Exploring the genetic architecture and the chromatin organisation of breast cancer

Olivia Leavy

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy of the
University of London

September 2017

Faculty of Epidemiology and Population Health
Department of Non-communicable Disease Epidemiology
London School of Hygiene and Tropical Medicine

Funded by Breast Cancer Now

Declaration

I, Olivia Leavy, confirm that the work presented in this thesis is my own and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the thesis.

The CHi-C data used in chapter 6, based on cell-lines, was collected and prepared by Dr. Olivia Fletcher and her team at the Institute of Cancer Research (ICR), London, in order for me to analyse it. The negative binomial regression analysis results produced in this thesis have been included in a research paper, with Dr Fletcher and her team, and submitted to a journal. The paper is currently under review.

The individuals in the UK2, BBCS and COGS datasets used throughout this thesis, have been collected, genotyped and prepared for analysis by external researchers. I did not perform the imputation of the BBCS and UK2 studies.

BMI summary data based on 2010-2015 meta-analysis of GWAS data conducted by the Genetic Investigation of Anthropometric Traits (GIANT) consortium, which is available in the public domain, was used to conduct the analyses in chapter 4.

Signed: [redacted]

Date: 14th September 2017

Ethical Approval

The project investigators for each of the studies used in this thesis, UK2 GWAS, BBCS GWAS and COGS, have confirmed to me that the approvals for their studies have been granted, although the specific approval numbers were not available to me. The names of the ethical bodies that have approved the UK2 GWAS, BBCS GWAS and each of the BCAC studies used in the COGS are given in **Appendix 1: Tables 1,2 & 3**. The analyses performed in this thesis were covered by the original approvals.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Frank Dudbridge, for his continued guidance and support over the last four years. Thank you for your encouragement, for giving me the opportunity to attend various conferences and meetings, and for always making time for me. It has been an honour to learn from, and work with you.

I would also like to thank Dr Olivia Fletcher from the Institute of Cancer Research (ICR) for giving me the opportunity to analyse CHi-C data.

Thanks to my funder, Breast Cancer Now, for making this thesis possible.

Thank you to my friends, old and new, for the support you have given me. To all of my officemates/friends, both at LSHTM and the University of Leicester, thanks for all your advice and understanding, for welcoming me, and for joining me for my 100 cups of tea a day. My school friends, thank you for the being such great friends, who never fail to make me laugh. Ellie, thanks for always being there for me over the years, and for the many dinner dates during stressful times. Sarah, thank you for being a lovely friend, housemate, and support during my degree.

Lastly, I would like to thank Ryan, my siblings and my parents. Ryan, thank you for offering me constant encouragement, understanding and support throughout my time at university. Thank you Daniel, Liam and Alex for being there for me when I have needed it most. Mum & Dad, thank you both for everything you have done for me. You have always been there for me, and have offered continuous love & support over the years, for that I am extremely grateful.

Abstract

With breast cancer being a highly prevalent complex disease that affects many women worldwide, research over the years has focused on establishing underlying breast cancer risk factors. Understanding how, and why the disease develops will potentially reduce the number of women developing breast cancer, or increase the number of women being diagnosed at an earlier stage of development. The disease has been shown to be a highly polygenic trait, so in order to learn more about the disease, this thesis focuses on the polygenic basis of breast cancer. Two breast cancer GWAS, the UK2 and BBCS, and the COGS were used to conduct the analyses presented in this thesis.

Using current chip heritability estimation methods, it was estimated that just under half of the genetic variation explained on the liability scale could be explained by genotyped SNPs. Common SNPs (MAF > 0.1) were shown to explain a substantial proportion of this variation, and the variance explained by each chromosome was shown to be linearly related to chromosome length, which indicated that variation is spread evenly across the genome. With BMI and age at menarche shown to be breast cancer risk factors, it was examined whether a shared polygenic basis exists between breast cancer and BMI, and whether there was evidence to suggest that breast cancer polygenic scores interact with either BMI, age at menarche or individual SNPs, to have an effect on breast cancer risk. With many susceptibility loci mapping to non-protein-coding regions of the genome, it was also tested whether individual genome-wide significant loci interact with other regions of the genome to influence breast cancer risk.

These results give further insight into the polygenic architecture of breast cancer, and provide further evidence that a large number of genetic variants explain much of the genetic variation in breast cancer.

Contents

List of Tables

List of Figures

List of Abbreviations

ALS	Amyotrophic lateral sclerosis
AVENGEME	Additive variance explained and number of genetic effects method of estimation
BBCS	British Breast Cancer Study
BMI	Body mass index
CDG-PGC	Cross-Disorder Group of the Psychiatric Genomics Consortium
CHi-C	Capture Hi-C
CHiCAGO	Capture Hi-C Analysis of Genomic Organization
CI	Confidence interval
D.f.	Degrees of freedom
ER	Estrogen receptor
GCTA	Genome-wide complex trait analysis
REML	Genomic restricted maximum likelihood method
GWAS	Genome-wide association study
iCOGS	Illumina Collaborative Oncological Gene-Environment Study
ISC	International Schizophrenia Consortium
LD	Linkage disequilibrium
LDSC	LD score regression
MAF	Minor allele frequency
MS	Multiple sclerosis
NHS	National Health Services
PRS	Polygenic risk score
QC	Quality control
SNP	Single-nucleotide polymorphism

Chapter 1 Introduction

In the United Kingdom, over 50,000 women a year are diagnosed with breast cancer, making it the most common type of cancer amongst British women (1). The disease is also common worldwide, with the number of women diagnosed with breast cancer ever increasing. With many women developing the disease, the importance of research into the underlying risk factors is evident. However, discovering all factors associated with breast cancer risk has not been an easy task as both environmental and genetic factors influence disease risk, making it a complex disease. Not only is breast cancer a complex disease, it is also a polygenic trait whereby many genetic mutations affect disease risk. With breast cancer having a polygenic basis, genetic epidemiology has been used to gain a better understanding of how environmental factors and genes influence disease risk in the human population (2). Improving our knowledge of the genetic risk factors and how they interact with the environment, will enable the development of individual breast cancer risk prediction, and risk-stratified screening to be implemented in the future. Thus, improving the chances of early diagnosis for breast cancer, which in turn could decrease the number of deaths from the disease. Understanding the aetiology of breast cancer could help to develop novel treatments, as well as improve the effectiveness of existing treatments, as it becomes possible to assign treatments based on a persons' DNA.

Over the years, various studies and statistical methods have been developed and used to identify breast cancer susceptibility genes and alleles linked to breast cancer risk. Approaches used have changed due to both technological advances and study costs. The price of genotyping over the years has been decreasing, meaning that a larger number of individuals can be genotyped today, then they could when the first GWAS was conducted. The first section of this chapter gives a brief introduction of the genetic studies that have been used to make such breast cancer discoveries. The remainder of the chapter highlights the forms of bias that can affect results, and what can be done to reduce the risk of such bias occurring.

1.1 Overview of the genetic epidemiology of breast cancer

With developments in both technology and statistical techniques over time, we have seen advances in the field of genetic epidemiology in regard to breast cancer. In this section, a brief explanation of the main study types that have been used to help shape our understanding of this disease in women of European descent, will be given. The choice of study design is dependent on the aim of the study. Earlier studies tended to focus on establishing whether a phenotype was familial, and if this was found, researchers then quantified how much variation in the phenotype could be explained by genetic variation. Research then began trying to explain some of the genetic variation by investigating where in the genome risk variants were located, and which variants were risk variants.

1.1.1 Familial aggregation

The aim of a familial aggregation study is to establish whether a specific disease clusters in families which, if shown, could indicate that inherited genetic factors influence disease risk (3, 4). In order to examine for the presence of familial aggregation, the family histories for cases and controls are compared (5). If the disease is shown to be more prominent amongst the relatives of the case subjects than the control subjects, then it is possible that the disease aggregates in families because of inherited genetic factors that influence disease risk. However, it is possible that this aggregation could also be due to a shared environment, or a combination of environmental and inherited genetic factors (gene-environment interactions) (5-7). Familial aggregation studies have shown that breast cancer does aggregate in families. Disease risk increases for women with a family history of the disease, and for first-degree relatives, the relative risk doubles (8-10).

1.1.2 Twin and family studies

With breast cancer shown to aggregate in families, research then focused on estimating how much phenotypic variation could be explained by genetic variation. Both twins and family members were used to establish whether the observed aggregation could be due to genetics, and/or shared environment (11). The heritability of a trait, this being the amount of phenotypic variation that is due to genetic variation, can be estimated and used to assess whether future genetic studies should be carried out. If the heritability estimate indicates that genes do not influence disease risk (heritability estimate = 0), then there would be no justification for carrying out future genetic studies.

There are two main types of heritability, broad-sense heritability and narrow-sense heritability. Broad-sense heritability (H^2) can be defined as the proportion of phenotypic variation (V_P) that is due to additive genetic variation (V_a), dominance genetic variation (V_d) and epistatic variation (V_{ep}) (12):

$$H^2 = \frac{V_a + V_d + V_{ep}}{V_P}$$

Additive variation is the proportion of phenotypic variance caused by the additive effect of alleles, whereas the dominance genetic variation is the genetic variation caused by dominant alleles. Epistatic variation is caused by the joint effect of multiple loci, for example some of the variation in a phenotype may be explained by two loci interacting. Narrow-sense heritability (h^2) on the other hand is the proportion of genetic variation that is due to additive genetic variation (V_a) only. Narrow-sense heritability can be defined as (13):

$$h^2 = \frac{V_a}{V_P}$$

With breast cancer being a complex disease with a binary outcome, a multifactorial estimation method can be used to estimate the disease heritability. A heritability measure for a binary trait can either be on the observed, 0/1, scale (h_o^2), or the unobserved liability scale (h_{liab}^2). On the observed scale, individuals are coded as either 0 or 1, these being individuals who are not observed to have the trait, or have the trait, respectively. For the liability scale, individuals each have a continuous normally distributed score, with the trait being observed in an individual if their liability score exceeds a certain liability threshold, T. The liability threshold, T, is equal to the quantile function of (1-K) for the normal distribution, with K being the disease prevalence. The liability scale is often used when predicting trait heritability for binary traits.

The sibling recurrence risk ratio (λ_S), the given risk for a sibling of an affected individual divided by the population prevalence (6), can be used to produce a narrow-sense heritability estimate on the unobserved liability scale. Wray et al. (14) have produced a narrow-sense heritability estimate for breast cancer on this scale, using an estimate of the sibling recurrence risk ratio and the following equation (6, 14):

$$h_{liab}^2 = \frac{2[\,T - T_1\sqrt{1-\left(T^2-T_1^2\right)\left(1-\frac{T}{i}\right)}\,]}{i + T_1^2(1-T)},$$

where, $T_1$is the quantile function of (1-$\lambda_S K$) for the normal distribution, and $\mathrm{i}=\mathrm{z}/\mathrm{K}$,where z is the normal density at T.

Wray et al. estimated the narrow-sense heritability for breast cancer on the unobserved liability scale to be 44%. The narrow-sense liability scale estimate was estimated for European women whilst assuming a prevalence = 3.6% and λ_S = 2.5 (6, 14, 15). With twin-study heritability estimates for complex-diseases tending to be ~50% (16), the estimate produced for breast cancer is similar to what is typically observed.

1.1.3 Family-based linkage analysis

With breast cancer being shown to have a genetic basis, studies were then conducted to try and establish where in the genome disease genes were located, in order to gain a better understanding of disease risk. Family-based linkage analysis was an analysis used to identify the chromosomal location of disease genes, using data based on families with numerous affected family members, and examining for linkage between loci. The analysis is based on the finding that during meiosis, loci that are physically close together on the chromosome tend to be linked. This would mean that the loci are inherited together by the offspring, from the parent, more often than expected if inheritance were independent under Mendel's second law (17-19).

Using family-based linkage analyses, BRCA1 and BRCA2 gene mutations were discovered to be associated with breast cancer risk in the 1990's (20). Results indicated that between 57%-65% of women with a BRCA1 gene mutation, and 45-49% of women with a BRCA2 gene mutation will develop breast cancer before the age of 70 years (21, 22). The highly penetrant genetic variants found in the BRCA1 and BRCA2 genes are however rare in the general population, with less than 1% of the population actually having either a BRCA1 or BRCA2 mutation (23). It has also been estimated that mutations in these two genes account for ~16% of the familial breast cancer risk (24-26), this being the risk of disease that aggregates in families. With BRCA1 and BRCA2 gene mutations estimated to not explain a large proportion of familial breast cancer risk, and shown to be rare in the general population, research then focussed on examining the genetic variation of breast cancer in the general population.

1.1.4 Population based genetic association studies

Population based genetic association studies have been fundamental for the development of our understanding of complex diseases in the general population. They have enabled associations between genetic variants and phenotypes, within different populations, to be discovered. With this form of association study being population based, collecting samples should be easier than family-based studies. In this subsection, the focus will be on candidate gene studies and genome-wide association studies (GWAS).

1.1.4.1 Candidate gene study

A candidate gene study typically involves analysing between 5-50 single-nucleotide polymorphisms (SNPs) within a specific gene. SNPs, a common type of genetic variation in humans, are single base mutations. A candidate gene region is a gene region of interest, specifically chosen for analysis based on either linkage analysis results, which have indicated that the disease variant could be within that specific region, or through existing knowledge of the biology of the disease (27, 28). Candidate gene studies have identified that genetic mutations within the genes *ATM, CHEK2, BRIP1, PALB2,* and *RAD50*, genes known to be involved in DNA repair, cause an increase in breast cancer risk (29). The main advantage of candidate gene studies is that they can cost less to conduct than other study types, notably GWAS, as only specific pre-selected SNPs will genotyped and analysed. However, by focussing on specific gene regions, you miss out on capturing the genetic variation across the rest of the genome. Genetic variants that map to other genes may not be considered if the function of the gene is either unknown or not fully understood, therefore hindering progress, especially if a disease is polygenic.

1.1.4.2 Genome-wide association study (GWAS)

With breast cancer being a polygenic disease, research adopted a hypothesis-free approach in order to identify potentially causal genetic variants across the genome. In recent years, with advances in technology, the completion of the Human Genome Project, the International HapMap Project, the 1000 Genomes Project, and the decrease in genotyping costs, it has become possible to undertake large scale association studies (30). One such example is a GWAS, which involves analysing SNPs across the whole genome, testing whether any individual SNPs could be associated with the phenotype of interest (31). The cost to genotype genetic variants across the whole genome is very expensive, so a set of tagging SNPs that reflect the genome are used instead. A tag SNP is a SNP that is used to represent a group of

SNPs that are in high linkage disequilibrium (LD), these being SNPs that are highly correlated with each other. If a tag SNP is shown to be associated with the phenotype, then it is likely that the true disease variant will be tagged to that SNP, if the tag SNP itself is not the causal variant. GWAS are hypothesis-free as focus is not on analysing specific tag SNPs that map to certain regions of the genome, or choosing which tag SNP to analyse based on the results of previous analyses. No biological assumptions are made, and an understanding of the phenotype is not needed in order to conduct a GWAS.

In 2003, before GWAS were published, the Human Genome Project produced an outline of the average human genome based on the genome sequence for a small number of individuals (32). The project aimed to discover the order and sequence of all base pairs across the genome (over 3 billion base pairs), and identify the genes across the whole genome (33, 34). The output produced by the Human Genome Project cannot be used to identify genetic differences between individuals, therefore it does not provide much help when testing for associations between genetic variants and disease risk. The output produced does however show that humans have fewer protein-coding genes than previously thought. Before the project began, the estimated number of human genes was thought to be as high as 120,000 (35, 36), but once the final draft was published in 2003, this number fell to approximately 19,000 genes (37).

With researchers using GWAS to identify causal variants, genetic variants across the genome needed to be identified, in order to establish tag SNPs. In 2002, the International HapMap Project was launched, with the aim of the project being to map genetic variants across the human genome. Over 1 million SNPs had been identified once the third, and last, phase of this project was released in 2009 (38-40). In 2008 the 1000 Genomes Project was launched, another project that aimed to map common genetic variants (frequency > 1%) in humans, using whole-genome sequencing, deep exome sequencing and dense microarray genotyping on individuals taken from different populations (41). Once the project ended in 2015, over 84 million SNPs had been sequenced for 2,504 individuals from 26 populations (41).

In 2005, one of the first GWAS using a SNP array was published by Klein et al. (42, 43), with the study focussing on age-related macular degeneration. With only 96 cases and 50 controls being analysed, compared to recently conducted studies, this was a very small study. In 2007, the first breast cancer GWAS was published (44), and since then, more than 90 individual genetic variants have been shown to be associated with breast cancer risk (44-51). However, the effect sizes of many of the individual variants discovered so far have been small. The odds ratios for GWAS hits tend to be less than 1.3, which makes them individually unsuitable for risk prediction, as prediction accuracy would be compromised by such low effect sizes (43, 52, 53). Also, it has been estimated that collectively discovered common genetic variants only explain ~16% of the familial risk of breast cancer (51), which was the same estimate given for the mutations within BRCA1 and BRCA2 alone.

With much of the genetic variation in many disease yet to be explained, and effect sizes for discovered SNPs being small, it has been widely accepted that GWAS have been underpowered to detect many associated genetic variants. To improve power, in order to detect associated variants, a much larger number of genotyped individuals are needed. Genotyping is expensive, so to be able to increase the number of individuals genotyped, yet keep cost down, the number of SNPs genotyped has to be reduced. To reduce the number genotyped SNPs, yet to make sure that the variants genotyped are informative, researchers have used custom arrays. Loci are included on the custom array if they have previously been shown to have some form of relationship with the phenotype of interest. Some researchers have decided which SNPs to genotype, based on findings from GWAS, along with other candidate regions of interest. The Illumina Collaborative Oncological Gene-Environment Study (iCOGS) array (45, 51) is a custom array, which has been used to genotype over 200,000 SNPs, in more than

100,000 women (51).The SNPs were chosen to be genotyped on the array, based on findings from prostate, ovarian and breast cancer GWAS. Using the iCOGS custom array to genotype breast cancer cases and controls, researchers have been able to increase the number of breast cancer susceptibility variants identified, and replicate previous GWAS results (45, 51). The familial risk estimate (~16%), given previously, included the associated breast cancer variants that were discovered using the iCOGS array. Many of the susceptibility variants discovered to date, have been identified using genetic data produced using the iCOGS array.

In the near future, it is expected that many more variants will be discovered, as study sample sizes are set to increase further. Recently a new custom array has been developed, the OncoArray. The custom array is being used to genotype over 500,000 SNPs in ~450,000 individuals, with this including women who have been diagnosed with breast cancer (54). The SNPs genotyped include over 200,000 tagging SNPs, which provide genome-wide coverage for the majority of common variants (54). The genotyped SNPs also consist up to date susceptibility variants and loci that have previously been identified through various breast, ovary, colon, prostate or lung cancer studies (54). The UK Biobank have just released genotype data, which includes SNPs that offer genome-wide coverage, for up to 500,000 individuals. There are set to be many studies published, for a wide range of traits and diseases, using this data. In 2012 the 100,000 Genomes Project, a project that aims to whole-genome sequence over 70,000 NHS (National Health Service) patients by 2017, was launched (55). The project aims to whole-genome sequence patients who had been diagnosed with cancer, or a rare disease, with the families of patients with rare diseases also being sequenced. The data will be used by the NHS as part of a new genomic medicine service, and will also be available for use in research. With the release of these datasets, we will see many more susceptibility SNPs being discovered for breast cancer, as well as other complex diseases and traits.

From the genetic studies conducted to date, it is evident that many genetic variants influence breast cancer risk. However, the variants discovered so far do not collectively explain a large amount of the genetic variation for breast cancer. It might be that some of the “missing” heritability for a disease could be explained by many common causal variants of low effects (12). A polygenic score analysis can be used to test whether a trait is affected by a combination of SNP effects. The analysis is based on the polygenic model proposed by Fisher (56), a model whereby disease is affected by many small effects. This analysis has been used in breast cancer studies (52, 57) to further understand how a combination of genetic variants influence disease risk.

1.1.4.3 Polygenic risk score (PRS) analysis

In recent years, polygenic risk scores (PRS) have been used to explore the polygenic basis of many diseases and traits. Both training and replication samples are needed in order to examine whether multiple SNPs with small effect sizes affect disease risk. The two samples could be based on two different studies containing cases with the same disease, two different studies that each contain cases with a different disease, or one study that has been split into two studies internally. Choice of training and replication sample depends on the aim of the analysis.

A polygenic risk score analysis tests whether the genetic effects estimated from a training sample, can be used to predict risk of disease in the replication sample, by constructing a PRS and testing whether the PRS is associated with the replication sample trait. If an association is shown, then the result suggests that the training sample SNP effects, can be used to predict the replication trait. If the same trait is used across the samples, and an association is shown, then this would suggests that disease has a polygenic basis. Studies, including schizophrenia (58) and BMI (59) studies, have used polygenic scoring to explore whether a collection of genetic variants are associated with disease risk, in order show that a trait has a polygenic basis (60). To conduct a polygenic score analysis, a PRS for each subject in the replication sample, j, is constructed using training sample SNP effects, $\hat{\beta}_i$, for each independent training sample SNP, i.

If the training sample trait is binary, each $\hat{\beta}_i$ is estimated using a logistic regression model. For a quantitative trait, a linear regression model is used.

For each subject j, a PRS is constructed such that:

$$PRS_j = \sum_{i=1}^{m} \hat{\beta}_i x_{ij}$$

Where, m is the total number of independent SNPs, and x_{ij} are the allele dosages at SNP i, for individual j.

In breast cancer, Machiela et al. (57) have used polygenic score analysis to assess whether a breast cancer prediction model, based on identified susceptibility breast cancer genetic variants, could be improved further by including common genetic variants in the model. The analysis was based on 1,145 breast cancer cases and 1,142 controls, genotyped as part of the Nurses' Health Study. The subjects were split into ten roughly equal subsets, with one set assigned as the training set, and the other nine the replication set. A polygenic score was constructed for the individuals in the combined replication set, using the estimated SNP effects from the training set. This was repeated until all ten individual subsets have been used as the training set, with the remaining nine sets as the replication set. Each polygenic score analysis was therefore conducted ten times. A risk score was first constructed using 13 susceptibility breast cancer risk variants, and on average, there was found to be a significant association between the PRS and breast cancer outcome in the replication sample (*p*-value = 5.83 x 10^{-17}). Machiela et al. then gradually included up to ~60,000 SNPs into the score, this included SNPs that had not yet been shown to be associated with breast cancer risk. Machiela et al. then used the area under the ROC (receiver operating characteristic) curve to assess whether including a larger number of SNPs into the score, improved predictive accuracy for breast cancer outcome. By averaging the AUC over the tenfold cross-validation analyses, and observing that none of the polygenic scores which included the common variants were significantly associated with breast cancer outcome in the replication sample, Machiela et al. (57) concluded that there was no evidence to suggest that breast cancer risk prediction could be improved by adding a larger number of common variants to the model. The sample size for this study was originally small, and was made even smaller by splitting the sample into approximately 10 subsets, therefore each subset analyses would have been underpowered to detect an association between score and breast cancer outcome. The authors acknowledged that this, along with the original sample size, would have meant that they would have had low power to detect any score and breast cancer outcome associations.

Using a much larger sample size, Mavaddat et al. (52) have constructed a polygenic score based on the SNP effects for 77 published susceptibility genetic variants, estimated using ~90,000 women of European descent, who were genotyped as part of the Collaborative Oncological Gene-Environment Study (COGS) (45). Mavaddat et al. examined whether breast cancer risk could be stratified by PRS, with the polygenic score being constructed using only the 77 published breast cancer susceptibility loci. It was shown that by computing the odds of disease using odds ratios, and then adjusting by the PRS that the odds do change. This was shown to be the case for women who had a family history of breast cancer, and for women without a family history of the disease. These therefore suggests that a breast cancer PRS can be used to predict breast cancer risk.

Polygenic scores are more useful in predicting disease risk than individual SNPs because of their small effect sizes. Studies have tended to construct scores using only susceptibility variants, however it could be more beneficial to produce a polygenic score based on all genotyped SNPs (60). Many associated loci may not be reaching genome-wide significance because the studies used are underpowered to detect associations, because of their sample size. This means that unidentified disease associated SNPs are being left out of many PRS analyses, therefore jeopardising the accuracy of the analysis results. Machiela et al. (57) did not find evidence that a less restricted polygenic score was better at predicting breast cancer risk than a score based on genome-wide significant SNPs, but this study would have suffered from being underpowered. For other complex diseases, such as schizophrenia (61) and multiple sclerosis (62), polygenic scores constructed using both a larger number of individuals and a larger number of genetic variants, have been shown to be associated with disease outcome. Therefore, it would be beneficial to examine whether, in a larger study, a score based on a larger number of SNPs can be used to predict breast cancer outcome.

1.1.5 Discussion

In this chapter so far, the genetic epidemiology of breast cancer has been briefly discussed. As it become cheaper to collect and analyse larger samples, and as our understanding of the disease changes, our choice of preferred study design will also change. We have learnt, like many complex diseases, that breast cancer is a polygenic disease whereby many genetic variants of small effect influence disease risk. For many diseases, polygenic scoring has been used to gain a better understanding of the polygenic basis of the disease. Known susceptibility variants have not been shown to explain a large amount of the estimated genetic variation for most complex diseases. With it being difficult to identify all genetic risk factors associated with disease risk, and with breast cancer being a polygenic disease, much more of the genetic variation for the disease may be explained if we were to consider all genotyped SNPs in analyses (60).

In the next section of this chapter, an overview of the quality control (QC) procedure that should be implemented before analysing genetic data will be given.

1.2 Quality control and linkage disequilibrium removal

GWAS, and other SNP based studies, are large-scale studies that involve thousands of genetic variants and subjects which, due to their large size, makes them susceptible to bias. It is important to check and remove various forms of bias, so that a meaningful conclusion can be made about the phenotype of interest. The inclusion of either SNPs or subjects that incur bias could cause inaccurate results, so QC should be implemented on both subjects and SNPs before data is analysed, in order to reduce the risk of this happening. Various tests and methods have been developed to identify and correct for bias when analysing genetic data. In this section, the different types of bias that GWAS data is prone to, and how to test and correct for it will be discussed. For the majority of the QC, PLINK version 1.90 (63) has been used to test and filter out the SNPs or subjects that fail QC.

If a specific analysis assumes that there is independence across SNPs, these include a polygenic score analysis and a GWAS analysis, then the correlation between SNPs should be assessed. LD-thinning methods, implemented in PLINK, can be used to both measure the correlation between SNPs, and then reduce the level of correlation between SNPs. LD-thinning methods are explained later in this section.

First QC is applied to the genotyped SNPs as it is better to remove troublesome SNPs first, as this could potentially save having to remove subjects. With reduced statistical power being an issue for many genetic studies, it would be best to try and retain as many subjects in each study as possible.

1.2.1 SNP quality control

1.2.1.1 Minor allele frequency (MAF)

For a given SNP in a specific population, the minor allele frequency (MAF) is the frequency of the least common allele. SNPs with a low MAF should be removed as they have reduced statistical power, and could be more susceptible to errors (64).

The MAF is calculated as:

$$\text{MAF} = \frac{freq(AA) \times 2 + freq(Aa) \times 1 + freq(aa) \times 0}{2N} ,$$

where, A is the minor allele and a is the alternative allele, $freq(AA)$ is defined as the number of individuals with genotype AA, $freq(Aa)$ is the number of individuals with genotype Aa, $freq(aa)$ is the number of individuals with genotype aa ,and N is the total number of individuals.

PLINK can be used to calculate the MAF and remove any SNPs with a MAF less than a stated value using the “--freq” and “--maf” options. SNPs in this thesis were retained for analysis if they had a MAF greater than 5%.

1.2.1.2 Missing rates

The SNP missing rate can be used to measure the amount of genotype information missing for a given SNP. This measure of missingness can be used to assess genotyping quality, with high missingness potentially indicating poor reliability in the data (64). The PLINK “--geno” option can be used to filter out SNPs with a missing rate greater than a given threshold. For the QC carried out in this thesis, the threshold used was 5%, which meant that only the SNPs with a 95% genotyping rate were retained.

1.2.1.3 Hardy-Weinberg equilibrium

An important concept in population genetics is the Hardy-Weinberg equilibrium (HWE), a concept first described by Hardy and Weinberg separately in 1908 (65, 66). The HWE theorem states that for a given large population of diploid organisms and non-overlapping generations, both the genotype and allele frequencies should remain constant from generation to generation when the conditions of no mutation, no migration and no selection hold (67).

Through calculation of the probabilities for genotype arrangements, the exact test can be used to examine whether SNPs deviate from HWE. Wigginton et al (68) state that if we know the number of heterozygous Aa genotypes, and the number of a alleles and A alleles for a given genetic variant, whilst assuming that each individual has $2N$ alleles for N individuals, then we can calculate the number of AA aa homozygous genotypes using the following formulae:

$$n_{AA} = \frac{n_A - n_{Aa}}{2}$$

and

$$n_{aa} = \frac{n_a - n_{Aa}}{2}$$

Where, n_A is the number of the A allele, n_a is the number of a allele and n_{AA} and n_{aa} are the number of AA and aa homozygous genotypes, respectively.

If the number of heterozygotes are known, then under the assumption of HWE, the probability of observing n_{Aa} heterozygotes for N individuals with n_A minor alleles is (68):

$$P(N_{Aa} = n_{Aa}|N, n_A) = \frac{2^{n_{Aa}} N!}{n_{AA}!\, n_{Aa}!\, n_{aa}!} \times \frac{n_A!\, n_a!}{(2N)!}$$

with the exact test *p*-value being the sum of these probabilities.

PLINK can be used to filter out by using “--hwe” option, and stating an exact test *p*-value threshold. This option in PLINK can be used to remove SNPs that have an exact test *p*-value less than a specified threshold, with the null hypothesis that HWE holds, failing to hold for those SNPs. If deviation from HWE is identified, then it could be indicative that there are problems with either genotyping, population stratification or inbreeding (68). Population stratification is explained later in this chapter. Care should be taken when testing for deviations from HWE as an observed deviation could actually be caused by an association between the genotypes and phenotype, and because of problems with either genotyping, population stratification or inbreeding (31). For the QC carried out in this thesis, the exact test *p*-value threshold used was 5 x 10^{-6}.

1.2.2 Sample quality control

Once QC has been implemented on SNP data, focus moves onto establishing whether bias is present amongst the individuals genotyped in the study.

1.2.2.1 Discordant sex information

With the analyses conducted in this thesis aiming to improve our understanding of breast cancer in women, the genetic data used to conduct the analysis should only contain individuals who are women. Individuals should be excluded if, genetically, they are not female. The “--check-sex” option in PLINK can be used to assess whether the number of X chromosomes matches the gender of the subject. Women have two X chromosomes (XX), so individuals in a study shown not to have two X chromosomes should be removed from the dataset, and not used in any of the analyses conducted in this thesis.

1.2.2.2 Missing genotypes

DNA quality can be assessed by measuring the number of genotypes each individual has missing. If genotyped individuals are found to have a high number of missing genotypes, it would suggest that DNA quality might be poor (64). The “--mind" option in PLINK can be used to filter out the individuals that have too many missing genotypes, based on a given missing rate. Individuals who have a missing genotype rate greater than the stated rate are removed from the study. For the studies used in this thesis, any individual with a missing rate > 5% was not included in the analysis.

1.2.2.3 Heterozygosity

Heterozygosity is the proportion of non-missing genotype calls where, for a given genotype, the two alleles are different (heterozygous) (69). If heterozygosity is greater than the expected heterozygosity, it is an indication that either the quality of the sample is poor, or that the data is contaminated. If heterozygosity is low, it could be an indication that either inbreeding or population stratification are present (70). The “--het" option in PLINK can be used to produce an output file which presents the observed number of homozygous genotypes (o_{hom}), and the number of non-missing genotypes per individual (n_{gen}). These two variables can be used to calculate the observed heterozygosity rate per individual, using the following formula:

$$\frac{n_{gen} - o_{hom}}{n_{gen}}$$

The heterozygosity rate, for each individual, is then examined. If the rate is ± 3 standard deviations away from the mean, then it is concluded that the heterozygosity is different from the expected heterozygosity for that individual (71).

1.2.2.4 Relatedness between subjects

For non-family based studies, it is assumed that the individuals within the study being analysed are not related to each other. Relatedness between subjects can easily arise in studies that contain a large number of individuals; it is not safe to simply assume that relatedness is not present, even if at the data collection stage related individuals were not recruited into the study. Relatedness between individuals can affect the accuracy of results, if the model or method used does not adjust for the relatedness between individuals. For example, when conducting a polygenic score analysis, if an individual in the training set is highly related to an individual in the replication set, the association between the polygenic risk score and the phenotype in the replication study could be inflated. Also analysing related individuals could cause results to reflect environmental effects, and not just the genetic effects (72).

KING version 1.4 (73) is a software package that can be used to measure the relatedness between a pair of individuals within a study, in order to determine whether two individuals are related to each other. KING is a robust relationship inference algorithm, robust because it adjusts for any population stratification present in the data. Both an estimate of the kinship coefficient (ϕ) and the probability of zero identity by descent (IBD) sharing (τ_0) are produced by KING, with both estimates used to assess the relatedness between a pair of individuals. The kinship coefficient, ϕ, is the probability that two alleles from each individual are IBD when chosen at random, with alleles being defined as IBD if they are inherited from the same ancestor (31). The probability of zero IBD sharing,τ_0, is the probability that two individuals share zero IBD alleles. The similarity between individuals can be measured based on the number of common alleles between the individuals for each genotype, this measure is known as identity by state (IBS).

Assume that there is HWE amongst SNPs, and that only an $IBD_{ij} = 0$ can produce an $IBS_{ij} = 0$ for a pair of individuals i and j. Then the proportion of SNPs with zero IBS can be estimated as (73):

$$\Pr\left(IBS_{ij} = 0\right) = \Pr\left(BB, bb \mid IBD_{ij} = 0\right) \times \Pr\left(IBD_{ij} = 0\right) = 2p^2(1-p)^2\tau_{0_{ij}}$$

Where, p is the reference allele (B) frequency for a SNP, b is the alternative allele, IBS_{ij} is the number of IBS alleles between individuals i and j.

This can then be used to estimate the probability of zero IBD between the two individuals i and j, such that (73):

$$\hat{\tau}_{0_{ij}} = \frac{N_{BB,bb}}{\sum_m 2\hat{p}_m^2(1-\hat{p}_m)^2}$$

Where, $N_{BB,bb}$ is defined as the total number of SNPs, where the genotypes between the two individuals are different homozygotes (BB,bb). The number of SNPs between two individuals, where there is no missing genotypes in either pair, can be defined as m.

The genotype frequencies for the whole sample can be used to estimate the allele frequency $\hat{p}_m$ at the m-th SNP, such that (73):

$$\hat{p}_m = \frac{ind_{BB} + ind_{Bb}/2}{ind_{BB} + ind_{Bb} + ind_{bb}}$$

Where, at the m-th SNP, ind_{BB} , ind_{Bb} and ind_{bb} are defined as the total number of individuals with genotypes BB, Bb and bb, respectively.

Assuming HWE, and that population stratification may be present, the genetic distance between two individuals, i and j, in terms of the kinship coefficient can be modelled as (73):

$$E(X^{(i)} - X^{(j)})^2 = 4E(P(1-P))(1-2\phi_{ij})$$

Where $X^{(i)}$ and $X^{(j)}$ are the genotype scores, for individuals i and j, respectively, with this being defined by the number of the reference alleles for an individual. It is assumed that P is the allele frequency for a randomly chosen SNP for an individual, and it is possible for P to vary if population stratification is present.

When measuring the relationship across different families, the kinship coefficient can be estimated as (73):

$$\hat{\phi}_{ij} = \frac{1}{2} - \frac{1}{4}\frac{\sum_m \left(X_m^{(i)} - X_m^{(j)}\right)^2}{N_{Bb}^{(i)}} = \frac{N_{Bb,Bb} - 2N_{BB,bb}}{2N_{Bb}^{(i)}} + \frac{1}{2} - \frac{1}{4}\frac{N_{Bb}^{(i)} + N_{Bb}^{(j)}}{N_{Bb}^{(i)}}$$

Where, $N_{Bb,Bb}$ is defined as the number of SNPs where both individuals of a subject pair are heterozygous, with $N_{Bb}^{(i)}$ and $N_{Bb}^{(j)}$ being the total numbers of heterozygotes for individuals, i and j, respectively.

Manichaikul et al. (73) present a table that defines the relationship of a pair of individuals based on both the kinship coefficient and the probability of zero IBD sharing:

Relationship	Kinship coefficient (ϕ)	Inference criteria	Probability of zero IBD sharing (τ_0)	Inference criteria
Monozygotic twin	$\frac{1}{2}$	$> \frac{1}{2^{\frac{3}{2}}}$	0	< 0.1
Parent-offspring	$\frac{1}{4}$	$(\frac{1}{2^{\frac{5}{2}}}, \frac{1}{2^{\frac{3}{2}}})$	0	< 0.1
Full sibling	$\frac{1}{4}$	$(\frac{1}{2^{\frac{5}{2}}}, \frac{1}{2^{\frac{3}{2}}})$	$\frac{1}{4}$	(0.1,0.365)
Second degree	$\frac{1}{8}$	$(\frac{1}{2^{\frac{7}{2}}}, \frac{1}{2^{\frac{5}{2}}})$	$\frac{1}{2}$	(0.365,1-$\frac{1}{2^{\frac{3}{2}}}$)
Third degree	$\frac{1}{16}$	$(\frac{1}{2^{\frac{9}{2}}}, \frac{1}{2^{\frac{7}{2}}})$	$\frac{3}{4}$	(1-$\frac{1}{2^{\frac{3}{2}}}$, 1-$\frac{1}{2^{\frac{5}{2}}}$)
Unrelated	0	$< \frac{1}{2^{\frac{9}{2}}}$	1	> 1-$\frac{1}{2^{\frac{5}{2}}}$

Table 1-1: Relationship inference criteria based on the kinship coefficient and the probability of zero IBD sharing

Source: Edited version of table in Manichaikul et al. (73) - page 2868

Both the estimated ϕ and τ_0 can be used to define the relationship between two subjects. A pair of subjects where the estimated ϕ is greater than $\frac{1}{2^{\frac{5}{2}}}$ and the estimated probability of zero IBD sharing is greater than 0.1, are defined as being first-degree relatives (Table 1-1). One way to manage the relatedness between subjects is to remove one individual from every first-degree relative pair. This was the method chosen to prevent related subjects being included in the analysis.

1.2.2.5 Population stratification

When using cases and controls to conduct a GWAS, in order to test for an association between an allele and a phenotype of interest, an allele is identified as being associated with a trait if it is more frequent in cases than controls. Occasionally, allele frequency differences between cases and controls may be due to ancestry differences, and not through an association with the phenotype. This is known as population stratification and can occur when there is both a difference in allele frequency between sub-populations, and a difference in disease prevalence. It is important to make sure that population stratification is not present in the data being analysed, as this could cause spurious associations. One way to examine whether population stratification is present in data is to estimate, and examine, the genomic inflation factor for the data. An estimate of the genomic inflation factor can be produced by taking the median chi-squared test statistic (χ_1^2) across all SNPs, and dividing it by the expected median under the null distribution (27, 74). If the estimate is greater than one, then it is an indication that population stratification may be present in the data.

It is also possible to visually identify population stratification by plotting eigenvectors/principal components that represent the data. Principal-component (PC) analysis, a procedure used to convert a set of potentially correlated variables into a set of linearly uncorrelated variables using an orthogonal transformation, can be used to create the eigenvectors/principal components. The first and second eigenvector, the eigenvectors that explain most of the variation in the data, are plotted against each other. If there is shown to be more than one separate cluster of subjects, then the plot would suggest that population stratification is present. More than one cluster may also be visualised when plotting additional eigenvectors, such as the second and third eigenvector, which again would suggest that population stratification is a problem.

If population stratification is shown to be present in the data, genomic control and PC analysis are two different methods that can be used to adjust for it, in order to reduce the effect it has on the result. The genomic control method adjusts for population stratification by dividing the chi-squared statistic for each individual SNP association by the estimated genomic inflation factor. To control for population stratification using PC analysis, eigenvectors/principal components are created and then included as covariates in the regression model used to test the association between alleles and a trait (74).

1.2.3 Linkage disequilibrium (LD) removal

When conducting a GWAS or a polygenic score analysis, as independence across SNPs is assumed, it is important to check that SNPs used in the score are independent. Causal SNPs tend to tag the SNPs that they are in LD with, and increase the tagged SNPs association with the phenotype. This means that if a tag SNP is discovered to be associated with the phenotype, it is not necessarily the causal SNP, or truly associated with the phenotype of interest. If one were to conduct a polygenic score analysis without first removing high LD between SNPs, the SNP effects used to construct a polygenic score may be inflated, which could cause inaccurate results.

Foulkes (31) state that under the assumption of independence between two loci, the expected haplotype distribution should be as follows:

$$n_{11} = Np_A p_B \qquad n_{12} = Np_A p_b$$

$$n_{21} = Np_a p_B \qquad n_{22} = Np_a p_b$$

where, n is the number of individuals and $\mathrm{N} = 2\mathrm{n}$, as each individual, n, has two homologous chromosomes. The alleles on locus 1 and locus 2 are defined as Aa and Bb, respectively. Also, p_A and p_a denote the population frequencies for alleles A and a, with p_B and p_b denoting the population frequencies for B and b.

If locus 1 and locus 2 are correlated, then the observed counts will be different to those expected when assuming independence, such that (31):

$$n_{11} = N(p_A p_B + D) \qquad n_{12} = N(p_A p_b - D)$$

$$n_{21} = N(p_a p_B - D) \qquad n_{22} = N(p_a p_b + D)$$

With, scalar D representing the difference when independence cannot be assumed.

If D were to be close to 0, then the observed counts would be close to the expected counts under independence. This would then indicate little or no departure from LD.

D can be expressed in terms of both the joint probability of A and B, and the product of the individual allele probabilities, such that:

$$D = p_{AB} - p_A p_B$$

With, $D \neq 0$ if there is LD present

Lewontin (75) proposed a rescaled version of D, known as D' which can be expressed as:

$$D' = \frac{|D|}{D_{max}}$$

where, D_{max} is the theoretical maximum for the observed allele frequencies, which can be given by:

$$D_{max} = \begin{cases} \min(p_A p_b, p_a p_B) & D > 0 \\ \min(P_A p_B, p_a p_b) & D < 0 \end{cases}$$

The correlation coefficient can also be used to express LD, such that (76):

$$r^2(p_a, p_b, p_{ab}) = \frac{(p_{ab} - p_a p_b)^2}{p_a(1 - p_a) p_b(1 - p_b)},$$

with an r^2 close to 0 suggesting low correlation, and an r^2 close to 1 suggesting high correlation between SNPs.

There are two main LD-thinning methods commonly used to reduce LD between SNPs, LD-based pruning and LD-based clumping. Both methods can be implemented using PLINK, with both methods retaining one SNP from a group of SNPs that have been identified as being in LD. The methods do however differ in how the retained SNP is chosen, but both methods can use r^2 to measure LD between SNPs.

LD-based pruning is implemented using PLINKs' "--indep-pairwise" command and stating a window size, the number of SNPs to shift the window by, and an r^2 value. The window size is measured in SNPs and specifies the number of SNPs within a subset, with the LD between the SNPs in each window being measured. This is then repeated for each shift in SNPs. If the r^2 between any of the SNPs is greater than the specified r^2, then one SNP from each correlated group will be randomly retained.

LD-based clumping is similar to LD-based pruning except that SNPs are first ranked by their individual association with the phenotype of interest. In a correlated group of SNPs, the SNP with the strongest association with the phenotype is retained. By retaining the SNP with the strongest association with the phenotype, it will reduce the risk of removing causal variants from the analysis.

1.3 Discussion

The methods and studies used to gain a better understanding of the underlying genetic architecture of a disease have changed over time due to improvements in our knowledge of the disease, sample size, data quality, technology and the reduction of genotyping costs. As sample sizes have been increasing, new developments in software and computational methods have enabled analyses to be performed on larger sample sizes, as well as a greater number of genetic variants. Even with sample sizes increasing, studies, such as GWAS, still suffer from being underpowered.

Over 90 individual genetic variants have been shown to be associated with breast cancer risk, but like most complex diseases, these variants collectively only explain a small proportion of the heritability for breast cancer (44-51). Much of the genetic variation for the breast cancer may be explained by a combination of SNPs that have a small effect on disease risk, that have not yet reached genome-wide significance. Polygenic scores can be used to test whether a combination of SNP effects is associated with a trait of interest, with the hope that once sample sizes are large enough, the scores can be used to accurately predict risk of disease. Assuming that a score is to be constructed using 1,000,000 SNPs, and that 1% of these SNPs have an effect on breast cancer risk, it has been estimated that we would need to collect a training sample of approximately 100,000 subjects, for a breast cancer polygenic score to accurately predict breast cancer risk (15). The estimated number training sample individuals needed then increases, if the proportion of SNPs that have an effect on disease increases (15).

Even though the cost of genotyping has been decreasing over the years, it is still considered to be expensive to genotype a sample. The high cost can make it hard to genotype enough individuals to have a highly powered study, in order to detect genome-wide significant associations, or to predict disease risk. Collecting a large number of cases is also difficult if the disease of interest is rare. In order to increase sample sizes in studies, consortium datasets have been generated. A consortium is a collaboration between institutions and researchers, with the goal of combining many studies and creating large datasets to be analysed. Depending on the analysis being carried out, the genetic data could include a combination of genome-wide significant SNPs, imputed SNPs, GWAS SNPs and SNPs genotyped on custom arrays. Variants genotyped on a custom array, such as the iCOGS array, are those that have been purposely selected. The selection of variants could be based on their relationship with the trait of interest. Being selective of which SNPs to genotype lowers the number of SNPs being genotyped, which decreases costs, thus enabling a larger number of subjects to be genotyped. Consortium data and custom arrays, have allowed samples of over 100,000 subjects to be studied. However, these larger samples have not always been genotyped for SNPs across the whole genome, meaning that parts of the genome have not been represented in analyses.

With sample sizes increasing, this being in terms of both the number of subjects and the number of genetic variants genotyped, the risk of bias occurring also increases. In order to improve the accuracy of results produced, it is important to make sure that the chances of bias occurring is reduced through QC measures.

1.4 Research questions and overview of thesis

The overall aim of this thesis is to gain a better understanding of the underlying polygenic architecture of breast cancer. With a better understanding of how many genetic mutations of small effect influence breast cancer risk, it will enable risk prediction to be possible in the future. Risk prediction in turn will allow the development, and implementation of risk stratification procedures. For instance, women who have been estimated to have a high risk of developing breast cancer, could be screened more frequently using established screening procedures, or screened at a younger age. This could increase the number of women detecting breast cancer in the early stages, which could decrease the number of women diagnosed with advanced stage breast cancer. Understanding the genetic mechanisms of the disease will also enable the development of breast cancer treatments, treatments that target certain genetic mutations, and personalised medicine.

The objective of this thesis is to:

1. Investigate whether a large number of SNPs could collectively explain the missing heritability for breast cancer.
2. Partition the genetic variation explained by genotyped SNPs to better understand how genetic variation is spread across the genome.
3. Examine whether there is evidence that a shared polygenic basis between breast cancer and body mass index exists.
4. Investigate whether there is evidence that PRS-body mass index, or PRS-age at menarche interactions exist.
5. Investigate whether there is evidence that any breast cancer derived PRS-SNP interactions exist.
6. Find evidence of physical interactions between known breast cancer loci, and other loci across the genome.

With breast cancer being a polygenic trait, in chapter 2, polygenic score analysis will be used to find evidence that confirms that breast cancer has a polygenic basis. It will be tested whether the estimated SNP effects from one breast cancer GWAS, can be used to predict breast cancer outcome in an independent breast cancer GWAS. If shown, this would indicate that breast cancer does have a polygenic basis, which would support other breast cancer findings. Once this has been examined, using three commonly used estimation methods, it will be estimated how much variation in breast cancer risk can be explained by a large number of common SNPs (MAF> 0.05). The estimates produced will be based on two European breast cancer GWAS, the UK2 study (48) and the British Breast Cancer Study (BBCS) (49, 77), as well as the Collaborative Oncological Gene-environment Study (COGS) (45). Estimates have been previously produced for both breast cancer and ER-negative breast cancer, but based on a smaller number of individuals. The estimates produced in this thesis will be based on all genotyped SNPs, not just genome-wide significant SNPs, and estimated using larger samples than those previously used to produce breast cancer chip heritability estimates. Producing breast cancer based chip heritability estimates for both GWAS and a study in which a custom array has been used, allows the genetic variation of breast cancer explained by GWAS SNPs to be compared to the variation explained by custom array SNPs. In breast cancer, this will be the first time that such a comparison has been made between a custom array, and a GWAS array. With the custom array allowing for more individuals to be genotyped, compared to a GWAS array, do the variants on the custom array explain nearly as much variation as a GWAS array? Or, is the difference large?

In chapter 3, this analysis will be taken further by partitioning the chip heritability estimates, in order to examine how genetic variation is spread across the genome. Genomic partitioning will be used to partition the genetic variation explained by the SNPs genotyped for each study, by chromosome, MAF and SNP annotation. In breast cancer, this will be the first time that partitioning analyses have been performed. This analysis is important as it has the potential to identify areas of the genome where causal variants are most likely to lie, if certain subsets are shown to explain a larger proportion of genetic variation, than other subsets.

In chapter 4, polygenic scores will be used to examine whether there is evidence to suggest that a shared polygenic basis exists between breast cancer, and body mass index (BMI). It will be the first time that this has been tested using polygenic scores, and if evidence of a shared polygenic basis is found, it could aid the development of novel treatments and procedures by enabling the two phenotypes to be studied together.

In chapter 5, polygenic scores and a case-only interaction analysis will be used to test whether there is evidence to suggest that either a PRS-BMI interaction, or a PRS-age at menarche interaction exist. Both BMI and age at menarche have been linked to breast cancer risk, but it is not known whether the presence of either of these risk factors influence the effect a breast cancer PRS has in predicting breast cancer risk. It will also be tested whether there is evidence to suggest that the effect a breast cancer derived PRS has on predicting breast cancer risk, is modified by any of the SNPs used to construct the score.

Finally, in chapter 6 it will be tested whether any significant physical interactions exist between known breast cancer susceptibility loci, and other loci positioned within 5Mb of the associated loci using Capture Hi-C (CHi-C) methods. With many of the discovered breast cancer susceptibility loci mapping to non-coding regions of the genome, it is not fully understood how the variants influence disease risk. This analysis aims to detect significant physical interactions that may explain how these significant loci influence breast cancer risk. With regard to the number of cell-lines and loci analysed at once, this is the largest CHi-C analysis to have been conducted for breast cancer and, as far as I am aware, it is also the largest CHi-C analysis to have been conducted for any disease.

With breast cancer being a polygenic trait, the work presented in this thesis focuses on using polygenic scores to gain a better understanding of the polygenic basis of breast cancer in individuals of European descent. Collectively, known genome-wide significant SNPs have not been shown to explain a large proportion of the genetic variation for the disease. It could be that a much larger proportion of the genetic variation in breast cancer is explained by SNPs that have not yet reached genome-wide significance. The analyses conducted in this thesis will be different to those previously conducted in breast cancer, as “deep” scores will be analysed. Machiela et al. (57) have previously assessed the ability breast cancer based polygenic scores have in predicting breast cancer risk, but with very small training and replication samples. Mavaddat et al.(52) have previously performed a PRS analysis and a PRS interaction analysis using a larger number of individuals, but their analysis was performed using a reduced number of SNPs. The analyses conducted in this thesis will be performed on the largest number of SNPs, genotyped in the largest number of individuals.

Chapter 2 Analysis of two breast cancer GWAS and the COGS

2.1 Introduction

2.1.1 Breast cancer datasets

Much of the analysis in this thesis has been performed on two European breast cancer GWAS, the UK2 study (48) and the British Breast Cancer Study (BBCS) (49, 77), and the Collaborative Oncological Gene-environment Study (COGS) (45). These studies have been used to explore the underlying polygenic architecture of breast cancer in order to further understand the genetic and environmental risk factors for the disease.

2.1.1.1 UK2 GWAS

In 2010, Turnbull et al. (48) first described the UK2 GWAS in a study that identified five previously unknown breast cancer susceptibility loci. Varghese et al. (78), a few years later, then used the study to find evidence of a shared genetic basis existing between breast cancer, and mammographic breast density.

The case-control GWAS consists of breast cancer cases collected through both the ICR Familial Breast and Ovarian Cancer Study (BOCS), and the Prospective study of Outcome in Sporadic versus Hereditary (POSH) breast cancer study. These studies recruited women of European descent using 23 UK based clinical genetics centres and UK oncology clinics. Women with breast cancer were included in the study if they had at least two affected first, or second-degree relatives in their family. However, women were excluded from the GWAS if they were found to have either BRCA1 or BRCA2 mutations. Controls were collected through the Wellcome Trust Case Control Consortium (WTCCC) study (79), which recruited controls from both the 1958 Birth Cohort, and the UK National Blood Service. Using an Illumina 670k array, 475,998

SNPs were genotyped in 3,628 cases, with an Illumina 1.2M array being used to genotype 5,190 controls.

2.1.1.2 British Breast Cancer Study (BBCS) GWAS

Johnson et al. (77) first described the British Breast Cancer Study (BBCS), and used it to find evidence that interactions between CHEK2*1100dekC, and other low-penetrance breast cancer susceptibility genes, possibly exist. Fletcher et al. (49) have also used the BBCS, along with other studies, to identify potential breast cancer risk loci and risk variants. Fletcher et al. identified 9q31.2 as a susceptibility risk locus, along with two genetic variants, rs3734805 and rs9383938.

The study contains breast cancer cases collected through UK based cancer registries, with the majority of the cases having been diagnosed with two primary breast cancers (bilateral breast cancer). The remaining cases were women who have at least two first-degree relatives who had previously been diagnosed with breast cancer. The controls used in the BBCS were the same controls used in the UK2 study. Therefore, controls were collected through the Wellcome Trust Case Control Consortium (WTCCC) study, and recruited from both the 1958 Birth Cohort and the UK National Blood Service. An Illumina 370k array was used to genotype 269,684 SNPs in 1,609 breast cancer cases, and an Illumina 1.2M array was used to genotype 5,190 controls.

For both GWAS, two versions of the data were available; one containing genotype and phenotype data for the GWAS in binary PED file format, and the other a MACH dosage file containing SNPs that had either been genotyped or imputed, for each GWAS subject. I did not perform the imputation, it was executed as part of the studies. The MACH dosage files were converted into best guess PLINK binary format using the "--dosage-mach" and "--make-bed" commands in GCTA.

In the analysis presented in this thesis, it should be assumed that genotyped SNPs have been used, unless stated. Imputed SNPs have been used when constructing risk scores across independent studies, in order to increase the number of SNPs in union between the two studies. This enabled a polygenic score to be constructed using a much larger number of SNPs than would have otherwise been possible if only genotyped SNPs were used. When imputed SNPs have been used in an analysis, it will be stated.

For both GWAS, I had data for genetic variants on chromosomes 1 to 22. For the UK2 GWAS, and not the BBCS GWAS, I also had data for genetic variants on the X chromosome. With there being only data for genetic variants on the X chromosome for one GWAS, the X chromosme was omitted from the analyses presented in this thesis. This was because for many of the analyses conducted in this thesis, only overlapping SNPs across the two studies were analysed.

2.1.1.3 Collaborative Oncological Gene-Environment Study (COGS)

In recent years, genetic variants genotyped on the iCOGS array, a custom array developed as part of the Collaborative Oncological Gene-Environment Study (COGS) project (80), has been used to identify breast, ovarian and prostate cancer risk factors. SNPs were chosen for inclusion on the custom array if they were thought to be somewhat related to any of the three cancers, this being based on a meta-analysis of breast, ovarian and prostate cancer GWAS results, as well as other studies. The array was used to genotype 211,115 SNPs in subjects from 52 BCAC (Breast Cancer Association Consortium) studies. In total, 41 out of the 52 BCAC studies contained women of European ancestry, nine studies contained populations of Asian ancestry, and two studies contained women of African-American ancestry. Based on European ancestry, the COGS data used to perform the analysis in this thesis contained 199,961 genotyped SNPs for 48,154 European cases and 43,612 European controls.

2.1.2 Estimating the genetic variation explained by common SNPs for polygenic traits

Since the first GWAS, many traits and diseases have been shown to have a polygenic basis, making it difficult to establish all associated genetic variants. With there being many variants that affect disease risk, it has been hard to discover them all. It has not helped that GWAS are expensive to conduct in a large number of individuals, and that many of the studies conducted to date have not been large enough to detect many of the associated genetic variants with small effect sizes. However, researchers have developed methods to estimate the proportion of phenotypic variance that can be explained by genotyped SNPs, without having to first discover the genetic variants associated with the trait (81). These estimates are known as chip, or SNP, heritability estimates, and are particularly useful as they enable the potential a genotyping array has in explaining the heritability of a trait, to be evaluated. Estimates can be produced using unrelated individuals, which is an advantage as it is easier to collect a larger number of unrelated individuals, than it is related individuals. Also, using unrelated individuals reduces the risk of shared environments inflating the chip heritability estimate (60). The estimates also allow researchers to assess whether a certain group of variants, for example SNPs mapping to a specific chromosome, explain more phenotypic variation than other groups. This type of analysis is known as genome partitioning, and will not be discussed further in this chapter, but will instead will discussed in chapter 3.

For many polygenic traits and diseases, chip heritability estimates have been produced and have shown that a fairly large proportion of the variation in a trait, can be explained by genotyped SNPs not yet reaching genome-wide significance. An early study conducted by Yang et al.(82) showed that a large proportion of the heritability for human height could be explained by common SNPs using when using the genomic-relatedness based restricted maximum-likelihood (GREML) (83), implemented as part of the genome-wide complex trait analysis software (GCTA). They estimated that 45% (se = 8%) of the phenotypic variance for height could be explained by 294,831 SNPs, genotyped in 3,925 individuals of European descent. This estimate was much larger than the estimated 5% explained by the combination of genome-wide significant SNPs, published before the analysis was conducted, and suggested that over 50% of the heritability for height could be explained by common SNPs. Otowa et al.(84) used both GREML and another estimation method, LD score regression (LDSC) (85), to produce chip heritability estimates for anxiety disorder. LDSC is a method that uses summary data to produce chip heritability estimates. In this study, anxiety disorder was defined by five phenotypes, these being; generalized anxiety disorder, panic disorder, phobias, social phobia, agoraphobia, and specific phobias. Based on 3,695 European individuals from the Rotterdam Study Cohort, Otowa et al. estimated that 13.8% (se = 18%) of the variation in liability to anxiety disorder could be explained by genotyped SNPs, when using GREML. An LDSC chip heritability estimate was produced using summary statistics, based on a meta-analysis of over 18,000 individuals and 995,869 SNPs, across nine cohorts. Using LDSC, they estimated that 9.5% (se = 3.7%) of the variation in liability to anxiety disorder could be explained by genotyped SNPs. These results showed that approximately a third of the genetic variation in anxiety disorders could be explained by common SNPs.

Chip heritability estimates, on the unobserved liability scale, have also been produced for breast cancer. Lu et al.(86) have produced a chip heritability estimate based on 489,247 genotyped SNPs, in 1,081 breast cancer cases and 1,085 controls. Using GREML, it was estimated that 13% (95% CI:[0%-56%]) of the variation in liability to breast cancer could be explained by genotyped SNPs. Assuming that the heritability of breast cancer on the unobserved liability scale is 44%, then approximately 30% of the genetic variation in liability to breast cancer could be explained by these genotyped SNPs. However, the 95% CI for the estimate was quite wide which, with only ~2,000 individuals being used in the analysis to produce the estimate, would have been due to the limited sample size. Therefore, the estimate could in fact be much larger than 13%. Sampson et al.(87) have also produced a chip heritability estimate for breast cancer, but instead they have focussed on ER-negative breast cancer. The estimate they produced was based on GWAS SNPs, genotyped in 1,998 ER-negative breast cancer cases and 3,263 controls. Using GREML, they estimated that 9.6% (95% CI: [0%-19.9%]) of the variation in liability to ER-negative breast cancer, could be explained by genotyped SNPs.

The non-breast cancer studies mentioned, are only a very small sample of the chip heritability studies that have been conducted to date. Studies in general have shown that a relatively large proportion of variation for a trait can be explained by currently genotyped SNPs, compared to the proportion of phenotypic variation that can be explained by SNPs reaching genome-wide significance. This common finding indicates that much of the missing heritability for many phenotypes, may be explained by SNPs not yet reaching genome-wide significance.

2.2 Methods

2.2.1 Testing whether a breast cancer PRS is associated with breast cancer risk

Many genetic variants have been shown to be associated with breast cancer risk, which implies that breast cancer has a polygenic basis. Polygenic scores can be used to provide additional evidence to infer that a disease, or trait, is polygenic. A training and replication sample are needed, these can either be based on two separate studies of the same phenotype, or a single study split into two samples. The two samples are used to perform a polygenic score analysis, and if there is found to be an association between the polygenic score and phenotype, the result suggests that the phenotype has a polygenic basis.

Two independent GWAS were used to perform the polygenic score analyses conducted in this thesis. Using the "--score" command in PLINK, the SNP effects from one GWAS were used to construct a PRS for the women in the remaining independent GWAS. The SNP effects were estimated using a logistic regression model, with the relevant number of ancestry principal components for the training sample, included as covariates in the model. The principal components were included in the model in order to reduce the presence of population stratification. A logistic regression model was then used to examine the relationship between the breast cancer PRS and breast cancer risk, with the relevant number of ancestry principal components for the replication sample included as covariates in the model.

2.2.2 Approaches to estimate chip heritability

Using data currently available, either publicly or through their own collection/collaboration, many researchers have attempted to identify causal variants in order to explain the heritability for many individual complex diseases. As explained in chapter 1, the genetic variants discovered for many diseases fall short of explaining a large proportion of the heritability for the disease. With that in mind, methods have been developed in order to estimate the proportion of phenotypic variance that can be explained by genotyped SNPs. Producing an accurate chip heritability estimate will enable researches to assess the potential current genotyped variants have of explaining “missing heritability”, without first having to establish which variants are causal variants. As mentioned in the previous section, methods typically used to estimate chip heritability include GREML (83) and LDSC (85). Palla et al.(88) have also developed an estimation method, known as the additive variance explained and number of genetic effects method of estimation (AVENGEME) (88). This method uses polygenic scores, and methodology previously given by Dudbridge (15), to produce a chip heritability estimate.

In this section, an overview of the methods that have been used in this thesis to produce chip heritability estimates on a liability scale will be given. The estimates produced were based on SNPs that have been genotyped for either the BBCS, UK2 GWAS study, or the COGS.

2.2.2.1 LD score regression (LDSC)

LD score regression (LDSC) (85), is a chip heritability estimation method that can be implemented using the web interface LD hub (89), which enables one to upload genotype data, and perform LDSC. The method produces an observed chip heritability estimate (h_o^2) by regressing the chi-squared test statistic (χ^2) for each individual SNP, i, against an LD score (ℓ_i), such that (85):

$$E[\chi^2|\ell_i] = Na + 1 + \frac{Nh_o^2\ell_i}{M}$$

Where, N is the number of subjects, M is the total number of SNPs, $\frac{h_o^2}{M}$ is the average observed chip heritability and a is the inflation from population stratification/cryptic relatedness, with $Na + 1$ being the intercept. The LD score,(ℓ_i) for a specific SNP, i, is estimated by summing the measures of r^2 associated with that SNP, for its relationship with other SNPs. LD scores are estimated under the belief that a SNP that is in LD with many SNPs, could have a higher univariate association statistic, than a SNP that is in LD with fewer SNPs (60).

From this regression an estimate of the slope is produced, which can then be multiplied by M/N to produce a heritability estimate on the observed scale.

For case-control studies, the proportion of cases in each study tends to be higher than the proportion of cases in the general population, meaning that cases are over represented in the data. This is known as ascertainment bias, and should be adjusted for. For binary traits, the chip heritability on the observed scale estimate and its standard error can be transformed to the liability scale, and adjusted for ascertainment using the following equations given by Lee et al (90):

$$h_l^2 = \widehat{h_o^2}\,\frac{K(1-K)}{z^2}\,\frac{K(1-K)}{P(1-P)}$$

$$SE(h_l^2) = \sqrt{var(h_o^2)\,\frac{K(1-K)}{z^2}\,\frac{K(1-K)}{P(1-P)}}$$

Where, P is the sampling fraction, K the population prevalence, z the normal density height at threshold T, h_l^2 is the chip heritability on a liability scale and h_o^2 is the chip heritability estimate on the observed scale.

2.2.2.2 Genomic-relatedness based restricted maximum-likelihood (GREML) method

GCTAs GREML uses both a genetic relationship matrix (GRM), $\boldsymbol{A}$, and a linear mixed model, to estimate the genetic variation explained by all genotyped SNPs.

The genetic correlation between each pair of individuals, j and k, is measured using the GRM, $\boldsymbol{A}$, which can be estimated from SNPs, such that (83):

$$A_{jk} = \frac{1}{M}\sum_{i=1}^{M}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-\ p_i)}$$

Where, x_{ij} is the number of copies that the j^{th} individual has of the reference allele for the i^{th}SNP, x_{ik} is the number of copies that the k^{th} individual has of the reference allele for the i^{th}SNP, and p_i is the frequency of the reference allele. The total number of SNPs is denoted as M.

The GRM between each pair of individuals is then used to help produce a chip heritability estimate. A GREML chip heritability estimate is produced using a restricted maximum likelihood analysis of the following mixed model (83):

$$\boldsymbol{y} = \ \boldsymbol{X\beta} + \boldsymbol{g} + \ \boldsymbol{\varepsilon}$$

with,

$$var(\boldsymbol{y}) = \ \boldsymbol{A}_g\sigma_g^2 + \boldsymbol{I}\sigma_\varepsilon^2$$

With, $\boldsymbol{y}$ being an n x 1 vector of phenotypes with sample size n, and $\boldsymbol{X}$ an incidence matrix for $\boldsymbol{\beta}$, a vector of fixed effects. In this thesis, $\boldsymbol{\beta}$ shall be a vector of fixed effects for the eigenvectors used to adjust for population stratification. An n x 1 vector of the random genetic effects of all SNPs for all the individuals is denoted by $\boldsymbol{g}$, with $\boldsymbol{g} \sim N(0, \boldsymbol{A}_g\sigma_g^2)$, where $\boldsymbol{A}_g$is the GRM estimated for the same SNPs and σ_g^2 is the variance explained by the SNPs. The vector of residual effects is denoted by ε, with $\varepsilon \sim N(0, \boldsymbol{I}\sigma_\varepsilon^2)$, where $\boldsymbol{I}$ is an n x n identity matrix.

The genetic relationship matrix is used to adjust for any relatedness between individuals in order to improve the accuracy of the chip heritability estimate. By adjusting for subject relatedness, the estimated variation explained should then be based on the variation of SNPs alone.

The variance explained by SNPs on the observed scale is transformed to the liability scale, and adjusted for ascertainment using the equations given by Lee et al (90). The GREML model allows for correlation between the analysed genotyped SNPs, so LD-clumping and LD-pruning does not have to be carried out before using this method to estimate chip heritability.

2.2.2.3 AVENGEME

AVENGEME (88), a method developed by Palla & Dudbridge (88) based on methods described by Dudbridge (15), can also be used to estimate chip heritability. This method consists of a group of R functions, one of which can be used to estimate the proportion of trait variance explained by the genetic effects in the training sample (σ_1^2), the genetic covariance between the training and replication sample (σ_{12}) and the proportion of markers with no effect on the training trait (π_{01}). These estimates can be generated using the R function, "estimatePolygenicModel", and *z*-scores that have been produced from testing the association between multiple polygenic scores and the replication trait.

Dudbridge (15) presents a model where a pair of traits, $\boldsymbol{Y} = (Y_1, Y_2)'$, can be expressed as a linear combination of m genetic effects, with a pair of random errors $\boldsymbol{E} = (E_1, E_2)'$ which include both environmental and un-modelled genetic effects:

$$\boldsymbol{Y} = \boldsymbol{\beta}'\boldsymbol{G} + \boldsymbol{E} = \left(\sum_{i=1}^{m} \beta_{i1}\, G_i + \; E_1\,, \sum_{i=1}^{m} \beta_{i2}\, G_i + \; E_2\right)'$$

Where, $\boldsymbol{\beta}$ is an m x 2 matrix of coefficients and $\boldsymbol{G}$ is a m-vector of coded genetic markers. It is assumed that the genetic markers are independent and standardised, and $\boldsymbol{E}$ is independent of $\boldsymbol{G}$. Estimated marker effects are given as $\hat{\beta}_{i1}$ and $\hat{\beta}_{i2}$, for $i=$ $1,\ldots m,$ with m being the total number of markers.

Assuming independence across the two samples, then either set of genetic effect estimates can be used to create a polygenic score ($\hat{S}_1$ or $\hat{S}_2$).

For the training sample:

$$\hat{S}_1 = \sum_{i=1}^{m} \hat{\beta}_{i2} G_i$$

and for the replication sample:

$$\hat{S}_2 = \sum_{i=1}^{m} \hat{\beta}_{i1} G_i$$

The polygenic score can then be used to test for an association with Y_1 and Y_2 , respectively.

Focussing on $\hat{S}_2$, and assuming that Y_1 and Y_2 are binary traits, $\hat{\beta}_{i1}$ can be produced using a logistic regression model, where the dependent variable is Y_1 and the independent variable is each G_i for $m.$

The association between $\hat{S}_2$ and Y_2 is then tested using logistic regression model, with the independent variable being $\hat{S}_2$ and the dependent variable Y_2. To estimate the proportion of trait variance explained by the genetic effects in the training sample (σ_1^2), the genetic covariance between the training and replication sample (σ_{12}) and the proportion of markers with no effect on the training trait (π_{01}), multiple $\hat{S}_2$ and Y_2 associations should be tested. The number of *z*-scores/*p*-values produced should be greater than or equal to the number of parameters being estimated. The multiple *z*-scores are produced by testing the association between $\hat{S}_2$ and Y_2 , with multiple $\hat{S}_2$

being constructed using SNPs from different SNP *p*-value intervals. The SNP *p*-value intervals are created based on each SNPs association with Y_1.

Palla et al.(88) show that the variance explained by the polygenic score in the regression of Y_2 on $\hat{S}_2$ can be given as:

$$R^2_{\hat{S}_2\,,Y_2} = \frac{mcov(\hat{\beta}_{i1}, \beta_{i2})^2}{var(\hat{\beta}_{i1})var(Y_2)},$$

and the asymptotic non-centrality parameter of the χ^2_1 test for the association between Y_2 on $\hat{S}_2$ can also be given as (88):

$$\lambda = \frac{n_2 R^2_{\hat{S}_2\,,Y_2}}{(1 - R^2_{\hat{S}_2\,,Y_2})}$$

where, n_2 is the sample size of the replication sample.

From this, the expectation of the *Z*-test can be estimated as (88):

$$\mu = \sqrt{\frac{n_2 R^2_{\hat{S}_2\,,Y_2}}{(1 - R^2_{\hat{S}_2\,,Y_2})}},$$

The observed *z*-scores, along with the expectation of the *Z*-test and maximum-likelihood estimation can then used to find the σ_1^2, σ_{12} and π_{01} values that maximize the following log-likelihood function (88):

$$\ell(\sigma_1^2, \sigma_{12}, \pi_{01}) = \sum_{i=1}^{k} log\varphi(Z_i - \mu(\sigma_1^2, \sigma_{12}, \pi_{01};\ d_i)\,)\,,$$

with $d_1, \ldots, d_k$ being the SNP selection *p*-value interval, for a set of $k$intervals, with the value of $k$being equal to or greater than the number of parameters estimated. Multiple z-scores, Z_i, are estimated using SNPs within each $d_1, \ldots, d_k$ interval. The expectation of the *Z*-test is then represented by $\mu(\sigma_1^2, \sigma_{12}, \pi_{01};\ d_i)$.

To produce a chip heritability based an ensemble of markers genotyped as part of a study, study subjects can be randomly split into roughly equal internal training and replication samples, with the Y_1 and Y_2 then being the same trait. The effect sizes for all independent markers in the training sample, including SNPs that have not yet been shown to be associated with the trait, are estimated using either a logistic or a linear regression model. Principal components can be included in the model if population stratification needs to be adjusted for. The SNP p-value intervals, $d_1, \dots, d_k$, are created by assigning SNPs to different *p*-value threshold groups, based on their individual association with the training trait. The individual effect sizes for each marker are then used to construct a polygenic score for each subject in the replication sample, for the different SNP *p*-value thresholds, and the association between each score and Y_2 is then tested. The *z*-scores produced are then used to estimate the additive genetic variance in the training sample (σ_1^2), which is the chip heritability when Y_1 and Y_2 are the same trait, as well as the proportion of null markers with no effect on the trait in the training sample (π_{01}). With the training and replication samples assumed to have the same genetic model, the replication variance and covariance can be fixed to equal the variance explained on the training sample, and the proportion of null SNPs in the replication sample can also be set to equal the proportion of null SNPs in the training sample. For a binary trait, both the prevalence and sampling fractions for the two samples are needed. When Y_1 or Y_2 are binary, the chip heritability estimates are transformed from the observed scale to the unobserved liability scale, with ascertainment also being adjusted for, using the equation given by Lee et al (90).

AVENGEME can also be used to estimate the power of the χ_1^2 test of the association between the polygenic score and Y_2, using the “polygenescore” R function. The estimates produced using “estimatePolygenicModel”, along with the total number of SNPs, sample size of training and replication sample, sampling fraction for both samples and prevalence, can be used to estimate $R^2_{\hat{S}_2, Y_2}$. This can then be used to estimate the asymptotic non-centrality parameter of the χ_1^2 test for the association between Y_2 on $\hat{S}_2$ (λ). Power can be estimated by estimating the distribution function for the chi-squared distribution, with the estimated λ as the non-centrality parameter.

2.3 Quality control and LD removal

2.3.1 SNP quality control

As explained in the previous chapter, QC should be carried out on genetic data before analysis is performed. QC was executed on the European subjects and autosomal SNPs within each study, first using PLINK to identify and remove any SNPs that have a MAF less than 5%, missing rates greater than 5%, or a significant departure from HWE (*p*-value < 5x10^{-6}). After SNP QC, 483,185 SNPs, 268,776 SNPs and 172,995 SNPs were retained in the UK2, BBCS and COGS, respectively.

2.3.2 Sample quality control

Individuals with discordant sex information, missing genotype rates greater than 5% and a heterozygosity rate ± 3 standard deviations away from the mean heterozygosity rate, were identified and removed using PLINK. With there being shared controls between the UK2 and BBCS GWAS, duplicate/MZ twin subject pairs were expected to be discovered across the two GWAS. Also, the BBCS and UK2 GWAS contributed to COGS, so overlap between COGS subjects and the subjects in the two GWAS was also expected. The relationship between subjects within and across each study was assessed using KING Version 1.4. The results from this assessment are given in Table 2-1. No first-degree relative pairs were identified within either the UK2 or BBCS GWAS. Within COGS, both across and within the various BCAC studies, no duplicate/MZ twin pairs were identified. However, 100 parent-offspring related pairs were discovered, with the majority of these pairs having differing disease status. For every case-control related pair, the control was excluded in order to retain cases in the analysis. For every control-control and case-case pair, a subject was randomly removed whilst trying to remove subjects as evenly as possible across the BCAC studies. There were found to be 29 BBCS-COGS duplicate/MZ twin pairs, 84 UK2-COGS duplicate/MZ twin pairs and 5,190 BBCS-UK2 duplicate/MZ twin subject pairs, all with the same disease status.

The duplicate/MZ twin pairs between the two GWAS were all controls, which meant that one control from each subject pair was removed from either GWAS, whilst ensuring that the sampling fraction within each study was close to a half in order to retain power (15). For the COGS-GWAS pairs, the COGS subject from each pair was removed. This was because there were a larger number of subjects in the COGS, than there were in either GWAS.

Between the studies, parent-offspring subject pairs were also identified, with 74 of these being BBCS-COGS pairs and 52 being UK2-COGS pairs. The majority of these pairs had differing disease status. No parent-offspring relatives were, however, identified across the two GWAS. The control from each parent-offspring pair was removed, however, if neither subject was a control, the COGS subject was removed. After exclusions, 3,628 cases and 3,581 controls remained in the UK2 study, and 1,609 cases and 1,609 controls in the BBCS, making the sampling fractions for the UK2 and the BBCS 0.503 and 0.5 respectively. For COGS, 48,069 cases and 43,481 controls remained after exclusions (sampling fraction = 0.525).

Study 1	Study 2	No. MZ twin pairs	No. parent-offspring pairs
BBCS	BBCS	0	0
UK2	UK2	0	0
COGS	COGS	0	100
BBCS	UK2	5,190	0
BBCS	COGS	29	74
UK2	COGS	84	52

Note:
when study 1 = study 2, this is the number of related subject pairs within the study

MZ twin pair: Kinship coefficient > $\frac{1}{2^{\frac{3}{2}}}$ and prob of zero IBD sharing < 0.1

Parent-offspring pair: Kinship coefficient ($\frac{1}{2^{\frac{5}{2}}}$, $\frac{1}{2^{\frac{3}{2}}}$) and prob of zero IBD sharing < 0.1

Table 2-1: Subject overlap within and between each study

Both PC analysis and the genomic inflation statistic were then used to assess whether population stratification was present amongst the European subjects in any of the three studies. For the BBCS, with both the lack of clustering when plotting the first two principal components (Figure 2-1), and the genomic inflation statistic for the data being close to 1 (λ = 1.015), it would seem that the BBCS does not suffer from population stratification. However, for the UK2 study, some clustering of subjects and a slight separation between cases and controls was found when plotting the first two principal components (Figure 2-2). The genomic inflation statistic was also slightly greater than 1 (λ = 1.113), therefore, both the PC plot and the genomic inflation statistic suggest that population stratification may exist in the UK2 GWAS.

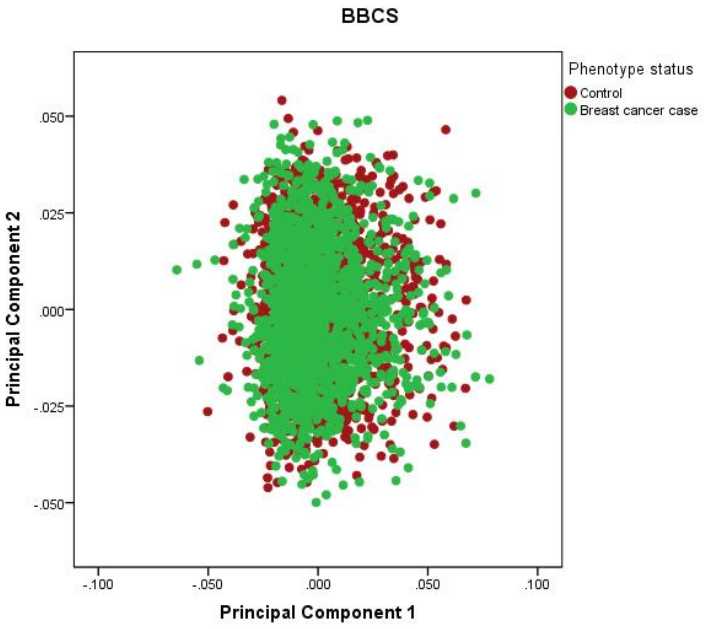


Figure 2-1: BBCS GWAS principal-component plot

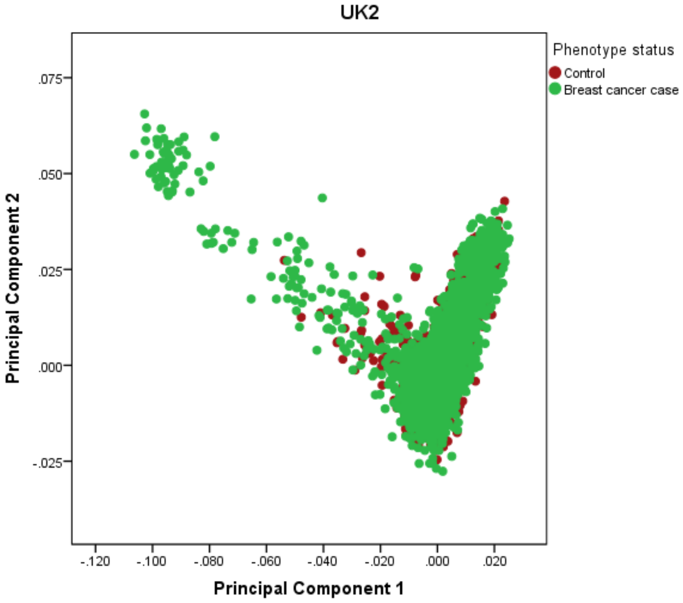


Figure 2-2: UK2 GWAS principal-component plot

To examine how many PCs were needed in order to adequately adjust for any population stratification present in the data, PCs were created and the data was adjusted by the PCs, one at a time, to see what effect this had on the genomic inflation statistic. When adjusting the UK2 study by ten PCs, the genomic inflation statistic decreased from 1.113 to 1.035, meaning that the inflation statistic was much closer to 1. Even though both Figure 2-1 and the genomic inflation statistic for the BBCS GWAS indicated that population stratification might not be present, the genomic inflation statistic can be decreased further by adjusting the data by four PCs (1.015 to 1.014). Adjusting the studies by a larger number of PCs, did not reduce the inflation statistic further for either study.

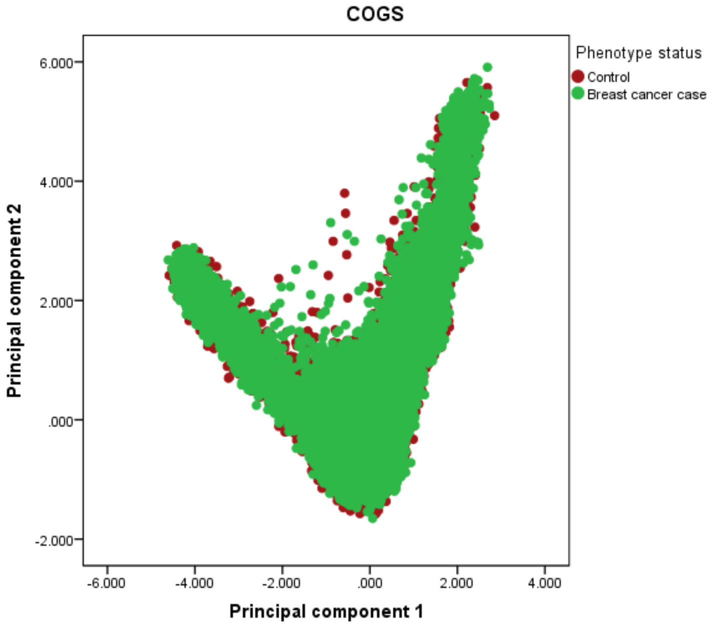


Figure 2-3: COGS principal-component plot

For the COGS, there was shown to be some clustering amongst subjects when plotting the first two principal components, which would indicate that population stratification may be present (Figure 2-3). The genomic inflation statistic for the data was close to 2 (λ = 1.980), but did decrease further when adjusting the data by nine PCs, and for study (λ = 1.335). Study was adjusted for by creating n-1 dummy variables, with n being the number of BCAC studies. When Michailidou et al (45, 51) analysed the COGS, they adjusted the data by nine PCs and for BCAC study in order to reduce the inflation statistic. They found that adjusting by more than nine PCs did not further decrease the inflation statistic.

The genomic inflation statistic remained high, even after adjusting for nine PCs, which could indicate that population stratification had not been correctly adjusted for. However, this result is consistent with multiple studies drawn from across Europe. Yang et al (91) have previously explained that genomic inflation should be expected when conducting a GWAS, especially a large-scale GWAS, even after adjusting for population stratification. The measure was proposed before GWAS, and before it was known that many genetic variants affect disease risk. It assumes that the test statistic for each SNP, apart from the few SNPs that are truly associated with the trait, should follow the distribution under the null hypothesis of no association (91). We now know that for many complex diseases, disease risk is influenced by many SNPs, which individually have a small effect on risk of disease. Therefore, many more SNPs than previously thought are in fact associated with the trait, meaning many SNPs will have a larger chi-squared statistic than previously expected. Then, with the iCOGS array being hugely enriched for associated SNPs, the SNPs in the COGS study, on average, will have many more SNPs with a larger chi-squared statistic, than many other studies.

After QC, 3,628 cases, 3,581 controls and 483,185 SNPs remained in the UK2 study, 1,609 cases, 1,609 controls and 268,776 SNPs in the BBCS, and 48,069 cases, 43,481 controls and 172,995 SNPs remained in COGS after exclusions.

The subjects and SNPs retained in the studies after QC were used to produce both GREML, and LDSC chip heritability estimates. To produce AVENGEME chip heritability estimates, polygenic score analyses needed to be conducted. Internal training and replication sets for each study were needed, and there needed to be independence across SNPs. To create the training and replication samples, each study, after QC, was split into approximately equal sized internal training and replication sets. Subjects were randomly assigned to a training or replication sample, whilst ensuring the sampling fraction was close to a half, and that the two samples were roughly equal in size. For the UK2 GWAS, 3,604 subjects were allocated to the training sample (sampling fraction = 0.498), and 3,605 subjects to the replication sample (sampling fraction = 0.508). For the BBCS GWAS, 1,609 subjects were allocated to the training sample (sampling fraction = 0.500), and 1,609 subjects to the replication sample (sampling fraction = 0.500). For COGS, 45,768 subjects (sampling fraction = 0.526) were assigned to the training sample, and 45,782 subjects (sampling fraction = 0.524) to the replication sample.

To ensure independence across SNPs, the LD amongst the SNPs within each study training sample was measured, and a LD-thinning technique was used to reduce high LD. With the full studies being used to conduct various analyses in this thesis, this includes testing whether breast cancer is polygenic using polygenic score analysis, the whole studies after QC were also LD-thinned.

2.3.3 LD-thinning

As discussed in chapter 1, there are two different LD-thinning methods that can be used to deal with high correlation between SNPs; LD-based clumping and LD-based pruning. Both methods were used, separately, on the SNPs retained after QC.

LD-based pruning was used to identify SNPs with a pairwise estimate of LD greater than 0.2 (r^2 > 0.2), using a sliding window of 50 SNPs, while sliding across the genome 5 SNPs at a time. An r^2 > 0.2 was used to prune the SNPs, as this tends to be the constraint used when pruning (71). For every group of correlated SNPs, one SNP was randomly retained whilst the others were removed. After QC and LD pruning, 90,907 SNPs in the UK2 study, 75,259 SNPs in the BBCS, and 142,816 SNPs in COGS remained. As LD-pruning randomly retains a SNP from a group of SNPs in high LD, there was no need to separately LD-prune the training sample SNPs, once the study itself has been LD-pruned. The LD-pruned SNPs for the whole QC study, were retained in the internal training samples. Therefore, the internal training samples for a study, contained the same number of SNPs as the whole LD-pruned study sample.

To conduct LD-based clumping, *p*-values for the association between each SNP and breast cancer were needed, in order to rank the SNPs by their association with the trait. A logistic regression model, with ancestry principal components, was used to test the association between each SNP and breast cancer outcome. LD-based clumping was then used to identify correlated SNPs, with a r^2 > 0.1. The most significant SNP from each group of correlated SNPs, based on the given *p*-values, was retained for further analysis. After QC and LD clumping, 83,702 SNPs in the UK2 study, 67,379 SNPs in the BBCS, and 44,181 SNPs in COGS remained. After LD clumping the internal training sets, 83,851 SNPs remained in the training set for the UK2 study, 67,654 SNPs in the BBCS training set and 44,181 SNPs in the COGS training set.

The number of SNPs retained after LD-based pruning were consistently larger than the number of SNPs retained after LD-clumping across all three breast cancer studies. This was because the r^2 threshold used varied, with r^2 > 0.1 being used for LD-clumping and r^2 > 0.2 for LD-pruning.

2.4 Analysis of polygenic scores

The first analysis performed, involved testing whether there was evidence that the SNP effects from one GWAS, could be used to predict breast cancer risk in an independent GWAS. If there was shown to be evidence, it would suggest that breast cancer has a polygenic basis. Polygenic score analysis, and the BBCS and UK2 GWAS were used conduct this analysis. One breast cancer GWAS was set as the training GWAS, and the other was the replication GWAS. A polygenic score for each subject in the replication GWAS was constructed using the SNP effects from the training GWAS, which were estimated whilst adjusting for the relevant number of principal components to control for population stratification. Different subsets of SNPs, based on their individual association with breast cancer outcome in the training sample, were created. Polygenic scores were constructed for the testing sample individuals, using the SNP effects for the SNPs within each subset ($p \leq 1$, $p \leq 0.7$, $p \leq 0.4$, $p \leq 0.1$, $p \leq 0.05$, $p \leq 0.01$, and $p \leq 0.001$). The association between each polygenic score and breast cancer outcome, in the replication GWAS, was then tested using a logistic regression model. Replication sample ancestry principal components were also included in the model, in order to adjust for population stratification in the data. This polygenic score analysis was done bi-directionally, so that each GWAS was used as both the testing and replication set.

There were found to be a low number of SNPs in union between the two GWAS, so to increase the number of SNPs in the analysis, imputed SNPs for the replication GWAS were incorporated in the analysis. The SNPs present in the training GWAS, were extracted from the imputed SNPs for the replication sample. The imputed SNPs were converted to PLINK best guess genotype format, and merged with the replication GWAS.

The results from this analysis suggest there to be a significant association between the breast cancer polygenic scores, and breast cancer outcome (Table 2-2). A PRS for

BBCS subjects, based on UK2 SNP effects, was shown to be associated with breast cancer status when using all SNPs (*p* ≤ 1, *p*-value = 8.05e-07), and also for the more stringent *p*-value thresholds. Significant associations were also shown in the other direction, with the PRS for UK2 subjects, based on BBCS SNP effects, shown to be associated with breast cancer status in the UK2 GWAS. Including non-significant SNPs in the scores, did not have much of an effect on the significance of the association. The PRS based on all genotyped SNPs was still shown to be associated with breast cancer outcome (*p* ≤ 1, *p*-value = 2.67e-07). All associations between PRS and breast cancer outcome were significant, regardless of the SNPs used in the score. For each *p*-value threshold the area under the ROC curve (AUC) was also computed using the “pROC” package in R (92) (Table 2-2). An AUC percentage close to 100% would indicate that the risk score excellently predicts breast cancer status for women in the replication sample. Unsurprisingly, the polygenic scores constructed are currently poor predictors of breast cancer status (AUC: ~55% - 61%). This was to be expected as sample sizes are not yet large enough to produce SNP effects that are accurate enough to be used in risk prediction.

With a significant association between each constructed polygenic score and breast cancer outcome observed for each SNP threshold, including the polygenic score constructed using all independent genotyped SNPs, it was then investigated whether SNPs that are more significantly associated with breast cancer are driving the observed associations. To do this, the SNPs with a *p*-value ≤ 0.001 in the training sample were excluded from each of the polygenic scores and then the association between each score and breast cancer outcome in the replication was tested.
Even after removing SNPs with a *p*-value ≤ 0.001 from the scores, a significant association between each of the breast cancer polygenic scores and breast cancer outcome in the independent replication sample was still observed (Table 2-3). Significant associations were observed in both directions, with even the least stringent

polygenic score still being shown to be associated with breast cancer outcome in the replication sample. Again, the polygenic scores constructed are presently poor predictors of breast cancer status (AUC: ~54% - 60%).

The results from these analyses indicate that a polygenic score based on GWAS SNPs could be used to predict breast cancer risk for women in an independent GWAS as significant associations between each score and breast cancer outcome in the replication sample was observed. These results therefore suggest that breast cancer has a polygenic basis. However, the computed AUC values suggest that the constructed polygenic scores are currently poor predictors of breast cancer outcome in the replication sample.

Training	Replication	SNP threshold*	No. SNPs**	*p*-value	*AUC (%)*
UK2 GWAS	BBCS GWAS	$p \leq 1$	82,704	8.05e-07	55.34
		$p \leq 0.7$	70,692	4.88e-07	55.41
		$p \leq 0.4$	50,893	3.75e-07	55.46
		$p \leq 0.1$	18,645	9.17e-07	55.19
		$p \leq 0.05$	10,734	1.59e-07	55.76
		$p \leq 0.01$	2,849	9.35e-08	55.92
		$p \leq 0.001$	377	2.85e-07	55.84
BBCS GWAS	UK2 GWAS	$p \leq 1$	63,328	2.67e-07	60.54
		$p \leq 0.7$	53,563	2.49e-07	60.54
		$p \leq 0.4$	38,030	4.9e-08	60.62
		$p \leq 0.1$	13,156	1.2e-07	60.52
		$p \leq 0.05$	7,355	9.83e-07	60.43
		$p \leq 0.01$	1,808	1.64e-07	60.54
		$p \leq 0.001$	221	2.97e-07	60.52

* Training sample SNPs association with breast cancer in training sample
** The number of SNPs used in polygenic score analysis
Note: AUC = Area under the ROC curve

Table 2-2: Association between PRS and breast cancer outcome using two independent GWAS

Training	Replication	SNP threshold*	No. SNPs**	*p*-value	*AUC (%)*
UK2 GWAS	BBCS GWAS	*0.001* < *p* ≤ 1	82,327	1.89e-05	54.78
		0.001 < *p* ≤ 0.7	70,315	1.21e-05	54.85
		0.001 < *p* ≤ 0.4	50,516	1.04e-05	54.88
		0.001 < *p* ≤ 0.1	18,268	6.12e-05	54.43
		0.001 < *p* ≤ 0.05	10,357	2.79e-05	54.81
		0.001 < *p* ≤ 0.01	2,472	3.36e-04	54.41
BBCS GWAS	UK2 GWAS	*0.001* < *p* ≤ 1	63,107	6.18e-06	60.40
		0.001 < *p* ≤ 0.7	53,342	5.88e-06	60.40
		0.001 < *p* ≤ 0.4	37,809	1.49e-06	60.47
		0.001 < *p* ≤ 0.1	12,935	9.31e-06	60.33
		0.001 < *p* ≤ 0.05	7,134	1.37e-04	60.22
		0.001 < *p* ≤ 0.01	1,587	4.95e-04	60.18

* Training sample SNPs association with breast cancer in training sample
** The number of SNPs used in polygenic score analysis
Note: AUC = Area under the ROC curve

Table 2-3: Association between PRS and breast cancer outcome using two independent GWAS - removing SNPs with *p*-value < 0.001 in the training set

2.5 Estimating chip heritability

The concept of heritability, and the various forms of this measure, were briefly discussed in chapter 1. Much of the heritability for breast cancer is said to be missing, but with advances in computation and statistical methods, we are now able to estimate the proportion of phenotypic variation that can be explained by genotyped SNPs (chip heritability). A chip heritability estimate can be produced using unrelated individuals, and without having to first identify all associated SNPs, which is useful as GWAS are currently underpowered to detect all associated variants with current sample sizes. Three different methods, AVENGEME, GREML and LDSC, have been used to estimate the variation in liability to breast cancer that can be explained by genotyped SNPs. Estimates have been produced for the two breast cancer GWAS, and the COGS.

As well as the chip heritability, the proportion of SNPs that have no effect on breast cancer will also be estimated for each study, using AVENGEME.

2.5.1 Heritability explained by GWAS SNPs

With the studies analysed being case control studies, and with the disease of interest being binary, the prevalence was needed in order to estimate chip heritability on a liability scale, and to adjust for ascertainment bias. The prevalence for breast cancer has been given as ~0.036 (6, 15), however the prevalence for bilateral breast cancer was not widely known. Treating the probability of developing the first primary breast cancer tumour, as independent to the probability of developing a second primary breast cancer tumour, the probability of developing two primary breast cancer tumours could be equal to the square of the primary breast cancer prevalence, which is ~0.001. However, a woman who has been diagnosed with breast cancer once, has an increased risk of developing a second primary breast cancer tumour, which therefore means that the risk of developing the second tumour is higher than developing the first primary breast cancer tumour (93). Therefore, the prevalence of bilateral breast cancer would be slightly higher than 0.001, meaning that the chip heritability estimate produced assuming this prevalence, would be a lower bound estimate. This prevalence value was assumed for each of the three studies, even for the COGS as it contained a mixture of breast cancer cases, which included familial and bilateral cases. The prevalence was also assumed for the UK2 study, as it contained familial breast cancer cases, which also meant that the prevalence value should be smaller than the general breast cancer prevalence.

In order to produce a h_l^2 (chip heritability on the liability scale) estimate using AVENGEME, multiple *z*-scores, produced after conducting multiple polygenic score analyses, were needed. Scores were constructed using the training SNPs within each *p*-value threshold subset. The *z*-scores produced for each subset, and used to conduct the analysis, are given in the **Appendix 2: Table 1 & 2**.

The AVENGEME h_l^2 estimates based on GWAS SNPs retained after LD pruning ranged from 17% to 19%, whereas the estimates based on GWAS SNPs retained after LD clumping ranged from 16% to 21% (Table 2-4). The h_l^2 estimate based on UK2 SNPs retained after LD clumping were larger than the estimate produced when using SNPs retained after LD pruning, but the same was not shown for the BBCS GWAS, as LD pruned SNPs produced a higher h_l^2 estimate. Both the GREML and LDSC h_l^2 estimates were smaller than the AVENGEME h_l^2 estimates, with LDSC producing much smaller estimates than both AVENGEME and GREML (h_l^2: 5.7% - 6.5%). This is not the first time that LDSC chip heritability estimates have been found to be smaller than GREML estimates. Yang et al. (94) state that when using the same data sets, chip heritability estimates produced using LDSC have tended to be smaller than those produced using GREML. Yang et al. believed that it is possible that this could be to do with errors when using a reference panel to estimate LD scores.

The 95% CI for the h_l^2 estimates produced for the BBCS GWAS were much wider than those produced for the UK2 GWAS, thus indicating that the BBCS estimates produced were less precise. With the sample size for the BBCS being smaller than the UK study, this would explain why the 95% CIs for h_l^2 estimates are wider for the BBCS, than they are for the UK2 study. On the Wiki FAQ page for LDSC, this is further confirmed as Bulik-Sullivan (95) warn that LDSC can produce very noisy estimates, if the sample size used is less than 5,000 subjects. The 95% CIs for the GREML h_l^2 estimates were narrower than those produced when using the other two estimation methods, suggesting that the estimates produced using this method were more precise, than those produced using the other two methods.

	AVENGEME h_l^2 (95% CI)		GREML h_l^2 (95% CI)	LDSC h_l^2 (95% CI)
GWAS	LD-pruning	LD-clumping		
UK2	0.171 (0.112, 0.229)	0.209 (0.152, 0.265)	0.143 (0.110, 0.176)	0.057 (0.000, 0.114)
BBCS	0.188 (0.070, 0.307)	0.158 (0.047, 0.272)	0.108 (0.037, 0.179)	0.065 (0.000, 0.192)

Notes: By assuming normally distributed estimators, the 95% confidence intervals for GREML and LD score regression were converted from the standard errors given for each h_l^2 estimate.

Table 2-4: Chip heritability estimates (h_l^2) for GWAS SNPs

The estimated proportion of markers that have no effect on breast cancer risk (π_{01}), based on SNPs genotyped for either GWAS, was also estimated when using AVENGEME to estimate h_l^2 (Table 2-5). The estimated π_{01} for the UK2 GWAS SNPs was larger than the estimate produced for the BBCS, when using LD pruned SNPs to conduct the analysis. However, the opposite was shown when conducting the analysis on LD clumped SNPs. The results mainly suggest that less than 10% of the GWAS SNPs have an effect on breast cancer risk, but with the 95% CIs for all estimates being very wide, and the BBCS π_{01} estimate being much smaller than the other estimates (π_{01} = 0.485), it meant that a reasonable conclusion about the data could not be made.

	π_{01} (95% CI)	
GWAS	LD-pruning	LD-clumping
UK2	0.934 (0.000, 0.976)	0.900 (0.000, 0.960)
BBCS	0.485 (0.000, 0.994)	0.980 (0.000, 0.998)

Table 2-5: Estimated proportion of null SNPs (π_{01}) for GWAS

2.5.2 Heritability explained by custom array SNPs

With it estimated that up to 20.9% of the variation in liability to breast cancer, could be explained by genotyped GWAS SNPs, the next obvious step was to estimate how much variation in liability to breast cancer can be explained by custom array SNPs. Estimates were produced using AVENGEME, GREML and LDSC. A prevalence of 0.001 was assumed, and in order to produce AVENGEME h_l^2 estimates, multiple *z*-scores, from regressing a PRS and breast cancer outcome for different *p*-value thresholds, were produced **(Appendix 2: Table 3)**.

With ~90,000 individuals in the COGS, the sample size was too large for GCTA to compute the genetic relationship matrix needed for GREML to estimate h_l^2. To produce a GREML h_l^2 estimate, a subset of 10,000 subjects were used, with these 10,000 randomly extracted whilst maintaining similar proportions across the BCAC studies to those for the whole of COGS.

The results from this analysis suggested that up to 15% of the variation in liability to breast cancer, could be explained by the custom array SNPs (Table 2-6). A surprising find was that the AVENGEME estimate based on LD-pruned SNPs was observed to be larger than the estimate produced after LD-clumping SNPs. By retaining the SNP with the strongest association with breast cancer outcome in each LD block, you would expect these SNPs to explain a larger amount of the variation in disease, than those randomly retained. There were over three times as many SNPs retained after LD-pruning, than there were after LD-clumping, which might explain why the LD-pruning estimate was larger than the LD-clumping estimate. The GREML estimate, based on a reduced number of individuals, but a larger number of SNPs than those retained after LD-thinning, was much closer to the AVENGEME LD-clumped estimate, than the LD-pruned estimate.

The h_l^2 estimates based on GWAS SNPs were larger than the estimates based on the custom array SNPs. But with many more variants being genotyped for a GWAS, this is not surprising. With the h_l^2 estimates based on GWAS SNPs having a wider 95% CI than the estimates based on the custom array, the GWAS h_l^2 estimates were less precise. This was to be expected as the number of women genotyped for the COGS, was much larger than the number of women genotyped for either GWAS. The LDSC h_l^2 estimate was roughly double both the AVENGEME estimate based on LD-clumped SNPs, and the GREML estimate. It was fairly close to the AVENGEME estimate based on LD-pruned SNPs, but it also had a relatively wide 95% CI. The standard error for a LDSC h_l^2 estimate will usually be fairly large, if less than 200,000 SNPs are used to produce an estimate (95). With the number of SNPs analysed being under 200,000, the LDSC h_l^2 estimate based on SNP genotyped for the COGS, was therefore fairly imprecise.

AVENGEME h_l^2 (95% CI)		GREML h_l^2 (95% CI)	LDSC h_l^2 (95% CI)
LD-pruning	LD-clumping		
0.143 (0.137, 0.150)	0.059 (0.055, 0.063)	0.078 (0.060,0.096)	0.146 (0.091,0.201)

Notes: By assuming normally distributed estimators, the 95% confidence intervals for GREML and LD score regression were converted from the standard errors given for each h_l^2 estimate.

Table 2-6: Chip heritability (h_l^2) estimates for COGS

Using LD-clumped COGS SNPs, it was estimated that approximately 70% of the SNPs genotyped on the iCOGS array were null SNPs (Table 2-7). The estimate was slightly lower than the estimate produced using LD-pruned COGS SNPs, where it was estimated that approximately 79% of the SNPs genotyped on the iCOGS array were null SNPs.

Both π_{01} estimates were mainly lower than the estimates produced for the genotyped GWAS SNPs, which was to be expected as the SNPs on the iCOGS array had been chosen for their association with breast cancer, based on the results from previous studies. This meant that the array was enriched for breast cancer associated SNPs, so the proportion of SNPs with an effect on breast cancer risk should be higher. In addition, the COGS SNPs retained after LD-clumping would contain, on average, a larger number of associated breast cancer SNPs, than would be retained after LD-pruning. Therefore, it could be expected that the estimated proportion of null SNPs for COGS, based on LD-clumped SNPs, would be lower than the estimate produced using LD-pruned SNPs. The estimates produced in Table 2-7 for COGS were more precise than those produced for the two GWAS (Table 2-5) as the 95% CIs for the COGS estimates were shown to be narrower than the 95% CI for the GWAS estimates. With COGS having a much larger sample size, this was expected.

π_{01} *(95% CI)*	
LD-pruning	LD-clumping
0.788 (0.762, 0.810)	0.696 (0.636, 0.743)

Table 2-7: Estimated proportion of null SNPs (π_{01}) for COGS

2.6 Discussion

The aim of this chapter was to find evidence to suggest that a large number of SNPs, collectively explain a proportion of the heritability for breast cancer. With published genome-wide significant loci explaining a small proportion of the genetic variation for the disease, the remaining genetic variation for the disease needs to be accounted for. Using two breast cancer GWAS, the BBCS and the UK2 study, and up to 82,704 autosomal SNPs, significant polygenic components for breast cancer were observed. A breast cancer polygenic score, based on all autosomal SNPs, was shown to be significantly associated breast cancer outcome in an independent sample (*p*-value < 0.05). The same was observed when decreasing the number of SNPs in the polygenic score, by applying stricter inclusion thresholds. Even when the polygenic score was constructed using as few as 221 SNPs, the score was still significantly associated with breast cancer outcome. Other studies have tended to focus on genome-wide significant loci when examining whether a polygenic score can be used to predict disease risk in an independent sample. This is through the belief that increasing the number of SNPs in the score, by including SNPs not reaching genome-wide significance, it will cause the score to become noisy, which could reduce predictive power. However, if the number of SNPs used in the score is reduced, there is a risk of being too stringent, which could lead to true associations being excluded. This in itself could also lead to a reduction in power to detect an association between the score and trait. With significant associations between score and breast cancer outcome observed when including a large number of SNPs in the score, it suggests that power was not compromised. When the number of SNPs was reduced, the association was still significant, so again, this would suggest that power was not compromised.

Even though the association results indicated that the SNP effects from one breast cancer GWAS could be used to predict breast cancer risk in an independent GWAS, sample sizes are currently not large enough to make accurate risk predictions using the polygenic scores derived in this analysis. This was affirmed by the computed AUC values which were approximately 54% - 60%. The individual SNP effect estimates are not yet accurate enough, but risk prediction will be feasible once the sample sizes increase. These results do however indicate that a large number of SNPs collectively explain the variation in breast cancer risk, and this was still observed after removing SNPs with a *p*-value ≤ 0.001 (lowest threshold), thus suggesting that breast cancer is a polygenic trait.

Also in this chapter, three different methods were used to estimate chip heritability on the liability scale, with each one producing a different estimate with varying 95% CIs. Common GWAS SNPs were estimated to explain between 12%-48% of the genetic variation in liability to breast cancer (liability scale heritability = ~44%). SNPs genotyped on the iCOGS array were found to capture between 13%-33% of the genetic variation in liability to breast cancer. These results shown that SNPs that have not been shown to reach genome-wide significance, do explain some of the genetic variation in breast cancer risk. These results were similar to those produced for other complex diseases, where common SNPs have been shown to explain approximately a third of the variation in a trait. By using larger sample sizes than used by Lu et al.(86) to produce a breast cancer h_l^2 (h_l^2: 13%, 95% CI:[0%-56%]), more precise estimates have been produced. The 95% CI for the published breast cancer h_l^2 estimate was quite wide, but the point estimate, considering the width of the 95% CI, is not too dissimilar to the h_l^2 estimates produced for the BBCS and UK2 studies. After estimating the genetic variation that can be explained by genotyped SNPs, there is still shown to be some unexplained genetic variation in breast cancer liability. The variation may be explained by rare causal SNPs that have an MAF lower than those picked up on the GWAS array, or by interactions.

Focussing on the 95% CIs alone, out of the three methods, GREML was shown to produce the most precise GWAS based h_l^2 estimate. However, when including more individuals in the analysis, by producing h_l^2 estimates based on COGS, AVENGEME was shown to produce a more accurate h_l^2 estimate. The GREML method could not be applied to the full COGS sample, so a subset of 10,000 subjects was used to produce a GREML h_l^2 estimate. This was the disadvantage of using this method, as it meant that I could not take advantage of the large COGS sample. LDSC was also found to be an unsuitable method to estimate h_l^2 from COGS data because the number of SNPs genotyped on the iCOGS array was less than 200,000, which meant that estimate precision was compromised. AVENGEME produced the most accurate h_l^2 estimate for the COGS, and, compared to the other two methods, it was able to handle both the large sample size and the reduced number of SNPs well. Palla & Dudbridge (88) have shown, using simulations, that the accuracy of AVENGEME estimates can be improved by clumping SNPs with a r^2 = 0.1, rather than pruning SNPs, or using a less stringent r^2 threshold when clumping. We can see from looking at the results in both Table 2-4 and Table 2-6, that the h_l^2 estimates based on LD clumped SNPs have narrower 95% CIs, than the 95% CIs for the estimates produced using LD-pruning SNPs. As an LD-thinning method, LD-clumping SNPs is generally preferred over LD-pruning. This is because it allows the SNPs with the strongest association with the trait, and possibly the SNPs that are most likely to be the causal SNP, out of a group of SNPs in high LD to be retained in an analysis.

The h_l^2 AVENGEME estimates suggest that GWAS SNPs explain a larger proportion of the genetic variation in breast cancer risk, compared to the SNPs genotyped on the iCOGS array. This was expected, as there were many more SNPs genotyped for the GWAS, then there were for the COGS. The estimated proportion of null markers present in either GWAS, were different to the proportion of null markers estimated to be present in the COGS. The proportion of null SNPs in the COGS, was estimated to be smaller than the proportions estimated for either GWAS. This was anticipated, as the iCOGS array is enriched for breast cancer associated SNPs. The h_l^2 and π_{01} estimates produced after LD-clumping the GWAS SNPs, ranged from 16%-21% and 90%-98%, respectively. These AVENGEME estimates are consistent with the AVENGEME estimates produced for other complex disease. Using published association results for polygenic scores and meta-analyses data, Palla & Dudbridge (88) have used AVENGEME to produce h_l^2 and π_{01} for five diseases; rheumatoid arthritis, celiac disease, myocardial infarction, type II diabetes and schizophrenia. The h_l^2 estimates ranged from 13%-34%, with the π_{01} estimate ranging from 85%-97%, which is similar to the estimates that were produced for the two breast cancer GWAS. Even though AVENGEME had not produced the most precise h_l^2 estimates for GWAS SNPs, out of the three methods used in this analysis presented in this chapter, the user does benefit from being able to also estimate the proportion of null SNPs in a study. Similar to LDSC, AVENGEME can produce h_l^2 estimates based on summary statistics, but unlike LDSC, AVENGEME has been shown to work well with studies where fewer than 200,000 SNPs have been genotyped. AVENGEME has also been shown to handle large samples, in terms of sample size, whereas GREML has not. Overall, AVENGEME has been shown to produce relatively precise h_l^2 estimates, considering sample size, and has been shown to work well with both GWAS and the larger COGS.

Another method that can be used estimate chip heritability is LDAK (Linkage Disequilibrium-Adjusted Kinship) (96, 97). The chip heritability method is not as widely used as GREML or LDSC, and has not been used to produce estimates in this thesis, which could be considered a limitation. Unlike the other methods used in this thesis, LDAK estimates chip heritability under the assumption that SNPs in regions of low LD contribute more genetic variation in disease, than SNPs in high LD regions. On the other hand GREML, for example, assumes that the genetic variation explained by a group of SNPs is influenced by the number of SNPs in the group, with each SNP explaining the same amount of genetic variation. When using LDAK, the LD between SNPs is taken into consideration. If the two SNPs are in high LD, LDAK would expect these SNPs to contribute half of the genetic variation of that explained by two SNPs that are not in LD. It is assumed that the signal of a SNP should be virtually captured by the SNPs it is in high LD with, so it may not be necessary to include this SNP when estimating the genetic variance explained. For a group of SNPs, LDAK allocates a weight to each SNP, with SNPs in low LD regions being assigned a higher weight than those in high LD regions. A SNP is given a weighting of zero if it is in high LD with nearby SNPs. In order to use LDAK to estimate chip heritability, raw genotype data is needed, summary statistics cannot be used.

This results in this chapter show the importance of continuing to use, and increase the size of, GWAS in order to identify genetic variants associated with breast cancer risk. Larger sample sizes are also needed to improve estimate precision, which is evident from looking at the estimates produced in this chapter. The chip heritability estimates produced using smaller sample sizes, have tended to have wider 95% CI, or larger standard errors. Samples therefore need to be as large as possible, in order to improve precision. The h_l^2 estimates produced, however, do show that GWAS have the potential to identify many more genetic variants, once sample sizes increase, as this will improve the power to detect the associated SNPs. At the time of writing this thesis, the OncoArray was under development (54) and the UK Biobank data, based on 500,000 individuals had announced its release. With the release of these large datasets, in terms of SNPs and the number of individuals genotyped, comes the exciting prospect of discovering many more breast cancer associated loci, if the estimates produced in this chapter are anything to go by.

Chapter 3 Genome partitioning of genetic variation for breast cancer

3.1 Introduction

In the previous chapter, the variation in liability to breast cancer that could be explained by genotyped SNPS (h_l^2) was estimated for the two breast cancer GWAS, and the COGS. In this chapter, the chip heritability contribution for different SNP subsets will be estimated in order to partition the h_l^2 estimates produced in chapter 2 by minor allele frequency (MAF), chromosome and SNP annotation. Partitioning the variance explained by genotyped SNPs will improve our understanding of how this variation is spread across the genome. If sections of the genome are found to explain more variation for breast cancer then other regions, then it could indicate areas of the genome where causal variants are more likely to be positioned.

3.1.1 Literature on genetic partitioning

Partitioning the genetic variance explained by GWAS SNPs allows for the identification of areas of the genome that could harbour causal variants. Genomic partitioning studies have been carried out on a variety of complex traits including schizophrenia (58), Alzheimer's disease, multiple sclerosis and endometriosis (98). From reviewing the partitioning studies, it was evident that genetic variance was commonly partitioned by MAF, chromosome or SNP annotation.

3.1.1.1 Partitioning by SNP MAF

Stratifying SNPs based on their MAF, and estimating the genetic variation that can be explained by the stratified SNPs, allows us to better understand how genetic variation is distributed across different MAFs. From reviewing the partitioning studies that have been carried out, it was evident that GCTAs GREML was the method that tended to be used when partitioning genetic variation by MAF (98, 99). GREML can be used to stratify all genotyped SNPs by their MAF, and produce an estimate of the variance explained by each MAF group.

Using GREML, Lee et al.(58) have partitioned the estimated genetic variation in liability for schizophrenia ($h_l^2 = $23%, se = 1%) by MAF, in order to explore whether common variants play an important role in the genetic basis of the disease. SNPs were assigned to one of five MAF bins; 0.01-0.1, 0.1-0.2, 0.3-0.4 and 0.4-0.5, with the genetic variation explained by the SNPs in each bin estimated. The 0.01-0.1 bin was estimated to explain the least amount of genetic variation (2%, se= 1%) compared to the other MAF bins, which each explained ~5% of the genetic variation (se = 1%). The authors believed this could have been due to the reduced number of SNPs in that bin, as SNPs with a MAF< 0.01 were removed in a QC step before undertaking the analysis, therefore causing less common SNPs to be under-represented. With this result, Lee at al. concluded that a considerable proportion of the genetic variation in liability was due to common causal variants.

Another study, conducted by Lee et al. (98), has also estimated h_l^2 for different traits using GWAS SNPs, and then partitioned this by SNP MAF. Three traits were examined; Alzheimer's disease, multiple sclerosis and endometriosis. GREML was used to produce h_l^2 estimates, based on GWAS SNPs retained after QC. The h_l^2 estimates for Alzheimer's disease, multiple sclerosis and endometriosis were 26% (se = 4%, 488,532 SNPs and 10,135 individuals), 24% (se = 3%, 499,757 SNPs and 7,139 individuals) and 30% (se = 3%, 293,474 SNPs and 3,557 individuals), respectively. SNPs for each of the traits were then assigned to one of the following MAF bins: MAF < 0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4 and 0.4-0.5, and then the h_l^2 for each bin was estimated using GREML. As expected, summing the h_l^2 estimate across the MAF bins for all three traits produced a summed estimate that was similar to the overall h_l^2 estimate produced for the trait. The 0.3-0.4 MAF bin was shown to explain the most genetic variation for all three traits, compared to the other MAF bins (Alzheimer's: h_l^2 = 8%, se

= 3%, MS: h_l^2 = 9%, se = 3% and endometriosis: h_l^2 = 8%, se = 3%). Common SNPs with a MAF > 0.1 were shown to explain a large proportion of the genetic variation for each of the three traits, with the proportion being approximately 90%.

Sieradzka et al. (99) have also used MAF partitioning, this time to investigate whether common genetic variants are important in the aetiology of psychotic experiences. Sieradzka et al. used three approaches to estimate SNP heritability, for each of following psychotic experiences; paranoia, hallucinations, cognitive disorganization (CD), grandiosity, anhedonia and negative symptoms (NS). MAF-stratification (partitioning by MAF) was one of three approaches used to estimate SNP heritability. Six MAF bins were used in the analysis to stratify SNPs; MAF< 0.05, 0.05-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4 and 0.4-0.5. GREML was then used to estimate the genetic variance explained by the SNPs in each MAF bin, for each psychotic experience. The sum of the chip heritability estimates produced for each partitioned MAF bin were shown to be fairly consistent with the overall chip heritability estimate produced for each psychotic experience. For two psychotic experiences, CD and Anhedonia, the MAF bin that explained the most genetic variation was the MAF < 0.05 bin (CD: 8.9%, se = 8% and Anhedonia: 21.4%, se = 8%). Over 40% of the genetic variation for Hallucinations and NS was estimated to be explained by SNPs within the 0.2-0.3 MAF bin (Hallucinations: 5.6%, se = 9% and NS: 4.7%, se = 9%). For Grandiosity and Paranoia, it was SNPs within the 0.3-0.4 and the 0.4-0.5 MAF bin, respectively, that explained the most genetic variation (Grandiosity: 12.1 %, se = 9% and Paranoia: 15.7%, se = 8%). The results from this analysis showed that, for the majority of the adolescent psychotic experiences analysed, SNPs with an MAF > 0.05 explain a larger proportion of the genetic variation in disease.

3.1.1.2 *Partitioning by chromosome*

Partitioning genetic variation by chromosome can be used to investigate whether specific chromosomes explain more of the genetic variation for disease, than other chromosomes. If this is shown to be the case, it could indicate that the chromosome harbours a larger number of causal variants. The analysis can also be used to examine whether the variance explained by a chromosome is proportional to its length (Mb). If there is found to be a relationship between chromosome length and the variance explained, it would imply that the disease has a polygenic basis, as the results would suggests that polygenic effects are spread evenly across the genome. When the chromosome group estimates are summed, the summed estimate should be close to the overall chip heritability estimate. If the two are not close, it indicates that population stratification may be present in the data. Population stratification could cause LD between chromosomes, which in turn would cause the genetic variation explained by each chromosome, when individually estimated, to be overestimated, as the variation explained by one chromosome could include the variation from other chromosomes (98, 100).

In a study conducted by Yang et al.(100), previously mentioned in chapter 2, the genetic variation for height, von Willebrand factor (vWF), QT interval (QTi) and BMI were estimated using GREML. They found that ~45% (se = 2.9%, 11,576 individuals) of the phenotypic variation in height, ~17% (se = 2.9%, 11,558 individuals) in BMI, ~25% (se = 5.1%, 6,641 individuals) in vWF and ~21% (se = 5%, 6,567 individuals) in QTi could be explained by 565,040 autosomal SNPs. The genetic variance for the four traits was then partitioned by chromosome and regressed against chromosome length (Mb), using a linear regression model. The results from this analysis suggested that the variance explained by each chromosome was proportional to chromosome length for both height and QTi, as there was shown to be a strong linear relationship between the two variables (height: *p* = 1.4 x 10^{-6} and R^2= 0.695, QTi: *p* = 1.1 x 10^{-3} and R^2= 0.422). The same however could not be shown for vWF and BMI, where the linear association between the two variables for the two traits was non-significant, with a small R^2 (vWF: *p* = 0.524 and R^2= 0.021, BMI: *p* = 0.214 and R^2= 0.076). Yang et al.

concluded from these results that even though there was shown to be a linear relationship between the estimated variance for a chromosome and chromosome length, the relationship was imperfect. Some chromosomes of similar length were found not to explain a similar proportion of variance, with variability across chromosomes being observed. This was especially the case for vWF and BMI, where a non-significant linear association between chromosome length and genetic variation explained was observed. Yang et al. explained that for vWF, much of the genetic variation for the trait is explained by a common SNP that maps within a gene (*ABO*). This would mean that the genetic variation for this trait is not as evenly spread across the genome, as the other traits.

As well as partitioning h_l^2 by MAF, Lee et al. (58) have also partitioned the h_l^2 for schizophrenia by chromosome using GREML. When summing up the genetic variation explained by each individual chromosome, the estimated h_l^2 for schizophrenia was 26%. This was compared to the overall h_l^2 estimate for schizophrenia, which was estimated to be 23%. From this, Lee at al. concluded that there was little evidence of population stratification being present in the data. Lee et al. also tested whether there was a significant linear relationship between chromosome length, and the genetic variation explained by a chromosome. They found evidence to suggest that a significant linear relationship between the two exists (*p* = 2.6 x 10^{-8} and R^2= 0.89), thus suggesting that schizophrenia has a polygenic basis.

Having partitioned the h_l^2 by MAF for Alzheimer's disease, MS and endometriosis, Lee et al.(98) also partitioned the h_l^2 estimates by chromosome, using GREML. For all three traits, the authors stated that they found the sum of the individual h_l^2 estimates for each chromosome, to be similar to the overall h_l^2 estimate. This meant that there was no evidence to suggest that population stratification affected the data. Lee et al. found that for MS and endometriosis, the estimated h_l^2 for each chromosome was linearly related to chromosome length (MS: *p* = 0.007 and R^2= 0.31 and endometriosis: *p* = 0.003 and

R^2= 0.37). The same was not initially shown for Alzheimer's (*p* = 0.49 and R^2= 0.024), until chromosome 19 was removed, which then made the relationship significant (*p* = 0.02 and R^2= 0.25). Therefore, a linear relationship between the genetic variance explained by a chromosome and chromosome length was shown for all three traits, when chromosome 19 was omitted from the Alzheimer's analysis.

3.1.1.3 *Partitioning by SNP annotation*

The genetic variation for a trait can be partitioned by SNP annotation, with SNP annotation being the function or effect that a SNP has. Yang et al. (100) have also partitioned genetic variation by SNP annotation, by partitioning h_l^2 onto intergenic and genic regions of the genome. SNPs were assigned to either intergenic or genic regions. Three different genic boundaries were used, ±0 kb, ±20 kb and ±50 kb, with these being based on the SNPs distance from protein coding genes. This meant that three different partitioning analyses were performed for each of the four traits, one analysis for each differently defined genic group. Consistently, genic SNPs were found to explain a larger proportion of the variation for each trait, even for the different genic boundaries (±0 kb, ±20 kb and ±50 kb). Yang et al. then partitioned the h_l^2 for intergenic and genic regions onto chromosome. The results from this analysis mainly showed that proportionally, genic regions explain more variation than intergenic regions across the chromosomes. On chromosome 9, the genic region was, however, shown to explain a much larger proportion of the genetic variation in vWF, than the intergenic region. As mentioned previously, it was known by Yang et al. that *ABO* on chromosome 9 explained a large amount of genetic variation for the trait, so this would explain why the genic region for this chromosome explained a much larger proportion of the genetic variance for vWF (101).

Lee et al.(98) partitioned the genetic variation for Alzheimer's disease, MS and endometriosis by two categories, SNPs located in genes and SNPs not located in genes. For endometriosis, the estimated h_l^2 explained by SNPs located in genes was the same as the estimated h_l^2 explained by SNPs not located in genes (Genes: h_l^2 = 13%, se = 3% and not in genes: h_l^2 = 13%, se = 3%). However, for Alzheimer's disease and MS, the estimated h_l^2 explained by SNPs located in genes was larger than the estimated h_l^2 explained by SNPs not located in genes (Alzheimer's - Genes: h_l^2 = 15%, se = 3% and not in genes: h_l^2 = 9%, se = 3%, MS - Genes: h_l^2 = 19%, se = 3% and not in genes: h_l^2 = 11%, se = 3%).

Gusev et al.(102) have partitioned the genetic variation explained by regulatory and coding variants for eleven diseases: rheumatoid arthritis, Crohn disease, type 1 diabetes, ulcerative colitis, MS, ankylosing spondylitis, schizophrenia, bipolar disorder, coronary artery disease, hypertension and type 2 diabetes. The genome was annotated based on six categories: coding, untranslated region (UTR), promoter, DNaseI hypersensitivity sites (DHSs), intronic and intergenic. SNPs were then assigned to one of six categories, with each SNP assigned to only one category. Gusev et al. analysed both genotyped SNPs and 1000 Genomes imputed SNPs, as well as simulating data for both. The genetic variation for each category was estimated using GREML, and after conducting a meta-analysis of the results across all eleven traits, DHSs SNPs were shown on average to explain ~79% (se=8%) of the heritability for imputed SNPs and 38% (se=4%) for genotyped SNPs. Using LDAK, instead of GREML, Speed et al.(97) have also estimated the proportion of genetic variation that can be explained by DHSs SNPs. Genetic variation was partitioned for ten diseases and for nine of the ten diseases, the same data as that used by Gusev et al. was analysed. Speed et al. observed that, on average, DHSs explained 25% of chip heritability. This was much lower than the average estimate of ~79% estimated by Gusev et al. The average then reduced very slightly to 24% when averaging over 42 traits instead of ten. This result shows that the average amount that DHSs SNPs contribute to chip heritability varies depending on the method used, with the LDAK method suggesting that DHSs contribute much less than originally estimated when using GREML.

3.2 Partitioning analyses

The chip heritability contribution for different SNP subsets were estimated in order to partition the chip heritability estimates produced for breast cancer in chapter 2. AVENGEME was able to handle both the smaller GWAS studies, and the larger COGS study, as well as the number of SNP genotyped in the studies, which made it an appropriate method to conduct the genome partitioning analysis. Also, AVENGEME had never been used to conduct a genome partitioning analysis, so it was a great opportunity to conduct the first genomic partitioning analysis using both polygenic risk scores and AVENGEME. For each partitioning analysis, the SNPs from each study were grouped, and both chip heritability and the proportion of null SNPs were estimated for each subset.

3.2.1 Genetic variance partitioned by MAF

In order to estimate the proportion of genetic variation in breast cancer liability that can be explained by common SNPs, UK2, BBCS and COGS SNPs were partitioned by MAF. The SNPs retained in each study after QC and LD-clumping (r^2 > 0.1) were assigned to one of five MAF bins; MAF<0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4 and 0.4-0.5. The association between SNP and breast cancer risk was then tested for the SNPs within each MAF bin, whilst adjusting for the relevant number of principal components for each study. Within each MAF bin, the SNPs were then grouped by their *p*-value and a polygenic score was constructed. The association between the polygenic score and breast cancer outcome was then tested, for each *p*-value group. From this, multiple *z*-scores were produced, which were then used to yield both a chip heritability estimate, and an estimate of the proportion of null SNPs for each MAF bin ($h^2_{l_{bin}}$ and $\pi_{01_{bin}}$). The multiple *z*-scores for each *p*-value interval, within each MAF bin for all three studies can be found in **Appendix 3**.

3.2.1.1 GWAS

Summing the ${h_l^2}_{bin}$ estimates across all five MAF bins produced a summed estimate close to the overall h_l^2 estimates for the BBCS and UK2 GWAS given in Table 2-4 (Table 3-1). This was expected, and actually hoped for, as this had been preciously observed for other complex diseases (98) (99) and because the same SNPs had been used in both analyses.

For the UK2 GWAS, the estimated ${h_l^2}_{bin}$ ranged from 0.012 to 0.067 across the MAF range, and 0.019 to 0.049 for the BBCS GWAS. The largest proportion of breast cancer variation captured by GWAS SNPs was observed in the 0.1-0.2 MAF bin, with the UK2 0.1-0.2 bin producing a ${h_l^2}_{bin}$ estimate of 0.067 (95% CI: [0.036, 0.099]), and 0.049 for the BBCS GWAS (95% CI: [0.000, 0.110]). For both studies, the MAF bin that contained the largest proportion of SNPs was also the 0.1-0.2 bin. The second largest ${h_l^2}_{bin}$ estimate, for both the UK2 and BBCS, was produced based on the SNPs assigned to the 0.2-0.3 MAF bin (UK2: ${h_l^2}_{bin}$ = 0.060, 95% CI: [0.035, 0.085] and BBCS: ${h_l^2}_{bin}$ = 0.048, 95% CI: [0.000, 0.099]). Both ${h_l^2}_{bin}$ estimates for this MAF bin were fairly close to the largest ${h_l^2}_{bin}$ estimate, therefore, much of the genetic variation in breast cancer liability could be explained by SNPs with a MAF between 0.1 and 0.3. For the UK2 GWAS, approximately 94% of the estimated genetic variance for breast cancer, on the liability scale, could be explained by common SNPs (MAF > 0.1), and for the BBCS GWAS the percentage was ~88%. Therefore, for both GWAS, the results showed that common genotyped SNPs with MAF > 0.1 capture a large proportion of the genetic variation in liability for breast cancer.

GWAS	MAF bin	No. SNPs (Proportion)	$h^2_{l\,bin}$ (95% CI)	Overall h^2_l*
UK2	< 0.1	20,800 (24.8%)	0.012 (0.000, 0.039)	
	0.1-0.2	25,715 (30.7%)	0.067 (0.036, 0.099)	
	0.2-0.3	15,382 (18.3%)	0.060 (0.035, 0.085)	
	0.3-0.4	11,678 (13.9%)	0.038 (0.016, 0.059)	
	0.4-0.5	10,276 (12.3%)	0.041 (0.021, 0.061)	
	Total	83,851	0.218	0.209
BBCS	< 0.1	14,114 (20.9%)	0.019 (0.002, 0.074)	
	0.1-0.2	20,168 (29.8%)	0.049 (0.000, 0.110)	
	0.2-0.3	13,301 (19.7%)	0.048 (0.000, 0.099)	
	0.3-0.4	10,559 (15.6%)	0.022 (0.000, 0.069)	
	0.4-0.5	9,512 (14.1%)	0.023 (0.001, 0.062)	
	Total	67,654	0.161	0.158

* AVENGEME h^2_l estimates produced in table 2.3

Table 3-1: Partitioning LD clumped GWAS SNPs by MAF

The proportion of null markers within each MAF bin ($\pi_{01_{bin}}$) was also estimated (Table 3-2). For both the UK2 and BBCS GWAS, the MAF bin estimated to have the smallest $\pi_{01_{bin}}$ was the 0.1-0.2 MAF bin (UK2 $\pi_{01_{bin}}$ = 0.000, 95% CI: [0.000, 0.956] and BBCS $\pi_{01_{bin}}$ = 0.241, 95% CI: [0.000, 1.000]). The UK2 $\pi_{01_{bin}}$ estimate for the 0.1-0.2 MAF bin was found to be extremely small, with the estimate suggesting that less than 1% of the SNPs within that MAF bin had no effect on breast cancer risk. With the 95% CI for this estimate being extremely wide (95% CI: [0.000, 0.956]), along with the other $\pi_{01_{bin}}$ estimates produced for the UK2 MAF bins, a reasonable conclusion based on $\pi_{01_{bin}}$ estimates could not be made.

The BBCS $\pi_{01_{bin}}$ estimate for the 0.1-0.2 MAF bin was larger than the UK2 $\pi_{01_{bin}}$ estimate, with the estimate suggesting that approximately 24% of the SNPs within the bin were null. Similarly, the 95% CI for the $\pi_{01_{bin}}$ estimate was very wide (95% CI: [0.000, 1.000]), and this was also found to be true for the majority of the BBCS MAF bins, so the precision of these estimates had to be questioned. However, for two of the MAF bins, MAF < 0.1 and 0.4-0.5, the 95% CI were fairly narrow. The $\pi_{01_{bin}}$ estimates for those two MAF bins indicated that over 99% of the genotyped SNPs within the two

MAF bins were null SNPs. With the 95% CIs for the estimated $\pi_{01_{bin}}$ produced for the other MAF bins being very wide, it is difficult to comment on how these two $\pi_{01_{bin}}$ estimates compare.

GWAS	MAF bin	No. SNP (Proportion)	$\pi_{01_{bin}}$ (95% CI)
UK2	< 0.1	20,800 (24.8%)	0.991 (0.000, 0.998)
	0.1-0.2	25,715 (30.7%)	0.000 (0.000, 0.956)
	0.2-0.3	15,382 (18.3%)	0.843 (0.000, 0.957)
	0.3-0.4	11,678 (13.9%)	0.856 (0.000, 0.981)
	0.4-0.5	10,276 (12.3%)	0.877 (0.000, 0.974)
	Total	83,851	
BBCS	< 0.1	14,114 (20.9%)	0.998 (0.897, 1.000)
	0.1-0.2	20,168 (29.8%)	0.241 (0.000, 1.000)
	0.2-0.3	13,301 (19.7%)	0.870 (0.000, 1.000)
	0.3-0.4	10,559 (15.6%)	0.986 (0.000, 0.999)
	0.4-0.5	9,512 (14.1%)	0.993 (0.928 0.998)
	Total	67,654	

Table 3-2: Estimated proportion of null SNPs ($\pi_{01_{bin}}$) for MAF bins based on GWAS SNPs

3.2.1.2 COGS

For the COGS, the $h^2_{l\,bin}$ estimates produced for each MAF bin ranged from 0.010-0.018 (**Table 3-3**). The summed h^2_l estimate across all five MAF bins was fairly close to the h^2_l estimate produced in Table 2-6 (h^2_l clumped **=** 0.059, 95% CI: [0.055, 0.063]). A reasonably large proportion of the h^2_l was shown to be explained by common SNPs with an MAF > 0.1 (78%). Just like the BBCS and UK2 study, the majority of the COGS SNPs used to estimate h^2_l had a MAF between 0.1 and 0.2. In this instance, the $h^2_{l\,bin}$ estimate produced for each MAF bin seemed to reflect the number of SNPs within the bin, with the 0.1-0.2 MAF bin explaining the largest proportion of the overall estimate h^2_l (12,564 SNPs, $h^2_{l\,bin}$= 0.018, 95% CI:[0.016, 0.020]). Dividing the $h^2_{l\,bin}$ by the number of SNPs within the bin, produced a per-SNP $h^2_{l\,bin}$ for each MAF bin, which was shown to be similar across the MAF bins. This would imply that the $h^2_{l_{bin}}$ estimate for an MAF bin is influenced by the number of SNPs assigned to the bin.

MAF bin	No. SNPs (Proportion)	$h^2_{l_{bin}}$ (95% CI)	h^2_l*	Per-SNP $h^2_{l_{bin}}$
< 0.1	9,937 (22.5%)	0.012 (0.010, 0.014)		1.21e-06
0.1-0.2	12,564 (28.4%)	0.018 (0.016, 0.020)		1.43e-06
0.2-0.3	8,465 (19.2%)	0.012 (0.010, 0.014)		1.42e-06
0.3-0.4	6,944 (15.7%)	0.010 (0.008, 0.011)		1.44e-06
0.4-0.5	6,271 (14.2%)	0.013 (0.011, 0.014)		2.07e-06
Total	44,181	0.064	0.059	

* h^2_l given in Table 2-6, based on clumped SNPs

Table 3-3: Partitioning COGS SNPs by MAF

Compared to the BBCS and UK2 GWAS, the 95% CIs for the $\pi_{01_{bin}}$ estimates produced for the COGS were a lot narrower, which would suggest that these $\pi_{01_{bin}}$ estimates were more precise than those given Table 3-2. The $\pi_{01_{bin}}$ estimate produced for the 0.4-0.5 MAF bin was the largest estimate produced across all five bins ($\pi_{01_{bin}}$= 0.930, 95% CI: [0.751, 1.000]). However, all $\pi_{01_{bin}}$ estimates were estimated to be close to 90%, so the estimated proportion of null SNPs in each bin were similar across the different MAF bins. Assuming that the underlying model is correct, a proportion of the SNPs are assumed to be null, and the remaining SNP effects are normally distributed on the standardised genotype scale (15). Therefore, approximately, 90% of the SNPs within each bin have been estimated to be null SNPs, which could be an unexpectedly high result considering the iCOGS array is enriched for breast cancer SNPs. However, not all the SNPs genotyped on the custom array have been shown to be associated with breast cancer. SNPs were included on the array if they have been shown to be associated with breast cancer, thought to have an association with breast cancer, shown to be associated with either ovarian or prostate cancer, or thought to be associated with either ovarian or prostate cancer.

MAF bin	SNP (Proportion)	$\pi_{01_{bin}}$ (95% CI)
< 0.1	9,937 (22.5%)	0.878 (0.833, 1.000)
0.1-0.2	12,564 (28.4%)	0.911 (0.878,1.000)
0.2-0.3	8,465 (19.2%)	0.924 (0.817, 1.000)
0.3-0.4	6,944 (15.7%)	0.891 (0.776,1.000)
0.4-0.5	6,271 (14.2%)	0.930 (0.751,1.000)
Total	44,181	

Table 3-4: Estimated proportion of null SNPs ($\pi_{01_{bin}}$) for each MAF bin based on COGS SNPs

3.2.2 Genetic variance partitioned by chromosome

Susceptibility loci for breast cancer have been identified across the genome, with SNPs on each chromosome contributing to the overall chip heritability for the disease. In this section, the estimated proportion of variation in breast cancer liability that can be explained by GWAS SNPs will be partitioned by chromosome. To perform this analysis, SNPs retained after QC and LD-clumping for each study were grouped by the chromosome they are positioned on. For each chromosome group, internal training and replication samples were used to construct a polygenic score for the replication individuals, based on the SNP effects estimated using the training sample. Within each chromosome group, the association between each SNP and breast cancer outcome was tested using a logistic regression model, with a number of principal components included in the model in order to adjust for population stratification. The number of principal components used was study dependent. Within each chromosome group, the SNPs were then grouped by their association *p*-value and a polygenic score was constructed for each *p*-value interval, based on the SNPs and their effect size within the interval. The association between each PRS and breast cancer risk in the replication sample was then tested using a logistic regression model, with principal components included in the model. This was performed for each chromosome, with the *z*-scores produced given in **Appendix 4: Tables 1,2 & 3**. The z-scores were then used to produce both a $h^2_{l\,bin}$ and $\pi_{01_{bin}}$ estimate for each chromosome subset using AVENGEME.

3.2.2.1 UK2 study

Summing the $h^2_{l\,bin}$ estimates across the 22 chromosomes produced a summed estimate similar to the h^2_l estimate produced in chapter 2, indicating that population stratification has been adjusted for (h^2_l clumped = 0.209 vs. $h^2_{l\,bin}$ sum = 0.220) (Table 3-5). Collectively, SNPs mapping to either chromosome 5 or chromosome 10 were observed to have a larger $h^2_{l\,bin}$ estimate compared to the other chromosomes (chr 5 $h^2_{l\,bin}$= 0.028, 95% CI: [0.014, 0.043] and chr 10 $h^2_{l\,bin}$ = 0.028, 95% CI: [0.014, 0.041]) (Table 3-5 and Figure 3-1). Interestingly, when looking at the number of published genome-wide significant breast cancer SNPs that map to each chromosome (45, 51), there were found to be more published SNPs mapping to chromosomes 5 and 10 than the other chromosomes (Table 3-6). Currently, nine published genome-wide significant breast cancer SNPs map to chromosome 2, which makes the $h^2_{l\,bin}$ estimate for chromosome 2 unrealistic (chr 2 $h^2_{l\,bin}$= 0.000, 95% CI: [0.000, 0.014]. The 95% CI for this estimate does however suggest that the $h^2_{l\,bin}$ estimate can range from 0 to 0.014. Therefore both this, and the fact that genome-wide significant SNPs have been found to map to this chromosome, suggest that the proportion of genetic variation explained by SNPs on this chromosome is not zero. Published breast cancer SNPs were found to be present on all chromosomes, so it is unlikely that any of the chromosomes explain none of the genetic variation in breast cancer liability. Also, considering how small the $h^2_{l\,bin}$estimates are, the 95% CI for all $h^2_{l\,bin}$ estimates could be considered fairly wide.

It was apparent from Figure 3-1 that the $h^2_{l\,bin}$ estimate for a chromosome was not necessarily reflective of chromosome length (Mb). If this were true, the $h^2_{l\,bin}$ would be shown to decrease, as the chromosome number increases. To formally test the relationship between $h^2_{l\,bin}$ estimate and chromosome length, a linear regression model was fitted to the data (Figure 3-2). From this, a weak linear relationship between chromosome length and the variation in liability explained by each chromosome was observed (R^2 = 0.114) (Figure 3-2). However, the relationship was non-significant, meaning that there was no evidence to suggest that the coefficient was different from 0 (*p*-value = 0.124).

Chromosome	No. SNPs	${h_l^2}_{bin}$ (95% CI)	h_l^2*
1	6,694	0.016 (0.000, 0.032)	
2	6,473	0.000 (0.000, 0.014)	
3	5,551	0.008 (0.000, 0.023)	
4	5,100	0.016 (0.002, 0.030)	
5	5,156	0.028 (0.014, 0.043)	
6	5,120	0.009 (0.000, 0.023)	
7	4,596	0.016 (0.003, 0.030)	
8	4,283	0.015 (0.001, 0.028)	
9	3,900	0.008 (0.000, 0.021)	
10	4,364	0.028 (0.014, 0.041)	
11	4,105	0.010 (0.000, 0.022)	
12	4,181	0.015 (0.003, 0.028)	
13	3,157	0.004 (0.000, 0.015)	
14	2,845	0.001 (0.000, 0.012)	
15	2,705	0.002 (0.000, 0.012)	
16	2,852	0.008 (0.001, 0.020)	
17	2,731	0.016 (0.006, 0.027)	
18	2,719	0.000 (0.000, 0.011)	
19	2,067	0.000 (0.000, 0.009)	
20	2,435	0.011 (0.001, 0.021)	
21	1,363	0.000 (0.000, 0.007)	
22	1,454	0.008 (0.000, 0.016)	
Total	83,851	0.220	0.209

* AVENGEME h_l^2 estimates produced in table 2.3

Table 3-5: Partitioning UK2 clumped SNPs by chromosome

Chromosome	Number of published SNPs*
1	9
2	9
3	5
4	2
5	11
6	9
7	4
8	7
9	5
10	10
11	5
12	4
13	2
14	5
15	1
16	4
17	3
18	3
19	3
20	1
21	1
22	4

* Taking into consideration lead SNPs only

Table 3-6: Published breast cancer genome-wide significant SNPs on each chromosome

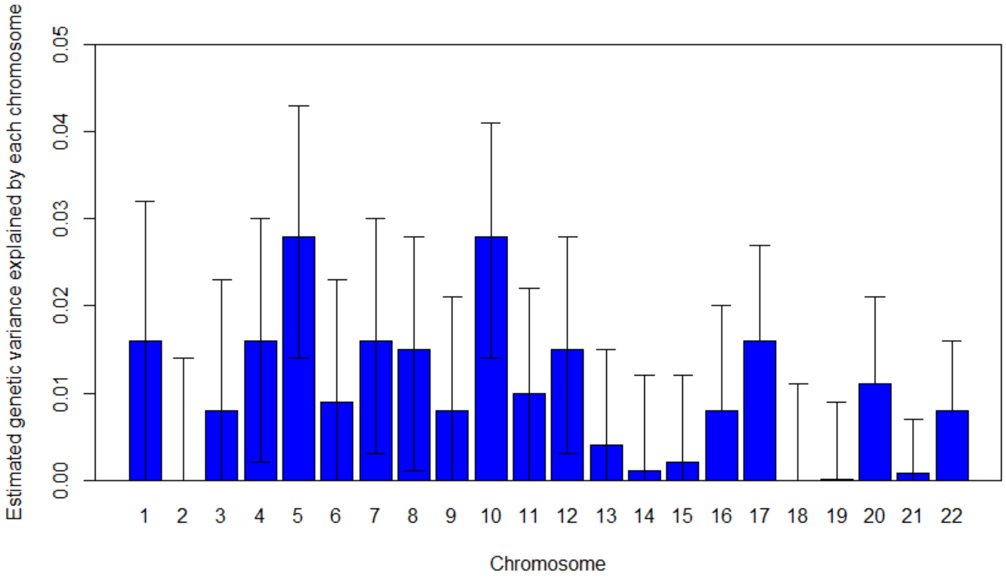


Figure 3-1: Proportion of variance in liability explained by UK2 clumped SNPs from each chromosome (95% CI error bars)

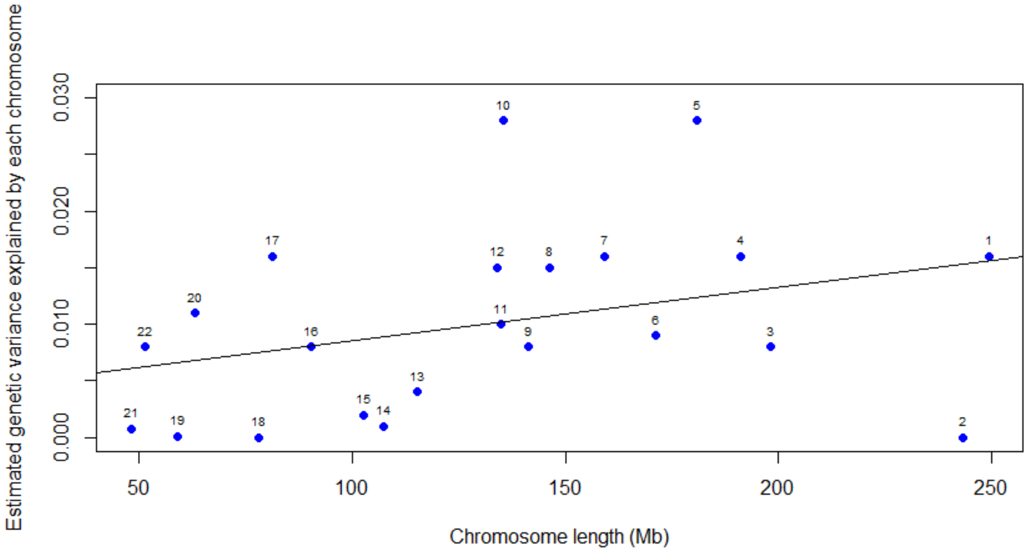


Figure 3-2: Proportion of variance in liability explained by UK2 clumped SNPs from each chromosome against chromosome length (Mb)

The proportion of null SNPs within each chromosome was also estimated (Table 3-7). For chromosomes 2 and 18, the estimated $h^2_{l\ bin}$ and $\pi_{01_{bin}}$ were identical (chr 2: $h^2_{l\ bin}$= 0.000, $\pi_{01_{bin}}$ =0.988 and chr 18: $h^2_{l\ bin}$ = 0.000, $\pi_{01_{bin}}$ = 0.998), as these were the values that maximised the likelihood numerically. Across the chromosomes, there was found to be much variation in the $\pi_{01_{bin}}$ estimates produced, with the estimates found to range from 0-1. Also, over half of the estimates had a very wide 95% CI, meaning that a reasonable conclusion based on these estimate could not be made. The $\pi_{01_{bin}}$ estimates for chromosome 10, chromosome 16, chromosome 18 and chromosome 19 had the narrowest 95% CIs. For these chromosomes, the $\pi_{01_{bin}}$ estimates ranged from 0.782-0.998, with the majority being over 0.930. With the other chromosomes $\pi_{01_{bin}}$ estimates found to have wide 95% CIs, a reasonable conclusion cannot be drawn from this analysis.

Chromosome	No. SNPs	$\pi_{01_{bin}}$ (95% CI)
1	6,694	0.000 (0.000, 0.985)
2	6,473	0.998 (0.530, 1.000)
3	5,551	0.953 (0.000, 0.981)
4	5,100	0.000 (0.000, 0.979)
5	5,156	0.886 (0.342, 0.968)
6	5,120	0.000 (0.000, 0.979)
7	4,596	0.626 (0.000, 0.976)
8	4,283	0.893 (0.000, 0.974)
9	3,900	0.001 (0.001, 0.971)
10	4,364	0.937 (0.837, 0.975)
11	4,105	0.931 (0.000, 0.973)
12	4,181	0.811 (0.000, 0.974)
13	3,157	0.951 (0.000, 1.000)
14	2,845	0.991 (0.000, 1.000)
15	2,705	0.005 (0.005, 0.957)
16	2,852	0.981 (0.847, 1.000)
17	2,731	0.927 (0.686, 0.957
18	2,719	0.998 (0.853, 1.000)
19	2,067	0.782 (0.772, 0.922)
20	2,435	0.259 (0.000, 0.952)
21	1,363	1.000 (0.000, 1.000)
22	1,454	0.685 (0.000, 0.916)
Total	83,851	

Table 3-7: Proportion of null UK2 SNPs ($\pi_{01_{bin}}$) within each chromosome

3.2.2.2 British Breast Cancer Study

For the BBCS GWAS, summing the $h^2_{l\ bin}$ estimates across the 22 chromosomes did not produce a summed estimate close to the h^2_l estimate produced in Table 2-4 (Table 3-8). The summed $h^2_{l\ bin}$ estimate was 0.256, whereas the overall h^2_l estimate was estimated to be 0.158, which meant that there was a difference of 0.098 between the two estimates. A difference between the two would indicate that population stratification is a problem in the data. In general, the 95% CIs for each chromosome estimate were wider than those for the UK2 estimates. With the BBCS study sample size being just under half the size of the UK2 study, the result confirms that sample size affects estimate precision. When producing 22 separate estimates and then summing them together, it would just take a small amount of variation in the ${h_l^2}_{bin}$ estimates, for the summed ${h_l^2}_{bin}$ estimate to be different to the overall h_l^2 estimate, which in itself is fairly variable (h_l^2 95% CI: [0.047,0.272]). Therefore, when summing the ${h_l^2}_{bin}$ estimates across 22 chromosome, the inaccuracy of each binned estimate becomes more apparent. With population stratification being adjusted for, and the genomic inflation statistic being close to 1 (1.014), population stratification should not be present in the data.

The ${h_l^2}_{bin}$ estimates for chromosomes 5, chromosome 6, chromosome 14, chromosome 15, chromosome 17, chromosome 18, chromosome 19 and chromosome 20 were approximately 0, which would indicate that SNPs mapping to these chromosomes either do not contribute to the heritability for breast cancer, or have little effect on breast cancer risk. However, with the 95% CI intervals for these estimates being wide, it could not be concluded that SNPs mapping to these chromosomes do not explain any of the estimated h_l^2. Nonetheless, it would have been unwise to have drawn this conclusion as genetic mutations within genes such as *TERT*, *MAP3K1* and *RAD51*, are located on these chromosomes, and have been shown to be associated with breast cancer risk (44, 103, 104). Also, each chromosome has been shown to have at least one published genome-wide significant breast cancer SNP mapping to it. Chromosome 2 SNPs were estimated to explain a larger proportion of the genetic variation in breast cancer, compared to the other chromosomes (Table 3-8 and Figure 3-3) (${h_l^2}_{bin}$= 0.042, 95% CI: [0.010, 0.075]). For the UK2 GWAS however, the ${h_l^2}_{bin}$ estimate for SNPs genotyped on chromosome 2 was approximately 0. The 95% CI for this estimate suggested that this estimate could in fact be non-zero as the 95% CI had an upper limit that was equal to 0.014 (95% CI: [0.000, 0.014]). The upper 95% CI limit was still not as large as 0.042, which was the ${h_l^2}_{bin}$ estimate produced for BBCS SNPs on chromosome 2. Located on chromosome 2 is the *CASP8* gene, which has been linked to breast cancer risk, meaning that a ${h_l^2}_{bin}$ estimate of 0 for this chromosome would be very unlikely (105).

Chromosome	No. SNPs	${h_l^2}_{bin}$ (95% CI)	h_l^2*
1	5,313	0.015 (0.000, 0.049)	
2	5,333	0.042 (0.010, 0.075)	
3	4,585	0.025 (0.000, 0.055)	
4	4,128	0.041 (0.012, 0.069)	
5	4,190	0.000 (0.000, 0.015)	
6	4,176	0.000 (0.000, 0.021)	
7	3,632	0.017 (0.001, 0.043)	
8	3,551	0.007 (0.000, 0.031)	
9	3,168	0.002 (0.000, 0.026)	
10	3,482	0.008 (0.001, 0.030)	
11	3,207	0.026 (0.001, 0.051)	
12	3,366	0.037 (0.014, 0.064)	
13	2,515	0.011 (0.000, 0.034)	
14	2,275	0.000 (0.000, 0.024)	
15	2,122	0.000 (0.000, 0.012)	
16	2,220	0.007 (0.000, 0.0270)	
17	2,102	0.000 (0.000, 0.018)	
18	2,286	0.000 (0.000, 0.009)	
19	1,645	0.000 (0.000, 0.013)	
20	1,937	0.000 (0.000, 0.015)	
21	1,173	0.012 (0.000, 0.028)	
22	1,248	0.005 (0.000, 0.021)	
Total	67,654	0.256	0.158

* AVENGEME h_l^2 estimates produced in table 2.3

Table 3-8: Partitioning BBCS clumped SNPs by chromosome

When plotting the estimated ${h_l^2}_{bin}$ for each chromosome, there was no obvious relationship between the estimated ${h_l^2}_{bin}$ and chromosome length (Figure 3-3). However, from assessing Figure 3-4 and fitting a linear regression model to the data, to test whether a significant linear relationship between the two variables existed, it was apparent that there was a significant linear relationship between chromosome length and the estimated h_l^2 for SNPs on each chromosome (R^2 = 0.315, *p*-value = 0.007). Therefore, the results suggested that breast cancer has a polygenic basis.

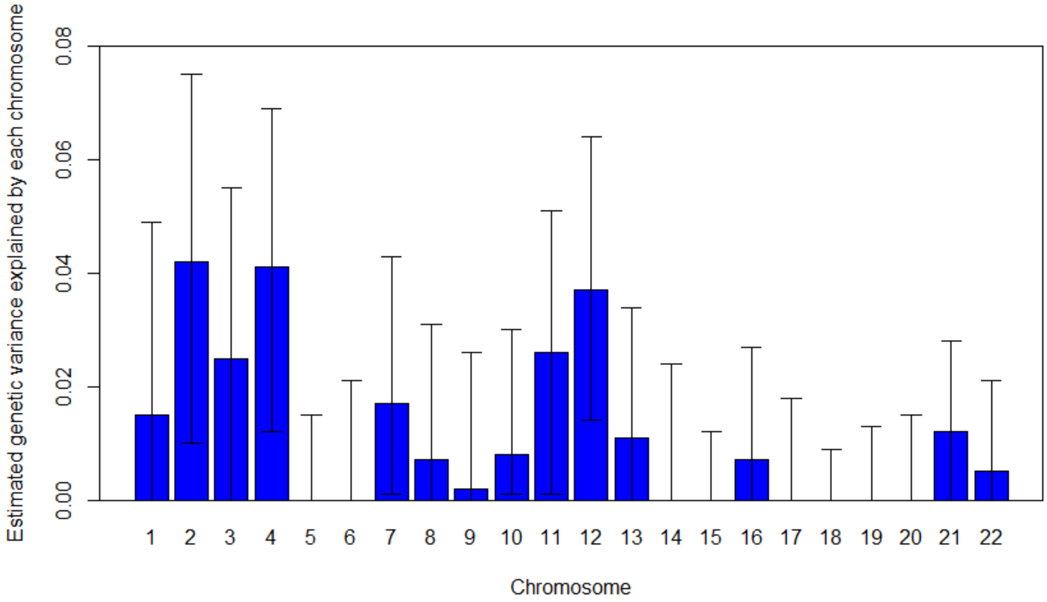


Figure 3-3: Proportion of variance in liability explained by BBCS clumped SNPs from each chromosome

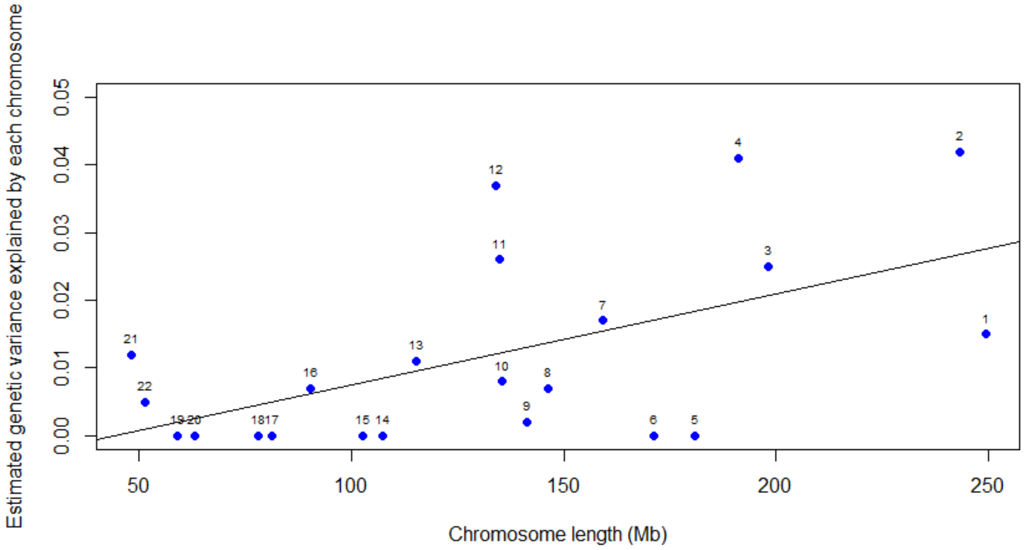


Figure 3-4: Proportion of variance in liability explained by BBCS clumped SNPs from each chromosome against chromosome length (Mb)

The proportion of null SNPs within each chromosome was also estimated for the BBCS GWAS (Table 3-9). For chromosomes 5, chromosomes 6, chromosomes 14, chromosomes 15, chromosomes 17 and chromosomes 20, and chromosomes 18 and chromosomes 19, the same $h^2_{l\,bin}$ and $\pi_{01_{bin}}$ estimates were produced as these were the values that maximised the likelihood numerically. The $\pi_{01_{bin}}$estimate for chromosome 2, the chromosome estimated to explain the most variation (Table 3-8), was estimated to be 0.122 (95% CI: [0.122, 0.982]) (Table 3-9). This was a low $\pi_{01_{bin}}$, which matched the lower bound value of the 95% CI, and suggested that 87.8% of the SNPs mapping to chromosome 2 had an effect on breast cancer risk. With the 95% CI for this estimate being wide, and with most of the other autosomal chromosomes having a wide 95% CI, a reasonable conclusion based on the $\pi_{01_{bin}}$ estimates could not made.

Chromosome	No. of SNPs	$\pi_{01_{bin}}$ (95% CI)
1	5,313	0.992 (0.000, 1.000)
2	5,333	0.122 (0.122, 0.982)
3	4,585	0.919 (0.000, 0.992)
4	4,128	0.000 (0.000, 0.962)
5	4,190	0.998 (0.000, 1.000)
6	4,176	0.998 (0.776, 1.000)
7	3,632	0.986 (0.000, 1.000)
8	3,551	0.988 (0.000, 1.000)
9	3,168	1.000 (0.000, 1.000)
10	3,482	0.997 (0.944, 1.000)
11	3,207	0.000 (0.000, 0.986)
12	3,366	0.948 (0.807, 0.987)
13	2,515	0.714 (0.000, 0.714)
14	2,275	0.998 (0.000, 1.000)
15	2,122	0.998 (0.426, 1.000)
16	2,220	0.977 (0.000, 1.000)
17	2,102	0.998 (0.444, 1.000)
18	2,286	1.000 (0.466, 1.000)
19	1,645	1.000 (0.782, 1.000)
20	1,937	0.998 (0.816, 1.000)
21	1,173	0.007 (0.007,0.955)
22	1,248	0.002 (0.002, 0.958)
Total	67,654	

Table 3-9: Proportion of null BBCS SNPs (π_{01bin}) within each chromosome

3.2.2.3 COGS

In addition to partitioning the h_l^2 by chromosome for the two GWAS, the estimated h_l^2 explained by SNPs genotyped for the COGS was also partitioned by chromosome. Summing the ${h_l^2}_{bin}$ estimates for each chromosome, produced an estimate that was close to the h_l^2 estimate produced for the COGS, given in Table 2-6 (Table 3-10). The SNPs mapping to chromosome 10 were estimated explain the largest proportion of genetic variation in liability for breast cancer, compared to the SNPs mapping to other autosomal chromosomes (${h_l^2}_{bin}$ = 0.006, 95% CI: [0.005, 1.000]) (Table 3-10) (Figure 3-5). Again, this is one of the chromosomes, the other being chromosome 5, that currently has the most published genome-wide significant breast cancer SNPs mapping to it. However, the 95% CI for this estimate was fairly wide, so the precision of the ${h_l^2}_{bin}$ estimate needs to be questioned. Apart from chromosomes 1, chromosome 2 and chromosome 3, the 95% CIs for the ${h_l^2}_{bin}$ estimates were shown to be very wide (Table 3-10 and Figure 3-6). The 95% CIs for the ${h_l^2}_{bin}$ estimates were so wide, that the upper limits for the majority of chromosomes could not be seen when visualising the estimates and their 95% CIs (Figure 3-6).

From plotting the estimated ${h_l^2}_{bin}$ for each chromosome and chromosome length, it seemed like there was a linear relationship between the two variables (Figure 3-7). A linear regression model was used to formally test the relationship, and the estimated ${h_l^2}_{bin}$ for each chromosome and chromosome length were shown to have a significant linear relationship (R^2 = 0.498, *p*-value = 0.00025). With SNPs on the iCOGS array being genotyped for their relationship with breast, ovarian and prostate cancer or based on previous analyses, and not because they tag most of the genome, some parts of the genome may be underrepresented on the array. With the SNP not necessarily representing the genome, it might be better to assess the relationship between the number of SNPs genotyped for each chromosome, then chromosome length. When testing this relationship, there was shown to be a significant linear relationship between the number of SNPs on a chromosome and the estimated ${h_l^2}_{bin}$ for each chromosome (R^2 = 0.642, *p*-value = 7.36e-06) (Figure 3-8). The significant linear relationship observed, suggest that the genetic variation in liability to breast cancer explained by COGS SNPs, is spread evenly across the genome.

Chromosome	No. of SNPs	h_l^2 (95% CI)	h_l^2*
1	3,406	0.004 (0.003, 0.005)	
2	3,480	0.005 (0.003, 0.006)	
3	2,876	0.005 (0.004, 0.006)	
4	2,559	0.002 (0.001, 1.000)	
5	2,632	0.005 (0.004, 1.000)	
6	2,783	0.005 (0.004, 1.000)	
7	2,398	0.002 (0.001, 1.000)	
8	2,351	0.004 (0.003, 1.000)	
9	2,055	0.003 (0.002, 1.000)	
10	2,498	0.006 (0.005, 1.000)	
11	2,224	0.004 (0.003, 1.000)	
12	2,212	0.005 (0.004, 1.000)	
13	1,688	0.000 (0.000, 1.000)	
14	1,496	0.002 (0.002, 1.000)	
15	1,383	0.001 (0.001, 1.000)	
16	1,437	0.003 (0.002, 1.000)	
17	1,475	0.001 (0.001, 1.000)	
18	1,371	0.001 (0.001, 1.000)	
19	1,170	0.001 (0.001, 1.000)	
20	1,197	0.001 (0.001, 1.000)	
21	690	0.000 (0.000, 1.000)	
22	800	0.001 (0.001, 1.000)	
Total	44,181	0.063	0.059

* AVENGEME h_l^2 estimates produced in table 2.5

Table 3-10: Partitioning COGS SNPs by chromosome

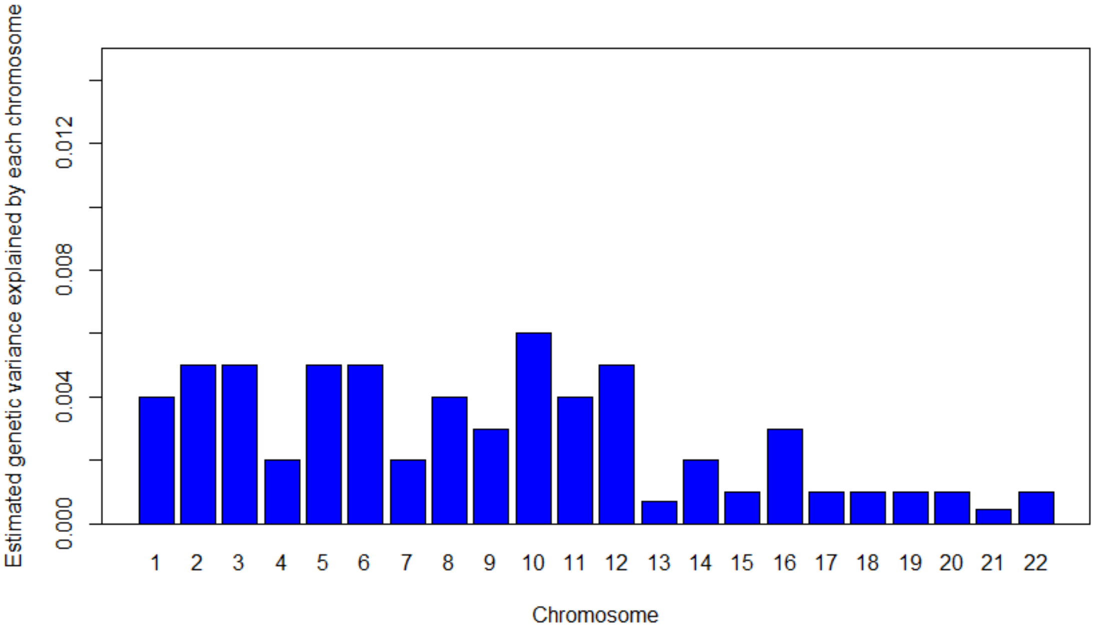


Figure 3-5: Proportion of variance in liability explained by COGS clumped SNPs on each chromosome

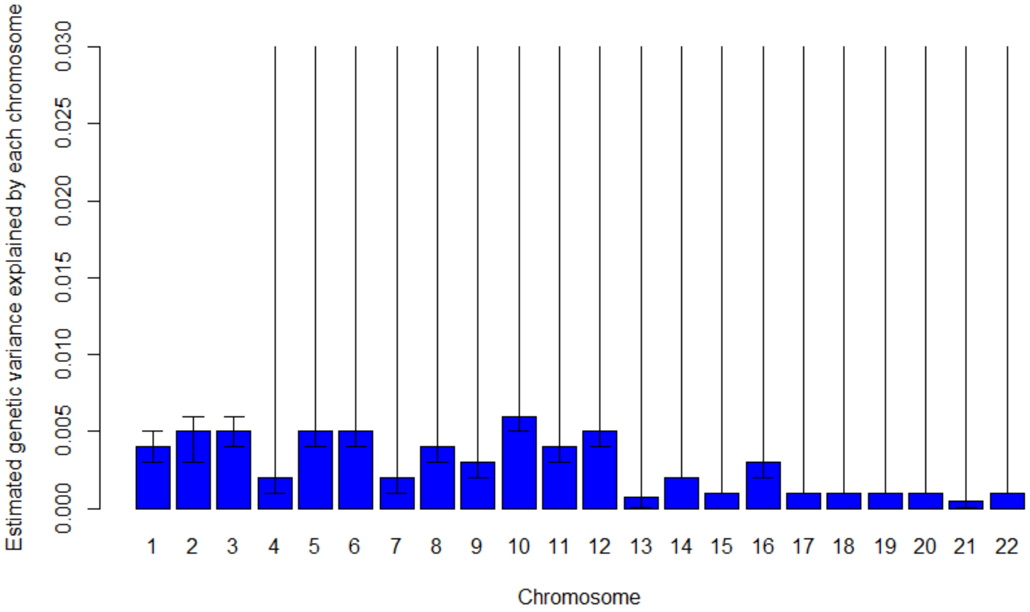


Figure 3-6: Proportion of variance in liability explained by COGS clumped SNPs on each chromosome (95% CI error bars)

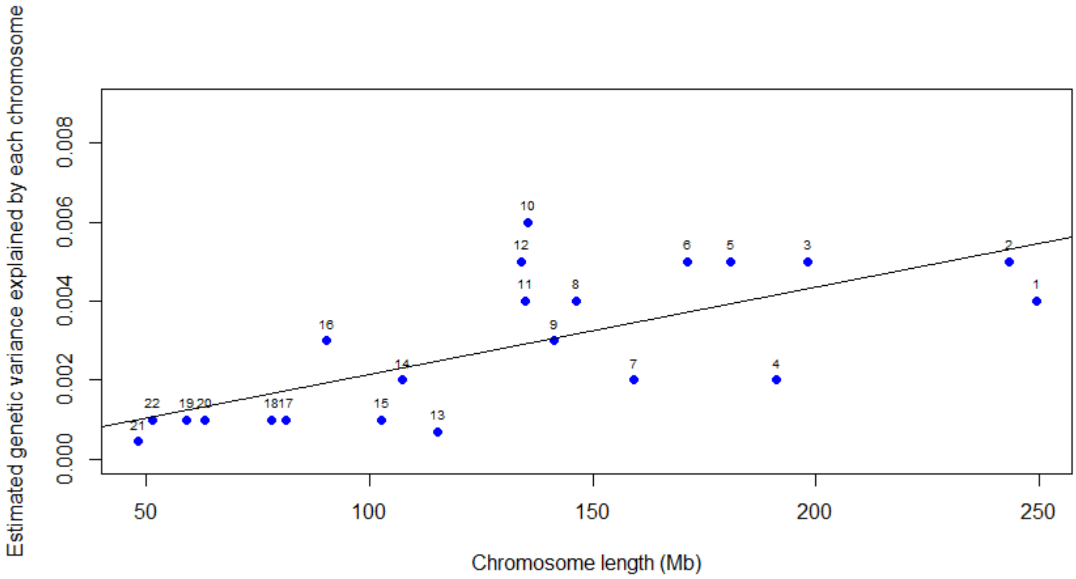


Figure 3-7: Proportion of variance in liability explained by COGS clumped SNPs from each chromosome against chromosome length (Mb)

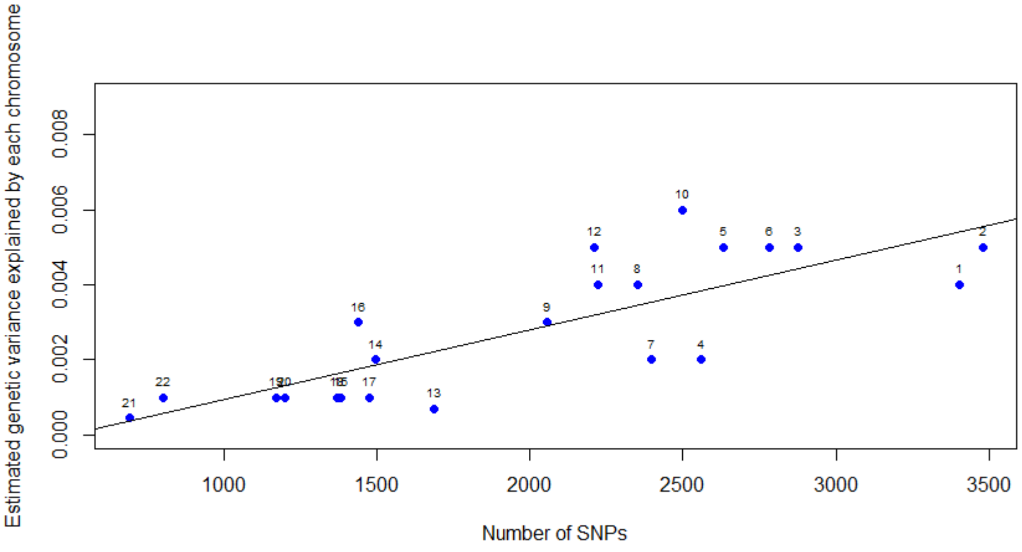


Figure 3-8: Proportion of variance in liability explained by COGS clumped SNPs from each chromosome against the number of SNPs

For each chromosome, the proportion of iCOGS SNPs with no effect on breast cancer risk, was also estimated (Table 3-11). The $\pi_{01_{bin}}$ estimates were all observed to be greater than 0.567, with approximately half of the $\pi_{01_{bin}}$ estimates having fairly narrow 95% CIs. The 95% CIs were narrower than those for the BBCS and UK2 $\pi_{01_{bin}}$ chromosome estimates. Again, this would have been because the COGS had a larger sample size than the BBCS and UK2 GWAS. Some of the $\pi_{01_{bin}}$ and $h_{l\ bin}^2$ estimates were identical as these were the values that maximised the likelihood numerically. Chromosome 16 was estimated to have the largest proportion of null SNPs mapping to a chromosome, compared to the other chromosomes, with the result suggesting that ~96% of the SNPs mapping to this chromosome have no effect on breast cancer risk ($\pi_{01_{bin}}$: 0.959 (95% CI: [0.959, 1.000]) (Table 3-11). The $\pi_{01_{bin}}$ estimates produced for COGS cannot be compared to those produced for the BBCS and UK2 GWAS, because the estimates have very wide 95% CIs. But, it can be said that most of the $\pi_{01_{bin}}$ estimates were observed to be less than 0.95, with the proportions shown to be fairly consistent across the chromosomes.

Chromosome	No. of SNPs	$\pi_{01_{bin}}$ (95% CI)
1	3,406	0.841 (0.539, 1.000)
2	3,480	0.921 (0.549, 1.000)
3	2,876	0.813 (0.453, 1.000)
4	2,559	0.950 (0.385, 1.000)
5	2,632	0.906 (0.402, 1.000)
6	2,783	0.786 (0.435, 1.000)
7	2,398	0.874 (0.343, 1.000)
8	2,351	0.895 (0.895, 1.000)
9	2,055	0.847 (0.847, 1.000)
10	2,498	0.923 (0.370, 1.000)
11	2,224	0.890 (0.890 1.000)
12	2,212	0.830 (0.830, 1.000)
13	1,688	0.936 (0.936, 1.000)
14	1,496	0.916 (0.916, 1.000)
15	1,383	0.621 (0.621, 1.000)
16	1,437	0.959 (0.959, 1.000)
17	1,475	0.940 (0.940, 1.000)
18	1,371	0.719 (0.719, 1.000)
19	1,170	0.923 (0.923, 1.000)
20	1,197	0.567 (0.567, 1.000)
21	690	0.955 (0.955, 1.000)
22	800	0.877 (0.877, 1.000)
Total	44,181	

Table 3-11: Proportion of null COGS SNPs ($\pi_{01_{bin}}$) within each chromosome

3.2.3 Genetic variance partitioned by SNP annotation

The estimated h_l^2 for the two GWAS was partitioned by SNP annotation to examine whether, across the genome, SNPs mapping to intergenic regions explain a larger proportion of the estimated h_l^2**,** compared to SNPs that map to gene regions. The estimated h_l^2 for the COGS was also partitioned by SNP annotation to explore whether the SNPs genotyped on the iCOGS array that map to intergenic regions, explain a greater proportion of the h_l^2**,** compared to the other SNPs. The majority of breast cancer susceptibility loci identified so far, map to non-coding, intergenic regions of the genome (106). Even though a larger number of susceptibility loci have been shown to map to intergenic regions of the genome, it is still possible that variants residing within genes explain more of the genetic variation in breast cancer. This analysis therefore aimed to test whether SNPs mapping to intergenic regions of the genome explain a larger proportion of h_l^2, compared to SNPs mapping elsewhere in the genome.

GWAS and COGS SNPs retained after QC and LD-clumping were annotated using the ENSEMBL Variant Effect Predictor (VEP) web interface (107). The web interface enables you to upload a list of SNP identifiers, these being the SNPs you wish to annotate, for a given species. ENSEMBL's VEP then reports the effect of each SNP that has been annotated. Not all of the SNPs uploaded have been annotated by ENSEMBL, so annotation information was not provided for some of the SNPs. This meant that the total number of SNPs analysed for each study in this section, were lower than the total number of SNPs used for the previous analyses conducted in this chapter.

The SNP information for annotated SNPs was exported into an Excel spreadsheet, which was then used to group the SNPs. Each SNP was identified as being 1 of the 18 SNP annotations given in Table 3-12. To examine whether gene variants explained a larger proportion of the h_l^2 compared to other SNPs, annotated SNPs were grouped into three categories; intergenic variants, regulatory variants and gene variants (Table 3-12). Each SNP belonged to a unique annotation group, which meant that none of the SNPs overlapped. The SNPs were grouped with the aid of the sequence ontology tree diagram (108) given in **Appendix 5: Diagram 1**.

SNP annotation	Annotation group
3 prime UTR variant	Gene variant
Intron variant	Gene variant
Intergenic variant	Intergenic variant
Downstream gene variant	Intergenic variant
Upstream gene variant	Intergenic variant
Splice donor variant	Gene variant
Non coding transcript exon variant	Gene variant
Regulatory region variant	Regulatory variant
Missense variant	Gene variant
Synonymous variant	Gene variant
Splice region variant	Gene variant
Stop gained	Gene variant
5 prime UTR variant	Gene variant
Stop lost	Gene variant
TF binding site variant	Regulatory variant
Splice acceptor variant	Gene variant
Start lost	Gene variant
Stop retained variant	Gene variant

Table 3-12: SNP annotation groups

Once the SNPs had been grouped, the internal training and replication samples for each study were used to construct a threshold PRS for the replication sample individuals, based on the SNPs within each annotation group. A logistic regression model, with principal components included as covariates, was used to estimate the SNP effects for the SNPs within each annotation group. These SNPs effects were grouped by their *p*-value, and then for each *p*-value threshold within an annotation group, a polygenic score was constructed for the women in the replication sample. A logistic regression model, with principal components included as covariates, was then used to test the association between each polygenic score and breast cancer risk. The *z*-scores produced, given in **Appendix 6: Tables 1,2 & 3**, were then used to estimate both the $h^2_{l\,bin}$ and $\pi_{01_{bin}}$ for each annotation group.

3.2.3.1 GWAS

When annotating the GWAS SNPs, it was found that a larger number of the SNPs were annotated as being gene variants, compared to intergenic and regulatory variants (Table 3-13). This was surprising as there are a lot more non-genic SNPs than genic SNPs in the genome, so one would not have expected most of the SNPs to be genic SNPs. The SNPs annotated were the genotyped SNPs retained after LD clumping. It could be that the SNPs that have the strongest association with breast cancer in an LD block, tend to be genic. To assess whether a greater number of genic SNPs were retained in the analysis because of clumping, SNPs retained after LD pruning were annotated. It was found that even after randomly pruning SNPs, the majority of SNPs were annotated as being genic, which indicates that this finding was not a result of clumping the SNPs. It was therefore not clear why a greater number of SNPs were found to be genic, than non-genic.

Using the annotations given by VEP, gene variants were collectively estimated to explain a larger proportion of the h^2_l, compared to intergenic and regulatory variants (UK2: $h^2_{l\,bin}$= 0.117, 95% CI: [0.074, 0.160] and BBCS: $h^2_{l\,bin}$= 0.150, 95% CI: [0.056, 0.245]). The 95% CI for the BBCS gene variant $h^2_{l\,bin}$ estimate was wider than the UK2 $h^2_{l\,bin}$ estimate, but there was little difference in the actual gene variant $h^2_{l\,bin}$ estimates produced.

With more SNPs being annotated as gene variants, than either intergenic or regulatory variants, the $h^2_{l\,bin}$ estimates were divided by the number of SNPs within each annotation group to produce an approximate estimate of the per-SNP ${h_l^2}_{bin}$. This was conducted in order to assess whether individually, gene variants do explain more genetic variation, or whether it was just because the gene variant group contained the largest number of SNPs. Per UK2 SNP, SNPs mapping to intergenic regions were estimated to explain a slightly larger proportion of the h_l^2, than SNPs mapping to either gene or regulatory regions. However, the difference between the per-SNP ${h_l^2}_{bin}$ estimates across the three groups was minuscule. So per-UK2 SNP, it could not be concluded whether one type of variant explained more h_l^2 than another.

Per-BBCS SNP, SNPs mapping to gene regions explained a larger proportion of the h_l^2, than SNPs mapping to either intergenic or regulatory regions. There was shown to be a slightly more noticeable difference between the BBCS per-SNP ${h_l^2}_{bin}$ estimate for each annotation group, then there was between the UK2 per-SNP ${h_l^2}_{bin}$ estimates. With the largest per-SNP ${h_l^2}_{bin}$ estimate varying between the two GWAS, a conclusion based on these results could not be made.

GWAS	Annotation group	No. SNPs	${h_l^2}_{bin}$ (95% CI)	${h_l^2}_{bin}$ per-SNP
UK2	Intergenic variant	32,406	0.089 (0.054, 0.125)	2.75e-06
	Regulatory variant	3,757	0.009 (0.000, 0.022)	2.50e-06
	Gene variant	47,642	0.117 (0.074, 0.160)	2.45e-06
	Total	83,805	0.215	
BBCS	Intergenic variant	25,119	0.017 (0.000, 0.074)	6.74e-07
	Regulatory variant	2,916	0.005 (0.000, 0.024)	1.71e-06
	Gene variant	35,916	0.150 (0.056, 0.245)	4.18e-06
	Total	63,951	0.172	

Table 3-13: Partitioning GWAS SNPs by SNP annotation

The $\pi_{01_{bin}}$ was also estimated for each annotation group, for each GWAS (Table 3-14). For the UK2 GWAS, the gene variant group was estimated to have the smallest proportion of null SNPs, compared to the other two annotation groups ($\pi_{01_{bin}}$= 0.000, 95% CI: [0.000, 0.949]) (Table 3-14). The estimate suggests that < 1% of the SNPs do not have an effect on breast cancer risk. However, the 95% CI for this estimate was found to be very wide, so the precision of the estimate was questioned. It was also estimated that ~6% of the SNPs in the regulatory group did not have an effect on breast cancer risk, again, the 95% CI for this estimate was found to be very wide. Similarly to the UK2 GWAS, the gene variant group for the BBCS was estimated to contain the lowest proportion of null SNPs, with the estimate suggesting that ~95% of the SNPs within the group have no effect on breast cancer risk ($\pi_{01_{bin}}$= 0.948, 95% CI: [0.000, 0.995]). The estimates produced for both the regulatory and intergenic group suggest that none of the SNPs in either group have an effect on breast cancer risk. This was an unrealistic result, but with extremely wide 95% CI, the estimate was not very precise. With the 95% CIs for all three groups being extremely wide, a conclusion could not be made based on these estimates.

	Annotation group	No. SNPs	$\pi_{01_{bin}}$ (95% CI)
UK2 GWAS	Intergenic variant	32,406	0.943 (0.745, 0.983)
	Regulatory variant	3,757	0.056 (0.056, 0.997)
	Gene variant	47,642	0.000 (0.000, 0.949)
	Total	83,805	
BBCS GWAS	Intergenic variant	25,119	1.000 (0.000, 1.000)
	Regulatory variant	2,916	1.000 (0.000, 1.000)
	Gene variant	35,916	0.948 (0.000, 0.995)
	Total	63,951	

Table 3-14: Proportion of GWAS null SNPs ($\pi_{01_{bin}}$) within each annotation group

3.2.3.2 COGS

For the COGS, a large majority of the genotyped SNPs were annotated as being genic SNPs, with the largest proportion of h_l^2 then being explained by genic SNPs (Table 3-15).. The second largest group, the intergenic variant group, was also the group that explained the second largest proportion of h_l^2. The ${h_l^2}_{bin}$ estimate based on regulatory variants was found to be quite small in comparison to the ${h_l^2}_{bin}$ estimates produced for intergenic and gene variants. The 95% CIs for each ${h_l^2}_{bin}$ estimate was found to be a lot narrower than the 95% CIs for the SNP annotation group ${h_l^2}_{bin}$ estimates produced in the previous section, for the two GWAS.

When estimating the ${h_l^2}_{bin}$ per-SNP, there was shown to be little difference in the per-SNP estimates, with genic SNPs only estimated to explain a tiny bit more h_l^2 than intergenic or regulatory region SNPs.

Annotation group	No. SNPs	${h_l^2}_{bin}$ (95% CI)	${h_l^2}_{bin}$ per-SNP
Intergenic variant	16,933	0.022 (0.020, 0.024)	1.30e-06
Regulatory variant	1,969	0.003 (0.002, 0.004)	1.52e-06
Gene variant	24,606	0.039 (0.036, 0.042)	1.58e-06
Total	43,508	0.064	

Table 3-15: Partitioning COGS SNPs by SNP annotation

The annotation group estimated to have the largest $\pi_{01_{bin}}$ was the regulatory variant group ($\pi_{01_{bin}}$ = 0.906, 95% CI: [0.832, 0.963]) (Table 3-16). This estimate suggested that ~91% of the genotyped regulatory SNPs had no effect on breast cancer risk. The gene variant group was observed to have the smallest $\pi_{01_{bin}}$, with the estimate suggesting that ~85% of the genotyped genic SNPs have no effect on breast cancer risk ($\pi_{01_{bin}}$ = 0.854, 95% CI: [0.828, 0.879]). The annotation group estimated to have the smallest $\pi_{01_{bin}}$ was also the gene variant group for the two GWAS, but for the

COGS, the 95% CI for the estimate was much narrower. All in all, the 95% CIs for all COGS SNP annotation group $\pi_{01_{bin}}$ estimates were narrow, which indicated that the estimates produced were fairly precise.

Annotation group	No. of SNPs	$\pi_{01_{bin}}$ (95% CI)
Intergenic variant	16,933	0.879 (0.846, 0.906)
Regulatory variant	1,969	0.906 (0.832, 0.963)
Gene variant	24,606	0.854 (0.828, 0.879)
Total	43,508	

Table 3-16: Proportion of COGS null SNPs ($\pi_{01_{bin}}$) within each annotation group

Overall, a limited interpretation could be made between the COGS and GWAS annotation group $h^2_{l\,bin}$ and $\pi_{01_{bin}}$ estimates because the 95% CIs for most of the GWAS estimates were very wide. Based on the COGS results, the results suggested that overall gene variants explain a higher proportion of the h^2_l for the COGS. Also, per marker, SNPs mapping to a gene regions were estimated to explain more genetic variation than regulatory and intergenic variants, but only a very tiny bit more.

3.3 Partitioning the COGS by cancer type

As previously mentioned, the iCOGS array is a custom array where, based on previous study results, SNPs were genotyped for their association with either breast, prostate or ovarian cancer. In order to explore how much of the estimated h_l^2 can be explained by the SNPs genotyped for their association with breast cancer, the h_l^2 was partitioned by cancer type. SNPs retained after QC and LD-clumping were split into two groups; breast cancer SNPs and prostate/ovarian cancer SNPs. SNPs genotyped for their association with breast cancer were grouped together, and the SNPs not associated with breast cancer were grouped together (prostate/ovarian cancer SNPs).

For each partitioned group, internal COGS training and replication samples were used to construct an interval PRS for the replication sample subjects. The SNP effects for each SNP, within a *p*-value interval, within the partitioned group, were used to construct an interval PRS. The SNP effects were estimated using a logistic regression model, with nine principal components and study included as covariates in the model. The association between each polygenic score and breast cancer outcome, in the replication sample, was tested in order to produce multiple *z*-scores **(Appendix 7: Table 1)**. These *z*-scores, whilst assuming a prevalence of 0.001, were then used to estimate the ${h_l^2}_{bin}$ and $\pi_{01_{bin}}$ for the two groups.

Summing the ${h_l^2}_{bin}$ estimates for each group, produced a summed estimate close to the h_l^2 estimate produced for the COGS (Table 3-17). There was found to be little difference between the ${h_l^2}_{bin}$ estimate based on breast cancer SNPs, and the ${h_l^2}_{bin}$estimate based prostate/ovarian SNPs (breast cancer ${h_l^2}_{bin}$= 0.031 vs. prostate/ovarian cancer ${h_l^2}_{bin}$= 0.029). However, with there being approximately 60% more prostate/ovarian SNPs than breast cancer SNPs, breast cancer SNPs, per-SNP, explained a larger proportion of the COGS based h_l^2 estimate for breast cancer. The estimated $\pi_{01_{bin}}$ for each group was approximately 70%, with the 95% CIs for the two $\pi_{01_{bin}}$ estimates being rather narrow (breast cancer $\pi_{01_{bin}}$ = 0.698, 95% CI: [0.633, 0.750] and prostate/ovarian: $\pi_{01_{bin}}$ = 0.682, 95% CI: [0.565, 0.763]). Both the $h_{l\ bin}^{2}$ and $\pi_{01_{bin}}$ estimates were very similar across the two groups, which suggested that the SNPs genotyped for their relationship with either prostate or ovarian cancer risk, also explain a reasonable proportion of the genetic variation in breast cancer risk.

	No. SNPs	$h_{l\ bin}^{2}$ (95% CI)	$\pi_{01_{bin}}$ (95% CI)
Breast cancer SNPs	16,761	0.031 (0.028, 0.033)	0.698 (0.633, 0.750)
Prostate/ovarian cancer SNPs	27,420	0.029 (0.026,0.032)	0.682 (0.565, 0.763)
Total	44,181	0.060	

Table 3-17: Partitioning the COGS SNPs by cancer type

3.4 Discussion

In this chapter, the genetic variance explained by genotyped SNPs was partitioned by chromosome, MAF and SNP annotation, to examine how the genetic variation for breast cancer is spread across the genome. Understanding how genetic variation is spread across the genome, by computing the heritability contribution of various SNP subsets, could narrow down the search for causal variants and enable the development of drug targets.

Partitioning the two GWAS by MAF showed that common genetic variants, with a MAF greater than 0.1, explained a large proportion of the genetic variation for breast cancer (> 88%), which was found to be consistent with other MAF partitioning studies (98). If rarer SNPs are not in high LD with more common SNPs, then it is possible for rarer SNPs to be underrepresented on a GWAS array. Therefore, much more of the genetic variation in breast cancer could be explained by rarer SNPs, with an MAF < 0.1, than estimated in this chapter. Not including many of the rarer SNPs would also lead to an underestimation of the genetic variation in breast cancer that can be explained by genetic variants. But based on the chip heritability estimates produced in chapter 2, it can be said that common SNPs (MAF > 0.1) genotyped on the array, explain a larger proportion of the genetic variation in breast cancer risk that is explained by genotyped SNPs. The same result was also shown for COGS, with common SNPs explaining over 80% of the estimated liability scale chip heritability.

For the BBCS GWAS and the COGS, a significant linear relationship between the estimated genetic variance explained by a chromosome and chromosome length, or the number of SNPs on a chromosome, was observed. This observation was consistent with other studies carried out on other complex diseases, however the relationship for the BBCS GWAS was not as strong as observed for other complex diseases, such as endometriosis and MS, where an R^2 of 0.37 and 0.31 had been observed (98). A weak linear relationship was observed for the UK2 GWAS, but the association was found to be non-significant. With larger sample sizes, and therefore more precise estimations, the relationship between the estimated variance explained by a chromosome and chromosome length, may become statistically significant. With the whole genome not necessarily represented by the SNPs genotyped on the iCOGS array, it was thought that it might be better to assess the relationship between the number of SNPs genotyped for each chromosome and chromosome contribution. The linear relationship was found to be even stronger for the COGS when assessing the relationship between genetic variation contribution for a chromosome, and the number of SNPs genotyped for each chromosome. For some of the chromosomes, the genetic variation explained by SNPs mapping to the chromosomes was estimated to be zero. This was an unrealistic estimate as at least one published genome-wide significant breast cancer SNP has been shown to map to each chromosome. It was also found that the 95% CIs for many of the estimates produced when partitioning by chromosome, including the COGS estimates, were fairly wide which meant that the estimates were not entirely accurate. This inaccuracy could have been due to the reduced number of SNPs used in the analysis when partitioning.

As well as partitioning the SNPs by their MAF and chromosome position, SNPs were also grouped by SNP annotation in order to investigate whether SNPs residing in intergenic regions could explain a larger proportion of the genetic variation in breast cancer, compared to those mapping to gene regions. The reasoning behind this analysis was that to date a large number of identified breast cancer susceptibility loci have been found to map to intergenic regions of the genome. It had not yet be explored whether variants mapping to intergenic regions explain a larger proportion of the genetic variation in breast cancer, compared to other variants. Unfortunately, due to study sample sizes, a conclusion could not be drawn from the GWAS partitioning analyses as the 95% CIs for the SNP annotation subset chip heritability estimates were very wide. However, the 95% CIs for the COGS SNP annotation subset chip heritability estimates were fairly narrow, with the results suggesting that overall gene variants explain a higher proportion of the liability scale chip heritability. With there being a larger number of gene variants than intergenic and regulatory variants, the genetic variation explained per-marker was assessed and compared across the three subsets. Per-marker, there was shown to be little difference in the estimated per-SNP chip heritability estimates, so there was little evidence to suggest that variants mapping to intergenic regions, were more enriched than those mapping to genic or regulatory regions

For the COGS, there was found to be little difference between the proportion of liability scale chip heritability explained by breast cancer SNPs, and the proportion explained by prostate/ovarian cancer SNPs. However, with there being a smaller number of breast cancer SNPs than prostate/ovarian cancer SNPs, the results indicated that the breast cancer SNPs could be more enriched than the prostate/ovarian cancer SNPs. With COGS SNPs estimated to explain ~6% of the variation in liability to breast cancer, and breast cancer SNPs genotyped on the custom array explaining ~50% of this variation, it was observed that the remaining ~50% could explained by prostate/ovarian cancer associated SNPs. This result confirms the importance of not restricting breast cancer analyses to only the SNPs thought to be related to breast cancer, as “non-breast cancer” SNPs were shown to explain a similar proportion of the genetic variation in liability to breast cancer.

The analysis presented in this chapter has enabled me to examine how the genetic variation in breast cancer liability is spread across the genome for two breast cancer GWAS and the COGS. The results from the GWAS partitioning analyses provided further evidence that breast cancer is a polygenic disease, with there being some evidence that the genetic variation for the disease can be explained by SNPs spread across the genome. It also showed that a substantial proportion of the genetic variation in breast cancer liability could be explained by common SNPs (MAF > 0.1). Wide 95%

CIs meant that it hard to draw reasonable conclusions for many of the results produced, particularly for the proportion of null SNPs estimates and the SNP annotation results.

It was expected that there might be some differences between the GWAS and the COGS, in how genetic variation is spread across the genome, because of the SNPs genotyped. Genotyped GWAS SNPs are spread across most of the genome, whereas for COGS SNPs, some parts of the genome might be underrepresented. But the results from the COGS partitioning analyses suggested that genetic variation explained by the iCOGS SNPs was spread across the genome. The results for the COGS partitioning analyses also indicated that common SNPs (MAF > 0.1) on the array explained a large proportion of the genetic variation in breast cancer liability, and that per-SNP breast cancer related SNPs explained more genetic variation than “non-breast cancer” SNPs. The COGS SNP annotation partitioning analyses suggested that intergenic SNPs, per-SNP, did not explain a larger proportion of the genetic variation in breast cancer liability, compared to genic and regulatory SNPs.

At the time of writing, AVENGEME had not yet been used to partition chip heritability for any complex disease, and this was the first analysis to partition the genetic variation for breast cancer by either MAF, chromosome or SNP annotation. Due to large 95% CIs for many of the GWAS results, especially those produced for the BBCS GWAS, it was hard to draw a reasonable conclusion as the precision of the estimates had to be questioned. The 95% CIs for the $h^2_{l\,bin}$ and $\pi_{01_{bin}}$ estimates produced in this chapter varied from fairly narrow to extremely wide (95% CI: [0.000 to 1.000]). Both the number of SNPs and individuals used to conduct the polygenic score analyses would have had an effect on the precision of the AVENGEME estimates. With partitioning analyses, SNPs are binned accordingly and then the proportion of genetic variance explained by each bin is estimated. Binning SNPs and then constructing multiple PRS by thresholding the SNPs in each bin, in order to used AVENGEME to produce bin estimates, would have reduced the number of SNPs used in each PRS. This, along with the differences in sample size across the studies, would have affected the precision of the estimates and would explain why some 95% CIs are wider than others. With much larger studies being conducted and released in the near future, it would be beneficial to replicate and improve the precision of the estimates, and the partitioning results produced in this chapter, in a much larger breast cancer sample.

Chapter 4 Examining whether a shared polygenic basis between breast cancer and BMI exists

4.1 Introduction

With breast cancer being a complex disease, a disease that is influenced by both environmental and genetic risk factors, it has been difficult to establish all risk factors associated with the disease. A number of risk factors associated with breast cancer risk have, however, been identified. Some risk factors are modifiable, meaning that they can either be controlled or changed, with BMI being an example of such a factor. Modifiable risk factors are favoured in public health as it may be possible to reduce disease risk through lifestyle changes. For example, a postmenopausal woman with a high BMI could reduce her risk of breast cancer by lowering her BMI through exercise and diet. Risk factors can also be reproductive factors, which are usually considered non-modifiable, with such examples including age at menarche, parity and age at menopause (109). These factors are much harder to control, that is if it is even possible to control them. Breast cancer risk factors include a mixture of modifiable and reproductive factors, which include BMI, age at menarche, age at menopause, age at first pregnancy, age, use of oral contraceptive, family history of disease and use of hormone replacement therapy (110). Some of these risk factors have been shown to have a positive effect on breast cancer risk, whilst others have been shown to have a negative effect on disease risk. BMI has been shown to be associated with breast cancer risk, but the effect BMI has on breast cancer risk depends on menopausal status (111-116). Premenopausal women with a BMI > 22, have a reduced risk of developing breast cancer, whereas postmenopausal women are at an increased risk (116). Another breast cancer risk factor is age at menarche, this being the age at which a woman has her first menstrual cycle. The younger a woman's age at menarche, the higher her risk is of developing breast cancer (110, 117). For every one year decrease in age, a woman's breast cancer risk increases by 5% (117). The age at which a woman begins menopause also affects breast cancer risk, with women who have a late menopause, over the age of 55, being twice as likely to develop breast cancer than women who start menopause before the age of 45 (110). For every year increase in age, a woman's breast cancer risk increases by approximately 3% (117). The age at which a woman first gives birth is another breast cancer risk factor, with there being an increase in the lifetime risk of breast cancer if a woman does not give birth, or if her first birth is at a later age (110). There is a small increase in the relative risk of developing breast cancer for women who take the oral contraceptive pill (110). For women who take the oral contraceptive pill, or for up to 10 years after stopping, those who begin taking the contraceptive pill before the age of 20, have a higher risk of developing breast cancer than women who begin at an older age (110). Also, women that use hormone replacement therapy, have an increased risk of developing breast cancer (110). For every year of use, a woman's relative risk of breast cancer increases by a factor of 1.023 if she uses hormone replacement therapy, or for up to 4 years after she has stopped using it (110). Percent mammographic breast density (PMD) is another breast cancer risk factor, with women who have dense breast tissue in over 75% of their breast, being up to 5 times more likely to develop breast cancer, compared to women with a PMD < 5% (78, 118).

For a variety of complex polygenic diseases, studies have examined whether associated risk factors have a shared polygenic pathway with the disease of interest, which if shown, would suggest that there is an overlap in the genetic architecture between them. In this chapter, polygenic scoring has been used to examine whether there is evidence that BMI, a measure used to deem whether an individual is a healthy weight, and breast cancer have a shared polygenic basis. BMI is a measure that is based on an individual’s weight and height measurement, with the formula being weight in kilograms divided by height in metres squared. An adult is classified as underweight if they have a BMI < 18.5, within normal range if 18.5 ≤ BMI < 25, overweight if BMI ≥ 25 and obese if their BMI ≥ 30 (119). In 2015, the adult prevalence for obesity in England, Scotland, Wales and Northern Ireland was published as being 27%, 28.8%, 23.5% and 25%, respectively (120, 121). The prevalence of obesity has increased within each country in recent years, and is predicted to continue to increase in the future (122). This is worrying as BMI has not only been shown to be associated with breast cancer risk, but also with many other chronic diseases, such as type-2 diabetes (123) and cardiovascular disease (124). BMI is known to be heritable, with up to 97 BMI susceptibility risk loci having been identified at the time of conducting the analysis in this chapter (125). With both BMI and breast cancer having a polygenic basis, it examined whether there was evidence of a shared polygenic basis. If evidence of an overlap is found, it would suggest that the association between BMI and breast cancer can partly be explained by genetics. However, unlike Mendelian randomization, establishing genetic overlap between traits does not establish causality.

4.2 Previous literature on shared genetic basis

Polygenic scores, constructed using two trait populations, can be used to examine whether there is evidence that two traits have a shared polygenic basis (60). Multiple studies have examined whether a shared polygenic pathway exists between two or more traits using polygenic scores. An example of such a study, includes a study conducted by The International Schizophrenia Consortium (ISC) (61). Using European individuals, the ISC tested for evidence of a shared genetic basis existing between schizophrenia and bipolar disorder using polygenic score analysis. The training set comprised of 3,322 schizophrenia cases and 3,587 controls, and two independent bipolar disorder replication samples were used, the WTCCC (1,829 cases and 2,935 controls) and STEP-BD (Systematic Treatment Enhancement Program for Bipolar Disorder) (955 cases and 1,498 controls). Polygenic scores were constructed for the replication sample individuals for five p-value significance thresholds; $p < 0.1$, $p < 0.2$, $p < 0.3$, $p < 0.4$, $p < 0.5$, these being based on the individual significance of the schizophrenia SNPs. The association between these polygenic scores and bipolar disorder were then tested. The ISC found evidence that polygenic scores, constructed using an ensemble of schizophrenia SNPs that had not all reached genome-wide significance, were associated with bipolar disorder risk. The polygenic score based on schizophrenia SNPs with a *p*-value < 0.5, was associated with bipolar disorder in both replication samples (WTCCC: *p*-value = 1×10^{-12} and STEP-BD: *p*-value = 7×10^{-9}). With this finding, the ISC concluded that there was evidence of a shared genetic basis existing between schizophrenia and bipolar disorder. Additionally, they used polygenic scores to test whether there was evidence that schizophrenia had a shared polygenic basis with various non-psychiatric diseases (coronary artery disease, Crohn’s disease, hypertension, rheumatoid arthritis, type I diabetes and type II diabetes) (61). The association between the schizophrenia based polygenic score and each of the non-psychiatric traits was non-significant, for all *p*-value thresholds (*p*-value > 0.05).

Therefore, there was no evidence to suggest that a shared polygenic basis exists between schizophrenia, and any of the non-psychiatric diseases tested.

The Cross-Disorder Group of the Psychiatric Genomics Consortium (CDG-PGC) have examined whether there is evidence of a shared polygenic pathway between multiple traits (126). Using individuals of European ancestry, the CDG-PGC tested whether there was evidence of a shared polygenic basis existing between five psychiatric disorders: autism spectrum disorder (4,788 trio cases, 4788 trio pseudo controls, 161 cases and 526 controls), attention deficit-hyperactivity disorder (1,947 trio cases, 1947 trio pseudo controls, 840 cases and 688 controls), bipolar disorder (6,990 cases and 4,820 controls), major depressive disorder (9,227 cases and 7,383 controls) and schizophrenia (9,379 cases and 7,736 controls) using polygenic score analysis (126). Each disorder was used as the training sample, with the remaining disorders then used as the replication sample. Multiple polygenic scores were constructed for each replication sample individual, based on various training sample SNP *p*-value thresholds. A significant cross-disorder overlap was observed between bipolar disorder, major depressive disorder and schizophrenia. A significant polygenic overlap was not detected between either attention deficit-hyperactivity disorder or autism spectrum disorder, with other disorders.

Using UK Biobank data (N > 100,000) and large GWAS consortium data, both LDSC and polygenic score analysis were used by Hagenaars et al. (127) to find evidence of a shared polygenic basis existing between various cognitive functioning and educational traits (Biobank data), and 24 health related phenotypes (GWAS consortium summary data). The cognitive and educational traits included reaction time, verbal-numerical reasoning, memory and educational attainment. Health related traits included various vascular-metabolic diseases, neuropsychiatric disorders, brain measures, physical and physiological measures and life-course cognitive traits and proxies. For the polygenic score analysis, the GWAS summary data for the 24 heath related phenotypes were used to construct multiple polygenic scores, for five different *p*-value threshold: $p < 0.01$, $p < 0.05$, $p < 0.1$, $p < 0.5$ and all GWAS SNPs, for the individuals in the UK Biobank sample. The association between the multiple polygenic scores and the UK Biobank phenotypes; reaction time, memory, verbal-numerical reasoning and educational attainment, was tested. LDSC was used to measure the level of genetic overlap between the traits by estimating the genetic correlation between the traits. Hagenaars et al. observed significant genetic correlations and significant associations between the cognitive functioning and educational traits (UK Biobank), and many of the health related outcomes (GWAS consortium). From this the authors concluded that the results indicated that there was genetic overlap between cognitive functions, and physical and mental health diseases. A study conducted by Bulik-Sullivan et al. (128) also assessed the correlation between multiple traits using LDSC. Bulik-Sullivan et al. tested for significant correlations between 24 traits using GWAS summary data, with each trait having a sample of at least 10,000 individuals. The traits studied included BMI, ulcerative colitis, schizophrenia, bipolar disorder and years of education. After adjusting for multiple-testing, a number of significant correlations were detected, with these including significant correlations between ulcerative colitis and childhood obesity, anorexia nervosa and BMI and anorexia nervosa and schizophrenia.

Not all studies have succeeded in finding evidence of a shared genetic basis between traits. A study conducted by Goris et al.(129) failed to find evidence of a shared genetic basis existing between multiple sclerosis (MS) (4,088 cases and 7,144 controls) and amyotrophic lateral sclerosis (ALS) (3,762 cases), when using a polygenic score analysis. The authors constructed polygenic scores for different *p*-value thresholds, based on the GWAS SNPs for one trait, for subjects in the remaining independent sample. The association between the score and independent replication trait was then tested. The associations were observed to be non-significant, which meant that Goris et al. found no evidence for a shared polygenic basis existing between MS and ALS.

For breast cancer, a genetic overlap has been demonstrated. PMD, the dense area of the breast divided by the total breast area, is a known risk factor for breast cancer and evidence has been found to suggest that PMD and breast cancer have a shared genetic basis (78, 130, 131). Varghese et al. (78) found evidence of a shared genetic basis existing between PMD and breast cancer using a polygenic score analysis. Using a published meta-analysis of five mammographic breast density GWAS, Varghese et al. constructed ten different polygenic scores using 1%-10% of the PMD SNPs, these being the top 1%-10% (up to 50,899 SNPs) of PMD SNPs after ranking the SNPs by their association with PMD. Each polygenic score was then tested for its association with breast cancer outcome in 3,628 breast cancer cases and 5,190 controls, using a logistic regression model. There was shown to be a significant association between the scores constructed using the top 3%-10% of SNPs and breast cancer risk, which indicated that through a large number of common variants, PMD and breast cancer have a shared genetic basis. A couple of years later, a meta-analysis was conducted by Lindstrom et al. (130), which focussed on loci associated with either of the three mammographic density phenotypes; dense area, non-dense area or percent density. From the identified genome-wide significant loci for all three phenotypes, Lindstrom et al. found that four mammographic density associated loci had previously been associated breast cancer. To add to this, Lindstrom et al. discovered four novel loci associated with the mammographic density phenotype, which were already known to be associated with breast cancer risk. With this result, Lindstrom et al. concluded that the analysis further proved that there was a shared genetic basis between breast cancer and mammographic density. To date, however, there have been no studies of genetic overlap between other known risk factors and breast cancer, using a large number of SNPs.

4.3 BMI data

BMI summary data, together with the BBCS GWAS, UK2 GWAS and the COGS, were used to assess whether there was evidence of a polygenic overlap in the genetic architecture between BMI and breast cancer. The BMI summary data used to conduct the analysis in this chapter, was collected as part of a BMI meta-analysis conducted by the Genome-wide Investigation of ANThropometric measures (GIANT) consortium, and is available in the public domain (125). The summary data was based on up to 339,224 subjects, extracted from 125 studies: 82 BMI GWAS and 43 Metabochip studies. The Metabochip is a custom Illumina iSelect genotyping array with approximately 200,000 genetic variants of interest genotyped on it (132). The genetic variants genotyped were those that had been identified as being related with either metabolic, cardiovascular and/or anthropometric traits (132). For each of the genetic variants, beta effects and *p*-values, for their association with BMI, were given.

4.4 Proposed method

LDSC and polygenic score analysis were used in this chapter to test whether there was evidence of a shared polygenic basis existing between breast cancer and BMI. For the polygenic score analysis, multiple polygenic score analyses were conducted, first using published genome-wide significant SNPs and then a large number of SNPs en-masse.

4.4.1 Genetic correlation between breast cancer and BMI

The genetic correlation between breast cancer and BMI was estimated using LDSC, via the web interface LD hub.

LDSC estimates the correlation between two traits using the following equation (128):

$$E[z_{1j}z_{2j}] = \frac{\sqrt{N_1N_2}\,\rho_g}{M}\,\ell_j + \frac{\rho N_s}{\sqrt{N_1N_2}}$$

Where, z_{1j} and z_{2j} are the *z*-scores for study 1 and study 2 for SNP *j* and the sample sizes of the two studies are denoted as N_1 and N_2. The genetic correlation between the two traits is denoted as ρ_g, ℓ_j is the LD score and the total number of markers is given as M. The phenotypic correlation is denoted as ρ and this is between the number of overlapping samples (N_s).

The genetic covariance is estimated by regressing the *z*-scores for study 1 and study 2 (z_{1j} and z_{2j}) against $\ell_j\sqrt{N_{1j}N_{2j}}$ and then multiplying this by M, where N_{1j} and N_{2j} are the sample sizes for SNP *j* in study 1 and study 2, respectively.

Similar to when using LDSC to produce a chip heritability estimate, the subjects and SNPs retained in the studies after QC were used to estimate the genetic correlation between breast cancer and BMI.

4.4.2 Shared polygenic basis (genome-wide significant SNPs)

Polygenic scores were constructed for the subjects in the breast cancer studies, using the SNP effects of up to 97 published BMI genome-wide significant loci (125), and the "--score" command in PLINK. Four separate analyses were performed, one with each breast cancer study (BBCS, UK2 and COGS) as the replication sample, and the fourth combining the two breast cancer GWAS to form a larger GWAS replication set. A reduced number of the published BMI SNPs had been genotyped in the three breast cancer studies, so to increase the number of SNPs in the polygenic score analysis, the BMI SNPs were extracted from the imputed GWAS SNPs (PLINK best guess) for the GWAS subjects and used in the replication set. However, this was only possible for the two GWAS, as I only had access to the genotyped COGS SNPs, so not as many SNPs were included in the BMI-COGS PRS analysis. The association between the polygenic scores and breast cancer outcome, for the breast cancer replication sample, was then tested using a logistic regression model.

A PRS was also constructed based on up to 94 published genome-wide significant breast cancer SNPs, these being SNPs discovered by Michailidou et al.(51), or in studies published prior to this study. The SNP effects were estimated using either the combined GWAS, in order to increase the sample size to improve the accuracy of the SNP effect estimates, or the COGS. Therefore, two analyses were conducted, with BMI being the replication sample trait. Having BMI as the replication trait, meant that the "--score" command in PLINK could not be used, as individual genotype data would be needed. With there being no individual genotype information for the replication trait, only summary data, a different way of constructing the score was required. To overcome this problem "grs.summary", an R function from the "gtx" R package (133), was used to construct a polygenic score and test for its association with BMI. This R function uses training sample SNP effects, the aligned SNP effects in the replication sample, and the standard errors for the SNP effects in the replication sample to do this.

The “grs.summary” R function uses the PRS formula, given in chapter 1, to test the association between a PRS and a replication trait, which in this case is BMI. To test the association between score and BMI, a linear regression model can be used, such that:

$$y_j = y_0 + \alpha PRS_j + \varepsilon_j$$

Where y_j is the replication trait BMI for individual j, y_0 is the constant, PRS_j is the polygenic score for individual j, α is the PRS regression coefficient, and ε_j is the error term.

Dastani et al. (134) state that the PRS regression coefficient (α) can be estimated by using the following equation:

$$\hat{\alpha} \cong \frac{\sum_{i=1}^{m} \hat{\beta}_i \widehat{w}_i \hat{s}_i^{-2}}{\sum_{i=1}^{m} \hat{\beta}_i^{2} \hat{s}_i^{-2}}$$

With,

$$se(\hat{\alpha}) \cong \sqrt{\frac{1}{\sum_{i=1}^{m} \hat{\beta}_i^{2} \hat{s}_i^{-2}}}$$

where, m is the total number of SNPs used to construct the polygenic score, $\hat{\beta}_i$ refers to the i-th SNP effect estimated for the training sample trait using the training sample, $\widehat{w}_i$ is the i-th SNP effect for the replication trait estimated from the replication sample (summary data) and $\hat{s}_i$ is the corresponding replication sample standard error estimate for the i-th SNP. It is assumed that independence across all m SNPs used in the PRS holds.

The nested chi-squared test statistic (χ_1^2) for the association between BMI and the PRS is then estimated as (135):

$$\chi_1^2 \cong \left(\frac{\hat{\alpha}}{se(\hat{\alpha})}\right)^2$$

For each analysis, regardless of the method used, the same effect allele for each SNP was used across the training and replication samples.

4.4.3 Shared polygenic basis (en-masse and p-value thresholding)

For the en-masse analysis, polygenic scores were constructed using the SNP effect of SNPs genotyped for each of the breast cancer studies. Polygenic scores were not constructed using the BMI SNP effects, as LD-removal could not be performed on the summary data at the time the analysis was conducted. Without LD-removal, independence across SNPs could not be assumed. Four separate analyses were conducted, one with each breast cancer study (BBCS, UK2 and COGS) as the training sample, and the fourth analysis combining the two breast cancer GWAS, in order to improve the accuracy of the SNP effect estimates. The replication sample trait in these analyses was BMI, which meant that that the PLINK "--score" command could not be used. So for these analyses, the R function "grs.summary" was used. For the analyses conducted using GWAS, BMI consortium dataset SNPs were extracted from the imputed GWAS SNPs (PLINK best guess), and used in the training set. This was carried out, in order to increase the number of SNPs present in the polygenic score analysis Before estimating the SNP effects for the training sample SNPs, both QC and LD clumping ($r^2 < 0.1$) were performed on each training sample. The SNP effects were then estimated for each training study using a logistic regression model, with the relevant number of principal components for each study included as covariates in the model, in order to adjust for population stratification. The training sample SNP effects were grouped according to their strength of association with breast cancer outcome, using the following *p*-value thresholds: $p \le 1$, $p \le 0.7$, $p \le 0.4$, $p \le 0.1$, $p \le 0.05$, $p \le 0.01$ and $p \le 0.001$. For each group of SNPs, "grs.summary" was used, along with the corresponding BMI beta coefficients and standard errors for the same SNPs, extracted from the BMI consortium summary data, to produce an association *p*-value. For each analysis, the same effect allele for each SNP was used across the training and replication samples.

4.4.3.1 Variance explained by the polygenic score (en-masse and p-value thresholding)

When using the “grs.summary” function, the Nagelkerke’s pseudo R^2 variance was also estimated and produced for each of the scores, for each *p*-value threshold. The measure can be used to quantify the proportion of variance that can be explained in the replication sample, by the training sample derived polygenic score. In the “grs.summary” function, the chi-square statistic produced for the association between replication trait and the PRS, is divided by the number of subjects in the replication study, to produce a pseudo R^2 estimate:

$$pseudo\ R^2 = \frac{\chi_1^2}{N}$$

Where, *N* is the number of individuals in the replication sample.

4.5 Shared polygenic basis analysis

4.5.1 Genetic correlation between breast cancer and BMI

Using all three breast cancer studies separately, LDSC, implemented using LD hub, was used to estimate the genetic correlation between breast cancer and BMI. The results suggested that the strength of the correlation between breast cancer and BMI were not different from zero for each of the breast cancer studies (correlation *p*-values > 0.05) (Table 4-1). Considering the estimated correlation strength between breast cancer and BMI, the standard errors were also observed to be fairly large. With the BBCS GWAS having less than 5,000 individuals, and the number of SNPs in the analysis being less than 200,000 for the COGS, the correlation estimates were expected to be fairly noisy (95, 128).

Overall, the results from the correlation analyses suggested that breast cancer and BMI were not significantly correlated (*p*-value > 0.05), thus there was no evidence to suggest that breast cancer and BMI have a shared polygenic basis.

		BMI	
Breast cancer study	Correlation	Standard error	*p*-value
UK2 GWAS	-0.0725	0.1073	0.4992
BBCS GWAS	0.0625	0.1298	0.6302
COGS	-0.0412	0.0522	0.4301

Table 4-1: Genetic correlation between breast cancer and BMI

4.5.2 Shared polygenic basis (genome-wide significant SNPs)

4.5.2.1 Published genome-wide BMI SNPs

In the following analyses PLINKs “--score" command and a logistic regression model, with the relevant PCs included as covariates in the model to adjust for population stratification, were used to examine whether there was evidence to suggest that a shared polygenic basis exists between BMI and breast cancer.

Polygenic scores were constructed for the breast cancer study subjects using the SNP effects extracted from the GIANT consortium summary BMI data, for published genome-wide significant BMI SNP (97 SNPS). Not all published SNPs were represented in the polygenic score, as a number of the BMI genome-wide significant SNPs had not be either genotyped or imputed for the breast cancer studies. The association between various polygenic scores and breast cancer outcome, in the breast cancer studies, was then tested. A non-significant association between the BMI derived polygenic score and breast cancer outcome was observed for both breast cancer GWAS (UK2 *p*-value = 0.951 and BBCS *p*-value = 0.799) (Table 4-2). Therefore, genetic overlap between BMI and breast cancer was not observed, when using known genome-wide significant BMI SNPs.

Training sample	Replication sample	No. SNPs	*p*-value
BMI	UK2	35	0.951
BMI	BBCS	24	0.799

Table 4-2: Shared genetic basis - BMI and breast cancer GWAS (logistic regression model)

The replication sample used in the analysis was increased, by combining the UK2 GWAS and the BBCS GWAS, to see whether this improved the significance of the association. Increasing the sample size of the replication set did not improve the significance of the association, there was still found to be a non-significant association between the BMI derived score and breast cancer outcome in the combined breast cancer GWAS sample (*p*-value = 0.972) (Table 4-3).

Training sample	Replication sample	No. SNPs	*p*-value
BMI	Combined GWAS	35	0.972

Table 4-3: Shared genetic basis - BMI and breast cancer GWAS combined (logistic regression model)

The association between the BMI derived polygenic score and breast cancer risk was then tested, with the COGS as the replication sample. A polygenic score, constructed for the COGS subjects using BMI SNPs, was not significantly associated with breast cancer risk in the COGS (*p*-value = 0.806) (Table 4-4). The number of SNPs used to construct the polygenic score for the COGS subjects, was considerably less than the number used to construct the scores for the GWAS individuals (Table 4-2and Table 4-3). In this section, no evidence was found to suggest that BMI and breast cancer have a shared polygenic basis. With only a small number of SNPs included in the polygenic scores, the genetic signal across the two traits could be underrepresented and restricted. There may be SNPs that affect both traits, but the small number of BMI SNPs analysed might have no effect on breast cancer risk.

Training sample	Replication sample	No. SNPs	*p*-value
BMI	COGS	10	0.806

Table 4-4: Shared genetic basis - BMI and COGS (logistic regression model)

4.5.2.2 Published genome-wide breast cancer SNPs

Next, breast cancer derived polygenic scores were constructed using the SNP effects for published genome-wide significant breast cancer SNPs, estimated using the subjects in the combined UK2/BBCS GWAS and the COGS. The two GWAS were combined for this analysis to improve the accuracy of the SNP effect estimates used in the polygenic score. Again, not all published breast cancer SNPs were used in the analysis, as the SNPs were either not imputed for either breast cancer GWAS, or not present in the BMI consortium summary data. For the following analyses in this section, the "grs.summary" R function was used to test the association between the breast cancer derived polygenic scores and BMI. The breast cancer SNP effects were estimated using a logistic regression model, with the relevant number of PCs for each study included as covariates in the model, in order to correct for any population stratification present in the data.

There was found to be a weak significant association between the combined GWAS breast cancer derived polygenic score and BMI (p-value = 0.035) (Table 4-5), but the association was not as strong as previously observed between other traits.

	Published breast cancer SNPs	
Training sample	No. SNPs	*p*-value
Combined GWAS	68	0.035

Table 4-5: Published breast cancer SNPs (GWAS) and BMI (grs.summary method)

Increasing both the training sample size and the number of SNPs used in the polygenic score, by using the COGS as the training sample, improves the strength of the association between the breast cancer derived polygenic score and BMI (*p*-value = 0.009) (Table 4-6). Increasing the sample size of the training sample should have improved the accuracy of the SNP effect estimates using in the score, compared to those produced using the combined breast cancer GWAS, as a larger number of individuals had been used to produce the estimates.

	Published breast cancer SNPs	
Training sample	No. SNPs	*p*-value
COGS	71	0.009

Table 4-6: Published breast cancer SNPs (COGS) and BMI (grs.summary method)

With significant associations observed between two different polygenic scores, based on genome-wide significant breast cancer SNPs, and BMI, the results would indicate that there could be shared genetic overlap between breast cancer and BMI. In the next section, the analysis has been taken further by investigating whether a significant association can be detected when using all genotyped SNPs, en-masse.

4.5.3 Shared polygenic basis (en-masse and p-value thresholding)

Using a much larger number of SNPs, than used in the previous analyses, it was tested whether evidence of a shared polygenic basis between the two traits can also be found when not restricting the score to genome-wide significant SNPs only. The association between multiple PRS, derived using SNPs within different *p*-value thresholds, and BMI was tested using the “grs.summary” R function.

There was found to be a non-significant association between the UK2 derived breast cancer polygenic score, based on all SNPs (p-value threshold $p \leq 1$, SNPs = 100,109)**,** and BMI (*p*-value = 0.560) (Table 4-7). For each of the UK2 derived polygenic scores, the association between the score and BMI was non-significant (*p*-value > 0.05), this was even observed for the score derived using only SNPs with a $p \leq 0.001$. There was also shown to be non-significant association between the BBCS derived breast cancer polygenic score, based on all SNPs (p-value threshold $p \leq 1$, SNPs = 94,666)**,** and BMI (*p*-value = 0.055) (Table 4-8). The association was, however, borderline non-significant (*p*-value = 0.055), but the strength of the association did not improve further when being more stringent and restricting the SNPs used in the score. Therefore, the results in Table 4-7 and Table 4-8 did not provide evidence of a shared polygenic basis existing between breast cancer and BMI.

p-value threshold	No. SNPs	*p*-value
p ≤ 1	100,109	0.560
p ≤ 0.7	88,215	0.567
p ≤ 0.4	66,794	0.575
p ≤ 0.1	27,326	0.610
p ≤ 0.05	16,854	0.299
p ≤ 0.01	5,108	0.344
p ≤ 0.001	853	0.366

Table 4-7: Performing grs.summary for different *p*-value intervals (UK2)

p-value threshold	No. SNPs	*p*-value
p ≤ 1	94,666	0.055
p ≤ 0.7	83,458	0.062
p ≤ 0.4	62,092	0.217
p ≤ 0.1	23,788	0.456
p ≤ 0.05	13,962	0.822
p ≤ 0.01	3,761	0.450
p ≤ 0.001	519	0.858

Table 4-8: Performing grs.summary for different *p*-value intervals (BBCS)

To test whether a significant association could be achieved by increasing the sample size of the training set, the two breast cancer GWAS were combined. The association between the combined GWAS breast cancer score and BMI was nonetheless still non-significant, when not restricting the SNPs used in the score (p-value threshold *p* ≤ 1, SNPs = 96,667) (*p*-value = 0.073) (Table 4-9). A significant association was observed between the combined breast cancer GWAS derived score and BMI for four of the *p*-value thresholds (*p* ≤ 0.4, *p* ≤ 0.1, *p* ≤ 0.05 and *p* ≤ 0.01), but not the *p* ≤ 0.001 threshold (*p*-value = 0.943, SNPs = 873). Like the results produced Table 4-5 and Table 4-6, the strength of the associations were not as strong as seen in other studies, but it was evidence that many common genetic variants of small effect could contribute to BMI.

p-value threshold	No. SNPs	*p*-value
p ≤ 1	96,667	0.073
p ≤ 0.7	85,543	0.065
p ≤ 0.4	64,497	0.043
p ≤ 0.1	26,009	0.029
p ≤ 0.05	15,927	0.023
p ≤ 0.01	4,776	0.028
p ≤ 0.001	873	0.943

Table 4-9: Performing grs.summary for different p-value intervals (combined GWAS)

To increase the sample size of the training set further, the SNP effects in the polygenic score were based on the SNPs genotyped for the COGS, but with this came a decrease in the number of SNPs used to derive the polygenic scores. The number of SNPs in union between those genotyped on the iCOGS array and those genotyped for the GIANT consortium was 41,386, which was approximately half the number of SNPs in union between the breast cancer GWAS and the GIANT consortium SNPs. This could therefore have a negative effect on the results, as many SNPs across the genome might not be represented in the analyses, leading to loss of genetic signal.

There was a non-significant association between the COGS derived breast cancer polygenic score, based on all SNPs (p-value threshold *p* ≤ 1, SNPs = 41,386)**,** and BMI (*p*-value = 0.715) (Table 4-10). For each of the COGS derived polygenic scores, the association between the score and BMI was non-significant (*p*-value > 0.05), even for the score derived using only SNPs with a *p* ≤ 0.001. These results therefore did not provide evidence of a shared polygenic basis existing between breast cancer and BMI.

p-value threshold	No. SNPs	*p*-value
$p \leq 1$	41,386	0.715
$p \leq 0.7$	34,396	0.759
$p \leq 0.4$	24,482	0.618
$p \leq 0.1$	9,560	0.199
$p \leq 0.05$	5,816	0.275
$p \leq 0.01$	1,954	0.081
$p \leq 0.001$	526	0.064

Table 4-10: Performing grs.summary for different p-value intervals (COGS)

Combined GWAS derived polygenic scores were the only scores shown to have a significant association with BMI, none of the other polygenic scores derived in this section were observed to have a significant association with BMI. By increasing the size of the breast cancer training sample further, and including as many genetic variants in the score as there are in Table 4-7, Table 4-8 and Table 4-9, a much lower *p*-value could potentially be observed.

As well as testing the association between multiple PRS and BMI, the variation in BMI explained by the breast cancer derived polygenic scores was also estimated (Pseudo R^2). The combined GWAS derived score explained the largest amount of variation in BMI (Figure 4-1) **(Appendix 8)**. It was estimated that up to 0.0017% of the variation in BMI can be explained by genotyped breast cancer SNPs. Considering the combined GWAS polygenic score had the greatest association with BMI, this was not a surprising result as the Pseudo R^2 is based on the chi-squared statistic. The BMI variance explained by the combined GWAS score was also not as large as seen with other traits, for example, a schizophrenia based polygenic score has been shown to explain over 0.4% of the variation in bipolar disorder (61). The limited variance explained in BMI was not surprising as the associations observed in other studies have typically been stronger than those observed in this analysis. As the Pseudo R^2 is based on the chi-squared statistic, one would expect the Pseudo R^2 to be larger in the other studies.

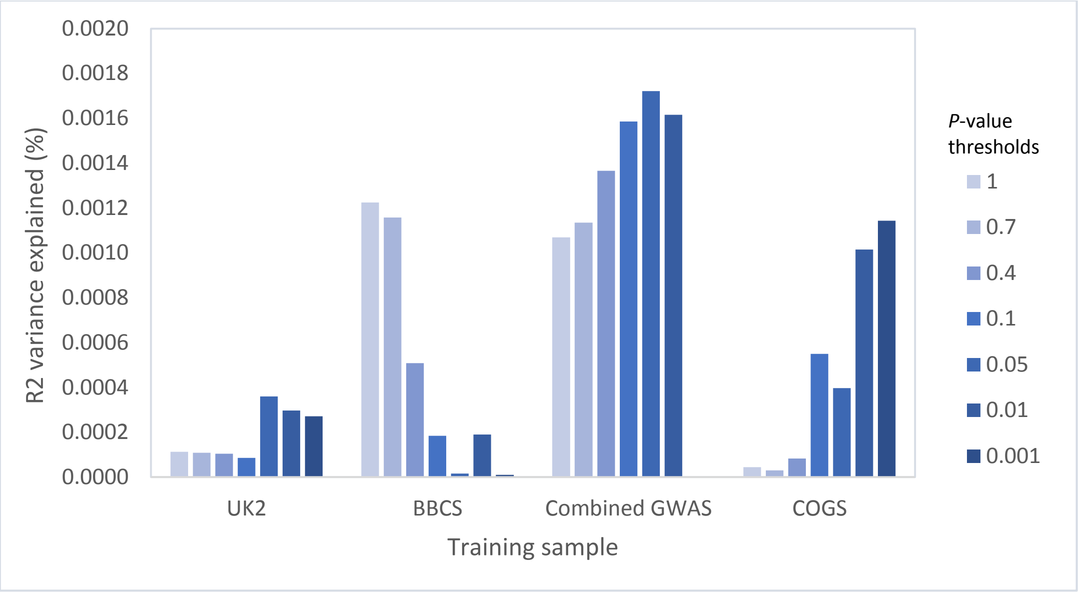


Figure 4-1: Polygenic score analysis - estimating the variance explained by the score in BMI

4.6 Discussion

We know that BMI is risk factor for breast cancer, and that both BMI and breast cancer have a polygenic basis, but we do not know whether there is a polygenic overlap between the two traits. Finding evidence of a genetic overlap between the two phenotypes could help improve risk prediction and the development of treatments, by enabling the two phenotypes to be studied together (60). Studies have tended to use either polygenic score analysis, LDSC or both methods, to test whether there is a shared polygenic basis between two traits. Bivariate GCTA (136) is another method that is used to test whether a shared genetic basis between two traits exists, but genotype data for both traits is needed (137). With summary data being used for one of the traits, bivariate GCTA could not be used.

Using both LDSC and polygenic score analysis, I have tested whether there was evidence to suggest that breast cancer and BMI have a shared polygenic basis. As shown in chapter 2, LDSC tends to work best on larger samples, and the breast cancer studies used were not as large as those used in other shared polygenic basis studies. As stated previously in this chapter, studies have been conducted by Bulik-Sullivan et al.(128) and Hagenaars et al.(127) in order to examine for significant correlations between multiple traits using LDSC, with the use of either GWAS summary data or UK Biobank data. Each trait in the Bulik-Sullivan et al. study had a sample size of at least 10,000 individuals, so the individual studies were much larger than the breast cancer GWAS used to conduct the analyses in this chapter. Both autism spectrum disorder and infant head circumference had a sample size of approximately 10,000 individuals, which was a similar size to the combined breast cancer GWAS sample size. Many of the significant associations detected by Bulik-Sullivan et al., tended to be between traits with larger sample sizes. Hagenaars et al. (127) also found significant genetic correlations when using up to 112,151 UK Biobank individuals and GWAS SNPs to perform their analyses, this again being a much larger sample than the GWAS sample sizes used to perform the analyses in this chapter. The COGS used in this chapter was just as large sample size wise, but the number of genetic variants was limited. From the LDSC correlation analysis conducted in this chapter, the results suggested that there was no evidence of a shared polygenic basis between breast cancer and BMI. The estimated correlation was very weak and non-significant for each of the breast cancer studies, therefore suggesting that the null hypothesis that the estimated correlation coefficient is equal to zero cannot be rejected. It was also found, as expected because of the number of SNPs and subjects analysed, that the correlation estimates produced were fairly noisy.

With sample sizes being too small to gain statistical power using LDSC, polygenic scores were also used. First the analysis focused on published genome-wide SNPs that have been individually identified for breast cancer and BMI. A polygenic score constructed using up to 35 genome-wide significant BMI SNPs, failed to be shown to be associated with breast cancer outcome in an independent sample. However, when using up to 71 genome-wide significant breast cancer SNPs to construct a polygenic score in an independent sample, the score was significantly associated with BMI (*p*-value BBCS/UK2= 0.035 and *p*-value COGS = 0.009). With the combined GWAS *p*-value being borderline significant, the significance of the association was not as significant as one would hope.

An en-masse polygenic score approach was then adopted, in which a polygenic score was constructed for multiple *p*-value thresholds, based on each SNPs association with breast cancer outcome in the training sample. These polygenic scores were then tested for their association with BMI in the independent sample (GIANT consortium summary data). The only breast cancer scores shown to be associated with BMI were the combined GWAS derived polygenic scores, based on SNPs with $p \leq 0.4$, $p \leq 0.1$, $p \leq 0.05$ and $p \leq 0.01$. The *p*-values were again, not as small as one would hope for, but nonetheless still significant. For the sample sizes and the number of SNPs used in these analyses, the statistical power to detect a genetic correlation was under 50%, with the power being considerably lower for the COGS (~6%). By increasing the number of breast cancer cases in future analyses, the power to detect an association between the score and independent outcome will improve.

The results from the analyses presented in this chapter indicate that when focussing on known BMI susceptibility genetic variants alone, there is no evidence of a shared polygenic basis between BMI and breast cancer. However, there is evidence to suggest that there is a shared polygenic basis between breast cancer and BMI when polygenic scores are based on published breast cancer susceptibility variants. By being less stringent on the breast cancer GWAS SNPs used to construct the polygenic score, significant associations were also observed, providing further evidence that breast cancer and BMI have a shared polygenic basis.

This was the first study to assess whether breast cancer and BMI could have a shared polygenic basis, which means that the analyses presented in this chapter were novel. A limitation was that BMI was the only risk factor analysed, no other breast cancer risk factors were examined. As stated previously in this chapter, there are many known breast cancer risk factors. Unfortunately at the time of conducting the analysis, I did not have access to data on other breast cancer risk factors. Another limiting factor is that BMI has a varying effect on breast cancer risk, depending on a woman’s menopausal status. As premenopausal women with a BMI greater than 22 have a decreased risk of breast cancer, compared to postmenopausal women with a BMI greater than 22, it would have been better to stratify by menopausal status and then maybe a more pronounced association may have been detected. The GWAS data used to conduct the analyses did not have this information available, so I was unable to stratify by menopausal status. Age could be used as a surrogate for menopausal status, but again, this information was not available for all individuals. An additional limitation is that for the shared polygenic basis analysis based on genome-wide significant BMI

SNPs, only a limited number of BMI SNPs were present in each breast cancer studies. This meant that the polygenic scores constructed in each of the analyses were based on a restricted number of SNPs, therefore the genetic signal across the two traits could be underrepresented and limited. The number of SNPs used in a score could have been increased using proxy SNPs. If a published BMI SNP was not found to have been genotyped in the breast cancer study, it may be that a SNP that is in high LD with the published SNP could have been genotyped instead. The proxy SNP could have be used to represent the absent published SNP in the score, thus increasing the number of SNPs used in the score. Another way of increasing the number of SNPs present in the score would have been by imputing the breast cancer studies for the published genome-wide significant BMI SNPs myself.

With the findings suggesting that breast cancer and BMI may have a shared polygenic basis, future breast cancer shared polygenic basis analyses should be conducted. Many of the shared polygenic basis studies conducted recently have analysed multiple traits, using large samples and GWAS SNPs. The same should happen in breast cancer, across breast cancer and many known breast cancer risk factors, using a much larger samples and GWAS SNPs. This could provide further insight into the genetic architecture of breast cancer, which could aid the future development of treatments and disease prevention strategies.

Chapter 5 Testing for evidence of PRS-environmental factor interactions and PRS-SNP interactions

5.1 Introduction

For most diseases, including breast cancer, both environmental and genetic risk factors have been shown to influence disease risk. The influence that a factor has, be it environmental or genetic, on disease risk may be modified by another factor through an interaction. In recent years, interaction studies have been performed in order to improve the identification of individuals, within specific populations, who may be at an increased risk of developing a given disease. With much of the heritability for many complex diseases, including breast cancer, being unexplained, it could be that the effect a genetic factor has on disease is either heightened or reduced with the exposure to specific environmental factors (138). In this chapter, an environmental factor will refer to any non-genetic disease risk factor.

In breast cancer, and other complex diseases, interaction studies have been used to find evidence for the existence of gene-environmental factor interactions. In recent years, with many diseases being shown to be highly polygenic, studies have begun to adopt a polygenic approach to test for gene-environmental interactions, believing that more than one genetic variant may be involved in an interaction (139). Peyrot et al.(140) have performed such a study. They examined whether there was evidence that a polygenic score for major depressive disorder, based on different *p*-value thresholds, could be modified by childhood trauma. Childhood trauma is a major risk factor for major depressive disorder, and in this study the risk factor was measured as a score that ranged from 0-8 based on four domains: emotional neglect, psychological abuse, physical abuse and sexual abuse. A total of 32,870 SNP effects, based on a meta-analysis conducted by the Psychiatric Genomics Consortium using 7,544 cases and 7,754 controls, were used in the analysis. For 1,645 major depressive disorder cases and 340 controls, a polygenic score was constructed for multiple *p*-value thresholds. Peyrot et al. tested for both a departure from multiplicativity and a departure from additivity, to test for interactions. From this, Peyrot et al. found evidence for interaction effects between multiple polygenic scores and childhood trauma, which helped the authors to conclude that individuals are at an increased risk of developing major depressive disorder if they have both a high PRS and have been exposed to childhood trauma. Another major depressive disorder and childhood trauma interaction study was published a couple of years later, this time by Mullins et al.(139), but with the additional risk factor adult stressful life events. Mullins et al. tested whether a polygenic score for major depressive disorder interacted with either adult stressful life events or childhood trauma. A polygenic score was constructed using SNPs effects from summary data based on a mega-analysis on major depressive disorder (7,615 cases and 7,931 controls), conducted by the Psychiatric Genomics Consortium. Polygenic scores were constructed for 1,605 major depressive disorder cases and 1,064 controls with stressful life event data, and 240 major depressive disorder cases and 272 controls with childhood trauma data. Polygenic scores were then constructed for multiple *p*-value thresholds for up to 87,737 SNPs. Mullins et al. found evidence of a significant interaction existing between the major depressive disorder derived polygenic scores and childhood trauma, but no significant interactions were found between any of the polygenic scores and stressful life events. Salvatore et al.(141) have also conducted a PRS x environmental interaction study. Salvatore et al. examined whether there was evidence that a polygenic score for alcohol problems could be modified by either parental knowledge or peer deviance. A polygenic score was constructed for multiple *p*-value thresholds using up to 1,231,165 SNPs. The SNP effects, used to construct the polygenic scores for 1,162 subjects from an independent alcohol problem study (FinnTwin12), were estimated using GWAS results based on the Avon Longitudinal

Study of Parents and Children (ALSPAC) study, which included 4,304 European individuals. The polygenic score constructed using SNPs with a $p \leq 0.05$ was estimated to explain the largest proportion of variation in alcohol problems, so Salvatore et al. then used this polygenic score to test for a PRS-parental knowledge and PRS-peer deviance interaction. A multiple regression analysis was used with alcohol problems as the dependent variable, and sex, PRS, environment factors and PRS x environmental factor as covariates in the model. The results from this analysis indicated that PRS-environmental factor interactions, with both factors, existed as both interaction were found to be significant (PRS x parental knowledge *p*-value = 0.02 and PRS x peer deviance *p*-value = 0.04).

More recently, an interaction study based on UK Biobank data was conducted by Tyrrell et al.(142) to examine whether a polygenic score for BMI, constructed using 69 BMI associated SNPs, interacted with any one of the twelve measures of obesogenic environments and behaviour factors, to have an effect on obesity. Such factors included TV watching, vigorous activity and sedentary behaviour. The SNP effects used in the polygenic score were based on SNPs from the GIANT consortium, which contained up to 339,224 individuals. The UK Biobank study was used to provide information on obesogenic environmental factors for up to 120,000 individuals. Using a linear regression model with BMI as the dependent variable, Tyrrell et al. found evidence an interaction between the BMI polygenic score and three of the twelve obesogenic measures: self-reported TV watching, self-reported physical activity and the Townsend deprivation index (measure of social-economic position).

However, not all studies conducted have successfully found evidence of PRS-environmental factor interactions. A recent study conducted by Trotta et al. (143) tested whether there was evidence that a polygenic score for schizophrenia, interacted with childhood adversity to have an effect on psychosis. The polygenic score for schizophrenia was constructed using genome-wide significant SNP effects, extracted from a large mega-analysis conducted by the Schizophrenia Working Group of the

Psychiatric Genomics Consortium. A polygenic score was constructed for multiple *p*-value thresholds, for 80 psychosis cases and 110 controls. To test for a PRS-childhood adversity interaction, a logistic regression model was used, with psychosis outcome as the dependent variable and the PRS, childhood adversity and the PRS-childhood adversity interaction as independent variables. Age, sex, educational level and ten principal components were also adjusted for. The interaction between the polygenic score and childhood adversity was non-significant (*p-value = 0.632),* therefore Trotta et al. failed to find evidence of an interaction. The authors concluded that the effect the schizophrenia polygenic score had on psychosis was not modified by the presence of childhood adversity. However, with very small number of subjects used in this study, the power to detect a significant interaction would have been very low. With the study being underpowered, it is possible that with a much larger replication sample, an interaction may be detected. These are just a number of the interaction studies that have been conducted to date.

In breast cancer, a number of SNP-environmental interaction studies have been conducted, but there has been an absence of PRS-environmental interaction studies. Campa et al. (144) have examined, using 8,576 breast cancer cases and 11,892 controls, whether there was evidence that 17 published susceptibility breast cancer loci individually interact with a number of established breast cancer risk factors, to have an effect on breast cancer risk. The study focused on nine breast cancer risk factors: age at menarche, parity, age at menopause, use of hormone replacement therapy, family history, height, BMI, smoking status and alcohol consumption. Two models for each SNP-environmental factor pair were constructed, the first model with breast cancer as the outcome and both the SNP and environmental factor as independent variables, whilst adjusting for age, study, ethnicity and country. The second model contained the same variables, but with an additional SNP-environmental factor interaction. A likelihood ratio test was then used to compare the goodness of fit of the two models, for each SNP-environmental factor pair. No significant interactions were detected, so the authors concluded that there was no evidence that these known common variants strongly modified the associations between the environmental risk factors and breast cancer. A larger study, conducted using data from 24 BCAC studies involving up to 34,793 breast cancer cases and 41,099 controls of European ancestry, was later performed by Nickels et al. (145). They investigated whether interactions existed between 23 breast cancer susceptibility SNPs, and 10 environmental risk factors (age at menarche, parity, breast feeding, BMI, height, oral contraceptive use, menopausal hormone therapy use, alcohol consumption, cigarette smoking and physical exercise). Using a similar approach to Campa et al. (144), they identified significant interactions between a number of published breast cancer risk loci and age at menarche, parity, breast feeding, BMI, height, oral contraceptive use, menopausal hormone therapy use, alcohol consumption, cigarette smoking and physical activity.

Mavaddat et al.(52) have examined whether there is evidence that a polygenic score, based on 77 published genome-wide significant SNPs, interacts with either age or family history to have an effect on either the occurrence of overall breast cancer outcome, estrogen receptor (ER)-positive breast cancer or ER-negative breast cancer. Polygenic scores were constructed for 46,450 breast cancer cases, 27,074 ER-positive breast cancer cases, 7,413 ER-negative breast cancer cases and 42,599 controls, taken from BCAC, using SNP effects based on previously published results produced for 77 genome-wide significant SNPs. Multiple age group intervals were formed, and a logistic regression model was used to test whether interactions were significant. For overall breast cancer and ER-positive breast cancer, there was shown to be a significant interaction between age and PRS, but the same could not be shown for PRS and family history. A problem with this analysis was that a large number of the published genome-wide significant SNPs have been previously identified using BCAC data. Using SNP effects based on published results to construct polygenic scores for the women in the BCAC sample would then mean that there will be some subject overlap in the training and replication sample, meaning that the two samples would not completely independent. This could have therefore inflated the association between the score and the breast cancer outcome in the replication sample, and caused inaccuracies in the interaction analysis. This omission should be considered when drawing a conclusion based on the analyses conducted for this study.

SNP-SNP interactions are another form of interaction that tends to be examined. With broad-sense heritability including epistasis effects, the differences in the broad-sense and the narrow-sense heritability estimates could be explained partially by SNP-SNP interactions. For breast cancer, Mavaddat et al.(52) have tested whether individual SNPs interact with each other to have an effect on breast cancer risk. The SNPs used were 77 known genome-wide significant SNPs breast cancer SNPs, published at the time the analysis was conducted. There was found to be no evidence of SNP x SNP interactions existing between any of the 77 SNPs. With breast cancer being a polygenic trait, it is possible that interactions between a combination of SNPs and an individual SNP may exist. This has not been examined yet in breast cancer, but a PRS-SNP interaction could show that the effect a polygenic score has on a trait is modified by a single SNP.

At the time of conducting the analyses presented in this chapter, there had been no PRS-environmental factor interaction studies conducted using an en-masse polygenic score for breast cancer. In the studies that had been conducted, the PRS had been constructed using only SNPs that had been identified as genome-wide significant, either independently, or in a PRS.

5.2 Proposed method

In this chapter, two main analyses have been performed. The first analysis focuses on establishing whether the effect a breast cancer polygenic score has on breast cancer outcome is modified by either BMI or age at menarche. The second analysis tested whether the effect a breast cancer polygenic score has on breast cancer outcome is modified by individual genotyped SNPs. To date, breast cancer interaction studies have tended to focus on a limited number of SNPs, these being the susceptibility SNPs discovered by the time the research was conducted. It has been tested whether known susceptibility SNPs interact with an environmental factor individually, or whether a PRS x environmental factor interaction is present, when combining the known susceptibility SNPs. With breast cancer being a polygenic trait, and with much of the genetic variation in breast cancer yet to be explained by individual genetic variants, an en-masse polygenic approach was instead adopted for the following analyses. With many associated genetic variants yet to be discovered, the en-masse approach will enable unknown associated SNPs to be represented in the analyses. As an en-masse polygenic approach has been used, the analyses have mainly focused on the two breast cancer GWAS (UK2 and BBCS), so that as many SNPs as possible across the genome could be represented in the analysis.

In the following analyses, the PRS-environmental interactions tested were PRS-age at menarche and PRS-BMI. Both age at menarche and BMI are breast cancer risk factors, and in chapter 2 it was observed that breast cancer polygenic scores, for different *p*-value thresholds, could be used to predict breast cancer outcome. With that in mind, it was tested whether there was evidence to suggest that the two risk factors interact with breast cancer polygenic scores, to influence breast cancer outcome. Out of the many known breast cancer risk factors, these two risk factors were chosen because there is evidence that individual breast cancer genetic variants interact with either one of these risk factors (145), and the data for these factors were available in one of the GWAS. Both BMI and age at menarche data were present in the BBCS GWAS, but not the UK2 GWAS. Not only were the number of BMI and age at menarche subjects limited because the data was only available for subjects in one of the GWAS, but not all of the women present in the BBCS had this information. Only approximately half of the cases in the GWAS had either BMI or age at menarche data, with none of the controls having this information. With a very small of subjects with either BMI or age at menarche information, and with larger studies, previously conducted, being unable to detect significant interactions, a case-only approach was adopted in order to improve the statistical power to detect any present interactions.

5.2.1 Case-only design

Interaction studies are known to suffer from reduced statistical power, with sample sizes needing to be larger than most other study types in order to detect a present interaction. As stated in the preceding chapters, the UK2 and BBCS GWAS are both relatively small, sample size wise, but they do offer genome-wide coverage. Research has suggested that a case-only analysis can be more effective at detecting interactions, than case-control studies of the same size (146). However, caution should be taken when using only cases to detect an interaction (147). A systematic review and meta-analysis investigating bias in both case-only gene-environment interactions studies, and gene-gene interaction studies was executed by Dennis et al.(147). Generally case-only studies do not incur more bias than case-control studies, but the review did find that correlation between the genotype and environmental factor was the main cause of bias in case-only interaction studies. With a case-only study, two assumptions should hold in order to reduce the risk of bias. The first assumption is that the interacting factors being tested should be independent of each other in the general population, and the second assumption is that the disease of interest should be rare.

Assuming that we wish to test for an interaction between two factors, a logistic regression model can be fitted to the data such that:

$$logitP(Y = 1) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G \times E$$

Where, Y is the disease (binary), β_0 is the intercept, β_1 and β_2 are the coefficients for the genetic factor (G) and the risk factor (E), respectively. The interaction between G and E is given as $G \times E$, with the coefficient β_3. The probability of disease is denoted as P ($P \in [0,1]$), and $\ln\left(\frac{P}{1-P}\right)$ is the log-odds ($\ln\left(\frac{P}{1-P}\right) \in [-\infty, +\infty]$).

The interaction term (β_3) can be defined as:

$$\beta_3 = \frac{OR_{G \times E}}{OR_G OR_E}$$

Where, $OR_{G \times E}$, OR_G and OR_E are the odds of risk for individuals with both G and E present together, just G present and just E present, respectively.

From this, the odds ratio for the G and E in cases ($OR_{case-only}$) can be written as:

$$OR_{case-only} = \beta_3 \times OR_{control-only}$$

Where, $OR_{control-only}$ the odds ratio for the G and E in controls.

If it can be assumed that G and E are independent and that the disease is rare, then $OR_{control-only}$ = 1, which then means that (148, 149):

$$OR_{case-only} = \beta_3$$

So, β_3 can be estimated using only cases.

Assuming that G is a binary variable, a logistic regression model can be fitted to the data for cases only, such that (149):

$$logitP(G = 1) = \beta_0 + \beta_1 E$$

With a continuous G, this being the polygenic score, a linear regression model can instead be fitted to the case data:

$$G = \beta_0 + \beta_1 E$$

If there is shown to be a linear relationship between the G and E, then this could be evidence that an interaction exists. This approach can also be applied when testing for PRS × SNP interactions.

Independence between the G and E can be a strong assumption to make, and it is thought that many case-only interaction analyses are conducted even though the assumption may not hold (150). It was uncertain whether each of the polygenic scores and either age at menarche, BMI and the genotyped SNPs are truly independent in the population. However, as explained later in this chapter, information on the environmental factors were only available for a small proportion individuals in the studies used in this thesis, all of which were breast cancer cases. This meant that it was not possible to conduct a case-control interaction analysis, therefore a case-only interaction analysis was performed instead.

The prevalence throughout this thesis has been assumed to be 0.1%, which could be considered rare. Therefore, this makes it more likely that the assumption holds.

5.2.2 PRS x environmental factor interactions

Two breast cancer risk factors, one modifiable risk factor (BMI) and the other a reproductive risk factor (age at menarche), were included in the PRS x risk factor interaction analysis. A case-only approach was adopted for this analysis. A polygenic score was constructed for women who had been diagnosed with breast cancer in the BBCS GWAS, who also had either a BMI measure or a menarche age. The SNP effects used in the polygenic score were estimated using the remaining BBCS cases who did not have either BMI or age at menarche information, the BBCS controls and the UK2 individuals. A logistic regression model was used to estimate the individual

SNP effects used in the score, whilst including ancestry principal components as covariates in the model to adjust for population stratification. The SNPs were then sorted by their association significance with breast cancer into multiple threshold groups, these being: $p \leq 1$, $p \leq 0.7$, $p \leq 0.4$, $p \leq 0.1$, $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.001$. A polygenic score, based on the SNP effects for the SNPs in each threshold group, was then constructed for either the BMI BBCS cases, or the age at menarche BBCS cases. The scores were then tested for their association with either BMI or age at menarche using a linear regression model, whilst adjusting for four ancestry principal components:

$$PRS_{case-only} = \beta_0 + \beta_1 \mathrm{BMI}_{case-only} + \beta_2 PC_1 + \beta_3 PC_2 + \beta_4 PC_3 + \beta_5 PC_4$$

$$PRS_{case-only} = \beta_0 + \beta_1 \mathrm{AM}_{case-only} + \beta_2 PC_1 + \beta_3 PC_2 + \beta_4 PC_3 + \beta_5 PC_4$$

Where, $\mathrm{BMI}_{case-only}$ and $\mathrm{AM}_{case-only}$ are the BBCS case sample with BMI and age at menarche data, respectively. The ancestry principal components for the BBCS cases are defined as PC_1, PC_2 , PC_3 and PC_4, with $\beta_2$to β_5 being the corresponding coefficients for PC_1 to PC_4.

For a separate PRS-environmental factor analysis, the COGS was used to estimate the SNP effects used to construct the polygenic score for the BBCS BMI and age at menarche breast cancer cases. A logistic regression model was used to estimate the individual SNP effects used in the score, with nine ancestry principal components and sub study included as covariates in the model. The SNPs were then organised by their significance with breast cancer into multiple threshold groups, the same thresholds used in the GWAS analysis. The SNP effects for the SNPs in each threshold group were then used to construct a PRS for either the BMI BBCS cases, or the age at menarche BBCS cases. The scores were then tested for their association with either BMI or age at menarche using a linear regression model, with four principal components included in the model.

In order to increase the number of SNPs present in the polygenic scores, the UK2 and COGS training SNPs were extracted from the imputed BBCS SNPs for the age at menarche and BMI BBCS cases. The imputed BMI and age at menarche BBCS breast cancer case sets were then used as the replication sample in the polygenic score analysis.

If a significant association between the polygenic score and environmental factor is shown, it would suggest that an interaction exists between the environmental factor and the breast cancer polygenic score.

5.2.3 PRS x SNP interactions

In chapter 2, it was observed that polygenic scores based on GWAS SNPs were significantly associated with breast cancer risk in an independent GWAS. It was shown that a PRS constructed using all independent GWAS SNPs en-masse, regardless of their individual significance with the trait, was significantly associated (*p*-value < 0.05) with breast cancer risk in an independent GWAS. This was also the case when being more stringent on the SNPs used to construct the polygenic scores. The significant associations observed were bi-directional, polygenic scores for both GWAS were shown to significantly predict risk of breast cancer in the other GWAS. This analysis was taken further by examining whether any of the SNPs used to construct a polygenic score, significantly interact with the polygenic score to have an effect on breast cancer risk. With each of the scores being significantly associated with breast cancer risk in an independent sample, does the presence of a certain SNP in the score, modify the effect the score has on the trait (PRS x SNP interaction)? A case-only approach was used to examine whether there was evidence of such interactions present.

For this analysis, multiple polygenic scores were constructed using individual SNP effects estimated using BBCS cases and controls. The scores constructed were based on SNPs, which included imputed BBCS SNPs based on the UK2 GWAS SNPs, with an individual $p \leq 1$, $p \leq 0.7$, $p \leq 0.4$, $p \leq 0.1$, $p \leq 0.05$, $p \leq 0.01$ and $p \leq 0.001$. Individual training sample SNP effects were estimated using a logistic regression model, with four ancestry principal components included as covariates in the model.

For each polygenic score, the SNPs effects of m *SNPs* were used to construct a score in the replication sample cases (UK2 GWAS) using the PRS formula:

$$\widehat{PRS}_{m\ (case-only)} = \sum_{j=1}^{m} \hat{\beta}_{j1} G_{j\ (case-only)}$$

Let $\hat{\beta}_{j1}$ be the SNP effect for SNP j based on the training sample (BBCS GWAS) estimated using a logistic regression model with four principal components included as covariates, and $G_{j(case-only)}$ be the coded allele, with 0,1 and 2 for SNP j for only the cases in the replication sample (UK2 GWAS).

To test whether one of the SNPs used to construct a score significantly interacts with the score, the SNP effect for the SNP is subtracted from polygenic score, for each replication sample individual.

The polygenic score used in the interaction analysis, excluding the SNP effect of the tested SNP, is then:

$$\begin{aligned} \widehat{PRS}_{m-1\ (case-only)} &= \widehat{PRS}_{m\ (case-only)} - \mathrm{SNP}_a \\ &= \left(\sum_{j=1}^{m} \hat{\beta}_{j1} G_{j\ (case-only)} \right) - \hat{\beta}_{a1} G_{a\ (case-only)} \end{aligned}$$

Where, SNP_a is a tested SNP that was initially used to construct the score, $\hat{\beta}_{a1}$ is the SNP effect for SNP_a and $G_{a\ (case-only)}$ is the coded allele (0,1 and 2) for SNP_a for replication cases only. This analysis is performed under the assumption that there is no LD between the individual SNP being tested and all the SNPs present in the polygenic score.

The interaction between $\widehat{PRS}_{m-1}$ and SNP_a can then be tested using a linear regression model:

$$\widehat{PRS}_{m-1\,(case-only)} = \beta_0 + \beta_1 \mathrm{SNP}_{a\,(case-only)} + \beta_2 PC_1 + \ldots + \beta_{11} PC_{10}$$

Where, β_0 is the intercept and β_1 is the coefficient for SNP_a for case subjects only. The first ten principal components used to minimise any population stratification present are denoted as PC_1 to PC_{10}, with $\beta_2$to β_{11} being the corresponding coefficients for PC_1 to PC_{10}.

By testing for multiple interactions between a threshold polygenic score and the SNPs within that same threshold, multiple comparisons should be corrected for. One would expect, at a 5% significance level, that 5% of the interactions tested are significant by chance alone. Therefore, to adjust for multiple comparisons, the false discovery rate (FDR) was used.

5.3 PRS x risk factor interaction analysis

In order to perform the PRS x risk factor analyses, two independent samples were needed and one of them, the replication sample, needed to contain individuals with BMI or age at menarche data. The main focus of these interaction analyses was to establish whether there was evidence of an interaction existing between a breast cancer polygenic score, that represents SNPs across the genome, and either BMI or age at menarche. With the data I had access to, BMI and age at menarche data was only available for BBCS subjects, I did not have this information for the UK2 or COGS subjects. This meant that only a small number of individuals could then be used to test for an interaction, so in order to improve the statistical power to detect an interaction, a case-only approach was implemented. The BBCS cases with either BMI or age at menarche data were assigned to the replication sample, and the remaining BBCS cases, those without either BMI or age at menarche data, and controls were assigned to the training set. The UK2 GWAS was combined with BBCS training sample to increase the number of individuals in the training sample, in order to improve the precision of the SNP effect estimates used to construct the polygenic score. For the BMI interaction analysis, this meant that the combined GWAS training sample consisted of 4,316 cases and 5,190 controls, with the replication set containing 921 BBCS cases. For the age at menarche interaction analysis, the combined GWAS training set contained 4,312 cases and 5,190 controls, with the replication set containing 925 BBCS cases.

To increase the training sample size further, and test whether a score enriched for breast cancer associated SNPs interacts with either breast cancer risk factors, the COGS was used as the training sample in a separate interaction analysis. For the BMI analysis with COGS subjects as the training sample, the training sample consisted of all European COGS subjects (48,064 cases and 43,486 controls), with the replication sample containing 921 BBCS cases. For the age at menarche analysis with the COGS subjects as the training sample, the training sample again consisted of 48,064 cases and 43,486 controls, with the replication sample containing 925 BBCS cases. Even though the COGS training set data was external to the BBCS GWAS, the number of BBCS subjects in the replication sample could not be increased further as the BMI and age at menarche information was only available for those BBCS subjects. BMI in the replication sample was found to range from 16.57 to 47.22 when including the women that are considered to be outliers ("underweight" to "obese") (Figure 5-1), and the mean BMI for women in the replication sample was calculated to be 26.71, which is just within the "overweight" interval (25 ≤ BMI < 30). The age at which a woman in the replication sample has her first menstruation cycle ranged from 9 to 20 years when including women that are considered to be outliers (Figure 5-2), with the mean age being 13 years.

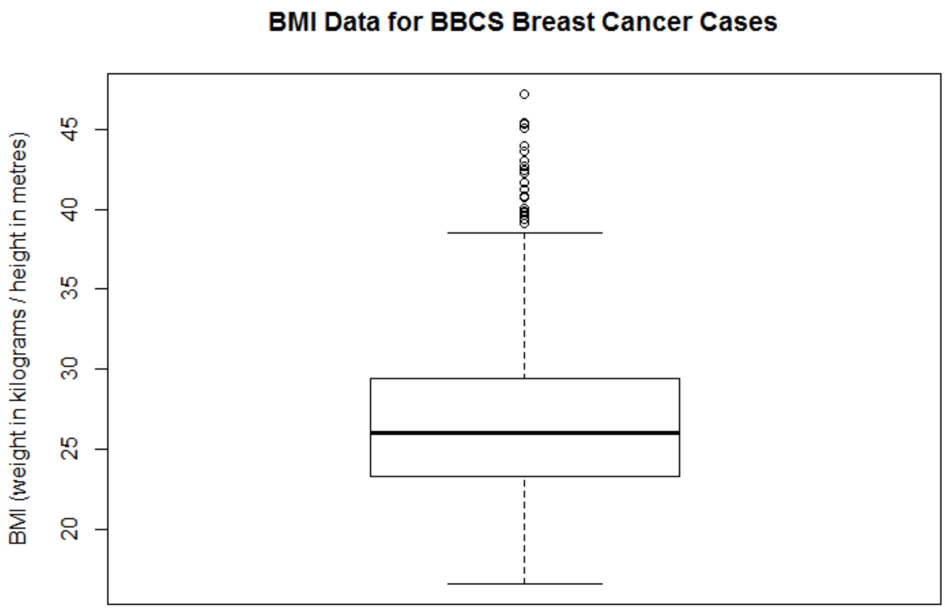


Figure 5-1: Boxplot of BMI distribution for replication sample

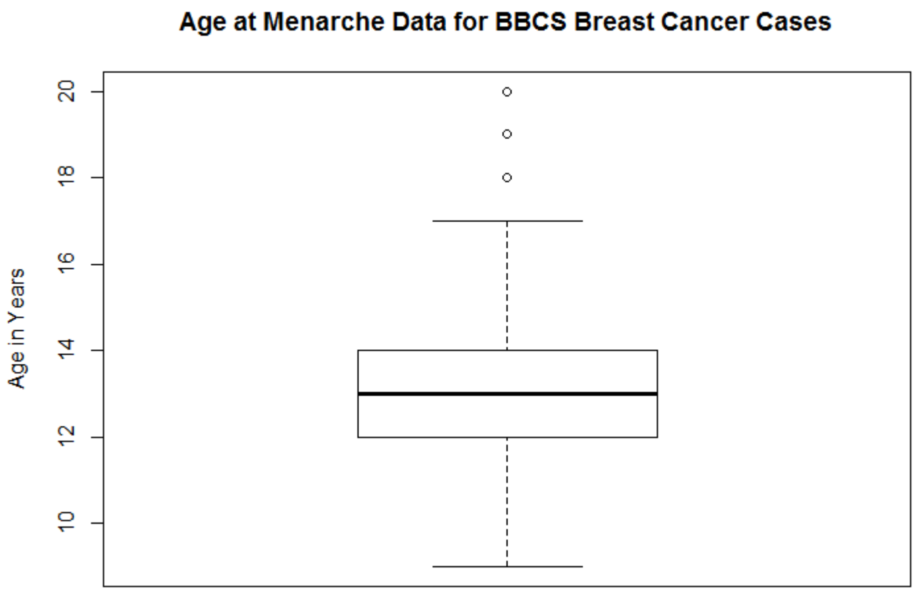


Figure 5-2: Boxplot of age at menarche distribution for replication sample

The SNPs used to construct the multiple polygenic scores were the SNPs retained after QC and LD-clumping (r^2 > 0.1). In order to increase the number of SNPs used to construct the polygenic scores for the BBCS cases, imputed SNPs were used. The UK2 SNPs retained after QC and LD-clumping (r^2 > 0.1) that had not been genotyped in the BBCS, were extracted from the BBCS imputed SNPs for all BBCS individuals. This then meant that up to 82,823 SNPs were used to construct the polygenic scores for the BBCS cases. Imputed SNPs were also used in the analyses when the COGS was the training sample. The COGS SNPs retained after QC and LD-clumping (r^2 > 0.1), that had not been genotyped in the BBCS, were extracted from the BBCS imputed SNPs for the BBCS cases used in the replication sample. Using imputed SNPs in the COGS based analyses meant that up to 41,651 SNPs were used to construct the polygenic scores for the BBCS BMI or age at menarche cases. After computing the scores, a linear regression model was used to test whether a breast cancer risk score was linearly associated with either BMI or age at menarche, with four principal components included as covariates in the model.

There was shown to be a non-significant linear association between age at menarche and the breast cancer polygenic score derived from 82,823 UK2/BBCS SNPs ($p \leq 1$, *p*-value = 0.602) (Table 5-1). The same was shown for each *p*-value threshold, none of the UK2/BBCS derived polygenic scores for the BBCS age at menarche subjects had a significant linear association with age at menarche in the BBCS cases. A significant linear association between age at menarche and the breast cancer polygenic score derived using 41,651 COGS SNPs ($p \leq 1$, *p*-value = 0.020) was however observed, suggesting that an interaction between the two exists (Table 5-1). A significant association between age at menarche and PRS was also observed for the score constructed using SNPs with a *p*-value ≤ 0.7 and a *p* -value ≤ 0.4 (association *p*-value = 0.016 and association p-value = 0.042). The significant associations observed, were however not as significant as one would have hoped. Nonetheless, the results still suggested a significant PRS x age at menarche interaction, when there was less of a restriction on the SNPs included in the polygenic score and when the score was based on the SNP effects estimated using the COGS.

The linear association between BMI and the breast cancer polygenic score derived from 82,823 UK2/BBCS SNPs (*p* ≤ 1, *p*-value = 0.153) was observed to be non-significant (Table 5-2). This was also found to be the case for most of the scores constructed using UK2/BBCS SNP estimates. A significant association was, however, observed between BMI and the breast cancer polygenic score derived using 377 GWAS SNPs (*p* ≤ 0.001, *p*-value = 0.040). The result suggested a significant PRS x BMI interaction existed when there was a restriction on the BBCS/UK2 SNPs included in the polygenic score. The same was not shown for any of the COGS derived polygenic scores, so there was no evidence to suggest an interaction between BMI and any of the breast cancer polygenic scores existed, when using COGS SNP effects.

		Age at menarche		
Training set	Replication set	*p*-value threshold	No. SNPs	*p*-value
UK2/BBCS	BBCS cases	*p* ≤ 1	82,823	0.602
		p ≤ 0.7	70,783	0.658
		p ≤ 0.4	50,954	0.799
		p ≤ 0.1	18,675	0.822
		p ≤ 0.05	10,751	0.967
		p ≤ 0.01	2,853	0.361
		p ≤ 0.001	377	0.208
COGS	BBCS cases	*p* ≤ 1	41,651	0.020
		p ≤ 0.7	34,575	0.016
		p ≤ 0.4	24,590	0.042
		p ≤ 0.1	9,597	0.427
		p ≤ 0.05	5,833	0.527
		p ≤ 0.01	1,962	0.959
		p ≤ 0.001	529	0.131

Table 5-1: Linear regression: breast cancer polygenic score and age at menarche

			BMI	
Training set	Replication set	p-value threshold	No. SNPs	*p*-value
UK2/BBCS	BBCS cases	$p \leq 1$	82,823	0.153
		$p \leq 0.7$	70,783	0.190
		$p \leq 0.4$	50,954	0.211
		$p \leq 0.1$	18,675	0.200
		$p \leq 0.05$	10,751	0.494
		$p \leq 0.01$	2,853	0.774
		$p \leq 0.001$	377	0.040
COGS	BBCS cases	$p \leq 1$	41,651	0.838
		$p \leq 0.7$	34,575	0.828
		$p \leq 0.4$	24,590	0.795
		$p \leq 0.1$	9,597	0.130
		$p \leq 0.05$	5,833	0.169
		$p \leq 0.01$	1,962	0.287
		$p \leq 0.001$	529	0.309

Table 5-2: Linear regression: breast cancer polygenic score and BMI

5.4 PRS x SNP interaction analysis

To examine whether individual SNPs modify the combined effect of SNPs on breast cancer risk, significant PRS x SNP interactions were tested for. All possible pair-wise interactions between individual SNPs and the PRS were tested, with the training sampling consisting of BBCS subjects and the replication sample represented by UK2 study cases. In order to maintain a large number of SNPs in the PRS, only the two GWAS were analysed. Imputed SNPs were also included in the analysis, to further increase the number of SNPs in the analysis. SNP effects were estimated using the BBCS subjects and a logistic regression model, with four principal components included as covariates.

Before correcting for multiple comparisons, 3,539 significant interactions between the PRS constructed using all independent GWAS SNPs, minus the one tested SNP, and individual SNPs were observed. After correcting for multiple comparisons using a FDR of 5%, no significant interactions were observed when using all SNPs in a score. For the $p \leq 0.001$ interval there were found to be 217 significant interactions, out of 220 tested PRS x SNP interactions, when testing at a 5% significant level. This meant that approximately 99% of the tested interactions were observed to be significant, which was a lot greater than the 5% that would be expected by chance. Even after correcting for multiple comparisons using a FDR of 5%, all 217 significant interactions were still observed. For the $p \leq 0.01$ interval, 432 significant interactions were observed, 89 of which were still significant after adjusting for multiple comparisons. After adjusting the association *p*-values by the FDR for the polygenic scores and SNP interaction tests constructed using SNPs with *p*-value thresholds $p \leq 0.05$ to $p \leq 1$, no SNPs were found to be significantly associated with the polygenic score constructed using the remaining SNPs within the same bin. Therefore, for these bins no evidence was found to suggest that individual SNPs interact with the constructed polygenic scores.

Focusing on the intervals where significant interactions were still observed after correcting for multiple testing (*p* ≤ 0.01 and *p* ≤ 0.001), there were found to be 17 common SNPs between those significant after FDR in *p* ≤ 0.01 and *p* ≤ 0.001 bins. None of the 17 common SNPs have previously been shown to be associated with breast cancer or other traits at genome-wide significance, in fact none of the 89 single SNPs in the *p* ≤ 0.01 interval have currently been found to reach genome-wide significance for any traits. Seven of the 217 SNPs observed to significantly interact with the PRS constructed using SNPs with a *p* ≤ 0.001 are published genome-wide significant breast cancer SNPs (Table 5-4).

With approximately 99% of the SNP x PRS interactions being found to be significant after adjusting by an FDR of 5% for the *p* ≤ 0.001 bin, this may suggest that there could be SNPs in the score which are highly correlated with the individual SNP being tested. None of the individual SNPs were found to be highly correlated with the remaining SNPs used to construct the PRS (all correlations were $r^2 < 0.2$), meaning that it was unlikely that the linear association was driven by correlation between the individual SNPs and those used in the score. This was also found to be the case for the *p* ≤ 0.01 analysis as none of the 89 SNPs were found to be highly correlated with the remaining SNPs used to construct the PRS (all correlations were $r^2 < 0.2$).

The results indicate that it could be possible that individual SNPs modify the combined effect of SNPs on breast cancer risk, with some of the individual SNPs having previously been observed to be associated with breast cancer risk. With each polygenic score in chapter 2 having been shown to be significantly associated with breast cancer risk in an independent sample, the results from the PRS x SNP interaction analysis suggest that the presence of a certain SNPs in either the *p* ≤ 0.01 or *p* ≤ 0.001 score, could modify the effect the score has on breast cancer risk. However, this is a case-only analysis, so it would be best to see whether the interactions replicate in a case-control setting.

PRS	No. SNPs*	Sig. interactions**	FDR**
$p \leq 1$	66,339	3,539	0
$p \leq 0.7$	55,786	2,959	0
$p \leq 0.4$	39,217	2,139	0
$p \leq 0.1$	13,442	887	0
$p \leq 0.05$	7,474	622	0
$p \leq 0.01$	1,813	432	89
$p \leq 0.001$	220	217	217

** The number of SNPs with a p*-value less than or equal to the given PRS threshold
*** p-value < 0.05*

No. SNPs-1 = the number of SNPs used to construct the PRS

For the FDR, the total no. SNPs in the PRS were used as the number of tests

Table 5-3: Testing for sig. interactions between SNPs and $\widehat{PRS}_{m-1\,(case-only)}$

SNP	Chromosome	Position
rs11249433	1	10566215
rs13387042	2	217905832
rs12655019	5	56195790
rs865686	9	110888478
rs1219648	10	123346190
rs10995190	10	64278682
rs3803662	16	52586341

Table 5-4: Published genome-wide significant breast cancer SNPs found to significantly interact with PRS

5.5 Discussion

In chapter 2, it was observed that polygenic scores constructed using breast cancer GWAS SNPs from one GWAS were associated with breast cancer status in an independent GWAS. To investigate this further, I tested whether there was evidence that the effect a breast cancer polygenic score has on breast cancer risk could be modified by either BMI or age at menarche. For other complex diseases, with a polygenic basis, evidence of PRS-environmental factor interactions have been established (139-142). Individual breast cancer susceptibility variants have been previously shown to interact with BMI and age at menarche, but this was the first time that it has been tested whether an en-masse breast cancer PRS interacts with either risk factor.

Initially, for an interaction analysis, the two breast cancer GWAS would have been considered small, sample size wise. The size of the replication GWAS was reduced further as only a limited number of BBCS cases had either BMI or age at menarche information. As only cases had BMI or age at menarche data, and to improve the power to detect significant interactions, a case-only approach was implemented. To conduct a case-only interaction analysis, it is assumed that the disease being studied is rare and that in the population the gene and environment factors being tested are independent. The problem with assuming independence is that, typically, there is uncertainty as to whether the assumption holds (149). Therefore, great care should be taken when drawing a conclusion based on the results of a case-only interaction analysis for this reason. Even though there was some uncertainty as to whether the assumption of independence holds between the breast cancer polygenic scores and BMI, age at menarche and the genotyped SNPs, a case-only analysis was conducted because information on the environmental factors were only available for a small proportion of cases genotyped in the studies used in this thesis. The interactions should also be tested using a case-control interaction analysis in a much larger number of individuals.

For the case-only interaction analysis conducted in this chapter, multiple polygenic scores were constructed for the BBCS cases who had either age at menarche or BMI information, for different *p*-value thresholds. A linear regression model was then used to model a PRS and an environmental factor, with a significant association providing evidence that a significant interaction exists. As none of the UK2/BBCS derived polygenic scores had a significant linear relationship with age at menarche, there was no evidence to suggest that a polygenic score constructed using GWAS SNPs interacts with age at menarche. A number of the scores constructed using SNPs genotyped on the iCOGS custom array were, however, shown to be significantly associated with age at menarche, thus suggesting that an interaction exists. The scores derived using COGS SNPs with a *p*-value ≤ 1, *p*-value ≤ 0.7 and *p*-value ≤ 0.4 were shown to be significantly associated with age at menarche ($p < 0.05$). When being more stringent on the choice of SNPs used to construct the polygenic score, the associations become non-significant. The results suggest that the breast cancer scores constructed using a large number of independent genotyped SNPs, could interact with age at menarche to have an effect on breast cancer risk. However, no significant associations were observed between BMI and any of the polygenic scores constructed, using either UK2/BBCS SNPs, or COGS SNPs. The PRS x environmental factor analyses conducted in this chapter would have only had up to 25% power to detect a PRS association with BMI. Therefore, the analyses should be replicated in a larger sample, preferably a sample with a greater number of individuals with BMI and age at menarche information.

BMI and age at menarche are not the only environmental factors that have been identified as breast cancer risk factors. Further analyses should therefore be conducted to examine whether other breast cancer risk factors, such as percent mammographic density, interact with breast cancer polygenic scores. Unfortunately, at the time of performing the analyses conducted in this chapter, I did not have access to data that would have enabled me to investigate whether interactions between PRS and other breast cancer risk factors exist. The data for other breast cancer risk factors was available, but it would have been too time consuming to apply for it, and this would have delayed my analyses.

In this chapter, it was also examined whether any of the genotyped GWAS SNPs interacted with a polygenic score to have an effect on breast cancer risk. Significant associations were found, and surprisingly for the $p \leq 0.001$ interval it was found that approximately 99% of the tested interactions were observed to be significant, which was a lot greater than the 5% that would be expected by chance. Even after adjusting for multiple testing using an FDR < 5%, a number of significant associations were observed for the scores constructed using SNPs with a $p \leq 0.01$ and $p \leq 0.001$. For the other intervals, no significant interactions were observed after adjusting for multiple testing. After measuring the correlation between the individual SNPs and those used in the PRS for the significant interactions, none of the individual SNPs were found to be highly correlated with the remaining SNPs used to construct the PRS. Only 17 SNPs were found to significantly interact with the PRS based on remaining SNPs with a $p \leq 0.01$ and $p \leq 0.001$, with none of the SNPs shown to significantly interact with any of the other scores ($p \leq 1$, $p \leq 0.7$, $p \leq 0.4$, $p \leq 0.1$ and $p \leq 0.05$) after correcting for multiple testing. There was therefore no evidence to suggest that these SNPs interacted with other PRS, just those based on SNPs with a $p \leq 0.01$ and/or $p \leq 0.001$. None of the individual SNPs were found to be highly correlated with any of the remaining SNPs used to construct the $p \leq 0.01$ and $p \leq 0.001$ scores, therefore suggesting that correlation between SNPs is not driving the significant linear association, and that it is possible that these SNPs are interacting with the scores. With this being a case-only analysis, it should be tested whether the same can be shown when conducting a case-control interaction analysis.

Chapter 6 Analysis of breast cancer susceptibility loci by Capture Hi-C (CHi-C)

6.1 Introduction

As it becomes possible to genotype a larger number individuals, for a larger number of genetic variants, it is expected that the number of breast cancer susceptibility loci identified will increase. However, it is not even clear for many of the breast cancer susceptibility variants identified to date which variant is the causal variant, or how disease risk is influenced by the variant. Many of the breast cancer susceptibility loci identified so far map to non-protein-coding regions of the genome, or regions of the genome that contain no genes (gene deserts), thus making it difficult to understand their function (106). Not understanding the underlying biological mechanism for susceptibility variants hinders the development of breast cancer prevention methods and treatments (151). This has not just been the case in breast cancer, the majority of susceptibility loci identified for other complex diseases so far also map to non-coding regions or gene-deserts (152-154).

Through studying genome structure, it is possible to gain a better understanding of the functions of these loci. It has been suggested that loci mapping to non-coding regions of the genome, could have an effect on disease risk through physical interactions with other loci across the genome, with these other loci not necessarily being positioned close to the susceptibility loci (106). It could be that when DNA is coiled up in its 3D structure, that regions of the genome that are not next to each other linearly, come together and physically interact in 3D space.

In this chapter, the Capture Hi-C (CHi-C) procedure (106) has been used to identify physical interactions between known breast cancer susceptibility loci (bait), and other loci across the genome (target). The analysis aims to identify physical interactions, which could potentially explain the underlying biological mechanisms of how the susceptibility loci effect breast cancer risk.

6.2 Capture Hi-C

CHi-C is a chromatin procedure used to test for physical chromatin interactions between a capture region, this being a pre-specified genomic location where a locus of interest maps to, and an unrestricted area of the genome. It has been hypothesised that a number of disease susceptibility variants mapping to non-coding regions of the genome, or gene-deserts, could be physically interacting with other loci to have an effect on disease risk. The CHi-C procedure enables this hypothesis to be explored. CHi-C is just one example of a Chromosome Conformation Capture (3C) based method, which can be used to test whether such physical interactions exist. There are various 3C-based methods used to test for said interactions, including Circularized Chromosome Conformation Capture (4C), Chromosome Conformation Capture Carbon Copy (5C), Hi-C and CHi-C, with the type of interactions tested at any one time differing between them (Figure 6-1) (155). 3C is an approach used to test for interactions between a single pair of loci (one-by-one approach). 4C is used to test for interactions between a single locus and multiple other loci (one-by-all approach) and 5C is used to test for interactions between many loci and their targets, but both within specific regions (many-by-many approach). Hi-C is used to test for interactions between any loci across the genome, known as an all-by-all approach. Capture Hi-C is an extension of the Hi-C method but differs in that a many-by-all approach is used and that the resolution of the analysis is improved, which allows for the analysis of GWAS risk loci (106).

To form CHi-C libraries using specific cell-lines, the first step is to covalently cross-link DNA-DNA using formaldehyde (Figure 6-2) (156). This formulates chromatin crosslinking which fixes the cells so that the points where the loci are physically interacting in the 3D structure are fixed together. After this a restriction enzyme, such as HindIII, is used to cut the fixed chromatin into pieces (fragments). The ends of the chromatin pieces are then joined together by DNA ligation and purified to remove crosslinks (reverse crosslink). A label is added to each purified pair, with these pairs being known as di-tags. The DNA itself is also sheared into fragments using the same restriction enzyme. The di-tags where at least one end maps to the capture region are retained for analysis. There are two ends to every di-tag, one end is the bait fragment, and the other end is the target fragment. Depending on the analysis being carried out, usually based on where the target fragment maps to, a number of di-tags will be further excluded. The analysis itself involves testing for significant physical interactions between a capture region fragment and a fragment mapping to another region of the genome. Interactions can be classified as being either cis-interactions or trans-interactions. Cis-interactions are those where the two fragments forming a di-tag map to the same chromosome, whereas trans-interactions are two fragments that map to different chromosomes.

The formation of the CHi-C libraries analysed in this chapter, as explained in this section, were conducted by Dr. Fletcher and her team at the ICR.

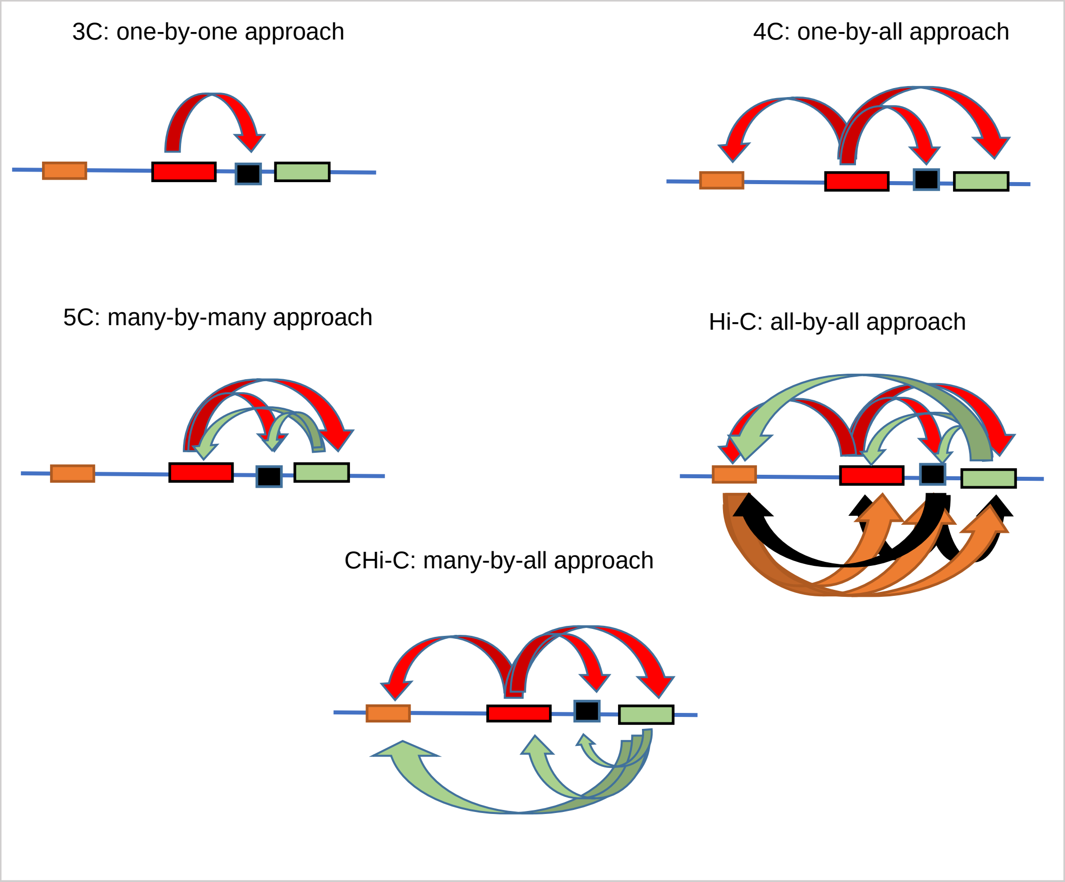

3C: one-by-one approach
4C: one-by-all approach
5C: many-by-many approach
Hi-C: all-by-all approach
CHi-C: many-by-all approach

Figure 6-1: Different 3C-based method interaction approaches

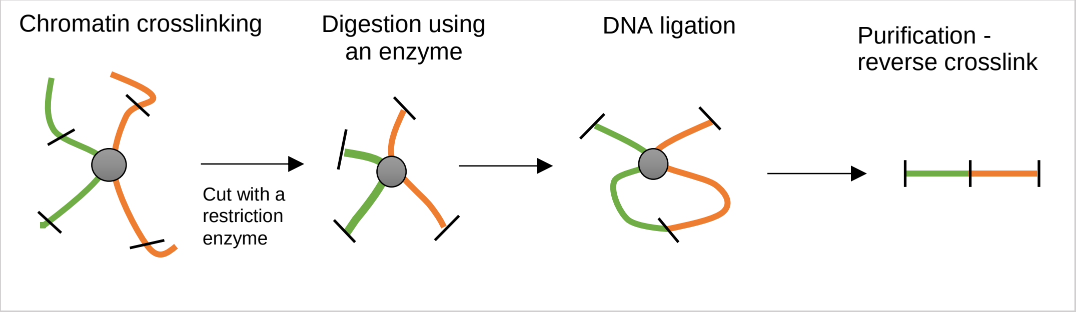

Chromatin crosslinking
Digestion using an enzyme
DNA ligation
Purification - reverse crosslink
Cut with a restriction enzyme

Figure 6-2: Crosslinking, digestion and ligations steps

6.2.1 Previous Literature

CHi-C is a method that has been used to identify significant physical interactions between established disease susceptibility risk loci and other, seemingly unrelated, regions of the genome for diseases such as breast cancer (106), colorectal cancer (157) and numerous autoimmune diseases (158).

Dryden et al. (106) have used the CHi-C procedure to test whether long-range physical interactions exist for three breast cancer loci (2q35, 8q24.21 and 9q31.2), that each map to gene deserts, with other loci in the genome. All three loci had been previously shown to be associated with ER-positive breast cancer, but had not been found to be strongly associated with ER-negative breast cancer risk. Three control loci were also included in the analysis, these being randomly selected gene-poor regions of a similar size to the breast cancer loci, but with no known association with breast cancer risk. Three different cell-lines were used to conduct the analysis, two of which were breast cancer cell-lines (BT483 and SUM44) and the other a non-breast cancer cell-line (GM06990), which was set as the control. For each cell-line, two biological replicates were generated, these being two different libraries produced for the same cell-line. Di-tags were generated for each biological replicate, for each cell-line. Dryden et al. examined whether significant physical interactions existed between fragments mapping to one of the capture loci and fragments mapping to another locus, this being either within the capture region (capture-to-capture interactions), but not interacting with itself, or 5Mb either side of the capture region (bait-to-5Mb interactions). For each cell line, Dryden et al. tested whether physical interactions between fragments occurred more often than expected by chance alone using the negative binomial regression method. This method first involved filtering out interactions deemed to be noise for each cell-line and biological replicate, using a truncated negative binomial distribution and the trans-chromosomal interaction counts for each bait fragment. The trans-chromosomal interaction counts for a bait fragment were the number of times the bait fragment interacted with another locus on a different chromosome. Then for each cell-line a negative binomial regression model was fitted to what was considered to be genuine signal. Significant bait-to-5Mb and capture-to-capture interactions were identified, with some target ends shown to map to protein-coding genes.

Using the same CHi-C method and analysis as described by Dryden et al (106), Martin et al. (158) examined whether significant physical interactions existed between fragments mapping to susceptibility loci for four autoimmune diseases: Rheumatoid arthritis, type 1 diabetes, psoriatic arthritis and juvenile idiopathic arthritis, and other regions of the genome. The two cell-lines, human B (GM12878) and T (Jurkat), were used to conduct the analysis as these were the most relevant cell-lines for these four diseases. Using the negative binomial regression method, significant bait-to-5Mb physical interactions and capture-to-capture physical interactions were tested for. Martin et al. identified many significant physical interactions and found that for a number of these interactions, the target end mapped to candidate genes or to the other autoimmune disease. The authors concluded that future work should be carried out in order to characterise the functionality of the identified interactions.

Jager et al. (157) have also used the CHi-C procedure to examine for significant physical cis and trans-interactions for 14 susceptibility colorectal cancer loci in three different colorectal cancer cell-lines (LS174T, LoVo and Colo205). The authors took a different approach and instead used the continuous Weibull distribution to perform their analysis, instead of the negative binomial distribution. Using this approach, Jager et al. identified a mixture of significant physical cis and trans-interactions for many of the loci.

6.3 Capture Hi-C data

The CHi-C data analysed consisted of loci that has previously been identified, in publications up until the year 2013, as being associated with overall breast cancer risk, and/or ER-positive breast cancer risk and/or ER-negative breast cancer risk. The loci analysed had either been discovered by Michailidou et al.(45), or in previous published studies. Not all published loci identified by the year 2013 were analysed, as some loci were not included in the analysis because of sequencing problems. 63 associated breast cancer loci were analysed, the remaining loci were based on six random SNPs and three random regions of the genome, 50kb, 100kb and 500kb in length to act as controls. These control loci/regions had not been identified as being associated with breast cancer risk.

CHi-C libraries were generated from seven cell-lines, two of these being estrogen receptor (ER)-positive breast cancer cell-lines (T47D and ZR751), two ER-negative breast cancer cell-lines (BT20 and MDAMB231), a normal breast epithelial cell-line (Bre80) and two control non-breast cancer cell-lines: a liver cancer cell-line (HepG2) and a lymphoblastoid cell-line (GM06990). With a number of the published loci analysed shown to be strongly only associated with either ER-positive breast cancer or ER-negative breast cancer, ER-positive and ER-negative cell-lines were used to generate some of the CHi-C libraries. When generating the libraries two biological replicates were produced for each cell-line, which meant that in total 14 CHi-C libraries were created. Each biological replicate for each cell-line were sequenced, and up to 71 million di-tags, with both ends uniquely mapping to the human reference genome, were generated. The locus of interest is defined as the capture region, and fragments mapping within this region are known as bait fragments. The fragments that these bait fragments pair with, are the target fragments. Together these paired fragments form a di-tag, with one end of the di-tag mapping to the capture region and the other end mapping to the target region. The fragments were created by partitioning regions of the genome into many sections, these sections being the same length as the enzyme used to split up the region (HindIII). These di-tags, along with the formed libraries, were generated by Dr. Olivia Fletcher and her team from the ICR.

Some of the fragments combined to form a di-tag, may not actually physically interact with each other, so their interaction count would be zero. For other fragments pairs, the two fragments do physically interact a number of times, meaning their interaction count would be greater than zero. It is however possible for bait and target fragments to physically interact by chance, not necessarily for biological reasons. Therefore, the objective of the analysis was to only acknowledge di-tags where the number of times the two fragments physically interact, is greater than would be expected by chance alone. These physical interactions would likely signify an interaction of biological importance.

Locus	SNPs*	ER status**	Locus no.
22q12.1	rs17879961, rs132390	positive	1
22q13.1	rs6001930	both	2
21q21.1	rs2823093	positive	3
21q21.2	rs200691	neither	4
20q13.13	rs6125607	neither	5
19p13.1	rs8170	negative	6
19p13.11	rs4808801	positive	8
19q13.31	rs3760982	both	9
18q11.2	rs527616	positive	10
18q11.2	rs1436904	positive	11
17q22	rs6504950	positive	12
16q12.1	rs3803662	positive	13
16q12.2	rs17817449	both	14
16q23.2	rs13329835	both	16
14q13.3	rs2236007	both	17
14q24.1	rs2588809	positive	18
14q24.1	rs999737	both	19
14q32.11	rs941764	both	20
13q13.1	rs11571833	negative	21
12p13.1	rs12422552	both	22
12p11.22	rs10771399	both	23
12q22	rs17356907	both	24
12q24.21	rs1292011	positive	25
11p15.5	rs3817198	positive	26
11q13.1	rs3903072	both	27
11q13.3	rs554219, rs78540526	positive	28
11q24.3	rs11820646	both	30
10p15.1	rs2380205	both	31
10p12.31	rs11814448, rs7072776	positive	32
10q21.2	rs10995190	both	33
10q22.3	rs704010	both	34
10q23.1	rs7071985	neither	35
10q25.2	rs7904519	negative	36
10q26.13	rs2981579	positive	38
9p21.3	rs1011970	both	39
9q31.2	rs10759243	positive	40
9q31.2	rs865686	positive	41
8p12	rs9693444	positive	42
8q21.11	rs6472903	both	43
8q21.11	rs2943559	both	44
8q24.21	rs13281615	both	45

Locus	SNPs*	ER status**	Locus no.
8q24.21	rs11780156	both	46
7q35	rs720475	both	47
6p25.3	rs11242675	both	48
6p23	rs204247	positive	49
6q14.1	rs17529111	negative	50
6q22.31	rs1337863	neither	51
6q25.1	rs12662670, rs2046210	negative	52
5p15.33	rs10069690	negative	53
5p12	rs10941679	positive	55
5q11.2	rs889312	positive	56
5q11.2	rs10472076, rs1353747	both	57
5q33.3	rs1432679	both	58
4q24	rs9790517	positive	59
4q34.1	rs6828523	positive	60
3p26.1	rs6762644	both	61
3p24.1	rs4973768	positive	62
3p24.1	rs12493607	positive	63
2p24.1	rs12710696	both	64
2q14.2	rs4849887	negative	65
2q31.1	rs2016394	positive	66
2q31.2	rs1550623	positive	67
2q33.1	rs1045485	neither	68
2q35	rs13387042	positive	69
2q35	rs16857609	both	70
1p36.22	rs616488	negative	71
1p31.1	rs66916276	neither	72
1p13.2	rs11552449	both	73
1p11.2	rs11249433	positive	74
2p25.1	500 kb***	neither	77
5q31.2	100 kb***	neither	78
1p13.3	50 kb***	neither	79

* Breast cancer associated SNPs mapping to the locus

** ER status for associated breast cancer
*** random region

Table 6-1: Loci used for CHi-C analyses

6.4 Methods

The CHi-C procedure was used to examine for significant long-range physical interactions at 72 loci, which included 63 breast cancer susceptibility loci, six random SNPs and three random genome regions. Two separate analyses were performed on all seven cell-lines separately. This was the first time that the CHi-C procedure had been used to simultaneously analyse such a large number of capture regions, especially in breast cancer where only three capture regions had been previously assessed. Dryden et al. had previously focussed their analysis on six capture loci, three of these being strongly associated with ER-positive breast cancer, with the three remaining loci acting as control loci. In this CHi-C analysis, 63 breast cancer susceptibility loci and nine control loci were analysed using seven cell-lines. The analysis was also conducted using a larger number of cell-lines as seven cell-lines had been analysed, which is over double the number used by Dryden et al.

The first analysis involved testing for significant physical bait-to-5Mb interactions, these being defined as interactions where one end of the di-tag mapped to one of 72 capture loci, and the other mapping to an area within 5 Mb of the capture region ("bait-to-5Mb" analysis). For the second analysis, it was tested whether significant physical interactions existed between fragments which both map within the capture region ("capture-to-capture" analysis). A number of the target capture loci mapped to regions of the genome that overlapped with other capture loci **(Appendix 9: Table 1)**. This meant that some of the target fragments for one locus were found to map within another capture locus. Therefore, technically the interaction was not just a capture-to-capture interaction, it was also an interaction between a bait fragment and a target fragment of a different locus. For this reason, significant physical interactions were removed from the capture-to-capture analysis results if it was found that a target fragment mapped to an overlapping capture region, and were instead included in the bait-to-5Mb analysis results.

The method used and developed by Dryden et al. (106) was used to perform both the bait-to-5Mb and capture-to-capture analyses. In this thesis, this method shall be referred to as the negative binomial regression method. Significant physical interactions found to occur across most cell-lines, when using this method, were investigated further using Ensembl (159). The genome browser was used to examine whether any genes of biological significant mapped within the target end of the recurring significant physical interactions. To further assess the plausibility of the observed significant physical interactions, a second method, CHiCAGO (Capture Hi-C Analysis of Genomic Organisation)(160), was also used to test for significant physical interactions. At the time of conducting the analysis, CHiCAGO was a newly developed CHi-C method. CHiCAGO was used to perform a bait-to-5Mb analysis, using the same CHi-C libraries as those used when performing the analysis with the negative binomial regression method. CHiCAGO, however, was not used to examine for capture-to-capture physical interactions as it was unable to adjust for the bias that materialises when conducting this type of interaction analysis.

6.4.1 Brownian and technical noise

The overall aim of the analyses performed in this chapter was to establish significant physical interactions between bait and target fragments. Occasionally two fragments will physically interact randomly, with there being no biological reasoning behind the interaction. These physical interactions are noise, and do not reflect true signal. There are two main sources of noise that need to be considered where performing a CHi-C analysis, Brownian noise and technical noise (106, 160). Brownian noise is the noise attributed to fragment pairs that physically interact randomly, by chance. These interactions are dependent on distance, with the physical interaction count increasing as the distance between fragment pairs decreases. Technical noise on the other hand is not dependent on distance, it is noise that is made up of fragment pairs that have interacted due to experimental bias, with this including bias that has resulted from sequencing errors. Both forms of noise should be accounted for when testing for significant interactions, in order for a meaningful conclusion to be made. Both the negative binomial regression method and CHiCAGO have taken these noise components into consideration, but have done so differently.

6.4.2 Negative binomial regression method

The negative binomial regression method was the main method used to conduct both bait-to-5Mb and capture-to-capture interaction analyses conducted in this thesis. Di-tags, where one end mapped to a capture region fragment and the other to a non-capture region fragment within 5Mb (bait-to-5Mb), were analysed separately and differently to the di-tags when both ends of the di-tag mapped to the capture region (capture-to-capture). These two analyses were performed separately for all seven cell-lines.

6.4.2.1 Significant bait-to-5Mb interactions

Once di-tags were established and libraries generated for each cell-line by Dr. Fletcher (161), the first step was to separate the fragments pairs considered to be noise from the true signal. Dryden et al. (106) have previously separated noise from real signal by assessing the interactability of each fragment analysed. They deemed interactability as the tendency a fragment has of interacting with other fragments. The interactability of a bait fragment can be measured by counting the number of trans-chromosomal interactions that a bait fragment has, this being the number of physical interactions it has with other fragments that map to a different chromosome. It is assumed that these collisions represent random interactions, with it being expected that across all captured fragments, the counts are similar if bias is not present (106). The counts for each captured fragment (bait) were used to classify each fragments interactability as either low or high, with low counts suggesting that the bait fragment represents stochastic noise, and high counts suggesting genuine signal. These two components, stochastic noise and genuine signal, together form a bimodal distribution. A distribution that can be used to model count data is needed to model the trans-chromosomal interaction counts. The Poisson and the negative binomial distribution are two distributions that can be used to model count data. The Poisson distribution would be a suitable distribution to model count data if the mean and variance are equal, but when modelling on the trans-chromosomal interaction counts, this cannot be assumed to be true as the variance tends to be larger than the mean. The negative binomial distribution on the other hand is a lot more flexible, and does not assume that the mean and variance are equal. This therefore makes it an ideal distribution to be used when data is overdispersed.

Individually, for each cell-line and biological replicate, a truncated negative binomial model was fitted to what was thought to be genuine signal based on the trans-chromosomal counts, this being the second component of the bimodal distribution. A histogram of the trans-chromosomal counts for di-tags present on one cell-line, for one biological replicate was used to help decide what threshold should be used to separate noise from genuine signal. An example of such a histogram can be observed in Figure 6-3, with the plot being fairly similar to the plots produced for each biological replicate, for each cell-line. From Figure 6-3, it is apparent that the counts form a bimodal distribution, in which there are two components. A peak in frequency can be observed when the trans-interaction count for a fragment is small, this is considered as the noise component of the bimodal distribution. The histograms were used to help set the truncation point for the truncated negative binomial for the second component of the bimodal distribution. The truncation point was used to filter out any bait fragments that have a trans-chromosomal count in the lowest 5% of the negative binomial distribution. Fragments with trans-chromosomal counts under this threshold were excluded from any further analyses, as these were regarded as noise. In the example presented in Figure 6-3, the truncation point was set to 4,000, with the threshold then fixed to 2,585.

This meant that bait fragments with a trans-interaction count less than 2,585, were excluded from the analysis.

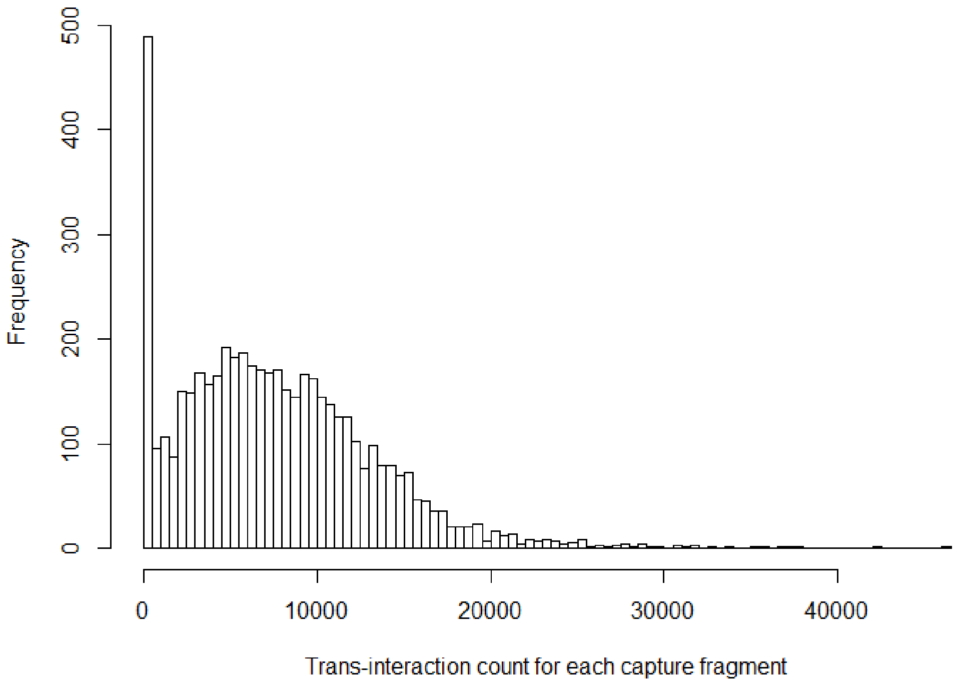


Figure 6-3: Histogram of trans-interaction counts

For each cell-line, the di-tags that had not been excluded from each of the two replicates were then analysed together. This meant that in total, seven different bait-to-5Mb interaction analyses were conducted. The filtered di-tags for each cell-line were split into 1 percentile bins, these being based on distance, in order to smooth the data. The “glm.nb” function in R was then used to fit the negative binomial regression model to the physical interaction counts for each fragment pair in the filtered dataset, for each bin. Experimental bias is corrected for by including the natural log of the trans-chromosomal counts for the two biological replicates separately as a covariate in the model. To also correct for the distance between interacting fragments, the natural log of distance (distance between the mid-points of the two fragments that form a fragment pair) was also included as a covariate in the model.

The observed interaction counts were then compared to those under the negative binomial regression model in order to obtain *p*-values for each fragment pair. The FDR was then used to adjust the *p*-values to account for multiple testing, with an FDR < 1% being used to signify a significant physical interaction.

6.4.2.1 Significant capture-to-capture interactions

A slightly different approach was used to test for significant capture-to-capture physical interactions, than the one used for the bait-to-5Mb analysis. With two ends of each capture-to-capture di-tag mapping to the capture region, the interactability of both ends of the di-tag were examined. For each cell-line, a histogram of the trans-interactions counts for both the bait and target fragments were assessed, for each biological replicate. Similarly to the bait-to-5Mb analysis, these histograms were used to set the truncation point for the truncated negative binomial, in order to set the threshold used to filter out the bait and target fragments with a trans-interaction count in the lowest 5% of the negative binomial distribution.

For each cell-line, the di-tags that have not been excluded from each of the two replicates were then analysed together, with the filtered data for each cell-line also being split into 1 percentile bins based on distance. A negative binomial regression model was then fitted to each bin, with experimental bias being corrected for by including the product of the natural log of the trans-interaction counts for both ends of each di-tag as covariates in the model, for the two biological replicates separately. Distance was also corrected for by taking the natural log of the distance between interacting fragments and including it as a covariate in the model. Similarly to the bait-to-5Mb analysis, the FDR was then used to adjust the *p*-values to account for multiple testing, with an FDR < 1% being used to signify a significant physical interaction.

6.4.3 CHiCAGO

After conducting the CHi-C analysis using the negative binomial regression method, CHiCAGO, an R package developed by Cains et al.(160), was used to also analyse the CHi-C. The package was used to examine whether significant bait-to-5Mb interactions detected using the negative binomial regression method, were also identified when using a different method. CHiCAGO can be used to detect significant bait-to-5Mb physical interactions, not capture-to-capture interactions, as it does not adjust for the interactability of both fragment ends. This meant that only the bait-to-5Mb results were compared.

6.4.3.1 Significant bait-to-5Mb interactions

CHiCAGO uses two count distributions, the Poisson and the negative binomial distribution, to create a two-component model in order to model the interaction count distribution (160).

Let X_{bt} be the number of physical interaction counts between target end t and bait end b.

Under the null hypothesis, it can be assumed that (160):

$$X_{bt} = B_{bt} + T_{bt}$$

Where, X_{bt} consists of two components, Brownian noise (B_{bt}) and technical noise (T_{bt}).

Brownian and technical bias are treated as two different entities, with the Poisson distribution used to model technical noise (T_{bt}) and the negative binomial distribution used to model Brownian noise B_{bt}, such that (160):

$$T_{bt} \sim Pois(\lambda_{bt})$$

$$B_{bt} \sim NB(\mu_{bt}, r)$$

where, λ_{bt} is the mean trans-chromosomal interaction count between each bait-target fragment pair consisting of a bait fragment b and a target fragment t. Each fragment is binned according to their trans-chromosomal interaction count, and λ_{bt} is estimated as the mean trans-chromosomal interaction count across two bins, one of which contains bait fragment b and the other that contains target fragment t. The dispersion parameter of the negative binomial distribution is r, with this being estimated by finding the r that maximises the likelihood of the negative binomial regression model.

Cairns et al.(160) define μ_{bt} as (160):

$$\mu_{bt} = s_b s_t f(d_{bt})$$

with, s_b being the bait fragment specific bias and s_t the target fragment specific bias. $f(d_{bt})$ represents the frequency of bait-target interactions over distance d_{bt}, which is the distance between the midpoint of bait fragment b and target fragment t, and is dependent on the distance between the bait fragment and the target fragment.

To estimate μ_{bt}, $f(d_{bt})$ is first estimated, then s_b and then s_t is finally estimated.

The distance from the centre of a given bait fragment b, is split up into 20kb bins. The average interaction count for target end fragments falling within individual 20kb bins, $\bar{X}_{bin_b}$, is calculated, whilst ignoring bait-to-bait fragment pairs and fragment pairs where the interaction count is equal to zero.

The geometric mean count over all bins at distance d_{bin} is then used to estimate $f(d_{bin})$, which can then be used to estimate $f(d_{bt})$ by fitting a cubic function on a log-log scale, and then extrapolating beyond distance d_{bin}.

Cairns et al.(160) estimate s_b using both $\bar{X}_{bin_b}$ and $f(d_{bin})$:

$$\hat{s}_b = median_{bin} \frac{\bar{X}_{bin_b}}{\hat{f}(d_{bin})}$$

Target end fragments are pooled together based on the number of non-zero trans-chromosomal interactions the target fragments are involved in. CHiCAGO then assumes that the target ends in each pooled group have the same target bias, s_t, with s_t then being estimated by taking the median bait and target fragment interaction count across the pooled bins.

When working with more than one biological replicate, *k*, the X_{bt} for each replicate, X_{btk}, is calculated. The overall X_{bt} is then calculated by taking the nearest integer of the following equation (160):

$$X_{bt} = \frac{\sum_k s_k X_{btk}}{\sum_k s_k}$$

Where, s_k is a sample-specific scaling factor, such that (160):

$$s_k = median_b(\frac{M_{bk}}{G_b})$$

With, M_{bk} being the number of fragments present within 1.5mb of each bait fragment, divided by the number of other ends present within 1.5Mb of the bait fragment. The geometric mean of M_{bk} across the replicates is then taken to estimate G_b.

Using the Delaporte model, it is then tested whether the observed interaction counts are greater than those expected under the model. The *p*-values obtained are then weighted in order to adjust for multiple testing, as well as the tendency for interactions to occur more when fragments are closer together than when they are further apart. Many more long-range interactions will be tested than shorter range interactions, which would cause there to be many type-1 errors amongst the long-range interactions (160). Therefore, CHiCAGO assigns and adjusts each *p*-value by a weight, which is allocated to a pair of fragments based on how likely it is that the pair of fragments will interact, with this being established by the distance between two fragments. A larger weight is given to fragment pairs that are closer together, with the maximum weight being assigned when the distance between the bait and target fragment is zero. The *p*-values are divided by their weights to construct weighted *p*-values, which means that the *p*-values for close fragment pairs will get smaller.

These *p*-values are then converted into log-transformed scores where:

$$s_{tb} = max(0, -\log Q_{tb} - \log w_{max})$$

where, Q_{tb} is the weighted *p*-value and w_{max} is the maximum weight values, this being the value of the weight when the distance between the bait and target fragment is zero.

Cairns et al.(160) suggest that a score greater than 5 should be used to indicate that a physical interaction is significant.

6.5 Negative binomial regression method analyses

The Negative binomial regression method was used to test for physical bait-to-5Mb interactions and bait-to-bait interactions. The significant interactions detected using the negative binomial regression method have been included in a journal paper that is currently under review (161).

6.5.1 Bait-to-5Mb interaction analysis

The analysis conducted in this section aimed to identify significant physical interactions between individual HindIII bait fragments that mapped to 1 of 72 capture regions, and individual HindIII target fragments mapping within 5 Mb of the corresponding capture region, using the negative binomial regression, in all seven cell-lines.

In total, 50 of the 72 capture loci had at least one bait HindIII fragment that significantly interacted with a target HindIII fragment within 5Mb of the capture region, in at least one cell-line (FDR<0.01) (Table 6-2). For some of the loci, no significant interactions were observed in any of the cell-lines. These loci were not included in Table 6-2. In all seven cell-lines, bait fragments mapping to 14q24.1 (locus 18), 11q13.3 (locus 28), 10p12.31 (locus 32), 3p26.1 (locus 61), 2q31.1 (locus 66), 2q35 (locus 70), and the random 500 kb region were shown to significantly interact (FDR < 0.01) with fragments mapping to a target region within 5Mb of the bait fragment. The loci 14q24.1, 11q13.3 and 2q31.1 have all been shown to be associated with ER-positive breast cancer risk, and both 3p26.1and 2q35 have been shown to be associated with both ER-positive and ER-negative breast cancer. None of the bait fragments mapping to capture loci associated with ER-negative breast cancer, were found to have significant interactions across all cell-lines. Fragments mapping to two loci associated with ER-negative breast cancer, 19q13.1 (locus 6) and 5p15.33 (locus 53), were shown to only significantly interact with target fragments in the ER-negative cell-lines.

Focus then moved onto establishing whether any of the target end HindIII fragments were frequently interacted with. For a number of the significant bait-to-5Mb interactions, the target end of the interacting pair mapped to the same HindIII fragment in all seven cell-lines (Table 6-3). Bait fragments mapping to the capture locus 14q24.1 (locus 18) significantly interacted with two consecutive HindIII fragments targets (69,255,090-69,263,908 bp), and the single HindIII fragment (69,280,294-69,288,644 bp), in all seven cell-lines. In six cell-lines, bait fragments mapping to this capture locus were found to also significantly interact with two HindIII fragments that map either side of the target HindIII fragment, 69,280,294-69,288,644 bp (69,276,282-69,280,293 bp and 69,288,645-69,296,090 bp). In at least six cell-lines, fragment mapping to the capture locus 11q13.3 (locus 28) significantly interacted with six consecutive HindIII target end fragments (68,843,286-68,882,444 bp), as well as four consecutive HindIII fragments (68,886,551-68,910,597 bp). Also, in at least six cell-lines, bait fragments mapping to 10p12.31 (locus 32) were found to significantly interact with two consecutive HindIII fragments (23,274,447-23,280,039 bp). A bait fragment within the capture locus 3p26.1 (locus 61) was observed to significantly interact with three consecutive target end HindIII fragments in all seven cell-lines (5,086,339-5,113,690 bp), with the same being shown for the capture locus 2q35 (locus 70) (217,552,337-217,565,782 bp). Finally, in all seven cell-lines bait fragments mapping to the 500kb region were found to significantly interact with the target end fragment, 11,272,832-11,276,344 bp. 2q35 bait fragments were also shown to significant interact with a target fragment mapping to 11,272,095-11,272,831 bp, in six of the cell-lines.

Concentrating on the significant interactions that were present in at least six cell-lines, Ensembl (159), the genome browser, was used to examine whether any known breast cancer or cancer related genes reside within the same region as the target ends. The gene, *IGFBP5*, has been shown to have an important role in breast cancer and maps to 217,552,337-217,565,782 bp. This gene has been shown to have a role in breast cancer metastasis, but the exact role of *IGFBP5* is not fully understood (162). Target fragments (69,255,090-69,263,908 bp) were found to map to *ZFP36L1*, a protein coding gene that has fairly recently been linked to breast cancer (163). The protein coding genes *ARMC3* and *TPCN2* are positioned within the target fragments mapping to 23,274,447-23,280,039 bp and 68,843,286-68,882,444 bp, respectively. *ARMC3* functions include metastasis and tumour initiation (164), and *TPCN2* is an ion transport gene that contains SNPs that have been shown to be associated with pigmentation traits (165).

With some of the significant physical interactions observed across most of the cell-lines and target ends mapping to protein coding genes, the interactions would seem plausible.

			Cell-lines						
			ER+ve	ER+ve	ER-ve	ER-ve	Lymph	LC	Normal
Locus	Locus no.	ER status*	T47D	ZR751	BT20	MDA	GM0	HepG2	Bre80
22q12.1	1	positive	0	1	0	0	0	0	5
21q21.2	4	neither	2	4	0	0	0	0	0
20q13.13	5	neither	2	5	0	43	0	2	265
19p13.1	6	negative	0	0	76	0	0	0	0
19p13.11	8	positive	0	0	42	14	3	16	5
19q13.31	9	both	0	0	0	0	2	0	1
18q11.2	11	positive	0	24	0	0	0	0	3
17q22	12	positive	231	30	0	1	1	0	63
16q12.2	14	both	298	229	11	18	0	0	10
16q23.2	16	both	80	7	0	0	0	0	0
14q13.3	17	both	50	61	0	0	0	0	64
14q24.1	18	positive	240	62	19	33	283	14	51
13q13.1	21	negative	14	0	32	0	17	1	2
12p13.1	22	both	0	8	0	4	11	0	23
12q24.21	25	positive	3	1	3	0	0	6	8
11p15.5	26	positive	1	0	183	16	0	21	0
11q13.1	27	both	0	274	0	20	0	0	0
11q13.3	28	positive	319	26	81	342	43	69	91
10p12.31	32	positive	6	10	10	50	41	56	62
10q22.3	34	both	7	4	0	32	0	1	9
10q26.13	38	positive	15	14	0	0	0	0	14
9p21.3	39	both	0	15	0	NA	11	0	68
9q31.2	40	positive	25	0	0	6	0	0	0
9q31.2	41	positive	24	31	0	0	0	0	11
8p12	42	positive	0	0	0	13	0	0	38
8q21.11	44	both	114	1,728	0	0	33	0	10
8q24.21	45	both	1,007	4	0	5	85	0	4
8q24.21	46	both	17	48	0	103	6	0	75
6p25.3	48	both	4	0	0	0	0	0	2
6p23	49	positive	0	0	0	8	0	0	0
6q22.31	51	neither	3	0	0	0	0	0	3
6q25.1	52	negative	0	9	0	0	1	0	0
5p15.33	53	negative	0	0	13	1	0	0	0
5p12	55	positive	0	24	0	0	0	0	0
5q11.2	56	positive	101	47	0	8	1	2	53
5q11.2	57	both	0	1	0	0	0	0	7
5q33.3	58	both	20	31	0	0	0	0	0
4q24	59	positive	0	10	0	0	122	0	0
3p26.1	61	both	497	193	147	161	140	5	179
3p24.1	62	positive	80	1	0	0	0	0	2
3p24.1	63	positive	161	0	0	0	0	0	19
2p24.1	64	both	13	3	0	0	0	0	0

Locus	Locus no.	ER status*	T47D	ZR751	BT20	MDA	GM0	HepG2	Bre80
2q31.1	66	positive	33	24	238	5	16	58	56
2q35	69	positive	4	6	12	0	0	0	0
2q35	70	both	262	201	83	9	10	65	19
1p36.22	71	negative	1	0	1	0	1	0	1
1p31.1	72	neither	47	65	0	1	4	0	110
1p13.2	73	both	2	1	2	0	0	0	2
1p11.2	74	positive	0	307	0	0	0	0	0
500kb**	77	neither	390	80	36	186	61	1	218

* ER status for associated breast cancer
** random region
Abbreviations: Lymph = Lymphoblastoid, LC = Lung cancer, Normal = Normal breast epithelial, MDA = MDAMB231 and GM0 = GM06990

Table 6-2: No. of significant near-cis interactions (<5Mb) (FDR < 0.01) using the negative binomial regression

Locus	Locus No.	Target fragment (bp)	Cell-lines with significant interactions (FDR < 0.01)**						
14q24.1	18	69,255,090-69,257,981	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
14q24.1	18	69,257,982-69,263,908	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
14q24.1	18	69,272,186-69,275,832	Bre80	GM0	HepG2	MDA	T47D	ZR571	
14q24.1	18	69,276,282-69,280,293	Bre80	GM0	HepG2	MDA	T47D	ZR751	
14q24.1	18	69,280,294-69,288,644	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
14q24.1	18	69,288,645-69,296,090	Bre80	BT20	GM0	MDA	T47D	ZR751	
11q13.3	28	68,843,286-68,856,786	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.3	28	68,856,787-68,858,830	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.3	28	68,858,831-69,961,119	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.3	28	68,861,120-68,873,431	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
11q13.3	28	68,873,432-68,882,444	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
11q13.3	28	68,886,551-68,886,942	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
11q13.3	28	68,886,943-68,891,662	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
11q13.3	28	68,891,663-68,903,868	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.3	28	68,903,869-68,910,597	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.3	28	69,060,151-69,065,192	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.3	28	69,065,355-69,075,253	Bre80	BT20	GM0	HepG2	MDA	T47D	
10p12.31	32	23,274,447-23,277,011	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
10p12.31	32	23,277,012-23,280,039	Bre80	BT20	GM0	MDA	T47D	ZR751	
5q11.2	56	55,563,961-55,567,802	Bre80	GM0	HepG2	MDA	T47D	ZR751	
3p26.1	61	5,025,725-5,026,309	Bre80	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,026,310-5,027,008	Bre80	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,027,009-5,028,985	Bre80	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,028,986-5,031,964	Bre80	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,044,212-5,059,881	Bre80	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,086,339-5,095,364	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
3p26.1	61	5,095,365-5,098,932	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
3p26.1	61	5,098,933-5,113,690	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
3p26.1	61	8,679,662-8,683,220	BT20	GM0	HepG2	MDA	T47D	ZR751	
3p26.1	61	8,773,633-8,780,364	Bre80	BT20	GM0	MDA	T47D	ZR751	
2q31.1	66	172,445,057-172,452,793	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q31.1	66	172,540,029-172,543,826	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q31.1	66	172,543,827-172,548,319	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q31.1	66	172,664,112-172,667,431	Bre80	BT20	HepG2	MDA	T47D	ZR751	
2q35	70	217,546,853-217,552,336	Bre80	BT20	HepG2	MDA	T47D	ZR751	
2q35	70	217,552,337-217,560,726	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
2q35	70	217,560,727-217,563,272	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
2q35	70	217,563,273-217,565,782	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
500kb*	77	11,272,095-11,272,831	Bre80	BT20	GM0	MDA	T47D	ZR751	
500kb *	77	11,272,832-11,276,344	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
500kb*	77	12,855,737-12,861,599	Bre80	BT20	GM0	MDA	T47D	ZR751	
500kb*	77	12,861,952-12,864,091	Bre80	BT20	GM0	MDA	T47D	ZR751	
500kb*	77	12,864,092-12,865,712	Bre80	BT20	GM0	MDA	T47D	ZR751	
500kb*	77	13,069,699-13,073,697	Bre80	BT20	GM0	MDA	T47D	ZR751	
500kb*	77	13,139,748-13,143,874	Bre80	BT20	GM0	MDA	T47D	ZR751	

Locus		Target fragment (bp)	Cell-lines with significant interactions (FDR < 0.01)**					
500kb*	77	15,697,232-15,701,484	Bre80	BT20	GM0	MDA	T47D	ZR751
500kb*	77	15,701,485-15,705,850	Bre80	BT20	GM0	MDA	T47D	ZR751

*random region **Significant interactions between that specific capture region and target fragment
Abbreviations: MDA = MDAMB231 and GM0 = GM06990

Table 6-3: Common locus-target end interactions for six or more cell-lines (negative binomial regression)

6.5.2 Capture-to-capture interaction analysis

After conducting the bait-to-5Mb interaction analysis, it was tested whether any significant physical interactions existed between individual HindIII bait fragments mapping to 1 of 72 capture regions, and individual HindIII target fragments mapping within the corresponding capture region bait (FDR<0.01). A number of the capture loci overlapped with other capture loci, meaning that they partly mapped to the same location in the genome. The overlapping loci are given in **Appendix 9: Table 1**. For this analysis, any bait fragments found to significantly interact with a target fragment that was positioned on more than one locus, were removed from the capture-to-capture results and included in the bait-to-5Mb results.

In total, bait fragments mapping to 41 capture loci were found to significantly interact with target fragments mapping within the capture region (FDR<0.01) (

Table 6-4). Only one locus, 22q13.1 (locus 6), was found to have bait fragments that significantly interacted with capture target fragments in all seven cell-lines. The 22q13.1 locus has been shown to be associated with both ER-positive breast cancer and ER-negative breast cancer, which is interesting as interactions have been shown in all seven cell-lines, which are a mixture of ER-positive, ER-negative and normal/non-breast cancer cell-lines. To investigate this further, it was examined whether there were any common significant capture-to-capture fragment interactions for this locus, in all seven cell-lines (Table 6-5). There was only found to be one common interaction in all seven cell-lines, this was with the target HindIII fragment positioned at 41,042,083-41,042,910 bp. At the time of writing, no genes or associated breast cancer variants had yet been mapped to this region.

Three loci were observed as having significant capture-to-capture interactions in six cell-lines, these were 11q13.1 (locus 27), 2q33.1 (locus 68) and the random 500 kb region. Fragments mapping to 11q13.1 were found to consecutively interact with seven target fragments within the region 65,533,743-65,580,061 bp. The gene, *OVOL1* is a transcription factor that maps within this target region (166). For the same capture region, fragments were found to significantly interact with multiple consecutive target fragments. Bait fragments interacted with three consecutive target fragments (65,590,364-65,607,749 bp), as well as two consecutive target fragments (65,616,145-65,627,751 bp), another three consecutive fragments (65,704,518-65,719,776 bp) and finally two consecutive fragments positioned at 65,725,868-65,754,245 bp. This capture region was shown to have the largest number of significant capture-to-capture interactions in six cell-lines. The gene, *SNX32*, maps within to the target region 65,616,145-65,627,751 bp, and the gene, *SART1,* maps to 65,725,868-65,754,245 bp. Within both of these genes are SNPs that have been shown to be associated with breast cancer risk. The SNP rs656040 (65,621,057 bp) resides in *SNX32* and has been shown to be associated with breast cancer risk, but a genome-wide association has not observed (167). The gene *SART1* has been linked to the maintenance of normal mitosis (168). For the locus 2q33.1, bait fragments within this capture region were shown to significantly interact with two consecutive target fragments (202,040,793-202,050,567 bp), as well as another two consecutive target fragments (202,067,137-202,073,354 bp) in six cell-lines. Two genes, *CFLAR* and *CASP10* map to the 202,040,793-202,050,567 bp target region, with the *CASP10* gene also mapping within the 202,067,137-202,073,354 bp target region. *CASP10* is a gene known to cause apoptosis to happen, which is linked to both the origin and the progression of cancer (169). Variants mapping on both *CASP10* and *CFLAR* have been shown to interact with variants mapping to *CASP8* (170, 171), a gene shown to be associated with breast cancer risk and apoptosis (105, 169, 172-175).

Other significant capture-to-capture interactions that were observed in six of the cell-lines, included a bait fragment mapping to 11q15.5 interacting with target fragments based at 65,657,866-65,663,289 bp, 65,664,081-65,668,293 bp and 65,669,066-65,692,121 bp. The gene, *FOSL1*, maps to both the 65,657,866-65,663,289 bp target region, and the 65,664,081-65,668,293 bp target region. This gene has been previously shown to be involved in transformation, proliferation and metastasis in many forms of cancer, and research suggests that this gene may be an important prognostic marker for breast cancer therapy (176). *DRAP1*, a protein-coding gene, maps within the 65,669,066-65,692,121 bp target region, but to my knowledge this gene has not yet been linked to cancer.

Bait fragments mapping to 2q33.1 were shown to significantly interact with fragments based at 202,015,774-202,019,501 bp and 202,020,801-202,025,048 bp, in six of the cell-lines. *CFLAR* was found to map to these target regions. Bait fragments based on the random 500 kb region were found to significantly interact with target fragments based at 11,727,964-11,733,174 bp and 11,918,735-11,921,166 bp in six of the cell-lines. No genes were found to be positioned on these regions. With the majority of the genes found to map to the target regions shown to play a role in either breast cancer or cancer, many of the significant interactions have been shown to be both meaningful and plausible. Findings are summarised in Table 6-6.

Locus	Locus no.	ER status	T47D	ZR751	Bre80	Hep	BT	GM0	MDA
22q12.1	1	positive	0	1	0	0	0	0	0
22q13.1	2	both	5	12	11	3	2	10	5
20q13.13	5	neither	0	0	0	0	52	0	0
19p13.1	6	negative	0	0	2	4	8	3	3
19p13.11	8	positive	1	0	3	2	1	0	0
14q24.1	18	positive	1	6	0	1	7	0	1
14q24.1	19	both	0	1	0	0	15	0	0
14q32.11	20	both	0	0	1	0	0	0	1
13q13.1	21	negative	0	0	1	0	1	0	0
12p13.1	22	both	0	0	0	0	0	1	0
12p11.22	23	both	0	0	15	0	0	0	1
12q22	24	both	0	0	1	0	0	0	0
11p15.5	26	positive	0	0	0	3	16	0	14
11q13.1	27	both	221	0	118	210	247	99	448
11q13.3	28	positive	0	0	0	0	21	0	2
10p12.31	32	positive	0	1	5	0	2	10	0
9p21.3	39	both	0	41	0	0	0	0	NA
9q31.2	41	positive	0	2	0	0	0	0	0
8p12	42	positive	0	0	1	0	0	0	1
8q21.11	44	both	0	13	0	0	0	0	1
8q24.21	46	both	0	0	0	0	0	0	1
7q35	47	both	0	0	3	0	0	0	0
6p25.3	48	both	0	0	1	0	0	0	1
6p23	49	positive	0	0	5	0	0	0	0
6q25.1	52	negative	35	0	0	0	0	0	0
5p15.33	53	negative	0	0	0	2	1	0	0
5p12	55	positive	0	66	21	1	1	0	0
4q24	59	positive	11	0	14	8	0	33	1
3p26.1	61	both	36	29	6	0	0	0	0
3p24.1	62	positive	0	84	51	0	0	0	0
2p24.1	64	both	2	0	0	0	0	0	1
2q14.2	65	negative	0	0	0	4	0	0	0
2q31.1	66	positive	0	0	0	0	3	1	0
2q33.1	68	neither	32	0	143	101	101	169	27
2q35	70	both	21	75	0	0	17	0	0
1p36.22	71	negative	2	0	46	0	2	0	16
1p31.1	72	neither	3	9	10	0	0	0	5
1p13.2	73	both	0	1	3	0	1	0	0
500kb**	77	neither	241	36	83	57	0	9	15
100kb**	78	neither	0	0	20	0	0	0	1
50kb**	79	neither	0	0	2	0	0	0	0

* ER status for associated breast cancer ** random region

Abbreviations: MDA = MDAMB231, Hep = HepG2, BT = BT20 and GM0 = GM06990

Table 6-4: No. of significant capture-to-capture interactions (FDR < 0.01) using the Negative binomial regression

Locus	Locus No.	Target fragment (bp)	Cell-lines with significant interactions (FDR < 0.01)**						
22q13.1	6	41,042,083-41,042,910	Bre80	BT20	GM0	HepG2	MDA	T47D	ZR751
11q13.1	27	65,533,743-65,538,089	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,538,090-65,541,319	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,541,320-65,542,524	Bre80	BT20	HepG2	MDA	MDA	T47D	
11q13.1	27	65,542,525-65,560,509	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,560,510-65,566,871	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,566,872-65,577,503	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,577,504-65,580,061	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,590,364-65,596,909	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,596,910-65,600,556	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,600,557-65,607,749	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,616,145-65,627,750	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,627,751-65,646,743	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,657,866-65,663,289	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,664,081-65,668,293	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,669,066-65,692,121	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,704,518-65,705,449	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,705,450-65,712,307	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,712,308-65,719,776	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,725,868-65,736,505	Bre80	BT20	GM0	HepG2	MDA	T47D	
11q13.1	27	65,736,506-65,754,245	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q33.1	68	202,015,774-202,019,501	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q33.1	68	202,020,801-202,025,048	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q33.1	68	202,040,793-202,047,685	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q33.1	68	202,047,686-202,050,567	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q33.1	68	202,067,137-202,068,586	Bre80	BT20	GM0	HepG2	MDA	T47D	
2q33.1	68	202,068,587-202,073,354	Bre80	BT20	GM0	HepG2	MDA	T47D	
500kb*	77	11,727,964-11,733,174	Bre80	GM0	HepG2	MDA	T47D	ZR751	
500kb*	77	11,918,735-11,921,166	Bre80	GM0	HepG2	MDA	T47D	ZR751	

*random region **Significant interactions between that specific capture region and target fragment
Abbreviations: MDA = MDAMB231 and GM0 = GM06990

Table 6-5: Common capture-to-capture interactions across cell-lines using the negative binomial regression

Capture locus	SNP	Target fragment/s (Mb)	Genes*	Gene information
11q13.3	rs3903072	65,533,743-65,580,061	*OVOL1*	Transcription factor
11q13.3	rs3903072	65,616,145-65,627,751	*SNX32*	Associated with BrCA
11q13.3	rs3903072	65,725,868-65,754,245	*SART1*	Associated with BrCA Maintenance of normal mitosis
11q15.5		65,657,866-65,663,289 65,664,081-65,668,293	*FOSL1*	Transformation, proliferation and metastasis in many types of cancers
2q33.1	rs1045485	202,040,793-202,050,567	*CFLAR/CASP10*	Variants shown to interact with CASP8 (gene linked to apoptosis and BrCa)
2q33.1	rs1045485	202,067,137-202,073,354	*CASP10*	Apoptosis- linked to origin & progression of cancer & variants shown to interact with CASP8 (gene linked to apoptosis and BrCa)
2q33.1	rs1045485	202,015,774-202,019,501 202,020,801-202,025,048	*CFLAR*	Variants shown to interact with CASP8 (gene linked to apoptosis and BrCa)

* protein coding genes mapping to the target fragment/s region

BrCa = Breast cancer

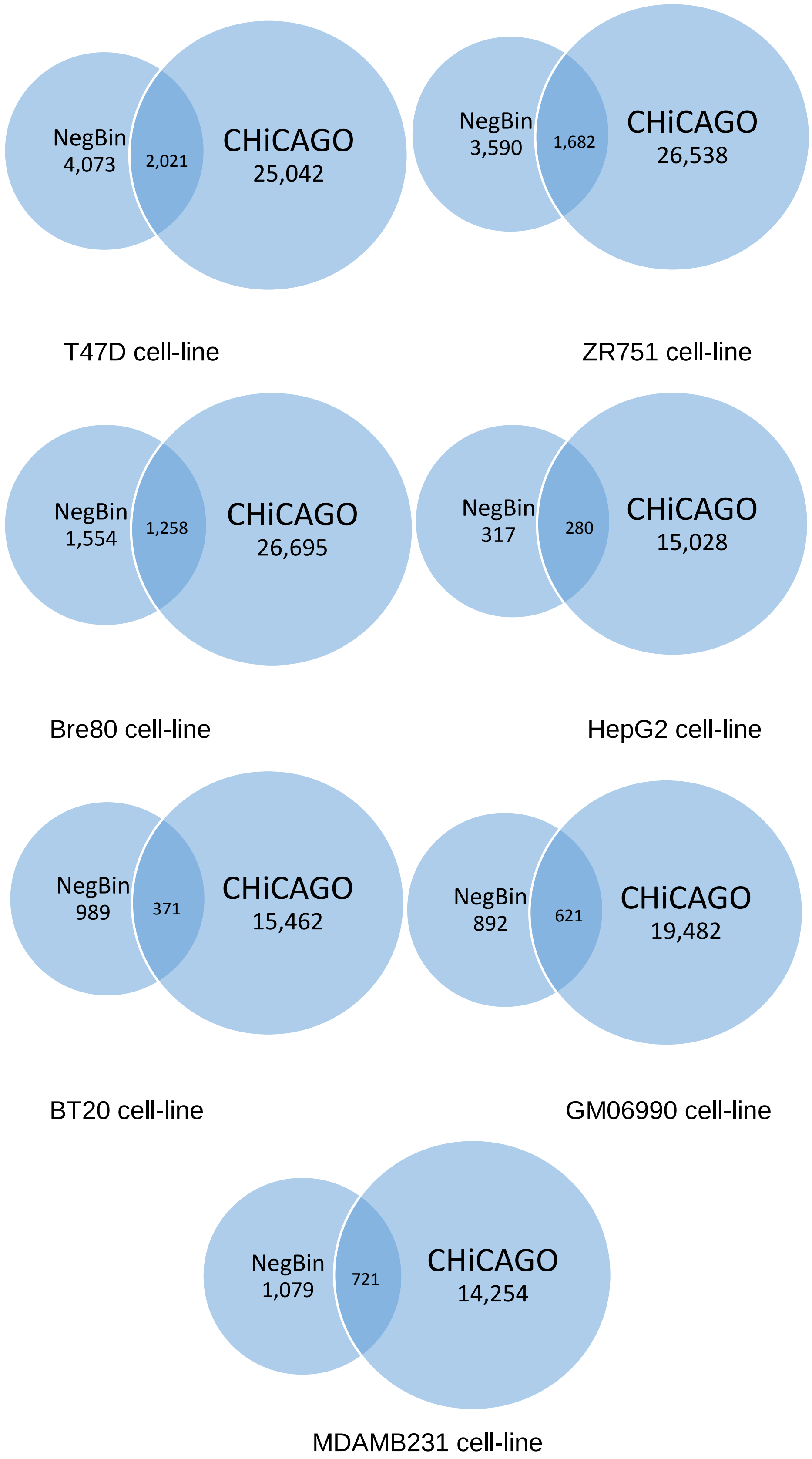
Table 6-6: Plausible significant capture-to-capture interactions

6.6 CHiCAGO analysis

6.6.1 Bait-to-5Mb interaction analysis

Once the CHi-C analysis had been conducted using the negative binomial regression method, CHiCAGO was then used to test for significant bait-to-5Mb interactions. Once the analysis was conducted, the number of significant physical interactions across all capture loci were summed for the two methods separately, for each cell-line. A larger number of significant interactions were observed in each cell-line when using CHiCAGO, compared to the number observed when using the negative binomial regression method (Figure 6-4). With the analyses conducted involving a large number of loci (72 loci), any differences in the number of significant interactions detected for each locus, by each method, would accumulate across the loci. This could explain the large difference in the number of interactions between the two methods. However, the two methods differ in how they detect significant interactions and control for type-1 errors, so it was expected that there would be a difference in the number of significant interactions detected. CHiCAGO uses two count distributions to test for significant interactions, the Poisson distribution and the negative binomial distribution, with this being known as the Delaporte distribution. An interaction was deemed significant if it occurred more often than expected under the Delaporte distribution, with *p*-values weighted based on the distance between the two “interacting” fragments. The negative binomial regression method on the other hand only uses the negative binomial distribution to test for significant interactions, and *p*-values are adjusted for multiple testing using the FDR. With the two methods controlling for type-1 errors differently, one using the FDR and the other using weights, the methods will be controlling for these errors at a different rate. Therefore, it would have been better to compare the two methods on a common scale, but the weights used to weight the CHiCAGO derived *p*-values were not given when conducting the analysis, so the original *p*-values could not be computed.

The two methods did identify a number of the same significant interactions, with 50% of the negative binomial regression interactions also being detected by CHiCAGO for the T47D cell-line, 49% for the ZR751 cell-line, 81% for the Bre80, 89% for the HepG2, 38% for the BT20 cell-line, 70% for the GM0 cell-line and 67% for the MDA cell-line. In six or more cell-lines, fragments mapping to many of the capture loci were shown to significantly interact with certain target HindIII fragments for both methods (FDR < 0.01 and score > 5) (Table 6-7). Fragments mapping to the capture locus 14q24.1 (locus 18) were observed to significantly interact with two consecutive target fragments mapping to 69,255,090-69,263,908 bp, as well as two consecutive target fragments mapping to 69,276,282-69,288,644 bp. For both methods, bait fragments mapping to 11q13.3 (locus 28) were shown to significantly interact with two consecutive target fragments mapping to 68,843,286-68,858,830 bp, another two consecutive target fragments mapping to 68,861,120-68,882,444 bp, as well three consecutive target fragments mapping to 68,886,551-68,903,868 bp. Fragments mapping to 10p12.31 (locus 32) significantly interacted with two consecutive target fragments 23,274,447-23,280,039 bp, and fragments mapping to 5q11.2 (locus 56) significantly interacted with four consecutive target fragments 5025725-5031964 bp, in both methods. Finally fragments mapping to 3p26.1 (locus 61) significantly interacted with three consecutive target fragments mapping to 5,086,339-5,113,690 bp, and fragments mapping to 2q35 (locus 70) were observed to significantly interact with four consecutive target fragments mapping to 217,546,853-217,565,782 bp. With these physical bait-to-5Mb fragment interactions being significant for both methods, and with known breast cancer or cancer genes mapping within some of the target regions, the interactions seem plausible. A summary of the most plausible bait-to-5Mb interactions are given in Table 6-8.



T47D cell-line

ZR751 cell-line

Bre80 cell-line

HepG2 cell-line

BT20 cell-line

GM06990 cell-line

MDAMB231 cell-line

Figure 6-4: Venn diagrams to compare the no. significant interactions in each cell-line when using the negative binomial regression method (NegBin) and CHiCAGO

Locus	Locus No.	Target fragment (bp)	Cell-lines with significant interactions**						
14q24.1	18	69,255,090-69,257,981	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
14q24.1	18	69,257,982-69,263,908	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
14q24.1	18	69,272,186-69,275,832	Bre	GM0	Hep	MDA	T47D	ZR751	
14q24.1	18	69,276,282-69,280,293	Bre	GM0	Hep	MDA	T47D	ZR751	
14q24.1	18	69,280,294-69,288,644	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
11q13.3	28	68,843,286-68,856,786	Bre	BT20	GM0	Hep	MDA	T47D	
11q13.3	28	68,856,787-68,858,830	Bre	BT20	GM0	Hep	MDA	T47D	
11q13.3	28	68,861,120-68,873,431	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
11q13.3	28	68,873,432-68,882,444	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
11q13.3	28	68,886,551-68,886,942	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
11q13.3	28	68,886,943-68,891,662	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
11q13.3	28	68,891,663-68,903,868	Bre	BT20	GM0	Hep	MDA	T47D	
11q13.3	28	69,060,151-69,065,192	Bre	BT20	GM0	Hep	MDA	T47D	
11q13.3	28	69,065,355-69,075,253	Bre	BT20	GM0	Hep	MDA	T47D	
10p12.31	32	23,274,447-23,277,011	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
10p12.31	32	23,277,012-23,280,039	Bre	BT20	GM0	MDA	T47D	ZR751	
5q11.2	56	55,563,961-55,567,802	Bre	GM0	Hep	MDA	T47D	ZR751	
3p26.1	61	5,025,725-5,026,309	Bre	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,026,310-5,027,008	Bre	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,027,009-5,028,985	Bre	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,028,986-5,031,964	Bre	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,044,212-5,059,881	Bre	BT20	GM0	MDA	T47D	ZR751	
3p26.1	61	5,086,339-5,095,364	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
3p26.1	61	5,095,365-5,098,932	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
3p26.1	61	5,098,933-5,113,690	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
2q31.1	66	172,540,029-172,543,826	Bre	BT20	GM0	Hep	MDA	T47D	
2q31.1	66	172,664,112-172,667,431	Bre	BT20	Hep	MDA	T47D	ZR751	
2q35	70	217,546,853-217,552,336	Bre	BT20	Hep	MDA	T47D	ZR751	
2q35	70	217,552,337-217,560,726	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
2q35	70	217,560,727-217,563,272	Bre	BT20	GM0	Hep	MDA	T47D	ZR751
2q35	70	217,563,273-217,565,782	Bre	BT20	Hep	MDA	T47D	ZR751	
500 kb*	77	11,272,832-11,276,344	Bre	BT20	GM0	Hep	MDA	T47D	

* 500 kb random region ** Negative binomial regression: FDR < 0.01 and CHiCAGO: score > 5
Abbreviations: MDA = MDAMB231, GM0 = GM06990, Bre80 = Bre and Hep = HepG2

Table 6-7: Common locus-target end interactions for six or more cell-lines (detected by both the negative binomial regression method and CHiCAGO)

Capture locus	Target fragment/s (Mb)	Genes*	Gene information
14q24.1	69,255,090-69,257,981 69,257,982-69,263,908	*ZFP36L1*	Linked to BrCa
11q13.3	68,843,286-68,856,786 68,856,787-68,858,830	*TPCN2*	Ion transport
10p12.31	23,274,447-23,277,011	*ARMC3*	Metastasis and tumour initiation
10p12.31	23,277,012-23,280,039	*ARMC3*	Metastasis and tumour initiation
2q35	217,552,337-217,560,726	*IGFBP5*	BrCa metastasis

* protein coding genes mapping to the target fragment/s region

BrCa = Breast cancer

Table 6-8: Plausible significant bait-to-5Mb interactions

6.7 Discussion

Many genetic variants have been identified as being associated with breast cancer risk, but how they influence disease risk is not quite understood as many of the variants map to non-coding regions of the genome. The CHi-C procedure can be used to establish significant physical interactions between loci that are not necessarily close on the genome, but come into contact physically when they are in their 3D state.

In this chapter a large-scale CHi-C analysis was performed using 72 loci and seven cell-lines, and to my knowledge this was the largest CHi-C analysis performed to date. CHi-C analyses tend to focus on testing for physical interactions for a small number of susceptibility loci, therefore making this analysis unique as up to 72 loci have been analysed in seven cell-lines. There was found to be evidence of both bait-to-5Mb interactions, and capture-to-capture interactions for many of the loci analysed across various cell-lines. Some of the significant physical interactions were present across all cell-lines, and a number of the target end fragments were found to map to genes known to have a role in cancer and/or breast cancer. This therefore provided evidence that it is possible that some of the susceptibility loci could be interacting with other regions of the genome to have an effect on breast cancer risk.

Three of the analysed loci have already been analysed by Dryden et al. (106) in a previous CHi-C analysis using two breast cancer cell-lines, BT483 and SUM44, and the control non-breast cancer cell-line, GM06990. The loci analysed were 2q35 (rs13387042), 8q24.21 (rs13281615) and 9q31.2 (rs865686). For 2q35, Dryden et al. identified 20 (BT483), 45 (SUM44) and zero (GM06990) significant bait-to-5Mb physical interactions (FDR < 0.01). For 8q24.21, they identified three (BT483), zero (SUM44) and 108 (GM06990) significant bait-to-5Mb interactions (FDR < 0.01). For 9q31.2, four (BT483), zero (SUM44) and zero (GM06990) significant bait-to-5Mb interaction were identified. Similarly, for the GM06990 cell-line in my analysis, zero significant interaction peaks were identified for 2q35 and the same for 9q31.2. No interactions were observed for all three loci for the HepG2 cell-line, a liver cancer cell-line and no interactions were observed for 2q35 in the Bre80 cell-line (normal breast epithelial cell-line). For 8q24.21, 85 significant interaction peaks were identified, which was fairly similar to the number of interactions found by Dryden et al. (106). In the four breast cancer cell-lines analysed in this chapter, significant interactions were detected for all three loci, but not in all four cell-lines. For 9q31.2, 24 (T47D), 31 (ZR751), zero (BT20) and zero (MDAMB231) significant bait-to-5Mb interactions (FDR < 0.01) were detected. For 8q24.21, 1,007 (T47D), four (ZR751), zero (BT20) and five (MDAMB231) significant bait-to-5Mb interactions (FDR < 0.01) were detected. For 2q35, four (T47D), six (ZR751), twelve (BT20) and five (MDAMB231) significant bait-to-5Mb interactions (FDR < 0.01) were detected. It was rather surprising how many significant interactions were detected for the 8q24.21 locus in the T47D cell-line, as the number detected was much larger than detected in other cell-line in this analysis, and in the analysis conducted by Dryden et al. It was also the largest number of interactions detected across all loci, therefore, this is an inconsistent result.

CHiCAGO, another method used to analyse CHi-C data, was also used to test for near-cis interactions. The results from this analysis were compared to those produced when using the negative binomial regression, and it was found that the number of significant interactions identified by CHiCAGO exceeded the number discovered when using the negative binomial regression. It was expected that the two methods would identify a different number of significant interactions as the methods varied in the distribution used and how they adjusted for the various forms of bias. CHiCAGO used weights to adjust the *p*-values for multiple testing and bias associated with the distance between bait and target fragments, whereas the negative binomial method adjusted *p*-values using the FDR. A fairly recent comparison was made between CHiCAGO and an alternative regression model that modelled technical and Brownian noise differently by Eijsbouts (177), with Eijsbouts focussing on the interactions of promoters only (promoter CHi-C). Using two different promoter CHi-C datasets, that were similar biologically, Eijsbouts tested whether CHiCAGO would identify a similar number of significant interactions between the two datasets. Between the two datasets, CHiCAGO detected more significant interactions in one of the datasets than the other, so there was thought to be a problem with design of the detection algorithm or the parameters used by CHiCAGO. With this Eijsbouts used an alternative regression method, which similar to CHiCAGO, also uses the lengths of interacting fragments to predict interaction counts. Trans-chromosomal interaction counts are also used and the regression method also allows the predictors in the model to be estimated simultaneously, whereas the CHiCAGO parameters are estimated in a set order. Bins between the two methods were set differently, with the bins in the regression model able to contain interactions that span different distance ranges. The regression method did perform differently, Eijsbouts found that the regression model was more consistent in detecting interactions between the two promoter CHi-C datasets, then CHiCAGO, however, there was still some inconsistency. CHiCAGO, therefore, has been previously shown to detect a different number of significant interactions to an alternative method. With the there being a large number of loci analysed in this chapter, any differences in the number of interactions detected across the two methods would acuminate. It would have been better if I was able to compare the two methods by their p-values before adjusting by the FDR or a weight, but the CHiCAGO weights used were not given for each *p*-value.

A number of interesting significant physical interactions were, however, detected for both methods and in most of the cell-lines. Bait fragments mapping to 14q24.1 were observed to significantly interact with target fragments in *ZFP36L1,* a gene that has been previously linked to breast cancer (163). Bait fragments mapping to 10p12.31 and 2q35 were found to significantly interact with target fragments mapping to *ARMC3* and *IGFBP5*, respectively, with both genes being linked to cancer metastasis.

Some of the significant capture-to-capture physical interactions detected by the negative binomial method were detected in most cell-lines. A number of the target end fragments mapped to genes that have previously been linked to breast cancer.

Analysis should next focus on establishing the importance of the other end fragments that the loci analysed have been shown to significantly interact with, especially the physical interactions that were shown to be significant in both methods. From just focussing on the interactions present across most of the cell-line, the results seem plausible as target end fragments have been shown to map to genes that have been linked to breast cancer or cancer.

Chapter 7 Summary of thesis and future research

7.1 Thesis summary

In recent years, GWAS have been used to discover many variants associated with a variety of complex diseases and traits. In breast cancer alone, over 90 genetic variants have been shown to be associated with disease risk. By estimating the heritability for many complex diseases, and then quantifying how much of the estimated heritability can be explained by variants associated with disease, it has been found that much of the heritability is unexplained. It has been widely hypothesised that there are many more variants associated with disease risk, but with small effect sizes, they are currently not reaching genome-wide significance as current GWAS are believed to be underpowered to detect such associations. The statistical power of a GWAS is affected by study sample size, and it is believed that as study size continue to increase, there will be many more susceptibility variants discovered. To increases sample sizes, and to improve the power to detect genome-wide significant loci, researchers have been collaborating and combining individual GWAS to form large consortia. In the last few years, large consortia have helped to discover many of the susceptibility variants that have been discovered to date.

Due to the polygenic nature of many diseases, research has begun focusing on the combined effect multiple genetic variants have on disease risk. Polygenic scoring has been used to assess whether genetic signal is present amongst an ensemble of SNPs. With research suggesting that diseases are polygenic, polygenic scores have also been used to examine whether there is evidence of a shared polygenic basis existing between seamlessly independent traits, whether PRS x risk factor interactions exists, and to estimate how much genetic variation in a trait could be explained by genotyped SNPs.

The aim of this thesis was to explain some of the “missing heritability” for breast cancer using polygenic scores, constructed using genotyped GWAS and custom array SNPs. A polygenic score analysis was used in chapter 2 to find evidence that the estimated SNP effects from one breast cancer GWAS, could be used to predict breast cancer risk in an independent GWAS. For different SNP inclusion thresholds, in both directions, the constructed PRS were shown to be significantly associated with breast cancer outcome in an independent GWAS. This result helps to confirm that breast cancer has a polygenic basis. With even more evidence found to strengthen the case that breast cancer has a polygenic basis, the polygenic nature of the disease should be considered when conducting genetic analyses.

Once it was shown that many genetic variants influence breast cancer risk, it was then estimated how much of the genetic variation for breast cancer on the liability scale, could be explained by genotyped SNPs. With known susceptibility variants not explaining a lot of the genetic variation in breast cancer, I wanted to examine the potential genotyped SNPs had in explaining the heritability of breast cancer. Few estimates for breast cancer have been produced based on all genotyped SNPs. Previous chip heritability estimates were either estimated using a smaller sample size than the studies used in the thesis, or were estimated for a subtype of breast cancer instead of overall breast cancer (86, 87). Using much larger breast cancer studies, researchers have estimated the contribution genome-wide significant SNPs have in explaining the familial risk of breast cancer (45, 51). Estimating the genetic variation in breast cancer that can be explained by SNPs, and not just the familial risk, allows an assessment to be made of the ability the SNPs have in explaining the genetic variation in breast cancer for the general population. Breast cancer is not completely familial, so estimating the genetic variation explained in breast cancer on the liability scale allows for a better understanding of the disease. Using three different estimation methods, which differ in how they estimate chip heritability, the variance explained by genotyped GWAS SNPs and custom array SNPs was estimated. GREML uses the genetic relatedness between unrelated individuals and a mixed linear model to estimate chip heritability, whereas LDSC uses the LD between SNPs, and AVENGEME uses polygenic score analysis results and maximum likelihood estimation to estimate chip heritability. Based on GWAS SNPs, chip heritability estimates indicated that genotyped GWAS SNPs explain up to half of the genetic variation in breast cancer liability (~16%-21%). Custom array SNPs explained a smaller proportion of the variation in breast cancer liability, with SNPs genotyped on the array explaining between ~6%-14% of the genetic variation on the liability scale. This was to be expected, as a smaller number of independent SNPs had been genotyped on this array. The chip heritability estimates varied across the different estimation methods, with GREML and AVENGEME estimates shown to be more precise than LDSC estimates. The chip heritability estimates produced show that with increased sample sizes, GWAS have the potential to identify many more associated SNPs that collectively explain a larger proportion of the genetic variation in breast cancer risk, than can be explained by the genome-wide significant SNPs identified to date.

The next natural step was to partition the estimated chip heritability for each study, in order to develop a better understanding of how genetic variation is spread across the genome. This was the first time that a genome partitioning analysis had been conducted for breast cancer. Being able to partition the chip heritability estimates was an advantage of producing estimates on the liability scale, as opposed to the familial risk based estimates, usually reported in breast cancer. The chip heritability estimates were partitioned by MAF, chromosome and SNP annotation using polygenic score analysis and AVENGEME. With the AVENGEME method having never been used to conduct a genome partitioning analysis, and the method accuracy either being similar or better than the other estimation methods, it was used to perform the partitioning analyses.

Partitioning by MAF showed that over 78% of the estimated chip heritability could be explained by common SNPs within each study. It is therefore evident that a large proportion of the genetic variation in breast cancer liability can be explained by common SNPs with an MAF > 0.1. The finding was consistent with other genome partitioning studies that had been conducted for other traits, meaning that the observed result was plausible.

When partitioning the chip heritability estimates by chromosome, a weak linear association between the genetic variance explained by a chromosome, and chromosome length (Mb) was observed. This result suggests that the genetic variation for the breast cancer is spread evenly across the genome, which, again, suggests that breast cancer is a polygenic disease. The association was, however, not significant for the UK2 GWAS. When partitioning the estimated genetic variation in breast cancer liability by chromosome for the COGS, there was also shown to be a positive significant linear association between the genetic variance explained by a chromosome, and the number of SNPs genotyped for each chromosome. These results are similar to other published chromosome partitioning studies, but a stronger linear association has usually been observed. With the per chromosome estimates produced being fairly imprecise, it could be that by improving precision by increasing sample size, as GWAS sample sizes were smaller than those used in the published studies, the association might strengthen.

Partitioning the chip heritability based on SNP annotation led to inconsistent results with large confidence intervals, which meant that a reasonable conclusion could not be made. For the UK2 GWAS, per-SNP estimates were fairly similar across the three annotation groups. UK2 SNPs mapping to intergenic regions of the genome were estimated to explain slightly more of the genetic variation in breast cancer liability, than SNPs mapping to either gene or regulatory regions. However, there was actually little difference between the three per-SNP estimates. For the BBCS, SNPs mapping to gene regions were shown, per-SNP, to explain a larger proportion of the genetic variation in breast cancer liability compared to SNPs mapping to either intergenic or regulatory regions. Again, the differences between these estimates, per-SNP, was actually fairly small, but slightly larger than observed when performing this analysis on the UK2 GWAS. For the COGS, per-SNP, SNPs mapping to regulatory regions were estimated to explain a larger amount of genetic variation in breast cancer liability, compared to the SNPs mapping to either intergenic or gene regions. The actual differences, per-SNP, between the three estimates were again fairly small. With there being only a small difference between the per-SNP estimates for each study, and the preciseness of the estimates being questioned, it was difficult to draw a conclusion from this partitioning analysis.

The SNPs present on the iCOGS array, were chosen based on previous breast cancer, ovarian and prostate cancer GWAS results. The chip heritability estimate for the COGS was partitioned based on the cancer type the SNPs were related to on the array. COGS SNPs were separated into two groups, breast cancer related SNPs and SNPs related to either prostate or ovarian cancer. The genetic variation explained by the SNPs in each group was then estimated. When partitioning by related cancer type, there was shown to be little difference between the two estimates produced, but per-SNP, “breast cancer SNPs” were estimated to explain more of the genetic variation in breast cancer liability than “prostate/ovarian cancer SNPs”. This result indicates that “prostate/ovarian cancer SNPs” do make up a proportion of the genetic signal in breast cancer, but much more of the genetic signal can be attributed to “breast cancer SNPs”.

Overall, the precision of many of the chip heritability estimates produced for individual subsets based on GWAS had to be questioned, as the 95% CIs for the subset estimates tended to be wide. When the COGS chip heritability estimate was partitioned, the subset estimate 95% CIs were found to be much narrower than those for the GWAS. The SNPs genotyped for the COGS in total explained a smaller proportion of the genetic variation in breast cancer liability than all the genotyped GWAS SNPs. Therefore, partitioning GWAS SNPs would provide much more insight into the genetic architecture of breast cancer, but with current estimates lacking precision, much larger sample sizes would be needed for a more precise conclusion to be made based on GWAS. The results do however suggest that breast cancer is a polygenic disease, with the much of the genetic variation in the disease being explained by common SNPs (MAF > 0.1) across the genome.

In chapter 4, both LDSC and polygenic score analysis was used to investigate whether there was evidence to suggest that BMI, a breast cancer risk factor, has a shared polygenic basis with breast cancer. This was the first time that it has been tested whether there is evidence of a shared polygenic basis existing between the two phenotypes. It is possible that many shared genetic variants across the two traits could explain why BMI and breast cancer are associated. Evidence of a shared polygenic basis could enable BMI and breast cancer to be studied together, which could potentially aid the development of new treatments, or help to identify women at an increased risk of developing the disease. Summary BMI data from the GIANT consortium was used, along with the breast cancer studies, to examine whether there was evidence of a genetic overlap between the two phenotypes. Using LDSC, via the web interface LD hub, the genetic correlation between breast cancer and BMI was estimated. The results from the correlation analyses suggested that breast cancer and BMI were not significantly correlated (*p*-value > 0.05). Therefore from this analysis there was no evidence to suggest that breast cancer and BMI have a shared polygenic basis. It was also tested whether a polygenic score for women in the breast cancer studies, constructed using the BMI summary data for published BMI susceptibility variants, was associated with breast cancer risk. The results from this analysis indicated that there was no associated between the BMI derived polygenic score and breast cancer outcome. It was then tested whether a polygenic scores based on published breast cancer SNPs could be used to predict BMI, with significant associations being observed. To examine this further, en-masse breast cancer derived polygenic scores were conducted for multiple *p*-value thresholds, and tested for their association with BMI. Evidence of a shared polygenic basis between breast cancer and BMI was observed, as some of the scores were shown to be associated with BMI, but a number of these significant associations could be considered borderline significant. The associations observed were not as significant as those detected for other shared genetic basis studies, but with the sample sizes having up to 50% power to detect genetic correlation, increasing the size of the breast cancer study used could improve the strength of the association.

Multiple significant associations between different breast cancer derived polygenic scores, and breast cancer outcome were observed in chapter 2. The next objective was to test for PRS x risk factor interactions, to examine whether there was evidence to suggest that breast cancer risk factors could be modifying the effect these scores have on breast cancer risk. BMI and age at menarche, factors that have been shown to be significantly associated with breast cancer risk in previous studies, were the breast cancer risk factors analysed. For this analysis, a case-only approach was adopted in order to improve the power to detect any significant associations, as the number of GWAS individuals with either BMI or age at menarche data was very small. In one of the analyses conducted, the breast cancer cases from the BBCS, with either age at menarche or BMI data, were assigned to the replication sample, with the remaining BBCS subjects being combined with UK2 subjects to form the training sample. For another analysis, the COGS study was set as the training sample, with the BBCS age at menarche/BMI cases assigned to the replication sample. Significant polygenic score and age at menarche linear associations were observed for COGS derived polygenic scores. Scores constructed using either all independent COGS SNPs, SNPs with a breast cancer association *p*-value ≤ 0.7 or a *p*-value ≤ 0.4 were observed to have a significant linear association with age at menarche. The results suggested that these breast cancer polygenic scores could be interacting with age at menarche to have an effect on breast cancer risk. A significant linear relationship between PRS and age at menarche was not shown for the more stringent *p*-value thresholds. This indicated that the effect that an en-masse PRS has on breast cancer risk, could potentially be modified by a woman's age at menarche. For BMI, only one significant linear association was observed, and that was for the most stringent *p*-value threshold (*p*-value ≤ 0.001), estimated using the combined GWAS sample. This result suggests that an interaction between this breast cancer PRS and BMI may exist, but as with many of the other associations, the association was borderline significant. With many of the associations observed being borderline significant, this analysis should be repeated in a larger sample to assess whether there is in fact evidence of either PRS x BMI or PRS x age at menarche interactions.

In addition to testing whether two breast cancer risk factors interacted with multiple breast cancer derived polygenic scores, it was examined whether the effect a breast cancer polygenic score has on breast cancer could be modified by individual genotyped SNPs. Independent SNPs were assigned to a *p*-value threshold bin, based on their individual significance with breast cancer. A polygenic score was constructed for each bin, with each SNP within the same bin being removed from the score. It was tested whether the removed SNP had a significant linear association with the newly formed score. After adjusting for multiple testing using an FDR of 5%, SNP x PRS significant interactions based on SNPs with a *p* ≤ 0.01 and *p* ≤ 0.001 were still observed, but not for the other SNP intervals. Therefore, there was some evidence to suggest that interactions could exist, between some of the SNP and PRS combinations tested. With the analysis being a case-only analysis, and the number of significant interactions greatly exceeding 5% (99%) for *p* ≤ 0.01, it should be further investigated whether these interactions are significant when conducting a case-control interaction analysis.

GWAS to date have been used to identify over 90 susceptibility breast cancer loci, but with many of these loci mapping to gene-deserts or non-coding regions of the genome, understanding fully how they affect breast cancer risk, or trying to pin point the causal variant, has been difficult. In chapter 6, CHi-C analysis was used to test whether susceptibility loci affect breast cancer risk through significant physical interactions with other loci, which map to genes or coding regions. Significant long-range interactions were observed for a large number of the loci, with some of the significant interactions occurring in most of the cell-lines. A number of the significant interactions seemed plausible as they occurred across most, if not all, of the cell-lines, with the target end mapping to a gene known to be somewhat associated with either breast cancer or cancer. The results therefore indicate that other loci could explain how genetic variants in gene deserts, or non-coding regions, affect breast cancer risk.

7.2 Strengths

This thesis has explored the underlying polygenic architecture of breast cancer using current approaches and methods. So far, breast cancer studies have tended to estimate the genetic variation in the familial breast cancer risk explained by discovered breast cancer susceptibility loci. One study has produced a chip heritability estimate for breast cancer, but the estimates produced and presented in this thesis were estimated using larger sample sizes (86). Estimating chip heritability allows the potential genotyped SNPs have in explaining the genetic variation of a trait, to be assessed without first having to discover all associated SNPs. This is a major advantage, as currently sample sizes are not large enough discover many of the associated variants, so these estimates allow us to examine whether much more of the genetic variation in a trait can be explained by SNPs currently not reaching genome-wide significance. Producing chip heritability estimates for both GWAS and custom array SNPs allows the two types of array to be compared, in terms of how much potential the SNPs genotyped on the array have in explaining the genetic variation in breast cancer. The estimates produced in this thesis indicate that we should continue to increase GWAS sample sizes, as much more of the genetic variation in breast cancer liability can be explained by GWAS SNPs, than the SNPs on the custom array.

To my knowledge, this was the first time that chip heritability estimates for breast cancer have been partitioned by either MAF, chromosome or SNP annotation. Therefore conducting these analyses provided insight into how the genetic variation for the disease is spread across the genome. At the time of writing AVENGEME had never been used to perform a genomic partitioning analysis, therefore the work presented in this thesis showed that it can be used to conduct this type of analysis. Future partitioning analyses, however, should be conducted using a larger sample than used in this thesis. With sample sizes increasing and more summary data becoming available, AVENGEME will be a great method to use to estimate chip heritability and conduct a partitioning analysis because it seems to provide reliable estimates with large studies, and can be used with summary data. LDSC can also be used on summary data, but AVENGEME is more reliable than LDSC with smaller studies, and can be used to estimate other parameters, such as the proportion of null SNPs.

For the majority of the analyses conducted in this thesis "deep" scores were used, which meant that all genotyped SNPs were represented in the analysis, regardless of their individual association with breast cancer. Most breast cancer studies to date have conducted their analyses using only genome-wide significant SNPs, meaning that much of the genetic signal across the genome has not been captured in their analyses. In this thesis, en-masse and more stringent polygenic scores were used to conduct various polygenic score analyses. This enabled much more of the genetic signal to be captured and represented in the analyses, than if only genome-wide significant SNPs had been used, as there are many more associated breast cancer SNPs, just studies have been underpowered to detect many of them.

The CHi-C analysis, presented in chapter 6, is a recent approach that has been used to gain a better understanding of the underlying biological mechanisms behind how susceptibility loci affect disease risk. The CHi-C analysis conducted for this thesis, and for a publication in collaboration with Dr. Fletcher, is the largest breast cancer CHi-C analysis to date (161). With over 60 capture loci and seven cell-lines analysed, to my knowledge, it is also currently one of the largest CHi-C analysis to have been performed for any complex disease. Two different methods were used to test for significant physical interactions between capture and target fragments, with a number of the physical interactions being found to be significant when using both methods, and across most cell-lines. This reinforces the plausibility of the results produced.

With the majority of the analyses conducted in this thesis having never been performed for breast cancer, it has meant that the results produced have enabled us to improve our knowledge of the disease.

7.3 Limitations

After estimating and partitioning the chip heritability estimates for each study and finding 95% CIs to be wide, it became clear that the number of individuals in the two GWAS were not large enough to produce fairly accurate estimates. The number of cases and controls in the GWAS would have led to some of the studies being underpowered, with it being estimated for the shared polygenic basis analysis that less than 50% power was achieved. With interaction analyses already suffering from being underpowered, compared to other analyses, the PRS x risk factor interaction analyses would have been especially underpowered as the replication sample sizes were very small. A case-only approach was adopted in order to improve the power to detect an association, but even after applying this approach, the studies would have still been underpowered. The case-only approach improves power but this is compared to a case-control study of the same size. Only 921 subjects had BMI information, and all 921 of them were breast cancer cases. Conducting a case-only analysis using the 921 cases, would be better powered than a case-control analysis with a combined total of 921 cases and controls. With BMI and age at menarche information only available for case subjects anyway, a case-only approach had to be applied regardless. The PRS x risk factor analyses conducted had achieved under 25% power, meaning that, unsurprisingly, the studies were underpowered.

To date, many modifiable and non-modifiable factors have been shown to be associated with breast cancer risk. In the shared polygenic basis analysis only one risk factor was analysed, and for the PRS x risk factor analyses, two risk factors were analysed. Information on other breast cancer risk factors were not provided for the subjects in the GWAS data I had access to. Due to time constraints, in regard to applying for and analysing new data, other PRS x risk factor interaction and shared polygenic basis analyses were not conducted. If I had more time, I would have applied for percent mammographic density GWAS data from the Markers of Density (MODE) consortium (78). Using this data, Varghese et al.(78) have previously shown that polygenic scores based on SNP effects estimated using this study were associated with breast cancer outcome, suggesting genetic overlap between the two phenotypes. Lindstrom et al.(130) have previously tested whether there are shared loci between breast cancer and three different mammographic density phenotypes; dense area, non-dense area and percent density. It would be interesting to test whether there is evidence of a shared polygenic basis between dense area and non-dense area, and breast cancer as it could indicate that these measures are driving the shared polygenic basis between percent mammographic density and breast cancer. Summary data on menopause age and age at menarche has been made available in the public domain by the Reproductive Genetics (ReproGen) Consortium (178-180). Therefore, I could have been tested whether there was evidence of a shared polygenic basis existing between breast cancer and either of the two non-modifiable risk factors.

For the BMI shared polygenic basis analysis that was conducted in this thesis, only genome-wide significant BMI SNPs were used to a score for the women in the breast cancer studies, whereas an en-masse approach would have been better. A reduced number of SNPs were used in the analysis because at the time of conducting the analysis, I was unable to find a way to make sure that all the SNPs used in the score were independent, based on summary data. It is however possible to LD clump summary SNPs using the LD information for the SNPs, based on a reference panel such as 1000 Genomes Project data. PLINK can be used, along with the reference panel, to LD thin a list of SNPs. The estimated summary data weights for the LD thinned SNPs could then be used to construct a polygenic score for the women in the breast cancer studies. With these results, AVENGEME could then be used to estimate how much variation in breast cancer liability could be explained by BMI SNPs.

7.4 Conclusion & future work

This year with the release of new genetic data from the UK Biobank for up to 500,000 genotyped individuals, 13,000 of which are breast cancer cases (181), as well as the OncoArray with data on breast cancer and various other cancers, and potentially data from the 100,000 genomes, we are soon going to experience an influx of large genetic studies for a variety of complex diseases. Not only will these genetic datasets enable researchers to access and analyse a larger number of individuals, but also a greater number of genetic variants. With larger studies, and more consortium summary data becoming available over the next few years, we will gain an even better understanding of the genetic basis of many complex diseases, including breast cancer. With AVENGEME being able to handle large sample sizes and summary data, future breast cancer analyses could explore whether with an increase in sample size, the precision of chip heritability estimates for partitioned subsets of SNPs improve. Larger studies could also be used to re-assess whether there is evidence of PRS x risk factor interactions, or a shared polygenic basis, between breast cancer and various other traits. The analysis conducted in the thesis should not only be replicated in a larger sample, but it ought to be also tested whether the same results can be shown within a different population. With larger studies, the analyses performed in this thesis could also be conducted for different subtypes of cancer, such as ER-negative breast cancer, which is a poor prognosis cancer subtype.

The polygenic score analyses presented in this thesis could be conducted on not only women, but also on men to gain a better understanding of male specific breast cancer. The disease is a lot rarer in males, with approximately 1% of people being diagnosed with breast cancer being male (182), yet the number of men being diagnosed with the disease is increasing. With there being an increase in the number of men being diagnosed with breast cancer, a better understanding of the disease is needed. Male breast cancer is also thought to be polygenic, with many SNPs of small effect being associated with the disease. A number of the SNPs shown to be associated with female breast cancer risk, have also been shown to be associated with male specific breast cancer (183). Therefore, it would also be beneficial to assess whether there is a shared polygenic basis between male and female breast cancer, which if shown could enable the two diseases to be studied together.

Sample sizes are currently not large enough to gain the accuracy needed for polygenic scores to be able to be used to predict an individual's risk of breast cancer. Risk scores are however being combined with environmental factor data and methylation data, to enable stratified risk prediction. The FORECEE (Female Cancer Prediction Using Cervical Omics to Individualise Screening and Prevention) project is a European collaboration that is currently running, with the aim to identify women at high risk of either breast, cervical, endometrial or ovarian cancer using cervical smear cells (184). It is hoped that by modelling a woman's risk of any of these four cancers, based on both environmental and genetic data, that women will be given a score and advised on how they can lower their risk. This score will be used to stratify women for observation, for example, women at an increased risk could be asked to have more frequent mammogram scans.

With breast cancer being a complex disease, it has been difficult to fully understand what causes the disease, and discover all associated risk factors. With many complex diseases to date shown to be influenced by many genetic variants, polygenic analyses have become important in helping to establish how genetics influence disease risk. The results presented in this thesis help to confirm that breast cancer is a polygenic trait, whereby many genetic variants with small effect sizes influence disease risk. By estimating the proportion of genetic variation in breast cancer liability that can be explained by GWAS SNPs, it has been shown that much of the "missing heritability" in breast cancer can be explained by GWAS SNPs. Therefore, as sample sizes increase, we should continue to find more SNPs associated with disease risk and account for more of the genetic variation in disease, but it is also expected that the additional genetic variants discovered will have even smaller effect sizes than we are currently observing (185). Therefore, identifying causal SNPs may become even more challenging (185).

With current sample sizes, breast cancer polygenic scores have been shown to be associated with breast cancer outcome in an independent sample. Therefore, with increased sample size, also comes the possibility of using breast cancer polygenic scores to predict an individual’s breast cancer risk. There is some evidence to suggest that breast cancer and BMI may have a shared polygenic basis, that breast cancer PRS may be modified by breast cancer risk factors, and that physical interactions between risk loci and other loci across the genome exists. Access to larger samples will further improve our understanding of the underlying polygenic basis for breast cancer. For risk prediction and precision medicine to become a possibility, and for present and future findings to be utilised, it is fundamental that breast cancer studies continue to increase the number of SNPs and women genotyped.

References

1. Breast Cancer Now: What is breast cancer? Available from: http://breastcancernow.org/about-breast-cancer/what-is-breast-cancer.
2. Thomas DC. Statistical Methods in Genetic Epidemiology: Oxford University Press; 2004.
3. Sahebi L, Dastgiri S, Ansarin K, Sahebi R, Mohammadi SA. Study Designs in Genetic Epidemiology. ISRN Genetics. 2013;2013:1-8.
4. Matthews AG, Finkelstein DM, Betensky RA. Analysis of familial aggregation studies with complex ascertainment schemes. Stat Med. 2008;27(24):5076-92.
5. Petersen GM. Familial Aggregation: Sorting Susceptibility From Shared Environment. Journal of the National Cancer Institute. 2000;92(14):1114-5.
6. Risch N. The Genetic Epidemiology of Cancer: Interpreting Family and Twin Studies and Their Implications for Molecular Genetic Approaches. Cancer Epidemiol Biomarkers Prev. 2001;10:733-41.
7. Committee on Assessing Interaction Among Social B, and Genetic Factors in Health. , Hernandez LM, Blazer DG. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. 2006:44-62.
8. Easton DF. Familial risks of breast cancer. Breast Cancer Research. 2002;4(5):179.
9. Key TJ, Verkasalo PK, Banks E. Epidemiology of breast cancer. The Lancet Oncology. 2001;2(3):133-40.
10. 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. 1997;71(5):800-9.
11. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer-analyses of cohorts of twins from Sweden, Denmark, and Finland. The New England Journal of Medicine. 2000;343(2):78-85.
12. Zaitlen N, Kraft P. Heritability in the genome-wide association era. Hum Genet. 2012;131(10):1655-64.
13. Wray NR, Visscher PM. Estimating Trait Heritability. Nature Education. 2008;1(1):29.
14. Wray NR, Yang J, Goddard ME, Visscher PM. The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet. 2010;6(2):e1000864.
15. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013;9(3):e1003348.
16. Polderman TJ, Benyamin B, de Leeuw CA, Sullivan PF, van Bochoven A, Visscher PM, et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet. 2015;47(7):702-9.
17. Stefan M, Pulst MD. Genetic Linkage Analysis. Arch Neurol. 1999;56(6):667-72.
18. Ho DWH, Chan D, Cheung KMC, Sham P, Song Y-Q. (ii) Family-based linkage and case control association studies. Current Orthopaedics. 2008;22(4):245-50.
19. Dawn Teare M, Barrett JH. Genetic linkage studies. The Lancet. 2005;366(9490):1036-44.

20. Pharoah PD, Dunning AM, Ponder BA, Easton DF. Association studies for finding cancer-susceptibility genetic variants. Nat Rev Cancer. 2004;4(11):850-60.
21. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet. 2003;72(5):1117-30.
22. Chen S, Parmigiani G. Meta-analysis of BRCA1 and BRCA2 penetrance. J Clin Oncol. 2007;25(11):1329-33.
23. 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.
24. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nat Genet. 2008;40(1):17-22.
25. 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).
26. Group ABCS. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. British Journal of Cancer. 2000;83(10):1301-98.
27. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7(10):781-91.
28. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet. 2002;3(5):391-7.
29. Ripperger T, Gadzicki D, Meindl A, Schlegelberger B. Breast cancer susceptibility: current knowledge and implications for genetic counselling. Eur J Hum Genet. 2009;17(6):722-31.
30. Reid-Lombardo KM, Petersen GM. Understanding genetic epidemiologic association studies Part 1: fundamentals. Surgery. 2010;147(4):469-74.
31. Foulkes AS. Applied Statistical Genetics with R: For Population-Based Association Studies: Springer; 2009.
32. Institute NHGR. An Overview of the Human Genome Project: What was the Human Genome Project? 2012 [updated November 8, 2012]. Available from: http://www.genome.gov/12011238.
33. Hood L, Rowen L. The Human Genome Project: big science transforms biology and medicine. Genome Medicine. 2013.
34. Huang KG, Murray FE. Entrepreneurial experiments in science policy: Analyzing the Human Genome Project. Research Policy. 2010;39(5):567-82.
35. Consortium IHGS. Initial sequencing and analysis of the human genome. Nature. 2001;409.
36. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene Index analysis of the human genone estimates approximately 120,000 genes. Nature Genetics. 2000;25.
37. Chi KR. The dark side of the human geneome. Nature. 2016;538.
38. International HapMap C. A haplotype map of the human genome. Nature. 2005;437(7063):1299-320.
39. Pemberton TJ, Wang C, Li JZ, Rosenberg NA. Inference of Unexpected Genetic Relatedness among Individuals in HapMap Phase III. The American Journal of Human Genetics. 2010;87(4):457-64.

40. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010;467(7311):52-8.
41. Consortium TGP. A global reference for human genetic variation. Nature 2015;526.
42. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308(5720):385-9.
43. Stringer S, Wray NR, Kahn RS, Derks EM. Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. PLoS One. 2011;6(11):e27964.
44. 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.
45. 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. Nat Genet. 2013;45(4):353-61, 61e1-2.
46. Fletcher O, Johnson N, Gibson L, Coupland B, Fraser A, Leonard A, et al. Association of genetic variants at 8q24 with breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2008;17(3):702-5.
47. 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. Nat Genet. 2009;41(5):585-90.
48. 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. Nat Genet. 2010;42(6):504-7.
49. Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, et al. Novel Breast Cancer Susceptibility Locus at 9q31.2: Results of a Genome-Wide Association Study. JNCI Journal of the National Cancer Institute. 2011;103(5):425-35.
50. 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. Nat Genet. 2012;44(3):312-8.
51. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015;47(4):373-80.
52. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst. 2015;107(5).
53. Manolio TA. Bringing genome-wide association findings into clinical use. Nat Rev Genet. 2013;14(8):549-58.
54. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, et al. The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev. 2016.
55. Griffin BH, Chitty LS, Bitner-Glindzicz M. The 100 000 Genomes Project: What it means for paediatrics. Arch Dis Child Educ Pract Ed. 2016.
56. Fisher RA. The correlations between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society pf Edinburgh. 1918;52(2):399-433.

57. Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol. 2011;35(6):506-14.
58. Lee SH, DeCandia TR, Ripke S, Yang J, Schizophrenia Psychiatric Genome-Wide Association Study C, International Schizophrenia C, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet. 2012;44(3):247-50.
59. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7-24.
60. Dudbridge F. Polygenic Epidemiology. Genet Epidemiol. 2016;40(4):268-72.
61. International Schizophrenia C, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52.
62. International Multiple Sclerosis Genetics C, Bush WS, Sawcer SJ, de Jager PL, Oksenberg JR, McCauley JL, et al. Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. Am J Hum Genet. 2010;86(4):621-5.
63. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
64. Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. Curr Protoc Hum Genet. 2011;Chapter 1:Unit1 19.
65. Hardy GH. Mendelian Proportions in a Mixed Population. Science. 1908:49-50.
66. Weinberg W. On the demonstration of heredity in man (trans 1963). Papers on Human Genetics. 1908:4-15.
67. Chen JJ. The Hardy-Weinberg principle and its applications in modern population genetics. Frontiers in Biology. 2010;5(4):348-53.
68. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76(5):887-93.
69. Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010;34(6):591-602.
70. Weale ME. Quality control for genome-wide association studies. Methods Mol Biol. 2010;628:341-72.
71. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5(9):1564-73.
72. Wray NR, Lee SH, Mehta D, Vinkhuyzen AA, Dudbridge F, Middeldorp CM. Research review: Polygenic methods and their application to psychiatric traits. J Child Psychol Psychiatry. 2014;55(10):1068-87.
73. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.
74. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7):459-63.

75. Lewontin RC. The Interaction of Selection and Linkage. I. Optimum Models. Genetics. 1964;50:757-82.
76. Devlin B, Risch N. A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping. Genomics. 1995;29:311-22.
77. Johnson N, Fletcher O, Naceur-Lombardelli C, dos Santos Silva I, Ashworth A, Peto J. Interaction between CHEK2*1100delC and other low-penetrance breast-cancer susceptibility genes: a familial study. The Lancet. 2005;366(9496):1554-7.
78. 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 Res. 2012;72(6):1478-84.
79. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78.
80. Garcia-Closas M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet. 2013;45(4):392-8, 8e1-2.
81. Sun J, Kranzler HR, Bi J. Refining multivariate disease phenotypes for high chip heritability. BMC Med Genomics. 2015;8 Suppl 3:S3.
82. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9.
83. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76-82.
84. Otowa T, Hek K, Lee M, Byrne EM, Mirza SS, Nivard MG, et al. Meta-analysis of genome-wide association studies of anxiety disorders. Mol Psychiatry. 2016;21(10):1391-9.
85. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291-5.
86. Lu Y, Ek WE, Whiteman D, Vaughan TL, Spurdle AB, Easton DF, et al. Most common 'sporadic' cancers have a significant germline genetic component. Hum Mol Genet. 2014;23(22):6112-8.
87. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI, et al. Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. J Natl Cancer Inst. 2015;107(12):djv279.
88. Palla L, Dudbridge F. A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait. Am J Hum Genet. 2015;97(2):250-9.
89. Zheng JIE, Erzurumluoglu M, Elsworth B, Howe L, Haycock P, Hemani G, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. 2016.
90. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011;88(3):294-305.

91. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011;19(7):807-12.
92. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. Package 'pROC'. 2017.
93. Howe HL, Weinstein R, Alvi R, Kohler B, Ellison JH. Women with multiple primary breast cancers diagnosed within a five year period, 1994-1998. Breast Cancer Research and Treatment. 2005;90(3):223-32.
94. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. Nat Genet. 2017;49(9):1304-10.
95. Bulik-Sullivan B. FAQ [updated 30th Jan 2015]. Available from: https://github.com/bulik/ldsc/wiki/FAQ.
96. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. Am J Hum Genet. 2012;91(6):1011-21.
97. Speed D, Cai N, Consortium U, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. Nat Genet. 2017;49(7):986-92.
98. Lee SH, Harold D, Nyholt DR, Consortium AN, International Endogene C, Genetic, et al. Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Hum Mol Genet. 2013;22(4):832-41.
99. Sieradzka D, Power RA, Freeman D, Cardno AG, Dudbridge F, Ronald A. Heritability of Individual Psychotic Experiences Captured by Common Genetic Variants in a Community Sample of Adolescents. Behav Genet. 2015;45(5):493-502.
100. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011;43(6):519-25.
101. Smith NL, Chen MH, Dehghan A, Strachan DP, Basu S, Soranzo N, et al. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. Circulation. 2010;121(12):1382-92.
102. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet. 2014;95(5):535-52.
103. Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat Genet. 2013;45(4):371-84, 84e1-2.
104. 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). Nat Genet. 2009;41(5):579-84.
105. 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. Nat Genet. 2007;39(3):352-8.
106. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. 2014;24(11):1854-68.

107. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.
108. Sequence Ontology: Variant Annotation tools [updated 27 September 2013; cited 2016]. Available from: http://www.sequenceontology.org/so_wiki/index.php/Variant_Annotation_tools.
109. Rudolph A, Chang-Claude J, Schmidt MK. Gene-environment interaction and risk of breast cancer. Br J Cancer. 2016;114(2):125-33.
110. McPherson K, Steel CM, Dixon JM. Breast cancer-epidemiology, risk factors, and genetics BMJ : British Medical Journal. 2000;321(7261):624-8.
111. Tehard B, Clavel-Chapelon F. Several anthropometric measurements and breast cancer risk: results of the E3N cohort study. Int J Obes (Lond). 2006;30(1):156-63.
112. Michels KB, Terry KL, Willett WC. Longitudinal Study on the Role of Body Size in Premenopausal Breast Cancer. Archives of Internal Medicine. 2006;166.
113. Sweeney C, Blair CK, Anderson KE, Lazovich D, Folsom AR. Risk factors for breast cancer in elderly women. Am J Epidemiol. 2004;160(9):868-75.
114. Kawai M, Minami Y, Kuriyama S, Kakizaki M, Kakugawa Y, Nishino Y, et al. Adiposity, adult weight change and breast cancer risk in postmenopausal Japanese women: the Miyagi Cohort Study. Br J Cancer. 2010;103(9):1443-7.
115. Suzuki R, Rylander-Rudqvist T, Ye W, Saji S, Wolk A. Body weight and postmenopausal breast cancer risk defined by estrogen and progesterone receptor status among Swedish women: A prospective cohort study. Int J Cancer. 2006;119(7):1683-9.
116. Bhaskaran K, Douglas I, Forbes H, dos-Santos-Silva I, Leon DA, Smeeth L. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. The Lancet. 2014;384(9945):755-65.
117. Cancer CGoHFiB. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. The Lancet Oncology. 2012;13(11):1141-51.
118. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2006;15(6):1159-69.
119. WHO. Obesity: Preventing and managing the global epidemic. World Health Organization, 2004.
120. England PH. UK and Ireland prevalence and trends 2017 [cited 2016]. Available from: https://www.noo.org.uk/NOO_about_obesity/adult_obesity/UK_prevalence_and_trends.
121. Fuller E, Mindell J, Prior G. Health Survey for England 2015: Health, social care and lifestyles. 2016 [11/01/2017]. Available from: http://www.content.digital.nhs.uk/catalogue/PUB22610.
122. McPherson K, Marsh T, Brown M. Foresight: Tackling obesities: Future choices - modelling furture trends in obesity & their impact on health. 2007.
123. Schienkiewitz A, Schulze MB, Hoffmann K, Kroke A, Boeing H. Body mass index history and risk of type 2 diabetes: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)–Potsdam Study. The american journal of clinical nutrition 2006;84(2):427-33.

124. Dudina A, Therese Cooney M, De Bacquer D, De Backer G, Ducimetière P, Jousilahti P, et al. Relationships between body mass index, cardiovascular mortality, and risk factors: a report from the SCORE investigators European Journal of Preventive Cardiology. 2011;18(5):731 - 42

125. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.

126. Cross-Disorder Group of the Psychiatric Genomics C. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. The Lancet. 2013;381(9875):1371-9.

127. Hagenaars SP, Harris SE, Davies G, Hill WD, Liewald DC, Ritchie SJ, et al. Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. Mol Psychiatry. 2016;21(11):1624-32.

128. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47(11):1236-41.

129. Goris A, van Setten J, Diekstra F, Ripke S, Patsopoulos NA, Sawcer SJ, et al. No evidence for shared genetic basis of common variants in multiple sclerosis and amyotrophic lateral sclerosis. Hum Mol Genet. 2014;23(7):1916-22.

130. Lindstrom S, Thompson DJ, Paterson AD, Li J, Gierach GL, Scott C, et al. Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. Nat Commun. 2014;5:5303.

131. Rauh C, Hack CC, Haberle L, Hein A, Engel A, Schrauder MG, et al. Percent Mammographic Density and Dense Area as Risk Factors for Breast Cancer. Geburtshilfe und Frauenheilkunde. 2012;72(8):727–33.

132. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet. 2012;8(8):e1002793.

133. Johnson T. Genetics ToolboX - Package ‘gtx’ 2015 [cited 2016]. Available from: https://cran.r-project.org/web/packages/gtx/gtx.pdf.

134. Dastani Z, Hivert MF, Timpson N, Perry JR, Yuan X, Scott RA, et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. PLoS Genet. 2012;8(3):e1002607.

135. Johnson T. Efficient Calculation for Multi-SNP Genetic Risk Scores. American Society of Human Genetics Annual Meeting; an Francisco,2012.

136. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28(19):2540-2.

137. Lee PH, Baker JT, Holmes AJ, Jahanshad N, Ge T, Jung JY, et al. Partitioning heritability analysis reveals a shared genetic basis of brain anatomy and schizophrenia. Mol Psychiatry. 2016;21(12):1680-9.

138. Manuck SB, McCaffery JM. Gene-environment interaction. Annu Rev Psychol. 2014;65:41-70.

139. Mullins N, Power RA, Fisher HL, Hanscombe KB, Euesden J, Iniesta R, et al. Polygenic interactions with environmental adversity in the aetiology of major depressive disorder. Psychol Med. 2016;46(4):759-70.

140. Peyrot WJ, Milaneschi Y, Abdellaoui A, Sullivan PF, Hottenga JJ, Boomsma DI, et al. Effect of polygenic risk scores on depression in childhood trauma. Br J Psychiatry. 2014;205(2):113-9.
141. Salvatore JE, Aliev F, Edwards AC, Evans DM, Macleod J, Hickman M, et al. Polygenic scores predict alcohol problems in an independent sample and show moderation by the environment. Genes (Basel). 2014;5(2):330-46.
142. Tyrrell J, Wood AR, Ames RM, Yaghootkar H, Beaumont RN, Jones SE, et al. Gene-obesogenic environment interactions in the UK Biobank study. Int J Epidemiol. 2017.
143. Trotta A, Iyegbe C, Di Forti M, Sham PC, Campbell DD, Cherny SS, et al. Interplay between Schizophrenia Polygenic Risk Score and Childhood Adversity in First-Presentation Psychotic Disorder: A Pilot Study. PLoS One. 2016;11(9):e0163319.
144. Campa D, Kaaks R, Le Marchand L, Haiman CA, Travis RC, Berg CD, et al. Interactions between genetic variants and breast cancer risk factors in the breast and prostate cancer cohort consortium. J Natl Cancer Inst. 2011;103(16):1252-63.
145. Nickels S, Truong T, Hein R, Stevens K, Buck K, Behrens S, et al. Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors. PLoS Genet. 2013;9(3):e1003284.
146. Pierce BL, Ahsan H. Case-only genome-wide interaction study of disease risk, prognosis and treatment. Genet Epidemiol. 2010;34(1):7-15.
147. Dennis J, Hawken S, Krewski D, Birkett N, Gheorghe M, Frei J, et al. Bias in the case-only design applied to studies of gene-environment and gene-gene interaction: a systematic review and meta-analysis. Int J Epidemiol. 2011;40(5):1329-41.
148. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic regression models and case-only designs for assessing susceptability in population-based case-control studies. Stat Med. 1994;13:153-62.
149. Albert PS, Ratnasinghe D, Tangrea J, Wacholder S. Limitations of the case-only design for identifying gene-enviroment interactions. American Journal of Epidemiology. 2001;154(8):687-93.
150. Hodgson ME, Olshan AF, North KE, Poole CL, Zeng D, Tse C, et al. The case-only independence assumption: associations between genetic polymorphisms and smoking among controls in two population-based studies. International Journal of Molecular Epidemiology and Genetics. 2012;3(4).
151. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet. 2008;17(R2):R156-65.
152. McGovern A, Schoenfelder S, Martin P, Massey J, Duffus K, Plant D, et al. Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. Genome Biol. 2016;17(1):212.
153. Martin P, McGovern A, Massey J, Schoenfelder S, Duffus K, Yarwood A, et al. Identifying Causal Genes at the Multiple Sclerosis Associated Region 6q23 Using Capture Hi-C. PLoS One. 2016;11(11):e0166923.
154. Schierding W, Cutfield WS, O'Sullivan JM. The missing story behind Genome Wide Association Studies: single nucleotide polymorphisms in gene deserts have a story to tell. Front Genet. 2014;5:39.

155. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14(6):390-403.
156. Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326(5950).
157. Jager R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat Commun. 2015;6:6178.
158. Martin P, McGovern A, Orozco G, Duffus K, Yarwood A, Schoenfelder S, et al. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat Commun. 2015;6:10069.
159. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44(D1):D710-6.
160. Cairns J, Freire-Pritchett P, Wingett SW, Varnai C, Dimond A, Plagnol V, et al. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biol. 2016;17(1):127.
161. Baxter J, Leavy OC, Dryden NH, Maguire S, Johnson N, Fedele V, et al. Capture Hi-C analysis of 63 breast cancer risk loci identifies putative target genes and causal variants. 2017.
162. Akkiprik M, Feng Y, Wang H, Chen K, Hu L, Sahin A, et al. Multifactorial roles of insulin-like growth factor binding protein 5 in breast cancer. Breast Cancer Research. 2008;10(212).
163. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016;534(7605):47-54.
164. Li X, Liu B, Ji CN, Kang Y, Mao Y. Cloning and expression of ARMC3_v2, a novel splicing variant of the human ARMC3 gene. Genetika. 2006;42(7).
165. Kocarnik JM, Park SL, Han J, Dumitrescu L, Cheng I, Wilkens LR, et al. Pleiotropic and sex-specific effects of cancer GWAS SNPs on melanoma risk in the population architecture using genomics and epidemiology (PAGE) study. PLoS One. 2015;10(3):e0120491.
166. Roca H, Hernandez J, Weidner S, McEachin RC, Fuller D, Sud S, et al. Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. PLoS One. 2013;8(10):e76773.
167. Hamdi Y, Soucy P, Adoue V, Michailidou K, Canisius S, Lemaçon A, et al. Association of breast cancer risk with genetic variants showing differential allelic expression: Identification of a novel breast cancer susceptibility locus at 4q21. Oncotarget. 2016;7(49):80140-63.
168. Olson JE, Wang X, Goode EL, Pankratz VS, Fredericksen Z, Vierkant RA, et al. Variation in genes required for normal mitosis and risk of breast cancer. Breast Cancer Research and Treatment. 2010.
169. Engel C, Versmold B, Wappenschmidt B, Simard J, Easton DF, Peock S, et al. Association of the variants CASP8 D302H and CASP10 V410I with breast and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers. Cancer Epidemiol Biomarkers Prev. 2010;19(11):2859-68.
170. Frank B, Hemminki K, Wappenschmidt B, Meindl A, Klaes R, Schmutzler RK, et al. Association of the CASP10 V410I variant with reduced familial breast

cancer risk and interaction with the CASP8 D302H variant. Carcinogenesis. 2006;27(3):606-9.
171. Micheau O, Tschopp J. Induction of TNF Receptor I-Mediated Apoptosis via Two Sequential Signaling Complexes. Cell. 2003;114(2):181-90.
172. Broeks A, Schmidt MK, Sherman ME, Couch FJ, Hopper JL, Dite GS, et al. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. Hum Mol Genet. 2011;20(16):3289-303.
173. Lin WY, Camp NJ, Ghoussaini M, Beesley J, Michailidou K, Hopper JL, et al. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. Hum Mol Genet. 2015;24(1):285-98.
174. Camp NJ, Parry M, Knight S, Abo R, Elliott G, Rigas SH, et al. Fine-mapping CASP8 risk variants in breast cancer. Cancer Epidemiol Biomarkers Prev. 2012;21(1):176-81.
175. Yin M, Yan J, Wei S, Wei Q. CASP8 polymorphisms contribute to cancer susceptibility: evidence from a meta-analysis of 23 publications with 55 individual studies. Carcinogenesis. 2010;31(5):850-7.
176. Lu D, Chen S, Tan X, Li N, Liu C, Li Z, et al. Fra-1 promotes breast cancer chemosensitivity by driving cancer stem cells from dormancy. Cancer Res. 2012;72(14):3451-6.
177. Eijsbouts C. Consistent Detection of Chromatin Contacts - a new model for the interpretation of Capture Hi-C data. Cambridge: University of Cambridge; 2016.
178. Perry JR, Day F, Elks CE, Sulem P, Thompson DJ, Ferreira T, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature. 2014;514(7520):92-7.
179. Day FR, Ruth KS, Thompson DJ, Lunetta KL, Pervjakova N, Chasman DI, et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. Nat Genet. 2015;47(11):1294-303.
180. Day FR, Thompson DJ, Helgason H, Chasman DI, Finucane H, Sulem P, et al. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. Nat Genet. 2017;49(6):834-41.
181. Biobank U. [updated 13 September 2016 cited 2017]. Available from: http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=100092.
182. Orr N, Lemnrau A, Cooke R, Fletcher O, Tomczyk K, Jones M, et al. Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk. Nat Genet. 2012;44(11):1182-4.
183. Orr N, Cooke R, Jones M, Fletcher O, Dudbridge F, Chilcott-Burns S, et al. Genetic variants at chromosomes 2q35, 5p12, 6q25.1, 10q26.13, and 16q12.1 influence the risk of breast cancer in men. PLoS Genet. 2011;7(9):e1002290.
184. Pashayan N, Reisel D, Widschwendter M. Integration of Genetic and Epigenetic Markers for Risk Stratification: Opportunities and Challenges. Personalized Medicine. 2016;13(2):93-5.
185. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5-22.

Appendices

Appendix 1: Ethical approval for studies used in thesis

Table 1: Ethical approval committees for the individual studies in UK2

Study	Acronym	Country	Approval Committee
ICR Familial Breast and Ovarian Cancer Study	BOCS	UK	The London Multi-Centre Research Ethics Committees
Prospective Study of Outcomes in Sporadic Versus Hereditary Breast Cancer	POSH	UK	South West Multi-centre Research Ethics Committee

Table 2: Ethical approval committees for the BBCS

Study	Acronym	Country	Approval Committee
British Breast Cancer Study	BBCS	UK	South East Multi-Centre Research Ethics Committee

Table 3: Ethical approval committees for the individual studies in BCAC (European descent)

Study	Acronym	Country	Approval Committee
Australian Breast Cancer Family Study	ABCFS	Australia	The University of Melbourne Health Sciences Human Ethics Sub-Committee (HESC)
Amsterdam Breast Cancer Study	ABCS	Netherlands	Leiden University Medical Center (LUMC) Commissie Medische Ethiek and Protocol Toetsingscommissie van het Nederlands Kanker Instituut/Antoni van Leeuwenhoek Ziekenhuis
Bavarian Breast Cancer Cases and Controls	BBCC	Germany	Friedrich-Alexander-Universitat Erlangen-Nurnberg Medizinische Fakultat Ethik-Commission
British Breast Cancer Study	BBCS	UK	South East Multi-Centre Research Ethics Committee
Breast Cancer In Galway Genetic Study	BIGGS	Ireland	Galway University College Hospital Clinical Research Ethical Committee
Breast Cancer Study of the University Clinic Heidelberg	BSUCH	Germany	Medizinische Fakultat Heidelberg Ethikkommission
CECILE Breast Cancer Study	CECILE	France	Comite Consultatif de Protection des Personnes dans la Recherche Biomedicale de Bicetre
Copenhagen General Population Study	CGPS	Denmark	Kobenhavns Amt den Videnskabsetiske Komite
Spanish National Cancer Centre Breast Cancer Study	CNIO-BCS	Spain	Hospital Universitario La Paz Comite Etico de Investigacion Clinica
California Teachers Study	CTS	USA	UC Irvine: Office of Research Institutional Review Board
DEMOKRITOS	DEMOKRITOS	Greece	National Centre of Scientific Research "Demokritos" Ethics Committee and Aristotle University of Thessaloniki Medical School Ethics Committee
ESTHER Breast Cancer Study	ESTHER	Germany	Ruprecht-Karls-Universitat Medizinische Fakultat Heidelberg Ethikkommission
Gene Environment Interaction and Breast Cancer in Germany	GENICA	Germany	Rheinische Friedrich-Wilhelms-Universitat Medizinische Einrichtungen Ethik-Kommission
Helsinki Breast Cancer Study	HEBCS	Finland	Helsingin ja uudenmaan sairaanhoitopiiri (Helsinki University Central Hospital Ethics Committee)
Hannover-Minsk Breast Cancer Study	HMBCS	Belarus	Medizinische Hochschule Hannover Ethik-Kommission
Karolinska Breast Cancer Study	KARBAC	Sweden	Lokala Forskningsetikkommitten Nord
Kuopio Breast Cancer Project	KBCP	Finland	Pohjois-Savon Sairraanhoitopiirin Kuntayhtyma Tutkimuseettinen Toimikunta
Kathleen Cuningham Foundation Consortium for Familial Breast Cancer/Australian Ovarian Cancer Study	kConFab/AOCS	Australia	kConFab: The Queenland Institute of Medical Research Human Research Ethics Committee (QIMR-HREC) AOCS: Peter MacCallum Cancer Centre Ethics Committee
Leuven Multidisciplinary Breast Centre	LMBC	Belgium	Commissie Medische Ethiek van de Universitaire Ziekenhuizen Kuleuven
Mammary Carcinoma Risk Factor Investigation	MARIE	Germany	Ruprecht-Karls-Universitat Medizinische Fakultat Heidelberg Ethikkommission
Milan Breast Cancer Study Group	MBCSG	Italy	Comitato Etico Indipendente della Fondazione IRCCS "Istituto Nazionale dei Tumori"

Study	Acronym	Country	Approval Committee
Mayo Clinic Breast Cancer Study	MCBCS	USA	Mayo Clinic IRB
Melbourne Collaborative Cohort Study	MCCS	Australia	The Cancer Council Victoria Human Research Ethics Committee
Multi-ethnic Cohort	MEC	USA	University of Southern California Health Sciences Campus IRB
Montreal Gene-Environment Breast Cancer Study	MTLGEBCS	Canada	McGill University IRB
Norwegian Breast Cancer Study	NBCS	Norway	Regional Komite for Medisinsk Forskningsetikk (Helseregion III Universitetet I Bergen, Universitetet I Oslo, Helseregion Sor, Helseregion II, and Ost-Norge)
Nashville Breast Health Study	NBHS	USA	Institutional review boards of Vanderbilt University Medical Center
Oulu Breast Cancer Study	OBCS	Finland	Ethical Committee of the Medical Faculty of University of Oulu and Northern Ostrobothnia Hospital District Ethical Committee
Ontario Familial Breast Cancer RegistrY	OFBCR	Canada	Mount Sinai Hospital Research Ethics Board
Leiden University Medical Centre Breast Cancer Study	ORIGO	Netherlands	Medical Ethical Committee and Board of Directors of the Leiden University Medical Center (LUMC)
The Stefanie Spielman Breast Bank and the Columbus Area Control Sample Bank, Ohio State University	OSU	USA	OSU Cancer Institutional Review Board
NCI Polish Breast Cancer Study	PBCS	Poland	National Institute of Health (NIH) IRB
Karolinska Mammography Project for Risk Prediction of Breast Cancer - Case-Control Study	pKARMA	Sweden	Regionala Etikprovningsnamnden i Stockholm (Regional Ethical Review Board in Stockholm)
Rotterdam Breast Cancer Study	RBCS	Netherlands	Medische Ethische Toetsings Commissie Erasmus Medisch Centrum
Roswell Park Cancer Center biorepository, Roswell Park Cancer Institute	RPCI	USA	RPCI Institutional Review Board
Singapore and Sweden Breast Cancer Study	SASBAC	Sweden	Regionala Etikprovningsnamnden i Stockholm (Regional Ethical Review Board in Stockholm)
Sheffield Breast Cancer Study	SBCS	UK	South Sheffield Research Ethics Committee
Study of Epidemiology and Risk factors in Cancer Heredity	SEARCH	UK	Multi Centre Research Ethics Committee (MREC)
Städtisches Klinikum Karlsruhe Deutsches Krebsforschungszentrum Study	SKKDKFZS	Germany	Ethics Committee of the Medical Faculty Heidelberg
IHCC-Szczecin Breast Cancer Study	SZBCS	Poland	Komisji Bioetycznej Pomorskiej Akademii Medycznej
UK Breakthrough Generations Study	UKBGS	UK	South East Multi-Centre Research Ethics Committee

Appendix 2: z-scores for multiple threshold bins

Table 1: LD-clumped SNPs (internal GWAS)

GWAS	*p*-value threshold	No. of SNPs	*z*-score
UK2	1-0.7	12,465	1.481
	0.7-0.4	20,279	1.522
	0.4-0.1	32,900	4.214
	0.1-0.05	7,746	3.853
	0.05-0.01	7,820	2.656
	0.01-0.001	2,354	3.135
	< 0.001	287	2.413
	Total	83,851	
BBCS	1-0.7	10,865	-0.241
	0.7-0.4	16,942	-0.057
	0.4-0.1	26,359	1.371
	0.1-0.05	6,004	3.066
	0.05-0.01	5,699	0.288
	0.01-0.001	1,603	1.451
	< 0.001	181	2.161
	Total	67,654	

Table 2: LD-pruned SNPs (internal GWAS)

GWAS	*p*-value threshold	No. of SNPs	*z*-score
UK2	1-0.7	26,969	0.540
	0.7-0.4	27,005	1.183
	0.4-0.1	27,368	3.065
	0.1-0.05	4,691	2.219
	0.05-0.01	3,895	3.138
	0.01-0.001	873	2.841
	< 0.001	107	1.538
	Total	90,907	
BBCS	1-0.7	22,355	0.142
	0.7-0.4	22,609	1.554
	0.4-0.1	22,739	0.988
	0.1-0.05	3,711	3.547
	0.05-0.01	3,100	0.972
	0.01-0.001	675	-0.292
	< 0.001	71	1.422
	Total	75,259	

Table 3: LD-pruned and LD-clumped SNPs (internal COGS)

LD-removal	*p*-value threshold	No. of SNPs	*z*-score
Pruning	1-0.7	40,183	31.255
	0.7-0.4	39,928	9.438
	0.4-0.1	43,831	20.042
	0.1-0.05	8,334	17.372
	0.05-0.01	7,557	21.298
	0.01-0.001	2,413	24.152
	< 0.001	570	23.249
	Total	142,816	
Clumping	1-0.7	7,526	37.507
	0.7-0.4	10,948	5.077
	0.4-0.1	16,562	15.205
	0.1-0.05	3,876	10.511
	0.05-0.01	3,783	14.574
	0.01-0.001	1,239	17.261
	< 0.001	247	15.486
	Total	44,181	

Appendix 3: *z*-scores for multiple MAF bins

Table 1: LD-clumped UK2 GWAS (internal GWAS)

GWAS	MAF bin	Interval	No. SNPs (Proportion)	*z*-score
UK2	< 0.1	1-0.7	20,800 (24.8%)	-0.389
		0.7-0.4		0.042
		0.4-0.1		0.276
		0.1-0.05		-0.181
		0.05-0.01		0.666
		0.01-0.001		1.083
		< 0.001		0.462
	0.1-0.2	1-0.7	25,715 (30.7%)	0.059
		0.7-0.4		1.423
		0.4-0.1		3.221
		0.1-0.05		2.237
		0.05-0.01		0.958
		0.01-0.001		0.732
		< 0.001		1.145
	0.2-0.3	1-0.7	15,382 (18.3%)	1.632
		0.7-0.4		1.664
		0.4-0.1		1.787
		0.1-0.05		2.876
		0.05-0.01		2.281
		0.01-0.001		2.231
		< 0.001		1.114
	0.3-0.4	1-0.7	11,678 (13.9%)	1.304
		0.7-0.4		-0.689
		0.4-0.1		2.242
		0.1-0.05		3.923
		0.05-0.01		0.032
		0.01-0.001		1.467
		< 0.001		1.219
	0.4-0.5	1-0.7	10,276 (12.3%)	1.860
		0.7-0.4		0.874
		0.4-0.1		2.745
		0.1-0.05		0.011
		0.05-0.01		2.413
		0.01-0.001		1.821
		< 0.001		1.520
		Total	83,851	

Table 2: LD-clumped BBCS GWAS (internal GWAS)

GWAS	MAF bin	Interval	No. SNPs (Proportion)	*z*-score
BBCS	< 0.1	1-0.7	14,114 (20.9%)	0.210
		0.7-0.4		-0.885
		0.4-0.1		1.438
		0.1-0.05		0.030
		0.05-0.01		0.588
		0.01-0.001		0.073
		< 0.001		2.336
	0.1-0.2	1-0.7	20,168 (29.8%)	0.170
		0.7-0.4		-0.755
		0.4-0.1		0.923
		0.1-0.05		2.731
		0.05-0.01		0.182
		0.01-0.001		0.137
		< 0.001		-0.269
	0.2-0.3	1-0.7	13,301 (19.7%)	-0.442
		0.7-0.4		1.230
		0.4-0.1		0.906
		0.1-0.05		0.708
		0.05-0.01		0.956
		0.01-0.001		0.753
		< 0.001		0.314
	0.3-0.4	1-0.7	10,559 (15.6%)	-0.118
		0.7-0.4		1.130
		0.4-0.1		-0.023
		0.1-0.05		1.162
		0.05-0.01		0.069
		0.01-0.001		0.529
		< 0.001		0.910
	0.4-0.5	1-0.7	9,512 (14.1%)	-0.402
		0.7-0.4		-0.611
		0.4-0.1		0.077
		0.1-0.05		2.308
		0.05-0.01		-1.121
		0.01-0.001		1.530
		< 0.001		2.015
		Total	67,654	

Table 3: LD-clumped COGS (internal COGS)

Study	MAF bin	Interval	No. SNPs (Proportion)	*z*-score
COGS	< 0.1	1-0.7		1.306
		0.7-0.4		1.932
		0.4-0.1		7.412
		0.1-0.05		5.910
		0.05-0.01		5.612
		0.01-0.001		7.013
		< 0.001		12.961
	0.1-0.2	1-0.7		1.127
		0.7-0.4		3.958
		0.4-0.1		8.327
		0.1-0.05		4.647
		0.05-0.01		7.558
		0.01-0.001		10.347
		< 0.001		22.066
	0.2-0.3	1-0.7		0.141
		0.7-0.4		2.571
		0.4-0.1		6.423
		0.1-0.05		4.371
		0.05-0.01		7.190
		0.01-0.001		6.878
		< 0.001		19.796
	0.3-0.4	1-0.7		-0.773
		0.7-0.4		2.282
		0.4-0.1		5.912
		0.1-0.05		4.064
		0.05-0.01		5.778
		0.01-0.001		8.495
		< 0.001		14.051
	0.4-0.5	1-0.7		1.211
		0.7-0.4		0.108,
		0.4-0.1		7.023
		0.1-0.05		5.255
		0.05-0.01		7.638
		0.01-0.001		6.633
		< 0.001		25.987
		Total		

Appendix 4: z-scores for multiple chromosome bins

Table 1: LD-clumped UK2 GWAS (internal GWAS)

GWAS	Interval	Chr.	No. SNPs	z-score	Chr.	No. SNPs	z-score
UK2	$0.7 < P \leq 1$	1	6,694	0.465	12	4,181	0.865
	$0.4 < P \leq 0.7$			0.998			0.624
	$0.1 < P \leq 0.4$			1.165			0.654
	$0.05 < P \leq 0.1$			1.225			0.865
	$0.01< P \leq 0.05$			1.096			2.029
	$0.001 < P \leq 0.01$			-0.926			1.527
	$P \leq 0.001$			0.535			-0.579
	$0.7 < P \leq 1$	2	6,473	0.696	13	3,157	-0.917
	$0.4 < P \leq 0.7$			1.485			0.354
	$0.1 < P \leq 0.4$			-0.406			-0.071
	$0.05 < P \leq 0.1$			-0.751			0.559
	$0.01< P \leq 0.05$			-0.797			0.560
	$0.001 < P \leq 0.01$			0.543			1.018
	$P \leq 0.001$			0.366			-0.553
	$0.7 < P \leq 1$	3	5,551	0.993	14	2,845	1.285
	$0.4 < P \leq 0.7$			0.357			-1.443
	$0.1 < P \leq 0.4$			-0.061			0.086
	$0.05 < P \leq 0.1$			0.746			1.346
	$0.01< P \leq 0.05$			1.400			-0.182
	$0.001 < P \leq 0.01$			-0.502			0.038
	$P \leq 0.001$			0.715			0.210
	$0.7 < P \leq 1$	4	5,100	0.886	15	2,705	-1.530
	$0.4 < P \leq 0.7$			1.023			-0.010
	$0.1 < P \leq 0.4$			2.378			1.463
	$0.05 < P \leq 0.1$			0.858			0.642
	$0.01< P \leq 0.05$			-0.503			-1.124
	$0.001 < P \leq 0.01$			0.819			-0.205
	$P \leq 0.001$			-0.666			-0.742
	$0.7 < P \leq 1$	5	5,156	-1.049	16	2,852	-0.114
	$0.4 < P \leq 0.7$			-1.750			1.720
	$0.1 < P \leq 0.4$			2.332			0.317
	$0.05 < P \leq 0.1$			3.071			1.269
	$0.01< P \leq 0.05$			2.071			-0.072
	$0.001 < P \leq 0.01$			2.076			1.285
	$P \leq 0.001$			1.428			2.608
	$0.7 < P \leq 1$	6	5,120	2.130	17	2,731	-0.830
	$0.4 < P \leq 0.7$			1.708			-0.494
	$0.1 < P \leq 0.4$			1.519			1.852
	$0.05 < P \leq 0.1$			0.579			0.338
	$0.01< P \leq 0.05$			-1.161			2.567
	$0.001 < P \leq 0.01$			-0.487			1.503
	$P \leq 0.001$			-0.474			2.331
	$0.7 < P \leq 1$	7	4,596	1.944	18	2,719	-0.847
	$0.4 < P \leq 0.7$			-0.036			0.012
	$0.1 < P \leq 0.4$			1.733			-1.590
	$0.05 < P \leq 0.1$			1.056			-0.314
	$0.01< P \leq 0.05$			0.498			1.219
	$0.001 < P \leq 0.01$			1.184			1.443
	$P \leq 0.001$			0.482			-0.154

GWAS	Interval	Chr.	No. SNPs	*z*-score	Chr.	No. SNPs	*z*-score
UK2	$0.7 < P \leq 1$	8	4,282	1.234	19	2,067	1.568
	$0.4 < P \leq 0.7$			0.861			1.517
	$0.1 < P \leq 0.4$			0.873			0.173
	$0.05 < P \leq 0.1$			0.367			-0.812
	$0.01< P \leq 0.05$			1.285			-0.284
	$0.001 < P \leq 0.01$			0.722			-0.342
	$P \leq 0.001$			2.779			-2.192
	$0.7 < P \leq 1$	9	3,900	-1.534	20	2,435	0.208
	$0.4 < P \leq 0.7$			0.223			0.153
	$0.1 < P \leq 0.4$			1.692			1.784
	$0.05 < P \leq 0.1$			1.105			-0.430
	$0.01< P \leq 0.05$			-0.191			2.257
	$0.001 < P \leq 0.01$			-0.010			1.019
	$P \leq 0.001$			-0.137			-1.017
	$0.7 < P \leq 1$	10	4,364	-0.307	21	1,363	0.446
	$0.4 < P \leq 0.7$			-0.762			0.397
	$0.1 < P \leq 0.4$			2.097			0.093
	$0.05 < P \leq 0.1$			2.483			0.518
	$0.01< P \leq 0.05$			1.898			-0.840
	$0.001 < P \leq 0.01$			2.266			0.265
	$P \leq 0.001$			3.840			1.382
	$0.7 < P \leq 1$	11	4,105	0.118	22	1,454	0.417
	$0.4 < P \leq 0.7$			-0.661			1.018
	$0.1 < P \leq 0.4$			0.105			1.707
	$0.05 < P \leq 0.1$			2.415			-0.342
	$0.01< P \leq 0.05$			0.392			1.147
	$0.001 < P \leq 0.01$			1.584			0.265
	$P \leq 0.001$			-0.214			1.259

Table 2: LD-clumped BBCS GWAS (internal GWAS)

GWAS	Interval	Chr.	No. SNPs	*z*-score	Chr.	No. SNPs	*z*-score
BBCS	0.7 < *P* ≤ 1	1	5,313	0.765	12	3,366	3,366
	0.4 < *P* ≤ 0.7			1.976			-0.410
	0.1 < *P* ≤ 0.4			-0.261			0.948
	0.05 < *P* ≤ 0.1			0.236			1.484
	0.01< *P* ≤ 0.05			0.790			1.455
	0.001 < *P* ≤ 0.01			0.239			2.115
	P ≤ 0.001			1.701			2.443
	0.7 < *P* ≤ 1	2	5,333	-0.489	13	2,515	-0.533
	0.4 < *P* ≤ 0.7			-0.942			1.832
	0.1 < *P* ≤ 0.4			2.174			-0.486
	0.05 < *P* ≤ 0.1			3.061			0.154
	0.01< *P* ≤ 0.05			0.869			0.555
	0.001 < *P* ≤ 0.01			-0.428			1.467
	P ≤ 0.001			0.412			-0.046
	0.7 < *P* ≤ 1	3	4,585	-1.164	14	2,275	-0.947
	0.4 < *P* ≤ 0.7			-0.347			-0.826
	0.1 < *P* ≤ 0.4			0.407			-0.067
	0.05 < *P* ≤ 0.1			2.532			-0.576
	0.01< *P* ≤ 0.05			1.127			-1.109
	0.001 < *P* ≤ 0.01			0.060			1.662
	P ≤ 0.001			0.560			1.119
	0.7 < *P* ≤ 1	4	4,128	1.116	15	2,122	-0.533
	0.4 < *P* ≤ 0.7			-0.518			1.832
	0.1 < *P* ≤ 0.4			3.128			-0.486
	0.05 < *P* ≤ 0.1			1.634			0.154
	0.01< *P* ≤ 0.05			-0.153			0.555
	0.001 < *P* ≤ 0.01			0.938			1.467
	P ≤ 0.001			-0.009			-0.046
	0.7 < *P* ≤ 1	5	4,190	-0.317	16	2,220	-0.947
	0.4 < *P* ≤ 0.7			-2.100			-0.826
	0.1 < *P* ≤ 0.4			-0.381			-0.067
	0.05 < *P* ≤ 0.1			-0.208			-0.576
	0.01< *P* ≤ 0.05			-0.462			-1.109
	0.001 < *P* ≤ 0.01			0.195			1.662
	P ≤ 0.001			-0.815			1.119
	0.7 < *P* ≤ 1	6	4,176	-0.850	17	2,102	-0.483
	0.4 < *P* ≤ 0.7			-0.616			0.818
	0.1 < *P* ≤ 0.4			0.127			-1.253
	0.05 < *P* ≤ 0.1			0.377			0.960
	0.01< *P* ≤ 0.05			-0.621			-0.100
	0.001 < *P* ≤ 0.01			-0.672			-2.177
	P ≤ 0.001			-0.540			-0.113
	0.7 < *P* ≤ 1	7	3,632	0.161	18	2,286	-0.746
	0.4 < *P* ≤ 0.7			0.165			-1.619
	0.1 < *P* ≤ 0.4			0.686			0.145
	0.05 < *P* ≤ 0.1			-0.708			1.447
	0.01< *P* ≤ 0.05			1.471			0.813
	0.001 < *P* ≤ 0.01			0.573			0.009
	P ≤ 0.001			2.006			0.340
	0.7 < *P* ≤ 1	8	3,551	-0.005	19	1,645	0.066
	0.4 < *P* ≤ 0.7			-0.571			0.974
	0.1 < *P* ≤ 0.4			-0.433			-0.160
	0.05 < *P* ≤ 0.1			0.325			0.880
	0.01< *P* ≤ 0.05			-0.232			-1.388
	0.001 < *P* ≤ 0.01			2.049			-1.110
	P ≤ 0.001			-0.109			0.757

GWAS	Interval	Chr.	No. SNPs	*z*-score	Chr.	No. SNPs	*z*-score
BBCS	0.7 < *P* ≤ 1	9	3,168	-0.239	20	1,937	0.028
	0.4 < *P* ≤ 0.7			-0.774			0.125
	0.1 < *P* ≤ 0.4			1.288			-1.030
	0.05 < *P* ≤ 0.1			0.809			-0.876
	0.01< *P* ≤ 0.05			-1.001			-1.093
	0.001 < *P* ≤ 0.01			-0.019			-1.319
	P ≤ 0.001			1.565			0.057
	0.7 < *P* ≤ 1	10	3,482	-0.257	21	1,173	1.794
	0.4 < *P* ≤ 0.7			-0.004			-0.074
	0.1 < *P* ≤ 0.4			0.458			-0.161
	0.05 < *P* ≤ 0.1			-0.956			-0.975
	0.01< *P* ≤ 0.05			0.355			-1.119
	0.001 < *P* ≤ 0.01			0.524			0.742
	P ≤ 0.001			2.208			-0.362
	0.7 < *P* ≤ 1	11	3,207	2.809	22	1,248	0.325
	0.4 < *P* ≤ 0.7			1.281			0.780
	0.1 < *P* ≤ 0.4			-0.163			-0.566
	0.05 < *P* ≤ 0.1			2.021			1.545
	0.01< *P* ≤ 0.05			0.607			-1.727
	0.001 < *P* ≤ 0.01			2.024			-0.625
	P ≤ 0.001			-1.821			-0.548

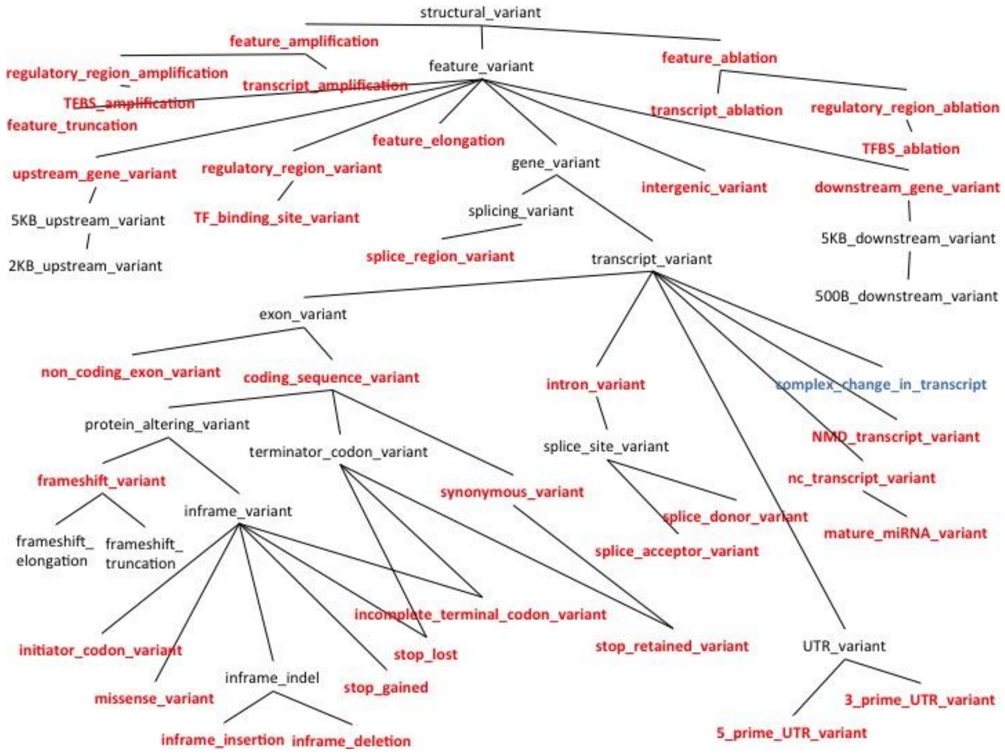
Table 3: LD-clumped COGS (internal COGS)

	Interval	Chr	No. SNPs	*z*-score	Chr.	No. SNPs	*z*-score
COGS	0.7 < *P* ≤ 1	1	3,406	0.397	12	2,212	0.843
	0.4 < *P* ≤ 0.7			1.648			1.797
	0.1 < *P* ≤ 0.4			3.960			5.471
	0.05 < *P* ≤ 0.1			2.472			4.243
	0.01< *P* ≤ 0.05			5.255			3.857
	0.001 < *P* ≤ 0.01			2.555			5.724
	P ≤ 0.001			6.353			10.146
	0.7 < *P* ≤ 1	2	3,480	0.880	13	1,688	0.045
	0.4 < *P* ≤ 0.7			2.397			0.219
	0.1 < *P* ≤ 0.4			4.287			2.244
	0.05 < *P* ≤ 0.1			3.379			0.418
	0.01< *P* ≤ 0.05			3.426			0.211
	0.001 < *P* ≤ 0.01			4.278			1.200
	P ≤ 0.001			11.430			1.886
	0.7 < *P* ≤ 1	3	2,876	0.834	14	1,496	0.611
	0.4 < *P* ≤ 0.7			2.092			1.962
	0.1 < *P* ≤ 0.4			5.222			1.754
	0.05 < *P* ≤ 0.1			3.443			1.806
	0.01< *P* ≤ 0.05			5.109			2.403
	0.001 < *P* ≤ 0.01			4.395			4.417
	P ≤ 0.001			8.376			8.033
	0.7 < *P* ≤ 1	4	2,559	1.624	15	1,383	0.811
	0.4 < *P* ≤ 0.7			0.414			0.952
	0.1 < *P* ≤ 0.4			0.869			4.100
	0.05 < *P* ≤ 0.1			0.025			0.263
	0.01< *P* ≤ 0.05			3.496			1.272
	0.001 < *P* ≤ 0.01			2.841			2.863
	P ≤ 0.001			6.170			1.615
	0.7 < *P* ≤ 1	5	2,632	0.489	16	1,437	1.223
	0.4 < *P* ≤ 0.7			1.248			0.402
	0.1 < *P* ≤ 0.4			5.149			1.755
	0.05 < *P* ≤ 0.1			1.190			2.030
	0.01< *P* ≤ 0.05			4.290			3.232
	0.001 < *P* ≤ 0.01			5.640			2.484
	P ≤ 0.001			12.370			14.253
	0.7 < *P* ≤ 1	6	2,783	0.260	17	1,475	0.697
	0.4 < *P* ≤ 0.7			2.024			0.475
	0.1 < *P* ≤ 0.4			5.374			1.657
	0.05 < *P* ≤ 0.1			4.192			1.740
	0.01< *P* ≤ 0.05			5.694			2.056
	0.001 < *P* ≤ 0.01			4.947			1.795
	P ≤ 0.001			8.600			5.242
	0.7 < *P* ≤ 1	7	2,398	0.663	18	1,371	1.153
	0.4 < *P* ≤ 0.7			0.671			0.592
	0.1 < *P* ≤ 0.4			2.700			2.793
	0.05 < *P* ≤ 0.1			1.692			1.138
	0.01< *P* ≤ 0.05			2.434			1.992
	0.001 < *P* ≤ 0.01			3.261			1.959
	P ≤ 0.001			3.755			1.715
	0.7 < *P* ≤ 1	8	2,351	0.130	19	1,170	0.372
	0.4 < *P* ≤ 0.7			0.549			2.804
	0.1 < *P* ≤ 0.4			2.324			1.550
	0.05 < *P* ≤ 0.1			3.896			1.206
	0.01< *P* ≤ 0.05			2.253			2.281
	0.001 < *P* ≤ 0.01			6.821			2.028
	P ≤ 0.001			8.768			5.520

	Interval	Chr.	No. SNPs	*z*-score	Chr.	No. SNPs	*z*-score
COGS	0.7 < *P* ≤ 1	9	2,055	2.221	20	1,197	0.995
	0.4 < *P* ≤ 0.7			1.177			0.742
	0.1 < *P* ≤ 0.4			3.728			2.591
	0.05 < *P* ≤ 0.1			3.876			1.902
	0.01< *P* ≤ 0.05			4.198			1.385
	0.001 < *P* ≤ 0.01			2.909			3.535
	P ≤ 0.001			7.864			0.018
	0.7 < *P* ≤ 1	10	2,498	0.757	21	690	0.578
	0.4 < *P* ≤ 0.7			3.050			0.432
	0.1 < *P* ≤ 0.4			4.276			1.202
	0.05 < *P* ≤ 0.1			2.973			0.689
	0.01< *P* ≤ 0.05			4.072			0.871
	0.001 < *P* ≤ 0.01			5.963			1.027
	P ≤ 0.001			18.465			2.996
	0.7 < *P* ≤ 1	11	2,224	0.408	22	800	0.757
	0.4 < *P* ≤ 0.7			1.356			0.952
	0.1 < *P* ≤ 0.4			4.399			1.650
	0.05 < *P* ≤ 0.1			2.498			3.250
	0.01< *P* ≤ 0.05			3.984			1.927
	0.001 < *P* ≤ 0.01			6.018			1.685
	P ≤ 0.001			11.859			4.835

Appendix 5: SNP annotation tree

Diagram 1: Term tree for the Variant Effect Predictor (VEP) tool



Source: http://www.sequenceontology.org/so_wiki/index.php/Variant_Annotation_tools

Appendix 6: z-scores for multiple SNP annotation bins

Table 1: LD-clumped UK2 GWAS (internal GWAS)

Annotation group	Interval	*z*-score	No. SNPs
Intergenic variant	0.7 < *P* ≤ 1	1.276	32,406
	0.4 < *P* ≤ 0.7	0.172	
	0.1 < *P* ≤ 0.4	2.124	
	0.05 < *P* ≤ 0.1	2.872	
	0.01< *P* ≤ 0.05	2.590	
	P ≤ 0.01	3.467	
Regulatory variant	0.7 < *P* ≤ 1	0.547	3,757
	0.4 < *P* ≤ 0.7	0.544	
	0.1 < *P* ≤ 0.4	0.641	
	0.05 < *P* ≤ 0.1	0.508	
	0.01< *P* ≤ 0.05	1.602	
	P ≤ 0.01	-0.101	
Gene variant	0.7 < *P* ≤ 1	0.735	47,642
	0.4 < *P* ≤ 0.7	1.742	
	0.1 < *P* ≤ 0.4	3.839	
	0.05 < *P* ≤ 0.1	2.709	
	0.01< *P* ≤ 0.05	1.091	
	P ≤ 0.01	2.230	

Table 2: LD-clumped BBCS GWAS (internal GWAS)

Annotation group	Interval	*z*-score	No. SNPs
Intergenic variant	$0.7 < P \leq 1$	1.115	25,119
	$0.4 < P \leq 0.7$	-0.815	
	$0.1 < P \leq 0.4$	1.428	
	$0.05 < P \leq 0.1$	0.710	
	$0.01 < P \leq 0.05$	-1.832	
	$P \leq 0.01$	1.649	
Regulatory variant	$0.7 < P \leq 1$	-0.638	2,916
	$0.4 < P \leq 0.7$	0.329	
	$0.1 < P \leq 0.4$	0.024	
	$0.05 < P \leq 0.1$	0.229	
	$0.01 < P \leq 0.05$	-0.705	
	$P \leq 0.01$	1.438	
Gene variant	$0.7 < P \leq 1$	1.115	35,916
	$0.4 < P \leq 0.7$	0.521	
	$0.1 < P \leq 0.4$	1.033	
	$0.05 < P \leq 0.1$	2.664	
	$0.01 < P \leq 0.05$	1.230	
	$P \leq 0.01$	1.745	

Table 3: LD-clumped COGS (internal COGS)

Annotation group	Interval	*z*-score	No. SNPs
Intergenic variant	$0.7 < P \leq 1$	0.547	16,933
	$0.4 < P \leq 0.7$	1.377	
	$0.1 < P \leq 0.4$	8.203	
	$0.05 < P \leq 0.1$	6.888	
	$0.01 < P \leq 0.05$	8.227	
	$P \leq 0.01$	21.447	
Regulatory variant	$0.7 < P \leq 1$	-0.496	1,969
	$0.4 < P \leq 0.7$	2.417	
	$0.1 < P \leq 0.4$	1.803	
	$0.05 < P \leq 0.1$	3.749	
	$0.01 < P \leq 0.05$	1.581	
	$P \leq 0.01$	8.838	
Gene variant	$0.7 < P \leq 1$	1.842	24,606
	$0.4 < P \leq 0.7$	4.677	
	$0.1 < P \leq 0.4$	12.789	
	$0.05 < P \leq 0.1$	6.872	
	$0.01 < P \leq 0.05$	11.910	
	$P \leq 0.01$	30.265	

Appendix 7: *z*-scores for COGS cancer type

Table 1: Multiple *z*-scores for breast cancer and ovarian or prostate SNPs (COGS)

Cancer type	Interval	*z*-score	No. SNPs
Breast	$0.7 < P \leq 1$	37.136	16,761
	$0.4 < P \leq 0.7$	2.363	
	$0.1 < P \leq 0.4$	10.459	
	$0.05 < P \leq 0.1$	8.324	
	$0.01< P \leq 0.05$	12.308	
	$0.001 < P \leq 0.01$	14.880	
	$P \leq 0.001$	14.360	
Ovarian or prostate	$0.7 < P \leq 1$	8.873	27,420
	$0.4 < P \leq 0.7$	4.599	
	$0.1 < P \leq 0.4$	11.562	
	$0.05 < P \leq 0.1$	6.993	
	$0.01< P \leq 0.05$	9.237	
	$0.001 < P \leq 0.01$	10.367	
	$P \leq 0.001$	7.627	

Appendix 8: Genetic variance explained by PRS in BMI

Table 1: Genetic variance explained by breast cancer PRS in BMI (Pseudo R2)

p-value threshold	Training sample	R2 variance explained (%)
1	BBCS	0.001225
0.7	BBCS	0.001158
0.4	BBCS	0.000509
0.1	BBCS	0.000185
0.05	BBCS	0.000017
0.01	BBCS	0.000191
0.001	BBCS	0.00001
1	UK2	0.000113
0.7	UK2	0.000109
0.4	UK2	0.000105
0.1	UK2	0.000087
0.05	UK2	0.00036
0.01	UK2	0.000298
0.001	UK2	0.000272
1	Combined GWAS	0.001069
0.7	Combined GWAS	0.001135
0.4	Combined GWAS	0.001366
0.1	Combined GWAS	0.001585
0.05	Combined GWAS	0.001721
0.01	Combined GWAS	0.001616
0.001	Combined GWAS	0.000002
1	COGS	0.000044
0.7	COGS	0.000031
0.4	COGS	0.000083
0.1	COGS	0.00055
0.05	COGS	0.000398
0.01	COGS	0.001015
0.001	COGS	0.001144

Appendix 9: Overlapping loci

Table 1: Overlapping target loci

Locus	Locus no.	Overlaps	Locus	Locus no.	Overlaps
22q12.1	1		9q31.2	41	40
22q13.1	2		8p12	42	
21q21.1	3		8q21.11	43	44
21q21.2	4		8q21.11	44	43
20q13.13	5		8q24.21	45	46
19p13.1	6		8q24.21	46	45
19p13.11	8		7q35	47	
19q13.31	9		6p25.3	48	
18q11.2	10		6p23	49	
18q11.2	11		6q14.1	50	
17q22	12		6q22.31	51	
16q12.1	13		6q25.1	52	
16q12.2	14		5p15.33	53	
16q23.2	16		5p12	55	
14q13.3	17		5q11.2	56	57
14q24.1	18		5q11.2	57	56
14q24.1	19		5q33.3	58	
14q32.11	20		4q24	59	
13q13.1	21		4q34.1	60	
12p13.1	22		3p26.1	61	
12p11.22	23		3p24.1	62	63
12q22	24		3p24.1	63	62
12q24.21	25		2p24.1	64	
11p15.5	26		2q14.2	65	
11q13.1	27		2q31.1	66	67
11q13.3	28		2q31.2	67	66
11q24.3	30		2q33.1	68	
10p15.1	31		2q35	69	70
10p12.31	32		2q35	70	69
10q21.2	33		1p36.22	71	
10q22.3	34	35	1p31.1	72	
10q23.1	35	34	1p13.2	73	
10q25.2	36		1p11.2	74	
10q26.13	38		2p25.1	77	
9p21.3	39		5q31.2	78	
9q31.2	40	41	1p13.3	79	