University of Dundee

MASTER OF SCIENCE

Understanding low-penetrance genetic risk for breast cancer

Merrick, Christopher Brian

Award date:
2013

Awarding institution:
University of Dundee

MASTER OF SCIENCE

Understanding low-penetrance genetic risk for breast cancer

Christopher Brian Merrick

2013

University of Dundee

"Understanding Low-Penetrance Genetic Risk for Breast Cancer"

Christopher Brian Merrick

MSc by Research in Surgery and Oncology

College of Medicine, Dentistry and Nursing

University of Dundee

June 2013

Table of Contents

List of Tables

List of Figures

List of Equations

Acknowledgements

I would first like to thank Dr Jonathan Berg for his extensive support and guidance throughout this project and during authoring of this thesis.

Thanks must also go to Dr Roger Tavendale for his knowledge and guidance during the lab based elements of the project; to Emma Gellatly for her assistance in using the BOADICEA web program; to Professor Andrew Evans, Dr Sarah Vinnicombe and Patsy Whelehan for their interpretation of mammographic imaging; to Dr Lee Baker for his advice in the statistical analysis of the data and to Proferssor Rami Aboud and the rest of the thesis monitoring committee for their ongoing input and encouragement.

Finally I would wish to thank Hannah-Leigh Gray for her motivational words and unrestricted emotional support.

Declaration

“I declare that the content of this project report is my own work and has not previously been submitted for any other assessment. The report is written in my own words and conforms to the University of Dundee’s Policy on plagiarism and academic dishonesty. Unless otherwise indicated, I have consulted all of the references cited in this report.”

Signed _____

Date _____

List of Abbreviations

ANKRD16	Ankyrin repeat domain 16
ANOVA	Analysis of variance analysis
ATM	Ataxia telangiectasia mutated
AUC	Area under curve
AUROC	Area under receiving operating characteristic curve
BCDDP	Breast Cancer Detection Demonstration Project
BCLC	Breast Cancer Linkage Consortium
BCRAT	Breast Cancer Risk Estimation Tool
BI-RADS	Breast Imaging-Reporting and Data System
BMI	Body mass index
BOADICEA	Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm
BRCA1	Breast cancer susceptibility gene 1
BRCA2	Breast cancer susceptibility gene 2
CASH	Cancer and Steroid Hormone Study
CASP8	Cysteine-aspartic acid protease 8
CCND1	Cyclin D1
CDCV	Common-disease common-variant
CDNK2A/B	Cyclin dependent kinase inhibitor 2A/B
CHEK2	Checkpoint kinase 2
COX11	Cytochrome C assembly protein 11
DCIS	Ductal carcinoma in situ
DF	Degrees of Freedom
dNDP	Deoxyribonucleotide di-phosphate
dNTP	Deoxyribonucleotide tri-phosphate
ER	Oestrogen receptor
ESR1	Oestrogen receptor 1
FBXO18	F-box protein helicase 18
FGF10	Fibroblast growth factor 10
FGFR2	Fibroblast growth factor receptor 2
GH1	Growth hormone 1
GS:3D	Generation Scotland DNA Donor Databank
GWAS	Genome Wide Association Study
HER2	Human epidermal growth factor receptor 2
HIP	Health Insurance Plan
HSD	Honestly significant difference
HWE	Hardy-Weinberg equilibrium

IBIS	International Breast Cancer Intervention Study
IGFBP3	Insulin growth factor binding protein 3
IHC	Immunohistochemistry
LD	Linkage disequilibrium
LOH	Loss of heterozygosity
LSP1	Lymphocyte specific protein 1
MAF	Mean allele frequency
MAP2K/Mek	Mitogen-activated protein kinase kinase
MAP3K/Raf	Mitogen-activated protein kinase kinase kinase
MAPK/Erk	Mitogen-activated protein kinase
MAPKAPK2	Mitogen-activated protein kinase-activated protein kinase 2
Mdn2	Murine double minute oncoprotein
MMST	Malmö mammographic screening trial
MRI	Magnetic resonance imaging
MRPS30	Mitochondrial ribosomal protein S30
MSP	Methylation specific polymerase chain reaction
MTOR	Mammalian target of rapamycin
MYEOV	Myeloma overexpressed gene
NEK10	Never in mitosis related kinase 10
NHS	National Health Service
NICE	National Institute of Health and Care Excellence
NIMA	Never-in mitosis A
NOTCH2	Notch homolog 2
OR	Odds ratio
PCR	Polymerase chain reaction
PDCP9	Programmed cell death protein 9
PI3K	Phosphoinositide 3-kinase
PR	Progesterone receptor
PTEN	Phosphatase and tensin homolog
Q-Q	Quantile-quantile
RAF	Risk allele frequency
Rb	Retinoblastoma protein
ROC	Receiving operating characteristic
RR	Relative risk
SAP	Shrimp alkaline phosphatase
SLC4A7	Solute carrier family 4, sodium bicarbonate co-transporter, member 7
SNP	Single nucleotide polymorphism
STXBP4	Syntaxin binding protein 4

TNM	Tumour, node, metastasis
TNRC9	Trinucleotide repeat-containing 9
TOM1L1	Target of myb1-like1
TOX3	TOX high mobility group box family member 3
TP53	Tumour protein 53
UTR	Un-translated region
UV	Ultraviolet
VBDM	Visual breast density measurement
WHO	World Health Organisation
ZM1Z1	Zinc finger MIZ-domain containing 1
ZNF365	Zinc finger protein 365

Summary

National screening for breast cancer using mammographic imaging, while shown to decrease breast cancer mortality, is associated with over-diagnosis and over-treatment. NICE has produced guidelines on the management of patients at increased risk of breast cancer due to a strong family history and/or mutations in rare, high-penetrance breast cancer genes such as *BRCA1* and *BRCA2*. In recent years, GWASs have identified a number of more common low-penetrance susceptibility loci for breast cancer. Although each individual locus confers a relatively lower increase in risk, it is has been shown that when combined under a log-additive model they provide a modest level of risk discrimination in European populations. Genotyping for 18 of these loci was carried out in 2,301 Scottish women (870 women with breast cancer, 385 women with a strong family history and 1046 population controls), using a single iPLEX™ Assay as part of the MassARRAY® System by Sequenom®. Polygenic risk across 18 loci was found to follow a log-normal distribution with a mean close to zero in the Scottish population, with higher means for those with a family history and for those with breast cancer. The discriminatory accuracy of the polygenic risk profile was shown by an AUROC = 0.602, which is consistent with other risk estimation models. Polygenic risk was not found to correlate with other established risk factors such as breast tissue density or family history risk as determined by the BOADICEA risk estimation tool. However, there were stronger associations of polygenic risk in both ER-positive disease and for those diagnosed at a younger age. Further research involving a larger polygenic risk profile may yet show stronger discriminatory accuracy, especially if used in conjunction with breast tissue density and other established risk estimation tools. In conclusion, this research has provided further evidence to support the use of genotype data in breast cancer risk discrimination at a population level.

1. Introduction

1.1 Breast Cancer in Scotland

In Scotland, as in the rest of the UK, breast cancer remains the most commonly diagnosed cancer in women, accounting for 28.9% of all female cancer cases in 2010 (1). In this year there were 4,457 newly diagnosed cases, an increase of 12.0% since 2000. In accordance with the increased incidence, lifetime risk of breast cancer has increased from 10.3% in the years 2003-2007 to 10.6% in the years 2005-2009. This equates to an estimated 1 in 9 Scottish women developing breast cancer in their lifetime. As with many other cancers, risk is highest in older age groups, with around 81% of breast cancers diagnosed in those over the age of 50. From the years 2003-2007, the 5-year survival rate relative to the population for women diagnosed with breast cancer across all ages was 85.9%. Despite this relatively high survival rate compared to other cancers, it remains the second highest cause of cancer death in Scottish women, no doubt due in part to its high incidence. In 2010 there were 1,032 recorded deaths due to breast cancer in Scottish women.

1.2 Prognostic Factors for Breast Cancer Outcome

A number of pathological factors have been found to influence an individual's prognosis of breast cancer. These include TNM (tumour, node and metastasis) staging, histological grading and typing, mitotic figure counts, hormone receptor status and human epidermal growth factor receptor 2 (HER2) status (2). Hormone receptor status refers whether or not oestrogen receptors (ER) and/or progesterone receptors (PR) are found to be present within the breast tumour tissue.

The effects of joint hormone receptor status on prognosis of breast cancer were investigated using data from 155,175 women over the age of 30 diagnosed with primary breast cancer from 1990 to 2001 (3). This was performed using a Cox proportional hazards model within categories of age of diagnosis, year of diagnosis, patient's race/ethnicity, histological tumour type, stage, grade, size and axillary lymph node status. Increased breast cancer mortality was demonstrated across all age ranges when compared to joint positive disease (ER+/PR+), with PR-negative disease (ER+/PR-) having a 1.2-1.5 fold increase, ER-negative disease (ER-/PR+) having a 1.5-2.1 fold increase and joint negative disease (ER-/PR-) having a 2.1-2.6 fold increase in mortality. Increased risk of mortality for hormone receptor negative disease was found to be largely independent of demographic and pathological characteristics with the exception of tumour grading. Women with low grade ER-/PR- tumours had a 2.6-3.1 fold increased risk of mortality, where has those with high grade ER-/PR- tumours had a 2.1-2.3 fold increase when compared to patients with ER+/PR+ tumours of same grading.

1.3 Breast Cancer Screening

The principles that are needed for adequate and appropriate disease screening are detailed in the 1968 World Health Organisation (WHO) report: "Principles and Practices of Screening for Disease" (4). To summarise, screening must be able to offer suitable identification of disease in which there is a recognisable early stage that if acted upon can improve disease prognosis than if it otherwise went undetected. In addition, it must also be cost effective and acceptable to the population. The WHO report applied these principles to breast cancer and hypothesised that prognosis of breast cancer could be improved through population screening using mammographic imaging and suggested the need for detailed investigation.

1.3.1 Early Trials of Breast Cancer Screening

The first evidence of the benefits of mammography was demonstrated in a study started in 1963 by the Health Insurance Plan of Greater New York (HIP) (5). The HIP study aimed to determine the value of periodic screening using clinical examination and mammography in reducing breast cancer mortality using control and study groups of around 31,000 women each aged between 40 and 64 years at the time of enrolment. The study group was invited to a total of four annual screening rounds. At ten years follow-up since study entry the mortality in the study group was shown to be reduced by around 30% compared to the control group (6).

The benefits of repeated mammography alone as a method of reducing breast cancer mortality was later investigated in the Malmö Mammographic Screening Trial (MMST) (7). The study group comprised of 21,088 women from the city of Malmö aged between 45 and 69 who were invited for mammographic screening every 18-24 months to complete a total of five rounds of screening, resulting in a mean follow up of 8.8 years. The control group comprised of 21,195 age matched women, also from Malmö, who received no invitation for screening. The number of breast cancer diagnoses was higher in the study group with 588 cases diagnosed compared to 447 cases in the control group. Across all ages there was found to be no significant difference in mortality (63 and 66 deaths in the study and control groups respectively giving a relative risk (RR) = 0.96, 95% CI = 0.68-1.35). In those aged 55 and above there was a reduction in breast cancer mortality of 20% (35 and 44 deaths in the study and control groups respectively) although again this was not statistically significant (RR = 0.79, 95% CI = 0.51-1.24).

In Sweden, a number of further randomised trials were undertaken in the provinces of Kopparberg and Ostergötland (8) (known collectively as the Swedish two county trial) and the cities of Stockholm and Gothenburg (9, 10).

An overview of these Swedish randomised trials in addition to the Malmö trial aimed to more thoroughly evaluate the efficacy of breast cancer screening (11). This included a total of 282,777 women who were followed for 5-13 years and showed a statistically significant 24% (95% CI = 13-34%) reduction in breast cancer mortality among those invited to screening using mammography compare to those who were not invited. Reduction in breast cancer mortality was largest in those aged from 50-69 years at 29% with a non-significant 13% reduction observed in those aged from 40-49 years.

The first randomised controlled trial in the UK was carried out in Edinburgh, one of several centres initially included in the UK Seven-year Trial of Breast Screening (12). The Edinburgh Trial aimed to recruit 65,000 women aged 45-64, with the intervention group being offered invitation to screening that consisted of clinical examination and mammography alternate years for 7 years. Results after 10-years of follow up demonstrated a non-significant 18% reduction in breast cancer mortality in the 22,944 women in the intervention group compared to the 21,344 women in the control group (RR = 0.82, 95% CI = 0.61-1.11) (13). Further follow-up at 14-years showed a greater reduction in breast cancer mortality once corrected for socioeconomic status (RR = 0·79, 95% CI = 0·60-1·02), which became significant once deaths after diagnoses made 3 years after the end of the trial period were discounted (RR = 0·71, 95% CI = 0·53-0·95) (14).

1.3.2 UK National Screening Programme

The National Health Service (NHS) Breast Cancer Screening Programme first began in 1988, after recommendations made by Professor Sir Patrick Forest in his 1986 Breast Cancer Screening Report (15). This report focused on preliminary data from the Swedish and Edinburgh screening trials. The report concluded that the optimum age for women to be invited for screening is 50-64 due to poor response rates in those aged 65 and over, although those women

may still be offered screening on demand. Evidence for the optimum screening interval was said to be insufficient, with a suggestion for an initial interval of 3 years, whilst kept under review. Currently, the NHS Breast Cancer Screening Programme invites all women in the UK aged 50-70 for mammography every 3 years.

1.3.3 Reviews of Breast Cancer Screening

A review of eight trials of breast cancer screening, was conducted by the Cochrane Breast Cancer Group (16) and are summarised in Table 1. The Edinburgh trial was excluded from analysis due to poor randomisation, leaving seven trials that included around 600,000 women. The authors subsequently classified three trials as being adequately randomised with the remaining four trials as sub-optimally randomised. The outcomes of each trial are summarised in Table 2 and Table 3 at approximately 7 and 13 years follow-up respectively. The review found that the effects on breast cancer mortality at 13 years follow-up were RR = 0.81 (95% CI = 0.74-0.87) in the combined seven trials. The authors however found that breast cancer mortality was an unreliable outcome, with no effect shown on all cancer mortality at 10 years (RR = 1.02, 95% CI = 0.95-1.10) or all-cause mortality at 13 years (RR 0.99, 95% CI = 0.95-1.03) in the adequately randomised trials. Additionally, it was found that the screened groups had considerably higher numbers of lumpectomies and mastectomies (RR = 1.31, 95% CI = 1.22-1.42) in the two trials that measured this outcome. The authors concluded that there is a relative risk reduction of breast cancer mortality of 15%, which equates to an absolute risk reduction of 0.05%. However screening led to a 30% increase in relative risk of over-diagnosis and over-treatment, which is comparable to an absolute risk increase of 0.5%. Overall this means that for every 2000 women invited for screening over 10 years, one will have her life prolonged while 10 healthy women who would otherwise go undiagnosed will be treated.

Table 1 - Breast Cancer Screening Trials

Trial (Year)	Age	Study Size	Randomisation	Intervention Used	Outcomes Measured
New York/Hip Trial (1963)	40-64	Approx 62,000	Sub-optimal	Clinical examination and mammography annually for 4 cycles	Total mortality Breast cancer mortality Surgical interventions Radiotherapy
Malmö Trial (1976)	45-69	42,283	Adequate	Mammography every 18-24 months for 5 cycles	Total mortality Breast cancer mortality Surgical interventions Radiotherapy Chemotherapy
Malmö II Trial (1978)	40-50	17,730	Sub-optimal	As above	As above
Two-County Trial (1977)	40-74	163,008	Sub-optimal	Mammography every 2-3 years for 2-3 cycles	Total mortality Breast cancer mortality
Edinburgh Trial (1978)	45-64	54,654*	Inadequate	Mammography and physical examination alternate years for 7 years	Total mortality Breast cancer mortality Radiotherapy
Canadian Trial (1980)	40-49[a] 50-59[b]	50,713[a] 39,405[b]	Adequate	Clinical examination and mammography annually for 4-5 cycles	Total mortality Breast cancer mortality Surgical interventions
Stockholm Trial (1981)	40-64	133,06*	Sub-optimal	Mammography every 2 years for 2 cycles	Total mortality Breast cancer mortality Surgical interventions
Göteborg Trial (1982)	39-49[a] 50-59[b]	22,941[a] 25,498[b]	Sub-optimal	Mammography every 18 months for 4-5 cycles	Total mortality Breast cancer mortality
UK Age Trial (1991)	39-41	160,840	Adequate	Mammography annually for 7 cycles	Total mortality Breast Cancer mortality

[a, b] These trials were subdivided based upon age, see Table 1
***Numbers are inconsistently reported**

Table 2 - Outcomes of Breast Cancer Screening Trials, 7-years Follow-Up

Trial	Breast Cancer Mortality RR (95% CI)	Breast Cancer Mortality (age < 50) RR (95% CI)	Breast Cancer Mortality (age > 50) RR (95% CI)	Overall Mortality RR (95% CI)
New York/Hip Trial (1963)	0.65 (0.49, 0.86)	0.82 (0.26, 2.00)	0.65 (0.46, 0.92)	0.95 (0.87, 1.04)
Malmö Trial (1976)	0.96 (0.68, 1.35)	1.29 (0.74, 2.25)	0.80 (0.51, 1.24)	0.99 (0.93, 1.05)
Malmö II Trial (1978)	0.75 (0.46, 1.24)	0.75 (0.46, 1.24)	-	1.15 (0.99, 1.33)
Two-County Trial (1977)	0.66 (0.46, 0.94)[c] 0.77 (0.54, 1.10)[d]	0.79 (0.32, 1.93)[c] 1.13 (0.48, 2.67)[d]	0.63 (0.43, 0.93)[c] 0.70 (0.47, 1.04)[d]	1.03 (0.96, 1.10)[c] 0.99 (0.94, 1.05)[d]
Canadian Trial (1980)	1.36 (0.83, 2.21)[a] 0.97 (0.62, 1.52)[b]	1.36 (0.83, 2.21)[a]	0.97 (0.62, 1.52)[b]	1.02 (0.82, 1.27)[a] 1.01 (0.85, 1.20)[b]
Stockholm Trial (1981)	0.71 (0.14, 1.07)	0.80 (0.39, 1.63)	0.59 (0.36, 0.98)	0.91 (0.85, 0.99)
Göteborg Trial (1982)	0.73 (0.26, 2.00)[a] 0.90 (0.53, 1.54)[b]	0.73 (0.26, 2.00)[a]	0.90 (0.53, 1.54)[b]	1.17 (0.95, 1.43)[a] 0.93 (0.82, 1.06)[b]
UK Age Trial (1991)	0.83 (0.66, 1.04)	0.83 (0.66, 1.04)	-	0.96 (0.89, 1.04)
Total	**0.81 (0.72, 0.90)**	**0.89 (0.77, 1.04)**	**0.72 (0.62, 0.85)**	**0.99 (0.96, 1.02)**

[a, b] These trials were subdivided based upon age see Table 1
[c, d] Cities of Koppaberg and Östergötland respectively as part of Two-County Trial

Table 3 - Outcomes of Breast Cancer Screening Trials, 13-years Follow-Up

Trial	Breast Cancer Mortality RR (95% CI)	Breast Cancer Mortality (age < 50) RR (95% CI)	Breast Cancer Mortality (age > 50) RR (95% CI)	Overall Mortality RR (95% CI)
New York/Hip Trial (1963)	0.83 (0.70, 1.00)	0.78 (0.38, 1.37)	0.78 (0.60, 1.01)	0.99 (0.94, 1.05)
Malmö Trial (1976)	0.81 (0.61, 1.07)	0.52 (0.22, 1.20)	0.86 (0.64, 1.16)	0.98 (0.93, 1.04)
Two-County Trial (1977)	0.58 (0.45, 0.76)[c] 0.76 (0.61, 0.95)[d]	0.72 (0.38, 1.37)[c] 1.03 (0.58, 1.84)[d]	0.55 (0.42, 0.73)[c] 0.71 (0.56, 0.91)[d]	1.03 (0.99, 1.08)[c] 1.00 (0.96, 1.04)[d]
Canadian Trial (1980)	0.97 (0.74, 1.27)[a] 1.02 (0.78, 1.33)[b]	0.97 (0.74, 1.27)[a]	1.02 (0.78, 1.33)[b]	1.00 (0.87, 1.14)[a] 1.06 (0.96, 1.18)[b]
Stockholm Trial (1981)	0.73 (0.50, 1.06)	0.96 (0.48, 1.91)	0.64 (0.41, 1.01)	-
Göteborg Trial (1982)	0.75 (0.58, 0.97)	0.70 (0.46, 1.06)	0.83 (0.60, 1.15)	0.89 (0.83, 0.95)
UK Age Trial (1991)	0.83 (0.66, 1.04)	0.83 (0.66, 1.04)	-	0.96 (0.89, 1.04)
Total	**0.81 (0.74, 0.87)**	**0.84 (0.73, 0.96)**	**0.77 (0.69, 0.86)**	**0.99 (0.97, 1.01)**

[a, b] These trials were subdivided based upon age see Table 1
[c, d] Cities of Koppaberg and Östergötland respectively as part of Two-County Trial

More recently, an independent review of the benefits and harms of screening was undertaken by the Independent UK Panel on Breast Cancer Screening (17). The Panel examined the same trials as the Cochrane review although they did not discount the Edinburgh trial and instead focused on the UK setting. A meta-analysis of these trials found a RR = 0.80 (95% CI = 0.73-0.89) for breast cancer mortality, equating to a relative risk reduction of 20%. The Panel concluded that this is a reasonable estimate even when taking into consideration potential internal biases and relevancy of the trials. Potential estimates of over-diagnosis were taken from the trials that did not invite the control group for screening once the active trial period had finished (Malmö and Canada Trials). When expressed as a proportion of cancers diagnosed in the invited group during the screening period, excess incidence in these trials was found to be 19% (95% CI = 15-13%). Taking both of these into consideration, the Panel concluded that for every 10,000 women in the UK screened from the age of 50 to 70, 43 breast cancer deaths would be prevented and 129 cases of breast cancer would be over-diagnosed.

1.3.4 High Risk Screening

Guidelines have been produced by the National Institute of Health and Care Excellence (NICE) that can be used to help identify and classify women at increased risk of breast cancer due to either a positive family history or certain genetic factors (18). In these guidelines, women are either classified as average, moderate or high risk as summarised in Table 4. Approximately 20% of all breast cancers are accounted for by women of moderate risk, while less than 1% is accounted for by high risk women (18). Women found to be at increased risk are offered annual mammography from the age of 40. High risk women may be offered additional imaging with magnetic resonance imaging (MRI) or risk-reducing surgery such as mastectomy. However these will often rely on additional factor such as an overly dense breast pattern or a confirmed mutation within a known breast cancer gene.

Table 4 - NICE Guidelines for Breast Cancer Risk

Classification of risk is dependent upon a women's 10-year risk at age 40, lifetime risk and/or risk of mutations in the high penetrance breast cancer genes *BRCA1*, *BRCA2* or *TP53*. Risk estimates are based solely on family history and genetic data and not from additional risk factors such as BMI or parity. Several examples of a family history for each risk category are shown within the full guidelines and quick reference guide provided by NICE. This includes number and kind of relatives affected with breast or ovarian cancer and the age at which they were diagnosed.

Risk Category	10 year-risk at age 40-49	Lifetime Risk	Mutation Risk	Screening	Possible Interventions
AVERAGE	< 3%	< 17%	-	Mammography every 3 years from age 50	-
MODERATE	3-8%	17-30%	-	Annual mammography from age 40	-
HIGH	> 8%	> 30%	> 20%	Annual mammography from age 40 +/- annual MRI surveillance from age 30-39	Genetic counselling, risk-reducing surgery eg bilateral mastectomy/oopohrectomy

1.4 Family History Risk

A family history of breast cancer is well established as being a significant risk factor for development of the disease. Pharoah et al carried out a meta-analysis and systematic review of evidence for the effect of family history on the breast cancer risk (19). A total of 52 case-control studies and 22 cohort studies were pooled to provide estimates of relative risk. When compared to having no family history of breast cancer, having any relative affected gave a RR = 1.9, 95% CI = 1.7-2.0 for all ages. Unsurprisingly, the risk was found to be higher if a first degree relative was affected (RR = 2.1, 95% CI = 2.0.2.2) and lower if a second degree relative was affected (RR = 1.5, 95% CI = 1.4-1.6). Risk was found to be even higher if more than one family member was affected, with having both an

affected mother and affected sister giving a RR = 3.6 (95% CI 2.5-5.0). If either the individual was younger than 50 or affected family member was diagnosed before this age then risk is slightly higher. Conversely, the risk is slightly lower if either age is above 50. Although such an association may be partly due to shared environmental factors, data from the Swedish twin registry of over 10,000 twin pairs suggests that genetic factors may account for approximately 32% of individual variation in breast cancer susceptibility (20).

1.5 Models of Disease Inheritance

The common disease-common variant (CDCV) hypothesis proposes that the genetic variation of inheritable common diseases such as breast cancer occurs through a number of high frequency alleles ie common variants in the population (21). It is this hypothesis that formed the rationale for genome-wide association studies (GWASs), which aim to identify common variants across the genome that associate with a given trait or disease. These are typically single nucleotide changes in the genome's DNA sequence known as single nucleotide polymorphisms (SNPs). While GWASs have been effective in identifying such common variants for a number of diseases they somewhat contradict the CDCV hypothesis due to the problem of "missing heritability", ie the variants identified through GWASs are only able to explain a small proportion of inherited risk (22). Other such models proposed to explain the genetic component of risk include the infinitesimal model (23), the rare allele model (24) and the broad sense heritability model (25).

The infinitesimal model proposes that heritability is composed of a large number of common variants of small-effect across the entire range of possible allele frequencies. In this case, genetic variance can result from hundreds or thousands of individual loci. The "missing heritability" from GWASs is purported to be attributable to variants that individually infer a RR of less than 1.1 and explain far less than 1% of heritable risk (26). The term "infinitesimal" signifies

the notion that the heritability is not necessarily missing but is instead unable to be detected through the significance thresholds needed to identify risk alleles with high confidence (27). This has been demonstrated through meta-analyses of GWASs investigating height and body mass index (BMI) (28, 29). Such studies have shown that is unlikely that more than a few hundred loci will be confirmed and such loci are unlikely to explain even half of the total genetic variance.

In contrast, the rare-allele model comprises of a large number of large-effect rare variants contributing to heritability. Such rare variants have an allele frequency that is typically less than 1% and generally increase risk two-fold or higher above the population risk, with penetrance not necessarily requiring to be anywhere near 100%. The severity of disease under this model can then subsequently be modified either by other loci or by environmental factors (30). The effects of such variants however fail to explain the level of variance detected through GWASs.

The broad sense heritability model proposes that both common and rare variants are alone insufficient to explain missing heritability and instead relies on some combination of genotype, environmental and epigenetic interactions. Such interactions have been widely documented (31), with notably increasing numbers of studies demonstrating inheritance of epigenetic effects (32, 33). This model suggests that GWASs are unable to capture heterogeneity of effect sizes at a family level that would reflect these interactions and instead only detect the average effects of alleles across a whole population.

1.6 Genetic Risk

The inherited predisposition to breast cancer has been widely reported, with genetic factors falling into three broad categories, namely: rare, high-penetrance susceptibility genes; rare, medium-penetrance susceptibility genes; and common, low-penetrance susceptibility loci. These categories are summarised in Table 5 in terms of their population frequency and the increased relative risk of breast cancer caused by the high-risk variants.

Table 5 - Genetics of Breast Cancer Susceptibility

Genetic Factor	Examples	Allele Frequency in the Population	Relative Risk of Breast Cancer
Rare, high-penetrance genes	*BRCA1, BRCA2, TP53, PTEN*	Less than 0.1%	10-20 fold increase
Rare, mid-penetrance genes	*CHEK2, ATM*	Less than 0.6%	2-4 fold increase
Common, low-penetrance loci	*FGFR2, MAP3K1, TOX3, LSP1*	Ranging from 5-85%	1.10-1.25 fold increase

1.6.1 Rare, High-Penetrance Susceptibility Genes

BRCA1 and *BRCA2* are tumour suppressor genes involved in DNA double strand break repair through their interaction with RAD51 (34). The contribution to inherited breast cancer by mutations in these genes was assessed through linkage and mutation analysis of 237 families with a strong family history of breast cancer (four or more affected members), collected by the Breast Cancer Linkage Consortium (BCLC) (35). Breast cancer was found to be linked to *BRCA1* mutations in 52% (95% CI = 42-62%) of families and to *BRCA2* mutations in 32% (95% CI = 22-43%) of families. Linkage to neither gene was found in 16% (95% CI = 6-28%) of families suggesting further predisposition genes may be involved in familial risk. The penetrance of breast cancer for

BRCA1 mutations was estimated to be 49% by 50 years of age and 71% by 70 years of age. Comparatively, *BRCA2* mutations gave an estimated penetrance of 28% by age 50 and 84% by age 70.

BRCA1 and *BRCA2* mutation prevalence was examined in 617 women with breast cancer, 254 of which were diagnosed before the age of 36 with the remaining diagnosed between the ages of 36 and 45 (36). Mutations were detected in 5.9% (3.5% *BRCA1* and 2.4% *BRCA2*) of patients diagnosed before the age of 36 and in 4.1% (1.9% *BRCA1* and 2.2% *BRCA2*) of patients diagnose at a later age. Prevalence data from the BCLC was used to estimate a prevalence of *BRCA1* and *BRCA2* mutations of 1.3% and 1.5% respectively in breast cancers across all ages. The prevalence in the general population was estimated to be around 0.11% and 0.12% respectively, equating to around 1 in 450 women.

A number of other high penetrance susceptibility genes have been identified in addition to *BRCA1* and *BRCA2*. Li-Fraumeni syndrome is a disease characterised by early onset sarcomas, brain and breast tumours and is caused by mutations in the tumour protein 53 gene (*TP53*) (37). Mutations in the phosphatase and tensin homolog gene (*PTEN*), a tumour suppressor gene, causes a similar cancer syndrome known as Cowden’s disease of which breast cancer is a feature (38). Despite their dramatic increase in breast cancer risk, such high penetrance mutations only account for around 25% of familial breast cancer risk, with the majority due to BRCA1 and BRCA2 mutations (8% each) (39).

1.6.2 Rare, Mid-Penetrance Susceptibility Genes

Mutations in the cell-cycle checkpoint kinase 2 gene (*CHEK2*) have been reported in families with Li-Fraumeni Syndrome (40). More specifically, a 1100delC mutation within exon 10 has been reported in Finnish families with a phenotype suggestive of Li-Fraumeni Syndrome that includes breast cancer (41). The CHEK2 protein is directly involved in the p53 pathway, by phosphorylating p53 in response to DNA damage, which leads to G1 phase cell cycle arrest (42).

A cohort study of 1,035 breast cancer cases and 1,885 population matched controls revealed a frequency of the 1100delC of 2.0% and 1.4% in the case and control populations respectively (43). Although this difference was not found to be statistically significant (P = 0.182), there was a significantly higher frequency found in 358 breast cancer patients with a positive family history (3.1%) giving an odds ratio (OR) = 2.27 (95% CI = 1.11-4.63, P = 0.021) compared with population controls. Further analysis of the 1100delC mutation in an independent set of 507 patients with familial breast cancer who do not have any *BRCA1* or *BRCA2* mutations again showed an elevated frequency (5.5%) giving an OR = 4.2 (95% CI = 2.4-7.2, P = 0.0002) when compare to controls. Additionally, decreased expression of CHEK2 protein was found in breast tumours from patients with the 1100delC mutation using tissue microarray analysis.

The ataxia-telangiectasia mutated gene (*ATM*) encodes for a protein kinase that is involved in DNA double-strand break repair through downstream phosphorylation of BRCA1, p53 and CHEK2 (44). The involvement of *ATM* in breast cancer was first reported in female relatives of patients with the autosomal recessive condition ataxia-telangiectasia, who were found to have an elevated risk of breast cancer (45). Further evidence was provided by screening 433 familial breast cancer cases and 521 controls for mutations within the *ATM*

gene, with 12 mutations discovered in the case group and 2 mutations discovered in the controls (P = 0.0047) (46). Compared to non-carriers, heterozygous carriers of *ATM* mutations were found to have an increased RR = 2.37 (95% CI = 1.51-3.78, P = 0.0003).

1.6.3 Common, Low-Penetrance Susceptibility Loci

1.6.3.1 FGFR2

The fibroblast growth receptor 2 gene (*FGR2)* was first identified as a breast cancer susceptibility gene in a three stage genome-wide association study by Easton et al (47). The first stage contained a panel of 266,722 SNPs, selected as known common variants across the entire genome (48), genotyped in 408 breast cancer cases and 400 controls from a UK population. Approximately 5% of these SNPS were then selected for the second stage where they were genotyped in a further 3,990 invasive breast cancer cases and 3,916 controls. The third and final stage examined the 30 most significant SNPs in 21,860 invasive breast cancer cases and 22,578 controls from 22 additional case-control studies, where a total of six SNPs were found to show an association that were significant at $P \leq 10^{-5}$. The SNP with the most significant association (combined $P_{trend} = 2 \times 10^{-76}$) was found to be rs2981582, which lies almost entirely within intron 2 of *FGFR2* as part of a 25kb block in linkage disequilibrium (LD). Haplotype analysis of the seven known SNPs within this LD block revealed multiple haplotypes at increased risk of breast cancer that contained the minor allele of rs2981582, suggesting that the causal variant was either rs2981582 itself or a strongly correlated variant. Fine-scale mapping of this region using an African American population of 1,253 invasive breast cancer cases and 1,245 controls revealed rs2981578 as the causal variant (49). This was made possible due to the relatively weaker linkage disequilibrium across this region in populations of African ancestry (50). When adjusted for age of diagnosis and family history, rs2981578 was found to have a per-allele OR = 1.24, (95% CI 1.04-1.47) with a risk allele frequency (RAF) = 0.49 for a European population.

The effects of rs2981578 on clinically important tumour characteristics were investigated by Garcia-Closas et al using data from a total of 20 studies that included a combined 23,039 breast cancer cases and 26,273 controls of European and Asian origin (51). There was found to be a stronger association for ER-positive breast cancer (per allele OR = 1.31, 95% CI = 1.27-1.36) than ER-negative breast cancer (per allele OR = 1.08, 95% CI = 1.03-1.14) when compared to controls with $P_{heterogeneity} = 0.013$ (after permutation analysis adjustment for multiple testing). Additionally, there were found to be stronger associations for rs2981578 with PR-positive and low grade tumours ($P_{heterogeneity} = 0.001$ and < 0.001) respectively than their opposite counterparts when compared to controls. No significant difference in association was found with tumour node status, tumour size or stage at diagnosis.

FGFR2 is a member of family of receptor tyrosine kinases (RTKs) and is involved in both normal mammary gland development (52) and tumorgenesis through its involvement in cell differentiation, division and migration, in addition to angiogenesis (53). Expression of *FGFR2* has been found to be amplified in small subsets of breast tumours. A series of 387 breast tumours was found to have increased DNA expression in 11.5% of samples (54) whereas a separate series of 103 breast tumours had amplified mRNA expression in 4% of samples (55). The role of increased expression of *FGFR2* in breast cancer development was further implicated through inhibition of FGFR signalling by using a low molecular weight compound to selectively block its tyrosine kinase activity (56). This led to down-regulation of cyclin D1 and cyclin D2, with subsequent reduction in cyclin D/cdk4 activity preventing G_1-S phase transition and the halting of cell proliferation.

A trend of increased mRNA expression of *FGFR2* was observed for rare homozygotes of rs2981578 using data from gene expression microarrays (57). This was confirmed using real-time (RT) polymerase chain reaction (PCR) data to compare rare and common homozygotes using a Wilcox signed-rank test ($P = 0.028$). This was proposed to occur through altered binding affinity for Oct-

1/Runx2 and C/EBPβ transcription factors. The effect of rs2981578 on FGFR2 protein expression in was investigated using immunohistochemistry staining in the nucleus and cytoplasm of breast tumour and surrounding normal tissue from 40 breast cancer cases (58). Cytoplasmic staining was found to be consistent throughout but despite variable levels of nuclear staining within the samples there was found to be no correlation with rs2981578, suggesting its role in breast cancer risk may be more complex than its direct effect on *FGFR2* mRNA expression.

1.6.3.2 MAP3K1

A number of other risk loci were identified through the three-stage GWAS conducted by Easton et al, one of which is the SNP rs889312 that lies within a LD block of approximately 280 kb that contains the mitogen-activated protein kinas2 kinase kinase 1 gene (*MAP3K1)* (47). This SNP gave a per allele OR = 1.13 (95% CI = 1.10-1.16), combined $P_{trend} = 7 \times 10^{-20}$ and a RAF = 0.28 based on a UK population. Garcia-Closas et al also found that rs889312 had stronger associations with ER-positive tumours than ER-negative tumours although this was not shown to be statistically significant ($P_{heterogeneity} = 0.99$). Similarly there was no significant difference in association with either PR status, tumour node status, tumour size or stage at diagnosis. Nordgard et al found expression of MAP3K1 mRNA varied significantly between both breast tumour sub-types and rs889312 genotype in a set of 112 breast tumours ($P = 5.2 \times 10^{-5}$ and 0.0045 respectively) (59).

The MAP3K1 protein is involved in the mitogen-activated protein kinase (MAPK) pathway, which downstream events includes regulation of genes involved cell proliferation and apoptosis (60). The MAPK pathway is involved in breast cancer development through both steroid hormone dependant and independent (ie ER-positive and negative) tumours. ER-positive tumours may involve the MAPK pathway either through: phosphorylation of the ER by MAPK to enhance its transcriptional efficiency; stimulation of growth factors by estradiol, which

increases MAPK levels; or activation of MAPK by membrane associated ER through a non-genomic protein cascade sequence. MAPK activation can also occur in response to growth factor stimuli such as epithelial growth factor (EGF), insulin-like growth factor 1 (IGF-1), insulin, prolactin and transforming growth factors α and β (TGF-α, -β) (60). Additionally, the MAPK pathway is linked with HER2 receptor status with activating mutations in the MAPK pathway found to be associated with HER2 positive breast tumours (61).

1.6.3.3 TOX3, 8q24, LSP1

The final three susceptibility loci identified by Easton et al's three stage GWAS were that of the TOX high mobility group box family member 3 gene (*TOX3*), 8q24 and the lymphocyte specific protein 1 gene (*LSP1*) (47). The SNP rs12443621 lies within a LD block that contains the 5' end of *TOX3* (also known as *TNRC9,* tri-nucleotide repeat containing 9) and gave a per allele OR = 1.11 (95% CI = 1.08-1.14), combined $P_{trend} = 2 \times 10^{-19}$ and a RAF = 0.46. Although *TOX3*'s function remains unknown, it has been shown that *TOX3* gene expression is higher in breast cancer cases with bone metastases than those without in a series of 107 primary breast tumours (62).

The 8q24 region contains rs13281615, which gave a per allele OR = 1.08 (95% CI = 1.05-1.11), combined $P_{trend} = 5 \times 10^{-12}$ and a RAF = 0.40 (47). This SNP is not located within or near any genes (80kb upstream and 250kb downstream) and so the basis for its association with breast cancer susceptibility remains obscure.

The SNP rs381718 is located within intron 11 of *LSP1* and gave a per allele OR = 1.07 (95% CI = 1.04-1.11), combined $P_{trend} = 3 \times 10^{-9}$ and a RAF = 0.30 (47). LSP1 is an F-actin bundling cytoskeletal protein involved in leukocyte chemotaxis and neutrophil emigration through the endothelium (63). It has been

shown that LSP1 is a major substrate of MAPK activated protein kinase 2 (MAPKAPK2) in the p38 MAPK pathway (64), with MAPKAPK2 implicated in lung tumourgenesis (65).

The SNPs rs13281615 in the 8q24 region and rs381718 within *LSP1* were also both investigated by Garcia-Closas et al for clinical and pathological associations (51). The only significant associations found were an increased association for ER positive and low grade tumours, both with rs13281615 (adjusted $P_{heterogeneity}$ = 0.038 and 0.016 respectively).

1.6.3.4 5p12

A GWAS by Stacey et al that included a total of 6,145 cases and 33,016 controls identified two further SNPs (rs4415084 and rs10941679) at 5p12 associated with an increased risk of breast cancer (66). The more common risk allele, rs44150484, gave a per allele OR = 1.16 (95% CI = 1.10-1.21) and a RAF = 0.40 based on the full data set from a European population. Most interestingly, this association of increased risk was only significant when compared to controls for ER-positive and not ER-negative breast cancers with a per allele OR = 1.23 (95% CI = 1.16-1.30) and 0.98 (95% CI= 0.88-1.10) respectively.

A potential gene that may be involved in this increased risk is the fibroblast growth factor-10 gene (*FGF10*) located 274kb distal to rs4415084 on 5p12. *FGF10* is a known oncogene with amplified mRNA levels found in approximately 10% of human breast cancers (67). FGF10 acts as a key ligand for FGFR2-B, through which it is believed to impart its effects on the regulation of cell proliferation, migration and differentiation (68). However, a recombination hotspot separates rs4415084 from *FGF10* (69) and Stacey et al reported they

were unable to detect a signal in *FGF10* that would account either the rs4415084 or rs10941679 signal (data not published) (66).

The two SNPs identified are both situated in a strong block of LD approximately 310 kb long. The only known gene to exist in this region is the mitochondrial ribosomal protein S30 gene (*MRPS30,* also known as *PDCP9*, programmed cell death protein 9) that is implicated in pro-apoptotic events (70). *MRPS30* is found to be up-regulated in infiltrating ductal carcinomas but not normal breast tissue (71) and forms part of the gene expression profile used to distinguish between ER-positive and ER-negative tumours (72).

1.6.3.5 NOTCH2, RAD51L1

Thomas et al discovered two novel risk loci in a three-stage GWAS of 9,770 breast cancer cases and 10,799 controls of mixed European descent (73). The first SNP (rs11249443) is found at 1p11.2 gave a per allele OR = 1.16 (95% CI = 1.09-1.24), combined $P_{trend} = 6.74 \times 10^{-10}$ and a RAF = 0.39 based on the combined control population. The second SNP (rs999737) is found at 14q24.1 and gave a per allele OR = 0.94 (95% CI = 0.88-0.99), combined $P_{trend} = 1.74 \times 10^{-10}$ and a RAF = 0.76.

The SNP rs11249443 lies within a large LD block that contains several pseudogenes, which in turn lies distal to the promoter of the Notch homolog 2 gene (*NOTCH2*). NOTCH signalling plays key roles in breast cancer development through its effects on cell proliferation, survival and differentiation (74). More specifically, NOTCH2 signalling is believed to function as a tumour suppressor via up regulation of PTEN or down regulation of the phosphatidylinositol 3-kinase (PI3K)/Akt/mammalian target of rapamycin (mTOR) pathway (75).

The increased risk of rs11249443 was found to be more strongly associated with ER-positive than ER-negative breast cancers with this SNP (P_{trend} = 0.001) (73). The expression of *NOTCH2* has been shown to be associated with ER status and rs11249433 genotype in a group of 108 breast tumour samples of various subtypes (76). ER positive tumours had relatively higher expression than ER negative tumours when compared to tumours with *TP53* mutations (1.38 vs 1.29, P = 0.0059). Within the ER positive sub group, heterozygotes of rs11249433 had relatively higher expression than non-risk homozygotes although this was higher still than high-risk homozygotes (1.52 and 1.11 respectively, P = 0.0062). A similar association was found in purified peripheral monocytes from 60 healthy control samples but not a total of 76 normal breast tissue samples (P = 0.015 and 0.381 respectively). No strong association was found between *NOTCH2* expression and rs11249433 in either ER negative or *TP53* mutation subgroups (P = 0.458 and 0.947 respectively).

The second SNP, rs999737, maps to a LD block 70kb in length that sits entirely within intron 12 of the RAD51-like 1 gene (*RAD51L1* also known as *RAD51B*). RAD51L1 is one of five paralogs that directly interacts with *RAD51* to catalyse key reactions involved in homologous recombination (HR) (77). Germ-line copy number variation of 14q24.1, which contains *RAD51L1*, has been found in a number of pedigrees containing Li-Fraumeni syndrome suggesting that *RAD51L1* may play a role in determining the syndrome's clinical variation of which a predisposition to breast cancer is a prominent feature (78). Unlike the other SNP identified by Thomas et al, rs999737 was not found to have any stronger associations based upon ER status (P_{trend} = 0.20).

1.6.3.6 ESR1

A multi-stage GWAS of 1,505 breast cancer cases and 1,522 controls from a Chinese population identified a novel breast cancer susceptibility locus at chromosome 6q25.1 (47). The SNP rs2046120 was subsequently validated in

an independent European study population of 1,591 cases and 1,466 controls, giving a per allele OR = 1.15 (95% CI = 1.03-1.28, P = 0.01) and a RAF = 0.34. This risk was found to be much lower than that of the Chinese population.

The SNP rs2046120 lies 180 kb upstream from the transcription start site of exon 1 of the oestrogen receptor 1 gene (*ESR1*) (79). *ESR1* encodes oestrogen receptor α (ERα), a receptor that plays a key role in the development of both pre- and post-menopausal breast cancer through regulation of oestrogen signalling (80, 81). *ESR1* expression was found to be amplified in 20.6% of breast cancers from a tissue microarray of over 2,000 breast cancer samples (82). Out of the subset of tumours that showed *ESR1* amplification, 99% of these were found to be ER positive compared to 66.6% of tumours found to be ER-positive in the subset without *ERS1* amplification (P < 0.0001). Additionally, survival following adjuvant tamoxifen monotherapy was found to be significantly longer in women with ER-positive disease with *ESR1* amplification than those without *ESR1* amplification (P = 0.023). Finally, *ESR1* amplification was noted in benign and precancerous breast lesions, suggesting that *ESR1* may have a role in the early development of a large subset of breast cancers.

1.6.3.7 CASP8

A case-control study of 999 breast cancer cases and 996 controls from the UK was used to investigate variations in the cysteine-aspartic acid protease 8 gene (*CASP8*) that might account act as low-penetrance susceptibility loci (83). Out of the four SNPs examined, only the effects of rs1045495 were able to be replicated in an independent UK population of 2,192 cases and 2,262 controls. The SNP rs1045495 was found to give be protective of breast cancer risk with a combined adjusted OR = 0.83 (95% CI = 0.74-0.94) for heterozygotes and 0.58 (95% CI = 0.39-0.88) for rare homozygotes (P_{trend} = 0.0002). Further evidence for the protective effect of rs1045495 was demonstrated using data from 16,423 cases and 17,109 controls pooled from 14 studies (84). This study gave a per

allele OR = 0.88 (95% CI = 0.84-0.92), P_{trend} = 1.1 x 10^{-7} and a RAF = 0.87 based upon a European population.

CASP8 encodes for a protein that acts as an important initiator of cell apoptosis in response to DNA damage and external cell death signalling (85). Down-regulations or absence of *CASP8* has been shown to associate with childhood brain tumours and is thought to occur through arrest of apoptosis (86, 87). The associations of *CASP8* and breast cancer were investigated using methylation specific polymerase chain reaction (MSP) techniques on four breast cancer and two non-cancer breast cell lines (88). In the four breast-cancer cell lines the promoter region of CASP8 was found to be methylated, which was not found to be the case in the two non-cancer lines. This methylation resulted in lower levels of mRNA and protein expression in the breast cancer cell lines compared to the non-cancer lines (relative level below 2.0).

1.6.3.8 2q35

A GWAS of 1,600 breast cancer cases and 11,563 controls from an Icelandic population was first to identify a susceptibility locus at 2q35 (89). The associations of the SNP rs13387042 was then replicated in five sample sets to give a total sample set of 4,554 breast cancers and 17,577 controls from various European populations. Risk was found to be confined to ER-positive breast cancers with a per allele OR = 1.22 (95% CI = 1.14-1.39), P_{trend} = 4.3 x 10^{-9}.

Associations of breast cancer risk with rs13387042 were further examined in subsets of differing hormone receptor status in 32,611 cases and 35,969 controls combined from 25 studies (90). Across all hormone receptor subtypes there was found to be a significant association with an overall per allele OR = 1.12 (95% CI = 1.09-1.15), P_{trend} = 1.0 x 10^{-19}. However there were found to be

slightly stronger associations for ER and PR positive disease when compared to their negative counterparts (per allele OR = 1.14 vs 1.09 and 1.15 vs 1.10 respectively). The causal genetic mechanism for this association is yet to be elicited as rs11387042 is located within a 90kb LD block that contains no known genes.

1.6.3.9 ZNF365

A two-stage GWAS carried out by Turnbull et al identified a further five novel breast cancer susceptibility loci using a combined 16,235 breast cancer cases and 17,120 controls from UK and European populations (91). The SNP rs10995190 located at 10q21 gave a per allele OR = 0.86 (95% CI = 0.82-0.91), $P_{trend} = 5.1 \times 10^{-15}$ and a RAF = 0.85. This association was found to only be significant with ER-positive breast cancer (OR = 0.83, 95% CI = 0.77-0.90, $P_{trend} = 4.1 \times 10^{-6}$) with little association related to ER-negative disease (OR = 0.91, 95% CI = 0.80-1.05, $P_{trend} = 0.19$).

The SNP rs10995190 is found within intron 4 of the zinc finger protein 365 gene (*ZNF655*), which encodes for at least four different protein isoforms (designated ZNF365A-D) (92). Each of these isoforms are expressed throughout different tissues in the body with only ZNF365A (also known as SU48) known to be expressed in the breast as it is ubiquitously expressed in low-levels throughout the body with significant expression in brain tissue. ZNF365A mRNA expression was demonstrated in human osteosarcoma, cervical cancer, breast epithelium and pancreatic cancer cell lines (93). Additionally, it was found to localise to the centrosome and it is speculated to contribute to the malignant transformation of cells through abnormal chromosome segregation.

1.6.3.10 11q13

Another susceptibility locus identified by Turnbull et al was that of the SNP rs614367 found at 11q13, giving a per allele OR = 0.15 (95% CI = 1.10-1.20), $P_{trend} = 3.2 \times 10^{-15}$ and a RAF = 0.15 (91). Again the association was largely exclusive to ER-positive breast cancer (OR = 1.17, 95% CI = 1.09-1.25, $P_{trend} = 2.4 \times 10^{-5}$) with no significant association to ER-negative disease (OR = 1.07, 95% CI = 0.95-1.22, P = 0.25).

Although rs614367 lies within a LD block that contains no identified genes, regions of 11q13 have been found to be amplified in 23% of samples from a panel of 389 primary breast tumours (94). Flanking this block are a number of plausible candidate genes. The myeloma overexpressed gene (*MYEOV*) has also been found to be amplified in breast tumours, along with somatic alterations of the cyclin D1 gene (*CCND1*), a gene involved within cell cycle control (95). Two oncogenic members of the fibroblast growth receptor (FGF) family (*FGF3* and *FGF4*) are involved in a number of cellular processes through direct interaction with distinct isoforms of FGFR2 (96). This suggests that there may be a possible link with the well-established *FGFR2* locus.

1.6.3.11 CDKN2A/B

The third locus identified by Turnbull et al was that of 9p21 containing the SNP rs1011970, which gave a per allele OR = 1.09 (95% CI = 1.04-1.14), $P_{trend} = 2.5 \times 10^{-8}$ and a RAF = 0.16 (91). Consistent with many other risk loci, associations were stronger with ER-positive (OR = 1.09, 95% CI = 1.01-1.17, $P_{trend} = 0.0222$) than ER-negative disease (OR = 1.00, 95% CI = 0.89-1.14, $P_{trend} = 0.94$). The SNP rs1011970 lies within a 180kb LD block that contains two separate genes that encode for cyclin-dependant kinase inhibitors (*CDKN2A* and *CDKN2B*).

CDKN2A encodes for cell-cycle regulatory proteins p14 (ARF) and p16 (INK4). Cell growth can be limited by protein p16 through the phosphorylation of the tumour suppressor retinoblastoma protein (Rb), which halts progression of the cell cycle (97). The protein p14 prevents degradation of p53 degradation induced by murine double minute oncoprotein (Mdm2) (98). Additionally, the interaction between p14 and Mdm2 has been shown to lead to accumulation of Rb (99). The expression of the p14 locus of *CDKN2A* was observed to be reduced in 26 out of 100 primary breast tumours, with the majority of these (77%) found to contain at least one genetic or epigenetic alteration (100). *CDKN2B* encodes for the tumour suppressor protein p15 (INK4B), that has been shown to act as an effector of transforming growth factor-beta (TGF-beta), potentially inducing G1-phase cycle arrest (101). Finally, there has been reported deletions of the *CDKN2A/B* locus in a number of cancers including melanomas, gliomas, lung cancers and leukaemias (102).

1.6.3.12 10q22, 10p15

The final two loci identified in the GWAS by Turnbull et al were that of 10q22 and 10p15 containing the SNPs rs704010 and rs2380205 respectively (91). The SNP rs704010 gave a per allele OR = 1.07 (95% CI = 1.03-1.11), P_{trend} = 3.7 x 10^{-9} and a RAF = 0.39. The SNP rs2380205 gave a per allele OR = 0.94 (95% CI = 0.91-0.98), P_{trend} = 4.6 x 10^{-7} and a RAF = 0.57. Unlike the other loci identified, neither SNP was found to associate more strongly for ER-positive breast cancer although no firm conclusions could be drawn due to a lack of ER-negative breast cases for this analysis.

The SNP rs704010 is located with a 20kb LD block that lies 90kb downstream from the zinc finger MIZ-domain containing gene (*ZMIZ1*). The ZMIZ1 protein (also known as ZIMP10) has been shown to act as a co-activator of the androgen receptor (AR), a receptor that plays a key role in male sexual

development and in prostate cell growth and survival through mediation of androgens (103).

The 105kb LD block in which rs2380205 is located contains the ankyrin repeat domain 16 gene (*ANKRD16*) of unknown function and the F-box protein helicase 18 gene (*FBXO18)*. *FBXO18* (also known as *FBH1*) codes for a member of the DNA helicase family of enzymes that are involved in the unwinding of nucleic acid strands, which subsequently mediates the replication, repair and recombination of DNA (104). Additionally, the functional F-box motif contained with FBXO18 protein is involved in regulation of the cell cycle by catalysing ubiquitin-mediated proteolysis (105)

1.6.3.13 NEK10, SLC4A7

Ahmed et al identified two further novel breast cancer susceptibility loci a two-stage GWAS that featured a combined 37,012 cases and 40,069 controls from 33 studies consisting of European populations (106). The locus 3p24 contained the SNP rs4976768 that gave a per allele OR = 1.11 (95% CI = 1.08-1.13), $P_{trend} = 1.4 \times 10^{-18}$ and a RAF= 0.46. Associations were stronger for ER positive (OR = 1.12, 95% CI = 1.09-1.16) than ER negative disease (OR = 1.06, 95% CI = 1.01-1.12, $P_{heterogeneity} = 0.022$). The never-in-mitosis related kinase 10 gene (*NEK10)* and the solute carrier family 4, sodium bicarbonate co-transporter, member 7 gene (*SLC4A7*) are the only two known genes found within this locus.

The protein encoded by *NEK10* is part of the never-in-mitosis A (NIMA) family of protein kinases (Nek) that are suggested to play a role in cell cycle control, with members Nek2, Nek7 and Nek9 implicated in mitosis regulation (107). Nek8 has been observed to be overexpressed in primary breast tumours when compared to normal breast tissue using quantitative real-time polymerase chain

reaction (qRT-PCR) techniques (108). Nek10 however has been shown to play a role in cellular responses to ultraviolet (UV) irradiation by forming a complex with MAP3K1 (109). This complex subsequently promoted phosphorylation and activation of MAP2K resulting in cell cycle arrest, but was only found to form when under UV irradiation and not in response to growth factors.

Decreased expression of SLC4A7, a substrate of tyrosine kinase, has been shown in breast tumour tissue and breast cancer cell lines when compared to normal breast tissue (110). SLC4A7 is thought to be responsible for the maintenance of cellular pH levels through its bicarbonate transporter, which is bound to the cell membrane. It was suggested that a loss of SLC4A7 may lead to a more acidic microenvironment that is more favourable to the growth and development of cancer cells in comparison to normal cells.

1.6.3.14 COX11

The second locus identified by Ahmed et al was that of 17q23.2 found to contain rs6504950, giving a per allele OR = 0.95 (95% CI = 0.92-0.97), P_{trend} = 1.4 x 10^{-8} and a RAF = 0.73 (106). Association was found to be exclusive to ER positive breast cancer (OR = 0.94, 95% CI = 0.91-0.97) with little association for ER negative disease (OR = 1.03, 95% CI = 0.98-1.09, $P_{heterogeneity}$ = 7.8 x 10^{-4}). The SNP is locate with the syntaxin binding protein 4 gene (*STXBP4*), which itself lies within a 300kb LD block that also contains the cytochrome C assembly protein 11 *(COX11*) and the target of myb1-like1 (*TOM1L1*) genes. Out of these only *COX11* however has been found to have altered expression in lymphoblastoid cell lines associated with rs6504950 (111). *COX11* therefore remains the most likely gene candidate in the absence of any direct associations with either breast or cancer development. The only known function of *COX11* is the role it plays in the function of the mitochondrial respiratory chain (112).

Table 6 - Allele Frequencies, Relative Risks and Associations of 18 Breast Cancer Loci

Locus	Genes Involved (possible)	Allele	Reference	Risk Allele Freq	Per Allele Relative Risk	Mechanisms Involved (possible)	Receptor Status Associations	Other Findings/ Associations
10q26	*FGFR2*	rs2981578	Udler et al. 2009	0.49	1.24	MAPK pathway	Stronger associations with ER+ breast cancer	
16q12	*TOX3*	rs12443621	Easton et al. 2007	0.46	1.21			*TOX3* expression higher in bone metastases
5p12	*(FGF10, MRPS30, PDCP9)*	rs4415084	Stacey et al. 2008	0.40	1.19	(MAPK pathway, cell apoptosis)		
1p11	*NOTCH2*	rs11249433	Thomas et al. 2009	0.39	1.16	NOTCH signaling	Stronger associations with ER+ breast cancer	*NOTCH2* loci associated with Type 2 diabetes
10q21	*ZNF365*	rs10995190	Turnbull et al. 2010	0.85	1.16	Chromosome segregation during cell cycle	Stronger associations with ER+ breast cancer	
14q24	*RAD51L*	rs999737	Thomas et al. 2009	0.76	1.15	Double-strand break repair/BRCA2		
6q25	*ESR1*	rs2046210	Zheng et al. 2009	0.34	1.15	Oestrogen signaling		*ESR1* amplified in precancerous lesions
11q13	*(MYEOV, CCND1, FGF3- 4)*	rs614367	Turnbull et al. 2010	0.15	1.15	(Cell cycle control, MAPK pathway)	Stronger associations with ER+ breast cancer	
2q33	*CASP8*	rs1045485	Cox et al. 2007	0.85	1.14	Apoptosis		
2q35		rs13387042	Milne et al. 2009	0.49	1.12		Stronger associations with ER+ and PR+ breast cancer	
5q11	*MAP3K1*	rs889312	Easton et al. 2007	0.28	1.11	MAPK Pathway		
3p24	*(NEK10, SLC4A7)*	rs4973768	Ahmed et al. 2009	0.46	1.11	(MAPK Pathway, alteration of cellular microenvironment)	Stronger associations with ER+ breast cancer	
9p21	*CDKN2A/B*	rs1011970	Turnbull et al. 2010	0.17	1.09	Cell cycle arrest, p53	Stronger associations with ER+ breast cancer	
8q24		rs13281615	Easton et al. 2007	0.40	1.08		Stronger associations with ER+ breast cancer	
11p15	*LSP1*	rs3817198	Easton et al. 2007	0.31	1.07			
10q22	*(ZMIZ1)*	rs704010	Turnbull et al. 2010	0.39	1.07			
10p15	*(ANKRD16, FBXO18)*	rs2380205	Turnbull et al. 2010	0.43	1.06	(Helicase activity, cell cycle regulation)		
17q	*COX11*	rs6504950	Ahmed et al. 2009	0.73	1.05	Mitochondrial respiratory chain	Stronger associations with ER+ breast cancer	

1.6.3.15 Newly Discovered Risk Loci

The total number of breast cancer susceptibility loci has increased dramatically since the discovery of the 18 loci summarised in Table 6. Such loci identified through GWAS include: 19q13 (rs8170 and rs2363956) (113); *TERT-CLPTM1L* (telomerase reverse transcriptase-Cleft lip and palate transmembrane protein 1-like protein) at 5p12 (rs10069690) (114); 12p11 (rs10771399), 12q24 (rs1292011) and 21q21 (rs2823093) (115); 20q11 (rs2284378) and 6q14 (rs17530068) (116).

The largest number of breast cancer susceptibility loci to date were identified through a meta-analysis of 9 GWAS including a total of 10,052 cases and 12,575 controls from a European population (117). An initial 29,807 SNPs were selected from the meta-analysis and subsequently genotyped in an independent European population of 45,290 cases and 41,880 controls combined from 41 studies in the Breast Cancer Association Consortium (BCAC). In total, 41 novel susceptibility loci were identified throughout the genome. Assuming that all of the identified loci can be combined under a multiplicative fashion, the genetic profile would define 5% of the population at a risk 2-3 times higher than that of the population with 1% of the population being at 3 times greater risk. Under NICE guidelines this would correspond approximately to the moderate and increased risk categories of risk respectively (18).

1.7 Breast Tissue Density

The appearance of breasts under radiological imaging is found to vary among women due to differences in the composition of tissue and the relative radiographic attenuation of fat, stroma and epithelium (118). Fat tissue will appear darker on a mammogram as it is more radiographically lucent than other tissue. In contrast, connective tissue and epithelium appear lighter as they

radiographically dense. This contrast of tissue densities is often referred to as mammographic density.

The association between mammographic density and increased breast cancer risk was first described by Wolfe in 1976 (119). Wolfe classified mammographic appearances of breast tissue in 5.284 women under four categories: N1, P1, P2 and DY, with DY being classified as a general increase in mammographic density when compared to the other groups. It was found that the incidence of breast cancer over 2 ½ years of follow-up was highest in the DY group (2.2%) with lower incidences in the other groups (0.1%, 0.4% and 1.7% for the N1, P1 and P2 groups respectively.

A meta-analysis of studies investigating the relationship between mammographic density and breast cancer risk included over 14,000 cases and 226,000 non-cases from a total of 42 studies (120). It was found that the quantitative measure of percentage density of breast had stronger associations of increased breast cancer risk than Wolfe classification in general non-symptomatic populations. Women with a percentage density of 75% or more were found to have a RR = 4.64 (95% CI = 3.64-5.91) when compared to those with a percentage density of less than 5% when using pre-diagnostic mammography. This increase in risk associated with increased breast tissue density was also consistent across different bands of percentage density and was found to be independent of both age and menopausal status.

Associations of mammographic density and breast cancer risk were further investigated in a Spanish population-based case-control study (121). This included 1,172 breast cancer cases and 4,688 non-case controls matched by year of entry into screening, age and geographic location. Women with a percentage density over 75% were found to have similar increased risks for either ductal carcinoma in situ (DCIS) (OR = 3.47, 95% CI = 1.46-8.27) or invasive tumours (OR = 2.95, 95% CI = 2.01-4.35) when compared to women

with a percentage density less than 10%. Additionally, no differences in increased risk were seen with either hormone receptor status or HER2 status. However, increased risk was found to be particularly higher in tumours detected during screening intervals (OR = 7.72, 95% CI = 4.02-14.81) compared to those detected through screening (OR = 2.17, 95% CI = 1.40-3.36). Overall increased risk due to percentage density was found to persist for at least seven years following last mammographic screening.

Numerous factors have been shown to influence breast density including age, body mass index (BMI), parity, menopausal status, hormone replacement therapy and alcohol consumption. However these factors are only thought to account for around 37% and 19% of the variability in breast density for pre- and post-menopausal women respectively (122). Evidence for the involvement of a genetic component in variation of breast density was demonstrated in twin studies that used data from a total of 571 monozygotic and 352 dizygotic twins from Australian and North American populations (123). Assuming a classic twin model of inheritance, these studies were able to show that heritability accounted for 60% (95% CI = 54-66%) and 63% (95% CI = 59-67%) variation of breast density in Australian and North American populations respectively. Follow-up analysis of data from the twin studies demonstrated that breast density follows a symmetrical unimodal distribution (close to normal) once adjusted for age (124). Such a distribution is consistent with other established polygenic traits, suggesting a polygenic model of inheritance with an additive effect from many low-penetrance loci.

Lindström et al carried out a meta-analysis of five GWAS consisting of a total of 4,877 women of European ancestry investigated potential involvement of SNPs in the variation of breast density (125). The previously discovered susceptibility loci at *ZNF365* (rs10995190) was found to account for a 1.8% (95% CI = 1.2-2.5%) change in breast density once adjusted for age and BMI, which was in the same direction as its association with increased breast cancer risk (ie the risk allele corresponded with higher breast density). However, this association

was calculated to only account for 0.5% of the total variability in breast density. Subsequent analysis demonstrated that the association of rs10995190 with breast cancer risk became slightly attenuated once corrected for mammographic density. A per allele OR = 0.85 (95% CI = 0.76-0.96, P = 0.008) was found prior before correction with a per allele OR = 0.90 (95% CI = 0.80-1.01, P = 0.09), suggesting that the genetic variation in *ZNF365* may influence breast cancer risk through breast density although may still act on both independently. Other known breast cancer susceptibility loci were examined in this same meta-analysis with two SNPs showing an association, namely rs2049210 in *ESR1* (P = 0.005) and rs3817198 in *LSP1* (P = 0.04). A similar analysis involving 19,895 Caucasian women from 10 countries again found an association with rs3817198 (P = 0.001) and percentage density once corrected for age, BMI, case status and menopausal status (126). Additionally, there was an association found with the previously un-investigated rs10483813 in *RAD51L* (P = 0.003). As with rs10995190, all of these associations were in the expected direction that corresponds to their association with breast cancer risk.

To further explore the shared genetic basis of breast cancer risk and percentage breast density, the most strongly associated genetic variants from the meta-analysis carried out by Lindström et al were combined to form a single risk score (127). This risk score was calculated for 3,628 breast cancer cases and 5,190 controls using the top 10% of SNPs most strongly associated with percentage density (50,899 SNPs). Woman in the highest decile of distribution of risk score were found to be at higher risk of breast cancer when compared to those in the lowest decile (OR = 1.31, 95% CI = 1.08-1.59). This suggests a shared genetic component across a large number of common loci, although it is yet to be elicited whether or not they act through shared or independent pathways.

1.8 Risk Estimation Models

Whilst NICE guidelines offer examples of family histories to aid with classification of breast cancer risk, they concede that such a method of “manual risk assessment” is only able to offer a crude estimation of breast cancer risk (18). For those women identified as being at increased risk, NICE recommends performing a thorough family history to allow for more accurate risk estimation, with computerised risk assessment models as potential aids in conjunction with careful clinical assessment.

1.8.1 Gail Model/BCRAT

The Gail Model, also known as the Breast Cancer Risk Assessment Tool (BCRAT), is one of the earliest risk estimation models currently used in clinical practice and is summarised in Table 7 along with other models. It was originally developed using case-control data from the Breast Cancer Detection Project (128), consisting of 2,852 breast cancer cases and 3,146 controls from the United States (129). The Gail model estimates an individual’s risk of breast cancer over a given time interval based upon that individual’s age and established risk factors. These include age at menarche, age at first live birth, number of previous breast biopsies (regardless if positive or negative) and the number of affected first-degree relatives, with an unconditional logistic regression model used to calculate a combined relative risk (130). Although such a model takes into account a relatively high amount of personal information, it fails to recognise additional family history data such as second-degree relatives or mutation status, both of which are recommended by NICE guidelines for manual risk assessment.

1.8.2 Claus Model

Claus et al were first to develop a model that more strongly examined the involvement of genetic factors using data from the Cancer and Steroid Hormone Study (CASH) (131). This included 4,730 US women with breast cancer aged 20-54 and 4,688 controls matched geographically and by 5-year age banding. Information regarding the occurrence of breast cancer in the individual's family was obtained using interviews. This only included mothers and sisters of the individual with daughters being excluded due only two daughters being reported in each group. POINTER software (132) was used to perform complex segregation analysis, with subsequent goodness-of-fit testing of the results providing evidence for an autosomal dominant allele with high penetrance for breast cancer risk. The predicted allele frequency in the population was found to be 0.33% with a total lifetime risk of approximately 92% for carriers. The highest proportion of cases predicted to carry the allele was found in those aged 20-29 (36%) with proportions gradually decreasing until the age of 80 and over (1%). Claus et al used these parameters to develop their model, which calculates individual's risk of breast cancer based upon the age of onset of breast cancer in first and second-degree affected relatives (133). The calculated risk provided is the absolute lifetime risk or 10-year risk up to the age of 80. While the Claus Model has a stronger appreciation for the established conventions of increased breast cancer risk due to a family history when compared to the Gail Model, it fails to both represent a multifactorial model of disease and include established high-penetrance breast cancer susceptibility genes.

1.8.3 BOADICEA Model

The development of new breast cancer risk estimation models with a stronger genetic component was made possible through the identification of the breast cancer susceptibility genes *BRCA1* and *BRCA2*. Family history data (including both breast and ovarian cancer) from 1,488 women diagnosed with breast cancer under the age of 55 was used to investigate a model that could best

explain familial breast cancer risk not attributable to either *BRCA1* or *BRCA2* (134). Segregation analysis was used to identify a hypothetical "*BRCA3*" gene, which had an estimated allele frequency of 24% (95% CI = 14-41%) and penetrance of 42% by age 70. Further segregation analysis was then applied to 156 high risk families that contained at least two breast cancer cases, at least one of which diagnosed before the age of 50 (135). The model of best fit was one that included a polygenic component with a mean of 0 and variance of 1.67, along with an estimated population frequency of 0.051% (95% CI = 0.021-0.125%) and 0.068% (95% CI = 0.033-0.141%) for *BRCA1* and *BRCA2* mutations respectively. This polygenic model was subsequently developed into what is now known as the BOADICEA model (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) (136). This model is able to estimate both breast cancer risk and that risk of *BRCA1/BRCA2* mutations based upon the age of diagnosis of breast and ovarian cancers in all relative of the individual assessed.

The original BOADICEA model was improved using data from three additional population studies, providing a combined dataset of 2,785 families with breast cancer, 301 of which contained *BRCA1* mutations and 236 contained *BRCA2* mutations (137). This extended model now allows for risks from male breast, prostate and pancreatic cancer as well as multiple cancers in a single affected family member. Additionally, the polygenic component was changed from that of constant variance to a variance that decreased with age and was found to more accurately reflect familial risk of breast cancer from epidemiological studies (138).

1.8.4 Tyrer-Cuzick/IBIS Model

The Tyrer-Cuzick model, sometimes referred to as the International Breast Intervention Study (IBIS) model, estimates breast cancer risk using both familial and personal risk factors (139). The model was developed using data from

national cancer incidence rates in the UK, published risk figures of *BRCA1* and *BRCA2* mutations (35, 140) and a Swedish population-based study that included daughters of mothers diagnosed with breast or ovarian cancer (141). Two autosomal loci are assumed for the genetic component of the model, one of which includes *BRCA1* and *BRCA2* mutation status while the other is a hypothetical low-penetrance susceptibility gene of dominant inheritance. Data from the Swedish population-based study was used to determine the hypothetical gene having a population allele frequency = 0.1139 and RR = 13.0377. An individual's risk of breast cancer is calculated based upon the probability of a given genotype across these loci, which itself is dependent upon an individual's family history of breast and ovarian cancer. This risk is then modified by the relative risks of personal factors that includes age at menarche, parity, age at menopause, BMI and presence of benign breast disease. The Tyrer-Cuzick model, like BOADICEA, can also provide estimates for the risk of a *BRCA1* or *BRCA2* mutation.

Table 7 - Risk Factors Used in Risk Estimation Models

Model	Personal Risk Factors	Family History	Mutation Status	Statistical Methodology	Absolute Risk Estimation
Gail/BCRAT	Yes	Limited - number of first degree relatives	-	Unconditional logistic regression	5-year and lifetime risk
Claus	-	Moderate – age of diagnosis, first and second-degree relatives	-	Hypothetical dominant high-penetrance gene	10-year and lifetime risk
BOADICEA	-	Extensive – age of diagnosis, all relatives, includes other cancers associated with *BRCA1/BRCA2*	*BRCA1*, *BRCA2*	Age-variable polygenic component	10-year and lifetime risk including *BRCA1/2* status
Tyrer-Cuzick/IBIS	Yes	Extensive – age of diagnosis, all relatives, includes ovarian cancer	*BRCA1, BRCA2*	Hypothetical dominant low-penetrance gene	10-year and lifetime risk including *BRCA1/2* status

1.8.5 Evaluation of Risk Estimation Models

The predictive value of the Tyrer-Cuzick model in comparison to the older Gail and Claus models was evaluated using a study population of 1,933 women undergoing mammographic screening in the UK (142). 1,217 of these women were part of the routine national screening programme while the remaining 1,933 women underwent mammography every 12-18 months due to a positive family history. A total of 64 cancers were diagnosed during the mean follow-up time of 5.27 years, 52 of which were diagnosed in the women undergoing more routine mammography with a mean follow-up of 6.39 years. Risk estimation models were evaluated depending on the ratio of expected breast cancer cases to the number of observed breast cancer cases as shown in Table 8 and Table 9. The Tyrer-Cuzick model was able to best predict the number of breast cancers in both populations, with the Gail and Claus models underestimating breast cancer risk.

Receiver operating characteristic (ROC) curves were generated to evaluate each model's accuracy at identifying individual cases from a population. The area under the ROC curve (AUROC), also known as the concordance-statistic (C-statistic) was calculated as shown in Table 10 using data from the 1,933 women with a strong family history. The Tyrer-Cuzick model was again found to perform better than the other models, while they all still showed significant discriminatory ability. Such ability however may be over-estimated due to a relatively small number of cases in a high risk population.

Table 8 - Comparison of Expected to Observed Breast Cancer Cases in a Total Study Population (n=3,150)

Model	Observed (O)	Expected (E)	E/O	95% CI
Gail	64	44.3037	0.69	0.54 – 0.90
Claus	64	48.5565	0.76	0.59 – 0.99
Tyrer-Cuzick	64	69.5653	1.09	0.85 – 1.41

Adapted data from: Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, et al. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. J Med Genet. 2003 Nov;40(11):807-14.

Table 9 - Comparison of Expected to Observed Cases of Breast Cancer in a 12-18 Month Screening Programme (n=1,933)

Model	Observed (O)	Expected (E)	E/O	95% CI
Gail	52	25.0312	0.48	0.37 – 0.64
Claus	52	29.1489	0.56	0.43 – 0.75
Tyrer-Cuzick	52	46.4621	0.89	0.62 – 1.08

Adapted data from: Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, et al. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. J Med Genet. 2003 Nov;40(11):807-14.

Table 10 - AUROC of Risk Assessment Models

Model	AUROC	95% CI
Gail	0.735	0.666 – 0.803
Claus	0.716	0.648 – 0.784
Tyrer-Cuzick	0.762	0.700 – 0.824

Adapted data from: Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, et al. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. J Med Genet. 2003 Nov;40(11):807-14.

The performance of the Gail and Tyrer-Cuzick models were further evaluated using prospective data from 1,857 women from 938 families (143). This cohort included both women at average and above average risk as determined by family history. Over a mean follow-up of 8.1 years 83 women developed breast cancer. Discrimination was assessed using AUROC, with the Tyrer-Cuzick model showing superior performance (AUROC = 69.5%, 95% CI = 63.8-75.2%) compared to the Gail model (AUROC = 63.2%, 95% CI = 57.6-68.9%).

The BOADICEA model's ability to identify the incidence of breast cancer cases and *BRCA1*/*BRCA2* mutations was evaluated using retrospective predictions in 640 women and 263 screened families (144). The ratios of observed to expected number of *BRCA1*, *BRCA2* and either *BRCA1* or *BRCA2* mutations were found to be 1.43 (95% CI = 1.05-1.90), 0.63 (95% CI = 0.34-1.08) and 1.12 (95% CI = 0.86-1.44) respectively, indicating an underestimation for incidence of *BRCA1* mutations. Discrimination ability between carriers and non-carriers as measured by AUROC was 0.83 (95% CI = 0.76-0.88). Incidence of breast cancer was also partially underestimated with a ratio of observed to expected cases of 1.41 (95% CI = 0.91-2.08). Discriminatory ability was found to be similar to other risk estimation models with an AUROC of 0.62 (95% CI = 0.52-0.73).

1.9 Expanding Risk Estimation Models

Pharoah et al used family history data from a case-based population of 1,484 breast cancer patients from the Anglian Breast Cancer Study Group (145) to identify potential models of risk that explain familial cases of breast cancer not attributable to *BRCA1* and *BRCA2* mutations (146). A polygenic model was found to have best fit and was hypothesized to follow a log-normal distribution of relative risk with a standard deviation = 1.2, with such a distribution suggested to be sufficient to discriminate between high and low risk groups. It was further proposed that if all susceptibility loci were identified, half of the

population at highest risk would account for 88% of all affected individuals. This is relatively higher than the ability of currently identified risk factors (age at menarche etc) to risk stratify women, with half of the population at highest risk accounting for only 62% of cases.

The implications of using such a polygenic approach with regards to risk prediction in breast cancer were further examined by Pharoah et al, using all 2,187 possible combinations of seven known risk loci under a log-additive model (147). They found that the distribution of relative risk in this model population followed a log-normal distribution with a mean just below zero as predicted by their polygenic model. Comparatively, the distribution was shifted to the right with a mean just above 0 for breast cancer cases. Under such a distribution, around 5% of women at the lowest risk would never meet a threshold for mammographic screening (designated as the 10-year risk at age 50 ie 2.3%), whereas around 5% of women at the highest risk would meet the threshold for screening at age 41.

The use of 10 common risk loci as a means of risk assessment in comparison and addition to previously established risk factors was examined by Wacholder et al (148). A combined cohort of 5,590 breast cancer cases and 5,998 controls aged between 50 and 79, predominantly from a US population, was used for receiving operating characteristic (ROC) curve analysis. The AUROC (also known as the C-statistic), was then used as a means of risk discrimination. A polygenic risk profile using 10 loci was found to perform similarly to a restricted Gail model, with a C-statistic of 58.9% and 58.0% respectively. When the polygenic profile was combined with the components used in the restricted Gail model there was found to be a modest increase in risk discrimination with a C-statistic of 61.8%. In addition to improved discriminatory ability, the addition of polygenic information resulted in over half of the subjects moving into a different quintile of risk, with 32.5% of individuals moving into a higher risk quintile and 20.4% moving into a lower quintile.

Darabi et al investigated the performance of an 18 locus genotype when used in combination to the Gail model along with BMI and mammographic percentage density (149). Genotyping was carried out in postmenopausal women aged 50-74 who were born in Sweden, 1,569 of which were diagnosed with breast cancer and 1,730 were healthy controls. Out of these, 1,022 cases and 868 controls had breast density measurements available. The addition of an 18 locus genotype to the Gail model was found to increase the AUC from 54.8% to 61.5%. When the 18 locus genotype was added to the Gail model already combined with percentage density and BMI measurements, the AUC increased from 60.4% to 61.9%. This would suggest that an 18 locus genotype offers more discriminatory utility than percentage density and BMI although no direct comparisons were made. Additionally, the discriminatory ability of an 18 locus genotype in isolation was not examined.

2. Hypotheses

The key question of this research was:

Can we use genetic information from low penetrance susceptibility loci to identify woman who will develop breast cancer more precisely?

To fully answer this to the best of our ability we also required asking the following:

- What is the distribution of risk from polygenic risk loci across populations?
- How effective is this polygenic risk profile at discriminating between cases and controls?
- Does it correlate with other risk factors such as family history or breast tissue density?
- Does it correlate with clinical factors such as ER status and age of diagnosis?

These questions formed the overall hypothesis of this research, namely that genotype data from low-penetrance risk loci will indeed be able to improve breast cancer risk discrimination at a population level.

Sub-hypotheses include:

- Polygenic risk across 18 loci will follow a log-normal distribution in the Scottish population;
- The mean will be close to zero within the population and will be higher in those with a strong family history and higher still in those with breast cancer;
- Discriminatory accuracy of 18 loci at the individual level will be close to that of other established risk estimation tools;
- Polygenic risk will be highest in those diagnosed at younger ages;
- Polygenic risk will be higher in those with ER positive breast cancer;
- Polygenic risk will correlate with both family history risk as determined through the BOADICEA risk estimation tool and breast tissue density.

3. Plan of Investigation

To answer these questions and test these hypotheses, 2,301 women from the Tayside population in Scotland will be genotyped across 18 loci using the MassArray® System by Sequenom®. 870 of these women have been previously diagnosed with breast cancer between the ages of 35 and 85 (case group), 385 women have a positive family history of breast cancer (increased risk group) and 1,046 women are population controls (control group).

The control group will be obtained from the Generation Scotland Donor DNA Databank (GS:3D) resource (150). This includes healthy participants between the ages of 17 and 70 without any previous diagnosis of malignancy including breast cancer. A full list of exclusion criteria from participation in the GS:3D resource is available at http://www.biomedcentral.com/1471-2350/11/166.

The relative risks from each loci will then be combined into a single genetic risk score under a log-additive model. Family history risk and breast tissue density will then be measured for women in the increased risk group in whom sufficient information is available.

4. Experimental Procedures

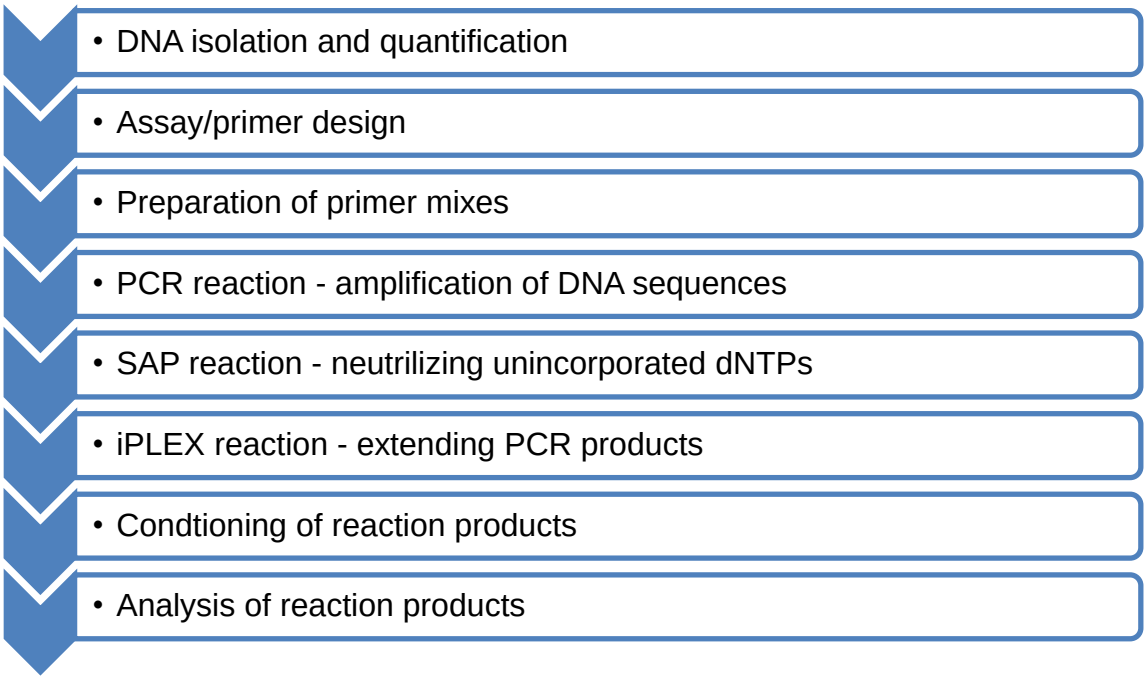
4.1 Isolation and Quantitation of DNA

Genomic DNA had been previously isolated and dried in 384-well plates for population controls and breast cancer cases at a concentration of 10ng per well. Genomic DNA was isolated from the increased risk group through the Human Genetics Service at Ninewells Hospital and stored in -20 degrees centigrade freezers. These genomic DNA samples were then diluted by a factor of 10 and then quantitated using a UV spectrophotometer at wavelengths of 260nm and 280nm. The genomic DNA was then further diluted to a concentration of 10ng/µl with 1 µl from each sample dispensed into a 384-well plate before being left to dry overnight. A number of no DNA controls were present in each plate to ensure the samples were free from contamination during the genotyping process.

4.2 Genotyping using MassARRAY® System by Sequenom®

The genotyping of samples was carried out using the MassARRAY System by Sequenom. The system uses a polymerase chain reaction (PCR) to amplify gene products containing the SNPs of interest from previously isolated and quantitated genomic DNA. The amplification products are then treated to neutralise unincorporated deoxyribonucleotide triphosphates (dNTPs). Subsequent extension of the amplification products is carried out by an iPLEX® reaction. It is this reaction that allows for allele discrimination due to mass spectroscopy of the reaction products. A flowchart of this process is shown in Figure 1.

Figure 1 - Overview of experimental procedures



4.3 Assay Design for Amplification and iPLEX™ Reactions

Assay design was initiated using MassArray Online Design Tools by Sequenom (available from https://www.mysequenom.com/Tools). Sequences containing the 18 SNPs of interested were generated using the rs Sequence Retriever application. The PreXTEND application then reformats the sequences to demark the PCR amplification primers. These primers are validated against a Golden Path genome build to ensure a unique amplification product. PCR primer design was successful for all SNPs of interest.

Assay design was then completed and further validated using MassARRAY Assay Design 4.0 Software (ASSAY Designer) by Sequenom. The ASSAY designer uses he PCR primer demarked SNP sequences from the Online Design Tools to generate PCR and extension primers (for use in amplification and iPLEX reactions respectively) in assays of a desire multiplex (in this case

18-plex). Assays are then validated for possible dimer formation and false priming potential of primers. Two SNPs (rs2380205 and rs1045485) failed assay design due to these factors.

SNPs that failed assay design were substituted with a proxy SNP based on linkage disequilibrium with values of r^2 = 2 and D' = 1. The proxies were identified using the SNP Annotation and Proxy Search (SNAP) web based program that was developed by the Broad Institute (available from hhtp://www.broadinstitute.org/mpg/snap/Idsearch.php) using data from the International HapMap Project (151). The identified proxy SNPs were rs12253826 and rs75325449 as proxies for rs2380205 and rs1045485 respectively. Subsequent PCR and extension primer assay design was successful with the inclusion of the proxy SNPs. The PCR and extension primer sequences are shown in the Appendix.

4.4 Preparing Primer Mixes

Separate primer mixes for both the amplification and iPLEX reactions are required to be prepared before each reaction can take place. The PCR primer mix consists of forward and reverse primers in equal concentrations in equal concentrations for each SNP of interest. The quantity for each forward and reverse primer is given in Table 11. The iPLEX primer mix requires differing concentrations of each extension primer. This is to compensate for signal to noise ratio when analysing the samples. The extension primers were divided into three groups based on their mass – low, medium and high. The quantity for each extension primer is given in Table 12. A total of 1500µl of both PCR and iPLEX primer mixes were prepared and then divided into 500µl aliquots before being stored in the -20 degrees centigrade freezer.

Table 11 - PCR Primer Mix

Reagent	Stock Conc (µM)	Required Conc (µM)	No. Primers	Required Volume of Each Primer (µl)	Total Volume Required 1500µl of Primer Mix (µl)
Forward Primer	50	0.5	18	15	270
Reverse Primer	50	0.5	18	15	270
Water, HPLC grade	n/a	n/a	n/a	n/a	960
Total					1500

Table 12 - iPLEX Primer Mix

Reagent	Stock Conc (µM)	Required Conc (µM)	No. Primers	Required Volume of Each Primer (µl)	Total Volume Required 1500µl of Primer Mix (µl)
Low Mass Primer	400	5	6	18.75	112.5
Medium Mass Primer	400	10	6	37.50	225
High Mass Primer	400	15	6	56.25	337.5
Water, HPLC grade	n/a	n/a	n/a	n/a	825
Total					1500

3.5 Amplifying DNA for iPLEX™ Genotyping

Genomic DNA containing the SNPs of interest was amplified using PCR. The PCR cocktail was prepared using the values in Table 13. 5µl of the PCR cocktail mix was manually pipetted into each well of each plate before they were sealed, vortexed and briefly centrifuged at 1,000RPM. The plates were then subjected to thermocycling under the conditions shown in Figure 2.

Table 13 - PCR Cocktail Mix

Reagent	Final Conc in 5µl rxn	Volume for 1 rxn (µl)	Volume for 384rxns + 20% overhang (µl)
Water, HPLC grade	N/A	2.8	1290.24
10x PCR buffer with 20mM $MgCl_2$	2mM	0.5	230.4
$MgCl_2$ (25 mM)	2mM	0.4	184.32
dNTP mix (25mM each)	500µM	0.1	46.08
Primer mix (0.5µM each)	0.1µM	1.0	460.8
PCR enzyme (5U/µl)	1 unit	0.2	92.16
DNA (10ng/µl)	10ng/rxn	Already loaded	Already loaded
Total		5.0	2304

Figure 2 - Thermocycling conditions for PCR

- 94°C for 2 minutes
- 95°C for 30 seconds
- 56°C for 30 seconds
- 72°C for 30 seconds
- 72°C for 5 minutes
- Hold at 4°C

45 cycles

4.6 Neutralising Unincorporated dNTPs (SAP Treatment)

Unincorporated dNTPs from the amplification products are neutralised using shrimp alkaline phosphatase (SAP) treatment. The SAP treatment is able to render dNTPs unavailable for future reactions by cleaving a phosphate and converting them to deoxyribonucleotide diphosphates (dNDPs). The SAP enzyme solution was prepare using the values in Table 14. 2µl of the SAP enzyme solution was manually pipetted into each well of each plate before they were sealed, vortexed and briefly centrifuged at 1,000RPM. The plate was then subjected to thermocycling under the conditions in Figure 3.

.

Table 14 - SAP Enzyme Solution

Reagent	Final Conc in 5µl rxn	Volume for 1 rxn (µl)	Volume for 384rxns + 20% overhang (µl)
Water, HPLC grade	N/A	1.53	705.024
SAP bugger (10x)	2mM	0.17	71.536
SAP Enzyme (1.7U/µl)	2mM	0.30	126.24
Total		2.00	841.6

Figure 3 - Thermocycling conditions for SAP reaction

- 37ºC for 40 minutes
- 85ºC for 5 minutes
- Hold at 4ºC

4.7 iPLEX Reaction (Extend Reaction)

The iPLEX reaction extends the reaction products with extension primers, which are then terminated by a mass modified nucleotide present in the termination mix (A, T, C and G, each of differing mass). The resultant masses of products can then be used to differentiate the alleles. The iPLEX reaction cocktail is prepared using the values in Table 15. 2µl of the iPLEX reaction cocktail was manually pipetted into each well of each plate before they were sealed, vortexed and briefly centrifuged at 1,000RPM. The plates were then subjected to thermocycling under the conditions in Figure 4.

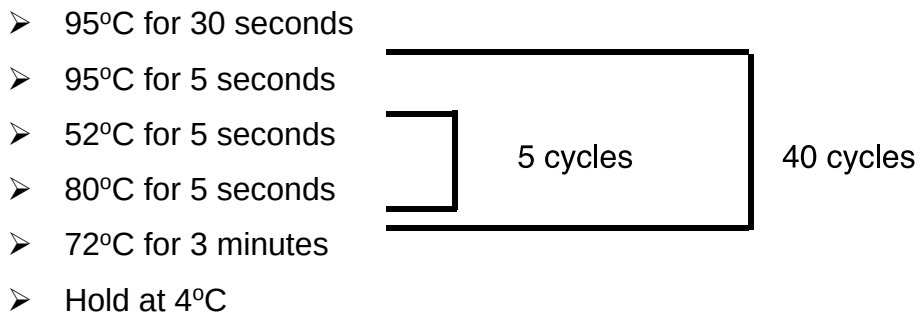
.

Table 15 - iPLEX Reaction Cocktail

Reagent	Final Conc in 5µl rxn	Volume for 1 rxn (µl)	Volume for 384rxns + 20% overhang (µl)
Water, HPLC grade	N/A	0.619	285.24
10x iPLEX buffer plus (10x)	0.222X	0.200	92.16
iPLEX termination mix	2mM	0.200	92.16
Prime mix (5µM: 10µM: 15µM)*	0.1µM	0.940	433.15
iPLEX enzyme	1 unit	0.041	18.89
Total		2.00	921.60

*For each primer of Low, Medium and High Mass respectively

Figure 4 - Thermocycling conditions for iPLEX Reaction



4.8 Conditioning the Reaction Products

The reaction products are conditioned to remove salts prior to mass spectroscopy and allele discrimination. 6mg SpectroCLEAN resin was spread into each well of a 384-well dimple plate. Excess resin was scraped clean off the plate and left to dry for 20 minutes. In the meantime 16µl of nanopure water added to each well of one of the 384-well sample plates. A single sample plate is placed upside down on the dimple plate and both are then inverted to allow the resin to fall out of the dimple plate and into the sample plate. The sample plate was sealed, vortexed and briefly centrifuged at 1,000RPM before being placed on a rotator and rotated about 360º for five minutes at room temperature. The sample plate was then centrifuged at 3200g for five minutes to complete conditioning. This was repeated for each sample plate.

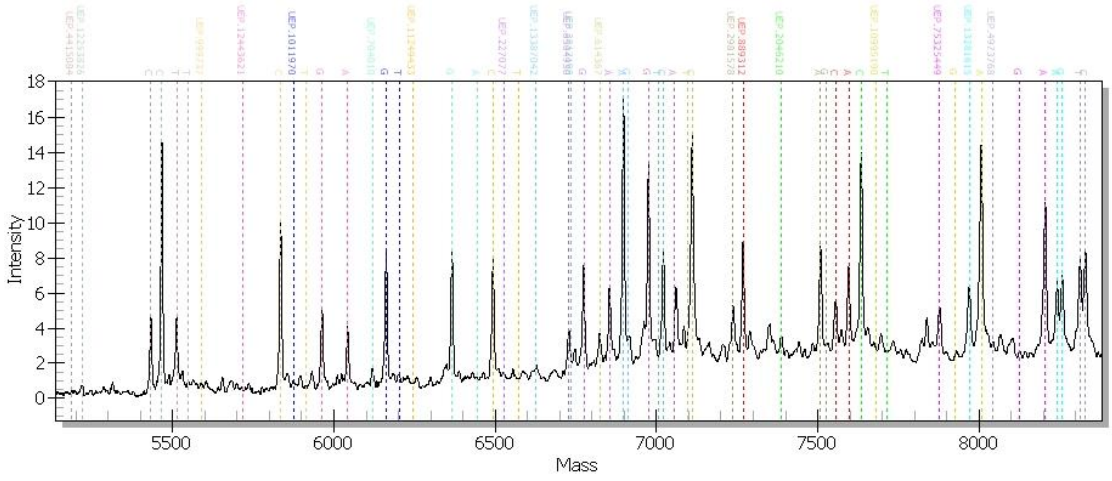
4.9 Analysis of Reaction Products

The reaction products are analysed by matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) mass spectrometry. This is used to distinguish between the alleles of the extension primers for each individual SNP of interest. Data from the assay file needs to be imported into the MassARRAY Typer Server to allow for allelic discrimination. The reaction products are first nano-dispensed into a SpectroCHIP® array using the MassARRAY Nanodispenser. The assay file created previously is applied to a virtual sample plate that represents the samples from the 384-well plate of reaction products. An experimental file is then created linking the virtual sample plate to the SpectroCHIP®. The spectra of data acquired by the MassARRAY Analyzer and transferred to the MassARRAY Typer Server. Examples of spectra acquired are shown in Figure 5. The data can then be displayed and analysed using the MassARRAY TyperAnalyzer. Call cluster plots are used for allelic discrimination. Examples of the call cluster plots produced are outlined in Figure 6.

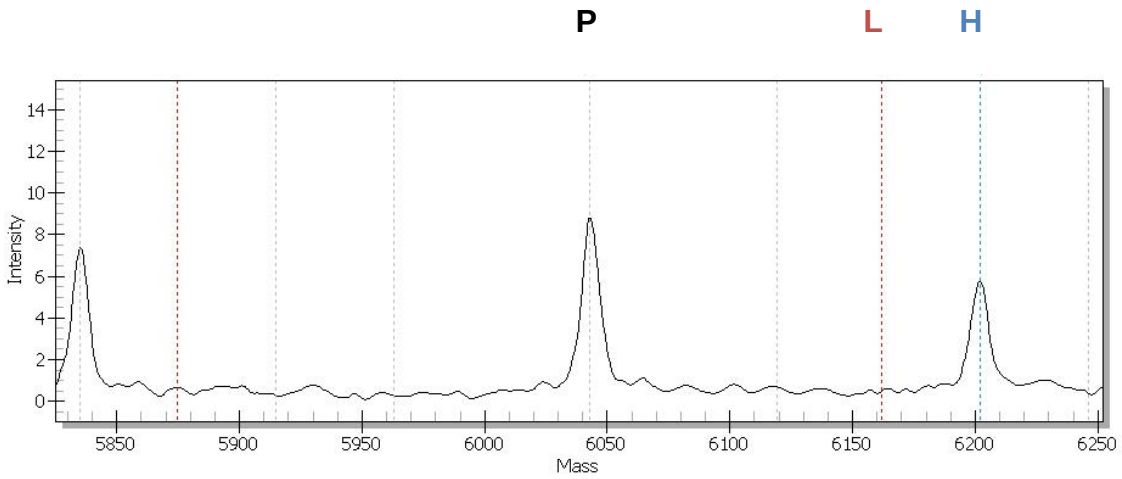
Figure 5 - Genotype Spectra

The graphs below demonstrate the genotype spectra from the MassARRAY analyser acquired from a single sample. Each graph is a plot of mass against the intensity of the signal received, with each peak corresponds to either an un-extended primer (P), low mass allele (L) or high mass allele (H). In the examples below the L and H represent G and T genotypes respectively. If it is blue it signifies a peak is present while red signifies no peak.

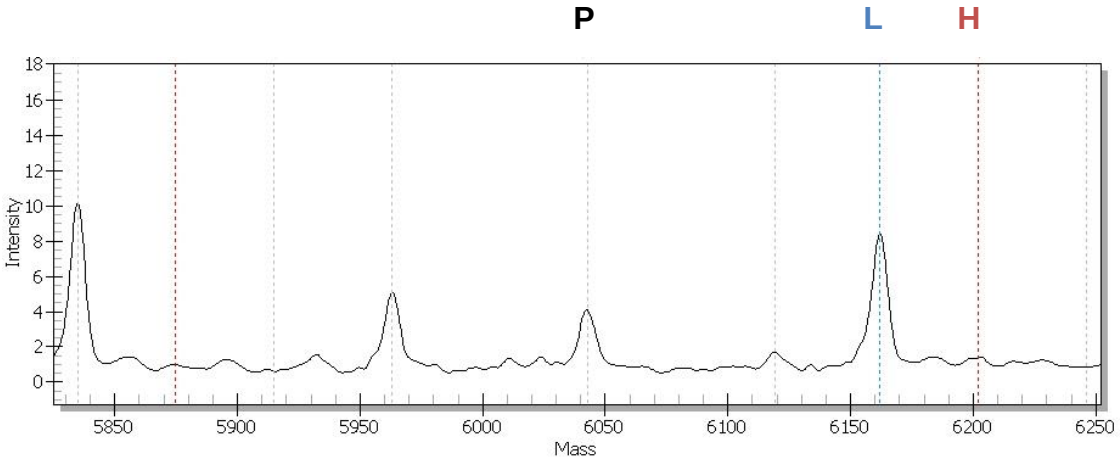
a) Complete spectra for all SNPs of interest



b) Spectra of rs1011970 showing TT genotype (a single peak at H)



c) Spectra of rs1011970 showing GG genotype (a single peak at L)



d) Spectra of rs1011970 showing GT genotype (a peak at L and H)

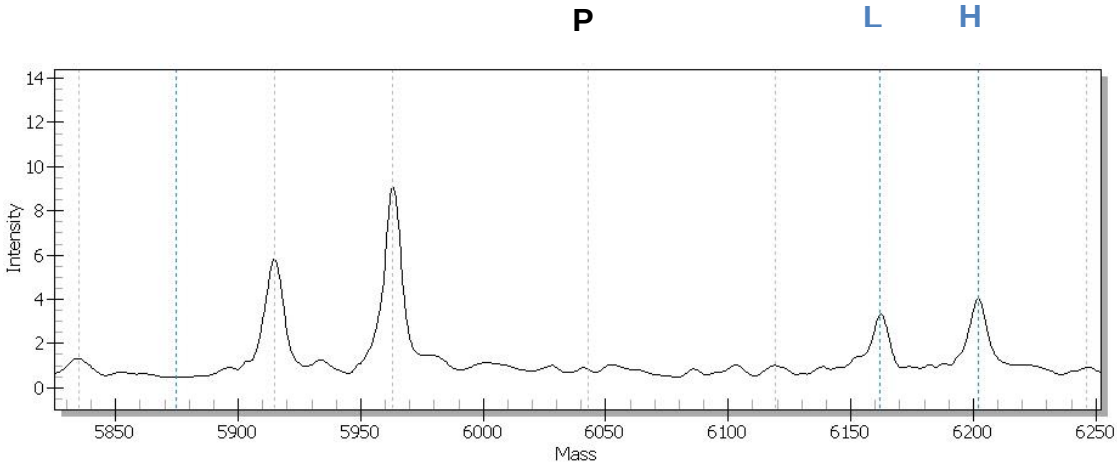
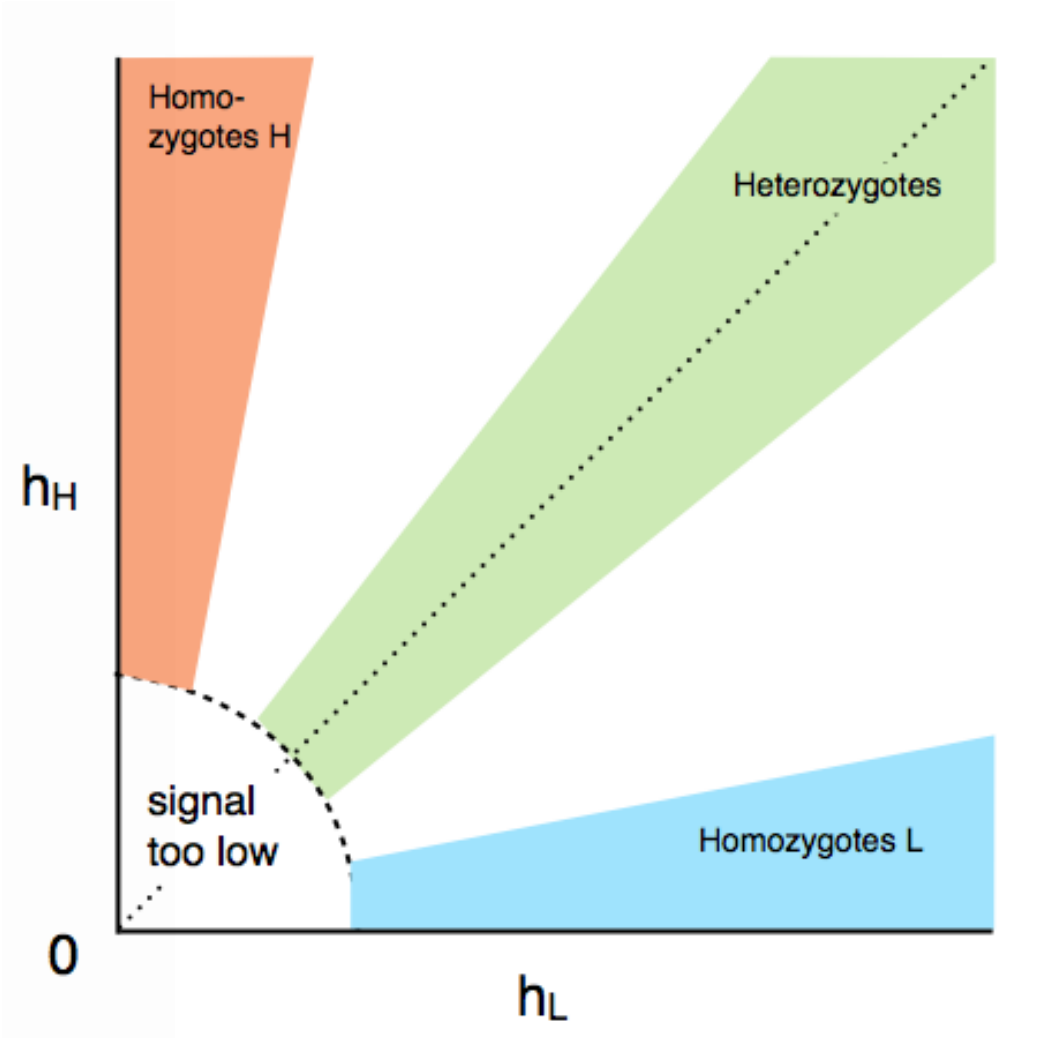


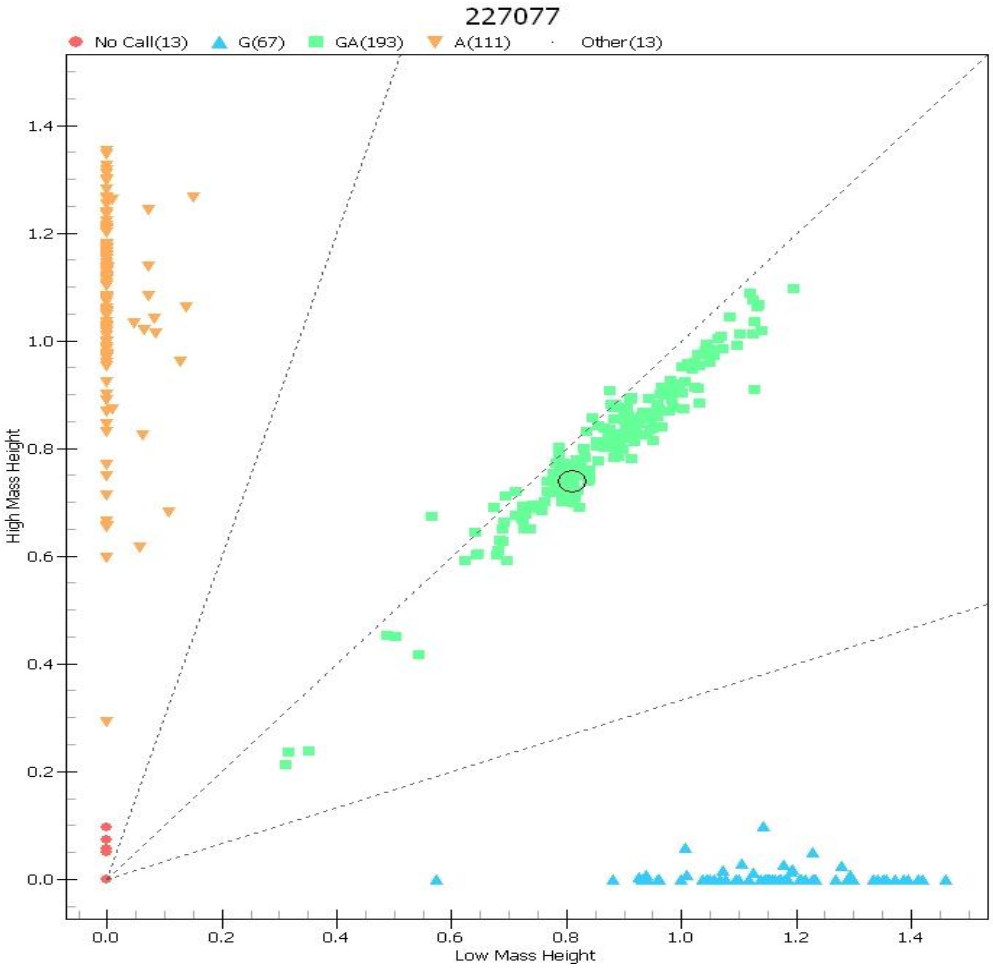
Figure 6 - Allelic Discrimination from Call Cluster Plots

Allelic discrimination occurs via a plot of the log height of the low mass allele peak (h_L) against the log height of the high mass allele peak (h_H). The areas where each sample lie determines the call of the genotype. Samples that lie along the X-axis are called as homozygotes for the L allele, samples that lie along the Y-axis are called as homozygotes for the H allele and samples that lie along the 45 degree line are called as heterozygotes. In a), these areas are represented as blue, orange and green respectively. Samples that lie too close to 0 on the graph are no-calls as the signal is too low (as determined by signal to noise ratio). In b), samples are represented by dots of the same colours as shown in a) with red dots representing no-calls.

a) Areas for genotype calls



b) Example call cluster plot



5. Materials

5.1 PCR Reaction

DNA source plate: 384-well deep-well PCR plate containing 10ng of dry DNA

100 mM dNTPs (QIAGEN)

25 mM $MgCl_2$ (QIAGEN)

PCR Primer Mix: Forward and reverse primers, 0.5µm each (Integrated DNA Technologies)

5U/µl Taq DNA Polymerase (QIAGEN)

10x PCR buffer (QIAGEN)

Water, HPLC grade

5.2 SAP Treatment

10x SAP buffer (Sequenom)

1.7 U/µl SAP enzyme (Sequenom)

Water, HPLC grade

5.3 iPLEX Reaction

iPLEX Enzyme (Sequenom)

10x iPLEX buffer (Sequenom)

iPLEX termination mix (Sequenom)

Extension Primer Mix: 5µM, 10µM, 15µM (Integrated DNA Technologies)

Water, HPLC grade

5.4 Conditioning

SpectroCLEAN resin (Sequenom)

Water, HPLC grade

Roto-shake Genie (Scientific Industries)

5.5 Miscellaneous

NanoDrop 8000 (Thermo Scientific)

ErgoOne Single Channel Pipette: P1, P10, P20, P100, P1000 (STARLAB)

TipOne Graduated Tips: 10µl, 200µl, 1000µl (STARLAB)

Thermo-Fast 384-well PCR plate (Thermo Scientific)

Adhesive PCR Film (Thermo Scientific)

Microcentrifuge 5415 D (Eppendorf)

Mixer Vortex Whirlimixer (Fisherbrand)

4-15 High Capacity Centrifuge (Sigma)

Veriti 384-well Thermal Cycler (Applied Biosystems)

MassARRAY Nanosispenser (Sequenom)

MassARRAY Analyzer 4 (Sequenom)

6. Statistical Methods

6.1 Exceptions from Dataset

Any samples that failed due to spotting errors were removed from all analysis. Any risk loci that failed to achieve > 95% coverage across remaining samples were then removed. Finally any samples that failed to achieve at least 75% genotype coverage across remaining risk loci were also removed from all subsequent analysis.

Unless otherwise specified all statistical analysis was performed using IBM® SPSS® Statistics Version 20 with a significance level (α) = 0.05 ie the null hypothesis is rejected at $p < 0.05$.

6.2 Assessing Hardy-Weinberg Equilibrium

Genotypes acquired using the MassARRAY® system by Sequenom® where assessed for Hardy-Weinberg Equilibrium (HWE) across each of the 18 risk loci in each individual study population. This was carried out with chi-square (x^2) testing outlined in Equation 1.

The x^2 value was plotted on a x^2 distribution graph with a degree of freedom (DF) = 1 to provide a two-tailed p-value. A x^2 value > 3.84 gives a two-tailed p-value < 0.05, which does not prove consistent with Hardy-Weinberg Equilibrium.

6.3 Calculating Combined Genetic Risk Across 18 Loci

The risk conferred from each loci was adjusted to take into account risk relative to the Scottish population using Equation 2. This uses the odds ratio (OR) from the original publications and the genotype frequencies derived from the control population dataset. For the loci in which HWE was consistent then the risk allele frequency (RAF) was used otherwise the genotype frequencies were used instead.

The total genetic risk across the 18 loci was then calculated by multiplying the relative risks derived from Equation 2 under a log-additive model as suggested by Pharoah et al (147) to provide a single genetic risk score. This was then log transformed under base 10 for subsequent analysis.

Equation 1 - Chi-Square Test

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where:

x^2 = Chi-Square test statistic

O_i = observed number of each genotype

E_i = expected number of each genotype based on Hardy-Weinberg from allele frequencies using the following formula:

$$p + q = 1$$

$$p^2 + 2pq + q^2 = 1$$

Where:

p = risk allele frequency

q = non-risk allele frequency

p^2 = risk homozygote frequency

$2pq$ = heterozygote frequency

q^2 = non-risk homozygote frequency

Equation 2 - Calculating Risk from Loci Relative to Population

$$R_p = p^2OR^2 + 2pqOR + q^2$$

$$RR = OR^{RA} /\ R_p$$

Where:

R_p = baseline population risk

p^2 = risk homozygote frequency in controls

$2pq$ = heterozygote frequency in controls

q^2 = non-risk homozygote frequency in controls

RR = risk relative to population

RA = number of risk alleles

6.4 Genetic Risk Distribution Across Groups

To determine whether or not genetic risk across 18 loci followed a log-normal distribution the log-genetic risk score was subjected to normality testing using the Shapiro-Wilk test statistic (*W*) and normal quantile-quantile (Q-Q) plots. The Shapiro-Wilk test statistic tests the null hypothesis that a sample is from a normally distributed distribution and is shown in Equation 3.

Equation 3 - Shapiro-Wilk Test Statistic

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where:

W = test statistic

$x_{(i)}$ = is the ith order sample

a_i = constants generated from a normal distribution of sample size n

6.5 Differences in Risk Distribution Between Groups

One-way analysis of variance analysis (ANOVA) was used to determine differences in genetic risk distribution between the different study groups. This is outlined in Equation 4 with the F statistic being plotted on an $F(d_1, d_2)$-distribution curve where d_1 represents degrees of freedom between groups and d_2 represents degrees of freedom within groups (2 and 2148 respectively). The null hypothesis states that the means of genetic risk in each group is equal.

Equation 4 - One-way ANOVA

$$F = \frac{\sum_i n_i(\overline{Y}_i - \overline{Y})^2/(K-1)}{\sum_{ij}\left(Y_{ij} - \overline{Y}_i\right)^2/(N-K)}$$

Where:

F = test statistic

Y = log-genetic risk score

K = number of groups

N = total number of observations

n_i = number of observations for i^{th} group

Y_{ij} = the j^{th} log-genetic risk score of the i^{th}

Tukey's honestly significant difference (HSD) test was subsequently used to identify which groups' means were significantly different. This single-step, post-hoc test is used in conjunction with a significant one-way ANOVA result. The null hypothesis states that the means of genetic risks between two groups tested is equal in a similar to an independent T-test. However, Tukey's HSD test is more conservative and is able to correct for multiple testing.

Significant results from Tukey's HSD test were then subjected to a size of effect calculation using Cohen's d as outline in Equation 5. A value of Cohen's d = 0.2, 0.5 and 0.8 are classified as "small", "medium" and "large" effect sizes respectively as suggested by Cohen (152).

Equation 5 - Cohen's d effect size equation

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s}$$

Where:

d = Cohen's d

s = pooled standard deviation as calculated by:

$$s = \sqrt{\frac{\sum(X_1 - \bar{X}_1)^2 + \sum(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

6.6 Discriminatory Accuracy of Risk Loci

To investigate the accuracy of the 18 risk loci at discriminating between breast cancer cases and population controls a ROC curve was used. The ROC curve is a plot of the false positive rate (1 – specificity) against the true positive rate (sensitivity) at various thresholds of the test score, in this case the genetic risk score derived from 18 risk loci. The area under the ROC curve (AUROC also known as the C-statistic) was then calculated. The null hypothesis states that the area under the curve is not greater than 0.5 ie the discriminatory accuracy is no better than by chance.

6.7 Correlating Age at Diagnosis with Risk Loci

Direct correlation between age at diagnosis in breast cancer cases and genetic risk across 18 loci was first investigated using Spearman's rank correlation coefficient (ρ_s). This statistic assesses how well the relationship between the two variables (age at diagnosis and genetic risk) can be attributed to a monotonic function, assuming that they don't both follow a normal distribution. The value of ρ_s will range between -1 and 1, with the extreme values corresponding to a perfect correlation between variables. The null hypothesis states there is no correlation ($\rho_s = 0$).

In order to investigate more subtle correlations, cases were subdivided into 5 groups based on age at diagnosis with each group representing a 10-year age band (35-44, 45-54, 55-64, 67-74 and 75-85 years old at diagnosis). Differences in genetic risk between these groups were then assessed using one-way ANOVA as previously outlined in Equation 4. The null hypothesis is that the means of genetic risk in each group is equal.

The AUROC analysis used previously to investigate the accuracy at discriminating between breast cancer cases and population controls was then applied to each of the individual age at diagnosis banded groups. This analysis was used to determine if discriminatory accuracy was higher for identifying breast cancer cases in younger age groups compared to population controls. This was complemented using independent sample T-tests comparing the means of genetic risk between the different age-banded breast cancer case groups and population controls.

6.8 Correlating Oestrogen Receptor Status with Risk Loci

Associations between ER status in breast cancer cases and genetic risk across 18 loci were first investigated using one-way ANOVA as shown in Equation 4. This tested for significant differences in means of genetic risk between those with ER-positive and ER-negative breast cancer cases from the control population.

AUROC analysis was again used to investigate the discriminatory accuracy of discriminating between breast cancer cases of different ER status and population controls. This was used to determine if discriminatory accuracy was higher for identifying breast cancer cases that were ER-positive.

A sub-analysis then further subdivided ER status based upon age of onset above and below the age of 55 to identify whether or not effects of different ER status were more prominent in either younger or older age groups.

6.9 Calculating Family History Risk

The family history component of breast cancer risk was calculated for 275 individuals from the increased risk group using the BOADICEA risk estimation web application (available from https://pluto.srl.cam.ac.uk/cgi-bin/bd2/v2/bd.cgi). Pedigrees were generated from all available family history data found within patient' cases notes, collected from breast cancer risk clinic appointments. The absolute 10-year risk of breast cancer at age 40 was used for each individual.

6.10 Correlating Genetic Risk with Family History Risk

To allow for correlation analysis, the genetic risk was converted from a relative risk to an absolute 10-year risk at age 40 using the ABSRISK program (developed by Dupont and Plummer (153) available from http://biostat.mc.vanderbilt.edu/wiki/Main/RelativeToAbsoluteRisks). The ABSRISK program used the mean from 2004-2008 of Scottish female breast cancer incidence and mortality data from the Information Service Department (ISD) Scotland (154) and total female mortality data from the General Registers Office (GR) for Scotland (155).

Correlation analysis of the absolute 10-year risks at age 40 from genotype data (genetic risk) and the BOADICA web program (family history risk) was performed using Spearman's rank correlation coefficient (ρ_s). This statistic assesses how well the relationship between two variables (in this cases genetic risk and family history risk (can be attributed to a monotonic function, assuming they don't both follow a normal distribution. The value of ρ_s ranges from -1 to 1,

with the extreme values corresponding to a perfect correlation between variables. The null hypothesis states there is no correlation ($\rho_S = 0$).

6.11 NICE Risk Classification

Individuals from the increased risk group were categorised as being as either average, moderate or high risk based on their absolute 10-year risk at age 40 under NICE guidelines. This was done separately using three different methods. First genetic risk from genotype data, then family history risk from BOADICEA and finally a combined risk calculated from multiplying the relative risk acquired through genotyping by the absolute risk from BOADICEA.

Assessing agreement between different methods was undertaken using Cohen's kappa statistic (κ). This statistic measures agreement between two methods or observers classifying samples into mutually exclusive categories and is defined in Equation 6. A value of κ = 1 implies perfect agreement between observers (eg BOADICEA vs combined risk) and a value of κ = 0 suggests agreement is not better than what may be obtained due to chance. The null hypothesis suggests that there is no agreement (κ = 0).

Equation 6 - Cohen's kappa statistic

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Where:

κ = test statistic

$\Pr(a)$ = the relative observed agreement between observers

$\Pr(e)$ = the hypothetical probability of chance agreement

6.12 Correlating Breast Tissue Density with Risk Loci

Mammographic breast tissue density measurements were available for 339 individuals from the increased risk group. Each individual's most recent digital mammogram was read independently by two consultant radiologists who scored breast tissue density on a visual breast density measurement (VBDM) analogue scale from 0-100 and by Breast Imaging-Reporting and Data System (BI-RADS) breast composition category (1, 2, 3 or 4, with 1 being least dense and 4 being most dense).

Correlations between genetic risk and breast tissue density were first examined using the VBDM and Spearman's rank correlation coefficient (ρ_s). One-way ANOVA was then used to detect differences between means of genetic risk as determined by BI-RADS score.

7. Results I – Distributions of Genetic Risk

7.1 Genotyping Across 18 Loci

The number of samples across the different study groups that were discarded from subsequent analysis due to spotting errors and more than 4 incomplete loci (< 75% total loci complete) are shown in Table 16. When disregarding spotting errors the proportion of complete loci across the total study population is 90.9% and when including those with less than 4 incomplete loci (> 75% total loci complete) the total working dataset consists of 97.3% of those sampled.

The genotypes of the 18 loci across different study groups are shown in Table 17, Table 18 and Table 19 along with assessment for HWE.

Table 16 - Call Rates of Genotyping Across Study Groups

Study group	Number genotyped	Spotting error (%)	Incomplete > 4 loci (%)	Incomplete ≤ 4 loci (%)	Complete (%)	Working dataset (%)
Cases	870	14 (1.6%)	27 (3.10%)	35 (4.02%)	793 (91.1%)	828 (95.2%)
Controls	1046	50 (4.78%)	28 (2.68%)	27 (2.58%)	941 (90.0%)	968 (92.5%)
FHx	385	26 (6.75%)	30 (7.79%)	79 (20.5%)	276 (71.7%)	355 (92.2%)
All	2301	90 (3.91%)	85 (3.69%)	141 (6.13%)	2010 (87.4%)	2151 (93.5%)
If spotting errors are disregarded						
Cases	856	-	27 (3.15%)	35 (4.01%)	793 (92.6%)	828 (96.7%)
Controls	996	-	28 (2.81%)	27 (2.71%)	941 (94.5%)	968 (97.2%)
FHx	359	-	30 (8.36%)	79 (22.0%)	276 (76.8%)	355 (98.9%)
All	2211	-	85 (3.84%)	141 (6.38%)	2010 (90.9%)	2151 (97.3%)

Table 17 - Genotypes of Control Group n = 968

Risk Loci	Low-risk homozygote (*frequency*)	Heterozygote (*frequency*)	High-risk homozygote (*frequency*)	Unsuccessful calls (*frequency*)	Risk allele Frequency	HWE p-value
FGFR2	268 *(0.277)*	474 *(0.490)*	225 *(0.232)*	1 *(0.001)*	0.478	0.582
TOX3	281 *(0.290)*	481 *(0.497)*	206 *(0.213)*	0	0.461	0.995
5p12	350 *(0.362)*	477 *(0.493)*	140 *(0.145)*	1 *(0.001)*	0.391	0.271
NOTCH2	361 *(0.373)*	453 *(0.468)*	136 *(0.140)*	18 *(0.019)*	0.382	0.749
ZNF365	22 *(0.023)*	222 *(0.229)*	724 *(0.748)*	0	0.863	0.312
RAD51L	83 *(0.086)*	375 *(0.387)*	506 *(0.523)*	4 *(0.004)*	0.719	0.257
ESR1	376 *(0.388)*	448 *(0.463)*	143 *(0.148)*	1 *(0.001)*	0.380	0.612
11q13	673 *(0.695)*	268 *(0.277)*	26 *(0.027)*	1 *(0.001)*	0.165	0.912
CASP8	16 *(0.017)*	234 *(0.242)*	718 *(0.742)*	0	0.863	0.537
2q35	228 *(0.236)*	472 *(0.488)*	267 *(0.276)*	1 *(0.001)*	0.520	0.490
MAP3K1	503 *(0.520)*	370 *(0.382)*	84 *(0.087)*	11 *(0.011)*	0.281	0.180
NEK10, SLC4A7	409 *(0.423)*	494 *(0.510)*	61 *(0.063)*	4 *(0.004)*	0.320	$3x10^{-8}$
CDKN2A/B	653 *(0.675)*	290 *(0.300)*	24 *(0.025)*	1 *(0.001)*	0.175	0.217
8q24	362 *(0.374)*	449 *(0.464)*	156 *(0.161)*	1 *(0.001)*	0.393	0.398
LSP1	442 *(0.457)*	407 *(0.420)*	116 *(0.120)*	3 *(0.003)*	0.331	0.138
10q22	366 *(0.378)*	467 *(0.482)*	134 *(0.138)*	1 *(0.001)*	0.380	0.439
10p15	350 *(0.362)*	454 *(0.469)*	164 *(0.169)*	0	0.404	0.418
COX11	78 *(0.081)*	412 *(0.426)*	477 *(0.493)*	1 *(0.001)*	0.706	0.402

Table 18 - Genotypes of Case Group n = 828

Risk Loci	Low-risk homozygote (*frequency*)	Heterozygote (*frequency*)	High-risk homozygote (*frequency*)	Unsuccessful calls (*frequency*)	Risk allele Frequency	HWE p-value
FGFR2	180 *(0.217)*	429 *(0.518)*	217 *(0.262)*	2 *(0.002)*	0.522	0.241
TOX3	196 *(0.237)*	413 *(0.499)*	219 *(0.264)*	0	0.514	0.962
5p12	284 *(0.343)*	422 *(0.510)*	122 *(0.147)*	0	0.402	0.085
NOTCH2	266 *(0.321)*	378 *(0.457)*	154 *(0.186)*	30 *(0.036)*	0.430	0.343
ZNF365	15 *(0.018)*	191 *(0.231)*	622 *(0.751)*	0	0.867	0.939
RAD51L	45 *(0.054)*	288 *(0.348)*	490 *(0.592)*	5 *(0.006)*	0.770	0.753
ESR1	307 *(0.371)*	400 *(0.483)*	121 *(0.146)*	0	0.388	0.614
11q13	570 *(0.688)*	234 *(0.283)*	24 *(0.029)*	0	0.170	0.999
CASP8	18 *(0.022)*	155 *(0.187)*	655 *(0.791)*	0	0.885	0.017
2q35	179 *(0.216)*	422 *(0.510)*	227 *(0.274)*	0	0.529	0.513
MAP3K1	420 *(0.507)*	340 *(0.411)*	58 *(0.070)*	10 *(0.012)*	0.279	0.334
NEK10, SLC4A7	229 *(0.277)*	407 *(0.492)*	186 *(0.225)*	6 *(0.007)*	0.474	0.841
CDKN2A/B	560 *(0.676)*	242 *(0.292)*	26 *(0.031)*	0	0.178	0.981
8q24	277 *(0.335)*	375 *(0.453)*	176 *(0.213)*	0	0.439	0.021
LSP1	368 *(0.444)*	374 *(0.452)*	86 *(0.104)*	0	0.330	0.528
10q22	283 *(0.342)*	409 *(0.494)*	136 *(0.164)*	0	0.411	0.564
10p15	245 *(0.296)*	409 *(0.494)*	174 *(0.210)*	0	0.457	0.891
COX11	75 *(0.091)*	324 *(0.391)*	429 *(0.518)*	0	0.714	0.223

Table 19 - Genotype of Increased Risk Group n = 355

Risk Loci	Low-risk homozygote (*frequency*)	Heterozygote (*frequency*)	High-risk homozygote (*frequency*)	Unsuccessful calls (*frequency*)	Risk allele Frequency	HWE p-value
FGFR2	104 *(0.294)*	173 *(0.489)*	77 *(0.218)*	1 *(0.003)*	0.462	0.751
TOX3	99 *(0.281)*	174 *(0.494)*	79 *(0.224)*	3 *(0.008)*	0.472	0.878
5p12	127 *(0.359)*	173 *(0.489)*	54 *(0.153)*	1 *(0.003)*	0.397	0.695
NOTCH2	131 *(0.378)*	151 *(0.435)*	65 *(0.187)*	8 *(0.023)*	0.405	0.071
ZNF365	10 *(0.028)*	85 *(0.239)*	260 *(0.732)*	0	0.852	0.346
RAD51L	50 *(0.141)*	103 *(0.291)*	201 *(0.568)*	1 *(0.003)*	0.713	$6x10^{-8}$
ESR1	141 *(0.397)*	173 *(0.487)*	41 *(0.115)*	0	0.359	0.269
11q13	254 *(0.718)*	90 *(0.254)*	10 *(0.028)*	1 *(0.003)*	0.155	0.556
CASP8	4 *(0.011)*	82 *(0.231)*	269 *(0.758)*	0	0.873	0.414
2q35	89 *(0.251)*	190 *(0.535)*	76 *(0.214)*	0	0.482	0.176
MAP3K1	142 *(0.502)*	121 *(0.428)*	20 *(0.071)*	72 *(0.203)*	0.284	0.397
NEK10, SLC4A7	91 *(0.259)*	179 *(0.509)*	82 *(0.233)*	3 *(0.008)*	0.487	0.740
CDKN2A/B	225 *(0.634)*	119 *(0.335)*	11 *(0.031)*	0	0.199	0.317
8q24	122 *(0.345)*	165 *(0.466)*	67 *(0.189)*	1 *(0.002)*	0.422	0.400
LSP1	156 *(0.444)*	154 *(0.439)*	41 *(0.117)*	4 *(0.011)*	0.336	0.750
10q22	137 *(0.386)*	167 *(0.470)*	51 *(0.144)*	0	0.379	0.993
10p15	117 *(0.363)*	145 *(0.450)*	60 *(0.186)*	33 *(0.093)*	0.411	0.207
COX11	30 *(0.085)*	146 *(0.416)*	175 *(0.499)*	4 *(0.011)*	0.707	0.954

7.2 Combined Genetic Risk Across 18 Loci

The genetic risk conferred from each loci relative to the population (as calculated in Equation 2) is shown in Table 20.

Table 20 - Risks from Loci Relative to Population

				Relative Risk to Population		
Risk Loci	**Per-Allele OR***	**RAF in controls**	**Baseline population risk**	**Non-risk homozygotes**	**Heterozygotes**	**Risk homozygotes**
FGFR2	1.24	0.48	1.24	0.80	1.00	1.24
TOX3	1.21	0.46	1.20	0.83	1.01	1.22
5p12	1.19	0.39	1.15	0.87	1.03	1.23
NOTCH2	1.16	0.38	1.13	0.89	1.03	1.20
ZNF365	1.16	0.86	1.30	0.77	0.90	1.04
RAD51L	1.15	0.72	1.23	0.81	0.94	1.08
ESR1	1.15	0.38	1.12	0.90	1.03	1.18
11q13	1.15	0.17	1.05	0.95	1.09	1.26
CASP8	1.14	0.86	1.26	0.80	0.91	1.03
2q35	1.12	0.52	1.13	0.89	0.99	1.11
MAP3K1	1.11	0.28	1.06	0.94	1.04	1.16
NEK10, SLC4A7	1.11	0.32	1.07	0.93	1.04	1.15
CDKN2A/B	1.09	0.18	1.03	0.97	1.06	1.15
8q24	1.08	0.39	1.06	0.94	1.02	1.10
LSP1	1.07	0.33	1.05	0.96	1.02	1.09
10q22	1.07	0.38	1.05	0.95	1.02	1.09
10p15	1.06	0.40	1.05	0.95	1.01	1.07
COX11	1.05	0.71	1.07	0.93	0.98	1.03

*From original papers

7.3 Genetic Risk Distribution

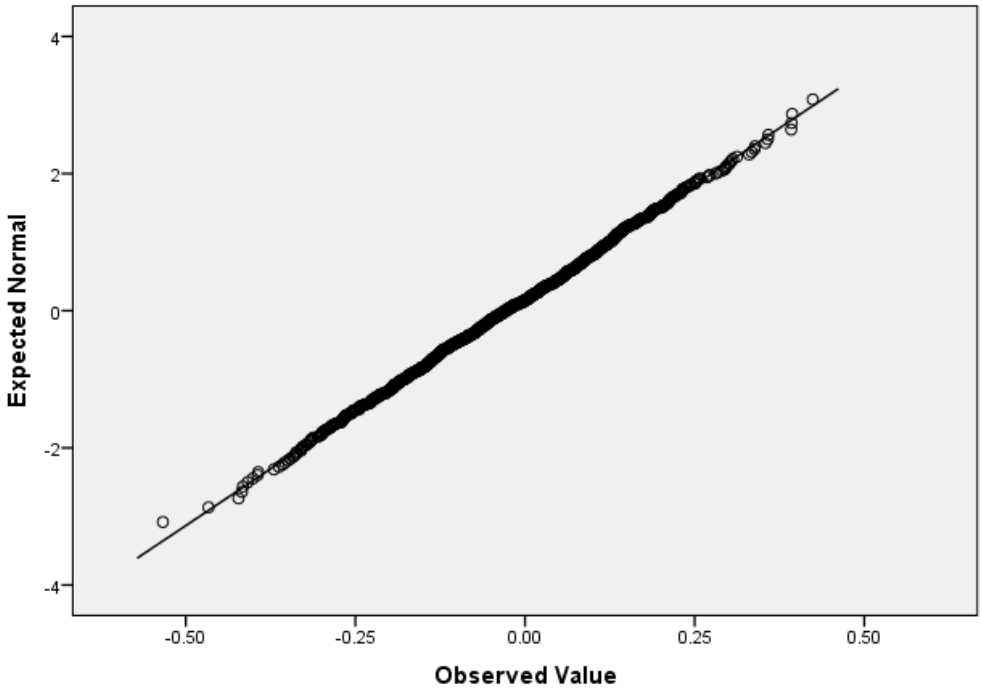
The descriptors for the distribution of log genetic risk across the different study groups is shown in Table 21. The null hypothesis for the Shapiro-Wilk test statistic was not rejected across each of the groups indicating they follow a normal distribution. This is further demonstrated by the linearity of the normal Q-Q plots shown in Figure 7. The groups showed similar standard deviations of 0.15096, 0.15043 and 0.15393 for the control, increased risk and case groups respectively. The control population was found to have a mean log genetic risk just below 0, with the distribution shifted to the right for the increased risk group and further to the right still for the case group.

Table 21 - Descriptors of Log Genetic Risk Distribution by Study Groups

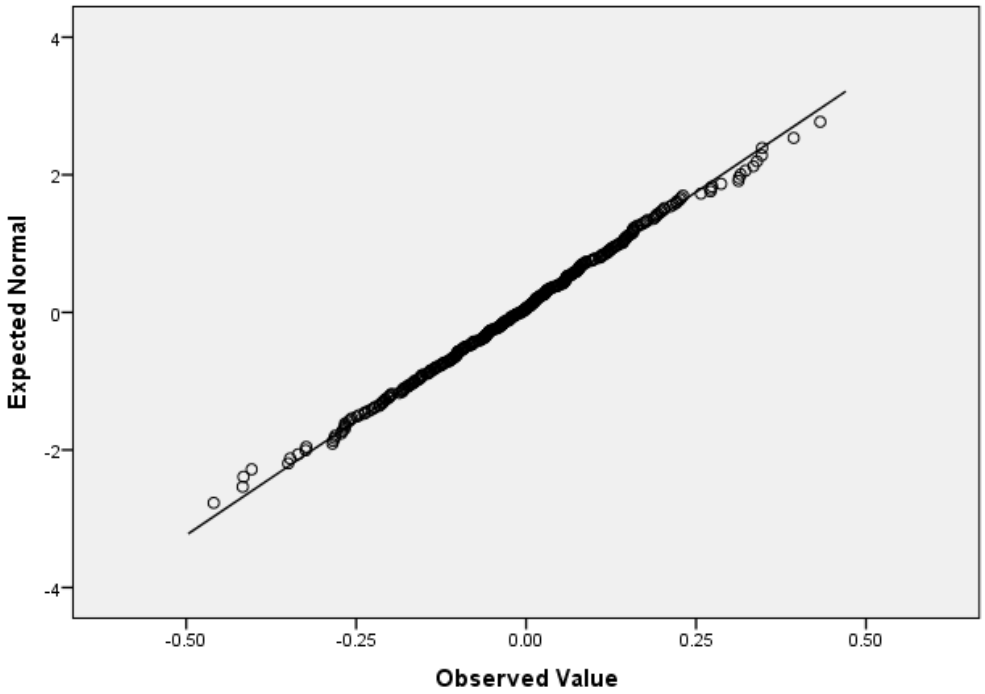
	Controls	Increased Risk	Cases
Sample size	968	355	828
Mean	-0.0270	-0.0131	0.0310
95% CI	-0.0365, -0.0174	-0.0288, 0.0026	0.0205, 0.0415
Standard Deviation	0.15096	0.15043	0.15393
Skewness	-0.017	-0.014	0.184
Kurtosis	-0.112	0.138	0.026
Shapiro-Wilk Test	0.913	0.922	0.063

Figure 7 - Normal Q-Q Plots of Log Genetic Risk

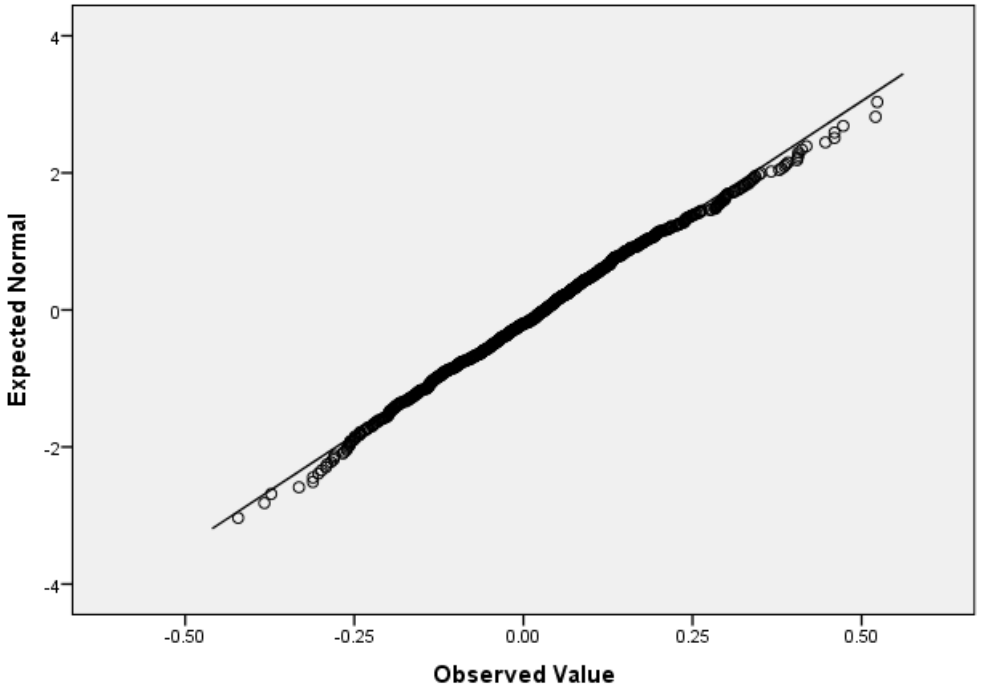
a) Control group



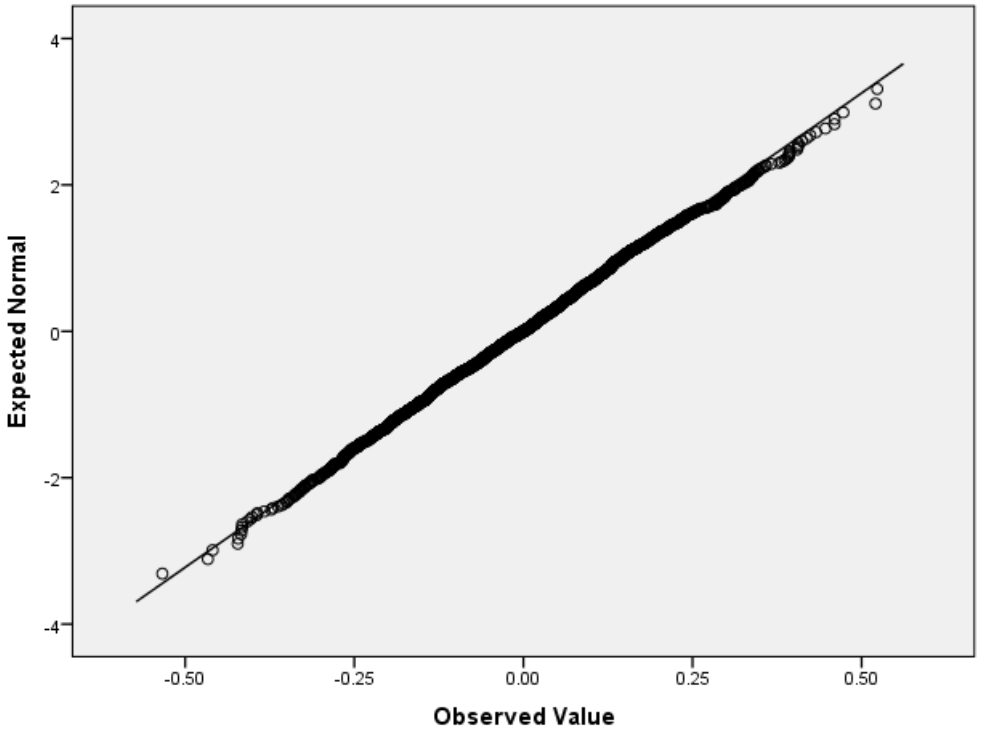
b) Increased risk group



c) Case group



d) Total study population



7.4 Differences in Risk Distribution Between Study Groups

One-way ANOVA testing rejected the null hypothesis of equal means across the study groups with $P < 0.001$. The results of post-hoc analysis with Tukey's HSD and Cohen's d effect size is shown in Table 22.

There was found to be no significant difference in mean between the control and increased risk groups. However significant differences were found between control and case groups, and increased risk and case groups (both $P < 0.001$) with small-medium and small effect sizes respectively.

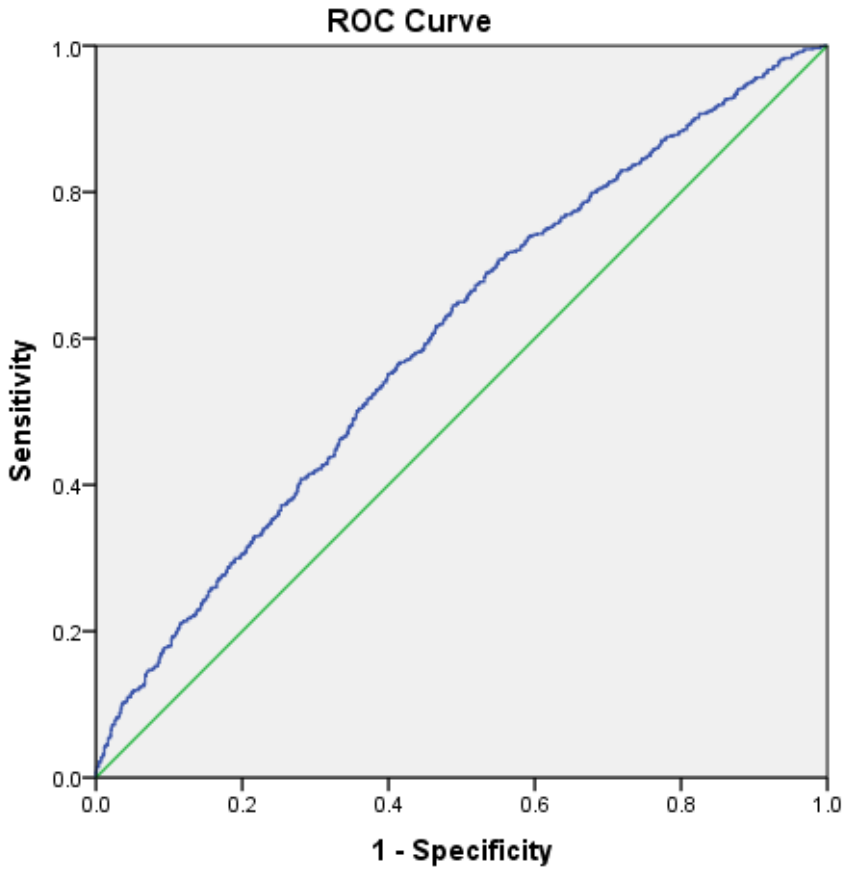
Table 22 - Differences in Genetic Risk Between Groups

Groups	Difference in Mean	95% CI	Tukey's HSD p-value	Cohen's (d)	Effect Size
Control – Increased Risk	0.01389	-0.0082, 0.0360	0.304	0.0922	-
Control – Cases	0.05800	0.0411, 0.0749	< 0.001	0.3804	Small-medium
Increased Risk – Cases	0.04410	0.0215, 0.0667	< 0.001	0.2898	Small

7.5 Discriminatory Accuracy Across 18 Loci

ROC analysis of the discriminatory accuracy across the 18 loci demonstrated a blue curved line as shown in Figure 8, with a green diagonal reference line demonstrating an AUC = 0.5. The null hypothesis was rejected (P < 0.001) demonstrating that genetic risk across 18 loci performs better than by chance with an AUC = 0.602 (95% CI 0.575, 0.628)

Figure 8 - ROC Curve



8. Results II – Correlations with Clinical Characteristics

8.1 Correlating Genetic Risk with Age at Diagnosis

Spearman's rank correlation coefficient did not demonstrate any direct correlation between age at diagnosis in breast cancer cases and genetic risk across 18 loci (P = 0.472).

One-way ANOVA analysis based on 10-year age banding (35-44, 45-54, 55-64, 65-74 and 75-85 years old at diagnosis) failed to demonstrate any significant difference in means of genetic risk across groups (P = 0.090).

Mean genetic risk across 10-year age banding along with independent T-testing and AUROC analysis compared to population controls is shown in Table 23. Each age banded group was found to have a significant difference in mean of genetic risk from population controls. The largest genetic risk and C-statistic was found in the youngest age group (35-44 year olds). Lower scores were demonstrated in the older age groups with an apparent trend for a decrease in genetic risk as age of diagnosis increases.

Table 23 - Genetic Risk in Cases by Age at Diagnosis in 10-year Bands

Age band	N	Mean (95% CI)	Difference in mean (95% CI)*	*P-value**	C-statistic (95% CI)*	*P-value**
35-44	87	0.0390 (0.0048, 0.0732)	0.06594 (0.03262, 0.09927)	*< 0.001*	0.620 (0.559, 0.681)	*< 0.001*
45-54	175	0.0375 (0.0134, 0.0615)	0.06444 (0.0395, 0.08903)	*< 0.001*	0.607 (0.562, 0.652)	*< 0.001*
55-64	163	0.0312 (0.0097, 0.0527)	0.0581 (0.03338, 0.08298)	*< 0.001*	0.611 (0.566, 0.656)	*< 0.001*
65-74	158	0.0279 (0.0033, 0.0525)	0.05487 (0.02932, 0.08041)	*< 0.001*	0.589 (0.542, 0.636)	*< 0.001*
75-85	74	0.0233 (-0.0105, 0.0571)	0.05031 (0.01466, 0.08595)	*0.006*	0.590 (0.526, 0.654)	*0.010*

*Compared to population controls

8.2 Correlating Genetic Risk with ER Status

Differences between the mean of genetic risk between ER-positive and ER-negative breast cancer cases from population controls along with AUROC analysis is shown in Table 24. Both ER-positive and ER-negative breast cancer cases were found to have significant differences in mean of genetic risk from population controls, with a higher difference found in ER-positive cases.

Table 24 - Genetic Risk in Cases by Oestrogen Receptor Status

ER Status	N	Mean (95% CI)	Difference in mean (95% CI)*	*P-value**	C-statistic (95% CI)*	*P-value**
+ve	417	0.0324 (0.0178, 0.0469)	0.05934 (0.03387, 0.07416)	*< 0.001*	0.601 (0.570, 0.633)	*< 0.001*
-ve	120	0.0221 (-0.0072, 0.0514)	0.04906 (0.02015, 0.07796)	*0.001*	0.586 (0.530, 0.641)	*0.002*

*Compared to population controls

Table 25 - Genetic Risk in ER-Positive Breast Cancer by Age at Diagnosis

Age	N	Mean (95% CI)	Difference in mean (95% CI)*	*P-value**	C-statistic (95% CI)*	*P-value**
< 55	143	0.0426 (0.0162, 0.0890)	0.06955 (0.04281, 0.09629)	*< 0.001*	0.619 (0.570, 0.667)	*< 0.001*
≥ 55	274	0.0271 (0.0096, 0.0445)	0.05401 (0.0419, 0.07670)	*< 0.001*	0.592 (0.556, 0.629)	*< 0.001*

*Compared to population controls

Table 26 - Genetic Risk in ER-Negative Breast Cancer by Age at Diagnosis

Age	N	Mean (95% CI)	Difference in mean (95% CI)*	*P-value**	C-statistic (95% CI)*	*P-value**
< 55	48	0.0112 (-0.0352, 0.0576)	0.03851 (-0.00578, 0.08207)	*0.089*	0.560 (0.478, 0.641)	*0.163*
≥ 55	72	0.0294 (-0.0092, 0.0679)	0.05633 (0.01992, 0.09274)	*0.002*	0.603 (0.531, 0.675)	*0.004*

*Compared to population controls

Differences between mean of genetic risk between breast cancer cases diagnosed above and below the age of 55 from population controls are shown in Table 25 and Table 26 for ER-positive and ER-negative disease respectively. A larger genetic risk was found in younger onset than older onset ER-positive breast cancer when compared to population controls. In contrast, a larger genetic risk was found in older onset than younger onset ER-negative disease and in particular, no significant difference was found in mean of genetic risk between young onset ER-negative disease compared to population controls (P = 0.089).

8.3 Correlation of Genetic Risk with Family History Risk

Spearman's rank correlation coefficient did not demonstrate any direct correlations between genetic risk and family history risk as determined by the BOADICA risk estimation tool (P = 0.964). Additionally, there was found to be no significant agreement of NICE risk categorisation based upon 10 year risk at age 40 using Cohen's kappa statistic (P = 0.717). This is demonstrated in Table 27, where only 48.1% of individuals were found in the same NICE category under both methods.

Table 27 - NICE Risk Categorisation

		BOADICEA NICE Categorisation			
		AVERAGE	**MODERATE**	**HIGH**	**Total**
Genetic Risk NICE Categorisation	**AVERAGE**	108	106	12	226
	MODERATE	4	5	0	9
	HIGH	0	0	0	0
	Total	112	111	12	235

8.4 Correlating Genetic Risk with Breast Tissue Density

Spearman's rank correlation coefficient did not demonstrate any direct correlations between genetic risk and breast tissue density as measured by a visual breast density measurement scale (P = 0.919).

The means of genetic risk by BI-RADS grouping are shown in Table 28. One-way ANOVA failed to show any significant difference in mean of genetic risk across groups (P = 0.486).

Table 28 - Genetic Risk by BI-RADS Score

BI-RADS	N	Mean	95% CI	SD
1	156	-0.0102	-0.0343, 0.0140	0.15267
2	111	-0.0161	-0.0441, 0.0119	0.14867
3	36	-0.0216	-0.0702, 0.0269	0.14342
4	8	0.0665	-0.0506, 0.1837	0.14013

9. Discussion

9.1 The MassARRAY System was successful in genotyping the study population

Genotyping of 18 risk loci was produced using a single iPLEX assay as part of the MassARRAY System by Sequenom. A small proportion of samples failed genotyping due to spotting errors at the time of transfer from 384 well plate to the SpectroCHIP® array using the MassARRAY Nanodispenser (3.91% across whole study population). Taking this into consideration, the final working dataset represented 97.3% of all samples successfully transferred for genotyping, which proved sufficient for statistical analysis. Call rates across loci ranged from 93.58-98.01% giving a total call rate across all loci of 97.11%, which is found to be consistent with publicised call rates (156).

The majority of loci across study populations were found to be in HWE (shown in Table 17, Table 18 and Table 19) with the exception of the *NEK10/SLC4A7* locus in the control group, the *CASP8* and *8q24* loci in the case group and the *RAD51L* locus in the increased risk group ($P < 0.05$). This relatively low number of loci across the different study groups however would be consistent with multiple testing and is reassuring for the accuracy of genotype calls made using the MassARRAY system.

9.2 Genotyping using a single SNP assay is economically viable

At the number of loci investigated in this study, the MassARRAY system is both a practical and cost-effective method for genotyping large numbers of individuals. By combining genotyping of 18 risk loci into a single iPLEX assay, less genetic material is required for analysis and data is made available through

a single output source. This is in contrast to TaqMan® Allelic Discrimination Assays by Applied Biosystems™, which will require approximately 18 duplicates of the same genetic sample (as some level of multiplexing is possible) to genotype across the same number of risk loci.

Due to the process used for genotyping using the MassARRAY system, the cost of consumables required for one 384 well plate of samples is approximately £600, irrespective of the size of the iPLEX assay used and not including the cost of primers. An approximate comparison of reagent/primer costs shows that TaqMan assays would cost around £1700 (£0.20-£0.30 per genotype per assay) to genotype 384 samples across 18 loci whereas a single iPLEX assay would only cost around £250 (£0.03-£0.04 per genotype per sample) (157). Even when considering additional consumables (such as the SpectroCHIP) the MassARRAY system is still considerable cheaper at around £650 (£0.08-£0.10 per genotype per sample). These cost comparisons fail to take into account the availability and cost of additional equipment required, which need to be considered for any long-term cost analysis.

With the discovery of additional risk loci for breast cancer there is the need for expansion of the original 18 loci single iPLEX assay. It is claimed that a single iPLEX assay can genotype up to 40 SNPs although there is little literature to support this. The problems faced when designing iPLEX assays are the potential of dimer formation between primers and false priming potential, both of which increase as a greater number of SNPs are added to the design process. The current number of risk loci currently published for breast cancer would require approximately a further 3 iPLEX assays to accommodate for them (at a level of approximately 15-20 SNPs per assay). Other genotyping technologies would need to be more thoroughly investigated for practicality and cost-effectiveness in comparison to multiple iPLEX assays as part of the MassARRAY System.

9.3 Polygenic risk across 18 loci follow a log-normal distribution

Genotype data across 18 risk loci from 968 individuals from the Scottish population has demonstrated that relative risk of breast cancer from low-penetrance loci follows a log normal distribution. The mean of this distribution was found to be just below zero (actual mean = -0.0270) as previously demonstrated by Pharoah et al using data modelled from seven risk loci (147). This distribution of risk was found to be similar for those at increased risk due to a positive family history of breast cancer (355 individuals) and in breast cancer cases (855 individuals) although these distributions were shifted to the right as expected (actual mean = -0.0131 and 0.0310 respectively).

If 18 risk loci were a true representation of the relative risk due to genetic factors across a population then the expected mean of the control population would lie close to zero and not just below as has been demonstrated. Although the relative risks from each allele were corrected using allele frequencies in the control group (ie a Scottish population), this fails to take into account the relative risks of each loci as they were first discovered in the literature (predominantly European populations). The allele frequencies in the Scottish population have been found to be different from those reported in the literature in many of the risk loci (see Table 17, Table 18 and Table 19). It may be that the Scottish population are at a lower polygenic risk of breast cancer than the total European population. Otherwise more risk loci or a larger study population may be required to give a more accurate representation of polygenic risk.

The distribution of risk from the case population was found to be significantly different from both the control and increased risk populations however there was no significant difference between the control and increased risk population. This somewhat suggests that polygenic risk across 18 loci are not a fair representation of risk conferred through a strong family history (discussed in

more detail below), although there are some limitations to the increased risk group dataset that may not provide an accurate representation. The increased risk group all had a positive family history, were aged from 40-50 and did not have a prior diagnosis of breast cancer. If they were diagnosed before the age of 40 or before recruitment then although they had a strong family history they would not have been included in the increased risk group. This may give an artificially lower mean of polygenic risk for this group. Regardless it is still reassuring that the distribution of risk is higher than the control population as was expected.

9.4 A polygenic risk profile performs similarly to established risk models

An 18-locus polygenic risk profile was found to have limited discriminatory ability for identifying cases of breast cancer with an AUROC = 0.602. This risk profile performs similarly to other established risk stratification models within European populations that are based upon family history data and to the Gail model, which uses predominantly personalised information.

It is clear that an 18-locus polygenic risk profile has relatively poor discriminatory ability for individualised risk stratification however may be beneficial for population wide screening strategies. In this instance it has benefits over family history risk stratification alone. Firstly, it does not rely on an individual's direct recall of information regarding their family history of breast cancer. This may at first seem a trivial point but it would be important to appreciate that not every individual is as aware of their family's medical history as would be required for full stratification of their risk using this information, especially at points beyond first-degree relations. In contrast genetic information does not require such recall. In a similar respect, although processing genetic

information is a far lengthier process overall than that of a family history, it requires far less time from the perspective of a patient and is more able to be routinely performed while already undertaking a screening programme ie a blood sample could be taken at the point of mammography/clinical examination. Other aspects that need to be considered are the costs and resources that are required for each modality of risk stratification, which have been discussed elsewhere.

9.5 Genetic risk may be higher for those at younger age of diagnosis

Previous studies have failed to investigate any potential relationship between age of onset of breast cancer and combined genetic risk across low-penetrance risk loci. A polygenic risk score was found to have (somewhat limited) predictive value (ie AUROC > 0.5) across all age groups. This risk score was found to be highest in the youngest age group (age 35-44) and lowest in the oldest age group (age 75-85) as would be expected for a multifactorial inheritance model of disease. Unfortunately no direct correlation or trend could be demonstrated with the relatively small sample sizes available.

This is the first study to demonstrate the predictive value of genotype data in identifying younger women (ie below 50) at increased risk of breast cancer. This is important to recognise as such women may otherwise not be part of the National Breast Cancer Screening Programme. These women may even benefit from additional screening as recommended by current NICE clinical guidelines. There would however be difficulty in identifying such women in the first place if they otherwise did not have a strong family history of breast cancer and this would need to be considered if such genotyping is to be employed as a strategy.

9.6 Genetic risk is highest for oestrogen receptor positive disease

A significant number of the 18 risk loci were found to have stronger associations for ER-positive disease when first identified through GWAS (see APPENDIX). Combined genetic risk across the 18 loci was found to be predictive for both ER-positive and ER-negative disease. Slightly stronger predictive ability was found for ER positive disease however this is not statistically different from the predictive ability of ER negative disease, which is again likely due to a limited sample size.

An improved predictive value for ER positive breast cancer over ER negative disease may have implications for preventative treatment. Tamoxifen and raloxifene are selective oestrogen receptor modulators that have been shown to be beneficial in reducing the risk of ER-positive breast cancer (48% reduction in incidence) but have been shown to have no effect for ER-negative disease (158). Such treatments are associated with added risks of endometrial cancer and thromboembolic disease. Due to the potential for serious adverse events such as these, it is recommended that such preventative treatment only be used for those at highest risk of breast cancer who are also at lowest risk of adverse events. The use of a polygenic risk profile may therefore help identify women who are most likely to receive benefit from this particular treatment.

Such use of genetic information may be improved through the development of a polygenic risk profile specific to ER-positive disease ie one that includes only risk loci associated with ER-positive disease. However, the number of risk loci available within this study is likely to be too low to give any additional improvement if the risk loci are restricted in this way. Several of the new breast cancer risk loci that have been identified are also found to have stronger associations for ER-positive than ER-negative breast cancer, with a proportion having no increased risk whatsoever for ER-negative disease (117). A new

polygenic profile could then be developed that using these risk loci, that can maintain the approximate number of SNPs used in this study.

Interestingly, further stratification of the ER-positive and ER-negative breast cancer cases by age of diagnosis revealed that the association for higher genetic risk in younger ages of onset was only true in ER-positive cases and not ER-negative disease. More specifically, genetic risk in young onset ER-negative women was not found to be significantly different from population controls. This is somewhat surprising considering that ER-negative disease is more common in younger ages of onset and so overall genetic risk may be expected to be higher for these women. However, this is in keeping with younger onset breast cancer not being driven predominantly by hormonal factors, which may carry more of a genetic component. Additionally, this finding highlights the relatively poor discriminatory ability of this particular polygenic profile in identifying specifically ER-negative disease. This suggests a need in identifying risk loci specific to ER-negative breast cancer, especially considering its poorer clinical outcomes.

9.7 Genotype data may help identify disease pathways

The underlying mechanisms (and even possible causative genes) for many of the established susceptibility loci have yet to be fully elicited. The identification of risk loci specific to ER-positive breast cancer suggests that there may be different underlying pathways in disease development of breast cancer based upon ER status. In a similar sense, loci with equal associations for both ER positive and ER negative disease suggests that some underlying disease pathways may be shared. Demonstrating associations between susceptibility loci and underlying pathological characteristics and tumour subtypes may help elicit such underlying mechanisms, which subsequently may guide treatment and preventative strategies. With the exception of hormone receptor status, there is currently very little literature demonstrating any such associations. This

however is not to say these associations do not exist but instead for many loci they have yet to be fully examined.

An alternative approach for eliciting underlying mechanisms could be performed using statistical analysis techniques that aim to identify interactions between individual susceptibility loci. For example, a number of loci are suggested to be involved in MAPK, apoptotic or cell cycle regulatory pathways (see Appendix). If such an interaction were found to occur between loci of unknown function and those in already established disease pathways, it may suggest that they fall within the same or interconnected pathways of breast cancer development. Additionally, such an approach may even reveal novel breast cancer pathways by identifying interactions between loci of established function.

One such method is to use logistic regression, a parametric approach that relates independent variables (eg genotyping across loci) to a binary outcome (eg diagnosis of breast cancer). This method however is less able to deal with high-dimensionality interactions (ie interactions beyond pairs). This is because as the number of loci rises, more cells within the contingency table contain no observations (ie are empty). Although this can be overcame with larger sample sizes this can often become impractical. An alternative method is that of Multifactor Dimensionality Reduction (MDR), a non-parametric and model-free approach that has been shown to reveal a high-order interaction between four SNPs found with oestrogen-metabolism genes in sporadic breast cancer case-control data (159). MDR effectively reduces the number of dimensions to one by pooling multi-locus genotypes into high and low-risk groups, which is then subjected to cross-validation and permutation testing to evaluate its ability to classify or predict disease status.

9.8 Genetic risk does not correlate with other established risk factors

To fully evaluate the benefits of a polygenic risk profile in estimating breast cancer risk, correlations with other established heritable risk factors needs to be considered. Genotype data from the increased risk group has demonstrated no apparent correlations between an 18 locus genotype and either breast tissue density or family history risk as calculated using the BOADICEA risk estimation tool.

There are some limitations to our current analysis of breast tissue density with respect to polygenic risk within our dataset. Breast tissue density itself is influenced by menopausal status, BMI, parity, use of hormone replacement therapy (HRT) and age of the patient. Despite this, these factors only account or around 20-30% of age adjusted variance (160). It may still be possible that a combination of low-penetrance polygenic risk loci may have limited associations with breast tissue density, but these associations are not strong enough across the 18 loci examined to be immediately apparent within our analysis above other possible confounding risk factors. Such associations may be of more benefit when considering underlying mechanisms of normal breast tissue and breast cancer development but may be less benefit when considering cancer risk. This is because such confounding factors of breast tissue density are themselves also associated with breast cancer risk. Breast tissue density may then be accepted as a partially surrogate marker of cancer risk due to “non-genetic” factors, which can be used in combination with polygenic risk.

The lack of correlation between polygenic risk across 18 loci and the risk conferred by the BOADICEA risk estimation tool demonstrates that each method act somewhat independently from one another. Despite the inclusion of a polygenic component into the BOADICEA risk estimation tool, this component must only cover a small proportion of the overall polygenic risk of breast cancer.

This is not overly surprising considering the vast number of new low-penetrance risk loci discovered in recent years. Further investigation would be needed to assess whether these additional loci are likely to show any correlation with BOADICEA's conferred risk. However until such investigation is undertaken it would be reasonable to allow for the combining of genotype data from the 18 risk loci examined in this study into the current BOADICEA model.

9.9 Approaches to improving national breast cancer screening

Any screening strategy gains most benefit from identifying those who are likely to gain most benefit and those who are not. In the case of breast cancer screening it would be prudent to identify individuals at highest risk who would gain most benefit and conversely those at lowest risk who may be subject to greater levels of harm associated with screening. So far there is no such strategy to classify women other than by their age (ie those aged 50-70) and for those with either a strong family history or known high risk mutation (such as *BRCA1* or *BRCA2*).

Several researchers have shown that the inclusion of genetic data into risk estimation models can improve upon their discriminatory accuracy but overall still not to a sufficiently high enough level for individual risk stratification. Such risk stratification however may be of benefit at the population level for identifying those women at the highest and lowest levels of risk. Based on the log-normal distribution of risk across 18 loci in the population, approximately 1.5% of individuals would be classified as increased risk under NICE guidelines based upon genotype information alone. Such people may benefit from additional screening at a younger age and from the additional clinical information that may be provided through genotype information.

Pharoah et al suggested a threshold value of absolute 10-year risk of breast cancer for which mammographic screening may be appropriate (147). This was calculated to be 2.3% based upon breast cancer incidence and all-cause mortality data from England and Wales. This would equate to a relative risk of 0.77, of which it is estimated to account for 20% of the population based upon risk distribution across seven loci. If similar epidemiological data was assumed for a Scottish population, this would account for 28.4% of the population based upon risk distribution across 18 loci. Under such a distribution, polygenic information may therefore be of more benefit in restricting use of mammographic screening and thus preventing over diagnosis and treatment rather than identifying the highest risk individuals.

10. Conclusion

This study has given further evidence for the utility of low-penetrance susceptibility loci in breast cancer risk stratification. Polygenic risk across 18 loci followed a log-normal distribution across the Scottish population, with a higher mean for both those at increased risk due to family history and for those diagnosed with breast cancer. Polygenic risk across 18 loci was not found to correlate with either family history risk as determined by the BOADICEA risk estimation tool or with breast tissue density as determined through digital mammography. This suggests that they may be used in combination for improved discrimination of risk. Women diagnosed with breast cancer at a younger age have a higher polygenic risk, which suggests that those at highest polygenic risk may benefit from additional screening. Additionally, polygenic risk is more strongly associated with ER positive disease, which has implications on the use of tamoxifen as a preventative treatment. Further research with the addition of a larger polygenic risk profile in combination with other risk factors is needed to fully investigate the potential of genotype data in breast cancer risk discrimination at a population level. This may subsequently improve the efficacy of a national mammographic screening programme by identifying those at highest and lowest risk.

11. Appendix

PCR and Extension Primer Sequences

SNP ID	Forward PCR Primer	Reverse PCR Primer	Extension Primer
4415084	ACGTTGGATGCACATACCTCTACCTCTAGC	ACGTTGGATGTGACCAGTGCTGTATGTATC	TCCTGATGACTTGAGCA
12253826	ACGTTGGATGAGGATTCAGTGAGCTCAGGA	ACGTTGGATGTAGAGACGGGGTTTCCGTG	GCTCAGGAGTTCAAGAC
999737	ACGTTGGATGGGTCCCGTTCACATGATATG	ACGTTGGATGCACCAAGGACTTATGGACAG	ACATGATATGAATGGGGC
12443621	ACGTTGGATGGATTCCCTTAGAAATAAGGAG	ACGTTGGATGGACGTTTTATATGCATTAGGC	AATACCTACCTCAAGTTCA
1011970	ACGTTGGATGAAGATACAGGTGGAACTGGG	ACGTTGGATGAGAACTGATAGGGAGCCAGC	TGGAACTGGGCCAGTGTTT
704010	ACGTTGGATGGTAAGAGTCTGGGCAGCTTG	ACGTTGGATGTACTGCCACGTCTTACAACC	AGACCTGACCTGAAAATAGC
11249433	ACGTTGGATGAAAAAGCAGAGAAAGCAGGG	ACGTTGGATGTGAGTCACTGTGCTAAGGAG	GAAAGCAGGGCTGGGTTTAA
13387042	ACGTTGGATGGAACAGCTAAACCAGAACAG	ACGTTGGATGGGAAGATTCGATTCAACAAGG	gggAGAAAGAAGGCAAATGGA
6504950	ACGTTGGATGCCAGGGTTTGTCTACCAAAG	ACGTTGGATGCTGAATCACTCCTTGCCAAC	tcGTCTACCAAAGGCAGGATAC
3817198	ACGTTGGATGTTCCCTAGTGGAGCAGTGG	ACGTTGGATGTCTCACCTGATACCAGATTC	ctCTGACTCTAGTGAAATGAGC
614367	ACGTTGGATGTGGCTGTTTGGGGCCTAAAG	ACGTTGGATGTCCTTGGCTTTTTCCTCCAG	tGGGCCTAAAGAGATGTAATGC
2981578	ACGTTGGATGTGAAGCTTTTACCTCTATGC	ACGTTGGATGTTAAGAGCCGGCGCCATCAC	tccTTTACCTCTATGCAAATATGC
889312	ACGTTGGATGGAAGGAGTCGTTGAGTTTTC	ACGTTGGATGATCTCTGAGATGCCCCTGCT	ccTGTAGTCTCTTAATTTGCACAT
2046210	ACGTTGGATGATCAGGGTGCCTCAACTGTC	ACGTTGGATGCCTCACACATACATACAGTC	TGAATCTTTTATTTCAGGTAGATG
10995190	ACGTTGGATGTGTTCTGATTGGCTTGCCAC	ACGTTGGATGCAATGGTTGTGTCCAAGTGC	tGTGGGAGTTCATTTCACACTAAAA
75325449	ACGTTGGATGAAGCACTCCCTTTTCCCCAC	ACGTTGGATGTTACTCCTGCAAGAATGGTC	CACACCAAAATATATTGTTTTTTTAC
13281615	ACGTTGGATGGGTAACTATGAATCTCATC	ACGTTGGATGATCACTCTTATTTCTCCCCC	gtTAACTATGAATCTCATCAAAAGAA
4973768	ACGTTGGATGCAAAATGATCTGACTACTCC	ACGTTGGATGAATCACTTAAAACAAGCAG	gTAAGAGCAAAGGTAACTCATGTTTA

12. References

1. ISD. Cancer Statistics: Breast Cancer. Information and Statistics Department Scotland; [17th August 2012]; Available from: http://www.isdscotland.org/health-topics/cancer/cancer-statistics/breast/.

2. Fitzgibbons PL, Page DL, Weaver D, Thor AD, Allred DC, Clark GM, et al. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. Archives of pathology & laboratory medicine. 2000;124(7):966-78. Epub 2000/07/11.

3. Dunnwald LK, Rossing MA, Li CI. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast cancer research : BCR. 2007;9(1):R6. Epub 2007/01/24.

4. Wilson JM, Jungner YG. Principles and practice of screening for disease. Geneva: World Health Organisation, 1968.

5. Shapiro S, Strax P, Venet L. Evaluation of periodic breast cancer screening with mammography. Methodology and early observations. JAMA : the journal of the American Medical Association. 1966;195(9):731-8. Epub 1966/02/28.

6. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. Journal of the National Cancer Institute. 1982;69(2):349-55. Epub 1982/08/01.

7. Andersson I, Aspegren K, Janzon L, Landberg T, Lindholm K, Linell F, et al. Mammographic screening and mortality from breast cancer: the Malmo mammographic screening trial. BMJ. 1988;297(6654):943-8. Epub 1988/10/15.

8. Tabar L, Gad A. Screening for breast cancer: the Swedish trial. Radiology. 1981;138(1):219-22. Epub 1981/01/01.

9. Frisell J, Glas U, Hellstrom L, Somell A. Randomized mammographic screening for breast cancer in Stockholm. Design, first round results and comparisons. Breast cancer research and treatment. 1986;8(1):45-54. Epub 1986/01/01.

10. Bjurstam N, Bjorneld L, Duffy SW, Smith TC, Cahlin E, Eriksson O, et al. The Gothenburg breast screening trial: first results on mortality, incidence, and mode of detection for women ages 39-49 years at randomization. Cancer. 1997;80(11):2091-9. Epub 1997/12/10.

11. Nystrom L, Rutqvist LE, Wall S, Lindgren A, Lindqvist M, Ryden S, et al. Breast cancer screening with mammography: overview of Swedish randomised trials. Lancet. 1993;341(8851):973-8. Epub 1993/04/17.

12. Roberts MM, Alexander FE, Anderson TJ, Forrest AP, Hepburn W, Huggins A, et al. The Edinburgh randomised trial of screening for breast cancer: description of method. British journal of cancer. 1984;50(1):1-6. Epub 1984/07/01.

13. Alexander FE, Anderson TJ, Brown HK, Forrest AP, Hepburn W, Kirkpatrick AE, et al. The Edinburgh randomised trial of breast cancer screening: results after 10 years of follow-up. British journal of cancer. 1994;70(3):542-8. Epub 1994/09/01.

14. Alexander FE, Anderson TJ, Brown HK, Forrest AP, Hepburn W, Kirkpatrick AE, et al. 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. Lancet. 1999;353(9168):1903-8. Epub 1999/06/17.

15. Forrest AP. Breast Cancer Screening: Report to the Health Ministers of England, Wales, Scotland and Northern Ireland. HMSO, 1986.

16. Gøtzsche P, Nielsen M. Screening for breast cancer with mammography (Review). Cochrane Database of Systematic Reviews. 2011(4).

17. The benefits and harms of breast cancer screening: an independent review. Lancet. 2012;380(9855):1778-86. Epub 2012/11/03.

18. NICE CG41 Familial breast cancer: the classification and care of women at risk of familial breast cancer in primary, secondary and tertiarty care. Available from: http://guidance.nice.org.uk/CG41.

19. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA. Family history and the risk of breast cancer: a systematic review and meta-analysis. International journal of cancer Journal international du cancer. 1997;71(5):800-9. Epub 1997/05/29.

20. Ahlbom A, Lichtenstein P, Malmstrom H, Feychting M, Hemminki K, Pedersen NL. Cancer in twins: genetic and nongenetic familial risk factors. Journal of the National Cancer Institute. 1997;89(4):287-93. Epub 1997/02/19.

21. Hemminki K, Forsti A, Bermejo JL. The 'common disease-common variant' hypothesis and familial risks. PloS one. 2008;3(6):e2504. Epub 2008/06/19.

22. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53. Epub 2009/10/09.

23. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nature reviews Genetics. 2008;9(4):255-66. Epub 2008/03/06.

24. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature reviews Genetics. 2010;11(6):415-25. Epub 2010/05/19.

25. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature reviews Genetics. 2010;11(6):446-50. Epub 2010/05/19.

26. Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature genetics. 2010;42(7):570-5. Epub 2010/06/22.

27. Gibson G. Hints of hidden heritability in GWAS. Nature genetics. 2010;42(7):558-60. Epub 2010/06/29.

28. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature genetics. 2010;42(11):937-48. Epub 2010/10/12.

29. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832-8. Epub 2010/10/01.

30. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nature genetics. 2008;40(6):695-701. Epub 2008/05/30.

31. Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nature reviews Genetics. 2009;10(8):565-77. Epub 2009/07/09.

32. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. Nature. 2007;447(7143):433-40. Epub 2007/05/25.

33. Jablonka E, Raz G. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. The Quarterly review of biology. 2009;84(2):131-76. Epub 2009/07/18.

34. Karran P. DNA double strand break repair in mammalian cells. Current opinion in genetics & development. 2000;10(2):144-50. Epub 2000/04/08.

35. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. American journal of human genetics. 1998;62(3):676-89. Epub 1998/04/29.

36. Peto J, Collins N, Barfoot R, Seal S, Warren W, Rahman N, et al. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. Journal of the National Cancer Institute. 1999;91(11):943-9. Epub 1999/06/08.

37. Malkin D, Li FP, Strong LC, Fraumeni JF, Jr., Nelson CE, Kim DH, et al. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science. 1990;250(4985):1233-8. Epub 1990/12/10.

38. Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science. 1997;275(5308):1943-7. Epub 1997/03/28.

39. Easton DF. How many more breast cancer predisposition genes are there? Breast cancer research : BCR. 1999;1(1):14-7. Epub 2001/03/16.

40. Bell DW, Varley JM, Szydlo TE, Kang DH, Wahrer DC, Shannon KE, et al. Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. Science. 1999;286(5449):2528-31. Epub 2000/01/05.

41. Vahteristo P, Tamminen A, Karvinen P, Eerola H, Eklund C, Aaltonen LA, et al. p53, CHK2, and CHK1 genes in Finnish families with Li-Fraumeni syndrome: further evidence of CHK2 in inherited cancer predisposition. Cancer research. 2001;61(15):5718-22. Epub 2001/08/02.

42. Chehab NH, Malikzay A, Appel M, Halazonetis TD. Chk2/hCds1 functions as a DNA damage checkpoint in G(1) by stabilizing p53. Genes & development. 2000;14(3):278-88. Epub 2000/02/16.

43. Vahteristo P, Bartkova J, Eerola H, Syrjakoski K, Ojala S, Kilpivaara O, et al. A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer. American journal of human genetics. 2002;71(2):432-8. Epub 2002/07/03.

44. Shiloh Y. ATM and related protein kinases: safeguarding genome integrity. Nature reviews Cancer. 2003;3(3):155-68. Epub 2003/03/04.

45. Swift M, Reitnauer PJ, Morrell D, Chase CL. Breast and other cancers in families with ataxia-telangiectasia. The New England journal of medicine. 1987;316(21):1289-94. Epub 1987/05/21.

46. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature genetics. 2006;38(8):873-5. Epub 2006/07/13.

47. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007;447(7148):1087-93. Epub 2007/05/29.

48. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. Science. 2005;307(5712):1072-9. Epub 2005/02/19.

49. Udler MS, Meyer KB, Pooley KA, Karlins E, Struewing JP, Zhang J, et al. FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. Human molecular genetics. 2009;18(9):1692-703. Epub 2009/02/19.

50. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A. Linkage disequilibrium patterns of the human genome across populations. Human molecular genetics. 2003;12(7):771-6. Epub 2003/03/26.

51. Garcia-Closas M, Hall P, Nevanlinna H, Pooley K, Morrison J, Richesson DA, et al. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS genetics. 2008;4(4):e1000054. Epub 2008/04/26.

52. Jackson D, Bresnick J, Rosewell I, Crafton T, Poulsom R, Stamp G, et al. Fibroblast growth factor receptor signalling has a role in lobuloalveolar development of the mammary gland. Journal of cell science. 1997;110 (Pt 11):1261-8. Epub 1997/06/01.

53. Dickson C, Spencer-Dene B, Dillon C, Fantl V. Tyrosine kinase signalling in breast cancer: fibroblast growth factors and their receptors. Breast cancer research : BCR. 2000;2(3):191-6. Epub 2001/03/16.

54. Adnane J, Gaudray P, Dionne CA, Crumley G, Jaye M, Schlessinger J, et al. BEK and FLG, two receptors to members of the FGF family, are amplified in subsets of human breast cancers. Oncogene. 1991;6(4):659-63. Epub 1991/04/01.

55. Penault-Llorca F, Bertucci F, Adelaide J, Parc P, Coulier F, Jacquemier J, et al. Expression of FGF and FGF receptor genes in human breast cancer. International journal of cancer Journal international du cancer. 1995;61(2):170-6. Epub 1995/04/10.

56. Koziczak M, Holbro T, Hynes NE. Blocking of FGFR signaling inhibits breast cancer cell proliferation through downregulation of D-type cyclins. Oncogene. 2004;23(20):3501-8. Epub 2004/04/30.

57. Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, et al. Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. PLoS biology. 2008;6(5):e108. Epub 2008/05/09.

58. Martin AJ, Grant A, Ashfield AM, Palmer CN, Baker L, Quinlan PR, et al. FGFR2 protein expression in breast cancer: nuclear localisation and correlation with patient genotype. BMC research notes. 2011;4:72. Epub 2011/03/23.

59. Nordgard SH, Johansen FE, Alnaes GI, Naume B, Borresen-Dale AL, Kristensen VN. Genes harbouring susceptibility SNPs are differentially expressed in the breast cancer subtypes. Breast cancer research : BCR. 2007;9(6):113. Epub 2007/11/27.

60. Santen RJ, Song RX, McPherson R, Kumar R, Adam L, Jeng MH, et al. The role of mitogen-activated protein (MAP) kinase in breast cancer. The Journal of steroid biochemistry and molecular biology. 2002;80(2):239-56. Epub 2002/03/19.

61. Creighton CJ, Hilger AM, Murthy S, Rae JM, Chinnaiyan AM, El-Ashry D. Activation of mitogen-activated protein kinase in estrogen receptor alpha-positive breast cancer cells in vitro induces an in vivo molecular phenotype of estrogen receptor alpha-negative human breast tumors. Cancer research. 2006;66(7):3903-11. Epub 2006/04/06.

62. Smid M, Wang Y, Klijn JG, Sieuwerts AM, Zhang Y, Atkins D, et al. Genes associated with breast cancer metastatic to bone. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2006;24(15):2261-7. Epub 2006/04/26.

63. Liu L, Cara DC, Kaur J, Raharjo E, Mullaly SC, Jongstra-Bilen J, et al. LSP1 is an endothelial gatekeeper of leukocyte transendothelial migration. The Journal of experimental medicine. 2005;201(3):409-18. Epub 2005/02/03.

64. Huang CK, Zhan L, Ai Y, Jongstra J. LSP1 is the major substrate for mitogen-activated protein kinase-activated protein kinase 2 in human neutrophils. The Journal of biological chemistry. 1997;272(1):17-9. Epub 1997/01/03.

65. Liu B, Yang L, Huang B, Cheng M, Wang H, Li Y, et al. A functional copy-number variation in MAPKAPK2 predicts risk and prognosis of lung cancer. American journal of human genetics. 2012;91(2):384-90. Epub 2012/08/14.

66. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics. 2008;40(6):703-6. Epub 2008/04/29.

67. Theodorou V, Boer M, Weigelt B, Jonkers J, van der Valk M, Hilkens J. Fgf10 is an oncogene activated by MMTV insertional mutagenesis in mouse mammary tumors and overexpressed in a subset of human breast carcinomas. Oncogene. 2004;23(36):6047-55. Epub 2004/06/23.

68. Chioni AM, Grose R. Negative regulation of fibroblast growth factor 10 (FGF-10) by polyoma enhancer activator 3 (PEA3). European journal of cell biology. 2009;88(7):371-84. Epub 2009/05/05.

69. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581-4. Epub 2004/04/24.

70. Cavdar Koc E, Ranasinghe A, Burkhart W, Blackburn K, Koc H, Moseley A, et al. A new face on apoptosis: death-associated protein 3 and PDCD9 are mitochondrial ribosomal proteins. FEBS letters. 2001;492(1-2):166-70. Epub 2001/03/15.

71. Grigoriadis A, Mackay A, Reis-Filho JS, Steele D, Iseli C, Stevenson BJ, et al. Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data. Breast cancer research : BCR. 2006;8(5):R56. Epub 2006/10/04.

72. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530-6. Epub 2002/02/02.

73. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, 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):579-84. Epub 2009/03/31.

74. Guo S, Liu M, Gonzalez-Perez RR. Role of Notch and its oncogenic signaling crosstalk in breast cancer. Biochimica et biophysica acta. 2011;1815(2):197-213. Epub 2011/01/05.

75. Graziani I, Eliasz S, De Marco MA, Chen Y, Pass HI, De May RM, et al. Opposite effects of Notch-1 and Notch-2 on mesothelioma cell survival under hypoxia are exerted through the Akt pathway. Cancer research. 2008;68(23):9678-85. Epub 2008/12/03.

76. Fu YP, Edvardsen H, Kaushiva A, Arhancet JP, Howe TM, Kohaar I, et al. NOTCH2 in breast cancer: association of SNP rs11249433 with gene expression in ER-positive breast tumors without TP53 mutations. Molecular cancer. 2010;9:113. Epub 2010/05/21.

77. Li X, Heyer WD. Homologous recombination in DNA repair and DNA damage tolerance. Cell research. 2008;18(1):99-113. Epub 2008/01/02.

78. Shlien A, Tabori U, Marshall CR, Pienkowska M, Feuk L, Novokmet A, et al. Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(32):11264-9. Epub 2008/08/08.

79. Kos M, Reid G, Denger S, Gannon F. Minireview: genomic organization of the human ERalpha gene promoter region. Mol Endocrinol. 2001;15(12):2057-63. Epub 2001/12/04.

80. Micheli A, Muti P, Secreto G, Krogh V, Meneghini E, Venturelli E, et al. Endogenous sex hormones and subsequent breast cancer in premenopausal women. International journal of cancer Journal international du cancer. 2004;112(2):312-8. Epub 2004/09/08.

81. Key T, Appleby P, Barnes I, Reeves G. Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies. Journal of the National Cancer Institute. 2002;94(8):606-16. Epub 2002/04/18.

82. Holst F, Stahl PR, Ruiz C, Hellwinkel O, Jehan Z, Wendland M, et al. Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. Nature genetics. 2007;39(5):655-60. Epub 2007/04/10.

83. MacPherson G, Healey CS, Teare MD, Balasubramanian SP, Reed MW, Pharoah PD, et al. Association of a common variant of the CASP8 gene with reduced risk of breast cancer. Journal of the National Cancer Institute. 2004;96(24):1866-9. Epub 2004/12/17.

84. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, et al. A common coding variant in CASP8 is associated with breast cancer risk. Nature genetics. 2007;39(3):352-8. Epub 2007/02/13.

85. Hengartner MO. The biochemistry of apoptosis. Nature. 2000;407(6805):770-6. Epub 2000/10/26.

86. Zuzak TJ, Steinhoff DF, Sutton LN, Phillips PC, Eggert A, Grotzer MA. Loss of caspase-8 mRNA expression is common in childhood primitive neuroectodermal brain tumour/medulloblastoma. Eur J Cancer. 2002;38(1):83-91. Epub 2001/12/26.

87. Takita J, Yang HW, Bessho F, Hanada R, Yamamoto K, Kidd V, et al. Absent or reduced expression of the caspase 8 gene occurs frequently in neuroblastoma, but not commonly in Ewing sarcoma or rhabdomyosarcoma. Medical and pediatric oncology. 2000;35(6):541-3. Epub 2000/12/07.

88. Wu Y, Alvarez M, Slamon DJ, Koeffler P, Vadgama JV. Caspase 8 and maspin are downregulated in breast cancer cells due to CpG site promoter methylation. BMC cancer. 2010;10:32. Epub 2010/02/06.

89. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nature genetics. 2007;39(7):865-9. Epub 2007/05/29.

90. Milne RL, Benitez J, Nevanlinna H, Heikkinen T, Aittomaki K, Blomqvist C, et al. Risk of estrogen receptor-positive and -negative breast cancer and single-nucleotide polymorphism 2q35-rs13387042. Journal of the National Cancer Institute. 2009;101(14):1012-8. Epub 2009/07/02.

91. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nature genetics. 2010;42(6):504-7. Epub 2010/05/11.

92. Gianfrancesco F, Esposito T, Ombra MN, Forabosco P, Maninchedda G, Fattorini M, et al. Identification of a novel gene and a common variant associated with uric acid nephrolithiasis in a Sardinian genetic isolate. American journal of human genetics. 2003;72(6):1479-91. Epub 2003/05/13.

93. Wang Q, Du X, Meinkoth J, Hirohashi Y, Zhang H, Liu Q, et al. Characterization of Su48, a centrosome protein essential for cell division. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(17):6512-7. Epub 2006/04/18.

94. Karlseder J, Zeillinger R, Schneeberger C, Czerwenka K, Speiser P, Kubista E, et al. Patterns of DNA amplification at band q13 of chromosome 11 in human breast cancer. Genes, chromosomes & cancer. 1994;9(1):42-8. Epub 1994/01/01.

95. Janssen JW, Cuny M, Orsetti B, Rodriguez C, Valles H, Bartram CR, et al. MYEOV: a candidate gene for DNA amplification events occurring centromeric to CCND1 in breast cancer. International journal of cancer Journal international du cancer. 2002;102(6):608-14. Epub 2002/11/26.

96. Ornitz DM, Xu J, Colvin JS, McEwen DG, MacArthur CA, Coulier F, et al. Receptor specificity of the fibroblast growth factor family. The Journal of biological chemistry. 1996;271(25):15292-7. Epub 1996/06/21.

97. Serrano M, Hannon GJ, Beach D. A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. Nature. 1993;366(6456):704-7. Epub 1993/12/16.

98. Zhang Y, Xiong Y, Yarbrough WG. ARF promotes MDM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the Rb and p53 tumor suppression pathways. Cell. 1998;92(6):725-34. Epub 1998/04/07.

99. Chang DL, Qiu W, Ying H, Zhang Y, Chen CY, Xiao ZX. ARF promotes accumulation of retinoblastoma protein through inhibition of MDM2. Oncogene. 2007;26(32):4627-34. Epub 2007/02/14.

100. Silva J, Dominguez G, Silva JM, Garcia JM, Gallego I, Corbacho C, et al. Analysis of genetic and epigenetic processes that influence p14ARF expression in breast cancer. Oncogene. 2001;20(33):4586-90. Epub 2001/08/09.

101. Hannon GJ, Beach D. p15INK4B is a potential effector of TGF-beta-induced cell cycle arrest. Nature. 1994;371(6494):257-61. Epub 1994/09/15.

102. Nobori T, Miura K, Wu DJ, Lois A, Takabayashi K, Carson DA. Deletions of the cyclin-dependent kinase-4 inhibitor gene in multiple human cancers. Nature. 1994;368(6473):753-6. Epub 1994/04/21.

103. Beliakoff J, Sun Z. Zimp7 and Zimp10, two novel PIAS-like proteins, function as androgen receptor coregulators. Nuclear receptor signaling. 2006;4:e017. Epub 2006/07/25.

104. Matson SW, Bean DW, George JW. DNA helicases: enzymes with essential roles in all aspects of DNA metabolism. BioEssays : news and reviews in molecular, cellular and developmental biology. 1994;16(1):13-22. Epub 1994/01/01.

105. Kim J, Kim JH, Lee SH, Kim DH, Kang HY, Bae SH, et al. The novel human DNA helicase hFBH1 is an F-box protein. The Journal of biological chemistry. 2002;277(27):24530-7. Epub 2002/04/17.

106. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nature genetics. 2009;41(5):585-90. Epub 2009/03/31.

107. Quarmby LM, Mahjoub MR. Caught Nek-ing: cilia and centrioles. Journal of cell science. 2005;118(Pt 22):5161-9. Epub 2005/11/11.

108. Bowers AJ, Boylan JF. Nek8, a NIMA family kinase member, is overexpressed in primary human breast tumors. Gene. 2004;328:135-42. Epub 2004/03/17.

109. Moniz LS, Stambolic V. Nek10 mediates G2/M cell cycle arrest and MEK autoactivation in response to UV irradiation. Molecular and cellular biology. 2011;31(1):30-42. Epub 2010/10/20.

110. Chen Y, Choong LY, Lin Q, Philp R, Wong CH, Ang BK, et al. Differential expression of novel tyrosine kinase substrates during breast cancer development. Molecular & cellular proteomics : MCP. 2007;6(12):2072-87. Epub 2007/09/15.

111. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nature genetics. 2007;39(10):1217-24. Epub 2007/09/18.

112. Carr HS, Maxfield AB, Horng YC, Winge DR. Functional analysis of the domains in Cox11. The Journal of biological chemistry. 2005;280(24):22664-9. Epub 2005/04/21.

113. Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, 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):885-92. Epub 2010/09/21.

114. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nature genetics. 2011;43(12):1210-4. Epub 2011/11/01.

115. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nature genetics. 2012;44(3):312-8. Epub 2012/01/24.

116. Siddiq A, Couch FJ, Chen GK, Lindstrom S, Eccles D, Millikan RC, et al. A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. Human molecular genetics. 2012;21(24):5373-84. Epub 2012/09/15.

117. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nature genetics. 2013;45(4):353-61. Epub 2013/03/29.

118. Johns PC, Yaffe MJ. X-ray characterisation of normal and neoplastic breast tissues. Physics in medicine and biology. 1987;32(6):675-95. Epub 1987/06/01.

119. Wolfe JN. Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer. 1976;37(5):2486-92. Epub 1976/05/01.

120. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2006;15(6):1159-69. Epub 2006/06/16.

121. Pollan M, Ascunce N, Ederra M, Murillo A, Erdozain N, Ales-Martinez JE, et al. Mammographic density and risk of breast cancer according to tumor characteristics and mode of detection: a Spanish population-based case-control study. Breast cancer research : BCR. 2013;15(1):R9. Epub 2013/01/31.

122. Vachon CM, Kuni CC, Anderson K, Anderson VE, Sellers TA. Association of mammographically defined percent breast density with epidemiologic risk factors for breast cancer (United States). Cancer causes & control : CCC. 2000;11(7):653-62. Epub 2000/09/08.

123. Boyd NF, Dite GS, Stone J, Gunasekara A, English DR, McCredie MR, et al. Heritability of mammographic density, a risk factor for breast cancer. The New England journal of medicine. 2002;347(12):886-94. Epub 2002/09/20.

124. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, et al. Mammographic breast density as an intermediate phenotype for breast cancer. The lancet oncology. 2005;6(10):798-808. Epub 2005/10/04.

125. Lindstrom S, Vachon CM, Li J, Varghese J, Thompson D, Warren R, et al. Common variants in ZNF365 are associated with both mammographic density and breast cancer risk. Nature genetics. 2011;43(3):185-7. Epub 2011/02/01.

126. Vachon CM, Scott CG, Fasching PA, Hall P, Tamimi RM, Li J, et al. Common breast cancer susceptibility variants in LSP1 and RAD51L1 are associated with mammographic density measures that predict breast cancer risk. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2012;21(7):1156-66. Epub 2012/03/29.

127. Varghese JS, Thompson DJ, Michailidou K, Lindstrom S, Turnbull C, Brown J, et al. Mammographic breast density and breast cancer: evidence of a shared genetic basis. Cancer research. 2012;72(6):1478-84. Epub 2012/01/24.

128. Baker LH. Breast Cancer Detection Demonstration Project: five-year summary report. CA: a cancer journal for clinicians. 1982;32(4):194-225. Epub 1982/07/01.

129. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute. 1989;81(24):1879-86. Epub 1989/12/20.

130. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. American journal of epidemiology. 1985;122(5):904-14. Epub 1985/11/01.

131. Claus EB, Risch N, Thompson WD. Genetic analysis of breast cancer in the cancer and steroid hormone study. American journal of human genetics. 1991;48(2):232-42. Epub 1991/02/01.

132. Lalouel JM, Morton NE. Complex segregation analysis with pointers. Human heredity. 1981;31(5):312-21. Epub 1981/01/01.

133. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. Cancer. 1994;73(3):643-51. Epub 1994/02/01.

134. Antoniou AC, Pharoah PD, McMullan G, Day NE, Ponder BA, Easton D. Evidence for further breast cancer susceptibility genes in addition to BRCA1 and BRCA2 in a population-based study. Genetic epidemiology. 2001;21(1):1-18. Epub 2001/07/10.

135. Antoniou AC, Pharoah PD, McMullan G, Day NE, Stratton MR, Peto J, et al. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. British journal of cancer. 2002;86(1):76-83. Epub 2002/02/22.

136. Antoniou AC, Pharoah PP, Smith P, Easton DF. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. British journal of cancer. 2004;91(8):1580-90. Epub 2004/09/24.

137. Antoniou AC, Cunningham AP, Peto J, Evans DG, Lalloo F, Narod SA, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. British journal of cancer. 2008;98(8):1457-66. Epub 2008/03/20.

138. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet. 2001;358(9291):1389-99. Epub 2001/11/14.

139. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Statistics in medicine. 2004;23(7):1111-30. Epub 2004/04/02.

140. Narod SA, Ford D, Devilee P, Barkardottir RB, Lynch HT, Smith SA, et al. An evaluation of genetic heterogeneity in 145 breast-ovarian cancer families. Breast Cancer Linkage Consortium. American journal of human genetics. 1995;56(1):254-64. Epub 1995/01/01.

141. Anderson H, Bladstrom A, Olsson H, Moller TR. Familial breast and ovarian cancer: a Swedish population-based register study. American journal of epidemiology. 2000;152(12):1154-63. Epub 2000/12/29.

142. Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C, et al. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. Journal of medical genetics. 2003;40(11):807-14. Epub 2003/11/25.

143. Quante AS, Whittemore AS, Shriver T, Strauch K, Terry MB. Breast cancer risk assessment across the risk continuum: genetic and nongenetic risk factors contributing to

differential model performance. Breast cancer research : BCR. 2012;14(6):R144. Epub 2012/11/07.

144. Stahlbom AK, Johansson H, Liljegren A, von Wachenfeldt A, Arver B. Evaluation of the BOADICEA risk assessment model in women with a family history of breast cancer. Familial cancer. 2012;11(1):33-40. Epub 2011/11/30.

145. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. British journal of cancer. 2000;83(10):1301-8. Epub 2000/10/25.

146. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. Nature genetics. 2002;31(1):33-6. Epub 2002/05/02.

147. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction, and targeted prevention of breast cancer. The New England journal of medicine. 2008;358(26):2796-803. Epub 2008/06/27.

148. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR, et al. Performance of common genetic variants in breast-cancer risk models. The New England journal of medicine. 2010;362(11):986-93. Epub 2010/03/20.

149. Darabi H, Czene K, Zhao W, Liu J, Hall P, Humphreys K. Breast cancer risk prediction and individualised screening based on common genetic variation and breast density measurement. Breast cancer research : BCR. 2012;14(1):R25. Epub 2012/02/09.

150. Kerr SM, Liewald DC, Campbell A, Taylor K, Wild SH, Newby D, et al. Generation Scotland: Donor DNA Databank; A control DNA resource. BMC medical genetics. 2010;11:166. Epub 2010/11/26.

151. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24(24):2938-9. Epub 2008/11/01.

152. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

153. Dupont WD, Plummer WD, Jr. Understanding the relationship between relative and absolute risk. Cancer. 1996;77(11):2193-9. Epub 1996/06/01.

154. ISD. Breast Cancer Statistics. Information and Statistics Department Scotland; Available from: http://www.isdscotland.org/Health-Topics/Cancer/Cancer-Statistics/Breast/.

155. GRO. Deaths Time Series Data. General Register Office for Scotland; [cited 2011 12th April]; Available from: http://www.gro-scotland.gov.uk/statistics/theme/vital-events/deaths/time-series.html.

156. Ragoussis J. Genotyping technologies for genetic research. Annual review of genomics and human genetics. 2009;10:117-33. Epub 2009/05/21.

157. Ragoussis J. Genotyping technologies for all. Drug Discovery Today: Technologies. 2006;3(2):115-22.

158. Cuzick J, Powles T, Veronesi U, Forbes J, Edwards R, Ashley S, et al. Overview of the main outcomes in breast-cancer prevention trials. Lancet. 2003;361(9354):296-300. Epub 2003/02/01.

159. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. American journal of human genetics. 2001;69(1):138-47. Epub 2001/06/19.

160. Vachon CM, Kushi LH, Cerhan JR, Kuni CC, Sellers TA. Association of diet and mammographic breast density in the Minnesota breast cancer family cohort. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2000;9(2):151-60. Epub 2000/03/04.