Genetic characterisation of radiation-induced breast cancer in survivors of Hodgkin's lymphoma

Magda Meier

A thesis submitted for the degree of

Master of Philosophy (M.Phil.)

The Institute of Cancer Research

University of London

Declaration

I, Magda Meier, declare that the work presented in this thesis is my own.

Where information has been derived from other sources, or work has been performed with the assistance of others, this has been clearly indicated.

Date: 30/08/19

Signed: MEMeier

Table of Contents

Abstract

Hodgkin's lymphoma (HL) accounts for less than one percent of new cancer cases every year in the UK. Although historically HL patients had a poor prognosis, decades of treatments for this hematopoietic neoplasm have led to improved survival rates. However, as with many childhood cancers, the price of success includes adverse treatment-related effects such as cardiovascular disease, neurological disorders, and elevated risk of second cancers in later life. In the case of HL, radiotherapy is thought to be a significant factor that underpins the high rate of second cancers amongst HL survivors relative to the general population. Although it occurs almost exclusively in women, breast cancer (BC) is the most common second cancer in HL survivors, who have a six-fold increased risk compared to the general population. Despite both the high incidence of BC in this setting and rapidly accruing knowledge regarding genetic determinants of sporadic and familial BC, the genetics of radiation-induced BC are largely understudied and it remains unclear whether some women are genetically predisposed to developing BC following radiation exposure. The aim of this thesis is to determine whether germline variation contributes to the aetiology of radiation-induced BC. Firstly, I performed a Genome-Wide Association Study (GWAS) of HL survivors who received radiotherapy as part of their HL treatment to identify common polymorphisms associated with radiation-induced BC. This GWAS identified a total of 72 SNPs forming ten independent signals. Secondly, I investigated whether a BC polygenic risk score (PRS) is associated with risk of radiation-induced BC. This PRS successfully

demonstrated a trend of increasing BC risk with increasing PRS scores, thus indicating that a PRS may be a useful method for identifying high risk and low risk individuals. Finally, I performed a targeted sequencing analysis of DNA repair genes in order to investigate whether rare variants in these genes influence predisposition to HLBC. This approach identified of total of 10 rare variants that are associated with radiation-induced BC, thereby providing an indication that DNA repair defects may play a role in the aetiology of radiation-induced BC. In identifying both common and rare variants underlying radiation-induced BC, this project contributes to improving the current understanding of this disease and provides new insights that may help develop new prevention and treatment strategies in the future.

Acknowledgments

I would like to thank my supervisor, Dr Nick Orr, for his help and guidance during the last three years and for giving me the opportunity to work at the Institute of Cancer Research (ICR). Nick, thank you for this opportunity and for supporting me, both in London and from Belfast. I would also like to thank my second supervisor, Professor Anthony Swerdlow, and my on-site supervisor, Dr Rachel Natrajan. I would like to thank the ICR as well as Breast Cancer Now for funding my project and for allowing me to work here, it has been a great honour.

I would like to acknowledge everyone who has helped me with the scientific work presented herein. Katarzyna Tomczyk performed the blood extractions for the Complex Trait Genetics team (CTG) and helped me to make the library preparations for the NGS step of this project. Dr Sarah Maguire provided bioinformatics support by performing the analyses of the NGS raw data, as well as providing guidance during the imputation step of the GWAS.

I would like to thank everyone who was a part of my everyday life at the ICR. First of all, I would like to thank all the members of my lab. Kasia, thank you for your guidance and training, you always took the time to explain every detail to me in the wet lab, I can't thank you enough. Tereza, thank you for all the support you gave me, you were always there when I had a question and your advice was always so helpful. You are a true mentor to me and you have

taught me how to be a better scientist. Lena, thank you for all your support, I really enjoyed our conversations together, we had a lot of fun. Sarah, you have taught me so much in both bioinformatics and wet-lab practice, thank you very much. Secondly, I would like to thank all the members of Tony's lab, in particular Jane, Penny, Michael and Aislinn. Jane and Penny, thank you for all your support, you always made me feel very welcome in Sutton and I enjoyed working with you very much. Michael, thank you very much for all our meetings and discussions regarding statistics. You always took the time to explain things to me, I really appreciate it. A special thanks to Aislinn for providing me with the clinical data I needed to complete this project, thank you so much. I would also like to thank Minouk, Lauren, Hannah and June, who always made me feel very welcome in Sutton, I always enjoyed our lunches together. Thirdly, I would like to thank everyone in Dr Olivia Fletcher's lab. Olivia, thank you very much for the support you gave me. You always made me feel at ease during our journal clubs and your advice was always very helpful, I can't thank you enough. Nikki, thank you for your help and advice, it was always a pleasure to work with you. Vita, I really enjoyed our few months working together, it was a lot of fun. You were always there if I needed any advice, thank you so much. Joe, thanks for all your support. I enjoyed our quizzes and chats.

A big thanks to all the people I met during my time at the ICR, in particular Kunzah, Nadia and Laura. Our chats and numerous coffees were always a highlight for me, your support has meant so much to me, thank you. Maria, I really enjoyed our chats and debates on the 6th floor, thank you. I would also

like to thank Nikiana and Isaac Garcia-Murillas for their advice and kindness. I would like to thank Michael Hazel and Fabrice from IT for their help every time I had a laptop issue, you always took the time to quickly help me, thank you.

Finally, I would like to thank my family for being supportive and for always believing in me, especially my Mum and step-dad. Mummy, thank you so much, your support is always so meaningful to me. You always make your best to understand me and support in any way you can, I can't thank you enough. Didier, merci pour ton aide. Ton soutien est toujours très important pour moi et tes messages m'ont toujours beaucoup aidé. I would also like to thank my brothers and sisters, especially Lois and Luc, who have always believed in me and who have always been role models for me. Lois, merci pour ton soutien dans les moments difficiles, tes conseils et tes vidéos marrantes m'ont beaucoup aidé. Luc, merci pour tes conseils, tes visites à Londres étaient toujours un moment de détente pour moi et un énorme soutien. Finally, I would like to thank Justin. Justin, you have been at my side every day. Your love, humour, support and understanding have meant so much to me, I could not have done any of this without you by my side.

List of tables

List of figures

List of abbreviations

ABVD	Adriamycin, Bleomycin, Vinblastine, Dacarbazine
ADHD	Attention Deficient Hyperactivity Disorder
BC	Breast Cancer
BCAC	Breast Cancer Association Consortium
BER	Base Excision Repair
cHL	Classical Hodgkin's Lymphoma
CTG	Complex Trait Genetics (team)
DCIS	Ductal carcinoma in situ
EBV	Epstein-Barr Virus
EDTA	Ethylenediaminetetraacetic acid
eQTL	Quantitative trait locus
ExAC	Exome Aggregation Consortium
FBC	Female Breast Cancer
GATK	Genome Analysis Tool Kit
GG-NER	Global Genome Nucleotide Excision Repair
GTEx	Genotype-Tissue Expression

GWAS	Genome-Wide Association Study
HL	Hodgkin’s Lymphoma
HLBC	Hodgkin’s Lymphoma followed by Breast Cancer
HR	Hazard Ratio
HRS	Hodgkin and Reed-Sternberg (HRS) cells
HWP	Hardy Weinberg Principle
IBD	Identity By Descent
JMJs	Jumonji-containing domain proteins
LDCHL	Lymphocyte Depleted Classical Hodgkin’s Lymphoma
LRCHL	Lymphocyte Rich Classical Hodgkin’s Lymphoma
MAF	Minor Allele Frequency
MCCHL	Mixed Cellularity Classical Hodgkin’s Lymphoma
MOPP	Mustargen, Oncovin, Procarbizine, Prednisone
MPG	Molecular and Population Genetics (team)
NCI	National Cancer Institute
NER	Nucleotide Excision Repair
NGS	Next-Generation Sequencing
NRS	National Recall Study
NSCHL	Nodular sclerosis Classical Hodgkin’s Lymphoma
OR	Odds Ratio

PCA	Principal Component Analysis
PRS	Polygenic Risk Score
QC	Quality Control
QQ	Quantile-Quantile (plot)
SIR	Standard Incidence Rate
SKAT-O	Optimised Sequence Kernel association Test
SNP	Single Nucleotide Polymorphism
SNV	Single-Nucleotide Variant
TPU	Tumour Profiling Unit (team)
VAT	Variant Association Tools
WSS	Weighted Sum Statistic

Chapter 1 - Introduction

1.1 Hodgkin's lymphoma

1.1.1 Epidemiology of Hodgkin's lymphoma

Hodgkin's lymphoma (HL) is a hematopoietic neoplasm that affects the lymphatic system. It originates in lymph nodes which are distributed throughout the body (Figure 1) (1). HL also occurs in organs such as the liver, lungs and bones, but this is comparatively rare (2). Most commonly, tumours originate in the lymph nodes surrounding the neck, specifically the supraclavicular, the axillary and the mediastinal zones (Figure 2). Patients often present with supradiaphragmatic lymphadenopathy (enlarged or swollen nodes) as well as symptoms including fever, weight loss and night sweats.

HL is a relatively rare cancer, with approximately 2,100 new cases per year that account, annually, for less than one percent of cancer incidences in the UK (2). Notably, HL is not to be mistaken with non-Hodgkin's lymphoma, which is a more common type of lymphoma affecting over 13,000 individuals every year (2). Although it affects individuals of all ages, age-distribution rates indicate two peaks: the first occurs in young adults aged between 20 and 24, while the second occurs in adults aged between 75 and 79 (Figure 3). The

incidence of HL has been increasing since the early 1990s, with an overall increase of 32% in men and 40% in women (Figure 4) (2).

Historically, HL has long been associated with Epstein-Barr virus (EBV), a gamma herpesvirus that causes glandular fever (infectious mononucleosis). Previous studies have shown that the virus is present in tumour cells, although this observation is infrequent in Western Europe and North America (3,4). While EBV infection appears to be associated with certain subtypes of HL, the absolute risk of developing HL following an EBV infection is low, with approximately one case arising per 1,000 EBV infected individuals (5,6).

Immunosuppressed individuals are known to have elevated risks of HL. The incidence of HL is high in individuals who have received organ transplantation, as well as in individuals who suffer from immune conditions such as rheumatoid arthritis (7). Individuals infected with HIV have a significantly higher risk of developing HL than the general population (8). Although HL risk has also been linked with smoking, the data supporting this association is somewhat conflicting, with some studies suggesting a protective effect while others support a deleterious effect (9,10). A recent analysis of smoking status in HL patients suggests that smoking may, in fact, be associated with specific histological subtypes of HL (11). In addition, alcohol has been shown to have a protective effect towards HL risk (12). Obesity, on the other hand, is associated with an increased HL risk (13).

Evidence of an inherited genetic influence on susceptibility to HL is well established, with over 40 SNPs having been found to be associated with predisposition to the disease (14). The strongest evidence is found in the human leukocyte antigen (HLA) region, which plays the essential role of helping the immune system distinguish body proteins from foreign proteins made by viruses or bacteria (14). The HLA region is the most polymorphic region in the human genome and is comprised of several sub-regions called HLA class I, II and III. Susceptibility loci have been found in a number of HLA genes, in particular *HLA-A*, *HLA-B* and *HLA-DRB1* (14–18). Given the essential role of these genes in the immune system, these associations indicate that HLA genes most likely have direct functional relevance to HL carcinogenesis. Interestingly, a number of associations in the HLA regions are dependent upon EBV status, suggesting that immune responses to EBV influence HL susceptibility (15,17). In addition, a number of non-HLA genes have been associated with HL. Most of these also map to genes that play a role in the immune system. These include genes with regulatory roles, such as *IL13, IL4RA and STAT6* (19–22).

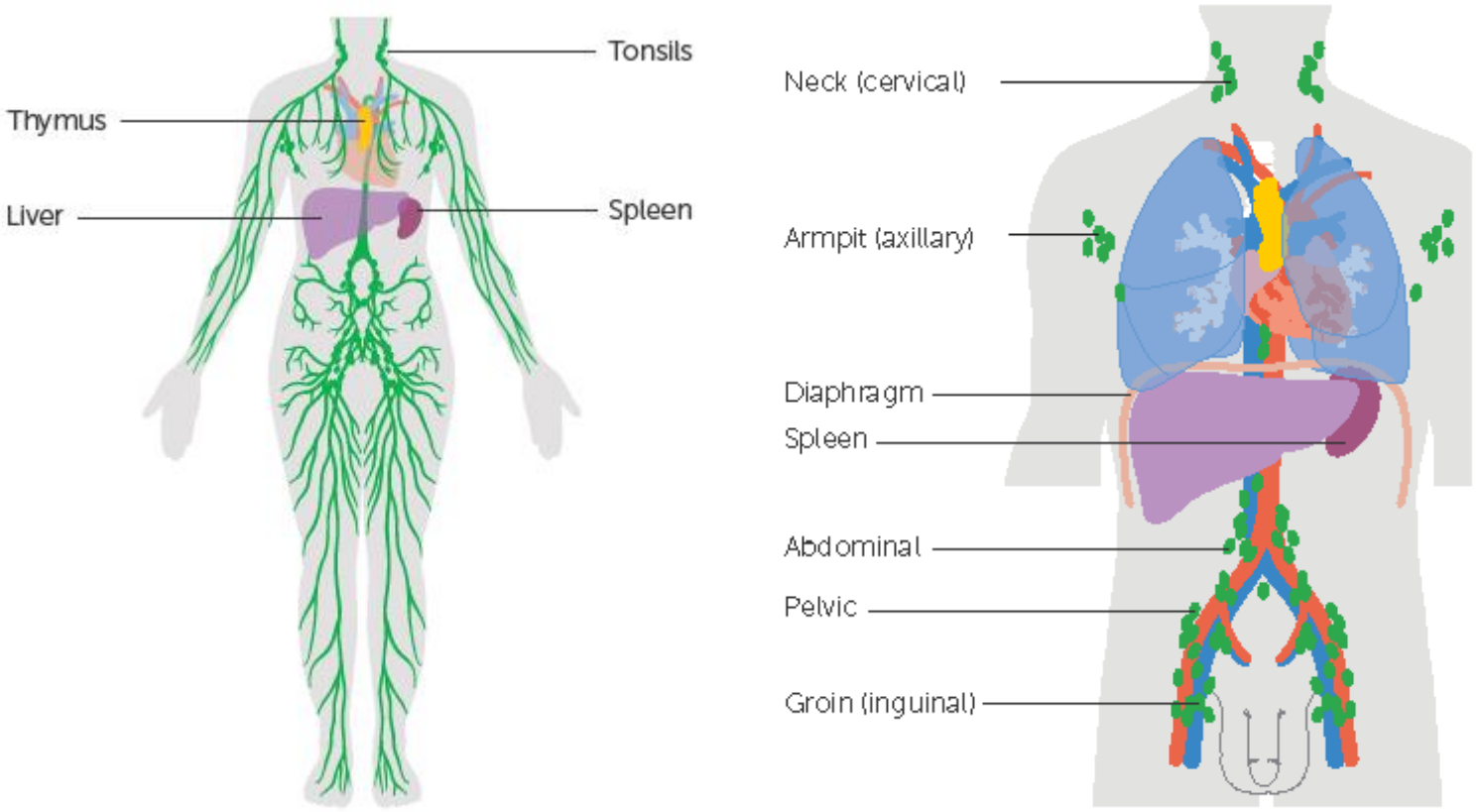

Figure 1. The lymphatic system and its distribution of lymph nodes. The left diagram shows the lymphatic system in green. The right diagram depicts the lymph nodes where Hodgkin's lymphoma tumours often originate. The lymph nodes are depicted in green. These diagrams were obtained from Cancer Research UK (2).

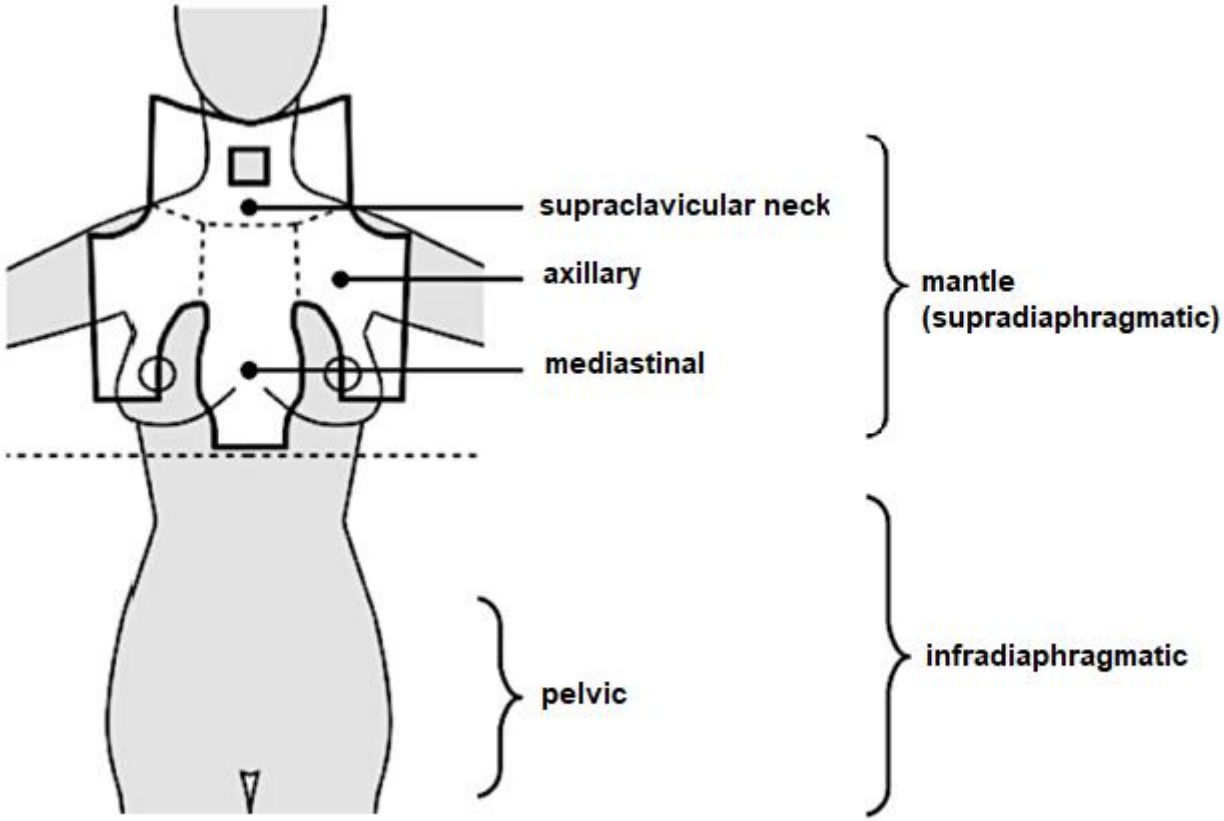

Figure 2. Fields where Hodgkin's lymphoma commonly originates. Modified from De Bruin *et al* (23).

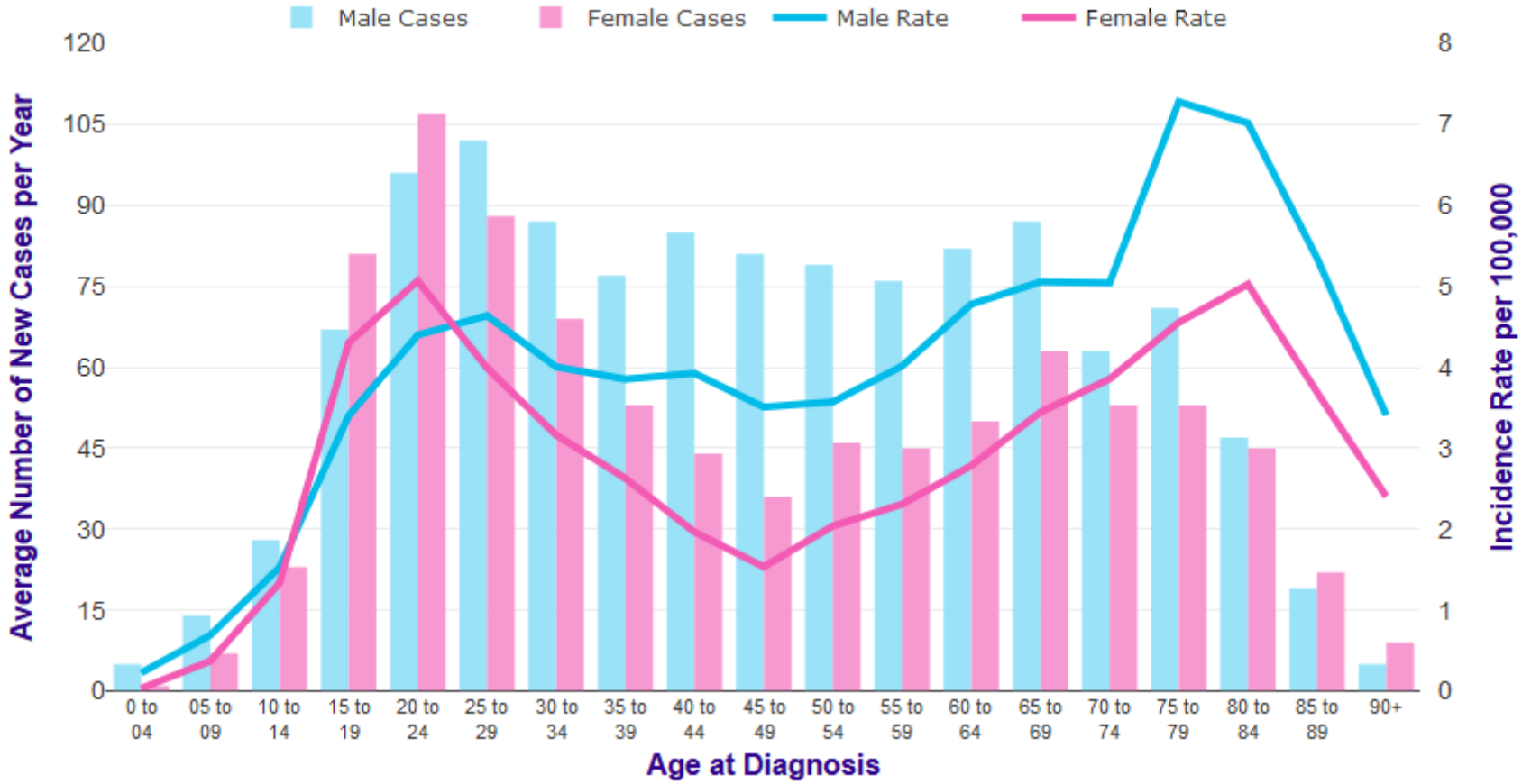

Figure 3. Average number of new HL cases per year and age-specific incidence rates per 100,000 in the UK between 2013 and 2015. This plot was obtained from Cancer Research UK (2)

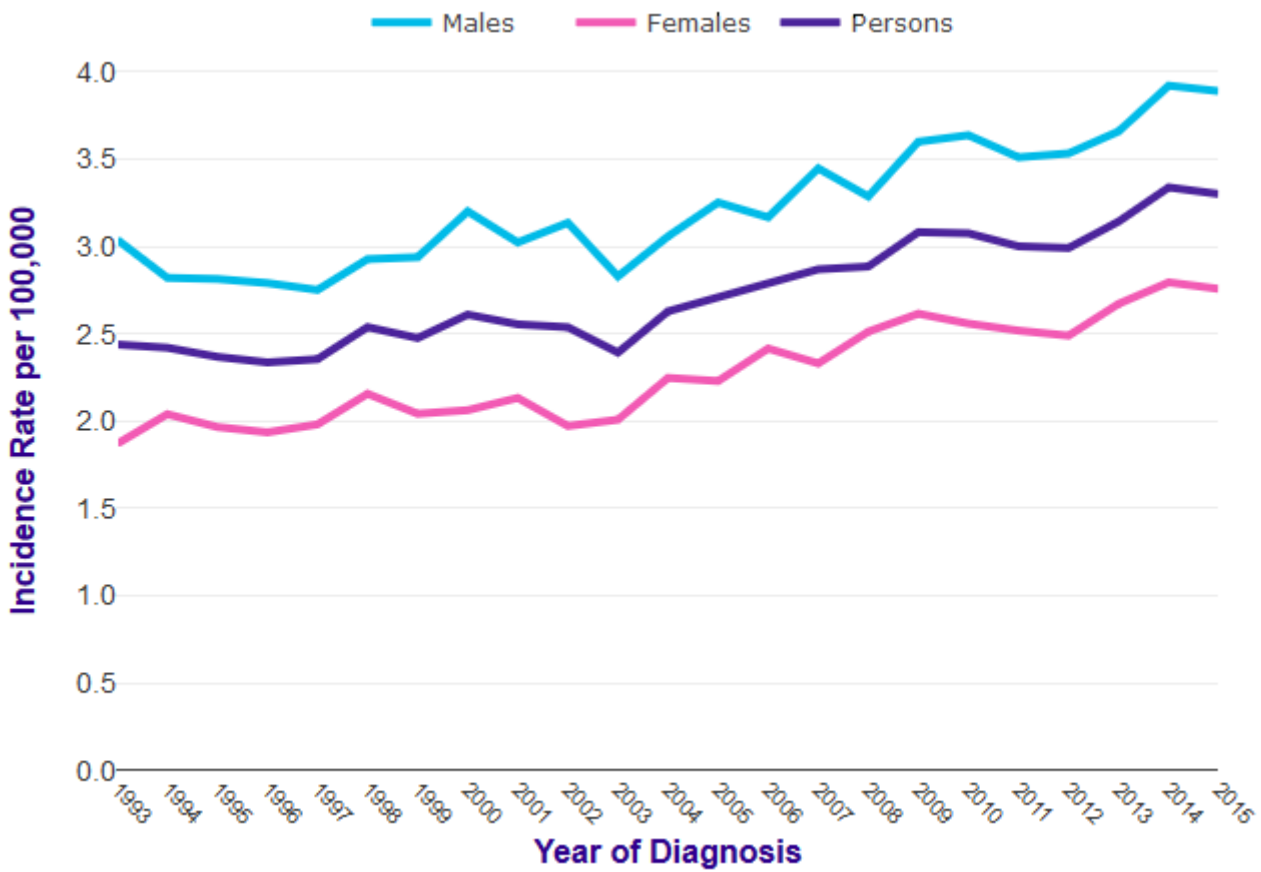

Figure 4. Age-standardised incidence rates of HL in the UK between 1993 and 2015. This plot was obtained from Cancer Research UK (2).

1.1.2 Pathology of Hodgkin's lymphoma

HL is characterised by the presence of Hodgkin and Reed-Sternberg (HRS) cells, which are derived from B lymphocytes. These cells have a distinctive appearance compared to surrounding cells. They are abnormally large (30-50 microns) and are often multinucleated (24) (Figure 5). Despite being a defining pathological characteristic of the disease, HRS cells typically constitute only a small minority of the cells that comprise HL tumours. Surrounding them is a reactive inflammatory cellular environment that includes neutrophils, eosinophils, plasma cells and lymphocytes.

HL tumours are categorised into subtypes based on HRS morphology and their surrounding cell composition, as well as immunohistochemistry (25,26). As such, there are two main categories of HL: classical Hodgkin's lymphoma (cHL) and nodular lymphocyte predominant Hodgkin's lymphoma (NLPHL). Classical Hodgkin's lymphoma (cHL) is further subdivided into four types: nodular sclerosis classical Hodgkin's lymphoma (NSCHL), mixed cellularity classical Hodgkin's lymphoma (MCCHL), lymphocyte rich classical Hodgkin's lymphoma (LRCHL) and lymphocyte depleted classical Hodgkin's lymphoma (LDCHL). The most common subtype is NSCHL which accounts for approximately 70% of all cHL cases in Western Europe and North America (26). NSCHL is characterised by neoplastic lacunar type HRS cells surrounded by an inflammatory environment of band-forming sclerosis. It is usually diagnosed at an early stage, when only one lymph node is affected.

MCCHL accounts for 20-25% of cHL cases, but is more frequent in developing countries, as well as in individuals with weakened immune systems. In this setting, HRS cells are surrounded by a mixed inflammatory environment without band-forming sclerosis (27). LRCHL comprises approximately 5% of all cHL cases. It is characterised by small lymphocytes and very few HRS cells. Finally, LDCHL represents the rarest subtype of cHL, accounting for less than 1% of all cHL cases in Western Europe and North America. This subtype is characterised by the presence of HRS in a relatively low inflammatory environment. LDCHL is often associated with HIV infection and tends to be a more aggressive form of the disease compared to other subtypes of cHL.

Upon diagnosis, HL tumours are staged-based on the Ann Arbor descriptive system (28). This system classifies tumours into limited (stages I and II, both non-bulky) or advanced (stage II bulky and stages III and IV). The term bulky is used to describe tumours that are at least ten centimetres (four inches) long. The proportions of cases diagnosed at each stage are relatively similar (Table 1) (2). Approximately half of all cases present with stage I or II disease.

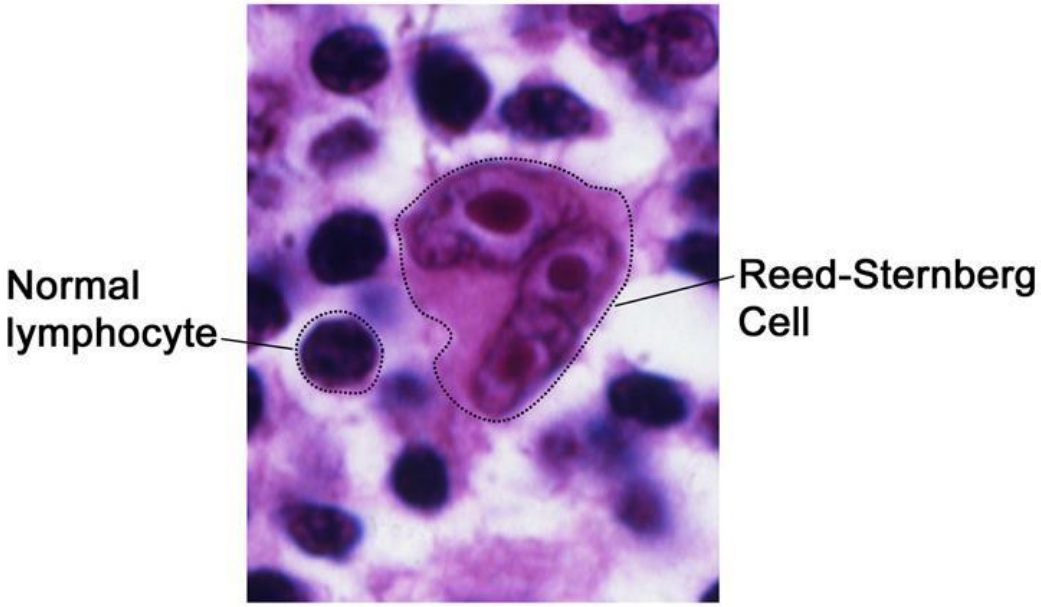

Figure 5. Hodgkin and Reed-Sternberg (HRS) cell. The HRS cell is multinucleated and abnormally large compared to a normal lymphocyte. This image was obtained from the National Cancer Institute (NCI) (29).

Table 1. Proportion of cases diagnosed at each HL stage in the UK between 2004 and 2008. This data, which was obtained from Cancer Research UK, reflects the proportions for all patients diagnosed between the age of 15 and 99 (2).

Stage at diagnosis	Proportion of cases
Stage I	24.4%
Stage II	30.8%
Stage III	15.4%
Stage IV	12.8%
Stage not known	16.7%

1.1.3 Treatment of Hodgkin's lymphoma

During the past 50 years, HL has progressed from being a largely fatal and incurable disease to becoming one of the most curable of all cancers. Prior to the development of combination chemotherapy (discussed later), the five-year survival rate for HL was less than 10% (30). In contrast, the ten-year survival rate in the UK is currently 80% (2). This large increase is due, predominantly, to the improvement of treatment strategies for HL.

Radiotherapy and chemotherapy have both been used to treat HL for many decades. Radiotherapy has been used for treatment since the 1900s but the manner in which it is used has considerably evolved over the decades, as a better understanding of the spread patterns of ionising radiation and their side-effects have accumulated (Figure 6). The development of the high-linear accelerator in the 1950s has enabled radiotherapy to be applied to increasingly precise areas. Total nodal radiotherapy, which involved the irradiation of zones of the lymphatic system was widely used during the 1960s. It was subsequently replaced by extended field radiotherapy, which involves irradiation of more precise areas of the lymphatic system (e.g. supradiaphragmatic or infradiaphragmatic). Both of these techniques were largely successful, enabling many patients to be successfully treated for HL, including individuals with advanced disease (31,32). However, despite these early successes, it became apparent that extended field radiotherapy caused long term complications including respiratory conditions, a weakened

immune system and infertility (discussed in Section 1.1.3) (33,34). To address the growing need for safer treatments, involved field radiotherapy and involved site radiotherapy were developed. The first targets a zone surrounding the affected node(s) while the latter enables even more precise targeting and aims to minimize irradiation of surrounding organs (Figure 6). Both techniques take advantage of Position Emission Tomogrophy (PET) scans to precisely direct the radiotherapy.

Chemotherapy has also been extensively used to treat HL. Early attempts, which yielded relatively poor responses, involved using single agents such as cyclophosphamide, chlorambucil or mechlorethamine (30). The first significant advance in chemotherapy came in the 1970s, with the development of combination therapy. The first combination therapy was the MOPP regimen (mustargen, oncovin, procarbizine, prednisone). It was used for newly diagnosed HL patients and resulted in remission rate of 81% (35). MOPP proved to be more effective than extended field radiotherapy for early stage disease (35). However, MOPP-related complications were significant and included bone marrow damage. This led to the development of a new regimen called ABVD (adriamycin, bleomycin, vinblastine, dacarbazine). While the success of MOPP and ABVD were similar, the reduced toxicity of the ABVD regimen led it to becoming the standard regimen used to treat HL (36).

The decision to use radiotherapy and/or chemotherapy is dependent upon the stage of HL, the subtype and prognosis factors such as whether or not the disease is bulky. After growing concerns of radiation-related toxicity and the development of ABVD, clinicians began to explore treatments that combined both, with the ultimate goal of minimising radiotherapy use by substituting it with chemotherapy. This approach was used for patients with early disease (37). In comparison, advanced HL is now often treated with chemotherapy alone (38). Chemotherapy is also used and followed by hematopoietic stem cell transplantation in patients who have relapsed (39). In more recent years, a number of alternative treatment approaches have also been explored to treat HL. One approach has been to target PD-1, a cell surface receptor involved in programmed cell death that is expressed by HRS cells and their surrounding inflammatory cells. Targeting PD-1 has been shown to significantly affect HRS cells and early clinical responses are promising (40). Another approach under development involves using brentuximan vedotin, an antibody drug conjugate that targets CD30, which is expressed on the surface of HRS cells, thereby inhibiting their proliferation and resulting in apoptosis (41,42).

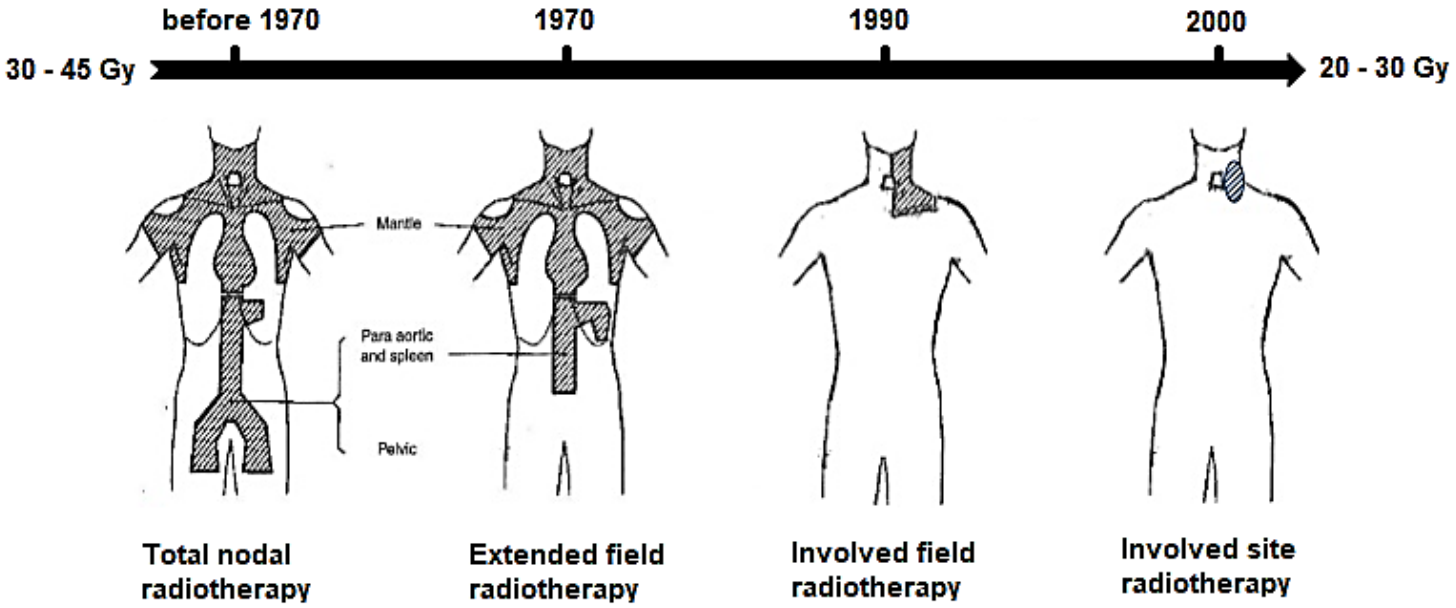

Figure 6. Evolution of the use of ionising radiation for treating HL. This diagram represents the four main radiation techniques used to treat HL. The shaded areas represent the irradiated zones. The black arrow represents the approximate chronology of when each technique was introduced, as well as the approximate evolution of dosages used (in practice, dosages vary according to the tumour stage, age of the patient, etc.). Source unknown.

1.1.4 Prognosis of Hodgkin's lymphoma and adverse effects of treatment

For newly incident HL cases, survival rates are very high. This is most likely due to the development of increasingly effective and safer treatments and also to factors such as improved diagnosis of the disease and more precise methods for classification of HL subtypes. The overall five-year survival rate for HL is currently 85% in England and Wales (2). Rates vary with age at HL diagnosis and HL subtype. Five-year survival rates are highest for patients treated at less than 39 years of age and they then decrease with increasing age at diagnosis. In addition, patients with LDCHL and MCCHL have a significantly worse diagnosis compared to patients who have NSCHL. Patients with LRCHL have the best prognosis (43).

Due to the high survival rates, adverse effects of treatment are common in HL survivors. Many treatment-related complications occur years after treatment for HL. HL survivors have a high risk for coronary heart disease which presents a median of 19 years after HL diagnosis and is thought to be attributable to the use of radiotherapy during HL treatment (44). Survivors are also known to suffer from congestive heart failure and valvular dysfunction, which are most likely caused by a combination of the use of radiotherapy and anthracycline chemotherapy agents such as adriamycin (which is part of the chemotherapy regimen ABVD) (45). The use of anthracyclines has also been associated with non-ischaemic cardiomyopathy in HL survivors (46). Moreover, hyperthyroidism and Graves' disease, conditions that both affect

the thyroid, have been associated with the use of bleomycin (which is part of the chemotherapy regimen ABVD) in patients (47). Another important adverse effect of HL treatment is infertility. This is due to the use of alkylating chemotherapy agents such as those in MOPP that can impair gonads, leading to infertility (48). Since the ABVD regimen does not contain any alkylating agents, it is not associated with such adverse effects. In rare cases, HL treatment has also been associated with a number of neurological disorders including dropped head syndrome, a disorder characterised by patients having difficulties lifting their head. Another disorder, brachial plexopathy, is characterised by shoulder and arm pain followed by sensory loss. Both dropped head syndrome and brachial plexopathy have been associated with mantle radiotherapy (49,50). Mantle radiotherapy has also been associated with episodic neurological dysfunction in which patients suffer from visual disturbances, monocular visual loss and hemianopia (blindness over half the field of vision) (51). Chemotherapy agents such as adriamycin, vinblastine and mechlorethamine have also been associated with a range of neurological complications, including cerebral infarction, encephalopathy and hearing loss (52,53).

1.2 Second cancers in Hodgkin's lymphoma survivors

Amongst the most severe adverse effects of HL treatment is the development of second cancers in long-term survivors: haematological neoplasms and

solid tumours are the biggest cause of mortality in these individuals (54,55). Second cancers are defined as malignancies that develop in a different organ to the original cancer. While not directly related to the original cancer (i.e. not resulting from metastasis), they are believed to occur as a consequence of treatment for the original cancer. This is evidenced by that fact that individuals who have had one cancer are more likely to develop a second cancer compared to the general population. Indeed, the standardised incidence ratio (SIR) for any solid cancer in HL survivors has been estimated to be 4.2, with a 30-year cumulative incidence of 29% (56). Risks for thyroid cancer, oesophageal cancer, lung cancer, and leukaemia are between five to ten times higher in HL survivors when compared to the general population. The risk of a second cancer remains high for at least 35 years following completion of HL treatment (SIR for ≥ 35 years = 3.9 [2.8 - 5.4]). Importantly, despite the development of more effective and safer treatment techniques over the last decades, the cumulative incidences of second solid cancers do not significantly differ between patients treated in the 1970s compared to those treated in the 1990s (Figure 7).

Despite occurring almost exclusively in women, breast cancer (BC) is the most common second cancer in HL survivors, with a 30-year cumulative incidence of 16.6% (56). Women treated for HL at an early age and having received supradiaphragmatic radiotherapy as part of their treatment have the highest risk of developing BC (57). More specifically, women who received supradiaphragmatic radiotherapy as their sole treatment have been estimated to have a relative six-fold increased risk of BC compared to women

from the general population. This high risk is almost as high as that of *BRCA* mutation carriers (58). There is a direct relationship between the risk of developing BC following HL and the dose of radiotherapy used to treat HL: the higher the dose of radiotherapy received, the higher the risk of developing BC (57). This suggests that the BC is radiation-induced. Indeed, BC risk has been shown to be twice as high in patients treated with mantle radiotherapy compared to those who received treatment in other areas of the body. Similarly, risks of BC are greater in women having received radiotherapy in two areas of the mantle compared to women who only received treatment in one area of the mantle (57). Cumulative risk estimates show that the risk of developing BC following HL remains high for up to 40 years following treatment for HL. In particular, for a follow-up for 40 years, the cumulative of patients having received over 40Gy of mantle radiotherapy is 48% (57). Given these high risks and that new cases are still arising amongst HL survivors, long-term monitoring is warranted.

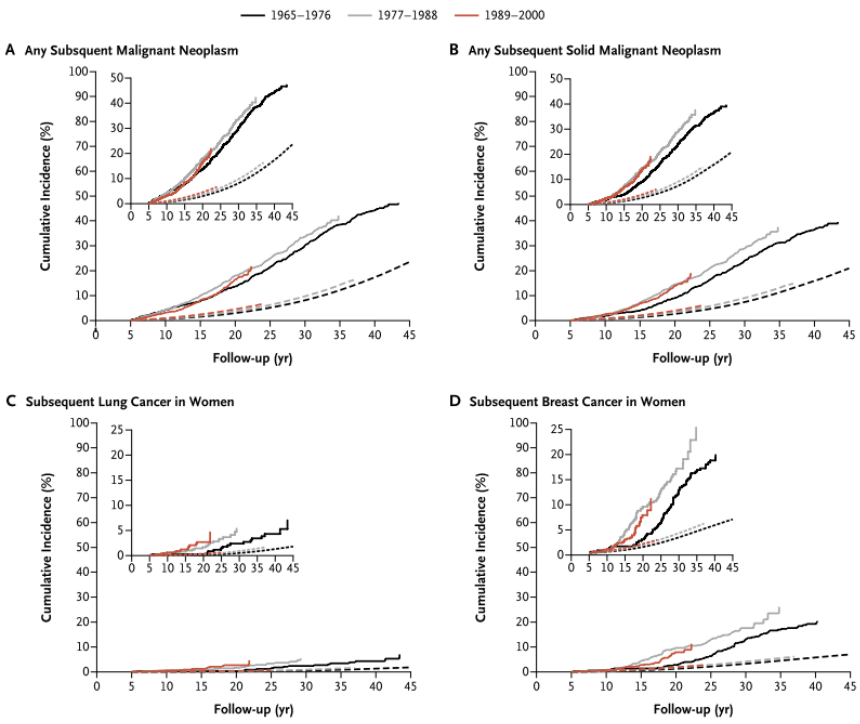

Figure 7. Cumulative incidence of second cancers in HL survivors according to treatment period, with death as competing risk. The data used for this study was obtained from a Dutch cohort comprised of 2207 men and 1968 women who had all been treated for HL and had survived for at least five years following treatment. The majority of patients (60.5%) were treated with both radiotherapy and chemotherapy. A total of 27.3% were treated with radiotherapy only and a total of 12.1% were treated with chemotherapy only. Samples were sorted into treatment period categories in a manner that would make them approximately equally sized and that would approximately reflect the changes in HL treatment over the decades (i.e. smaller dosages of radiotherapy and increase of combination with chemotherapy). The solid lines represent the observed incidence and the dashed lines represent the expected incidence in the general population. The inset graphs represent the same data on enlarged y-axes. Modified from Schaapveld *et al* (56).

1.3 Radiation-induced breast cancer

1.3.1 Non-genetic risk factors of radiation-induced breast cancer

A number of epidemiological studies have investigated the non-genetic risk factors that underlie predisposition to HLBC. As discussed in Section 1.2, increasing dose of radiotherapy greatly increases the risk of developing BC following treatment for HL. Notably, the risk conferred by ionising radiation is reduced when radiotherapy is used in conjunction with alkylating chemotherapy or pelvic radiotherapy (Table 2). This is due to the fact that alkylating chemotherapy and pelvic radiotherapy both damage ovaries, an event which, aside from leading to infertility, is associated with early menopause and thereby provides a protective effect against BC risk (59–61).

Aside from dose, age at treatment also plays an important role in HLBC risk. Women treated at a young age have the highest risk of developing HLBC, with risks being highest for women treated during puberty (62–65). Relative risks for HLBC are raised five-fold for children treated at the age of ten and 22-fold for women treated between the ages of ten and 14 (Table 3) (57). This peak of risk coinciding with puberty is thought to be due to the fact that radiation exposure to proliferating breast tissue during thelarche is particularly harmful.

Finally, the role of menarche status and pregnancy have also been investigated in the context of HLBC. Women treated with radiotherapy close to their age of menarche (both before and after) have been found to have raised BC risks, with women treated within six months of menarche having the highest risk (Table 4) (61). In addition, while it has previously been suggested that HLBC risk increased for women treated for HL during a first-time pregnancy, more recent evidence has suggested this is in fact not the case (61,66).

Understanding the non-genetic risk factors underlying HLBC has led to increased awareness of how treatments should be modified and which groups of HL survivors are at high risk of developing BC. However, the knowledge of these factors alone does not fully explain why some HL survivors go on to develop BC and some do not. This has led to speculation that HLBC may be underlain by genetic risk factors.

Table 2. Risks of BC in HL survivors in relation to treatment modality. Modified from Swerdlow *et al* (57). The data used for this study is from the National Recall Study (NRS study). Data was obtained for 5002 women treated for HL with supradiaphragmatic radiotherapy, of whom 373 later developed breast cancer. Abbreviations: AER = absolute excess risk per 10,000, RT = radiotherapy, CT = chemotherapy, SIR = standardised incidence ratio, u/k = unknown.

Treatment modality	SIR	95% CI	AER	95% CI
Supradiaphramatic RT	6.0	5.2 - 7.0	42.9	35.3 - 51.4
Supradiaphramatic RT plus alkylating CT	4.8	4.0 - 5.6	28.0	22.4 - 34.3
Supradiaphramatic RT plus ≥ 5 Gy pelvic RT	1.4	0.5 - 4.4	4.7	-8.2 - 35.8
Supradiaphramatic radiotherapy plus alkylating CT plus ≥ 5 Gy pelvic RT	3.8	2.4 - 6.1	32.2	14.0 - 58.5
Supradiaphramatic radiotherapy plus u/k if alkylating CT and/or ≥ 5 Gy pelvic RT	4.2	3.1 - 5.5	30.6	20.1 - 43.6

Table 3. Risks of BC in HL survivors in relation to age at first supradiaphragmatic radiotherapy. Modified from Swerdlow *et al* (57). The data used for this study is from the National Recall Study (NRS study). Data was obtained for 5002 women treated for HL with supradiaphragmatic radiotherapy, of whom 373 later developed breast cancer. Abbreviations: AER = absolute excess risk per 10,000, SIR = standardised incidence ratio.

Age at treatment	SIR	95% CI	AER	95% CI
0 - 9	5.4	0.8 - 38.2	7.5	-1.5 - 49.4
10 -14	22.0	14.9 - 32.6	65.1	41.1 - 97.6
15 - 19	14.3	12.0 - 17.2	58.6	47.7 - 71.1
20 - 24	5.4	4.5 - 6.6	30.7	23.7 - 38.8
25 - 29	3.2	2.5 - 4.0	22.1	15.1 - 30.5
30 - 35	2.4	1.8 - 3.0	19.4	11.4 - 29.2

Table 4. Risks of BC in HL survivors in relation to menarche timing and supradiaphragmatic radiotherapy. Modified from Cooke *et al* (61). The data used for this study is a subset of the National Recall Study (NRS study). Data was obtained for 2497 women treated for HL with supradiaphragmatic radiotherapy, of whom 260 later developed breast cancer.

Duration between menarche and supradiaphragmatic radiotherapy	Cases	Controls	Total	OR	95% CI	P-value
Supradiaphragmatic radiotherapy at least 5 years before menarche	1	10	11	0.94	0.10 - 8.46	0.95
Supradiaphragmatic radiotherapy 2 - 5 years before menarche	6	16	22	4.08	1.27 - 13.14	0.02
Supradiaphragmatic radiotherapy 0.5 - 2 years before menarche	7	12	19	4.90	1.60 - 14.98	0.005
Supradiaphragmatic radiotherapy within 6 months of menarche	9	14	23	5.52	1.97 - 15.46	0.001
Supradiaphragmatic radiotherapy 0.5 - 2 years after menarche	9	29	38	3.47	1.40 - 8.58	0.007
Supradiaphragmatic radiotherapy 2 - 5 years after menarche	45	200	245	2.38	1.43 - 3.97	0.001
Supradiaphragmatic radiotherapy 5 - 10 years after menarche	67	577	644	1.33	0.89 - 1.98	0.16
Supradiaphragmatic at least 10 years after menarche	99	1271	1370	1.00	NA	NA
Never had radiotherapy	1	6	7	2.14	0.20 - 22.56	0.53
Unknown timing of menarche	16	102	118	1.67	0.90 - 3.11	0.11

1.3.2 Genetic risk factors of radiation-induced breast cancer

While genetic susceptibility has been widely researched in sporadic and familial BC, the genetics of HLBC are largely understudied. This is predominantly due to a paucity of studies with the requisite data and biological material. A small number of studies have nevertheless attempted to analyse the effects of common loci in the context of HLBC germline predisposition (Table 5).

One of the first Single Nucleotide Polymorphisms (SNPs) to be associated with HLBC was rs1219648, which maps to the fibroblast growth factor receptor 2 (*FGFR2*) locus at 10q26.13 (Table 5). In a pooled analysis of two case-control studies totalling 232 HLBC cases and 461 controls, Ma *et al* inspected associations of known sporadic female BC predisposition SNPs (14 SNPs at the time of their study) with HLBC (67). The allele frequencies of rs1219648 were found to be significantly different between HL survivors who developed BC and HL survivors who did not. The minor allele of rs1219648 was overrepresented in cases and its effect was considerably larger than in sporadic BC (for HLBC: odds ratio (OR) = 1.59 [1.26-2.02], P-value = 1.11×10^{-4}; for sporadic BC: OR = 1.26 [1.13-1.40], P-value = 1.50×10^{-5}), although confidence intervals were large. This association between rs1219648 and risk of HLBC constitutes one of the first observations that established FBC loci also influence HLBC susceptibility and that their effects may be different

in the context of radiation exposure (68). The risk conferred by rs1219648 was particularly high in women treated before the age of 20 (OR = 1.70 [1.16-2.50]), as well as in women who had not received alkylating chemotherapy or pelvic radiotherapy (OR = 1.79 [1.25-2.57]). Intriguingly, copy loss of *FGFR2* has been shown to confer increased resistance to radiotherapy in prostate cancer (69).

A second study of HL survivors looked at associations of common SNPs with second cancers (including but not limited to BC) in a genome-wide association study (GWAS) of 158 cases and 153 controls (70) (Table 5). Cases were defined as individuals having had HL followed by a second cancer (including but not limited to BC). This GWAS identified two SNPs that were associated with second cancers: rs4946728 and rs1040411, both of which localise to 6q21 and form a haplotype in the vicinity of the tumour suppressor positive regulatory domain zinc finger protein 1 (*PRDM1*). The SNPs conferred large effects (rs4946728 OR = 3.32 [2.25-4.90], P-value = 5.99×10^{-10}; rs1040411 OR = 2.39 [1.73-3.30], P-value = 1.18×10^{-7}). These associations did not differ between BC and other second cancers (P_{het} = 0.41 for rs4946728 and P_{het} = 0.58 for rs1040411), nor between males and females (P_{het} = 0.83 for rs4846728 and P_{het} = 0.29 for rs1040411). Quantitative trait locus (eQTL) was performed in order to determine whether the SNPs were associated with the expression of *PRDM1* and any other gene within five megabases. This revealed that increasing risk allele dosages of both SNPs were significantly associated with lower expression levels of *PRDM1*. Upon treating cells lines that were homozygous for either the risk haplotype or the

protective haplotype with ionising radiation, it was shown that the levels of PRDM1 protein was not induced in cells homozygous for the risk haplotype (P-value = 0.19).

More recently, Morton *et al* performed a GWAS of radiation-induced BC in a cohort of survivors of various childhood cancers including (but not limited to) HL that comprised 207 cases and 2774 controls that detected three SNPs that were associated with radiation-induced BC (71) (Table 5). The first, rs4342822 located at 1q41, was associated with an almost two-fold raised risk of BC in patients who received at least 10 Gy radiation during their treatment (hazard ratio (HR) = 1.92 [1.49- 2.44], P-value = 7.09×10^{-9}). The authors also found evidence of associations between two rare variants, rs74949440 and rs17020562, mapping to 11q23 and 1q32.3 respectively, and risk of radiation-induced breast cancer. Similarly to rs4342822, the effect of both SNPs varied according to radiation exposure: rs74949440 was associated with risk only in women who received at least 10 Gy (HR = 2.59 [1.62-4.16), P-value = 5.84×10^{-8}) while rs17020562 was associated with risk only in women who received less than 10 Gy (HR = 44.52 [15.06-131.62], P-value = 6.68×10^{-8}).

While these previous studies represent an important advance in the understanding of HLBC susceptibility, they suffer from a number of limitations. Firstly, all three studies lack statistical power due to their small sample sizes. This is further evidenced by that the fact that both GWAS

studies (Best *et al* and Morton *et al*) do not replicate each other's findings or the SNP identified by Ma *et al*. While this does not completely invalidate their results, it does limit the credibility of the findings of all three studies. Secondly, while Ma *et al* and Best *et al* replicate their results in independent datasets, this is not the case in the GWAS performed by Morton *et al.* Such findings will need to be validated in order to be more certain of the contribution of these SNPs. In addition, one of the SNPs identified by Morton *et al*, rs17020562, is rare (MAF = 0.0005). Given the fact that a GWAS of the size reported by Morton *et al* are not sufficiently powered for testing associations with rare SNPs, and the fact that the HR identified for this SNP is strikingly high (HR = 44.52), this finding should be interpreted with caution. Thirdly, the design of the three studies is not entirely suitable for studying HLBC. While Best *et al* and Morton *et al* do use a broad approach in their study, their analyses are limited by the lack of homogeneity in their datasets. Indeed, both are comprised of HL survivors who have developed various second cancers which include BC. This approach means that they are analysing genetic susceptibility to second cancers in HL survivors, rather than just HLBC. Although the study of second cancers in general is an important area of research, the heterogeneity of the samples likely adds noise to the datasets, resulting in a loss of statistical power. With new cases of HLBC still arising, and no evidence of decreasing risks in patients treated in the 1990s and 2000s (as discussed in Section 1.2), there remains a need to further investigate HLBC in order to obtain a better understanding of this disease. Knowledge of the variants underlying HLBC would not only further our understanding of this neoplasm but also facilitate improved stratifications of

BC risk in HL survivors. This in turn would help devise prophylactic interventions in HL survivors with particularly high risks of BC. In this study, I aim to further characterise the genetic profile of HLBC and provide further insights into the aetiology of this disease.

Table 5. Summary of SNPs previously found to be associated with HLBC. The data for sporadic BC was obtained from BCAC (72,73).

Study	No. cases	No. controls	Cytoband	rsid	Nearest Gene	Minor allele	Major allele	Risk allele	MAF	MAF cases	MAF controls	HR/ OR	95% CI	P-value	Sporadic BC OR	Sporadic BC 95% CI	Sporadic BC P-value
Ma *et al* 2012	232	461	10q26.13	rs1219648	*FGFR2*	G	A	G	0.38	0.46	0.34	1.59	1.26-2.02	1.11×10^{-4}	1.26	1.25-1.28	5.14×10^{-302}
Best *et al 2011*	158	153	6q21	rs4946728	*PRDM1*, *ATG5*	A	C	A	0.75	0.85	0.64	3.32	2.25-4.90	5.99×10^{-10}	1.01	0.99-1.02	0.37
Best *et al 2011*	158	153	6q21	rs1040411	*PRDM1*, *ATG5*	A	G	A	0.55	0.65	0.44	2.39	1.73-3.30	1.18×10^{-7}	1.00	0.99-1.01	1.00
Morton *et al* 2017	207	2774	1q41	rs4342822	*PRXO1*	G	T	G	0.47	0.66	0.46	1.92	1.49-2.44	7.09×10^{-9}	1.01	0.99-1.02	0.31
Morton *et al* 2017	207	2774	11q33.2	rs74949440	*TAGLN*	T	C	T	0.025	0.09	0.02	2.59	1.62-4.16	5.84×10^{-8}	1.02	0.98-1.06	0.34
Morton *et al* 2017	207	2774	1q32.3	rs17020562	*RPS6KC1*	C	T	C	0.0032	0.04	0.0005	44.52	15.06-131.62	6.68×10^{-8}	1.00	0.90-1.13	0.90

1.4. Aims and objectives

The purpose of this thesis is to characterise the germline risk factors of HLBC in order to gain a better understanding of the aetiology of this disease. The insights obtained will add to the understanding of HLBC and help design effective prevention and treatment strategies for future HL patients. This project bases itself on the largest cohort of HL patients worldwide. This dataset therefore provides me with a unique opportunity to ratify previous findings as well as discover new potential associations.

This project has two main objectives. The first is to assess the role of common germline variants in HLBC susceptibility. To address this, I performed a GWAS to identify novel SNPs associated with HLBC. Additionally, I investigated whether a BC Polygenic Risk Score (PRS) was associated with HLBC. My second objective was to assess whether rare germline variants in DNA repair genes are involved in HLBC susceptibility by performing targeted Next-Generation Sequencing (NGS) of over 200 genes with roles in DNA repair.

Chapter 2 - Polygenic susceptibility to radiation-induced breast cancer

2.1 Introduction

Genetic susceptibility to sporadic and familial BC has been extensively studied. To this date, a total of 172 loci have been associated with sporadic female BC and account for approximately 18% of familial risk (73,74). While some of these loci were identified with candidate gene and candidate variant studies, most were discovered after the development of large collaborative consortiums such as the Breast Cancer Association Consortium (BCAC), which enabled studies using thousands of samples (75–78). Particular successes include the identification of five BC risk loci (locating to *FGFR2*, 8q24, *LSP1*, *TNRC9*, *MAP3K1*) in one the first large GWAS comprised of 21,860 cases and 22,578 controls, performed by Easton *et al* (77). This was soon followed by the discovery of many other variants, often using meta-analyses that combined different studies.

As more SNPs were found to be associated with sporadic BC, it became clear that BC is underlain by both common variants with small effects, rare variants with large effects and a spectrum of low-frequency variants with intermediate effects (Figure 8) (79). When considering HLBC in this context, this raises the question as to whether radiation-included BC is also underlain by both

common and rare variants. Moreover, the discovery of rare variants led to the gradual realisation that a number of FBC loci are associated with DNA repair. Examples include *ATM*, *BRCA1, BRCA2*, *RAD50*, etc (80–83). Given the fact that HL treatment most always includes supradiaphragmatic radiotherapy, it would seem that any mutation in genes that are involved in repairing the DNA damage caused by the radiation would cause high risks of carcinogenesis. I therefore hypothesise that some HL women may be genetically predisposed to developing BC following radiation exposure and that some of these predisposing variants may be in radiation response genes, including DNA repair genes.

This chapter presents two analyses designed to identify common germline variants that are associated with HLBC. The first analysis is a GWAS of women who developed BC subsequent to a diagnosis of HL earlier in life. Cases were women who developed BC or Ductal carcinoma in situ (DCIS) following treatment for HL. Controls were HL survivors who did not develop BC. This dataset totals 766 samples, of which 261 are cases and 505 are controls. While previous studies have already attempted to find such variants, my dataset constitutes an opportunity to confirm previous findings, but also to detect new associations that were not observable in other cohorts. This is due not only to its larger size, but also to a lack of heterogeneity in the clinical background of participants. Unlike the previous GWAS performed by Morton *et al* which was based on a cohort of patients having undergone various childhood cancers (including but not limited to HL), my dataset is comprised of women who have all been treated specifically for HL and who have all

received supradiaphragmatic radiotherapy to the mantle area (71). This lack of heterogeneity in clinical background increases the statistical power of my study to detect associations with radiation-induced BC.

The second analysis involves assessing the performance of a recently published BC PRS by Mavaddat *et al* in the context of HLBC (84,85). By using the information from loci known to be associated with sporadic FBC, Mavaddat *et al* showed that one could stratify BC risk in women both with or without a family history of BC. The PRS showed that the OR for BC per 1 standard deviation of PRS score was 1.61 [1.57-1.65] and that the lifetime risk women in the top percentile of the PRS was 32.6%. Importantly, women in the top 19% of the PRS reached a ten-year absolute BC risk of 2.6% by age 40, instead of 47 as in the general population. Given that 47 is the age at which women become eligible for the UK Breast Screening Program, these results indicate that these women are at high risk of developing BC and should begin screening earlier. Recently, some attempts have been made to incorporate further factors into PRS scores. For example, a recent study attempted to combine family history alongside genotypes for 77 known susceptibility SNPs to further stratify individuals (86). As additional susceptibility variants are discovered, PRS scores could be accurate enough to help clinicians identify high risk individuals and develop prevention strategies such as earlier screening strategies for high risk patients.

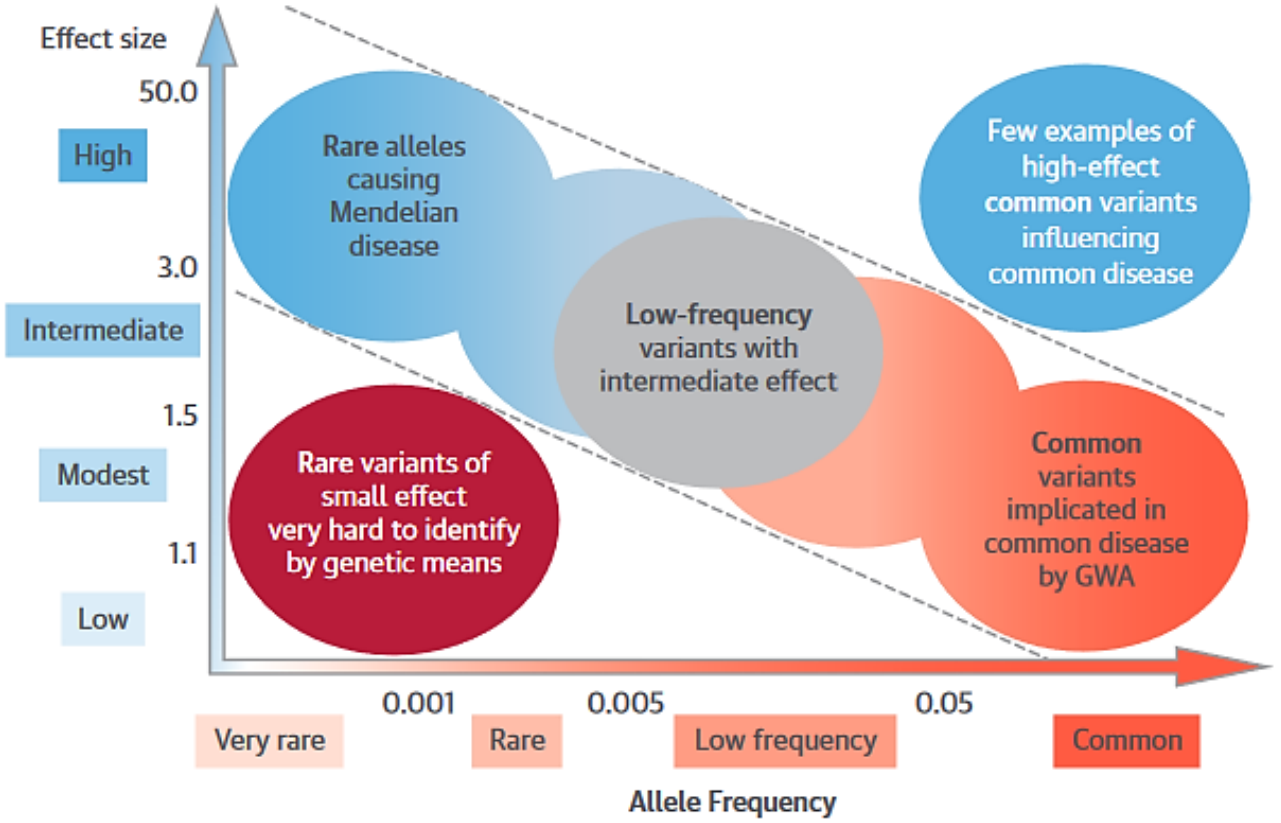

Figure 8. Genetic variant allele frequencies and their strength of genetic effect. This plot was obtained from from Assimes *et al* (79).

2.2 Materials and methods

2.1.1 Sample collection

The samples utilised for this project are from the National Study of Breast Cancer in Women Treated by Radiotherapy for Hodgkin's Lymphoma, also called the National Recall Study or NRS. The NRS began in 2003 and is ongoing (57). Its aims are to investigate the risk factors for HLBC in order to improve treatment decisions and inform prevention strategies. Participants are women from the UK who developed HL before the age of 36 and received mantle radiotherapy as part of their treatment for HL. A total of 5002 individuals have been recruited to the NRS. Each participant was provided with a detailed BC-orientated questionnaire about non-genetic and behavioural risk factors for the disease. Notably, this collected epidemiological data was used to study the non-genetic risk factors for HLBC mentioned in the Section 1.3.1 (Tables 2, 3 and 4). In addition, a subset of patients (915 individuals) also gave consent for providing blood samples. Ethylenediaminetetraacetic acid (EDTA) anti-coagulated blood samples were collected from these individuals and constitute the dataset for this project.

2.1.2 DNA extraction and genotyping

DNA extractions were performed by the ICR Complex Trait Genetics (CTG) and Molecular and Population Genetics (MPG) Teams. Blood samples were received in 9 ml EDTA tubes (3 tubes per patient) and were centrifuged to separate the plasma, buffy coat and erythrocytes. Buffy coats were obtained by aspiration, aliquoted into 300 µl cryostraws and were then stored in vapour phase LN2 until required. DNA extraction was performed using Qiagen QIAamp DNA mini kits (Qiagen) according to the manufacturer's instructions. DNA samples were quantified using Picogreen (Invitrogen) and were normalised to a final working concentration of 50 ng/µl in 200 mL of nuclease-free water. A total of 539 samples were genotyped using Illumina Infinium Human-600K arrays (Illumina) and 364 samples were genotyped using Illumina Infinium Oncorray-500K arrays (see Table 5 for numbers of cases and controls per array). The Illumina Infinium Human-600K array comprises 657,366 SNPs selected to maximise genome-wide coverage of common variants. The Illumina Infinium Oncorray-500K comprises 499,170 SNPs, of which 250,000 comprise a genome-wide backbone for imputation of unobserved variants, while the remainder were selected on the basis of association with breast, ovarian, lung, colorectal or prostate cancer. Illumina genotyping was outsourced to a commercially available service provided by the Wellcome Trust Centre for Human Genetics (Oxford, UK).

2.1.3 Quality control (QC)

Genotype data quality control (QC) was conducted in accordance to Laurie *et al* (87) and Turner *et al* (88). QC was performed separately for each genotyping array. The majority of QC analyses were performed with GLU (Genotype Library and Utilities) (http://code.google.com/p/glu-genetics), PLINK (89) and R (90). SNPs were clustered using Genome Studio (Illumina). SNPs with Gentrain scores (a statistical measure based upon cluster shape and relative distance between each cluster) of less than 0.15 were excluded. GLU was used to identify and remove SNPs and samples with completion rates of less than 95%. SNPs whose genotype frequencies differed significantly (P-value $< 1 \times 10^{-5}$) from those expected under the Hardy-Weinberg Principle (HWP) were identified using an exact test implemented in GLU (91). Since deviation from HWP might be expected for disease associated variants, SNPs that were in violation of HWP were flagged, but were not discarded.

A Principal Component Analysis (PCA) was conducted using the Eigensoft package (92). Samples in the study were compared to genotype data from the HapMap project corresponding to individuals of Western European, East Asian and West African ancestry (93). HapMap samples were subjected to QC in order to remove SNPs and samples with low completion rates. SNPs that were not present in both the study dataset and the Hapmap dataset were removed, as were strand-ambiguous SNPs. Both datasets were then merged. The top ten principal components for each individual were obtained using the smartPCA function in Eigensoft. K-means cluster analysis was applied to the first two principal components in order to identify ancestry

outlier samples for exclusion. Concordance between samples was assessed with GLU. If unexpected duplicate samples were found, the sample with the highest completion rate was retained. Identity by Descent (IBD) analysis was used to identify and exclude related individuals. Samples with low levels of X chromosome heterozygosity (i.e. samples with suspected Y chromosomes) were identified and excluded with PLINK. In order to perform a high quality imputation of genotypes, datasets were further cleaned to retain only autosomal biallelic SNPs: indels, CNVs, SNPs with multiple mappings, SNPs on sex chromosomes were all excluded.

2.1.4 Imputation of genotypes

Imputation of unobserved genotypes was performed in order to increase the SNP density of my dataset. Due to the fact that the Illumina Infinium Human-600K and Illumina Infinium Oncorray-500K arrays share few SNPs in common (less than 30% of their total), the imputation was carried out separately on each dataset before merging them. Using the SHAPEIT software, genotyped SNPs were used to estimate haplotype frequencies in the dataset (94). Following this, IMPUTE2 was used to infer missing genotypes using 1000 Genomes project Phase 3 data as a reference (95). Imputation was performed separately on each dataset. The resulting datasets were then merged. Prior to merging the datasets, the coordinates of the Infinium Human-600K array dataset were converted from build 36 to build 37, so that both datasets had the same build. Following the merging of the

datasets, poorly imputed SNPs were discarded. The quality of imputation was assessed with the info score, provided by IMPUTE2 (96). This metric varies between 0 and 1, where 1 indicates that a SNP has been imputed with a high degree of certainty. SNPs with an info score below 0.3 were discarded. Finally, uncommon SNPs, defined as having a MAF (Minor Allele Frequency) of less than 2% were discarded.

2.1.5 Clinical data assessment

Since risk of HLBC is influenced by a number of non-genetic factors including age at HL treatment, radiotherapy dosage and the use of alkylating radiotherapy and/or pelvic radiotherapy, information on these factors was obtained in order to include these variables as covariates in association analyses. Participants' age at HL diagnosis was used as a proxy for age at HL treatment since it was not possible to obtain the exact dates of treatment. Although the timing of menarche (in relation to the timing of treatment for HL) plays a role in HLBC risk (as discussed in Section 1.3.1), information regarding menarche could not be obtained. Confirmation was sought as to whether each woman had been treated with radiotherapy for HL before the age of 40; this threshold was chosen because risk of HLBC is high in women that were treated before age 40, after which it drops considerably (97). Additionally, all cases must have had HL before BC and the BC must have occurred at least five years subsequent to the diagnosis of HL (the

assumption being that otherwise the BC may be sporadic and not radiation-induced). Subjects who did not meet these criteria were discarded from association analysis.

2.1.6 Association analysis

Association analyses were primarily conducted using Cox regression in ProbABEL (98). This type of cohort survival analysis addresses a specific issue about individuals who previously had HL: the idea that controls may become cases as time passes. It is analogous to conditional matched case-control analysis with matching on survival time. By assigning cases and controls into time categories based on their duration of follow-up, one can effectively avoid comparing samples that have completely different durations of follow-up. This analysis allows the comparison of cases and controls who have had similar durations of follow-up and therefore should have similar, if not comparable, risks of HLBC. The duration of follow-up was defined as the time between the date of HL diagnosis and the date of BC diagnosis for cases and the time between the date of HL diagnosis and the date of death (or the last date of follow-up, which is 01/01/17) for controls. As well as the duration of follow-up, the risk factors described in section 2.1.5 were also added into the Cox regression model: the age at HL diagnosis, the dosage of radiotherapy received and the use of alkylating radiotherapy and/or pelvic radiotherapy. Preliminary Cox regressions using just these four covariates (without the genotypes) indicated that the risk of HLBC was significantly

higher in the Illumina Infinium Human-600K dataset compared to the Illumina Infinium Oncorray-500K dataset, suggesting a batch effect was present. The genotyping array was therefore added as a covariate to the Cox regression model. To run the Cox regression in ProbABEL, each covariate was formatted into categories. Categorical variables, such as the genotyping array, were coded as simple binary variables (“0” and “1”). For continuous variables, such as the dosage of radiotherapy received, samples were sorted into categories in a manner that would make them approximately equally sized (Table 6).

2.1.7 Comparison of HLBC ORs and sporadic FBC ORs for known FBC loci

In accordance with Orr *et al*, the comparison of the ORs obtained in the GWAS to the published ORs of loci associated with sporadic FBC, was performed with a Chi-Squared test (99). This test is based on the assumption that both sets of ORs are log-normally distributed. It also assumes that the difference between the estimated log ORs is normally distributed, with a mean of zero and variance equal to the sum of the squared standard errors of the two estimates. Using this method, a Chi-Squared test statistic was obtained for each OR comparison as well as for all combined comparisons.

Table 6. Covariate categories used for the Cox regression association testing. The number of samples in this table represent the numbers obtained after QC (see Section 2.3.2 for details of sample exclusions).

Covariate	Categories	No. cases	No. controls	Total of samples	Percentage of total samples (%)
Genotyping array	"0": Illumina Infinium Oncorray-500K	96	243	339	44.3
	"1" : Illumina Infinium Human-600K	165	262	427	55.7
Age at HL diagnosis	"0": 5-14 years old	17	21	38	5.0
	"1": 15-19 years old	86	124	210	27.4
	"2": 20-24 years old	74	172	246	32.1
	"3": 25-29 years old	44	111	155	20.2
	"4": 30-35 years old	40	77	117	15.3
Dosage of radiotherapy received	"999": unknown	29	54	83	10.8
	"0": 6-34 Gy	35	78	113	14.8
	"1": 35-42 Gy	140	258	398	51.9
	"2": ≥ 43 Gy	57	115	172	22.5
Alkylating chemotherapy and/or pelvic radiotherapy	"0": No	116	119	315	41.1
	"1": Yes	145	306	451	58.9
Duration of follow-up	"0": 6-19 years	72	14	86	11.2
	"1": 20-24 years	51	44	95	12.4
	"2": 25-29 years	64	107	171	22.3
	"3": 30-35 years	43	120	163	21.3
	"4": 35-39 years	22	105	127	16.6
	"5": 40-55 years	9	115	124	16.2

2.1.8 Polygenic Risk Score (PRS) construction

The PRS was constructed based on the 313 SNP model of Mavaddat *et al* for sporadic BC (85). This PRS is comprised of three types of SNPs. Firstly, it comprises the 172 published FBC predisposition SNPs. Secondly, SNPs that are correlated with the 172 published SNPs and which, based upon stepwise forward regression analysis, contribute additional information in addition to the published variants are also included. Finally, SNPs that are specifically associated with either oestrogen-positive or oestrogen-negative BC are included. The formula used for calculating the scores is as follows:

$$\mathbf{PRS = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \beta_n x_n}$$

In this formula, beta (β) represents the per-allele log odds ratio (OR) for BC associated with the minor allele for SNP_k and x_k is the number of alleles for the same SNP (0, 1 or 2). To avoid overfitting, the published ORs for sporadic BC were used to calculate the PRS. Genotype posterior probabilities from IMPUTE2 were converted to risk allele dosages for the purpose of calculating the PRS. Samples were sorted into percentiles according to their PRS score. Logistic regression was then performed to assess the association between percentiles of PRS score and risk of HLBC. In accordance with Mavaddat *et al* (84), the following percentile groups were used: <1%, 1-5%, 5-10%, 10-20%, 20-40%, 40-60% (reference group), 60-80%, 80-90%, 90-95%, 95-99% and >99%. The PRS distributions in HLBC cases and controls were

compared to those of sporadic BC cases and controls from the Generations Study (100).

2.3 Results

2.3.1 Genotyping quality control

Inspection of Gentrain scores and locus completion rates indicated that the Infinium Human-600K dataset had a larger number of poorly performing SNP assays (Table 7). A total of 1,074 SNPs had genotype proportions that deviated significantly from those expected under Hardy-Weinberg equilibrium (235 from the Infinium Human-600K dataset and 839 from the Infinium Oncorray-500K dataset). PCA analysis identified a total of 22 ancestry outliers in the Infinium Human-600K dataset and 13 ancestry outliers in the Infinium Oncorray-500K dataset (Figure 9). A total of 3 samples (all from Infinium Human-600K dataset) were identified as being related (Figures 10 and 11) and one sample (from the Infinium Oncorray-500K dataset) displayed unexpected low levels of X chromosome heterozygosity. No samples displayed low or high levels of heterozygosity (which can be an indication of sample contamination or inbreeding).

2.3.2 Sample exclusions

Assessment of the clinical data revealed that a number of samples did not fulfil the criteria required by this project. A total of 24 samples were removed from the Infinium Human-600K dataset and a total of 16 samples were

removed from the Infinium Oncorray-500K dataset (Table 8). Following QC, a total of 766 samples (261 cases and 505 controls) and 6,788,016 SNPs were available for association testing.

Table 7. Summary of quality control results for SNPs (prior to imputation).

	Infinium Human-600K	**Infinium Oncorray-500K**
Pre QC total	**657,366**	**499,170**
SNPs with a low Gentrain score (<0.15)	- 95,896	- 156
SNPs with low completion rates (< 95%)	- 11,236	- 775
SNPs with mapping issues	NA	- 864
Indels	NA	- 10,208
CNVs	- 58,865	NA
SNPs with differences between array versions	- 10	NA
Monomorphic SNPs	- 12,803	- 49,264
SNPs on X, Y, XY, MT chromosomes	- 14,469	- 15,282
Duplicates and triplicates SNPs and SNPs with multiple mappings	- 8	- 1,004
Unmapped SNPs	- 2,997	- 855
Post QC total	**461,082**	**420,762**

Table 8. Summary of quality control results for samples.

	Infinium Human-600K	**Infinium Oncorray-500K**
Pre-QC total	**535**	**380**
Samples with low completion rates	- 45	- 1
PCA analysis	- 22	- 13
Sample duplicates	- 14	- 10
Related samples (Identity By Descent - IBD)	- 3	NA
Gender check (low levels of X chromosome heterozygosity)	0	- 1
Low/high heterozygosity	0	0
Post-QC total	**451**	**355**
Clinical data exclusions	- 24	- 16
Total of testable samples	**427**	**339**
No. Cases	**165**	**96**
No. Controls	**262**	**243**

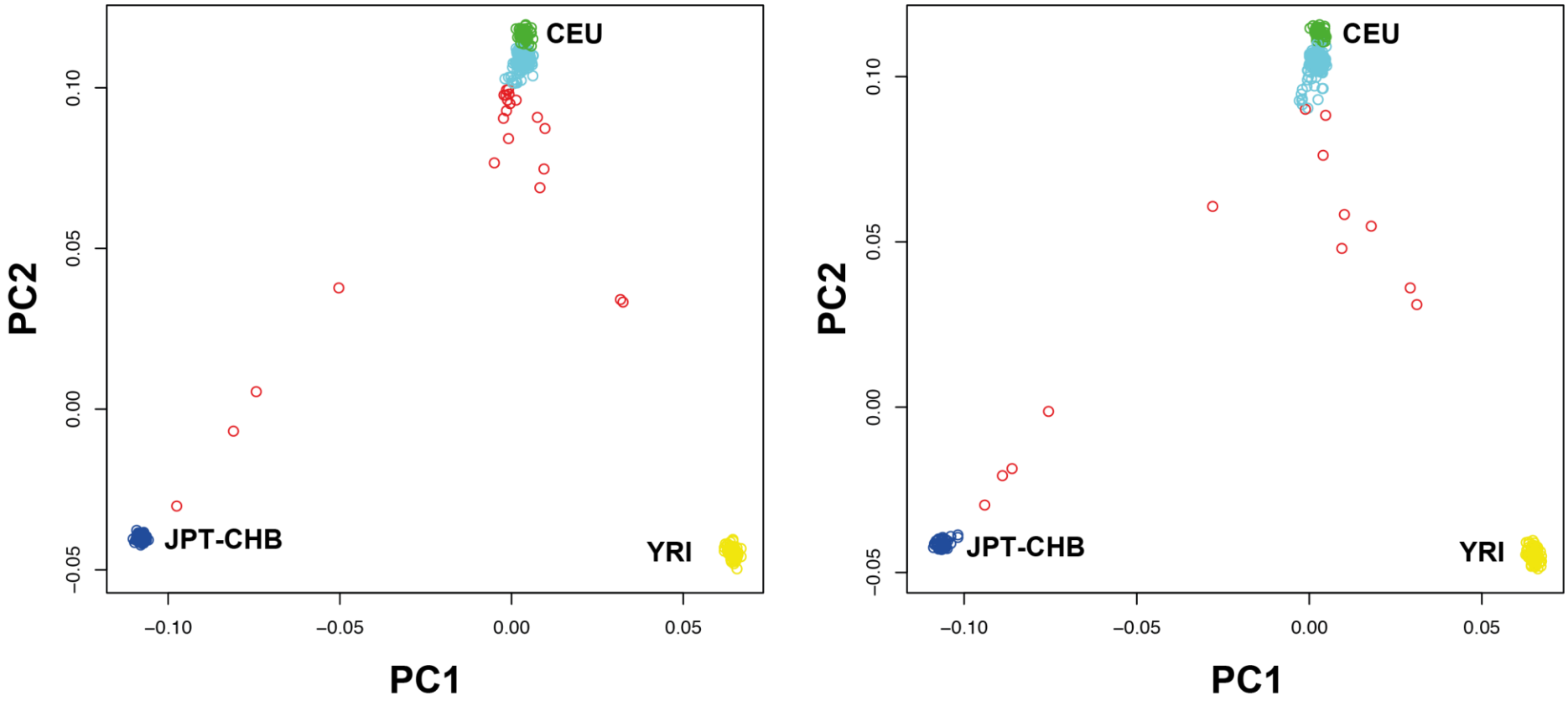

Figure 9. PCA analysis of inferred ancestry using genotype data from the Hapmap project. The left plot represents the results for Infinium Human-600K dataset, the right plot represents the results for the Infinium Oncorray-500K dataset. Reference populations representing Western Europe (CEU), Yoruba from West Africa (YRI), and Chinese and Japanese from East Asia (JPT-CHB). Ancestry outliers were identified using kmeans cluster analysis and are represented in red.

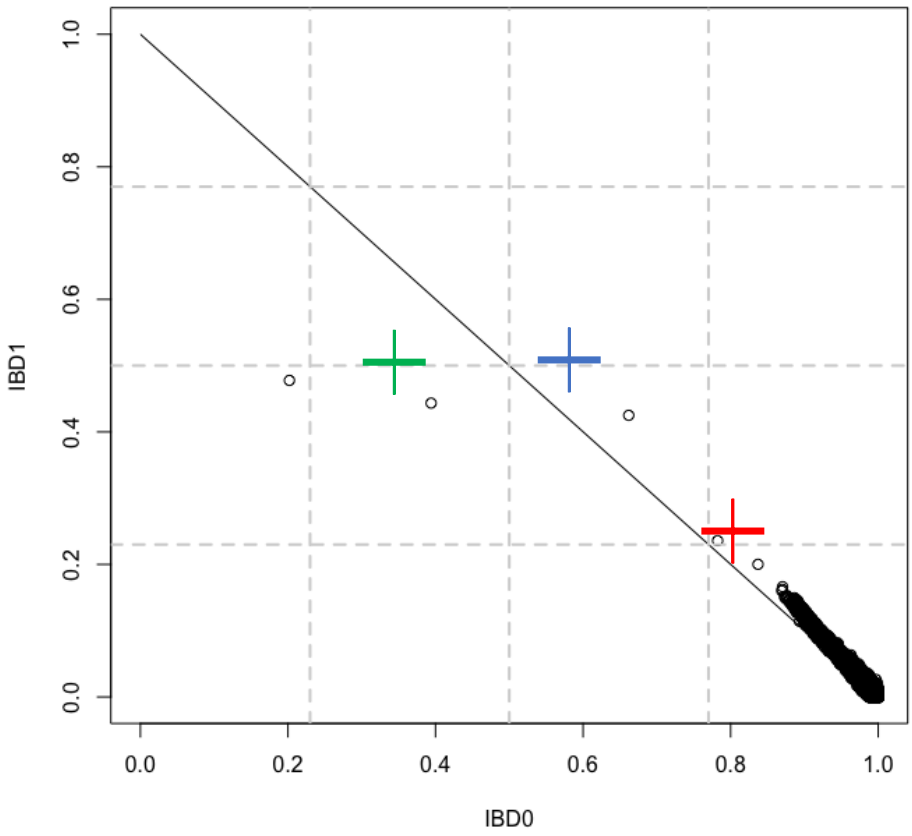

Figure 10. IBD scatterplot of samples from the Infinium Human-600K dataset. Each point represents a pair of samples plotted by degrees of relatedness. The x-axis represents the proportion of loci that share zero alleles (IBD0) while the y-axis represents the proportion of loci that share one allele (IBD1). Monozygotic twins or duplicate samples are expected to localise at IBD1 = IBD0 = 0. Parent-offspring pairs are expected to localise at IBD0 = 0 and IBD1 = 1. Full siblings, half-siblings and first cousins are expected to localise to the areas indicated by green, blue and red crosses respectively.

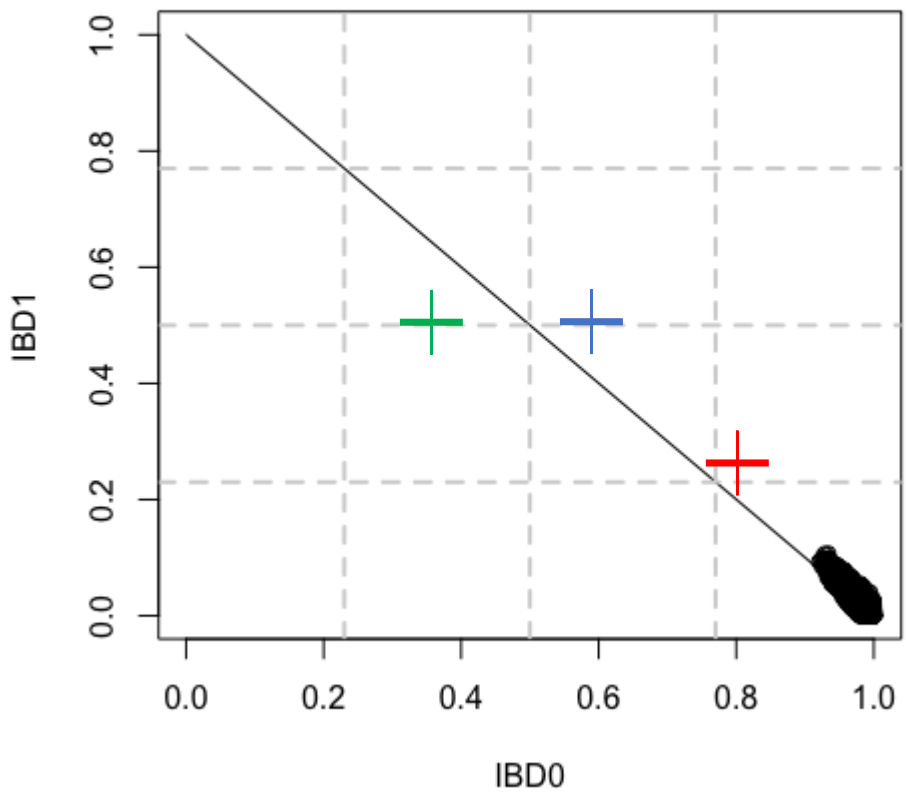

Figure 11. IBD scatterplot of samples from the Infinium Oncorray-500K dataset. Each point represents a pair of samples plotted by degrees of relatedness. The x-axis represents the proportion of loci that share zero alleles (IBD0) while the y-axis represents the proportion of loci that share one allele (IBD1). Monozygotic twins or duplicate samples are expected to localise at IBD1 = IBD0 = 0. Parent-offspring pairs are expected to localise at IBD0 = 0 and IBD1 = 1. Full siblings, half-siblings and first cousins are expected to localise to the areas indicated by green, blue and red crosses respectively.

2.3.3 Common inherited variants predisposing to radiation-induced breast cancer

The association analyses were conducted on 766 samples (261 cases and 505 controls) and 6,788,016 SNPs. This small dataset meant that this study is powered to detect associations of common variants with large effect sizes. More specifically, it is powered to detect variants with a relative risk approximately > 1.30 using a P-value threshold of 0.05 and approximately > 1.80 using a P-value threshold of P-value of less than 5×10^{-8} (Figures 12, 13 and 14). Visual inspection of quantile-quantile (QQ) plots revealed little evidence of test statistic inflation due to population stratification or other cryptic sources of confounding and this observation was supported by a lambda statistic of only 1.03 (Figure 15). No SNPs attained genome-wide significance (P-value < 1×10^{-8}). However, 72 SNPs were associated with HLBC at P-value < 5×10^{-6} (Supplementary Table 1).

Three distinct regions of association were observed localising to 5q21.3, 9p24.1 and 20p12 (Figure 14). At 5q21.3, the minor allele of SNP rs200924286, was associated with an increased risk of HLBC (HR = 1.60 [1.33-1.94], P-value = 1.75×10^{-6}) (Table 9). An analysis of all SNPs 100 Kb proximal or distal to rs200924286, conditioned on rs200924286 genotype, provided no evidence of independent associations at this locus (Figure 16). SNP rs200924286 is not associated with sporadic BC (OR = 1.00 [0.99-1.02]. P-value =0.66) (72,73). It resides in a large (> 2 Mb) gene desert with little evidence of regulatory activity as evidenced by a paucity of histone marks

that are indicative of enhancer elements, transcription factor (TF) binding sites or deoxyribonuclease (DNase) hypersensitivity sites.

At 9p24.1, SNP rs9792577 was the most significant of 21 correlated variants (HR = 0.65 [0.55-0.78], P-value = 3.31 x 10^{-6}) and localised to an intron of Lysine-specific demethylase 4C, *KMD4C*. Conditional analysis provided no support for additional associations at this locus (Figure 17). SNP rs9792577 is not associated with sporadic BC (OR = 0.99 [0.98-1.01], P-value = 0.26) (Table 9) (72,73). A global analysis of gene expression of 48 distinct tissue types from Genotype-Tissue Expression (GTEx) data indicated that rs9792577 is an eQTL for *KMD4C* (P-value = 6.40 x 10^{-5}) (101). However, an eQTL analysis of *KMD4C* expression in female breast tissue (n = 94) from GTEx was not significant (P-value = 0.75) (Figure 18). *KDM4C* is most expressed in the cerebellum (Figure 19). It is also expressed in a number of cancers, including thyroid, colorectal and testis cancer (Figure 20). Overall, the expression of *KDM4C* appears mostly consistent between normal tissues and cancer tissues.

At 20p12, SNP rs671426 was the most significant of 15 correlated variants (HR = 0.63 [0.51-0.77], P-value = 3.40 x 10^{-6}) and localised to an intron of *SNAP25-AS1*, which encodes an antisense RNA belonging to the synoptomal nerve-associated protein 25 (*SNAP25)* gene family. SNP rs671426 is not associated with sporadic BC (OR = 0.99 [0.99-1.01], P-value = 0.34) (Table 9). Conditional analysis did not support the presence of

independent risk SNPs at this locus (Figure 21) and there was no evidence of eQTL associations in GTEx data.

The remaining seven association signals with P-values less than 5 x 10^{-6} mapped to 1q36.13, 6p12.3, 7q22.1, 10p12.2, 12q24.21, 14q12 and 18q11.2 (Table 9). Of these, five were not associated with sporadic BC (Table 9) (72,73). Four of the seven SNPs mapped to gene deserts with little evidence of regulatory activity (10p12.2, 12q24.1, 14q12, 18q11.2). SNP rs147418674 at 1q36.13 was localised in the intron of *CROCC*. *CROCC* encodes a protein that is part of the ciliary rootlet, a cytoskeleton-like structure that is part of a cilium. It is also in the vicinity of *MIR3675*, which encodes a microRNA of unknown function, and *RNU1-2*, a small nuclear RNA of unknown function. SNP rs891238519 at 6p12.3 was located in the intron of *PKHD1*, a gene which encodes a transmembrane receptor-like protein and is known to cause polycystic kidney disease and is involved in colorectal cancer (102,103). SNP rs73156689 was located in the vicinity of E3 ubiquitin-protein ligase (*TRIM4*), gap junction gamma-3 protein (*CJC3*), zinc-alpha-2-glycoprotein (*AZGP1*) and zinc-alpha-2-glycoprotein pseudogene 1 (*AZGP1P1*). A global analysis of gene expression in 48 tissues from GTEx indicated that rs73156689 was associated with *TRIM4* and *PVRIG* expression. However, an eQTL analysis of *TRIM4* and *PVRIG* in female breast tissue (n =94) from GTEx did not detect an eQTL association (*TRIM4* P-value = 0.358; PVRIG P-value = 0.869) (Figure 22). There was evidence of significant heterogeneity in the effects conferred by each SNP upon either sporadic BC or radiation-induced BC (Table 10).

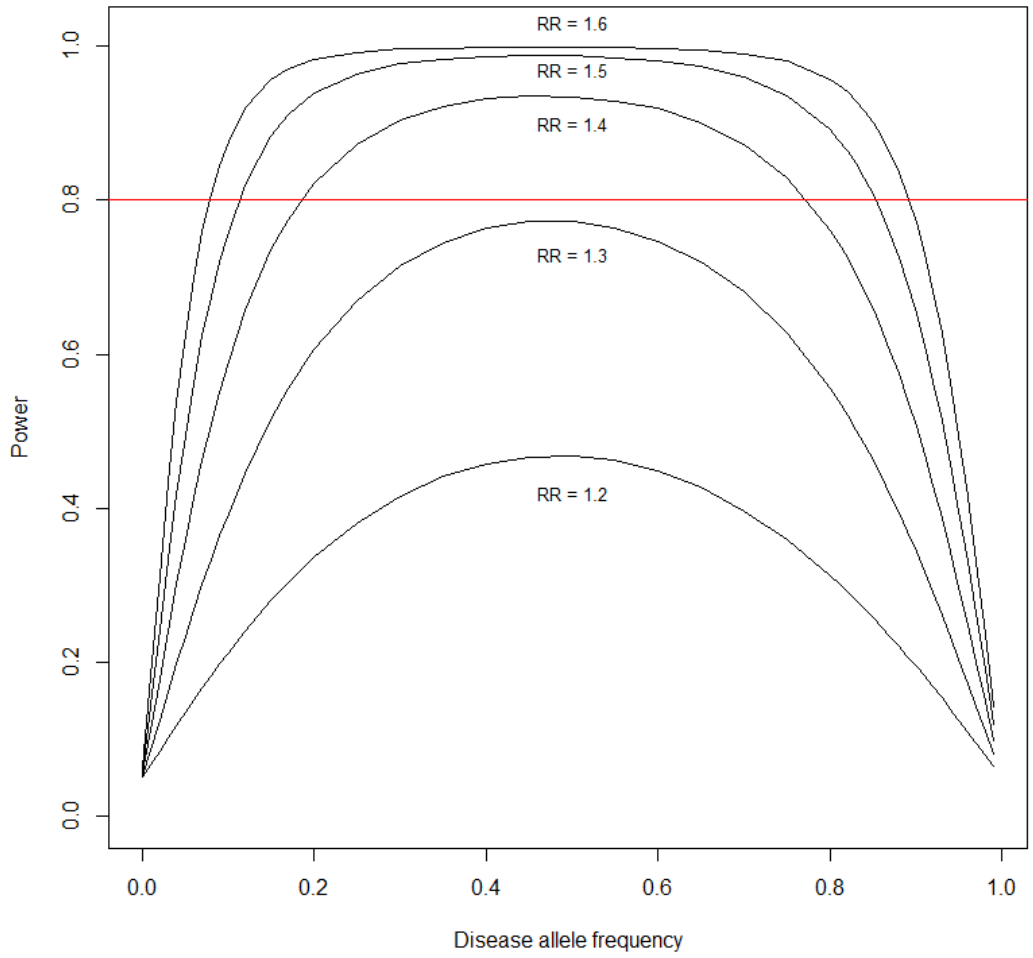

Figure 12. Power analysis (P-value < 0.05 threshold). The y-axis represents the study power while the x-axis represents the disease allele frequency. The red line represents the 80% power threshold. Each black line corresponds to a different effect size (relative risk or RR). This power calculations was based on the following parameters: number of cases = 261, number of controls = 505, multiplicative model, P-value threshold < 0.05).

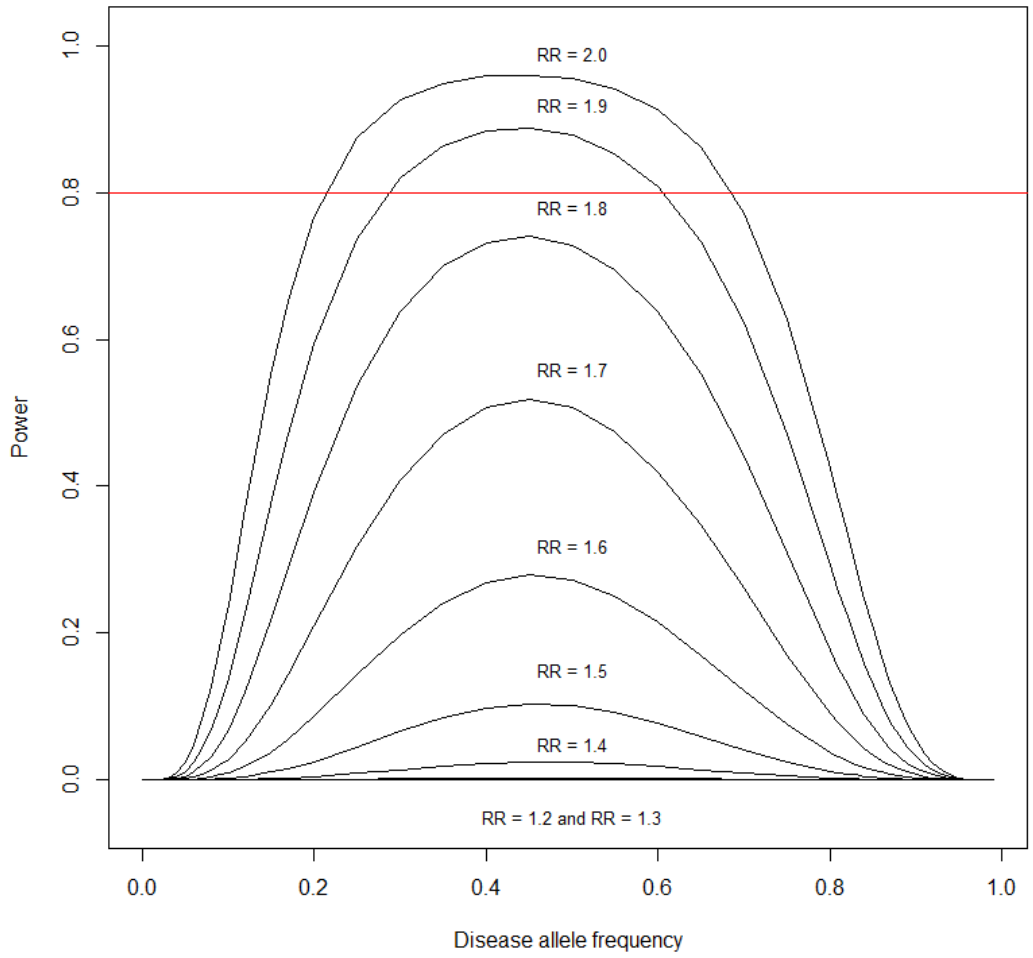

Figure 13. Power analysis (P-value < 5 x 10^{-8} threshold). The y-axis represents the study power while the x-axis represents the disease allele frequency. The red line represents the 80% power threshold. Each black line corresponds to a different effect size (relative risk or RR). This power calculations was based on the following parameters: number of cases = 261, number of controls = 505, multiplicative model, P-value threshold < 5 x 10^{-8}).

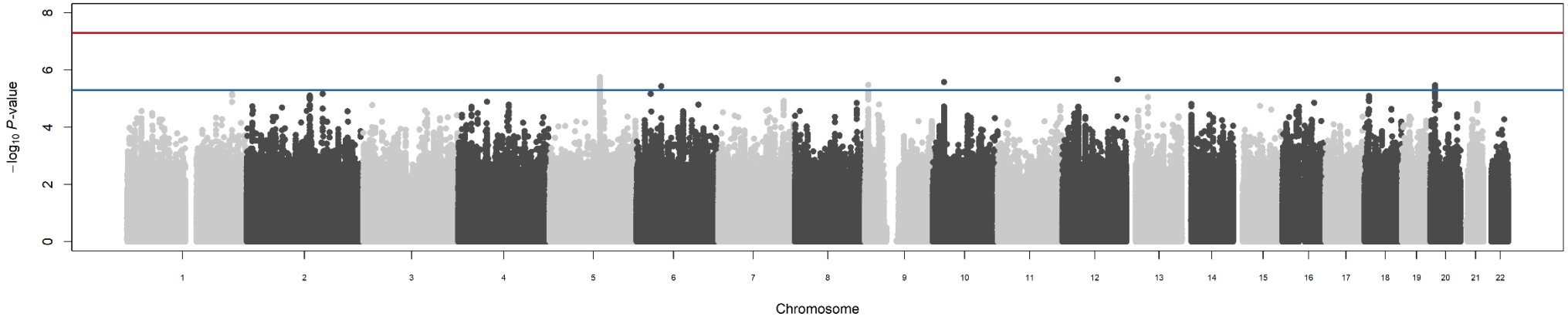

Figure 14. Manhattan plot of association of 6,788,016 SNPs with the risk of HLBC. Each circle represents a SNP plotted according to its chromosome and physical location. The y-axis shows the negative $\log_{10}$ of the P-value. The blue and red lines represent the 5×10^{-6} and 5×10^{-8} significance thresholds respectively.

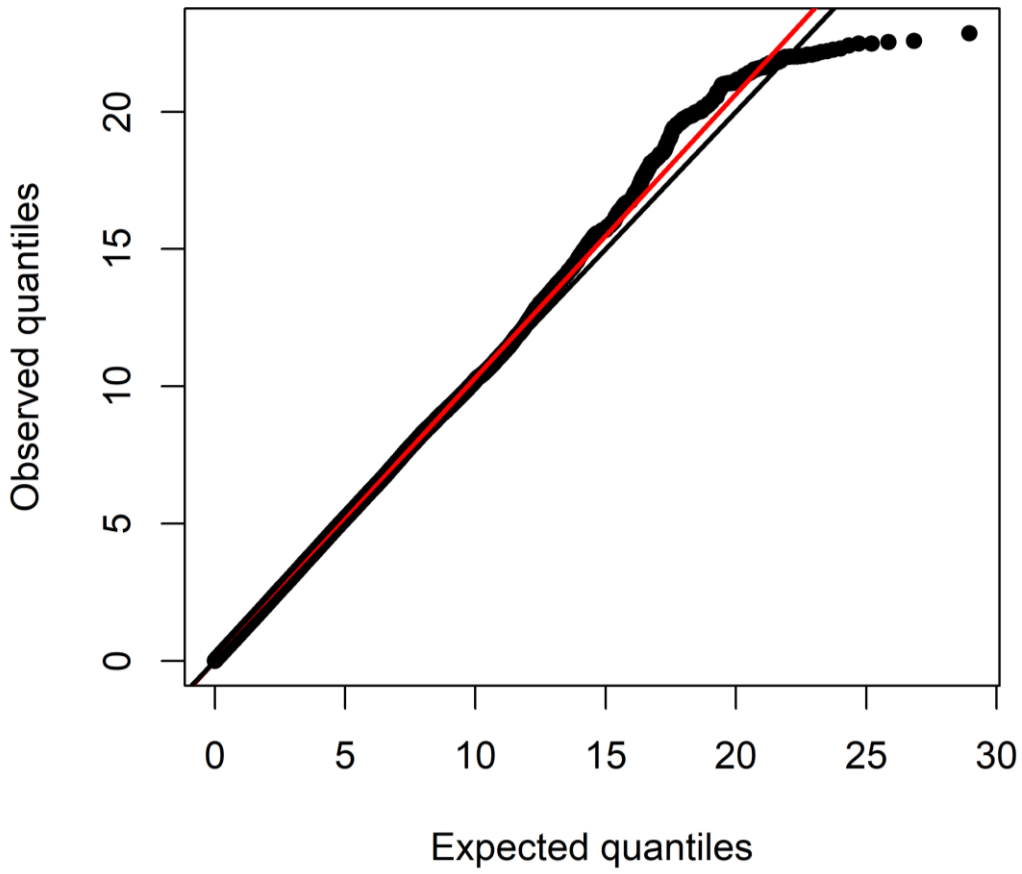

Figure 15. QQ plot of test statistics. The observed test statistics (y-axis) from the association test for 6,788,016 SNPs are plotted against the expected test statistics values under the null distribution (x-axis). The black line represents the expected distribution of test statistics assuming no true associations at any SNP locus. The red line represents the minimal inflation of the test statistics ($\lambda = 1.03$) observed in the GWAS data.

Table 9. Independent signals of association with the risk of HLBC in HL survivors. Results are ordered by chromosome. The data for sporadic BC was obtained from BCAC (72,73).

Cytoband	Nearest gene (±50kb)	rsid	Minor allele	Major allele	Risk allele	MAF	MAF cases	MAF controls	Hazard ratio	95% CI	P-value	Sporadic BC OR	Sporadic BC 95% CI	Sporadic BC P-value
1p36.13	*MIR3675, RNU1-2 CROCC*	rs147418674	G	T	G	0.06	0.09	0.04	2.25	1.55-3.26	7.67×10^{-6}	1.03	0.99-1.07	0.14
5q21.3	NA	rs200924286	GTGTGTA	G	GTGTGTA	0.28	0.34	0.24	1.60	1.33-1.94	1.75×10^{-6}	1.00	0.99-1.02	0.66
6p12.3	*PKHD1*	rs891238519	CGTGT	C	C	0.20	0.14	0.23	0.57	0.44-0.73	3.69×10^{-6}	0.99	0.98-1.01	0.64
7q22.1	*TRIM4, GJC3, AZGP1, AZGP1P1*	rs73156689	A	G	A	0.06	0.10	0.04	2.03	1.50-2.75	2.62×10^{-6}	1.02	0.99-1.05	0.14
9p24.1	*KDM4C*	rs9792577	A	G	G	0.48	0.40	0.52	0.65	0.55-0.78	3.31×10^{-6}	0.99	0.98-1.01	0.26
10p12.2	NA	rs35697184	TA	T	T	0.20	0.14	0.23	0.56	0.44-0.73	2.62×10^{-6}	0.99	0.98-1.01	0.31
12q24.21	NA	rs838308	T	C	C	0.20	0.14	0.23	0.57	0.44-0.73	2.12×10^{-6}	0.99	0.98-1.01	0.71
14q12	NA	rs191202224	T	G	T	0.06	0.10	0.04	1.83	1.33-2.53	6.19×10^{-4}	1.02	0.98-1.05	0.36
18q11.2	NA	rs1668135	G	T	T	0.22	0.16	0.25	0.65	0.51-0.83	3.57×10^{-4}	0.99	0.98-1.01	0.40
20p12.2	*SNAP25-AS1*	rs671426	C	T	T	0.30	0.22	0.34	0.63	0.51-0.77	3.40×10^{-6}	0.99	0.99-1.01	0.34

Table 10. Test of heterogeneity of the effects of the independent signals of association in HLBC individuals and sporadic BC individuals. The heterogeneity tests were performed using ORs in order to be able to compare the datasets directly and were run using METAL (104). The data for sporadic BC was obtained from BCAC (72,73). Results are ordered by chromosome.

Cytoband	Nearest gene (±50kb)	rsid	Minor allele	Major allele	Risk allele	MAF	MAF cases	MAF controls	OR	95% CI	P-value	Sporadic BC OR	Sporadic BC 95% CI	Sporadic BC P-value	Heterogeneity test P-value
1p36.13	*MIR3675, RNU1-2 CROCC*	rs147418674	G	T	G	0.06	0.09	0.04	4.11	2.25-7.50	4.32×10^{-6}	1.03	0.99-1.07	0.14	2.90×10^{-6}
5q21.3	NA	rs200924286	GTGTGTA	G	GTGTGTA	0.28	0.34	0.24	1.71	1.34-2.19	1.54×10^{-5}	1.00	0.99-1.02	0.66	1.80×10^{-5}
6p12.3	*PKHD1*	rs891238519	CGTGT	C	C	0.20	0.14	0.23	0.56	0.42-0.74	4.80×10^{-5}	0.99	0.98-1.01	0.64	5.60×10^{-5}
7q22.1	*TRIM4, GJC3, AZGP1, AZGP1P1*	rs73156689	A	G	A	0.06	0.10	0.04	3.32	2.06-5.36	8.99×10^{-7}	1.02	0.99-1.05	0.14	6.20×10^{-7}
9p24.1	*KDM4C*	rs9792577	A	G	G	0.48	0.40	0.52	0.61	0.49-0.76	1.51×10^{-5}	0.99	0.98-1.01	0.26	1.16×10^{-5}
10p12.2	NA	rs35697184	TA	T	T	0.20	0.14	0.23	0.53	0.40-0.70	5.19×10^{-6}	0.99	0.98-1.01	0.31	4.02×10^{-6}
12q24.21	NA	rs838308	T	C	C	0.20	0.14	0.23	0.56	0.43-0.72	1.27×10^{-5}	0.99	0.98-1.01	0.71	1.46×10^{-5}
14q12	NA	rs191202224	T	G	T	0.06	0.10	0.04	3.43	2.04-5.78	3.58×10^{-6}	1.02	0.98-1.05	0.36	2.78×10^{-6}
18q11.2	NA	rs1668135	G	T	T	0.22	0.16	0.25	0.54	0.41-0.70	4.54×10^{-6}	0.99	0.98-1.01	0.40	6.00×10^{-6}
20p12.2	*SNAP25-AS1*	rs671426	C	T	T	0.30	0.22	0.34	0.59	0.47-0.75	9.52×10^{-6}	0.99	0.99-1.01	0.34	1.32×10^{-5}

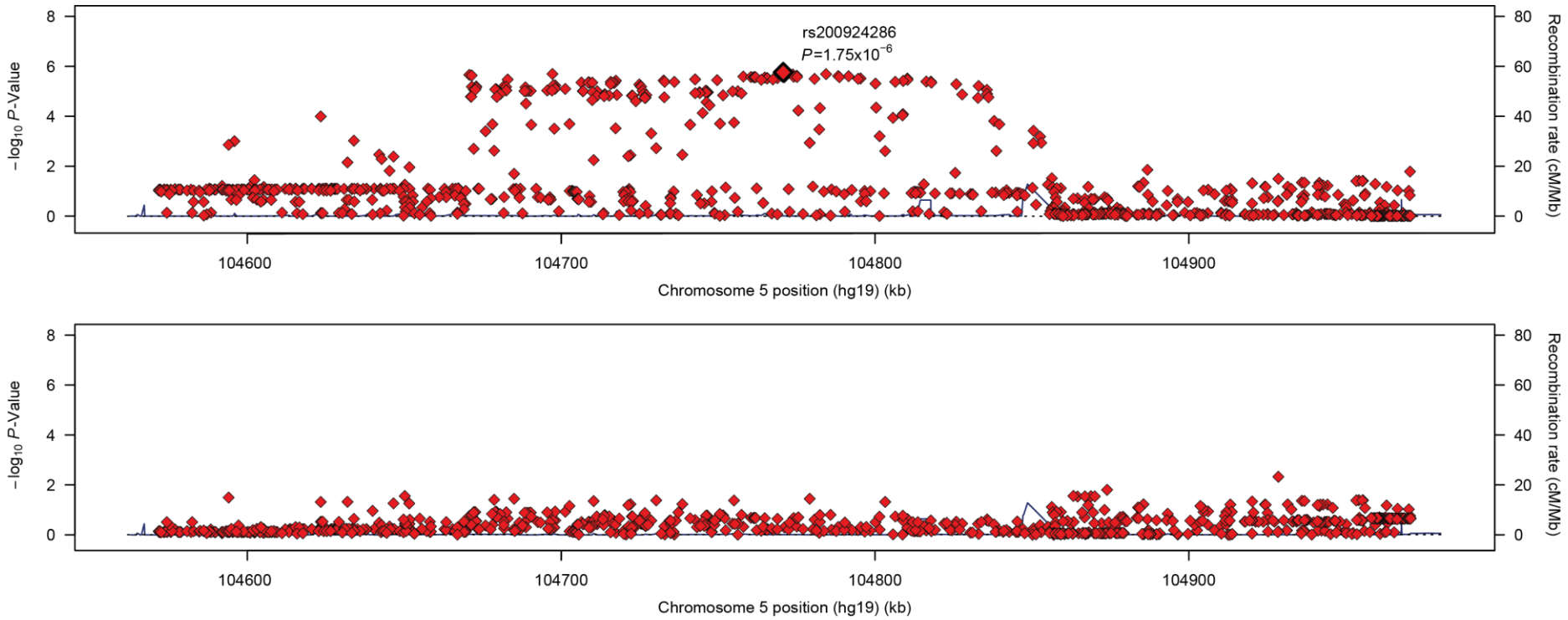

Figure 16. Regional association plots for chromosome 5. The two plots demonstrate a forward stepwise regression analysis where the top associated SNP of the first plot, which was obtained in the GWAS, is used as a covariate in the following analysis until no more significant signals are detected. The x-axis shows the chromosomal position and the y-axis shows the $-\log_{10}$ P-values. SNPs are represented as red diamonds. The top associated SNP in the GWAS, rs20092486, is shown on the first plot as a large diamond. The blue line represents the recombination rate (cM/Mb).

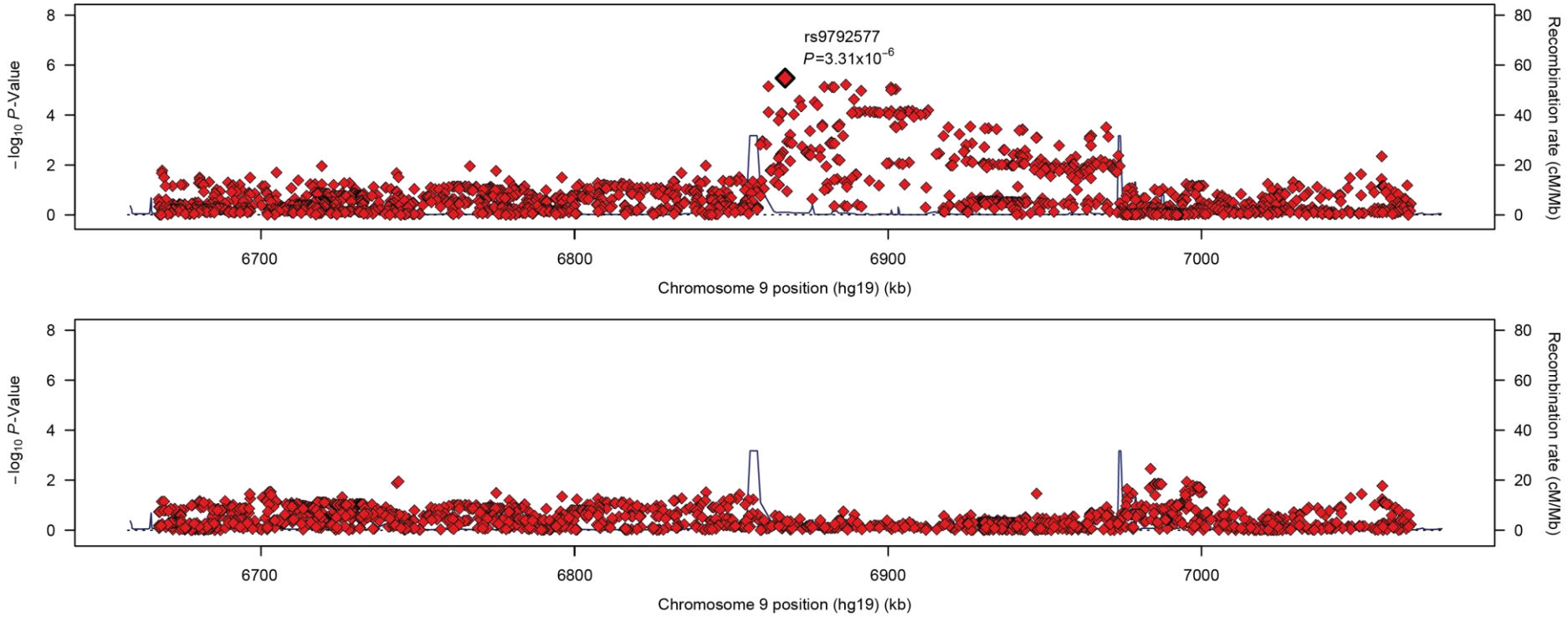

Figure 17. Regional association plots for chromosome 9. The two plots demonstrate a forward stepwise regression analysis where the top associated SNP of the first plot, which was obtained in the GWAS, is used as a covariate in the following analysis until no more significant signals are detected. The x-axis shows the chromosomal position and the y-axis shows the $-\log_{10}$ P-values. SNPs are represented as red diamonds. The top associated SNP in the GWAS, rs9792577, is shown on the first plot as a large diamond. The blue line represents the recombination rate (cM/Mb).

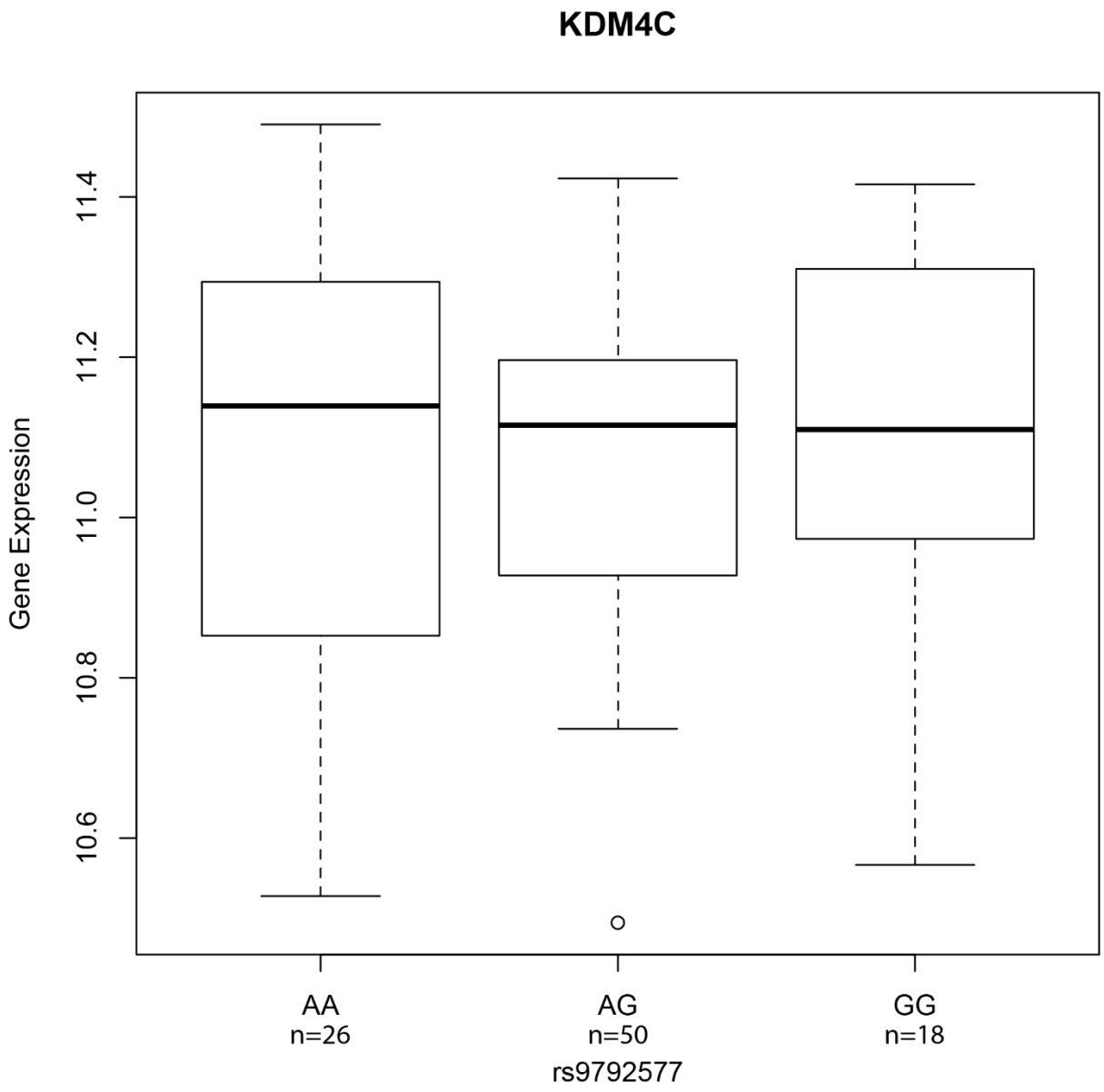

Figure 18. eQTL analysis from normal female breast tissue of rs9792577 and KDM4C. The data was obtained from the Genotype-Tissue Expression (GTEx) database (101).

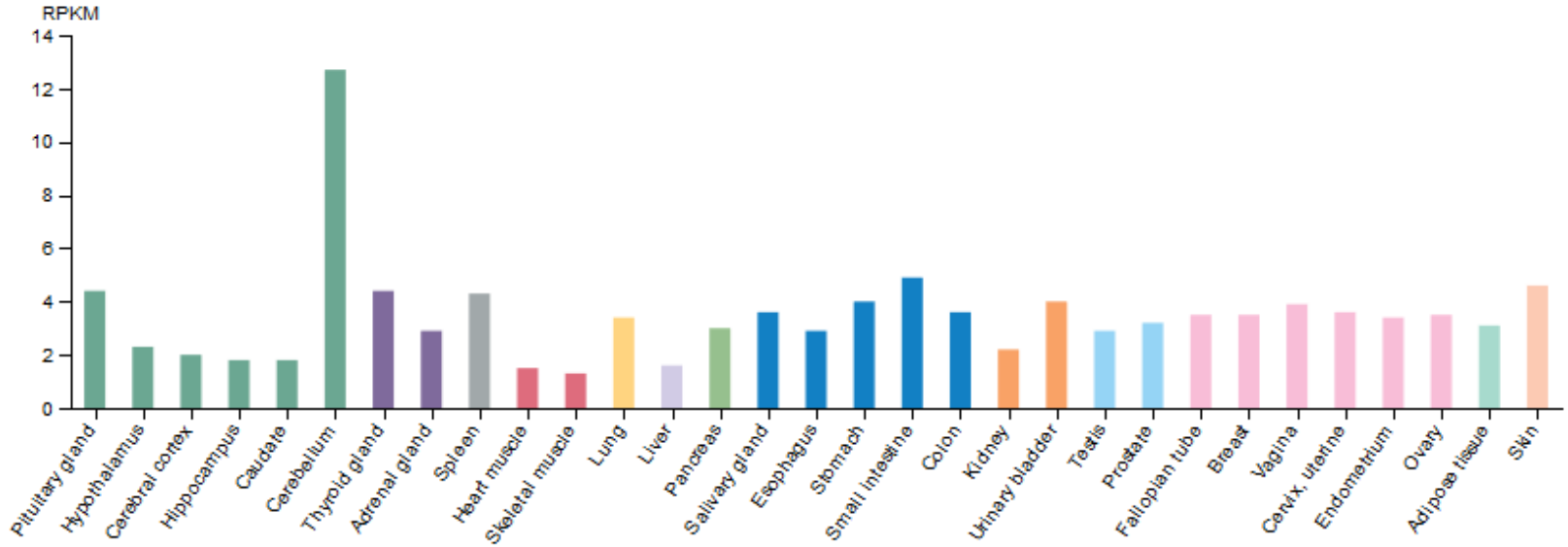

Figure 19. KM4C expression in healthy tissues. This data was obtained from the Genotype-Tissue Expression (GTex) database (101)

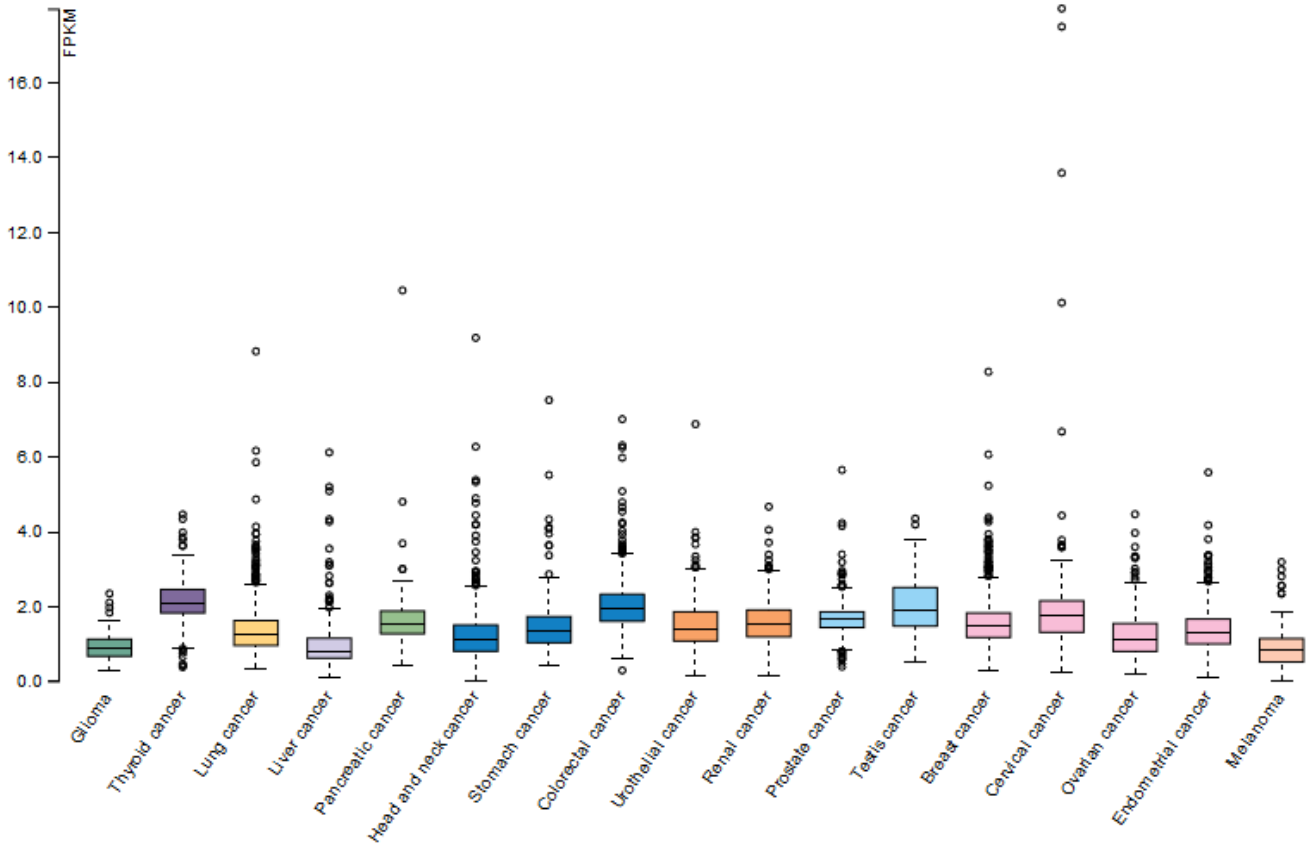

Figure 20. KM4C expression in cancer tissues. This data was obtained from The Cancer Genome Atlas (TCGA) (105)

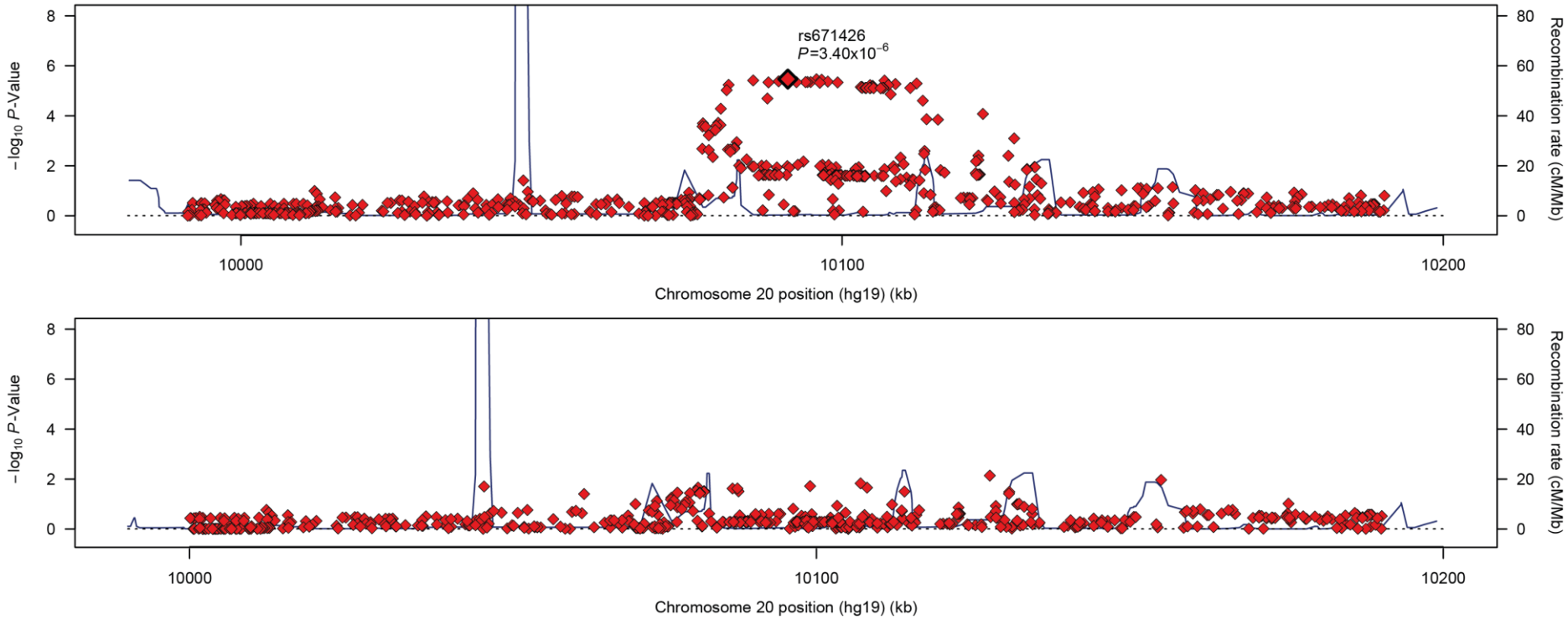

Figure 21. Regional association plots for chromosome 20. The two plots demonstrate a forward stepwise regression analysis where the top associated SNP of the first plot, which was obtained in the GWAS, is used as a covariate in the following analysis until no more significant signals are detected. The x-axis shows the chromosomal position and the y-axis shows the $-\log_{10}$ P-values. SNPs are represented as red diamonds. The top associated SNP in the GWAS, rs671426, is shown on the first plot as a large diamond. The blue line represents the recombination rate (cM/Mb).

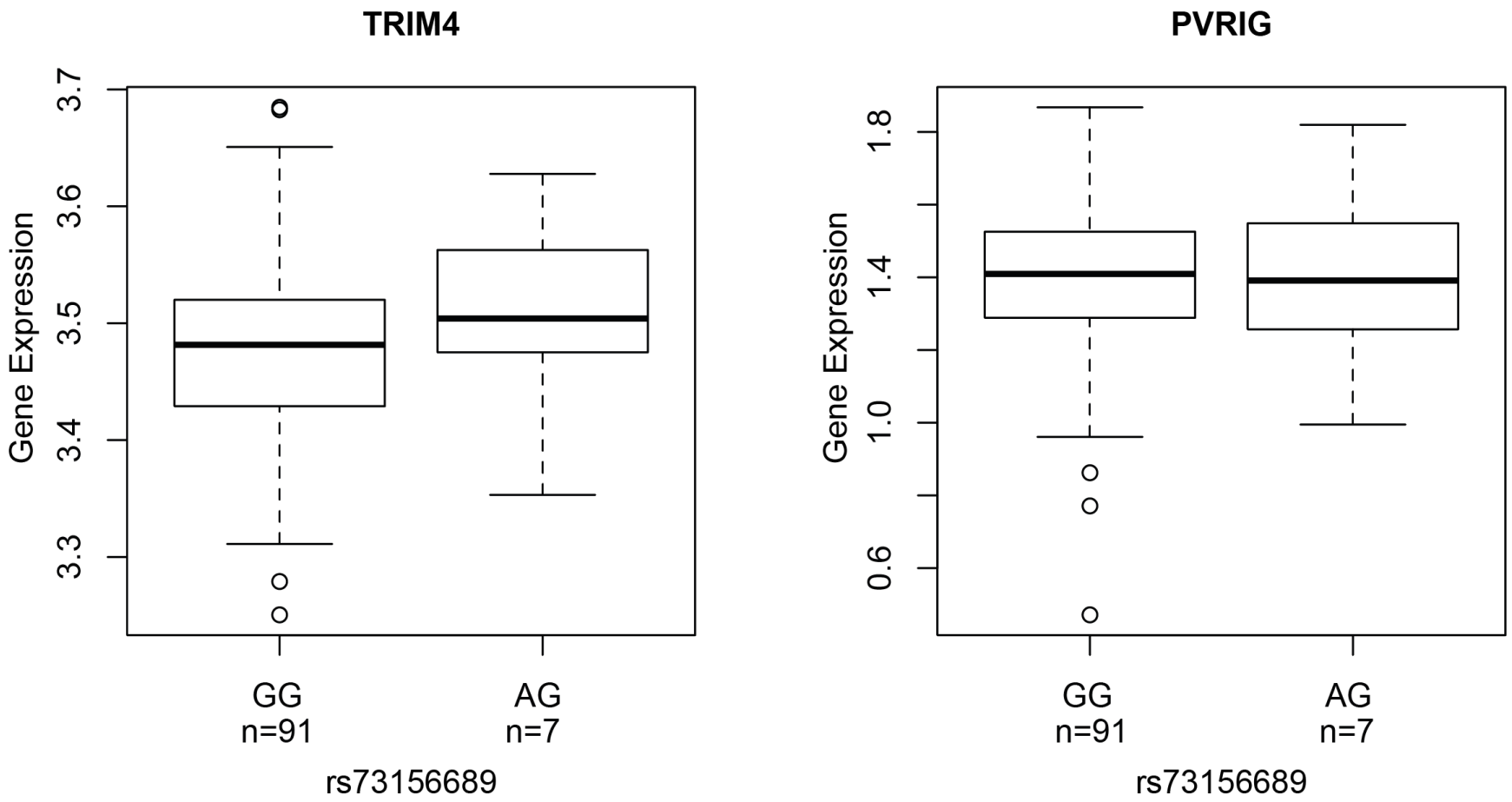

Figure 22. eQTL analysis from normal female breast tissue of rs73156689 and TRIM4 and PVRIG. The data was obtained from the Genotype-Tissue Expression (GTEx) database.

2.3.4 Association of previously published loci with HLBC

SNPs rs17020562 at 1q41, rs4342822 at 1q32.3 and rs74949440 at 11q23.3, all recently reported by Morton *et al*, were non-significant in my GWAS (for rs17020562 P-value = 0.051 ; for rs4342822 P-value = 0.68; for rs74949440 P-value = 0.67) (Table 11) (71). Their HRs were found to be almost twice as small compared to those obtained by Morton *et al* (for rs17020562 HR = 21.08; for rs4342822 HR = 1.04; for rs74949440 HR =0.90). All three SNPs are not associated with sporadic BC (Table 11).

SNPs rs4946728 at 6q21 and rs1040411 at 6q21, previously reported by Best *et al*, were non-significant in my GWAS (for rs4946728, P-value = 0.93; for rs1040411, P-value = 0.69) (Table 11) (70). Their HRs were also much smaller than those obtained by Best *et al* (for rs4946728 HR = 1.00; for rs1040411 HR = 1.16). Both SNPs are not associated with sporadic BC (Table 11).

SNP rs1219648 (at 10q26.13) which annotates to *FGFR2* and was reported by Ma *et al*, was also found to be significant (HR = 1.29 [1.07-1.54], P-value = 5.6×10^{-3}) (Table 11) (67). This effect is smaller than the one reported by Ma *et al* (OR = 1.73 [1.29-2.34], P-value = 2.73×10^{-4}), and instead is much closer to the one reported in sporadic BC (OR = 1.26 [1.25-1.28], P-value = 5.14×10^{-302}]) (Table 11).

Table 11. Association of SNPs previously associated with HLBC. Results are ordered by chromosome. The data for sporadic BC was obtained from BCAC (72,73).

Cytoband	Nearest Gene (±50kb)	rsid	Minor allele	Major allele	Risk allele	MAF	MAF cases	MAF controls	Hazard ratio	95% CI	P-value	Study HR or OR	95% CI	P-value	Sporadic BC OR	Sporadic BC 95% CI	Sporadic BC P-value
1q32.3	NA	rs17020562	C	T	C	0.001	0.0023	0.00046	21.08	2.29-193.94	0.051	Morton et al 2017 HR = 44.52	15.06-131.62	6.68×10^{-8}	1.00	0.90-1.13	0.90
1q41	*AK092251*	rs4342822	G	T	G	0.39	0.43	0.37	1.04	0.87-1.28	0.68	Morton et al 2017 HR = 1.92	1.49-2.44	7.09×10^{-9}	1.00	0.99-1.02	0.31
6q21	NA	rs4946728	A	C	A	0.36	0.32	0.38	1.00	0.83-1.29	0.93	Best *et al* OR = 3.32	2.25-4.90	5.99×10^{-10}	1.01	0.99-1.02	0.37
6q21	NA	rs1040411	A	G	A	0.19	0.13	0.22	1.16	0.87-1.23	0.69	Best *et al* OR = 2.39	1.73-3.30	1.18×10^{-7}	1.00	0.99-1.01	1.00
10q26.13	*FGFR2*	rs1219648	G	A	G	0.39	0.46	0.37	1.29	1.07-1.54	0.0056	Ma *et al* 2012 OR = 1.73	1.29-2.34	2.73×10^{-4}	1.26	1.25-1.28	5.14×10^{-302}
11q23.3	*TAGLN*	rs74949440	T	C	C	0.03	0.031	0.034	0.90	0.54-1.50	0.67	Morton et al 2017 HR = 2.59	1.62-4.16	5.84×10^{-8}	1.02	0.98-1.06	0.34

2.3.5 Association of known FBC loci with HLBC

Although none of the tested 172 FBC loci were genome-wide significant in the GWAS, 11 FBC SNPs had a P-value of less than 0.05 (Table 12 and Supplementary Table 2). Of particular interest is the observation of two associated SNPs annotating to Fibroblast Growth Factor 2 (*FGFR2*): rs2981578 and rs35054928. Although both SNPs are correlated with rs1219648, the SNP identified by Ma *et al* (for rs2981578 and rs1219648, D' = 0.97; for rs35054928 and rs1219648, D' = 0.97), their ORs did not appear to differ from the ORs in sporadic BC (Table 12). This suggests that the signal at 10q26.13 may be a false positive in my GWAS and that this region may not be associated with HLBC.

Three of the associated SNPs (rs16857609, rs6964587 and rs10760444) had HRs going in opposite directions to the ORs for sporadic BC, although their confidence intervals were large and very close to one. The remaining eight had HRs in the same direction as the ORs for sporadic BC. Among these, five had ORs that were considerably larger than the OR for sporadic BC. These included rs4971059 at 1q22, rs4951011 at 1q32.1, rs6596100 at 5q31.1, rs3757322 at 6q25.1 and rs10816625 at 9q31.2.

The Chi-squared test used to compare the ORs for HLBC to those in sporadic BC identified a total of 11 SNPs that had significantly different ORs in HLBC compared to sporadic BC (Table 13). Of these, five were the same as the associated FBC SNPs identified in the Cox regression analyses: rs4971059 at 1q22, rs4951011 at 1q32.1, rs16857609 at 2q35, rs6596100 at 5q31.1 and rs6964587 at 7q21.2. While the differences in the ORs are significant for 11 FBC SNPs, the overall P-value of the Chi-Squared test was not significant (P-value = 0.77).

Table 12. Known FBC loci found to be significantly associated (P-value < 0.05) in the GWAS. Results are ordered by chromosome. Results are ordered by chromosome. The data for sporadic BC was obtained from BCAC (72,73).

Cytoband	Nearest gene (±50kb)	rsid	Minor allele	Major allele	Risk allele	MAF	MAF cases	MAF controls	Hazard ratio	95% CI	P-value	Sporadic BC OR	Sporadic BC 95% CI	Sporadic BC P-value
1q22	*TRIM46, KRTCAP2, AX746485, MUC1, THBS3, MTX1, MIR92B*	rs4971059	A	G	A	0.35	0.40	0.33	1.28	1.07-1.53	7.5×10^{-4}	1.04	1.03-1.06	4.86×10^{-11}
1q32.1	*ZC3H11A, LAX1, ZBED6*	rs4951011	G	A	G	0.18	0.20	0.16	1.30	1.05-1.61	1.98×10^{-2}	1.04	1.02-1.06	$5\ 1.27\ 10^{-5}$
2q35	*DIRC3*	rs16857609	T	C	C	0.24	0.22	0.26	0.80	0.65-0.98	2.75×10^{-2}	0.95	0.91-0.94	1.82×10^{-25}
5q31.1	*HSPA4*	rs6596100	T	C	C	0.23	0.19	0.25	0.75	0.60-0.94	1.09×10^{-2}	0.96	0.94-0.97	7.74×10^{-9}
6q25.1	*CCDC170*	rs3757322	G	T	G	0.33	0.37	0.30	1.26	1.05-1.51	1.22×10^{-2}	1.09	1.08-1.11	3.32×10^{-41}
7q21.2	*AKAP9*	rs6964587	T	G	G	0.38	0.34	0.39	0.82	0.68-0.90	2.66×10^{-2}	0.96	0.95-0.97	8.95×10^{-11}
9q31.2	NA	rs10816625	G	A	G	0.06	0.08	0.05	1.52	1.13-2.04	9.58×10^{-3}	1.11	1.09-1.14	5.04×10^{-18}
9q33.3	*LMX1B*	rs10760444	G	A	A	0.44	0.40	0.46	0.84	0.70-1.00	4.84×10^{-2}	0.97	0.96-0.98	9.06×10^{-9}
10q26.13	*FGFR2*	rs2981578	C	T	C	0.46	0.49	0.44	1.26	1.06-1.51	1.08×10^{-2}	1.23	1.22-1.25	1.31×10^{-245}
10q26.13	*FGFR2*	rs35054928	GC	G	GC	0.4	0.44	0.38	1.30	1.09-1.56	3.71×10^{-3}	1.27	1.25-130	3.50×10^{-154}
22q13.1	*PLA2G6*	rs738321	G	C	C	0.39	0.36	0.40	0.84	0.70-1.00	4.38×10^{-2}	0.95	0.94-0.97	1.04×10^{-13}

Table 13. Chi-squared test comparing ORs of FBC loci obtained in the GWAS with those obtained in published studies on sporadic BC. Only the SNPs with significantly different ORs are shown. Results are ordered by chromosome.

Cytoband	Nearest gene (±50kb)	SNP	Major allele	Minor allele	Risk allele	GWAS OR	GWAS 95% CI	Sporadic BC OR	Sporadic BC 95% CI	χ^2	χ^2 P-value
1q22	*TRIM46, KRTCAP2, AX746485, MUC1, THBS3, MTX1, MIR92B*	rs4971059	G	A	A	1.37	1.09-1.72	1.04	1.03-1.06	5.36	2.08×10^{-2}
1q32.1	*ZC3H11A, LAX1, ZBED6*	rs4951011	A	G	G	1.38	1.04-1.82	1.04	1.02-1.06	3.90	4.70×10^{-2}
2q35	*DIRC3*	rs16857609	C	T	T	0.80	0.62-1.02	0.95	0.91-0.94	5.04	2.48×10^{-2}
5q11.1	NA	rs72749841	T	C	T	1.31	0.94-1.81	1.07	1.05-1.10	4.18	4.10×10^{-2}
5q31.1	*HSPA4*	rs6596100	C	T	C	0.71	0.55-0.92	0.96	0.94-0.97	4.55	3.29×10^{-2}
6q23.1	*L3MBTL3*	rs6569648	T	C	T	1.22	0.96-1.55	1.05	1.04-1.07	4.42	3.55×10^{-2}
7q21.2	*AKAP9*	rs6964587	G	T	T	0.79	0.63-0.99	0.96	0.95-0.97	5.34	2.09×10^{-2}
9q33.1	*ASTN2*	rs1895062	A	G	A	1.21	0.97-1.51	1.05	1.04-1.06	4.85	2.76×10^{-2}
9q31.2	NA	rs10816625	A	G	G	1.73	1.12-2.69	1.11	1.09-1.14	3.90	4.82×10^{-2}
9q33.3	*LMX1B*	rs10760444	A	G	G	0.77	0.61-0.96	0.97	0.96-0.98	6.85	8.87×10^{-3}
20q11.22	*RALY*	rs2284378	C	T	T	0.76	0.60-0.95	0.99	0.97-1.00	5.68	1.72×10^{-2}

2.3.6 Association between risk of HLBC and a 313 SNP BC Polygenic Risk Score

Prior to the calculation of PRS scores, the total number of risk alleles per sample were calculated in order to determine if cases and controls cluster separately. Such an analysis was previously performed for sporadic BC by Mavaddat *et al* (84) and provides an illustration of how cases tend to carry more risk alleles than controls (Figure 23). If the 313 FBC loci used for this PRS play a role in HLBC susceptibility and indeed allow to stratify BC risk in HLBC women, then one would expect a separate clustering of cases and controls, such as the one observed in Mavaddat *et al*. My results certainly indicate such a clustering, albeit a discrete one. This is encouraging, as it suggests that the PRS may be effective in stratifying HL survivors with particularly high risks versus HL survivors with comparatively lower risks. However, one must be cautious in the interpretation of these results. One may have a lower number of risk alleles, but have alleles with strong effects and therefore a high PRS score. Therefore, to be able to truly assess the impact of the 313 FBC loci, one must calculate the PRS itself. Nevertheless, this risk allele analysis is a first indication that cases and controls may be stratified using a PRS based on the 313 FBC loci previously used to stratify sporadic BC risk by Mavaddat *et al*.

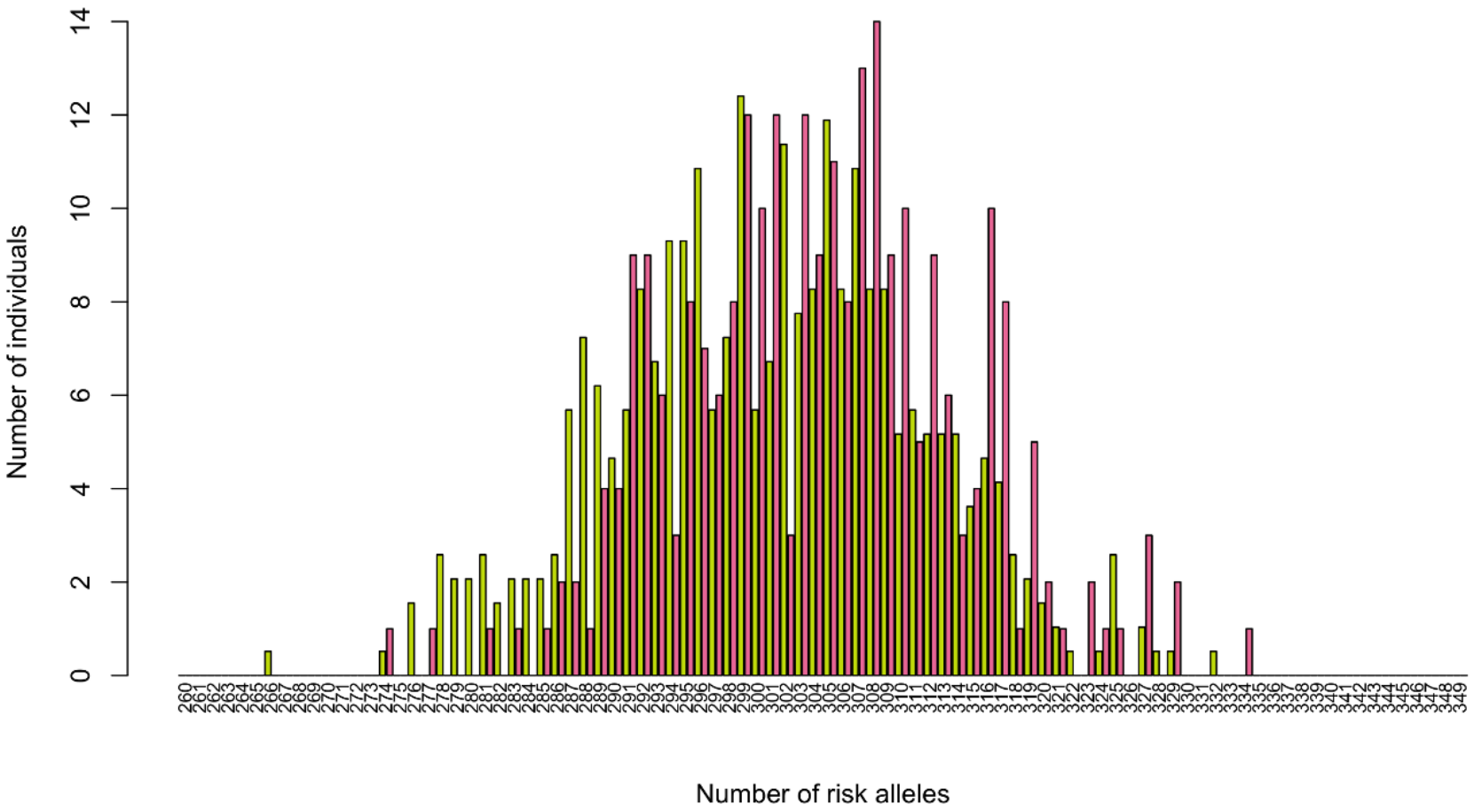

Figure 23. Histogram of risk allele count in HLBC cases and controls. Cases are in pink and controls are in green.

The results of the logistic regression suggest that there is a trend of increasing risks (i.e. increasing ORs) with increasing PRS scores (Table 14). The OR per standard deviation increased in the PRS was 1.92 [1.45-2.48], with a P-value of 5.00×10^{-7}.

Following the analysis of the PRS trends, the next aim was to compare my PRS scores to those obtained for sporadic BC. PRS scores for sporadic BC were obtained for 1087 cases and 697 controls from the Breast Cancer Now Generations Study (100). Distributions of PRS scores were plotted as density curves and separated by phenotypic status (Figure 24). While HLBC cases do tend to have higher PRS scores than HLBC controls, the difference in PRS distributions for all four groups appears small. For instance, the distribution of scores for HLBC cases largely overlapped that of BC controls, suggesting that their risks are very similar. Nevertheless, both groups of HLBC tend to have lower PRS scores than the two groups of sporadic BC. Overall, these observations suggest that the influence of common genetic variants may play a more limited role in HLBC than in sporadic BC. In other words, HL survivors (cases and controls) may have BC risks that are similar to BC controls. This would imply that external exposure, namely ionising radiotherapy, may be the single most important factor affecting HLBC risk.

Table 14. Logistic regression on the HLBC PRS scores.

	HLBC			Mavaddat *et al*	
Percentile	**OR**	**95% CI**	**P-value**	**OR**	**95% CI**
<1%	0.24	0.03-0.01	0.19	0.29	0.23-0.37
1-5%	0.25	0.08-0.75	1.34×10^{-2}	0.42	0.37-0.47
5-10%	0.32	0.12-0.80	1.53×10^{-2}	0.55	0.50-0.61
10-20%	0.48	0.25-0.90	2.14×10^{-2}	0.65	0.60-0.70
20-40%	0.72	0.45-1.17	0.18	0.80	0.76-0.85
40-60%	1.00	(referent)	(referent)	1.00	(referent)
60-80%	1.15	0.72-1.82	0.56	1.18	1.12-1.24
80-90%	1.52	0.87-2.65	0.14	1.48	1.39-1.57
90-95%	0.98	0.47-2.05	0.96	1.69	1.56-1.82
95-99%	1.39	0.64-3.02	0.41	2.20	2.03-2.38
>99%	1.68	0.41-7.00	0.47	2.81	2.43-3.24

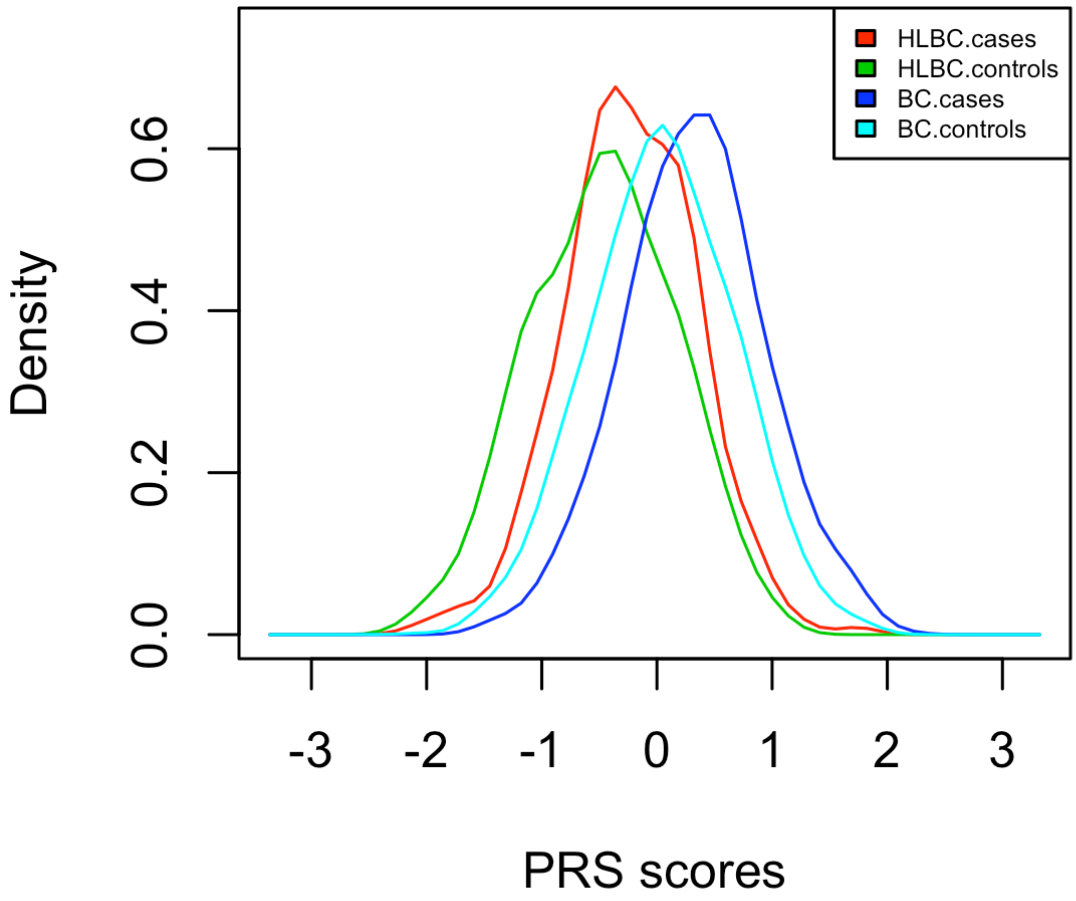

Figure 24. Kernel density curves of PRS scores for HLBC and sporadic BC. PRS scores for sporadic BC were obtained from the Breast Cancer Association Consortium (100).

2.4 Discussion

Despite a lack of genome-wide significant hits, my GWAS identified a number of nominally significant associations representing novel regions underlying HLBC susceptibility. A total of 72 SNPs, locating to ten independent regions were found to be associated with HLBC. The majority of these SNPs locate to three peaks, located at 5q21.3, 9p24.1 and 20p12.2. I have also identified the contribution of a number of known FBC loci, some of which appear to have significantly different effects in HLBC compared to sporadic BC. Notably, this GWAS does not validate all the findings of previous studies on HLBC susceptibility. Nevertheless, caution is required for the interpretation of all these results. Due to its sample size, this GWAS remains underpowered and therefore may not be able to detect all potential associations, including associations found in other cohorts.

Following the associations analyses, I demonstrated that a PRS originally constructed for sporadic BC that can stratify HLBC cases and controls. This is reflected both by looking at the number of risk alleles per phenotype and by a logistic regression on the PRS scores. When comparing my PRS scores to those of healthy control individuals and individuals with sporadic BC, it appears that the PRS scores are consistently lower in my cohort. The most likely reason for these observations is that the influence of common variants is not as great as it is in sporadic BC and that instead, it is the environment exposures, namely radiation exposure, that underlie HLBC.

The results from the GWAS and PRS will need to be validated. This will be done by combining my dataset with the samples from collaborators at the Division of Cancer Epidemiology and Genetics of the National Cancer Institute (NCI). Their cohort is made up of two cohorts called the Childhood Cancer Survivor Study and the St. Jude Lifetime Cohort. Individuals who participated in these studies are female survivors of childhood cancers such as Hodgkin's lymphoma, non-Hodgkin's lymphoma, leukaemia, neuroblastoma etc. While the clinical background of such patients is less homogeneous than my present cohort, genotyping of these women recently revealed the presence of two SNPs potentially associated with radiation-induced breast cancer, as discussed in Section 1.3.2 (106). As such, these patients represent the best available dataset in which to replicate the findings of my GWAS and PRS and the increase in sample size will certainly lead to an increase in statistical power.

Chapter 3 - Rare inherited variants underlying radiation-induced breast cancer

3.1 Introduction

Investigation of germline predisposition to sporadic BC has revealed that the disease is underlain by common low-penetrance SNPs, moderate penetrance variants and rare pathogenic variants. The latter two categories are comprised of susceptibility genes such as *BRCA1*, *BCRA2*, *PALB2*, *CHEK2* and *ATM*, which have been identified in approximately five percent of the general population and 30 to 40% of familial BC cases (107–111). The discovery of these genes have highlighted a relationship between BC and DNA repair, as mutations in those genes have been associated with deficiencies in repair DNA damage (80–83).

Predicated on the observation that the genetic architecture of predisposition to BC comprises both a large number of common variants with small effects and a currently unknown complement of low-frequency variants with intermediate effects, I hypothesised that a proportion of HLBC cases may be attributable to coding mutations in genes that are implicated in the cellular response to radiation-induced DNA damage. Rare variants, in particular those occurring in DNA repair related-genes are known to play a role in sporadic BC (as discussed in Section 2.1) and therefore there is a high prior

probability that they may also influence predisposition to HLBC. The capacity to repair DNA damage is a mechanism which can lead to carcinogenesis if it is impaired (112). The fact that the vast majority of HL patients are treated with supradiaphragmatic radiotherapy means that this is a group of women who have most likely accumulated significant DNA damage and for whom a compromised capacity for DNA repair may have long term consequences. It seems therefore particularly vital for these women to have functioning repair mechanisms and any deficiency could lead to deleterious mutations being left unrepaired. I therefore hypothesise that a subset of HLBC women may have deficient DNA repair mechanisms that lead to mutations being left unrepaired after exposure to radiation. Such unrepaired mutations would then provide a basis for carcinogenesis. To assess this hypothesis, I performed a targeted NGS sequencing study to obtain the sequences of HLBC cases and controls for 410 genes implicated in DNA repair.

3.2 Materials and methods

3.2.1 Sample preparation

The samples used for this analysis were the same as those used in Chapter 2. For samples that had insufficient DNA remaining upon completion of the GWAS, fresh aliquots of buffy coat were recovered from LN2 storage and DNA was extracted. These extractions were performed by using Qiagen DNA Purification from Blood or Body Fluids Spin kits (Qiagen). DNA

concentrations were initially assessed using a Qubit 3.0 Fluorometer and subsequently confirmed using PicoGreen.

3.2.2 Library preparation

The library preparation protocol was designed by Dr Sarah Maguire in collaboration with the ICR Oncogenetics team. As part of this protocol, custom Illumina-compatible adapters with a total of 96 unique indexes were generated. This enables 96 samples to be pooled and sequenced together on a single lane, while remaining identifiable in downstream bioinformatic analyses. The indexing of each sample therefore provides a cost-effective manner of sequencing many individuals. Stock DNA samples were diluted to a concentration of 500ng in 130µL of low TE buffer (Tris-EDTA; 10mM Tris base, 0.1mM EDTA). Samples were sheared on a Covaris M220 in Covaris 96 microTUBE well plates.

Following the shearing of DNA samples, fragment overhangs were repaired by eluting DNA fragments into a master mix comprised of T4 DNA ligase buffer, deoxynucleotide triphosphates (dNTPs), an end-repair enzyme mix and water (Table 15). Samples were incubated for 30 minutes at 20°C. This step, and each subsequent step, was followed by a clean-up step which consisted of binding the DNA fragments to SeraMag SpeedBeads suspended in an 8.9% polyethylene glycol (PEG8000) solution, discarding the supernatant and performing an ethanol wash. DNA was dried for two minutes

at room temperature. In order to allow binding of adapters to the repaired DNA fragments, a poly(A) tail was added to each end of the fragments. The DNA fragments were eluted into a master mix comprised of 10x New England Biolabs (NEB) NEB2 buffer, 10mM deoxyadenosine triphosphate (dATPs), Klenow fragment (3' to 5'' excision) and water (Table 16). Samples were incubated for 30 minutes at 37°C for the A-tail ligation reaction to take place. Next, adapters were ligated to the DNA fragments. The DNA fragments were eluted into a master mix comprised of 10x T4 DNA ligase buffer, T4 DNA Ligase, water and the 96 adapters with unique indexes (Table 17). Samples were incubated at 16°C overnight to allow the ligation reaction to take place. Adapter sequences were completed and libraries were amplified by PCR for six cycles (see conditions below) using a Kapa HiFi master mix comprised of Kapa HiFi Ready Mix, primers and water (Table 18). Following this, a final clean-up was performed and the libraries were eluted into Tris-HCI buffer (EB buffer) and transferred to a clean 96 well plate.

95°C	45 seconds	
98°C	15 seconds	X6 cycles
60°C	30 seconds	
72°C	1 minute	
72°C	3 minutes	
10°C	Hold	

Following library preparation, samples were quantified using an Agilent High Sensitivity DNA Kit (Agilent) on an Agilent Bionanalyzer to assess the size distribution of the fragments. For each run (96 samples), distributions were checked for eleven randomly selected samples. Average sizes were 350-400

base pairs (Figure 25). Next, the samples were quantified using a Kapa Library Quantification Kit (Roche). This quantification was performed on a Roche LightCycler 480. Samples were pooled in batches of 48 prior to target enrichment. Equimolar ratios of each library based on qPCR quantification (approximately 21 ng) were combined to a 1 µg total quantity as input for the hybridisation reaction.

Table 15. Composition of the end-repair master mix used for repairing the overhangs of sheared DNA fragments during library preparation.

End repair master mix	Volume per sample
H_20	41.75µl
10x T4 DNA ligase buffer	5µl
10mM dNTP mix	2µl
End repair enzyme Mix	1.25µl
Total Volume	**50µl**

Table 16. Composition of the A-tail master mix used for ligating poly(A) tails to the repaired DNA fragments during library preparation.

A-tail master mix	Volume per sample
H_20	43.25µl
10x NEB buffer	5µl
10mM dATP mix	1µl
Klenow Fragment (3' to 5' excision)	0.75µl
Total Volume	**50µl**

Table 17. Composition of the adapter ligation master mix used for ligating adapters to the DNA fragments during library preparation. A volume of 2µl of 120 base pair adapters containing a 7 base pair P7 index (15µM) was added to each sample separately, bringing the total reaction volume to 30µl.

Adapter ligation master mix	Volume per sample
H_20	24µl
10x T4 DNA ligase buffer	3µl
T4 DNA Ligase	1µl
Total Volume	**28µl**

Table 18. Composition of the Kapa HiFi master mix used for completing adapter sequences and minimally amplifying the libraries during library preparation.

KAPA HiFi PCR master mix	Volume per sample
H_20	11.5µl
2x KAPA HiFi Ready Mix	12.5µl
P5/P7 primer mix (50µM)	1µl
Total Volume	**25µl**

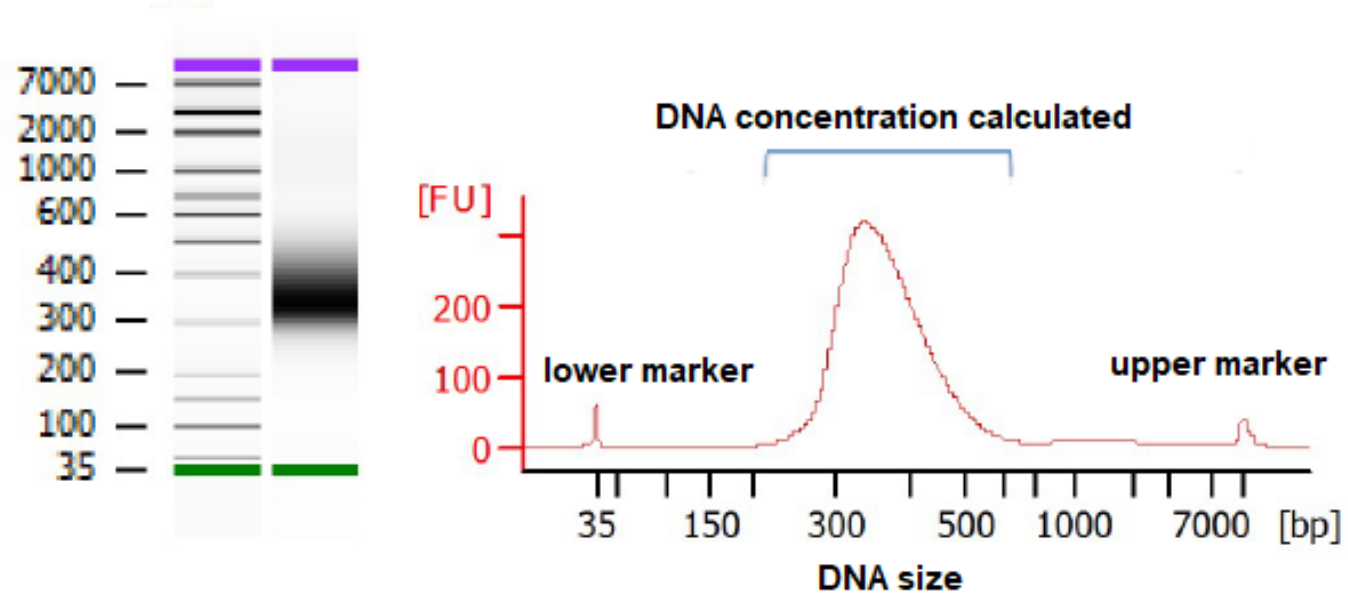

Figure 25. Example of the size distribution of DNA fragments obtained with the Bioanalyzer during library preparation (prior to target enrichment). For the plot on the right, the x-axis represents the DNA size (base pairs) and the y-axis represents the signal intensity (FU). The concentration of DNA was calculated using the area shown by the blue bracket. Positions of the lower and upper markers are indicated on the plot. They are also shown in the form of a gel on the left. Lower and upper markers are represented by green and purple lines respectively and the DNA appears as a strong dark band.

3.2.3 Target enrichment design

A SureSelect Custom Target Enrichment Panel was designed by Dr Sarah Maguire to enable target capture of 410 genes. This gene panel, totalling a 1.5Mb sequence, predominantly contains DNA repair genes (i.e. genes known to be involved some way with DNA repair), as well as BC predisposition genes and prostate cancer genes. It was designed using Agilent's eArray software and involved generating 120-mer RNA baits tiling the target regions with an average of eight baits per target.

3.2.4 Target enrichment

Target enrichment was performed based on the SureSelect protocol (Agilent). Pooled genomic libraries were lyophilised in a vacuum concentrator at 37°C and reconstituted in 3.4 µl of water. Genomic libraries were mixed with Block Mix (5.6µl) to make blocked genomic libraries. These were added to a 96 well plate which was then sealed with strip caps. Blocked genomic libraries were incubated on a thermocycler at 95°C for five minutes and then at 65°C for five minutes. Hybridisation buffer mix was added to capture mix to make the capture library hybridisation mix (Table 19). This was mixed by vortexing at high speed for five seconds and centrifuging briefly. A volume of 20 µl of capture library hybridisation mix was added to each well with a blocked library and mixed quickly by pipetting to make the enrichment

reactions. The enrichment reactions were overlaid with mineral oil and sealed with strip caps. Enrichment reactions were incubated at 65°C for 72 hours.

3.2.5 Capture of enriched libraries

For each enrichment reaction, 50 µl of Dynal MyOne Streptavadin T1 magnetic beads (Invitrogen) were washed, three times, with 200 µl of SureSelect Binding Buffer. The beads were then re-suspended in 200 µl of Binding Buffer per reaction and transferred to individual micro-centrifuge tubes. While the enrichment reaction plate was kept at 65°C on a thermocycler, enrichment reactions were transferred to the prepared beads. Samples and beads were mixed briefly by inverting the tubes. Beads and samples were incubated on a nutator for 30 minutes at room temperature. Following this, the bead and sample mix was gently centrifuged and placed on a magnet until the supernatant cleared. The supernatant was removed and discarded. The beads were re-suspended in 500 µl of SureSelect Wash Buffer 1 and vortexed to mix. The Wash buffer 1 and beads mix was then incubated at room temperature for 15 minutes, with vortexing occasionally being applied. The Wash Buffer 1 and beads mix was gently centrifuged and placed on a magnet until the supernatant cleared. The supernatant was removed and discarded. Beads were re-suspended in 500µl of SureSelect Wash Buffer 2 (pre-warmed to 65°C) and vortexed to mix. The Wash Buffer 2 and beads mix was incubated at 65°C for ten minutes, vortexing occasionally. The Wash Buffer 2 and beads mix was gently centrifuged and

placed on a magnet until the supernatant cleared. The supernatant was removed and discarded. The Wash Buffer 2 wash procedure was repeated for a total of three times. The beads were re-suspended in Kapa HiFi PCR mix (Table 20) and moved to a PCR plate for off bead PCR using the same conditions detailed for the PCR reaction discussed in Section 3.2.2, but this time for 13 cycles.

Clean-up of the PCR reaction was carried out using SeraMag Speedbeads. A volume of 3 µl of SeraMag SpeedBeads per enrichment reaction was placed on a magnet and the supernatant was removed. The beads were re-suspended in 950 µl of TE buffer, vortexed and placed on a magnet. The supernatant was then removed. TE wash was performed a total of three times. The beads were re-suspended in 40 µl of 20% PEG solution per enrichment reaction and placed in a separate micro-centrifuge tube. The PCR reaction plate was then placed on a magnet for five minutes until the supernatant cleared. 50 µl of the amplified library was added to the prepared SpeedBeads. The full volume was then mixed by pipetting 15 times and incubated at room temperature for five minutes. The amplified enriched libraries and beads were placed on a magnet for five minutes until the supernatant cleared after which it was removed and discarded. A volume of 150 µl of 70% ethanol was used to wash the beads. This was incubated for 30 seconds and ethanol was removed. The beads were then dried at room temperature for five minutes. To elute the enriched libraries, beads were re-suspended in 30 µl of EB buffer and incubated at room temperature for five minutes. The EB and bead mix was then placed on a magnet for three

minutes and the supernatant, representing the final captured library for each pool of 48 samples, was transferred to a new micro-centrifuge tube.

Libraries were quantified a final time with the Bioanalyser and by qPCR using the Kapa Illumina Library Quantification Kit (Roche). Two pools of 48 samples were combined in equimolar concentrations (96 samples in total) and were subjected to 100 cycles of paired-end sequencing on a single lane of a Hiseq 2500. Sequencing was performed by the ICR Tumour Profiling Unit (TPU).

Table 19. Composition of the capture mix, hybridisation mix and block mix used to enrich the targets for sequencing.

Capture mix	Volume per reaction
Bait library	2µl
RNAase block	5µl
Total	**7µl**

Hybridisation buffer	Volume per reaction
Sure Select Hyb 1	6.63µl
Sure Select Hyb 2	0.27µl
Sure Select Hyb 3	2.65µl
Sure Select Hyb 4	3.45µl
Total	**13µl**

Block mix	Volume per reaction
Sure Select Block 1	2.5µl
Sure Select Block 2	2.5µl
Index Block Reaction	0.6µl
Total	**5.6µl**

Table 20. Composition of the Kapa HiFi master mix used to amplify the captured libraries for sequencing.

Master mix	Volume per reaction
2x Kapa HiFi ReadyMix	25µl
P5/P7 primer mix (50µl)	2µl
H_20	23µl
Total	**50µl**

3.2.6 Quality control, alignment and variant calling

Samples were aligned to build hg19 of the human genome using the Burrows-Wheeler Aligner (BWA) version 0.7.12 (113). Contaminated samples were identified using VerifyBamID (114). Variants were called using the Genome Analysis Toolkit (GATK) following best practise guidelines for germline samples (115) including mapping the raw reads to the reference genome, marking duplicates and base recalibration (116). Genotype likelihoods were calculated separately for each sample and this was followed by joint genotyping and variant recalibration. Variants with low depth (DP <8) and low quality (GQ <20) were discarded.

Concordance between the sequencing data and GWAS array data was assessed with the Genotype Library and Utilities (GLU) package (117). Samples which were discordant were removed from further analyses. Samples with > 20% missing data were identified and removed using Variant Association Tools (VAT) (118). Similarly, variants with over 10% of data missing were excluded. Due to the nature of capture arrays, it is common for samples to have different missing rates at different target regions. For this reason, relatively lenient cut-off criteria were used at this stage. Nevertheless, further filtration took place in a gene specific manner when running the association tests. Functional annotation of the variants was performed using ANNOVAR and dbNSFP (119,120). Rare variants were defined as having a MAF of less than 1% in either the 1000 Genomes Project European samples

or the Exome Aggregation Consortium (ExAC) non-Finnish European samples (121,122). The analyses were focussed exclusively on coding variants. Coding rare variants were defined as nonsynonymous Single Nucleotide Variants (SNVs), insertion/deletions (indels), stopgain or splice variants that localised to exons. Truncating stop gain or indel rare variants were defined as those that result in the premature incorporation of a termination codon in the open reading frame of a gene. Using VAT, variants were aggregated by gene in order to conduct association tests (118). The aggregation was performed by implementing the Gene- or Region-based Analysis of Variants of Intermediate and Low frequency (GRANVIL) method developed by Morris & Zeggini (123). This method, which consists in aggregating the number rare variants used as regressors in a logistic regression model, was used to aggregate all variants falling within the exon regions of the genes. Splicing variants were not considered as none were obtained following QC (see Section 3.3.1).

3.2.7 Statistical analyses

All association tests were carried out using VAT (118). Per gene, only variants and samples with a completion rate of over 95% were included in the tests. In order to address the fact that the effects of coding rare variants could be in both directions (i.e. some variants may be protective while others may be deleterious) and that few variants may be causal, this group was subjected to an optimised sequence kernel association test (SKAT-O) in order to

identify gene-level associations. With the assumption that truncating variants have a high prior probability of being pathogenic, a weighted sum statistic (WSS) burden test was applied to this group in order to obtain gene-level associations (124). This test, developed by Madsen and Browning, consists in weighting variants inversely by their frequency in controls prior to aggregating them, thus up-weighting the “rarer” variants. This method is based on the assumption that the rarer the variant, the larger its risk effect. The “rarer” variants therefore have the highest weights and have an increased contribution to the collapsed variant score (125).

3.3 Results

3.3.1 Quality control results

Inspection of the sequencing results indicated that five samples were not concordant with SNP calls from the GWAS genotyping arrays. In addition, a total of five samples were found to be contaminated. Sequencing failed for 38 samples and five samples had low completion rates (Table 21). Average median coverage was 87X meaning that each base of the capture region was covered a median of 87 times (Table 22). The percentage of duplicate reads was low (12.4%) indicating that no over-amplification of libraries occurred. The average percentage of on-target reads (30.4%) was in line with on-target reads for a panel of this size (Table 22). A small number of variants were removed due to low completion rates (a threshold of 10% was used in accordance with the GATK best practise) (Table 23) (115). Multiallelic variants were removed, due to the fact it would not be possible with the used methods to determine which allele is disease-causing. In addition, complex indels (i.e. large indels of over 5 base pairs that commonly result from sequencing errors and misalignment) were removed.

All QC procedures resulted in a total of 898 testable samples, of which 303 were cases and 595 were controls. The total number of testable rare variants obtained after QC was 2,089, of which 139 were truncating. No splicing variants were obtained after QC. Coding rare variants (2,089 variants) and

truncating rare variants (139 variants) were aggregated separately into genes. This resulted in a total of 252 testable genes, comprised of 219 genes with rare coding variants and 33 genes with rare truncating variants.

Table 21. Summary of sample QC for the rare variant analyses.

	No. cases	No. controls
Pre-QC total	321	630
Contaminated samples	- 1	- 4
Failed concordance QC	- 2	- 3
Failed sequencing	- 14	- 24
Samples with >20% data missing	- 1	- 4
Post-QC total	303	595

Table 22. Sequencing metrics for the rare variant analyses. Averages were shown for all samples which passed QC.

QC metric	Average for all samples
Total reads	4,199,102
Mapped reads	4,023,679
Paired reads	3,970,949
On-target reads	1,577,391
Mean coverage	87.8 X
Median coverage	87.2 X
Duplicate reads (%)	12.4 %
Percentage of on-target reads	30.4 %
Concordance rate with array sample	99.3 %

Table 23. Summary of variant QC for the rare variant analyses.

	Removed	Total
Pre-QC total	-	17,103
Multiallelic variants and complex indel errors	206	16,897
Variants with >10% data missing	1211	15,686
ExAC or 1000 Genome Project MAF >0.01	559	15,127
Total rare variants	-	15,127
Total coding rare variants	13,038	**2089**
Total truncating rare variants	1950	**139**

3.3.2 Rare inherited variants predisposing to radiation-induced breast cancer

Of the 219 genes with coding rare variants, eight genes reached statistical significance (P-value < 0.05) in the SKAT-O tests (Table 24 and Supplementary Table 3). For *RECQL5*, *CDKN1B*, *HSCB*, *RAD23B* and *UBE2N* there was an excess of coding mutations in cases compared to controls, indicating that they were associated with an increased risk of HLBC; coding mutations in *MED12, OGG1* and *TOP2A* were associated with a decreased risk of HLBC.

RecQ like helicase 5 (*RECQL5*) is a DNA helicase that plays an important role in maintaining genomic stability. This gene, which belongs to the *RECQ* gene family, produces a protein that has a number of roles in replication, recombination and DNA repair. In the context of the latter, the RECQL5 protein is known to be able to displace *RAD51* from single stranded DNA, hence preventing aberrant homologous recombination (126,127). Due to its vital role in dissociating RAD51-DNA filaments, *RECQL5* is characterised as an oncogene and represents a promising biomarker for therapeutic interventions. It is known to be amplified in 20-25% of breast cancers and increased expression has been shown to lead to significant dysregulation of DNA repair (128).

CDKN1B is a cyclin-dependent kinase inhibitor. The role of this enzyme is to bind to cyclin E-CDK2 or cyclin D-CDK4 complexes and prevent their activation in order to control the cell cycle progression at G1. *CDKN1B* is therefore a vital tumour suppressor gene that is necessary for preserving genomic integrity and preventing aberrant cells from progressing through the cell cycle. Loss of *CDKN1B* is known to cause increased genomic instability and induce radio-resistance in luminal BCs (129).

HscB mitochondrial iron-sulphur cluster cochaperone (*HSCB*) encodes a J-type co-chaperone. This gene is involved in the synthesis of iron-sulphur clusters which are protein co-factors important for the redox reactions of the mitochondrial electron transport chain. In the context of DNA repair, *HSCB* has been associated with increased susceptibility to Friedreich ataxia, a disease characterised by elevated amounts of double-strand breaks (130).

UV excision repair protein RAD23 homolog B (*RAD23B*) encodes a protein that is involved in nucleotide excision repair (NER). This protein forms a complex with the xeroderma pigmentosum group C (XPC) and can recognize DNA lesions caused by a variety of sources including UV and chemotherapeutics drugs like cisplatin) (131). By recognising these lesions, the XPC-RAD23 complex is the initiator of the global genomic nucleotide excision repair (GG-NER) pathway (one of the two sub-pathways of NER). *RAD23B* has been shown to be downregulated in highly invasive BCs and

immunohistochemistry shows that it is significant associated with grade 3 BCs (132).

UBE2N encodes an E2 ubiquitin-conjugating enzyme. The role of this enzyme is to ubiquitinise abnormal and short-lived proteins, targeting them for degradation. This ubuquitinisation results in a number of ubiquitin cascades required for DNA repair to take place (133). *UBE2N* is implicated in DNA repair following events such as UV exposure and ionising radiation. One example is its implication in Rad6/Rad18-dependent post-replication repair and translesion synthesis after exposure to UV and ionizing radiation. Deficiencies in *UBE2N* have been associated with a comprised double-strand break repair using homologous recombination (134). In addition, it has been shown to be required for H2A-histone ubiquitination following DNA damage as well as for *TP53BP1* and *BRCA1* foci formation to lesions (135).

MED12 encodes a protein called the mediator complex subunit 12. This subunit belongs to the CDK8 subcomplex which modulates interactions between Polymerase II and Mediator, thereby regulating transcription initiation (136). *MED12* has been associated with breast fibroadenomas (FAs), breast phyllodes as well as prostate cancer (137,138).

The 8-oxogunanine DNA glycosylase (*OGG1*) gene encodes an enzyme that is involved in Base Excision Repair (BER). This enzyme is responsible for cleaving 8-oxoguanine (8-oxoG) bases, which are mutated bases that occur

as a result of exposure to reactive oxygen species (139). Due to its vital role in DNA repair, *OGG1* represents a vital tumour suppressor. It is associated with a number of cancers including lung, oesophageal cancer and prostate cancer (139–141) . In particular, *OGG1* has been associated with ovarian cancer risk in *BRCA1* mutation carriers. This association was discovered as part of a study investigating of polymorphisms in genes involved in BER, in relation to *BRCA1* and *BRCA2* status (142). The SNP rs2304277 of *OGG1* was found to be significantly associated with ovarian cancer risk in *BRCA1* mutation carriers (HR = 1.12 [1.03-1.21], P-value = 4.86×10^{-3}).

DNA topoisomerase II alpha (*TOP2A*) encodes an enzyme that is involved in modifying the topological states of DNA during processes such as DNA transcription and replication. In particular, it is involved in chromosomal condensation and chromatid separation, and provides relief of torsional stress during transcription and replication (143,144) . *TOP2A* acts by catalysing the transient breaking and re-joining of two strands of DNA (145). This gene is known to be associated with ataxia-telangiectasia, a known DNA repair disorder (146). It has also been shown to be amplified in or deleted in BC (147).

Table 24. Significant genes from association testing (SKAT-O) of 219 genes with coding rare variants. Only the statistically significant genes are shown. Results are ordered according to P-value.

Gene	Sample Size	Case MAF	Control MAF	Beta	P-value
RECQL5	824	0.0014	0.00076	7.02	0.0043
CDKN1B	891	0.0017	0.00	1.99	0.0072
HSCB	894	0.0050	0.00	1.99	0.0074
RAD23B	894	0.0017	0.00	1.33	0.019
UBE2N	894	0.0017	0.00	1.33	0.019
MED12	842	0.0014	0.0032	-2.38	0.027
OGG1	894	0.00064	0.0019	-6.45	0.032
TOP2A	890	0.00	0.0015	-3.03	0.046

Of the 33 genes with truncating rare variants, two genes, *RBBP8* and *FANCA*, reached statistical significance in the WSS burden tests (Table 25 and Supplementary Table 4). After correction for multiple testing, no genes were significantly associated with HLBC. There was no evidence of an excess of either coding or loss of function mutations amongst established breast cancer predisposition genes including *BRCA1/2*, *ATM*, *CHEK2* or *PALB2*.

RB binding protein 8 endonuclease (*RBBP8*), encodes an endonuclease called CtIP. This endonuclease is known to be involved in double-strand breaks repair (148). It does so by promoting homologous recombination (HR), classical non-homologous end-joining (c-NHEJ) and alternative non-homologous end-joining (alt-NHEJ), thus maintaining genome integrity (149–152). Due to its role in the repair of double-strand breaks, *RBBP8* is believed to interact with *BRCA1* and its loss in BC leads disruptions in homologous recombination repair and sensitisation to PARP inhibitors (153).

The second gene hit, FA complementation group A (*FANCA*) belongs to a well-known family of genes associated with defective DNA repair and genome instability: the Fanconi anaemia complementation group (*FANC*). This gene is believed to encode a DNA repair protein that plays a role in post-replication repair or during cell cycle checkpoints. Mutations in *FANCA* can severely disrupt DNA repair. Mutations in FANCA are thought to be a cause of Fanconi anaemia (FA) disease and germline mutations in *FANCA* have

been shown to result in an increased sensitivity to DNA damaging agents (154). *FANCA* has also been shown to directly interact with *BRCA1* (155).

Table 25. Significant genes from association testing (weighted burden test) on 33 genes with truncating rare variants. Only the statistically significant genes (P-value < 0.05) are shown. Results are ordered according to P-value.

Gene	Sample Size	Case MAF	Control MAF	Beta	P-value
RBBP8	883	0.00093	0.00058	1.98	0.019
FANCA	890	0.0010	0.00082	1.99	0.019

3.4 Discussion

The targeted sequencing of HLBC cases and controls identified a total of ten genes associated with radiation-induced BC, of which eight were identified by testing genes with coding rare variants and two by testing genes with truncating rare variants. Although only nominally significant and requiring independent validation, the discovery of associations using this custom panel suggests that a subset of HLBC cases may have rare moderately penetrant susceptibility variants in genes involved in DNA damage responses. Deficiency in repairing DNA damage would certainly increase the risk of carcinogenesis in individuals being treated with radiotherapy. The observations of rare variants associated with HLBC suggest that non-genetic risk factors (age at HL treatment, dose of radiotherapy etc) may not the sole factors explaining why some HL survivors go on to develop BC and some do not. Instead, it is the combination of these non-genetic factors and deficient DNA repair mechanisms that could lead to carcinogenesis. This means that while HL treatments may continue to improve and become safer, the risk conferred by rare variants will persist and the cumulative risks of radiation-induced BC in HL survivors could remain at similar rates to their current values. According to Swerdlow *et al*, for a follow-up of 40 years, the cumulative risk for women who received 40Gy of mantle radiotherapy is currently 48% (65). These high risks therefore warrant a continued need for improved prevention and treatment strategies.

However, one must be cautious in the interpretation of these results for a number of reasons. First of all, similarly to the GWAS, this analysis is largely underpowered. And this is reflected by the large P-values obtained (i.e. all are close to 0.05). Although this means that many more samples would be required to detect genes with small effects, the dataset used in this project is the largest worldwide cohort of HL survivors and this analysis is the first attempted investigation of the role of rare variants in HLBC. However, since a number of rare variants underlying sporadic and familial BC have very large effects (for example *BCRA1/2*, *PALB2*, *CHEK2* etc), this analysis was also justified by the fact it could potentially have detected variants with very high effects.

To address the limitations of this study, the results will have to be validated. This will be done by combining this dataset with the one from collaborators at the Division of Cancer Epidemiology and Genetics of the National Cancer Institute (NCI) (the same collaborators mentioned in Section 2.4). Moreover, whole exome sequencing will be used instead of targeted sequencing, enabling a less biased approach that requires no prior hypothesis about the types of genes underlying HLBC. Whole genome sequencing (WGS) will eventually be required in order to assess the role of non-coding variants.

Chapter 4 - Conclusions and future directions

The aim of this thesis was to identify germline variants that confer an increased risk of HLBC in order to gain a better understanding of the aetiology of this disease. This was achieved by assessing the contribution of both common and rare polymorphisms to HLBC risk in the largest cohort of HL survivors worldwide.

Firstly, I conducted a GWAS that identified a total of total 72 SNPs associated with HLBC (P-value < 5 x 10^{-6}). The significant SNPs form a total of 10 independent signals at 1q36.13, 5p21.3, 6p12.3, 7q22.1, 9p24.1, 10p12.2, 12q24.21, 14q.12, 18q11.2 and 20p12.2. Among these loci, three peaks of association were discovered at 5q21.3, 9p24.1 and 20p12. Most of the top SNPs for these signals had elevated HRs (HR > 1.5). While their exact functional role in HLBC remains to be determined, these SNPs represent novel signals that could shed light on the aetiology of HLBC. Of the six SNPs previously associated with HLBC, only one was significant in the GWAS (rs1219648 located to *FGFR2* at 10q26.13).

I identified a total of 11 known FBC loci associated with HLBC risk (P-value < 0.05), a number of which had effects that were significantly larger than those for sporadic BC suggesting that these SNPs had larger effects upon BC risk in the context of radiation exposure.

Despite being the largest GWAS ever performed on HL survivors, this study, like previous studies on radiation-induced BC, suffers from a lack of statistical power due to small sample size. It is also worth remembering that previous studies such as Morton *et al* did not have a cohort that was as homogeneous as the one used in this project, making direct comparisons of results with them more difficult. Aside from sample size, the GWAS performed in my study is limited by some additional considerations. For instance, it was also not possible to obtain clinical data regarding menarche status at the time of HL treatment, a known risk factor for HLBC (as discussed in Section 1.3.2). The clinical data that was obtained was not always fully complete, with some patients having received an unknown dose of radiotherapy. The incompleteness of the data used as the covariates constrained the choice of categories used for the association test (as each category requires a minimal number of cases and controls in it). Additionally, the design of the Cox regression could have potentially benefited from one-to-one matching of cases and controls, based on their duration of follow-up. However, due to the sample size, it was not possible to find matching cases and controls for each exact duration of follow-up. Finally, the sampling process itself may suffer from a bias caused by the fact that patients that are particularly ill (i.e. patients who may suffer from the most aggressive forms of HLBC) may have refused to participate in the National Recall study due to their health condition. This could mean that the cohort used in this project does not fully represent the true distribution of HLBC patients and therefore that the analyses in this project may be missing potential variants that tend to underlie the most

severe cases of HLBC. The results of the GWAS will need be validated. Plans to do so through collaboration with Morton *et al* are underway. Following this, direct genotyping of the identified signals in the combined datasets, as well as those identified in each separate one, will need to be performed, in order to truly replicate findings.

Next, I demonstrated that by using the 313 SNPs used by Mavaddat *et al* in a sporadic BC PRS score, it is possible to stratify HLBC cases and controls. I also constructed a PRS based on the 313 SNPs which demonstrated a trend (albeit not statistically significant) of increasing HLBC risk with increasing PRS scores. While the results of the logistic regression were not significant, the ORs were consistent with those of the PRS developed by Mavaddat *et al* for sporadic FBC implying a substantial overlap in polygenic susceptibility between HLBC and sporadic BC. This PRS could therefore be used in similar way as it is currently used by clinicians for healthy women with no prior history of cancer. More specifically, it could be used to identify HL survivors who should begin breast screening at an earlier date than the usual recommended age. HL patients currently join the national screening program, which consists in a mammography once every three years, at the age of 50 (although they are eligible from age 47) (156). However, given the fact that women in the top percentiles of the PRS distribution are likely at a very high risk, a number of HL survivors should be screened sooner and for longer (65). Potential guidelines to follow would be those from the National Comprehensive Network (NCCN) and the Childhood Oncology Group (COG), which are based in the United States (157,158). The NCCN recommends that women

who underwent chest or axillary irradiation undergo breast cancer screening eight to ten after HL treatment or at age 40, whichever comes first. In addition women who received who received chest irradiation between ages the ten and 30 should receive MRIs in addition to mammography (157,159). The COG recommends yearly breast self-examination from puberty until the age of 25, and then a self-examination every six months as well as a yearly mammogram and breast MRI beginning eight years after irradiation or at age 25, whichever occurs last (158,159).

It is important to remember that the PRS analysis suffers from the same statistical limitations as the GWAS and just like my GWAS results, this PRS will need to be validated in another dataset. This will be performed following the merging of my dataset with Morton *et al*'s cohort. The PRS will also most likely have to be updated, as more FBC SNPs are identified, in order to determine if the observed stratifications and trends are maintained. In addition, there are a number of further analyses that could be performed using PRS scores on the cohort. Indeed, since this PRS may not be entirely suited for HLBC, other PRS scores may be better suited for identifying HL survivors at risk of developing HLBC. One possibility could be to use a "best-fit" approach, whereby PRS scores are calculated using different P-value thresholds (i.e. by selecting different numbers of SNPs based on different P-value thresholds) and the PRS which results in the best stratification is selected (160). A completely different approach could be to select SNPs involved in a particular pathway or mechanism, calculate a PRS score and determine whether or not these SNPs alone can stratify cases and controls.

In the context of HLBC, one could select SNPs known to be involved in radiosensitivity and DNA damage response. The results of such a PRS could then give an indication as to whether or not DNA damage plays a role in HLBC (although statistical limitations due to small sample size would still apply in my dataset).

Finally, I performed a targeted NGS sequencing of HLBC cases and controls, identifying a total of ten genes that were nominally associated with radiation-induced BC, of which eight had coding rare variants and two had truncating rare variants. The associated genes were *RECQL5*, *CDKN1B*, *HSCB*, *RAD23B*, *UBE2N*, *MED12*, *OGG1* and *TOP2A*. Although the discovery of these associations support the hypothesis that a subset of women have mutations in DNA repair genes that predispose them to radiation-induced BC, one must be very cautious in the interpretation of these results, as this analysis is underpowered. Importantly, it is possible that additional rare variants, whether involved in DNA repair or not, play a role in HLBC susceptibility. Therefore, a less targeted sequencing approach of HL survivors could reveal further insights into the aetiology of the disease.

This project provides a broad overview of genetic susceptibility to radiation-induced BC in HL survivors. By assessing the contribution of both common and rare variants, I have identified novel candidate HLBC predisposition variants that warrant future validation and that might ultimately provide new

insights into the aetiology of HLBC. To continue to gain a better understanding of HLBC, it will be necessary to further investigate the disease in a number of ways. Firstly, as previously mentioned, by combining my cohort with collaborators at the Division of Cancer Epidemiology and Genetics of the National Cancer Institute (NCI), I hope to gain statistical power to further investigate common polymorphisms underlying HLBC. The construction of a best-fit PRS, as well as a radiosensitivity-based PRS, could potentially be a useful clinical tool to identify HL survivors in need of early breast screenings. In addition, exome sequencing of the combined datasets will allow for a less biased approach to studying rare variants and could reveal further insights into their role in HLBC. Secondly, HLBC tumours are much understudied. The cohort used for this project has the potential to answer questions pertaining to HLBC tumours in a variety of ways. Indeed, as part of the recruitment process for this cohort (the National Recall Study cohort), a subset of women gave consent to give pathological samples taken from surgery, in the form of formalin-fixed paraffin-embedded blocks (FFPE blocks). During my time at the ICR, with the help of Jane Lehiban and Penny Coulson from Professor Swerdlow's team, I collected over 1,000 tumour blocks from 260 patients. This constitutes the largest collection of HLBC tumours worldwide and will allow for these malignancies to be investigated on a scale that has not been possible before. Histological staining of these tumours will enable tumours to be classified according to the main BC subtypes in order to determine whether HLBC tumours are enriched for a particular subtype. Given the large sample size available, the results from this classification could ratify or contradict those of previous studies which were

based on much smaller datasets (161,162). Furthermore, WGS and RNA sequencing of the HLBC tumour blocks will provide a wealth of information about radiation-induced BC pathology. By looking at WGS data, one could determine whether HLBC tumours have specific mutational signatures, as was done for many types of cancers including sporadic BC by Alexandrov *et al* (163). A particular focus would be to determine if HLBC tumours present mutational signatures 3 and 6, which are associated with *BRCA1/2* mutations and DNA mismatch repair (MMR) deficiency respectively. Indeed, the presence of these signatures in HLBC tumours would be a further indication that deficiencies in DNA repair play a role in the aetiology of HLBC. In addition, by determining which signatures are present in HLBC tumours, one could compare their distribution with those found in sporadic BC (which include Signature 3) in order to gain more insights as to whether HLBC truly differs from sporadic BC. WGS data would also enable the investigation of larger variants such as CNVs, microsatellites and chromosomal aberrations and to look for markers of genomic instability, as was done by Broeks *et al* (164). Furthermore, RNAseq analysis of the HLBC tumours could reveal additional insights as to whether or not HLBC truly differs from sporadic BC. For instance, by building a gene expression profile as was done in Perou *et al*, one could directly compare HLBC to the other subtypes of BC and truly determine if HLBC should be classified as separate subtype (165,166). Finally, a number of HL patients in the NRS cohort developed bilateral BCs, either sequentially or simultaneously. The comparison of bilateral tumours could reveal crucial insights into HLBC mechanisms. If bilateral tumours are found to be mostly concordant, in terms of mutational profiles or expression

profiles, this could be an indication that HLBC tumours are not constrained to a specific subtype and rather develop in way that is not necessarily predictable. Conversely, if tumour pairs are found to resemble each other, this would suggest that radiation-induced BC tumours develop following specific mechanisms and pathways. Knowledge of such patterns could play a crucial way in the way HLBC is treated, as it could help clinicians decide which pathways to target and which treatments to use. While there remains much to discover about HLBC tumours, it is becoming increasingly clear that there is a need to characterise second cancers occurring in childhood cancer survivors. By identifying both common and rare polymorphisms underlying HLBC, this project contributes to the better understanding of radiation-induced BC and provides new insights that will likely lead to better prevention and improved strategies for treating this cancer.

Bibliography

1. Küppers R. The biology of Hodgkin's lymphoma. Nat Rev Cancer. 2009;9:15–27.

2. Cancer Research UK [Internet]. Available from: https://www.cancerresearchuk.org/

3. Ambinder RF, Browning PJ, Lorenzana L, Leventhal BG, Cosenza H, Mann RB, et al. Epstein-Barr Virus and Childhood Hodgkin's Disease in Honduras and the United States. Am Soc Hematol. 1993;81(2):462–7.

4. Chang KL, Albújar PF, Chen YY, Johnson RM, Weiss LM. High prevalence of Epstein-Barr virus in the Reed-Sternberg cells of Hodgkin's disease occurring in Peru. Blood. 1993;81(2):496–501.

5. Glaser SL, Lin RJ, Stewart SL, Ambinder RF, Jarret RF, Brousset P, et al. Epstein-Barr Virus-Associated Hodgkin 's Disease : Epidemiologic Characteristics in International Data. Int J Cancer. 1997;70:375–82.

6. Hjalgrim H, Askling J, Rostgaard K, Hamilton-Dutoit S, Frisch M, Zhang J-S, et al. Characteristics of Hodgkin's Lymphoma after Infectious Mononucleosis. N Engl J Med. 2003;349(14):1324–32.

7. Landgren O, Engels EA, Pfeiffer RM, Gridley G, Mellemkjaer L, Olsen JH, et al. Autoimmunity and susceptibility to Hodgkin lymphoma: A population-based case-control study in Scandinavia. J Natl Cancer Inst. 2006;98(18):1321–30.

8. Patel P, Hanson DL, Sullivan PS, Novak RM, Moorman AC, Tong TC, et al. Incidence of Types of Cancer among HIV-Infected Persons Compared with the General Population in the United States, 1992–2003. Annakls Intern Med. 2008;148(10):728–36.

9. Monnereau A, Troussard LO, Berthou C, Fenaux P, Soubeyran P, Marit G, et al. Cigarette smoking, alcohol drinking, and risk of lymphoid neoplasms: results of a French case–control study. Cancer Causes Control. 2008;19(10):1147–60.

10. Briggs NC, Irene Hall H, Brann EA, Moriarty CJ, Levine RS. Cigarette smoking and risk of Hodgkin's disease: A population-based case-control study. Am J Epidemiol. 2002;156(11):1011–20.

11. Kamper-Jørgensen M, Rostgaard K, Glaser SL, Zahm SH, Cozen W, Smedby KE, et al. Cigarette smoking and risk of Hodgkin lymphoma and its subtypes: a pooled analysis from the International Lymphoma Epidemiology Consortium (InterLymph). Ann Oncol. 2013;24:2245–55.

12. Gorini G, Stagnaro E, Fontana V, Miligi L, Ramazzotti V, Amadori D, et al. Alcohol consumption and risk of Hodgkin's lymphoma and multiple myeloma: A multicentre case-control study. Ann Oncol. 2007;18(1):143–8.

13. Ingham II RR, Reagan JL, Dalia S, Furman M, Merhi B, Nemr S, et al. The Relationship Between Obesity and Lymphoma: A Meta-Analysis of Prospective Cohort Studies. Blood. 2011;118(21):5198.

14. Kushekhar K, Van Den Berg A, Nolte I, Hepkema B, Visser L,

Diepstra A. Genetic associations in classical Hodgkin lymphoma: A systematic review and insights into susceptibility mechanisms. Cancer Epidemiol Biomarkers Prev. 2014;23(12):2737–47.

15. Niens M, Jarrett RF, Hepkema B, Nolte IM, Diepstra A, Platteel M, et al. HLA-A*02 is associated with a reduced risk and HLA-A*01 with an increased risk of developing EBV+ Hodgkin lymphoma. Blood. 2007;110(9):3310–5.

16. Klitz W, Aldrich CL, Fildes N, Horning SJ, Begovich a B. Localization of predisposition to Hodgkin disease in the HLA class II region. Am J Hum Genet. 1994;54(3):497–505.

17. Huang X, Kouprie N, Kooistra W, Nolte I, van den Berg A, Visser L, et al. HLA Associations in Classical Hodgkin Lymphoma: EBV Status Matters. PLoS One. 2012;7(7):1–9.

18. Hjalgrima H, Rostgaarda K, Johnsonb PCD, Lake A, Shield L, Little A-M, et al. HLA-A alleles and infectious mononucleosis suggest a critical role for cytotoxic T-cell response in EBV-related Hodgkin lymphoma. Proc Natl Acad Sci. 2010;107(14):6400–5.

19. Urayama KY, Jarrett RF, Hjalgrim H, Diepstra A, Kamatani Y, Chabrier A, et al. Genome-wide association study of classical hodgkin lymphoma and epstein-barr virus status-defined subgroups. J Natl Cancer Inst. 2012;104(3):240–53.

20. Cozen W, Timofeeva MN, Li D, Diepstra A, Hazelett D, Delahaye-Sourdeix M, et al. A meta-analysis of Hodgkin lymphoma reveals 19p13.3 TCF3 as a novel susceptibility locus. Nat Commun. 2014;5.

21. Monroy CM, Cortes AC, Lopez MS, Jr. AMD, Etzel CJ, Younes A, et al. Hodgkin disease risk: Role of genetic polymorphisms and gene–gene interactions in inflammation pathway genes. Mol Carcinog. 2010;36–46.

22. Butterbach K, Beckmann L, de Sanjosé S, Benavente Y, Becker N, Foretova L, et al. Association of JAK-STAT pathway related genes with lymphoma risk: Results of a European case-control study (EpiLymph). Br J Haematol. 2011;153(3):318–33.

23. De Bruin ML, Sparidans J, van't Veer MB, Noordijk EM, Louwman MWJ, Zijlstra JM, et al. Breast Cancer Risk in Female Survivors of Hodgkin's Lymphoma: Lower Risk After Smaller Radiation Volumes. J Clin Oncol. 2009;27(26):4239–46.

24. Küppers R, Hansmannb M-L. The Hodkin and Reed/Sternberg cell. Int J Biochem Cell Biol. 2005;37(3):511–7.

25. Nancy B, Harris L, Jaffe ES, Stein H, Banks PM, Chan JKC, et al. A Revised European-American Classification of Lymphoid Neoplasms : Blood J. 1994;84(5):1361–92.

26. Swerdlow SH, Campo E, Harris NL, Jaffe ES, Pileri SA, Stein H, et al. WHO Classification of Tumours: Pathology and Genetics. Tumours of Haematopoietic and Lymphoid Tissues. Vol. 25, Clinical and Laboratory Haematology. 2003. 201-201 p.

27. Gulley BML, Eagan PA, Quintanilla-martinez L, Picado AL, Smir BN, Childs C, et al. Epstein-Barr Virus DNA Is abundant and monoclonal in the Reed-Sternberg cells of Hodgkin's Disease: association with

mixed cellularity subtype hispanic american ethnicity. Blood. 1994;83(6):1595–602.

28. Smithers DW. Summary of Papers Delivered at the Conference on Staging in Hodgkin's Disease (Ann Arbor). Cancer Res. 1992;31(11):1869–70.

29. National Cancer Institute [Internet]. Available from: https://www.cancer.gov

30. Jacobs EM, Peters FC, Luce JK, Zippin C, Wood DA. Mechlorethamine HCl and Cyclophosphamide in the Treatment of Hodgkin's Disease and the Lymphomas. J Am Med Assoc. 1968;203(6):392–8.

31. Kaplan HS. The Radical Radiotherapy of Regionally Localized Hodgkin's Disease. Radiology. 1962;78(4):553–61.

32. Rosenberg SA, Kaplan HS. Evidence for an Orderly Progression in the Spread of Hodgkin's Disease. Cancer Res. 1966;26(June):1225–31.

33. Rosenberg SA, Kaplan HS. The evolution and summary results of the Stanford randomized clinical trials of the management of Hodgkin's disease: 1962–1984. Radiat Oncol. 1985;11(1):5–22.

34. Kaplan HS, Rosenberg SA. Extended-Field Radical Radiotherapy in Advanced Hodgkin's Disease: Short-Term Results of 2 Randomized Clinical Trials. Cancer Res. 1966;26(June):1268–76.

35. Longo DL, Gladstein E, Duffey PL, Young RC, Hubbard SM, Wesley

WJU, et al. Combination Chemotherapy in the Treatment of Advanced Hodgkin's Disease. Ann Intern Med. 1970;73(6):881–95.

36. Canellos GP, Niedzwiecki D. Long-Term Follow-up of Hodgkin's Disease Trial. N Engl J Med. 2002;346:1417–8.

37. Nissen N, Nordentof A. Radiotherapy versus combined modality treatment of Stage I and II Hodgkin's disease. Cancer Treat Rep. 1982;66(4):799–803.

38. Carde P, Karrasch M, Fortpied C, Brice P, Khaled H, Casasnovas O, et al. Eight Cycles of ABVD Versus Four Cycles of BEACOPP escalated Plus Four Cycles of BEACOPP baseline in Stage III to IV, International Prognostic Score ≥ 3, High-Risk Hodgkin Lymphoma: First Results of the Phase III EORTC 20012 Intergroup Trial. J Clin Oncol. 2016;34(17):2028–36.

39. Majhail NS, Weisdorf DJ, Defor TE, Miller JS, McGlave PB, Slungaard A, et al. Long-term results of autologous stem cell transplantation for primary refractory or relapsed Hodgkin's lymphoma. Biol Blood Marrow Transplant. 2006;12(10):1065–72.

40. Millenson MM, Armand P, Lesokhin AM, Zhu L, Rodig SJ, Scott EC, et al. PD-1 Blockade with Nivolumab in Relapsed or Refractory Hodgkin's Lymphoma. N Engl J Med. 2014;372(4):311–9.

41. Okeley NM, Miyamoto JB, Zhang X, Sanderson RJ, Benjamin DR, Sievers EL, et al. Intracellular activation of SGN-35, a potent anti-CD30 antibody-drug conjugate. Clin Cancer Res. 2010;16(3):888–97.

42. Francisco JA, Cerveny JA, Meyer DL, Mixan BJ, Klussman K, Chace DF, et al. cAC10-vcMMAE, an anti-CD30–monomethyl auristatin E conjugate with potent and selective antitumor activity. Blood. 2003;102:1458–65.

43. Allemani C, Sant M, De Angelis R, Marcos-Gragera R, Coeberg JW, EUROCARE Working Group. Hodgkin disease survival in Europe and the U.S. Cancer. 2006;107(2):352–60.

44. van Nimwegen FA, Schaapveld M, Cutter DJ, Jnaus CPM, Krol ADG, Hauptmann M, et al. Radiation Dose-Response Relationship for Risk of Coronary Heart Disease in Survivors of Hodgkin Lymphoma. J Clin Oncol. 2015;34(3):235–43.

45. Aleman BMP, Belt-dusebout AW Van Den, Bruin ML De, Veer MB Van, Baaijens MHA, Boer JP De, et al. Late cardiotoxicity after treatment for Hodgkin lymphoma. Blood. 2016;109(5):1878–87.

46. Monsez J-J, Charniot J-C, Vignat N, Artigou J-Y. Cardiac side-effects of cancer chemotherapy. Int J Cardiol. 2010;144(1):3–15.

47. Hancock SL, Cox RS, McDougall IR. Thyroid diseases after treatment of Hogkin's disease. N Engl J Med. 1991;325:599–605.

48. Van Der Kaaij MAE, Heutte N, Meijnders P, Abeilard-Lemoisson E, Spina M, Moser EC, et al. Premature ovarian failure and fertility in long-term survivors of Hodgkin's lymphoma: A European Organisation for Research and Treatment of Cancer Lymphoma Group and Groupe d'ÉTude des Lymphomes de l'Adulte Cohort Study. J Clin Oncol. 2012;30(3):291–9.

49. Furby A, Behin A, Lefaucheur J-P, Beauvais K, Marcorelles P, Mussini J-M, et al. Late-onset cervicoscapular muscle atrophy and weakness after radiotherapy for Hodgkin disease: a case series. J Neurol Neurosurg Psychiatry. 2010;81:101–4.

50. Malow B, Dawson D. Neuralgic amyotrophy in association with radiation therapy for Hodgkin's disease. Neurology. 1991;41(3):440–1.

51. Feldmann E, Posner JB. Episodic Neurologic Dysfunction in Patients With Hodgkin's Disease. J Am Med Assoc. 1986;43(12):1227–33.

52. Schachter S, Freeman R. Transient ischemic attack and adriamycin cardiomyopathy. Neurology. 1982;32(12):1380–1.

53. Segal G, Duckert L. Reversible mechlorethamine-associated hearing loss in a patient with Hodgkin's disease. Cancer. 1986;57(6):1098–91.

54. van Eggermond AM, Schaapveld M, Lugtenburg PJ, Krol ADG, de Boer JP, Zijlstra JM, et al. Risk of multiple primary malignancies following treatment of Hodgkin lymphoma. Blood. 2014;124:319–27.

55. Aleman BMP, van der Belt-Dusebout AW, Klokman WJ, van't Veer MB, Bartelink H, van Leeuwen FE. Long-Term Cause-Specific Mortality of Patients Treated for Hodgkin's Disease. J Clin Oncol. 2003;18:3431–9.

56. Schaapveld M, Aleman BMP, van Eggermond AM, Janus CPM, Krol ADG, van der Maazen RWM, et al. Second Cancer Risk Up to 40 Years after Treatment for Hodgkin's Lymphoma. N Engl J Med.

2015;373(26):2499–511.

57. Swerdlow AJ, Cooke R, Bates A, Cunningham D, Falk SJ, Gilson D, et al. Breast cancer risk after supradiaphragmatic radiotherapy for Hodgkin's lymphoma in England and Wales: A national cohort study. J Clin Oncol. 2012;30(22):2745–52.

58. Horwich A, Swerdlow AJ. Second primary breast cancer after Hodgkin's disease. Br J Cancer. 2004;90(2):294–8.

59. Travis LB, Hill DA, Dores GM, Gospodarowicz M, Van Leeuwen FE, Holowaty E, et al. Breast Cancer Following Radiotherapy and Chemotherapy Among Young Women With Hodgkin Disease. J Am Med Assoc. 2003;290(4).

60. van Leeuwen F, Klokman W, Stovall M, Dahler E, van't Veer M, Noordijk E, et al. Roles of radiation dose, chemotherapy, and hormonal factors in breast cancer following Hodgkin's disease. JNCI J Natl Cancer Inst. 2003;95(13):971–80.

61. Cooke R, Jones ME, Cunningham D, Falk SJ, Gilson D, Hancock BW, et al. Breast cancer risk following Hodgkin lymphoma radiotherapy in relation to menstrual and reproductive factors. Br J Cancer. 2013;108(11):2399–406.

62. Swerdlow AJ, Barber JA, Vaughan Hudson G, Cunningham D, Gupta RK, Hancock BW, et al. Risk of Second Malignancy After Hodgkin's Disease in a Collaborative British Cohort: The Relation to Age at Treatment. J Clin Oncol. 2000;18(3):498–509.

63. van Leeuwen F, Klokman W, Veer M, Hagenbeek. A, Krol A, Vetter U, et al. Long-term risk of second malignancy in survivors of Hodgkin's disease treated during adolescence or young adulthood. J Clin Oncol. 2000;18(3):487–97.

64. Dores G, Metayer C, Curtis R, Lynch C, Clarke E, Glimelius B, et al. Second malignant neoplasms among long-term survivors of Hodgkin's disease: a population-based evaluation over 25 years. J Clin Oncol. 2002;20(16):3484–94.

65. Swerdlow AJ, Cooke R, Bates A, Cunningham D, Falk SJ, Gilson D, et al. Breast Cancer Risk After Supradiaphragmatic Radiotherapy for Hodgkin's Lymphoma in England and Wales: A National Cohort Study. J Clin Oncol. 2012;30(22):2745–52.

66. Chen J, Lee RJ, Tsodikov A, Smith L, Gaffney DK. Does radiotherapy around the time of pregnancy for Hodgkin's disease modify the risk of breast cancer? Radiat Oncol. 2004;58(5):1474–9.

67. Ma YP, van Leeuwen FE, Cooke R, Broeks A, Enciso-Mora V, Olver B, et al. FGFR2 genotype and risk of radiation-associated breast cancer in Hodgkin lymphoma. Blood. 2012;119(4):1029–31.

68. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci. 2008;105(11):4340–5.

69. Matsubara A, Teishima J, Mirkhat S, Yasumoto H, Mochizuki H, Seki M, et al. Restoration of FGF Receptor Type 2 Enhances

Radiosensitivity of Hormone-refractory Human Prostate Carcinoma PC-3 Cells. Anticancer Res. 2008;28(4B):2141–6.

70. Best T, Li D, Skol AD, Kirchhoff T, Jackson SA, Bhatia S, et al. Variants at 6q21 implicate PRDM1 in the etiology of therapy-induced second malignancies after Hodgkin lymphoma. Nat Med. 2011;17(8):941–3.

71. Morton LM, Sampson JN, Armstrong GT, Chen T-H, Hudson MM, Karlins E, et al. Genome-Wide Association Study to Identify Susceptibility Loci That Modify Radiation-Related Risk for Breast Cancer After Childhood Cancer. JNCI J Natl Cancer Inst. 2017;109(11):1–11.

72. Michailidou K, Hall P, Gonzalez-neira A, Ghoussaini M, Milne RL, Schmidt MK, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet. 2013;45(4):353–61.

73. Michailidou K. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92–4.

74. Michailidou K. Identification of ten variants associated with risk of estrogen- receptor-negative breast cancer. Nature. 2017;49(12):1767–78.

75. Kwon J, Han H, Lee D, Do J, Kim J, Kim M. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. Nat Commun. 2016;7(11375).

76. Cox A. A common coding variant in CASP8 is associated with breast

cancer risk. Nat Genet. 2007;39:352–8.

77. Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Ballinger DG, Struewing JP, et al. Group Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2009;447(7148):1087–93.

78. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, et al. GWAS study identifies alleles in FGFR2 associated with risk of breast cancer. 2012;39(7):870–4.

79. Assimes TL, Roberts R. Genetics: Implications for Prevention and Management of Coronary Artery Disease. J Am Coll Cardiol. 2016;68(25):2797–818.

80. Goodarzi AA, Noon AT, Deckbar D, Ziv Y, Shiloh Y, Löbrich M, et al. ATM Signaling Facilitates Repair of DNA Double-Strand Breaks Associated with Heterochromatin. Mol Cell. 2008;31(2):167–77.

81. Moynahan ME, Chiu JW, Koller BH, Jasin M. Brca1 Controls Homology-Directed DNA Repair. Mol Cell. 1999;4(4):511–8.

82. Oliver A, Swift S, Lord C, Ashworth A, Pearl L. Structural basis for recruitment of BRCA2 by PALB2. EMBO Rep. 2009;10(9):990–6.

83. Dolganov GM, Maser RS, Novikov A, Tosto L, Chong S, Bressan DA, et al. Human Rad50 is physically associated with human Mre11: identification of a conserved multiprotein complex implicated in recombinational DNA repair. Mol Cell Biol. 1996;16(9):4832–41.

84. Mavaddat N, Pharoah PDP, Michailidou K, Tyrer J, Brook MN, Bolla

MK, et al. Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. JNCI J Natl Cancer Inst. 2015;107(5):1–15.

85. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet. 2019;104(1):21–34.

86. Apicella C, Milne RL, Tsimiklis H, Phillips K. Breast cancer risk prediction using clinical models and 77 independent risk-associated SNPs for women aged under 50 years: Australian Breast Cancer Family Registry. 2017;25(2):359–65.

87. Laurie C, Doheny K, Mirel D. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010;34(6):591–602.

88. Turner S, Armstrong LL, Bradford Y, Carlson CS, Dana C, Crenshaw AT, et al. Quality Control Procedures for Genome Wide Association Studies. Curr Proc Hum Genet. 2011;68(1):1–24.

89. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007;81(3):559–75.

90. R Development Core Team. R: A Language and Environment for Statistical Computing [Internet]. Team RDC, editor. R Foundation for Statistical Computing. R Foundation for Statistical Computing; 2015. (R Foundation for Statistical Computing). Available from: http://www.r-

project.org

91. Wigginton JE, Cutler DJ, Abecasis GR. A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am J Hum Genet. 2005;76(5):887–93.

92. Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;8(38):904-009.

93. The International HapMap. A haplotype map of the human genome. Nature. 2005;437(7063):1299–320.

94. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. Am J Hum Genet. 2013;93(4):687–96.

95. Howie B, Marchini J, Stephens M. Genotype Imputation with Thousands of Genomes. G3 Genes, Genomes, Geneicts. 2011;1(6):457–70.

96. Zheng HF, Rong JJ, Liu M, Han F, Zhang XW, Richards JB, et al. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PLoS One. 2015;10(1):1–10.

97. Ng AK, Bernardo MVP, Weller E, Backstrand K, Silver B, Karen C, et al. Second malignancy after Hodgkin disease treated with radiation therapy with or without chemotherapy: long-term risks and risk factors. Blood. 2002;100(6):1989–96.

98. Aulchenko YS, Struchalin M V, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. BMC

Bioinformatics. 2010;11(134).

99. 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).

100. Swerdlow AJ, Jones ME, Schoemaker MJ, Hemming J, Thomas D, Williamson J, et al. The Breakthrough Generations Study: Design of a long-term UK cohort study to investigate breast cancer aetiology. Br J Cancer. 2011;105(7):911–7.

101. Genotype-Tissue (GTex) [Internet]. Available from: https://gtexportal.org/home/

102. Sweeney WE, Avner ED. Polycystic Kidney Disease, Autosomal Recessive. GeneReviews. 2001.

103. Sjöblom T, Jones S, Wood L, Parsons D, Lin J, Barber T, et al. The Consensus Coding Sequences of Human Breast and Colorectal Cancers. Science (80-). 2006;314(5797):268–74.

104. Willer CJ, Li Y, Abecasis GR, Overall P. METAL : fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1.

105. The Cancer Genome Atlas Program [Internet]. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

106. Morton LM, Swerdlow AJ, Schaapveld M, Ramadan S, Hodgson DC,

Radford J, et al. Current knowledge and future research directions in treatment-related second primary malignancies. Eur J Cancer Suppl. 2014;12(1):5–17.

107. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet. 2010;39(2):165–7.

108. Miki Y, Swensen J, Shattuck-Eidens D, Futreal P, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science (80-). 1994;266(5182):66–71.

109. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion, Jonathan Collins N, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995;378:789–92.

110. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet. 2006;38:873–5.

111. Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, et al. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet. 2002;1:55–9.

112. Valko M, Izakovic M, Mazur M, Rhodes CJ, Telser J. Role of oxygen radicals in DNA damage and cancer incidence. Mol Cell Biochem. 2004;266(1):37–56.

113. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

114. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype data. Am J Hum Genet. 2012;91:839–48.

115. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human. Nat Genet. 2014;46(3):310–5.

116. Picard [Internet]. 2018. Available from: https://broadinstitute.github.io/picard/

117. Jacobs K. Genotype Library and Utilities (GLU) software package [Internet]. BioInformed LLC and the U.S. Department of Health & Human Services; 2009. Available from: http://code.google.com/p/glu-genetics

118. Leal S, Wang G. Association Analysis of Sequence Data using Variant Association Tools (VAT). 2014.

119. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):1–7.

120. Liu X, Jian X, Boerwinkle E. dbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions. Hum Mutat. 2011;32(8):894–9.

121. Consortium T 1000 GP. A global reference for human genetic variation. Nature. 2015;526:68–74.

122. Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.

123. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010;34(2):188–93.

124. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014;95(1):5–23.

125. Madsen BE, Browning SR. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLoS Genet. 2009;5(2).

126. Paliwal S, Kanagaraj R, Sturzenegger A, Burdova K, Janscak P. Human RECQ5 helicase promotes repair of DNA double-strand breaks by synthesis-dependent strand annealing. Nucleic Acids Res. 2014;42(4):2380–90.

127. Newman JA, Aitkenhead H, Savitsky P, Gileadi O. Insights into the RecQ helicase mechanism revealed by the structure of the helicase domain of human RECQL5. Nucleic Acids Res. 2017;45(7):4231–43.

128. Olson HC, Davis L, Kiianitsa K, Khoo KJ, Liu Y, Knijnenburg TA, et al. Increased levels of RECQ5 shift DNA repair from canonical to alternative pathways. Nucleic Acids Res. 2018;46(18):9496–509.

129. Berton S, Cusan M, Segatto I, Citron F, D'Andrea S, Benevol S, et al. Loss of p27kip1 increases genomic instability and induces radio-resistance in luminal breast cancer cells. Sci Rep. 2017;7(1):1–14.

130. Sun G, Gargus JJ, Ta DT, Vickery LE. Identification of a novel candidate gene in the iron-sulfur pathway implicated in ataxia-susceptibility: Human gene encoding HscB, a J-type co-chaperone. J Hum Genet. 2003;48(8):415–9.

131. Van der Spek PJ, Eker A, Rademakers S, Visser C, Sugasawa K, Masutani C, et al. XPC and human homologs of RAD23: Intracellular localization and relationship to other nucleotide excision repair complexes. Nucleic Acids Res. 1996;24(13):2551–9.

132. Linge A, Maurya P, Friedrich K, Baretton G, Kelly S, Henry M, et al. Identification and Functional Validation of RAD23B as a Potential Protein in Human Breast Cancer Progression. J Proteome Res. 2014;13(7):3212–22.

133. Wang B, Elledge SJ. Ubc13/Rnf8 ubiquitin ligases control foci formation of the Rap80/Abraxas/Brca1/Brcc36 complex in response to DNA damage. Proc Natl Acad Sci. 2007;104(52):20759–63.

134. Barber LJ, Wang X, Murakawa Y, Yamamoto K, Boulton SJ, Yamada K, et al. A Critical Role for the Ubiquitin-Conjugating Enzyme Ubc13 in Initiating Homologous Recombination. Mol Cell. 2007;25(5):663–75.

135. Oldreive C, Nakada S, Durocher D, Taylor AMR, Kolas NK, Panier S, et al. The RIDDLE Syndrome Protein Mediates a Ubiquitin-Dependent

Signaling Cascade at Sites of DNA Damage. Cell. 2009;136(3):420–34.

136. Narayanan DL, Phadke SR. A novel variant in MED12 gene: Further delineation of phenotype. Am J Med Genet Part A. 2017;173(8):2257–60.

137. Piscuoglio S, Murray M, Fusco N, Marchiò C, Loo FL, Martelotto LG, et al. MED12 somatic mutations in fibroadenomas and phyllodes tumors of the breast. Histopathology. 2015;67(5):719–29.

138. Barbieri CE, Baca SC, Lawrence, Michael S Demichelis, Francesca Blattner M, Theurillat, Jean-Philippe White TA, Stojanov P, Van Allen E, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet. 2012;44:685–9.

139. Lu R, Nash HM, Verdine GL. A mammalian DNA repair enzyme that excises oxidatively damaged guanines maps to a locus frequently lost in lung cancer. Curr Biol. 1997;7(6):397–407.

140. Kubo N, Morita M, Nakashima Y, Kitao H, Egashira A, Saekii H, et al. Oxidative Dna damage in human esophageal cancer: clinicopathological analysis of 8-hydroxydeoxyguanosine and its repair enzyme. Dis esophagus. 2014;27(13):285–93.

141. Paz-Elizur T, Kruspy M, Blumenstein S, Elinger D, Schechtman E, Livneh Z. DNA Repair Activity for Oxidative Damage and Risk of Lung Cancer. J Natl Cancer Inst. 2003;95(17):1312–9.

142. Osorio A, Milne RL, Kuchenbaecker K, Vaclová T, Pita G, Alonso R,

et al. DNA Glycosylases Involved in Base Excision Repair May Be Associated with Cancer Risk in BRCA1 and BRCA2 Mutation Carriers. PLoS Genet. 2014;10(4).

143. Pendleton M, Lindsey Jr RH, Felix CA, Grimwade D, Osheroff N. Topoisomerase II and leukemia. Ann N Y Acad Sci. 2014;1310(1):98–110.

144. Dykhuizen EC, Hargreaves DC, Miller E, Cui K, Korshunov A, Kool M, et al. mSWI/SNF (BAF) Complexes Facilitate Decatentation of DNA by Topoisomerase IIα. Nature. 2013;487(7451):624–7.

145. Roca J. The mechanisms of DNA topoisomerases. Trends Biochem Sci. 1995;156–60.

146. Tamaichi H, Sato M, Porter ACG, Shimizu T, Mizutani S, Takagi M. Ataxia telangiectasia mutated-dependent regulation of topoisomerase II alpha expression and sensitivity to topoisomerase II inhibitor. Cancer Sci. 2013;104(2):178–84.

147. Engstrøm MJ, Ytterhus B, Vatten LJ, Opdahl S, Bofin AM. TOP2A gene copy number change in breast cancer. J Clin Pathol. 2014;67(5):420–5.

148. Sartori AA, Lukas C, Coates J, Mistrik M, Fu S, Bartek J, et al. Human CtIP promotes DNA end resection. Nature. 2007;450:509–14.

149. Hoa NN, Kobayashi J, Omura M, Hirakawa M, Yang SH, Komatsu K, et al. BRCA1 and CtIP are both required to recruit Dna2 at double-strand breaks in homologous recombination. PLoS One.

2015;10(4):1–17.

150. Quennet V, Beucher A, Barton O, Takeda S, Löbrich M. CtIP and MRN promote non-homologous end-joining of etoposide-induced DNA double-strand breaks in G1. Nucleic Acids Res. 2011;39(6):2144–52.

151. Makharashvili N, Tubbs AT, Yang S-H, Wang H, Barton O, Zhou Y, et al. Catalytic and non-catalytic roles of the CtIP endonuclease in double-strand break end resection. Mol Cell. 2014;54(6):1022–33.

152. Bennardo N, Cheng A, Huang N, Stark JM. Alternative-NHEJ is a mechanistically distinct pathway of mammalian chromosome break repair. PLoS Genet. 2008;4(6).

153. Wang, Junhui Ding, Qianshan Fujimori, Hiroaki Motegi, Akira Miki Y, Mitsuko M. Loss of CtIP disturbs homologous recombination repair and sensitizes breast cancer cells to PARP inhibitors. Oncotarget. 2015;7(7):7701–14.

154. Wilkes DC, Sailer V, Xue H, Cheng H, Collins CC, Gleave M, et al. A germline FANCA alteration that is associated with increased sensitivity to DNA damaging agents. Cold Spring Harb Mol Case Stud. 2017;3(5):1–12.

155. Folias A, Matkovic M, Bruun D, Reid S, Hejna J, Grompe M, et al. BRCA1 interacts directly with the Fanconi anemia protein FANCA. Hum Mol Genet. 2002;11(21):2591–7.

156. Ralleigh G, Given-Wilson R. Breast cancer risk and possible

screening strategies for young women following supradiaphragmatic irradiation for Hodgkin's disease. Clin Radiol. 2004;59(8):647–50.

157. National Comprehensive Cancer Network [Internet]. Available from: https://www.nccn.org

158. Chilhood Oncology Group [Internet]. Available from: http://www.survivorshipguidelines.org/

159. Ng AK. Current survivorship recommendations for patients with Hodgkin lymphoma: focus on late effects. Hematology. 2014;124(23):3373–80.

160. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics. 2015;31(9):1466–8.

161. Castiglioni F, Terenziani M, Carcangiu ML, Miliano R, Aiello P, Bertola L, et al. Radiation Effects on Development of HER2-Positive Breast Carcinomas. Clin Cancer Res. 2007;13(1):46–51.

162. Horst KC, Hancock SL, Ognibene G, Chen C, Advani RH, Rosenberg S a, et al. Histologic subtypes of breast cancer following radiotherapy for Hodgkin lymphoma. Ann Oncol. 2014;25(4):848–51.

163. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014;24:52–60.

164. Broeks A, Braaf LM, Wessels LFA, van de Vijver M, De Bruin ML, Stovall M, et al. Radiation-Associated Breast Tumors Display a Distinct Gene Expression Profile. Int J Radiat Oncol Biol Phys.

2010;76(2):540–7.

165. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SSJ, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.

166. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol. 2011;5(1):5–23.

Supplementary materials

Supplementary Table 1. Significant associations (P-value < 5 x 10^{-6}) obtained from the Cox regression association analyses of 766 samples (261 cases and 505 controls) on 6,788,016 SNPs. Results are ordered according to P-value.

Chr	Position	rsid	Minor allele	Major allele	Risk allele	MAF	MAF Cases	MAF Controls	Hazard ratio	95% CI	P-value
5	104770755	rs200924286	GTGTGTA	G	GTGTGTA	0.28	0.34	0.24	1.60	1.33-1.94	1.75 x 10^{-6}
5	104697206	rs2431958	T	G	T	0.25	0.32	0.22	1.59	1.32-1.91	2.02 x 10^{-6}
5	104784402	rs10073707	C	T	C	0.25	0.31	0.21	1.59	1.32-1.91	2.06 x 10^{-6}
5	104773783	rs6881607	C	G	C	0.29	0.36	0.26	1.57	1.31-1.88	2.11 x 10^{-6}
12	114944187	rs838308	T	C	C	0.20	0.14	0.23	0.57	0.44-0.73	2.12 x 10^{-6}
5	104670730	rs71622899	G	GTATGTA	G	0.25	0.32	0.22	1.59	1.32-1.91	2.19 x 10^{-6}
5	104671468	rs658042	T	G	T	0.25	0.32	0.22	1.59	1.32-1.91	2.33 x 10^{-6}
5	104775100	rs6887756	C	T	C	0.29	0.36	0.26	1.56	1.30-1.88	2.37 x 10^{-6}
5	104788257	rs13178200	C	T	C	0.29	0.36	0.26	1.56	1.30-1.87	2.45 x 10^{-6}
5	104791577	rs4580751	T	A	T	0.29	0.36	0.26	1.56	1.30-1.87	2.49 x 10^{-6}
5	104758041	rs10479326	G	T	G	0.30	0.36	0.26	1.57	1.30-1.88	2.57 x 10^{-6}
5	104774091	rs4295364	G	C	G	0.25	0.32	0.22	1.58	1.31-1.90	2.61 x 10^{-6}
10	23896241	rs35697184	TA	T	T	0.20	0.14	0.23	0.56	0.44-0.73	2.62 x 10^{-6}
5	104775164	rs6883579	G	A	G	0.25	0.32	0.22	1.58	1.31-1.90	2.68 x 10^{-6}
5	104760475	rs4476697	C	T	C	0.25	0.32	0.22	1.59	1.32-1.92	2.68 x 10^{-6}
5	104771109	rs10037034	C	A	C	0.29	0.36	0.26	1.56	1.30-1.30	2.71 x 10^{-6}

5	104788643	rs10038910	C	T	C	0.25	0.32	0.22	1.58	1.31-1.90	2.71×10^{-6}
5	104761350	rs7723599	T	C	T	0.25	0.32	0.22	1.59	1.32-1.92	2.72×10^{-6}
5	104761581	rs10479327	G	T	G	0.25	0.32	0.22	1.59	1.32-1.92	2.72×10^{-6}
5	104761949	rs9327941	T	C	T	0.25	0.32	0.22	1.59	1.32-1.92	2.73×10^{-6}
5	104764881	rs11242551	C	T	C	0.25	0.32	0.22	1.59	1.32-1.92	2.81×10^{-6}
5	104765659	rs10040650	G	C	G	0.29	0.36	0.26	1.56	1.30-1.87	2.94×10^{-6}
5	104768926	rs34620659	C	T	C	0.25	0.31	0.21	1.58	1.31-1.91	2.99×10^{-6}
5	104765585	rs10065459	T	C	T	0.29	0.36	0.26	1.56	1.30-1.87	3.03×10^{-6}
5	104765644	rs10044124	C	T	C	0.29	0.36	0.26	1.56	1.30-1.87	3.05×10^{-6}
5	104795394	rs4131494	A	C	A	0.29	0.36	0.26	1.56	1.30-1.87	3.06×10^{-6}
5	104810307	rs6862074	G	A	G	0.29	0.36	0.26	1.56	1.30-1.88	3.07×10^{-6}
5	104794603	rs4479819	G	A	G	0.29	0.36	0.26	1.56	1.30-1.87	3.17×10^{-6}
5	104780726	rs10063846	C	A	C	0.25	0.32	0.22	1.58	1.31-1.90	3.18×10^{-6}
9	6867111	rs9792577	A	G	A	0.48	0.40	0.52	0.65	0.55-0.78	3.31×10^{-6}
5	104742680	rs7715902	T	G	T	0.25	0.31	0.22	1.57	1.31-1.90	3.33×10^{-6}
5	104767470	rs12522619	G	A	G	0.29	0.36	0.26	1.56	1.30-1.87	3.35×10^{-6}
5	104682941	rs249558	T	C	T	0.25	0.32	0.22	1.57	1.31-1.90	3.38×10^{-6}
20	10090970	rs671426	C	T	T	0.30	0.22	0.34	0.63	0.51-0.77	3.40×10^{-6}
5	104763766	rs9327942	C	T	C	0.30	0.36	0.26	1.56	1.30-1.87	3.45×10^{-6}
20	10095741	rs489548	G	A	A	0.30	0.22	0.34	0.63	0.51-0.77	3.46×10^{-6}
5	104810352	rs6888357	T	C	T	0.25	0.31	0.21	1.58	1.31-1.91	3.47×10^{-6}
5	104749687	rs7707957	G	A	G	0.25	0.32	0.22	1.57	1.31-1.90	3.58×10^{-6}
20	10096596	rs518701	T	C	C	0.30	0.22	0.34	0.63	0.51-0.77	3.61×10^{-6}
6	51790236	rs891238519	CGTGT	C	C	0.20	0.14	0.23	0.57	0.44-0.73	3.69×10^{-6}
5	104732641	rs4364361	T	C	T	0.30	0.36	0.26	1.55	1.29-1.86	3.70×10^{-6}
5	104717430	rs62369169	C	T	C	0.30	0.36	0.26	1.55	1.29-1.85	3.78×10^{-6}
5	104808969	rs4604172	G	T	G	0.25	0.31	0.21	1.58	1.31-1.90	3.85×10^{-6}

20	10085211	rs685573	T	A	A	0.30	0.22	0.33	0.63	0.51-0.77	3.85×10^{-6}
20	10107052	rs2207849	T	C	C	0.31	0.23	0.34	0.63	0.52-0.77	3.90×10^{-6}
5	104817801	rs4703344	C	T	C	0.24	0.31	0.21	1.58	1.31-1.90	4.07×10^{-6}
5	104816351	rs67029016	G	C	G	0.29	0.36	0.26	1.56	1.29-1.87	4.08×10^{-6}
5	104732996	rs4626314	T	C	T	0.30	0.36	0.26	1.55	1.29-1.85	4.16×10^{-6}
5	104806455	rs13155574	G	A	G	0.26	0.33	0.23	1.57	1.30-1.89	4.17×10^{-6}
20	10089441	rs637625	G	A	A	0.30	0.22	0.33	0.63	0.51-0.77	4.19×10^{-6}
5	104736048	rs4493645	C	A	C	0.30	0.36	0.26	1.55	1.29-1.85	4.20×10^{-6}
5	104715761	rs56196078	A	G	A	0.30	0.36	0.26	1.54	1.29-1.85	4.21×10^{-6}
20	10097626	rs665397	A	C	C	0.30	0.22	0.34	0.63	0.52-0.78	4.29×10^{-6}
5	104708701	rs10590294	CATATA	C	CATATA	0.25	0.31	0.22	1.56	1.30-1.88	4.35×10^{-6}
5	104706453	rs4703337	A	T	A	0.25	0.31	0.22	1.56	1.30-1.88	4.45×10^{-6}
5	104708922	rs7715654	A	G	A	0.25	0.31	0.22	1.56	1.30-1.88	4.47×10^{-6}
20	10094774	rs570383	T	C	C	0.30	0.22	0.33	0.63	0.52-0.78	4.48×10^{-6}
20	10094251	rs543895	A	T	T	0.30	0.22	0.33	0.63	0.52-0.78	4.48×10^{-6}
5	104708946	rs4262077	T	G	T	0.25	0.31	0.22	1.56	1.30-1.88	4.49×10^{-6}
5	104708659	rs7715185	A	T	A	0.25	0.31	0.22	1.56	1.30-1.88	4.50×10^{-6}
20	10093923	rs2423481	A	T	T	0.30	0.22	0.33	0.63	0.52-0.78	4.50×10^{-6}
5	104709915	rs7725129	T	A	T	0.25	0.31	0.22	1.56	1.30-1.88	4.50×10^{-6}
5	104817968	rs56328905	TG	T	TG	0.29	0.36	0.26	1.55	1.29-1.86	4.51×10^{-6}
20	10092415	rs590717	G	A	A	0.30	0.22	0.33	0.63	0.52-0.78	4.52×10^{-6}
20	10091214	rs563455	T	C	C	0.30	0.22	0.33	0.63	0.52-0.78	4.55×10^{-6}
5	104716910	rs1423206	G	C	G	0.25	0.31	0.22	1.56	1.30-1.88	4.58×10^{-6}
5	104716918	rs1423207	A	G	A	0.25	0.31	0.22	1.56	1.30-1.88	4.58×10^{-6}
5	104716309	rs4703338	T	C	T	0.25	0.31	0.22	1.56	1.30-1.88	4.58×10^{-6}
20	10087754	rs544718	G	T	T	0.30	0.22	0.33	0.63	0.52-0.78	4.61×10^{-6}
20	10099250	rs585439	A	G	G	0.30	0.23	0.34	0.63	0.52-0.78	4.64×10^{-6}

20	10096293	rs515940	T	C	C	0.30	0.22	0.33	0.63	0.52-0.78	4.79×10^{-6}
5	104800148	rs5870159	GA	G	GA	0.25	0.32	0.22	1.57	1.30-1.89	4.93×10^{-6}

Supplementary Table 2. Associations of the 172 known FBC loci obtained from the Cox regression association analyses of 766 samples (261 cases and 505 controls) on 6,788,016 SNPs. Results are ordered according to P-value.

Chr	Position	rsid	Minor allele	Major allele	Risk allele	MAF	MAF cases	MAF controls	Hazard ratio	95% CI	P-value
10	123340431	rs35054928	GC	G	GC	0.40	0.44	0.38	1.30	1.09-1.56	0.0037
1	155148781	rs4971059	A	G	A	0.35	0.40	0.33	1.28	1.07-1.53	0.0075
9	110837073	rs10816625	G	A	G	0.06	0.08	0.05	1.52	1.13-2.04	0.0096
10	123340311	rs2981578	C	T	C	0.46	0.49	0.44	1.26	1.06-1.51	0.011
5	132407058	rs6596100	T	C	C	0.23	0.19	0.25	0.75	0.60-0.94	0.011
6	151942194	rs3757322	G	T	G	0.33	0.37	0.30	1.26	1.05-1.51	0.012
1	203766331	rs4951011	G	A	G	0.18	0.20	0.16	1.30	1.05-1.61	0.020
7	91630620	rs6964587	T	G	G	0.38	0.34	0.39	0.82	0.68-0.98	0.027
2	218296508	rs16857609	T	C	C	0.24	0.22	0.26	0.80	0.65-0.98	0.026
22	38568833	rs738321	G	C	C	0.39	0.36	0.40	0.84	0.70-1.00	0.044
9	129396434	rs10760444	G	A	A	0.44	0.40	0.46	0.84	0.70-1.00	0.048
22	46283297	rs28512361	A	G	A	0.11	0.12	0.10	1.35	1.01-1.80	0.050
9	110895353	rs676256	C	T	T	0.39	0.37	0.41	0.84	0.71-1.00	0.053
14	68660428	rs2588809	T	C	T	0.17	0.19	0.16	1.25	1.00-1.56	0.056
18	29977689	rs117618124	C	T	T	0.04	0.03	0.05	0.64	0.39-1.06	0.063
20	32588095	rs2284378	T	C	C	0.33	0.30	0.35	0.84	0.69-1.01	0.065
7	101552440	rs71559437	A	G	G	0.14	0.13	0.15	0.80	0.61-1.04	0.088
8	124610166	rs58847541	A	G	A	0.15	0.18	0.14	1.23	0.97-1.54	0.090
22	40876234	rs6001930	C	T	C	0.10	0.12	0.09	1.26	0.97-1.65	0.094
2	217963060	rs34005590	A	C	C	0.05	0.04	0.05	0.68	0.43-1.10	0.095

19	18571141	rs4808801	G	A	A	0.37	0.34	0.38	0.86	0.72-1.04	0.11
7	144074929	rs720475	A	G	A	0.25	0.27	0.24	1.17	0.96-1.42	0.12
6	130349119	rs6569648	C	T	C	0.27	0.29	0.26	1.16	0.96-1.40	0.13
10	80841148	rs704010	T	C	T	0.40	0.42	0.39	1.14	0.96-1.36	0.13
19	13954571	rs2594714	A	G	A	0.24	0.26	0.23	1.17	0.95-1.44	0.14
5	49641645	rs72749841	C	T	C	0.17	0.19	0.16	1.21	0.94-1.54	0.14
20	48945911	rs6122906	G	A	G	0.21	0.23	0.20	1.17	0.95-1.44	0.14
2	227226952	rs12479355	G	A	A	0.22	0.20	0.23	0.85	0.69-1.06	0.15
6	28926220	rs9257408	C	G	G	0.38	0.36	0.39	0.88	0.73-1.05	0.15
19	17401404	rs67397200	G	C	G	0.30	0.33	0.28	1.15	0.95-1.38	0.16
22	29121087	rs17879961	G	A	G	0.00	0.00	0.00	5.31	0.82-34.28	0.16
1	121280613	rs11249433	G	A	G	0.45	0.46	0.44	1.13	0.95-1.33	0.16
11	803017	rs6597981	A	G	A	0.49	0.50	0.49	1.13	0.95-1.34	0.17
8	117209548	rs970715757	A	G	A	0.35	0.36	0.34	1.13	0.95-1.35	0.17
2	121245122	rs4849887	T	C	C	0.10	0.09	0.11	0.81	0.60-1.11	0.17
3	141112859	rs34207738	C	CTT	C	0.44	0.46	0.44	1.12	0.94-1.34	0.19
3	71532113	rs6805189	C	T	T	0.48	0.47	0.49	0.89	0.75-1.06	0.20
7	130667121	rs4593472	T	C	T	0.35	0.37	0.34	1.13	0.94-1.36	0.20
2	10135681	rs113577745	G	C	C	0.10	0.09	0.11	0.83	0.61-1.12	0.22
7	28356889	rs17156577	C	T	T	0.12	0.10	0.12	0.84	0.63-1.12	0.23
19	13158277	rs78269692	C	T	C	0.06	0.07	0.06	1.23	0.88-1.70	0.23
1	114448389	rs11552449	T	C	C	0.20	0.18	0.21	0.88	0.71-1.09	0.24
22	39359355	rs868638441	<CN0>	C	C	0.09	0.08	0.10	0.81	0.56-1.17	0.25
16	52599188	rs4784227	T	C	T	0.23	0.25	0.22	1.12	0.92-1.36	0.26
3	63967900	rs1053338	G	A	A	0.14	0.13	0.15	0.86	0.66-1.12	0.26
5	32567732	rs2012709	T	C	T	0.45	0.48	0.43	1.10	0.93-1.30	0.27

9	119313486	rs1895062	G	A	G	0.43	0.46	0.41	1.10	0.92-1.31	0.28
7	139942304	rs11977670	A	G	A	0.45	0.47	0.44	1.10	0.92-1.31	0.28
1	202187176	rs6678914	A	G	A	0.41	0.43	0.41	1.10	0.92-1.30	0.29
5	345109	rs116095464	C	T	C	0.06	0.07	0.06	1.21	0.85-1.72	0.30
8	76230301	rs6472903	G	T	T	0.17	0.15	0.17	0.88	0.69-1.12	0.30
10	95292187	rs140936696	CAA	C	CAA	0.17	0.18	0.17	1.14	0.90-1.44	0.30
1	46600917	rs1707302	A	G	G	0.35	0.33	0.36	0.91	0.76-1.09	0.31
5	1296255	rs3215401	AG	A	A	0.32	0.30	0.33	0.91	0.75-1.10	0.31
8	36858483	rs13365225	G	A	G	0.17	0.18	0.16	1.12	0.90-1.40	0.31
1	242034263	rs72755296	G	A	G	0.04	0.04	0.04	1.27	0.81-2.01	0.31
10	5886734	rs2380205	T	C	C	0.45	0.43	0.45	0.92	0.77-1.09	0.32
3	172285237	rs58058861	A	G	G	0.22	0.21	0.23	0.90	0.72-1.12	0.32
2	19320803	rs12710696	T	C	C	0.36	0.34	0.37	0.92	0.76-1.10	0.35
6	82128386	rs17529111	C	T	C	0.21	0.22	0.20	1.11	0.89-1.38	0.35
4	175846426	rs6828523	A	C	C	0.12	0.12	0.13	0.88	0.68-1.15	0.36
10	22315843	rs11814448	C	A	C	0.02	0.02	0.01	1.39	0.71-2.72	0.36
6	151952332	rs9397437	A	G	A	0.07	0.07	0.07	1.16	0.84-1.61	0.37
3	46866866	rs6796502	A	G	A	0.10	0.11	0.09	1.14	0.86-1.50	0.37
12	28174817	rs7297051	T	C	C	0.22	0.21	0.22	0.91	0.74-1.12	0.37
6	152437016	rs2747652	T	C	C	0.47	0.46	0.48	0.93	0.78-1.10	0.37
4	38816338	rs6815814	C	A	C	0.21	0.21	0.21	1.10	0.89-1.36	0.39
1	18807339	rs2992756	T	C	T	0.49	0.48	0.50	0.93	0.78-1.10	0.39
1	149927034	rs12048493	C	A	C	0.38	0.39	0.38	1.09	0.90-1.32	0.40
5	90732225	rs10474352	T	C	C	0.17	0.15	0.17	0.91	0.71-1.15	0.41
13	32972626	rs11571833	T	A	A	0.01	0.01	0.01	0.63	0.19-2.07	0.41
10	114773927	rs7904519	G	A	G	0.47	0.48	0.46	1.07	0.90-1.28	0.42

8	102478959	rs514192	A	T	T	0.30	0.29	0.30	0.93	0.77-1.12	0.43
8	29509616	rs9693444	A	C	C	0.33	0.32	0.33	0.93	0.78-1.11	0.43
11	69379161	rs75915166	A	C	C	0.06	0.05	0.06	0.85	0.57-1.27	0.43
2	29120733	rs4577244	T	C	C	0.24	0.23	0.24	0.92	0.75-1.13	0.43
16	87085237	rs4496150	A	C	C	0.23	0.21	0.24	0.92	0.75-1.14	0.44
6	149608874	rs9485372	A	G	G	0.18	0.17	0.19	0.92	0.73-1.15	0.45
1	204518842	rs4245739	C	A	A	0.25	0.24	0.26	0.93	0.76-1.13	0.46
8	129194641	rs11780156	T	C	T	0.19	0.20	0.18	1.09	0.87-1.36	0.47
13	73957681	rs6562760	A	G	G	0.24	0.23	0.24	0.93	0.76-1.14	0.48
19	19545696	rs2965183	A	G	A	0.33	0.35	0.33	1.07	0.89-1.27	0.48
5	56053723	rs62355902	T	A	T	0.18	0.19	0.18	1.08	0.87-1.36	0.48
17	29230520	rs146699004	G	GGT	G	0.25	0.26	0.25	1.07	0.88-1.30	0.49
5	16187528	rs13162653	T	G	T	0.46	0.48	0.45	1.06	0.89-1.26	0.50
12	96027759	rs17356907	G	A	A	0.30	0.30	0.31	0.94	0.78-1.13	0.51
16	54682064	rs28539243	A	G	A	0.50	0.49	0.51	0.94	0.79-1.12	0.52
1	50846032	rs140850326	C	CAA AGG GCA AGA TCT CCT TTTT	CAA AGG GCA AGA TCT CCT TTTT	0.46	0.45	0.46	0.95	0.80-1.12	0.52
19	46183031	rs71338792	AT	A	A	0.21	0.20	0.21	0.93	0.75-1.16	0.52
14	69034682	rs999737	T	C	T	0.26	0.26	0.25	1.07	0.88-1.30	0.53
9	22062134	rs1011970	T	G	T	0.18	0.18	0.18	1.07	0.86-1.33	0.54
8	76417937	rs2943559	G	A	G	0.08	0.08	0.08	1.11	0.80-1.53	0.54
3	30682939	rs12493607	C	G	C	0.32	0.33	0.31	1.06	0.88-1.27	0.55
3	99723580	rs9833888	T	G	G	0.24	0.23	0.24	0.94	0.77-1.15	0.55

1	88156923	rs17426269	A	G	A	0.16	0.17	0.16	1.07	0.85-1.36	0.56
20	5948227	rs16991615	A	G	A	0.06	0.07	0.06	1.11	0.78-1.58	0.56
2	202181247	rs1830298	C	T	C	0.26	0.27	0.26	1.06	0.88-1.27	0.57
11	129461171	rs11820646	T	C	T	0.41	0.42	0.40	1.05	0.88-1.25	0.57
17	44252468	rs2532263	A	G	G	0.22	0.21	0.22	0.94	0.75-1.18	0.57
16	53813367	rs17817449	G	T	G	0.40	0.40	0.40	1.05	0.88-1.25	0.59
12	120832146	rs206966	T	C	C	0.16	0.15	0.17	0.94	0.73-1.20	0.60
2	172972971	rs2016394	A	G	G	0.47	0.47	0.47	0.96	0.80-1.14	0.63
6	10456706	rs9348512	A	C	C	0.36	0.35	0.37	0.96	0.80-1.15	0.64
6	81094287	rs12207986	G	A	G	0.46	0.46	0.45	1.04	0.88-1.23	0.64
11	65583066	rs3903072	T	G	T	0.45	0.46	0.45	1.04	0.88-1.24	0.64
4	106084778	rs9790517	T	C	C	0.21	0.20	0.21	0.95	0.77-1.18	0.64
5	158244083	rs1432679	C	T	T	0.44	0.44	0.44	1.04	0.87-1.24	0.65
10	22032942	rs7072776	A	G	A	0.29	0.29	0.29	1.05	0.86-1.27	0.65
6	26680698	rs71557345	A	G	A	0.10	0.10	0.10	1.07	0.80-1.44	0.65
4	89243818	rs10022462	T	C	T	0.45	0.46	0.45	1.04	0.88-1.22	0.66
11	1909006	rs3817198	C	T	C	0.32	0.33	0.32	1.04	0.87-1.24	0.66
5	1279790	rs10069690	T	C	T	0.28	0.29	0.28	1.05	0.86-1.28	0.66
2	25129473	rs6725517	G	A	G	0.44	0.45	0.43	1.04	0.88-1.23	0.66
22	29621477	rs132390	C	T	T	0.03	0.03	0.03	0.89	0.53-1.51	0.67
5	44706498	rs10941679	G	A	G	0.28	0.29	0.27	1.04	0.86-1.27	0.67
14	91841069	rs941764	G	A	G	0.34	0.34	0.34	1.04	0.86-1.25	0.67
7	21940960	rs7971	G	A	A	0.32	0.31	0.32	0.96	0.80-1.16	0.68
18	42399590	rs6507583	G	A	G	0.07	0.08	0.07	1.07	0.78-1.48	0.68
5	169591487	rs4562056	T	G	G	0.36	0.36	0.36	1.03	0.87-1.23	0.70
4	84370124	rs1459137999	TAA	TA	TA	0.50	0.49	0.50	0.97	0.81-1.15	0.70

6	20621238	rs2223621	T	C	T	0.39	0.40	0.39	1.03	0.86-1.24	0.73
8	128355618	rs13281615	G	A	G	0.42	0.43	0.41	1.03	0.87-1.23	0.73
18	24570667	rs1436904	G	T	T	0.42	0.43	0.42	0.97	0.82-1.15	0.73
17	77781725	rs745570	G	A	A	0.49	0.48	0.50	0.97	0.82-1.15	0.73
3	4742276	rs6762644	G	A	A	0.41	0.40	0.41	0.97	0.82-1.15	0.73
21	16520832	rs2823093	A	G	A	0.26	0.26	0.26	1.03	0.85-1.25	0.74
1	145644984	rs12405132	T	C	C	0.36	0.36	0.37	0.97	0.81-1.16	0.75
1	118230221	rs7529522	C	T	T	0.25	0.25	0.25	0.97	0.79-1.18	0.75
16	53855291	rs11075995	A	T	A	0.23	0.24	0.22	1.03	0.84-1.26	0.75
2	111925731	rs71801447	C	CTT ATG TT	C	0.05	0.05	0.05	1.06	0.72-1.58	0.75
3	87037543	rs13066793	G	A	A	0.07	0.07	0.07	0.95	0.65-1.37	0.77
7	94113799	rs17268829	C	T	T	0.31	0.30	0.31	0.97	0.80-1.18	0.77
9	136151579	chr9:136151579	T	TGG TGC AGG CGC AGG AAA AAA TTG TGG CAA TTC CTCA	T	0.18	0.18	0.18	1.03	0.83-1.29	0.77
14	105212261	rs10623258	CTT	C	CTT	0.46	0.46	0.47	1.03	0.86-1.22	0.78
6	16399557	rs3819405	T	C	T	0.33	0.33	0.32	1.03	0.85-1.24	0.78
2	217920769	rs4442975	G	T	G	0.49	0.50	0.48	1.02	0.86-1.22	0.79
5	58337481	rs1353747	G	T	T	0.11	0.11	0.11	0.96	0.72-1.28	0.79

18	24337424	rs527616	C	G	G	0.35	0.35	0.35	1.02	0.85-1.23	0.80
1	41380440	rs4233486	C	T	T	0.32	0.32	0.32	0.98	0.81-1.18	0.80
19	44286513	rs3760982	A	G	G	0.46	0.45	0.46	1.02	0.86-1.21	0.81
17	40836389	rs72826962	T	C	C	0.02	0.02	0.02	1.09	0.54-2.22	0.81
12	115836522	rs1292011	G	A	G	0.41	0.43	0.41	1.02	0.85-1.22	0.82
10	64299890	rs10995201	G	A	A	0.15	0.15	0.15	0.97	0.74-1.27	0.83
16	80650805	rs13329835	G	A	G	0.22	0.23	0.22	1.02	0.83-1.25	0.83
1	201437832	rs3538942	T	C	C	0.06	0.06	0.06	0.96	0.65-1.42	0.85
16	56420987	rs2432539	A	G	A	0.40	0.40	0.40	1.02	0.85-1.21	0.86
5	81538046	rs7707921	T	A	A	0.25	0.24	0.26	0.98	0.80-1.20	0.86
10	123349324	rs45631563	T	A	T	0.04	0.04	0.04	1.04	0.66-1.65	0.86
10	123093901	rs11199914	T	C	C	0.31	0.30	0.31	0.99	0.82-1.19	0.88
15	91512067	rs2290203	A	G	G	0.21	0.21	0.21	1.02	0.83-1.25	0.89
5	111217786	rs6882649	G	T	G	0.32	0.32	0.31	0.99	0.82-1.19	0.89
3	27416013	rs4973768	T	C	T	0.46	0.46	0.46	1.01	0.85-1.20	0.90
12	14413931	rs12422552	C	G	C	0.26	0.26	0.26	1.01	0.83-1.23	0.90
8	106358620	rs12546444	T	A	T	0.10	0.10	0.10	1.02	0.75-1.38	0.91
22	42038786	rs73161324	T	C	C	0.05	0.05	0.05	0.98	0.64-1.49	0.91
1	217220574	rs11117758	A	G	A	0.23	0.23	0.22	0.99	0.81-1.21	0.91
9	110837176	rs13294895	T	C	C	0.19	0.19	0.19	0.99	0.79-1.24	0.93
4	126843504	rs77528541	T	G	G	0.14	0.14	0.14	0.99	0.76-1.29	0.93
14	93104072	rs11627032	C	T	T	0.26	0.26	0.26	0.99	0.81-1.21	0.93
10	9088113	rs67958007	T	TG	T	0.13	0.14	0.13	0.99	0.77-1.28	0.93
6	1318878	rs11242675	C	T	C	0.38	0.39	0.38	0.99	0.83-1.19	0.93
2	174212894	rs1550623	G	A	G	0.15	0.16	0.15	0.99	0.78-1.26	0.94
5	50195093	rs35951924	AT	A	AT	0.32	0.33	0.32	0.99	0.81-1.21	0.94

1	42137311	rs79724016	G	T	G	0.03	0.03	0.03	0.98	0.59-1.63	0.94
6	13722523	rs204247	G	A	G	0.41	0.42	0.40	0.99	0.83-1.18	0.94
5	58184061	rs10472076	C	T	C	0.37	0.37	0.36	0.99	0.83-1.19	0.95
11	69331642	rs554219	G	C	G	0.46	0.41	0.49	1.01	0.77-1.32	0.95
14	37132769	rs2236007	A	G	G	0.21	0.21	0.22	0.99	0.81-1.22	0.96
12	85009437	rs202049448	C	T	C	0.30	0.31	0.29	1.00	0.82-1.23	0.98
17	53209774	rs2787486	C	A	C	0.31	0.32	0.31	1.00	0.83-1.21	0.98
9	110306115	rs10759243	A	C	A	0.27	0.27	0.27	1.00	0.82-1.22	0.99
1	10566215	rs616488	G	A	A	0.34	0.33	0.34	1.00	0.83-1.20	1.00

Supplementary Table 3. Results of association testing (SKATO) for 219 genes with rare coding variants. Results are ordered according to P-value.

Gene	Sample Size	Case MAF	Control MAF	Beta	P-value
RECQL5	824	0.0014	0.00076	7.02	0.0043
CDKN1B	891	0.0017	0.00	1.99	0.0072
HSCB	894	0.0050	0.00	1.99	0.0074
RAD23B	894	0.0017	0.00	1.33	0.019
UBE2N	894	0.0017	0.00	1.33	0.019
MED12	842	0.0014	0.0032	-2.38	0.027
OGG1	894	0.00064	0.0019	-6.45	0.032
TOP2A	890	0.00	0.0015	-3.03	0.046
RAD1	894	0.0028	0.0011	1.97	0.052
ERCC6-PGBD3	892	0.0018	0.00078	5.26	0.057
RAD54L	890	0.0015	0.00072	4.96	0.058
GTF2H1	893	0.0017	0.0023	-1.76	0.060
CHEK1	894	0.0011	0.00028	1.98	0.060
MRPS34	894	0.010	0.0034	2.63	0.072
ATM	893	0.0013	0.00084	9.34	0.081
GTF2H5	894	0.0017	0.00	0.66	0.081
MNAT1	894	0.0017	0.00	0.66	0.081
MRPS18C	894	0.0017	0.00	0.66	0.081
PTEN	894	0.0017	0.00	0.66	0.081
UBC	894	0.0017	0.00	1.33	0.081
RAD9B	875	0.00	0.0012	-3.38	0.081
ATR	892	0.00087	0.0017	-6.77	0.084
XPA	890	0.0017	0.00028	1.65	0.089
TREX1	894	0.0029	0.0011	2.96	0.091
RAD54L2	894	0.00093	0.00084	0.58	0.098
XRCC4	893	0.0029	0.0011	2.96	0.11
POLL	894	0.0011	0.00037	2.63	0.11
MMS19	894	0.00083	0.0017	-5.47	0.11
ANTXR1	889	0.00	0.0013	-2.02	0.11
XRCC2	893	0.00	0.00084	-2.02	0.11
RB1	889	0.00033	0.0012	-3.38	0.12
ERCC8	882	0.0012	0.00084	1.22	0.12
EME1	894	0.0012	0.00076	1.61	0.13
XPC	894	0.0011	0.00061	1.95	0.13
AKT1	894	0.0011	0.00028	0.99	0.15
CDK7	894	0.0011	0.00028	0.99	0.15
NABP2	894	0.0011	0.00028	0.99	0.15
ALKBH3	892	0.00071	0.0012	-1.39	0.15
RPA4	894	0.0010	0.00034	1.32	0.15

FANCE	894	0.00	0.00084	-1.68	0.15
XRCC1	887	0.00038	0.0012	-3.70	0.16
ERCC3	894	0.00042	0.0011	-3.40	0.16
PER2	890	0.0010	0.00053	2.60	0.17
ERCC2	892	0.00028	0.00080	-3.73	0.18
LIG1	892	0.00083	0.0011	-1.45	0.18
MLH3	872	0.00055	0.00098	-3.83	0.18
POLN	885	0.0015	0.00077	3.29	0.19
RAD17	853	0.00092	0.0019	-3.08	0.19
RECQL4	886	0.0011	0.00072	4.85	0.21
TP53	894	0.00083	0.00042	1.63	0.22
XRCC5	893	0.0017	0.00070	2.29	0.22
TOPBP1	888	0.00078	0.00060	1.24	0.22
E2F1	893	0.00083	0.0023	-2.37	0.24
LRIG3	893	0.00077	0.00071	0.29	0.24
PMS2	881	0.00095	0.0016	-3.72	0.25
CCNE1	894	0.00055	0.0015	-2.38	0.26
RAD23A	893	0.0010	0.0015	-1.04	0.26
MBD4	886	0.0015	0.0025	-3.50	0.28
ERCC6	894	0.0012	0.00077	4.21	0.28
RPA1	893	0.00071	0.00072	-0.034	0.28
WDR48	894	0.0011	0.0022	-2.73	0.28
FNDC8	894	0.0011	0.0014	-0.36	0.28
MSH3	891	0.00071	0.0012	-4.13	0.30
BRCA1	894	0.0015	0.0011	4.80	0.30
LIG3	892	0.00045	0.00084	-1.71	0.31
BLM	874	0.0016	0.0012	3.24	0.31
CHEK2	889	0.0011	0.00075	2.55	0.32
DCLRE1C	869	0.0010	0.00051	1.31	0.32
RAD52	893	0.00	0.00084	-1.01	0.32
POLB	894	0.00	0.00084	-1.01	0.33
RNASEL	894	0.00042	0.0011	-2.04	0.34
CDK4	894	0.00083	0.00042	0.65	0.35
DCLRE1A	894	0.0014	0.0021	-3.48	0.35
POLM	894	0.0014	0.0020	-3.05	0.36
EPM2AIP1	893	0.00083	0.00084	-0.0078	0.37
FAAP100	875	0.0016	0.0012	3.54	0.38
ZAR1L	889	0.00	0.0011	-1.34	0.38
POLE	893	0.00069	0.00089	-3.22	0.38
CLK2	894	0.00071	0.00060	0.31	0.40
RPA2	893	0.0017	0.0025	-1.38	0.40
MUS81	894	0.00062	0.0013	-2.05	0.40
BIVM-ERCC5	874	0.00083	0.00089	-0.14	0.40
ALKBH2	894	0.00071	0.00060	0.31	0.40
ERCC4	886	0.0010	0.0016	-3.50	0.41

APEX1	894	0.00024	0.00072	-1.36	0.41
NEIL3	829	0.0029	0.0019	2.31	0.41
FANCA	889	0.0010	0.00082	3.52	0.42
NEIL2	894	0.00033	0.0010	-1.36	0.42
TMPRSS2	840	0.0024	0.0014	1.62	0.42
EME2	806	0.00067	0.0012	-2.01	0.42
ABL1	888	0.0015	0.0017	-0.70	0.42
BRIX1	893	0.00083	0.00042	0.33	0.43
CCNH	894	0.00083	0.00042	0.33	0.43
DMC1	894	0.00083	0.00042	0.33	0.43
RAD51	894	0.00083	0.00042	0.33	0.43
SLC23A3	894	0.00083	0.00042	0.33	0.43
FOXA1	888	0.00083	0.00042	0.32	0.44
PPM1D	894	0.0017	0.00084	1.31	0.44
RBBP8	867	0.00093	0.00058	2.26	0.44
FANCG	894	0.00071	0.00048	0.64	0.44
PALB2	885	0.0011	0.0017	-2.33	0.44
RDM1	892	0.00067	0.00067	-0.018	0.45
FANCI	894	0.0015	0.0011	4.14	0.45
RECQL	887	0.0019	0.0013	1.60	0.47
RAD54B	881	0.00051	0.00091	-2.03	0.48
NEIL1	835	0.00074	0.00085	-0.37	0.48
RAD51C	893	0.0020	0.0020	-0.067	0.48
HUS1	894	0.0017	0.0025	-1.04	0.48
ATRIP	871	0.0012	0.00093	0.99	0.49
BARD1	890	0.0020	0.0030	-1.10	0.50
SLX4	894	0.0014	0.0015	-1.29	0.51
FSBP	893	0.00042	0.0011	-1.02	0.51
NHEJ1	893	0.00092	0.00056	1.29	0.51
DCLRE1B	894	0.00066	0.0011	-1.72	0.52
ESR2	883	0.0012	0.0011	0.32	0.53
DDB1	894	0.00066	0.00051	0.32	0.54
FANCF	894	0.00066	0.00051	0.32	0.54
GTF2H4	876	0.00057	0.0014	-0.99	0.55
PMS1	887	0.0013	0.00091	1.91	0.56
CCNB1	893	0.00055	0.00070	-0.35	0.56
VARS2	894	0.00	0.00084	-0.67	0.56
CDK2	894	0.00	0.00084	-0.67	0.57
RFFL	893	0.00	0.00084	-0.67	0.57
FANCB	889	0.0012	0.00076	1.62	0.57
ERCC1	894	0.00	0.00084	-0.67	0.57
SPOP	894	0.00	0.00084	-0.67	0.57
MRPL36	894	0.00	0.00084	-0.67	0.57
POLQ	891	0.00079	0.0010	-3.63	0.57
MGMT	894	0.00	0.00084	-0.67	0.57
HCN3	894	0.00	0.00084	-0.67	0.57

MPG	885	0.00083	0.00051	1.30	0.57
ESR1	860	0.0012	0.0018	-2.07	0.57
MSR1	887	0.0012	0.0017	-2.10	0.58
APEX2	894	0.0017	0.0024	-1.40	0.58
GEN1	888	0.0012	0.0016	-2.19	0.60
AR	827	0.0014	0.0019	-0.32	0.61
CDH1	848	0.00089	0.00062	1.61	0.62
MSH5	894	0.0018	0.0021	0.95	0.62
XRCC6	891	0.00047	0.00072	-0.70	0.64
FANCC	819	0.0014	0.0011	0.97	0.64
CD3EAP	894	0.00062	0.0011	-1.38	0.65
PNKP	875	0.00083	0.0012	-1.41	0.66
APLF	891	0.0020	0.0027	-1.41	0.67
USP1	885	0.00090	0.00091	-0.024	0.67
FANCD2	885	0.0013	0.0015	0.67	0.68
CLSPN	893	0.0020	0.0016	2.23	0.68
MSH2	812	0.00049	0.00060	-0.010	0.68
TOE1	894	0.0017	0.00084	0.65	0.69
FZR1	894	0.00055	0.00056	-0.010	0.70
STK11	894	0.00055	0.000562	-0.010	0.70
GTF2H3	894	0.0012	0.00084	0.64	0.70
POLD1	808	0.0015	0.00095	1.68	0.70
BRIP1	854	0.0010	0.0014	-2.02	0.71
RAD50	892	0.0013	0.0010	2.15	0.71
HID1	843	0.00078	0.00056	1.35	0.71
UNG	835	0.00042	0.00084	-0.66	0.72
NDUFAF2	894	0.0017	0.00084	0.33	0.72
FANCL	880	0.0017	0.0020	-1.73	0.73
TET2	894	0.0017	0.0015	2.20	0.73
LIG4	889	0.00076	0.00077	-0.045	0.73
FAAP24	894	0.0025	0.0020	1.26	0.75
MDC1	894	0.00066	0.00063	0.26	0.76
CDC25C	894	0.00083	0.0013	-0.69	0.76
CCND1	894	0.00	0.00084	-0.34	0.76
CETN2	894	0.00	0.00084	-0.34	0.76
DHFR	894	0.00	0.00084	-0.34	0.76
EN2	894	0.00	0.00084	-0.34	0.76
ERCC5	894	0.00	0.00084	-0.34	0.76
HOXB9	894	0.00	0.00084	-0.34	0.76
N4BP2L1	894	0.00	0.00084	-0.34	0.76
PMEL	894	0.00	0.00084	-0.34	0.76
PRPF19	894	0.00	0.00084	-0.34	0.76
SAP30BP	894	0.00	0.00084	-0.34	0.76
SAPCD1	894	0.00	0.00084	-0.34	0.76
TSC2	894	0.00	0.00084	-0.34	0.76
VRK2	894	0.00	0.00084	-0.34	0.76

ZNF276	894	0.00	0.00084	-0.34	0.76
CDK1	893	0.00	0.00084	-0.34	0.76
DUT	893	0.00	0.00084	-0.34	0.76
ATRX	853	0.0013	0.0016	-1.49	0.78
NUDT1	894	0.00074	0.00075	-0.040	0.78
SETMAR	889	0.00047	0.00060	-0.35	0.79
XAB2	849	0.0015	0.0020	-1.62	0.79
CDKN1A	892	0.0012	0.0016	-1.06	0.80
LRIG1	894	0.00074	0.00078	-0.46	0.80
TP53BP1	893	0.0013	0.0013	0.44	0.80
TDP1	894	0.00083	0.00056	0.64	0.80
TDG	850	0.085	0.084	-0.049	0.80
FANCM	894	0.0011	0.0011	0.45	0.81
XRCC3	843	0.00083	0.00071	0.63	0.81
UVSSA	764	0.0015	0.0014	1.06	0.81
MDM4	894	0.0014	0.0011	0.62	0.83
RAD9A	892	0.00062	0.00074	-0.37	0.83
NBR1	892	0.0017	0.0016	0.56	0.83
KCNH2	820	0.00055	0.00076	-0.39	0.83
CASP8	894	0.00042	0.00063	-0.35	0.83
MSH4	848	0.00067	0.00093	-0.80	0.84
HOXB13	893	0.00062	0.00084	-0.71	0.84
MLH1	893	0.0014	0.0012	1.48	0.85
ERG	894	0.0017	0.0017	-0.010	0.87
BRCA2	889	0.0012	0.0011	2.16	0.88
WRN	892	0.0017	0.0016	1.78	0.88
HELQ	887	0.0012	0.0011	0.59	0.89
RAD51D	887	0.0013	0.0012	0.29	0.90
PARP1	893	0.0015	0.0017	-0.80	0.90
TKFC	894	0.00055	0.00084	-0.35	0.91
PURG	894	0.00055	0.00084	-0.35	0.91
MDM2	894	0.00055	0.00084	-0.35	0.91
PARP2	894	0.00071	0.00072	-0.030	0.91
CDC25A	879	0.00066	0.00068	-0.014	0.92
DDB2	894	0.0017	0.0017	-0.020	1.00
MRE11A	893	0.0019	0.0017	0.58	1.00
MSH6	884	0.0010	0.0011	-0.39	1.00
MUTYH	894	0.0011	0.0012	-0.76	1.00
NBN	880	0.0011	0.0011	-0.48	1.00
NTHL1	843	0.00085	0.00085	-0.014	1.00
PER1	894	0.00076	0.00081	-0.44	1.00
RAD51B	887	0.0017	0.0017	-0.073	1.00
SMUG1	894	0.00066	0.00084	-0.36	1.00

Supplementary Table 4. Results of association testing (weighted burden test) for 33 rare coding truncating genes. Results are ordered according to P-value.

Gene	Sample Size	Case MAF	Control MAF	Beta	P-value
RBBP8	883	0.00093	0.00058	1.98	0.019
FANCA	890	0.0010	0.00082	1.99	0.019
UVSSA	894	0.0015	0.0014	1.33	0.057
RAD9A	893	0.00062	0.00074	0.99	0.16
WRN	890	0.0017	0.0016	1.64	0.18
CHEK2	891	0.0011	0.00075	1.28	0.26
PMS1	894	0.0013	0.00091	0.33	0.31
XRCC3	894	0.00083	0.00071	0.33	0.33
ALKBH3	894	0.00071	0.0012	0.33	0.33
RECQL	894	0.0019	0.0013	0.33	0.33
POLQ	882	0.00079	0.0010	0.33	0.34
ERCC8	892	0.0012	0.00084	0.33	0.35
FANCM	890	0.0011	0.0011	0.31	0.46
BRCA2	892	0.0012	0.0011	-0.01	0.47
FANCI	891	0.0015	0.0011	-0.010	0.49
RAD54B	892	0.00051	0.00091	-0.012	0.49
BLM	894	0.0016	0.0012	-0.010	0.56
BRIP1	857	0.0010	0.0014	-0.34	0.61
RAD1	894	0.0028	0.0011	-0.35	0.61
ERCC6	894	0.0012	0.00077	-0.68	0.63
WDR48	894	0.0011	0.0022	-0.35	0.66
FANCL	885	0.0017	0.0020	-1.39	0.72
ERCC1	894	0.00	0.00084	-0.67	0.76
MED12	842	0.0014	0.0032	-0.68	0.78
MSH5	894	0.0018	0.0021	-0.67	0.78
NTHL1	878	0.00085	0.00085	-0.67	0.78
MRPL36	894	0.00	0.00084	-0.67	0.78
DCLRE1B	894	0.00066	0.0011	-0.67	0.79
RECQL4	891	0.0011	0.00072	-0.67	0.79
ERCC3	894	0.00042	0.0011	-1.01	0.85
NUDT1	894	0.00074	0.00075	-1.01	0.85
FANCC	819	0.0014	0.0011	-1.00	0.87
NEIL1	841	0.00074	0.00085	-1.35	0.90