Genetic Susceptibility to Endometrial Cancer

Timothy Hua Tse Cheng

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2015

Genetic Susceptibility to Endometrial Cancer

Timothy Hua Tse Cheng
Pembroke College, University of Oxford, D.Phil, Hilary 2015

Abstract

Endometrial cancer (EC) is the fourth most common cancer affecting women in the UK. Those with a family history of EC have an increased risk compared with the general population. Highly penetrant germline mutations in mismatch repair (MMR) genes and DNA polymerases account for only a small proportion of the familial aggregation.

The aim of this thesis is to investigate the genetic susceptibility to EC in the general population using cases and controls of European ancestry. A GWAS meta-analysis totalling 7,737 EC cases and 37,144 controls yielded five novel EC risk loci of genome-wide significance ($P < 5x10^{-8}$). In decreasing order of significance, these were at chromosomes 13q22.1 (rs11841589, near *KLF5*), 6q22.31 (rs13328298, in *LOC643623* and near *HEY2* and *NCOA7*), 8q24.21 (rs4733613, telomeric to *MYC*), 15q15.1 (rs937213, in *EIF2AK4*, near *BMF*) and 14q32.33 (rs2498796, in *AKT1*, near *SIVA1*). A second independent EC signal was found in the 8q24 locus.

The association found in a previous EC GWAS at *HNF1B* on chromosome 17 was replicated at a higher significance, with the most significant SNP being rs11263763. *CYP19A1* SNPs have previously been associated with EC and higher circulating levels of oestrogen from candidate studies, but I confirmed this locus to be genome-wide significant for the first time.

Functional annotation and *in vitro* studies for the EC risk loci at the intergenic region of chromosome 13q22 suggested that the functional SNP sits within a transcriptional repressor for *KLF5*, with the higher-risk allele reducing repressor activity.

The propensity for germline MMR and DNA polymerase muations to cause both EC and colorectal cancer (CRC) prompted me to search for common variants associated with both cancer phenotypes. An EC CRC GWAS meta-analysis showed little evidence of shared susceptibility loci. However, this meta-analysis revealed a novel genome-wide significant risk locus: rs3184504, a missense *SH2B3* SNP that has not previously been associated with either EC or CRC.

This thesis has enhanced the understanding of genetic susceptibility to sporadic EC and increased the number of genome-wide EC-associated variants to seven.

Acknowledgements

I would like to thank my supervisor Professor Ian Tomlinson for his encouragement and support. I have really enjoyed the scientific process and your guidance and approach to tackling challenges has been inspiring. Being part of this group has been a wonderful experience.

Members of the Tomlinson group have been very helfpul in this process. I am particularly grateful to Susanne and Annie for the help with the experimental work presented in chapter 5 and to Luke for help with editing the thesis. Others, including Dan, Claire Michal, Francesc, Shazia, David, Danny and Emily have kept the first floor cubicle area a fun and vibrant place despite the lack of windows.

I would also like to thank the Rhodes Trust and the Nuffield Department of Medicine for their financial support. This adventure in Oxford would not have occurred without the inspiring Dr Deanna Lee Rudgard.

I owe a huge debt of gratitude to my mum, dad and Naomi. Thank you so much for the encouragement and constant prayer.

Last but not least, I would like to thank my dearest Jessica for her love and support.

Preface

This thesis is a collection of the research work carried out by the author in the Molecular and Population Genetics Laboratory, Wellcome Trust Centre for Human Genetics, Department of Clinical Medicine, University of Oxford, between October 2010 and April 2015, under the supervision of Professor Ian Tomlinson.

I conducted the genotyping for the NSECG samples, and I conducted all of the experimental work and data analysis presented in this thesis. The only exception is the chromosome 13q22 luciferase assay work which I helped plan, but the experimental work was conducted by Dr Susanne Flach.

The work presented in chapter 4 and chapter 5 is currently being prepared for submission in a manuscript entitled "Five endometrial cancer risk loci identified through genome-wide association analysis." The work presented in chapter 6 is part of a manuscript entitled "Meta-analysis of genome-wide association studies identifies a common susceptibility polymorphism for colorectal and endometrial cancer near *SH2B3*."

I state that the research described herein is original, though the work of others is cited with appropriate references within the text. No part of this research has previously been submitted for a degree at this or any other University.

Contents

List of Figures

List of Tables

List of Abbreviations

A	Adenine
ANECS	Australian National Endometrial Cancer Study
bp	Base pair
C	Cytosine
ChIP	Chromatin immuno-precipitation
Chr	Chromosome
CI	Confidence interval
cM/Mb	Centimorgan/megabse
CNV	Copy number variant
COSMIC	Catalogue of Somatic Mutations in Cancer
DHS	DNase1 hypersensitivity site
DNA	Deoxyribonucleic acid
DZ	Dizygotic
EC	Endometrial cancer
EDM	Exonuclease domain mutations
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression quantitative trait loci
FAIRE	Formaldehyde-assisted isolation of regulatory elements
G	Guanine
GTEx	Genotype Tissue Expression Project

GWAS	Genome-wide association study
HapMap	International Halotype Map Project
HRT	Hormone replacement therapy
HWE	Hardy-Weinberg equilibrium
IBD	Idendity by descent
IBS	Identity by state
iCOGS	Collaborative Oncological Gene-environment Study
Indel	Insertion / deletion
kb	kilobase
LD	Linkage disequilibrium
LS	Lynch syndrome
LVSI	Lymphovascular space invasion
MAF	Minor allele frequency
mb	megabase
MZ	Monozygotic
NSECG	National Study of Endometrial Cancer Genetics
OCP	Oral contraceptive pills
OR	Odds ratio
PCA	Principle components analysis
PCR	Polymerase chain reaction
QQ	Quantile-quantile
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
SE	Standard error
SEARCH	Studies of Epidemiology and Risk factors in Cancer Heredity
SNP	Single nucleotide polymorphism
T	Thymine

TCGA	The Cancer Genome Atalas
UK	United Kingdom
USA	United States of America
WTCCC	Wellcome Trust case control consortium

Chapter 1

Introduction

1.1 Endometrial cancer

1.1.1 Background and epidemiology

Endometrial cancer (EC) is the fourth most common cancer in women in Europe (Ferlay *et al.*, 2013) and the United States (Siegel *et al.*, 2014), and the most common cancer of the female reproductive system. In many economically developing countries, EC is the second most common gynaecological cancer after cancer of the uterine cervix (Jemal *et al.*, 2011). The worldwide incidence of EC in 2012 was 320,000 (World Cancer Research Fund, www.wcrf.org).

In the United Kingdom, about 7,400 women are diagnosed with EC every year. The age-standardized EC incidence rates have risen from 12.8 per 100,000 in 1975 to 20.1 per 100,000 in 2011 (www.ons.gov.uk, figure 1.1).

The increased incidence is seen predominantly in older age groups: incidence among women aged below 55 years has decreased since 1975 but the incidence among women aged 65-79 has doubled. Rising EC rates have been attributed to the increasing prevalence of obesity (Bray *et al.*, 2005) and the use of oestrogen-only forms of hormonal replacement therapy (HRT), which was particularly common in the 1960s

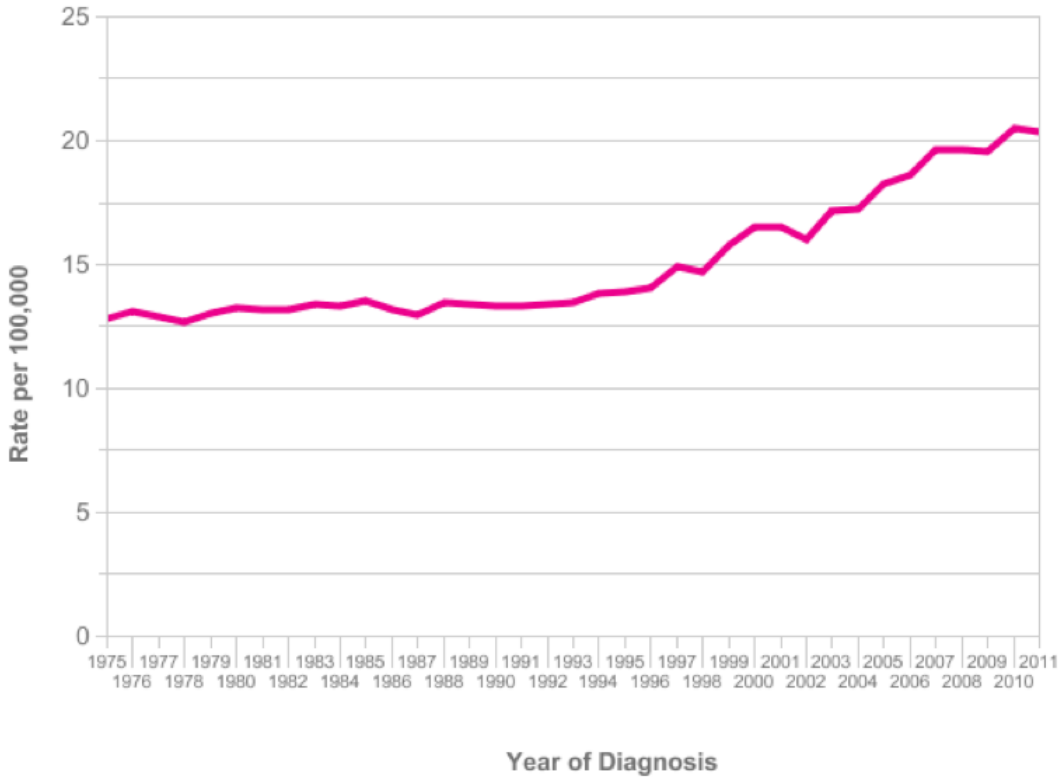

Figure 1.1: UK age-adjusted EC incidence rate per 100,000 females from 1975-2011 (www.ons.gov.uk).

and 1970s (Cogliano *et al.*, 2011; Dossus *et al.*, 2010).

The incidence of EC is related to age. From age 40, age-specific incidence rates rise sharply, reaching a maximum of 96 per 100,000 in the 70-74 age group. Unlike many other cancers, EC incidence then decreases in older age groups.

The incidence of EC are similar in White, Asian and Black females in the UK (www.ncin.org.uk) while in the US, the incidence is higher in White women than in Black, Hispanic, or in Asian women (Allard and Maxwell, 2009).

The 5 and 10-year survival rates for EC are 79.0% and 77.5% respectively (Quaresma *et al.*, 2014). This suggests that most patients can be considered cured after five disease-free years. The survival rate of EC is similar to that of breast cancer. It ranks fourth highest in 10-year survival rate for common cancers affecting females, though this rate decreases with age. However, the 5-year survival rate has increased steadily, from 58.8% in the 1970s to 79.0% in 2011. In the US, the survival rate is higher in White compared to African-American women and this may be due to the fact that African-American women are more likely to get aggressive cancer subtypes and to present at a later stage (Allard and Maxwell, 2009; Elshaikh *et al.*, 2013; Smotkin *et al.*, 2012).

Thus, EC is the 10th most common cause of cancer death among women in the UK and accounts for 3% of all female cancer deaths (www.ons.gov.uk).

1.1.2 Risk factors

The majority of known EC risk factors are thought to contribute to EC risk by the unopposed effects of oestrogen. Oestrogen causes the simple columnar epithelium of the endometrium to proliferate, and prolonged, unopposed oestrogen stimulation can cause endometrial hyperplasia and eventually, adenocarcinoma.

1.1.2.1 Endogenous oestrogen

There are factors in a woman's reproductive history that increase endogenous oestrogen levels. Nulliparous women have a 35-42% increased EC risk compared with parous women (Schonfeld *et al.*, 2013). Pregnancy is thought to reduce EC risk by increasing progesterone levels (Kim and Chapman-Davis, 2010) and by inducing physical changes to the uterus (Baird and Dunson, 2003). Decreased EC risk is linked specifically to a greater number of full-term pregnancies (Dossus *et al.*, 2010).

A longer menstrual lifespan causes a higher cumulative exposure to oestrogen: early menarche and late menopause have been associated with increased EC risk (Schonfeld *et al.*, 2013). EC risk can be up to 11% higher for women with early menarche (age<13), and up to 34% higher in women with menopause at age 50-54. EC risk in post-menopausal women is correlated with higher circulating levels of oestrogen and androgen and lower sex hormone-binding globulin, which leads to an increased proportion of unbound, biologically active sex hormones (Lukanova *et al.*, 2004; Nyholm *et al.*, 1993; Zeleniuch-Jacquotte *et al.*, 2001).

Some women suffer from infertility as a result of chronic anovulation, and apart from the EC risk conferred by nulliparity, there is unopposed oestrogen production that causes continued proliferation of the endometrium (Brinton *et al.*, 2005). This leads to an increased risk of endometrial hyperplasia and EC.

Other medical conditions can also influence endogenous oestrogen levels. An estimated 34% of ECs are linked to obesity (Parkin *et al.*, 2011). EC risk is increased by 81% for every 5 units' increase in body mass index (BMI), and EC risk is up to 2.5 times higher in obese women (BMI>30, Zhang *et al.*, 2014). Excess adipose tissue increases the peripheral aromatization of androgens to oestradiol. This causes a higher circulating level of oestrogen and affects the ovulatory cycle. In obesity, there is a decreased sensitivity to leptin and higher leptin levels may also directly contribute to EC risk (Wang *et al.*, 2014a). EC risk is up to 81% higher in women with diabetes

and metabolic syndrome (Esposito *et al.*, 2014; Zhang *et al.*, 2013). This risk may be explained by the prevalence of obesity in these patients (Luo *et al.*, 2014). Similarly, women with polycystic ovary syndrome have a 2.8 times increased EC risk (Barry *et al.*, 2014) and this may be due to the effects of unopposed oestrogen and the co-existence of metabolic syndrome in these women (Palomba *et al.*, 2009). Obesity, type 2 diabetes and metabolic syndrome are interlinked phenotypes that result in a series of endocrine changes including lower circulating levels of sex hormone-binding globulin, insulin resistance and chronic anovulation (Amant *et al.*, 2005; Potischman *et al.*, 1996). Oestrogen-secreting tumours are rare but may occur; an example is granulosa cell tumours arising from the ovary. ECs associated with granulosa cell tumours are generally early stage and well-differentiated (Evans *et al.*, 1980).

1.1.2.2 Exogenous oestrogen

Exogenous oestrogen can be taken in systemic doses as hormone replacement therapy (HRT) in post-menopausal women or as contraception (per oral or transdermal) for women of child-bearing age. There are various forms of HRT and contraception that contain oestrogen and progesterone, and it is the formulations that promote the unopposed effects of oestrogen that confer an increased EC risk. Oestrogen-only HRT increases EC risk by 2.3 times (Grady *et al.*, 1995), and the effects of combined HRT depend on the dosage of progesterone in the formulation (Brinton and Felix, 2014). Women who have used combined oral contraceptive pills (OCP) have a 26-43% reduced risk of EC, but this protective effect may be limited to those who have more than five years of OCP (Havrilesky *et al.*, 2013). The protective effect persists for 20 years after OCP cessation (Mueck *et al.*, 2010).

Tamoxifen is a selective oestrogen receptor modulator used in the treatment of oestrogen receptor-postive breast cancer. It is an antagonist of the oestrogen receptor in breast tissue, but it acts as a partial agonist in the endometrium. The risk of EC

in breast cancer survivors who have been treated with tamoxifen is increased three times compared to non-users (Early Breast Cancer Trialists' Collaborative Group *et al.*, 2011). The risk of breast cancer recurrence is balanced against the risk of developing EC in these patients, such that tamoxifen is generally perscribed for a limited period of time (< 5 years).

1.1.2.3 Other factors

Cohort studies have shown that EC risk is higher in women with a personal history of breast and ovarian cancer (Hemminki *et al.*, 2003; Jégu *et al.*, 2014; Youlden and Baade, 2011). The use of Tamoxifen in breast cancer patients may contribute to the increased EC risk seen in breast cancer survivors. Also, those with a family history of EC and other cancers have an increased risk of developing EC (section 1.2).

In a meta-analysis of 34 case-control and cohort studies, cigarette smoking was associated with a decreased risk of developing EC (Zhou *et al.*, 2008). It has been proposed that this is because smoking increases the hepatic metabolism of oestrogen, although the other health risks of smoking outweigh this benefit.

As noted in section 1.1.1, the risk of EC increases with age until around age 75; beyond this age, the risk of developing EC then decreases.

There is literature proposing links between EC and physical activity (Moore *et al.*, 2010), diet (Bravi *et al.*, 2009; Tang *et al.*, 2009) and aspirin use, (Neill *et al.*, 2013) although further evidence is necessary to elucidate these relationships.

1.1.3 Diagnosis

EC is cancer of the uterus arising from the endometrial lining of the uterine body. Malignancies can also arise from the uterine cervix (cervical cancer) and uterine smooth muscle or connective tissue (uterine sarcoma), but these are usually treated as separate clinical entities. The relatively low mortality rate of EC can be attributed

to the fact that most cases are diagnosed at an early stage, as the average age of diagnosis is 61 years. The presenting symptom in 90% of EC cases is abnormal uterine bleeding.

For women presenting with bleeding in the genital area, it is important to localize the anatomical source of bleeding. Although uterine bleeding is the most likely cause of bleeding in the genital tract, the bleeding may also arise from the lower genital tract (vulva, vagina and cervix) or, less commonly, from other parts of the upper genital tract (fallopian tubes and ovaries). The bleeding may also arise from a non-gynaecologic site such as the urological or digestive tracts.

The likely aetiology of uterine bleeding depends on age, reproductive status, and pattern of bleeding (cyclic or non-cyclic). Given the epidemiology, cases of uterine bleeding in post-menopausal women are carefully investigated for EC. Physical examination, including pelvic examination, is likely to be unremarkable in most cases that present at an early stage. Findings such as a fixed, enlarged pelvic mass may be signs of late-stage disease.

In the past, post-menopausal bleeding was most commonly investigated using dilatation and curettage (D&C), but current diagnostic strategies have evolved to include less invasive procedures including transvaginal ultrasound and endometrial biopsy. D&C is a preferred procedure for some gynaecologists because it can be therapeutic as well as diagnostic, but it has false-negative rates as high as 10% (Colombo *et al.*, 2013). Transvaginal ultrasound can visualize the homogeneity and thickness of the endometrial lining (Jacobs *et al.*, 2011) and can be used in conjunction with endometrial sampling, which has a sensitivity of up to 99% (Timmermans *et al.*, 2010). Hysteroscopy is used to directly visualize the endometrium and to obtain targeted biopsies. Therefore, it has a high accuracy in diagnosing a focalized lesion but may not be able to exclude EC in all cases (Clark *et al.*, 2002).

EC is a histological diagnosis based on endometrial tissue obtained from curettage,

endometrial biopsy, biopsy during hysteroscopy, or hysterectomy.

1.1.4 Screening

EC screening is not performed routinely for women in the general population. This is because the majority of cases present at an early stage with endometrial bleeding and there are no high-quality data in support of the efficacy of screening in reducing EC mortality. Although EC is an important treatable disease and it is advantageous to diagnose and treat at an earlier stage, this is balanced by the relatively short latent period, and also invasiveness or acceptance of the screening methods. Gynaecological exams and other screening tests would have to conducted at a short interval in order to detect asymptomatic EC, and screening is only warranted in women with a higly increased risk of developing EC. Tests such as transvaginal ultrasound and endometrial sampling are a part of the EC diagnostic procedure and they have also been assessed as screening methods (Jacobs *et al.*, 2011). These methods are sensitive, but carry the risk of false positive results. The efficacy of screening also depends on factors such as menopause status and endometrial thickness.

Women with Lynch syndrome, a high-penetrance Mendelian cancer syndrome that predisposes affected individuals to both EC and colorectal cancer. Lynch syndrome is discussed in detail in section 1.4 and affected women have a lifetime risk of up to 70% for developing EC and tend to develop EC at an earlier age (Hampel *et al.*, 2006). Asymptomatic women with Lynch syndrome are advised to undergo annual endometrial sampling starting at age 30, or 5-10 years prior to the earliest age of first diagnosis of Lynch-associated cancer of any kind in the family (Lindor *et al.*, 2006; Meyer *et al.*, 2009). These patients may also choose to undergo risk-reducing hysterectomy and bilateral salpingo-oophorectomy at an appropriate time to prevent endometrial and ovarian cancers.

1.1.5 Histopathological characteristics

EC can be broadly split into two subgroups based on histological appearance, molecular characteristics and clinical outcome (Bokhman, 1983; Felix *et al.*, 2010). Type 1 ECs account for 80-90% of EC cases and display well-differentiated endometrioid histology. These tumours are thought to be caused by unopposed oestrogen exposure and carry a favourable prognosis. They are seen as a malignant extension of atypical or complex endometrial hyperplasia and it is the histological evidence of stromal invasion that marks it as adenocarcinoma. Type 2 tumours acount for 10-20% of ECs and include poorly differentiated tumours and those with non-endometrioid histology including serous, clear-cell, mucinous, squamous, transitional, mesonephric and undifferentiated cells. Type 2 tumours typically present later, behave more aggressively, and carry a worse prognosis (Moore and Fader, 2011).

Serous and clear-cell tumours are the commonest form of type 2 tumours, accounting for up to 15% of ECs. These histologies are considered to be high grade and confer a worse prognosis because myometrial and vascular invasion is more common, and metastasis may occur in the absence of myometrial invasion (Boruta *et al.*, 2004). Serous ECs are thought to develop from endometrial intraepithelial carcinomas (EIC), with neoplastic transformation occurring in a background of endometrial atrophy (Sherman, 2000).

Molecular studies have shown that endometrioid tumours are more likely harbour mutations in *PTEN*, *KRAS* , *CTNNB1*, and mismatch repair genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*). Serous tumours show overexpression of p16, abnormal expression of p53 and variable loss of the oestrogen receptor (ER). Clear-cell tumours typically show loss of ER and wild-type p53 staining (Chiang and Soslow, 2014; Hecht and Mutter, 2006).

1.1.6 Staging and risk assessment

EC staging is based on the 2010 joint International Federation of Gynecology and Obstetrics (FIGO)/ Tumour Nodes Metastasis (TNM) classification system (Benedet *et al.*, 2000; Pecorelli, 2009). Evaluation includes: i) clinical parameters such as gynaecological examination, ii) biochemical tests such as blood counts and liver/ renal function tests, and iii) imaging modalities such as chest X-ray, transvaginal ultrasound, computed tomography and magnetic resonance imaging. The aim of staging is to assess the extent of local involvement (T), and to look for regional (N) and distant (M) tumour spread. Several factors have been identified which increase the risk of recurrence in early stage disease. These include histological grade (III) and type (non-endometrioid), >50% myometrial invasion, lymphovascular space invasion (LVSI), lymph node metastasis, and tumour diameter >2cm.

EC staging is an important step after diagnosis. It plays a central role in formulating a management plan for each patient.

1.1.7 Treatment

The primary goal of EC treatment is curative surgery by means of complete tumour excision. Additional measures are undertaken as necessary in order to reduce the chance of recurrence.

Patients with low grade endometrioid cancers confined to the endometrium have low-risk EC and their chance of recurrence is low. For these women, surgery with complete excision of the local disease is curative and sufficient. Total extrafascial hysterectomy with bilateral salpingo-oophorectomy with pelvic and para-aortic lymph node dissection is the standard treatment and staging procedure for EC (Pecorelli, 2009). Newer surgical techniques, such as minimally invasive surgery in suitable low-risk patients have achieved comparable disease-specific survival rates, sometimes with lower rates of peri-operative complications (Mourits *et al.*, 2010).

Patients with extra-endometrial involvement, high grade histology and LVSI are considered intermediate risk, while those with stage III disease or serous or clear-cell histology are considered high-risk. Patients with an increased risk may benefit from adjuvant therapy to improve disease control. Adjuvant therapy may consist of radiotherapy, platinum-based chemotherapy, or a combination of the two. For late stage EC, the aim is maximal surgical debulking. Palliative forms of surgery, radiotherapy and hormonal therapy are available when curative intent is not feasible.

The aim of follow-up visits after treatment is the control of any treatment side-effects and the early detection of recurrent disease (Salani *et al.*, 2011). This is based monitoring of symptoms, regular physical examinations, and other modalities such as CA-125 and imaging studies as appropriate. The majority of recurrences occur within three years of treatment (Fung-Kee-Fung *et al.*, 2006) and recurrences can occur at loco-regional or distant sites (Ben Arie *et al.*, 2012).

1.2 Heritability of endometrial cancer

Broadly speaking, occurrences of EC can be divided in to familial and sporadic cases. Familial EC cases are those where there are many family members affected and the pedigree displays a Mendelian inheritance pattern (section 1.3.3). Familial EC is discussed in section 1.4. The majority of incidences of EC are sporadic: these patients do not have a Mendelian pedigree. Sporadic EC cases may have several risk factors for EC (section 1.1.2). This thesis concentrates on genetic variants associated with sporadic EC.

Heritability describes the proportion of an observable phenotype that can be explained by genetic variation. EC can be treated as a readily ascertained binary phenotype based on histological diagnosis. Its heritability can be estimated by observing the presence of EC in individuals with varying degrees of relatedness. Twin

studies compare phenotype concordance between monozygotic (MZ) and dizygotic (DZ) twins, on the basis that MZ twins are genetically identical and DZ twins share half their genetic variation.

An EC twin study based on the Swedish twin registry with 11,659 women (Terry *et al.*, 1999) found that there was little concordance between MZ (1/53) and DZ twins (3/107) for EC. Another twin study based on twin registries in Sweden, Denmark and Finland totalling 44,788 pairs of twins (Lichtenstein *et al.*, 2000) also found limited concordance in EC for MZ (1/107) and DZ twins (5/245). Although these are the largest twin registries with cancer phenotype information available, the rarity of twins limits this approach, even in relatively common cancers such as EC. The low number of twins with EC limits the statistical power availble to identify the heritability of EC, but the presence of discordant twins is informative in that it suggests there are other risk factors present besides heredity.

However, it has been empirically observed in the clinical setting that women with a family history of EC have a slightly increased risk of developing EC compared with the general population. Cohort and case-control studies have been used to investigate and quantify the risk of developing EC for first degree relatives of EC cases compared with the general population. The occurrence of EC in first degree relatives of EC patients can be compared with cohort-specific EC rates (Goldgar *et al.*, 1994) and the occurrence of first degree EC family history in matched population controls (Gruber and Thompson, 1996; Lucenteforte *et al.*, 2009; Parazzini *et al.*, 1994). These data were used to calculate ORs and estimate familial relative risk and the results showed that there was an increased familial relative risk (1.32) and OR (1.5 - 2.8).

A recent meta-analysis of 3,871 cases and 49,475 controls from ten case-control studies and 33,510 EC cases from six cohort studies found that first degree family history of EC conferred a relative risk of 1.82 (95%CI 1.65-1.98) for developing EC. First degree family history referred with those with one or more relatives (mothers,

sisters and daughters) with EC. This meta-analysis also found that women with a family history of CRC had an increased chance of developing EC (OR=1.17, 95%CI 1.03-1.31).

Results from twin studies suggest that there are factors other than germline genetics that determine predisposition to EC. This concurs with evidence from epidemiological studies where prolonged exogenous oestrogen treatment can cause a relative risk of ∼10. However, case-control and cohort studies show a familial relative risk of ∼2, suggesting that there is a heritable genetic component to EC predisposition.

1.3 Study of Heredity

The resemblance between parents and offspring had long been noticed before the formal study of heredity, and inheritance patterns were studied before knowledge of what was actually being transmitted down the generations that gave rise to similar traits. Here, I will highlight several key developments in the field of genetics and discuss the search for the genetic aetiology of EC in parallel.

1.3.1 Molecular basis of heredity

Deoxyribonucleic acid (DNA) was first isolated by Friedrich Miescher from surgical bandages in 1869 and named 'nuclein' because it was found in the cell nucleus (Dahm, 2008). This molecule was found to be ubiquitous in animal and plant cells. Phoebus Levene discovered the nucleotide phosphate composition of DNA (Levene, 1919). Despite this, the function and importance of DNA was unknown.

Frederick Griffith demonstrated the 'transforming principle', where the heat-treated virulent (smooth strain) *Streptococcus pneumoniae* could mediate its effect via non-virulent (rough strain) colonies (Griffith, 1928; Lorenz and Wackernagel, 1994). In 1944, the Avery–MacLeod–McCarty experiment formally demonstrated that it was

DNA that was responsible for the bacterial transformation observed in Griffith's experiments (Avery *et al.*, 1944). Prior to this, proteins were thought to be responsible for carrying genetic information.

Based on the X-ray diffraction images by Rosalind Franklin and Raymond Gosling and knowledge of complimentary base pairing, James Watson and Francis Crick presented the double-helix model of DNA structure in 1953 (Watson and Crick, 1953). This eloquent description of DNA structure was groundbreaking in that the anti-parallel strands and complimentary base pairing allowed for the semi-conservative replication of genetic material and established DNA as the central molecule of genetics.

DNA molecules consist of four nucleotide bases; adenine (A), thymine (T), cytosine (C) and guanine (G) and a sugar phosphate backbone. The double helical structure is such that A is always paired by its complimentary base T, and C with G. Francis Crick went on to present the 'central dogma of molecular biology' which sets outs the transfer of genetic information from DNA to RNA to protein (Crick, 1970). The use of the word 'dogma' has been controversial; there are now increasing examples of variation in information transfer in biological systems but this remains a helpful framework for understanding the basis of molecular biology.

1.3.2 Darwinian natural selection

In 1859, Charles Darwin published *Origin of Species* where he postulated that modern species descended from common ancestors and that the process of natural selection guided evolutionary change. However, the mechanisms of inheritance were unknown at the time, and Darwin struggled to explain how phenotypic variation arises in a population to allow the process of natural selection.

1.3.3 Mendelian inheritance

Unbeknownst to Darwin, Gregor Mendel conducted detailed experiments about the inheritance patterns in pea plants in the 1860s. Mendel developed pure breeding lines of pea plants that produced round and wrinkled seeds. He found that crossing these plants would produce round seeds in all of the first daughter generation (F1). Crossing F1 plants, three quarters of second daughter generation (F2) plants would be round while a quarter would be wrinkled. Mendel explained these results with the laws of segregation and independent assortment, providing the theoretical basis for the existence of discrete hereditary factors, or genes. Prior to this, many scientists believed in 'blending' inheritance, where an offspring's phenotype would be intermediate between the two parents'.

1.4 Familial EC

In 1895, pathologist Aldred Scott Warthin noted the long history of cancer deaths (particularly of the colorectum, stomach and uterus) in a seamstress's family and he proceeded to document his findings and follow up members of the family. This family displayed autosomal dominant inheritance, and affected individuals sometimes developed cancers in multiple organs with early onset disease. By producing one of the first descriptions of the familial clustering of cancer, his work laid the foundations for the field of cancer genetics.

1.4.1 Lynch syndrome: background

In 1962, Henry Lynch encountered a patient with a similar family history to those described by Warthin, with multiple members affected by colorectal cancer and EC. More families were found to have autosomal dominant inheritance of cancer and these were referred to as displaying cancer family syndrome (Lynch and Krush, 1971). The

recognition of this clinical phenotype allowed the formulation of diagnosis and management guidelines for affected families. The term hereditary non-polyposis colorectal cancer (HNPCC) had also been used to differentiate this condition from familial adenomatous polyposis (FAP) but given the frequent extra-colonic manifestations, the preferred name for this cancer predisposition syndrome became Lynch syndrome (LS, Boland, 2005; Boland and Troncale, 1984).

LS is a autosomal dominant cancer predisposition syndrome that primarily affects the colon and endometrium. There is also an increased frequency of cancers of the stomach, small bowel, hepatobiliary system, ovary (Watson and Lynch, 1994), urologic tract, glioblastoma (Watson *et al.*, 2008), pancreas (Kastrinos *et al.*, 2009), breast (Win *et al.*, 2012), prostate (Bauer *et al.*, 2011) and adrenal cortex (Raymond *et al.*, 2013).

1.4.2 Linkage analysis for LS

In Mendel's law of independent assortment, two genes located on different chromosomes are inherited independently and the likelihood of observing two alleles together is a multiplication of the two allele frequencies. In the opposite scenario, two genetic variants can be located close to each other on the same chromosome. In the absence of recombination, these two variants are always inherited together such that the allele on one locus is linked, or can be inferred from the allele of another nearby locus.

Genetic linkage is the tendency of alleles that are close together to be inherited together. The closer the distance between two alleles, the less likely for recombination to occur between these two loci during meiosis. In 1980, Botstein *et al.* first proposed the use of restriction enzymes to create a genome-wide linkage map based on restriction fragment length polymorphisms (RFLPs, Botstein *et al.*, 1980). Using microsatellite linkage analysis for several large LS families, Peltomaki *et al.* was able to map the disease causing variant to chromosome 2p (Peltomäki *et al.*, 1993). Shortly

afterwards, Lindblom *et al.* were able to map a second LS locus to chromosome 3p by linkage analysis using RFLPs and microsatellite markers (Lindblom *et al.*, 1993).

The loci discovered by Peltomaki and Lindblom corresponded to *MSH2* and *MLH1* respectively. Not all families with LS showed linkage to these loci suggesting that there were other genomic locus causing LS. We now know that LS is caused by inactivating germline mutations in DNA mismatch repair (MMR) genes: *MLH1*, *MSH2*, *MSH6* and *PMS2*. MMR is a post-replicative proofreading and editing system that ensures genome integrity. Specifically, it repairs the single base pair and small insertion or deletion loops that can arise during polymerase replication of small repeat sequences (Hsieh and Yamane, 2008; Kunkel and Erie, 2005). DNA polymerase errors cause a 'nucleotide bulge' that persist in the absence of a proficient MMR system. In many cases, this results in the changes of CA dinucleotide repeats (Oki *et al.*, 1999). These errors are permanently fixed in subsequent replication cycles and thereby accumulate. Therefore, the MMR system acts as a tumour suppressor and its inactivation causes a high burden of somatic mutations across the genome (>12 mutations per 10^6 bases), known as the mutator phenotype.

1.4.3 Molecular and histological features of LS

Homozygous MMR mutations lead to a 100-fold reduction in MMR activity (Parsons *et al.*, 1993) but heterozygous cells display normal repair. In accordance with Knudson's two hit model of carcinogenesis (Knudson, 1985), LS tumours develop in individuals with heterozygous MMR mutations when the wild-type allele is knocked out somatically in the target tissue. Loss of heterozygosity can occur by deletion (Hemminki *et al.*, 1994), mutation (Liu *et al.*, 1995, 1994) or *MLH1* promoter methylation (Kane *et al.*, 1997).

The mutator phenotype causes variation in short tandem repeat sequences, or microsatellite instability (MSI), in LS-associated EC tumours. MSI is commonly

observed in EC tumours (present in up to 25%, Peltomäki and Vasen, 2004) and MSI in sporadic ECs is caused by *MLH1* promoter methylation (Simpkins *et al.*, 1999).

The histology of LS-associated EC tumours is diverse and can include endometrioid, clear cell, serous and undifferentiated subtypes. Other features associated with MSI include poor differentiation, mucinous features, signet ring cell differentiation, tumour cells growing in a medullary-type pattern, and inflammatory infiltrate at the tumor invading front or periphery (Hampel *et al.*, 2007).

In a series of 600 EC patients, MSI was present in 28% of EC tumours(Goodfellow, 2005). Patients with pathogenic germline MMR mutations have LS, but amongst those with no germline mutation detected, the MSI can be due to *MLH1* promoter methylation or two somatic mutations in MMR genes (Haraldsdottir *et al.*, 2014).

1.4.4 LS-associated EC

1.4.4.1 Life-time risk

Women with LS have a 30-70% lifetime risk of developing EC; this exceeds their risk for developing CRC (Lu *et al.*, 2005). LS patients with *MSH6* mutations have a higher risk of developing EC (up to 70%) than those with *MSH2* or *MLH1* mutations(up to 50%, Barrow *et al.*, 2013). The most commonly mutated gene in LS-associated EC is *MSH2*, accounting for 50-66% of these EC cases. The next most common are *MLH1* and *MSH6* which account for 24-40% and 10-13% respectively (Bonadona *et al.*, 2011). Germline deletions of the terminal end of the epithelial cell adhesion molecule *EPCAM*, located upstream of *MSH2*, results in the abnormal transcriptional elongation from *EPCAM* to *MSH2* and LS (Ligtenberg *et al.*, 2009). Women with LS caused by *EPCAM* deletions have a 12% risk of developing EC (Kempers *et al.*, 2011).

1.4.4.2 LS-associated and sporadic EC

Compared with sporadic cases of EC, the EC phenotype in women with LS has a younger average age of onset of 50 years (Aarnio *et al.*, 1995). Most cases of LS-associated EC present at an early stage and display endometrioid histology, though type 2 ECs are also seen (Carcangiu *et al.*, 2010). LS-associated non-endometrioid EC presents at a younger average age (47 years) compared with an average age of 65 in EC cases in the general population (Broaddus *et al.*, 2006; Garg *et al.*, 2009). Risk factors associated with unopposed oestrogen such as obesity and exogenous oestrogen use (section 1.1.2) are often absent in women with LS, where EC also does not arise in the background of endometrial hyperplasia. LS-associated EC tumours are also more likely to involve the low uterine segment (Westin *et al.*, 2008).

1.4.4.3 Clinical significance

LS accounts for 2-5% of all ECs (Kwon *et al.*, 2011; Lancaster *et al.*, 2007). Women with LS-associated EC have a 50% chance of being diagnosed with a second cancer within 15 years of the EC diagnosis (Lynch *et al.*, 1977; Mecklin and Järvinen, 1986). In 50% of women with LS, EC develops before CRC (Lu *et al.*, 2005) and thus the diagnosis of women with LS when presenting with EC is clinically significant for the management of their cancer risk. In terms of EC survival, there is no significant difference between sporadic cases and those which are associated with EC. The EC screening recommendations for women with LS are described in section 1.1.4.

LS is underdiagnosed amongst patients presenting with EC. It is estimated that 25-60% of LS patients would be missed based on the Amsterdam and Bethesda criteria (Hampel *et al.*, 2007; Resnick *et al.*, 2009). To address this issue, there are algorithms that advocate the universal screening of LS in EC patients (Resnick *et al.*, 2009; Weissman *et al.*, 2012).

The progression of LS research illustrates how the genetic cause of a rare, high-

penetrance Mendelian form of EC can be discovered. This leads to an improved understanding of the disease process and to progress in the diagnosis and clinical management of patients.

1.4.5 DNA polymerases

As with LS, germline exonuclease domain mutations (EDMs) were first noted in CRC pedigrees and were subsequently associated with a greatly increased risk of EC (Palles *et al.*, 2013). *POLE* and *POLD1* encode for catalytic subunits of the polymerases that act on the leading and lagging strand of the DNA replication fork respectively. The exonuclease (proof-reading) domain of these polymerases increases replication fidelity by recognizing and excising mispaired bases (Goldsby *et al.*, 2001; Simon *et al.*, 1991). The proof-reading function increases replication fidelity by $\sim$ 100 fold. Pathogenic EDMs cause a high-penetrance CRC and EC syndrome, with tumours that are microsatellite stable (MSS), but display an 'ultramutator phenotype' with over one million base substitutions per tumour. Germline *POLD1* EDMs including p.Ser478Asn predispose to early-onset EC; somatic missense *POLE* EDMs have been observed in up to 7% of ECs (Church *et al.*, 2013). *POLE*-mutatnt ECs are associated with good prognosis (Cancer Genome Atlas Research Network *et al.*, 2013).

The role of DNA polymerases in EC is a recent discovery and continuing research is necessary to delineate the incidence of germline polymerase mutations in EC and the clinical and pathological understanding of the condition. The similarities between MMR and DNA polymerase-associated EC and CRC prompted the search for genetic variants that predispose to both EC and CRC in chapter 6.

1.4.6 Cowden syndrome

Cowden syndrome, also known as multiple hamartoma syndrome, is a rare autosomal dominant disorder caused by an inactivating mutation in the tumour suppressor gene,

phosphatase and tensin homolog (*PTEN*). The subsequent overactivity of the mTOR pathway causes cell growth and proliferation resulting in the presence of hamartomatous tumours in multiple organ systems. There is also an increased risk of EC and tumours in the breast, skin, thyroid, brain and gastrointestinal tract (Tan *et al.*, 2012). Instead of presenting with EC, the diagnosis of Cowden syndrome is often made based on the incidental finding of mucocutaneous lesions. Women with Cowden syndrome have a 10-28% risk of developing EC (Blumenthal and Dennis, 2008; Tan *et al.*, 2012), although *PTEN* is not routinely tested in EC patients because of the syndrome's rarity.

1.4.7 Beyond familial studies

Linkage analysis has been successful in mapping causal genetic variants for monogenic phenotypes like LS that are highly penetrant and show a Mendelian pattern of inheritance. The Online Mendelian Inheritance in Man database (www.omim.org) catalogues known genetic diseases and links them to relevant genes in the human genome (Hamosh *et al.*, 2005). Linkage analysis has been less successful in identifying genetic variants for heritable diseases that are not inherited in a Mendelian pattern. Human height is an example of a quantitative phenotype that is highly heritable but does not follow a simple Mendelian pattern of inheritance.

R.A. Fisher suggested that quantitative traits are affected by a large number of Mendelian factors, each contributing a small amout to the phenotype (Fisher, 1918). If a large number of genes, each with a comparable, small effect contribute in an additive manner, the larger the number of contributing genes, the more a quantitative trait will approximate the normal distribution in a population. This can be extended to binary disease traits: the development of cancer could be a result of additive genetic risk factors that surpass a susceptibility threshold to put a particular individual at an increased risk. Many common diseases in humans including sporadic EC are probably

complex, polygenic phenotypes and other strategies are required in order to search for associated genetic variants.

1.5 Population-based studies

Instead of looking at pedigrees and related individuals, geneticists turned to unrelated individuals in order to search for genetic associations. This required several developments in the field of population genetics.

Population genetics is the study of the genetic composition of a population based on the distributions and changes of allele frequencies over time. This is subject to evolutionary processes such as natural selection, genetic drift, mutation and gene flow. The emergence of population genetics is based on the work of R.A Fisher, Sewall Wright and J.B.S. Haldane, which integrated the principles of Darwinian natural selection and Mendelian genetics.

1.5.1 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE, Hardy, 1908) is one of the most important principles in population genetics, and it is an extension of the ideas proposed by Darwin and Mendel. HWE is crucial to understanding of the nature of allele frequencies in populations and it also allows the modelling of evolutionary forces.

HWE describes the relationship between genotype and allele frequency in a sexually reproducing, diploid population. This relationship is theorized to remain stable from generation to generation in a large population undergoing random mating, in the absence of evolutionary forces. HWE considers a single biallelic autosomal locus where p and q represent the minor and major allele frequencies and $p + q = 1$. The expected genotype frequencies under HWE is:

$$p^2 + 2pq + q^2 = 1 \tag{1.1}$$

The expected genotype frequencies of the rare homozygote, heterozygote and common homozygote are p^2, $2pq$ and q^2 respectively. In a large population with random mating, the situation is akin to Mendel's pea plant experiment (section 1.3.3) and Hardy-Weinberg proportions are achieved in the first daughter generation (F1). If generations are non-overlapping, and parents died after mating, then HWE would be achieved after one round of mating, regardless of what the initial genotypic proportions were. In overlapping generations, several rounds of mating are required for Hardy-Weinberg proportions to occur and crucially, this proportion will remain stable (or in equilibrium) in subsequent generations in the absence of evolutionary forces.

HWE resolved some of the unsettled questions in genetics at the time. For example, some of Darwin's contemporaries questioned how population diversity could occur to allow natural selection, if sexual reproduction reduces genetic variation by 'blending' inheritance. Geneticists in the early 1900s questioned how recessive traits persisted in populations. HWE shows how allele frequencies can remain stable across multiple generations under the assumptions of Mendelian inheritance.

1.5.2 EC candidate gene studies

Candidate gene studies aim to quantify and compare the allele frequencies of polymorphisms in cases and controls in the general population. Before the advent of high-throughput genotyping, genome-wide coverage was impossible and investigators had to choose to look at particular regions of interest. The choice of variants and candicate genes was based on *a priori* biological knowledge and assumptions of causality between gene function and phenotype. In an ideal situation, allelic variation at a functional locus would cause marked change in a phenotype of interest detectable in

a case-control study. As discussed in section 2.3.2, HWE can act as a guide to expected genotype counts and allele frequencies in a population. Deviation from HWE could suggest sample selection bias, technical errors, or the presence of non-random mating or evolutionary forces at work at a particular locus.

Various candidate gene studies have been conducted to assess predisposition to EC. Known EC risk factors (section 1.1.2) suggest that oestrogen exposure plays an important role in EC pathogenesis, particularly for type 1 endometrioid histology tumours. Some candidate studies have concentrated on genes in oestrogen-related pathways. *CYP19A1* encodes for aromatase, which converts circulating androgens to oestradial. Two *CYP19A1* single nucleotide polymorphisms (SNPs, rs749292 and rs727479) were demonstrated to be associated with a 10-20% increase in circulating oestrogen levels in post-menopausal women (Haiman *et al.*, 2007). The alleles associated with higher circulating oestrogen levels were also associated with increased risk of EC (P=7.1x10^{-7} and 0.09 respectively, Setiawan *et al.*, 2009).

ESR1 encodes for oestrogen receptor 1, which is the predominant receptor mediating the effects of oestrogen in the endometrium. *ESR1* polymorphisms have been associated with breast cancer risk, particularly in oestrogen receptor-positive cases (Dunning *et al.*, 2009; Turnbull *et al.*, 2010a). Several studies have investigated the association between *ESR1* polymorphims and EC risk but these studies were underpowered and have produced inconsistent results (Einarsdóttir *et al.*, 2009; Wang *et al.*, 2012; Wedrén *et al.*, 2008; Weiderpass *et al.*, 2000).

Given the role of MMR genes in LS, other DNA repair genes have been investigated. These include *ERCC1* (Samulak *et al.*, 2011; Sobczuk *et al.*, 2012), *XPA*, *XPC* (Doherty *et al.*, 2011) and *RAD51* (Romanowicz-Makowska *et al.*, 2012), all of which are part of the nucleotide excision repair or base excision repair systems. Other studies have assessed variants in genes associated with cell-cycle control (Cai *et al.*, 2011), *TP53* (Gu *et al.*, 2011; Tang *et al.*, 2012) and obesity (Chen *et al.*, 2012).

Many of these candidate studies have a very small sample size (less than 100), and different studies have produced inconsistent results. Candidate gene studies of other phenotypes have been blightd a similar trend where many associations are not replicated in subsequent studies (Tabor *et al.*, 2002). This could be due to many reasons including lack of statistical power, and population stratification, which cannot be controlled for in candidate gene studies of different populations.

Perhaps the biggest limiting factor in candidate gene studies is that they rely heavily on the *a priori* knowledge of association between the target gene and disease process. For many complex traits, there is limited information about the biological and genetic basis of disease. Even if an investigator correctly selects a phenotype-related gene, it remains a challenge to choose a potential causal variant to genotype. Candidate genes studies attempt to identify variants using direct association where the candidate causal variant is directly genotyped in cases and controls. This is in contrast to indirect association studies, which makes use of the non-random relationship between alleles in different loci across the genome (linkage disequilibrium). In an indirect association study, a neargby variant which is inherited together with the causal variant may be used to tag nearby variants.

The *CYP19A1* study is an example of a candidate gene study that is well-powered (4,998 cases and 8,285 controls), and has a compelling biological link between the gene polymorphism and EC phenotype.

1.5.3 Human Genome Project

Beyond assessing genetic variants in a piecemeal manner, knowledge the entire human genetic sequence would allow scientists to study the genetic landscape and search for associations across the entire genome.

The Human Genome Project was an international collaborative project that aimed to determine the complete base sequence of the human genome. The planning began

in the 1980s, the first working draft was announced in 2000 (Lander *et al.*, 2001) and the project was declared complete in 2003 (build 35, Consortium, 2004). DNA was extracted from several anonymous donors of European descent and the euchromatic haploid genome consisting of about 3 billion base pairs was separated into bacterial artificial chromosomes and sequenced.

In the fifty years between the discovery of DNA structure and the first copy of the reference human genome, remarkable advances in molecular biology such as DNA sequencing (Sanger and Coulson, 1975) and polymerase chain reaction (PCR, Mullis and Faloona, 1987) paved the way for the Human Genome Project. The reference human genome provides invaluable insight into the genomic landscape, including the number and location of genes. There are continual updates providing increasing accuracy and fewer gaps. Furthermore, it enables further study of genomic function in terms of human genetic variation, disease states and comparison with other species.

1.5.4 Cataloguing human genetic variation

The International Haplotype Map (HapMap) Project was commenced in 2002 and it aimed to catalogue human genetic variation in populations of different ancestries (International HapMap Consortium, 2003). Building on the Human Genome Project, a catalogue of genetic variation led to a better understanding of population structure. The HapMap Project was also set up with the express aim of finding genetic variants related to common diseases.

As discussed in section 1.4.2, genetic linkage is the tendency of alleles located close together on a chromosome to be inherited together. The frequency at which two alleles on any chromosome are observed together varies between always (perfect linkage disequilibrium, LD) and the frequency expected based on chance (linkage equilibrium). A haplotype is a set of variants, or combination of alleles, that tend to be inherited together.

The International HapMap Consortium used various genotyping techniques in order to construct a genome-wide haplotype map using DNA from four different populations: 30 Yoruba people of Ibadan, Nigeria, 44 people from Tokyo, Japan, 45 Han Chinese people from Beijing, China and 30 people with Northern and Western European ancestry from Utah, USA.

SNPs are a common form of genetic variation that can be reliably ascertained. There are two alleles in the vast majority of SNPs since a triallelic SNP would require an additional mutational event. The HapMap Project has greatly increased the number of known SNPs to several million; most of these lie in non-coding regions. SNPs in coding regions can be separated in to non-synonymous (missense) variants that affect amino acid sequence and synonymous SNPs which do not.

The International HapMap Project allowed population geneticists to assess LD structure across the genome and study the impact of evolutionary forces in different populations (McVean *et al.*, 2005).

Subsequent phases of the International Hapmap Project have increased the number of known genetic variants (International HapMap 3 Consortium *et al.*, 2010). The 1000 Genomes Project continues in the same vein with focus on rarer genetic variants (1000 Genomes Project Consortium *et al.*, 2010). The 1000 Genomes Project aims to bridge the gap between rare, high-penetrance genetic variants and common genetic variants. This is achieved by sequencing 1,092 individuals from diverse ancestries and it is hoped that this will lead to the discovery of more than 95% of SNPs, copy number variants (CNVs) and indels with minor allele frequencies (MAF) as low as 1% across the genome, and 0.1% in gene regions.

1.6 Genome-wide association studies

1.6.1 Background

The *common disease-common variant* hypothesis suggests that common germline genetic variants can explain at least part of the inherited predisposition to common disease phenotypes such as sporadic EC. As mentioned before, complex phenotypes are thought to be mediated through variants in multiple loci, each contributing a modest increase in risk and interacting to produce the phenotype. This concurs with the observation that the genetic basis of common diseases have been largely resistant to previous efforts based on techniques such as positional cloning and linkage studies.

Commercial tag-SNP arrays allow high-density genotyping of a several hundred thousand to several million polymorphic SNPs across the whole genome. The selection of tag-SNPs on arrays was aided by knowledge of LD structure from the International HapMap Project (section 1.5.4). The genotype of a single tag-SNP can represent neighbouring SNPs in the same haplotype block and a non-redundant selection of tag-SNPs can successfully interrogate a large proportion of the common genetic variation across the genome. Though causal variants may not be genotyped directly in SNP arrays, tag-SNPs are in LD with surrounding SNPs so that they act as proxy markers to the nearby causal variants. Instead of occurring randomly and uniformly across the genome, human recombination is concentrated into short (1–2 kb) hotspots that occur every 100–200 kb (Jeffreys *et al.*, 2001; McVean *et al.*, 2004). This further aids the design of SNP arrays and renders association studies feasible.

Genome-wide association studies (GWAS) have been successful at identifying common genetic variants associated with a modest increase in risk for complex traits. Much like how the careful selection and genotyping individuals of different ancestries in the HapMap project can represent the allele frequencies in different populations (section 1.5.4), the genotyping of large numbers of unrelated individuals with a par-

ticular phenotype can yield representative allele frequencies for this phenotypic population. Unrelated samples are used because multiple members of same family can skew results away from population allele frequencies.

GWAS are comparable to candidate gene studies in that it attempts to identify differences in allele frequencies between cases and controls in the general population. However, as the name suggests, GWAS allows investigators to look for associations across the whole genome and to do so without prior assumptions about biological function or where these associations may be located. The density of genotyping in SNP arrays also allows robust quality control methods and it is possible to discover variants in novel pathways not previously known to be biologically related to the phenotype.

The affordability of commercial tag-SNP arrays allows the genotyping of large numbers of individuals to assess genome-wide variation from germline DNA derived from peripheral blood lymphocytes. Thousands of blood samples can be collected in the clinic setting and matched phenotypic information can be gathered using questionnaires. The larger the number of individuals genotyped, the more accurately it represents the true population allele frequency at a particular locus. Binary traits such as the presence or absence of a disease can be assessed using a case-control study design. An allele at a particular genomic locus can be seen as the independent variable that influences the susceptibility or predisposition to developing a phenotype, the dependent variable. Comparison of allele frequencies between cases and controls can indentify SNPs where there is a marked difference between the two groups, and the allele is significantly associated with the phenotype.

1.6.2 Considerations

The comparison between cases and controls is most robust when populations are very similar apart from the case-control status. In the absence of confounders, differences

in the case-control status can be attributable to independent variable (genotype). As noted above (section 1.5.4), the allele frequency of some SNPs can vary greatly between different populations and this can give rise to population stratification. The incidence, clinical and pathological features of a phenotype can vary in populations with different ancestries. A particular disease phenotype may have different genetic aetiologies in different populations. There is some evidence for this being the case in EC from epidemiology studies in the US (section 1.1.1). All samples in the experiments described in this thesis were of European ancestry. Other GWAS have made use of large samples from different populations in order to discover and refine the genetic association signals (Al Olama *et al.*, 2014; Rosenberg *et al.*, 2010). Also, other factors that may influence phenotypic outcome can be accounted for by the use of co-variates in association testing.

Misclassification bias is where samples in a study are erroneously analysed as being cases or controls. EC is a relatively well-defined phenotype with a histological diagnosis and early presentation for the majority of cases. Some of the samples from population-based controls may have undiagnosed EC, or indeed endometrial hyperplasia and may go on to develop EC. However, the effects of misclassification bias is low unless the phenotype is very common in the general population (i.e. obesity, hypertension, McCarthy *et al.*, 2008). In terms of case selection, some studies have aimed to improve study power by enriching for more severe phenotypes. In the context of cancer, cases can be chosen with an earlier age of diagnosis, and more aggresive disease.

In a GWAS, association testing is done for each individual SNP and this means that there are up to millions of tests being done. The commonly used P-value cut-off of 0.05 rejects the null-hypothesis of no association as the observed result would occur by chance only 5% of the time. However, when a large number of tests are done, this P-value cutoff would lead to many false-positive associations. There are various

methods of taking multiple testing into account, one of the most conservative is the Bonferroni correction (Dunn, 1959), where the P-value threshold of significance is divided by the number of tests. For example, the P-value cutoff of 0.05 can be divided by 1 million (SNPs being tested) to yield the commonly cited P-value threshold for genome-wide significance: $P < 5x10^{-8}$. The aim of a stringent cutoff is to reduce the number of false-positive results and to increase the chance of an association being replicable.

Although affordable, the economic cost of high-density genotyping for a large number of samples remains sizeable, and studies are often designed so that there is an initial discovery phase with tag-SNP arrays. This is followed by one or more replication phases where a smaller number of SNPs with promising P-values are brought forward for genotyping in additional samples to see if the differing allele frequencies in cases and controls are replicated. The statistical power to detect associations can be calculated at the planning stages to aid the study design. Generally, a study has a high statistical power to detect associations if there is a large number of samples, and if the causal variant is common and has a large effect size.

1.6.3 Examples

The first successful GWAS was for age-related macular degeneration. 116,204 SNPs were genotyped in 96 cases and 50 controls, and two associated SNPs were found (Klein *et al.*, 2005). Since then, there have been a large number of studies for different phenotypes. The Wellcome Trust Case Control Consortium (WTCCC) study in 2007 with 14,000 cases and 30,00 controls, was an important proof of concept and elucidated the methodology in identifying novel variants while reducing bias and confounding factors (Wellcome Trust Case Control Consortium, 2007). WTCCC phase 2 controls have become publically available for use in GWAS of different phenotypes and these were included in this EC study.

GWAS of other sporadic incidences of cancer, such as CRC have yielded many associations across the genome (Houlston *et al.*, 2010; Tomlinson *et al.*, 2008). Some of these associations lie in intergenic regions and further fine-mapping and genotyping have identified several independent associations near the bone morphogenetic protein signalling pathway (Tomlinson *et al.*, 2011). Intergenic GWAS signals are thought to exert their effects on phenotype by regulating the expression of nearby genes. One CRC-associated SNP, rs16969681 lies 16.8kb upstream of *GREM1* and within a *GREM1* enhancer. The minor allele increases the enhancer activity, and elevated *GREM1* expression is associated with the development of colorectal tumours (Lewis *et al.*, 2014). Chapter 5 of this thesis describes the effort to characterize the regulatory potential of an intergenic EC-associated SNP using similar methods.

As of Februrary 2015, the National Human Genome Research Institute GWAS catalog includes 2,111 publications and 15,396 associated SNPs for hundreds of phenotypes.

The reduction in cost of genotyping arrays and the established methodology has lead to a wide range of phenotypes being studied. It could be argued that even if there is only a small heritable component to a particular phenotype, a GWAS with a large sample size may provide the statistical power to detect significant associations. The converse may be true in that a GWAS may not identify any associations, or may identify associations that explain only a small part of the variance for a highly heritable trait.

Figure 1.2 shows the relationship between allele frequency and effect size for selected phenotypes (Bush and Moore, 2012). There is a general negative correlation between effect size and allele frequency in that variants with large effect sizes are usually rare. GWAS have been successful in identifying variants in the right bottom corner of the of the diagram, which are common and confer only a modest increase in risk. Although many GWAS SNPs have a small effect size, they still provide an

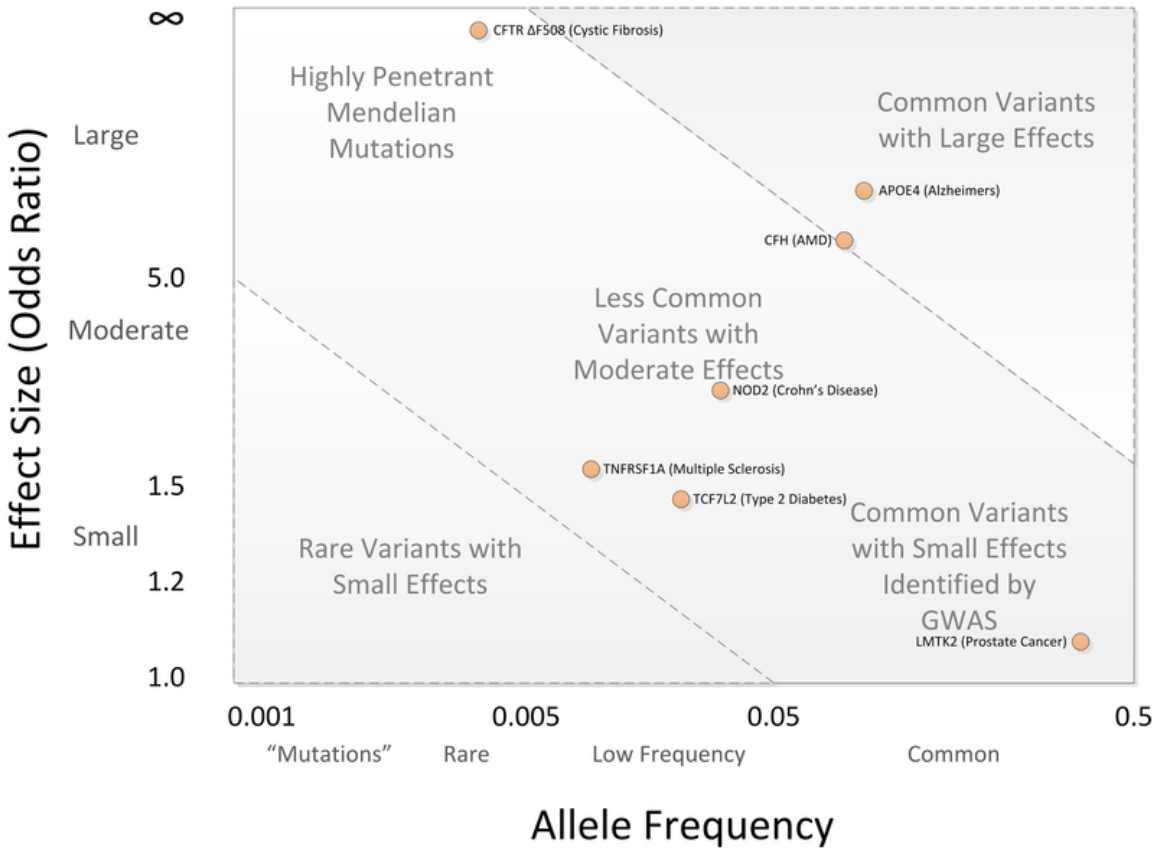

Figure 1.2: Effect size and allele frequency for genetic variants associated with different phenotypes. X-axis is allele frequency, Y-axis is effect size in terms of odds ratio (Bush and Moore, 2012).

improved understanding of genetic aetiology in that associations have often implicated genes or pathways that were not previously suspected to be involved with the phenotype. Chapter 3 describes an EC GWAS and chapter 4 is the meta-analysis with two other EC GWAS and two replication phases.

In a GWAS with 1,265 EC cases in the discovery phase and genotyping in an additional 3,957 cases and 6,886 controls found a single tag-SNP associated with EC at genome-wide significance (Spurdle *et al.*, 2011). This SNP is located in the second intron of *HNF1B* and has been associated with other phenotypes including type 2 diabetes.

1.7 Imputation

Knowledge of the genome-wide haplotype structure allowed the development of SNP arrays that capture the majority of common genetic variation. Thus, based on LD, SNP arrays capture information on the genetic variants across the genome beyond SNPs that are directly genotyped. The 1000 Genomes Project provides a more detailed catalogue of genetic variation, and genotypes from tag-SNP arrays can be phased and compared with the haplotypes of populations of a similar ancestry in 1000 Genomes. Imputation makes use of a fine-scale recombination map and a reference haplotype map (for example, from 1000 Genomes) in order to 'fill-in' missing genotypes in the study samples (Howie *et al.*, 2009; Marchini *et al.*, 2007).

These filled-in, or imputed genotypes are expressed in terms of probabilities of a SNP's genotype being homozygous for the reference, heterozygous or homozygous for the alternative allele. Many imputation softwares make use of hidden Markov models, incorporating information from neighbouring SNPs in order to determine the likelihood of observing a particular state or genotype at a particular locus. Imputation output also includes quality control parameters such as information scores. A region

or sample is most likely to be well-imputed if the haplotype of the reference panel highly resembles that of the study sample. European populations are noted to have larger LD blocks on average compared with African populations, and thus imputation is able to more accurately fill in a larger number of SNPs in Europeans (Shifman *et al.*, 2003). Often, different genotyping arrays are used for different phases of a GWAS, and imputation allows association testing of SNPs beyond the overlapping SNPs that are directly genotyped on both arrays. More SNPs can be assessed with the use of imputation, which allows fine-mapping of an association signal. In an ideal situation, an imputed SNP in partial LD with a genotyped SNP may display a higher significance and larger effect size than the genotyped SNP, making it a better candidate causal SNP. Imputation has been employed in the analyses described in chapters 4 and 6.

1.8 Thesis outline

EC is a significant cause of morbidity and mortality in women, with rising incidence rates since the 1990s. Epidemiological risk factors mainly related to the unopposed effects of oestrogen explains part of the increased risk seen in some women, but a familial relative risk of $\sim$2 suggests that there is a heritable component to EC risk. Rare, high-penetrance Mendelian syndromes such as LS can only explain a small proportion of the familial aggregation and the genetic aetiology of the majority of sporadic EC cases remains largely unknown.

This thesis explores the susceptibility to EC in women of European ancestry using population-based methods. Chapter 3 describes the National Study of Endometrial Cancer Genetics (NSECG) GWAS which is a discovery scan aimed at detecting novel common variants associated with EC. The sample size of the NSECG study means that it only has the power to detect common variants with a large effect size (OR>1.5) at high significance levels.

Chapter 4 is a EC meta-analysis of NSECG with two other GWAS (ANECS and SEARCH), and two replication phases. NSECG, ANECS and SEARCH are imputed to provide high-density genome-wide coverage for 2,212 EC cases and 9,544 controls. The first iCOGS replication phase consists of 4,330 cases and 26,849 controls genotyped on a custom array with $\sim$ 200,000 SNPs prioritized from previous cancer GWAS. The second replication phase consisted of 1195 cases and 795 non-overlapping controls genotyped for the top five SNP. All samples were of European ancestry. This meta-analysis discovered five novel EC risk loci.

The most significant of the novel EC risk loci lies on chromosome 13q22, downstream of *KLF5*. *KLF5* has been known to play a role in uterine development, homeostasis and tumourigenesis. Chapter 5 describes the characterization of the association signal at 13q22 using data from public databases and also *in vitro* experiments. This is done to assess the epigenetic landscape and potential regulatory mechanism at this locus.

Rare, high-penetrance germline mutations in MMR and DNA polymerase genes put individuals at a greatly increased risk of developing EC and CRC. This raises the possibility of shared genetic aetiology between EC and CRC for some cases. Chapter 6 describes the search for common variants associated with both cancer phenotypes.

The aims, overview and discussion of every section are provided in each results chapter. The final chapter summarizes the findings of this thesis and considers the direction of future development in this field.

Chapter 2

Materials and Methods

2.1 Population Cohorts

Population-based genetic studies require the collection of large numbers of unrelated samples with well-define phenotypes. Germline DNA is then extracted from peripheral blood lymphocytes and this is followed by accurate genotyping and statistical analysis. This section describes the study design in the case and control population cohorts used. All studies have the relevant institutional review board approval in each country in accordance with the principles embodied in the Declaration of Helsinki, and informed consent was obtained from all participants. A total of 7,737 cases and 37,144 controls were included in the meta-analysis. Cases and controls were matched based on geographical location and case-control clustering in principal components analysis (PCA).

2.1.1 National Study for Endometrial Cancer Genetics

The National Study for Endometrial Cancer Genetics (NSECG) consisted of women with the histological diagnosis of endometrial cancer were identified by collaborating clinicians throughout England from 2008. Each patient filled out a detailed ques-

tionnaire and provided blood for DNA extraction. All samples were of self-reported White European ancestry and 925 samples were successfully genotyped using Illumina 660W Quad SNP arrays over a 3-month period from June to August 2011 and all the genotyping was done by our scientific officer Angela Jones and myself, according to Illumina protocol. Of the 925 endometrial cancer samples, 87% had endometroid histology, while the others 13% is comprised of serous, clear cell and other subtypes.

In the endometrial cancer GWAS and meta-analysis described in Chapter 3 and 4, NSECG cases were matched with UK1-CORGI controls, Scotland 1 controls, and BC58 controls, ensuring that each control cohort was not used more than once. For the analysis with colorectal cancer in Chapter 6, NSECG cases were matched with CGEMS breast controls. Additional NSECG cases were used as part of the ECAC COGS initiative, where samples were from the same collection and processing scheme. These cases were matched with controls from the British Breast Cancer Study, Sheffield Breast Cancer Study and UK Breakthrough Generations Study. These control groups consisted of cancer-free women in the UK of White European ancestry recruited as participants of screening programmes with no breast lesions, and friends and relatives of participants of breast cancer studies.

2.1.2 Australian National Endometrial Cancer Study

The Australian National Endometrial Cancer Study (ANECS, Spurdle *et al.*, 2011) is an Australian population-based case-control family study of cancer of the uterine corpus with women aged 18-79 identified through hospitals and state-based cancer registries, all with endometrioid histology. Women were newly diagnosed with histologically confirmed primary endometrial cancer between July 2005 and December 2007, with a participation rate of 63%. Female controls, with no personal history of endometrial cancer or hysterectomy were recruited from two sources: The population-based Hunter Community Study (McEvoy *et al.*, 2010) and parents of participants in

an adolescent twin study (McGregor *et al.*, 1999). All participants were of White European ancestry and 606 cases and 3,083 controls were genotyped using the Illumina 610K arrays.

Additional ANECS cases were used as part of the ECAC COGS initiative, where samples were from the same collection and processing scheme. Further controls were obtained from the Austrailian Electoral Roll, with women matched to the age and geographic distribution of the cases, and also the Australian Red Cross Blood Service in Queensland.

2.1.3 Studies of Epidemiology and Risk factors in Cancer Heredity

Studies of Epidemiology and Risk factors in Cancer Heredity (SEARCH, Spurdle *et al.*, 2011) women with endometrial cancer aged 18-69 (average age diagnosis 58 years) ascertained through the Eastern Cancer Registration (www.ecric.org.uk). Information on tumor pathology characteristics was provided by the Eastern Cancer Registration and Information Centre and was derived from clinical pathology reports for all patients. Genotyping was done using Illumina 610K arrays. There was a total of 681 cases, all of endometrioid histology. These cases were matched with National Blood Service Controls from the Wellcome Trust Case Control Consortium described below. Additional SEARCH cases and controls were used as part of the ECAC COGS initiative. Controls were female, also between the ages of 18-69 at the time of recruitment and matched to cases in geographical profile. They had no history of cancer at the time of recruitment.

2.1.4 Wellcome Trust Case Control Consortium

The Wellcome Trust Case Control Consortium (WTCCC) (Wellcome Trust Case Control Consortium, 2007) is a collaboration between research groups across the UK that has pioneered the use of common controls in the study of different phenotypes. These controls are individuals from the UK of European ancestry and are genotyped on Illumina 1.2M arrays and are publicly available. Controls are from two sources: i) the National Blood Service (NBS) where samples were obtained form UK blood transfusions services repository of anonymised DNA samples, and ii) the 1958 Birth Cohort (BC58) which is from the National Child Development Study (Power and Elliott, 2006), an epidemiological survey based on all individuals born in England, Wales and Scotland during one week in 1958 (www.b58cgene.sgul.ac.uk/followup.php).

2.1.5 Endometrial Cancer Association Consortium

The Endometrial Cancer Association Consortium included 4,330 endometrial cancer cases collected from eight studies from seven countries: UK, USA, Belgium, Germany, Norway, Sweden and Australia, all of European ancestry. These samples were genotyped using a custom Illumina Infinium iSelect array with 211,155 SNPs designed by the Collaborative Oncological Gene-environment Study (COGS) initiative. The SNPs on this array were chosen based on regions of interest from previous breast, prostate, ovarian and endometrial cancer studies, rather than on genome-wide coverage (Eeles *et al.*, 2013; Michailidou *et al.*, 2013; Pharoah *et al.*, 2013; Sakoda *et al.*, 2013).

26,849 healthy female controls with European ancestry were selected from controls genotyped by the Breast Cancer Association Consortium (BCAC) and Ovarian Cancer Association Consortium (OCAC) part of the iCOGS project which were genotyped on the same arrays. These were matched with EC cases based on geographical location and PCA clustering.

The eight groups used in the iCOGS analysis were i) ANECS ii) SEARCH iii)

NSECG iv) Mayo Endometrial Cancer Study (MECS) v) Leuven Endometrial Cancer Study (LES) vi) Bavarian Endometrial Cancer Study (BECS) vii) Molecular Markers in Treatment of Endometrial Cancer (MoMaTEC) and viii) Cancer Hormone Replacement Epidemiology (CAHRES). All cases and controls are of European ancestry.

Descriptions of the NSECG, ANECS and SEARCH cohorts are as above.

The Mayo Endometrial Cancer Study (MECS) is a hospital based prospective biobank collection. The majority of patients seen at Mayo Clinic, Minnesota, USA, with primary endometrial cancer diagnosed at age 18 and older are enrolled and clinical data were abstracted from electronic medical records. The collection was started in 2006 and contains blood and fresh frozen tissue. Cancer-free women presenting for general medical examination were used as controls.

The Leuven Endometrial Study (LES) is a hospital based case-control study in Belgium. Endometrial cancer cases were identified by active surveillance of electronic patient files at the Leuven University Hospital, and consited of White women aged 27-80 years at diagnosis. Clinical data for endometrial cancer patients were recorded during interview at the time of diagnosis, and from pathology reports. All medical records were reviewed by trained abstractors and pathology reports compatible with primary, invasive, epithelial endometrial adenocarcinoma of all stages (I –IV) and of all grades.

The Bavarian Endometrial Cancer Cases and Controls Study (BECS) is a single-center case-control study, conducted between 2002 and 2008 in Germany, and aims to investigate genetic and epidemiological risk factors for endometrial cancer. Cases were incident cases referred to the University Hospital Erlangen, by surrounding practioners, or prevalent cases that were outpatients in follow-up care approached within 6 years after treatment for primary endometrial cancer in the same hospital. Epidemiological information for cases and controls was collected by a structured questionnaire, which was completed during an interview and clinical data for the cases was obtained

from clinical health records.

Molecular Markers in Treatment of Endometrial Cancer (MoMaTEC) cases were recruited from an unselected patient population primarily treated for endometrial carcinoma at Haukeland University Hospital, Bergen, Noway between 2001 and 2009. This is the referral hospital for Hordaland county. Clinical Information for cases including age, FIGO stage, histologic subtype, grade and prognosis was extracted from medical records.

The Cancer Hormone Replacement Epidemiology Study (CAHRES) is a population based case-control study was conducted among Swedish women aged 50-74 years, who were residing in Sweden between January 1st 1994 and December 31st 1995. Endometrial cancer cases were identified through the nation-wide cancer registries in Sweden and this study has also formerly referred to known as the Singapore and Sweden Breast/Endometrial Cancer Study (SASBAC) with details of the population selection process published previously (Weiderpass *et al.*, 1999). Histological specimens were reviewed and re-classified by the study pathologist and participants provided detailed questionnaire information.

2.1.6 Colorectal Cancer GWAS Cohorts

Controls used in the NSECG study were from colorectal cancer (CRC) GWAS where controls were cancer-free individuals from the same populations. This section describes the population cohorts in these CRC cancer GWAS and these CRC cases were also used in chapter 6, which describes the efforts to find SNPs the predispose to both endometrial cancer and CRC.

2.1.6.1 UK1-CORGI

UK Phase 1 (Tomlinson *et al.*, 2008) comprised of 888 cases with colorectal neoplasia ascertained through the Colorectal Tumour Gene Identification (CoRGI) consortium.

Cases had CRC at age 75 or below, most cases also had a family history of CRC. Recruitment was done via genetics clinics and hospitals in England and known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic *MUTYH* mutation carriers were excluded. The 998 controls were spouses or partners unaffected by cancer and without a personal family history (to second degree relative level) of colorectal neoplasia. These samples were genotyped using Illumina Hap550 SNP genotyping arrays by other members of the Tomlinson Group in 2007. UK1-CORGI controls was used in the NSECG GWAS in Chapter 3 as there was overlap in the collection centres for these samples and NSECG cases. UK1-CORGI cases and controls were used as a single group in the analysis in Chapter 6.

2.1.6.2 Scotland 1

Scotland Phase 1 (Tenesa *et al.*, 2008) included 973 CRC cases which were population-based incident case series aged <55 at diagnosis. Known dominant polyposis syndromes, HNPCC/Lynch syndrome or bi-allelic MUTYH mutation carriers were excluded. Control subjects were sampled from the Scottish population NHS registers, matched by age (±5 years), gender and area of residence within Scotland. Genotyping was done using Illumina HumanHap300 and Illumina HumanHap240S arrays. Of the 998 controls available, 392 which were genetically comparable to the English NSECG samples were used in the analysis in chapter 3. Scotland 1 cases and controls were used as a single group in the analysis in chapter 6.

2.1.6.3 VICTOR-QUASAR

VICTOR and QUASAR (VQ, COGENT Study *et al.*, 2008) comprised of 1,894 CRC cases (mean age of diagnosis 62.5 years) from the VICTOR and QUASAR2 trials. VICTOR was a randomised controlled trial for rofecoxib (VIOXX) in colorectal cancer patients after potentially curative therapy. QUASAR2 was a multicentre study of

capecitabine and bevacizumab as adjuvant CRC treatment. These samples were genotyped using Illumina HumanHap300, Illumina HumanHap270, Illumina Human 1.2MDuo arrays.

2.1.6.4 Colon Cancer Family Registry

The Colon Cancer Family Registry (CCFR Haile *et al.*, 1999) recruited newly diagnosed CRC cases reported to population cancer registries in the USA, Canada and Australia (Seattle, Ontario and Australasian Colorectal cancer registry). Population-based controls are used from the same populations and genotyping was done with Illumina Hap 550.

2.1.6.5 Cancer Genetic Markers of Susceptibility

Publicly-available controls of European ancestry were obtained from the Cancer Genetic Markers of Susceptibility studies (CGEMS) for breast cancer and prostate cancer. The CGEMS breast controls (Hunter *et al.*, 2007) were from the Nurses' Health Study which is a longitudinal study of women who provided blood samples between 1989 and 1990. These women were cancer-free at blood collection and were followed for incident disease until 2004. 1141 samples were genotyped using Illumina Hap550. CGEMS breast cancer controls were matched with NSECG EC cases in chapter 6.

CGEMS prostate controls (Yeager *et al.*, 2007) were from the Prostate, Lung, Colon and Ovarian (PLCO, www.cancer.gov/prevention/plco) Cancer Screening Trial which recruited 155,000 people aged 55 to 74 with no history of these cancers. The prostate study selected men with no history of prostate cancer from this group, and genotyping was done on Illumina Hap550 for 1,101 men. CGEMS prostate cancer controls were matched with CFR CRC cases in chapter 6.

2.2 Laboratory Methods

2.2.1 DNA extraction from peripheral blood lymphocytes

Peripheral venous blood samples were collected in EDTA tubes for all NSECG patients and stored at $-80\,^{\circ}$C. After thawing at room temperature, Maxwell ®16 (Promega UK Ltd) Blood DNA Purification kits were used according to manufacturer's instructions. Briefly, 420 μl of whole blood and 230 μl of elution buffer was pipetted into the cartridge for each sample and up to 16 samples could be processed in a 30 minute DNA extraction run. DNA was then quantified using spectrophotometry (section 2.2.3) and stored in a $4\,^{\circ}$C cold room.

2.2.2 Polymerase Chain Reaction (PCR) and gel electropheresis

Polymerase chain reactions (PCR) is one of the most important techniques and it is used to amplify a segment of DNA. PCR was performed for genotyping and sequencing. A typical 25 μl PCR reaction contained 30-50 ng of DNA, 0.4 pmoles of each primer, 1x Mg^{2+} free PCR buffer (Promega), 2 mM $MgCl_2$ (Promega), 0.2 mM of each dNTP (Amersham) and 1 unit of Taq DNA polymerase (Bioline). A standard PCR reaction consisted of an initial denaturation step at 95^{o}C for 5 min, 35 cycles of 94^{o}C for 1 min, 55^{o}C for 1 min and 72^{o}C for 1 min and a final extension step at 72^{o}C for 10 min. PCR reactions were carried out on a Tetrad PTC-225 thermocycler (MJ Research) and G-STORM GS4 thermal cycler. Genomic sequence and annotation data were downloaded from the Ensembl site (www.ensembl.org) and the University of California Santa Cruz genome site (genome.ucsc.edu). In all cases, PCR primers were designed using the Primer3 programme (frodo.wi.mit.edu/primer3). Stock primers were diluted to 100 μM, with working stocks diluted at 1:10.

PCR products were analysed using agarose gel electrophoresis. The agarose gels

were stained with ethidium bromide and were run alongside a 1 kb DNA ladder using Tris-Borate EDTA (TBE) running buffer.

2.2.3 Determination of Nucleic acid quantity and quality

A micro-volume ultraviolet (UV)-visible spectrophotometer (Nanodrop 2000, ThermoScientific) was used to determine the quantity and quality of DNA and RNA. The heterocyclic rings of nucleotides in nucleic acids absorb UV light and the wavelength of maximum absorption is 260nm. The instrument was first blanked using water or the corresponding elution buffer, then 1 μl of the sample was loaded onto an optical pedestal and the absorbed light from the sample was measured at 260 nm (A260). The appropriate constant of 50 and 40 for DNA and RNA respectively was used to obtain sample concentration. The absorption at 260 nm and 280 nm also yielded a ratio (A260/A280) which provided an estimation of purity with an ideal value of greater than or equal to 1.8. Compared to other forms of nucleic acid quantification, Nanodrop has the advantages of being quick to use and providing a measure for sample purity.

2.2.4 KASPar SNP Genotyping

KASPar (KBiosciences, Hertfordshire, UK) is a genotyping method based on competitive-allele specific PCR. Two forward primers ($\sim$20bp long) are designed such that they differ at the position of the SNP, and each allele is tagged by the sequence corresponding to VIC and FAM dyes. There is a common reverse primer targeting the sequence just downstream of the SNP. Primers are designed using PrimerPicker. Primers are made up to 100μM and used to make the assay mix with volumes as listed in table 2.1

The assay mix is then used to make up the assay mix for the 36-cycle PCR reaction. Genotyping is typically done on a 96-well plate with two negative controls

Table 2.1: KASPar assay mix. Primers at 100μM, volumes required for 100μl assay mix.

Primer	Volume (μl)
Allele1	12
Allele 2	12
Common primer	30
Water	46

(no DNA). The composition of master mix for a 96-well plate is listed in table 2.2. Each well contains 9μl master mix and 1 μl DNA at 10ng/μl.

Table 2.2: KASPar master mix

	Volume (μl)
Assay Mix	14.7
KASPar 2X reaction buffer	505.0
Water	405.3

After the PCR cycles are complete, the plates are analysed using the ABI 7900HT which detects the VIC and FAM flourescence in each well (figure 2.1). For SNPs with a MAF of more thn ∼10%, there are usually enough samples for the clustering to differentiate the reference homozygotes, heterozygotes and alternative homozygotes. For SNPs with a low MAF, DNA samples of known rare homozygotes are used to aid clustering.

2.2.5 DNA Microarray SNP Genotyping

The availability and affordability of commercial SNP arrays have been an important part of enabling GWAS. Researchers are able to interrogate hundreds of thousands, to millions of SNPs across the genome that capture the majority of common genetic variation. This is done for thousands of samples that provide the statistical power necessary for association studies. The design of many of these SNP arrays were based on SNP and LD information for European populations and thus are ideal for use on

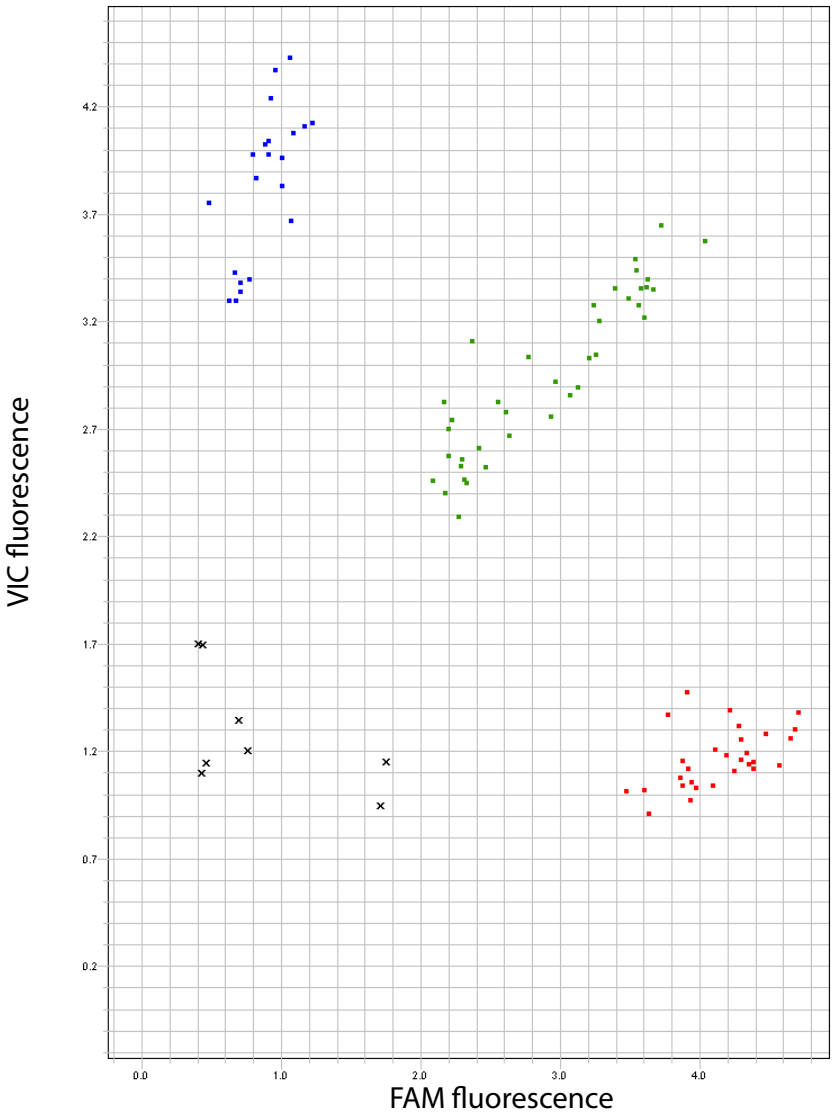

Figure 2.1: KASPar allelic discrimination plot. VIC and FAM fluoresence for each well is quantified after PCR. Reference homozygotes, heterozygotes and alternate homozygotes are in blue, green and red respectively. Black crosses represent no-DNA controls.

samples with European ancestry.

Illumina SNP arrays are based on beadchip technology that makes use of silica beads coated with oligo-nucleotide probes that target a specific locus. Complementary DNA sequences bind to the probes and there is a single-base extension of one of four labelled nucleotides. During the scanning process, a laser passes through the beadchip and the signal intensity for each allele, for each SNP is detected by the scanner.

The NSECG GWAS samples were genotyped using Illumina 660W Quad arrays according to manufacturer's protocols. Processed chips were read using the Illumina iScan and genotypye calls were made using the Illumina Genome Studio data analysis software and converted in to PLINK format for analysis. The genotyping was carried out by myself and Angela Jones.

2.2.6 Cell culture

Endometrial cancer cell lines were stored in -80^{o}C liquid nitrogen freezers and stocks were thawed in a 37^{o}C water bath. Cell lines were maintained in media as detailed in table 2.3 and grown in incubators set to 37^{o}C with 5% CO_2. Cell culture protocols were carried out in a class II laminar flow hood under sterile conditions and all cell lines were confirmed mycoplasma negative. Media was changed every 2-3 days and cells were passaged before they were confluent and cell lines were grown until they were 70-80% confluent. DNA, RNA were extracted from cell lines and used for *KLF5* quantification, chromatin immunoprecipitation and FAIRE. After use, cells were kept in freezing medium (90%FBS, 10%DMSO) and stored in -80^{o}C liquid nitrogen freezers.

2.2.6.1 Ishikawa and ECC-1 cells

Results from Ishikawa and ECC-1 cells are listed separately in publicly available ENCODE (ENCODE Project Consortium *et al.*, 2007) tier 3 data but STR-profiling

Table 2.3: Cell culture conditions. Table details medium used for each cell line. FBS, % foetal bovine serum; PS, % penicillin-streptomycin added to medium.

Cell line	Media	FBS	PS
AN3CA	DMEM	10	1
ARK-2	DMEM/F12	10	1
ECC-1	RPMI	10	1
EFE-184	RPMI	10	1
HEC-1B	DMEM:F12	10	1
HEC-116	MEM	10	1
Ishikawa	DMEM	10	1
MFE-296	40% RPMI 1640 + 40% DMEM	10	1
NOU1	RPMI 1640	10	1
RL95-2	DMEM:F12	10	1
SNG-II	DMEM	10	1

has shown that these two cell lines are very similar (Korch *et al.*, 2012). Based on the results presented by Korch et al., the International Cell Line Authentication Committee (ICLAC) recommended in the 2013 Database of Cross-Contaminated or Misidentified Cell Lines that ECC-1 be re-identified as Ishikawa cells. Our in vitro functional analysis for the 13q22 locus made use of our supply of Ishikawa cells for FAIRE and ChIP experiments and ECC-1 cells for luciferase reporter assays. Both cell lines displayed identical genotypes for rs9600103 and rs11841589, similar *KLF5* expression levels, and 20x sequencing using the Ion AmpliSeq Comprehensive Cancer Panel confirmed that variants in Ishikawa and ECC-1 are 90% concordant (based on 3,004 exonic SNVs in 409 cancer-related genes).

2.2.7 DNA and RNA extraction from cell lines

Cells were snap frozen with dry ice after centrifugation and DNA and RNA was extracted using DNeasy DNA extraction (Qiagen) and RNeasy minikit (Qiagen) according to manufacturer's instructions. Nucleic acids were then quantified using spectrophotometry (section 2.2.3).

2.2.8 RNA reverse transcribed into cDNA

RNA was first treated with DNase1 (Thermo Scientific) to remove contaminating genomic DNA and complementary DNA (cDNA) was reverse transcribed from RNA, using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems), according to the manufacturer's protocol. In brief, this process is comparable to PCR described earlier (section 2.2.2) but reverse transcriptase is used and random hexamer primers are used to bind non-specifically to all transcripts. cDNA was stored in $-20\,^{\circ}\mathrm{C}$

2.2.9 Quantitative real time PCR(qRT-PCR)

The absolute expression of *KLF5* was quantified using qRT-PCR using the ABI 7900HT cycler (Applied Biosystems) for each of the EC cell lines. The critical threshold was manually set at 0.2 and standard protocols were used to calculate relative expression with the ddCT method as described by (Livak and Schmittgen, 2001), using *GAPDH* as an endogenous control. TaqMan Gene Expression Assays were used for *KLF5* and *GAPDH* and these pre-prepared assays contain primer pairs, and probe that have been tested and optimized by Applied Biosystems. There is a FAM reporter dye that is attached to the 5' end of the probe and a non-fluorescent quencher is at the 3' end. A typical reaction contained 10 ng of purified cDNA, 0.5 μl of Taqman gene expression assay, 5 μl of TaqMan FAST Universal Mastermix and dH_20 to the total volume of 10 μl. Samples were run in duplicate or triplicate on the ABI 7900HT cycler under the following conditions: 95^oC for 20 s followed by 40 cycles of denaturation at 95^oC for 1 s and simultaneous annealing and extension at 60^oC for 20 s.

2.2.10 SYBR-Green RT-PCR

The SYBR-Green Dye I is used to detect PCR product by binding to double-strand DNA. A typical reaction contained 10 ng of purified cDNA, 10 μl of FAST SYBR-Green Mastermix, 300 nM of both forward and reverse primer and dH_20 to a total volume of 20 μl. The samples were run on the ABI PRISM 7900HT with the following conditions: 95^{o}C for 20 s followed by 40 cycles of denaturation at 95^{o}C for 1 s and simultaneous annealing and extension at 60^{o}C for 20 s. As the target sequence of DNA is amplified, the SYBR-Green dye I binds to the newly formed ds DNA and thus the fluorescence intensity is proportional to the amount of PCR product produced. To test the specificity of the primers and DNA sequence being amplified, I checked for non-specific PCR product formation by producing a disociation curve for each pair of primers. After the PCR as described above, samples were run with the following conditions: 95^{o}C for 15 s followed by 50^{o}C for 1 min and 95^{o}C for 15 s. The dissociation curve shows the changes in fluorescence plotted against time and a single narrow peak as seen in figure 2.2 suggests the presence of the specific, expected PCR product.

2.2.11 Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)

Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE) is a technique that allows the identification of nucleosome-depleted regions of the genome. This was used to assessed the epigenetic landscape of chromosome 13q22. The vast majority of the human genome is nucleosome bound and it is the genomic regions that are nuclosome depleted that are more likely to be transcriptionally active. Various experimental techniques are use to identify nucleosome depleted regions such as nuclueases. FAIRE makes use of reversible cross-linking to identify genomic regions that nucleosome

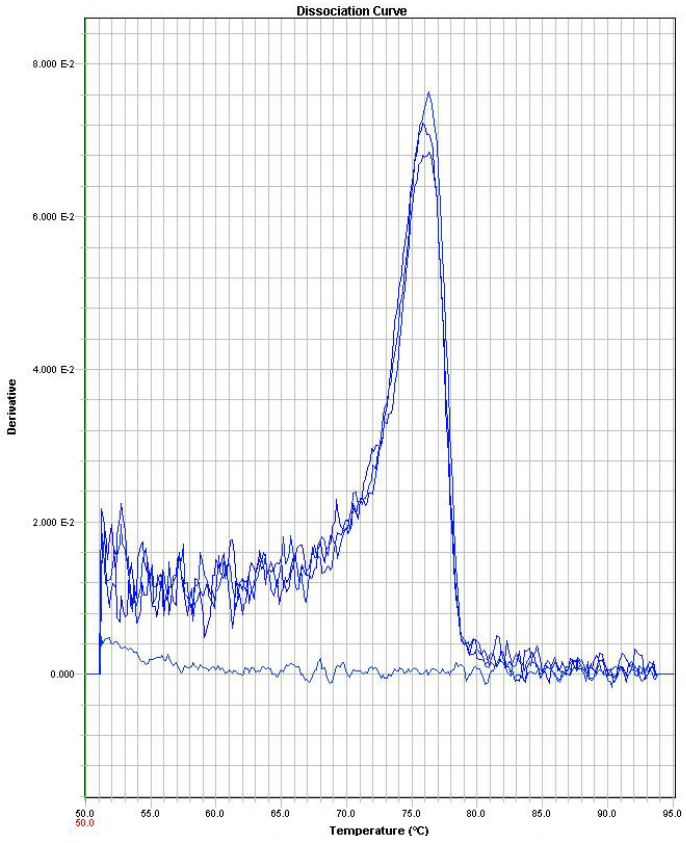

Figure 2.2: SYBR-Green RT-PCR dissociation curve with single discrete peak, in three technical replicates and a negative control.

bound and the aqueous phase in phenol-chloroform extraction allows for the isolation of nucleosome-free DNA. Quantitative PCR with primers specific to the intergenic 13q22 region are used to regions that are nucleosome-free compared to neighbouring regions.

The method described here was adapted from (Giresi *et al.*, 2007). Cross-linking was done on a rocker in room temperature with 1% formaldehyde (Sigma life sciences) was added to approximately 10^8 cells for 5 minutes, after which 115 mM glycine (Sigma life sciences) was added to inhibit the cross-linking. For each cell line, a non-crosslinked control was prepared in parallel for all of the remaining steps. Cells were then rinsed twice in 4 °C phosphate buffered saline solution (PBS), cell scrapers were used to detach cells adhered to the pretri dish surface and cells were transferred to a 15ml tube.

This was then put in the centrifuge and spun at 7000 g for 5 minutes, the supernatant was removed and the cell pellet was resuspended in successive lysis buffers: Lysis buffer I (50 mM HEPES-KOH, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% tritonX-100), lysis buffer II (10 mM tris-HCl, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA) and lysis buffer III (10 mM tris-HCl, 2100 mM NaCl, 1 mM EDTA, 0.1% sodium deoxycholate, 0.5%N-lauroylsarcosine). Cells are incubated on a rocker at 4 °C for 10 minutes in each lysis buffer after which they are spun down at 1300 g for 5 minutes so that the supernatant can be removed. The cells were then sonicated using the Bioruptor (Diagenode) in 7 to 15 30-second cycles to generate fragments that are 100 - 1000 bp in size. After sonication, the cells were centrifuged at 7000 g for 10 minutes, the supernatant was taken and gel electrophoresis in 1% agarose confirmed the size of the DNA fragments.

For phenolchloroform purification, Eppendorf phase-lock gel tubes were first centrifuged at 13000 rpm for 1 min. 300 μl of sonicated DNA product and 300 μl of phenol-chloroform (Sigma) is added to each phase-lock tube which is vortexed and

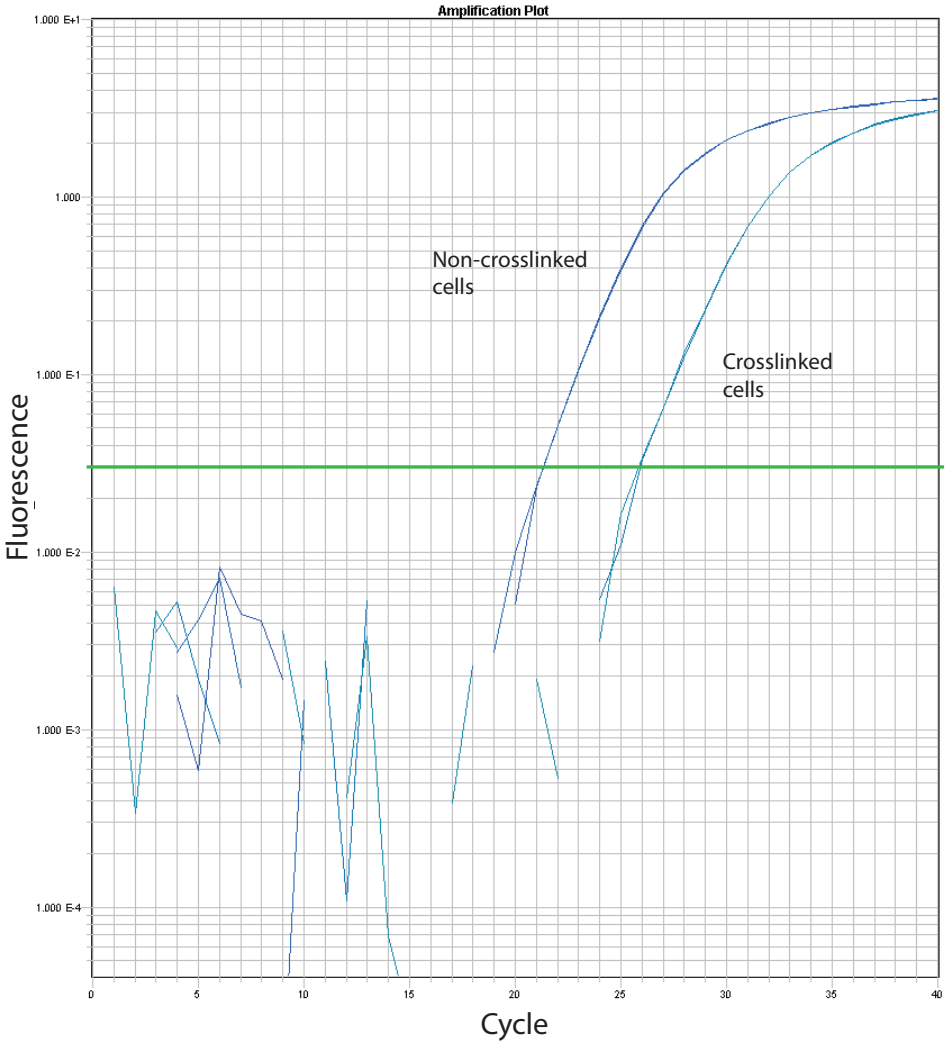

Figure 2.3: FAIRE qPCR aplification plot of SYBR-green fluorescence after successive PCR cycles for crosslinked and non-crosslinked cells. Critical threshold is marked by green horizontal line.

spun for 5 min at 13000 rpm. A further 300 μl of phenol-chloroform is added, followed by vortex and centrifugation as above, 300 μl of300 μl of chloroform is added with the vortex and centrifugation repeated. The aqueous phase was transferred to a fresh eppendorf and 1 μl glycogen (Roche), 270 mM sodium acetate and 66% ethanol were added in that order for each tube. Incubation occurred overnight, after which the DNA was spun at 15000 g for 30 minutes and washed with cold 70% ethanol. After another spin at 15000 g for 5 minutes, the supernatant was removed and the dried pellet was resuspended in 100 μl of elution buffer (Qiagen). The DNA was then quantified using spectrophotometry (section 2.2.3).

50ng of DNA from paired crosslinked and non-crosslinked cells was analyzed in duplicate by SYBR-green quantitative PCR (qPCR) (section 2.2.10, figure 2.3) using primers at roughly 1kb intervals in the 13q22 region downstream of *KLF5*. The ddCt method (Livak and Schmittgen, 2001) was used to normalize the results to the input DNA from the non-crosslinked cells to calculate the fold increase of enriched areas relative to the *RHO* promoter as a negative control. There were two technical and two biological replicates for each cell line.

2.2.12 Cross-linked chromatin immunoprecipitation (ChIP)

The epigenetic landscape of chromosome 17q22 was further assessed using chromatin immunoprecipitation (ChIP). ChIP investigates the interaction between DNA and proteins. The use of two antibodies that target histone marks linked with activation would lead to identification of genomic regions that are likely to be functionally active.

A 3-day protocol was conducted for cross-linked chromatin immunoprecipitation, which was performed on endometrial cancer cell lines grown to 70% confluence in 30 ml medium in 15 cm petri dishes. Cross-linking was done on a rocker at room temperature where 1% formaldehyde (Sigma life sciences) was added to the cells for 10 minutes. After the 10 minute period, 115 mM glycine (Sigma life sciences) was

added to quench the formaldehyde and stop the cross-linking reaction.

Cells were then rinsed twice in 4 °C phosphate buffered saline solution (PBS), cell scrapers were used to detach cells adhered to the pretri dish surface and cells were transferred to a 15ml tube. This was then put in the centrifuge and spun at 700 g for 5 minutes, the supernatant was removed and the cell pellet was resuspended in lysis buffer (1% sodium dodecyl sulfate (SDS) (Fluka Analytical), 10 mM EDTA (Ambion), 50mM Tris-HCl(Ambion)) and incubated for 10 minutes. The cells were then sonicated using the Bioruptor (diagenode) in 7 to 15 30-second cycles to generate fragments that are 1000 - 1500 bp in size. After sonication, the cells were centrifuged at 700 g for 10 minutes, the supernatant was taken and gel electrophoresis in 1% agarose confirmed the size of the DNA fragments. The fragmented DNA was then diluted ten times to the immuno-precipitation dilutaion buffer (1% tritonX-100 (Sigma life sciences), 2 nM EDTA (Ambion), 20 mM Tris-HCl (Ambion), 150 mM sodium chloride and each cell line was separated into four tubes: input chromatin, no-antibody-control and one tube for each antibody. 5 μl of anti-dimethyl-histone H3 Lys4 (H3K4Me2, Millipore 07-030) and anti-acetyl-histone H4 (H4Ac, Millipore 06-866) were added to the antibody tubes and along with the no-antibody-control, incubated overnight at 4 °C for immunoprecipitation. The input chromatin is kept refrigerated at 4 °C until the reverse cross-linking of day 2. Phenylmethylsulfonyl fluoride (PMSF, Sigma life sciences, 0.1%) and protease inhibitor is added to the lysis buffer and IP dilution buffer to deactivate proteases, while sodium butyrate (Sigma-Aldrich, 0.2%) is added to these solutions to inhibit histone deacetylases.

For day two of the protocol, 5 μl of protein A dynabeads (Novex 10001D) is added to each tube and incubated for 4 hours. A series of washes were done using Tris/Sucrose/EDTA (TSE) I (1% tritonX-100, 2 mM EDTA, 20 mM Tris-Hcl, 150 mMNaCl, 0.1 %SDS), TSE II (1% tritonX-100, 2 mM EDTA, 20 mM Tris-HCl, 500 mMNaCl, 0.1% SDS), Buffer III (0.25 M lithium chloride, 1 mM EDTA, 10 mM

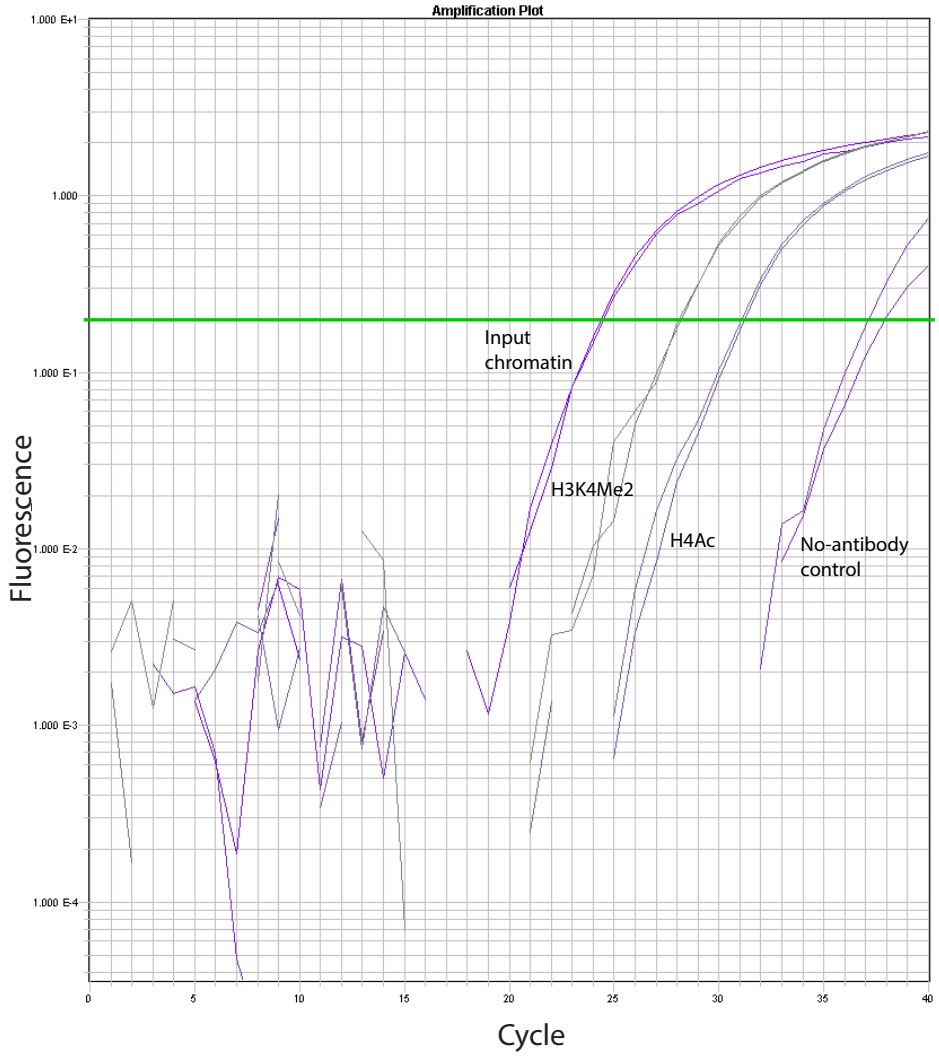

Figure 2.4: ChIP qPCR aplification plot of SYBR-green fluorescence after successive PCR cycles for input chromatin, anti-H3K4Me2, anti-H4Ac and no-antibody control. Critical threshold is marked by green horizontal line.

Tris-HCl, 1% tergitol-type NP-40, 1% sodium deoxycholate) and tris-EDTA (Tris-EDTA 1X). PMSF and sodium butyrate were added to all wash buffers in the same concentrations as above. The series of washes were as follows: TSE I once (on/off), TSE II four times (on/off, 5 minutes, 10 minutes twice), Buffer III once (10 minutes) and TE thrice (on/off, 5 minutes, 10 minutes). 1ml of each buffer was added for each wash and a MagnaRack magnetic separation rack (Invitrogen CS15000) was used to retain the dynabeads between each wash. 300 μl of extraction solution (1% SDS 0.1 M sodium bicarbonate) was added the dynabeads were removed after a 30 minute incubation. Then 0.7 M sodium chloride chloride was added and reverse cross-linking occurred overnight at 65 °C.

Day three involves purification of the DNA using QIAquick PCR purification kit (Qiagen) according to manufacturer's protocol. 1μl of DNA was analyzed in duplicate or triplicate by SYBR-green qPCR as above (section 2.2.11, figue 2.4). The ddCt method (Livak and Schmittgen, 2001) was used to normalize the results to the input chromatin to calculate the fold increase of enriched areas relative to the *RHO* promoter as a negative control. The no-antibody-control provided another quality control for the specificity of the antibodies used. There were two technical and two biological replicates for each cell line.

2.2.13 Luciferase reporter assays

Luciferase reporter assays are used to investigate the regulatory activity of a segment of DNA. For luciferase reporter assays, the regions chr:13 73,810,509-73,813,452 around rs9600103 and chr13:73,813,268-73,816,290 around rs11841589 were cloned into the pGL3-Promoter vector (Promega) to test for regulatory effects in Ishikawa/ECC-1 cells. These cells were selected because they express *KLF5*, showed evidence of a DHS, FAIRE and anti-H3K4Me2 enrichment at rs9600103 and were readily transfectable. Site-directed mutagenesis was used so that both the high- and low-risk

alleles of rs9600103 and rs11841589 were tested. After sequencing to verify the correct insert sequences, cells were transiently co-transfected using lipofectamine with the appropriate pGL3-Promoter constructs, and with the Renilla luciferase pGL4.75 vector (Promega) as a control for transfection efficiency.

After 48 hours, luciferase activity was measured (Dual-Glo Luciferase Assay System, Promega), and after subtracting background from lipofectamine-only controls, firefly luciferase activity from enhancer regions was normalized to the Renilla luciferase values for each sample. Levels of firefly luciferase activity were compared with a control plasmid consisting of an empty pGL3 and also a noncoding 2.2-kb stretch of plasmid sequence (taken from the pENTR1A plasmid, Invitrogen) cloned into the pGL3-Promoter vector that we had previously used as a length of DNA with no regulatory activity (Lewis *et al.*, 2014). Luciferase activity experiments had three or four replicates, each performed on three occasions (total of 11 assays).

2.3 Statistical Methods

2.3.1 Power calculations for genetic association studies

The power to detect a genetic association can be estimated by taking into account the sample size, allele frequencies and the effect size of the genetic variant. This forms an important part of planning in a GWAS and also for assessing if the allele frequencies and effect size of associations seen are expected and if the power was present to detect such an association given the sample size. Power can be calculated using a summarised non-centrality parameter in a non-central χ^2 distribution (Patnaik, 1949), which varies from zero to one.

The non-centrality parameter (λ) is calculated by equation 2.1:

$$\lambda = \frac{2r^2 N_{case} N_{ctrl} (F_{case} - F_{ctrl})^2}{NF(1-F)} \tag{2.1}$$

N is the number of cases and controls, F if the frequency of the risk allele, and r^2 is the LD between the marker of interest and the causal variant. Generally, a study is well-powered to detect associations at a level of 0.8 (de Bakker *et al.*, 2005; **?**). Power calculations in this thesis were performed using an R script written by Damjan Vukcevic.

2.3.2 Hardry-Weinberg Equilibrium

As described in section 1.5.1, HWE is an important principle in population genetics and it describes the expected relationship between genotype and allele frequencies in a population in the absence of non-random mating and evolutionary forces. For a single bi-allelic locus in a sexually reproducing diploid population, this can be summarized with the a simple quadratic equation (equation 1.1). Evolutionary forces are always present in real populations, although for the majority of SNPs across the genome that are not under selection, Hardy-Weinberg proportions are observed in genoytpes and allele frequencies.

In the context of genetic association studies, marked departures from the frequencies expected under HWE can be used to look for genotyping an other errors (Hosking *et al.*, 2004). The genotype frequency of genotyped SNPs can be compared to their expected frequencies under HWE with the Pearson's χ^2 test at 1 degree of freedom. Significant deviation from HWE was defined as P$< 1\text{x}10^{-6}$ in cases and controls for GWAS analysis.

2.3.3 Measures of linkage disequilibrium

Linkage disequilibrium (LD) is the non-random association of multiple genomic loci due to co-inheritance of from the same ancestral chromosome (section 1.5.4 Reich *et al.*, 2001). Two alleles which are always observed together are in perfect LD. Two alleles which are observed together as one would expect given the allele frequencies

are in linkage equilibrium. There are two common measures of LD: D' and r^2.

Considering two bi-allelic loci, A and B, with alleles $A1$, $A2$, $B1$ and $B2$. The frequency of $A1$ (P_{A1}) is $P_{A1B1} + P_{A1B2}$. The frequency of the $A1B1$ haplotype is P_{A1B1}. This is summarized in table 2.4:

Table 2.4: Allele and haplotype frequencies for loci A and B.

	$B1$	$B2$	Total
$A1$	P_{A1B1}	P_{A1B2}	P_{A1}
$A2$	P_{A2B1}	P_{A2B2}	P_{A2}
Total	P_{B1}	P_{B2}	

The difference between the observed haplotype frequency and the expected frequency by chance is D.

$$D = P_{A1B1} - P_{A1}P_{B1} \tag{2.2}$$

D can be scaled by dividing by the theoretical maximum D_{max} to yield D' (Lewontin, 1964). A D' value of -1 or 1 shows that there is no evidence of recombination between the two loci. D' values are known to fluctuate if there is a small number of samples and with rarer alleles.

The coefficient r^2 is the most commonly used in association studies and is defined as:

$$r^2 = \frac{D^2}{P_{A1B1}P_{A1B2}P_{A2B1}P_{A2B2}} \tag{2.3}$$

The r^2 coefficient is used throughout this thesis. r^2 can vary between 0 and 1. A value of 0 suggests that the two allele are independent while 1 suggests that there is complete dependency.

2.3.4 Cryptic relatedness

Cryptic relatedness describes kinship amongst cases and controls that is unknown to the investigator. This can be reduced by careful case-control selection and the availability of genome-wide SNP genotypes allows the assessment of related based on the genetic similarity between samples. Identity by state (IBS) is used to detect samples that are genetically more different than you would expect in a homogenous population. Identity by descent (IBD) identifies samples which are very genetically similar and can be calculated in a genome-wide manner by making use of IBS information. These calculations were done using PLINK (Purcell *et al.*, 2007).

$$\hat{\pi} = \frac{P(Z = 1)}{2} + P(Z = 2) \tag{2.4}$$

Equation 2.4 shows the calculation for $\hat{\pi}$ which is a measure of relatedness between pairs of samples. P(Z=1) and P(Z=2) are the probabilities of two samples sharing one and two alleles respectively. The expected probability of IBD states are expressed in terms of allele frequency and are averaged over all SNPs. A $\hat{\pi}$ score of 1 denotes that all alleles are shared IBD, and thus represent identical twins or more likely, dupicate samples. A $\hat{\pi}$ cutoff of 0.1875 was used which means that included samples were less related than second-degree relatives.

2.3.5 Principal components analysis

Principal component analysis (PCA) is a statistical method to extract the main sources of variation in a multi-variate dataset. SNPs from the GWAS datasets are pruned with a r^2 cutoff of 0.1 such that the remaining SNPs are independent markers across the whole genome. Certain regions of the genome which are known to display long-range LD is also excluded. 40,000 or more SNPs are used in the PCA by EIGENSTRAT software (Price *et al.*, 2006). The variance in the genotype ma-

trix undergoes orthogonal linear transformation such that 10 principal components (PCs) are calculated for each sample in the GWAS dataset. The greatest variance is accounted for by the first PC and then the second, and so forth. It is possible to visualize the results by plotting serial scatterplots of the first PC against the second, the second against the third and so on. There is a greater spread seen in PC plots of the first few principal components and by the 4th or 5th PC, all samples are part of a large indistinguishable central cluster.

There are many reasons that may explain differences between the genetic composition of samples. Samples with different population ancestries would cluster separately in PCA and thus PCA is an important method for control population stratification in GWAS. PCA can be conducted with HapMap or 1000 Genomes samples with different ancestries to see if the study population cluster with European populations as expected. Scatterplots of the first several PCs generally have a large central cluster and samples that deviate from the centroid in plots of the first 4 PCs are excluded. It is important that cases and controls overlap in the centroid in PCA scatterplots as this would suggest that there are no marked genetic differences across most of the genome. The presence of population stratification may lead to increased false positive results (type 1 error).

2.3.6 Pearson's χ^2 test

The genotype counts in cases and controls can be tested for association using the Pearson's χ^2 test. The Pearson's χ^2 test compares the distribution of observed values with that of expected values under the χ^2 distribution. Genotype counts in cases and controls can be put in a contingency table and the row and column totals are used to calculate expected genotype counts. The χ^2 test statistic summarizes the difference between the observed and expected genotypes (equation 2.5).

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} \tag{2.5}$$

O_i and E_i are the observed and expected values in each cell of the contingency table. Given the degrees of freedom, the level of significance can be represented using P-values. A low P-value would suggest marked differences between the observed and expected values and that the null-hypothesis of no association should be rejected. 3x2 contingency tables can be used to assess genotypes with two degrees of freedom. Alternatively, 2x2 contingency tables can be used for allelic, dominant or recessive models with one degree of freedom. Observed and expected values would differ greatly if HWE does not hold, and this produces a high χ^2 statistic.

Association testing in the NSECG GWAS (chapter 3) was done using PLINK software (Purcell *et al.*, 2007).

2.3.7 Logistic regression

In a case-control study, logistic regression measures the relationship between the binary phenotype and independent variable: genotype. Other independent variables can be included in the log-linear model. Factors such as age, family history or principal components can be included as covariates as the addition of this information may improve the correlation between genotype and phenotype. Covariates were not used in the association testing in chapters 4 and 6.

Association testing in chapters 4 and 6 was done using SNPTEST (Marchini and Howie, 2010) where the additive model was used. SNPTEST was also used to conduct step-wise conditional testing for the 1.5 regions containing genome-wide significant novel risk loci.

Association testing conditioning on the top SNP would mitigate the effects of this SNP and reveal the presence of any residual association signal that may be hidden.

SNPs that are in high LD with the top SNP will be less significant after conditioning for the top SNP, but other SNPs which are not in LD may remain, suggesting an independent association signal. Conditional testing can be done conditioning on the lead SNP, and the second most significant SNP, and so on, until no association signal remains.

2.3.8 Genomic inflation

Genomic inflation is the systemic overdispersion of the test statistic. This is seen by an unusually large number of SNPs that display some degree of significance in association testing. This is important in GWAS because of the large number of tests being done, and the vast majority of SNPs in the genome are not expected to be associated with the phenotype of interest. The genomic inflation factor (λ) is the ratio between the median of the observed test statistic distribution and the expected median. A λ of close to one suggests that there is no significant genomic inflation and generally, a λ of less than 1.05 is acceptable in GWAS. As the number of samples in a GWAS increases, reaching the tens of thousands and beyond, λ also increases. The genomic inflation factor is interpreted in conjunction with the quantile-quantile plot, which compares the observed and expected vales for each SNP and provides a visual representation of the same data.

2.3.9 Imputation

As discussed in section 2.3.9, SNPs that are not directly genotyped on a SNP array can be inferred with the aid of the haplotype structure of a reference panel, and a fine-scale recombination map. Imputation was done using IMPUTE2 (Howie *et al.*, 2009) which employs hidden Markov models in order to compute the probability of observing a particular genotype at a particular locus. An effective population size of 20,000 was used as per recomendations. The quality of the imputation can be

assessed using information scores and only SNPs with an info score of more than 0.9 are used for subsequent analysis.

Imputation is most accurate when the haplotype structure of the reference panel highly resembles that of the reference panel. For this reason, 1000 Genomes includes individuals from diverse ancestries and the appropriate haplotypes can be used for imputation. The Tomlinson group has access to high-coverage whole-genome sequences for 196 UK individuals and these were phased and used as an additional reference panel. Complete Genomics sequencing offers higher coverage (>30x) compared with 1000 Genomes (4-6x) and the addition of UK samples to the reference panel is beneficial because the majority of samples in the EC meta-analysis are from the UK. Other studies have also found increased imputation accuracy with the addition of additional sequencing as an additional reference panel (Timpson *et al.*, 2014).

2.3.10 Inverse variance fixed-effects meta-analysis

Meta-analysis is a method used to aggregate results from multiple studies in order to determine the overall effect. This is based on summary statistics from each study for each genotyped or well-imputed SNP: the effect size and standard error. If the effects from multiple studies concur, this means that the observation is made in a larger population. There is an increased statistical power to detect associations with larger sample sizes.

There are different methods for meta-analyses and a commonly used one is the inverse-variance method (Fleiss, 1993). Studies are combined using a weighted average and the weight of each study can be assigned based on variance, which reflects sample size. The weight of the i^{th} study (W_i) can be given as

$$W_i = \frac{1}{\delta_i^2} \tag{2.6}$$

where δ_i^2 is the variance of the i^{th} study. The combined effect of N studies (β) is

$$\beta = \frac{\sum_{i=1}^{N} \beta_i W_i}{\sum_{i=1}^{N} W_i} \tag{2.7}$$

where β_i is the effect size of the i^{th} study. The variance for the combined effect (δ^2) is

$$\delta^2 = \frac{1}{\sum_{i=1}^{N} W_i} \tag{2.8}$$

The standard error of the combined effect is the square-root of the variance (δ). The significance of the overall effect can be calculated using a χ^2 test.

A fixed effect meta-analysis assumes that all samples come from a homogenous population and that all studies share a common effect size. Thus the size of the study is the only factor that determines how closely it approximates the true effect size, with larger studies being more reliable.

In a scenario where there is significant heterogeneity, the random effect model allows for a distribution of different true effect sizes across different studies (DerSimonian and Laird, 1986). The combined effect of the random effect model represents the mean of the population of true effects. As well as weighting each study based on intra-study variance (like the fixed effect model), it also takes into account inter-study variance.

The fixed and random effects models vary in how much weight they assign to each study and this produces marked differences on the contribution of studies with small sample sizes or extreme effects. When there is little heterogeneity of effect sizes in a meta-analysis, both methods will produce comparable results. Thus, meta-analysis results should be interpreted in conjunction with measures of between-study heterogeneity (section 2.3.11).

Fixed effect meta-analysis was used for both chapters 4 and 6. For the EC meta-

analysis in chapter 4, the prevalence of EC is comparable in samples of European ancestry from the Europe, US and Australia, and similar study designs were employed. All samples used in the EC CRC meta-analysis were also of European ancestry and there was no specific hypothesis to support varying effects in EC and CRC. SNPs that show a comparable effect size in both phenotypes and all study groups would be easier to interpret. Meta-analysis was conducted for all genotype and well-imputed SNPs (info >0.9) using GWAMA (Mägi and Morris, 2010).

2.3.11 Measures of heterogeneity

In the context of genetic association studies, it is important to assess the heterogeneity in a meta-analysis because a highly significant result could be driven by a spurious extreme result by a single study. There are two common measures of heterogeneity in meta-analyses.

Cochran's statistic (Q) is defined as:

$$Q = \sum_{i=1}^{N} W_i(\beta - \beta_i)^2 \tag{2.9}$$

where W_i is the weight of the i^{th} study, β is the combined effect, and β^2 is the effect size of the i^{th} study. The χ^2 test can be used to obtain a P-value to assess the significance of the heterogeneity. The Cochran statistic is underpowered if there are very few studies and with small studies. An alternative is the I^2 heterogeneity metric which is defines as:

$$I^2 = \frac{Q - (N-1)}{Q} \tag{2.10}$$

where Q is the Cochran's statistic and N is the number of studies. I^2 measures the proportion of total variation across studies due to heterogeneity beyond chance, and can vary from 0-1. Generally, I^2 values of more than 0.5 signify large heterogeneity.

GWAMA calculates both these statistic and the I^2 metric has been used in this these because a proportion is easier to interpret compared with a cut-off score (Huedo-Medina *et al.*, 2006). The heterogeneity measures are also interpreted in conjunction with the forest plot, which is a helpful visual representation of the effect sizes and confidence intervals of different studies.

2.3.12 Functional annotation of associated variants

Associated variants identified in GWAS can be annotated using public databases that provide different measures of functional significance. Variants identified in GWAS are generally thought to be tag-SNPs and the causal variant is in LD with the tag-SNP. By annotating the most significant SNP and other variants that are in partial LD, one can investigate if these variants have been identified as potentially functional based on other experiments.

Haploreg concentrates on annotating non-coding regions by identifying nearby SNPs that are in LD based on 1000 Genomes data (Ward and Kellis, 2012). Annotation includes predicting chromatin state in different cell types, conservation across mammals and effects on regulatory motifs. RegulomeDB is a similar annotation software that makes use of manual annotations, computational predictions and high-throughput experimental data to derive a Regulome DB score that ranges from 1a to 6. 1a is the highest score in terms of regulatory potential and is characterized by correlation to expression data, transcription factor binding and DNAse hypersensitivity. These programs are helpful in assessing variants for regulatory potential and querying multiple publicly available databases allows one to look for different types of evidence that can guide further functional work.

Chapter 3

NSECG GWAS

3.1 Introduction

3.1.1 Background

As discussed in section 1.6, GWAS is a powerful method for identifying genetic variants associated with common phenotypes such as EC. When the genotyping and data analysis was conducted for this study, an EC GWAS (Spurdle *et al.*, 2011) with cases and controls from Australia and the UK had recently identified a single risk loci; rs4430796, a tag-SNP located in the intron of *HNF1B* on chromosome 17. Some of the samples from the National Study of the Genetics of Endometrial Cancer (NSECG) were previously used in the replication phases in the study by Spurdle *et al.* but were not genotyped using arrays in the discovery phase of that study. Thus the genotyping of NSECG samples using Illumina 660W Quad arrays can be seen as an independent discovery scan to identify additional variants associated with EC.

3.1.2 Aims and Overview

This chapter describes the effort to investigate the genetic susceptibility to sporadic EC in individuals of British Caucasian ancestry using 925 NSECG EC cases and

2,507 cancer-free controls. This GWAS identified one SNP with a P-value of less than 10^{-6}: rs4430796 (OR=0.74, CI%95 0.65–0.83, P=9.94x10−7. This replicates the result described by Spurdle *et al.* and further lends weight to the significance of the chromosome 17q12 *HNF1B* locus in EC predisposition. The second SNP of interest, rs9313548 (OR: 0.1.28, 95%CI 1.15 – 1.42, P=$7.22\text{x}10^{-6}$) is located at chromosome 5q35 near *FGF18*. Association testing with 801 cases of endometrioid histology yielded slightly more significant results for rs4430796 and rs9313548 (P=$2.90\text{x}10^{-7}$ and $3.47\text{x}10^{-6}$ respectively).

3.2 Population cohorts

The NSECG study consists of women with a histological diagnosis of EC identified by collaborating clinicians in England. Details of the study can be found in section 2.1 and 1,000 samples were genotyped using the Illumina 660WQuad arrays (section 2.2.5) of which 925 samples passed all quality control measures. These cases were analysed with 2,507 controls from the CORGI, Scotland phase 1 and WTCCC2 1958 Birth cohort (BC58) controls (see section 2.1). MSI status and genetic testing results were not available for all NSECG EC cases, but samples were excluded if they had a diagnosis of LS from pathology reports or clinical records.

Samples used in the NSECG GWAS analysis consisted of 925 NSECG EC cases and 2,507 cancer-free controls, 895 from the CORGI study, 391 from Scotland 1 and 1,221 from the BC58.

3.3 Power calculations for NSECG

The power to detect associations can be calculated given the number of cases and controls for a binary trait. The methodology used for calculations is described in section 2.3.1, and figure 3.1 shows the power to detect associations at different control

allele frequencies for the NSECG dataset.

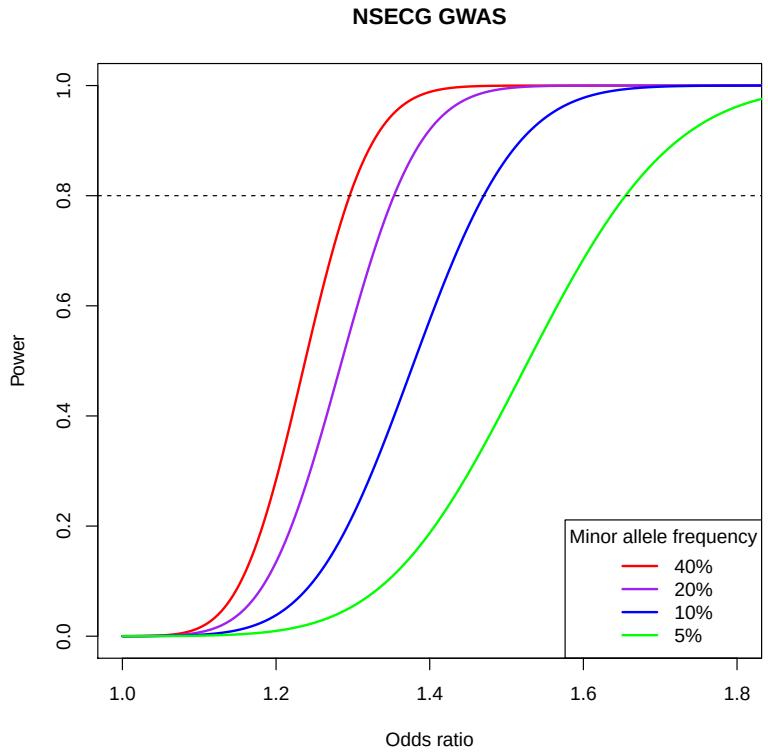

Figure 3.1: Graph showing the power to detect associations at P=10^{-4} in NSECG given different minor allele frequencies of controls.

This study has 57% power to detect associations at P=10^{-4} for a SNP with a MAF of 0.4 and OR of 1.25. Assuming the same allele frequency and effect size, this study only has 8.4% power to detect associations at genome-wide significance (P<5x10^{-8}).

3.4 GWAS quality control measures

Stringent quality control measures are required in GWAS in order to avoid the danger of spurious associations. There are strict parameters which should be met and deviation from this can suggest errors or biases anywhere in the pipeline from the collection of samples to the processing of samples and analysis. Quality control measures can

be broadly separated into those conducted in a per sample basis to those conducted in a per SNP basis. Generally, per sample quality control measures are conducted before per SNP quality control measures in order to first remove any samples that might have grossly failed genotyping.

3.4.1 Sample quality control

The following quality control measures assess the quality of the DNA samples used.

3.4.1.1 Concordance of duplicate samples

Duplicates of samples ENDO12G3 and ENDO12G6 were included. The duplicates were processed on different days and on a different array plate. All 4 samples had an individual call rate of more than 99% and comparing the genotype calls between duplicates, there was a concordance of more than 99.9%. This suggests a high accuracy in genotyping and consistency in the genotyping process with no observable difference in processing between different batches. The two samples with the higher call rate were kept for subsequent analysis.

3.4.1.2 Missingness and heterozygosity

Samples with more than 5% missing genotypes were excluded from further analysis. Individuals with a high proportion of missing genotypes may suggest a suboptimal genotyping process or poor DNA quality, and missing data can skew association testing. Nineteen EC cases were removed because of low genotyping. Extreme heterozygosity of autosomal SNPs may suggest contamination or inbreeding and the proportion of missing genotypes was plotted against heterozygosity rates ((N-O)/N; where N is the number of observed genotypes, O is number of homozygotes). Figure 3.2 shows that there are several samples that are excluded both because of low genotyping call rates and increased heterozygosity and this may suggest contamination

present in these samples. Samples with slightly lower heterozygosity and >95% call rates were kept since significant inbreeding is less likely to be present because these samples subsequently passed IBD QC analysis.

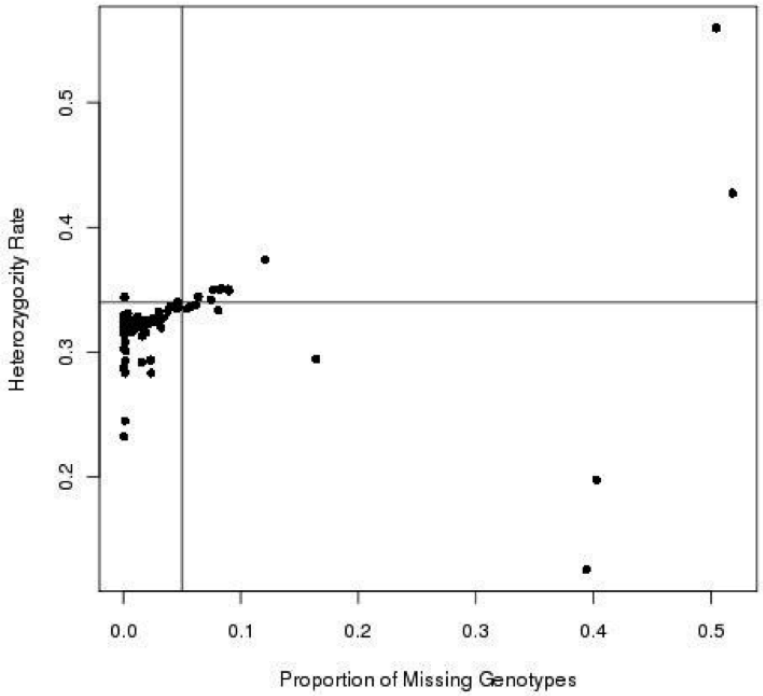

Figure 3.2: Plot of heterozygosity against missingness in NSECG samples. Cut-offs for heterozygosity and missingness were set at 0.34 and 0.05 respectively

3.4.1.3 Sex-check based on chromosome X homozygosity

The sex of each sample is estimated using the genotype data on chromosome X excluding the pseudo-autosomal regions, and this is compared with the sex provided in the phenotype files. The X chromosome inbreeding/homozygosity estimate (or F-score in PLINK) approximates the homozygosity of the X chromosome. Since males have only a single version of chromosome X, they should be homozygous for all SNPs not in pseudo-autosomal regions and yield an F-score of 1. Females have a lower rate of homozygozity given the presence of two alleles for each SNP and therefore

the presence of heterozygotes. This score can be affected by genotyping error and in practice, a male call is made if F is more than 0.8; a female call is made if F is less than 0.2. Any discrepancy between the imputed and recorded sex status highlights possible plating errors, sample mix-ups or contamination. This is especially important as we are studying a female specific disease phenotype and a discordant imputed sex status can flag up male controls mis-assigned as cases. In our study, no sample in either cases or controls, was excluded based on imputed sex.

3.4.1.4 Pairwise identity by descent (IBD)

Each individual is compared to every other and estimates of pairwise IBD identifies individuals who are genotypically more similar to each other then we would expect by chance (section 2.3.4). In the presence of duplicates and family members, the genotypes of these individuals are over-represented and not an accurate reflection of the allele frequencies in the larger population, to which we want to generalize the results to. The maximum relatedness between any pair of individuals is less than that of a second-degree relative and thus a cut-off of 0.1875 was used. In our sample set, 9 individuals were excluded for this reason.

3.4.1.5 Principal components analysis

Confounding is a source of bias in GWA studies caused by underlying differences between the case and control groups apart from their case/control status. Population stratification describes the genotypic differences between cases and controls due to systematic ancestry differences. It is therefore important to exclude individuals of divergent ancestry so that the differences in allele frequency driving the association are indeed due to disease status and not heterogeneous ancestry. PCA has been shown to be an effective way to control for population stratification in GWAS . PCA is a multivariate statistical method that uses orthogonal transformation of observations

in a data matrix to produce linearly uncorrelated variables, called principal components (section 2.3.5). PCA can also detect other confounding factors that could have caused systematic differences between the case and control group and thus in an ideal situation, the case and control individuals should cluster tightly and overlap for scatter plots of the first few principal components. Samples that clustered away from the main centroid in the first 4 PCs were excluded.

3.4.2 SNP Quality Control

The following quality control measures assess the quality of genotyped SNPs.

3.4.2.1 Genotypic call rate

SNPs will only be kept if the call rate per genotype is above 95%. SNPs with low call rates may indicate that they were poorly designed and that the arrays cannot accurately determine the genotypes for these SNPs. In commercial SNP arrays, there is usually more than one set of probes for each SNP in order to increase the accuracy of the genotyping but there is a subset of SNPs that fail because of nucleotide unspecificity.

3.4.2.2 Hardy-Weinberg Equilibrium

SNPs that show extensive deviation from Hardy-Weinberg equilibrium (HWE) are removed and the threshold used in this study was $P<10^{-6}$. HWE is expected to hold in stable populations in the absence of selection (section 2.3.2), and significant deviation indicates an unexpected relationship between allele frequency and genotype counts. This can arise due to failed array genotyping or faulty clustering and genotype calling. While controls should be well within HWE, deviation from HWE in cases may also suggest selection, which can exist in the presence of a phenotypic association.

3.4.2.3 Minor allele frequency

Rare SNPs with a minor allele frequency (MAF) of less than 1% suggest a low number of rare homozygotes and heterozygotes and genotype-calling algorithms are ill-equipped to make these calls accurately unless there is a large number of samples for clustering. SNPs with a low minor allele frequency may also cause spurious associations because the signals were driven by a small number of individuals who possess the rare variant and the power to detect association with rare variants is low.

3.4.2.4 Evaluation of cluster plots for top SNPs

Cluster plots were checked manually for SNPs that show promising results from association analysis. If distinct groups were seen, then it was manually re-clustered for that marker whereas markers showing poor differentiation between the three genotypes were zeroed and excluded from analysis.

3.5 Control of population stratification

As stated above, the control of population stratification is one of the most important aspects of a GWAS in that the differences in ancestry can account for differences in allele frequency in cases and controls. This study aims to look at susceptibility to EC in an English population of Caucasian ancestry, and recruiting questionaires helped us to screen out any individuals who did not identify themselves as "White-British." PCA (section 2.3.5) is a useful method for controlling population stratification and we conducted multiple rounds of PCA in order to exclude outlying samples that deviated significantly from the main cluster.

40,382 SNPs that were present in all case and control datasets were used for PCA in order to control for population stratification. These SNPs pass all of the quality control measures described above, and in addition, regions of long range linkage

disequilibrium were excluded. The SNP set was pruned so that nearby SNPs in LD with ar^2 of more than 0.2 were excluded since relatedness between the markers can influence the principal components model.

PCA was first conducted on the NSECG cases in a dataset combined with HapMap haplotype data to detect large-scale continental level ancestry differences. The four populations in the HapMap Project include haplotype information from 30 Yoruba Trios from Nigeria (YRI), 30 trios of northern and western European ancestry (CEU), 44 unrelated individuals from Tokyo Japan (JPT) and 45 unrelated Han Chinese individuals from Beijing, China (CHB). Figure 3.3 shows that all NSECG cases clustered well within the European CEU cluster except for four outlying samples that deviated significantly toward African ancestry and 11 outliers deviating toward the Asian ancestry cluster. Exclusions were made at the dividing line as shown in figure 3.3 and samples that deviated slightly from the main cluster are further assesed in subsequent PCAs.

Cancer-free controls were available from the UK-CORGI and Scotland phase 1 studies. In addition, publicly available WTCCC2 controls were used base on a birth cohort and blood donation services. Some WTCCC controls were matched with SEARCH cases in the GWAS by Spurdle *et al.* and thus I used 1,221 BC58 controls that were not used in the SEARCH analysis. Figure 3.4 shows the PCA with NSECG cases and controls from CORGI, Scotland phase 1 and BC58. Samples were excluded as shown and the remaining 925 NSECG cases and 2,507 controls cluster tightly for the first two principal components as shown, and also for subsequent principal components, suggesting that there is no significant population stratification in this study. Thus, principal components were not included as covariates in the association analysis.

It is worth noting that not all the controls have been screened for EC, and male controls are also included in the analysis although EC only affects females. The

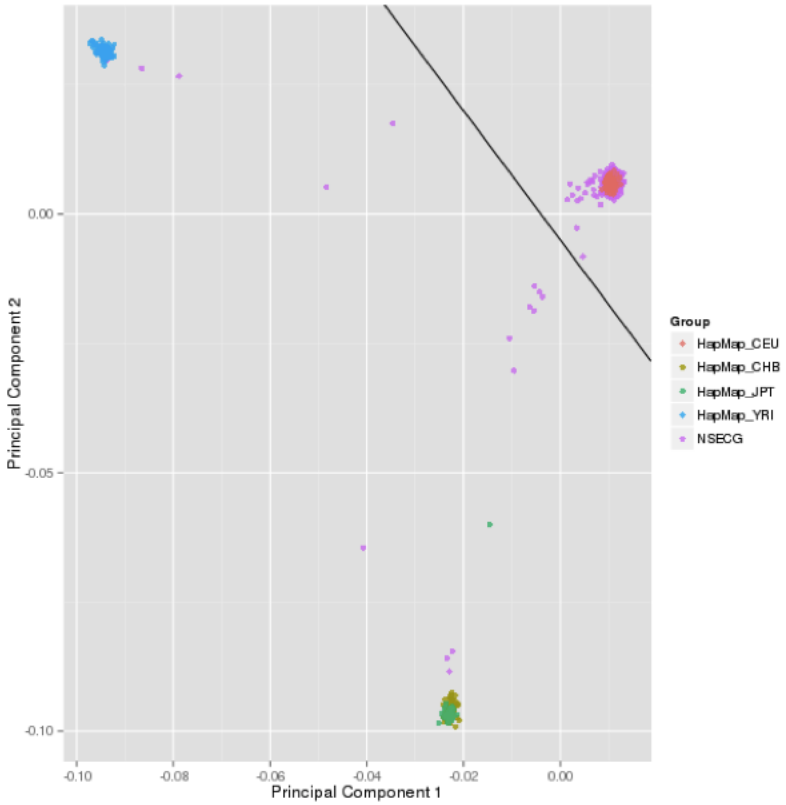

Figure 3.3: PCA of NSECG samples with HapMap samples of European (CEU), Yoruba African (YRI), Japanese (JPT) and Chinese (CHB) ancestry. The first two principal components are used in this scatter plot, NSECG samples below the dividing line were excluded.

prevalence of EC is relatively low in the general population and therefore the chances of misclassification bias is low. In addition, only autosomal variants are assessed in the GWAS, there is no evidence that inheritance of autosomal variants in influenced by sex, and thus inclusion of a larger number of controls of both sexes provides an improved approximation of allele frequencies in the general population.

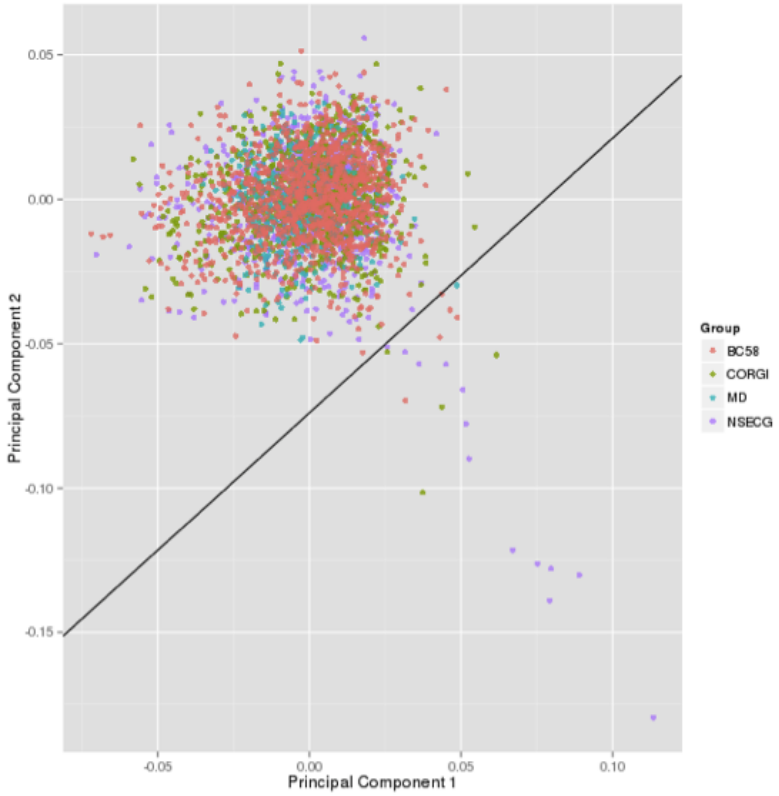

Figure 3.4: PCA of NSECG cases and CORGI, Scotland phase 1(MD) and 1958 birth cohort (BC58) controls. This is a scatter plot of the first two principal components. Samples below the dividing line were excluded.

3.6 Results

Association testing was performed using PLINK software, with allelic ORs calculated using the allelic test. The quantile-quantile plot (QQ plot) sorts all of the SNPs in order from highest to lowest P-values, and plots the expected P-values against the observed values on a logarithmic scale. The y=x guideline allows the comparison of

observed and expected values, and a marked upward deviation from the guideline would suggest that there is a propensity for false positive results. Figure 3.5 demonstrates that the majority of markers follow the null distribution with only a handful of SNPs in the upper tail showing evidence of association. The genomic inflation factor (λ_{GC}, see section 2.3.8, based on the median χ^2 statistic, is 1.02 and this suggests that there is no significant population stratification that is unaccounted for and that there is a low chance of systematic false positive associations in this analysis.

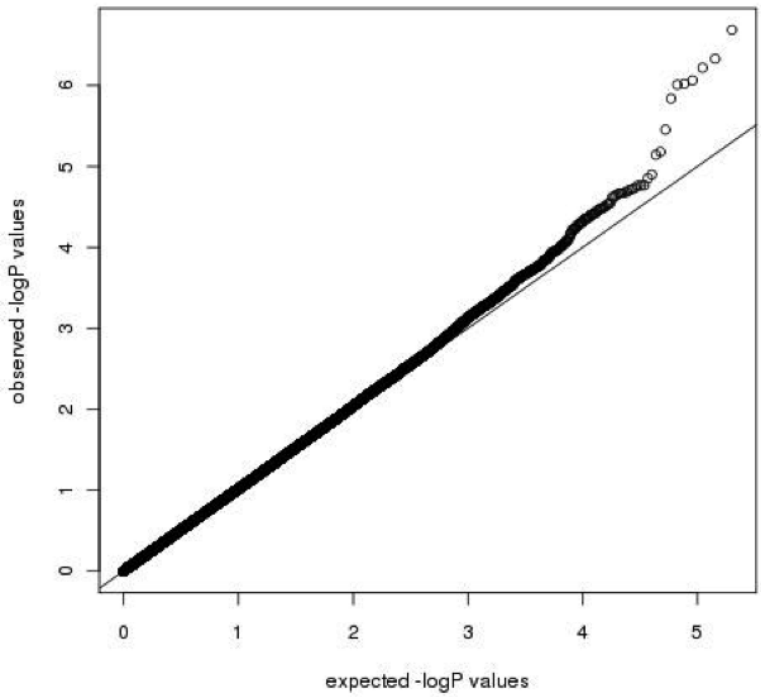

Figure 3.5: Quantile-quantile plot of NSECG association testing results comparing observed and expected levels of significance in -$\log_{10}$ P-values.

The P-values from association testing were plotted for all SNPs across the genome in the Manhattan plot in figure 3.6. The $-log_{10}$ of the P-values are plotted against the SNP's position in the genome and thus SNPs with the lowest P-values, the highest significance where the null hypothesis of no association should be rejected, appear highest in the plot. As expected, the vast majority of SNPs tested showed little or

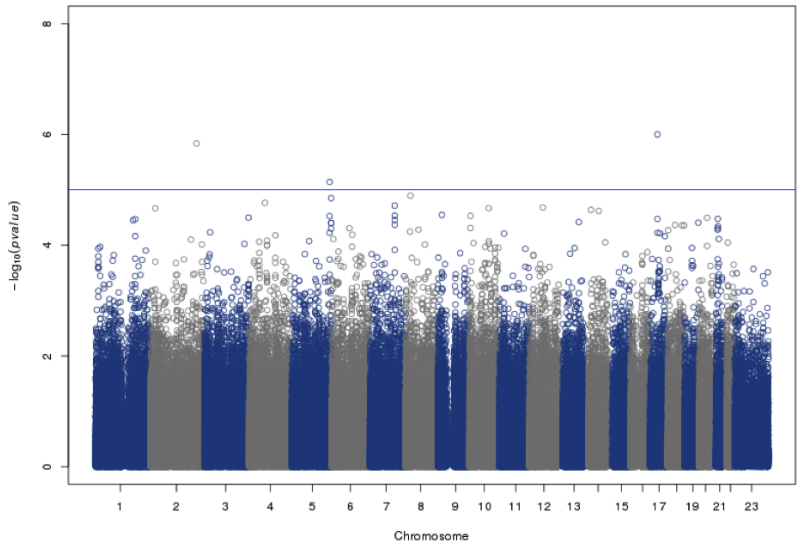

Figure 3.6: NSECG Manhattan plot: Y-axis is -$\log_{10}$ P-values from EC association testing. X-axis is genomic position, with chromosomes marked. Horizonal line represents P=10−5.

no association with EC. The 15 most significant SNPs are listed in table 3.1.

There were two SNPs with a P-value of less than 10^{-5} and the most significant SNP is rs4430796 (OR: 0.74, CI95: 0.65 – 0.83, P=9.94x10^{-7}) which is located at the second intronic region of *HNF1B* at 17q12, 12kb away from the transcription start site. Figure 3.7 is the regional association plot which shows the location of the top SNP in relation to *HNF1B* and the genes nearby. There are two SNPs in linkage disequilibrium (with an r^2 of more than 0.4) close to rs4430796 with P-values of less than 10^{-4}, which gives support to the existence of an important variant in this region and also suggests that the haplotype structure in NSECG is similar to that of European samples in the 1000 Genomes Project.

The second SNP with a P-value of lower than 10^{-5} is rs9313548 (OR: 0.1.28, 95%CI 1.15 – 1.42, P=7.22x10^{-6}) which is located in an intergenic region of chromosome 5q35, 76kb downstream from *FGF18* and 128kb downstream of *NPM1*. Figure 3.8 is the regional association plot of rs9313548 and that shows several nearby SNPs in partial LD with P-values of less than 0.001 and also the relation of these SNPs to

Table 3.1: Top 15 SNPs in NSECG GWAS. Chr, chromosome; Allele, minor allele; AF, allele frequency; L95, lower; U95, upper 95% confidence interval.

SNP	Chr	Position	Nearby gene	Allele	Case AF	Control AF	OR	L95	U95	P
rs4430796	17	36,098,040	*HNF1B*	G	0.422	0.497	0.74	0.65	0.83	9.94E-07
rs9313548	5	170,961,300	*FGF18*	T	0.532	0.471	1.28	1.15	1.42	7.22E-06
rs11135721	8	23,150,878	*R3HCC1*	T	0.425	0.367	1.27	1.14	1.42	1.27E-05
rs2913788	5	177,757,764	*COL23A1*	C	0.377	0.435	0.78	0.70	0.88	1.41E-05
rs17688758	4	71,997,259	*SLCA4*	C	0.114	0.081	1.47	1.23	1.75	1.71E-05
rs2523017	7	111,807,776	*DOCK4*	A	0.143	0.187	0.72	0.62	0.84	1.92E-05
rs348649	12	62,462,075	*FAM19A2*	A	0.413	0.357	1.27	1.14	1.41	2.08E-05
rs2492742	10	88,005,754	*GRID1*	A	0.045	0.074	0.59	0.47	0.76	2.14E-05
rs4533498	2	22,688,256	*KLHL29*	G	0.127	0.092	1.44	1.21	1.70	2.16E-05
rs4981793	14	31,305,647	*COCH*	C	0.255	0.308	0.77	0.68	0.87	2.30E-05
rs2318310	14	66,397,186	*FUT8*	T	0.280	0.333	0.78	0.69	0.87	2.40E-05
rs2811806	9	17,964,514	*SH3GL2*	C	0.227	0.182	1.32	1.16	1.50	2.84E-05
rs2708600	7	112,005,565	*ZNF277*	A	0.354	0.410	0.79	0.71	0.88	2.91E-05
rs1064891	10	6,276,574	*PFKFB3*	C	0.456	0.400	1.26	1.13	1.40	2.94E-05
rs7374260	3	196158332	UBXN7	C	0.5233	0.4667	1.254	1.127	1.396	3.16E-05

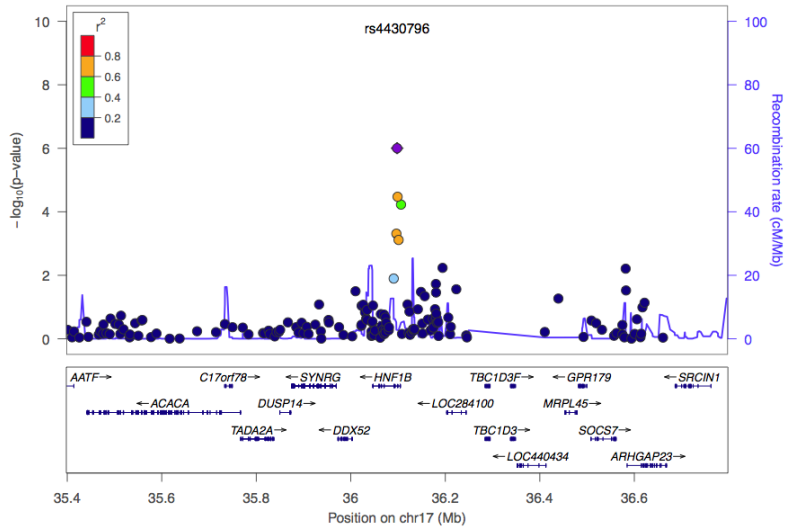

Figure 3.7: Regional association plot of chromosome 17q12 around rs4430796. The top SNP rs4430796 is marked as a purple diamond, and the dot color represents the LD with the top SNP. Y-axis is -$\log_{10}$ P-values from the NSECG GWAS. The blue line shows recombination rates in cM/Mb.

nearby genes.

ECs with endometrioid-only histology have been postulated to be oestrogen driven and follow a different clinical course (section 1.1.5), and thus association testing was also done with the 795 EC cases with endometrioid-only histology. The two most significant SNPs were rs4430796 (OR: 0.71, CI95%: 0.63 - 0.81, P=2.90x10^{-7}) and rs9313548(OR: 1.31, CI95%: 1.17 - 1.46, P=3.47x10^{-6}). These are the same SNPs as the all histologies analysis and there were no other SNPs that were found with P<10^{-4} which are not listed in table 3.1.

3.7 Discussion

The NSECG GWAS identified two SNPs with a P-value of less than 10^{-5}; the more significant is rs4430796, which is located in the intronic region of *HNF1B* on chromosome 17q12.

HNF1B (hepatocyte nuclear factor 1 homeobox B) encodes a member of the

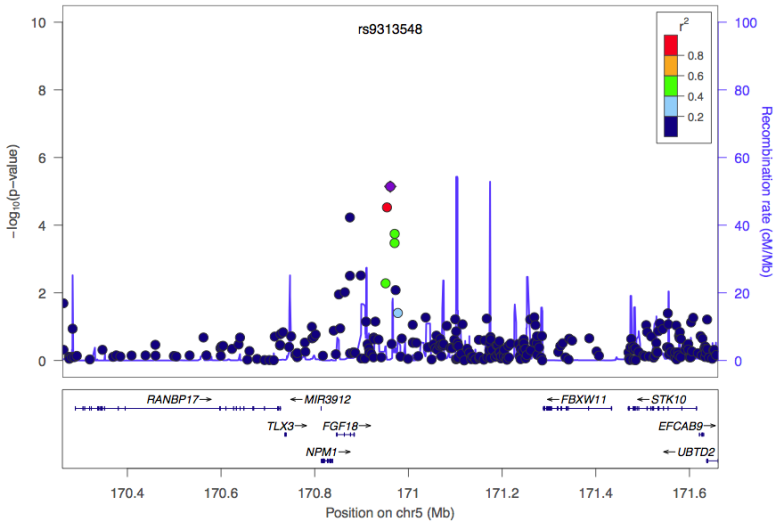

Figure 3.8: Regional association plot of chromosome 5q35 around rs9313548. The top SNP rs9313548 is marked as a purple diamond, and the dot color represents the LD with the top SNP. Y-axis is -$\log_{10}$ P-values from the NSECG GWAS. The blue line shows recombination rates in cM/Mb.

homeodomain-containing superfamily of transcription factors and rs4430796 has previously been associated with certain subtypes of ovarian cancer (Pharoah *et al.*, 2013; Shen *et al.*, 2013) , prostate cancer (Berndt *et al.*, 2011) and type 2 diabetes (Gudmundsson *et al.*, 2007). Although an increased body mass index (BMI) is a risk factor for both type 2 diabetes and EC, and type 2 diabetes has been associated with an increased risk of EC (section 1.1.2), the association is in opposite directions suggesting different causal mechanisms. Of interest to gynecological system, recurrent microdeletion at 17q12 is a cause of Mayer-Rokitansky-Kuster-Hauser (MRKH) syndrome, which is characterized by congenital aplasia of the uterus and upper vagina due to anomalous development of the Mullerian ducts. *HNF1B* deletions are associated with incomplete Mullerian duct fusion (Bernardini *et al.*, 2009) and Mullerian duct aplasia causing combined uterine and renal malformations (Oram *et al.*, 2010).

FGF18 is a mitogenic member of the fibroblast growth factor family, which are tyrosine kinases that regulate cell proliferation and metabolism. This gene encodes a phosphoprotein which moves between the nucleus and the cytoplasm and the gene

product is thought to be involved in several processes including regulation of the ARF/p53 pathway. FGF signalling has been shown to increase the mitotic rate of tumour cells, promote tumour angiogenesis and enhances tumorigenicity of cancer stem cells (Beenken and Mohammadi, 2009; Turner and Grose, 2010). Multiple drugs have been developed in order to block the FGF pathway including monoclonal antibodies (Sun *et al.*, 2007) and FP-1039, soluble decoy receptor which targets the mitogenic members of the FGF family including FGF18

The NSECG GWAS has provided further evidence for the association of *HNF1B* SNP rs4430796 with EC and found other promising risk loci including chromosome 5q35 near *FGF18*. Further genotyping and meta-analysis with other EC GWAS would increase the power to detect associations and a collaboration has been formed with the ANECS and SEARCH GWAS studies in order to identify more risk loci associated with EC.

Chapter 4

Endometrial cancer GWAS meta-analysis

4.1 Introduction

4.1.1 Background

GWAS have been successful in identifying common variants that are associated with complex traits and the combining of different GWAS of the same phenotype provide additional power to detect associated variants (Evangelou and Ioannidis, 2013). The added sample size increases power but may also introduce heterogeneity, in that samples may display more diverse geographical locations, ancestries, sample recruitment methods and processing. There have been other EC GWAS and EC is a phenotype that can be consistently ascertained via pathology reports and cancer registries. The meta-analysis conducted in this chapter makes use of samples from the UK, Europe, USA and Australia, all of European ancestry in order to reduce genetic heterogeneity.

4.1.2 Aims and overview

This chapter describes the meta-analysis of NSECG with a re-analysis of a previously published EC GWAS consisting of ANECS and SEARCH datasets (1287 cases, 5584 controls, Spurdle *et al.*, 2011) and two replication phases. The first replication phase consists of 4,330 EC cases and 26,849 controls from Europe, the United States and Australia analysed in eight groups as the Endometrial Cancer Association Consortium (ECAC) branch of the Collaborative Oncological Gene-environment Study (COGS) initiative. The second replication phase made use of non-overlapping NSECG samples in order to directly genotype and validate SNPs that were near or surpassing genome-wide significance. Figure 4.1 is a flow diagram that summarizes the EC meta-analysis study design and results.

I discovered five novel endometrial cancer risk loci mapping to chromosomes 13q22.1 (rs11841589, near *KLF5* and *KLF12*), 6q22.31 (rs13328298, in *LOC643623* and near *HEY2* and *NCOA7*), 8q24.21 (rs4733613, telomeric to *MYC*), 15q15.1 (rs937213, in *EIF2AK4*, near *BMF*) and 14q32.33 (rs2498796, in *AKT1*, near *SIVA1*). The association of the chromosome 15q15 *CYP19A1* from previous candidate gene studies was confirmed to be genome wide-significant for the first time.

4.2 Additional cohorts

In order to identify additional EC risk loci, a meta-analysis was performed using EC cases and controls from a previous EC GWAS (Spurdle *et al.*, 2011) and two further replication phases as detailed in figure 4.2. Detailed descriptions about each study cohort can be found in section 2.1. WTCCC2 controls were previously used in the SEARCH study and National blood service (NBS) and 1958 birth cohorts controls were allocated in the meta-analysis such that each control group was used only once.

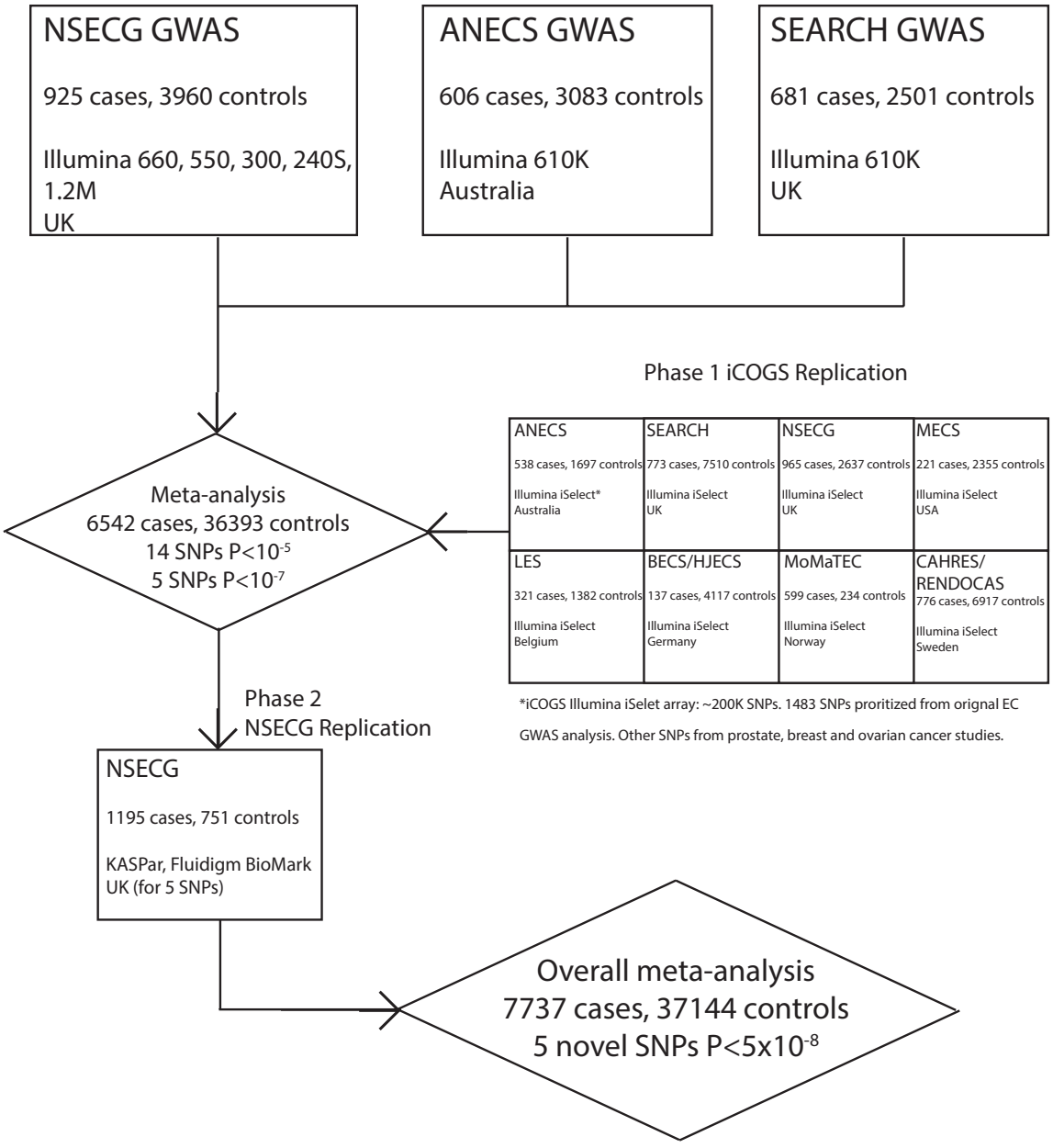

Figure 4.1: The NSECG GWAS was meta-analysed with the ANECS and SEARCH GWAS and phase 1 iCOGS replication (eight groups). This meta-analysis of 6,542 cases and 36,393 controls yielded 14 regions with at least one SNP $P < 10^{-5}$, of which five regions contained SNPs with $P < 10^{-7}$. Five SNPs were brought forward to the phase 2 NSECG replication and were confirmed as novel genome-wide significant risk loci (P<5x10^{-8}) in the overall meta-analysis of 7,737 cases and 37,144 controls.

	Study	Case sampling frame	Control sampling frame	Genotyping	EC Cases	Endometoid	Controls
	New GWAS						
1	NSECG	UK National Study of Endometrial Cancer Genetics	UK; population based cases	Illumina 660W	925	795	
	UK1-CORGI	UK Colorectal Tumour Gene Identification Consortium	England; partners of cases with no colorectal tumours	Illumina 550			894
	SP1	Scotland1	Scotland; cancer free controls from NHS registers	Illumina 300, 240S			392
	BC58	UK 1958 Birth Cohort	UK; population based controls, born in 1958	Illumina 1.2M			2,674
	Previous GWAS						
2	ANECS	Australian National Endometrial Cancer Study	Australia; population based cases	Illumina 610K	606	606	
	QIMR	Queensland Institute of Medical Research	Australia; parents of participants in adolescent twin study	Illumina 610K			1,846
	HCS	Hunter Community Study	Australia; population-based controls	Illumina 610K			1,237
3	SEARCH	UK Studies of Epidemiology and Risk factors in Cancer Heredity	England; population based cases via cancer registries	Illumina 610K	681	681	
	NBS		UK; population-based from National Blood Service	Illumina 1.2M			2,501
	iCOGS Replication						
4	ANECS	Australian National Endometrial Cancer Study	Australia; population based cases	Infinium iSelect	373	217	
	NECS	Newcastle Endometrial Cancer Study	Australia; hospital-based cases	Infinium iSelect	165	116	
	ABCFS	Australian Breast Cancer Family Study	Australia; from electoral rolls	Infinium iSelect			443
	AOCS	Australian Ovarian Cancer Study	Australia; population-based, from electoral rolls	Infinium iSelect			817
	MCCS	Melbourne Collaborative Cohort Study	Australia; random sample from initial cohort	Infinium iSelect			437
5	SEARCH	UK Studies of Epidemiology and Risk factors in Cancer Heredity	England; population based cases	England; population based controls	Infinium iSelect	773	620	7,510
6	NSECG	National Study of Endometrial Cancer Genetics	England; population based cases	Infinium iSelect	965	839	
	BBCS	British Breast Cancer Study	UK; friend, non-blood relative of cases	Infinium iSelect			1,353
	SBCS	Sheffield Breast Cancer Study	UK; women mammography screening negative results	Infinium iSelect			835
	UKBGS	UK Breakthrough Generations Study	UK; women without breast lesions selected from BGS cohort	Infinium iSelect			449
7	MECS	Mayo Endometrial Cancer Study	USA; Hospital based cases	Infinium iSelect	221	163	
	MCBCS	Mayo Clinic Breast Cancer Study	USA; Cancer-free women	Infinium iSelect			1,762
	MCOCCCS	Mayo Clinic Ovarian Cancer Case-Control Study	USA; Cancer-free women	Infinium iSelect			593
8	LES	Leuven Endometrial Cancer Study	Belgium; hospital based cases	Infinium iSelect	321	219	
	LMBC	Leuven Multidisciplinary Breast Centre	Belgium; controls from blood donors	Infinium iSelect			1,382
9	BECS/HJECS	Bavarian Endometrial Cancer Study/	Germany; hospital-based cases, population-based cases	Infinium iSelect	137	112	
	BBCC	Bavarian Breast Cancer Cases and Controls/ Hannover-Jena Endometrial Cancer Study	Germany; healthy women from newspaper advertisement	Infinium iSelect			441
	BSUCH	Breast Cancer Study of the University Clinic Heidelberg	Germany; female blood donors	Infinium iSelect			920
	ESTHER	ESTHER Breast Cancer Study	Germany; random sample from routine health check-up	Infinium iSelect			486
	GC-HBOC	German Consortium for Hereditary Breast & Ovarian Cancer	Germany; KORA study	Infinium iSelect			138
	GENICA	Gene Environment Interaction and Breast Cancer in Germany	Germany; random address sample	Infinium iSelect			420
	MARIE	Mammary Carcinoma Risk Factor Investigation	Germany; randomly drawn from population registries	Infinium iSelect			1,712
10	MoMaTEC	Molecular Markers in Treatment of Endometrial Cancer	Norway; population based cases	Norway; female blood donors	Infinium iSelect	599	505	
	NBCS	Norwegian Breast Cancer Study	Norway; BC screening negative	Infinium iSelect			234
11	CAHRES/RENDOCAS	Cancer Hormone Replacement Epidemiology	Sweden; population based cases	Infinium iSelect	543	512	1,345
	RENDOCAS	Registry of Endometrial Cancer in Sweden	Sweden; hospital based cases	Infinium iSelect	233	205	
	KARBAC	Karolinska Breast Cancer Study	Sweden; blood donors	Infinium iSelect			585
	pKARMA	Karolinska Mammography Project for Risk Prediction of Breast Cancer	Sweden; mammography screening negative	Infinium iSelect			4,987
	Replication phase 2						
12	NSECG	National Study of Endometrial Cancer Genetics	UK; population based cases	KASPar, Fluidigm	1,195	1,045	
	UK1/CORGI	UK Colorectal Tumour Gene Identification Consortium	England; partners with no colorectal tumours	KASPar, Fluidigm			751
	OVERALL META-ANALYSIS TOTAL				**7,737**	**6,635**	**37,144**

Figure 4.2: Table displaying three EC GWAS: NSECG, ANECS, SEARCH and eight iCOGS groups. Sampling frame, genotyping platform and sample numbers. A total of 7,737 EC cases, 6,635 with endometroid only histology, and 37,144 controls.

4.2.1 Previous EC GWAS

The Australian National Endometrial Cancer Study (ANECS) is an Australian population-based case-control family study of cancer of the uterine corpus with 608 EC cases of endometrioid histology and 3,083 controls. The UK Studies of Epidemiology and Risk factors in Cancer Heredity (SEARCH) is an ongoing population-based study with 681 EC cases of endometroid histology matched with 2,501 NBS controls. All samples were genotyped using Illumina 610K arrays except NBS, which was genotyped using Illumina 1.2M arrays. Standard quality control measures were conducted by collaborators as detailed in the publication (Spurdle *et al.*, 2011) and I also checked for all quality control measures described in chapter 3.

4.2.2 Phase 1 iCOGS replication

The phase 1 iCOGS replication consists of a further 4,330 EC cases and 26,849 controls from Europe, the United States and Australia who were genotyped using a custom Illumina Infinium iSelect array as part of the Collaborative Oncological Gene-environment Study (COGS) initiative (Eeles *et al.*, 2013; Michailidou *et al.*, 2013; Pharoah *et al.*, 2013; Sakoda *et al.*, 2013). The SNPs on this array were chosen based on regions of interest from previous breast, prostate, ovarian and endometrial cancer studies, rather than for genome-wide coverage. This array also included 1,483 SNPs selected on the basis of the most promising signals from the previous EC GWAS (ANECS and SEARCH). The controls were healthy females participating in the Breast Cancer Association Consortium (BCAC, Michailidou *et al.*, 2013) and Ovarian Cancer Association Consortium (OCAC, Pharoah *et al.*, 2013) parts of the iCOGS project, and were matched and analyzed with EC cases in eight groups based on geographical location and case-control clustering in PCA. All samples were of European ancestry as confirmed by PCA and samples were excluded if they clustered away from the centroid in the first four PCs (figure 4.3).

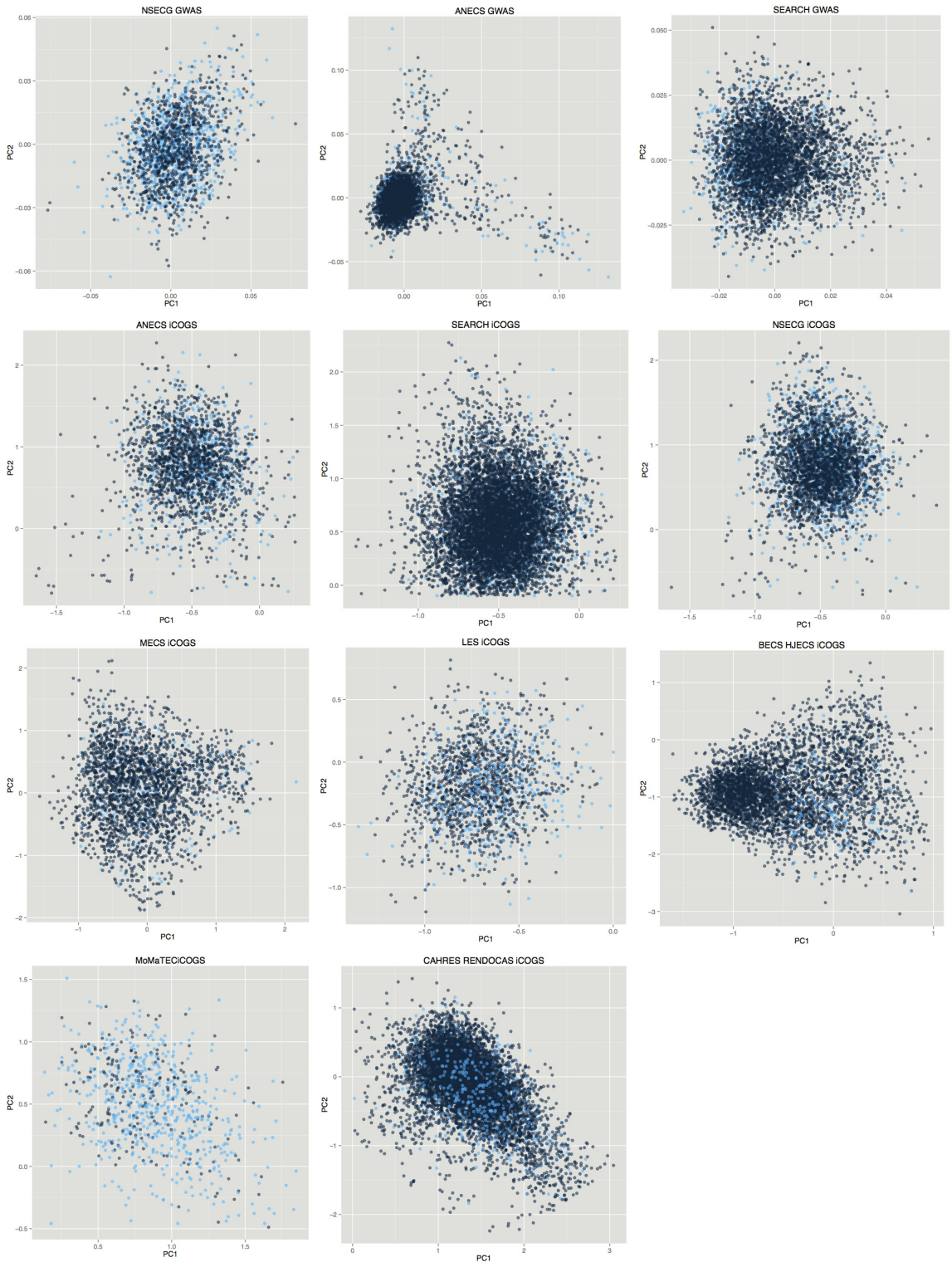

Figure 4.3: Plots of the first two principal components (PCs) in each study. EC cases are represented by blue dots, whereas controls are in black. Samples were excluded if they clustered away from the centroid in the first four PCs and these are plots of the samples used in the analysis.

4.2.3 NSECG replication

In the meta-analysis of three EC GWAS(NSECG, ANECS and SEARCH) and eight groups in the iCOGS replication, I identified five novel regions with at least one SNP with $P_{meta}<10^{-7}$. After regional imputation, five SNPs were genotyped in the phase 2 NSECG replication using 1,195 non-overlapping EC samples from NSECG and 751 UK CORGI controls using competitive allele-specific PCR (KASPar, KBiosciences) and the Fluidigm BioMark HD System. These samples were collected as part of the NSECG and CORGI studies as described previously (section 2.1), but had not been used in the NSECG GWAS or the NSECG part of iCOGS genotyping.

4.3 Power calculations for EC meta-analysis

The NSECG, ANECS, SEARCH and iCOGS meta-analysis totalled 6,542 cases and 36,393 controls and figure 4.4 shows the power plot for this analysis. This study has 74% power to detect a SNP with MAF 0.2, OR 1.15 at genome-wide significance (P=5x10^{-8})

4.4 Meta-analysis of NSECG, previous EC GWAS and iCOGS

Case-control analysis for the three GWAS and eight iCOGS groups were performed using frequentist tests with an additive effect logistic regression model. There was little evidence of systematic overdispersion of the test statistic (λ_{GC}=1.002-1.038 in each of the groups, figure 4.5). All genotyped SNPs were used for the λ_{GC} calculations and quantile-quantile (QQ) plots for the NSECG, ANECS and SEARCH GWAS. SNPs within 500kb or in LD (r^2>0.2) with previously prioritized EC SNPs were excluded. The remaining ∼100,000 SNPs were used for λ_{GC} calculation and QQ

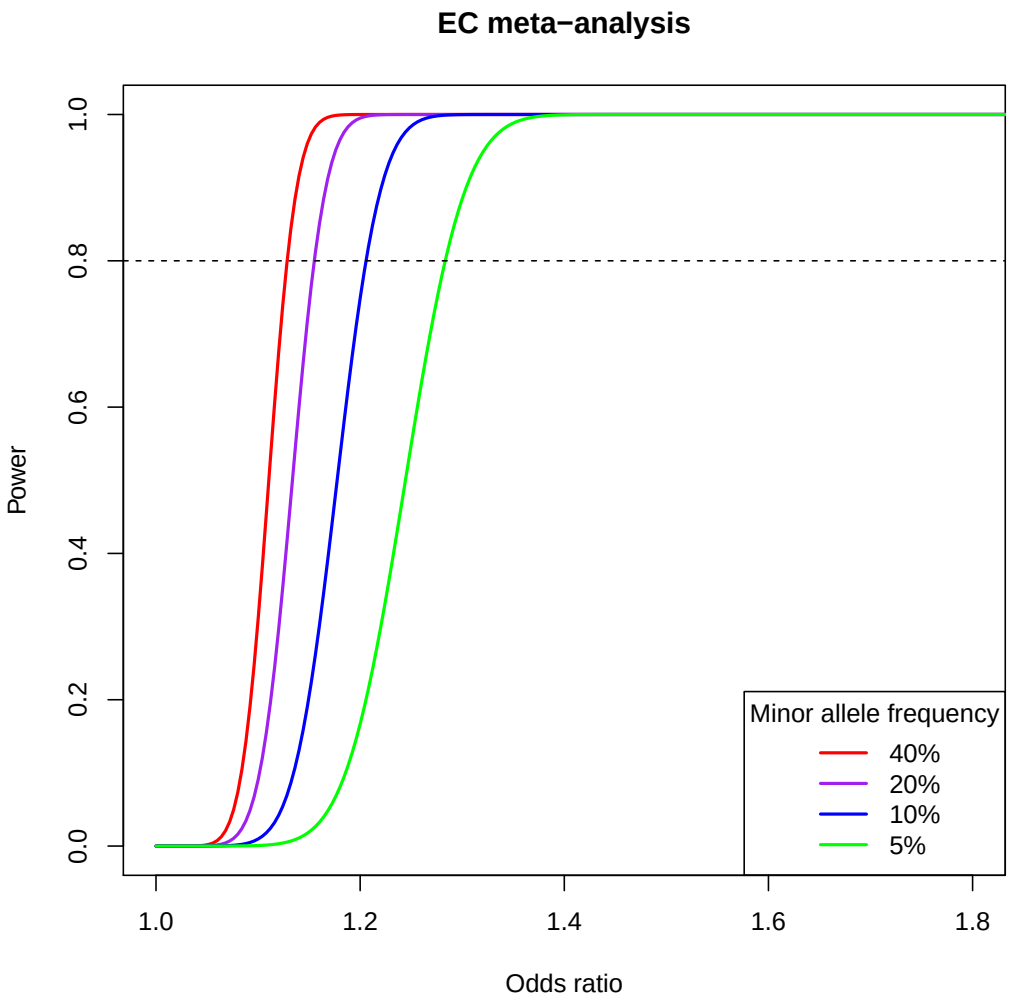

Figure 4.4: Power plot EC meta-analysis with 6,542 cancer cases and 36,393 controls. The graph displays the power to detect associations at P=5x10^{-8} given the odds ratio and minor allele frequency.

plots for the eight iCOGS groups.

Fixed-effects, inverse variance-weighted meta-analysis (section 2.3.10) was conducted for all typed and well-imputed SNPs (info score>0.90) in a total of 6,542 EC cases and 36,393 controls of European ancestry. Heterogeneity was assessed using the I^2 statistic (section 2.3.11).The genotyping arrays for NSECG, ANECS and SEARCH were designed for genome-wide coverage, and thus genome-wide imputation was conducted for these studies. The iCOGS was designed based on previously prioritized cancer SNPs and thus, only genotyped SNPs were used in this phase 1 meta-analysis.

For the 14 regions that contained at least one SNP with a P-value $<10^{-5}$, I performed imputation of the 1.5Mb surrounding region using two reference panels (1000 Genomes 2012 release and 196 high-coverage whole genome-sequenced UK individuals, section 2.3.9). The inclusion of UK high-coverage sequence data as a reference panel allowed me to more accurately infer the haplotype structure in these regions of interest (Timpson *et al.*, 2014; Whiffin *et al.*, 2013). Table 4.1 displays phase 1 meta-analysis results . An imputed SNP with a more significant P-value is included only if it has an IMPUTE2 information score of more than 0.9. I employed this stringent info score cutoff because I wanted to improve the chances of replicating allele frequencies in cases and controls in the phase 2 replication genotyping.

4.5 NSECG replication

There are five regions novel SNPs with $P_{meta}<10^{-7}$ in the meta-analysis including the phase 1 replication, and the most significant SNP from each region was genotyped using competitive allele-specific PCR (KASPar, section 2.2.4) and Fluidigm BioMark. Fluidigm BioMark HD System genotyping was carried out by collaborators in Brisbane using standard protocols. Table 4.2 shows the NSECG replication results. The case and control allele frequencies were similar to previous studies and the ORs and

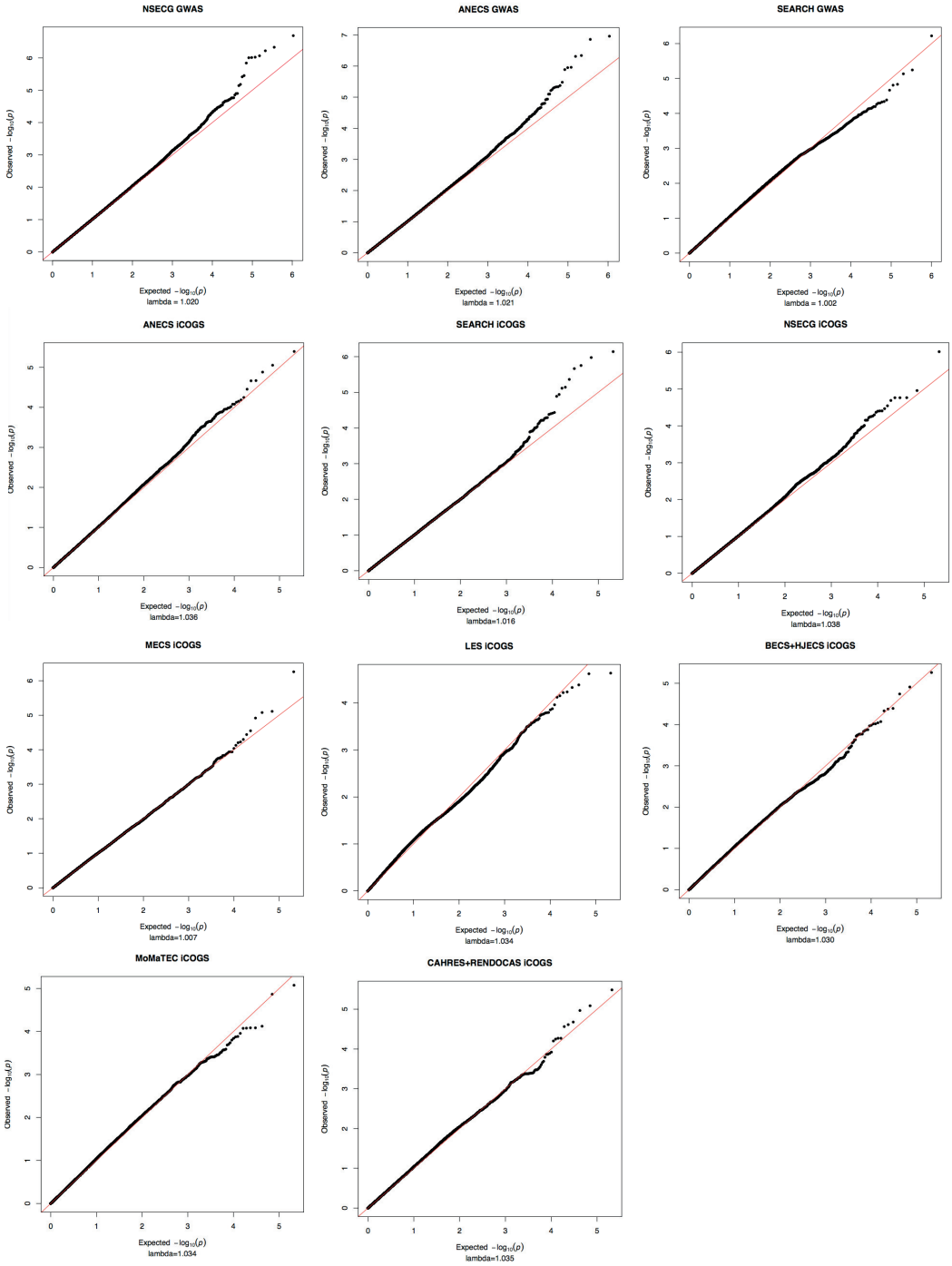

Figure 4.5: EC GWAS and iCOGS quantile-quantile plot: -$\log_{10}$ P-values (y-axis) are plotted against the expected -$\log_{10}$ P-values expected under the null hypothesis (x-axis). The red line denotes the expectation under no deviation from the null hypothesis. The QQ plots show little evidence of genomic inflation and the λ_{GC} for each study are NSECG GWAS 1.020, ANECS GWAS 1.021, SEARCH GWAS 1.002, ANECS iCOGS 1.036, SEARCH iCOGS 1.016, NESCG iCOGS 1.038, MECS iCOGS 1.007, LES iCOGS 1.034, BECS iCOGS 1.030, MoMaTEC iCOGS 1.034, CAHRES RENDOCAS iCOGS 1.037. All SNPs for the NSECG, ANECS and SEARH GWAS are included.

Table 4.1: Meta-analysis and regional imputation for 14 risk loci P<10^{-5} identified by Phase 1 meta-analysis of genome-wide imputed GWAS and genotyped iCOGS datasets. The top SNP genotyped in iCOGS and top SNP after regional imputation are listed for each locus. Imputed SNPs are only included if info score>0.9 in all datasets, imputed SNPs are in bold. Pos, position in build 19; Ref, reference allele; I^2, heterogeneity I^2.

Locus	Nearby gene(s)	Top iCOGS genotyped SNP in meta-analysis						Top SNP after regional imputation					
		SNP	Pos	Ref	Allelic OR (95%CI)	P	I^2	SNP	Pos	Ref	Allelic OR (95%CI)	P	I^2
13q22	*KLF5, KLF12*	rs11841589	73,814,891	G	1.16 (1.11-1.21)	6.89E-11	0.25	rs11841589	73,814,891	G	1.16 (1.11-1.21)	6.89E-11	0.25
6q22	*NCOA7, HEY2*	rs2747714	126,007,620	A	1.13 (1.08-1.17)	3.35E-09	0.00	**rs13328298**	126,016,580	G	1.13 (1.08-1.17)	2.78E-09	0.00
8q24	*MYC*	rs4733613	129,599,278	G	0.84 (0.79-0.89)	5.64E-09	0.02	rs4733613	129,599,278	G	0.84 (0.79-0.89)	5.64E-09	0.02
15q15	*EIF2AK4, BMF*	rs937213	40,322,124	T	0.90 (0.86-0.93)	3.81E-08	0.41	rs937213	40,322,124	T	0.90 (0.86-0.93)	3.81E-08	0.41
14q32	*AKT1, SIVA1*	rs3001371	105,242,831	C	0.89 (0.85-0.93)	1.33E-07	0.00	**rs2498796**	105,243,220	G	0.89 (0.85-0.93)	8.66E-08	0.02
12q24	*SH2B3, ATXN2*	rs3184504	111,884,608	C	1.11 (1.07-1.15)	2.16E-07	0.00	rs3184504	111,884,608	C	1.11 (1.07-1.15)	2.16E-07	0.00
17q21	*SKAP1*	rs3944039	46,205,070	C	1.10 (1.06-1.15)	1.63E-06	0.50	**rs1452666**	46,307,750	C	1.11 (1.06-1.15)	9.49E-07	0.48
2p16	*BCL11A*	rs6732518	60,708,597	T	1.11 (1.06-1.16)	2.26E-06	0.00	**rs7579014**	60,707,894	A	1.11 (1.07-1.16)	5.34E-07	0.00
12p12	*SSPN*	rs17467365	26,431,039	C	1.27 (1.15-1.41)	2.82E-06	0.33	rs17467365	26,431,039	C	1.27 (1.15-1.41)	2.82E-06	0.33
3q13	*LSAMP*	rs4378954	115,650,448	C	1.16 (1.09-1.24)	7.20E-06	0.31	rs4378954	115,650,448	C	1.16 (1.09-1.24)	7.20E-06	0.31
9q31	*KLF4*	rs4978670	110,826,546	A	0.91 (0.88-0.95)	7.33E-06	0.20	rs4978670	110,826,546	A	0.91 (0.88-0.95)	7.33E-06	0.20
8q21	*STAU2*	rs4237005	74,391,780	G	0.88 (0.83-0.93)	8.02E-06	0.00	rs4237005	74,391,780	G	0.88 (0.83-0.93)	8.02E-06	0.00
11q13	*CCDC88B*	rs71456310	64,128,494	C	1.14 (1.08-1.21)	8.69E-06	0.00	**rs12808002**	64,110,932	C	1.14 (1.08-1.20)	2.18E-06	0.00
3q21	*EEFSEC*	rs3021461	128,095,652	G	0.90 (0.86-0.94)	9.96E-06	0.41	rs3021461	128,095,652	G	0.90 (0.86-0.94)	9.96E-06	0.41

standard errors were used in an overall meta-analysis totalling 7,737 cases and 37,144 controls. In summary, the five novel EC risk loci were directly genotyped in the NSECG replication and either genotyped or imputed with an info score of more than 0.9 in all datasets.

4.6 Genotyping platform concordance

A small number of samples were genotyped in multiple platforms as part of quality control. Table 4.3 shows that samples genotyped on different platforms showed a concordance of >98.5%.

4.7 Novel EC risk loci

The five SNPs genotyped in the NSECG replication were meta-analysed with the previous results and all five were confirmed to be associated with EC at a genome-wide significant level (table 4.4). The five novel loci mapped to chromosomes 13q22.1 (rs11841589, near *KLF5* and *KLF12*), 6q22.31 (rs13328298, in *LOC643623* and near *HEY2* and *NCOA7*), 8q24.21 (rs4733613, telomeric to *MYC*), 15q15.1 (rs937213, in *EIF2AK4*, near *BMF*) and 14q32.33 (rs2498796, in *AKT1*, near *SIVA1*). Analysis was also done using the 6,635 EC cases with endometrioid only histology and those results are displayed in table 4.4. The 6q22 locus was the only one to show increased significance using endometrioid histology cases only. Analysis of 5,590 endometrioid histology EC cases for other SNPs across the genome in the phase 1 meta-analysis did not reveal any novel SNPs of a high association significance not listed in table 4.1. Conditional analysis was conducted for each locus and chromosome 8q24 is the only one to show evidence of a second independent signal. The five novel genome-wide significant EC risk loci are presented in descending levels of significance in the following sections followed by the results for 15q21 candidate gene *CYP19A1* and the

Table 4.2: Phase 2 NSECG replication and overall meta-analysis results. Ref, reference allele; AF, allele frequency. All SNPs are either typed or imputed with info>0.9 in all arrays including iCOGS, NSECG phase 2 replication was genotyped using KASPar (rs4733613, rs937213, rs2498796) and Fluidigm (rs13328298, rs11841589).

Locus	Nearby gene(s)	SNP	Position	Ref	Case AF	Phase 1 meta-analysis Control AF	OR (95%CI)	P	*I2*	Case AF	Phase 2 replication Control AF	OR (95%CI)	P	Overall P
6q22	*NCOA7, HEY2*	rs13328298	126,016,580	G	0.600	0.573	1.13 (1.08-1.17)	2.78E-09	0.00	0.616	0.569	1.22 (1.02-1.46)	0.032	3.73E-10
8q24	*MYC*	rs4733613	129,599,278	G	0.856	0.877	0.84 (0.79-0.89)	5.64E-09	0.02	0.853	0.867	0.89 (0.74-1.08)	0.219	3.09E-09
15q15	*EIF2AK, BMF*	rs937213	40,322,124	T	0.553	0.582	0.90 (0.86-0.93)	3.81E-08	0.41	0.549	0.570	0.92 (0.81-1.05)	0.524	1.77E-08
14q32	*AKT1, SIVA1*	rs2498796	105,243,220	G	0.677	0.705	0.89 (0.85-0.93)	8.66E-08	0.02	0.681	0.702	0.91 (0.79-1.04)	0.177	3.55E-08
13q22	*KLF5, KLF12*	rs11841589	73,814,891	G	0.757	0.729	1.16 (1.11-1.21)	6.89E-11	0.25	0.741	0.723	1.10 (0.91-1.33)	0.323	4.83E-11

Table 4.3: Genotyping concordance rates for different platforms in quality control duplicates. The number of overlapping SNPs and samples are listed. Concordance is presented as a percentage of concordant SNPs out of total SNPs.

Genotyping		Overlapping		Concordance
Platform1	Platform2	Samples	SNPs	
KASPar	Fluidigm	491	2	0.988
KASPar	Illumina 660W	23	6	0.986
KASPar	Illumina iSelect	9	3	1.000

known association at 17q12; *HNF1B*.

4.7.1 13q22 rs11841589 near *KLF5*

The most significant of the novel risk loci is rs11841589 (OR=1.15, 95%CI 1.11–1.21, P=4.83×10^{-11}) located on chromosome 13q22. This gene lies in a 608kb intergenic region flanked by two Kruppel-like factor subfamily of zinc finger proteins *KLF5* and *KLF12*. The forest plot and regional association plot are displayed in figures 4.6a and 4.7a. Conditional analysis suggested that there was a single signal at this locus. The nearest gene, *KLF5*, lies 163kb away, and is a transcription factor associated with cell cycle regulation. KLF5 is targeted by the tumour suppressor FBXW7 and is rapidly degraded via the ubiquitin proteasome pathway. I decided to pursue further functional analysis of 13q22, including *in vitro* exploration of the epigenetic landscape of this locus. This work is presented in chapter 5.

4.7.2 6q22 rs13328298 near *HEY2* and *NCOA7*

rs13328298 (OR=1.13, 95% CI 1.09–1.18, P=$3.73\text{x}10^{-10}$, figure 4.6b) on chromosome 6q22 lies in the long non-coding RNA *LOC643623*, 54kb upstream of helix-loop-helix transcription factor *HEY2* and 86kb upstream of nuclear receptor coactivator 7 (*NCOA7*, figure 4.7b). *HEY2* is a transcriptional repressor that is part of the Notch pathway, which maintains stem cells. Dysregulation of Notch has been associated

Table 4.4: Five novel risk loci associated with EC at P<$5x10^{-8}$ in the meta-analysis. Position in build19, Allelic OR, 95%CI, P-value and heterogeneity I^2 listed for all histologies and endometrioid-only meta-anlysis.

Locus	SNP	Position	Nearby gene(s)	Ref	Alt	MAF	All histologies			Endometroid histology		
							OR (95%CI)	P	I^2	OR (95%CI)	P	I^2
13q22.1	rs11841589	73,814,891	*KLF5, KLF12*	G	T	0.26	1.15 (1.11-1.21)	$4.83x10^{-11}$	0.19	1.16 (1.10-1.21)	$6.01x10^{-10}$	0.00
6q22.31	rs13328298	126,016,580	*HEY2, NCOA7*	G	A	0.42	1.13 (1.09-1.18)	$3.73x10^{-10}$	0.00	1.15 (1.11-1.20)	$1.02x10^{-11}$	0.00
8q24.21	rs4733613	129,599,278	*MYC*	G	C	0.13	0.84 (0.80-0.89)	$3.09x10^{-9}$	0.00	0.84 (0.79-0.89)	$7.70x10^{-9}$	0.09
15q15.1	rs937213	40,322,124	*EIF2AK, BMF*	T	C	0.42	0.90 (0.86-0.93)	$1.77x10^{-8}$	0.36	0.90 (0.86-0.94)	$2.22x10^{-7}$	0.30
14q32.33	rs2498796	105,243,220	*AKT1, SIVA1*	G	A	0.30	0.89 (0.85-0.93)	$3.55x10^{-8}$	0.00	0.88 (0.85-0.92)	$4.22x10^{-8}$	0.00

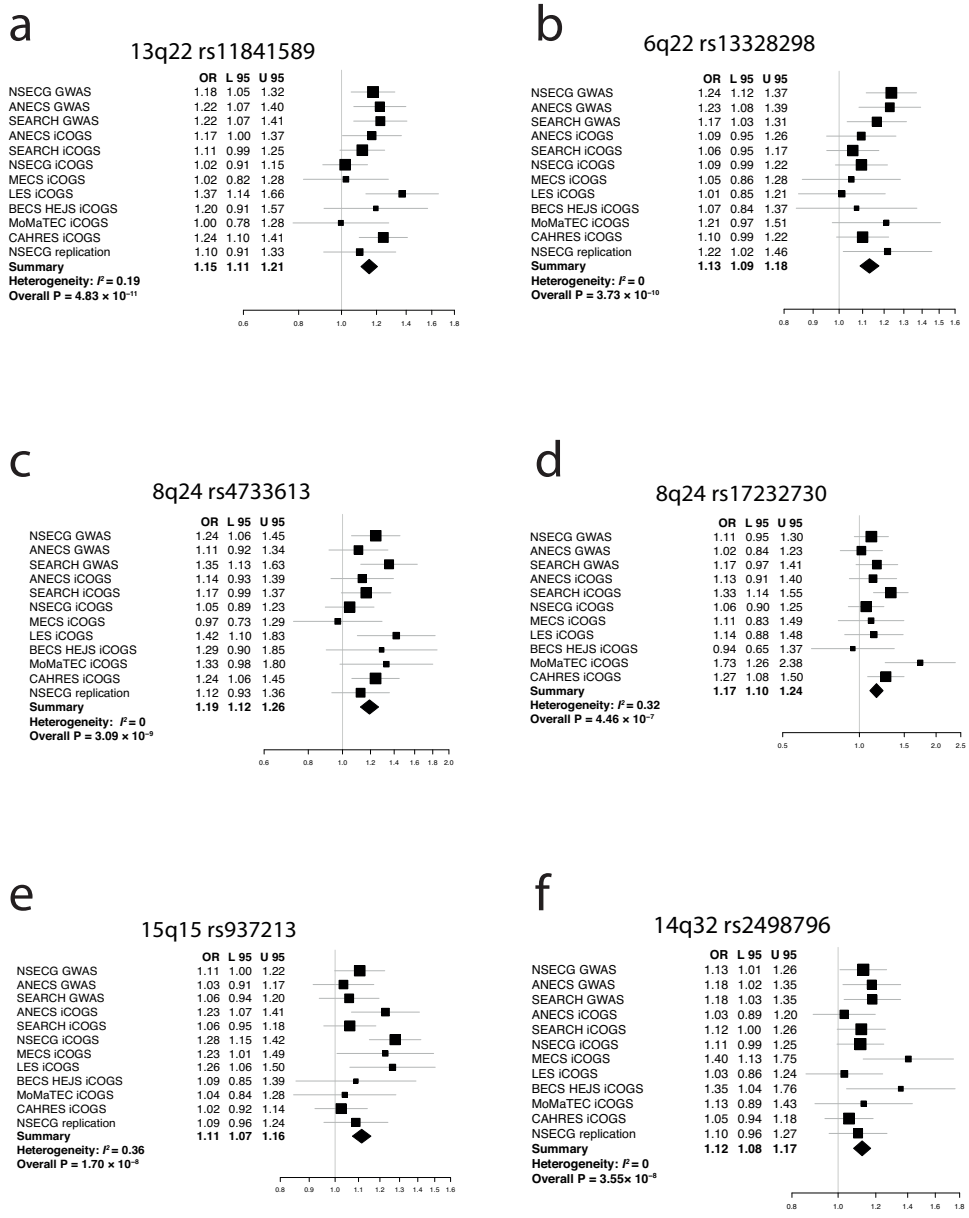

Figure 4.6: Forest plot for novel EC risk loci: odds ratios and 95% confidence intervals of each study of the meta-analysis are listed and shown in the adjacent plot. Black squares represent the point estimates of the odds ratio and have areas proportional to study size. Lines represent 95% confidence intervals. The diamond shows the summary statistic and the overall meta-analysis P-value and heterogeneity statistic is shown. Data from three EC GWAS and eight EC iCOGS and NSECG replication. The I^2 heterogeneity scores suggest that there is no marked difference in effects between studies. The SNPs represented are: a) rs11841589 (13q22), b) rs13328298 (6q22), c) rs4733613 (8q24), d) rs17232730 (8q24, pairwise r^2 0.02 with rs4733613), e) rs937213 (15q15) and f) rs2498796 (14q32).

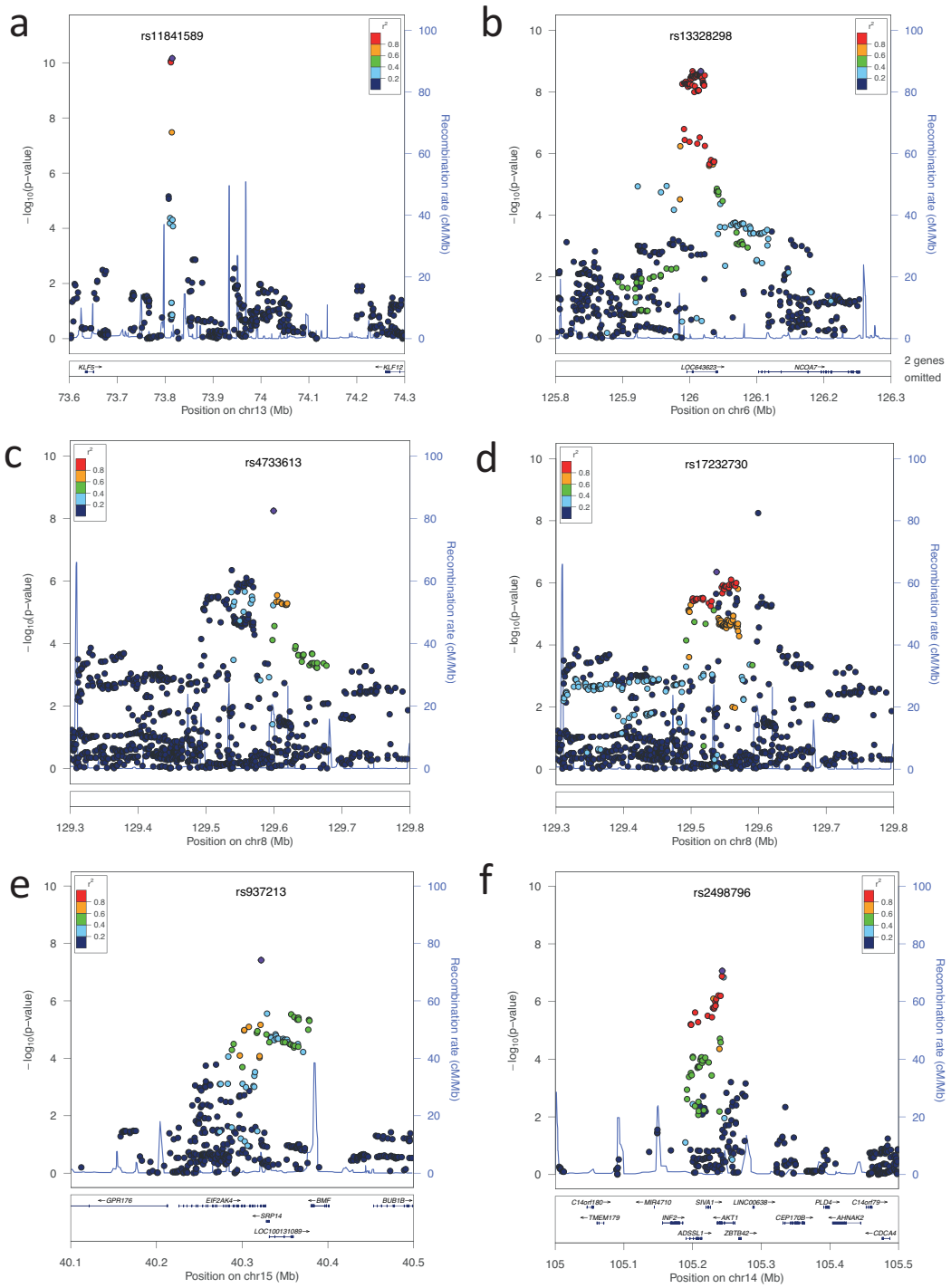

Figure 4.7: Regional association plots for novel risk loci: the -$\log_{10}$ P-values from the meta-analysis and regional imputation for NSECG, ANECS, SEARCH, and eight iCOGS groups are shown for SNPs at: a) 13q22.1, b) 6q22, c) and d) 8q24, e) 15q15 and f) 14q32.33. The top SNP is marked as a purple diamond, and dot colour represents the LD with the top SNP. All plotted SNPs are either genotyped or have an IMPUTE info score of more than 0.9 in all datasets. The blue line shows recombination rate in cM/Mb.

with different cancers (Katoh and Katoh, 2007). *NCOA7*, also known as oestrogen receptor-associated protein 140 (*ERAP140*), modulates the activity of oestrogen receptor via direct binding. Of the new EC loci, this was the only one where restricting the meta-analysis to endometrioid-only histology EC cases yielded a higher significance and effect size despite the reduced sample size (OR=1.15, 95% CI 1.11–1.20, P=1.02x10^{-11}).

Conditional analysis pointed toward a single association signal in this region and the association signal is part of a large LD block spanning 33.6kb (chr6:125,988,964-126,022,602) and 121 SNPs (pairwise r^2>0.8), encompassing most of *LOC643623* and the genomic region just upstream. Functional annotation of 45 highly correlated SNPs ($r^2 \geq 0.98$) identified several SNPs with potential function, either within Dnase1 hypersensitivity sites (DHS), or exhibiting transcription factor binding or altering motifs (i.e. rs2747720 and rs1739363). *HEY2* is expressed in moderate levels in the endometrium and *NCOA7* is expressed in low levels in glandular endometrial tissue (Human Protein Atlas, www.proteinatlas.org). The genotype of SNPs correlated with rs13328298 were not associated with the expression of *HEY2* or *NCOA7* in public datasets (table 4.12).

4.7.3 8q24 rs4733613 telomeric to *MYC*

The third novel locus is at chromosome 8q24; rs4733613 (OR=0.84, 95% CI 0.80–0.89, P=3.09x10^{-9}) lies 55kb-1.9mb telomeric to multiple SNPs identified in other GWAS of solid tumours, including bladder (Kiemeney *et al.*, 2008; Rothman *et al.*, 2010), breast (Easton *et al.*, 2007; Michailidou *et al.*, 2013), colon (Tomlinson *et al.*, 2008; Whiffin *et al.*, 2014), ovarian (Goode *et al.*, 2010) and prostate cancer (Eeles *et al.*, 2009; Gudmundsson *et al.*, 2009). rs4733613 lies in a gene desert; the closest gene *MYC* lies 846kb away and thus it lies further from *MYC* compared to other cancer GWAS SNPs and it is not in LD with any of the other cancer GWAS SNPs at 8q24

Table 4.5: Functional annotation of SNPs in LD (r^2>0.98) with rs13328298 (in bold) using Haploreg and RegulomeDB. Positions in build 19, LD r^2 with lead SNP. Enhancer, enhancer histone marks, Bound, protein bound; RegDB, RegulomeDB score. Annotated SNPS have RegDB score of 3a (TF binding+any motif+DNase peak) to 6 (other). Black spaces represent no data.

Locus	Position	LD	SNP	Enhancer	DHS	Bound	Altered motifs	GENCODE	dbSNP	RegDB
6q22	126,008,435	0.98	rs2747717				3 altered motifs	RP1-293L8.2	intronic	6
6q22	126,009,109	0.98	rs2747718				6 altered motifs	RP1-293L8.2	intronic	6
6q22	126,009,214	0.98	rs2747719				Pax-5,TCF4	RP1-293L8.2	intronic	
6q22	126,009,398	0.98	rs2797158			BATF	3 altered motifs	RP1-293L8.2	intronic	5
6q22	126,009,458	0.98	rs2747720		GM12878	BATF	4 altered motifs	RP1-293L8.2	intronic	3a
6q22	126,009,527	0.98	rs2747721		GM12878	BATF	9 altered motifs	RP1-293L8.2	intronic	4
6q22	126,009,629	0.98	rs2747722				4 altered motifs	RP1-293L8.2	intronic	
6q22	126,010,086	0.99	rs35069021				5 altered motifs	RP1-293L8.2	intronic	6
6q22	126,010,116	0.99	rs2797160				6 altered motifs	RP1-293L8.2	intronic	
6q22	126,010,904	0.99	rs34649676				CIZ,HNF1,Pou5f1	RP1-293L8.2	intronic	6
6q22	126,011,079	0.99	rs1739368				5 altered motifs	RP1-293L8.2	intronic	6
6q22	126,011,291	0.99	rs1739371					RP1-293L8.2	intronic	
6q22	126,011,325	0.99	rs1739372				4 altered motifs	RP1-293L8.2	intronic	
6q22	126,011,381	0.99	rs2797162				Pou3f2,p300	RP1-293L8.2	intronic	
6q22	126,011,509	0.99	rs1739373				3 altered motifs	RP1-293L8.2	intronic	6
6q22	126,012,262	0.99	rs1739378				Hsf,TEF-1	RP1-293L8.2	intronic	
6q22	126,014,157	0.99	rs1739347				Ascl2,Myc,Myf	RP1-293L8.2	intronic	6
6q22	126,014,573	0.99	rs1739348				PU.1	RP1-293L8.2	intronic	
6q22	126,014,916	0.98	rs1612249					RP1-293L8.2	intronic	4
6q22	126,014,984	0.99	rs1739349				Ik-3,Pax-4,Sox	RP1-293L8.2	intronic	5
6q22	126,015,057	0.99	rs1578793				4 altered motifs	RP1-293L8.2	intronic	5
6q22	126,015,469	0.99	rs1578794				4 altered motifs	RP1-293L8.2	intronic	5
6q22	126,015,954	0.99	rs6927161				3 altered motifs	RP1-293L8.2	intronic	6
6q22	126,016,003	0.99	rs6904992				Myc,Osr	RP1-293L8.2	intronic	
6q22	126,016,499	0.99	rs12717178				6 altered motifs	RP1-293L8.2	intronic	
6q22	126,016,580	-	**rs13328298**				14 altered motifs	RP1-293L8.2	intronic	
6q22	126,016,951	0.99	rs6933302				PRDM1,YY1	RP1-293L8.2	intronic	
6q22	126,017,029	0.99	rs6933471				3 altered motifs	RP1-293L8.2	intronic	6
6q22	126,017,141	0.98	rs6910786				9 altered motifs	RP1-293L8.2	intronic	6
6q22	126,017,155	0.99	rs6910933				4 altered motifs	RP1-293L8.2	intronic	6
6q22	126,017,481	0.99	rs6934435				16 altered motifs	RP1-293L8.2	intronic	6
6q22	126,017,551	0.99	rs201940333				10 altered motifs	RP1-293L8.2	intronic	6
6q22	126,017,691	0.99	rs1777226		H1-hESC		3 altered motifs	RP1-293L8.2	intronic	5
6q22	126,017,808	0.99	rs1739354				3 altered motifs	RP1-293L8.2	intronic	
6q22	126,018,114	0.99	rs1739355					RP1-293L8.2	intronic	6
6q22	126,018,270	0.99	rs1777225				E2A,LBP-1	RP1-293L8.2	intronic	
6q22	126,019,736	0.98	rs1739358					RP1-293L8.2	intronic	
6q22	126,019,738	0.98	rs77678056					RP1-293L8.2	intronic	
6q22	126,019,768	0.98	rs78602343				Foxp1,Nkx3	RP1-293L8.2	intronic	
6q22	126,020,703	0.98	rs1739362				Foxj2,Maf	RP1-293L8.2	intronic	
6q22	126,020,980	0.98	rs1739363	H1	H1-hESC, CD20+	EBF1	Pou3f3	RP1-293L8.2	intronic	4
6q22	126,021,030	0.98	rs1777222	H1	H1-hESC, CD20+	EBF1	Pax-4	RP1-293L8.2	intronic	4
6q22	126,021,277	0.98	rs984040				Foxp1	RP1-293L8.2	intronic	
6q22	126,021,328	0.98	rs984041				Pou2f2	RP1-293L8.2	intronic	
6q22	126,021,435	0.98	rs926853				GATA	RP1-293L8.2	intronic	

($r^2 < 0.02$).

Stepwise conditional logistic regression revealed a possible second independent EC signal (P_{meta}=4.46x10^{-7}, P_{cond}=1.29x10^{-5}) at rs17232730 (figure 4.7d, 62kb centromeric to rs4733613, pairwise r^2=0.02). This SNP showed a higher significance in endometrioid histology only analysis (P=2.22x10^{-7}). rs17232730 is in moderate LD with the ovarian cancer SNP rs10088218 (r^2=0.43), with both cancers sharing the same risk allele (figure 4.6). No further significant EC association signal remained at this locus after conditioning on both rs4733613 and rs17232730 (P_{cond} >0.04 for all SNPs), and there was specifically no evidence that the other known cancer SNPs in the region influence EC risk. The signal at rs4733613 remains highly significant in endometrioid only analysis (P=3.09x10^{-9}).

The plausibility of two independent signals at this locus is strengthened by the fact that i) the regional association plot for 8q24 (figure 4.7c and d) shows evidence for two distinct peaks which are not in LD and ii) the forest plots for rs4733613 and rs17232730 (figures 4.6c and d) show that there is a consistent risk allele for each SNP, but the effect sizes vary when one compares the two SNPs in the same study.

The functional annotation of rs4733613, rs17232730 and correlated SNPs showed that they may be of regulatory importance in terms of DHS, enhancer histone modifications, and transcription factor binding (i.e. rs1516977, RegulomeDB score 2b). A SNP moderately correlated with rs4733613 was marginally assocaited with the expression of *MYC* and *POU5F1B* in normal but not tumour endometrial data(table 4.12).

4.7.4 15q15 rs937213 near *BMF*

The fourth new locus in 15q15, rs937213 (OR=0.90, 95% CI 0.86–0.93, P=1.77x10^{-8}, figure 4.6e) lies within the intronic region of eukaryotic translation initiation factor 2-alpha kinase 4 (*EIF2AK4*) and 58kb downstream of apoptotic regulator Bcl-2-

Table 4.6: Conditional analysis for 8q24 locus. Positions in build 19 Ref, reference; Alt, alternative allele; MAF, minor allele frequency. There is evidence for a second independent signal at rs17232730 which is in partial LD with rs10088218 (marked with asterix), which is associated with ovarian cancer (all subtypes), with the association being more significant for cancers of serous histology.

SNP	Position	Ref	Alt	MAF	Pairwise r2 with		All histology meta-analysis		Conditioning on rs4733613		Conditioning on rs17232730	
					rs4733613	rs17232730	Allelic OR (95%CI)	P	Allelic OR (95%CI)	P	Allelic OR (95%CI)	P
rs4733613	129599278	G	C	0.126	-	0.017	0.84 (0.79-0.89)	5.64E-09		-	0.86 (0.81-0.91)	2.32 × 10-7
rs17232730	129537746	G	C	0.123	0.017	-	1.17 (1.10-1.24)	4.46E-07	1.14 (1.08-1.22)	1.29E-05	-	-
rs10088218*	129543949	G	A	0.128	0.016	0.43	1.14 (1.07-1.20)	1.65E-05	1.12 (1.05-1.18)	2.92E-04	1.01 (0.91-1.12)	0.818

Table 4.7: Functional annotation of lead SNP rs4733613 (in bold, which has no SNP with pairwise LD r^2>0.70) and SNPs in LD (r^2>0.70) with rs17232730 (in bold) using Haploreg and RegulomeDB. Positions in build 19, LD r^2 with lead SNP. Enhancer, enhancer histone marks, Bound, protein bound; PhastCons, conservation in vertebrates; RegDB, RegulomeDB score. Annotated SNPS have RegDB score of 2b (TF binding+any motif+DNase Footprint+DNase peak) to 6 (other). Black spaces represent no data. There are no SNPs with r^2>0.70 for rs4733613.

Locus	Position	LD	SNP	Enhancer	DHS	Bound	Altered Motifs	GENCODE	RegDB	PhastCons
8q24	129,599,278	-	**rs4733613**	HepG2				22kb 5' of RP11-89M16.1		
8q24	129,497,735	0.71	rs77156523				Znf143	RP11-89M16.1		
8q24	129,499,432	0.78	rs77241108				4 altered motifs	RP11-89M16.1	6	
8q24	129,501,028	0.79	rs113834169	3 cell types	Th1	USF2,USF1	4 altered motifs	RP11-89M16.1	3a	
8q24	129,501,760	0.79	rs76384007	HepG2		MAFK	3 altered motifs	RP11-89M16.1	3a	
8q24	129,502,569	0.79	rs10505515	HepG2	HRPEpiC		Foxa,Mef2	RP11-89M16.1	5	
8q24	129,503,700	0.79	rs17805799				10 altered motifs	RP11-89M16.1	6	
8q24	129,506,822	0.79	rs77683961	HepG2				RP11-89M16.1	5	
8q24	129,507,019	0.79	rs75706585				Nr2e3	RP11-89M16.1	5	
8q24	129,508,906	0.79	rs17231703					RP11-89M16.1		
8q24	129,514,190	0.79	rs17805924	HepG2			SP1	RP11-89M16.1	5	
8q24	129,517,137	0.79	rs17231857				4 altered motifs	RP11-89M16.1	5	
8q24	129,517,554	0.79	rs17231948				Ets	RP11-89M16.1		
8q24	129,517,938	0.79	rs76388889				Cdx,DMRT1	RP11-89M16.1		
8q24	129,526,064	0.81	rs17232172	3 cell types			E2F,Smad3	RP11-89M16.1		
8q24	129,529,129	0.81	rs1878378				HNF4	RP11-89M16.1		
8q24	129,530,492	0.81	rs1400485				Hoxb13	RP11-89M16.1		
8q24	129,537,624	0.88	rs79593486	NHLF			10 altered motifs	RP11-89M16.1	3b	
8q24	129,537,746	-	**rs17232730**	NHLF	8 cell types			RP11-89M16.1	4	
8q24	129,545,257	0.76	rs1516973	HepG2, HSMM				RP11-89M16.1		
8q24	129,546,233	0.76	rs1607120	NHLF	7 cell types		6 altered motifs	RP11-89M16.1	5	
8q24	129,546,982	0.76	rs10086718				11 altered motifs	RP11-89M16.1	6	
8q24	129,548,654	0.76	rs74866331				TLX1::NFIC	RP11-89M16.1	5	
8q24	129,549,261	0.76	rs1516977			ERALPHA_A	7 altered motifs	RP11-89M16.1	2b	Score=669; lod=696
8q24	129,550,041	0.76	rs10112057		Th1,Th2		4 altered motifs	RP11-89M16.1	5	
8q24	129,551,752	0.76	rs10088873	HSMM, HepG2			Pbx-1	RP11-89M16.1		
8q24	129,552,140	0.76	rs10098999	4 cell types	4 cell types	HNF4A,RXRA	4 altered motifs	RP11-89M16.1	4	
8q24	129,555,724	0.72	rs2138636	NHLF, HSMM	9 cell types		5 altered motifs	RP11-89M16.1	4	
8q24	129,555,928	0.73	rs17807904	HSMM	3 cell types		3 altered motifs	RP11-89M16.1	5	
8q24	129,556,514	0.73	rs10087367				7 altered motifs	RP11-89M16.1		
8q24	129,559,228	0.73	rs10098821				7 altered motifs	RP11-89M16.1		
8q24	129,560,314	0.72	rs10102835				17 altered motifs	RP11-89M16.1	6	
8q24	129,562,824	0.73	rs76076434				5 altered motifs	RP11-89M16.1	6	
8q24	129,563,362	0.72	rs78830272				5 altered motifs	RP11-89M16.1	6	
8q24	129,565,545	0.73	rs75370373	HepG2, K562	3 cell types	E2F6,MAX	3 altered motifs	RP11-89M16.1	3a	
8q24	129,566,898	0.73	rs7007074	3 cell types			11 altered motifs	RP11-89M16.1	5	Score=309; lod=24

modifying factor (*BMF*). EIF2AK4 is a kinase that phosphorylates the alpha subunit of eukaryotic translation initiation factor-2 and downregulates protein synthesis at times of cellular stress (Berlanga *et al.*, 1999). *BMF* encodes a protein that is part of the BCL2 protein family, which form hetero- or homo-dimeric apoptotic regulators and are involved in a wide range of cellular activities. The eQTL query SNP rs2412456 (r^2=0.58 with rs937213) was associated with the expression of *EIF2AK4* in tumour (P_{TCGA}=0.02) and *BMF* in normal endometrial tissue (P_{GTEx}=0.006, table 4.12). *EIF2AK4* is moderately expressed in endometrial tissue and *BMF* is moderately to highly expressed in glandular endometrial tissue (Human Protein Atlas, www.proteinatlas.org).

4.7.5 14q32 rs2498796 *AKT1*

The fifth locus, rs2498796 (OR=0.89, 95% CI 0.85–0.93, P=3.55x10^{-8},figure 4.6f) at 14q32, lies within the intronic region of the serine/threonine kinase gene *AKT1*, an oncogene of the PI3K/AKT/MTOR signaling pathway. This pathway affects cell survival and proliferation (Cantley, 2002) and is activated in endometrial tumours (Slomovitz and Coleman, 2012), where it is associated with aggressive disease and poor prognosis (Cohen *et al.*, 2010; Salvesen *et al.*, 2009; Shoji *et al.*, 2009). The regional association plot (figure 4.7f) reveals a series of SNPs in LD (r^2>0.8) in a 35kb region extending downstream of *AKT1* and including nearby *SIVA1*, of which three SNPs show strong evidence of functional activity (table 4.9, RegulomeDB score 1b) including transcrition factor binding, DHS, and eQTL effects for *SIVA1* in lymphoblastoid cell lines. SIVA1 is an apoptosis regulatory protein that inhibits p53 activity via HDM2-mediated ubiquitination (Du *et al.*, 2009). SIVA1 also enhances epithelial mesenchymal transition, which allows epithelial cells to acquire motility and invasiveness (Li *et al.*, 2011). *AKT1* is highly expressed in endometrial tissue while the corresponding data for *SIVA1* are not available (Human Protein Atlas,

Table 4.8: Functional annotation of SNPs in LD (r^2>0.70) with rs937213 (in bold) using Haploreg and RegulomeDB. Positions in build 19, LD r^2 with lead SNP. Enhancer, enhancer histone marks, Bound, protein bound; RegDB, RegulomeDB score. RegDB score of 5: TF binding or DNase peak. Black spaces represent no data.

Locus	Position	LD	SNP	Enhancer	DHS	Altered Motifs	GENCODE	dbSNP	RegDB
15q15	40,290,475	0.71	rs199884855			19 altered motifs	*EIF2AK4*	intronic	
15q15	40,302,243	0.72	rs3816900	HSMM		FXR,HNF1,RORalpha1	*EIF2AK4*	intronic	
15q15	40,302,441	0.72	rs12592831	HSMM	BE2_C	Myf	*EIF2AK4*	intronic	5
15q15	40,303,842	0.72	rs2412464	HSMM		Sin3Ak-20	*EIF2AK4*	intronic	
15q15	40,322,124	-	**rs937213**		FibroP	7 altered motifs	*EIF2AK4*	intronic	5

www.proteinatlas.org).

4.8 15q21 *CYP19A1* at genome-wide significance

CYP19A1 encodes for aromatase and previous candidate-gene studies have assessed EC risk in relation to *CYP19A1* variants (section 1.5.2). The rs727479-A allele has been associated with EC (Setiawan *et al.*, 2009) and increased circulating levels of oestrogen (Haiman *et al.*, 2007; Prescott *et al.*, 2012). My meta-analysis confirmed this SNP to be genome-wide significant for EC for the first time (OR=1.17, 95%CI 1.12–1.22, P=6.35x10^{-14}). The most significant SNP in this region in this meta-analysis is the *CYP19A1* intronic SNP rs2414098 (OR=1.17, 95%C 1.12–1.22, P=4.51x10^{-14}, pairwise LD with rs727479 r^2=0.85) and the forest plot (figure 4.8) shows that the rs2414098-C allele confers EC risk in all datasets (heterogeneity I^2=0). The regional association plot is displayed in figure 4.9 and conditional analyis confirms the presence of a single association signal (P>0.05 for all SNPs after conditioning for rs2414098). This SNP remains highly significant in endometrioid-only analyis (P=2.48x10^{-13}). The functional annotation in table 4.10 provide some evidence for enhancer histone modifications and transcription factor binding in these SNPs.

4.9 Replication of 17q12 *HNF1B*

The previously published EC GWAS found a single genome-wide significant hit at chromosome 17q12, rs4430796 in the second intron of *HNF1B* (Spurdle *et al.*, 2011). This study was based on ANECS and SEARCH samples, and with the addition of NSECG and iCOGS studies, rs4430796 was replicated at a higher level of significance (OR= 0.84, 95% CI 0.81–0.88, P=5.95x10^{-18}). It is of interest that this tag-SNP has also been associated with serous and clear-cell ovarian cancer (Pharoah *et al.*, 2013; Shen *et al.*, 2013), prostate cancer (Berndt *et al.*, 2011) and type 2 diabetes

Table 4.9: Functional annotation of SNPs in LD (r^2>0.70) with rs2498796 (in bold) using Haploreg and RegulomeDB. Positions in build 19, LD r^2 with lead SNP. Promoter and Enhancer, promoter and enhancer histone marks, Bound, protein bound; RegDB, RegulomeDB score. Annotated SNPS have RegDB score of 1b (eQTL+TF binding+any motif+DNase Footprint+DNase peak) to 6 (other). Black spaces represent no data.

Locus	Position	LD	SNP	Promoter	Enhancer	DHS	Bound	eQTL	Altered Motifs	GENCODE	dbSNP	RegDB
14q32	105,192,685	0.74	rs72715972	HepG2		3 cell types			5 altered motifs	*ADSSL1*	intronic	2b
14q32	105,196,230	0.79	rs80097179		K562		4 bound proteins		Znf143	*ADSSL1*	missense	4
14q32	105,197,354	0.78	rs60876857		K562				Roaz	*ADSSL1*	intronic	5
14q32	105,197,756	0.79	rs60798007		HepG2	40 cell types	9 bound proteins		9 altered motifs	*ADSSL1*	intronic	4
14q32	105,197,846	0.77	rs57098433		HepG2	30 cell types	8 bound proteins		NF-E2,PLZF	*ADSSL1*	intronic	4
14q32	105,203,678	0.79	rs7160733		Huvec, HepG2				SP1,ZEB1	*ADSSL1*	intronic	
14q32	105,208,057	0.8	rs4983384					Cortex, LCL	5 altered motifs	*ADSSL1*	intronic	1f
14q32	105,218,333	0.76	rs8006580		3 cell types	GM06990	SETDB1		Hic1			4
14q32	105,222,037	0.79	rs1132975	HepG2	GM12878			Cortex, LCL	RXRA	*SIVA1*	synonymous	6
14q32	105,226,075	0.79	rs45607139						17 altered motifs	*SIVA1*		2b
14q32	105,228,216	0.8	rs4983549						7 altered motifs	*SIVA1*		6
14q32	105,229,646	0.85	rs7158655	HepG2					13 altered motifs	*SIVA1*		5
14q32	105,230,225	0.84	rs144188214	HepG2					34 altered motifs	*SIVA1*		6
14q32	105,230,391	0.74	rs28368454		HepG2				HMG-IY,NRSF	*SIVA1*		6
14q32	105,231,196	0.79	rs66464514	HepG2					RP58,Sox,TAL1	*SIVA1*		6
14q32	105,231,640	0.77	rs4983550	HepG2					Hltf,Sox	*SIVA1*		
14q32	105,233,095	0.85	rs2498804	HepG2	H1, K562	HepG2	POL2	Cortex, LCL	Nkx2,Nkx3	*SIVA1*		1b
14q32	105,233,408	0.85	rs2498803		HepG2, K562	H7-hESC	7 bound proteins		3 altered motifs	*SIVA1*		4
14q32	105,233,421	0.85	rs2494730		HepG2, K562	H7-hESC	7 bound proteins		6 altered motifs	*SIVA1*		1b
14q32	105,234,442	0.78	rs2498802		HepG2	Fibrobl		Cortex	5 altered motifs	*SIVA1*		1f
14q32	105,235,558	0.79	rs2498801	HepG2	K562, H1	6 cell types	3 bound proteins		5 altered motifs			4
14q32	105,237,680	0.82	rs2494731				SETDB1	Cortex, Liver, LCL	Pax-5,RXRA	*AKT1*	intronic	1f
14q32	105,238,591	0.87	rs55839843			PrEC			4 altered motifs	*AKT1*	intronic	
14q32	105,240,784	0.89	rs2494733				ERALPHA_A,CTCF		8 altered motifs	*AKT1*	intronic	2b
14q32	105,242,228	0.85	rs2498797			4 cell types	3 bound proteins		6 altered motifs	*AKT1*	intronic	2b
14q32	105,242,831	0.98	rs3001371				ZNF263	LCL	RREB-1	*AKT1*	intronic	1b
14q32	105,242,966	0.77	rs2494735				ZNF263		5 altered motifs	*AKT1*	intronic	4
14q32	105,243,220	-	**rs2498796**						4 altered motifs	*AKT1*	intronic	5

Table 4.10: Functional annotation of SNPs in LD (r^2>0.80) with rs2414098 (in bold) using Haploreg and RegulomeDB. Positions in build 19, LD r^2 with lead SNP. Enhancer, enhancer histone marks, Bound, protein bound; RegDB, RegulomeDB score. Annotated SNPS have RegDB score of 3a (TF binding+any motif+DNase peak) to 6 (other). Black spaces represent no data.

Locus	Position	LD	SNP	Enhancer	DHS	Bound	eQTL	Altered Motifs	GENCODE	dbSNP	RegDB
15q21	51,517,100	0.84	rs12595627	4 cell types	Melano				*CYP19A1*	intronic	5
15q21	51,519,276	0.85	rs4775935	K562, HSMM				Bbx,Mef2,PLZF	*CYP19A1*	intronic	
15q21	51,524,292	0.85	rs2414095						*CYP19A1*	intronic	6
15q21	51,525,173	0.85	rs12592697		Osteobl,HCFaa				*CYP19A1*	intronic	5
15q21	51,529,835	0.85	rs2414097		WERI-Rb-1			4 altered motifs	*CYP19A1*	intronic	5
15q21	51,533,736	0.85	rs7173595					5 altered motifs	*CYP19A1*	intronic	6
15q21	51,534,055	0.85	rs7175531					4 altered motifs	*CYP19A1*	intronic	3a
15q21	51,534,547	0.85	rs724779					4 altered motifs	*CYP19A1*	intronic	6
15q21	51,536,141	0.83	rs10459592	GM12878, NHEK, HMEC	33 cell types	5 bound proteins		TFII-I	*CYP19A1*	intronic	4
15q21	51,537,127	0.83	rs10851499					Mef2,Mrg1::Hoxa9	*CYP19A1*	intronic	4
15q21	51,537,806	-	**rs2414098**					CEBPA,Pax-4	*CYP19A1*	intronic	6
15q21	51,538,810	0.81	rs12899068					ERalpha-a,GR,Pax-6	*CYP19A1*	intronic	

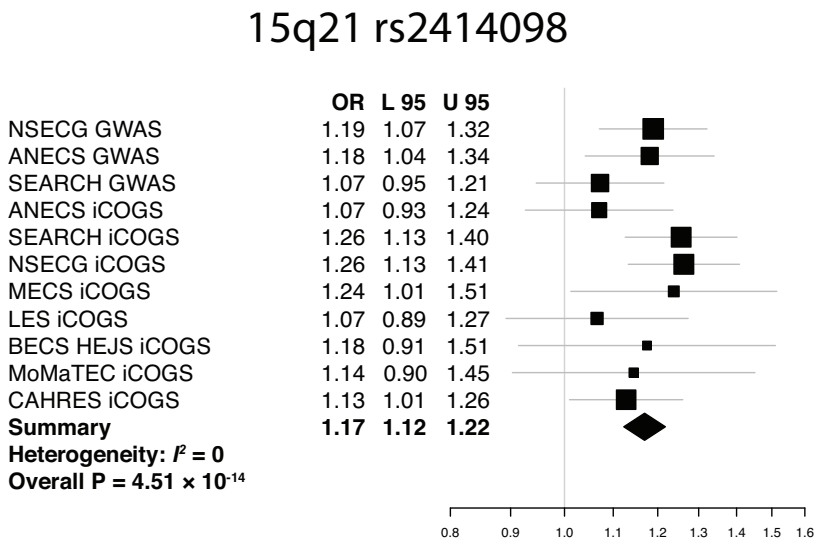

Figure 4.8: Forest plot for association between cancer risk and chromosome 15q21 rs2414098 genotype in three EC GWAS and eight EC iCOGS. Black squares represent the point estimates of the odds ratio and have areas proportional to study size. Lines represent 95% confidence intervals. The diamond shows the summary statistic. The overall meta-analysis P-value and heterogeneity statistic are shown.

(Gudmundsson *et al.*, 2007). This meta-analysis found that the most significant SNP in the 17q12 locus is rs11263763 (OR=0.83, 95%CI=0.80-0.87, P=2.78x10^{-19}, figure 4.10), which lies in the first intron of *HNF1B*. rs11263763 is 5,525bp from rs4430796 and is in high LD (r^2=0.94). It remains highly significant in endometrioid-only histology analysis (OR=0.83, 95%CI=0.80-0.87, P=6.51x10^{-17}). The regional association plot of this region is shown in figure 4.11 and there is no residual association signal (P>0.05) after conditioning on rs11263763. Functional annotation of rs11263763 and correlated SNPs provides evidence of transcription factor binding and promoter histone marks (table 4.11).

4.10 eQTL searches in endometrial tissue

Functional annotation tables listed in previous sections have included eQTL searches in public databases of all tissues. The GTEx and TCGA databases provide endometrial tissue genotype and expression data (see section 5.3) and SNPs in the highest

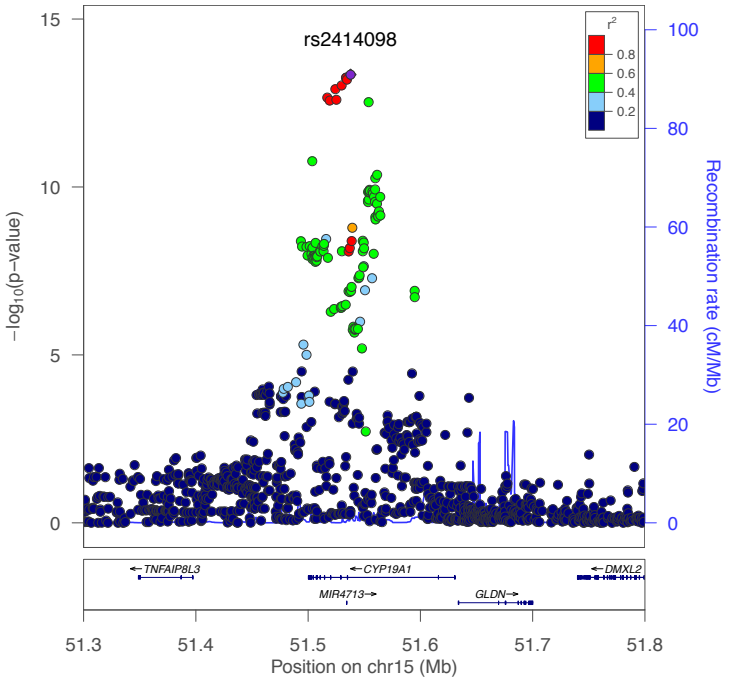

Figure 4.9: Regional association plot of chromosome 15q21 around rs2414098. The top SNP rs2414098 is marked as a purple diamond, and the dot colour represents the LD with the top SNP. Y-axis is -$\log_{10}$ P-values from imputation and meta-analysis for the eight EC and CRC GWAS. The blue line shows recombination rates in cM/Mb.

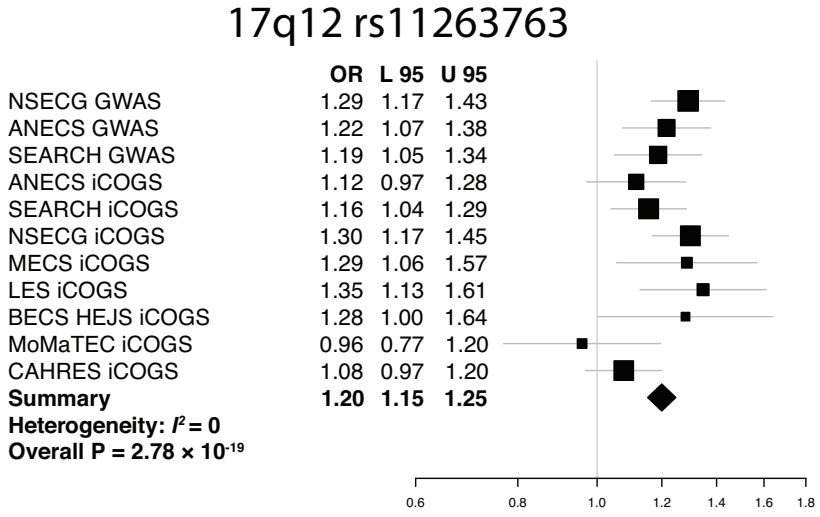

Figure 4.10: Forest plot for association between cancer risk and chromosome 17q12 rs11263763 genotype in three EC GWAS and eight EC iCOGS. Black squares represent the point estimate of the odds ratio and have areas proportional to study size. Lines represent 95% confidence intervals. The diamond shows the summary statistic. The overall meta-analysis P-value and heterogeneity statistic is shown.

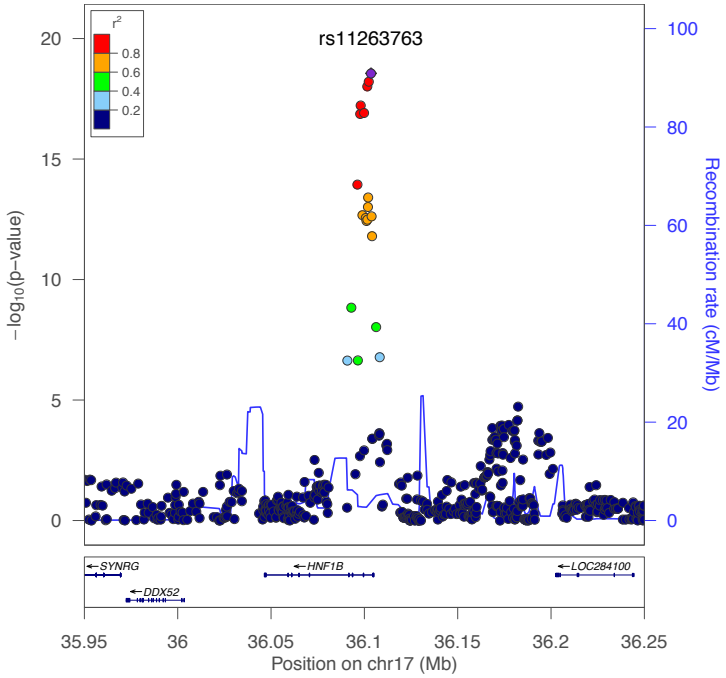

Figure 4.11: Regional association plot of chromosome 17q12 around rs11263763. The top SNP rs11263763 is marked as a purple diamond, and the dot colour represents the LD with the top SNP. Y-axis is -$\log_{10}$ P-values from imputation and meta-analysis for the eight EC and CRC GWAS. The blue line shows recombination rates in cM/Mb.

Table 4.11: Functional annotation of SNPs in LD (r^2>0.80) with rs11263763 (in bold) using Haploreg and RegulomeDB. Positions in build 19, LD r^2 with lead SNP. Enhancer, enhancer histone marks, Bound, protein bound; RegDB, RegulomeDB score. Annotated SNPS have RegDB score of 2b (TF binding+any motif+DNase Footprint+DNase peak) to 6 (other). Black spaces represent no data.

Locus	Position	LD	SNP	Promoter	DHS	Bound	Altered Motifs	GENCODE	dbSNP	RegDB
17q12	36,096,300	0.8	rs2005705				4 altered motifs	*HNF1B*	intronic	6
17q12	36,097,775	0.93	rs11263761				6 altered motifs	*HNF1B*	intronic	
17q12	36,098,040	0.94	rs4430796		Th1	STAT3	Maf	*HNF1B*	intronic	4
17q12	36,099,840	0.96	rs11651755		Osteobl		Crx,EWSR1-FLI1	*HNF1B*	intronic	5
17q12	36,099,952	0.97	rs10908278		Osteobl		43 altered motifs	*HNF1B*	intronic	5
17q12	36,101,586	0.97	rs8064454	HepG2, GM12878	HSMMtube,CD34+_Mobilized,HFF		Arid5a,CEBPB,Mef2	*HNF1B*	intronic	2b
17q12	36,102,381	0.98	rs11651052	8 cell types	26 cell types	CTBP2,SUZ12	NF-kappaB	*HNF1B*	intronic	3a
17q12	36,103,565	-	**rs11263763**	5 cell types	18 cell types	8 bound proteins		*HNF1B*	intronic	3a

LD with the GWAS SNP were queried to see if there was any association between genotype and expression of nearby genes (table 4.12). There are 32 and 54 samples available in GTEx and TCGA respectively, and this small sample size provides a very limited power to detect eQTLs. The most significant association was seen between rs11841589 and *KLF5* and these results are futher discussed in chapter 5.

4.11 Discussion

This meta-analysis has identified five novel EC risk loci at genome-wide significance at chromosomes 13q22.1 (rs11841589, near *KLF5* and *KLF12*), 6q22.31 (rs13328298, in *LOC643623* and near *HEY2* and *NCOA7*), 8q24.21 (rs4733613, telomeric to *MYC*), 15q15.1 (rs937213, in *EIF2AK4*, near *BMF*) and 14q32.33 (rs2498796, in *AKT1*, near *SIVA1*). I found a further independent 8q24.21 signal (rs17232730) . Several of the genes within the EC SNP regions have known or highly likely roles in tumourigenesis, including *KLF5*, *HEY2*, *NCOA7*, *BMF*, *AKT1* and *SIVA1*. Functional annotation of SNPs in LD with novel risk loci reveals regulatory potential including histone marks, transcription factor binding, DHS and altered motifs. Futher genotyping and fine-mapping may bring us closer to the causal variants. Endometrial tissue eQTL analysis would benefit from a larger number of samples to validate the association of the 13q22 locus with *KLF5* expression and a more comprehensive analysis of genome-wide significant endometrial eQTLs. This may be particularly helpful at the the 8q24 locus, which is located more than 800kb away from *MYC*, its closest gene.

Of the novel risk loci, rs4733613 on chromosome 8q24 has the lowest MAF of 0.12. This concurs with the expectations from the power calculations (section 4.3) in that this meta-analysis would have only 21% power to detect genome-wide significant associations for SNPs with a MAF of 0.1 (assuming an OR of 1.15). Further studies with additional EC cases and controls would provide additional power to detect

Table 4.12: eQTL analysis in endometrial tissue for seven genome-wide significant loci. The top GWAS SNP is listed along with the eQTL query SNP, and LD represents the pairwise LD in r^2 between the GWAS SNP and query SNP. Nearby genes are queried as listed. The P-values are listed for GTEx (normal endometrial tissue) and TCGA (tumour tissue). "-" denotes that expression level fo this gene is missing for this dataset.

Locus	GWAS SNP	Query SNP	LD	Gene	GTEx P	TCGA P
13q22.1	rs11841589	rs11841589	1	*KLF5*	0.02	0.0049
				KLF12	1.00	0.74
				PIBF1	0.90	-
				DIS3	0.60	0.56
				BORA	0.80	-
6q22.31	rs13328298	rs984040	0.98	*HEY2*	0.80	0.97
				NCOA7	0.80	0.81
				HDDC2	0.80	0.12
				TPD52L1	0.20	0.91
				HINT3	0.90	0.61
				TRMT11	0.50	0.75
8q24.21	rs4733613	rs16903272	0.6	*MYC*	0.03	0.60
				PVT1	0.60	-
				POU5F1B	0.05	-
	rs17232730	rs1400485	0.81	*MYC*	0.50	0.32
				PVT1	0.60	-
				POU5F1B	0.60	-
15q15.1	rs937213	rs2412456	0.58	*EIF2AK4*	0.50	0.02
				SRP14	0.40	0.23
				BMF	0.01	0.55
				BUB1B	0.60	0.10
				PAK6	-	0.11
				GPR176	0.70	0.40
14q32.33	rs2498796	rs2494731	0.82	*AKT1*	0.20	0.15
				SIVA1	0.70	0.62
				ZBTB42	0.20	-
15q21.2	rs2414098	rs10459592	0.83	*CYP19A1*	-	0.46
				GLDN	0.50	0.14
				DMXL2	0.80	0.39
17q12	rs11263763	rs11651755	0.96	*HNF1B*	0.50	0.71
				DDX52	0.90	0.94

associations, particularly any SNPs with MAF<0.1 and a modest effect size.

CYP19A1 has been confirmed to be associated wtih EC at a genome-wide significant level for the first time. *CYP19A1* was previously identified as being a candidate gene because it encodes for aromatase, which converts testosterone to estradiol. After the cessation of ovarian oestrogen production in menopause, endogenous oestrogens are primarily synthesized via the aromatase-mediated peripheral conversion of androgens in adipose tissue. Prolonged oestrogen exposure has long been seen as a risk factor for EC. Together with studies linking rs727479-A to increased circulating levels of estradiol in post-menopausal women, it is highly likely that variants in *CYP19A1* influence the risk of developing EC by affecting circulating oestradiol as an intermediate phenotype.

The EC association at *HNF1B* has been replicated in this analysis. The most significant SNP in this region is rs11263763 (P=2.78x10^{-19}), located at the first intron of *HNF1B*. This SNP was also found to be the most significant SNP in the fine-mapping analysis conducted by our collaborators, which makes use of a similar dataset with the addition of Asian GWAS samples (P=8.4x10^{-14}, Painter *et al.*, 2014). This SNP remains highly significant in endometrioid histology analysis (P=6.51x10^{-17}), but unlike the NSECG GWAS analysis, the significance level is slightly reduced when only using samples with endometrioid histology. The rs11263763-G EC protective allele was associated with a lower level of *HNF1B* expression (P=0.013) and falls within an extended promoter region defined by ENCODE. The *HNF1B* locus is a member of the homeodomain-containing superfamily of transcription factors and has been associated with serous and clear cell ovarian cancer (Pharoah *et al.*, 2013; Shen *et al.*, 2013), prostate cancer (Berndt *et al.*, 2011) and type 2 diabetes (Gudmundsson *et al.*, 2007). The minor alleles of *HNF1B* GWAS SNPs confers risk (serous ovarian cancer, type 2 diabetes) and protection (EC, prostate cancer, clear-cell ovarian cancer) in different phenotypes suggesting a complex genetic architecture in this region. It

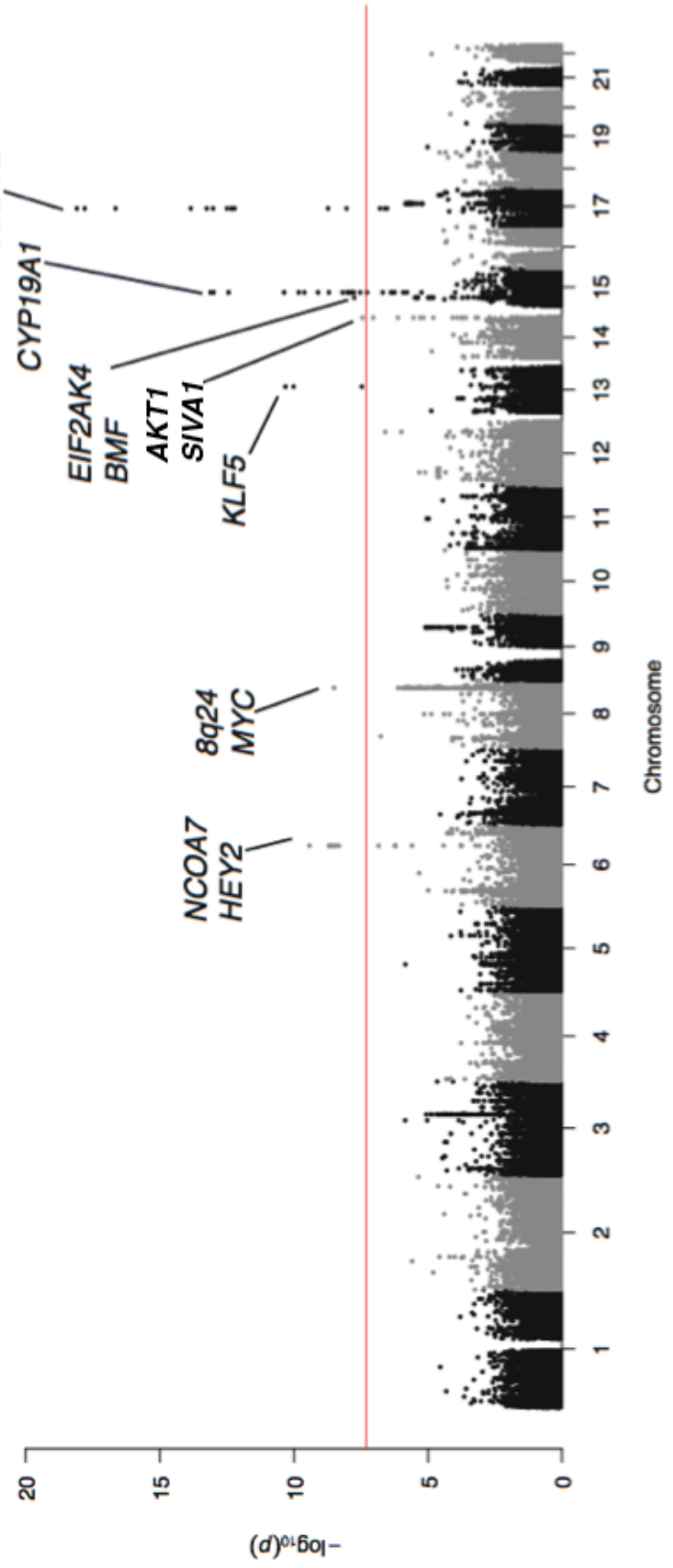

Figure 4.12: EC meta-analysis Manhattan plot: Y-axis is -$\log_{10}$ P-values plotted against chromosome and position. Red horizonal line represents genome-wide significance P=5x10^{-8}.

may be surprising that different phenotypes have different risk alleles, and this may be an example of a balancing mutation where the risk of developing one phenotype is balanced by protective effects for another phenoytpe. It may also be that the same tag-SNP is tagging different causal variants and highlight the different effects that variants can have in different tissues.

rs9313548 on chromosome 5q35 near *FGF18* was promising in the NSECG GWAS (P=7.22x10^{-6}), but this SNP had a different risk allele in the ANECS GWAS and was not genotyped in iCOGS. rs9313548 had a P-value of 0.027 in the NSECG, ANECS and SEARCH meta-analysis. This highlights the value of large sample sets for detecting and replicating associations.

My work brings the total number of genome-wide significant EC risk loci to seven (figure 4.12) and provide an enhanced insight into the genetic and biological basis of EC. The fact that the 8q24 (rs17232730) and 17q12 *HNF1B* risk loci are also observed in other cancer GWAS raises the possibility that multi-cancer GWAS may reveal common pathways that contribute to tumourigenesis in different tissues.

Chapter 5

Functional analysis of 13q22 locus

5.1 Introduction

5.1.1 Background

The meta-analysis in Chapter 4 has established seven loci associated with EC at a genome-wide significant level. Of the novel association loci, the SNP rs11841589 (OR=1.15, 95% CI 1.11–1.21, P=4.83x10^{-11}) at 13q22 also has a high level of significance in endometrioid histology only analysis (P = 8.70 $\times$ 10^{-11}) and lies 163kb downstream of *KLF5*, a promising candidate gene. *KLF5* is a transcription factor that is associated with cell cycle regulation and is targeted by the tumor suppressor *FBW7* and is rapidly degraded via the ubiquitin proteasome pathway. rs11841589 lies in a 608kb intergenic region between *KLF5* and *KLF12* and there are 3 other genes *PIBF1*, *DIS3* and *BORA* that are located within 500kb of this SNP. Therefore, I postulated that rs11841589 may be tagging a causal variant that mediates its effect on EC predisposition by acting as a regulatory element to one of the nearby genes and proceded to conduct further functional analysis for this region.

5.1.2 Aims and overview

The aim of this chapter is to further characterize the EC association at 13q22 using various methods including *in vitro* functional studies. Fine-mapping and functional annotation is employed to refine the association signal and to look for possible causal variants. eQTL searches are conducted to look for SNPs that are associated with the expression of nearby genes and this suggested that rs11841589 is associated with *KLF5* expression. The expression of *KLF5* is quantified in 11 EC cell lines and the epigenetic landscape of a 16kb region around rs11841589 is explored using formaldehyde-assisted identification of regulatory elements (FAIRE) and chromatin immunoprecipitation (ChIP) using antibodies against H3K4Me2 and panH4Ac antibodies. Luciferase reporter assays are performed to assess the regulatory potential of the genomic sequence around rs11841589 and rs9600103. Results suggest that the rs11841589-G EC risk allele is associated with an increased expression of *KLF5* in normal endometrial tissue and EC tumour cells and rs9600103 (a SNP in perfect LD with rs11841589) is a DNase1 hypersensitivity site (DHS), conserverved locus in vertebrates and is enriched in FAIRE and ChIP analysis. Luciferase assays show that the genomic sequence around rs9600103 could act as a suppressor in an allele-specific manner with the rs9600103-T EC protective allele exerting a stronger suppressor effect on gene expression.

5.2 Fine-mapping of 13q22 locus

The lead SNP, rs11841589 (P=4.83 $\times$ 10^{-11}) is located at 13q22 and regional impuation of 1.5mb around this SNP using 196 high coverage whole genome sequences and the 1000 Genomes reference panel revealed 5 SNPs, all in LD with rs11841589 (r^2 >0.98) that surpassed genome wide significance (table 5.3) . These SNPs are located in a 3kb region bound by rs9600103 (P=8.70 $\times$ 10^{-11}) and rs11841589 (chr13:

73,811,879-73,814,891, figure 5.1). There is no residual association signal after conditioning for the top SNP (P>0.04 for all SNPs in the 1.5mb region).

5.3 Expression quantitaitve trait locus analysis

5.3.1 Background

A quantitative trait locus is a sequence of DNA that contains or is linked to genes, and that contributes to an observable phenotype or quantitative trait. Expression quantitative trait loci (eQTL) refers to genomic loci, in this case SNPs, that regulate the expression levels of mRNA. Polymorphic genetic variation can influence gene expression using a variety of mechanisms including transcription regulation factors, alterations in splicing and RNA stability (Borel *et al.*, 2011; Pickrell *et al.*, 2010). Genetic variation that is located near to and far away from the target effector gene are known as cis and trans-eQTLs respectively. The advent of high-throughput sequencing including microarrays and deep sequencing have allowed investigators to measure both genetic variation and gene expression at a genomic level, allowing the assessment of association of all genomic loci to all genes. Many GWAS risk loci are located in intergenic regions and eQTL analysis has the potential to shed further light on how they influence gene expression and phenotype. Indeed, SNPs that are identified in association studies are more likely to be eQTLs than one would expect by chance (Nicolae *et al.*, 2010). I have conducted searches for the SNPs of genome-wide significance at the 13q22 locus in public cis-eQTL databases.

5.3.2 Public eQTL databases

There are various public eQTL databases that aim to provide information on eQTLs across the genome in different tissue types. Many of these databases are based on peripheral blood samples because these samples are relatively easy to obtain and

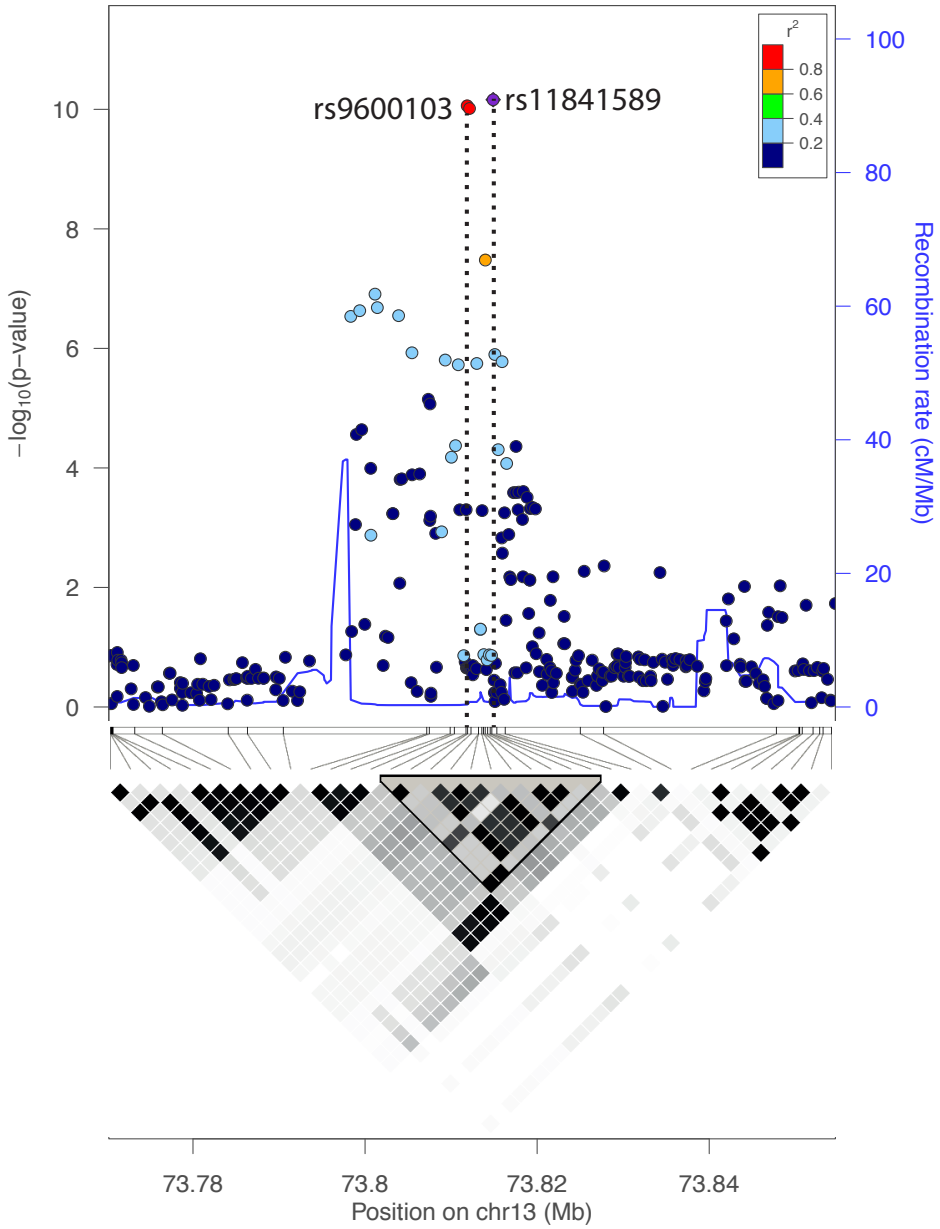

Figure 5.1: Regional association plot of chromosome 13q22 around rs11841589, from position 73,770,218 to 73,854,666. The top SNP rs11841589 is marked as a purple diamond, and the dot colour represents the LD withrs11841589. Y-axis is -$\log_{10}$ P-values from the meta-analysis and regional imputation for NSECG, ANECS, SEARCH and eight iCOGS groups. The blue line shows recombination rates in cM/Mb. Underneath is a LD plot of same region in greyscale with black squares representing SNPs in perfect LD (r^2=1) and white squares representing linkage equilibrium. The 3kb LD block bordered by rs9600103 and rs11841589 is outlined in black.

therefore large datasets are possible. I conducted searches for SNPs in the 13q22 region that surpass genome-wide significance, and identified no cis or trans eQTLs in multiple datasets (Fairfax *et al.*, 2012; Montgomery *et al.*, 2010; Stranger *et al.*, 2007; Westra *et al.*, 2013). The eQTLs with the strongest effects occur in multiple cell types and are usually located within 100kb of the transcriptional start site (Dixon *et al.*, 2007; Stranger *et al.*, 2007), although there are many regulatory elements that are tissue specific, and these eQTLs are more broadly distributed with weaker effects (Dimas *et al.*, 2009; ENCODE Project Consortium *et al.*, 2007; Heintzman *et al.*, 2009). This suggests that these 13q22 SNPs may not be a strong eQTLs across many different tissue types, but it could yet have an effect in endometrial tissue.

5.3.3 eQTL analyis in GTEx normal uterine tissue

The Genotype-Tissue Expresion Project (GTEx, (GTEx Consortium, 2013)) aims to catalog variation and associations in genotype and tissue-specific expression in humans. Tissues are obtained in low post-mortem intervals or in organ transplantation settings with strict inclusion and exclusion criteria such that the tissues used for analysis are largely disease-free. Expression and genotype data are generated using an Affymetrix Expression array and Illumina Human Omni 5M-Quad SNP array and there was matching genotype and uterine expression data for 32 samples. Quality control measures are detailed in the reference and the expression levels are represented as a rank normalized score. rs11841589 is typed in the Affymetrics arrays and a look-up revealed that the rs11841589-G risk allele is associated with an increased expression of *KLF5* with a P-value of 0.02 (figure 5.2).

Genes within 1mb of rs11841589 including *KLF12*, *PIBF1*, *DIS3* and *BORA* were also queried in the same manner and the mRNA expression levels of these genes were not correlated with rs11841589 genotype (P>0.6, table 5.1).

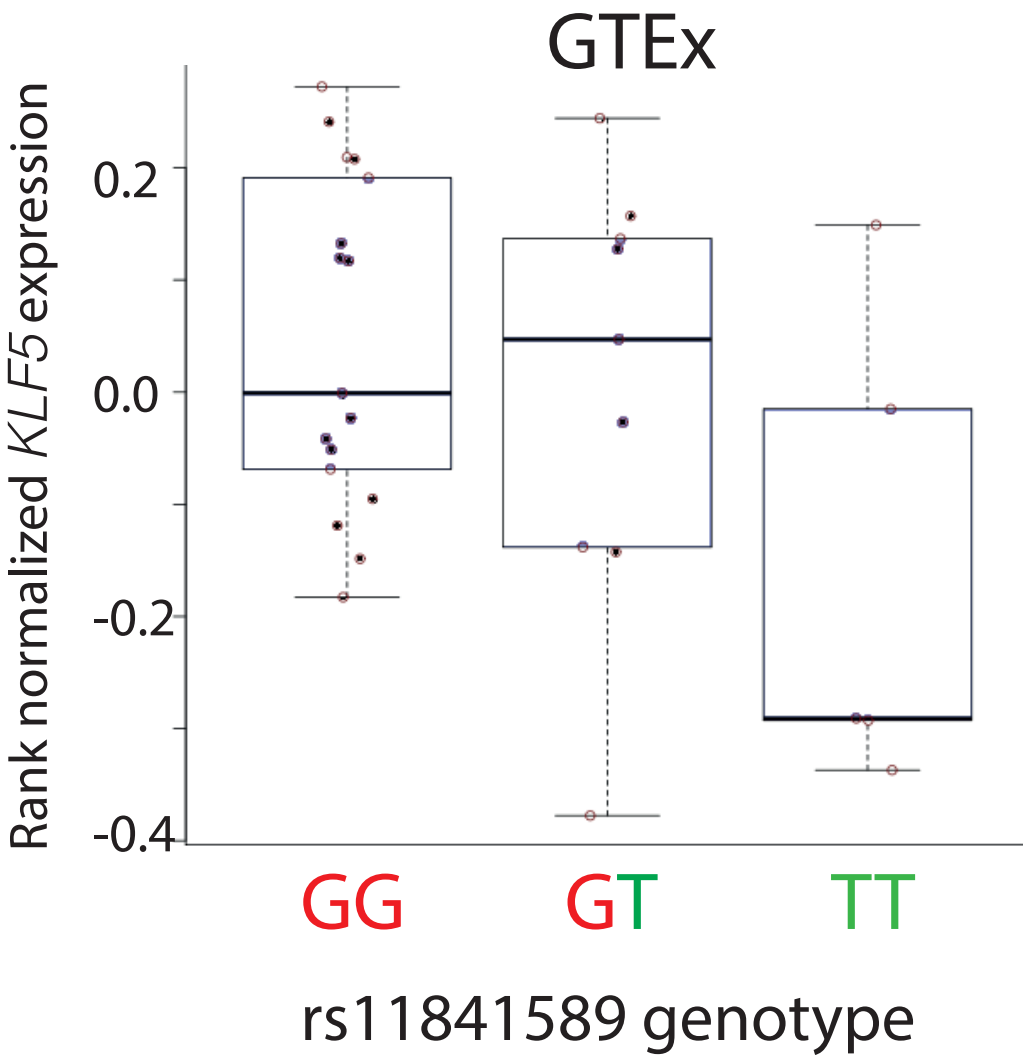

Figure 5.2: rs11841589 genotype and rank normalized *KLF5* expression in 32 normal uterine samples. The rs11841589-G EC risk allele (in red) is associated with increased expression (P=0.02).

Table 5.1: eQTL analysis for 13q22 SNP rs11841589 in normal endometrial (GTEx) and EC (TCGA) samples. The P-values represent the significance of association between rs11841589 genotype and cis-gene mRNA levels. "NA" denotes that expression data was missing for this dataset.

Locus	SNP	Position	Cis-gene	GTEx P	TCGA P
13q22.1	rs11841589	73,814,891	*KLF5*	0.02	0.0049
13q22.1	rs11841589	73,814,891	*KLF12*	1.00	0.74
13q22.1	rs11841589	73,814,891	*PIBF1*	0.90	NA
13q22.1	rs11841589	73,814,891	*DIS3*	0.60	0.56
13q22.1	rs11841589	73,814,891	*BORA*	0.80	NA

5.3.4 eQTL analysis in TCGA endometrial tumour samples

The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) aims to catalogue the genomic changes in different cancers including EC. There is genotype, CNV and expression data for 54 EC samples, generated using the Affymetrix 6.0 SNP array and AgilentG4502A arrays. Since there are marked copy number differences in tumour samples, somatic CNVs were taken into account by regressing gene expression to average copy number spanning the gene. Residual unexplained variance in gene expression was then regressed on the genotype of a particular SNP. A ruby script developed by Luke Freeman-Mills was employed to test SNPs within 1mb of each gene of interest based on the methods described by (Li *et al.*, 2013). A boxplot of residual *KLF5* expression against rs11841589 genotype shows that the rs11841589-G risk allele is significantly associated with increased expression (P=0.0049, figure 5.3). This concurs with results observed in normal endometrial tissue.

Expression of other genes within 1mb of rs11841589 showed no evidence of correlation (P>0.55, table 5.1).

Therefore, rs11841589 is an eQTL in endometrial tissue where the rs11841589-G EC risk allele is associated with higher expression of *KLF5* in both normal endometrial and EC tumour samples. Higher *KLF5* levels have been observed in EC compared to normal endometrial tissue and this analysis suggests that there is an

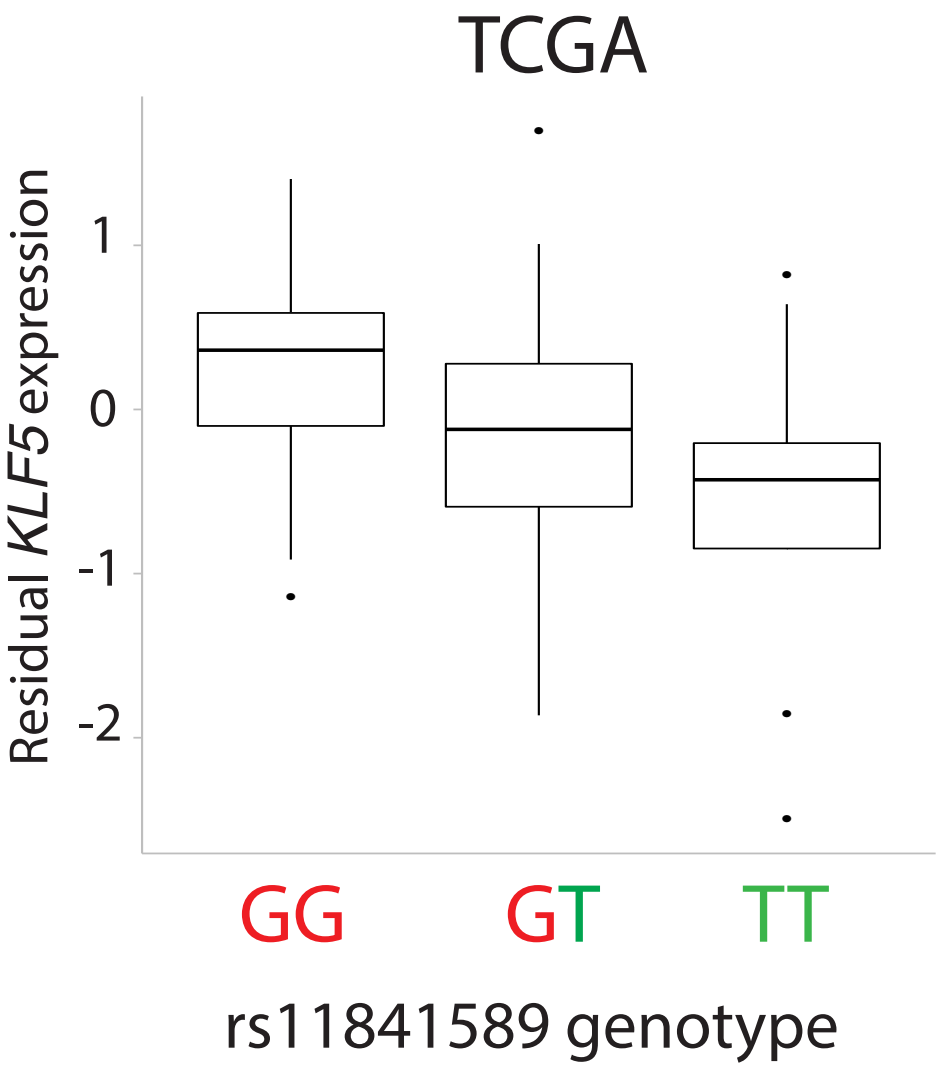

Figure 5.3: rs11841589 genotype and residual *KLF5* expression in 54 tumour samples. The rs11841589-G EC risk allele (in red) is associated with increased expression (P=0.0049.

allele-specific effect exerted by rs11841589. Much like for GWAS, the resolution of the eQTL signal is limited by the size of the LD block and thus this effect may be caused by any of the SNPs in the 13q22 locus correlated with rs11841589 in normal and tumour endometrial tissue. Evidence here suggests that this allele-specific effect may play a role in both predisposition to, and maintenance of the EC disease process. The GTEx analysis could only be conducted as a single-SNP lookup, while a the TCGA analysis revealed that rs11841589 was a local maximum in terms of eQTL significance.

5.4 Functional annotation of 13q22 locus

The UCSC genome browser (http://genome.ucsc.edu) is a public database that hosts various annotations of the genome and I conducted a search for possible functional elements in the rs9600103-rs11841589 locus (chr13:73,814,891-73,811,879). As with the eQTL analyis described above, many of the functional annotation tracks included data from experiments performed on different cell types, and there was no strong evidence for regulatory potential observed. Two EC cell lines (Ishikawa and ECC-1) were part of the tier 3 release of the ENCODE project (ENCODE Project Consortium *et al.*, 2007, www.encodeproject.org) and there is evidence of a DHS present at rs9600103 in these cell lines. Sensitivity to DNase1 suggest that DNA is not nucleosome-bound and that it is accessible to functional elements. The DHS peak is present in the F-Seq DHS density signal (Boyle *et al.*, 2008) in Ishikawa and ECC-1 cell lines, which are treated with oestrogen and tamoxifen, suggesting that the open chromatin state is not oestrogen-dependent (figure 5.8). An alternative F-Seq peak method (Boyle *et al.*, 2008) for designating contiguous regions surpassing the significance threshold revealed that the signal around rs9600103 (position 73,811,879) is statistically significant in all four Ishikawa and ECC-1 experiments. There is a second peak 7.5kb

telomeric to rs9600103 at position 73,819,586 that is only statistically significant in ECC-1 cells (table 5.2).

Table 5.2: F-Seq peak DHS positions showing contiguous regions that surpass the statistical significance threshold (P<0.05). Experiments conducted in Ishikawa and ECC-1 cells treated with oestrogen and tamoxifen, from ENCODE. rs9600103 is located at position 73,811,879. Pos, position on chromosome 13 in build 19; Start pos, start; End pos, end position of statistically significant peak; Length, length of peak; Peak signal, position of highest significance; P, P-value.

	Start pos	End pos	Length	Peak signal	Score	P
ECC1-DMSO	73,811,867	73,812,131	264	73,812,017	485	0.014
ECC1-oestrogen	73,811,988	73,812,098	110	73,812,042	481	0.038
Ishi-oestrogen	73,811,824	73,812,110	286	73,811,999	520	0.019
Ishi-oestrogen	73,819,472	73,819,686	214	73,819,575	525	0.017
Ishi-tamoxifen	73,811,810	73,812,124	314	73,812,006	536	0.013
Ishi-tamoxifen	73,819,487	73,819,716	229	73,819,597	536	0.013

A corresponding conservation peak is observed around rs9600103 in the 100 vertebrates PhastCons track from UCSC genome browser (Felsenstein and Churchill, 1996; Pollard *et al.*, 2010, figure 5.8), which means that this region is conserved amongst vertebrates, and lends further weight to the possibility that it is functionally significant.

Public databases such as Haploreg (Ward and Kellis, 2012), RegulomeDB (Boyle *et al.*, 2012) and ENCODE (www.encodeproject.org) are available for annotating variants. These were employed to look at SNPs in LD with rs11841589 (r^2>0.4, table 5.3). These SNPs lie 10-13kb away from pseudogene Ro-associated Y1 pseudogene 8 (RNY1P8) and show some evidence of altered transcription factor motifs but otherwise, there is no strong support for a functional role.

5.5 Characterization of EC cell lines

Cell culture was conducted as described in section 2.2.6 and DNA was extracted for genotyping using competitive allele-specific PCR (KASPar, section 2.2.4). RNA was

Table 5.3: Functional annotation of SNPs in LD (r^2>0.4) with rs11841589 using Haploreg and RegulomeDB. Pos, positions in build 19; LD, r^2 with rs11841589 (in bold); AF, allele frequency; Enhancer, enhancer histone mark; RegDB, RegulomeDB score. RegDB score 5 represents TF binding or DNase peak. Blank spaces represent no data.

SNP	Chr 13 Pos	LD	AF	Enhancer	GENCODE	Altered Motifs	RegDB
rs9600103	73,811,879	0.98	0.28		10kb 3' of *RNY1P8*	AP-1,Elf3,GR,PRDM1,SIX5	5
rs7981863	73,812,141	0.98	0.28		11kb 3' of *RNY1P8*	Rad21,SMC3	
rs7988505	73,813,435	0.98	0.27		12kb 3' of *RNY1P8*	CDP,Foxa,Nanog,Sox	
rs7989799	73,813,436	0.99	0.27		12kb 3' of *RNY1P8*		
rs9592895	73,813,982	0.75	0.33	NHEK	13kb 3' of *RNY1P8*	DMRT7,SIX5	
rs11841589	73,814,891	-	0.28		13kb 3' of *RNY1P8*	Pou2f2	

extracted for conversion to cDNA and the expression levels of *KLF5* in cell lines were quantificatified using RT-qPCR as described in section 2.2.9. Figure 5.4 shows the expression of *KLF5* in 11 EC cell lines with *GAPDH* used as an internal control and all expression levels reported in relation to the Ishikawa cell line using the ddCT method.

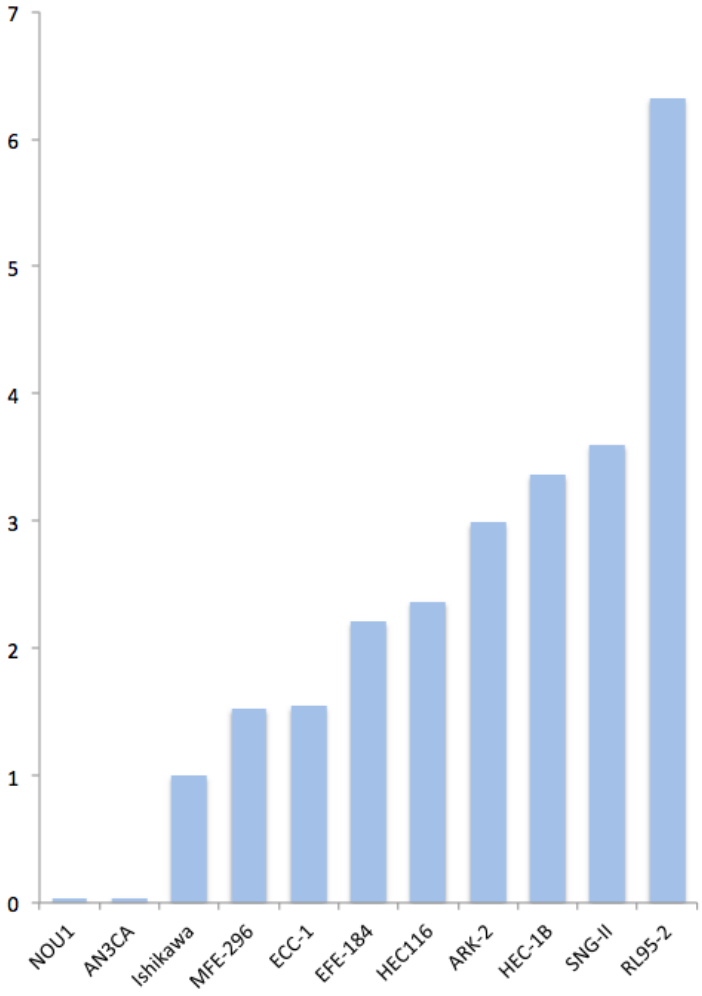

Figure 5.4: *KLF5* expression in 10 EC cell lines relative to Ishikawa cells. *GAPDH* was used as an internal control.

Nine of the eleven EC cell lines express significant levels of *KLF5* including Ishikawa, and ECC-1, the two cell lines used by ENCODE. The genotypes of the of the cell lines for rs11841589 and rs9600103 showed that these SNPs were in perfect

Table 5.4: EC cell lines: *KLF5* expression, rs11841589 and rs9600103 genotypes. *KLF5* expression levels are normalized using *GAPDH* as an internal control and expressed in relation to Ishikawa cells using the ddCT method.

	KLF5 expression	rs9600103	rs11841589
NOU1	0.03	TT	TT
AN3CA	0.04	AA	GG
Ishikawa	1.00	AA	GG
MFE-296	1.53	AA	GG
ECC-1	1.55	AA	GG
EFE-184	2.21	TA	GT
HEC-116	2.37	TA	GT
ARK2	2.99	TT	TT
HEC1B	3.36	TA	GT
SNGII	3.60	AA	GG
RL952	6.32	AA	GG

LD with comparable MAFs as population frequencies (table 5.4).

5.6 FAIRE

I performed formaldehyde-assisted identification of regulatory elements (FAIRE) for a 16kb region around rs9600103 and rs11841589 (chr13: 73,804,930-73,820,618) as described in section 2.2.11. FAIRE has been shown to identify regulatory regions by enriching for nucleosome-depleted chromatin regions that are more efficiently cross-linked and captured in the aqueous phase of the experimental procedure (Giresi *et al.*, 2007). SYBR green RT-PCR (section 2.2.9) with primers specific to the 16kb region in 13q22 was used to assess the relative enrichment after normalizing for *RHO* promoter levels and input uncrosslinked DNA using the ddCt method (Livak and Schmittgen, 2001). There is significant overlap between regions of regulatory potential identified by FAIRE and DHS-seq, though there are also regions that are only identified using either method (Song *et al.*, 2011). Three *KLF5*-expressing cell lines (Ishikawa, EFE-184 and ARK-2) and also AN3CA (which displayed minimal *KLF5* expression) were used in FAIRE experiments. Ishikawa cells were chosen because of the DHS seen in

ENCODE experiments and Ishikawa, EFE-184 and ARK-2 cells are homozygous risk, heterozygous and homozygous protective for rs9600103 and rs11841589, respectively (table 5.4). AN3CA cells were chosen because they have minimal *KLF5* and have the commoner (risk) rs11841589 and rs9600103 alleles.

Figure 5.5 shows that there are two enrichment peaks noted in the FAIRE analysis in the three *KLF5* expressing cell lines Ishikawa. The largest peak is observed in Ishikawa and ARK-2 cells at position 73,811,879, corresponding to the location of rs9600103. A second, smaller peak is located at rs6562748 (position 73,809,322) in EFE-184 and ARK-2. There are no regions of enrichment observed in the non-*KLF5* expressing AN3CA cells.

5.7 ChIP using anti-H3K4Me2 and anti-H4Ac

I performed Chromatin Immuno-Precipitation (ChIP) as described in section 2.2.11 for the same 16kb region around rs11841589 and rs9600103 using anti-dimethyl-histone H3 lysine 4 (H3K4Me2) and anti-acetyl-histone H4 (H4Ac) antibodies. H3K4Me2 and H4Ac are markers of transcription factor binding (Wang *et al.*, 2014b) and active chromatin respectively. ChIP using these antibodies identifies regions enriched for these histone modifications. The same four cell lines used for FAIRE were tested, and a similar ddCT method was used to calculate enrichment based on input chromatin and the *RHO* promoter. The no-antibody control for both the anti-H3K4Me2 and anti-H4Ac experiments showed minimal enrichment, which confirms the specificity and efficiency of the antibodies used.

ChIP results displayed in figure 5.5 show that there is a distinct H3K4Me2 peak centred around rs9600103 (position 73,811,879) in Ishikawa and ARK-2 cells. A similar but smaller peak is observed in ARK-2 cells in H4Ac ChIP, while a peak over rs7996454 (position 73,810,516) is seen in Ishikawa and EFE-184 cells. The results for

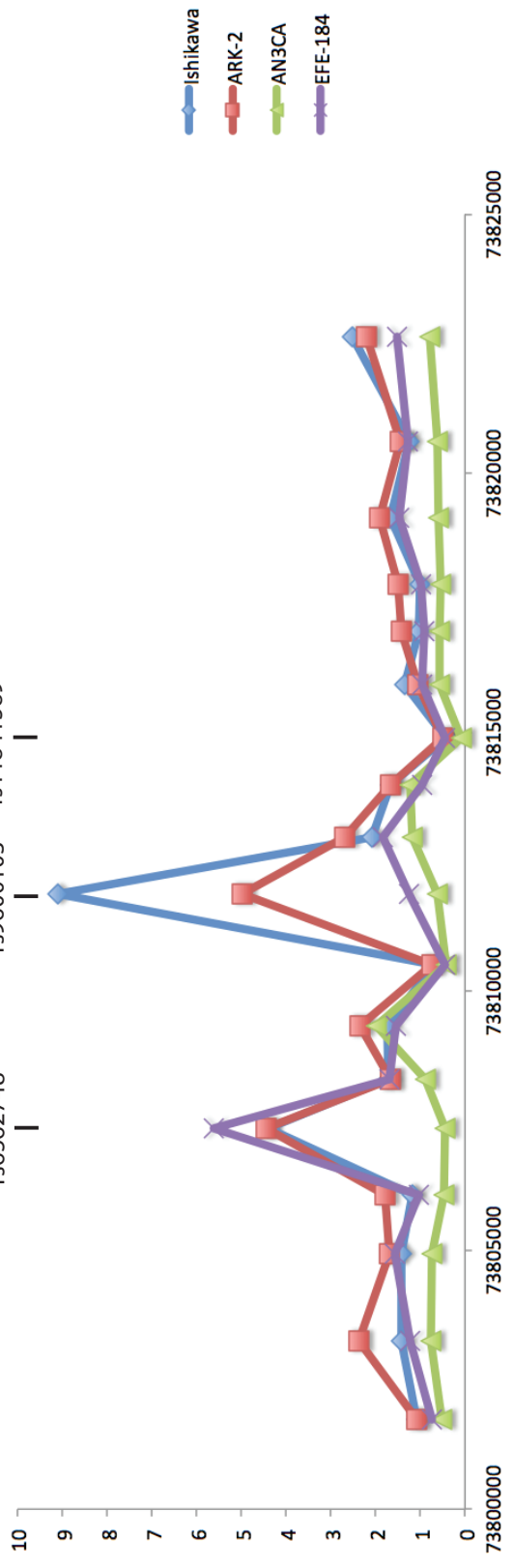

Figure 5.5: FAIRE for four EC cell lines Ishikawa, ARK-2, AN3CA and EFE-184 in 13q22 locus (position 73,804,930-73,820,618). Position on chromosome 13 is plotted against enrichment normalized to non-crosslinked genomic DNA, relative to the *RHO* promoter as a negative control. Black marks display the position of rs6562748, rs9600103 and rs11841589.

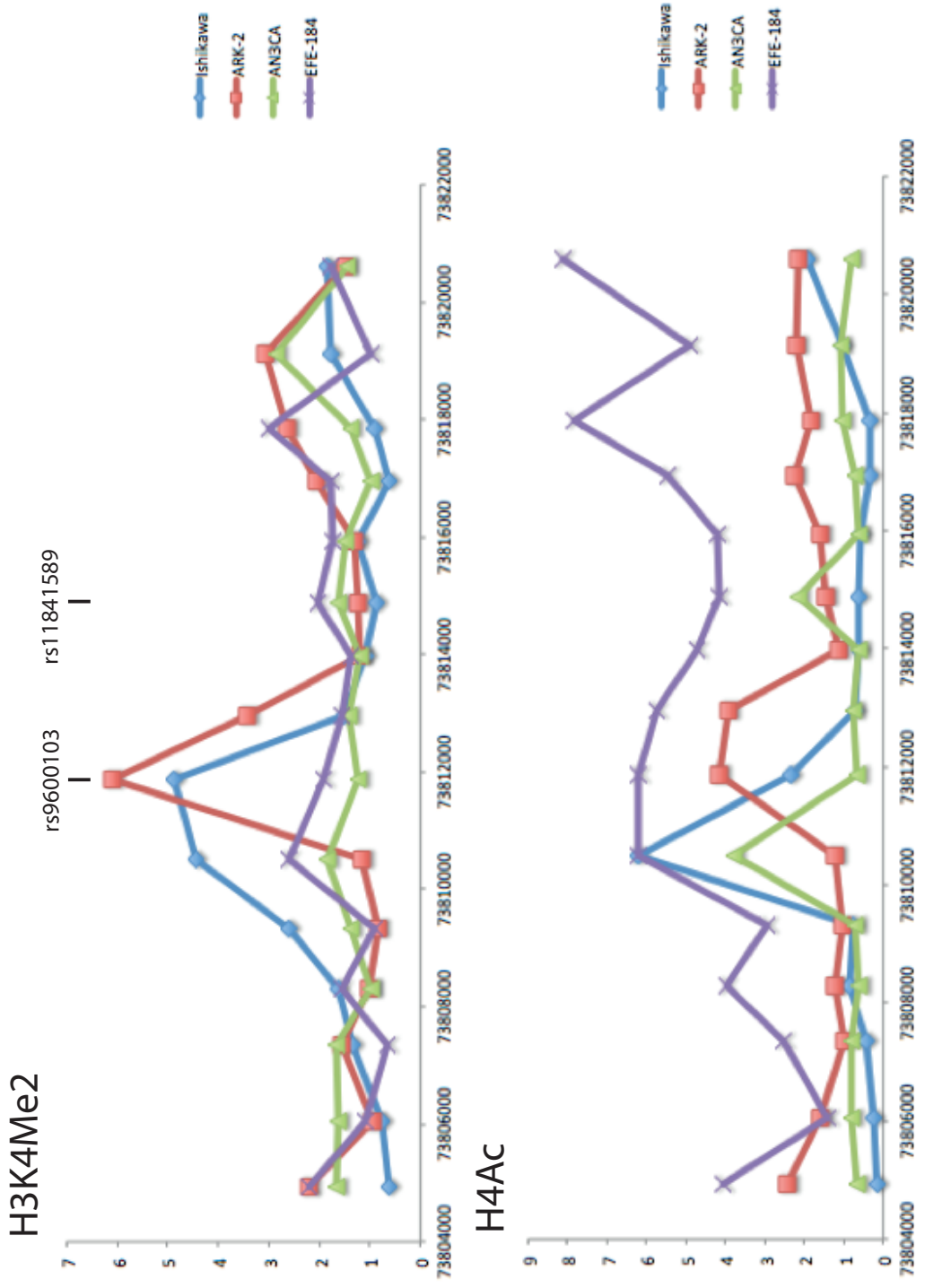

Figure 5.6: ChIP with anti-H3K4Me2 and H4Ac antibodies for four EC cell lines Ishikawa, ARK-2, AN3CA and EFE-184 in 13q22 locus (position 73,804,930-73,820,618). Position on chromosome 13 is plotted against enrichment normalized to sonicated input DNA, relative to the *RHO* promoter as a negative control. Black marks display the position of rs9600103 and rs11841589.

EFE-184 cells are more difficult to interpret as there is evidence of enrichment across an 8.7kb region from 73,811,879 to 73,820,618 that is not seen in the anti-H3K4Me2 ChIP or FAIRE. Ishikawa and to a lesser extent, AN3CA cells show an enrichment peak at position 73,810,516.

5.8 Luciferase reporter assay

The regulatory nature of the region around rs9600103 and rs11841589 was then investigated using allele-specific luciferase enhancer reporter assays as described in section 2.2.13. I planned and supervised the experimental work, which was conducted by Dr Susanne Flach.

For this experiment, we used ECC-1 cells, because these consistently gave very good transfection efficiency, expressed *KLF5* and contained a DHS. Two 3kb fragments: i) chr:13 73,810,509-73,813,452 centred around rs9600103 and ii) chr13:73,813,268-73,816,290 around rs11841589 were cloned into the pGL3-Promoter vector to test for regulatory effects. There is a 182bp overlap when comparing these two DNA segments and each fragment only contains either rs9600103 or rs11841589. Site directed mutagenesis was employed to introduce the alternative allele in the two SNP positions and fragment sequences are otherwise identical. Two controls were used: the empty pGL3 vector and a neutral 2.2kb sequence with no regulatory activity (Lewis *et al.*, 2014). There was no significant difference between the luciferase activity of the two controls (P=0.085) and the no-insert pGL3 was used as the control comparison with other fragments.

Resutls can be interpreted in two groups as presented in figure 5.7. Group 1, with fragments containing the rs9600103-T, rs11841589-T and rs11841589-G alleles all had similar activity. This activity was significantly lower than that of the pGL3 no-insert control and the fragment containing the rs9600103-A high-risk allele (figure 5.7).

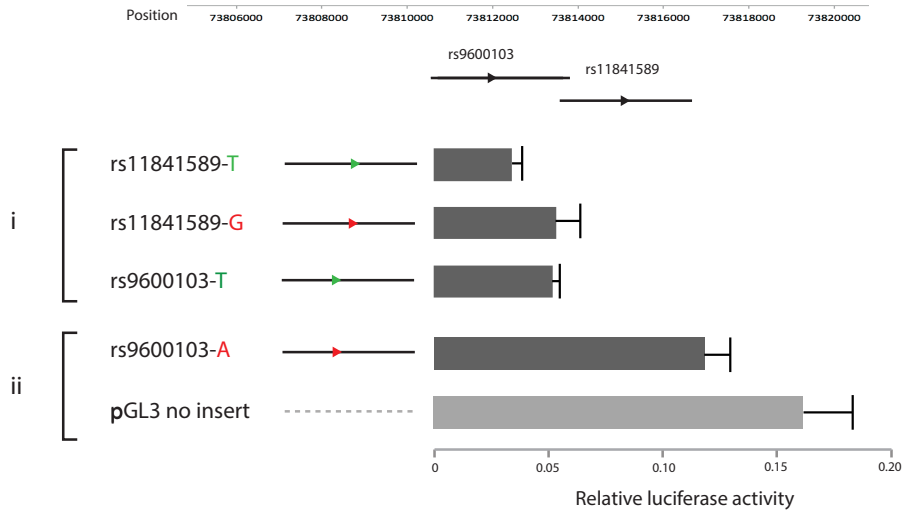

Figure 5.7: Luciferase reporter assay to analyse the activity of 3kb fragments containing either rs9600103 or rs11841589 as enhancers or repressors using the pGL3 promoter vector in Ishikawa/ECC-1 cells. Schematic diagram displays position of fragments on chromosome 13. Green arrows represent the low-risk and red arrows the high-risk alleles. We focused on the rs9600103 site owing to the presence of a DHS in the ENCODE data. The bar chart shows the average luciferase activity of experiments with 3 or 4 replicates, each performed on 3 occasions (total of 11 assays). Error bars represent the standard error of the mean. Data were normalized by subtraction of background luminescence and normalised to pGL4 Renilla activity. Group 1 consists of the rs11841589-T, rs11841589-G and rs9600103-T fragments. Group 2 consists of rs9600103-A fragment and no-insert pGL3 control. All intra-group pairwise t-test comparisons between group 1 and 2 are not significant, while all inter-group pairwise comparisons are significant ($P<0.05$, table 5.5).

Table 5.5: Pairwise t-test P-values for 13q22 luciferase assays. Group 1 consists of the rs11841589-T, rs11841589-G and rs9600103-T fragments. Group 2 consists of rs9600103-A fragment and no-insert pGL3 control. All paired t-tests between group 1 and 2 are significant ($P<0.05$).

	rs11841589-T	rs11841589-G	rs9600103-T	rs9600103-A	pGL3
rs11841589-T	-	0.058	0.128	0.045	0.008
rs11841589-G	-	-	0.979	0.039	0.013
rs9600103-T	-	-	-	0.018	0.014
rs9600103-A	-	-	-	-	0.226
pGL3	-	-	-	-	-

Together, these results suggest that the causal EC variation tagged by rs11841589 results, at least in part, from a regulatory site containing rs9600103. This result concurs with what we would expect, given that the rs9600103-A / rs11841589-G EC risk allele is associated with a higher expression of *KLF5* in eQTL analysis.

The strongest KLF5 expression is associated with the rs9600103 high-risk allele, and the data suggest that the region might act as a transcriptional repressor.

5.9 Discussion

The EC meta-analysis suggests that there is a single signal at chromosome 13q22, localized to a 3kb LD block bounded by rs9600103 and rs11841589 containing five correlated SNPs surpassing genome-wide significance. eQTL analysis revealed that the rs11841589-G risk allele is associated with a higher expression of *KLF5* in normal endometrial tissue and EC tumour tissue. rs11841589 genotype was not associated with expression of other genes within 500kb (*KLF12*, *PIBF1*, *DIS3* and *BORA*, P>0.56). This suggests that the GWAS signal may be exerting its effect by regulating the expression of *KLF5* in an allele-specific manner in normal uterine and endometrial tumour tissue. The uterine tissue eQTL for rs11841589 is based on a small sample size and it would be useful to know if this association is retained in larger endometrial eQTL studies. TCGA has recently released a RNA-seq expression data for about 500 endometrial tumour and adjacent tissue. Early analysis of this data does not display any correlation between rs11841589 genotype and *KLF5* expression.

Functional annotation of SNPs correlated with the top SNP did not reveal any strong regulatory elements across different cell types, although there is evidence of vertebrate conservation at the rs9600103 locus. In EC cell lines used in ENCODE, there is a statistically significant DHS present over the same position around rs9600103. This DHS is present when the EC cell lines were treated with oestrogen or tamoxifen,

suggesting that the open chromatin is not oestrogen-dependent.

There have been reports that the ECC-1 cell lines are very similar to the Ishikawa cells using short tandem repeats profiling (Korch *et al.*, 2012), and the International Cell Line Authentication Committee (ICLAC) have recomended that ECC-1 cells should be treated as Ishikawa cells. In line with this, the ENCODE project plan to rename the ECC-1 tracks as Ishikawa, and the DHS-seq results can be interpreted as biological repeats confirming the existence of a DHS at rs9600103. We have sequenced DNA from Ishikawa and ECC-1 cells using the Ion AmpliSeq Comprehensive Cancer Panel and 2,689 out of 3,004 SNVs observed across 409 cancer-associated genes share the same allele, which is highly suggestive that these two are in fact the same cell line.

Further FAIRE and ChIP experiments were conducted to assess the 13q22 epigenetic landscape. Figure 5.8 displays the 16kb region with the annotations and analysis overlayed.

Signals from FAIRE and anti-H3K4Me2 ChIP for *KLF5*-expressing cell lines Ishikawa and ARK-2 co-located with the conservation peak and DHS from ENCODE, providing strong evidence for open chromatin and transcription factor binding at rs9600103. Anti-H4Ac ChIP did not display a consistent signal across this region, with ARK-2 cells showing enrichment at rs9600103, but Ishikawa and, to a lesser extent, AN3CA cells displayed an H4Ac peak 2.6kb centromeric, at position 73,809,322. Results for EFE-184 cells are difficult to interpret as the enrichment signals vary greatly between the FAIRE, anti-H3K4Me2 and anti-H4Ac ChIP experiments. EFE-184 is a relatively slow-growing cell line and I struggled to have a similar confluency of cells in the 15cm petri dishes used for the experiments.

DNase-seq and FAIRE-seq experiments perfomed on the same cell lines showed that there is a a 30-40% overlap in regions identified using the two methods using ranked comparisons (Song *et al.*, 2011). Regions detected by both methods, or cross-

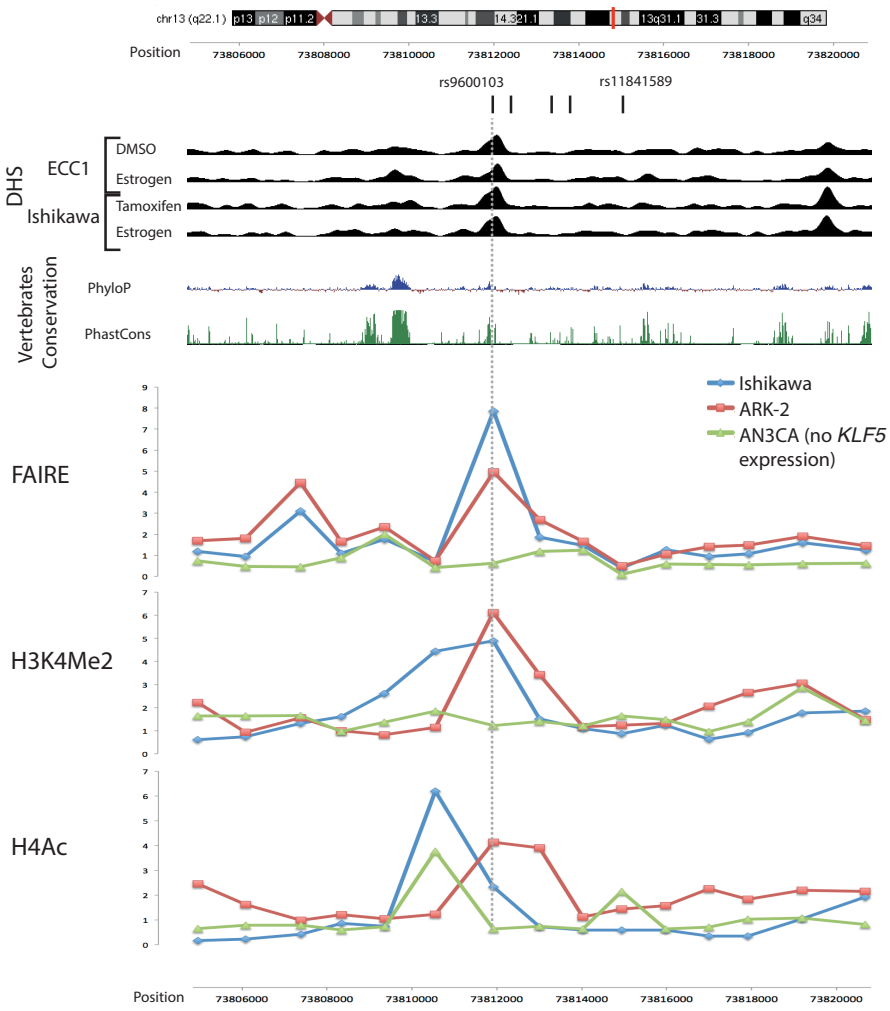

Figure 5.8: Combining results from FAIRE, ChIP with anti-H3K4Me2 and H4Ac antibodies, ENCODE and conservation data for 13q22 locus (position 73,804,930 - 73,820,618). EC cell lines ARK-2 (rs9600103-TT), Ishikawa (rs9600103-AA) and AN3CA (rs9600103-AA) are shown, with the y-axis displaying enrichment normalized to non-crosslinked genomic DNA/sonicated input DNA, relative to the *RHO* promoter as a negative control using the ddCt method. DNase1 hypersensitivity site (DHS) density signal in ENCODE EC cell lines ECC-1 and Ishikawa from experiments with cell lines treated with oestrogen and tamoxifen. 100 vertebrates conservation from UCSC Genome Browser.

validated sites are abundant in intergenic regions across the genome and show significant overlap to ChIP-seq data generated using antibodies to CTCF, MYC and Pol II. The presence of a strong signal in endometrial cells which is absent in larger public databses of different tissues suggests that the open-chromatin state at rs9600103, as well as its eQTL effects on *KLF5*, may be specific to endometrial tissue. The results here also highlight the value of employing multiple techniques to assess the epigenetic landscape of a region and co-occurring signals provide strong evidence for regulatory potential.

Allele-specific luciferase enhancer reporter assays suggest that there is a repressor effect exerted by the protective alleles in the rs9600103 and rs11841589 fragments. This effect is more marked in rs9600103 and together, these results suggest that the causal EC variation tagged by rs11841589 results, at least in part, from a regulatory site containing rs9600103. This result also concurs with the observation that the rs9600103-T EC protective allele is associated with a lower expression of *KLF5*. The data suggest that the region might act as a transcriptional repressor.

Luciferase reporter assay represent an artificial model system by which to assess the regulatory potential of a genetic sequence. The pGL3-promoter vector uses the strong viral SV40 promoter and it would be interesting to repeat the experiments here using the *KLF5* promoter. rs9600103 lies 160kb downstream of *KLF5* the vector could be made to more closely resemble the *in vivo* system by cloning the rs9600103 fragments downstream to the luc^+ gene. In addition, fragments of varying size around rs9600103 and the differences in luciferase activity would further shed light on where the causal variant may be located.

KLF5 is expressed in low levels in the glandular and stromal cells of the human endometrium (Human Protein Atlas, www.proteinatlas.org) and has previously been shown to be expressed at a higher level in EC (Simmen *et al.*, 2010). *KLF5* is a member of the Kruppel-like factor subfamily of zinc finger proteins and is a transcription

factor associated with cell proliferation and survival. Its expression levels are higher in actively dividing or proliferating cells and lower in differentiated or mature cells (Jiang *et al.*, 2008; McConnell *et al.*, 2007). In terms of cell survival, *KLF5* interacts with p53 to regulate the activity of an apoptotic inhibitor in acute lymphoblastic leukaemia cells. *KLF5* has been implicated as an oncogene and tumour suppressor in different cancers (Chen *et al.*, 2003; Diakiw *et al.*, 2012; McConnell *et al.*, 2009; Nandan *et al.*, 2008) while also playing a role in uterine development and homeostasis (Davis *et al.*, 2011; Mutter *et al.*, 2001; Shi *et al.*, 1999; Simmen *et al.*, 2010). *KLF5*-null mice show embryonic lethality (Ema *et al.*, 2008), but uterine deletion of *KLF5* allows normal uterine development, though with impaired fertility due to impairment of the implantation process (Sun *et al.*, 2012). Elevated *KLF5* levels are strongly correlated with activating *KRAS* mutations (Nandan *et al.*, 2008) and *KLF5* is known to be targeted for degradation by the tumour suppressor *FBXW7* (Liu *et al.*, 2010). *KRAS* and *FBXW7* are amongst the most common mutations seen in EC, present in 15% and 9% of cancers respectively (COSMIC, Forbes *et al.*, 2008).

Additional genotyping would aid fine-mapping and further functional work could be pursued to elucidate the relationship between the rs9600103 intergenic locus and *KLF5* expression with the use of techniques chromosome conformation capture. In addition, this opens the door to further investigation into the role of *KLF5* in EC susceptibility and tumourigenesis.

Chapter 6

Shared predisposition variants for EC and CRC

6.1 Introduction

6.1.1 Background

There is evidence of shared genetic aetiology in the predisposition to EC and CRC from high-penetrance Mendelian cancer syndromes. Lynch syndrome and polymerase proofreading-associated polyposis predispose affected individuals to developing both EC and CRC. It is unknown why pathogenic germline mutations in these mismatch repair (MMR) and polymerase genes and would cause tumourigenesis specifically in the human endometrium and colon, but this observation prompts one to ask if there are other genetic variants that contribute to the development of both EC and CRC.

6.1.2 Aims and Overview

The aim of this chapter is to look for genetic variants that are associated with predisposition to both EC and CRC using GWAS data. SNPs that have previously been associated with either EC or CRC are assessed in the other cancer phenotype and

all EC and CRC datasets are meta-analysed in search of novel predisposition loci to both EC and CRC. Genome-wide enrichment for EC and CRC is assessed using moderately significant SNPs that do not surpass genome-wide significance. The CRC predisposition SNP at the *TERT* locus is the most significant of previously-established associations in the other cancer phenotype and I have discovered two novel susceptibility loci at chromosome 12q24 (*SH2B3*) and 18q22 (*TSHZ1*) for EC and CRC.

6.2 MMR and polymerases in EC and CRC

As described in section 1.4, high-penetrance germline mutations in MMR genes *MLH1*, *MSH2*, *MSH6* and *PMS2* (Barrow *et al.*, 2013), and in DNA polymerases *POLD1* and *POLE* (Palles *et al.*, 2013) were first demonstrated to be associated with CRC risk, and female mutation carriers were subsequently shown to have a greatly increased risk of endometrial cancer (lifetime risk 30-70%, Dunlop *et al.*, 1997; Senter *et al.*, 2008), in some cases exceeding that for CRC.

The MMR system maintains genomic stability by correcting mismatched nucleotide pairs that arise during DNA replication. Pathogenic MMR mutations cause a microsatellite instability (MSI) phenotype in ECs and CRCs (Jass *et al.*, 1995). Bi-allelic *MLH1* promoter methylation (Herman *et al.*, 1998; Simpkins *et al.*, 1999) and a few somatic mutations in *MLH1* and *MSH2* (Haraldsdottir *et al.*, 2014) are seen in sporadic CRCs and ECs, causing the same MSI and hypermutator phenotype. Histologically, MMR-deficient CRCs and ECs are characterised by poor differentiation and the presence of mucinous and signet-cell features and tumour-infiltrating lymphocytes (Hampel *et al.*, 2006; Smyrk *et al.*, 2001).

POLE and *POLD1* encode subunits for polymerases that synthesise the leading and lagging strand of the DNA replication fork. The exonuclease (proofreading) do-

mains of these polymerases increase replication fidelity by recognising and excising mispaired bases (Goldsby *et al.*, 2001; Simon *et al.*, 1991). Germline missense mutations in the exonuclease domains of *POLD1* and *POLE* predispose to CRC and EC, and somatic *POLE* mutations occur in sporadic CRCs and ECs (Cancer Genome Atlas Network, 2012; Cerami *et al.*, 2012; Palles *et al.*, 2013). Polymerase exonuclease domain mutations (EDMs) do not cause MSI but lead to an ultramutator phenotype, with over one million base substitutions in some cancers.

Given the commonalities in genetic aetiology, mutator phenotype and histological featues in MMR and polymerase-associated EC and CRC, there may be other genes involved with DNA repair or otherwise, that predispose individuals to both cancer phenotypes. EC candidate gene studies investigating the role of genetic variants in DNA-repair genes have proved inconclusive (section 1.5.2). GWAS provides a powerful method to assess common variants across the genome and the increased sample size gives increased power to detect associations.

6.3 EC and CRC GWAS datasets

Eight GWAS datasets genotyped on various Illumina tag-SNP arrays were available. The three EC GWAS were described in chapter four: i) NSECG, ii) ANECS and iii) SEARCH (total 2,212 cases and 6,725 controls, (Spurdle *et al.*, 2011). The five CRC GWAS comprised: iv) CORGI, v) Scotland 1, vi) VICTOR/QUASAR2/BC58, vii) CFR1 and viii) CFR2 (total 5,725 cases and 6,671 controls, (Houlston *et al.*, 2010; Newcomb *et al.*, 2007; Tomlinson *et al.*, 2008; Yeager *et al.*, 2007). All samples were of European ancestry with the majority of samples from the UK, and others were from USA, western Europe and Australia. Standard quality control measures were performed for each GWAS, as described in the referenced publications, and details about each dataset are described in section 2.1 and summarized in figure 6.1 and 6.2.

Some of the control datasets, including the Wellcome Trust Case Control Consortium 2 (WTCCC2 Wellcome Trust Case Control Consortium, 2007), have previously been used in both CRC and EC GWAS. Controls were assigned proportionately to case data sets and were not used more than once. Moreover, any duplicate or cryptically related samples were excluded by pairwise identity by descent (IBD) analysis. Principal component analysis (PCA) was conducted for all samples together, to ensure that all individuals were of European ancestry and we therefore excluded all individuals who clustered outside the main centroid in pairwise plots of the first 4 PCs. Whole-genome imputation using two reference panels (1000 Genomes 2012 release and 196 high-coverage whole genome-sequenced UK individuals) yielded up to 6 million SNPs either typed or imputed with high quality (info score >0.9) for these GWAS datasets.

As described in section 2.1, 4,330 EC cases and 26,849 female controls were genotyped as part of the Endometrial Cancer Association Consortium (ECAC), with samples from seven countries: UK, USA, Belgium, Germany, Norway, Sweden and Australia. The controls were selected from healthy females participating in the Breast Cancer Association Consortium (BCAC) and Ovarian Cancer Association Consortium (OCAC) part of the iCOGS project and matched and analysed with cases in eight groups by geographical location and PCA. These samples were genotyped using a custom Illumina Infinium iSelect array with 200,000 SNPs designed by the COGS (Collaborative Oncological Gene-environment Study) initiative (Eeles *et al.*, 2013; Michailidou *et al.*, 2013; Pharoah *et al.*, 2013; Sakoda *et al.*, 2013). The SNPs on this array were chosen based on regions of interest from previous breast, prostate, ovarian and endometrial cancer studies, rather than on genome-wide coverage. We did not impute genotypes from the COGS studies, but included directly-genotyped SNPs in the discovery meta-analysis.

Case-control analysis for each GWAS data set was performed using frequentist tests with a logistic regression model after imputation. There was no evidence of sys-

	Study		Case sampling frame	Control sampling frame	Genotyping Platform	EC Cases	Controls
	EC GWAS						
1	NSECG	UK National Study of Endometrial Cancer Genetics	UK; population based cases		Illumina660WQuads	925	
	CGEMS breast	Cancer Genetic Markers of Susceptibility (Breast)		USA; population based cancer free controls from breast study	Illumina HumanHap550		1,141
2	ANECS	Australian National Endometrial Cancer Study	Australia; population based cases		Illumina 610K	606	
	QIMR	Queensland Institute of Medical Research		Australia; parents of participants in adolescent twin study	Illumina 610K		1,846
	HCS	Hunter Community Study		Australia; population-based controls	Illumina 610K		1,237
3	SEARCH	UK Studies of Epidemiology and Risk factors in Cancer Heredity	England; population based cases via cancer registries		Illumina 610K	681	
	NBS			UK; population based controls identified through National Blood Service	Illumina 1.2M		2,501
	EC GWAS total					**2,212**	**6,725**
	EC iCOGS						
4	ANECS	Australian National Endometrial Cancer Study	Australia; population based cases		Illumina Infinium iSelect	373	
	NECS	Newcastle Endometrial Cancer Study	Australia; hospital-based cases		Illumina Infinium iSelect	165	
	ABCFS	Australian Breast Cancer Family Study		Australia; from electoral rolls	Illumina Infinium iSelect		443
	AOCS	Australian Ovarian Cancer Study		Australia; population-based, from electoral rolls	Illumina Infinium iSelect		817
	MCCS	Melbourne Collaborative Cohort Study		Australia; random sample from initial cohort	Illumina Infinium iSelect		437
5	SEARCH	UK Studies of Epidemiology and Risk factors in Cancer Heredity	England; population based cases	England; population based controls	Illumina Infinium iSelect	773	7,510
6	NSECG	National Study of Endometrial Cancer Genetics	England; population based cases		Illumina Infinium iSelect	965	
	BBCS	British Breast Cancer Study		UK; friend, sister-in-law, daughter-in-law or other non-blood relative of breast cancer case	Illumina Infinium iSelect		1,353
	SBCS	Sheffield Breast Cancer Study		UK; women attending Sheffield Mammography Screening, with no breast lesion	Illumina Infinium iSelect		835
	UKBGS	UK Breakthrough Generations Study		UK; women without breast lesions selected from BGS cohort	Illumina Infinium iSelect		449
7	MECS	Mayo Endometrial Cancer Study	USA; Hospital based cases		Illumina Infinium iSelect	221	
	MCBCS	Mayo Clinic Breast Cancer Study		USA; Cancer-free women presenting for general medical examination	Illumina Infinium iSelect		1,762
	MCOCCCS	Mayo Clinic Ovarian Cancer Case-Control Study		USA; Cancer-free women presenting for general medical examination	Illumina Infinium iSelect		593
8	LES	Leuven Endometrial Cancer Study	Belgium; hospital based cases		Illumina Infinium iSelect	321	
	LMBC	Leuven Multidisciplinary Breast Centre		Belgium; controls from blood donors	Illumina Infinium iSelect		1,382
9	BECS/HJECS	Bavarian Endometrial Cancer Study/Hannover-Jena Endometrial Cancer Study	Germany; hospital and population based cases		Illumina Infinium iSelect	137	
	BBCC	Bavarian Breast Cancer Cases and Controls		Germany; healthy women >55yrs from newspaper advertisement	Illumina Infinium iSelect		441
	BSUCH	Breast Cancer Study of the University Clinic Heidelberg		Germany; female blood donors	Illumina Infinium iSelect		920
	ESTHER	ESTHER Breast Cancer Study		Germany; random sample from routine health check-up	Illumina Infinium iSelect		486
	GC-HBOC	German Consortium for Hereditary Breast & Ovarian Cancer		Germany; KORA study	Illumina Infinium iSelect		138
	GENICA	Gene Environment Interaction and Breast Cancer in Germany		Germany; random address sample	Illumina Infinium iSelect		420
	MARIE	Mammary Carcinoma Risk Factor Investigation		Germany; randomly drawn from population registries	Illumina Infinium iSelect		1,712
10	MoMaTEC	Molecular Markers in Treatment of Endometrial Cancer	Norway; population based cases	Norway; female blood donors	Illumina Infinium iSelect	599	
	NBCS	Norwegian Breast Cancer Study		Norway; attendees at Norwegian Breast Cancer Screening Program	Illumina Infinium iSelect		234
11	CAHRES/RENDOCAS	Cancer Hormone Replacement Epidemiology	Sweden; population based cases		Illumina Infinium iSelect	543	1,345
	RENDOCAS	Registry of Endometrial Cancer in Sweden	Sweden; hospital based cases		Illumina Infinium iSelect	233	
	KARBAC	Karolinska Breast Cancer Study		Sweden; blood donors	Illumina Infinium iSelect		585
	pKARMA	Karolinska Mammography Project for Risk Prediction of Breast Cancer		Sweden; cancer-free participants of mammography screening	Illumina Infinium iSelect		4,987
	EC iCOGS total					**4,330**	**33,766**

Figure 6.1: Table displaying EC GWAS and iCOGS studies; sampling frame, genotyping platform and sample numbers. A total of 4,330 EC cases and 33,766 controls.

	Study	Case sampling frame	Control sampling frame	Genotyping Platform	Cases	Controls
CRC GWAS						
1	UK1-CORGI	Colorectal Tumour Gene Identification Consortium	England; Genetics clinic-based, with family history of CRC	England; spouses and partners of cases with no personal or family history of colorectal neoplasia	Illumina Hap550	888	899
2	Scotland1	Scotland	Scotland; population based CRC cases, age <55	Scotland; from NHS registers matched by age and region	Illumina HumanHap300, 240S	973	998
3	VQ	VICTOR/QUASAR2	UK; CRC cases enrolled in chemotherapy clinical trials		Illumina HumanHap300, 270, 1.2MDuo	1,894	2,674
	WTCCC2 BC58	UK 1958 Birth Cohort		UK; population based controls, born within one week in 1958	Illumina 1.2M		
4	CFR1	Colon Cancer Family Registry Phase1	USA and Australia; cases from cancer registries	USA and Australia; population based controls, no family history	Illumina Human1M	1,175	999
5	CFR2	Colon Cancer Family Registry Phase 2	USA and Australia; cases from cancer registries		Illumina Human1M	795	
	CGEMS prostate	Cancer Genetic Markers of Susceptibility (Prostate)		USA; population based cancer free controls from prostate study	Illumina HumanHap550		1,101
CRC GWAS total						**5,725**	**6,671**

Figure 6.2: Table displaying CRC GWAS; sampling frame, genotyping platform and sample numbers. A total of 5,725 CRC cases and 6,672 controls.

tematic over-dispersion of the test statistic for any of the 16 studies χ^2 (λ_{GC}=1.01-1.04 based on weakly correlated SNPs, r^2<0.2). Fixed-effects, inverse variance weighted meta-analysis was conducted for the 6 million well-imputed SNPs in the eight CRC and EC GWAS (8,935 cases, 13,396 controls) across the genome. For the 200,000 SNPs genotyped on the COGS array, the additional 4,330 EC cases and 26,849 controls from ECAC were included in a meta-analysis of 16 studies yielding a total of 13,265 cases and 40,245 controls for these SNPs. SNPs with globally significant CRC/EC associations ($P_{meta} < 5x10^{-8}$) were identified and the regions examined using standard fine-mapping and annotation methods.

6.4 Power calculations for combined EC and CRC GWAS

The combined sample size of 10,055 cancer cases and 40,437 controls provides greatly increased power to detect common variants that have an effect on both EC and CRC. This power analysis assumes that allele frequency and OR or a SNP is stable for EC and CRC. Figure 6.3 shows the power plot for the combined EC and CRC analysis, and for a SNP with MAF 0.2 and OR 1.15, there is 97% power to detect an association at $P=5x10^{-8}$.

The average MAF of CRC GWAS SNPs is 0.32, and assuming that these SNPs have a modest effect on EC (OR=1.1), this analysis has 98% power to detect an association at P=0.05 and 52% power at $P=10^{-4}$. The average minor allele frequency of EC GWAS SNPs is 0.35, and assuming these SNPs have a similar effect on CRC, (OR=1.1), this analysis has 95% power to detect an association at P=0.05 and 38% power at $P=10^{-4}$.

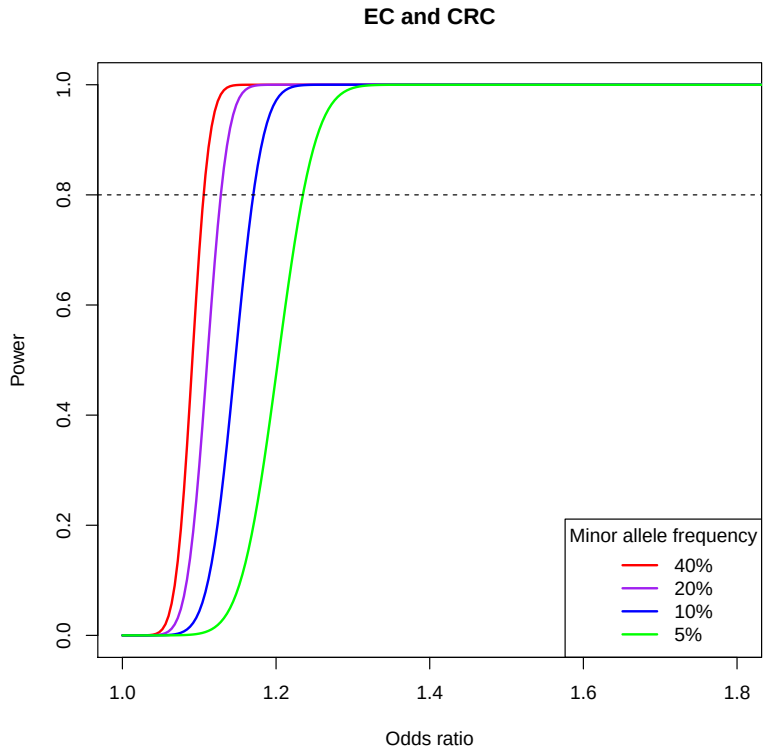

Figure 6.3: Power plot for combined EC and CRC analysis with 10,055 cancer cases and 40,437 controls. The graph displays the power to detect associations at P=5x10^{-8} given the odds ratio and minor allele frequency.

6.5 Assessment of established associations

The effects of 27 previously published tag-SNPs that have been formally associated with CRC risk in GWAS were investigated in EC. Seven EC SNPs were similarly investigated in CRC. All of these SNPs reach genome-wide significance in either CRC or EC with the exception of SNPs in the *TERT* region (for CRC, (Kinnersley *et al.*, 2012)), which is included owing to high-quality evidence from candidate-based studies. All SNPs were either discovered or replicated in European populations and were genotyped directly or had near-perfect proxies on the Illumina GWAS arrays used; many of these SNPs were also present on the iCOGS arrays.

The 34 previously-identified EC and CRC SNPs are listed in table 6.1 with the associated publications listed in table 6.2. 17 of the 34 SNPs showed the same direction of effect in both cancer types (that is, same nominal risk allele, irrespective of effect size); this was a non-significant deviation from randomness (sign test, P=1). rs2736100 lies in the intronic region of the telomerase reverse transcriptase *TERT*, and it or highly correlated SNPs have previously been associated with the risk of multiple different cancer types. Two other CRC SNPs (rs6691170 and rs10936599) were nominally associated with EC risk (P<0.05). Interestingly, the latter of these lies close to the telomerase RNA component *TERC* locus, and is a multi-cancer risk SNP (Chubb *et al.*, 2013; Figueroa *et al.*, 2014), which has been associated with longer telomeres (Codd *et al.*, 2013). No other previously-associated SNPs were significant at P<0.05 in the other cancer phenotype.

6.6 Novel predisposition loci for EC and CRC

In the fixed effect meta-analysis (section 2.3.10) of 8 EC and CRC GWAS, and the additional 8 iCOGS datasets for 200,000 SNPs, I searched for genome-wide significant SNPs ($P_{meta} < 5x10^{-8}$) that had not previously been identified in EC or CRC. I

Table 6.1: Established EC and CRC associations in other phenotype. Chr, chromosome; Pos, position; MAF, minor allele frequency; P, P-value in other cancer phenotype; OR, Odds ratio for minor allele; L95, lower; U95, and upper 95% confidence interval; Same direction of effects in both phenotypes; SNP genotyped in iCOGS with additional EC samples included. Significant SNPs (P<0.05) in bold.

Cancer		SNP	Chr	Pos	Nearby gene(s)	Minor Allele	MAF	P	OR	L95	U95	Same effect	iCOGs typed
CRC	1	rs1801133	1	11,856,378	MTHFR	A	0.34	0.686	0.99	0.92	1.06	Yes	No
CRC	2	rs10911251	1	183,081,194	LAMC1	C	0.43	0.236	1.04	0.97	1.12	No	No
CRC	3	**rs6691170**	1	222,045,446	DUSP10	T	0.37	0.023	1.09	1.01	1.17	Yes	No
CRC	4	**rs10936599**	3	169,492,101	TERC	T	0.24	0.033	0.92	0.84	0.99	Yes	No
CRC	5	**rs2736100**	5	1,286,516	TERT	A	0.5	0.000167	0.93	0.89	0.96	No	Yes
CRC	6	rs647161	5	134,499,092	PITX1	C	0.33	0.559	1.02	0.95	1.1	No	No
CRC	7	rs1321311	6	36,622,900	CDKN1A	A	0.24	0.925	1.00	0.92	1.08	No	No
CRC	8	rs16892766	8	117,630,683	EIF3H	C	0.09	0.134	0.95	0.88	1.02	No	Yes
CRC	9	rs6983267	8	128,413,305	MYC	T	0.46	0.143	1.03	0.99	1.07	No	Yes
CRC	10	rs10795668	10	8,701,219	GATA3	A	0.32	0.715	0.99	0.92	1.06	Yes	No
CRC	11	rs1035209	10	101,345,366	NKX2-3, SLC25A28	T	0.2	0.243	1.05	0.97	1.15	Yes	No
CRC	12	rs3824999	11	74,345,550	POLD3	T	0.49	0.647	0.98	0.92	1.05	Yes	No
CRC	13	rs3802842	11	111,171,709	COLCA1, COLCA2,POU2AF1	C	0.31	0.513	0.99	0.94	1.03	No	Yes
CRC	14	rs10774214	12	4,368,352	CCND2	T	0.38	0.171	1.05	0.98	1.13	Yes	Yes
CRC	15	rs3217810	12	4,388,271	CCND2	T	0.14	0.762	1.02	0.92	1.13	Yes	No
CRC	16	rs11169552	12	51,155,663	DIP2B, ATF1	T	0.26	0.963	1.00	0.93	1.08	No	No
CRC	17	rs4444235	14	54,410,919	BMP4	C	0.48	0.100	1.03	0.99	1.07	Yes	Yes
CRC	18	rs1957636	14	54,560,018	BMP4	T	0.41	0.961	1.00	0.96	1.04	No	Yes
CRC	19	rs16969681	15	32,993,111	GREM1	T	0.09	0.379	0.97	0.9	1.04	No	Yes
CRC	20	rs11632715	15	33,004,247	GREM1	A	0.48	0.332	1.04	0.97	1.11	Yes	No
CRC	21	rs9929218	16	68,820,946	CDH1, CDH3	A	0.29	0.679	0.98	0.91	1.06	Yes	No
CRC	22	rs4939827	18	46,453,463	SMAD7	C	0.46	0.229	0.98	0.94	1.02	Yes	Yes
CRC	23	rs10411210	19	33,532,300	RHPN2	T	0.09	0.202	1.04	0.98	1.12	No	Yes
CRC	24	rs961253	20	6,404,281	BMP2	A	0.37	0.975	1.00	0.96	1.04	No	Yes
CRC	25	rs4813802	20	6,699,595	BMP2	G	0.37	0.268	1.04	0.97	1.12	Yes	No
CRC	26	rs2423279	20	7,812,350	HAO1	C	0.24	0.897	1.01	0.93	1.09	Yes	No
CRC	27	rs4925386	20	60,921,044	LAMA5	T	0.3	0.064	1.07	1	1.16	No	No
EC	28	rs13328298	6	126,016,580	HEY2, NCOA7	A	0.43	0.518	0.98	0.94	1.03	No	No
EC	29	rs4733613	8	129,599,278	MYC	C	0.12	0.875	1.01	0.93	1.09	No	Yes
EC	30	rs11841589	13	73,814,891	KLF5, KLF12	T	0.29	0.789	1.01	0.95	1.07	Yes	Yes
EC	31	rs2498796	14	105,243,220	AKT1	A	0.31	0.556	1.02	0.96	1.08	No	No
EC	32	rs937213	15	40,322,124	EIF2AK, BMF	C	0.42	0.302	0.97	0.93	1.02	Yes	Yes
EC	33	rs2414098	15	51,537,806	CYP19A1	T	0.39	0.131	0.96	0.91	1.01	No	No
EC	34	rs11263763	17	36,103,565	HNF1B	G	0.48	0.737	1.01	0.96	1.06	Yes	No

Table 6.2: References for EC and CRC associations listed in table above. All associations have reached genome-wide significance apart from CRC 5 *TERT*, which is included due to evidence from candidate studies.

Cancer		Reference
CRC	1	Hubner et al. Int Journal Cancer2006
CRC	2	Peters et al. Gastroenterology 2013
CRC	3	Houlston et al. Nat Genet 2010
CRC	4	Houlston et al. Nat Genet 2010
CRC	5	Rafnar et al. Nat Genet 2009
CRC	6	Jia et al. Nat Genet 2013
CRC	7	Dunlop et al. Nat Genet 2012
CRC	8	Tomlinson et al. Nat Genet 2008
CRC	9	Tomlinson et al. Nat Genet 2007
CRC	10	Tomlinson et al. Nat Genet 2008
CRC	11	Whiffin et al. Hum Mol Genet 2014
CRC	12	Dunlop et al. Nat Genet 2012
CRC	13	Tenesa et al. Nat Genet 2008
CRC	14	Whiffin et al. Hum Mol Genet 2014
CRC	15	Peters et al. Gastroenterology 2013
CRC	16	Houlston et al. Nat Genet 2010
CRC	17	Houlston et al. Nat Genet 2008
CRC	18	Tomlinson et al. PLoS Genetics 2011
CRC	19	Tomlinson et al. PLoS Genetics 2011
CRC	20	Tomlinson et al. PLoS Genetics 2011
CRC	21	Houlston et al. Nat Genet 2008
CRC	22	Broderick et al. Nat Genet 2007
CRC	23	Houlston et al. Nat Genet 2008
CRC	24	Houlston et al. Nat Genet 2008
CRC	25	Tomlinson et al. PLoS Genetics 2011
CRC	26	Whiffin et al. Hum Mol Genet 2014
CRC	27	Houlston et al. Nat Genet 2010
EC	28	See chapter 4
EC	29	See chapter 4
EC	30	See chapter 4
EC	31	See chapter 4
EC	32	See chapter 4
EC	33	See chapter 4
EC	34	Spurdle et al. Nat Genet 2011

discovered a single genome-wide significant SNP rs3184504, on chromosome 12q24 (OR=1.10, 95%CI=1.07-1.13, P_{meta}=7.23x10^{-9}, heterogeneity I^2=0). This SNP is a missense variant (p.Trp262Arg) in exon 4 of *SH2B3*. The rs3184504-C allele is consistently the risk allele in all EC and CRC datasets including the iCOGS array where this SNP was included due to a promising but yet unproven association with breast cancer risk (figure 6.4).

The top 15 SNPs from this meta-analysis are displayed in table 6.3, and SNPs in LD with previous associations have not been listed. Some SNPs that have been previously been associated with EC or CRC (listed in table 6.1) came up as highly significant in this analysis but were less significant after including the GWAS of the other cancer phenotype.

There are SNPs that have previously been independently identified in GWAS of different phenotypes where the risk allele for one phenotype is the protective allele for another (Chung *et al.*, 2014; Sivakumaran *et al.*, 2011). In order to search for SNPs for which the same allele has differing directions of effect in CRC and EC, I conducted a fixed-effects meta-analysis with the odds ratios of all the CRC GWAS SNPs inverted.

In this analysis, I discovered rs12970291 on chromosome 18q22, where the major G allele is protective in CRC (OR=0.78, 95%CI=0.69-0.90, P=3.42x10^{-4}) and confers risk in EC (OR=1.24, 95%CI=1.11-1.38, P=1.11x10^{-4}). In meta-analysis, the rs12970291 association reached genome-wide significance (OR:1.26, 95%CI:1.16-1.38, P_{meta}=4.82x10^{-8}, figure 6.5). The significance and interpretation of this result is discussed later in this chapter, and it is clear that replication is important for this SNP in both EC and CRC in subsequent GWAS and meta-analyses.

Table 6.4 displays the top 15 SNPs in the EC CRC meta-analysis where the ORs of CRC SNPs are reversed. Thus, a particular allele for these SNPs is protective for either EC or CRC, and is a risk allele in the other cancer phenotype.

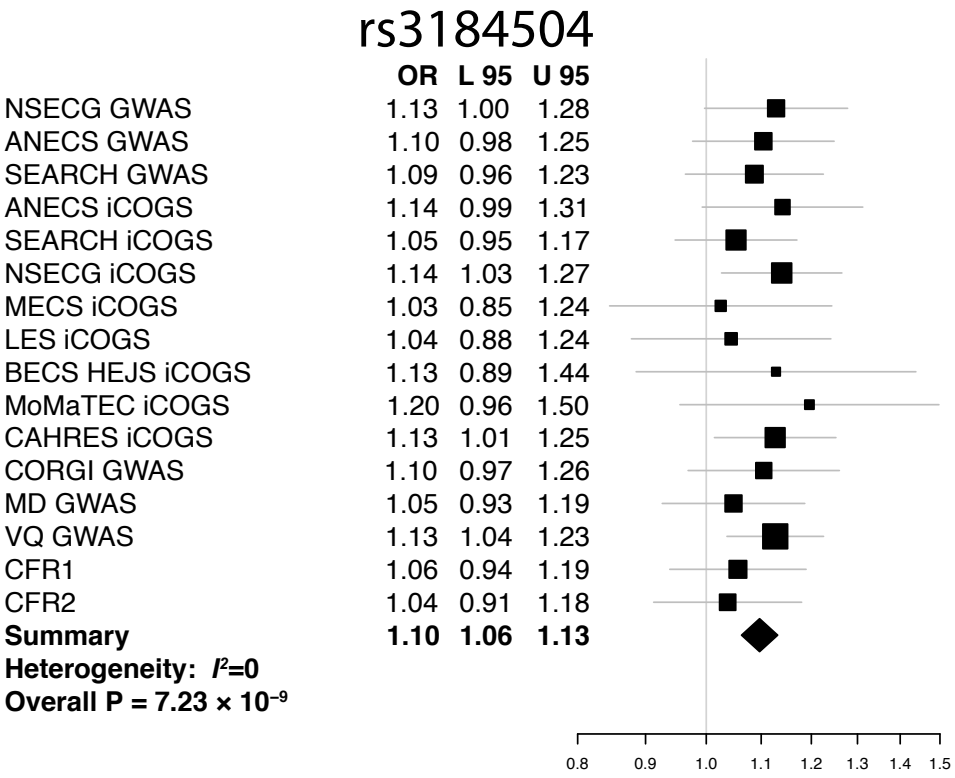

Figure 6.4: Forest plot for association between cancer risk and chromosome 12q24 rs3184504 genotype in EC GWAS, EC iCOGS and CRC GWAS. Black squares represent the point estimate of the odds ratio and have areas proportional to study size. Lines represent 95% confidence intervals. The diamond shows the summary statistic. The overall meta-analysis P-value and heterogeneity statistic are shown.

Table 6.3: Meta-analysis results for EC, CRC GWAS and EC iCOGS where that SNP is directly genotyped. Chr, chromosome; Pos, position; Ref, reference allele; AF, allele frequency; Odds ratio, upper and lower 95% confidence interval; P, meta-analysis P-value; I^2, heterogeneity I^2 statistic; iCOGS typed, genotyped in 8 additional iCOGS groups. Genome-wide significant SNP in bold.

SNP	Chr	Pos	Nearby gene	Ref	AF	OR	L95	U95	P	I^2	iCOGs typed
rs4378954	3	115,650,448	*LSAMP*	C	0.898	1.15	1.09	1.21	4.05E-07	0.10	Yes
rs17035310	4	106,064,754	*TET2*	C	0.876	1.13	1.08	1.18	6.10E-07	0.00	Yes
rs3181245	6	24,651,320	*TDP2*	C	0.514	1.12	1.07	1.16	2.23E-07	0.45	No
rs17503919	6	89,565,737	*RNGTT*	A	0.862	1.17	1.11	1.25	1.66E-07	0.00	No
rs10457678	6	139,122,240	*ECT2L*	A	0.773	1.10	1.06	1.14	4.12E-07	0.00	Yes
rs7740797	6	155,121,259	*SCAF8*	G	0.513	1.12	1.07	1.16	2.20E-07	0.36	No
rs10217586	9	22,121,349	*CDKN2B-AS1*	T	0.539	0.93	0.90	0.96	1.83E-06	0.74	Yes
rs76225372	10	19,320,086	*ARL5B*	A	0.912	1.21	1.13	1.31	1.68E-07	0.00	No
rs1512436	11	106,306,871	*GUCY1A2*	T	0.553	0.93	0.90	0.95	1.22E-06	0.00	Yes
rs2052678	12	29,833,329	*TMTC1*	G	0.861	0.86	0.81	0.91	1.38E-07	0.31	No
rs3184504	**12**	**111,884,608**	***SH2B3***	**C**	**0.520**	**1.10**	**1.06**	**1.13**	**8.23E-09**	**0.00**	**Yes**
rs12446552	16	11,720,066	*LITAF*	T	0.723	0.92	0.89	0.95	3.05E-06	0.24	Yes
rs9901225	17	40,755,811	*PSMC3IP*	T	0.551	1.11	1.06	1.16	1.13E-06	0.49	No
rs11085466	19	21,751,811	*ZNF429*	G	0.782	1.15	1.09	1.22	1.53E-07	0.00	No
rs6051080	20	25,975,674	*NANP*	A	0.649	1.12	1.08	1.17	1.23E-07	0.00	No

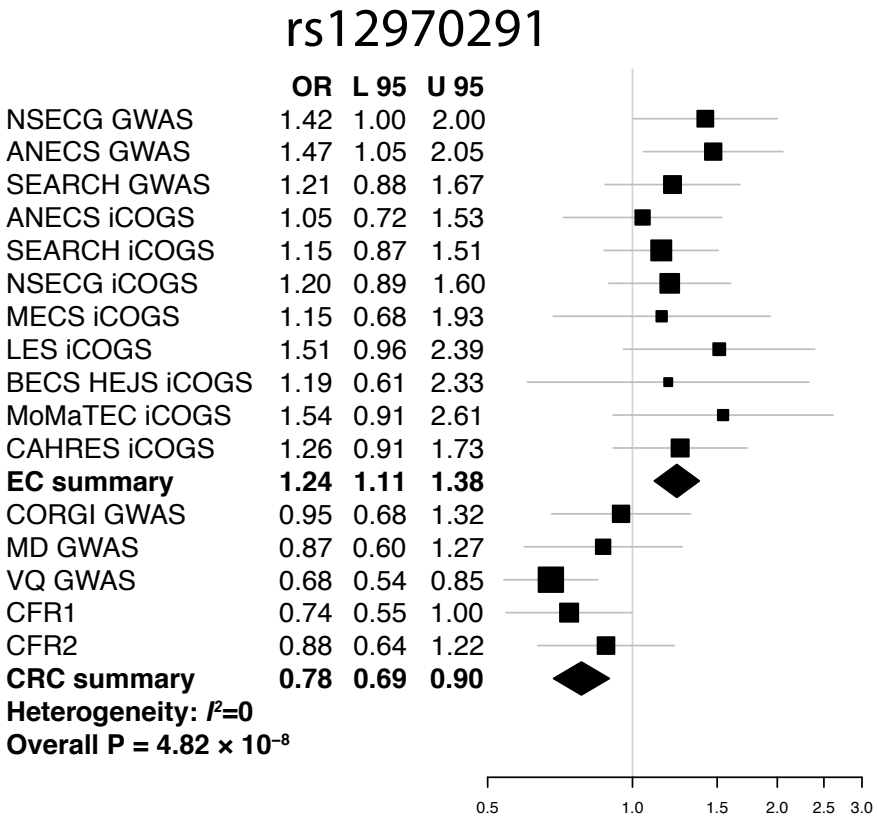

Figure 6.5: Forest plot for chromosome 18q22 rs12970291 in EC GWAS, EC iCOGS and CRC GWAS. Black squares represent the point estimate of the odds ratio and have areas proportional to study size. Lines represent 95% confidence intervals. The diamond shows the summary statistic in EC, CRC and overall meta-analysis. The overall meta-analysis P-value and heterogeneity statistic is shown.rs12970291-A is consistently the risk allele for EC and protective allele for CRC. The overall P-value is derived from meta-analysis of rs12970291 risk allele for both EC and CRC.

Table 6.4: Meta-analysis results for EC, CRC GWAS and EC iCOGS where that SNP is directly genotyped, with CRC ORs reversed. Chr, chromosome; Pos, position; Ref, reference allele, AF, allele frequency; Odds ratio, upper and lower 95% confidence interval; P, meta-analysis P-value; I^2, heterogeneity I^2 statistic; iCOGS typed, genotyped in 8 additional iCOGS groups. Genome-wide significant SNP in bold.

SNP	Chr	Pos	Nearby gene	Ref	AF	OR	L95	U95	P	I^2	iCOGs typed
rs2270539	1	78,047,628	*ZZZ3*	T	0.979	0.80	0.72	0.88	1.60E-05	0.00	Yes
rs34462569	2	166,552,918	*CSRNP3*	G	0.934	1.22	1.12	1.33	3.75E-06	0.03	No
rs1852266	7	48,340,033	*ABCA13*	C	0.741	0.92	0.89	0.96	2.27E-05	0.00	Yes
rs4727012	7	148,740,493	*EZH2*	C	0.891	1.18	1.10	1.26	9.64E-07	0.34	No
rs112541862	9	97,483,158	*FBP1*	A	0.962	1.30	1.16	1.45	4.10E-06	0.03	No
rs77922938	9	111,060,196	*KLF4*	G	0.953	0.85	0.79	0.91	8.33E-06	0.00	Yes
rs606460	11	75,248,457	*GDPD5*	G	0.725	1.11	1.06	1.17	3.58E-06	0.29	No
rs11607499	11	81,811,229	*PRCP*	T	0.954	1.27	1.15	1.40	1.75E-06	0.00	No
rs12817211	12	50,579,398	*LIMA1*	C	0.767	1.12	1.07	1.18	2.76E-06	0.50	No
rs3117967	13	22,508,254	*FGF9*	G	0.896	0.89	0.85	0.94	1.33E-05	0.17	Yes
rs1952157	14	34,289,438	*NPAS3*	C	0.690	1.08	1.05	1.12	6.26E-06	0.52	Yes
rs11150038	16	78,076,559	*CLEC3A*	G	0.944	1.27	1.15	1.39	9.89E-07	0.00	No
rs76525880	17	46,614,250	*HOXB1*	A	0.929	1.16	1.09	1.24	6.68E-06	0.58	Yes
rs12970291	**18**	**73,011,336**	***TSHZ1***	**T**	**0.965**	**1.26**	**1.16**	**1.38**	**4.82E-08**	**0.00**	**Yes**
rs4816890	21	24,420,502	NCAM2	T	0.637	0.90	0.86	0.94	2.13E-06	0.24	No

6.7 Fine-mapping and functional annotation of novel loci

Fine-mapping of the chromosome 12q24 and 18q22 locus was done using the eight EC and CRC GWAS, including SNPs that had been imputed with an info score greater than 0.9. The eight EC iCOGS groups were not included because this SNP array was not aimed at genome-wide coverage and use of the eight GWAS studies allows a fairer comparison between SNPs at a particular locus.

The regional association plot of chromosome 12q24 shows that rs3184504 is the most significant SNP, with an additional three SNPs in strong pairwise LD (r^2>0.9) with rs3184504 (figure 6.5) showing strong evidence of CRC-EC association ($P_{fine-mapping}$ $<10^{-5}$). Conditional analysis suggests the existence of a single association signal in this region and these four highly significant SNPs lie in a 68kb region that includes the genes *SH2B3* and *ATXN2* and their functional annotation is shown in table 6.5. None of the 4 SNPs was associated with the mRNA level of *SH2B3*, *ATXN2* or other nearby genes in public eQTL databases in endometrial or other tissues. There is however evidence of functional effects in terms of enhancer histone marks and motif changes to suggest that these correlated SNPs may be of regulatory importance.

Fine-mapping analysis of chromosome 18q22 identified a large number of SNPs in high pairwise LD with rs12970291 (r^2>0.85), in a 70kb region that includes the gene *TSHZ1*, 15.0kb centromeric to rs12970291 (figure 6.7a). Seventeen SNPs had a stronger disease association than rs12970291 in fine-mapping, with the lowest P value at rs35185115 ($P_{fine-mapping}$=1.08x10^{-6}, pairwise r^2 with rs12970291 = 1). Conditional analysis revealed a single association signal at this *SH2B3-ATNX2* region.

Fine-mapping of EC and CRC GWAS separately (figure 6.7 b and c) showed an association peak in the same LD block between 10.5-51.8kb downstream of *TSHZ1*. The co-occuring peak around 73,017,234 suggests that there may be a single causal variant

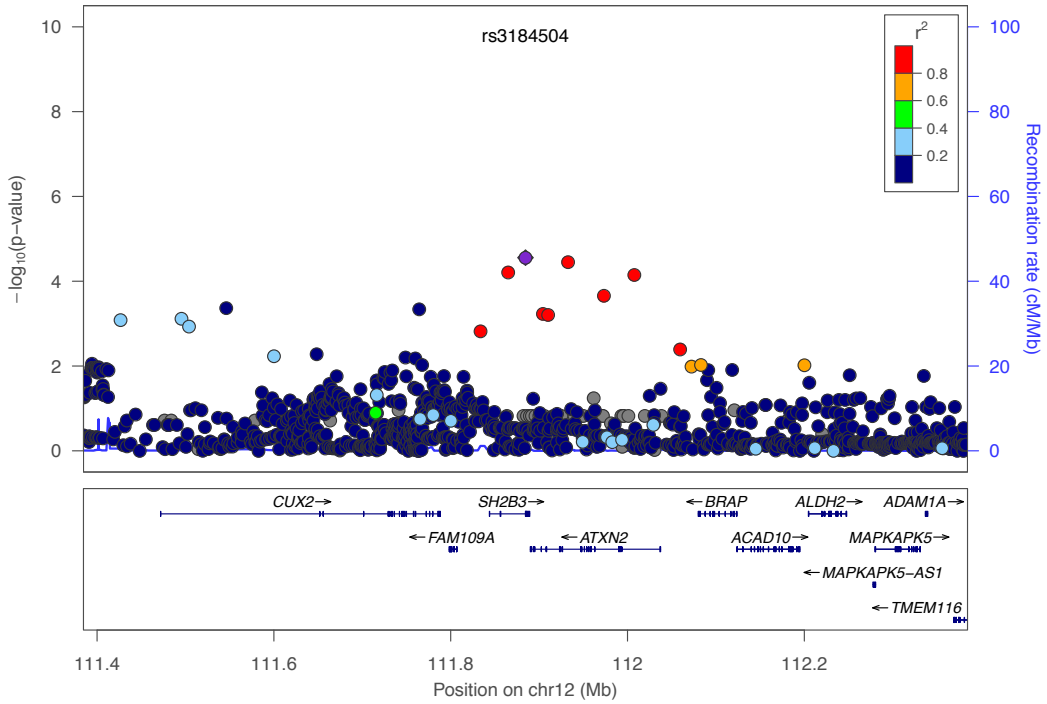

Figure 6.6: Regional association plot of chromosome 12q24 around rs3184504. The top SNP rs3184504 is marked as a purple diamond, and the dot color represents the LD with the top SNP. Y-axis is -$\log_{10}$ P-values from imputation and meta-analysis for the eight EC and CRC GWAS. The blue line shows recombination rates in cM/Mb.

Table 6.5: Functional annotation of SNPs in LD (r^2>0.4) with rs3184504 (in bold) using Haploreg and RegulomeDB. Positions in build 19, LD r^2 with rs3184504; ref, reference allele; AF: allele frequency, promoter and enhancer histone marks, DHS: DNase1 hypersensitivity sites, RegDB RegulomeDB score , RegDB score of 3a TF binding+any motif+DNase peak, 4 represents TF binding+DNase peak, 5 represents TF binding or DNase peak, 6 represents other. Blank spaces represent no data.

SNP	Chr 12 Pos	LD	Ref	AF	RefSeq genes	Location	Promoter	Enhancer	DHS	Altered Motifs	RegDB
rs10774624	111,833,788	0.81	G	0.53	10kb 5' of *SH2B3*	intergenic				Ets,Nr2f2	
rs7310615	111,865,049	0.93	C	0.53	*SH2B3*	intronic		GM12878, Huvec		ERalpha-a	
rs3184504	111,884,608	1	T	0.53	*SH2B3*	missense		Huvec, GM12878		HES1,Mtf1	3a
rs4766578	111,904,371	0.9	T	0.51	*ATXN2*	intronic		NHLF		7 motifs	6
rs10774625	111,910,219	0.9	A	0.52	*ATXN2*	intronic		K562		9 motifs	5
rs7137828	111,932,800	0.87	C	0.53	*ATXN2*	intronic				SP1	5
rs597808	111,973,358	0.75	A	0.51	*ATXN2*	intronic				7 motifs	
rs653178	112,007,756	0.78	C	0.53	*ATXN2*	intronic	8 organs	7 marks		Esr2	5
rs11065979	112,059,557	0.61	C	0.44	20kb 3' of *BRAP*	intergenic				Myf	5
rs11065987	112,072,424	0.51	A	0.42	7.5kb 3' of *BRAP*	intergenic		K562	K562	Mrg1,Hoxa9	3a
rs11065991	112,083,162	0.5	C	0.42	*BRAP*	intronic				Myc	4

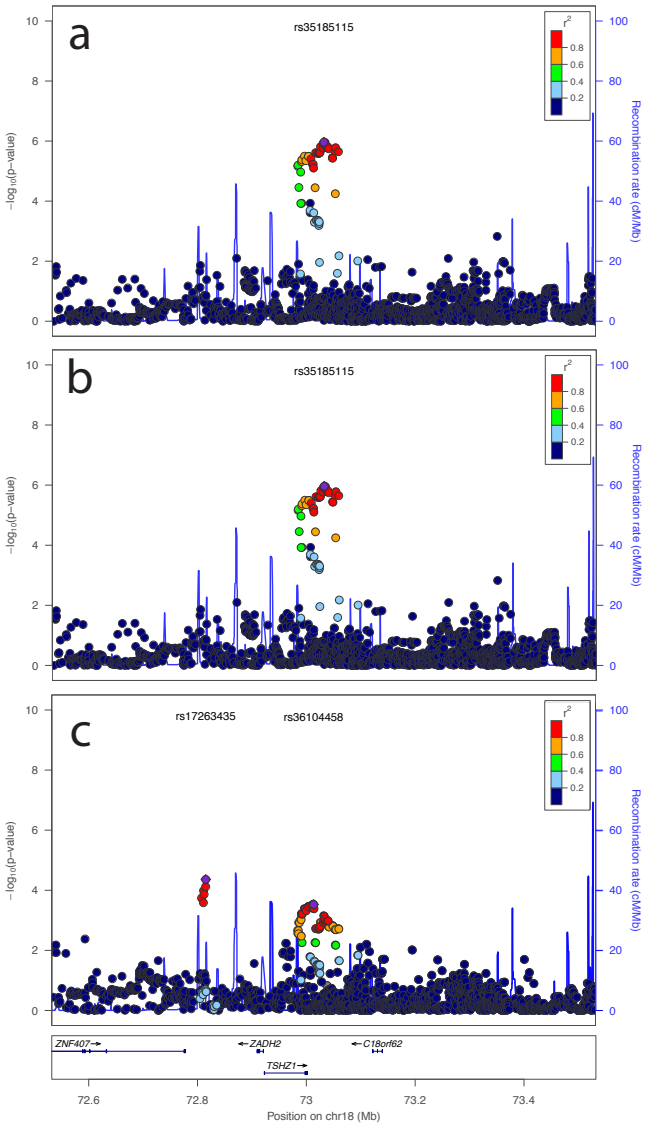

Figure 6.7: Regional association plot for chromosome 18q22 around rs1297029 for a: CRC and EC GWAS b: CRC GWAS and c: EC GWAS. rs12970291 ($P_{CRC/ECmeta}$=4.82x10^{8}, top SNP typed in iCOGS for CRC and EC analysis), rs35185115 ($P_{CRC/ECmeta}$=1.08x10^{-6}, top SNP typed or well-imputed in eight EC and CRC GWAS), rs71359086 (P_{CRC}=2.24×10^{-4}, top SNP in CRC GWAS), and rs36104458 (P_{EC}=4.35×10^{-5}, top SNP in EC GWAS) all lie between 10.5 - 51.8kb downstream of *TSHZ1* and are in perfect LD (r^2=1). An additional peak, rs17263435 (P_{EC}=4.35x10^{-5}) is observed in the EC GWAS, but is not significant in CRC (P=0.10). The top SNP in each plot is marked as a purple diamond, and the dot color represents the LD with the top SNP. Y-axis is -log$_{10}$ P-values, the blue line shows recombination rates in cM/Mb.

downstream of *TSHZ1* present in EC and CRC acting in opposite directions. An additional signal at rs17263435 (P_{EC}=4.35x10^{-5}, figure 6.7 c) for EC is not replicated in CRC (P_{CRC}=0.10). Several SNPs in the region have potential functional importance (table 6.6). Of particular note is the missense SNP rs3390274 (p.Ala468Thr, pairwise r^2 with rs17263435)=0.77 in the last exon of *TSHZ1*. SNPs with a pairwise LD of >0.4 with rs12970291 in the region were not significantly associated with mRNA level of *TSHZ1* or other nearby genes in public eQTL databases.

6.8 Genome-wide enrichment for EC and CRC

Beyond SNPs that are genome-wide significant for EC or CRC and those that are genome-wide significant in a combined meta-analysis, there may be a large number of SNPs with promising P-values (i.e. P<10^{-3}) for both phenotypes. Observing a large number of SNPs with promising P-values for both cancer phenotypes may suggest genome-wide enrichment and that there may be more SNPs that predispose to both EC and CRC, suggesting a common genetic aetiology. The selection of P-value cutoffs for such an analysis is a balance between including the largest number of SNPs that contribute to the phenotype however small the effect, while excluding SNPs that that do not contribute and essentially represent added noise. I assessed a range of P-value cutoffs (P=10^{-3} to P=0.05) employing methods as described by Rahmioglu *et al.* (2015) to see if there was a larger number of shared predisposition SNPs observed than one would expect by chance. The χ^2 test was used on the resulting two by two contingency table (figure 6.8).

Genome-wide enrichment analysis for 246,896 independent loci (r^2<0.1) below genome-wide significance showed no observable statistical enrichment between CRC and EC SNPs. There were no SNPs with a P-value of less than 10^{-4} in both CRC and EC. When using a EC and CRC cutoff of P <0.05 and 10^{-3}, the χ^2 test produced a

Table 6.6: Functional annotation of SNPs in LD (r^2>0.4) with rs12970291 (in bold) using Haploreg and RegulomeDB. Positions in build 19, LD r^2 with rs12970291; ref, reference allele; AF: allele frequency, promoter and enhancer histone marks, DHS: DNase1 hypersensitivity sites, RegDB RegulomeDB score , RegDB score of 2b represents TF binding+any motif+DNase Footprint+DNase peak, 4 represents TF binding+DNase peak, 5 represents TF binding or DNase peak, 6 represents other. Blank spaces represent no data.

SNP	Pos	LD	Ref	AF	RefSeq genes	Location	Promoter	Enhancer	DHS	Motifs	RegDB
rs12969532	72,984,262	0.52	C	0.03	*TSHZ1*	intronic				3 motifs	5
rs71359079	72,984,455	0.5	C	0.03	*TSHZ1*	intronic				6 motifs	6
rs12965321	72,986,176	0.59	G	0.02	*TSHZ1*	intronic				4 motifs	5
rs34981053	72,989,467	0.59	T	0.02	*TSHZ1*	intronic		Huvec, HSMM	35 cell types	3 motifs	5
rs28420499	72,989,742	0.5	G	0.03	*TSHZ1*	intronic		Huvec, HSMM	5 cell types	7 motifs	5
rs77026922	72,991,038	0.68	T	0.03	*TSHZ1*	intronic			7 cell types	13 motifs	5
rs56280961	72,991,193	0.4	C	0.02	*TSHZ1*	intronic			Medullo	Pax-5,Pax-6	5
rs34551261	72,991,577	0.68	G	0.03	*TSHZ1*	intronic			6 cell types	8 motifs	4
rs36046989	72,996,430	0.73	C	0.03	*TSHZ1*	intronic			ProgFib	4 motifs	4
rs33930274	72,998,899	0.77	G	0.03	*TSHZ1*	missense			4 cell types	Pou2f2	5
rs34242916	73,003,598	0.73	T	0.03	1.7kb 3' of *TSHZ1*	intergenic		HepG2	HepG2	NRSF	5
rs34704942	73,003,694	0.73	C	0.03	1.8kb 3' of *TSHZ1*	intergenic			4 cell types	5 motifs	5
rs12969489	73,008,360	0.81	C	0.03	6.5kb 3' of *TSHZ1*	intergenic				5 motifs	
rs9946654	73,011,336	0.81	T	0.03	9.4kb 3' of *TSHZ1*	intergenic				3 motifs	
rs36104458	73,012,489	0.81	A	0.03	11kb 3' of *TSHZ1*	intergenic			CLL	Gfi1	5
rs12962160	73,013,181	0.81	A	0.03	11kb 3' of *TSHZ1*	intergenic				HNF1,PLZF	
rs34227877	73,013,811	0.57	A	0.02	12kb 3' of *TSHZ1*	intergenic			Fibrobl	3 motifs	5
rs17056795	73,016,067	0.69	G	0.04	14kb 3' of *TSHZ1*	intergenic			Fibrobl	3 motifs	5
rs12970291	73,017,234	1	G	0.03	15kb 3' of *TSHZ1*	intergenic				9 motifs	5
rs12968532	73,020,946	1	C	0.03	19kb 3' of *TSHZ1*	intergenic				4 motifs	
rs35969565	73,022,413	1	C	0.03	21kb 3' of *TSHZ1*	intergenic				GLI	
rs12959779	73,023,486	1	G	0.03	22kb 3' of *TSHZ1*	intergenic			Medullo,Osteobl	9 motifs	5
rs12965434	73,024,810	1	G	0.03	23kb 3' of *TSHZ1*	intergenic					
rs12961817	73,024,904	1	C	0.03	23kb 3' of *TSHZ1*	intergenic			Melano	4 motifs	5
rs78784067	73,025,962	0.91	C	0.03	24kb 3' of *TSHZ1*	intergenic				Hmx,Sox	5
rs2581665	73,030,498	1	C	0.97	29kb 3' of *TSHZ1*	intergenic				3 motifs	5
rs12955930	7,3030,682	1	C	0.03	29kb 3' of *TSHZ1*	intergenic				3 motifs	5
rs35185115	73,032,159	1	A	0.03	30kb 3' of *TSHZ1*	intergenic			PanIslets	Foxp3,Mef2	5
rs34521731	73,032,447	0.91	TC	0.03	31kb 3' of *TSHZ1*	intergenic			3 cell types	11 motifs	
rs12969979	73,032,676	1	G	0.03	31kb 3' of *TSHZ1*	intergenic			LNCaP	YY1	5
rs12966313	73,032,803	1	C	0.03	31kb 3' of *TSHZ1*	intergenic				4 motifs	6
rs12956172	73,033,299	1	A	0.03	31kb 3' of *TSHZ1*	intergenic			3 cell types	3 motifs	2b
rs12959858	73,034,322	1	G	0.03	32kb 3' of *TSHZ1*	intergenic			FibroP,PanIslets	3 motifs	5
rs35181847	73,03,7586	1	C	0.03	36kb 3' of *TSHZ1*	intergenic				7 motifs	
rs12963413	73,039,236	1	C	0.03	37kb 3' of *TSHZ1*	intergenic				7 motifs	6
rs4453592	73,040,883	0.85	T	0.02	39kb 3' of TSHZ1	intergenic				3 motifs	
rs34468146	73,047,740	0.95	T	0.03	46kb 3' of TSHZ1	intergenic			Osteobl	Gm397,Mtf1	5
rs71359085	73,048,182	0.95	G	0.03	46kb 3' of TSHZ1	intergenic			Medullo	5 motifs	5
rs34240979	73,052,310	0.95	C	0.03	50kb 3' of TSHZ1	intergenic				12 motifs	5
rs12953745	73,053,195	0.66	C	0.02	51kb 3' of TSHZ1	intergenic			Th1,Medullo	Irf,SIX5	5
rs71359086	73,053,690	0.95	C	0.03	52kb 3' of TSHZ1	intergenic				5 motifs	5
rs12956810	73,058,905	0.95	C	0.03	57kb 3' of TSHZ1	intergenic			H1-hESC	RXRA	5

		Endometrial Cancer					
		$P<1x10^{-3}$	$P\geq1x10^{-3}$	$P<1x10^{-2}$	$P\geq1x10^{-2}$	$P<0.05$	$P\geq0.05$
Colorectal Cancer	$P<1x10^{-3}$	1	435	8	428	38	398
	$P\geq1x10^{-3}$	420	246040	3361	243099	14957	231503
		P=0.766		P=0.397		P=0.0208	
	$P<1x10^{-2}$	7	3346	50	3319	224	3129
	$P\geq1x10^{-2}$	414	243129	3303	240224	14771	228772
		P=0.589		P=0.524		P=0.138	
	$P<0.05$	24	14448	193	14279	913	13559
	$P\geq0.05$	397	232027	3176	229248	14082	218342
		P=0.888		P=0.524		P=0.222	

Figure 6.8: Table displaying genome-wide enrichment in CRC and EC. Multiple P-value cutoffs are used for EC (horizonal) and CRC (vertical edge). χ^2 tests are done using the number of SNPs in each cutoff category and test P-values are displayed in boxes below contingency table.

nominally significant P-value of 0.021, but there was otherwise no excess of significant SNPs observed.

6.9 Discussion

Colorectal carcinoma (CRC) is the fourth commonest cancer in the western world and cancer of the uterine corpus, or endometrial carcinoma (EC), is the fourth commonest cancer among women. Both are the cause of significant morbidity and mortality worldwide and high-penetrance Mendelian cancer syndromes provide evidence of a shared genetic aetiology in some cases. Using a combined CRC and EC GWAS meta-analysis, I have identified a region on chromosome 12q24.1 spanning two genes, *SH2B3* and *ATXN2*, which contains a SNP formally associated with cancer risk. Of the variants in this region, rs3184504 is of particular interest, because it is a non-synonymous change (TGG → CGG; p.Trp262Arg) in the pleckstrin homology domain of *SH2B3*, which is *a priori* a much stronger candidate than the spinocerebellar ataxia gene *ATXN2*. *SH2B3* is a member of the SH2B adaptor family of proteins

and is involved in a range of signalling activities by growth factor and cytokine receptors. It is a key negative regulator in cytokine signalling in haematopoiesis, and is expressed at a high level in the bone marrow and leucocytes, but at a low level in the normal colon and endometrium (EMBL-EBI expression atlas). Comparative genomics shows that the rs3184504 risk allele (C, Arg residue) is conserved in all primates and some vertebrates, and has a much lower allele frequency (∼0.5) in Europeans than in African, Asian and admixed American populations (∼1.0). Tryptophan and arginine possess a hydrophobic and positively charged side chain respectively and different *in silico* methods predict that the amino acid substitution has varying levels of impact (Grantham score=121 (range 0 – 215, (Grantham, 1974)), Polyphen2 score=0.12 (Adzhubei *et al.*, 2010) and SIFT score of 1.0 (Kumar *et al.*, 2009), CADD score PHRED-scaled: 5.53 (Kircher *et al.*, 2014), leaving open the possibility of a modest effect on protein function. The NHGRI GWAS Catalog shows that SNPs in the *SH2B3-ATNX2* region including rs3184504 and rs653178 have been previously associated with immune-mediated conditions: coeliac disease (Hunt *et al.*, 2008), rheumatoid arthritis (Stahl *et al.*, 2010), type 1 diabetes (Barrett *et al.*, 2009), autoimmune hepatitis(de Boer *et al.*, 2014) and also cardiovascular traits including coronary artery disease (Dichgans *et al.*, 2014) and blood pressure(Wain *et al.*, 2011). rs653178 has been linked to *SH2B3* expression in peripheral blood cell eQTL analysis ($P=9.24x10^{-12}$), although this association is not present in public eQTL data sets. Interestingly, rs3184504-T is generally the risk allele in autoimmune traits, suggesting opposing effects of the functional polymorphism on cancer and other traits, perhaps via shared effects on immune activation. A similar phenomenon has been found for the *HNF1B* SNP rs4430796 which has opposing effects on EC and type 2 diabetes risk(Painter *et al.*, 2014).

The *TERT-CLPTM1L* locus has been identified in multiple cancer susceptibility GWAS including breast (Haiman *et al.*, 2011), prostate (Kote-Jarai *et al.*, 2011), lung

(McKay *et al.*, 2008), pancreas (Petersen *et al.*, 2010), glioma (Shete *et al.*, 2009), testicular germ cell (Turnbull *et al.*, 2010b), and basal cell carcinoma (Stacey *et al.*, 2009). It is of interest that the CRC SNP rs2736100 also shows signs of significance in EC in our analysis (OR=1.08, CI95%=1.04-1.12, P=1.67x10^{-4}). In parallel with this study, our group have recently performed a detailed analysis of the *TERT-CLPTM1L* locus in EC which provided evidence that rs7705526 is associated with EC risk (P_{assoc}=7.7x10^{-5}), albeit at locus-specific rather than genome-wide significance thresholds (Carvajal-Carmona *et al.*, 2014). rs7705526 is moderately correlated with rs2736100 ($r^2 \sim 0.5$) but is poorly tagged in most Illumina GWAS arrays. Figure 6.9 shows the complex LD structure between these two SNPs and 4 other SNPs previously associated with CRC and EC to varying levels of significance (P=8.4x10^{-3} – 4.9x10^{-6}) at this locus.

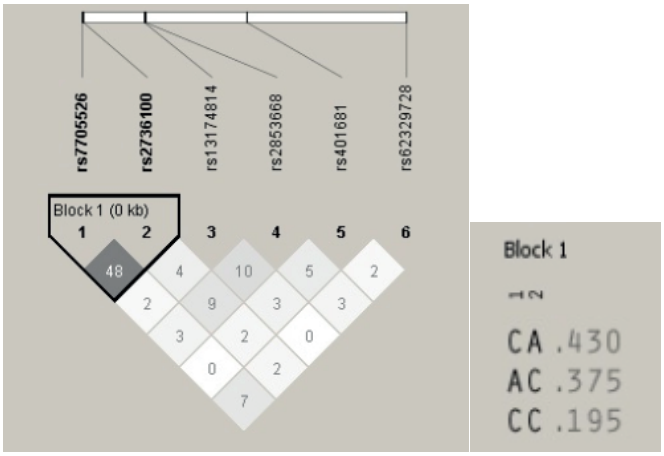

Figure 6.9: Haplotype structure in *TERT-CLPTM1L* in NSECG samples. rs2736100 was associated with CRC (Kinnersley *et al.*, 2012). rs401681 (Rafnar *et al.*, 2009) and rs2853668 (Peters *et al.*, 2012) have also been shown to be associated with CRC in previous studies of smaller scale. rs7705526, rs13174814 and rs62329728 have been recently associated with EC (Carvajal-Carmona *et al.*, 2014). There is partial LD (r^2=0.48) between rs2736100 and rs7705526.

The rs2736100-A allele is the risk allele for CRC and testicular germ cell tumour, while the same allele is protective for EC, glioma and lung cancer, suggesting that this variant has its effects in a tissue-specific manner.

The chromosome 8q24 locus has also been associated with many different tumours including cancer of the bladder (Kiemeney *et al.*, 2008; Rothman *et al.*, 2010), breast (Easton *et al.*, 2007; Michailidou *et al.*, 2013), ovary (Goode *et al.*, 2010) and prostate (Eeles *et al.*, 2009; Gudmundsson *et al.*, 2009). rs6983267 and rs4733613 at 8q24 have been independently associated with CRC and EC respectively, although neither of these SNPs are significant in the other cancer genotype (table 6.1). These SNPs lie 1.2mb apart and are not in LD; no other SNP in the 8q24 locus is highly significant in the EC-CRC meta-analysis. This sugests that the EC and CRC association signals are separate with no overlap despite being located in the same chromosomal band.

I have found evidence for a SNP (rs12970291, chromosome 18q22) that has opposing allelic effects on CRC and EC risk. The top candidate gene in this region is *TSHZ1* which encodes zinc finger homeodomain factor teashirt zinc finger family member 1, a protein involved in skin, skeletal, brain and gut development (Manfroid *et al.*, 2004) that is functionally related to the CRC gene *BMP* (Long *et al.*, 2001). One of several candidate SNPs near and within *TSHZ1* is the uncommon missense variant rs33930274 (p.Ala468Thr) in the last exon of *TSHZ1*, although the predicted functional consequences of this change are inconsistent (Grantham score=58, SIFT=0.0, Polyphen2=0.0, CADD score PHRED-scaled: 0.001).

Apart from the *SH2B3* and *TERT* SNPs, only two of 27 previously reported CRC SNPs, including one near *TERC*, showed any good evidence of association with EC and none of the known EC SNPs was associated with CRC risk ($P<0.05$). Power calculations suggested that this analysis was well powered (95-97% power, OR=1.1, MAF$\sim$0.3) to detect known GWAS SNPs in the other cancer phenotype. Genome-wide enrichment analysis showed no evidence of an excess of shared predisposition variants for both EC and CRC, suggesting that there may be relatively few common risk alleles contributing a modest effect to both cancer phenotypes. This is similar to the observation in breast and ovarian cancer, which also share high-penetrance risk

Mendelian causes but GWAS have identified relatively few shared common risk alleles. The discovery of *SH2B3* and *TSHZ1* associations hightlight the value of conducting multi-cancer GWAS for finding novel cancer associations as well as for shedding light on the haplotype structure of genomic loci such as the *TERT-CLPTM1L* region, which exerts effects on many different cancers.

A random effects meta-analysis could be conducted with the same data and it would be of interest to see if any additional significant loci are found. EC and CRC are different phenotypes and the random effects model would operate with the presumption that there are underlying differences in effects in the different studies.

Chapter 7

Discussion

7.1 The search for EC risk factors

EC is a common cancer with rising incidence rates over the past two decades in the UK. Known epidemiological risk factors are mainly related to prolonged exposure to the unopposed effects of oestrogen, particularly for type 1 ECs with endometrioid histology. Twin concordance studies suggest a limited role for inherited EC risk, but large case-control and cohort studies demonstrate a heritable component: those with a first degree relative with EC have a relative risk of $\sim$2. Clearly, environmental factors play an important role in EC predisposition. The search for the genetic aetiology of EC complements a wider search for EC risk factors, aiming to improve risk stratification and our understanding of the disease biology.

The study of familial cases of EC has greatly enhanced our understanding of the genetic aetiology of EC. Rare, high penetrance Mendelian syndromes caused by mutations in MMR genes, DNA polymerases or *PTEN* confer up to a 70% lifetime risk of developing EC. This has led to an improved diagnosis and clinical management for these patients and families, especially for LS. However, these Mendelian syndromes only explain a small part of the familial aggregation of EC.

Knowledge of these mutational pathways have been informative for EC tumours more generally. For example, MSI is found in up to 20% of sporadic EC tumours, with *MLH1* methylation being the cause in the majority of cases (Esteller *et al.*, 1998; Kanaya *et al.*, 2003). Somatic *POLE* EDMs are detected in microsatellite stable tumours associated with a 'hypermutator' phenotype (Church *et al.*, 2013). *PTEN* is mutated in 35% of EC tumours (COSMIC). The conditional loss of uterine *PTEN* in mice induces EC (Daikoku *et al.*, 2008) and provides a model for studying EC in mice. Molecular characteristics are also informative in terms of EC prognosis (Church *et al.*, 2013; Risinger *et al.*, 1998; Terada *et al.*, 2013; van Gool *et al.*, 2015).

7.2 Genetic aetiology of sporadic EC

The genetic cause of sporadic EC remains largely unknown. Candidate gene studies have tested hypotheses based on genes related to oestrogen and DNA repair, but most of these results have been inconclusive, with non-replicable results.

The aim of this thesis was to investigate the genetic aetiology of sporadic EC using genome-wide population-based methods. The NSECG GWAS (925 cases and 2,507 controls) replicated the association at *HNF1B* and provided further evidence of the significance of this locus in EC.

A meta-analysis with two other EC GWAS and two replication phases with a total of 7,737 cases and 37,144 controls yielded five novel risk loci of genome-wide significance ($P<10^{-5}$). In decreasing order of significance, these were at chromosomes 13q22.1 (rs11841589, near *KLF5*), 6q22.31 (rs13328298, in *LOC643623* and near *HEY2* and *NCOA7*), 8q24.21 (rs4733613, telomeric to *MYC*), 15q15.1 (rs937213, in *EIF2AK4*, near *BMF*) and 14q32.33 (rs2498796, in *AKT1*, near *SIVA1*). There is evidence for a second independent association signal at 8q24: rs17232730 (P_{meta}=4.46x10^{-7}, P_{cond}=1.29x10^{-5}), a SNP that is in moderate LD with the ovarian

cancer SNP rs10088218 (r^2=0.43).

All samples in the meta-analysis were of European ancestry and there is little evidence of allelic heterogeneity (I^2<0.36). All novel risk loci were either directly genotyped or imputed with an info score of more than 0.9 in all datasets. The novel SNPs had MAFs that varied from 0.13 to 0.42, and risk alleles of the novel loci conferred a modest increase in risk, with ORs ranging from 1.13 to 1.19.

As discussed in chapter 4, several of the genes within these regions have known or likely roles in tumourigenesis: *KLF5*, *HEY2*, *NCOA7*, *BMF*, *AKT1* and *SIVA1*.

HNF1B remains the locus most significantly associated with EC risk, and a recent fine-mapping paper further delineates the association signal at the first intron / extended promoter region (Painter *et al.*, 2014).

The association of *CYP19A1* with EC risk seen in candidate gene studies has been confirmed to be genome-wide significant in the EC GWAS meta-analysis. None of the other EC candidate SNPs previously investigated were significant in this analysis and *CYP19A1* provides a rare example of an association from a candidate study being replicated by GWAS. Several highly significant risk SNPs in the EC meta-analysis have also been associated with increased levels of circulating oestrogen in post-menopausal women. This suggests that the EC risk is mediated via the peripheral conversion of androgen to oestrogen by aromatase, causing increased oestrogen levels.

CYP19A1 variants can be considered alongside other epidemiological EC risk factors that influence endogenous levels of oestrogen. There may be other genetic variants that contribute toward EC risk via the unopposed effect of oestrogen, although candidate gene *ESR1* was not highly significant in this analysis. Of the novel risk loci, *NCOA7* modulates oestrogen receptor activity and the top SNP is more significant with endometrioid histology analysis. This may suggest that the effect of the 6q22 SNP may be related to oestrogenic effects.

Other novel EC risk loci implicate other cellular mechanisms such as cell cycle regulation and apoptosis. For example, it is interesting that *PTEN* (cause of Cowden syndrome, commonly mutated in EC) antagonizes the PI3K/AKT/MTOR pathway. The GWAS hit on 14q32 is in the intronic region of *AKT1*, which affects cell proliferation and is upregulated in endometrial tumours. It may be revealing to investigate other pathological processes in tumours that are not oestrogen-driven.

Functional analysis of the 13q22 locus illustrates how it is important to further characterize EC risk loci. The results of chapter 5 suggest that the genomic sequence around rs9600103 is transcriptionally active and that it may be a part of a transcriptional repressor affecting the expression of *KLF5* in an allele-specific manner. Chromosome conformation capture experiments could assess interaction of other genomic loci with the 13q22 signal. Preliminary analysis of TCGA RNA-seq data from 458 EC samples did not replicate the association between rs11841589 genotype and *KLF5* expression. The samples from GTEx are normal uterine tissue and further investigation into pathology reports is necessary to ascertain whether these are from the endometrium or other parts of the uterus. In future, genotype and expression data from larger biobanks could help establish the role of genetic variants in determining gene expression in the endometrium.

Besides trying to establish a stronger link between the intergenic 13q22 locus and *KLF5* expression, the role of *KLF5* in EC tumourigenesis can be investigated using animal models. At the time of writing, another member of our group is developing mice to investigate if heterozygous deletion of *KLF5* influences the *PTEN* deletion induced EC.

86% of EC cases displayed endometrioid histology only. Association testing with these samples produced a slightly increased significance compared to all histologies for rs13328298 (*HEY2/NCOA7*) and rs17232730 (8q24). It is unclear if the increased significance (despite a slightly reduced sample size) indicates an association with en-

dometrioid histology or if fluctuations in significance are due to chance. GWAS for endometrioid histology alone analysis did not identify any highly significant SNPs not present in the all histologies analysis. The low number of samples with other histologies limited the statistical power to detect associations though arguably, there is a more pressing need to investigate the causes of serous and clear-cell EC tumours. Non-endometrioid tumours confer a poorer prognosis, are not associated with oestrogen exposure, and have no known precursor lesion (such as endometrial hyperplasia) that can be detected and monitored. It is worth bearing in mind that LS-associated EC can display the full range of EC histologies and thus there may not be a strong correlation between histological subtypes and genetic aetiology.

7.3 Multi-cancer analysis

The strong predisposition to EC and CRC for women with MMR and DNA polymerase mutations prompted the search for other variants associated with both cancers. The missense *SH2B3* SNP rs3184504-C is associated with an increased risk to both EC and CRC.

The *TERT-CLPTM1L* locus has been associated with cancer of the breast, prostate, lung, pancreas, glioma, testicular germ cell, and basal cell carcinoma. There appears to be a complex haplotype structure in this region. The rs2736100-A allele is the risk allele for CRC and testicular germ cell tumour, while the same allele is protective for EC, glioma and lung cancer. rs12970291, near *TSHZ1* also has opposing allelic effects on CRC and EC risk. These SNPs may act in a tissue-specific manner where a particular allele confers risk in one tissue while being protective in another.

Apart from the loci listed above, there is minimal evidence of overlap between the SNPs that show promising P-values for EC and CRC across the genome.

Of the other risk loci, the *HNF1B* tag-SNP has been associated with serous and

clear-cell ovarian cancer and prostate cancer. The second EC risk locus at chromosome 8q24 is in moderate LD with ovarian cancer SNP rs10088218.

There have been more than 200 published cancer GWAS SNPs, and it would be a worthy pursuit to conduct a genome-wide meta-analysis that may identify variants that are associated with multiple cancers. As with the EC CRC meta-analysis, care should be taken so that there is no overlap in samples, since common control datasets have been used for different GWAS. Such an analysis may have increased statistical power to identify novel variants that are associated with multiple cancers. Variants that are over-represented in all cancer cases may point toward genes related to the hallmarks of cancer (Hanahan and Weinberg, 2011), and particularly metastasis, the process which defines malignancies. Some variants may be significant in subgroups of cancers; different subgroups could be tested based on histology, embryonic origin, or otherwise. However, there would be significant challenges in such an analysis because of the heterogeneity in cancer phenotypes and differences between studies. There may be significant misclassification bias based on the increased prevalence of undiagnosed cancers in controls, as well as issues around multiple testing and significance thresholds.

7.4 Conclusion and future developments

The results of this EC GWAS meta-analysis demonstrate that there are common variants that are highly associated with sporadic EC that explain a small proportion of the inherited risk of EC. These results are comparable with the trend seen in other cancer GWAS in that risk variants have a modest effect size (OR<1.2 for EC). There are planned efforts for further genotyping of EC samples, and a larger sample size may detect a larger number of novel EC associated variants. However, the total number of associated variants and the phenotypic variance explained may be limited by the

heritability of sporadic EC. Twin studies show that there is a low concordance in EC compared with other cancers such as CRC and prostate cancer, where an estimated 35-42% of sporadic cancer risk is inherited (Lichtenstein *et al.*, 2000). There are alternative methods for estimating heritability based on GWAS data (Yang *et al.*, 2010, 2011) and it would be of interest to establish the EC heritability that can be explained by common variants.

A recent EC GWAS (E2C2) which consisted of 7,077 cases and 16,343 controls replicated the association at *HNF1B* but did not find any other genome-wide significant associations (De Vivo *et al.*, 2014). The most notable difference between this study and the meta-analysis described in chapter 4 is that the E2C2 replication phase consisted of samples of African American, Asian, Latina, Hawaiian and European ancestry. Large population stratification may explain why the E2C2 study was unable to detect novel associations. For instance, I searched for the 13q22 and 6q22 EC risk loci in a Chinese EC GWAS and found that these variants were not replicated in the Chinese population. Epidemiological studies in the US have shown differences in EC incidence, histology and prognosis amongst different ethnic groups. Perhaps there are inter-ethnic differences in EC genetic aetiology and it may be interesting to investigate the differences in haplotype structure at EC risk loci between Europeans and other ancestries.

EC GWAS have demonstrated that most novel EC risk loci do not overlap with those of other cancers. This suggests that a small number of genes may affect multiple cancers, but on the whole, the genetic aetiology of different cancers is diverse. The identification of EC risk loci also allows comparison with other phenotypes. For example, the *SH2B3* SNP risk allele for EC and CRC is protective for a immune-mediated phenotypes such as coeliac disease, type 1 diabetes and rheumatoid arthritis. Although the effect size for each of these phenotypes is modest, this may be suggestive that selection is occurring at this location and there is a balance between the risk of

cancer and auto-immune disease.

In future, targeted re-sequencing and improved imputation reference panels should capture more of the genetic variation at EC risk loci and aid the discovery of causal variants. This process will be aided by functional annotation using various publicly available databases. Larger databases for data such as eQTL analysis may be helpful in linking associations to gene expression. *In vitro* methods such as those employed in chapter 5 can be conducted for the other novel risk loci, and implicated genes can be further studied using animal models.

As the cost of high-throughput sequencing becomes even more affordable, there is an improved capability to assess structural variants and rare variants. For rare variants, study designs could take advantage of selecting specific phenotypic subgroups or homogenous populations. Otherwise, the study of rare variants may struggle to demonstrate large effect sizes or genome-wide significance levels.

The identification of common variants associated with EC has achieved the aim of enhancing our understand of the genetic aetiology of EC, opening the door to further research into the novel risk loci. It seems unlikely that there are common SNPs with a large effect size in European populations not detected by this meta-analysis. However, there may be a number of variants that influence EC risk with a modest effect size that cannot be detected at genome-wide significance given the current SNP-array coverage and sample sizes. It seems unlikely that genetic variants associated with sporadic EC will be of individual predictive value for the general population although there may still be undiscovered rare, high-penetrance Mendelian causes of EC. Given the early presentation of most EC cases, genetic risk is unlikely to influence screening strategy unless it is a highly-penetrant Mendelian disease like LS.

Ultimately, establishing measures to reduce known risk factors such as oestrogen exposure and obesity may have the greatest impact on reducing EC incidence. The continuing role of EC genetics research is to highlight genes and pathways in-

volved with disease, and further research is warranted particularly for ECs with non-endometrioid histology, and for those who lack oestrogen-related risk factors.

References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Aarnio, M., Mecklin, J. P., Aaltonen, L. A., Nyström-Lahti, M., and Järvinen, H. J. (1995). Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome. *International Journal of Cancer. Journal International Du Cancer*, 64(6):430–433.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249.

Al Olama, A. A., Kote-Jarai, Z., Berndt, S. I., Conti, D. V., Schumacher, F., Han, Y., Benlloch, S., Hazelett, D. J., Wang, Z., Saunders, E., Leongamornlert, D., Lindstrom, S., Jugurnauth-Little, S., Dadaev, T., Tymrakiewicz, M., Stram, D. O., Rand, K., Wan, P., Stram, A., Sheng, X., Pooler, L. C., Park, K., Xia, L., Tyrer, J., Kolonel, L. N., Le Marchand, L., Hoover, R. N., Machiela, M. J., Yeager, M., Burdette, L., Chung, C. C., Hutchinson, A., Yu, K., Goh, C., Ahmed, M., Govindasami, K., Guy, M., Tammela, T. L. J., Auvinen, A., Wahlfors, T., Schleutker, J., Visakorpi, T., Leinonen, K. A., Xu, J., Aly, M., Donovan, J., Travis, R. C., Key, T. J., Siddiq, A., Canzian, F., Khaw, K.-T., Takahashi, A., Kubo, M., Pharoah, P., Pashayan, N., Weischer, M., Nordestgaard, B. G., Nielsen, S. F., Klarskov, P., Røder, M. A., Iversen, P., Thibodeau, S. N., McDonnell, S. K., Schaid, D. J., Stanford, J. L., Kolb, S., Holt, S., Knudsen, B., Coll, A. H.,

Gapstur, S. M., Diver, W. R., Stevens, V. L., Maier, C., Luedeke, M., Herkommer, K., Rinckleb, A. E., Strom, S. S., Pettaway, C., Yeboah, E. D., Tettey, Y., Biritwum, R. B., Adjei, A. A., Tay, E., Truelove, A., Niwa, S., Chokkalingam, A. P., Cannon-Albright, L., Cybulski, C., Wokołorczyk, D., Kluźniak, W., Park, J., Sellers, T., Lin, H.-Y., Isaacs, W. B., Partin, A. W., Brenner, H., Dieffenbach, A. K., Stegmaier, C., Chen, C., Giovannucci, E. L., Ma, J., Stampfer, M., Penney, K. L., Mucci, L., John, E. M., Ingles, S. A., Kittles, R. A., Murphy, A. B., Pandha, H., Michael, A., Kierzek, A. M., Blot, W., Signorello, L. B., Zheng, W., Albanes, D., Virtamo, J., Weinstein, S., Nemesure, B., Carpten, J., Leske, C., Wu, S.-Y., Hennis, A., Kibel, A. S., Rybicki, B. A., Neslund-Dudas, C., Hsing, A. W., Chu, L., Goodman, P. J., Klein, E. A., Zheng, S. L., Batra, J., Clements, J., Spurdle, A., Teixeira, M. R., Paulo, P., Maia, S., Slavov, C., Kaneva, R., Mitev, V., Witte, J. S., Casey, G., Gillanders, E. M., Seminara, D., Riboli, E., Hamdy, F. C., Coetzee, G. A., Li, Q., Freedman, M. L., Hunter, D. J., Muir, K., Gronberg, H., Neal, D. E., Southey, M., Giles, G. G., Severi, G., Breast and Prostate Cancer Cohort Consortium (BPC3), PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, COGS (Collaborative Oncological Gene-environment Study) Consortium, GAME-ON/ELLIPSE Consortium, Cook, M. B., Nakagawa, H., Wiklund, F., Kraft, P., Chanock, S. J., Henderson, B. E., Easton, D. F., Eeles, R. A., and Haiman, C. A. (2014). A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nature Genetics*, 46(10):1103–1109.

Allard, J. E. and Maxwell, G. L. (2009). Race disparities between black and white women in the incidence, treatment, and prognosis of endometrial cancer. *Cancer Control: Journal of the Moffitt Cancer Center*, 16(1):53–56.

Amant, F., Moerman, P., Neven, P., Timmerman, D., Van Limbergen, E., and Vergote, I. (2005). Endometrial cancer. *Lancet*, 366(9484):491–505.

Avery, O. T., Macleod, C. M., and McCarty, M. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEU-

MOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *The Journal of Experimental Medicine*, 79(2):137–158.

Baird, D. D. and Dunson, D. B. (2003). Why is parity protective for uterine fibroids? *Epidemiology (Cambridge, Mass.)*, 14(2):247–250.

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D. J., Stevens, H., Todd, J. A., Walker, N. M., Rich, S. S., and Type 1 Diabetes Genetics Consortium (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, 41(6):703–707.

Barrow, E., Hill, J., and Evans, D. G. (2013). Cancer risk in Lynch Syndrome. *Familial Cancer*, 12(2):229–240.

Barry, J. A., Azizia, M. M., and Hardiman, P. J. (2014). Risk of endometrial, ovarian and breast cancer in women with polycystic ovary syndrome: a systematic review and meta-analysis. *Human Reproduction Update*, 20(5):748–758.

Bauer, C. M., Ray, A. M., Halstead-Nussloch, B. A., Dekker, R. G., Raymond, V. M., Gruber, S. B., and Cooney, K. A. (2011). Hereditary prostate cancer as a feature of Lynch syndrome. *Familial Cancer*, 10(1):37–42.

Beenken, A. and Mohammadi, M. (2009). The FGF family: biology, pathophysiology and therapy. *Nature Reviews. Drug Discovery*, 8(3):235–253.

Ben Arie, A., Lavie, O., Gdalevich, M., Voldarsky, M., Barak, F., Schneider, D., Levy, T., Anteby, E., and Gemer, O. (2012). Temporal pattern of recurrence of stage I endometrial cancer in relation to histological risk factors. *European Journal of Surgical Oncology: The Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology*, 38(2):166–169.

Benedet, J. L., Bender, H., Jones, H., Ngan, H. Y., and Pecorelli, S. (2000). FIGO staging classifications and clinical practice guidelines in the management of gynecologic cancers. FIGO Committee on Gynecologic Oncology. *International Journal of Gynaecology and Obstetrics: The Official Organ of the International Federation of Gynaecology and Obstetrics*, 70(2):209–262.

Berlanga, J. J., Santoyo, J., and De Haro, C. (1999). Characterization of a mammalian homolog of the GCN2 eukaryotic initiation factor 2alpha kinase. *European journal of biochemistry / FEBS*, 265(2):754–762.

Bernardini, L., Gimelli, S., Gervasini, C., Carella, M., Baban, A., Frontino, G., Barbano, G., Divizia, M. T., Fedele, L., Novelli, A., Béna, F., Lalatta, F., Miozzo, M., and Dallapiccola, B. (2009). Recurrent microdeletion at 17q12 as a cause of Mayer-Rokitansky-Kuster-Hauser (MRKH) syndrome: two case reports. *Orphanet Journal of Rare Diseases*, 4:25.

Berndt, S. I., Sampson, J., Yeager, M., Jacobs, K. B., Wang, Z., Hutchinson, A., Chung, C., Orr, N., Wacholder, S., Chatterjee, N., Yu, K., Kraft, P., Feigelson, H. S., Thun, M. J., Diver, W. R., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G. L., Crawford, E. D., Haiman, C., Henderson, B., Kolonel, L., Le Marchand, L., Siddiq, A., Riboli, E., Travis, R. C., Kaaks, R., Isaacs, W., Isaacs, S., Wiley, K. E., Gronberg, H., Wiklund, F., Stattin, P., Xu, J., Zheng, S. L., Sun, J., Vatten, L. J., Hveem, K., Njølstad, I., Gerhard, D. S., Tucker, M., Hayes, R. B., Hoover, R. N., Fraumeni, J. F., Hunter, D. J., Thomas, G., and Chanock, S. J. (2011). Large-scale fine mapping of the HNF1b locus and prostate cancer risk. *Human Molecular Genetics*, 20(16):3322–3329.

Blumenthal, G. M. and Dennis, P. A. (2008). Germline PTEN mutations as a cause of early-onset endometrial cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 26(13):2234; author reply 2234.

Bokhman, J. V. (1983). Two pathogenetic types of endometrial carcinoma. *Gynecologic Oncology*, 15(1):10–17.

Boland, C. R. (2005). Evolution of the nomenclature for the hereditary colorectal cancer syndromes. *Familial Cancer*, 4(3):211–218.

Boland, C. R. and Troncale, F. J. (1984). Familial colonic cancer without antecedent polyposis. *Annals of Internal Medicine*, 100(5):700–701.

Bonadona, V., Bonaïti, B., Olschwang, S., Grandjouan, S., Huiart, L., Longy, M., Guimbaud, R., Buecher, B., Bignon, Y.-J., Caron, O., Colas, C., Noguès, C., Lejeune-Dumoulin, S., Olivier-Faivre, L., Polycarpe-Osaer, F., Nguyen, T. D., Desseigne, F., Saurin, J.-C., Berthet, P., Leroux, D., Duffour, J., Manouvrier, S., Frébourg, T., Sobol, H., Lasset, C., Bonaïti-Pellié, C., and French Cancer Genetics Network (2011). Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. *JAMA*, 305(22):2304–2310.

Borel, C., Deutsch, S., Letourneau, A., Migliavacca, E., Montgomery, S. B., Dimas, A. S., Vejnar, C. E., Attar, H., Gagnebin, M., Gehrig, C., Falconnet, E., Dupré, Y., Dermitzakis, E. T., and Antonarakis, S. E. (2011). Identification of cis- and trans-regulatory variation modulating microRNA expression levels in human fibroblasts. *Genome Research*, 21(1):68–73. Available from: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3012927/`.

Boruta, D. M., Gehrig, P. A., Groben, P. A., Bae-Jump, V., Boggess, J. F., Fowler, W. C., and Van Le, L. (2004). Uterine serous and grade 3 endometrioid carcinomas: is there a survival difference? *Cancer*, 101(10):2214–2221.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3):314–331.

Boyle, A. P., Guinney, J., Crawford, G. E., and Furey, T. S. (2008). F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics (Oxford, England)*, 24(21):2537–2538.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M., and Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, 22(9):1790–1797.

Bravi, F., Scotti, L., Bosetti, C., Gallus, S., Negri, E., La Vecchia, C., and Tavani, A. (2009). Coffee drinking and endometrial cancer risk: a metaanalysis of observational studies. *American Journal of Obstetrics and Gynecology*, 200(2):130–135.

Bray, F., Dos Santos Silva, I., Moller, H., and Weiderpass, E. (2005). Endometrial cancer incidence trends in Europe: underlying determinants and prospects for prevention. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 14(5):1132–1142.

Brinton, L. A. and Felix, A. S. (2014). Menopausal hormone therapy and risk of endometrial cancer. *The Journal of Steroid Biochemistry and Molecular Biology*, 142:83–89.

Brinton, L. A., Westhoff, C. L., Scoccia, B., Lamb, E. J., Althuis, M. D., Mabie, J. E., and Moghissi, K. S. (2005). Causes of infertility as predictors of subsequent cancer risk. *Epidemiology (Cambridge, Mass.)*, 16(4):500–507.

Broaddus, R. R., Lynch, H. T., Chen, L.-M., Daniels, M. S., Conrad, P., Munsell, M. F., White, K. G., Luthra, R., and Lu, K. H. (2006). Pathologic features of endometrial carcinoma associated with HNPCC: a comparison with sporadic endometrial carcinoma. *Cancer*, 106(1):87–94.

Bush, W. S. and Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822.

Cai, H., Xiang, Y.-B., Qu, S., Long, J., Cai, Q., Gao, J., Zheng, W., and Shu, X. O. (2011). Association of genetic polymorphisms in cell-cycle control genes and suscepti-

bility to endometrial cancer among Chinese women. *American Journal of Epidemiology*, 173(11):1263–1271.

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

Cancer Genome Atlas Research Network, Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., Robertson, A. G., Pashtan, I., Shen, R., Benz, C. C., Yau, C., Laird, P. W., Ding, L., Zhang, W., Mills, G. B., Kucherlapati, R., Mardis, E. R., and Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73.

Cantley, L. C. (2002). The phosphoinositide 3-kinase pathway. *Science (New York, N.Y.)*, 296(5573):1655–1657.

Carcangiu, M. L., Radice, P., Casalini, P., Bertario, L., Merola, M., and Sala, P. (2010). Lynch syndrome–related endometrial carcinomas show a high frequency of nonendometrioid types and of high FIGO grade endometrioid types. *International Journal of Surgical Pathology*, 18(1):21–26.

Carvajal-Carmona, L. G., O'Mara, T. A., Painter, J. N., Lose, F. A., Dennis, J., Michailidou, K., Tyrer, J. P., Ahmed, S., Ferguson, K., Healey, C. S., Pooley, K., Beesley, J., Cheng, T., Jones, A., Howarth, K., Martin, L., Gorman, M., Hodgson, S., National Study of Endometrial Cancer Genetics Group (NSECG), The Australian National Endometrial Cancer Study Group (ANECS), Wentzensen, N., Fasching, P. A., Hein, A., Beckmann, M. W., Renner, S. P., Dörk, T., Hillemanns, P., Dürst, M., Runnebaum, I., Lambrechts, D., Coenegrachts, L., Schrauwen, S., Amant, F., Winterhoff, B., Dowdy, S. C., Goode, E. L., Teoman, A., Salvesen, H. B., Trovik, J., Njolstad, T. S., Werner, H. M. J., Scott, R. J., Ashton, K., Proietto, T., Otton, G., Wersäll, O., Mints, M., Tham, E., RENDOCAS, Hall, P., Czene, K., Liu, J., Li, J., Hopper, J. L., Southey, M. C., Australian Ovarian Cancer Study (AOCS), Ekici, A. B., Ruebner, M., Johnson, N., Peto, J., Burwinkel, B., Marme, F., Brenner, H., Dieffenbach, A. K., Meindl, A., Brauch, H., The

GENICA Network, Lindblom, A., Depreeuw, J., Moisse, M., Chang-Claude, J., Rudolph, A., Couch, F. J., Olson, J. E., Giles, G. G., Bruinsma, F., Cunningham, J. M., Fridley, B. L., Børresen-Dale, A.-L., Kristensen, V. N., Cox, A., Swerdlow, A. J., Orr, N., Bolla, M. K., Wang, Q., Weber, R. P., Chen, Z., Shah, M., Pharoah, P. D. P., Dunning, A. M., Tomlinson, I., Easton, D. F., Spurdle, A. B., and Thompson, D. J. (2014). Candidate locus analysis of the TERT-CLPTM1l cancer risk region on chromosome 5p15 identifies multiple independent variants associated with endometrial cancer risk. *Human Genetics*.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., and Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404.

Chen, C., Bhalala, H. V., Vessella, R. L., and Dong, J.-T. (2003). KLF5 is frequently deleted and down-regulated but rarely mutated in prostate cancer. *The Prostate*, 55(2):81–88.

Chen, X., Xiang, Y.-B., Long, J.-R., Cai, H., Cai, Q., Cheng, J., Wen, W., Gao, Y.-T., Zheng, W., and Shu, X.-O. (2012). Genetic polymorphisms in obesity-related genes and endometrial cancer risk. *Cancer*, 118(13):3356–3364.

Chiang, S. and Soslow, R. A. (2014). Updates in diagnostic immunohistochemistry in endometrial carcinoma. *Seminars in Diagnostic Pathology*, 31(3):205–215.

Chubb, D., Weinhold, N., Broderick, P., Chen, B., Johnson, D. C., Försti, A., Vijayakrishnan, J., Migliorini, G., Dobbins, S. E., Holroyd, A., Hose, D., Walker, B. A., Davies, F. E., Gregory, W. A., Jackson, G. H., Irving, J. A., Pratt, G., Fegan, C., Fenton, J. A. L., Neben, K., Hoffmann, P., Nöthen, M. M., Mühleisen, T. W., Eisele, L., Ross, F. M., Straka, C., Einsele, H., Langer, C., Dörner, E., Allan, J. M., Jauch, A., Morgan, G. J., Hemminki, K., Houlston, R. S., and Goldschmidt, H. (2013). Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nature Genetics*, 45(10):1221–1225.

Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). GPA: A Statistical Approach to Prioritizing GWAS Results by Integrating Pleiotropy and Annotation. *PLoS Genetics*, 10(11). Available from: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4230845/`.

Church, D. N., Briggs, S. E. W., Palles, C., Domingo, E., Kearsey, S. J., Grimes, J. M., Gorman, M., Martin, L., Howarth, K. M., Hodgson, S. V., NSECG Collaborators, Kaur, K., Taylor, J., and Tomlinson, I. P. M. (2013). DNA polymerase e and d exonuclease domain mutations in endometrial cancer. *Human Molecular Genetics*, 22(14):2820–2828.

Clark, T. J., Voit, D., Gupta, J. K., Hyde, C., Song, F., and Khan, K. S. (2002). Accuracy of hysteroscopy in the diagnosis of endometrial cancer and hyperplasia: a systematic quantitative review. *JAMA*, 288(13):1610–1621.

Codd, V., Nelson, C. P., Albrecht, E., Mangino, M., Deelen, J., Buxton, J. L., Hottenga, J. J., Fischer, K., Esko, T., Surakka, I., Broer, L., Nyholt, D. R., Mateo Leach, I., Salo, P., Hägg, S., Matthews, M. K., Palmen, J., Norata, G. D., O'Reilly, P. F., Saleheen, D., Amin, N., Balmforth, A. J., Beekman, M., de Boer, R. A., Böhringer, S., Braund, P. S., Burton, P. R., de Craen, A. J. M., Denniff, M., Dong, Y., Douroudis, K., Dubinina, E., Eriksson, J. G., Garlaschelli, K., Guo, D., Hartikainen, A.-L., Henders, A. K., Houwing-Duistermaat, J. J., Kananen, L., Karssen, L. C., Kettunen, J., Klopp, N., Lagou, V., van Leeuwen, E. M., Madden, P. A., Mägi, R., Magnusson, P. K. E., Männistö, S., McCarthy, M. I., Medland, S. E., Mihailov, E., Montgomery, G. W., Oostra, B. A., Palotie, A., Peters, A., Pollard, H., Pouta, A., Prokopenko, I., Ripatti, S., Salomaa, V., Suchiman, H. E. D., Valdes, A. M., Verweij, N., Viñuela, A., Wang, X., Wichmann, H.-E., Widen, E., Willemsen, G., Wright, M. J., Xia, K., Xiao, X., van Veldhuisen, D. J., Catapano, A. L., Tobin, M. D., Hall, A. S., Blakemore, A. I. F., van Gilst, W. H., Zhu, H., Consortium, C., Erdmann, J., Reilly, M. P., Kathiresan, S., Schunkert, H., Talmud, P. J., Pedersen, N. L., Perola, M., Ouwehand, W., Kaprio, J., Martin, N. G., van Duijn, C. M., Hovatta, I., Gieger, C., Metspalu, A., Boomsma, D. I., Jarvelin, M.-R., Slagboom, P. E., Thompson, J. R., Spector, T. D., van der Harst, P., and Samani, N. J. (2013).

Identification of seven loci affecting mean telomere length and their association with disease. *Nature Genetics*, 45(4):422–427, 427e1–2.

COGENT Study, Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S., Colorectal Cancer Association Study Consortium, Carvajal-Carmona, L., Howarth, K., Jaeger, E., Spain, S. L., Walther, A., Barclay, E., Martin, L., Gorman, M., Domingo, E., Teixeira, A. S., CoRGI Consortium, Kerr, D., Cazier, J.-B., Niittymäki, I., Tuupanen, S., Karhu, A., Aaltonen, L. A., Tomlinson, I. P. M., Farrington, S. M., Tenesa, A., Prendergast, J. G. D., Barnetson, R. A., Cetnarskyj, R., Porteous, M. E., Pharoah, P. D. P., Koessler, T., Hampe, J., Buch, S., Schafmayer, C., Tepel, J., Schreiber, S., Völzke, H., Chang-Claude, J., Hoffmeister, M., Brenner, H., Zanke, B. W., Montpetit, A., Hudson, T. J., Gallinger, S., Campbell, H., and Dunlop, M. G. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genetics*, 40(12):1426–1435.

Cogliano, V. J., Baan, R., Straif, K., Grosse, Y., Lauby-Secretan, B., El Ghissassi, F., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Freeman, C., Galichet, L., and Wild, C. P. (2011). Preventable exposures associated with human cancers. *Journal of the National Cancer Institute*, 103(24):1827–1839.

Cohen, Y., Shalmon, B., Korach, J., Barshack, I., Fridman, E., and Rechavi, G. (2010). AKT1 pleckstrin homology domain E17k activating mutation in endometrial carcinoma. *Gynecologic Oncology*, 116(1):88–91.

Colombo, N., Preti, E., Landoni, F., Carinelli, S., Colombo, A., Marini, C., Sessa, C., and ESMO Guidelines Working Group (2013). Endometrial cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of oncology: official journal of the European Society for Medical Oncology / ESMO*, 24 Suppl 6:vi33–38.

Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome.

Nature, 431(7011):931–945. Available from: `http://www.nature.com/nature/journal/v431/n7011/full/nature03001.html`.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, 122(6):565–581.

Daikoku, T., Hirota, Y., Tranguch, S., Joshi, A. R., DeMayo, F. J., Lydon, J. P., Ellenson, L. H., and Dey, S. K. (2008). Conditional loss of uterine Pten unfailingly and rapidly induces endometrial cancer in mice. *Cancer Research*, 68(14):5619–5627.

Davis, H., Lewis, A., Spencer-Dene, B., Tateossian, H., Stamp, G., Behrens, A., and Tomlinson, I. (2011). FBXW7 mutations typically found in human cancers are distinct from null alleles and disrupt lung development. *The Journal of Pathology*, 224(2):180–189.

de Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics*, 37(11):1217–1223. Available from: `http://www.nature.com/ng/journal/v37/n11/abs/ng1669.html`.

de Boer, Y. S., van Gerven, N. M. F., Zwiers, A., Verwer, B. J., van Hoek, B., van Erpecum, K. J., Beuers, U., van Buuren, H. R., Drenth, J. P. H., den Ouden, J. W., Verdonk, R. C., Koek, G. H., Brouwer, J. T., Guichelaar, M. M. J., Vrolijk, J. M., Kraal, G., Mulder, C. J. J., van Nieuwkerk, C. M. J., Fischer, J., Berg, T., Stickel, F., Sarrazin, C., Schramm, C., Lohse, A. W., Weiler-Normann, C., Lerch, M. M., Nauck, M., Völzke, H., Homuth, G., Bloemena, E., Verspaget, H. W., Kumar, V., Zhernakova, A., Wijmenga, C., Franke, L., Bouma, G., Dutch Autoimmune Hepatitis Study Group, LifeLines Cohort Study, and Study of Health in Pomerania (2014). Genome-wide association study identifies variants associated with autoimmune hepatitis type 1. *Gastroenterology*, 147(2):443–452.e5.

De Vivo, I., Prescott, J., Setiawan, V. W., Olson, S. H., Wentzensen, N., Australian National Endometrial Cancer Study Group, Attia, J., Black, A., Brinton, L., Chen, C., Chen, C.,

Cook, L. S., Crous-Bou, M., Doherty, J., Dunning, A. M., Easton, D. F., Friedenreich, C. M., Garcia-Closas, M., Gaudet, M. M., Haiman, C., Hankinson, S. E., Hartge, P., Henderson, B. E., Holliday, E., Horn-Ross, P. L., Hunter, D. J., Le Marchand, L., Liang, X., Lissowska, J., Long, J., Lu, L., Magliocco, A. M., McEvoy, M., O'Mara, T. A., Orlow, I., Painter, J. N., Pooler, L., Rastogi, R., Rebbeck, T. R., Risch, H., Sacerdote, C., Schumacher, F., Scott, R. J., Sheng, X., Shu, X.-o., Spurdle, A. B., Thompson, D., Vanden Berg, D., Weiss, N. S., Xia, L., Xiang, Y.-B., Yang, H. P., Yu, H., Zheng, W., Chanock, S., and Kraft, P. (2014). Genome-wide association study of endometrial cancer in E2c2. *Human Genetics*, 133(2):211–224.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188.

Diakiw, S. M., Kok, C. H., To, L. B., Lewis, I. D., Brown, A. L., and D'Andrea, R. J. (2012). The granulocyte-associated transcription factor Krüppel-like factor 5 is silenced by hypermethylation in acute myeloid leukemia. *Leukemia Research*, 36(1):110–116.

Dichgans, M., Malik, R., König, I. R., Rosand, J., Clarke, R., Gretarsdottir, S., Thorleifsson, G., Mitchell, B. D., Assimes, T. L., Levi, C., O'Donnell, C. J., Fornage, M., Thorsteinsdottir, U., Psaty, B. M., Hengstenberg, C., Seshadri, S., Erdmann, J., Bis, J. C., Peters, A., Boncoraglio, G. B., März, W., Meschia, J. F., Kathiresan, S., Ikram, M. A., McPherson, R., Stefansson, K., Sudlow, C., Reilly, M. P., Thompson, J. R., Sharma, P., Hopewell, J. C., Chambers, J. C., Watkins, H., Rothwell, P. M., Roberts, R., Markus, H. S., Samani, N. J., Farrall, M., Schunkert, H., METASTROKE Consortium, CARDIoGRAM Consortium, C4D Consortium, and International Stroke Genetics Consortium (2014). Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke; a Journal of Cerebral Circulation*, 45(1):24–36.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., Gagnebin, M., Nisbett, J.,

Deloukas, P., Dermitzakis, E. T., and Antonarakis, S. E. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science (New York, N.Y.)*, 325(5945):1246–1250.

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R., and Cookson, W. O. C. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, 39(10):1202–1207.

Doherty, J. A., Weiss, N. S., Fish, S., Fan, W., Loomis, M. M., Sakoda, L. C., Rossing, M. A., Zhao, L. P., and Chen, C. (2011). Polymorphisms in nucleotide excision repair genes and endometrial cancer risk. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 20(9):1873–1882.

Dossus, L., Allen, N., Kaaks, R., Bakken, K., Lund, E., Tjonneland, A., Olsen, A., Overvad, K., Clavel-Chapelon, F., Fournier, A., Chabbert-Buffet, N., Boeing, H., Schütze, M., Trichopoulou, A., Trichopoulos, D., Lagiou, P., Palli, D., Krogh, V., Tumino, R., Vineis, P., Mattiello, A., Bueno-de Mesquita, H. B., Onland-Moret, N. C., Peeters, P. H. M., Dumeaux, V., Redondo, M.-L., Duell, E., Sanchez-Cantalejo, E., Arriola, L., Chirlaque, M.-D., Ardanaz, E., Manjer, J., Borgquist, S., Lukanova, A., Lundin, E., Khaw, K.-T., Wareham, N., Key, T., Chajes, V., Rinaldi, S., Slimani, N., Mouw, T., Gallo, V., and Riboli, E. (2010). Reproductive risk factors and endometrial cancer: the European Prospective Investigation into Cancer and Nutrition. *International Journal of Cancer. Journal International Du Cancer*, 127(2):442–451.

Du, W., Jiang, P., Li, N., Mei, Y., Wang, X., Wen, L., Yang, X., and Wu, M. (2009). Suppression of p53 activity by Siva1. *Cell Death and Differentiation*, 16(11):1493–1504.

Dunlop, M. G., Farrington, S. M., Carothers, A. D., Wyllie, A. H., Sharp, L., Burn, J., Liu, B., Kinzler, K. W., and Vogelstein, B. (1997). Cancer risk associated with germline DNA mismatch repair gene mutations. *Human Molecular Genetics*, 6(1):105–110.

Dunn, O. J. (1959). Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, 30(1):192–197. Available from: `http://projecteuclid.org/euclid.aoms/1177706374`.

Dunning, A. M., Healey, C. S., Baynes, C., Maia, A.-T., Scollen, S., Vega, A., Rodríguez, R., Barbosa-Morais, N. L., Ponder, B. A. J., SEARCH, Low, Y.-L., Bingham, S., EPIC, Haiman, C. A., Le Marchand, L., MEC, Broeks, A., Schmidt, M. K., ABCS, Hopper, J., Southey, M., ABCFS, Beckmann, M. W., Fasching, P. A., BBCC, Peto, J., Johnson, N., BBCS, Bojesen, S. E., Nordestgaard, B., CGPS, Milne, R. L., Benitez, J., CNIO-BCS, Hamann, U., Ko, Y., GENICA, Schmutzler, R. K., Burwinkel, B., GC-HBOC, Schürmann, P., Dörk, T., HABCS, Heikkinen, T., Nevanlinna, H., HEBCS, Lindblom, A., Margolin, S., KARBAC, Mannermaa, A., Kosma, V.-M., KBCS, Chen, X., Spurdle, A., kConFab and the AOCS Management Group, Change-Claude, J., Flesch-Janys, D., MARIE, Couch, F. J., Olson, J. E., for MCBCS, Severi, G., Baglietto, L., MCCS, Børresen-Dale, A.-L., Kristensen, V., NBCS, Hunter, D. J., Hankinson, S. E., NHS, Devilee, P., Vreeswijk, M., ORIGO, Lissowska, J., Brinton, L., PBCS, Liu, J., Hall, P., SASBAC, Kang, D., Yoo, K.-Y., SEBCS, Shen, C.-Y., Yu, J.-C., TWBCS, Anton-Culver, H., Ziogoas, A., UCIBCS, Sigurdson, A., Struewing, J., USRTS, Easton, D. F., Garcia-Closas, M., Humphreys, M. K., Morrison, J., Pharoah, P. D. P., Pooley, K. A., Chenevix-Trench, G., and BCAC (2009). Association of ESR1 gene tagging SNPs with breast cancer risk. *Human Molecular Genetics*, 18(6):1131–1139.

Early Breast Cancer Trialists' Collaborative Group, Davies, C., Godwin, J., Gray, R., Clarke, M., Cutter, D., Darby, S., McGale, P., Pan, H. C., Taylor, C., Wang, Y. C., Dowsett, M., Ingle, J., and Peto, R. (2011). Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet*, 378(9793):771–784.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., Wareham, N., Ahmed, S., Healey, C. S., Bowman, R., SEARCH collaborators, Meyer, K. B., Haiman, C. A.,

Kolonel, L. K., Henderson, B. E., Le Marchand, L., Brennan, P., Sangrajrang, S., Gaborieau, V., Odefrey, F., Shen, C.-Y., Wu, P.-E., Wang, H.-C., Eccles, D., Evans, D. G., Peto, J., Fletcher, O., Johnson, N., Seal, S., Stratton, M. R., Rahman, N., Chenevix-Trench, G., Bojesen, S. E., Nordestgaard, B. G., Axelsson, C. K., Garcia-Closas, M., Brinton, L., Chanock, S., Lissowska, J., Peplonska, B., Nevanlinna, H., Fagerholm, R., Eerola, H., Kang, D., Yoo, K.-Y., Noh, D.-Y., Ahn, S.-H., Hunter, D. J., Hankinson, S. E., Cox, D. G., Hall, P., Wedren, S., Liu, J., Low, Y.-L., Bogdanova, N., Schürmann, P., Dörk, T., Tollenaar, R. A. E. M., Jacobi, C. E., Devilee, P., Klijn, J. G. M., Sigurdson, A. J., Doody, M. M., Alexander, B. H., Zhang, J., Cox, A., Brock, I. W., MacPherson, G., Reed, M. W. R., Couch, F. J., Goode, E. L., Olson, J. E., Meijers-Heijboer, H., van den Ouweland, A., Uitterlinden, A., Rivadeneira, F., Milne, R. L., Ribas, G., Gonzalez-Neira, A., Benitez, J., Hopper, J. L., McCredie, M., Southey, M., Giles, G. G., Schroen, C., Justenhoven, C., Brauch, H., Hamann, U., Ko, Y.-D., Spurdle, A. B., Beesley, J., Chen, X., kConFab, AOCS Management Group, Mannermaa, A., Kosma, V.-M., Kataja, V., Hartikainen, J., Day, N. E., Cox, D. R., and Ponder, B. A. J. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093.

Eeles, R. A., Kote-Jarai, Z., Al Olama, A. A., Giles, G. G., Guy, M., Severi, G., Muir, K., Hopper, J. L., Henderson, B. E., Haiman, C. A., Schleutker, J., Hamdy, F. C., Neal, D. E., Donovan, J. L., Stanford, J. L., Ostrander, E. A., Ingles, S. A., John, E. M., Thibodeau, S. N., Schaid, D., Park, J. Y., Spurdle, A., Clements, J., Dickinson, J. L., Maier, C., Vogel, W., Dörk, T., Rebbeck, T. R., Cooney, K. A., Cannon-Albright, L., Chappuis, P. O., Hutter, P., Zeegers, M., Kaneva, R., Zhang, H.-W., Lu, Y.-J., Foulkes, W. D., English, D. R., Leongamornlert, D. A., Tymrakiewicz, M., Morrison, J., Ardern-Jones, A. T., Hall, A. L., O'Brien, L. T., Wilkinson, R. A., Saunders, E. J., Page, E. C., Sawyer, E. J., Edwards, S. M., Dearnaley, D. P., Horwich, A., Huddart, R. A., Khoo, V. S., Parker, C. C., Van As, N., Woodhouse, C. J., Thompson, A., Christmas, T., Ogden, C., Cooper, C. S., Southey, M. C., Lophatananon, A., Liu, J.-F., Kolonel, L. N., Le Marchand, L., Wahlfors, T., Tammela, T. L., Auvinen, A., Lewis, S. J., Cox, A.,

FitzGerald, L. M., Koopmeiners, J. S., Karyadi, D. M., Kwon, E. M., Stern, M. C., Corral, R., Joshi, A. D., Shahabi, A., McDonnell, S. K., Sellers, T. A., Pow-Sang, J., Chambers, S., Aitken, J., Gardiner, R. A. F., Batra, J., Kedda, M. A., Lose, F., Polanowski, A., Patterson, B., Serth, J., Meyer, A., Luedeke, M., Stefflova, K., Ray, A. M., Lange, E. M., Farnham, J., Khan, H., Slavov, C., Mitkova, A., Cao, G., UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology, UK ProtecT Study Collaborators, PRACTICAL Consortium, and Easton, D. F. (2009). Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nature Genetics*, 41(10):1116–1121.

Eeles, R. A., Olama, A. A. A., Benlloch, S., Saunders, E. J., Leongamornlert, D. A., Tymrakiewicz, M., Ghoussaini, M., Luccarini, C., Dennis, J., Jugurnauth-Little, S., Dadaev, T., Neal, D. E., Hamdy, F. C., Donovan, J. L., Muir, K., Giles, G. G., Severi, G., Wiklund, F., Gronberg, H., Haiman, C. A., Schumacher, F., Henderson, B. E., Le Marchand, L., Lindstrom, S., Kraft, P., Hunter, D. J., Gapstur, S., Chanock, S. J., Berndt, S. I., Albanes, D., Andriole, G., Schleutker, J., Weischer, M., Canzian, F., Riboli, E., Key, T. J., Travis, R. C., Campa, D., Ingles, S. A., John, E. M., Hayes, R. B., Pharoah, P. D. P., Pashayan, N., Khaw, K.-T., Stanford, J. L., Ostrander, E. A., Signorello, L. B., Thibodeau, S. N., Schaid, D., Maier, C., Vogel, W., Kibel, A. S., Cybulski, C., Lubinski, J., Cannon-Albright, L., Brenner, H., Park, J. Y., Kaneva, R., Batra, J., Spurdle, A. B., Clements, J. A., Teixeira, M. R., Dicks, E., Lee, A., Dunning, A. M., Baynes, C., Conroy, D., Maranian, M. J., Ahmed, S., Govindasami, K., Guy, M., Wilkinson, R. A., Sawyer, E. J., Morgan, A., Dearnaley, D. P., Horwich, A., Huddart, R. A., Khoo, V. S., Parker, C. C., Van As, N. J., Woodhouse, C. J., Thompson, A., Dudderidge, T., Ogden, C., Cooper, C. S., Lophatananon, A., Cox, A., Southey, M. C., Hopper, J. L., English, D. R., Aly, M., Adolfsson, J., Xu, J., Zheng, S. L., Yeager, M., Kaaks, R., Diver, W. R., Gaudet, M. M., Stern, M. C., Corral, R., Joshi, A. D., Shahabi, A., Wahlfors, T., Tammela, T. L. J., Auvinen, A., Virtamo, J., Klarskov, P., Nordestgaard, B. G., Røder, M. A., Nielsen, S. F., Bojesen, S. E., Siddiq, A., Fitzgerald, L. M., Kolb, S., Kwon, E. M., Karyadi, D. M., Blot, W. J., Zheng, W., Cai, Q., McDonnell, S. K., Rinckleb, A. E.,

Drake, B., Colditz, G., Wokolorczyk, D., Stephenson, R. A., Teerlink, C., Muller, H., Rothenbacher, D., Sellers, T. A., Lin, H.-Y., Slavov, C., Mitev, V., Lose, F., Srinivasan, S., Maia, S., Paulo, P., Lange, E., Cooney, K. A., Antoniou, A. C., Vincent, D., Bacot, F., Tessier, D. C., COGS–Cancer Research UK GWAS–ELLIPSE (part of GAME-ON) Initiative, Australian Prostate Cancer Bioresource, UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology, UK ProtecT (Prostate testing for cancer and Treatment) Study Collaborators, PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, Kote-Jarai, Z., and Easton, D. F. (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nature Genetics*, 45(4):385–391, 391e1–2.

Einarsdóttir, K., Darabi, H., Czene, K., Li, Y., Low, Y. L., Li, Y. Q., Bonnard, C., Wedrén, S., Liu, E. T., Hall, P., Liu, J., and Humphreys, K. (2009). Common genetic variability in ESR1 and EGF in relation to endometrial cancer risk and survival. *British Journal of Cancer*, 100(8):1358–1364.

Elshaikh, M. A., Munkarah, A. R., Robbins, J. R., Laser, B. S., Bhatt, N., Cogan, C., and Siddiqui, F. (2013). The impact of race on outcomes of patients with early stage uterine endometrioid carcinoma. *Gynecologic Oncology*, 128(2):171–174.

Ema, M., Mori, D., Niwa, H., Hasegawa, Y., Yamanaka, Y., Hitoshi, S., Mimura, J., Kawabe, Y.-i., Hosoya, T., Morita, M., Shimosato, D., Uchida, K., Suzuki, N., Yanagisawa, J., Sogawa, K., Rossant, J., Yamamoto, M., Takahashi, S., and Fujii-Kuriyama, Y. (2008). Krüppel-like factor 5 is essential for blastocyst development and the normal self-renewal of mouse ESCs. *Cell Stem Cell*, 3(5):555–567.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H.,

Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., Karaöz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., NISC Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J.,

Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Ruan, Y., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Esposito, K., Chiodini, P., Capuano, A., Bellastella, G., Maiorino, M. I., and Giugliano, D. (2014). Metabolic syndrome and endometrial cancer: a meta-analysis. *Endocrine*, 45(1):28–36.

Esteller, M., Levine, R., Baylin, S. B., Ellenson, L. H., and Herman, J. G. (1998). MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. *Oncogene*, 17(18):2413–2417.

Evangelou, E. and Ioannidis, J. P. A. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews. Genetics*, 14(6):379–389.

Evans, A. T., Gaffey, T. A., Malkasian, G. D., and Annegers, J. F. (1980). Clinicopathologic review of 118 granulosa and 82 theca cell tumors. *Obstetrics and Gynecology*, 55(2):231–238.

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F. O., and Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature Genetics*, 44(5):502–510.

Felix, A. S., Weissfeld, J. L., Stone, R. A., Bowser, R., Chivukula, M., Edwards, R. P., and Linkov, F. (2010). Factors associated with Type I and Type II endometrial cancer. *Cancer causes & control: CCC*, 21(11):1851–1856.

Felsenstein, J. and Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93–104.

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J. W. W., Comber, H., Forman, D., and Bray, F. (2013). Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *European Journal of Cancer (Oxford, England: 1990)*, 49(6):1374–1403.

Figueroa, J. D., Ye, Y., Siddiq, A., Garcia-Closas, M., Chatterjee, N., Prokunina-Olsson, L., Cortessis, V. K., Kooperberg, C., Cussenot, O., Benhamou, S., Prescott, J., Porru, S., Dinney, C. P., Malats, N., Baris, D., Purdue, M., Jacobs, E. J., Albanes, D., Wang, Z., Deng, X., Chung, C. C., Tang, W., Bas Bueno-de Mesquita, H., Trichopoulos, D., Ljungberg, B., Clavel-Chapelon, F., Weiderpass, E., Krogh, V., Dorronsoro, M., Travis, R., Tjønneland, A., Brenan, P., Chang-Claude, J., Riboli, E., Conti, D., Gago-Dominguez, M., Stern, M. C., Pike, M. C., Van Den Berg, D., Yuan, J.-M., Hohensee, C., Rodabough, R., Cancel-Tassin, G., Roupret, M., Comperat, E., Chen, C., De Vivo, I., Giovannucci, E., Hunter, D. J., Kraft, P., Lindstrom, S., Carta, A., Pavanello, S., Arici, C., Mastrangelo, G., Kamat, A. M., Lerner, S. P., Barton Grossman, H., Lin, J., Gu, J., Pu, X., Hutchinson, A., Burdette, L., Wheeler, W., Kogevinas, M., Tardón, A., Serra, C., Carrato, A.,

García-Closas, R., Lloreta, J., Schwenn, M., Karagas, M. R., Johnson, A., Schned, A., Armenti, K. R., Hosain, G. M., Andriole, G., Grubb, R., Black, A., Ryan Diver, W., Gapstur, S. M., Weinstein, S. J., Virtamo, J., Haiman, C. A., Landi, M. T., Caporaso, N., Fraumeni, J. F., Vineis, P., Wu, X., Silverman, D. T., Chanock, S., and Rothman, N. (2014). Genome-wide association study identifies multiple loci associated with bladder cancer risk. *Human Molecular Genetics*, 23(5):1387–1398.

Fisher, R. A. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Available from: `https://digital.library.adelaide.edu.au/dspace/handle/2440/15097`.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2(2):121–145.

Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J. W., Futreal, P. A., and Stratton, M. R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]*, Chapter 10:Unit 10.11.

Fung-Kee-Fung, M., Dodge, J., Elit, L., Lukka, H., Chambers, A., Oliver, T., and Cancer Care Ontario Program in Evidence-based Care Gynecology Cancer Disease Site Group (2006). Follow-up after primary therapy for endometrial cancer: a systematic review. *Gynecologic Oncology*, 101(3):520–529.

Garg, K., Shih, K., Barakat, R., Zhou, Q., Iasonos, A., and Soslow, R. A. (2009). Endometrial carcinomas in women aged 40 years and younger: tumors associated with loss of DNA mismatch repair proteins comprise a distinct clinicopathologic subset. *The American Journal of Surgical Pathology*, 33(12):1869–1877.

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17(6):877–885.

Goldgar, D. E., Easton, D. F., Cannon-Albright, L. A., and Skolnick, M. H. (1994). Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *Journal of the National Cancer Institute*, 86(21):1600–1608.

Goldsby, R. E., Lawrence, N. A., Hays, L. E., Olmsted, E. A., Chen, X., Singh, M., and Preston, B. D. (2001). Defective DNA polymerase-delta proofreading causes cancer susceptibility in mice. *Nature Medicine*, 7(6):638–639.

Goode, E. L., Chenevix-Trench, G., Song, H., Ramus, S. J., Notaridou, M., Lawrenson, K., Widschwendter, M., Vierkant, R. A., Larson, M. C., Kjaer, S. K., Birrer, M. J., Berchuck, A., Schildkraut, J., Tomlinson, I., Kiemeney, L. A., Cook, L. S., Gronwald, J., Garcia-Closas, M., Gore, M. E., Campbell, I., Whittemore, A. S., Sutphen, R., Phelan, C., Anton-Culver, H., Pearce, C. L., Lambrechts, D., Rossing, M. A., Chang-Claude, J., Moysich, K. B., Goodman, M. T., Dörk, T., Nevanlinna, H., Ness, R. B., Rafnar, T., Hogdall, C., Hogdall, E., Fridley, B. L., Cunningham, J. M., Sieh, W., McGuire, V., Godwin, A. K., Cramer, D. W., Hernandez, D., Levine, D., Lu, K., Iversen, E. S., Palmieri, R. T., Houlston, R., van Altena, A. M., Aben, K. K. H., Massuger, L. F. A. G., Brooks-Wilson, A., Kelemen, L. E., Le, N. D., Jakubowska, A., Lubinski, J., Medrek, K., Stafford, A., Easton, D. F., Tyrer, J., Bolton, K. L., Harrington, P., Eccles, D., Chen, A., Molina, A. N., Davila, B. N., Arango, H., Tsai, Y.-Y., Chen, Z., Risch, H. A., McLaughlin, J., Narod, S. A., Ziogas, A., Brewster, W., Gentry-Maharaj, A., Menon, U., Wu, A. H., Stram, D. O., Pike, M. C., Wellcome Trust Case-Control Consortium, Beesley, J., Webb, P. M., Australian Cancer Study (Ovarian Cancer), Australian Ovarian Cancer Study Group, Ovarian Cancer Association Consortium (OCAC), Chen, X., Ekici, A. B., Thiel, F. C., Beckmann, M. W., Yang, H., Wentzensen, N., Lissowska, J., Fasching, P. A., Despierre, E., Amant, F., Vergote, I., Doherty, J., Hein, R., Wang-Gohrke, S., Lurie, G., Carney, M. E., Thompson, P. J., Runnebaum, I., Hillemanns, P., Dürst, M., Antonenkova, N., Bogdanova, N., Leminen, A., Butzow, R., Heikkinen, T., Stefansson, K., Sulem, P., Besenbacher, S., Sellers, T. A., Gayther, S. A., Pharoah, P. D. P., and Ovarian Cancer Association Consortium (OCAC) (2010). A genome-wide association

study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nature Genetics*, 42(10):874–879.

Goodfellow, P. (2005). MSI in endometrial cancer: Prevalence and clinical applications. *International Journal of Gynecological Cancer*, 15(2):402–403. Available from: `http://onlinelibrary.wiley.com/doi/10.1111/j.1525-1438.2005.t01-1-abst_8.x/abstract`.

Grady, D., Gebretsadik, T., Kerlikowske, K., Ernster, V., and Petitti, D. (1995). Hormone replacement therapy and endometrial cancer risk: a meta-analysis. *Obstetrics and Gynecology*, 85(2):304–313.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science (New York, N.Y.)*, 185(4154):862–864.

Griffith, F. (1928). The Significance of Pneumococcal Types. *The Journal of Hygiene*, 27(2):113–159.

Gruber, S. B. and Thompson, W. D. (1996). A population-based study of endometrial cancer and familial risk in younger women. Cancer and Steroid Hormone Study Group. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 5(6):411–417.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585.

Gu, Y., Zhou, X., and Zhang, S. L. (2011). Meta-analysis of an association of codon 72 polymorphisms of the p53 gene with increased endometrial cancer risk. *Genetics and molecular research: GMR*, 10(4):3609–3619.

Gudmundsson, J., Sulem, P., Gudbjartsson, D. F., Blondal, T., Gylfason, A., Agnarsson, B. A., Benediktsdottir, K. R., Magnusdottir, D. N., Orlygsdottir, G., Jakobsdottir, M., Stacey, S. N., Sigurdsson, A., Wahlfors, T., Tammela, T., Breyer, J. P., McReynolds,

K. M., Bradley, K. M., Saez, B., Godino, J., Navarrete, S., Fuertes, F., Murillo, L., Polo, E., Aben, K. K., van Oort, I. M., Suarez, B. K., Helfand, B. T., Kan, D., Zanon, C., Frigge, M. L., Kristjansson, K., Gulcher, J. R., Einarsson, G. V., Jonsson, E., Catalona, W. J., Mayordomo, J. I., Kiemeney, L. A., Smith, J. R., Schleutker, J., Barkardottir, R. B., Kong, A., Thorsteinsdottir, U., Rafnar, T., and Stefansson, K. (2009). Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nature Genetics*, 41(10):1122–1126.

Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J. T., Thorleifsson, G., Manolescu, A., Rafnar, T., Gudbjartsson, D., Agnarsson, B. A., Baker, A., Sigurdsson, A., Benediktsdottir, K. R., Jakobsdottir, M., Blondal, T., Stacey, S. N., Helgason, A., Gunnarsdottir, S., Olafsdottir, A., Kristinsson, K. T., Birgisdottir, B., Ghosh, S., Thorlacius, S., Magnusdottir, D., Stefansdottir, G., Kristjansson, K., Bagger, Y., Wilensky, R. L., Reilly, M. P., Morris, A. D., Kimber, C. H., Adeyemo, A., Chen, Y., Zhou, J., So, W.-Y., Tong, P. C. Y., Ng, M. C. Y., Hansen, T., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Tres, A., Fuertes, F., Ruiz-Echarri, M., Asin, L., Saez, B., van Boven, E., Klaver, S., Swinkels, D. W., Aben, K. K., Graif, T., Cashy, J., Suarez, B. K., van Vierssen Trip, O., Frigge, M. L., Ober, C., Hofker, M. H., Wijmenga, C., Christiansen, C., Rader, D. J., Palmer, C. N. A., Rotimi, C., Chan, J. C. N., Pedersen, O., Sigurdsson, G., Benediktsson, R., Jonsson, E., Einarsson, G. V., Mayordomo, J. I., Catalona, W. J., Kiemeney, L. A., Barkardottir, R. B., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2007). Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nature Genetics*, 39(8):977–983.

Haile, R. W., Siegmund, K. D., Gauderman, W. J., and Thomas, D. C. (1999). Study-design issues in the development of the University of Southern California Consortium's Colorectal Cancer Family Registry. *Journal of the National Cancer Institute. Monographs*, (26):89–93.

Haiman, C. A., Chen, G. K., Vachon, C. M., Canzian, F., Dunning, A., Millikan, R. C., Wang, X., Ademuyiwa, F., Ahmed, S., Ambrosone, C. B., Baglietto, L., Balleine, R.,

Bandera, E. V., Beckmann, M. W., Berg, C. D., Bernstein, L., Blomqvist, C., Blot, W. J., Brauch, H., Buring, J. E., Carey, L. A., Carpenter, J. E., Chang-Claude, J., Chanock, S. J., Chasman, D. I., Clarke, C. L., Cox, A., Cross, S. S., Deming, S. L., Diasio, R. B., Dimopoulos, A. M., Driver, W. R., Dünnebier, T., Durcan, L., Eccles, D., Edlund, C. K., Ekici, A. B., Fasching, P. A., Feigelson, H. S., Flesch-Janys, D., Fostira, F., Försti, A., Fountzilas, G., Gerty, S. M., Gene Environment Interaction and Breast Cancer in Germany (GENICA) Consortium, Giles, G. G., Godwin, A. K., Goodfellow, P., Graham, N., Greco, D., Hamann, U., Hankinson, S. E., Hartmann, A., Hein, R., Heinz, J., Holbrook, A., Hoover, R. N., Hu, J. J., Hunter, D. J., Ingles, S. A., Irwanto, A., Ivanovich, J., John, E. M., Johnson, N., Jukkola-Vuorinen, A., Kaaks, R., Ko, Y.-D., Kolonel, L. N., Konstantopoulou, I., Kosma, V.-M., Kulkarni, S., Lambrechts, D., Lee, A. M., Marchand, L. L., Lesnick, T., Liu, J., Lindstrom, S., Mannermaa, A., Margolin, S., Martin, N. G., Miron, P., Montgomery, G. W., Nevanlinna, H., Nickels, S., Nyante, S., Olswold, C., Palmer, J., Pathak, H., Pectasides, D., Perou, C. M., Peto, J., Pharoah, P. D. P., Pooler, L. C., Press, M. F., Pylkäs, K., Rebbeck, T. R., Rodriguez-Gil, J. L., Rosenberg, L., Ross, E., Rüdiger, T., Silva, I. d. S., Sawyer, E., Schmidt, M. K., Schulz-Wendtland, R., Schumacher, F., Severi, G., Sheng, X., Signorello, L. B., Sinn, H.-P., Stevens, K. N., Southey, M. C., Tapper, W. J., Tomlinson, I., Hogervorst, F. B. L., Wauters, E., Weaver, J., Wildiers, H., Winqvist, R., Van Den Berg, D., Wan, P., Xia, L. Y., Yannoukakos, D., Zheng, W., Ziegler, R. G., Siddiq, A., Slager, S. L., Stram, D. O., Easton, D., Kraft, P., Henderson, B. E., and Couch, F. J. (2011). A common variant at the TERT-CLPTM1l locus is associated with estrogen receptor-negative breast cancer. *Nature Genetics*, 43(12):1210–1214.

Haiman, C. A., Dossus, L., Setiawan, V. W., Stram, D. O., Dunning, A. M., Thomas, G., Thun, M. J., Albanes, D., Altshuler, D., Ardanaz, E., Boeing, H., Buring, J., Burtt, N., Calle, E. E., Chanock, S., Clavel-Chapelon, F., Colditz, G. A., Cox, D. G., Feigelson, H. S., Hankinson, S. E., Hayes, R. B., Henderson, B. E., Hirschhorn, J. N., Hoover, R., Hunter, D. J., Kaaks, R., Kolonel, L. N., Le Marchand, L., Lenner, P., Lund, E., Panico, S., Peeters, P. H., Pike, M. C., Riboli, E., Tjonneland, A., Travis, R., Trichopoulos,

D., Wacholder, S., and Ziegler, R. G. (2007). Genetic variation at the CYP19a1 locus predicts circulating estrogen levels but not breast cancer risk in postmenopausal women. *Cancer Research*, 67(5):1893–1897.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–517.

Hampel, H., Frankel, W., Panescu, J., Lockman, J., Sotamaa, K., Fix, D., Comeras, I., La Jeunesse, J., Nakagawa, H., Westman, J. A., Prior, T. W., Clendenning, M., Penzone, P., Lombardi, J., Dunn, P., Cohn, D. E., Copeland, L., Eaton, L., Fowler, J., Lewandowski, G., Vaccarello, L., Bell, J., Reid, G., and de la Chapelle, A. (2006). Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients. *Cancer Research*, 66(15):7810–7817.

Hampel, H., Panescu, J., Lockman, J., Sotamaa, K., Fix, D., Comeras, I., LaJeunesse, J., Nakagawa, H., Westman, J. A., Prior, T. W., Clendenning, M., de la Chapelle, A., Frankel, W., Penzone, P., Cohn, D. E., Copeland, L., Eaton, L., Fowler, J., Lombardi, J., Dunn, P., Bell, J., Reid, G., Lewandowski, G., and Vaccarello, L. (2007). Comment on: Screening for Lynch Syndrome (Hereditary Nonpolyposis Colorectal Cancer) among Endometrial Cancer Patients. *Cancer Research*, 67(19):9603.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.

Haraldsdottir, S., Hampel, H., Tomsic, J., Frankel, W. L., Pearlman, R., de la Chapelle, A., and Pritchard, C. C. (2014). Colon and Endometrial Cancers with Mismatch Repair Deficiency can Arise from Somatic, Rather Than Germline, Mutations. *Gastroenterology*.

Hardy, G. H. (1908). MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science (New York, N.Y.)*, 28(706):49–50.

Havrilesky, L. J., Gierisch, J. M., Moorman, P. G., Coeytaux, R. R., Urrutia, R. P., Lowery,

W. J., Dinan, M., McBroom, A. J., Wing, L., Musty, M. D., Lallinger, K. R., Hasselblad, V., Sanders, G. D., and Myers, E. R. (2013). Oral contraceptive use for the primary prevention of ovarian cancer. *Evidence Report/Technology Assessment*, (212):1–514.

Hecht, J. L. and Mutter, G. L. (2006). Molecular and pathologic aspects of endometrial carcinogenesis. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 24(29):4783–4791.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112.

Hemminki, A., Peltomäki, P., Mecklin, J. P., Järvinen, H., Salovaara, R., Nyström-Lahti, M., de la Chapelle, A., and Aaltonen, L. A. (1994). Loss of the wild type MLH1 gene is a feature of hereditary nonpolyposis colorectal cancer. *Nature Genetics*, 8(4):405–410.

Hemminki, K., Aaltonen, L., and Li, X. (2003). Subsequent primary malignancies after endometrial carcinoma and ovarian carcinoma. *Cancer*, 97(10):2432–2439.

Herman, J. G., Umar, A., Polyak, K., Graff, J. R., Ahuja, N., Issa, J. P., Markowitz, S., Willson, J. K., Hamilton, S. R., Kinzler, K. W., Kane, M. F., Kolodner, R. D., Vogelstein, B., Kunkel, T. A., and Baylin, S. B. (1998). Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proceedings of the National Academy of Sciences of the United States of America*, 95(12):6870–6875.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., and Xu, C.-F. (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European journal of human genetics: EJHG*, 12(5):395–399.

Houlston, R. S., Cheadle, J., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., Spain, S. L., Broderick, P., Domingo, E., Farrington, S., Prendergast, J. G. D., Pittman, A. M.,

Theodoratou, E., Smith, C. G., Olver, B., Walther, A., Barnetson, R. A., Churchman, M., Jaeger, E. E. M., Penegar, S., Barclay, E., Martin, L., Gorman, M., Mager, R., Johnstone, E., Midgley, R., Niittymäki, I., Tuupanen, S., Colley, J., Idziaszczyk, S., COGENT Consortium, Thomas, H. J. W., Lucassen, A. M., Evans, D. G. R., Maher, E. R., CORGI Consortium, COIN Collaborative Group, COINB Collaborative Group, Maughan, T., Dimas, A., Dermitzakis, E., Cazier, J.-B., Aaltonen, L. A., Pharoah, P., Kerr, D. J., Carvajal-Carmona, L. G., Campbell, H., Dunlop, M. G., and Tomlinson, I. P. M. (2010). Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nature Genetics*, 42(11):973–977.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, 5(6). Available from: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2689936/`.

Hsieh, P. and Yamane, K. (2008). DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mechanisms of Ageing and Development*, 129(7-8):391–407.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., and Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods*, 11(2):193–206.

Hunt, K. A., Zhernakova, A., Turner, G., Heap, G. A. R., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L. C., Ryan, A. W., Panesar, D., Gwilliam, R., Takeuchi, F., McLaren, W. M., Holmes, G. K. T., Howdle, P. D., Walters, J. R. F., Sanders, D. S., Playford, R. J., Trynka, G., Mulder, C. J. J., Mearin, M. L., Verbeek, W. H. M., Trimble, V., Stevens, F. M., O'Morain, C., Kennedy, N. P., Kelleher, D., Pennington, D. J., Strachan, D. P., McArdle, W. L., Mein, C. A., Wapenaar, M. C., Deloukas, P., McGinnis, R., McManus, R., Wijmenga, C., and van Heel, D. A. (2008). Newly identified genetic risk

variants for celiac disease related to the immune response. *Nature Genetics*, 40(4):395–402.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A., Wang, J., Yu, K., Chatterjee, N., Orr, N., Willett, W. C., Colditz, G. A., Ziegler, R. G., Berg, C. D., Buys, S. S., McCarty, C. A., Feigelson, H. S., Calle, E. E., Thun, M. J., Hayes, R. B., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R. N., Thomas, G., and Chanock, S. J. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics*, 39(7):870–874.

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I. W., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Gibbs, R. A., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Bonnen, P. E., Gibbs, R. A., Gonzaga-Jauregui, C., Keinan, A., Price, A. L., Yu, F., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S. F., Zhang, Q., Ghori, M. J. R., McGinnis, R., McLaren, W., Pollack, S., Price, A. L., Schaffner, S. F., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., and McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58.

International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(6968):789–796.

Jacobs, I., Gentry-Maharaj, A., Burnell, M., Manchanda, R., Singh, N., Sharma, A., Ryan, A., Seif, M. W., Amso, N. N., Turner, G., Brunell, C., Fletcher, G., Rangar, R., Ford, K., Godfrey, K., Lopes, A., Oram, D., Herod, J., Williamson, K., Scott, I., Jenkins, H., Mould, T., Woolas, R., Murdoch, J., Dobbs, S., Leeson, S., Cruickshank, D., Skates, S. J., Fallowfield, L., Parmar, M., Campbell, S., and Menon, U. (2011). Sensitivity of transvaginal ultrasound screening for endometrial cancer in postmenopausal women: a case-control study within the UKCTOCS cohort. *The Lancet. Oncology*, 12(1):38–48.

Jass, J. R., Cottier, D. S., Jeevaratnam, P., Pokos, V., Holdaway, K. M., Bowden, M. L., Van de Water, N. S., and Browett, P. J. (1995). Diagnostic use of microsatellite instability in hereditary non-polyposis colorectal cancer. *Lancet*, 346(8984):1200–1201.

Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics*, 29(2):217–222.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90.

Jiang, J., Chan, Y.-S., Loh, Y.-H., Cai, J., Tong, G.-Q., Lim, C.-A., Robson, P., Zhong, S., and Ng, H.-H. (2008). A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nature Cell Biology*, 10(3):353–360.

Jégu, J., Colonna, M., Daubisse-Marliac, L., Trétarre, B., Ganry, O., Guizard, A.-V., Bara, S., Troussard, X., Bouvier, V., Woronoff, A.-S., and Velten, M. (2014). The effect of patient characteristics on second primary cancer risk in France. *BMC cancer*, 14:94.

Kanaya, T., Kyo, S., Maida, Y., Yatabe, N., Tanaka, M., Nakamura, M., and Inoue, M. (2003). Frequent hypermethylation of MLH1 promoter in normal endometrium of patients with endometrial cancers. *Oncogene*, 22(15):2352–2360. Available from: `http://www.nature.com/onc/journal/v22/n15/full/1206365a.html#bib6`.

Kane, M. F., Loda, M., Gaida, G. M., Lipman, J., Mishra, R., Goldman, H., Jessup, J. M.,

and Kolodner, R. (1997). Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer Research*, 57(5):808–811.

Kastrinos, F., Mukherjee, B., Tayob, N., Wang, F., Sparr, J., Raymond, V. M., Bandipalliam, P., Stoffel, E. M., Gruber, S. B., and Syngal, S. (2009). Risk of pancreatic cancer in families with Lynch syndrome. *JAMA*, 302(16):1790–1795.

Katoh, M. and Katoh, M. (2007). Integrative genomic analyses on HES/HEY family: Notch-independent HES1, HES3 transcription in undifferentiated ES cells, and Notch-dependent HES1, HES5, HEY1, HEY2, HEYL transcription in fetal tissues, adult tissues, or cancer. *International Journal of Oncology*, 31(2):461–466.

Kempers, M. J. E., Kuiper, R. P., Ockeloen, C. W., Chappuis, P. O., Hutter, P., Rahner, N., Schackert, H. K., Steinke, V., Holinski-Feder, E., Morak, M., Kloor, M., Büttner, R., Verwiel, E. T. P., van Krieken, J. H., Nagtegaal, I. D., Goossens, M., van der Post, R. S., Niessen, R. C., Sijmons, R. H., Kluijt, I., Hogervorst, F. B. L., Leter, E. M., Gille, J. J. P., Aalfs, C. M., Redeker, E. J. W., Hes, F. J., Tops, C. M. J., van Nesselrooij, B. P. M., van Gijn, M. E., Gómez García, E. B., Eccles, D. M., Bunyan, D. J., Syngal, S., Stoffel, E. M., Culver, J. O., Palomares, M. R., Graham, T., Velsher, L., Papp, J., Oláh, E., Chan, T. L., Leung, S. Y., van Kessel, A. G., Kiemeney, L. A. L. M., Hoogerbrugge, N., and Ligtenberg, M. J. L. (2011). Risk of colorectal and endometrial cancers in EPCAM deletion-positive Lynch syndrome: a cohort study. *The Lancet. Oncology*, 12(1):49–55.

Kiemeney, L. A., Thorlacius, S., Sulem, P., Geller, F., Aben, K. K. H., Stacey, S. N., Gudmundsson, J., Jakobsdottir, M., Bergthorsson, J. T., Sigurdsson, A., Blondal, T., Witjes, J. A., Vermeulen, S. H., Hulsbergen-van de Kaa, C. A., Swinkels, D. W., Ploeg, M., Cornel, E. B., Vergunst, H., Thorgeirsson, T. E., Gudbjartsson, D., Gudjonsson, S. A., Thorleifsson, G., Kristinsson, K. T., Mouy, M., Snorradottir, S., Placidi, D., Campagna, M., Arici, C., Koppova, K., Gurzau, E., Rudnai, P., Kellen, E., Polidoro, S., Guarrera, S., Sacerdote, C., Sanchez, M., Saez, B., Valdivia, G., Ryk, C., de Verdier, P.,

Lindblom, A., Golka, K., Bishop, D. T., Knowles, M. A., Nikulasson, S., Petursdottir, V., Jonsson, E., Geirsson, G., Kristjansson, B., Mayordomo, J. I., Steineck, G., Porru, S., Buntinx, F., Zeegers, M. P., Fletcher, T., Kumar, R., Matullo, G., Vineis, P., Kiltie, A. E., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., Rafnar, T., and Stefansson, K. (2008). Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nature Genetics*, 40(11):1307–1312.

Kim, J. J. and Chapman-Davis, E. (2010). Role of progesterone in endometrial cancer. *Seminars in Reproductive Medicine*, 28(1):81–90.

Kinnersley, B., Migliorini, G., Broderick, P., Whiffin, N., Dobbins, S. E., Casey, G., Hopper, J., Sieber, O., Lipton, L., Kerr, D. J., Dunlop, M. G., Tomlinson, I. P., Houlston, R. S., and Colon Cancer Family Registry (2012). The TERT variant rs2736100 is associated with colorectal cancer risk. *British Journal of Cancer*, 107(6):1001–1008.

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720):385–389.

Knudson, A. G. (1985). Hereditary cancer, oncogenes, and antioncogenes. *Cancer Research*, 45(4):1437–1443.

Korch, C., Spillman, M. A., Jackson, T. A., Jacobsen, B. M., Murphy, S. K., Lessey, B. A., Jordan, V. C., and Bradford, A. P. (2012). DNA profiling analysis of endometrial and ovarian cell lines reveals misidentification, redundancy and contamination. *Gynecologic Oncology*, 127(1):241–248.

Kote-Jarai, Z., Olama, A. A. A., Giles, G. G., Severi, G., Schleutker, J., Weischer, M.,

Campa, D., Riboli, E., Key, T., Gronberg, H., Hunter, D. J., Kraft, P., Thun, M. J., Ingles, S., Chanock, S., Albanes, D., Hayes, R. B., Neal, D. E., Hamdy, F. C., Donovan, J. L., Pharoah, P., Schumacher, F., Henderson, B. E., Stanford, J. L., Ostrander, E. A., Sorensen, K. D., Dörk, T., Andriole, G., Dickinson, J. L., Cybulski, C., Lubinski, J., Spurdle, A., Clements, J. A., Chambers, S., Aitken, J., Gardiner, R. A. F., Thibodeau, S. N., Schaid, D., John, E. M., Maier, C., Vogel, W., Cooney, K. A., Park, J. Y., Cannon-Albright, L., Brenner, H., Habuchi, T., Zhang, H.-W., Lu, Y.-J., Kaneva, R., Muir, K., Benlloch, S., Leongamornlert, D. A., Saunders, E. J., Tymrakiewicz, M., Mahmud, N., Guy, M., O'Brien, L. T., Wilkinson, R. A., Hall, A. L., Sawyer, E. J., Dadaev, T., Morrison, J., Dearnaley, D. P., Horwich, A., Huddart, R. A., Khoo, V. S., Parker, C. C., Van As, N., Woodhouse, C. J., Thompson, A., Christmas, T., Ogden, C., Cooper, C. S., Lophatonanon, A., Southey, M. C., Hopper, J. L., English, D. R., Wahlfors, T., Tammela, T. L. J., Klarskov, P., Nordestgaard, B. G., Røder, M. A., Tybjærg-Hansen, A., Bojesen, S. E., Travis, R., Canzian, F., Kaaks, R., Wiklund, F., Aly, M., Lindstrom, S., Diver, W. R., Gapstur, S., Stern, M. C., Corral, R., Virtamo, J., Cox, A., Haiman, C. A., Le Marchand, L., Fitzgerald, L., Kolb, S., Kwon, E. M., Karyadi, D. M., Orntoft, T. F., Borre, M., Meyer, A., Serth, J., Yeager, M., Berndt, S. I., Marthick, J. R., Patterson, B., Wokolorczyk, D., Batra, J., Lose, F., McDonnell, S. K., Joshi, A. D., Shahabi, A., Rinckleb, A. E., Ray, A., Sellers, T. A., Lin, H.-Y., Stephenson, R. A., Farnham, J., Muller, H., Rothenbacher, D., Tsuchiya, N., Narita, S., Cao, G.-W., Slavov, C., Mitev, V., Easton, D. F., Eeles, R. A., UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology, UK ProtecT Study Collaborators, The Australian Prostate Cancer BioResource, and PRACTICAL Consortium (2011). Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nature Genetics*, 43(8):785–791.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081.

Kunkel, T. A. and Erie, D. A. (2005). DNA mismatch repair. *Annual Review of Biochemistry*, 74:681–710.

Kwon, J. S., Scott, J. L., Gilks, C. B., Daniels, M. S., Sun, C. C., and Lu, K. H. (2011). Testing women with endometrial cancer to detect Lynch syndrome. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 29(16):2247–2252.

Lancaster, J. M., Powell, C. B., Kauff, N. D., Cass, I., Chen, L.-M., Lu, K. H., Mutch, D. G., Berchuck, A., Karlan, B. Y., Herzog, T. J., and Society of Gynecologic Oncologists Education Committee (2007). Society of Gynecologic Oncologists Education Committee statement on risk assessment for inherited gynecologic cancer predispositions. *Gynecologic Oncology*, 107(2):159–162.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker,

M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., Bastide, M. d. l., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. Available from: `http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html`.

Levene, P. A. (1919). The Structure of Yeast Nucleic Acid Iv. Ammonia Hydrolysis. *Journal of Biological Chemistry*, 40(2):415–424. Available from: `http://www.jbc.org/content/40/2/415`.

Lewis, A., Freeman-Mills, L., de la Calle-Mustienes, E., Giráldez-Pérez, R. M., Davis,

H., Jaeger, E., Becker, M., Hubner, N. C., Nguyen, L. N., Zeron-Medina, J., Bond, G., Stunnenberg, H. G., Carvajal, J. J., Gomez-Skarmeta, J. L., Leedham, S., and Tomlinson, I. (2014). A polymorphic enhancer near GREM1 influences bowel cancer risk through differential CDX2 and TCF7l2 binding. *Cell Reports*, 8(4):983–990.

Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49(1):49–67.

Li, N., Jiang, P., Du, W., Wu, Z., Li, C., Qiao, M., Yang, X., and Wu, M. (2011). Siva1 suppresses epithelial-mesenchymal transition and metastasis of tumor cells by inhibiting stathmin and stabilizing microtubules. *Proceedings of the National Academy of Sciences of the United States of America*, 108(31):12851–12856.

Li, Q., Seo, J.-H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., Brown, M., Tyekucheva, S., and Freedman, M. L. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*, 152(3):633–641.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England Journal of Medicine*, 343(2):78–85.

Ligtenberg, M. J. L., Kuiper, R. P., Chan, T. L., Goossens, M., Hebeda, K. M., Voorendt, M., Lee, T. Y. H., Bodmer, D., Hoenselaar, E., Hendriks-Cornelissen, S. J. B., Tsui, W. Y., Kong, C. K., Brunner, H. G., van Kessel, A. G., Yuen, S. T., van Krieken, J. H. J. M., Leung, S. Y., and Hoogerbrugge, N. (2009). Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nature Genetics*, 41(1):112–117.

Lindblom, A., Tannergård, P., Werelius, B., and Nordenskjöld, M. (1993). Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nature Genetics*, 5(3):279–282.

Lindor, N. M., Petersen, G. M., Hadley, D. W., Kinney, A. Y., Miesfeldt, S., Lu, K. H., Lynch, P., Burke, W., and Press, N. (2006). Recommendations for the care of individuals with an inherited predisposition to Lynch syndrome: a systematic review. *JAMA*, 296(12):1507–1517.

Liu, B., Nicolaides, N. C., Markowitz, S., Willson, J. K., Parsons, R. E., Jen, J., Papadopolous, N., Peltomäki, P., de la Chapelle, A., and Hamilton, S. R. (1995). Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. *Nature Genetics*, 9(1):48–55.

Liu, B., Parsons, R. E., Hamilton, S. R., Petersen, G. M., Lynch, H. T., Watson, P., Markowitz, S., Willson, J. K., Green, J., and de la Chapelle, A. (1994). hMSH2 mutations in hereditary nonpolyposis colorectal cancer kindreds. *Cancer Research*, 54(17):4590–4594.

Liu, N., Li, H., Li, S., Shen, M., Xiao, N., Chen, Y., Wang, Y., Wang, W., Wang, R., Wang, Q., Sun, J., and Wang, P. (2010). The Fbw7/human CDC4 tumor suppressor targets proproliferative factor KLF5 for ubiquitination and degradation through multiple phosphodegron motifs. *The Journal of Biological Chemistry*, 285(24):18858–18867.

Livak, K. J. and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods (San Diego, Calif.)*, 25(4):402–408.

Long, Q., Park, B. K., and Ekker, M. (2001). Expression and regulation of mouse Mtsh1 during limb and branchial arch development. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, 222(2):308–312.

Lorenz, M. G. and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. *Microbiological Reviews*, 58(3):563–602.

Lu, K. H., Dinh, M., Kohlmann, W., Watson, P., Green, J., Syngal, S., Bandipalliam, P., Chen, L.-M., Allen, B., Conrad, P., Terdiman, J., Sun, C., Daniels, M., Burke, T.,

Gershenson, D. M., Lynch, H., Lynch, P., and Broaddus, R. R. (2005). Gynecologic cancer as a "sentinel cancer" for women with hereditary nonpolyposis colorectal cancer syndrome. *Obstetrics and Gynecology*, 105(3):569–574.

Lucenteforte, E., Talamini, R., Montella, M., Dal Maso, L., Pelucchi, C., Franceschi, S., La Vecchia, C., and Negri, E. (2009). Family history of cancer and the risk of endometrial cancer. *European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP)*, 18(2):95–99.

Lukanova, A., Lundin, E., Micheli, A., Arslan, A., Ferrari, P., Rinaldi, S., Krogh, V., Lenner, P., Shore, R. E., Biessy, C., Muti, P., Riboli, E., Koenig, K. L., Levitz, M., Stattin, P., Berrino, F., Hallmans, G., Kaaks, R., Toniolo, P., and Zeleniuch-Jacquotte, A. (2004). Circulating levels of sex steroid hormones and risk of endometrial cancer in postmenopausal women. *International Journal of Cancer. Journal International Du Cancer*, 108(3):425–432.

Luo, J., Beresford, S., Chen, C., Chlebowski, R., Garcia, L., Kuller, L., Regier, M., Wactawski-Wende, J., and Margolis, K. L. (2014). Association between diabetes, diabetes treatment and risk of developing endometrial cancer. *British Journal of Cancer*, 111(7):1432–1439.

Lynch, H. T., Harris, R. E., Lynch, P. M., Guirgis, H. A., Lynch, J. F., and Bardawil, W. A. (1977). Role of heredity in multiple primary cancer. *Cancer*, 40(4 Suppl):1849–1854.

Lynch, H. T. and Krush, A. J. (1971). The cancer family syndrome and cancer control. *Surgery, Gynecology & Obstetrics*, 132(2):247–250.

Manfroid, I., Caubit, X., Kerridge, S., and Fasano, L. (2004). Three putative murine Teashirt orthologues specify trunk structures in Drosophila in the same way as the Drosophila teashirt gene. *Development (Cambridge, England)*, 131(5):1065–1073.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association

studies. *Nature Reviews Genetics*, 11(7):499–511. Available from: `http://www.nature.com/nrg/journal/v11/n7/abs/nrg2796.html`.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7):906–913.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369. Available from: `http://www.nature.com/nrg/journal/v9/n5/full/nrg2344.html`.

McConnell, B. B., Bialkowska, A. B., Nandan, M. O., Ghaleb, A. M., Gordon, F. J., and Yang, V. W. (2009). Haploinsufficiency of Krüppel-like factor 5 rescues the tumor-initiating effect of the Apc(Min) mutation in the intestine. *Cancer Research*, 69(10):4125–4133.

McConnell, B. B., Ghaleb, A. M., Nandan, M. O., and Yang, V. W. (2007). The diverse functions of Krüppel-like factors 4 and 5 in epithelial biology and pathobiology. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 29(6):549–557.

McEvoy, M., Smith, W., D'Este, C., Duke, J., Peel, R., Schofield, P., Scott, R., Byles, J., Henry, D., Ewald, B., Hancock, S., Smith, D., and Attia, J. (2010). Cohort profile: The Hunter Community Study. *International Journal of Epidemiology*, 39(6):1452–1463.

McGregor, B., Pfitzner, J., Zhu, G., Grace, M., Eldridge, A., Pearson, J., Mayne, C., Aitken, J. F., Green, A. C., and Martin, N. G. (1999). Genetic and environmental contributions to size, color, shape, and other characteristics of melanocytic naevi in a sample of adolescent twins. *Genetic Epidemiology*, 16(1):40–53.

McKay, J. D., Hung, R. J., Gaborieau, V., Boffetta, P., Chabrier, A., Byrnes, G., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E.,

Mates, D., Bencko, V., Foretova, L., Janout, V., McLaughlin, J., Shepherd, F., Montpetit, A., Narod, S., Krokan, H. E., Skorpen, F., Elvestad, M. B., Vatten, L., Njølstad, I., Axelsson, T., Chen, C., Goodman, G., Barnett, M., Loomis, M. M., Lubiński, J., Matyjasik, J., Lener, M., Oszutowska, D., Field, J., Liloglou, T., Xinarianos, G., Cassidy, A., EPIC Study, Vineis, P., Clavel-Chapelon, F., Palli, D., Tumino, R., Krogh, V., Panico, S., González, C. A., Ramón Quirós, J., Martínez, C., Navarro, C., Ardanaz, E., Larrañaga, N., Kham, K. T., Key, T., Bueno-de Mesquita, H. B., Peeters, P. H., Trichopoulou, A., Linseisen, J., Boeing, H., Hallmans, G., Overvad, K., Tjønneland, A., Kumle, M., Riboli, E., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Matsuda, F., Blanche, H., Gut, I., Heath, S., Lathrop, M., and Brennan, P. (2008). Lung cancer susceptibility locus at 5p15.33. *Nature Genetics*, 40(12):1404–1406.

McVean, G., Spencer, C. C. A., and Chaix, R. (2005). Perspectives on Human Genetic Variation from the HapMap Project. *PLoS Genet*, 1(4):e54. Available from: `http://dx.plos.org/10.1371/journal.pgen.0010054`.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science (New York, N.Y.)*, 304(5670):581–584.

Mecklin, J. P. and Järvinen, H. J. (1986). Clinical features of colorectal carcinoma in cancer family syndrome. *Diseases of the Colon and Rectum*, 29(3):160–164.

Meyer, L. A., Broaddus, R. R., and Lu, K. H. (2009). Endometrial cancer and Lynch syndrome: clinical and pathologic considerations. *Cancer Control: Journal of the Moffitt Cancer Center*, 16(1):14–22.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., Wang, Q., Dicks, E., Lee, A., Turnbull, C., Rahman, N., Breast and Ovarian Cancer Susceptibility Collaboration, Fletcher, O., Peto, J., Gibson, L., Dos Santos Silva, I., Nevanlinna, H., Muranen, T. A., Aittomäki, K., Blomqvist, C., Czene, K., Irwanto, A., Liu, J., Waisfisz, Q.,

Meijers-Heijboer, H., Adank, M., Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), van der Luijt, R. B., Hein, R., Dahmen, N., Beckman, L., Meindl, A., Schmutzler, R. K., Müller-Myhsok, B., Lichtner, P., Hopper, J. L., Southey, M. C., Makalic, E., Schmidt, D. F., Uitterlinden, A. G., Hofman, A., Hunter, D. J., Chanock, S. J., Vincent, D., Bacot, F., Tessier, D. C., Canisius, S., Wessels, L. F. A., Haiman, C. A., Shah, M., Luben, R., Brown, J., Luccarini, C., Schoof, N., Humphreys, K., Li, J., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Couch, F. J., Wang, X., Vachon, C., Stevens, K. N., Lambrechts, D., Moisse, M., Paridaens, R., Christiaens, M.-R., Rudolph, A., Nickels, S., Flesch-Janys, D., Johnson, N., Aitken, Z., Aaltonen, K., Heikkinen, T., Broeks, A., Veer, L. J. V., van der Schoot, C. E., Guénel, P., Truong, T., Laurent-Puig, P., Menegaux, F., Marme, F., Schneeweiss, A., Sohn, C., Burwinkel, B., Zamora, M. P., Perez, J. I. A., Pita, G., Alonso, M. R., Cox, A., Brock, I. W., Cross, S. S., Reed, M. W. R., Sawyer, E. J., Tomlinson, I., Kerin, M. J., Miller, N., Henderson, B. E., Schumacher, F., Le Marchand, L., Andrulis, I. L., Knight, J. A., Glendon, G., Mulligan, A. M., kConFab Investigators, Australian Ovarian Cancer Study Group, Lindblom, A., Margolin, S., Hooning, M. J., Hollestelle, A., van den Ouweland, A. M. W., Jager, A., Bui, Q. M., Stone, J., Dite, G. S., Apicella, C., Tsimiklis, H., Giles, G. G., Severi, G., Baglietto, L., Fasching, P. A., Haeberle, L., Ekici, A. B., Beckmann, M. W., Brenner, H., Müller, H., Arndt, V., Stegmaier, C., Swerdlow, A., Ashworth, A., Orr, N., Jones, M., Figueroa, J., Lissowska, J., Brinton, L., Goldberg, M. S., Labrèche, F., Dumont, M., Winqvist, R., Pylkäs, K., Jukkola-Vuorinen, A., Grip, M., Brauch, H., Hamann, U., Brüning, T., GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network, Radice, P., Peterlongo, P., Manoukian, S., Bonanni, B., Devilee, P., Tollenaar, R. A. E. M., Seynaeve, C., van Asperen, C. J., Jakubowska, A., Lubinski, J., Jaworska, K., Durda, K., Mannermaa, A., Kataja, V., Kosma, V.-M., Hartikainen, J. M., Bogdanova, N. V., Antonenkova, N. N., Dörk, T., Kristensen, V. N., Anton-Culver, H., Slager, S., Toland, A. E., Edge, S., Fostira, F., Kang, D., Yoo, K.-Y., Noh, D.-Y., Matsuo, K., Ito, H., Iwata, H., Sueta, A., Wu, A. H., Tseng, C.-C., Van Den Berg, D., Stram, D. O., Shu, X.-O., Lu, W., Gao, Y.-T., Cai, H., Teo, S. H., Yip, C. H., Phuah, S. Y., Cornes,

B. K., Hartman, M., Miao, H., Lim, W. Y., Sng, J.-H., Muir, K., Lophatananon, A., Stewart-Brown, S., Siriwanarangsan, P., Shen, C.-Y., Hsiung, C.-N., Wu, P.-E., Ding, S.-L., Sangrajrang, S., Gaborieau, V., Brennan, P., McKay, J., Blot, W. J., Signorello, L. B., Cai, Q., Zheng, W., Deming-Halverson, S., Shrubsole, M., Long, J., Simard, J., Garcia-Closas, M., Pharoah, P. D. P., Chenevix-Trench, G., Dunning, A. M., Benitez, J., and Easton, D. F. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics*, 45(4):353–361, 361e1–2.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–777.

Moore, K. N. and Fader, A. N. (2011). Uterine papillary serous carcinoma. *Clinical Obstetrics and Gynecology*, 54(2):278–291.

Moore, S. C., Gierach, G. L., Schatzkin, A., and Matthews, C. E. (2010). Physical activity, sedentary behaviours, and the prevention of endometrial cancer. *British Journal of Cancer*, 103(7):933–938.

Mourits, M. J. E., Bijen, C. B., Arts, H. J., ter Brugge, H. G., van der Sijde, R., Paulsen, L., Wijma, J., Bongers, M. Y., Post, W. J., van der Zee, A. G., and de Bock, G. H. (2010). Safety of laparoscopy versus laparotomy in early-stage endometrial cancer: a randomised trial. *The Lancet. Oncology*, 11(8):763–771.

Mueck, A. O., Seeger, H., and Rabe, T. (2010). Hormonal contraception and risk of endometrial cancer: a systematic review. *Endocrine-Related Cancer*, 17(4):R263–271.

Mullis, K. B. and Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology*, 155:335–350.

Mutter, G. L., Baak, J. P., Fitzgerald, J. T., Gray, R., Neuberg, D., Kust, G. A., Gentleman, R., Gullans, S. R., Wei, L. J., and Wilcox, M. (2001). Global expression changes of con-

stitutive and hormonally regulated genes during endometrial neoplastic transformation. *Gynecologic Oncology*, 83(2):177–185.

Mägi, R. and Morris, A. P. (2010). GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics*, 11(1):288. Available from: `http://www.biomedcentral.com/1471-2105/11/288/abstract`.

Nandan, M. O., McConnell, B. B., Ghaleb, A. M., Bialkowska, A. B., Sheng, H., Shao, J., Babbin, B. A., Robine, S., and Yang, V. W. (2008). Krüppel-like factor 5 mediates cellular transformation during oncogenic KRAS-induced intestinal tumorigenesis. *Gastroenterology*, 134(1):120–130.

Neill, A. S., Nagle, C. M., Protani, M. M., Obermair, A., Spurdle, A. B., Webb, P. M., and Australian National Endometrial Cancer Study Group (2013). Aspirin, nonsteroidal anti-inflammatory drugs, paracetamol and risk of endometrial cancer: a case-control study, systematic review and meta-analysis. *International Journal of Cancer. Journal International Du Cancer*, 132(5):1146–1155.

Newcomb, P. A., Baron, J., Cotterchio, M., Gallinger, S., Grove, J., Haile, R., Hall, D., Hopper, J. L., Jass, J., Le Marchand, L., Limburg, P., Lindor, N., Potter, J. D., Templeton, A. S., Thibodeau, S., Seminara, D., and Colon Cancer Family Registry (2007). Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 16(11):2331–2343.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888.

Nyholm, H. C., Nielsen, A. L., Lyndrup, J., Dreisler, A., Hagen, C., and Haug, E. (1993). Plasma oestrogens in postmenopausal women with endometrial cancer. *British Journal of Obstetrics and Gynaecology*, 100(12):1115–1119.

Oki, E., Oda, S., Maehara, Y., and Sugimachi, K. (1999). Mutated gene-specific phenotypes of dinucleotide repeat instability in human colorectal carcinoma cell lines deficient in DNA mismatch repair. *Oncogene*, 18(12):2143–2147.

Oram, R. A., Edghill, E. L., Blackman, J., Taylor, M. J. O., Kay, T., Flanagan, S. E., Ismail-Pratt, I., Creighton, S. M., Ellard, S., Hattersley, A. T., and Bingham, C. (2010). Mutations in the hepatocyte nuclear factor-1b (HNF1b) gene are common with combined uterine and renal malformations but are not found with isolated uterine malformations. *American Journal of Obstetrics and Gynecology*, 203(4):364.e1–5.

Painter, J. N., O'Mara, T. A., Batra, J., Cheng, T., Lose, F. A., Dennis, J., Michailidou, K., Tyrer, J. P., Ahmed, S., Ferguson, K., Healey, C. S., Kaufmann, S., Hillman, K. M., Walpole, C., Moya, L., Pollock, P., Jones, A., Howarth, K., Martin, L., Gorman, M., Hodgson, S., National Study of Endometrial Cancer Genetics Group (NSECG), CHIBCHA Consortium, De Polanco, M. M. E., Sans, M., Carracedo, A., Castellvi-Bel, S., Rojas-Martinez, A., Santos, E., Teixeira, M. R., Carvajal-Carmona, L., Shu, X.-O., Long, J., Zheng, W., Xiang, Y.-B., The Australian National Endometrial Cancer Study Group (ANECS), Montgomery, G. W., Webb, P. M., Scott, R. J., McEvoy, M., Attia, J., Holliday, E., Martin, N. G., Nyholt, D. R., Henders, A. K., Fasching, P. A., Hein, A., Beckmann, M. W., Renner, S. P., Dörk, T., Hillemanns, P., Dürst, M., Runnebaum, I., Lambrechts, D., Coenegrachts, L., Schrauwen, S., Amant, F., Winterhoff, B., Dowdy, S. C., Goode, E. L., Teoman, A., Salvesen, H. B., Trovik, J., Njolstad, T. S., Werner, H. M. J., Ashton, K., Proietto, T., Otton, G., Tzortzatos, G., Mints, M., Tham, E., RENDOCAS, Hall, P., Czene, K., Liu, J., Li, J., Hopper, J. L., Southey, M. C., Australian Ovarian Cancer Study (AOCS), Ekici, A. B., Ruebner, M., Johnson, N., Peto, J., Burwinkel, B., Marme, F., Brenner, H., Dieffenbach, A. K., Meindl, A., Brauch, H., The GENICA Network, Lindblom, A., Depreeuw, J., Moisse, M., Chang-Claude, J., Rudolph, A., Couch, F. J., Olson, J. E., Giles, G. G., Bruinsma, F., Cunningham, J. M., Fridley, B. L., Børresen-Dale, A.-L., Kristensen, V. N., Cox, A., Swerdlow, A. J., Orr, N., Bolla, M. K., Wang, Q., Weber, R. P., Chen, Z., Shah, M., French, J. D., Pharoah, P. D. P.,

Dunning, A. M., Tomlinson, I., Easton, D. F., Edwards, S. L., Thompson, D. J., and Spurdle, A. B. (2014). Fine-mapping of the HNF1b multicancer locus identifies candidate variants that mediate endometrial cancer risk. *Human Molecular Genetics*.

Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., Kemp, Z., Spain, S. L., Guarino, E., Guarino Almeida, E., Salguero, I., Sherborne, A., Chubb, D., Carvajal-Carmona, L. G., Ma, Y., Kaur, K., Dobbins, S., Barclay, E., Gorman, M., Martin, L., Kovac, M. B., Humphray, S., CORGI Consortium, WGS500 Consortium, Lucassen, A., Holmes, C. C., Bentley, D., Donnelly, P., Taylor, J., Petridis, C., Roylance, R., Sawyer, E. J., Kerr, D. J., Clark, S., Grimes, J., Kearsey, S. E., Thomas, H. J. W., McVean, G., Houlston, R. S., and Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, 45(2):136–144.

Palomba, S., Falbo, A., Zullo, F., and Orio, F. (2009). Evidence-based and potential benefits of metformin in the polycystic ovary syndrome: a comprehensive review. *Endocrine Reviews*, 30(1):1–50.

Parazzini, F., La Vecchia, C., Moroni, S., Chatenoud, L., and Ricci, E. (1994). Family history and the risk of endometrial cancer. *International Journal of Cancer. Journal International Du Cancer*, 59(4):460–462.

Parkin, D. M., Boyd, L., and Walker, L. C. (2011). 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. *British Journal of Cancer*, 105 Suppl 2:S77–81.

Parsons, R., Li, G. M., Longley, M. J., Fang, W. H., Papadopoulos, N., Jen, J., de la Chapelle, A., Kinzler, K. W., Vogelstein, B., and Modrich, P. (1993). Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell*, 75(6):1227–1236.

Patnaik, P. B. (1949). The non-central chi2- and F-distributions and their applications. *Biometrika*, 36(Pt. 1-2):202–232.

Pecorelli, S. (2009). Revised FIGO staging for carcinoma of the vulva, cervix, and endometrium. *International Journal of Gynaecology and Obstetrics: The Official Organ of the International Federation of Gynaecology and Obstetrics*, 105(2):103–104.

Peltomäki, P., Aaltonen, L. A., Sistonen, P., Pylkkänen, L., Mecklin, J. P., Järvinen, H., Green, J. S., Jass, J. R., Weber, J. L., and Leach, F. S. (1993). Genetic mapping of a locus predisposing to human colorectal cancer. *Science (New York, N.Y.)*, 260(5109):810–812.

Peltomäki, P. and Vasen, H. (2004). Mutations associated with HNPCC predisposition – Update of ICG-HNPCC/INSiGHT mutation database. *Disease Markers*, 20(4-5):269–276.

Peters, U., Hutter, C. M., Hsu, L., Schumacher, F. R., Conti, D. V., Carlson, C. S., Edlund, C. K., Haile, R. W., Gallinger, S., Zanke, B. W., Lemire, M., Rangrej, J., Vijayaraghavan, R., Chan, A. T., Hazra, A., Hunter, D. J., Ma, J., Fuchs, C. S., Giovannucci, E. L., Kraft, P., Liu, Y., Chen, L., Jiao, S., Makar, K. W., Taverna, D., Gruber, S. B., Rennert, G., Moreno, V., Ulrich, C. M., Woods, M. O., Green, R. C., Parfrey, P. S., Prentice, R. L., Kooperberg, C., Jackson, R. D., Lacroix, A. Z., Caan, B. J., Hayes, R. B., Berndt, S. I., Chanock, S. J., Schoen, R. E., Chang-Claude, J., Hoffmeister, M., Brenner, H., Frank, B., Bézieau, S., Küry, S., Slattery, M. L., Hopper, J. L., Jenkins, M. A., Le Marchand, L., Lindor, N. M., Newcomb, P. A., Seminara, D., Hudson, T. J., Duggan, D. J., Potter, J. D., and Casey, G. (2012). Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Human Genetics*, 131(2):217–234.

Petersen, G. M., Amundadottir, L., Fuchs, C. S., Kraft, P., Stolzenberg-Solomon, R. Z., Jacobs, K. B., Arslan, A. A., Bueno-de Mesquita, H. B., Gallinger, S., Gross, M., Helzlsouer, K., Holly, E. A., Jacobs, E. J., Klein, A. P., LaCroix, A., Li, D., Mandelson, M. T., Olson, S. H., Risch, H. A., Zheng, W., Albanes, D., Bamlet, W. R., Berg, C. D., Boutron-Ruault, M.-C., Buring, J. E., Bracci, P. M., Canzian, F., Clipp, S., Cotterchio, M., de Andrade, M., Duell, E. J., Gaziano, J. M., Giovannucci, E. L., Goggins, M., Hallmans, G., Hankinson, S. E., Hassan, M., Howard, B., Hunter, D. J., Hutchin-

son, A., Jenab, M., Kaaks, R., Kooperberg, C., Krogh, V., Kurtz, R. C., Lynch, S. M., McWilliams, R. R., Mendelsohn, J. B., Michaud, D. S., Parikh, H., Patel, A. V., Peeters, P. H. M., Rajkovic, A., Riboli, E., Rodriguez, L., Seminara, D., Shu, X.-O., Thomas, G., Tjønneland, A., Tobias, G. S., Trichopoulos, D., Van Den Eeden, S. K., Virtamo, J., Wactawski-Wende, J., Wang, Z., Wolpin, B. M., Yu, H., Yu, K., Zeleniuch-Jacquotte, A., Fraumeni, J. F., Hoover, R. N., Hartge, P., and Chanock, S. J. (2010). A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature Genetics*, 42(3):224–228.

Pharoah, P. D. P., Tsai, Y.-Y., Ramus, S. J., Phelan, C. M., Goode, E. L., Lawrenson, K., Buckley, M., Fridley, B. L., Tyrer, J. P., Shen, H., Weber, R., Karevan, R., Larson, M. C., Song, H., Tessier, D. C., Bacot, F., Vincent, D., Cunningham, J. M., Dennis, J., Dicks, E., Australian Cancer Study, Australian Ovarian Cancer Study Group, Aben, K. K., Anton-Culver, H., Antonenkova, N., Armasu, S. M., Baglietto, L., Bandera, E. V., Beckmann, M. W., Birrer, M. J., Bloom, G., Bogdanova, N., Brenton, J. D., Brinton, L. A., Brooks-Wilson, A., Brown, R., Butzow, R., Campbell, I., Carney, M. E., Carvalho, R. S., Chang-Claude, J., Chen, Y. A., Chen, Z., Chow, W.-H., Cicek, M. S., Coetzee, G., Cook, L. S., Cramer, D. W., Cybulski, C., Dansonka-Mieszkowska, A., Despierre, E., Doherty, J. A., Dörk, T., du Bois, A., Dürst, M., Eccles, D., Edwards, R., Ekici, A. B., Fasching, P. A., Fenstermacher, D., Flanagan, J., Gao, Y.-T., Garcia-Closas, M., Gentry-Maharaj, A., Giles, G., Gjyshi, A., Gore, M., Gronwald, J., Guo, Q., Halle, M. K., Harter, P., Hein, A., Heitz, F., Hillemanns, P., Hoatlin, M., Høgdall, E., Høgdall, C. K., Hosono, S., Jakubowska, A., Jensen, A., Kalli, K. R., Karlan, B. Y., Kelemen, L. E., Kiemeney, L. A., Kjaer, S. K., Konecny, G. E., Krakstad, C., Kupryjanczyk, J., Lambrechts, D., Lambrechts, S., Le, N. D., Lee, N., Lee, J., Leminen, A., Lim, B. K., Lissowska, J., Lubiński, J., Lundvall, L., Lurie, G., Massuger, L. F. A. G., Matsuo, K., McGuire, V., McLaughlin, J. R., Menon, U., Modugno, F., Moysich, K. B., Nakanishi, T., Narod, S. A., Ness, R. B., Nevanlinna, H., Nickels, S., Noushmehr, H., Odunsi, K., Olson, S., Orlow, I., Paul, J., Pejovic, T., Pelttari, L. M., Permuth-Wey, J., Pike, M. C., Poole, E. M., Qu, X., Risch, H. A., Rodriguez-Rodriguez, L., Rossing, M. A., Rudolph,

A., Runnebaum, I., Rzepecka, I. K., Salvesen, H. B., Schwaab, I., Severi, G., Shen, H., Shridhar, V., Shu, X.-O., Sieh, W., Southey, M. C., Spellman, P., Tajima, K., Teo, S.-H., Terry, K. L., Thompson, P. J., Timorek, A., Tworoger, S. S., van Altena, A. M., van den Berg, D., Vergote, I., Vierkant, R. A., Vitonis, A. F., Wang-Gohrke, S., Wentzensen, N., Whittemore, A. S., Wik, E., Winterhoff, B., Woo, Y. L., Wu, A. H., Yang, H. P., Zheng, W., Ziogas, A., Zulkifli, F., Goodman, M. T., Hall, P., Easton, D. F., Pearce, C. L., Berchuck, A., Chenevix-Trench, G., Iversen, E., Monteiro, A. N. A., Gayther, S. A., Schildkraut, J. M., and Sellers, T. A. (2013). GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nature Genetics*, 45(4):362–370, 370e1–2.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–772.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121.

Potischman, N., Swanson, C. A., Siiteri, P., and Hoover, R. N. (1996). Reversal of relation between body mass and endogenous estrogen concentrations with menopausal status. *Journal of the National Cancer Institute*, 88(11):756–758.

Power, C. and Elliott, J. (2006). Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology*, 35(1):34–41.

Prescott, J., Thompson, D. J., Kraft, P., Chanock, S. J., Audley, T., Brown, J., Leyland, J., Folkerd, E., Doody, D., Hankinson, S. E., Hunter, D. J., Jacobs, K. B., Dowsett, M., Cox, D. G., Easton, D. F., and De Vivo, I. (2012). Genome-wide association study of circulating estradiol, testosterone, and sex hormone-binding globulin in postmenopausal women. *PloS One*, 7(6):e37815.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909. Available from: `http://www.nature.com/ng/journal/v38/n8/abs/ng1847.html`.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3):559–575.

Quaresma, M., Coleman, M. P., and Rachet, B. (2014). 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study. *Lancet*.

Rafnar, T., Sulem, P., Stacey, S. N., Geller, F., Gudmundsson, J., Sigurdsson, A., Jakobsdottir, M., Helgadottir, H., Thorlacius, S., Aben, K. K. H., Blöndal, T., Thorgeirsson, T. E., Thorleifsson, G., Kristjansson, K., Thorisdottir, K., Ragnarsson, R., Sigurgeirsson, B., Skuladottir, H., Gudbjartsson, T., Isaksson, H. J., Einarsson, G. V., Benediktsdottir, K. R., Agnarsson, B. A., Olafsson, K., Salvarsdottir, A., Bjarnason, H., Asgeirsdottir, M., Kristinsson, K. T., Matthiasdottir, S., Sveinsdottir, S. G., Polidoro, S., Höiom, V., Botella-Estrada, R., Hemminki, K., Rudnai, P., Bishop, D. T., Campagna, M., Kellen, E., Zeegers, M. P., de Verdier, P., Ferrer, A., Isla, D., Vidal, M. J., Andres, R., Saez, B., Juberias, P., Banzo, J., Navarrete, S., Tres, A., Kan, D., Lindblom, A., Gurzau, E., Koppova, K., de Vegt, F., Schalken, J. A., van der Heijden, H. F. M., Smit, H. J., Termeer, R. A., Oosterwijk, E., van Hooij, O., Nagore, E., Porru, S., Steineck, G., Hansson, J., Buntinx, F., Catalona, W. J., Matullo, G., Vineis, P., Kiltie, A. E., Mayordomo, J. I., Kumar, R., Kiemeney, L. A., Frigge, M. L., Jonsson, T., Saemundsson, H., Barkardottir, R. B., Jonsson, E., Jonsson, S., Olafsson, J. H., Gulcher, J. R., Masson, G., Gudbjartsson, D. F., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2009). Sequence variants at the TERT-CLPTM1l locus associate with many cancer types. *Nature Genetics*, 41(2):221–227.

Rahmioglu, N., Macgregor, S., Drong, A. W., Harris, H. R., Randall, J. C., Prokopenko, I., International Endogene Consortium (IEC), The GIANT Consortium, Nyholt, D. R., Morris, A. P., Montgomery, G. W., Missmer, S. A., Lindgren, C. M., and Zondervan, K. T. (2015). Genome-wide enrichment analysis between endometriosis and obesity-related traits reveals novel susceptibility loci. *Human Molecular Genetics*, 24(4):1185–1199.

Raymond, V. M., Everett, J. N., Furtado, L. V., Gustafson, S. L., Jungbluth, C. R., Gruber, S. B., Hammer, G. D., Stoffel, E. M., Greenson, J. K., Giordano, T. J., and Else, T. (2013). Adrenocortical carcinoma is a lynch syndrome-associated cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 31(24):3012–3018.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Resnick, K. E., Hampel, H., Fishel, R., and Cohn, D. E. (2009). Current and emerging trends in Lynch syndrome identification in women with endometrial cancer. *Gynecologic Oncology*, 114(1):128–134.

Risinger, J. I., Hayes, K., Maxwell, G. L., Carney, M. E., Dodge, R. K., Barrett, J. C., and Berchuck, A. (1998). PTEN mutation in endometrial cancers is associated with favorable clinical and pathologic characteristics. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 4(12):3005–3010.

Romanowicz-Makowska, H., Smolarz, B., Połać, I., and Sporny, S. (2012). Single nucleotide polymorphisms of RAD51 G135c, XRCC2 Arg188his and XRCC3 Thr241met homologous recombination repair genes and the risk of sporadic endometrial cancer in Polish women. *The Journal of Obstetrics and Gynaecology Research*, 38(6):918–924.

Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature Reviews*

Genetics, 11(5):356–366. Available from: `http://www.nature.com/nrg/journal/v11/n5/abs/nrg2760.html`.

Rothman, N., Garcia-Closas, M., Chatterjee, N., Malats, N., Wu, X., Figueroa, J. D., Real, F. X., Van Den Berg, D., Matullo, G., Baris, D., Thun, M., Kiemeney, L. A., Vineis, P., De Vivo, I., Albanes, D., Purdue, M. P., Rafnar, T., Hildebrandt, M. A. T., Kiltie, A. E., Cussenot, O., Golka, K., Kumar, R., Taylor, J. A., Mayordomo, J. I., Jacobs, K. B., Kogevinas, M., Hutchinson, A., Wang, Z., Fu, Y.-P., Prokunina-Olsson, L., Burdett, L., Yeager, M., Wheeler, W., Tardón, A., Serra, C., Carrato, A., García-Closas, R., Lloreta, J., Johnson, A., Schwenn, M., Karagas, M. R., Schned, A., Andriole, G., Grubb, R., Black, A., Jacobs, E. J., Diver, W. R., Gapstur, S. M., Weinstein, S. J., Virtamo, J., Cortessis, V. K., Gago-Dominguez, M., Pike, M. C., Stern, M. C., Yuan, J.-M., Hunter, D. J., McGrath, M., Dinney, C. P., Czerniak, B., Chen, M., Yang, H., Vermeulen, S. H., Aben, K. K., Witjes, J. A., Makkinje, R. R., Sulem, P., Besenbacher, S., Stefansson, K., Riboli, E., Brennan, P., Panico, S., Navarro, C., Allen, N. E., Bueno-de Mesquita, H. B., Trichopoulos, D., Caporaso, N., Landi, M. T., Canzian, F., Ljungberg, B., Tjonneland, A., Clavel-Chapelon, F., Bishop, D. T., Teo, M. T. W., Knowles, M. A., Guarrera, S., Polidoro, S., Ricceri, F., Sacerdote, C., Allione, A., Cancel-Tassin, G., Selinski, S., Hengstler, J. G., Dietrich, H., Fletcher, T., Rudnai, P., Gurzau, E., Koppova, K., Bolick, S. C. E., Godfrey, A., Xu, Z., Sanz-Velez, J. I., D García-Prats, M., Sanchez, M., Valdivia, G., Porru, S., Benhamou, S., Hoover, R. N., Fraumeni, J. F., Silverman, D. T., and Chanock, S. J. (2010). A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature Genetics*, 42(11):978–984.

Sakoda, L. C., Jorgenson, E., and Witte, J. S. (2013). Turning of COGS moves forward findings for hormonally mediated cancers. *Nature Genetics*, 45(4):345–348.

Salani, R., Backes, F. J., Fung, M. F. K., Holschneider, C. H., Parker, L. P., Bristow, R. E., and Goff, B. A. (2011). Posttreatment surveillance and diagnosis of recurrence in women with gynecologic malignancies: Society of Gynecologic Oncologists recommendations. *American Journal of Obstetrics and Gynecology*, 204(6):466–478.

Salvesen, H. B., Carter, S. L., Mannelqvist, M., Dutt, A., Getz, G., Stefansson, I. M., Raeder, M. B., Sos, M. L., Engelsen, I. B., Trovik, J., Wik, E., Greulich, H., Bø, T. H., Jonassen, I., Thomas, R. K., Zander, T., Garraway, L. A., Oyan, A. M., Sellers, W. R., Kalland, K. H., Meyerson, M., Akslen, L. A., and Beroukhim, R. (2009). Integrated genomic profiling of endometrial carcinoma associates aggressive tumors with indicators of PI3 kinase activation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(12):4834–4839.

Samulak, D., Romanowicz-Makowska, H., Smolarz, B., Kulig, A., and Sporny, S. (2011). Association between Arg399gln polymorphism of X-ray repair cross-complementing 1 (XRCC1) gene and sporadic endometrial cancer in the Polish population. *European Journal of Gynaecological Oncology*, 32(5):491–495.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–448.

Schonfeld, S. J., Hartge, P., Pfeiffer, R. M., Freedman, D. M., Greenlee, R. T., Linet, M. S., Park, Y., Schairer, C., Visvanathan, K., and Lacey, J. V. (2013). An aggregated analysis of hormonal factors and endometrial cancer risk by parity. *Cancer*, 119(7):1393–1401.

Senter, L., Clendenning, M., Sotamaa, K., Hampel, H., Green, J., Potter, J. D., Lindblom, A., Lagerstedt, K., Thibodeau, S. N., Lindor, N. M., Young, J., Winship, I., Dowty, J. G., White, D. M., Hopper, J. L., Baglietto, L., Jenkins, M. A., and de la Chapelle, A. (2008). The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. *Gastroenterology*, 135(2):419–428.

Setiawan, V. W., Doherty, J. A., Shu, X.-O., Akbari, M. R., Chen, C., De Vivo, I., Demichele, A., Garcia-Closas, M., Goodman, M. T., Haiman, C. A., Hankinson, S. E., Henderson, B. E., Horn-Ross, P. L., Lacey, J. V., Le Marchand, L., Levine, D. A., Liang, X., Lissowska, J., Lurie, G., McGrath, M., Narod, S. A., Rebbeck, T. R., Ursin, G., Weiss, N. S., Xiang, Y.-B., Yang, H. P., Zheng, W., and Olson, S. H. (2009). Two estrogen-related variants in CYP19a1 and endometrial cancer risk: a pooled analysis in

the Epidemiology of Endometrial Cancer Consortium. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, 18(1):242–247.

Shen, H., Fridley, B. L., Song, H., Lawrenson, K., Cunningham, J. M., Ramus, S. J., Cicek, M. S., Tyrer, J., Stram, D., Larson, M. C., Köbel, M., PRACTICAL Consortium, Ziogas, A., Zheng, W., Yang, H. P., Wu, A. H., Wozniak, E. L., Woo, Y. L., Winterhoff, B., Wik, E., Whittemore, A. S., Wentzensen, N., Weber, R. P., Vitonis, A. F., Vincent, D., Vierkant, R. A., Vergote, I., Van Den Berg, D., Van Altena, A. M., Tworoger, S. S., Thompson, P. J., Tessier, D. C., Terry, K. L., Teo, S.-H., Templeman, C., Stram, D. O., Southey, M. C., Sieh, W., Siddiqui, N., Shvetsov, Y. B., Shu, X.-O., Shridhar, V., Wang-Gohrke, S., Severi, G., Schwaab, I., Salvesen, H. B., Rzepecka, I. K., Runnebaum, I. B., Rossing, M. A., Rodriguez-Rodriguez, L., Risch, H. A., Renner, S. P., Poole, E. M., Pike, M. C., Phelan, C. M., Pelttari, L. M., Pejovic, T., Paul, J., Orlow, I., Omar, S. Z., Olson, S. H., Odunsi, K., Nickels, S., Nevanlinna, H., Ness, R. B., Narod, S. A., Nakanishi, T., Moysich, K. B., Monteiro, A. N. A., Moes-Sosnowska, J., Modugno, F., Menon, U., McLaughlin, J. R., McGuire, V., Matsuo, K., Adenan, N. A. M., Massuger, L. F. A. G., Lurie, G., Lundvall, L., Lubiński, J., Lissowska, J., Levine, D. A., Leminen, A., Lee, A. W., Le, N. D., Lambrechts, S., Lambrechts, D., Kupryjanczyk, J., Krakstad, C., Konecny, G. E., Kjaer, S. K., Kiemeney, L. A., Kelemen, L. E., Keeney, G. L., Karlan, B. Y., Karevan, R., Kalli, K. R., Kajiyama, H., Ji, B.-T., Jensen, A., Jakubowska, A., Iversen, E., Hosono, S., Høgdall, C. K., Høgdall, E., Hoatlin, M., Hillemanns, P., Heitz, F., Hein, R., Harter, P., Halle, M. K., Hall, P., Gronwald, J., Gore, M., Goodman, M. T., Giles, G. G., Gentry-Maharaj, A., Garcia-Closas, M., Flanagan, J. M., Fasching, P. A., Ekici, A. B., Edwards, R., Eccles, D., Easton, D. F., Dürst, M., du Bois, A., Dörk, T., Doherty, J. A., Despierre, E., Dansonka-Mieszkowska, A., Cybulski, C., Cramer, D. W., Cook, L. S., Chen, X., Charbonneau, B., Chang-Claude, J., Campbell, I., Butzow, R., Bunker, C. H., Brueggmann, D., Brown, R., Brooks-Wilson, A., Brinton, L. A., Bogdanova, N., Block, M. S., Benjamin, E., Beesley, J., Beckmann, M. W., Bandera, E. V., Baglietto, L., Bacot, F., Armasu, S. M., Antonenkova, N., Anton-Culver,

H., Aben, K. K., Liang, D., Wu, X., Lu, K., Hildebrandt, M. A. T., Australian Ovarian Cancer Study Group, Australian Cancer Study, Schildkraut, J. M., Sellers, T. A., Huntsman, D., Berchuck, A., Chenevix-Trench, G., Gayther, S. A., Pharoah, P. D. P., Laird, P. W., Goode, E. L., and Pearce, C. L. (2013). Epigenetic analysis leads to identification of HNF1b as a subtype-specific susceptibility gene for ovarian cancer. *Nature Communications*, 4:1628.

Sherman, M. E. (2000). Theories of endometrial carcinogenesis: a multidisciplinary approach. *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, 13(3):295–308.

Shete, S., Hosking, F. J., Robertson, L. B., Dobbins, S. E., Sanson, M., Malmer, B., Simon, M., Marie, Y., Boisselier, B., Delattre, J.-Y., Hoang-Xuan, K., El Hallani, S., Idbaih, A., Zelenika, D., Andersson, U., Henriksson, R., Bergenheim, A. T., Feychting, M., Lönn, S., Ahlbom, A., Schramm, J., Linnebank, M., Hemminki, K., Kumar, R., Hepworth, S. J., Price, A., Armstrong, G., Liu, Y., Gu, X., Yu, R., Lau, C., Schoemaker, M., Muir, K., Swerdlow, A., Lathrop, M., Bondy, M., and Houlston, R. S. (2009). Genome-wide association study identifies five susceptibility loci for glioma. *Nature Genetics*, 41(8):899–904.

Shi, H., Zhang, Z., Wang, X., Liu, S., and Teng, C. T. (1999). Isolation and characterization of a gene encoding human Kruppel-like factor 5 (IKLF): binding to the CAAT/GT box of the mouse lactoferrin gene promoter. *Nucleic Acids Research*, 27(24):4807–4815. Available from: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC148782/`.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., and Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics*, 12(7):771–776. Available from: `http://hmg.oxfordjournals.org/content/12/7/771`.

Shoji, K., Oda, K., Nakagawa, S., Hosokawa, S., Nagae, G., Uehara, Y., Sone, K., Miyamoto, Y., Hiraike, H., Hiraike-Wada, O., Nei, T., Kawana, K., Kuramoto, H., Aburatani, H., Yano, T., and Taketani, Y. (2009). The oncogenic mutation in the pleck-

strin homology domain of AKT1 in endometrial carcinomas. *British Journal of Cancer*, 101(1):145–148.

Siegel, R., Ma, J., Zou, Z., and Jemal, A. (2014). Cancer statistics, 2014. *CA: a cancer journal for clinicians*, 64(1):9–29.

Simmen, R. C. M., Pabona, J. M. P., Velarde, M. C., Simmons, C., Rahal, O., and Simmen, F. A. (2010). The emerging role of Krüppel-like factors in endocrine-responsive cancers of female reproductive tissues. *The Journal of Endocrinology*, 204(3):223–231.

Simon, M., Giot, L., and Faye, G. (1991). The 3' to 5' exonuclease activity located in the DNA polymerase delta subunit of Saccharomyces cerevisiae is required for accurate replication. *The EMBO journal*, 10(8):2165–2170.

Simpkins, S. B., Bocker, T., Swisher, E. M., Mutch, D. G., Gersell, D. J., Kovatich, A. J., Palazzo, J. P., Fishel, R., and Goodfellow, P. J. (1999). MLH1 promoter methylation and gene silencing is the primary cause of microsatellite instability in sporadic endometrial cancers. *Human Molecular Genetics*, 8(4):661–666.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J. F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *American Journal of Human Genetics*, 89(5):607–618.

Slomovitz, B. M. and Coleman, R. L. (2012). The PI3k/AKT/mTOR pathway as a therapeutic target in endometrial cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 18(21):5856–5864.

Smotkin, D., Nevadunsky, N. S., Harris, K., Einstein, M. H., Yu, Y., and Goldberg, G. L. (2012). Histopathologic differences account for racial disparity in uterine cancer survival. *Gynecologic Oncology*, 127(3):616–619.

Smyrk, T. C., Watson, P., Kaul, K., and Lynch, H. T. (2001). Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma. *Cancer*, 91(12):2417–2422.

Sobczuk, A., Poplawski, T., and Blasiak, J. (2012). Polymorphisms of DNA repair genes in endometrial cancer. *Pathology oncology research: POR*, 18(4):1015–1020.

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B.-K., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R. M., Shibata, Y., Showers, K. A., Simon, J. M., Vales, T., Wang, T., Winter, D., Zhang, Z., Clarke, N. D., Birney, E., Iyer, V. R., Crawford, G. E., Lieb, J. D., and Furey, T. S. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, 21(10):1757–1767. Available from: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3202292/`.

Spurdle, A. B., Thompson, D. J., Ahmed, S., Ferguson, K., Healey, C. S., O'Mara, T., Walker, L. C., Montgomery, S. B., Dermitzakis, E. T., Australian National Endometrial Cancer Study Group, Fahey, P., Montgomery, G. W., Webb, P. M., Fasching, P. A., Beckmann, M. W., Ekici, A. B., Hein, A., Lambrechts, D., Coenegrachts, L., Vergote, I., Amant, F., Salvesen, H. B., Trovik, J., Njolstad, T. S., Helland, H., Scott, R. J., Ashton, K., Proietto, T., Otton, G., National Study of Endometrial Cancer Genetics Group, Tomlinson, I., Gorman, M., Howarth, K., Hodgson, S., Garcia-Closas, M., Wentzensen, N., Yang, H., Chanock, S., Hall, P., Czene, K., Liu, J., Li, J., Shu, X.-O., Zheng, W., Long, J., Xiang, Y.-B., Shah, M., Morrison, J., Michailidou, K., Pharoah, P. D., Dunning, A. M., and Easton, D. F. (2011). Genome-wide association study identifies a common variant associated with risk of endometrial cancer. *Nature Genetics*, 43(5):451–454.

Stacey, S. N., Sulem, P., Masson, G., Gudjonsson, S. A., Thorleifsson, G., Jakobsdottir, M., Sigurdsson, A., Gudbjartsson, D. F., Sigurgeirsson, B., Benediktsdottir, K. R., Thorisdottir, K., Ragnarsson, R., Scherer, D., Hemminki, K., Rudnai, P., Gurzau, E., Koppova, K., Botella-Estrada, R., Soriano, V., Juberias, P., Saez, B., Gilaberte, Y., Fuentelsaz,

V., Corredera, C., Grasa, M., Höiom, V., Lindblom, A., Bonenkamp, J. J., van Rossum, M. M., Aben, K. K. H., de Vries, E., Santinami, M., Di Mauro, M. G., Maurichi, A., Wendt, J., Hochleitner, P., Pehamberger, H., Gudmundsson, J., Magnusdottir, D. N., Gretarsdottir, S., Holm, H., Steinthorsdottir, V., Frigge, M. L., Blondal, T., Saemundsdottir, J., Bjarnason, H., Kristjansson, K., Bjornsdottir, G., Okamoto, I., Rivoltini, L., Rodolfo, M., Kiemeney, L. A., Hansson, J., Nagore, E., Mayordomo, J. I., Kumar, R., Karagas, M. R., Nelson, H. H., Gulcher, J. R., Rafnar, T., Thorsteinsdottir, U., Olafsson, J. H., Kong, A., and Stefansson, K. (2009). New common variants affecting susceptibility to basal cell carcinoma. *Nature Genetics*, 41(8):909–914.

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., Li, Y., Kurreeman, F. A. S., Zhernakova, A., Hinks, A., Guiducci, C., Chen, R., Alfredsson, L., Amos, C. I., Ardlie, K. G., BIRAC Consortium, Barton, A., Bowes, J., Brouwer, E., Burtt, N. P., Catanese, J. J., Coblyn, J., Coenen, M. J. H., Costenbader, K. H., Criswell, L. A., Crusius, J. B. A., Cui, J., de Bakker, P. I. W., De Jager, P. L., Ding, B., Emery, P., Flynn, E., Harrison, P., Hocking, L. J., Huizinga, T. W. J., Kastner, D. L., Ke, X., Lee, A. T., Liu, X., Martin, P., Morgan, A. W., Padyukov, L., Posthumus, M. D., Radstake, T. R. D. J., Reid, D. M., Seielstad, M., Seldin, M. F., Shadick, N. A., Steer, S., Tak, P. P., Thomson, W., van der Helm-van Mil, A. H. M., van der Horst-Bruinsma, I. E., van der Schoot, C. E., van Riel, P. L. C. M., Weinblatt, M. E., Wilson, A. G., Wolbink, G. J., Wordsworth, B. P., YEAR Consortium, Wijmenga, C., Karlson, E. W., Toes, R. E. M., de Vries, N., Begovich, A. B., Worthington, J., Siminovitch, K. A., Gregersen, P. K., Klareskog, L., and Plenge, R. M. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics*, 42(6):508–514.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., and Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics*, 39(10):1217–1224.

Sun, H. D., Malabunga, M., Tonra, J. R., DiRenzo, R., Carrick, F. E., Zheng, H., Berthoud,

H.-R., McGuinness, O. P., Shen, J., Bohlen, P., Leibel, R. L., and Kussie, P. (2007). Monoclonal antibody antagonists of hypothalamic FGFR1 cause potent but reversible hypophagia and weight loss in rodents and monkeys. *American Journal of Physiology. Endocrinology and Metabolism*, 292(3):E964–976.

Sun, X., Zhang, L., Xie, H., Wan, H., Magella, B., Whitsett, J. A., and Dey, S. K. (2012). Kruppel-like factor 5 (KLF5) is critical for conferring uterine receptivity to implantation. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1145–1150.

Tabor, H. K., Risch, N. J., and Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, 3(5):391–397. Available from: `http://www.nature.com/nrg/journal/v3/n5/full/nrg796.html`.

Tan, M.-H., Mester, J. L., Ngeow, J., Rybicki, L. A., Orloff, M. S., and Eng, C. (2012). Lifetime cancer risks in individuals with germline PTEN mutations. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 18(2):400–407.

Tang, N.-P., Li, H., Qiu, Y.-L., Zhou, G.-M., and Ma, J. (2009). Tea consumption and risk of endometrial cancer: a metaanalysis. *American Journal of Obstetrics and Gynecology*, 201(6):605.e1–8.

Tang, W., He, X., Chan, Y., and Luo, Y. (2012). Lack of association between p53 codon 72 polymorphism and endometrial cancer: a meta-analysis. *Cancer Epidemiology*, 36(3):e153–157.

Tenesa, A., Farrington, S. M., Prendergast, J. G. D., Porteous, M. E., Walker, M., Haq, N., Barnetson, R. A., Theodoratou, E., Cetnarskyj, R., Cartwright, N., Semple, C., Clark, A. J., Reid, F. J. L., Smith, L. A., Kavoussanakis, K., Koessler, T., Pharoah, P. D. P., Buch, S., Schafmayer, C., Tepel, J., Schreiber, S., Völzke, H., Schmidt, C. O., Hampe, J., Chang-Claude, J., Hoffmeister, M., Brenner, H., Wilkening, S., Canzian,

F., Capella, G., Moreno, V., Deary, I. J., Starr, J. M., Tomlinson, I. P. M., Kemp, Z., Howarth, K., Carvajal-Carmona, L., Webb, E., Broderick, P., Vijayakrishnan, J., Houlston, R. S., Rennert, G., Ballinger, D., Rozek, L., Gruber, S. B., Matsuda, K., Kidokoro, T., Nakamura, Y., Zanke, B. W., Greenwood, C. M. T., Rangrej, J., Kustra, R., Montpetit, A., Hudson, T. J., Gallinger, S., Campbell, H., and Dunlop, M. G. (2008). Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature Genetics*, 40(5):631–637.

Terada, K. Y., Black, M., Terada, L. H., Davis, J., and Shimizu, D. M. (2013). Survival of endometrial cancer patients with lymphatic invasion and deficient mismatch repair expression. *Gynecologic Oncology*, 129(1):188–192.

Terry, P., Baron, J. A., Weiderpass, E., Yuen, J., Lichtenstein, P., and Nyrén, O. (1999). Lifestyle and endometrial cancer risk: a cohort study from the Swedish Twin Registry. *International Journal of Cancer. Journal International Du Cancer*, 82(1):38–42.

Timmermans, A., Opmeer, B. C., Khan, K. S., Bachmann, L. M., Epstein, E., Clark, T. J., Gupta, J. K., Bakour, S. H., van den Bosch, T., van Doorn, H. C., Cameron, S. T., Giusa, M. G., Dessole, S., Dijkhuizen, F. P. H. L. J., Ter Riet, G., and Mol, B. W. J. (2010). Endometrial thickness measurement for detecting endometrial cancer in women with postmenopausal bleeding: a systematic review and meta-analysis. *Obstetrics and Gynecology*, 116(1):160–167.

Timpson, N. J., Walter, K., Min, J. L., Tachmazidou, I., Malerba, G., Shin, S.-Y., Chen, L., Futema, M., Southam, L., Iotchkova, V., Cocca, M., Huang, J., Memari, Y., McCarthy, S., Danecek, P., Muddyman, D., Mangino, M., Menni, C., Perry, J. R. B., Ring, S. M., Gaye, A., Dedoussis, G., Farmaki, A.-E., Burton, P., Talmud, P. J., Gambaro, G., Spector, T. D., Smith, G. D., Durbin, R., Richards, J. B., Humphries, S. E., Zeggini, E., Soranzo, N., UK1OK Consortium Members, and UK1OK Consortium Members (2014). A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nature Communications*, 5:4871.

Tomlinson, I. P. M., Carvajal-Carmona, L. G., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., Palles, C., Broderick, P., Jaeger, E. E. M., Farrington, S., Lewis, A., Prendergast, J. G. D., Pittman, A. M., Theodoratou, E., Olver, B., Walker, M., Penegar, S., Barclay, E., Whiffin, N., Martin, L., Ballereau, S., Lloyd, A., Gorman, M., Lubbe, S., COGENT Consortium, CORGI Collaborators, EPICOLON Consortium, Howie, B., Marchini, J., Ruiz-Ponte, C., Fernandez-Rozadilla, C., Castells, A., Carracedo, A., Castellvi-Bel, S., Duggan, D., Conti, D., Cazier, J.-B., Campbell, H., Sieber, O., Lipton, L., Gibbs, P., Martin, N. G., Montgomery, G. W., Young, J., Baird, P. N., Gallinger, S., Newcomb, P., Hopper, J., Jenkins, M. A., Aaltonen, L. A., Kerr, D. J., Cheadle, J., Pharoah, P., Casey, G., Houlston, R. S., and Dunlop, M. G. (2011). Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS genetics*, 7(6):e1002105.

Tomlinson, I. P. M., Webb, E., Carvajal-Carmona, L., Broderick, P., Howarth, K., Pittman, A. M., Spain, S., Lubbe, S., Walther, A., Sullivan, K., Jaeger, E., Fielding, S., Rowan, A., Vijayakrishnan, J., Domingo, E., Chandler, I., Kemp, Z., Qureshi, M., Farrington, S. M., Tenesa, A., Prendergast, J. G. D., Barnetson, R. A., Penegar, S., Barclay, E., Wood, W., Martin, L., Gorman, M., Thomas, H., Peto, J., Bishop, D. T., Gray, R., Maher, E. R., Lucassen, A., Kerr, D., Evans, D. G. R., CORGI Consortium, Schafmayer, C., Buch, S., Völzke, H., Hampe, J., Schreiber, S., John, U., Koessler, T., Pharoah, P., van Wezel, T., Morreau, H., Wijnen, J. T., Hopper, J. L., Southey, M. C., Giles, G. G., Severi, G., Castellví-Bel, S., Ruiz-Ponte, C., Carracedo, A., Castells, A., EPICOLON Consortium, Försti, A., Hemminki, K., Vodicka, P., Naccarati, A., Lipton, L., Ho, J. W. C., Cheng, K. K., Sham, P. C., Luk, J., Agúndez, J. A. G., Ladero, J. M., de la Hoya, M., Caldés, T., Niittymäki, I., Tuupanen, S., Karhu, A., Aaltonen, L., Cazier, J.-B., Campbell, H., Dunlop, M. G., and Houlston, R. S. (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature Genetics*, 40(5):623–630.

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S.,

Ghoussaini, M., Hines, S., Healey, C. S., Hughes, D., Warren-Perry, M., Tapper, W., Eccles, D., Evans, D. G., Breast Cancer Susceptibility Collaboration (UK), Hooning, M., Schutte, M., van den Ouweland, A., Houlston, R., Ross, G., Langford, C., Pharoah, P. D. P., Stratton, M. R., Dunning, A. M., Rahman, N., and Easton, D. F. (2010a). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature Genetics*, 42(6):504–507.

Turnbull, C., Rapley, E. A., Seal, S., Pernet, D., Renwick, A., Hughes, D., Ricketts, M., Linger, R., Nsengimana, J., Deloukas, P., Huddart, R. A., Bishop, D. T., Easton, D. F., Stratton, M. R., Rahman, N., and UK Testicular Cancer Collaboration (2010b). Variants near DMRT1, TERT and ATF7ip are associated with testicular germ cell cancer. *Nature Genetics*, 42(7):604–607.

Turner, N. and Grose, R. (2010). Fibroblast growth factor signalling: from development to cancer. *Nature Reviews. Cancer*, 10(2):116–129.

van Gool, I. C., Eggink, F. A., Freeman-Mills, L., Stelloo, E., Marchi, E., de Bruyn, M., Palles, C., Nout, R. A., de Kroon, C. D., Osse, E. M., Klenerman, P., Creutzberg, C. L., Tomlinson, I. P. M., Smit, V. T. H. B. M., Nijman, H. W., Bosse, T., and Church, D. N. (2015). POLE proofreading mutations elicit an anti-tumor immune response in endometrial cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*.

Wain, L. V., Verwoert, G. C., O'Reilly, P. F., Shi, G., Johnson, T., Johnson, A. D., Bochud, M., Rice, K. M., Henneman, P., Smith, A. V., Ehret, G. B., Amin, N., Larson, M. G., Mooser, V., Hadley, D., Dörr, M., Bis, J. C., Aspelund, T., Esko, T., Janssens, A. C. J. W., Zhao, J. H., Heath, S., Laan, M., Fu, J., Pistis, G., Luan, J., Arora, P., Lucas, G., Pirastu, N., Pichler, I., Jackson, A. U., Webster, R. J., Zhang, F., Peden, J. F., Schmidt, H., Tanaka, T., Campbell, H., Igl, W., Milaneschi, Y., Hottenga, J.-J., Vitart, V., Chasman, D. I., Trompet, S., Bragg-Gresham, J. L., Alizadeh, B. Z., Chambers, J. C., Guo, X., Lehtimäki, T., Kühnel, B., Lopez, L. M., Polašek, O., Boban,

M., Nelson, C. P., Morrison, A. C., Pihur, V., Ganesh, S. K., Hofman, A., Kundu, S., Mattace-Raso, F. U. S., Rivadeneira, F., Sijbrands, E. J. G., Uitterlinden, A. G., Hwang, S.-J., Vasan, R. S., Wang, T. J., Bergmann, S., Vollenweider, P., Waeber, G., Laitinen, J., Pouta, A., Zitting, P., McArdle, W. L., Kroemer, H. K., Völker, U., Völzke, H., Glazer, N. L., Taylor, K. D., Harris, T. B., Alavere, H., Haller, T., Keis, A., Tammesoo, M.-L., Aulchenko, Y., Barroso, I., Khaw, K.-T., Galan, P., Hercberg, S., Lathrop, M., Eyheramendy, S., Org, E., Sõber, S., Lu, X., Nolte, I. M., Penninx, B. W., Corre, T., Masciullo, C., Sala, C., Groop, L., Voight, B. F., Melander, O., O'Donnell, C. J., Salomaa, V., d'Adamo, A. P., Fabretto, A., Faletra, F., Ulivi, S., Del Greco, F. M., Facheris, M., Collins, F. S., Bergman, R. N., Beilby, J. P., Hung, J., Musk, A. W., Mangino, M., Shin, S.-Y., Soranzo, N., Watkins, H., Goel, A., Hamsten, A., Gider, P., Loitfelder, M., Zeginigg, M., Hernandez, D., Najjar, S. S., Navarro, P., Wild, S. H., Corsi, A. M., Singleton, A., de Geus, E. J. C., Willemsen, G., Parker, A. N., Rose, L. M., Buckley, B., Stott, D., Orru, M., Uda, M., LifeLines Cohort Study, van der Klauw, M. M., Zhang, W., Li, X., Scott, J., Chen, Y.-D. I., Burke, G. L., Kähönen, M., Viikari, J., Döring, A., Meitinger, T., Davies, G., Starr, J. M., Emilsson, V., Plump, A., Lindeman, J. H., Hoen, P. A. C. t., König, I. R., EchoGen consortium, Felix, J. F., Clarke, R., Hopewell, J. C., Ongen, H., Breteler, M., Debette, S., Destefano, A. L., Fornage, M., AortaGen Consortium, Mitchell, G. F., CHARGE Consortium Heart Failure Working Group, Smith, N. L., KidneyGen consortium, Holm, H., Stefansson, K., Thorleifsson, G., Thorsteinsdottir, U., CKDGen consortium, Cardiogenics consortium, CardioGram, Samani, N. J., Preuss, M., Rudan, I., Hayward, C., Deary, I. J., Wichmann, H.-E., Raitakari, O. T., Palmas, W., Kooner, J. S., Stolk, R. P., Jukema, J. W., Wright, A. F., Boomsma, D. I., Bandinelli, S., Gyllensten, U. B., Wilson, J. F., Ferrucci, L., Schmidt, R., Farrall, M., Spector, T. D., Palmer, L. J., Tuomilehto, J., Pfeufer, A., Gasparini, P., Siscovick, D., Altshuler, D., Loos, R. J. F., Toniolo, D., Snieder, H., Gieger, C., Meneton, P., Wareham, N. J., Oostra, B. A., Metspalu, A., Launer, L., Rettig, R., Strachan, D. P., Beckmann, J. S., Witteman, J. C. M., Erdmann, J., van Dijk, K. W., Boerwinkle, E., Boehnke, M., Ridker, P. M., Jarvelin, M.-R., Chakravarti, A., Abecasis, G. R., Gudnason,

V., Newton-Cheh, C., Levy, D., Munroe, P. B., Psaty, B. M., Caulfield, M. J., Rao, D. C., Tobin, M. D., Elliott, P., and van Duijn, C. M. (2011). Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nature Genetics*, 43(10):1005–1011.

Wang, P.-P., He, X.-Y., Wang, R., Wang, Z., and Wang, Y.-G. (2014a). High leptin level is an independent risk factor of endometrial cancer: a meta-analysis. *Cellular Physiology and Biochemistry: International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology*, 34(5):1477–1484.

Wang, Y., Cui, M., and Zheng, L. (2012). Genetic polymorphisms in the estrogen receptor-a gene and the risk of endometrial cancer: a meta-analysis. *Acta Obstetricia Et Gynecologica Scandinavica*, 91(8):911–916.

Wang, Y., Li, X., and Hu, H. (2014b). H3k4me2 reliably defines transcription factor binding regions in different cells. *Genomics*, 103(2-3):222–228.

Ward, L. D. and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*, 40(Database issue):D930–934.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Watson, P. and Lynch, H. T. (1994). The tumor spectrum in HNPCC. *Anticancer Research*, 14(4B):1635–1639.

Watson, P., Vasen, H. F. A., Mecklin, J.-P., Bernstein, I., Aarnio, M., Järvinen, H. J., Myrhøj, T., Sunde, L., Wijnen, J. T., and Lynch, H. T. (2008). The risk of extra-colonic, extra-endometrial cancer in the Lynch syndrome. *International Journal of Cancer. Journal International Du Cancer*, 123(2):444–449.

Wedrén, S., Lovmar, L., Humphreys, K., Magnusson, C., Melhus, H., Syvänen, A.-C., Kindmark, A., Landegren, U., Fermér, M. L., Stiger, F., Persson, I., Baron, J. A., and

Weiderpass, E. (2008). Estrogen receptor alpha gene polymorphism and endometrial cancer risk–a case-control study. *BMC cancer*, 8:322.

Weiderpass, E., Adami, H. O., Baron, J. A., Magnusson, C., Bergström, R., Lindgren, A., Correia, N., and Persson, I. (1999). Risk of endometrial cancer following estrogen replacement with and without progestins. *Journal of the National Cancer Institute*, 91(13):1131–1137.

Weiderpass, E., Persson, I., Melhus, H., Wedrén, S., Kindmark, A., and Baron, J. A. (2000). Estrogen receptor alpha gene polymorphisms and endometrial cancer risk. *Carcinogenesis*, 21(4):623–627.

Weissman, S. M., Burt, R., Church, J., Erdman, S., Hampel, H., Holter, S., Jasperson, K., Kalady, M. F., Haidle, J. L., Lynch, H. T., Palaniappan, S., Wise, P. E., and Senter, L. (2012). Identification of individuals at risk for Lynch syndrome using targeted evaluations and genetic testing: National Society of Genetic Counselors and the Collaborative Group of the Americas on Inherited Colorectal Cancer joint practice guideline. *Journal of Genetic Counseling*, 21(4):484–493.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Westin, S. N., Lacour, R. A., Urbauer, D. L., Luthra, R., Bodurka, D. C., Lu, K. H., and Broaddus, R. R. (2008). Carcinoma of the lower uterine segment: a newly described association with Lynch syndrome. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 26(36):5965–5971.

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., 't Hoen, P. A. C., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A. B., Hernandez,

D. G., Nalls, M. A., Homuth, G., Nauck, M., Radke, D., Völker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S. A., Enquobahrie, D. A., Lumley, T., Montgomery, G. W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R. C., Visscher, P. M., Knight, J. C., Psaty, B. M., Ripatti, S., Teumer, A., Frayling, T. M., Metspalu, A., van Meurs, J. B. J., and Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243. Available from: `http://www.nature.com/ng/journal/v45/n10/abs/ng.2756.html`.

Whiffin, N., Dobbins, S. E., Hosking, F. J., Palles, C., Tenesa, A., Wang, Y., Farrington, S. M., Jones, A. M., Broderick, P., Campbell, H., Newcomb, P. A., Casey, G., Conti, D. V., Schumacher, F., Gallinger, S., Lindor, N. M., Hopper, J., Jenkins, M., Dunlop, M. G., Tomlinson, I. P., and Houlston, R. S. (2013). Deciphering the genetic architecture of low-penetrance susceptibility to colorectal cancer. *Human Molecular Genetics*, 22(24):5075–5082.

Whiffin, N., Hosking, F. J., Farrington, S. M., Palles, C., Dobbins, S. E., Zgaga, L., Lloyd, A., Kinnersley, B., Gorman, M., Tenesa, A., Broderick, P., Wang, Y., Barclay, E., Hayward, C., Martin, L., Buchanan, D. D., Win, A. K., Hopper, J., Jenkins, M., Lindor, N. M., Newcomb, P. A., Gallinger, S., Conti, D., Schumacher, F., Casey, G., Liu, T., Swedish Low-Risk Colorectal Cancer Study Group, Campbell, H., Lindblom, A., Houlston, R. S., Tomlinson, I. P., and Dunlop, M. G. (2014). Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Human Molecular Genetics*, 23(17):4729–4737.

Win, A. K., Young, J. P., Lindor, N. M., Tucker, K. M., Ahnen, D. J., Young, G. P., Buchanan, D. D., Clendenning, M., Giles, G. G., Winship, I., Macrae, F. A., Goldblatt, J., Southey, M. C., Arnold, J., Thibodeau, S. N., Gunawardena, S. R., Bapat, B., Baron, J. A., Casey, G., Gallinger, S., Le Marchand, L., Newcomb, P. A., Haile, R. W., Hopper, J. L., and Jenkins, M. A. (2012). Colorectal and other cancer risks for carriers and noncarriers from families with a DNA mismatch repair gene mutation: a prospective

cohort study. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 30(9):958–964.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, 88(1):76–82.

Yeager, M., Orr, N., Hayes, R. B., Jacobs, K. B., Kraft, P., Wacholder, S., Minichiello, M. J., Fearnhead, P., Yu, K., Chatterjee, N., Wang, Z., Welch, R., Staats, B. J., Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cancel-Tassin, G., Cussenot, O., Valeri, A., Andriole, G. L., Gelmann, E. P., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Hoover, R., Hunter, D. J., Chanock, S. J., and Thomas, G. (2007). Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*, 39(5):645–649.

Youlden, D. R. and Baade, P. D. (2011). The relative risk of second primary cancers in Queensland, Australia: a retrospective cohort study. *BMC cancer*, 11:83.

Zeleniuch-Jacquotte, A., Akhmedkhanov, A., Kato, I., Koenig, K. L., Shore, R. E., Kim, M. Y., Levitz, M., Mittal, K. R., Raju, U., Banerjee, S., and Toniolo, P. (2001). Postmenopausal endogenous oestrogens and risk of endometrial cancer: results of a prospective study. *British Journal of Cancer*, 84(7):975–981.

Zhang, Y., Liu, H., Yang, S., Zhang, J., Qian, L., and Chen, X. (2014). Overweight, obesity and endometrial cancer risk: results from a systematic review and meta-analysis. *The International Journal of Biological Markers*, 29(1):e21–29.

Zhang, Z.-H., Su, P.-Y., Hao, J.-H., and Sun, Y.-H. (2013). The role of preexisting diabetes

mellitus on incidence and mortality of endometrial cancer: a meta-analysis of prospective cohort studies. *International Journal of Gynecological Cancer: Official Journal of the International Gynecological Cancer Society*, 23(2):294–303.

Zhou, B., Yang, L., Sun, Q., Cong, R., Gu, H., Tang, N., Zhu, H., and Wang, B. (2008). Cigarette smoking and the risk of endometrial cancer: a meta-analysis. *The American Journal of Medicine*, 121(6):501–508.e3.