal Hormone Therapy on Invasive Breast Cancer**. *Cancer Epidem Biomar* 2009, **18**(11):3079-3085.

241. Rebbeck TR, DeMichele A, Tran TV, Panossian S, Bunin GR, Troxel AB, Strom BL: **Hormone-dependent effects of FGFR2 and MAP3K1 in breast cancer susceptibility in a population-based sample of post-menopausal African-American and European-American women**. *Carcinogenesis* 2009, **30**(2):269-274.

242. Lee E, Schumacher F, Lewinger JP, Neuhausen SL, Anton-Culver H, Horn-Ross PL, Henderson KD, Ziogas A, Van den Berg D, Bernstein L *et al*: **The association of polymorphisms in hormone metabolism pathway genes, menopausal hormone therapy, and breast cancer risk: a nested case-control study in the California Teachers Study cohort**. *Breast Cancer Research* 2011, **13**(2).

243. Rudolph A, Chang-Claude J, Schmidt MK: **Gene-environment interaction and risk of breast cancer**. *British journal of cancer* 2016, **114**(2):125-133.

244. Travis RC, Reeves GK, Green J, Bull D, Tipper SJ, Baker K, Beral V, Peto R, Bell J, Zelenika D *et al*: **Gene-environment interactions in 7610 women with breast cancer: prospective evidence from the Million Women Study**. *Lancet* 2010, **375**(9732):2143-2151.

245. Hein R, Flesch-Janys D, Dahmen N, Beckmann L, Lindstrom S, Schoof N, Czene K, Mittelstrass K, Illig T, Seibold P *et al*: **A genome-wide association study to identify genetic susceptibility loci that modify ductal and lobular postmenopausal breast cancer risk associated with menopausal hormone therapy use: a two-stage design with replication**. *Breast Cancer Res Tr* 2013, **138**(2):529-542.

246. Rudolph A, Hein R, Lindstrom S, Beckmann L, Behrens S, Liu J, Aschard H, Bolla MK, Wang J, Truong T *et al*: **Genetic modifiers of menopausal hormone replacement therapy and breast cancer risk: a genome-wide interaction study**. *Endocr-Relat Cancer* 2013, **20**(6):875-887.

247. Osawa T, Muramatsu M, Wang F, Tsuchida R, Kodama T, Minami T, Shibuya M: **Increased expression of histone demethylase JHDM1D under nutrient starvation suppresses tumor growth via down-regulating angiogenesis**. *Proc Natl Acad Sci U S A* 2011, **108**(51):20725-20729.

248. Ibrahim SA, Yip GW, Stock C, Pan JW, Neubauer C, Poeter M, Pupjalis D, Koo CY, Kelsch R, Schule R *et al*: **Targeting of syndecan-1 by microRNA miR-10b promotes breast cancer cell motility and invasiveness via a Rho-GTPase- and E-cadherin-dependent mechanism**. *International journal of cancer* 2012, **131**(6):E884-896.

249. Pharoah PDP, Guilford P, Caldas C, Consortiu IGCL: **Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families**. *Gastroenterology* 2001, **121**(6):1348-1353.

250. Petridis C, Shinomiya I, Kohut K, Gorman P, Caneppele M, Shah V, Troy M, Pinder SE, Hanby A, Tomlinson I *et al*: **Germline CDH1 mutations in bilateral lobular carcinoma in situ**. *British journal of cancer* 2014, **110**(4):1053-1057.

251. Weigelt B, Horlings HM, Kreike B, Hayes MM, Hauptmann M, Wessels LF, de Jong D, Van de Vijver MJ, Vant Veer LJ, Peterse JL: **Refinement of breast cancer classification by molecular characterization of histological special types**. *The Journal of pathology* 2008, **216**(2):141-150.

252. Milne RL, Goode EL, Garca-Closas M, Couch FJ, Severi G, Hein R, Fredericksen Z, Malats N, Zamora MP, Perez JIA *et al*: **Confirmation of 5p12 As a Susceptibility Locus for Progesterone-Receptor-Positive, Lower Grade Breast Cancer**. *Cancer Epidem Biomar* 2011, **20**(10):2222-2231.

253. Claus EB, Stowe M, Carter D: **Family history of breast and ovarian cancer and the risk of breast carcinoma in situ**. *Breast Cancer Res Tr* 2003, **78**(1):7-15.

254. Kerlikowske K, Barclay J, Grady D, Sickles EA, Ernster V: **Comparison of risk factors for ductal carcinoma in situ and invasive breast cancer**. *Jnci-J Natl Cancer I* 1997, **89**(1):76-82.

255. Claus EB, Petruzella S, Matloff E, Carter D: **Prevalence of BRCA1 and BRCA2 mutations in women diagnosed with ductal carcinoma in situ**. *Jama* 2005, **293**(8):964-969.

256. Hall MJ, Reid JE, Wenstrup RJ: **Prevalence of BRCA1 and BRCA2 mutations in women with breast carcinoma In Situ and referred for genetic testing**. *Cancer prevention research* 2010, **3**(12):1579-1585.

257. Tie J, Wang YX, Tomasetti C, Li L, Springer S, Kinde I, Silliman N, Tacey M, Wong HL, Christie M *et al*: **Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer**. *Science translational medicine* 2016, **8**(346).

258. Tu ND, Hammet F, Mahmoodi M, Tsimiklis H, Teo ZL, Li R, Pope BJ, Terry MB, Buys SS, Daly M *et al*: **Mutation screening of PALB2 in clinically ascertained families from the Breast Cancer Family Registry**. *Breast Cancer Res Tr* 2015, **149**(2):547-554.

259. Dansonka-Mieszkowska A, Kluska A, Moes J, Dabrowska M, Nowakowska D, Niwinska A, Derlatka P, Cendrowski K, Kupryjanczyk J: **A novel germline PALB2 deletion in Polish breast and ovarian cancer patients**. *Bmc Med Genet* 2010, **11**.

260. Masciari S, Dillon DA, Rath M, Robson M, Weitzel JN, Balmana J, Gruber SB, Ford JM, Euhus D, Lebensohn A *et al*: **Breast cancer phenotype in women with TP53 germline mutations: a Li-Fraumeni syndrome consortium effort**. *Breast Cancer Res Treat* 2012, **133**(3):1125-1130.

261. Masciari S, Dillon D, Dick MG, Robson ME, Weitzel JN, Ford JM, Balmana J, Gruber SB, Euhus D, Garber JE: **Breast cancer phenotype in women with TP53 germ-line mutations: An LFS consortium effort.** *J Clin Oncol* 2011, **29**(15).

262. Schrader KA, Stratton KL, Murali R, Laitman Y, Cavallone L, Offit L, Wen YH, Thomas T, Shah S, Rau-Murthy R *et al*: **Genome Sequencing of Multiple Primary Tumors Reveals a Novel PALB2 Variant**. *J Clin Oncol* 2016, **34**(8):E61-E67.

263. Thompson ER, Gorringe KL, Rowley SM, Li N, McInerny S, Wong-Brown MW, Devereux L, Li J, Lifepool I, Trainer AH *et al*: **Reevaluation of the BRCA2 truncating allele c.9976A > T (p.Lys3326Ter) in a familial breast cancer context**. *Scientific reports* 2015, **5**:14800.

264. Campa D, Barrdahl M, Gaudet MM, Black A, Chanock SJ, Diver WR, Gapstur SM, Haiman C, Hankinson S, Hazra A *et al*: **Genetic risk variants associated with in situ breast cancer**. *Breast Cancer Research* 2015, **17**.

265. Purrington KS, Slager S, Eccles D, Yannoukakos D, Fasching PA, Miron P, Carpenter J, Chang-Claude J, Martin NG, Montgomery GW *et al*: **Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer**. *Carcinogenesis* 2014, **35**(5):1012-1019.

266. Narod S: **BREAST CANCER The importance of overdiagnosis in breast-cancer screening**. *Nature Reviews Clinical Oncology* 2016, **13**(1):4-5.

267. Willer CJ, Li Y, Abecasis GR: **METAL: fast and efficient meta-analysis of genomewide association scans**. *Bioinformatics* 2010, **26**(17):2190-2191.

268. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ: **LocusZoom: regional visualization of genome-wide association scan results**. *Bioinformatics* 2010, **26**(18):2336-2337.

269. Warren H, Dudbridge F, Fletcher O, Orr N, Johnson N, Hopper JL, Apicella C, Southey MC, Mahmoodi M, Schmidt MK *et al*: **9q31.2-rs865686 as a Susceptibility Locus for Estrogen Receptor-Positive Breast Cancer: Evidence from the Breast Cancer Association Consortium**. *Cancer Epidem Biomar* 2012, **21**(10):1783-1791.

270. Wapnir IL, Dignam JJ, Fisher B, Mamounas EP, Anderson SJ, Julian TB, Land SR, Margolese RG, Swain SM, Costantino JP *et al*: **Long-term outcomes of invasive ipsilateral breast tumor recurrences after lumpectomy in NSABP B-17 and B-24 randomized clinical trials for DCIS**. *Journal of the National Cancer Institute* 2011, **103**(6):478-488.

271. Thomas J, Hanby A, Pinder S, Ellis I, Macartney J, Clements K, Lawrence G, Bishop H, Sloane Project Steering G: **Implications of inconsistent measurement of ER status in non-invasive breast cancer: a study of 1,684 cases from the Sloane Project**. *The breast journal* 2008, **14**(1):33-38.

272. Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, Richesson DA, Bojesen SE, Nordestgaard BG, Axelsson CK, Arias JI *et al*: **Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics**. *PLoS genetics* 2008, **4**(4):e1000054.

273. Purrington KS, Slettedahl S, Bolla MK, Michailidou K, Czene K, Nevanlinna H, Bojesen SE, Andrulis IL, Cox A, Hall P *et al*: **Genetic variation in mitotic regulatory pathway genes is associated with breast tumor grade**. *Human molecular genetics* 2014, **23**(22):6034-6046.

274. Lambrechts D, Truong T, Justenhoven C, Humphreys MK, Wang J, Hopper JL, Dite GS, Apicella C, Southey MC, Schmidt MK *et al*: **11q13 is a susceptibility locus for hormone receptor positive breast cancer**. *Hum Mutat* 2012, **33**(7):1123-1132.

275. Arvold ND, Punglia RS, Hughes ME, Jiang W, Edge SB, Javid SH, Laronga C, Niland JC, Theriault RL, Weeks JC *et al*: **Pathologic characteristics of second breast cancers after breast conservation for ductal carcinoma in situ**. *Cancer* 2012, **118**(24):6022-6030.

276. Elsayegh N, Barrera AM, Muse KI, Lin H, Kuerer HM, Helm M, Litton JK, Arun BK: **Evaluation of BRCAPRO Risk Assessment Model in Patients with Ductal Carcinoma In situ Who Underwent Clinical BRCA Genetic Testing**. *Frontiers in genetics* 2016, **7**:71.

277. Petridis C, Brook MN, Shah V, Kohut K, Gorman P, Caneppele M, Levi D, Papouli E, Orr N, Cox A *et al*: **Genetic predisposition to ductal carcinoma in situ of the breast**. *Breast cancer research : BCR* 2016, **18**(1):22.

278. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, Wang Q, Dennis J, Dunning AM, Shah M *et al*: **Prediction of breast cancer risk based on profiling with common genetic variants**. *J Natl Cancer Inst* 2015, **107**(5).

279. Wellek S, Ziegler A: **Cochran-Armitage test versus logistic regression in the analysis of genetic association studies**. *Human heredity* 2012, **73**(1):14-17.

280. Purcell S, Cherny SS, Sham PC: **Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits**. *Bioinformatics* 2003, **19**(1):149-150.

281. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ *et al*: **PLINK: A tool set for whole-genome association**

and population-based linkage analyses. *American journal of human genetics* 2007, **81**(3):559-575.

282. Delaneau O, Marchini J, Zagury JF: **A linear complexity phasing method for thousands of genomes**. *Nat Methods* 2011, **9**(2):179-181.

283. Howie BN, Donnelly P, Marchini J: **A flexible and accurate genotype imputation method for the next generation of genome-wide association studies**. *PLoS genetics* 2009, **5**(6):e1000529.

284. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes**. *Nature genetics* 2007, **39**(7):906-913.

285. Kong Y: **Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies**. *Genomics* 2011, **98**(2):152-153.

286. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**(6):841-842.

287. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

288. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome research* 2010, **20**(9):1297-1303.

289. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M *et al*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nature genetics* 2011, **43**(5):491-+.

290. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al*: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**(15):2156-2158.

291. Wang K, Li MY, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic acids research* 2010, **38**(16).

292. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM: **Robust relationship inference in genome-wide association studies**. *Bioinformatics* 2010, **26**(22):2867-2873.

293. Ng PC, Henikoff S: **SIFT: predicting amino acid changes that affect protein function**. *Nucleic acids research* 2003, **31**(13):3812-3814.

294. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations**. *Nat Methods* 2010, **7**(4):248-249.

295. Kircher M, Witten DM, Jain P, ORoak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nature genetics* 2014, **46**(3):310-+.

296. Quang D, Chen YF, Xie XH: **DANN: a deep learning approach for annotating the pathogenicity of genetic variants**. *Bioinformatics* 2015, **31**(5):761-763.

297. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome research* 2002, **12**(6):996-1006.

298. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L *et al*: **Ensembl 2016**. *Nucleic acids research* 2016, **44**(D1):D710-716.

Appendix 1

Clinical information on cases downloaded from TCGA. AoD corresponds to age of diagnosis.

Sample	AoD	ER status	Menopause status	Sample	AoD	ER status	Menopause status
TCGA_BC0014	40	Positive	Premenopausal	TCGA_BC0061	62	Positive	Postmenopausal
TCGA_BC0103	44	Positive	Premenopausal	TCGA_BC0004	62	Positive	Postmenopausal
TCGA_BC0083	44	Positive	Postmenopausal	TCGA_BC0080	62	Positive	Postmenopausal
TCGA_BC0114	44	Positive	Unknown	TCGA_BC0036	62	Positive	Postmenopausal
TCGA_BC0095	44	Positive	Unknown	TCGA_BC0013	62	Positive	Postmenopausal
TCGA_BC0008	45	Positive	Premenopausal	TCGA_BC0062	62	Positive	Postmenopausal
TCGA_BC0040	46	Positive	Premenopausal	TCGA_BC0049	62	Positive	Unknown
TCGA_BC0031	46	Positive	Premenopausal	TCGA_BC0015	62	Positive	Postmenopausal
TCGA_BC0105	46	Positive	Premenopausal	TCGA_BC0100	63	Positive	Postmenopausal
TCGA_BC0022	46	Positive	Premenopausal	TCGA_BC0045	63	Positive	Postmenopausal
TCGA_BC0005	46	Positive	Premenopausal	TCGA_BC0041	63	Unknown	Postmenopausal
TCGA_BC0016	46	Positive	Premenopausal	TCGA_BC0021	63	Positive	Postmenopausal
TCGA_BC0025	46	Positive	Premenopausal	TCGA_BC0078	63	Negative	Postmenopausal
TCGA_BC0118	47	Positive	Premenopausal	TCGA_BC0099	63	Positive	Postmenopausal
TCGA_BC0039	48	Positive	Premenopausal	TCGA_BC0092	64	Positive	Postmenopausal
TCGA_BC0096	48	Positive	Postmenopausal	TCGA_BC0088	65	Positive	Postmenopausal
TCGA_BC0051	49	Positive	Premenopausal	TCGA_BC0018	65	Positive	Postmenopausal
TCGA_BC0050	49	Positive	Premenopausal	TCGA_BC0104	65	Positive	Postmenopausal
TCGA_BC0069	49	Unknown	Premenopausal	TCGA_BC0043	65	Positive	Postmenopausal
TCGA_BC0066	50	Positive	Premenopausal	TCGA_BC0024	66	Positive	Postmenopausal
TCGA_BC0109	50	Positive	Premenopausal	TCGA_BC0098	67	Positive	Postmenopausal
TCGA_BC0055	50	Positive	Unknown	TCGA_BC0017	68	Positive	Postmenopausal
TCGA_BC0119	51	Positive	Premenopausal	TCGA_BC0075	68	Positive	Postmenopausal
TCGA_BC0033	52	Positive	Postmenopausal	TCGA_BC0110	68	Positive	Postmenopausal
TCGA_BC0032	53	Positive	Postmenopausal	TCGA_BC0081	69	Positive	Postmenopausal
TCGA_BC0070	53	Positive	Postmenopausal	TCGA_BC0101	70	Positive	Postmenopausal
TCGA_BC0065	53	Positive	Postmenopausal	TCGA_BC0085	70	Positive	Postmenopausal
TCGA_BC0019	53	Positive	Postmenopausal	TCGA_BC0056	70	Positive	Postmenopausal
TCGA_BC0111	54	Positive	Postmenopausal	TCGA_BC0003	71	Positive	Postmenopausal
TCGA_BC0002	54	Positive	Postmenopausal	TCGA_BC0097	71	Positive	Postmenopausal
TCGA_BC0120	54	Positive	Postmenopausal	TCGA_BC0026	72	Unknown	Postmenopausal
TCGA_BC0073	55	Negative	Postmenopausal	TCGA_BC0071	72	Positive	Postmenopausal
TCGA_BC0064	56	Positive	Postmenopausal	TCGA_BC0077	72	Positive	Postmenopausal
TCGA_BC0037	56	Positive	Postmenopausal	TCGA_BC0048	73	Positive	Postmenopausal
TCGA_BC0035	56	Positive	Postmenopausal	TCGA_BC0116	74	Positive	Postmenopausal
TCGA_BC0023	56	Positive	Postmenopausal	TCGA_BC0006	74	Positive	Postmenopausal
TCGA_BC0074	57	Positive	Postmenopausal	TCGA_BC0082	75	Positive	Postmenopausal
TCGA_BC0094	58	Positive	Postmenopausal	TCGA_BC0052	75	Positive	Postmenopausal
TCGA_BC0034	59	Positive	Postmenopausal	TCGA_BC0027	78	Positive	Postmenopausal
TCGA_BC0029	59	Positive	Postmenopausal	TCGA_BC0089	78	Positive	Postmenopausal
TCGA_BC0010	59	Positive	Unknown	TCGA_BC0115	79	Positive	Postmenopausal
TCGA_BC0028	60	Positive	Postmenopausal	TCGA_BC0090	79	Positive	Postmenopausal
TCGA_BC0044	60	Positive	Postmenopausal	TCGA_BC0007	80	Positive	Postmenopausal
TCGA_BC0108	60	Positive	Postmenopausal	TCGA_BC0020	80	Positive	Postmenopausal
TCGA_BC0053	60	Positive	Unknown	TCGA_BC0076	80	Positive	Postmenopausal
TCGA_BC0087	61	Positive	Postmenopausal	TCGA_BC0030	80	Positive	Postmenopausal
TCGA_BC0058	61	Positive	Postmenopausal	TCGA_BC0067	81	Positive	Postmenopausal
TCGA_BC0046	61	Positive	Postmenopausal	TCGA_BC0093	84	Positive	Postmenopausal
TCGA_BC0009	61	Positive	Postmenopausal	TCGA_BC0113	84	Positive	Postmenopausal
TCGA_BC0042	61	Positive	Postmenopausal	TCGA_BC0091	85	Positive	Postmenopausal
TCGA_BC0072	61	Positive	Postmenopausal	TCGA_BC0112	87	Positive	Postmenopausal
TCGA_BC0038	62	Positive	Postmenopausal	TCGA_BC0079	90	Positive	Postmenopausal
TCGA_BC0068	62	Positive	Postmenopausal	TCGA_BC0102	90	Positive	Postmenopausal

Appendix 2

Table representing all primers used for the targeted sequencing experiment. The multiplexing approach is indicated in the first three columns where the stock wells and final wells are indicated for each primer pair. All forward primers had the adaptor sequence ACACTGACGACATGGTTCTACA attached to their 5 end whereas the reverse primers had TACGGTAGCAGAGACTTGGTCT attached to their 5 end to allow for sequencing primers to anneal.

Stock Plate	Stock Well	Final Well	Forward primer	Reverse primer	Amplicon ID
1	A01	A01	GTTCCATCTACCTTTCCCCCAC	CCCTTTCCAACCCCTCCCT	CDH1_t2_2
1	A02	A01	GCCAAACCGCTGTGACCC	CGCAAGCAGGAAGCGTTTG	IDE_t1_1
1	A03	A01	GGTCTCAGGGCTCAGCAG	CACACACCCACCTTCCTTGAG	PABPN1L_t2_2
1	A04	A01	CCCTGATGCCCTCCACCAT	GCTGCCGAAAAGAACCAACTTC	PABPN1L_t6_1
1	A05	A01	CGCCAGGCAGCAAGAGG	GTGTCAAGCTCCTCCAGGTC	ATRIP_t1_2
1	A06	A01	AGGCCTCGGTGAAGGG	CTTGTGCGGAGACAGAGAAGTG	ESR2_t4_3
1	A07	A01	CCTCTGCGCTCATGTTCCT	CACTGGCCATGCGGACA	PABPN1L_t5_1
1	A08	A01	GGCCCCAGAATCTCCTTGGT	GACGCGAGGGGAGTGGA	PABPN1L_t1_3
1	A09	A01	CCGGAACTCCACTGTTAGCTTAT	CCGCCCCGGAAATGACG	SRA1_t1_e2_4
1	B01	B01	AGACAGAGGGTCCATACTAAGCG	CAACTGGCCCGCGTGA	SRA1_t1_e1_1
1	B02	B01	CTGTCGGATACTTGGGGTG	GAACGGGTCGTCAGGGTC	ATRIP_t1_1
1	C01	C01	GACACAAAGCCAGGCCTAAAAC	AACTGGGTCCCGGTGTCG	PALB2_t1_1
1	C02	C01	AAGCCTTCCAAACAGGCTTAT	AGGAACGCGGCTGGAA	SRA1_t1_e2_1
1	C03	C01	CCTTGGTGGCAAACTCTATGTAGG	TGGAGGCCCACTTCAGC	PABPN1L_t4_1
1	C04	C01	CCCCGGACCCTGACGA	CACAGCCATCAAGGGGATCTG	ATRIP_t1_3
1	C05	C01	CGGAACTGCAAAGCACCTGT	GAATGCGTCCCTCGCAAGT	CDH1_t1_1
1	C06	C01	CTCACCGGTTCTGCCCTTG	TCCTTGAGGAGGACTCAAAACA	PABPN1L_t6_2
1	D01	D01	GCGGCGCTGTTGGTTTC	GCGGCCTCTCTCCAGGT	CDH1_t2_1
1	D02	D01	CTGCCTGCCCCTTCACC	GTGTCACCTCCCCTATAAGCC	PABPN1L_t3_1
1	D03	D01	CTAGGCTGAGCGGATTGTTAGG	CAGCAGCACCGTGTCCC	ATRIP_t12_1
1	D04	D01	GCCCCACGCACCTCTG	AGTATAAGCTAACAGTGGAGTTCCG	SRA1_t1_e2_2
1	E01	E01	TTCCAGCCGCGTTCCTTG	GTACGTGAAGCCGGGTGAG	SRA1_t1_e2_3
1	E02	E01	GCCCAACCTTAATGGAAACTGTG	GAGGCCATCAAGATGAAGGTGTG	PABPN1L_t2_1
1	F01	F01	AGTGGACACTGTCTCTAAGGAGC	TTGAGGCTCAGCAGTCTCATGG	CHEK2_t1_3
1	F02	F01	GGCTGTCCCAGAATGCAAGAAG	GTCCAGATGAAGCTCCCAGAATG	TP53_t1_2
1	F03	F01	GGAATCCTATGGCTTTCCAACCTA	CTCCCCCTCCTCTGTTGCT	TP53_t8_6
1	F04	F01	CATCGCAGTGGACTTCTGG	TGCCTGGACCAGCTCT	DCLRE1B_t1_2
1	F05	F01	AGGCTCCTGACACACTGGA	TGGAGATGCTGAATGCCCAC	ESR2_t8_1
1	F06	F01	TAGGTTCCAGTGTGTGTTCCAAG	CCAACCCTTCCCTTGATCTGC	ATRIP_t9_2
1	F07	F01	CACGGTGATGTTGCACAGACAG	TGCTGACCCCCGGCAT	ATRIP_t12_2
1	F08	F01	ACTAGCTGCTCGGGGCTC	GCGGTTTTACCCTGGCAATTC	ESR2_t4_4
1	F09	F01	CTTGGGGTGTCCAGAGAACTTG	TGGGGTTTCTTTCCTCACCTCTA	PABPN1L_t4_2
1	F10	F01	CATCCGTCACCCAGACCC	AAGGAGGAGAAAGAGGAGGAA	PABPN1L_t1_1
1	F11	F01	GTGTTAAGCCTGCTCTCTCTTCA	TTCCCACTGGCCCTCTTTTTG	CTNNA1_t16_2
1	F12	F01	AGTCTTCCAATGCCTGTTCCAAA	TTTTCAGTCCCCGCATCAGAG	SRA1_t2_2
1	G01	G01	CCATCCTGAAGGGCCCATAATC	CTCTACCAGCACGATGCCAAA	CHEK2_t1_2
1	G02	G01	TTTTATCGGGACGCCGTTGT	CACGGTGTGGTCCGAGTG	DCLRE1B_t1_1
1	G03	G01	TCTGTGGCAACAATCAGAGGTTT	TTTGTTCACTGCCCTCCTCTCTC	ATG2B_t42_3
1	G04	G01	TTCACTGGACTGAATCTGGTTGC	GGTGCTCCGCTTCTCTTAGC	ATRIP_t8_3
1	G05	G01	TCTACAGCCTGGGAAAGGAATCA	TCCAACGCAGCATGTTGGAAT	DCLRE1B_t4_2
1	G06	G01	GGAAAACTGATAGCCAGAAAGCC	CAGCTAGTGCTCACCCTCCT	ESR2_t4_2
1	G07	G01	CTCAGCCAGGTTCTCTTGCTC	GCTCCAGACGGTCTCCTCA	PABPN1L_t1_2
1	G08	G01	ACGGCCTATCGCAGAAGGA	GAATCTGACTTGTTGGGGAACCT	ATRIP_t12_3
1	G09	G01	GAGGTGGGACCTGCCCTA	TGGTAAGGAAAGGAGTGGTGCT	ATRIP_t13_1
1	G10	G01	AAATCATTGTGGACCCCTTGAGC	CCATGTACTCCGAAAGCAGGTC	CTNNA1_t6_2
1	G11	G01	GGGCCTCCGTGCACCT	CAGCTGATCGTCTTCTGTCTGG	CTNNA1_t13_1
1	G12	G01	TCAGACTCGACTGGGAAACTTGT	TTAAGCCATGTTGCTTGATCCTG	SRA1_t2_3
1	H01	H01	CCTACGATTGCTATCCTTCCCAC	CTCAGACTGTCCTGAAAGCCTCC	DCLRE1B_t4_4
1	H02	H01	CCTCACCTCTTTCTGTCTAGCTG	GCGGGAAGCATCGATAAACTCAT	CTNNA1_t12_1
1	H03	H01	AGGCCAACAGGGACCTGATATAC	GGGATGGTCAAGAGCTGGAAAT	CTNNA1_t5_3
1	H04	H01	CAAGATCTGGAGCAAAGATGAGC	TTCTGTCTCTACACACACAGGGA	ESR2_t5_2
1	H05	H01	GGAAGATGAAGGCACCAGAGAAA	GACTGATATTCAGGAGCCCCGAG	CTNNA1_t17_3
1	H06	H01	CATTTCCTCCCCCTTGTACAGTT	TGCTAAGTGCAGTCACAGAGAAG	ATRIP_t8_4
1	H07	H01	CTTAAGAAAACCCTCGCAGGGAC	GGCCCCCTTACCACCAGAG	ATRIP_t11_1
1	H08	H01	AGTGTGAAGGTAGGTTATGGAGC	GTGGGAACAGAGCTGAGGTG	ATRIP_t9_1
1	H09	H01	TTAACTTGCAGACACTTTTCCCA	CGATGCTTTGGTTTGGGTGAT	ESR2_t7_1
1	H10	H01	CTCCAAGAGAAGGATGTGGATGG	TGTGTGCGAGCACTGGAAAG	CTNNA1_t11_2
1	H11	H01	TCAGAGGTAAAGGACCACTCAAA	GGCCCATGAAACTTACCCTGAAT	CTNNA1_t15_2
1	H12	H01	ATGTGAGTGCCACACTGAACC	CACCTCGGCCTTGACCTTG	CTNNA1_t16_1
2	A01	A02	GGTGTAGGAGCTGCTGGTG	TCTGACTGCTCTTTTCACCCAT	TP53_t1_3
2	A02	A02	GAGAGGACTGGCTGGAGTTTG	ACTCACCTTTGTTGTTGGACACT	CHEK2_t1_4
2	A03	A02	AACCAGCCCTGTCGTCTCTC	CTGTGCAGCTGTGGGTTGATT	TP53_t2_1
2	A04	A02	GAGGCCAAGCAGCAGTACATT	TGGGCAGTGTAGGATGTGATTTC	CDH1_t10_2
2	A05	A02	TTCACAGTGGAGGAGAAGGC	GAAGCTCGGAGTAAGAGGAATGG	DCLRE1B_t4_3
2	A06	A02	CTTCAAGCACAGCCCACCT	GGAAGCCATACCCTGTGACTTTT	CTNNA1_t17_1
2	A07	A02	GGATGGAGCACAAGTGGTTTACT	CTGCATCCTTATCCTGTTGTCCC	PABPN1L_t7_1
2	A08	A02	GGTAGAGACCAACCCTGAGGAC	CTCCATCACTAAGCCTGTGAACC	ATRIP_t8_5
2	A09	A02	GACAGGAGCATCAGGAGGTTA	TCAGAGCAATGACTTCTGGCTT	ESR2_t7_2
2	A10	A02	CGCCATTTCTGTGACTCGTCTT	TTATGAAACTGCGGCTCGAGAAC	ATG2B_t42_2
2	A11	A02	TAGAGGAATCCTGCAGAAGAACG	CACCCTGGTGCTGTGAGG	CTNNA1_t5_2
2	A12	A02	GGGTCAGAGGGTCTATCTCTGG	CCCACAAGTTTCCCAGTCGAG	SRA1_t2_1
2	B01	B02	TCAGAGTCAGACAAAGACCAGGA	CTCAAGGGAAGGGAGCTGAAAAA	CDH1_t16_3
2	B02	B02	TTTGGAGAGACACTGCCAACTG	TCGAGGCAGCAAAGGCTC	CDH1_t11_2
2	B03	B02	GCTGCTTCTGGCCTTCTTTATCT	TCTCCGCCTCCTTCTTCATCATA	CDH1_t14_1
2	B04	B02	CAGGTGGTGCCCATTGTAAGT	GGGGATTCAAACACAACACCTTG	DCLRE1B_t4_5
2	B05	B02	AAAAAGCGAAGATTGCGGAACAG	CCTGTGCAGCCTGGTGG	CTNNA1_t14_2
2	B06	B02	CAGATTCTGCTGCTGGGGAAG	CTTAAGGCACTGGGTCAGGACA	ATRIP_t8_6
2	B07	B02	CCCTCTTTGCTTTTACTGTCCTCT	TCGGAGGAGGGAATCTCAGC	ESR2_t8_2
2	B08	B02	CGGTTTCATAACCCACAGATCCA	ccTGGATTAGACAGCGCACTAAA	CDH1_t3_3
2	B09	B02	CCACAGGTTCTTATGATGGGTCA	GCCGATGAAGAACTGTACAAGGG	ATRIP_t8_2
2	B10	B02	CTTTATGAGAAAGGGCACCCTGA	TTCTTAGGATGGGGTGGGGAC	ATRIP_t10_1
2	B11	B02	CAGTCTTGGTGGTGGTAAGAAGG	GAGGAGTCTGTTTTCAGAGGAGG	SRA1_t4_1
2	B12	B02	TGAAGAGGAATCCCAAAGTTCCA	CCCCTGTCATCTTCTGTCCCT	TP53_t1_1

2	C01	C02	CGGACGATGATGTGAACACCTAC	ATACACATTTGTCCTCCACACCC	CDH1_t7_2
2	C02	C02	ATCTGGGGTATCAGGTAGGTGTC	CCTGGAGTCGATTGATTAGAGCC	BRCA1_t23_1
2	C03	C02	CAGCTACATGTTGTTTGCTGGTC	ATGATTAGGGCTGTGTACGTGCT	CDH1_t11_1
2	C04	C02	AGGTGATAAAAGTGAATCTGAGGCATA	TGTTTGTGCCTGTCCTGGG	TP53_t5_3
2	C05	C02	AAATTGGTTTAGGGTCCCCCTTG	ATACCCTGCCTGGAAAAAGTTCA	DCLRE1B_t4_10
2	C06	C02	CACAGCTCATGGACCTCTACTTT	CCAGTGCGCCCTTCACC	ESR2_t4_1
2	C07	C02	TCATCTTGTCTTCTGCAGGTTCC	CTCCTTCCCAAACAAGCTTTCCA	ATRIP_t7_2
2	C08	C02	CCTGACCCGGGGAGGTAA	GACAGAACCAGGACAGTGATTGG	DCLRE1B_t2_1
2	C09	C02	GCTTCACACCAGGGACTCTTTT	AGCCATGACATTCTATAGCCCTG	ESR2_t1_2
2	C10	C02	GCACAATAAAATTAAACGAGCTTCCTC	GCCACAGAAGCAACGTCAAAC	ATG2B_t42_1
2	C11	C02	AGGAGTTGGATGACTCTGACTTT	GAAACAAGAAAGGGGACAGGGAA	CTNNA1_t13_2
2	C12	C02	ATACCTTGCACCAGTAGAGCCAT	CTTCCTATGTGGTGCTGTCTTCC	SRA1_t3_2
2	D01	D02	CATTTCAGCAATCTGAGGAACCC	CATTCCCCTGTCCCTCTCTCTT	BRCA1_t20_1
2	D02	D02	TTGTGTTTGCACAGTGCCTTTC	TATTGTCCTGAGTCATCCCTGTG	PALB2_t12_1
2	D03	D02	aaGAGGAATCCTTTAGCCCCCTG	CTCCCACGCTGGGGTATTG	CDH1_t9_1
2	D04	D02	ATCTGAAAGCGGCTGATACTGAC	CTTCTTGAAGCGATTGCCCCATT	CDH1_t16_2
2	D05	D02	TTCTTGTCCTGCTTGCTTACCTC	CCTTACTGCCTCTTGCTTCTCTT	TP53_t5_4
2	D06	D02	TCAACCTTTTTTCTCCAAAGGACT	GAGCCATGCTTTGGCTTTCC	CDH1_t15_2
2	D07	D02	GACCCAAAACCCAAAATGGC	CATGTGATGTCATCTCTCCTCC	TP53_t9_1
2	D08	D02	CAAGCCTTCTCTGGCTGTTAGAA	CAACTGCTTCTTGATCCGCAAAG	DCLRE1B_t4_6
2	D09	D02	ATATTCTGGTTCCATGTGTTGGG	ATCGGAGGATTATCGTTGGTGTC	CDH1_t8_1
2	D10	D02	GCTGCAGATTTCTCTTCATTGGC	CACAGGCTGCTTGCTCAAC	SRA1_t4_2
2	D11	D02	CACCATCCCAGTTCTGATTCTGC	GATCACCACTGAGCTACCAAGG	CDH1_t14_2
2	D12	D02	TTCAGATAGTAAGCCCCACAGTC	CAGAGGCCTGGGCAATTCTTC	ATRIP_t6_5
2	E01	E02	CCTCACAACCTCCGTCATGTG	CTTGTGCCCTGACTTTCAACTCT	TP53_t2_2
2	E02	E02	AACATAGCCCTGTGTGTATGACT	CAATTTCATCGGGATTGGCAGGG	CDH1_t15_1
2	E03	E02	CTGCTGATCCTGTCTGATGTGAA	TTGGGTCGTTGTACTGAATGGTC	CDH1_t12_2
2	E04	E02	TCAGGGCAGAATTGGATTAAGCA	ATTTTTGTCAGGGAGCTCAGGAT	CDH1_t7_1
2	E05	E02	TGGCTTAGAGAAGGAATATTTGATGGT	ACCTCATCAGAATGGTAGGAATAGC	BRCA2_t10_56
2	E06	E02	TCTTCTCTCCTTGTAGTCTTCCCA	GAGCATATACCTCAGCCAAAGGA	MME_t17_2
2	E07	E02	GCCACGTTTTACTGAGCAAGTAG	AGCAAATGATAAGAAAGAAGCTGTT	CTNNA1_t12_2
2	E08	E02	CACTAAAGGAGAAAGGTGCCCAG	GAGCACGGCTCCATATACATACC	ESR2_t1_3
2	E09	E02	CTTCACACGACCAGACTCCATAG	AGAAAGCCCTTCCTTTCCCTTTT	ESR2_t2_2
2	E10	E02	GTCACAGTCACAGGTAGGTTGTC	TTGGCCCTCAAGGCTCCTAT	PALB2_t5_3
2	E11	E02	ACAGAGACTCAAAGAAGGCCAAG	CAGTGGGGCCGTCAACATA	DCLRE1B_t4_7
2	E12	E02	GCCCCTTCTCCCATGTTTTCTT	GACAGACCCCTTAAAGACCTCCT	CDH1_t6_1
7	E01	E02	CTCCAGCATAGCCAACCACA	TCTCTTAGAAGCAGGTATGTGATGA	SRA1_t3_1
2	F01	F02	TCCTGGCATGTGTTTCTACAGAG	CTCAGTCTGTCTTGCCAGTGATA	PALB2_t5_2
2	F02	F02	CTGACTTGGTTGTGTCGATCTCT	AGACCTTTCTTTGGAAACCCTCT	CDH1_t8_2
2	F03	F02	CTGCCCTGCAGTGAATTTTGAAG	TATTCAGCGTGACTTTGGTGGAA	CDH1_t3_2
2	F04	F02	GGCAGCTGTCAAAAGAATTGAGG	GTCTTCTGTCCAAGTGCGTTTTC	CHEK2_t9_1
2	F05	F02	AGCAGAAGGGAAAGGAAAGGAAA	CTGCTGTCATCTGATCCTCCT	CHEK2_t2_1
2	F06	F02	GTCAGCGTGTGTGACTGTGAA	TTCCAGGAAATAAACCTCCTCCA	CDH1_t13_3
2	F07	F02	GTGAGTCTCAAAAGAGGGTGACT	GAGAACCATGGCCCATCCAG	DCLRE1B_t4_9
2	F08	F02	TCCAAAGCCTCAGGTCATAAACA	GAAATTGAAAGGTGGGGATCTGG	CDH1_t12_3
2	F09	F02	CAGCTTACAGTTGCCACCTTTTC	TCTCTGAGGTGACTACGTGAATG	CTNNA1_t11_1
2	F10	F02	cctCAGAGGACAGGGCTTTT	CCTCCCACTCTGCTATAGGACATAA	ESR2_t3_1
2	F11	F02	ACTGTGTGTAATATTTGCGTGCTT	GGCTGAGACAGGTGTGGA	BRCA2_t26_1
2	F12	F02	CGGGCTAGTGTCTTGCTGTATTC	tgttatggACCAGTGCTACTCCC	PALB2_t2_2
7	F01	F02	ATCATTAGTGGGGCTGCCTTG	ACATGTTATAGAAAACTCCTCATCCA	CTNNA1_t6_3
2	G01	G02	TTGCTAGACTTCTTGCCCCAGAT	GAGCTCAGACTAGCAGCTTCG	CDH1_t16_1
2	G02	G02	AGAGACCCCAGTTGCAAACC	CCCAGGCCTCTGATTCCTCA	TP53_t3_5
2	G03	G02	TGGATGTGCTGGATGTGAATGAA	AGTTGCTGCAAGTCAGTTGAAAA	CDH1_t10_3
2	G04	G02	GCAGAATTGCTCACATTTCCCAA	CTGGGTCTTTTCCCTTTCTCTCC	CDH1_t4_2
2	G05	G02	TGAGTGATAAACCAAACCCATGC	TGAAGTGACAGTTCCAGTAGTCC	BRCA1_t22_1
2	G06	G02	CTGCTCACACTTTCTTCCATTGC	TGACCCTGAATCTGATCCTTCTG	BRCA1_t15_3
2	G07	G02	CAGGTGATTTGATGAAGGCTGCT	AGACATAGGCCTGTATACTTACAACT	CTNNA1_t3_2
2	G08	G02	TGTTAATGTTCAGTGGAATCTCATGT	GCCTTCTTAGAACGACCTCTCTT	CTNNA1_t2_1
2	G09	G02	ATCCTAAGTCTGGCAGGAAATGT	CTATGCAACACCAGACAGGAAGT	GOLGB1_t19_1
2	G10	G02	TGTTTGGGGCTCTTGACCTTATC	CATTCCGGGCAACCAGATTCA	ATRIP_t8_1
2	G11	G02	GACAGCACCCACTGTTGAAGAT	GTCAGGTGAGACCCACAAAGAAT	ATRIP_t8_7
2	G12	G02	GGCACCCCTCTTCTAGCTACTG	CTGCAGTGTGGGCTCAATGTT	DCLRE1B_t4_11
7	G01	G02	TTGCAATGATCCACAAAAGATTGCT	AGGTAGCGCTCACTATGTCTTCT	IDE_t18_1
2	H01	H02	GCTGTGTCATCCAACGGGAAT	TTGGGGTCCAAAGAACCTAAGAG	CDH1_t6_2
2	H02	H02	GAGTTCACAACACAGCAGCAC	ACATCAAATGCCCCCACTTTACT	CHEK2_t15_5
2	H03	H02	AACTCCAATCAGAACCTTCCACC	TATTCCTGAGGACCAAGAACCTG	CHEK2_t1_1
2	H04	H02	CTAGACTTGGTCTGGTGGAAGGC	CCTGAGGCTTTGGATTCCTCTC	CDH1_t12_1
2	H05	H02	GTGCAGGCTGATTTTCTTTTTCC	ATTAGGATGTCTGGCACATGCAC	PALB2_t4_3
2	H06	H02	AGTCTTCGCACAGTGAAAACTAA	TCTACCAGGCTCTTAGCCAAAAT	BRCA2_t3_2
2	H07	H02	GCCAAGAACTTGATGAATGCTGT	CCCGTTTAATCTTGGTCTGTGTC	CTNNA1_t17_2
2	H08	H02	GGCAAGTCAGAAAGTCAGATGGA	TCTCTTTTGCACTTATGGATGCAC	MME_t1_2
2	H09	H02	TCCCTTCCTTGAAGATTCACTCG	CACCTGCCTACACACTTACCTTT	FMO2_t6_4
2	H10	H02	CCCAAGATTTTTCTTTAGGCCACC	AAGCTTTGAAATGCAGACCACAC	ESR2_t11_1
2	H11	H02	TTTCAAATGAGCAAGTTGGGGTG	CCATATTACTTTATACTCCTTTAAATACGGTT	PALB2_t5_4
2	H12	H02	ATACTGCTCTGTAGTGCTTCACC	AAACACCCTTTCCATTCAGCTCT	GOLGB1_t17_1
7	H01	H02	GGGGCTTTGTGGACAATCTTCTT	TCACTTACATGGTCTGCAATGGT	CTNNA1_t15_1
3	A01	A03	GGTGCCAGTCTTGCTCACAG	GGCCTGGGTTAAGTATGCAGAT	BRCA1_t21_1
3	A02	A03	AGAAGGCAAGGTTTTCTACAGCA	AAAATCCTGGGTGGATGTTACCC	CDH1_t5_2
3	A03	A03	GGAAAGACCCACAGCTAACATCA	CCGTGGCTGGCCTAACTTTTTT	CHEK2_t2_2
3	A04	A03	TTCTCCAGCCCAAGAATCTATCA	GAATTTGCAATCCTGCTTCGACA	CDH1_t13_2
3	A05	A03	TCAATGCCTTCTTGGTTTGGTCT	AATTTGAGTTATTCTGTGTATTAGAACTTTATTTT	IDE_t5_3
3	A06	A03	GTAGTGCAGATACCCAAAAAGTGG	GCTTCAAGAGGTGTACAGGCATC	BRCA2_t17_3
3	A07	A03	TCTATTGTTGGAGGTAGGAGAGG	TGTCGGGAAGGAAGAACCAG	DCLRE1B_t3_1
3	A08	A03	AAGCCCTGGAGGTTGGTGA	TCTAGTCTGGAGAAGTCACTGGG	DCLRE1B_t2_2
3	A09	A03	TTGGCTAATGCACTCTGAGAACT	ATTGCCACTGTCGTCCCATTT	CTNNA1_t14_1
3	A10	A03	AGGACAGTTGGCATCCAAAGAAT	CCAGCAGGAGACTCAGAACTAC	ATRIP_t7_1
3	A11	A03	GTCATGTCCTCTATGGACTTTTCC	AAATGACCCAGGTCTCAAATGCC	IDE_t21_3
3	A12	A03	AGCCTGCCCTGGAGGAA	CTCTTGACAGTCTATTTGGGATATTTATTT	PALB2_t12_2
8	A01	A03	GGCTATCTCCTTCAGCCTTTACC	TCTGCATCTGATGATGCTTTTTG	GOLGB1_t19_2
3	B01	B03	AGGTTGGACTGTTAGACCTGAAG	ATGGGCCTTTTTCATTTTCTGGG	CDH1_t4_1
3	B02	B03	CTCCTAAACTCCAGCAGTCCAC	TGGCAAGTTCAACATTATTCCCT	CHEK2_t11_2
3	B03	B03	GGCACAGCAGGCCAGT	CTCATCTTGGGCCTGTGTTATCT	TP53_t4_6
3	B04	B03	AGATTCCCACTTACCTCTGCGAA	CAGAACTGTTGCCATTGTGTCAG	PALB2_t7_2
3	B05	B03	AAAATAACCTAAGGGATTTGCTTTGT	TGAAACAAACTCCCACATACCACT	BRCA2_t4_1
3	B06	B03	AACGTCGTAATCACCACACTGAA	TGTGAGGATGCCAGTTTCTGC	CDH1_t9_2
3	B07	B03	CAGACACCACCATGGACATTCTT	CAACCTCTGCATTGAAAGTTCCC	BRCA1_t15_2
3	B08	B03	AGATGATCCCTGCTCTTCTGT	GCCAGTAGCAATCTTCCTGTGAT	CTNNA1_t3_3
3	B09	B03	AAACATTTTGTTGATTTATTAGGATTTTCCA	GCCCAGTAGCTGTACCAACTC	DCLRE1B_t4_1

3	B10	B03	TGCTACTGATTTCTTCCTGTTCCT	ACACCCAAGAACATTTTCCCCAC	PALB2_t4_15
3	B11	B03	CATCCTCCAAGATACCCCCCATA	ACAGAAGAAGCCTTAAGACACCC	CHEK2_t13_1
3	B12	B03	GAACCTGCTTCTCCAAGACTGC	CTCAGAACATTTCACCTCAGCAC	GOLGB1_t17_2
8	B01	B03	GAGACTCACTACATCAGACGGAA	CTGGAAGATTTACAGGCAACCCT	ATRIP_t3_2
3	C01	C03	CTCTGCAAAGGGGAGTGGAATAC	TCTCTTTCTCTTATCCTGATGGGT	BRCA1_t19_1
3	C02	C03	CCAATTTTGGTAAGCTGCCCATC	GCCGGTTGTAAAGAGCCATGTAT	PALB2_t7_1
3	C03	C03	CCAAGATGATTTTCTTAGGCTCTGG	TTTCTTTGCAGCCTGAAGTGATT	IDE_t25_1
3	C04	C03	GTAGCTATGACAAAGGCCAACCA	TCTTTAGTGGAGCAAAAGGGTGA	IDE_t23_1
3	C05	C03	TGAGGAGGGAGGACTGTCTCTAA	TCATCGAAGCCATGGGTGATTT	MME_t17_1
3	C06	C03	ATAACAAGTGTTGGAAGCAGGGA	TGACATTAAGGAAAGTTCTGCTGTT	BRCA1_t9_6
3	C07	C03	ACATTCACTGAAAATTGTAAAGCCTA	GTCTTACCGAAAGGGTACACAGG	BRCA2_t12_3
3	C08	C03	GTCTCAGCCCAGATGACTTCAAA	GTATTTGGTGCCACAACTCCTTG	BRCA2_t26_3
3	C09	C03	TCAATGCAGAGGTTGAAGATGGT	ACTTTGTAATTCAACATTCATCGTTGTG	BRCA1_t15_4
3	C10	C03	TCGTGCTGATATTTGTGTGAGGT	TGACTTGTCTAGGAAGGCAGTTG	PALB2_t4_2
3	C11	C03	TGCCCTGCAAGTGTGAGATTTTA	GTGTTAATGATGGGGCTGATGTG	ESR2_t5_1
3	C12	C03	GACACTGGGGAAACAGATCCATA	ATCTACAGGTGACAGGAAACTGA	DCLRE1B_t3_2
8	C01	C03	CTTTGCGGATCAAGAAGCAGTTG	GGACCCTAAACCAATTTCCTCCC	DCLRE1B_t4_8
3	D01	D03	CTACTGAATGCAAAGGACACCAC	TGCAGCGTTTATAGTCTGCTTTTA	BRCA1_t10_1
3	D02	D03	CAGAAATGAGAAACCACCAATCACA	TGAAACCCATTTCTACTCTTTTCTTC	CHEK2_t5_1
3	D03	D03	GCTCTTAACATGCTCCAGACCTT	AAATGCTGCTTCCACATAAAGCC	MME_t12_1
3	D04	D03	TCAGAAGGAGATAAAGGGGAAGGA	ATGGCTGAACTAGAAGCTGTGTT	BRCA1_t11_1
3	D05	D03	CCCCACACTGAGAACAGTATGAA	GTACCACTTGTCCTCCCAGTATT	SLC15A2_t4_2
3	D06	D03	CTGTGGTTATAAAGCCAGAAGCA	CACATATTTGGGTAGCTTGTTATACAT	ATRIP_t5_2
3	D07	D03	CTAAGATGGGGAAAGCAGGTGAA	AGAGGACCTTATTGTTCTACCAGGA	PALB2_t5_5
3	D08	D03	CTCCACGGCTACTTTCCTCTGG	ACACTCTTGATGGCAGGAATGAAA	PALB2_t4_6
3	D09	D03	ACCCAATTCAATGTAGACAGACG	TTTGTCCATGGTGTCAAGTTTCT	BRCA1_t6_5
3	D10	D03	AGCATGTTTCTTTTGCCTTCCAG	ATGCAAAACTGAACTATCCCTCC	ATRIP_t4_2
3	D11	D03	CCCCTCAACTTGCTCAAACAAATA	TGGACCCAATATAAGAGCACCTTG	MME_t21_1
3	D12	D03	CATTCACTTGACCTGCAGAGGG	CGGCACTTCTTTTCTGGATTCAT	MME_t22_1
8	D01	D03	CCCTTGTAAACACCAATAGTAAAGGG	AACATCTTCTACAGCAGCCACAA	CTNNA1_t2_2
3	E01	E03	CACAAGTTCGCTCTTTGGAGAAG	GAAACCGTAGAGGCCTTTTGACT	CDH1_t3_1
3	E02	E03	TTCTGCTTTTGCTCACCACTAGG	TGACAAGTTACACATCAAAACCCA	PALB2_t4_20
3	E03	E03	AGAATGCAGGTTTAATATCCACTTTG	TTGCAAATGTAAGTGGTGCTTC	BRCA2_t9_11
3	E04	E03	ACAGTGTTAGGTGAAAATGTGGA	ACTACCAAACAGACATGCAAAGC	IDE_t20_1
3	E05	E03	TGCGAATTAAGAAGAAACAAAGGC	ACACTCTGTCATAAAAGCCATCAGT	BRCA2_t14_2
3	E06	E03	CAGACTCTTCCAGCTGTTGCT	CAATTGGTGGCGATGGTTTTCTC	BRCA1_t14_2
3	E07	E03	TCCTTCTCACTCAACCATAAAGTGATT	CCTGGGCACCTTTCTCCTTTAG	ESR2_t1_1
3	E08	E03	CACTCACCACACTTCACCATTCC	CAGTGTTGTTGACCAGGAAGAGA	ESR2_t3_2
3	E09	E03	CCCAAGCTCTTTTGTCTGGT	GCCTGGGAACTCTCCTGTTC	BRCA2_t26_6
3	E10	E03	GACTACTGACAAGTCCATTTCCA	TGCAAAAACTTACCTGAGAATAAGAAA	SLC15A2_t20_1
3	E11	E03	ACGCTAGTTGTAGAAACAGCATC	GAGGGGAAATGCGTAGAAGGAAT	ESR2_t6_1
3	E12	E03	CTAGCAGGCACTGTCCCAC	GAATGGTGGCCTGTTAATTCTGG	CHEK2_t13_2
8	E01	E03	AGGATTATTGGGACTTTGCAGAAC	CCTAGGGCCCATTTTCTTTCGAT	MME_t22_2
3	F01	F03	TCATGCTGTTTACATTCACTAAGGC	AACCTACCTGTGACTGTGACTCT	PALB2_t5_1
3	F02	F03	AGAAGAAAACGGCATTTTGAGTGT	TGCAGTTATGCCTCAGATTCACTT	TP53_t6_1
3	F03	F03	ACAACAAAACCATATTTACCATCACG	CTGTGATGGCCAGAGAGTCTAAA	BRCA2_t20_2
3	F04	F03	ACATCAGTGACTGTGAAAAAGCA	GCCCCGTGAAGGGGAAG	CHEK2_t15_1
3	F05	F03	TCTAAGAAGGCCCATGTTTTGGC	AAAGGCCTGCCTCTCTTTACCTA	CTNNA1_t2_3
3	F06	F03	TCATGGCTGGATATTCATGGTGG	ATGACCTTTGTGCCTCTTCTTGC	ESR2_t1_4
3	F07	F03	TCCTTTGTAGCTTGCTCACACTT	TGGAGAGAATTGCAGCTGAAGAA	ATG2B_t17_2
3	F08	F03	CCCTAGGTTCCATTTTCCCAACT	TTAGTCTTGTGAGAGCAGCCAAG	FMO2_t6_5
3	F09	F03	TATACCAAATTCAGCCCACCTGT	TATGCAGATGGTTGAAGACACCC	IDE_t19_1
3	F10	F03	ACACTATTCAGGAGAAAATGAGAGCC	TGTGACATTTCCCATACCTGACC	IDE_t7_1
3	F11	F03	ATTGTTTTTATTGTGTGATACATGTTTACTT	AGCCAACTGTATTCCTTTTCCAGT	BRCA2_t15_1
3	F12	F03	CTTTGTTCTGGATTTCGCAGGTC	AAGTATTTCATTTTCTTGGTGCCAT	BRCA1_t11_2
8	F01	F03	TGACTTCATCTAATCACCTCCTACCA	TCATGAGAACCTTATGTGGAACCC	CHEK2_t11_1
3	G01	G03	GGAGTATAAAGTAATATGGATGAAGAAAGGC	AGAGACATCTTAAAGAGGGAAGCTG	PALB2_t5_6
3	G02	G03	TGCAATCAAAAGGGAGCAATAAGC	GGTCTCGGGAATGACATTACGTT	MME_t3_1
3	G03	G03	ACAAAGATGGAGACCTCGTTGAC	ACCACCAATAATACACTTAAGATTGAACA	MME_t18_2
3	G04	G03	GCTGTAATGAGCTGGCATGAGTA	TTCAGCTGCTTGTGAATTTTCTG	BRCA1_t9_42
3	G05	G03	AACTTTTCAGCATTTACCAGCAA	CATCAGGGTGCTGTAGGCAT	CTNNA1_t5_1
3	G06	G03	CAGCTGTGTGATTACTTACTGGA	GGTTTCCTGAAGCTATGTTCCTT	ESR2_t6_2
3	G07	G03	GAGATATGCCTTACACGTTCTGC	TCCATAGTCTCAGCATAGAAGCC	GOLGB1_t12_1
3	G08	G03	GATGGAACAGTGGAGGAGAACAT	AAAATGGAACCTAGGGGCTGGAT	FMO2_t6_3
3	G09	G03	TCCCAAATTTACTGAAGGGGTGT	TTCGTACCCTTAGGTTTGCACAG	IDE_t4_2
3	G10	G03	AGTTTGAATCCATGCTTTGCTCT	ACTGCAAATACAAACACCCAGGA	BRCA1_t9_2
3	G11	G03	GAAACTCCCACCACAGCACATA	ACCAATATTAAGCCTTAGTGGGTATC	CHEK2_t12_6
3	G12	G03	CAACAGAAAACTGGGAGCAAAAA	AGGACTTTTTTCCCATATTTAGAATTCAG	MME_t6_1
8	G01	G03	GCAGGTTATTAATGCTGCACTGG	ACATCTTCACATACTTCTCTGGG	CTNNA1_t10_2
3	H01	H03	TGACTACTGGCACTTTTGTTGAAG	TGATCAGTAAATAGCAAGTCCGT	BRCA2_t10_41
3	H02	H03	CCCAGTATTTAGCACACTCAGCA	CAGTCAGAAAAGCCACCTCACTA	IDE_t21_2
3	H03	H03	TTTTGGAGATCATAGCTCAAGCC	TCCATCTCAAAAGGCCATATCAC	IDE_t15_2
3	H04	H03	TGCTGCTATTTAGTGTTATCCAAGG	AAGGAGCCAACATAACAGATGGG	BRCA1_t9_39
3	H05	H03	TGAGCCAAATGTGTATGGGTGAA	CAGCCTATGGGAAGTAGTCATGC	BRCA1_t9_7
3	H06	H03	TTCCTTGTCACTCAGACCAACTC	TATTGGCAAAGGCATCTCAGGAA	BRCA1_t9_3
3	H07	H03	GGTGAGTTCTTATTTCAGTTACTGGTG	CCTAGTGGTGAGCAAAAGCAGAA	PALB2_t4_13
3	H08	H03	TCTGCCAAGTCATCTCTGCAAAA	CAGGTTCAAAGAGGGATGCTCA	ESR2_t2_1
3	H09	H03	ATGATGATGTCCCAAGTCGTCTA	ACGAGTGAATCTTCAAGGAAGGG	FMO2_t6_2
3	H10	H03	AGTGGCCATGGGTTTAAATGAGG	AGACAGAGAGGTGAATGCAGTTG	IDE_t4_1
3	H11	H03	TCTTACCCTCCATCTTCTGCAAAC	GGCTTAGGGCATTGTTTTGTTCC	PALB2_t9_2
3	H12	H03	GGCTTGTGATTTTGAAGCCCAG	CAGAGGGAGCAGCCAGTTTATTT	ATRIP_t4_1
8	H01	H03	GCTGTCCATGCAGGCAACATA	TTGCTCTTATGGTTGTTACCCAG	CTNNA1_t1_2
4	A01	A04	AGAAAATTGTGTTTTCACTTTACCCT	TCTTGACTTCTGGAACAATTGCC	PALB2_t13_2
4	A02	A04	TCAAAGGGCTCCACTGGTTTTTC	TCCTTGGATGATGATGCTTTCAC	PALB2_t5_8
4	A03	A04	TCAAGTTATTCAATTGCTAGTCATGGG	TCGAGGTACTCATTATTCAGTTTGTT	MME_t14_2
4	A04	A04	AAGTGACTTTTGGACTTTGTTTCTT	TCGGGAAACAAGCATAGAAATGGA	BRCA1_t9_17
4	A05	A04	TATTTTCTTTCCTCCCAGGGTCGT	TCAACCTCATCTGCTCTTTCTTG	BRCA2_t6_2
4	A06	A04	ACCTACATAAAACTCTTTCCAGAATGT	GCAATGGAAGAAAGTGTGAGCAG	BRCA1_t15_1
4	A07	A04	GCTCTCTCAACAGGACAATCATCT	TCTCTGGTAAGAAAACAGAAAATAATTTGTAA	SLC15A2_t4_4
4	A08	A04	GCTCTGTGTGACACTCCAGGT	CCTGTATTTTAGTTGAAGAAGCACCC	BRCA2_t16_2
4	A09	A04	AGGGTTATTTCAGTGAAGAGCAGT	GCTTACAATACGCAACTTCCACA	BRCA2_t21_8
4	A10	A04	ACTGCCAAATCTGCTTTCTTGAT	TGGGAAAACCTATCGGAAGAAGG	BRCA1_t9_35
4	A11	A04	AGCCAGGCTGTTTGCTTTTATTAC	CAGCTGAGAGGCATCCAGAAAA	BRCA1_t9_41
4	A12	A04	TCTTCAATGATAATAAATTCTCCTCTGTGT	GAGGGGCCAAGAAATTAGAGTCC	BRCA1_t9_5
9	A01	A04	AACCCTTTCATATTCATACCTTTCTCT	ATCTGAATGCCACTGAGAATGCC	CHEK2_t12_8
4	B01	B04	GCCAGAACCACCATCTTTCAGTA	AGTACCCGTTCCCTTGATGTCTA	BRCA1_t12_1
4	B02	B04	AAGTCCTCCATTTCTGTATCCATGC	ACCTGATGAAGACTTTGGACCTC	PALB2_t5_7
4	B03	B04	TGTGCCTCCAAACTTACAGGT	ACTGCCCAACCAGAAAAAGGT	PALB2_t4_10

4	B04	B04	GAGGTCCTTCTGCACGTAACTTC	GGAGGTGAGGACTGCATTTTCTA	IDE_t9_2
4	B05	B04	TCTTCAGAGTTTTGCAGCCATGA	AGTAGAACTGAGGTATGCTCCCA	IDE_t25_3
4	B06	B04	ACCCTCAATCAAAAAATGCTTCCA	GCCTATTTGTACCTTGAGCTCCT	IDE_t15_1
4	B07	B04	CCCATGTCCTCAAAGTTTGCATT	GCATTTCTGGACTCCTTGTAGGT	MME_t11_2
4	B08	B04	GTGGCATGATCTTTTACATAGGTTT	CCTGTGCAAAGTTCAAGAAAAATAGTTG	MME_t20_1
4	B09	B04	CTCACACAGGGGATCAGCATTC	ACAGCATGAGAACAGCAGTTTATT	BRCA1_t9_40
4	B10	B04	AGGAGAGAAAGGGAAAAGACCCA	CACCTTCAGCCATCCTGTTTCTC	CDH1_t5_1
4	B11	B04	AGCTGACCAATTAATCCAGGAAG	ACAACCTTCTCCCTTCTACATGC	WDR17_t19_2
4	B12	B04	GATCAGAAAGGGTCCCACTGC	TCCTCACATCACCCCATTTTTCC	PALB2_t9_3
9	B01	B04	TTCCTCAGCAAATGATCCTTCAG	CAATGTTCTCCTCCACTGTTCCA	FMO2_t6_1
4	C01	C04	ACTTACTGCCTCTCTTGCTGAAC	CCCAGGAGTGGTAGGTCTCATAA	CHEK2_t6_3
4	C02	C04	TGTTGAAGCAAGGTTCCGAGATA	AGTTTTTGAACTTGCTGCTGTCTT	IDE_t20_2
4	C03	C04	GTGCTTCTTGGACAGCTGAAAA	ATCAATAACAACAATGTATGTGGAAGT	MME_t6_2
4	C04	C04	TCCAGTTGCAGGTTCTTTACCTT	GGAGGAAGTCTTCTACCAGGCAT	BRCA1_t9_28
4	C05	C04	TGCAATGAAGCAGAAAACAAGC	ACGATGGCCTCCATATATACTTCT	BRCA2_t25_2
4	C06	C04	ACTGCACTGTGAAGAAAACAAGC	AGTCTACTAGGCATAGCACCGTT	BRCA1_t9_4
4	C07	C04	ACCAGAATATCTTTATGTAGGATTCAGAG	AGAAACTACCCATCTCAAGAGGA	BRCA1_t14_1
4	C08	C04	AACCTCAGAAACAGCAACTTGGA	AGTGTGAACAAACTGATAGTGTCCT	MME_t12_2
4	C09	C04	TGTTTTCAGTACCGTTCGAATCT	AGTTCCTTCTCGAGTGTCCCATA	WDR17_t13_2
4	C10	C04	CAGAGACAGGTGGGAGGAG	GGGAACATGGTTTTGACCTTTTTT	PALB2_t13_3
4	C11	C04	CTGAAGCTGCACATCATCCAGTT	ACAGAATTGAGGGCTCAGGTAAA	GOLGB1_t12_2
4	C12	C04	CCACTGAGCCTTGTACAGACTTTT	ACATATTAAGAGAGAAATCTTTGATGCAC	MME_t3_2
9	C01	C04	TCAGATGTTGGAATTTTTGTGTTAAATCT	CCACTAATGATGCTTTCCAGACG	CTNNA1_t6_1
4	D01	D04	ATGGAAACTGGCAGCTATGGAAT	GACAACTGGCTTGTGCAACATTT	BRCA2_t16_3
4	D02	D04	AGGAAGTCAGTTTGAATTTACTCAGTTT	CAAATTGCTTGCTGCTGTCTACC	BRCA2_t10_26
4	D03	D04	TCATCATCTTTGCTTATCAGCTCCT	TCCTTCAGACACAGCTACTTATG	CHEK2_t14_1
4	D04	D04	TTAATCTTCACAACAACCCTGTAAAAT	GCAAAGAAAACCAATTTTTGATGCC	PALB2_t10_1
4	D05	D04	TGTGTCATGTAATCAAATAGTAGATGTG	AGCAATTTCAACAGTCTAATCAATGTC	BRCA2_t7_1
4	D06	D04	ACAGTCAATATCAGAATAAACCAAAATGA	TCTCTTAACAATTTCCGGGCTGA	IDE_t17_1
4	D07	D04	ACTAGGACTGCTCCCACCA	ACTGGAAAGGTTAAGCGTCAATA	BRCA2_t26_7
4	D08	D04	TTCAATGCAAGTTTTTCAGGTCA	ACCTGCATTCTTCAAAGCTACAGA	BRCA2_t9_9
4	D09	D04	TGAAAAAGAGCAAGGTACTAGTGAA	GCACCACAGTCTCAATAGAAACAA	BRCA2_t10_51
4	D10	D04	GCTTGCGGGTGTCTTTAGTTC	TGGTCACTTGGTCTTTATTCTGGT	CDH1_t13_1
4	D11	D04	TGTCAGGGAGCTGAACTTCTCTA	AGCCAGTTTATTTCACAAACAAGATA	IDE_t19_2
4	D12	D04	ACCAAAAGCAACAGTTAAGGATTT	GTGGCTGTGGAGGTGGT	CDH1_t10_1
9	D01	D04	GTAAAATGTGCTCCCCAAAAGCA	GGAGTTGGTCTGAGTGACAAGG	BRCA1_t9_1
4	E01	E04	TTCCTGAGTTTTCATGGACAGCA	TCACTTGCTGAGTGTGTTTCTCA	BRCA1_t4_1
4	E02	E04	TGCAATTCAGTACAATTAGGTGGG	GCTTTCAAAACGAAAGCTGAACC	BRCA1_t9_30
4	E03	E04	ACTGAAAGGCTTTATACTCTTCTCATATT	CTCTGTTATTCTGTTTATCAAAGGACCC	CHEK2_t7_1
4	E04	E04	TAAAGGAACTGTAGTCGCCCCTG	CTGCTAGATCACCAGTAACTGAAA	PALB2_t4_11
4	E05	E04	AAGGTTTAAATTTTTACTTGCATCCTTATTT	CTGCAGAAAAACATTCTTGCACA	PALB2_t4_5
4	E06	E04	TCACACTGTGGGAAAAAGAACAA	GTGCCCAAAGAGCTGAAAAGATT	PALB2_t3_1
4	E07	E04	ATTCACTTCCCAAAGCTGCCTAC	CTCTCTAACCTTGGAACTGTGAG	BRCA1_t6_1
4	E08	E04	TGTCATGGACCACGTTTCAGATT	ACACTTCACCAAGTTGCCTACAA	CTNNA1_t8_2
4	E09	E04	AGTTACATGGCTTAAGTTGGGGAG	TGGACGTTCTAAATGAGGTAGATGA	BRCA1_t9_36
4	E10	E04	ACGTATGGCGTTTCTAAACATTGC	TTCTTCTTTTCCAGCCTTTCCAT	BRCA2_t15_2
4	E11	E04	ACTATATGACTGAATGAATATCTCTGGTT	GTGCTCTTTTGTGAATCGCTGAC	BRCA1_t18_1
4	E12	E04	TTTGGTTCTGTTTTTGCCTTCCC	AATAATGCTGAAGACCCCAAAGA	BRCA1_t9_20
9	E01	E04	TGCACCTGTGAGAGGATTAATGT	TTAGTATGTCAATATAACAAATACATGAAAGAATG	WDR17_t13_3
4	F01	F04	AGCTGACAGAGACAAAGATGAAGG	AGCATAATTTTTGGCTGCTTTGTTT	PALB2_t8_1
4	F02	F04	CAGACTTCCAGGACCTTCATGC	AGGAGTCGTATATAGCTAATCTCTGTG	IDE_t7_2
4	F03	F04	TCGGCTATCCTGATGACATTGTTT	CAACCTGTGGTTTCAGGCTACTT	MME_t14_5
4	F04	F04	AATTACAAAAACAAGTACAGAATAGGACT	GGGCCAATGTCATCTTGTTATACAG	MME_t9_1
4	F05	F04	AGCTTTTATGGAAGATGATGAACTGA	TAGTGATTGGCAACACGAAAGGT	BRCA2_t10_75
4	F06	F04	GACAACCCGAACGTGATGAAAAG	ACTTTAGGGTCTTTGCCCATTGA	BRCA2_t10_50
4	F07	F04	TAGAGCATGTGGTGTGATGTCTC	AGAACACGGACTTGTTTTTCCCA	CTNNA1_t10_1
4	F08	F04	CTGTGCCTGGCCTGATACAATTA	CTAGTCTCTTTTGTTGGGCCTCC	BRCA2_t19_1
4	F09	F04	TTTCATTCAAAAGTAAAAAGGTGAATCAAT	CTTCAAGGAAGCCACATATGCAA	IDE_t8_2
4	F10	F04	TGTGAAAAATCTAAAAACCAAGTGAAAG	GCTCCATTTAGACCTGAAAGGGTT	BRCA2_t9_4
4	F11	F04	GAAGCAAAATGTAATAAGGAAAAACTACAG	CCACTTTTGAATGTTGTACTGGGT	BRCA2_t10_4
4	F12	F04	AGTGGCGACCAGAATCCAAATC	TTCCTTGATACTGGACTGTCAAAA	BRCA2_t24_14
9	F01	F04	ACTGTGGTTAACTTCATGTCCCA	ACTGATTATGGCACTCAGGAAAGT	BRCA1_t9_19
4	G01	G04	TGCTTTAGATCGTTTGTCTTGTG	TAGCTGTAACTACAACCACCATC	FAM175A_t9_3
4	G02	G04	ATTTCTTTTTAGGAGAACCCTCAATCAA	GTCAGAATATTATATACCATACCTATAGAGGGAGA	BRCA2_t11_2
4	G03	G04	GCAGCAAGCAATTTGAAGGTACA	TTCACAGCTTTTTGCAGAGCTTC	BRCA2_t10_30
4	G04	G04	ACTGAAAGAAAGTGTCCCAGTTG	ACTAGTACCTTGCTCTTTTTCATCA	BRCA2_t10_49
4	G05	G04	ACATCAGCTACTTTGGCATTTGA	AATAAGCAGAAACTGCCATGCTC	BRCA1_t9_38
4	G06	G04	TTTTTATCAGATGTCTTCTCCTAATTGTG	CCAAGGCTCTTCTCTTTTTGCAG	BRCA2_t26_2
4	G07	G04	CCCTTTGAGAGTGGAAGTGACAA	GAAAATCTTTCTTTCTTTTGTTCTCTGTG	BRCA2_t9_5
4	G08	G04	TGGACTGGAAAAGGAATACAGTT	AGCCAACTTTTTAGTTCGAGAGAC	BRCA2_t15_3
4	G09	G04	GGCCTGCTCGCTGGTAT	AGAAATATATGGTAAGTTTCAAGAATACATCA	BRCA2_t18_2
4	G10	G04	CCAGCACAGAAAAACGAGATCCT	GAGGCTAGTTAGTAGCAGTGGGA	PALB2_t9_1
4	G11	G04	AACTTGTGGGCAGTTGGC	GCACCTTGAACACATTCCTCCTA	PALB2_t4_9
4	G12	G04	TAGGACTTGCCCCTTTCGTCTAT	AGCAGAAAACACAGAAAAATCTCCA	BRCA2_t24_11
9	G01	G04	GGCTTATCTTTCTGACCAACCAC	GCAACATTCTCTGCCCACTCTG	BRCA1_t9_16
4	H01	H04	TACCTCCACCTGTTAGTCCCATT	TGCAAGTTCTTCGTCAGCTATTG	BRCA2_t26_4
4	H02	H04	CCTTTAACTCTGAAACCAATTGTAGG	TTTGGAGCTTTGCTGCTGTTAT	PALB2_t6_2
4	H03	H04	CTGGCGCTTTGAAACCTTGAAT	GTGAGTCAGTGTGCAGCATTTG	BRCA1_t9_18
4	H04	H04	ATGAAACAGTTGTAGATACCTCTGAA	TGGTTCCACTTCAGATACAAATGAGT	BRCA2_t9_3
4	H05	H04	TTTTCAGTGCCTGTTAAGTTGGC	TCTTGGTCATTTGACAGTTCTGC	BRCA1_t8_1
4	H06	H04	AAGAAGTAGAACAGCGTGTGTTT	TTTGATCTCTGTCTCCAGCTGTC	ATRIP_t6_1
4	H07	H04	TTGCCACGTATTTCTAGCCTACC	TCTGAATATAGACTTTTTGATACCCTGA	BRCA2_t9_7
4	H08	H04	TTCTGCTCCGTTTGGTTAGTTCC	ACTGAAAATCTAATTATAGGAGCATTTGTT	BRCA1_t9_34
4	H09	H04	GGAATAGCCACATACAGAATGCC	AGATGTGTGTGTTGGTAACTTTGA	CHEK2_t10_1
4	H10	H04	TCACCAAAACCCTTCATCTTTTCA	GTCCAGCTAGTACACCACAAATCA	FAM175A_t9_2
4	H11	H04	TGCAAAAGTCCCAAAGTAGGAGAA	TGTAACAAGAGGCTCCAACAGTC	CTNNA1_t1_1
4	H12	H04	CCTCAGGTTGCAAAACCCCTAAT	TAAAGAAGCCAGCTCAAGCAATA	BRCA1_t9_11
9	H01	H04	ATTTGGCTTGTTACTCTTCTTGGC	AGTCAGTAGAAATCTAAGCCCACC	BRCA1_t9_27
5	A01	A05	AACTGAGGACCTAGAGGGAAAGC	ACCTAGAGACTGCTTTAGTGCAA	PALB2_t10_2
5	A02	A05	TTCTGAGGAATGCAGAGATGCTG	TAAAAGCCCCTAAACCCCACTTC	BRCA2_t10_29
5	A03	A05	TTGTAGAATGGCCTTAATCAAATGTT	AGGCCTTCCAAAAACACATTCAG	IDE_t21_1
5	A04	A05	CTCTACTGATTTGGAGTGAACTCTT	CCAGAAGTGATGAACTGTTAGGT	BRCA1_t9_37
5	A05	A05	CTTTCTTCAGAAGCTCCACCCTA	GAGATTGGTACAGCGGCAGAG	BRCA2_t2_2
5	A06	A05	GTAATGAGTCCAGTTTCGTTGCC	TGTGAACAAAAGGAAGAAAATCAAGGA	BRCA1_t9_15
5	A07	A05	GGATGCCATCTTCTTTACAGACCT	GACTGTACCTGTTGTCTTTTGGC	CTNNA1_t4_1
5	A08	A05	TCTTTCAAGGAATGAGCACCTCC	GCTGTTTACACTTAGATCTTGGTCTTT	ATG2B_t17_3
5	A09	A05	ACGTATGGAAAGTTCTAATAAATGTCAGC	CAAGTCTCCTGAAGACAAGCG	IDE_t2_1
5	A10	A05	ACATTTTTTCAGACTGCAAGTGGG	TGTTTCCTCATAACTTAGAATGTCCA	BRCA2_t10_47

5	A11	A05	AAAGCAGGCATAAGTGAATGGTC	AAAACGTATTTCTGGGGCTGTT	PALB2_t3_2
5	A12	A05	TGCTGGGCGGTGGTTTTTT	GCAGTGTACATGGTAAGCCTTCATA	MME_t11_3
10	A01	A05	TCTGTTCCAGAATGGTCAGAGGT	CTGAGGACAGAGGGCTTCTAACT	CTNNA1_t9_1
5	B01	B05	CTGTGATACTGAGAAAAGACAGTAGT	TGTCTGTTTTGTTGGGTTTTGTT	PALB2_t5_9
5	B02	B05	AGAACAAGAAGCTATATGACTGAATTCTT	TGATGTGACTTTTGTTTTCACAGAC	PALB2_t6_1
5	B03	B05	GGAGAGCTGACTTTAGTTAATGAGAGAA	GGAGGCTGTCATTCAGAGTCATT	PALB2_t4_4
5	B04	B05	ATTACCTGTCTAGGGCACCTTCT	AAGTCATTTTTGTGAACATATGCTTTTT	IDE_t3_2
5	B05	B05	CAGGAAGGACAGTGTGAAAATGA	AGAACATCCTTGGAAGTAGGAGT	BRCA2_t10_5
5	B06	B05	AGAAGCAGAAGATCGGCTATAAAA	CCTTAACAGCATACCACCCATCT	BRCA2_t17_2
5	B07	B05	GTTACAGCTGGGGAGGTCATGTT	TAAGGTACCCAAGGCCAAAAGAG	SLC15A2_t20_2
5	B08	B05	TACCCAGTCCCCCATGTATTAGG	AGAAATGGCAAAATGCTGACCTG	IDE_t13_1
5	B09	B05	GAACTTCTCCAGTGGCTTCTTCA	GAGTCCTCCTTCTGTGAGCAAA	BRCA2_t9_8
5	B10	B05	TGCAAATGCATACCCACAAACTG	ACAAACGATTTTACCACTGGCTATC	BRCA2_t10_60
5	B11	B05	TTAAAGGGAGGAGGGGAGAAA	AGGCTCTTTAGCTTCTTAGGACA	BRCA1_t17_1
5	B12	B05	TCATTAATACTGGAGCCCACTTC	TCTGCTAGAGGAAAACTTTGAGGA	BRCA1_t9_12
10	B01	B05	TACACTGGGAGAAAACATTGCTG	TCCATTTTCTATTCATCATTTACTGCTT	MME_t19_5
5	C01	C05	aaaaaaCCAAGCTGCTTCTTCTAA	TTCTTATAGGTCCTGTTGTTGGAG	IDE_t24_1
5	C02	C05	GACATGCTTGTATTTTTCAAACTTTTCT	ACACCAGATCATTTTGTAGTTTGG	MME_t7_1
5	C03	C05	CCAGCCACATTAAGCATTTGGAC	GAGCAGGACAAGGACCGAGAG	MME_t1_1
5	C04	C05	TGTGTTACAAAGAATGAATTAATGACCT	TTGCTAGTTAGCTGTCTGCTCTT	MME_t13_1
5	C05	C05	CTTTCCCAAAACATGGCACTCAC	AAGCTCTTTCTTTTCACCTGCAT	PALB2_t7_3
5	C06	C05	AAAAAATTGTAAGGGTTCTTACCTCGAC	AAGAAATCCCTTTTCTTCCCTTCCA	IDE_t5_2
5	C07	C05	AGATGTTCTTCTTGGAAAGGGTGT	GCACTCCCTGTTGGATGTTATGA	IDE_t6_2
5	C08	C05	CCTTTTTCTGGTTGGGCAGTTG	TGTCTGGGAAAAGACTAAAGGAACA	PALB2_t4_12
5	C09	C05	AGTTTCTCTTCTTTTTCTTCTCTTGGA	AAGGTAAAGAACCTGCAACTGGA	BRCA1_t9_24
5	C10	C05	CCTGCTTTTTTCCAGCCATTCAG	AATGCAAACTTTGAGGACATGGG	MME_t10_2
5	C11	C05	GGTATCTTGAAGTTGAGGAATGCTG	CACACAGAATGAGTTTTTCCCTCT	CTNNA1_t4_2
5	C12	C05	GCCTGTTCCATCTCAAATAATGAAG	CCCCAAGACTTCCATCTAAAATCCT	CTNNA1_t9_2
10	C01	C05	TCACATGTCAGGCTTATTCATAGAT	CATATATACAGTATGAGTTTACAGGGTTTTT	IDE_t12_2
5	D01	D05	GTTTCTTACCTTTCCACTCCTGGT	TCACTATCAGAACAAAGCAGTAAAGT	BRCA1_t13_2
5	D02	D05	TGAAATACTCCACACAGCAATGTA	GGTTTCTAAATAAGGGTGGCCGTA	IDE_t10_1
5	D03	D05	CCTCTAGCAGATTTTTCTTACATTTAGTTT	GAAAGATAAGCCAGTTGATAATGCC	BRCA1_t9_14
5	D04	D05	TCTGCCTCATACAGGCAATTCAG	AGCTAAAGAACTTGACCAAGACA	BRCA2_t6_1
5	D05	D05	AAACAACAATTACGAACCAAACCT	TTGTAGTTCTCCCCAGTCTACCA	BRCA2_t2_3
5	D06	D05	GGCCAGGGGTTGTGCTTTTTA	TTCGAGGCAGAGTGGATGTTTTT	BRCA2_t14_1
5	D07	D05	AGTACAGCAAGTGGAAAGCAAGT	CACAGTGCTCTGGGTTTCTCTTAT	BRCA2_t10_69
5	D08	D05	ACCAGAAGAATTGCATAACTTTTCCT	CTTTTCATCACGTTCGGGTTGTC	BRCA2_t10_48
5	D09	D05	AGCACAAATGGCTATAGGCTATCATT	TGAGAACTTAGAAAAACAATCAGCAC	CTNNA1_t7_1
5	D10	D05	AGTCTGGGTTTTATATCGTCTGC	GTCCTGGTATCCAGTGCATCG	ESR2_t10_1
5	D11	D05	ACAATTCTGACTTTGTTTCCTTGAAT	ACCACTTACTTATTGTTTTGGCTCC	WDR17_t11_1
5	D12	D05	TGGTGGGATCACTGATAAGAAGTA	TCCTTAAATATAAAGTCACAAATCAACAATTAAAA	IDE_t2_2
10	D01	D05	CTGGTTATTTAACAGATGAAAATGAAGTG	TTACTTGAAGATAAACTTATTGGATGTACC	BRCA2_t10_32
5	E01	E05	AGTAACACACAAAGTGGTCCCAG	ACTGGTTTGTTGGAAGAATGTGA	PALB2_t11_1
5	E02	E05	TGGATGGAGAAGACATCATCTGG	ACAATGTGTACCATATAACTAATTTTACCTT	BRCA2_t19_2
5	E03	E05	acacttggccCTGTCACTTTTTA	CAACTTCTAGCCTGTCGATTGTT	PALB2_t4_1
5	E04	E05	AGTTTTGGTTTTCATTTGCTGGT	CACCTGTAAGTTTGGAGGCACAA	PALB2_t4_8
5	E05	E05	ACAGATCCACAGCATGAAGAAGT	TGAATGTTGCAGGCTGGGTTAAT	IDE_t8_1
5	E06	E05	TCATTATACTATTCTCTACTTTTGTAATGCTTG	TGGAAGAATTTAGACCTGACTTAATAGA	IDE_t11_1
5	E07	E05	CCTGATACTTTTCTGGATGCCTCT	TCCACCTCCAAGGTGTATGAAGT	BRCA1_t9_43
5	E08	E05	TGTACAGAGAATAGTTGTAGTTGTTGA	AGGAAAGGCACATTCCATAGCTG	BRCA2_t16_1
5	E09	E05	CGCTTTTGCTAAAAACAGCAGAAC	TCAGACTGTTAATACAGATTTCTCTCCA	BRCA1_t9_8
5	E10	E05	TCACCTTGTGATGTTAGTTTGGA	AAAGACTTGCTTGGTACTATCTTCT	BRCA2_t10_66
5	E11	E05	GCTTTTCTCCCCATCTGTAAAGGA	AGAACAATTTCATCTTTAGTCAGCTAATC	ESR2_t9_1
5	E12	E05	ATTTGTGGTGTACTAGCTGGACT	TTGTCAGGCATTACGGACCTTTT	FAM175A_t9_4
10	E01	E05	GCTACTCCAACAGTTTAGTGCT	TTGGCAATTTCTTTTTCCAATTCCA	MME_t8_1
5	F01	F05	GTTTTATGCAGCAGATGCAAGGT	AACTAGTATTCTGAGCTGTGTGC	BRCA1_t16_1
5	F02	F05	TCCATATTGCTTATACTGCTGCT	TCAGGGAACTAACCAAACGGAG	BRCA1_t9_32
5	F03	F05	GCTTCATTACAAAACGCAAGACA	CCAGAGAAAGCAGATGAATTTACCA	BRCA2_t10_67
5	F04	F05	CAAATGAGGGTCTGCAACAAAGG	TGCTTGAAGATTTTTCCAAAGTCAG	BRCA2_t13_1
5	F05	F05	ACCGGTACAAACCTTTCATTGT	GAACAAGATGGCTGAAAGTCTGG	BRCA2_t23_1
5	F06	F05	TGTGCTCACTGTACTTGGAATGT	AGGCAACGAAACTGGACTCATTA	BRCA1_t9_13
5	F07	F05	ACGAAACACCCATAAAGAAAAAAGAACT	GTGGGAGCAGTCCTAGTGGAT	BRCA2_t26_5
5	F08	F05	CTTGATTCTGGTATTGAGCCAGT	GCTGCATTTTTATTTTTGCAGGGTG	BRCA2_t10_59
5	F09	F05	AGATGAAACGGACTTGCTATTTACT	TTTGCTCCGTTTTAGTAGCAGTT	BRCA2_t10_45
5	F10	F05	TCATTTGCATAGGAGATAATCATAGGAA	AATGTGTTAAAGTTCATTGGAACAGAA	BRCA1_t1_1
5	F11	F05	CTTCTTTTCCTCCTTGCTTCTTTT	GCAGCCAATTTTACTAAAGAAGATTATTG	IDE_t16_1
5	F12	F05	ATGGAGCAGAACTGGTGGG	AGTACATCTAAGAAATTGAGCATCCTT	BRCA2_t17_4
10	F01	F05	TGCAATTTCAGAATTGTTATTCAAAGG	TCATTGTTTTTAGATATTTTCCCACTATAAATCT	CHEK2_t4_2
5	G01	G05	TTGTTTCAGACTTTGAATAGCAGAG	GGGACAAAAGCTGTGAATATTGCT	CHEK2_t3_1
5	G02	G05	ccatcgtaagtcaagtagcatctgT	TCAAAGACAATGGCTCCTGGTTG	TP53_t7_e6_1_4
5	G03	G05	TTTGAACTTCCTGCATGCTCACT	TCACCTGTACTTACTAACTTTCAGCA	IDE_t3_3
5	G04	G05	AGAACAGACAACTCCTGGAAACG	AATGCGCCTCTATTTCCCTAACC	IDE_t14_1
5	G05	G05	GCAAGGCTGGTGCTTGTGA	AGGATCTGAGAAGTACATGAACGG	IDE_t24_3
5	G06	G05	TTTGCTCTTGAAAAACAGATTCCTT	TGGTTTTTAAGTGTGCCCCTTTT	IDE_t9_1
5	G07	G05	TGCTCTTCATTTGATTTCATTAGGAGT	TTTTGCTCCCAGTTTTCTGTTGC	MME_t5_1
5	G08	G05	TAAATCCATAGGCTACGGCTAAAC	TCAATGTTATAATTTAGAAAACGGCACA	MME_t9_2
5	G09	G05	AGTGATATTGAGAATATTAGTGAGGAAACT	ACAAAAGTGCCAGTAGTCATTTCA	BRCA2_t10_37
5	G10	G05	TCTAACACTGTGAAAAAGCCAACA	ACTCTGGAGACTAGACCAAACCA	IDE_t5_1
5	G11	G05	TGGAGAAAATACCCCTATTGCAT	AGATGCTGCTCTTCATCTCTCTT	BRCA2_t9_6
5	G12	G05	AGATATGGAGAGAAATCTGTATTAACAGTC	TCCAGTGATGAAAACATTCAAGCA	BRCA1_t9_10
10	G01	G05	ACTTGACTTGTGTAAACGAACCC	ACCTAGAGTCATTTTTATATGCTGCTTT	BRCA2_t10_12
5	H01	H05	AGCTAATAATGGAGCCACATAACACA	ACTCAGTCATAACAGCTCAAAGT	BRCA1_t2_1
5	H02	H05	ACCATATTCTGTAAGGACAGGACAAA	GATCACAGTGGCAATGGAACCTT	CHEK2_t4_1
5	H03	H05	CAGCAGTATTTCAGTCCTTGCAC	TGATACACCATGCACTGGGATTG	MME_t18_1
5	H04	H05	TGCAAAAAACTGGAGAAAGTATGGT	CCTATAAGCCAGAATCCAGAAGGC	BRCA1_t13_1
5	H05	H05	TCCCTTCTTTGGGTGTTTTATGCT	TGCTGCATTCTTCACTGCTTCAT	BRCA2_t20_1
5	H06	H05	TGCTACTCTCTACAGATCTTTCAGTT	ACCTGGTTCTTTTACTAAGTGTTCAAA	BRCA1_t9_22
5	H07	H05	AGCAGTTTCAGGACATCCATTTT	AGCCATGTCCATCAATGTTTTGC	BRCA2_t13_3
5	H08	H05	CCTGAGATGCATGACTACTTCCC	AGATTAGGGGTTTTGCAACCTGA	BRCA1_t9_9
5	H09	H05	TCGAGTGATTCTATTGGGTTAGGA	CAAGAAAGCAGATTTGGCAGTTCA	BRCA1_t9_33
5	H10	H05	AGTTTGTTCTACTTACTCCAAAGATTCA	GGGCTCTCCTCTTCTTTTTCCAA	BRCA2_t10_74
5	H11	H05	AACAATTGATGGTAAAGTGCTACACA	AGATATTACCTTTGGAAATAGCACTAAACT	WDR17_t11_2
5	H12	H05	ACTCACAAATTCATCCATCTAAGCA	AGTAGCCATAAAGATCATCAGCAA	CHEK2_t6_1
10	H01	H05	ACCCTCTAACTATAACTGAATCTTGGA	AAGAATGGGCTTAAGTTTGGTTAAAT	MME_t10_1
6	A01	A06	TCCCTGTGTAAGTGCATTTTGGT	TTTTAGAAAACACTTTCTCGGTGTAAT	BRCA2_t1_1
6	A02	A06	ATTTTTGCAGAATGTGAAAAGC	CAAAACAACAACAACAAAAAAACC	BRCA2_t8_4
6	A03	A06	CAGGCTAGGCTAAGCTATGATGT	AAATGGTTCTATGACTTTGCCTGA	TP53_t7_e6_1_1
6	A04	A06	TGTCATGATTCTGTTGTTTCAATGT	TCTACTGGCAGCAGTATATTTGTT	BRCA2_t10_38

6	A05	A06	TCAGAACTGAGCATAGTCTTCACT	TTTCTGAAGAACCACCTTCAACA	BRCA2_t10_70
6	A06	A06	AGTGTTTCTTGCTGTATGTTCGG	TTTTCCTACAATTAGCATTTATGAGCA	CHEK2_t3_2
6	A07	A06	GGTATGGTGTCGCCTCTTTTTC	GCCCATATATTGAGCAGAGATACT	IDE_t13_2
6	A08	A06	TGTTAAATTATGAAGCCATCTCTGTGG	AGTGAGCATGCAGGAAGTTCAAA	IDE_t3_1
6	A09	A06	GTAGTACATGGCATGCTGGTGAG	ACTTCAGAAAGGGGAAAGAAGGG	IDE_t17_2
6	A10	A06	TCTCATTCCCAGTATAGAGGAGACTT	TGGACAGGAAACATCATCTGCTT	BRCA2_t3_1
6	A11	A06	GAGGCATTGGATGATTCAGAGGA	GTTTCCAAACTAACATCACAAGGT	BRCA2_t10_64
6	A12	A06	AATAATTTTGTCTTCCAAGTAGCTAATGA	CCTCTGCAAGAACATAAACCAAATCTT	BRCA2_t10_11
11	A01	A06	TCTTTTGGGACAATTCTGAGGAAA	TCTGAAACTTTTTTGCTTTTTGGATCA	BRCA2_t10_3
6	B01	B06	ACAGACAGTTTCAGTAAAGTAATTAAGGA	CACTCTGAATGTCAGCAAAAACCT	BRCA2_t10_63
6	B02	B06	AGGCTTGTACAGCATATGTGGATT	AAAAAGCGGTTGACATTATTCAGT	MME_t8_2
6	B03	B06	CCGGTTTTAATATTTCTTCCCAAATCT	AAGGAGGGAAAGACCTGCTTCTA	MME_t15_1
6	B04	B06	TTTACCAGCCACGGGAGC	GGTAGCTCCAACTAATCATAAGAGAT	BRCA2_t23_2
6	B05	B06	ACTGCTATACGTACTCCAGAACA	GGTGAATAGTGAAGACTATGCTCAGTT	BRCA2_t10_68
6	B06	B06	TGATATATAATGTTCAGTTTTTAGTTCTTGCT	ACGTAAAAACCATTTCTTACCTGATT	MME_t2_3
6	B07	B06	GAAATTTTGTAACCAGATATTTTGAATGGT	GTGGGGAAGGGACATGTTAGC	ATRIP_t5_1
6	B08	B06	TTTCTCCATCTCCCCAACTACAT	AAATGTGATCTTCTCTGTTCCTCT	ATRIP_t3_1
6	B09	B06	AGGTTGACTTAGAATCTCACTTTCCTG	TGGCCAACTGCCCACAAG	PALB2_t4_7
6	B10	B06	AAACAGCAAAAAGTCCTGCAACT	TCTGGTTGACCATCAAATATTCCT	BRCA2_t10_54
6	B11	B06	CAGAATCCAAATTTTACCGCACCT	TGGTTGGTCTGCCTGTAGTAATC	BRCA2_t13_2
6	B12	B06	AAAGAAAAGAAGAAGAAGAAGAAGAAGAA	CCTTGTATTTTACAGATGCAAACAGC	BRCA1_t5_1
11	B01	B06	ACTTACTGCAAGTAGCTCAACAT	AGTAGTAACCAAGATAAAGCATCCA	FAM175A_t9_1
6	C01	C06	actgtgcccAAACACTACCTT	TCAGAATTGTCCCAAAAGAGCTA	BRCA2_t10_1
6	C02	C06	GAGGTAGCTTCAGAACAGCTTCA	AGGCTTGCTCAGTTTCTTTTGATT	BRCA2_t10_18
6	C03	C06	TCTTTAACTGTTCTGGGTCACAAAT	TGGAGTTTTAAATAGGTTTGGTTCGT	BRCA2_t2_1
6	C04	C06	TGGCGTCCATCATCAGATTTATATTC	ACTAACAAGCACTTATCAAAACTGAAA	BRCA2_t22_2
6	C05	C06	ACTAGCTCTTCACCCTGCAAAAA	TCTCGTTGTTTTCCTTAATTACTTTACTG	BRCA2_t10_61
6	C06	C06	GGGAAAAGAACAGGCTTCACCTA	ATCTGTCAGTTCATCATCTTCCAT	BRCA2_t10_73
6	C07	C06	CAGCTGCCCCAAAGTGTAAAG	TGCAGGACTTTTTGCTGTTTCT	BRCA2_t10_52
6	C08	C06	AAGTGTACAAGAGAATAAAAAGCAATCT	GGGCATACTCCTTAACTTCTTTCTCA	CTNNA1_t8_1
6	C09	C06	ACCTAGTCATGATTTCTAGAGGCAAAG	TCTTGAAGGTGATGCTACTCTCA	BRCA2_t10_7
6	C10	C06	CAACTGCATTCACCTCTCTGTCT	TGGTAGTTTTGTTTCTGATTCTGC	IDE_t4_3
6	C11	C06	CCCAAAGCTACACACACGAGATT	AGTTTTCTGAGCCTTCAAATGATGA	PALB2_t8_2
6	C12	C06	GGGACTACTACTATATGTGCATTGA	ACAGAGGACTTACCATGACTTGC	BRCA2_t8_3
11	C01	C06	TATTTATGTATATTCTCTCCTTTTTCTAGATGGT	AATTGATCTACAATGAATAAAAGTGTAAACAAAGC	MME_t2_5
6	D01	D06	ATGTTCTTGCAGAGGAGAACAAA	TGAAGCTACCTCCAAAACTGTGA	BRCA2_t10_14
6	D02	D06	CCGGACATTTTCTGGTCTGAGTT	ATCTGCTTTAATAAACTATGTAGTAGCTTTG	IDE_t11_2
6	D03	D06	AGTCTTTTGGCACGGTTTCTGTA	AGCATACATAGGGTTTCTCTTGGT	BRCA1_t5_2
6	D04	D06	CCTTGTTTCTATTGAGACTGTGGTG	TCAATGACTGAATAAGGGGACTG	BRCA2_t10_53
6	D05	D06	ATGTAGCACGCATTCACATAAGG	TGCAGATGAGACTGACTTATGAAGC	BRCA2_t10_65
6	D06	D06	TGGCTTATAAAATATTAATGTGCTTCTGT	AACTATCTTCTTCAGAGGTATCTACAAC	BRCA2_t9_1
6	D07	D06	AGCTTTTTCAAAATTTCTATTTCTGTTTCA	tcattGACCTGGGCTTTGATT	CHEK2_t7_4
6	D08	D06	TCAGAAAATAATCACTCTATTAAAGTTTCTCC	TTGTTTTCACAGGAACATCAGAAAAAG	BRCA2_t10_72
6	D09	D06	GGTCACTATTTGTTGTAAGTATTTTTGTTTA	TCTTGATTTTCTATTATCCTGTCAAATTCAT	BRCA2_t11_1
6	D10	D06	GGATGAGAAAGGCAAGCCTACAT	GGGAGTTTCTCACTACTTTCCCT	CHEK2_t8_1
6	D11	D06	CCGCCATACCCAGCCCTAT	TTAGTGTACTGTTCTGGGCTGCT	IDE_t22_1
6	D12	D06	AGTAGCAGTGCAGAAAGCAAAAG	GGTTCTCCACCTCTGCTATCAAT	MME_t4_2
11	D01	D06	CCCCTCAGATGTTATTTTCCAAGC	CTTTAAGATAGTCATCTGGTTTTCAGG	BRCA2_t10_25
6	E01	E06	ACAGTGATACTGACTTTCAATCCC	GGAATATTTTTGGTTAATTCAACATCAGATTCATA	BRCA2_t10_6
6	E02	E06	ACAAAAACAGCCCCAGAAATACG	TGTTGGTGTTTTTCTTCTTCCAGT	PALB2_t2_1
6	E03	E06	TGGTCCAAACTTTTCATTTCTGCTTT	AGTACAGTCTTTAGTTGGGGTGG	BRCA2_t25_1
6	E04	E06	TACTACAGGCAGACCAACCAAAG	TGGAGTTGTTTTTGTTAAACTGATGA	BRCA2_t13_4
6	E05	E06	CACCTGCATTTAGGATAGCCAGT	CTCTGAATCATCCAATGCCTCGT	BRCA2_t10_62
6	E06	E06	AAAGCAGCATATAAAAATGACTCTAGG	TGTTCAGAGAGCTTGATTTCCTT	BRCA2_t10_15
6	E07	E06	CACTTCTTCCATTGCATCTTTCTCA	TTTGTCGCTGCTAACTGTATGTT	BRCA2_t22_1
6	E08	E06	ACAGGTGATCATAAGGTCCACAG	ACACGTAAGTGAATGAATTAGCTACAA	ATRIP_t2_1
6	E09	E06	TGCTTTAATTGTGTGAGTGGGTTG	CACTACAGGTTGGGAACGATAGA	MME_t16_1
6	E10	E06	AGCCTTATTCACTAAAATTCAGGAGG	AAGAATACCCTAGATACTAAAAAATAAAGTCAA	BRCA2_t19_3
6	E11	E06	CAAAAACAGTACACAAGGCATTTTT	AAGGAGGTGCTCATTCCTTGA	ATG2B_t17_1
6	E12	E06	AAATCTCAATTCACCTAAAAACCTTCA	AGCAGAGCTTCTTTAAGCTGACC	WDR17_t19_1
11	E01	E06	TTAACCACACCCTTAAGATGAGC	CTGAGCTTGTTTCTTATCATTCAACA	BRCA2_t21_3
6	F01	F06	TGGAAATTAGGAAGGCCATGGAA	TCTGGATTTATAATCATTTTGTTAGTAAGGT	BRCA2_t21_10
6	F02	F06	TTTTTAAAGTGAATATTTTTAAGGCAGTTCT	CAGAGGAAAAGGTCTAGGGTCAG	BRCA2_t18_1
6	F03	F06	TGTTGATAAGAGAAACCCAGAGCA	ACCAACTGTTGTTTGTCTTGTTG	BRCA2_t10_71
6	F04	F06	AGATAATCAAAAGAAACTGAGCAAGCC	TCTGCCTTTTGGCTAGGTGTTAAAT	BRCA2_t10_22
6	F05	F06	TGCAAATGAGCAATTATGTTTGCATAG	ACTCCTAATGAAATCAAATGAAGAGCA	MME_t4_1
6	F06	F06	TGGAAAAGAATCAAGATGTATGTGCT	TCTTCAGAGTCTGGATTGACAGTTAT	BRCA2_t10_8
6	F07	F06	AGGACTCCTTATGTCCAAATTTAATTGAT	TTAGTTCTGATTTTTGGTCTTTCGG	BRCA2_t9_10
6	F08	F06	TTCAAACAGTACTATAGCTGAAAATGAC	TCAACATTCTTCAATACTGGCTCAA	BRCA2_t10_57
6	F09	F06	ATAGCTGCAAAGACCACATTGGA	CACATTCATCAGCGTTTGCTTCA	BRCA2_t9_2
6	F10	F06	ACAAGAGAAATACTGAAAATGAAGATAACA	AGTGTTTCCCTCCTTCATAAACTGG	BRCA2_t10_43
6	F11	F06	TCCTGGACTTGACCTAAATCACAA	CCAACAGTCTATTATAAACAATGAAAGTGA	MME_t20_4
6	F12	F06	AAAAATGAAAACTCTTAAACTAAATTTGTGC	TGTTGCCTCGCTTCACAG	CTNNA1_t3_1
6	G01	G06	TGTTAATAAAAATAAAACTTAACAATTTTCCCCTT	GCAAAGGTATAACGCTATTGTCAAATTC	BRCA2_t5_1
6	G02	G06	TGGGTTTTGATGTGTAACTTGTCAT	TGAATGAAATGTCACTGATTCTTTCTTAAAT	PALB2_t4_21
6	G03	G06	CCAAAACACAAATCTAAGAGTAATCCA	CGTTTACACAAGTCAAGTCTGTTTCA	BRCA2_t10_9
6	G04	G06	AGTCATTGAAAATTCAGCCTTAGC	TGTCATTTTCAGCTATAGTACTGTTTG	BRCA2_t10_55
6	G05	G06	GGTTGTAGTTACAGCTACTTTTAGAAACA	GCTTTGTAGAAGAAATGTTATATGTTGAACTG	FAM175A_t9_5
6	G06	G06	ACAAACTGCACATACATCCCTGA	TGGGGGGAAATTTTTTAGGATCTG	BRCA1_t7_1
6	G07	G06	AAAATACCGAAAGACCAAAAATCAGAAC	AAAAAACACAGAAGGAATCGTCATC	BRCA2_t9_12
6	G08	G06	TCTCTGAACATAACATTAAGAAGAGCA	ACTATTTTTACAATCAGAAACAACTACACT	BRCA2_t10_20
6	G09	G06	GGAAGTTGCGAAAGCTCAAGAAG	TTCTGGTTTCTGATCAAAGAAATTTACA	BRCA2_t10_46
6	G10	G06	AAACATTGATGGACATGGCTCTG	ACTGAAAGGCAAAAATTCATCACACA	BRCA2_t13_5
6	G11	G06	AGTCTGTCTGGACATAAACAAGCA	GACTCTCATTTGCTGGCTGGAC	PALB2_t13_1
6	G12	G06	TTACCTTCCAAGAGTTTTTGACATGA	TGAATAAATTTTAGAATCAGTGATCGCCTC	CHEK2_t5_2
11	G01	G06	GGAGAACCTCTACTCAAACTGTTACC	TGTCCAAGAAGCACCTAAAGCAA	MME_t5_2
6	H01	H06	TTGTTTCCTAGGCACAATAAAAGAT	TGTTCATTTATAAAAACGAGACTTTTC	BRCA2_t12_5
6	H02	H06	ACTCACCTGCAATAAGTTGCCTT	GCATTGTACCTGCCACAGTAGAT	BRCA1_t7_4
6	H03	H06	GGACAGCACGTATGTCATTAGCA	TTTTACTTACTCTGTATGCTTGACCA	MME_t19_1
6	H04	H06	TGTTTAGGTTTATTGCATTCTTCTGTGA	ACTGTAGTTTTTCCTTATTACATTTTGCT	BRCA2_t10_2
6	H05	H06	GAGCTTTCGCCAAATTCAGCTAT	TGCCAGTAAATTGTAACATTCGTC	BRCA2_t24_7
6	H06	H06	GCATGTCTAACAGCTATTCCTACCAT	TACAGTTTGTGGGTATGCATTTG	BRCA2_t10_58
6	H07	H06	TTTCCTACTGTGGTTGCTTCCAA	ATTTGCCTTTTGAGTATTCTTTCTACA	BRCA1_t3_1
6	H08	H06	AGACATAAAATGCTTCTAGATAAGCTAAC	AGGTGCAGTGTGTCCATTAAGAA	WDR17_t13_1
6	H09	H06	TTCAGAAAACTTACTGCTTCTTTGATAA	TCAAGTCGTTTATTTGGAAGTTGTT	IDE_t16_2
6	H10	H06	aCACAACATAAATAATGAAACCTCCAA	AGAACAAAAATGTTCCATTGCCAGA	IDE_t6_1
6	H11	H06	TTTTTATTCTCAGTTATTCAGTGACTTGT	AGTGTTTTTGCAGCTGTGTCATC	BRCA2_t17_1
6	H12	H06	GGATGGATGCCGAGACAAAAAAG	TGGAAACACTACATATTGAAGGAGC	MME_t13_2

Appendix 3

Publication: Germline *CDH1* mutations in bilateral lobular carcinoma *in situ*.

British Journal of Cancer 2014

DOI:10.1038/bjc.2013.792

Received 9 November 2013; Revised 18 November 2013; Accepted 20 November 2013

http://www.nature.com/bjc/journal/v110/n4/full/bjc2013792a.html

British Journal of Cancer

Appendix 4

Publication: Genetic Predisposition to *In Situ* and Invasive Lobular Carcinoma of the Breast

PLOS Genetics 2014

DOI:10.1371/journal.pgen.1004285

Received October 4, 2013; Accepted February 17, 2014; Published April 17, 2014

http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004285

PLOS Genetics

Appendix 5

Publication: Genetic predisposition to ductal carcinoma *in situ* of the breast

Breast Cancer Research 2016

DOI: 10.1186/s13058-016-0675-7

Received: 1 September 2015; Accepted: 6 January 2016; Published: 17 February 2016

https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-016-0675-7

Breast Cancer Research