Genetics and Genomics of Endometrial Cancer

GENETICS AND GENOMICS OF ENDOMETRIAL CANCER

MAXINE M. CHEN

A Dissertation Submitted to the Faculty of

The Harvard T.H. Chan School of Public Health

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Science

in the Department of Epidemiology

Harvard University

Boston, Massachusetts.

May 2016

Dissertation Advisor: Dr. Immaculata De Vivo Maxine M. Chen

Genetics and Genomics of Endometrial Cancer

Abstract

Endometrial cancer (EC) is the most common gynecological cancer among women in the developed world and is hypothesized to arise from excess estrogen exposure from established risk factors like estrogen-only hormone therapy and obesity. EC is divided into the common “estrogen-dependent” endometrioid subtype and the rare “estrogen-independent” non-endometrioid subtype. However, this broad categorization of EC is not sufficient based on evidence for EC heterogeneity. Furthermore, family history and hereditary syndromes also increase risk, suggesting a genetic component. This dissertation examines the genetic and genomic architecture of EC to provide insight into its etiology and heterogeneity.

In Chapter 1, a four-study EC genome-wide association study meta-analysis of 4,907 cases and 11,645 controls in women of European ancestry is presented. Four loci reached genome-wide significance. Our study identified one novel susceptibility locus at 6p22.3 and confirmed two previously discovered loci at 6q22.31 and 13q22.1. Genes near the 6p22.3 locus are implicated in malignancy and poor prognosis in many cancers, highlighting the potential importance of this region to general cancer susceptibility.

In Chapter 2, we conduct an exome-wide association study of EC. Using a new, commercially-developed exome array comprising ~260,000 putative functional exonic variants, we genotyped a multiethnic population of 3,067 women (1,169 EC cases and 1,898 controls) from the Epidemiology of Endometrial Cancer Consortium to test whether rare variants in coding regions are associated with EC risk. No variants reached global significance in this study. Larger studies are needed to detect associations between rare exonic variants and EC.

In Chapter 3, we combined targeted next-generation sequencing from archival EC tissue with clinical, immunohistochemical, and epidemiologic data to characterize EC in 37 women from the Nurses’ Health Study. Mutations most frequently occurred in *TP53*, *PTEN*, and *PIK3CA*. *TP53* mutations were seen in the majority of tumors that were p53 abnormal. Low grade correlated with frequency of *PTEN* and *PIK3CA* mutation. Our archival EC tissue had mutation profiles consistent with previous studies, supporting use of targeted sequencing panels on archival tissue for mutation detection. This comprehensive annotation of EC demonstrates the utility of integrating many data types in elucidating the spectrum of tumor heterogeneity.

Table of Contents

List of Figures

Chapter 1: GWAS meta-analysis of 16,852 women identifies new susceptibility locus for endometrial cancer.

Chapter 2: Exome-wide association study of endometrial cancer in a multiethnic population.

Chapter 3: Mutation analysis of endometrial cancer in a population-based study by targeted next-generation sequencing.

List of Tables

Chapter 1: GWAS meta-analysis of 16,852 women identifies new susceptibility locus for endometrial cancer.

Chapter 2: Exome-wide association study of endometrial cancer in a multiethnic population.

Chapter 3: Mutation analysis of endometrial cancer in a population-based study by targeted next-generation sequencing.

Acknowledgements

To

Dr. Immaculata De Vivo, my advisor, mentor, and advocate, who made my experience here truly formative; Drs. Peter Kraft, Ed Giovannucci, and Xihong Lin, who have expanded the boundaries of my knowledge through uncommon questions; the numerous collaborators and women that contributed to NSECG, SEARCH, ANECS, E2C2, and NHS, without whom these projects would not have existed; Marta Crous-bou, Jen Prescott, and Connie Chen Turman, for getting into in the programming weeds with me; Hardeep Ranu, Pati Soule, and the De Vivo lab members, for their ability to finesse sometimes-ornery lab equipment into cooperating; all the administrative assistants, also known as magicians of space-time; the denizens of rm 200, past and present, for conversations enlightening and hilarious; the friends I have made in my time here, sanity checks in the odd world of graduate school; Jean, Oscar, and Emily, my siblings and navigators; Andrew, for joining me on this thing called life; and my parents, for everything—

I am forever grateful,

thank you.

Chapter 1

GWAS meta-analysis of 16,852 women identifies new susceptibility locus for endometrial cancer.

Maxine M. Chen[1†], Tracy A. O'Mara[2†], Deborah J. Thompson[3], Jodie N. Painter[2], The Australian National Endometrial Cancer Study Group (ANECS)[2], John Attia[4,5], Amanda Black[6], Louise Brinton[6], Stephen Chanock[6], Chu Chen[7], Timothy H.T. Cheng[8], Linda S. Cook[9,10], Marta Crous-Bou[1,11], Jennifer Doherty[12], Christine M. Friedenreich[13], Montserrat Garcia-Closas[6,14], Mia M. Gaudet[15], Maggie Gorman[8], Christopher Haiman[16], Susan E. Hankinson[11,17], Patricia Hartge[6], Brian E. Henderson[16], Shirley Hodgson[18], Elizabeth G. Holliday[4], Pamela L. Horn-Ross[19], David J. Hunter[1], Loic Le Marchand[20], Xiaolin Liang[21], Jolanta Lissowska[22], Jirong Long[23], Lingeng Lu[24], Anthony M. Magliocco[25], Lynn Martin[8], Mark McEvoy[5], National Study of Endometrial Cancer Genetics Group (NSECG)[8], Sara H. Olson[21], Irene Orlow[21], Loreall Pooler[16], Jennifer Prescott[11], Radhai Rastogi[21], Timothy R. Rebbeck[26], Harvey Risch[24], Carlotta Sacerdote[27,28], Frederick Schumacher[16], Veronica Wendy Setiawan[16], Rodney J. Scott[4,29,30], Xin Sheng[16], Xiao-ou Shu[23], Constance Turman[1], David Van Den Berg[16], Zhaoming Wang[6], Noel S. Weiss[31], Nicholas Wentzensen[6], Lucy Xia[16], Yong-Bing Xiang[32], Hannah P. Yang[6], Herbert Yu[20], Wei Zhang[23], Paul D.P. Pharoah[33], Alison M. Dunning[33], Ian Tomlinson[3], Douglas F. Easton[3,33], Peter Kraft[1], Amanda B. Spurdle[2‡], Immaculata De Vivo[1,11‡]*

[1] Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA 02115, USA

[2] Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Herston, Brisbane, Queensland 4006, Australia

[3] Department of Public Health and Primary Care, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge CB1 8RN, UK

[4] Hunter Medical Research Institute, John Hunter Hospital, Newcastle, New South Wales 2305, Australia

[5] School of Medicine and Public Health, Centre for Clinical Epidemiology and Biostatistics, University of Newcastle, Newcastle, New South Wales 2308, Australia

[6] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA

[7] Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

[8] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK

[9] University of New Mexico, Albuquerque, NM 87131, USA

[10] Division of Cancer Care, Department of Population Health Research, Alberta Health Services, Calgary, AB, Canada

[11] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

[12] Geisel School of Medicine, Dartmouth College, Lebanon, NH 03755, USA

[13] Department of Cancer Epidemiology and Prevention Research, CancerControl Alberta, Alberta Health Services, Calgary, AB, Canada

[14] Division of Genetics and Epidemiology, The Institute of Cancer Research, London UK

[15] Epidemiology Research Program, American Cancer Society, Atlanta, GA 30329, USA

[16] University of Southern California, Los Angeles, CA 90033, USA

[17] Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA 01003, USA

[18] Department of Clinical Genetics, St George's, University of London, London SW17 0RE, UK

[19] Cancer Prevention Institute of California, Fremont, CA 94538, USA

[20] University of Hawaii Cancer Center, Honolulu, HI 96813, USA

[21] Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[22] Department of Cancer Epidemiology and Prevention, M Sklodowska-Curie Cancer Center and Institute of Oncology, Warsaw, Poland

[23] Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt - Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN 27232, USA

[24] Yale University School of Public Health, New Haven, CT 06510, USA

[25] H Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

[26] Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

[27] Center for Cancer Prevention (CPO-Piemonte), Turin, Italy

[28] Human Genetic Foundation (HuGeF), Turin, Italy

[29] Hunter Area Pathology Service, John Hunter Hospital, Newcastle, New South Wales 2305, Australia

[30] School of Biomedical Sciences and Pharmacy, University of Newcastle, Newcastle, New South Wales 2308, Australia

[31] University of Washington, Seattle, WA 19024, USA

[32] Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China

[33] Department of Oncology, Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge CB1 8RN, UK

[†] MMC and TAO contributed equally to this work.

[‡] ABS and IDV supervised this work equally.

* Corresponding author.

Corresponding Author: Immaculata De Vivo, Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115. Phone: +1 617-525-2094; Fax: +1 617-525-2008; Email: nhidv@channing.harvard.edu

Published in *Human Molecular Genetics;* 2016. doi: 10.1093/hmg/ddw092

Abstract

Endometrial cancer is the most common gynecological malignancy in the developed world. Although there is evidence of genetic predisposition to the disease, most of the genetic risk remains unexplained. We present the meta-analysis results of four GWAS (4,907 cases and 11,945 controls total) in women of European ancestry. We describe one new locus reaching genome-wide significance ($P < 5 \times 10^{-8}$) at 6p22.3 (rs1740828; $P = 2.29 \times 10^{-8}$, OR = 1.20), providing evidence of an additional region of interest for genetic susceptibility to endometrial cancer.

Introduction

Endometrial carcinoma (EC), which arises from the epithelial lining of the uterus, is the sixth most common cancer among females worldwide and the most common gynecological malignancy in developed countries(1). According to SEER data(2), between 2005 and 2011, 18.3% of women with EC in the United States did not survive five or more years after diagnosis. Incidence rates of EC in developed countries are increasing over time(3, 4), with most diagnoses made after age 55, making this a significant concern for older women in an aging population. A number of modifiable risk factors have been established, including obesity, estrogen-only post-menopausal hormone therapy, and reproductive history. However, not much is known about the genetic etiology of EC.

Evidence suggests a component of genetic predisposition to EC. Multiple studies have seen a greater than two-fold risk in those with a family history of EC(5–7) and risk for women with first-degree female relatives with early onset disease increases nearly three-fold(8). Additionally, women with Lynch Syndrome, a hereditary autosomal dominant genetic condition due to germline pathogenic variants in DNA mismatch repair genes, have an estimated lifetime risk of EC between 40-70%(9). Heritability estimates for EC are as high as 52%(10–12), though inconsistency in heritability estimates indicate the true value is likely lower.

Genome-wide association studies (GWAS) have discovered more than 1,500 common variants associated with a variety of cancer types(13). However, the statistical power of GWAS may be limited by the modest effect sizes of common variants and by inadequate sample sizes(14, 15). To date, three independent GWAS have been conducted to identify SNPs that contribute to EC risk. One GWAS found a significant association between rs4430796, in 17q12 near *HNF1B*, and EC risk(16). Fine-mapping of this region identified likely variants underlying this association in *HNF1B* intron 1(17). Analysis including a more comprehensive validation phase of this GWAS has since identified an additional 6 loci associated with EC risk at genome-wide levels of significance ((18), Cheng et al submitted). However, no other novel genome-wide significant loci associated with EC risk were identified by the two other published GWAS(14, 15).

Meta-analysis methods synthesize summary data from multiple independent studies, increasing power and reducing false-positive findings(19). We thus conducted a discovery meta-analysis of four GWAS datasets of women of European ancestry for a total of 4,907 cases and 11,945 controls, comprising the largest discovery data set for EC yet.

Results

Meta-analysis of GWAS Results for Risk of Endometrial Cancer.

Meta-analysis of GWAS results from the Australian National Endometrial Cancer Study (ANECS), the US Epidemiology of Endometrial Cancer Consortium (E2C2), the UK National Study of Endometrial Cancer Genetics (NSECG), and the UK Studies of Epidemiology and Risk factors in Cancer Heredity (SEARCH) in 4,907 cases and 11,945 controls of European ancestry examined 9,486,271 SNPs for association with risk of EC. No evidence of genomic inflation was observed in the meta-analysis (λ_{GC} = 1.013, Figure S1.1). After implementing quality control, including removal of SNPs with p-values for heterogeneity <0.05 from further consideration, a total of 137 SNPs clustered in four chromosomal regions reached genome-wide significance at $p < 5 \times 10^{-8}$ (Figure 1.1, Table S1.1).

This meta-analysis of four independent EC GWAS datasets identified four loci with genome-wide levels of significance (Table 1.1). Three loci have been discovered previously by analyses that included the ANECS, SEARCH, and NSECG GWAS datasets((16, 18), Cheng et al submitted): 17q12 near *HNF1B*, 13q22.1 near *KLF5* and 6q22.31 intronic to *LOC643623*. The direction of effect for all three previously identified loci in the E2C2 GWAS alone was consistent with that observed in the original studies (Figure 1.2). In the E2C2 GWAS alone, p-values for the most significant SNPs in 13q22.1 (rs9600103, E2C2 P = 1.74×10^{-5}) and 6q22.31 (rs2797160, E2C2 $P = 1.18 \times 10^{-6}$) exceeded the confirmation threshold of P =

0.017 based on a Bonferroni correction for three tests, representing an independent validation of these two previously reported EC GWAS hits.

The fourth locus at 6p22.3 is a novel risk region for EC, represented by rs1740828 (OR = 1.20, $P = 2.29 \times 10^{-8}$) (Table 1.1). This locus at 6p22.3 falls in an intergenic region between *SOX4* and *CASC15* (Figure 1.3). *SOX4* encodes a transcription factor involved in the regulation of several aspects of development(20). *CASC15* is a long intergenic noncoding RNA that has been identified as a neuroblastoma susceptibility locus(21, 22).

Conditional and joint analyses of these four regions did not identify any secondary association signals, indicating no additional independently associated SNPs after conditioning on the region's lead SNP.

Functional Annotation

Though the most significant risk-associated SNP at 6p22.3 is located in an intergenic region, it may be a marker for an underlying variant that may modulate or regulate nearby or distant genes. To pursue a putative functional role that variants at 6p22.3 may have in risk of EC, we annotated SNPs in LD ($r^2 > 0.2$ in EU 1000 Genomes) with the region's lead SNP, rs1740828, with publicly available data on relevant regulatory elements located near the susceptibility region. Candidate causal SNPs with log likelihood ratios of >1:100 compared with rs1740828 (r^2 between 0.2 and 0.5) overlap with putative enhancers defined by Hnisz(23) and PreSTIGE(24) for *SOX4*, *CASC15*, and *CDKAL1* (Figure 1.3). *CDKAL1* encodes for a methylthiotransferase and is a known type 2 diabetes susceptibility gene(25–27). ENCODE data also show these SNPs mapped to regions displaying evidence of enhancer-specific histone modification (mono-methylation of H3 lysine 4 (H3K4Me1) and H3 lysine 27 acetylation (H3K27Ac)), DNAseI hypersensitivity sites representative of open chromatin, and regions bound by transcription factors.

eQTL Analysis

In order to identify potential biological mechanisms underlying the association between the 6p22.3 locus and EC risk, we performed eQTL analysis using publicly available mRNA expression, somatic copy-number variation and methylation data of 408 EC tumor tissues and 30 adjacent normal endometrial tissues from TCGA. Expression levels of *SOX4*, *CASC15*, and *CDKAL1*, identified as potential target genes by cross reference to Hnisz and PreSTIGE data, were assessed in the analysis. After adjusting for multiple comparisons, no significant associations were seen between SNPs in the risk loci region (Chr6:21549085-21749085) and expression levels of any of these three genes (Table S1.2a, S1.2b). Associations between SNPs and gene expression were also explored using uterine-specific Genotype-Tissue Expression (GTEx) project data (www.gtexportal.org). Similarly, no significant associations were observed between risk SNPs and expression levels of the target genes (data not shown).

Discussion

Our EC GWAS meta-analysis, the largest discovery data set for EC yet, identified one new susceptibility locus at 6p22.3 and confirmed previously discovered loci at 6q22.31 and 13q22.1. The new locus at 6p22.3, represented by rs1740828, lies between two genes, *SOX4* and *CASC15*.

Assuming a log-additive association with risk, these four loci are estimated to account for ~4.4% of the familial relative risk of EC in women of European ancestry. This fraction is less than what has been discovered in studies with comparable sample sizes for cancers such as colorectal(28) and pancreatic cancer(29). It is likely that additional common variants with more modest effect sizes, as well as copy-number variants, rare variants, and indels not tagged by current genotyping arrays, have yet to be discovered, and will contribute to explaining familial endometrial cancer risk. Our meta-analysis was ≥80% powered to detect an association of the magnitude of rs1740828 for SNPs with MAF > 0.21, suggesting that even larger sample sizes would be needed to detect modest effects from lower frequency variants.

Functional annotation suggests that SNPs in LD with rs1740828 overlap putative enhancers for *SOX4*, *CASC15* and *CDKAL1*. Our eQTL results do not support regulation of these particular genes by SNPs falling within 100kb of the lead SNP of the 6p22.3 locus that we identified. However, this may be due to the lack of substantial eQTL data available for adjacent normal endometrial tissue or because eQTLs are context-dependent and may only be expressed in certain stages of cancer development or only when under particular stimuli. Comprehensive studies involving fine-mapping as well as functional analysis are needed to identify biological processes underlying our observed GWAS-identified risk signal at 6p22.3.

Of note, existing data suggest that the 6p22.3 region is relevant to cancer susceptibility in general, summarized in a review of genetic and biological studies reporting on the associations of *CASC15*, *CDKAL1*, and *SOX4* SNPs and gene expression with cancer risk and prognosis (Table S1.3). In larger studies(21, 30), SNPs in/near *CASC15* have been associated with neuroblastoma ($P<10^{-9}$), and increased *CASC15* expression has been implicated in melanoma progression(31). A GWAS of bladder cancer provided suggestive evidence of increased risk in the *CDKAL1* region (lead SNP rs4510656, $p=6.98 \times 10^{-7}$)(32). Given the established associations between EC risk and body-mass index (BMI)(33) and diabetes(34), it is notable that the *CDKAL1* region is also associated with diabetes risk and BMI(35). Furthermore, although the *SOX4* region has yet to be associated with cancer risk by GWAS to date, *SOX4* overexpression has been implicated in malignancy and poor prognosis in a variety of cancers, including chondrosarcoma(36) and cancers of the lung(37–39), prostate(40, 41), breast(42, 43), and endometrium(44). A meta-analysis of 10 studies with >1000 cancer patients reported that *SOX4* tumor overexpression is modestly correlated with poor overall survival(45).

In summary, our study has identified a new endometrial cancer risk locus at 6p22.3. Given previously published associations of SNPs in this region at either genome-wide or notable levels of significance (P<10-6) with other cancer types, our results also highlight this region as a potential general cancer susceptibility locus. Extensive fine-mapping and functional studies are required to identify the biological basis of cancer risk at this region.

Materials and Methods

Datasets. Four large genotyping studies, the Australian National Endometrial Cancer Study (ANECS), the US Epidemiology of Endometrial Cancer Consortium (E2C2), the UK National Study of Endometrial Cancer Genetics (NSECG), and the UK Studies of Epidemiology and Risk factors in Cancer Heredity (SEARCH), contributed a total of 16,852 women (4,907 cases, 11,945 controls) of European ancestry with confirmed EC diagnosis to the meta-analysis. We did not restrict by EC subtype in this analysis. Details of the participating studies and genotyping platforms used are provided in Table S1.4.

Briefly, 606 cases from ANECS(16) were compared to 3083 Australian controls from the Brisbane Adolescent Twin Study (QIMR Controls)(46, 47) (n=1846) and the Hunter Community Study(48) (n=1237). E2C2(49) is an NCI-supported international consortium of more than 45 studies created to investigate the etiology of EC. As previously described(15), four US-based cohort studies, 2 US-based case-control studies, and 1 Poland-based case-control study from the consortium contributed 2695 cases and 2777 controls to this analysis. Cases from NSECG(17) (n=925) were compared with 895 controls from the UK1/CORGI colorectal cancer study(50). Cases from SEARCH(16) (n=681) were compared to 5190 controls from the Wellcome Trust Case-Control Consortium(51).

Genotyping and Imputation. Within each study, genotyping was performed on specific Illumina platforms, as detailed in Table S1.4. Quality control methods agreed upon by all studies were implemented. Briefly, this involved exclusion of SNPs with call rates <95%, MAFs <1%, Hardy-Weinberg violation of at least $P < 10^{-12}$ for cases and $P < 10^{-7}$ for controls, or individuals who are genetically male, first-degree cryptic relations or duplicates, or with call rates <95%. All genotypes were imputed to the positive strand of the 1000 Genomes Project v3, phase 1 dataset with either Minimac(52) or IMPUTE2(53).

Statistical Analysis. Primary association analyses of single variants with EC risk were performed separately in each study using logistic regression implemented with SNPTEST v2(54) or ProbABEL(55), adjusting for relevant principal components and variables specific to the study. Summary statistics reported from each study were combined using fixed-effect meta-analysis with inverse variance weights in METAL(56). The p-value threshold to reach genome-wide significance in the meta-analysis was set to 5×10^{-8}. Heterogeneity across studies was assessed using Cochran's Q statistic. Conditional and joint analysis of summary-level associations, performed with GCTA(57), was used to determine the presence of secondary associations within chromosomal regions of size less than 500kb. The power to detect an association of equal magnitude to rs1740828, the most significant result in the meta-analysis, was calculated using QUANTO 1.2(58).

Functional Annotation. SNPs in linkage disequilibrium (LD), defined as $r^2 > 0.2$ in the European 1000 Genomes data, with the most significant SNP (rs1740828) were annotated using HaploregV2(59) and data from ENCODE(60) including promoter and enhancer histone marks, DNaseI hypersensitivity sites, bound proteins and altered motifs. Additionally, enhancer-gene pairs reported by Hnisz(23) and PreSTIGE(24) were cross-referenced against risk loci to identify likely enhancers overlapping SNPs in LD (r2>0.2) with rs1740828.

eQTL Analysis. To examine tissue-specific eQTLs, data from EC patients were accessed from The Cancer Genome Atlas (TCGA)(61). Normalised RNA-Seq, copy-number and methylation data were downloaded through the Cancer Browser (https://genome-cancer.ucsc.edu). Germline SNP genotypes (Affymetrix 6.0 arrays) were downloaded through the TCGA controlled access portal (https://tcga-data.nci.nih.gov/tcga/)

and QC performed. SNPs were excluded for call rate <95%, MAF <1% or deviations from HWE significant at 10^{-4}. Samples were excluded for low overall call rate (<95%), heterozygosity >3 standard deviations from the mean and non-female sex status (X-chromosome homozygosity rate >0.2). For duplicate samples or samples identified as close relatives by Identity-By-State probabilities >0.85, the sample with the lower call rate was excluded. To assess untyped SNPs, we imputed genotypes present in the 1000 Genomes dataset Phase 3v5 in the risk locus region (+/- 100kb of the lead SNP, rs1740828) for SNPs that were not genotyped by the Affymetrix 6.0 platform. Haplotypes were phased using the MaCH program(62) before running minimac for genotype imputation(53, 52), using the recommended parameters (20 iterations of the Markov sampler and 200 states). SNPs imputed with an $R^2 > 0.3$ and MAF > 0.01 were included in the eQTL analysis. Associations were assessed after Bonferroni correction for the total number of tests performed (number of SNP investigated = 2088, number of genes assessed= 3 and number of sample sets = 2), with a P-value < $4.0x10^{-6}$ required for statistical significance.

Thirty cancer tissue samples had adjacent normal endometrial tissues available with complete genotype and RNA-Seq data. Since gene expression in tumours is affected by acquired somatic alterations, we accounted for somatic copy-number variation and methylation in eQTL analysis of EC tissue. In total, 366 TCGA patients had complete genotype, RNA-Seq, copy-number and methylation data available for the analysis. Expression of *SOX4*, *CASC15* and *CDKAL1* (which were identified as target genes by cross-reference to Hnisz and PreSTIGE data) were adjusted for sequencing platform (Illumina GA or Illumina HiSeq) in adjacent normal EC, and adjusted for sequencing platform, copy-number variation and methylation in EC tissue. The associations between genotype and residual gene expression were evaluated using linear regression models by the mach2qtl program(62, 63).

Contribution to familial risk.

Contribution of known SNPs to familial relative risk under a multiplicative model was computed using the formula detailed in Eeles et al. 2013(64). We assumed the observed familial risk to first-degree relatives of EC cases was 2-fold, the loci had a log-additive association with risk, and the loci were not in LD.

Acknowledgements

ANECS, SEARCH, QIMR, HCS, NSECG and UK1/CORGI thank the many individuals who participated in the research studies. We also thank Kaltin Ferguson, Felicity Lose, Shahana Ahmed, Catherine Healey, Kyriaki Michailidou, Ella Barclay, Lynn Martin and the ANECS research team, the Eastern Cancer Registration and Information Centre and the SEARCH research team, the National Cancer Research Network, the numerous institutions and their staff who supported recruitment. We thank Nick Martin, Grant Montgomery, Dale Nyholt and Anjali Henders for access to GWAS data from QIMR Controls. We thank and acknowledge the contribution of our clinical and scientific collaborators and their staff. See http://www.anecs.org.au/ for full listing of the ANECS Group and other contributors to ANECS. SEARCH collaborators include Mitul Shah, Caroline Baynes, Don Conroy, Bridget Curzon, Patricia Harrington, Sue Irvine, Clare Jordan, Craig Luccarini, Rebecca Mayes, Hannah Munday, Barbara Perkins, Radka Platte, Anabel Simpson, Anne Stafford and Judy West. The NSECG Group comprises: M Adams, A Al-Samarraie, S Anwar, R Athavale, S Awad, A Bali, A Barnes, G Cawdell, S Chan, K Chin, P Comes, M Crawford, J Cullimore, S Ghaem-Maghami, R Gomall, J Green, M Hall, M Harvey, J Hawe, A Head, J Herod, M Hingorani, M Hocking, C Holland, T Hollingsworth, J Hollingworth, T Ind, R Irvine, C Irwin, M Katesmark, S Kehoe, G Kheng-Chew, K Lankester, A Linder, D Luesley, C B-Lynch,V McFarlane, R Naik, N Nicholas, D Nugent, S Oates, A Oladipo, A Papadopoulos, S Pearson, D Radstone, S Raju, A Rathmell, C Redman, M Rymer, P Sarhanis, G Sparrow, N Stuart, S Sundar, A Thompson, S Tinkler, S Trent, A Tristram, N Walji and R Woolas. QIMR thanks Margie Wright, Lisa Bowdler, Sara Smith, Megan Campbell and Scott Gordon for control sample collection and data collection.

ANECS recruitment was supported by project grants from the National Health and Medical Research Council of Australia (ID#339435), The Cancer Council Queensland (ID#4196615) and Cancer Council

Tasmania (ID#403031 and ID#457636). **SEARCH** recruitment was funded by a programme grant from Cancer Research UK [C490/A10124]. Case genotyping was supported by the National Health and Medical Research Council (ID#552402). Control data was generated by the Wellcome Trust Case Control Consortium (**WTCCC**), and a full list of the investigators who contributed to the generation of the data is available from the WTCCC website. We acknowledge use of DNA from the British 1958 Birth Cohort collection, funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02 - funding for this project was provided by the Wellcome Trust under award 085475. **NSECG** was supported by the EU FP7 CHIBCHA grant and **CORGI** by Cancer Research UK. Recruitment of the **QIMR controls** was supported by the National Health and Medical Research Council of Australia (NHMRC). The University of Newcastle, the Gladys M Brawn Senior Research Fellowship scheme, The Vincent Fairfax Family Foundation, the Hunter Medical Research Institute and the Hunter Area Pathology Service all contributed towards the costs of establishing the **Hunter Community Study**. T.O'M and J.N.P. are supported by NHMRC project grant (ID#APP1031333).

A.B.S. is supported by the National Health and Medical Research Council (NHMRC) Fellowship Scheme. D.F.E. is a Principal Research Fellow of Cancer Research UK. A.M.D. was supported by the Joseph Mitchell Trust. I.T. is supported by Cancer Research UK and the Oxford Comprehensive Biomedical Research Centre.

M.M.C. is supported by training grant 5T32CA009001-38 from the NCI.

The **Nurses' Health Study (NHS)** is supported by the NCI, NIH Grants Number UM1 CA186107, P01 CA087969, R01 CA49449, 1R01 CA134958, and 2R01 CA082838. The authors would like to thank the participants and staff of the Nurses' Health Study for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data. The authors would also like to thank Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School. Finally, the authors would also like to acknowledge Pati Soule and Hardeep Ranu for their laboratory assistance.

The **Connecticut Endometrial Cancer Study** was supported by NCI, NIH Grant Number 2R01 CA082838.

The **Fred Hutchinson Cancer Research Center (FHCRC)** is supported by NCI, NIH Grant Number 2R01 CA082838, NIH RO1 CA105212, RO1 CA 87538, RO1 CA75977, RO3 CA80636, NO1 HD23166, R35 CA39779, KO5 CA92002 and funds from the Fred Hutchinson Cancer Research Center.

The **Multiethnic Cohort Study (MEC)** is supported by the NCI, NHI Grants Number CA54281, CA128008 and 2R01 CA082838.

The **California Teachers Study (CTS)** is supported by NCI, NIH Grant Number 2R01 CA082838, R01 CA91019 and R01 CA77398, and contract 97-10500 from the California Breast Cancer Research Fund.

The **Polish Endometrial Cancer Study (PECS)** is supported by the Intramural Research Program of the NCI.

The **Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)** is supported by the Extramural and the Intramural Research Programs of the NCI.

Conflict of Interest Statement

No conflicts of interest to declare.

References

1. Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D. and Bray,F. (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer*, **136**, E359–386.

2. Howlader, N., Noone, A., Krapcho, M., Garshell, J., Miller, D., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., *et al.* (2014) *SEER Cancer Stat. Rev. 1975-2011.*

3. Duncan, M.E., Seagroatt, V. and Goldacre, M.J. (2012) Cancer of the body of the uterus: trends in mortality and incidence in England, 1985-2008. *BJOG*, **119**, 333–339.

4. Wartko, P., Sherman, M.E., Yang, H.P., Felix, A.S., Brinton, L.A. and Trabert, B. (2013) Recent changes in endometrial cancer trends among menopausal-age U.S. women. *Cancer Epidemiol.*, **37**, 374–377.

5. Hemminki, K., Vaittinen, P. and Dong, C. (1999) Endometrial Cancer in the Family-Cancer Database. *Cancer Epidemiol. Biomarkers Prev.*, **8**, 1005–1010.

6. 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. *Eur. J. Cancer Prev.*, **18**, 95–99.

7. Win, A.K., Reece, J.C. and Ryan, S. (2015) Family history and risk of endometrial cancer: a systematic review and meta-analysis. *Obstet. Gynecol.*, **125**, 89–98.

8. 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 Epidemiol. Biomarkers Prev.*, **5**, 411–417.

9. Meyer, L.A., Broaddus, R.R. and Lu, K.H. (2009) Endometrial Cancer and Lynch Syndrome: Clinical and Pathologic Considerations. *Cancer Control*, **16**, 14–22.

10. Schildkraut, J.M., Risch, N. and Thompson, W.D. (1989) Evaluating genetic association among ovarian, breast, and endometrial cancer: evidence for a breast/ovarian cancer relationship. *Am. J. Hum. Genet.*, **45**, 521–529.

11. 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. *N. Engl. J. Med.*, **343**, 78–85.

12. Lu, Y., Ek, W.E., Whiteman, D., Vaughan, T.L., Spurdle, A.B., Easton, D.F., Pharoah, P.D., Thompson, D.J., Dunning, A.M., Hayward, N.K., *et al.* (2014) Most common 'sporadic' cancers have a significant germline genetic component. *Hum. Mol. Genet.*, **23**, 6112–6118.

13. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., *et al.* (2014) The NHGRI-EBI Catalog of published genome-wide association studies.

14. Long, J., Zheng, W., Xiang, Y.B., Lose, F., Thompson, D., Tomlinson, I., Yu, H., Wentzensen, N., Lambrechts, D., Dörk, T., *et al.* (2012) Genome-wide association study identifies a possible susceptibility locus for endometrial cancer. *Cancer Epidemiol. Biomarkers Prev.*, **21**, 980–987.

15. 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., *et al.* (2014) Genome-wide association study of endometrial cancer in E2C2. *Hum. Genet.*, **133**, 211–224.

16. 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, *et al.* (2011) Genome-wide association study identifies a common variant associated with risk of endometrial cancer. *Nat. Genet.*, **43**, 451–454.

17. 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., *et al.* (2015) Fine-mapping of the HNF1B multicancer locus identifies candidate variants that mediate endometrial cancer risk. *Hum. Mol. Genet.*, **24**, 1478–1492.

18. Thompson, D.J., O'Mara, T.A., Glubb, D.M., Painter, J.N., Cheng, T., Folkerd, E., Doody, D., Dennis, J., Webb, P.M., Gorman, M., *et al.* (2016) CYP19A1 fine-mapping and Mendelian randomisation: estradiol is causal for endometrial cancer. *Endocr. Relat. Cancer*, **23**, 77–91.

19. Evangelou, E. and Ioannidis, J.P.A. (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.*, **14**, 379–389.

20. Prior, H.M. and Walter, M.A. (1996) SOX genes: architects of development. *Mol. Med.*, **2**, 405–412.

21. Maris, J.M., Mosse, Y.P., Bradfield, J.P., Hou, C., Monni, S., Scott, R.H., Asgharzadeh, S., Attiyeh, E.F., Diskin, S.J., Laudenslager, M., *et al.* (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.*, **358**, 2585–2593.

22. Russell, M.R., Penikis, A., Oldridge, D.A., Alvarez-Dominguez, J.R., McDaniel, L., Diamond, M., Padovan, O., Raman, P., Li, Y., Wei, J.S., *et al.* (2015) CASC15-S Is a Tumor Suppressor lncRNA at the 6p22 Neuroblastoma Susceptibility Locus. *Cancer Res.*, **75**, 3155–3166.

23. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A. and Young, R.A. (2013) Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, **155**, 934–947.

24. Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sallari, R., Lupien, M., Markowitz, S. and Scacheri, P.C. (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.*, **24**, 1–13.

25. Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., *et al.* (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.*, **39**, 770–775.

26. Wu, Y., Li, H., Loos, R.J.F., Yu, Z., Ye, X., Chen, L., Pan, A., Hu, F.B. and Lin, X. (2008) Common variants in CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8, and HHEX/IDE genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population. *Diabetes*, **57**, 2834–2842.

27. Takeuchi, F., Serizawa, M., Yamamoto, K., Fujisawa, T., Nakashima, E., Ohnaka, K., Ikegami, H., Sugiyama, T., Katsuya, T., Miyagishi, M., *et al.* (2009) Confirmation of multiple risk Loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population. *Diabetes*, **58**, 1690–1699.

28. Al-Tassan, N.A., Whiffin, N., Hosking, F.J., Palles, C., Farrington, S.M., Dobbins, S.E., Harris, R., Gorman, M., Tenesa, A., Meyer, B.F., *et al.* (2015) A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci. Rep.*, **5**, 10442.

29. Wolpin, B.M., Rizzato, C., Kraft, P., Kooperberg, C., Petersen, G.M., Wang, Z., Arslan, A.A., Beane-Freeman, L., Bracci, P.M., Buring, J., *et al.* (2014) Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat. Genet.*, **46**, 994–1000.

30. Diskin, S.J., Capasso, M., Schnepp, R.W., Cole, K.A., Attiyeh, E.F., Hou, C., Diamond, M., Carpenter, E.L., Winter, C., Lee, H., *et al.* (2012) Common variation at 6q16 within HACE1 and LIN28B influences susceptibility to neuroblastoma. *Nat. Genet.*, **44**, 1126–1130.

31. Lessard, L., Liu, M., Marzese, D.M., Wang, H., Chong, K., Kawas, N., Donovan, N.C., Kiyohara, E., Hsu, S., Nelson, N., *et al.* (2015) The CASC15 Long Intergenic Noncoding RNA Locus Is Involved in Melanoma Progression and Phenotype Switching. *J. Invest. Dermatol.*, **135**, 2464–2474.

32. Figueroa, J.D., Ye, Y., Siddiq, A., Garcia-Closas, M., Chatterjee, N., Prokunina-Olsson, L., Cortessis, V.K., Kooperberg, C., Cussenot, O., Benhamou, S., *et al.* (2014) Genome-wide association study identifies multiple loci associated with bladder cancer risk. *Hum. Mol. Genet.*, **23**, 1387–1398.

33. 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. *Int. J. Biol. Markers*, **29**, e21–29.

34. Liao, C., Zhang, D., Mungo, C., Tompkins, D.A. and Zeidan, A.M. (2014) Is diabetes mellitus associated with increased incidence and disease-specific mortality in endometrial cancer? A systematic review and meta-analysis of cohort studies. *Gynecol. Oncol.*, **135**, 163–171.

35. Wen, W., Cho, Y.S., Zheng, W., Dorajoo, R., Kato, N., Qi, L., Chen, C.H., Delahanty, R.J., Okada, Y., Tabara, Y., *et al.* (2012) Meta-analysis identifies common variants associated with body mass index in East Asians. *Nat. Genet.*, **44**, 307–311.

36. Lu, N., Lin, T., Wang, L., Qi, M., Liu, Z., Dong, H., Zhang, X., Zhai, C., Wang, Y., Liu, L., *et al.* (2015) Association of SOX4 regulated by tumor suppressor miR-30a with poor prognosis in low-grade chondrosarcoma. *Tumour Biol.*, **36**, 3843–3852.

37. Walter, R.F.H., Mairinger, F.D., Werner, R., Ting, S., Vollbrecht, C., Theegarten, D., Christoph, D.C., Zarogoulidis, K., Schmid, K.W., Zarogoulidis, P., *et al.* (2015) SOX4, SOX11 and PAX6 mRNA expression was identified as a (prognostic) marker for the aggressiveness of neuroendocrine tumors of the lung by using next-generation expression analysis (NanoString). *Future Oncol.*, **11**, 1027–1036.

38. Wang, D., Hao, T., Pan, Y., Qian, X. and Zhou, D. (2015) Increased expression of SOX4 is a biomarker for malignant status and poor prognosis in patients with non-small cell lung cancer. *Mol. Cell. Biochem.*, **402**, 75–82.

39. Zhou, Y., Wang, X., Huang, Y., Chen, Y., Zhao, G., Yao, Q., Jin, C., Huang, Y., Liu, X. and Li, G. (2015) Down-regulated SOX4 expression suppresses cell proliferation, metastasis and induces apoptosis in Xuanwei female lung cancer patients. *J. Cell. Biochem.*, **116**, 1007–1018.

40. Liu, P., Ramachandran, S., Ali Seyed, M., Scharer, C.D., Laycock, N., Dalton, W.B., Williams, H., Karanam, S., Datta, M.W., Jaye, D.L., *et al.* (2006) Sex-determining region Y box 4 is a transforming oncogene in human prostate cancer cells. *Cancer Res.*, **66**, 4011–4019.

41. Wang, L., Zhang, J., Yang, X., Chang, Y.W.Y., Qi, M., Zhou, Z., Zhang, J. and Han, B. (2013) SOX4 is associated with poor prognosis in prostate cancer and promotes epithelial-mesenchymal transition in vitro. *Prostate Cancer Prostatic Dis.*, **16**, 301–307.

42. Zhang, J., Liang, Q., Lei, Y., Yao, M., Li, L., Gao, X., Feng, J., Zhang, Y., Gao, H., Liu, D.X., *et al.* (2012) SOX4 induces epithelial-mesenchymal transition and contributes to breast cancer progression. *Cancer Res.*, **72**, 4597–4608.

43. Song, G.D., Sun, Y., Shen, H. and Li, W. (2015) SOX4 overexpression is a novel biomarker of malignant status and poor prognosis in breast cancer patients. *Tumour Biol.*, **36**, 4167–4173.

44. Huang, Y.W., Liu, J.C., Deatherage, D.E., Luo, J., Mutch, D.G., Goodfellow, P.J., Miller, D.S. and Huang, T.H.M. (2009) Epigenetic repression of microRNA-129-2 leads to overexpression of SOX4 oncogene in endometrial cancer. *Cancer Res.*, **69**, 9038–9046.

45. Chen, J., Ju, H.L., Yuan, X.Y., Wang, T.J. and Lai, B.Q. (2016) SOX4 is a potential prognostic factor in human cancers: a systematic review and meta-analysis. *Clin. Transl. Oncol.*, **18**, 65–72.

46. 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. *Genet. Epidemiol.*, **16**, 40–53.

47. Painter, J.N., Anderson, C.A., Nyholt, D.R., Macgregor, S., Lin, J., Lee, S.H., Lambert, A., Zhao, Z.Z., Roseman, F., Guo, Q., *et al.* (2011) Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat. Genet.*, **43**, 51–54.

48. McEvoy, M., Smith, W., D'Este, C., Duke, J., Peel, R., Schofield, P., Scott, R., Byles, J., Henry, D., Ewald, B., *et al.* (2010) Cohort Profile: The Hunter Community Study. *Int. J. Epidemiol.*, **39**, 1452–1463.

49. Olson, S.H., Chen, C., De Vivo, I., Doherty, J.A., Hartmuller, V., Horn-Ross, P.L., Lacey, J.V., Lynch, S.M., Sansbury, L., Setiawan, V.W., *et al.* (2009) Maximizing resources to study an uncommon cancer: E2C2--Epidemiology of Endometrial Cancer Consortium. *Cancer Causes Control*, **20**, 491–496.

50. Houlston, R.S., Cheadle, J., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Spain, S.L., Broderick, P., Domingo, E., Farrington, S., *et al.* (2010) Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.*, **42**, 973–977.

51. 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**, 661–678.

52. Fuchsberger, C., Abecasis, G.R. and Hinds, D.A. (2014) minimac2: faster genotype imputation. *Bioinformatics*, **31**, 782–784.

53. 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 Genet.*, **5**, e1000529.

54. Ferreira, T. and Marchini, J. (2011) Modeling interactions with known risk loci-a Bayesian model averaging approach. *Ann. Hum. Genet.*, **75**, 1–9.

55. Aulchenko, Y.S., Struchalin, M.V. and van Duijn, C.M. (2010) ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, **11**, 134.

56. Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.

57. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., *et al.* (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, **44**, 369–375, S1–3.

58. Gauderman, W.J. (2002) Sample Size Requirements for Association Studies of Gene-Gene Interaction. *Am. J. Epidemiol.*, **155**, 478–484.

59. Ward, L.D. and Kellis, M. (2011) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930–934.

60. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

61. The Cancer Genome Network (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.

62. Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816–834.

63. Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*, **10**, 387–406.

64. Eeles, R.A., Al Olama, A.A., Benlloch, S., Saunders, E.J., Leongamornlert, D.A., Tymrakiewicz, M., Ghoussaini, M., Luccarini, C., Dennis, J., Jugurnauth-Little, S., *et al.* (2013) Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.*, **45**, 385–391.

Table 1.1. Association results for loci reaching genome-wide significance with no evidence of significant study heterogeneity

Lead SNP	Chromosome	Position (hg19)	Nearby Gene	Description	Alleles	OR	P	RAF*
rs2797160	6q22.31	126010116	*LOC643623***	intronic	A/G	1.21	4.04E-13	0.578
rs9600103	13q22.1	73811879	*KLF5*	intergenic	A/T	1.23	3.76E-12	0.722
rs1740828	6p22.3	21649085	*SOX4*	intergenic	G/A	1.20	2.29E-08	0.516
rs11651052	17q12	36102381	*HNF1B*	intronic	G/A	1.16	1.18E-08	0.535

* Risk Allele Frequency

** uncharacterized gene region

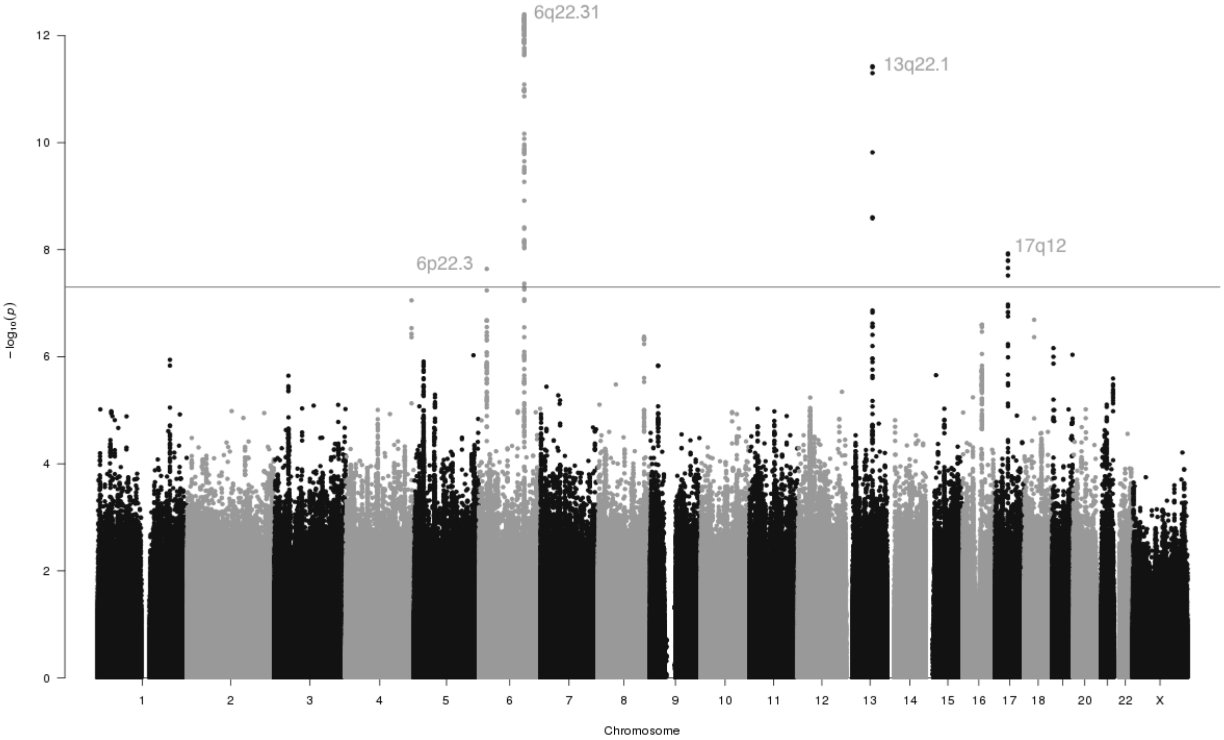

Figure 1.1. Manhattan plot of meta-analysis results for endometrial cancer in four cohorts. Association results between imputed and genotyped SNPs and risk of EC in women of European ancestry are depicted. Dashed line indicates the log of the threshold for genome-wide significance ($P < 5.0 \times 10^{-8}$).

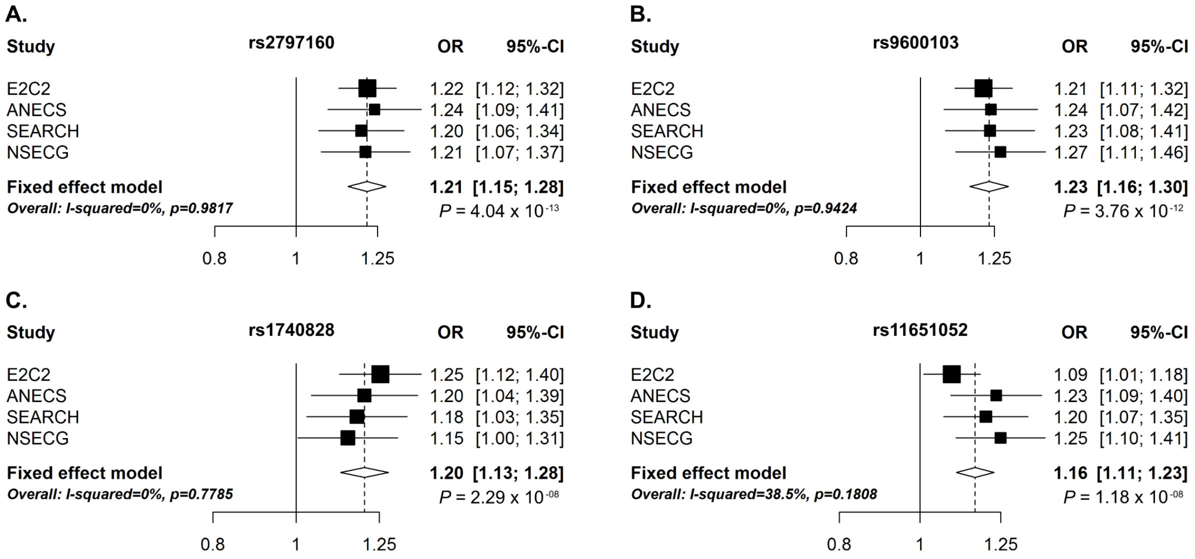

Figure 1.2. Forest plots of the odds ratios for the association between rs2797160, rs1740828, rs9600103, rs11651052 and endometrial cancer.

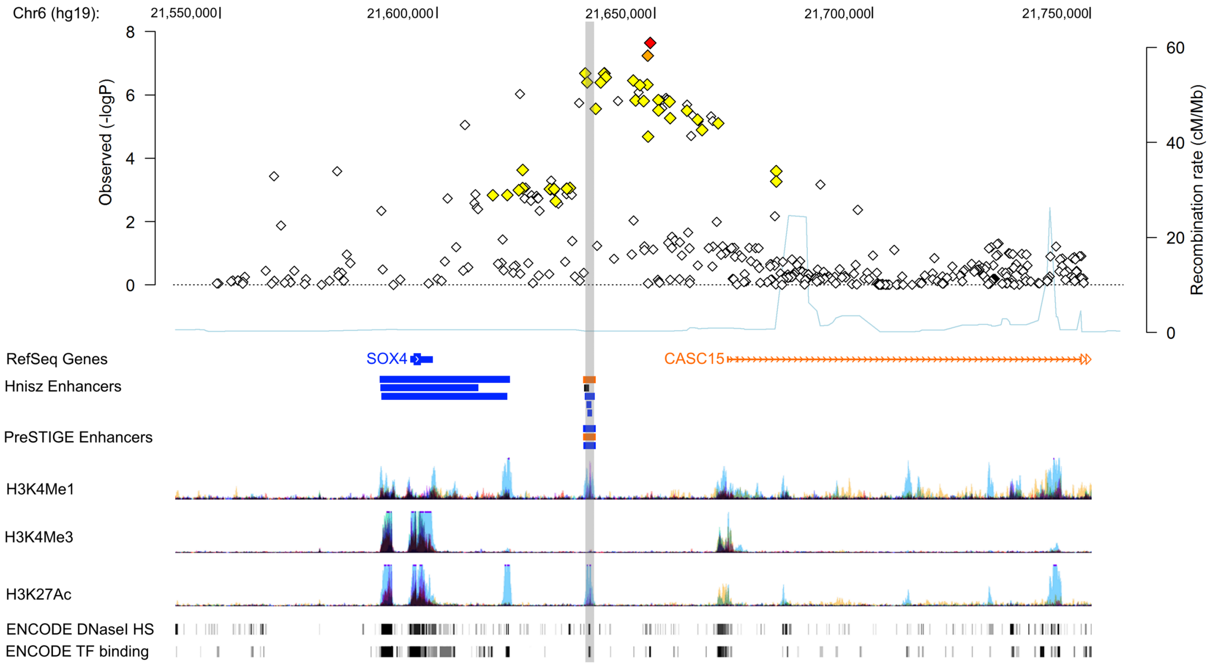

Figure 1.3. Regional association plot of 6p22.3 with annotation of genomic features, likely enhancers, and target genes. Association results for all SNPs in the 6p22.3 locus with EC risk from the meta-analysis are shown in the first panel. SNPs are plotted as the negative log of the P-value against relative position across the locus (base position [hg19] displayed across the top). The lead SNP, rs1740828, is shown as a red filled diamond. LD with surrounding SNPs are indicated by color (SNPs $0.5 \leq r^2 < 0.8$ are orange, $0.2 \leq r^2 <0.5$ are yellow, and $r^2 < 0.2$ are unfilled). There were no SNPs with an $r^2 \geq 0.8$ to the lead SNP. The second panel displays genes as identified by RefSeq. Likely enhancers predicted by Hnisz et al(23) and PreSTIGE(24) that overlap SNPs in LD ($r^2 > 0.2$) with the lead SNP are depicted as colored bars, where the color matches the schematic of its predicted target gene (the black bar is predicted to target *CDKAL1*, not shown in this figure). Histone modification associated with promoters (H3K4Me1) and enhancers (H3K4Me1 and H3K27Ac) from seven ENCODE Project cell types and DNaseI hypersensitivity sites (DHS) and transcription factor (TF) binding sites identified in 125 and 91 ENCODE Project cell types, respectively, are also displayed.

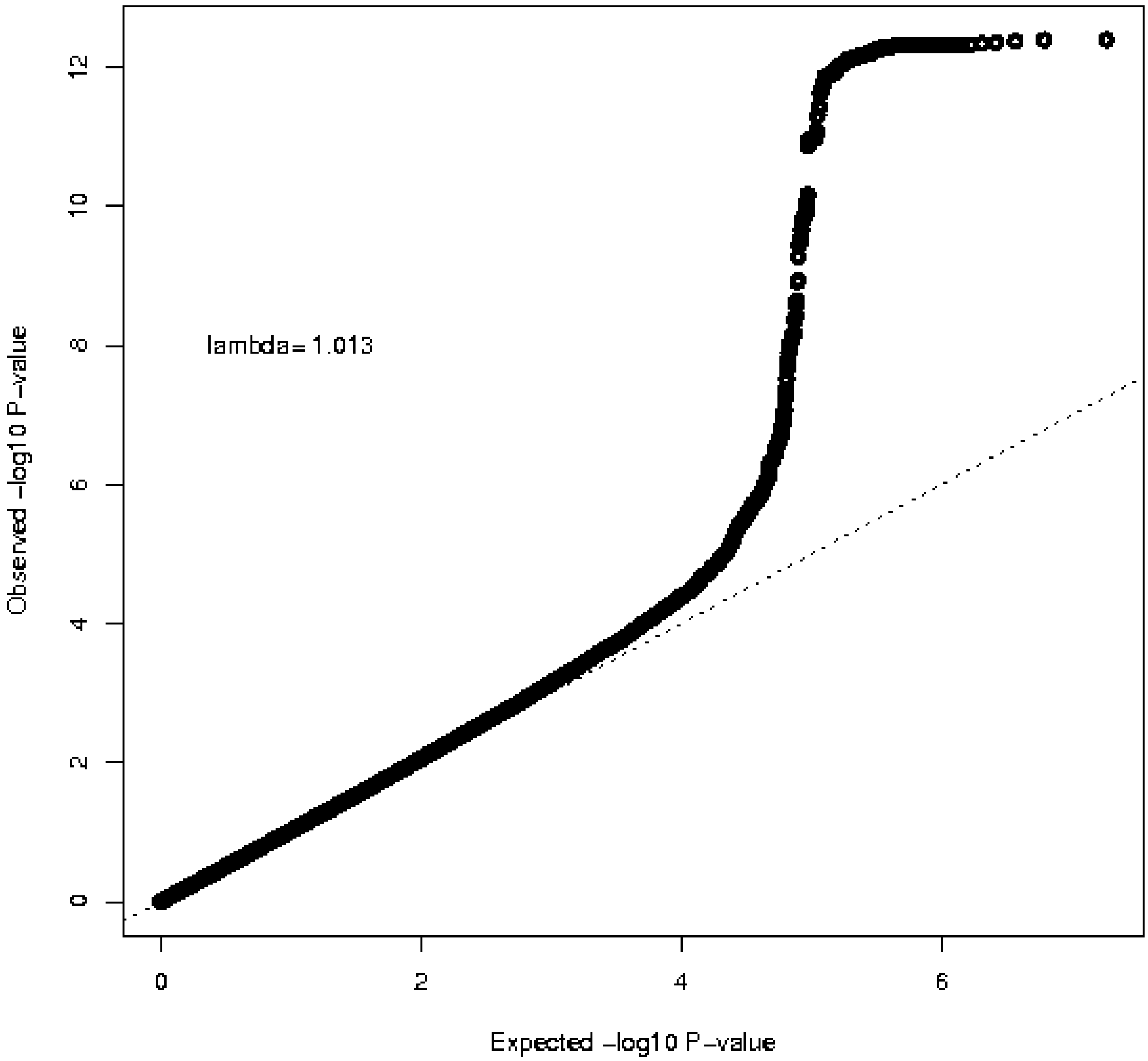

Figure S1.1. Quantile-quantile plot of association results from meta-analysis of imputed and genotyped SNPs and risk of endometrial cancer.

Table S1.1. Meta-analysis results for SNPs reaching genome-wide significance.

Marker Name	Chr.	Position	A1	A2	Beta Estimate	Standard Error	P-value	Direction of Effect in Each Contributing Study	Heterozygosity Chi-Square Statistic	Heterozygosity DF	Heterozygosity P-Value
6q22.31											
rs2797160	6	126010116	a	g	0.1942	0.0268	4.04E-13	++++	0.174	3	0.9817
rs1777226	6	126017691	a	c	-0.1939	0.0267	4.08E-13	----	0.158	3	0.984
rs1739354	6	126017808	c	g	-0.1937	0.0267	4.28E-13	----	0.154	3	0.9847
rs6910933	6	126017155	c	g	0.1935	0.0267	4.43E-13	++++	0.178	3	0.9811
rs6934435	6	126017481	t	g	0.1934	0.0267	4.62E-13	++++	0.155	3	0.9844
rs1739362	6	126020703	a	t	0.1934	0.0267	4.74E-13	++++	0.142	3	0.9863
rs1777225	6	126018270	t	c	0.1934	0.0267	4.74E-13	++++	0.168	3	0.9826
rs12717178	6	126016499	a	g	-0.1932	0.0267	4.79E-13	----	0.186	3	0.9798
rs6933302	6	126016951	t	c	0.1932	0.0267	4.79E-13	++++	0.186	3	0.9798
rs6933471	6	126017029	t	g	0.1932	0.0267	4.79E-13	++++	0.186	3	0.9798
rs1739355	6	126018114	a	g	0.1933	0.0267	4.80E-13	++++	0.148	3	0.9854
rs6927161	6	126015954	t	c	0.1932	0.0267	4.82E-13	++++	0.188	3	0.9795
rs6904992	6	126016003	a	g	-0.1932	0.0267	4.82E-13	----	0.188	3	0.9795
rs1739373	6	126011509	a	g	0.1932	0.0267	4.84E-13	++++	0.16	3	0.9837
rs1739349	6	126014984	c	g	-0.1931	0.0267	4.87E-13	----	0.188	3	0.9795
rs1578793	6	126015057	a	g	0.1931	0.0267	4.87E-13	++++	0.188	3	0.9795
rs1578794	6	126015469	t	c	0.1931	0.0267	4.87E-13	++++	0.188	3	0.9795
rs1739368	6	126011079	t	c	-0.193	0.0267	4.93E-13	----	0.185	3	0.98
rs1739347	6	126014157	t	c	-0.1931	0.0267	4.93E-13	----	0.187	3	0.9797
rs1739348	6	126014573	t	c	0.1931	0.0267	4.93E-13	++++	0.188	3	0.9795

Table S1ti1 (continued)ti Meta-analysis results for SNPs reaching genome-wide significancet

rs1739371	6	126011291	a	g	-0.193	0.0267	4.95E-13	----	0.183	3	0.9802
rs1739378	6	126012262	a	c	-0.1931	0.0267	4.95E-13	----	0.189	3	0.9793
rs2797162	6	126011381	t	g	-0.193	0.0267	5.06E-13	----	0.189	3	0.9793
rs1739372	6	126011325	a	g	-0.193	0.0267	5.08E-13	----	0.188	3	0.9796
rs984041	6	126021328	a	t	0.193	0.0267	5.09E-13	++++	0.142	3	0.9864
rs1739363	6	126020980	a	g	-0.193	0.0267	5.22E-13	----	0.139	3	0.9868
rs926853	6	126021435	a	t	0.1929	0.0267	5.27E-13	++++	0.144	3	0.9861
rs1777222	6	126021030	t	c	-0.1928	0.0267	5.47E-13	----	0.142	3	0.9863
rs2797154	6	126005197	a	g	-0.1926	0.0267	5.48E-13	----	0.108	3	0.9908
rs984040	6	126021277	t	c	0.1927	0.0267	5.67E-13	++++	0.142	3	0.9864
rs2747717	6	126008435	a	g	0.1923	0.0267	5.78E-13	++++	0.185	3	0.9799
rs2797159	6	126009557	a	g	-0.1923	0.0267	5.86E-13	----	0.168	3	0.9825
rs2747721	6	126009527	a	g	0.1921	0.0267	6.41E-13	++++	0.152	3	0.985
rs2747722	6	126009629	a	g	0.1918	0.0267	6.50E-13	++++	0.161	3	0.9836
rs2797158	6	126009398	a	g	-0.1918	0.0267	6.57E-13	----	0.162	3	0.9835
rs2747720	6	126009458	a	g	0.1918	0.0267	6.57E-13	++++	0.162	3	0.9835
rs2747718	6	126009109	a	c	-0.1917	0.0267	6.69E-13	----	0.157	3	0.9842
rs1777224	6	126019527	t	c	0.1921	0.0267	6.80E-13	++++	0.092	3	0.9928
rs2747719	6	126009214	t	c	-0.1916	0.0267	6.81E-13	----	0.162	3	0.9835
rs1418948	6	126007018	t	c	-0.1913	0.0266	6.90E-13	----	0.172	3	0.982
rs13328298	6	126016580	a	g	-0.1924	0.0268	7.20E-13	----	0.21	3	0.976
rs78602343	6	126019768	t	c	0.192	0.0268	7.34E-13	++++	0.154	3	0.9846
rs983543	6	126005767	a	g	0.191	0.0266	7.42E-13	++++	0.173	3	0.9818
rs76407388	6	126004194	a	g	-0.1974	0.0275	7.49E-13	----	0.831	3	0.8419
rs1739352	6	126005310	t	c	-0.191	0.0266	7.50E-13	----	0.174	3	0.9817
rs2747714	6	126007620	a	g	0.191	0.0266	7.58E-13	++++	0.176	3	0.9814
rs4897153	6	126003403	a	g	-0.1908	0.0266	7.84E-13	----	0.176	3	0.9813
rs6910786	6	126017141	a	t	0.1916	0.0267	7.85E-13	++++	0.179	3	0.981

Table S1t1 (continued)ti Meta-analysis results for SNPs reaching genome-wide significancet

rs1777194	6	126004883	a	g	0.1908	0.0266	7.87E-13	++++	0.175	3	0.9815
rs4897152	6	126002400	a	g	-0.1908	0.0266	7.92E-13	----	0.177	3	0.9812
rs1935979	6	126002774	a	g	-0.1906	0.0266	8.34E-13	----	0.177	3	0.9811
rs2747724	6	126004935	a	g	0.1905	0.0266	8.48E-13	++++	0.175	3	0.9816
rs1739357	6	126019655	t	g	0.191	0.0267	9.03E-13	++++	0.095	3	0.9925
rs9321050	6	126001568	a	g	-0.1902	0.0266	9.25E-13	----	0.148	3	0.9856
rs1739367	6	126004720	t	g	0.1901	0.0266	9.61E-13	++++	0.18	3	0.9808
rs1777197	6	126007401	a	g	-0.1909	0.0268	9.73E-13	----	0.107	3	0.991
rs1954360	6	126001064	a	g	-0.19	0.0266	9.81E-13	----	0.148	3	0.9855
rs1954361	6	126001423	c	g	-0.19	0.0266	9.82E-13	----	0.148	3	0.9855
rs9491471	6	125991715	t	c	0.1898	0.0266	1.02E-12	++++	0.168	3	0.9825
rs1935772	6	125994708	t	c	-0.1893	0.0266	1.18E-12	----	0.145	3	0.986
rs6904069	6	125995134	a	g	0.1892	0.0266	1.21E-12	++++	0.139	3	0.9868
rs4897151	6	125993202	t	g	-0.1893	0.0266	1.21E-12	----	0.162	3	0.9834
rs1832938	6	125988964	c	g	0.1895	0.0267	1.21E-12	++++	0.185	3	0.9799
rs2211419	6	125995533	a	g	-0.1892	0.0266	1.22E-12	----	0.136	3	0.9872
rs6940748	6	125994080	t	c	0.189	0.0266	1.29E-12	++++	0.129	3	0.9881
rs6569435	6	125998186	t	c	-0.189	0.0266	1.29E-12	----	0.136	3	0.9873
rs1418642	6	125999768	a	g	0.189	0.0266	1.30E-12	++++	0.135	3	0.9874
rs2211420	6	125995549	t	c	-0.189	0.0266	1.31E-12	----	0.139	3	0.9868
rs8180614	6	126000599	c	g	0.1888	0.0266	1.36E-12	++++	0.136	3	0.9872
rs1418641	6	125999854	t	c	0.1888	0.0266	1.36E-12	++++	0.137	3	0.9871
rs1418640	6	125999866	a	g	0.1888	0.0266	1.36E-12	++++	0.137	3	0.9871
rs4895798	6	126000162	a	g	-0.1888	0.0266	1.36E-12	----	0.137	3	0.9871
rs1832980	6	125997444	t	g	-0.1887	0.0266	1.39E-12	----	0.137	3	0.987
rs1935774	6	125996661	t	c	0.1877	0.0266	1.70E-12	++++	0.079	3	0.9942
rs1418639	6	125999940	t	c	0.1883	0.0267	1.78E-12	++++	0.074	3	0.9947
rs1935773	6	125996475	a	g	-0.1875	0.0266	1.95E-12	----	0.098	3	0.9921

Table S11 (continued) Meta-analysis results for SNPs reaching genome-wide significance

rs28629380	6	126004197	a	g	-0.1881	0.0268	2.10E-12	----	0.203	3	0.9771
rs1832979	6	125997436	a	g	-0.1866	0.0266	2.28E-12	----	0.075	3	0.9946
rs9401843	6	126004124	t	c	-0.1877	0.0268	2.33E-12	----	0.094	3	0.9925
rs1739364	6	126022383	a	g	-0.183	0.0268	8.21E-12	----	0.15	3	0.9852
rs12527010	6	125991507	a	g	0.18	0.0264	9.91E-12	++++	0.081	3	0.9941
rs2747715	6	126007719	a	t	0.1817	0.0267	1.05E-11	++++	0.126	3	0.9885
rs1630556	6	126013155	a	g	0.1822	0.0268	1.05E-11	++++	0.151	3	0.9851
rs1739380	6	126012858	t	c	0.1821	0.0268	1.06E-11	++++	0.151	3	0.9851
rs1777183	6	126011995	a	c	0.1821	0.0268	1.07E-11	++++	0.151	3	0.9851
rs1739375	6	126012013	t	c	0.1821	0.0268	1.07E-11	++++	0.151	3	0.9851
rs1739376	6	126012084	c	g	0.1821	0.0268	1.07E-11	++++	0.151	3	0.9851
rs1739377	6	126012236	t	c	0.1821	0.0268	1.07E-11	++++	0.151	3	0.9851
rs1739379	6	126012593	t	c	0.1821	0.0268	1.08E-11	++++	0.15	3	0.9852
rs1739374	6	126011825	t	c	-0.1821	0.0268	1.08E-11	----	0.151	3	0.9851
rs2747725	6	126012397	t	g	0.1821	0.0268	1.08E-11	++++	0.152	3	0.985
rs1777182	6	126013614	a	t	-0.182	0.0268	1.09E-11	----	0.148	3	0.9856
rs1777195	6	126006861	a	c	0.1806	0.0267	1.37E-11	++++	0.126	3	0.9885
rs1612249	6	126014916	a	c	-0.1948	0.0299	6.85E-11	--?-	0.133	2	0.9358
rs78229684	6	126007996	t	c	-0.1916	0.0295	8.47E-11	--?-	0.171	2	0.9181
rs1612274	6	126014907	a	c	-0.1759	0.0273	1.09E-10	----	0.976	3	0.8071
rs1739366	6	126007409	t	c	0.191	0.0297	1.18E-10	++?+	0.107	2	0.948
rs1343120	6	125992810	a	g	-0.1897	0.0295	1.32E-10	--?-	0.165	2	0.921
rs1739370	6	126011231	t	c	-0.1739	0.0271	1.32E-10	----	0.892	3	0.8274
rs1418951	6	125996185	a	g	-0.1894	0.0295	1.38E-10	--?-	0.136	2	0.9341
rs1739358	6	126019736	a	g	-0.1909	0.0298	1.49E-10	--?-	0.121	2	0.9414
rs77678056	6	126019738	a	g	0.1908	0.0298	1.52E-10	++?+	0.121	2	0.9413
rs926854	6	126021780	a	g	0.1907	0.0298	1.56E-10	++?+	0.315	2	0.8541
rs926855	6	126021782	a	g	0.1905	0.0298	1.63E-10	++?+	0.316	2	0.854

Table S1t1 (continued)ti Meta-analysis results for SNPs reaching genome-wide significancet

rs80303782	6	126004193	t	c	0.1941	0.0306	2.24E-10	++?+	0.522	2	0.7704
rs1418637	6	125992553	a	g	-0.1702	0.027	2.84E-10	----	1.004	3	0.8003
rs2211418	6	125995503	a	g	0.1861	0.0296	3.11E-10	++?+	0.101	2	0.9505
rs2326292	6	125999674	a	g	0.1696	0.027	3.26E-10	++++	0.895	3	0.8265
rs2797161	6	126010789	a	g	0.1927	0.0307	3.60E-10	++?+	0.122	2	0.9407
rs2747723	6	126010790	t	c	-0.1926	0.0307	3.62E-10	--?-	0.119	2	0.9423
rs1832937	6	125985934	a	g	0.1654	0.0267	5.41E-10	++++	0.029	3	0.9987
rs1777198	6	126007416	t	c	-0.1809	0.0298	1.22E-09	--?-	0.067	2	0.967
rs1777220	6	126022602	t	g	-0.1591	0.027	3.85E-09	----	1.022	3	0.7959
rs2226158	6	125995467	a	g	0.1733	0.0295	4.06E-09	++?+	0.081	2	0.9604
rs9491503	6	126031682	a	g	-0.1581	0.0273	6.60E-09	----	0.336	3	0.9531
rs1268093	6	126029235	a	g	-0.1575	0.0272	6.98E-09	----	0.384	3	0.9436
rs1268066	6	126035041	t	c	-0.1579	0.0273	7.11E-09	----	0.369	3	0.9466
rs1343121	6	126036184	t	c	-0.1579	0.0273	7.17E-09	----	0.366	3	0.9473
rs1269176	6	126029682	a	t	-0.1577	0.0273	7.17E-09	----	0.379	3	0.9446
rs6939969	6	126034563	t	c	-0.1574	0.0272	7.50E-09	----	0.359	3	0.9485
rs1268092	6	126029043	t	c	-0.1569	0.0272	7.90E-09	----	0.326	3	0.9551
rs6569437	6	126034540	t	g	0.1566	0.0272	8.78E-09	++++	0.327	3	0.9548
rs1268067	6	126036621	t	c	0.1566	0.0273	9.40E-09	++++	0.255	3	0.9682
rs6939865	6	126027318	a	c	0.169	0.0309	4.31E-08	++?+	0.025	2	0.9874
6p22.3											
rs1740828	6	21649085	a	g	-0.1829	0.0327	2.29E-08	----	1.094	3	0.7785
13q22.1											
rs9600103	13	73811879	a	t	0.2074	0.0299	3.76E-12	++++	0.39	3	0.9424
rs7981863	13	73812141	t	c	-0.2072	0.0299	3.93E-12	----	0.394	3	0.9416
rs11841589	13	73814891	t	g	-0.2066	0.0299	5.04E-12	----	0.521	3	0.9144
rs9592895	13	73813982	t	c	0.1801	0.0281	1.53E-10	++++	0.375	3	0.9453
rs7989799	13	73813436	a	t	-0.1981	0.0332	2.49E-09	--?-	0.132	2	0.9363

Table S1ti1 (continued)ti Meta-analysis results for SNPs reaching genome-wide significancet

rs7988505	13	73813435	c	g	0.1979	0.0332	2.60E-09	++?+	0.129	2	0.9375
17q12											
rs11651052	17	36102381	a	g	-0.1523	0.0267	1.18E-08	----	4.88	3	0.1808
rs4430796	17	36098040	a	g	0.1511	0.0265	1.23E-08	++++	5.805	3	0.1215
rs8064454	17	36101586	a	c	-0.1507	0.0267	1.60E-08	----	4.488	3	0.2133
rs11263763	17	36103565	a	g	0.1533	0.0271	1.62E-08	++++	4.614	3	0.2023
rs11651755	17	36099840	t	c	0.1487	0.0266	2.20E-08	++++	5.125	3	0.1629
rs11263761	17	36097775	a	g	0.1505	0.0272	3.05E-08	++++	5.081	3	0.1659

Abbreviations-- Chr.: Chromosome, A1: Allele 1 (minor allele), A2: Allele 2

Top SNP at each locus highlighted in grey

Table S1.2a. Results from eQTL analysis of 6p22 region: endometrial tumour tissue.

*TRAIT	MARKER	BP	ALLELES	FREQ1	RSQR	EFFECT2	STDERR	CHISQ	PVALUE
Top 25 SNP associations with CASC15 sorted by p-value									
CASC15	rs546083928	21596246	R,I	0.9877	0.6818	-1.251	0.508	6.0715	0.01374
CASC15	rs543173991	21629474	T,A	0.9884	0.7664	-1.118	0.468	5.7084	0.01688
CASC15	rs190176895	21629550	C,A	0.9877	0.7487	-1.031	0.462	4.9942	0.02543
CASC15	rs111411936	21626246	G,A	0.9877	0.7672	-1.021	0.458	4.9656	0.02586
CASC15	rs7762989	21624153	C,T	0.9883	0.8214	-0.998	0.459	4.719	0.02983
CASC15	rs71657670	21613526	R,D	0.9864	0.77	-0.912	0.44	4.2926	0.03828
CASC15	rs76501983	21631411	G,A	0.9854	0.8228	-0.81	0.396	4.1772	0.04097
CASC15	rs78923341	21612472	A,C	0.9872	0.895	-0.797	0.411	3.7544	0.05267
CASC15	rs111228858	21612309	G,A	0.9872	0.8945	-0.797	0.411	3.7507	0.05279
CASC15	rs79169915	21607457	G,A	0.9882	0.848	-0.871	0.45	3.7445	0.05298
CASC15	rs112650104	21611568	C,G	0.9871	0.889	-0.791	0.412	3.6915	0.05469
CASC15	rs111412021	21606278	T,C	0.9882	0.8163	-0.889	0.466	3.6363	0.05653
CASC15	rs1744875	21648534	C,A	0.4998	0.4732	-0.237	0.126	3.5344	0.06011
CASC15	rs10557323	21615430	R,D	0.9849	0.974	-0.668	0.361	3.4317	0.06396
CASC15	rs77222012	21614945	G,A	0.9849	0.9755	-0.668	0.361	3.4286	0.06408
CASC15	rs75127321	21613946	G,A	0.9849	0.9767	-0.666	0.36	3.4167	0.06454
CASC15	rs6935968	21613156	G,A	0.9849	0.9776	-0.665	0.36	3.4112	0.06475
CASC15	rs112538002	21611751	T,C	0.9848	0.97	-0.659	0.361	3.3284	0.06809
CASC15	rs113842280	21611571	T,G	0.9848	0.967	-0.656	0.361	3.2996	0.0693
CASC15	rs112553613	21616239	A,G	0.9848	0.9676	-0.656	0.361	3.2977	0.06938
CASC15	rs1744866	21631188	C,T	0.9885	0.3807	1.111	0.629	3.1167	0.0775
CASC15	rs78345714	21734098	T,C	0.9849	0.6839	0.611	0.349	3.0665	0.07992
CASC15	rs76944255	21582643	G,A	0.981	0.4687	-0.739	0.422	3.0606	0.08021
CASC15	rs80061387	21579793	T,C	0.9808	0.4585	-0.738	0.425	3.0156	0.08246
CASC15	rs2251647	21677746	C,A	0.9512	0.9786	0.348	0.201	2.9867	0.08395
Top 25 SNP associations with CDKAL1 sorted by p-value									
CDKAL1	rs79164129	21601917	T,C	0.9855	0.3283	0.747	0.266	7.88	0.004998
CDKAL1	rs2251647	21677746	C,A	0.9512	0.9786	0.231	0.088	6.7986	0.009123
CDKAL1	rs201884896	21657694	D,R	0.9455	0.3238	0.372	0.148	6.3512	0.01173
CDKAL1	rs7772335	21696185	C,T	0.8913	0.4593	-0.208	0.084	6.192	0.01283
CDKAL1	rs66647983	21633079	R,D	0.5368	0.3749	-0.132	0.059	5.0066	0.02525
CDKAL1	rs114455294	21693186	A,T	0.9731	0.5546	-0.287	0.132	4.7061	0.03006
CDKAL1	rs1740849	21619809	G,A	0.3637	0.9622	-0.079	0.038	4.2455	0.03936
CDKAL1	rs571708107	21634038	D,R	0.6795	0.8686	0.084	0.042	4.0624	0.04385
CDKAL1	rs75994264	21745903	C,T	0.9868	0.3369	-0.504	0.251	4.0444	0.04432
CDKAL1	rs60368679	21746134	C,T	0.9868	0.3206	-0.514	0.257	4.0132	0.04515
CDKAL1	rs534329540	21629251	R,I	0.4972	0.5784	-0.097	0.048	3.9977	0.04556
CDKAL1	rs1744855	21623715	G,A	0.3803	0.9708	-0.074	0.037	3.9652	0.04645
CDKAL1	rs1744861	21627986	A,T	0.3793	0.9676	-0.074	0.037	3.9605	0.04658
CDKAL1	rs1740837	21633917	C,T	0.6783	0.9174	0.079	0.04	3.8783	0.04891
CDKAL1	rs1740838	21632759	G,T	0.3851	0.7571	-0.082	0.042	3.7777	0.05194
CDKAL1	rs115733488	21695961	T,C	0.9892	0.5305	-0.357	0.184	3.7584	0.05254
CDKAL1	rs7754702	21696403	T,C	0.9482	0.5267	-0.192	0.099	3.742	0.05306

Table S1.2a (continued). Results from eQTL analysis of 6p22 region: endometrial tumour tissue.

CDKAL1	rs111405274	21694954	A,G	0.9508	0.5151	-0.204	0.106	3.7005	0.0544
CDKAL1	rs574039752	21629819	R,D	0.3639	0.9279	-0.074	0.039	3.6944	0.0546
CDKAL1	rs79337490	21695025	T,C	0.9508	0.5151	-0.203	0.106	3.6933	0.05463
CDKAL1	rs59493338	21646242	T,A	0.701	0.8468	0.083	0.044	3.6608	0.05571
CDKAL1	rs1744856	21625895	G,A	0.6341	0.4717	-0.106	0.055	3.6402	0.0564
CDKAL1	rs1740833	21646435	A,G	0.6798	0.9621	0.076	0.04	3.6377	0.05649
CDKAL1	rs111232506	21744922	C,T	0.9798	0.3015	-0.385	0.203	3.5898	0.05814
CDKAL1	rs7772692	21696544	A,G	0.9472	0.5318	-0.184	0.097	3.586	0.05827
Top 25 SNP associations with SOX4 sorted by p-value									
SOX4	rs72175369	21738076	R,D	0.9214	0.781	0.343	0.131	6.8263	0.008983
SOX4	rs7451817	21735497	A,C	0.9341	0.9124	0.337	0.133	6.4457	0.01112
SOX4	rs9358449	21732383	G,A	0.937	0.9915	0.311	0.129	5.8172	0.01587
SOX4	rs12206842	21734489	A,T	0.9372	1	0.31	0.129	5.7854	0.01616
SOX4	rs6941897	21671979	C,A	0.9897	0.6097	-0.865	0.361	5.7599	0.0164
SOX4	rs9358448	21731736	G,A	0.9305	0.9625	0.296	0.126	5.5305	0.01869
SOX4	rs9348484	21731460	C,G	0.93	0.9619	0.295	0.126	5.5252	0.01874
SOX4	rs80177376	21744978	C,T	0.9048	0.3545	0.389	0.176	4.8931	0.02696
SOX4	rs112039884	21729100	C,T	0.9274	0.9676	0.263	0.124	4.5154	0.03359
SOX4	rs71657670	21613526	R,D	0.9864	0.77	-0.705	0.334	4.4501	0.0349
SOX4	rs111442391	21730723	C,T	0.9726	0.3086	0.691	0.329	4.4135	0.03566
SOX4	rs75327712	21728829	G,A	0.9295	0.9778	0.257	0.125	4.2478	0.0393
SOX4	rs79949484	21728966	G,A	0.9319	0.8819	0.272	0.133	4.1475	0.0417
SOX4	rs10946466	21727456	C,A	0.9298	0.9928	0.248	0.124	3.9864	0.04587
SOX4	rs111412021	21606278	T,C	0.9882	0.8163	-0.698	0.354	3.8896	0.04859
SOX4	rs12189901	21726940	G,A	0.9308	0.9638	0.248	0.127	3.8181	0.0507
SOX4	rs570404489	21689796	R,D	0.7798	0.3015	-0.259	0.133	3.7747	0.05203
SOX4	rs79169915	21607457	G,A	0.9882	0.848	-0.657	0.342	3.7008	0.05439
SOX4	rs145545902	21735584	D,R	0.6454	0.9341	-0.127	0.068	3.5366	0.06003
SOX4	rs113455272	21616778	T,C	0.9824	0.9392	-0.468	0.254	3.4099	0.06481
SOX4	rs78923341	21612472	A,C	0.9872	0.895	-0.571	0.312	3.3522	0.06711
SOX4	rs111228858	21612309	G,A	0.9872	0.8945	-0.571	0.312	3.3493	0.06723
SOX4	rs112650104	21611568	C,G	0.9871	0.889	-0.568	0.312	3.3007	0.06925
SOX4	rs138380902	21731187	R,D	0.9079	0.9299	0.199	0.109	3.2972	0.0694
SOX4	rs6935968	21613156	G,A	0.9849	0.9776	-0.49	0.273	3.2116	0.07312

* **TRAIT**: eQTL for which we are testing the marker's association with; **MARKER**: SNP being tested; **BP**: Base position of marker within chromosome 6; **ALLELES**: Allele 1, Allele 2; **FREQ1**: Frequency of Allele 1; **RSQR**: Squared correlation between imputed and true genotypes; **EFFECT2**: Beta estimate using Allele 2 as the risk allele; **STDERR**: Standard error of beta estimate; **CHISQ**: Chi-square statistic of association test; **PVALUE**: P-value of association test.

Table S1.2b. Results from eQTL analysis of 6p22 region: endometrial normal tissue.

*TRAIT	MARKER	BP	ALLELES	FREQ1	RSQR	EFFECT2	STDERR	CHISQ	PVALUE
Top 25 SNP associations with CASC15 sorted by p-value									
CASC15	rs545611803	21688770	R,I	0.8158	0.533	1.287	0.671	3.6777	0.05514
CASC15	rs76714354	21572464	A,G	0.9805	0.3806	86.951	47.656	3.329	0.06807
CASC15	rs74926222	21568994	C,T	0.9816	0.3776	86.545	47.443	3.3277	0.06812
CASC15	rs80177376	21744978	C,T	0.9048	0.3545	0.877	0.481	3.3259	0.0682
CASC15	rs73737558	21577723	A,G	0.9837	0.4293	72.47	43.444	2.7827	0.09529
CASC15	rs150345835	21568080	R,D	0.983	0.3784	74.073	45.702	2.6269	0.1051
CASC15	rs7772335	21696185	C,T	0.8913	0.4593	1.561	0.97	2.5929	0.1073
CASC15	rs78584681	21743570	R,D	0.7497	0.6177	0.648	0.414	2.4505	0.1175
CASC15	rs6925407	21724100	G,C	0.9884	0.9399	-13.662	8.759	2.4329	0.1188
CASC15	rs116779637	21718919	G,T	0.9886	0.7935	-235.67	151.091	2.4329	0.1188
CASC15	rs840985	21744508	G,A	0.5605	0.3093	-0.795	0.532	2.2322	0.1352
CASC15	rs61215435	21721256	C,T	0.9893	0.7083	-66.484	44.79	2.2033	0.1377
CASC15	rs78265086	21720541	T,C	0.9826	0.8636	-105.08	71.153	2.1808	0.1397
CASC15	rs79883278	21720741	C,T	0.9819	0.8353	-76.859	52.548	2.1393	0.1436
CASC15	rs142355149	21692662	A,C	0.9804	0.36	8.946	6.32	2.0036	0.1569
CASC15	rs6933476	21720156	G,A	0.9819	0.8936	-120.77	85.908	1.9764	0.1598
CASC15	rs1744847	21695888	C,T	0.9409	0.4743	1.439	1.046	1.8942	0.1687
CASC15	rs141673420	21695454	T,C	0.9835	0.3745	-4.593	3.348	1.8823	0.1701
CASC15	rs73392015	21591766	C,T	0.9671	0.4088	9.202	6.973	1.7415	0.1869
CASC15	rs74831068	21693978	T,C	0.8189	0.778	0.908	0.691	1.7252	0.189
CASC15	rs7772258	21686985	T,C	0.8466	0.7118	0.982	0.756	1.6861	0.1941
CASC15	rs79232286	21689146	G,A	0.8483	0.7428	0.953	0.742	1.6462	0.1995
CASC15	rs7746995	21686064	G,A	0.9859	0.4629	40.582	31.656	1.6435	0.1999
CASC15	rs80035391	21685357	G,A	0.8973	0.5163	1.156	0.918	1.5882	0.2076
CASC15	rs1853345	21742346	T,C	0.73	0.9808	0.389	0.31	1.5737	0.2097
Top 25 SNP associations with CDKAL1 sorted by p-value									
CDKAL1	rs78584681	21743570	R,D	0.7497	0.6177	0.225	0.078	8.3539	0.003849
CDKAL1	rs74926222	21568994	C,T	0.9816	0.3776	24.845	8.929	7.7431	0.005392
CDKAL1	rs76714354	21572464	A,G	0.9805	0.3806	23.934	8.969	7.1217	0.007616
CDKAL1	rs80177376	21744978	C,T	0.9048	0.3545	0.241	0.09	7.0822	0.007785
CDKAL1	rs114750919	21725361	C,G	0.9867	0.8444	-245.64	102.084	5.79	0.01612
CDKAL1	rs73737558	21577723	A,G	0.9837	0.4293	19.641	8.176	5.7711	0.01629
CDKAL1	rs150345835	21568080	R,D	0.983	0.3784	20.644	8.601	5.7606	0.01639
CDKAL1	rs72175369	21738076	R,D	0.9214	0.781	0.189	0.08	5.5387	0.0186
CDKAL1	rs12189901	21726940	G,A	0.9308	0.9638	0.177	0.078	5.0977	0.02396
CDKAL1	rs75327712	21728829	G,A	0.9295	0.9778	0.174	0.078	5.0258	0.02497
CDKAL1	rs112039884	21729100	C,T	0.9274	0.9676	0.174	0.078	5.0258	0.02497
CDKAL1	rs12206842	21734489	A,T	0.9372	1	0.174	0.078	5.0251	0.02498
CDKAL1	rs10946466	21727456	C,A	0.9298	0.9928	0.174	0.078	5.0251	0.02498
CDKAL1	rs9358449	21732383	G,A	0.937	0.9915	0.174	0.078	5.0223	0.02502
CDKAL1	rs7451817	21735497	A,C	0.9341	0.9124	0.175	0.078	5.0085	0.02522
CDKAL1	rs79949484	21728966	G,A	0.9319	0.8819	0.191	0.086	4.9246	0.02648
CDKAL1	rs9348484	21731460	C,G	0.93	0.9619	0.173	0.078	4.8921	0.02698

Table S1.2b (continued). Results from eQTL analysis of 6p22 region: endometrial normal tissue.

CDKAL1	rs9358448	21731736	G,A	0.9305	0.9625	0.173	0.078	4.8904	0.02701
CDKAL1	rs7739131	21727148	C,T	0.7149	0.9794	0.13	0.064	4.1061	0.04273
CDKAL1	rs138380902	21731187	R,D	0.9079	0.9299	0.145	0.076	3.5995	0.0578
CDKAL1	rs2328627	21714362	G,A	0.9889	0.782	-168.8	90.064	3.5127	0.0609
CDKAL1	rs141898580	21709985	C,T	0.9816	0.8379	-18.782	10.029	3.5075	0.06109
CDKAL1	rs35578300	21741886	R,D	0.7077	0.8309	0.12	0.064	3.5039	0.06122
CDKAL1	rs115395409	21696639	C,T	0.99	0.5674	-2.586	1.382	3.5029	0.06126
CDKAL1	rs1853345	21742346	T,C	0.73	0.9808	0.108	0.058	3.4601	0.06287
Top 25 SNP associations with SOX4 sorted by p-value									
SOX4	rs7739131	21727148	C,T	0.7149	0.9794	0.808	0.311	6.7364	0.009446
SOX4	rs78584681	21743570	R,D	0.7497	0.6177	0.932	0.379	6.0368	0.01401
SOX4	rs145960980	21729330	D,R	0.4093	0.9758	-0.68	0.289	5.5456	0.01853
SOX4	rs9460669	21729251	A,G	0.4053	0.9459	-0.685	0.292	5.4967	0.01905
SOX4	rs7775506	21729962	C,T	0.4216	0.9976	-0.649	0.277	5.4863	0.01917
SOX4	rs2180419	21728317	A,G	0.4028	0.9843	-0.637	0.285	5.0049	0.02528
SOX4	rs1830667	21728089	G,T	0.4027	0.9868	-0.633	0.284	4.9629	0.0259
SOX4	rs1407655	21727902	G,A	0.4028	0.9882	-0.63	0.284	4.9262	0.02645
SOX4	rs7744078	21727822	G,C	0.4006	0.9944	-0.629	0.284	4.9087	0.02672
SOX4	rs7764209	21727755	A,T	0.4006	0.9953	-0.627	0.284	4.8963	0.02692
SOX4	rs7740084	21727531	A,G	0.4005	0.9983	-0.625	0.283	4.8695	0.02733
SOX4	rs10636012	21727515	I,R	0.4062	0.9929	-0.622	0.283	4.8435	0.02775
SOX4	rs11753001	21726506	T,C	0.9147	0.5112	1.556	0.715	4.7325	0.0296
SOX4	rs9460666	21726380	G,A	0.9296	0.3512	2.809	1.373	4.1831	0.04083
SOX4	rs7760462	21730877	G,C	0.4165	0.9846	-0.571	0.285	4.0301	0.04469
SOX4	rs9368332	21731094	G,A	0.4156	0.9974	-0.562	0.284	3.918	0.04777
SOX4	rs35578300	21741886	R,D	0.7077	0.8309	0.609	0.312	3.8106	0.05093
SOX4	rs80177376	21744978	C,T	0.9048	0.3545	0.859	0.441	3.8043	0.05112
SOX4	rs9460665	21724144	G,A	0.8084	1	0.692	0.359	3.7153	0.05392
SOX4	rs6925679	21724480	A,G	0.8086	1	0.692	0.359	3.7153	0.05392
SOX4	rs9466165	21724322	T,C	0.8089	0.9953	0.693	0.36	3.7068	0.05419
SOX4	rs6926491	21724670	G,A	0.8091	0.9926	0.693	0.36	3.7064	0.0542
SOX4	rs7772163	21726011	G,C	0.8088	0.9857	0.693	0.36	3.7057	0.05423
SOX4	rs9460671	21731289	G,C	0.41	0.9868	-0.547	0.284	3.7031	0.05431
SOX4	rs9460664	21723621	G,A	0.8102	0.9858	0.695	0.362	3.6967	0.05452

* **TRAIT**: eQTL for which we are testing the marker's association with; **MARKER**: SNP being tested; **BP**: Base position of marker within chromosome 6; **ALLELES**: Allele 1, Allele 2; **FREQ1**: Frequency of Allele 1; **RSQR**: Squared correlation between imputed and true genotypes; **EFFECT2**: Beta estimate using Allele 2 as the risk allele; **STDERR**: Standard error of beta estimate; **CHISQ**: Chi-square statistic of association test; **PVALUE**: P-value of association test.

Table S1.3. Studies of 6p22.3 genes and cancer using SNP or expression data.

Gene	Study	Study Type	Brief Description	# of Cases	Top SNP (Risk Allele)	Effect	P	R2 to our top SNP	Significant	PMID
CASC15	Maris et al. N Engl J Med. 2008	SNP	GWAS of clinically aggressive neuroblastoma.	1251	rs6939340 (G)	OR: 1.40 (1.26-1.56)	7.01E-10	0.002	Y	18463370
	Diskin et al. Nat Genet. 2012	SNP	GWAS of neuroblastoma.	2101	rs9295536 (A)	OR: 1.357	7.80E-16	0.003	Y	22941191
	Latorre et al. Cancer Epidemiol Biomarkers Prev. 2012	SNP	GWAS of neuroblastoma in African Americans.	390	rs9295536 (C)	OR: 0.90 (0.73-1.10)	0.3	0.003	N	22328350
	He et al. Tumour Biol. 2015	SNP	Hospital-based case control study (Taqman assay) of neuroblastoma in Han Chinese.	201	rs6939340 (A)	OR: 0.53 (0.38-0.74)	AG: 0.006, AA: 0.030	0.002	Y	26307394
	Lessard et al. J Invest Dermatol. 2015	Expression	Tissue microarray with FFPE melanoma LN metastasis specimens: high CASC15 expression is associated with lower 10-year disease free survival	141	N/A	NR	0.002	N/A	Y	26016895
CDKAL1	Kuruma et al. World J Gastroenterol. 2014	SNP	Case-control study (SNPtype assay) of pancreatic cancer in Japanese.	360	rs2206734 (A)	RR- AG: 1.18 (0.85-1.64), AA: 1.21 (0.78-1.89)	NR	0	N	25516658
	Ma et al. Diabetes Res Clin Pract. 2014	SNP	Case-control study (Sequenom MassARRAY) of cancer incidence in Han Chinese with T2D	429	rs7756992 (G)	HR: 0.80 (0.65-1.00)	0.048	0	Y	24468095
	Sainz et al. J Clin Endocrinol Metab. 2012	SNP	Case-control study (KASPar assay) of colorectal cancer risk	1782	rs7754840 (C)	OR: 0.94 (0.85–1.04)	0.03 in males	0	Y	22419714
	Meyer et al. Cancer Epidemiol Biomarkers Prev. 2010	SNP	Case-control study (TaqMan assay) of prostate cancer incidence.	397	rs7754840 (C)	HR: 1.01 (0.87, 1.18)	NR	0	N	20142250
	Figueroa et al. Hum Mol Genet. 2014	SNP	GWAS of bladder cancer	7697	rs4510656 (C)	OR 0.89	6.98E-07	0	Not GWS	24163127
SOX4*	Chen et al. Clin Transl Oncol. 2015	Expression	Meta-analysis of 10 studies with >1000 cancer patients: SOX4 overexpression correlated with poor overall survival	1348	N/A	HR: 1.67 (1.01-2.78)	NR	N/A	Y	26250764
	Song et al. Tumour Biol. 2015	Expression	mRNA and protein expression of SOX4 in breast cancer and adjacent normal: overexpression is an unfavorable prognostic factor regardless of stage, tumor size, metastasis	148	N/A	HR: 1.67 (1.04-2.66)	OS: 0.033	N/A	Y	25592378
	Walter et al. Future Oncol. 2015	Expression	mRNA expression of SOX4 in neuroendocrine tumors of the lung.	60	N/A	N/A	OS: 0.0002	N/A	Y	25804118
	Lu et al. Tumour Biol. 2015	Expression	SOX4 overexpression associated with poor prognosis as measured by tumour recurrence in chondrosarcoma patients.	92	N/A	HR: 3.67 (0.28-48.37)	RFS: 0.035	N/A	Y	25572678
	Wang et al. Mol Cell Biochem. 2015	Expression	SOX4 mRNA and protein expression were markedly higher in NSCLC tissues than in normal lung tissues.	168	N/A	HR: 3.21 (2.06-5.02)	OS: <0.001	N/A	Y	25567207
	Zhou et al. J Cell Biochem. 2015	Expression	High expression levels of SOX4 mRNA were correlated with worse overall survival in Xuanwei females with lung cancer.	96	N/A	NR	OS: < 0.001	N/A	Y	25565486
	Huang et al. 2009 Cancer Res. 2009	Expression	SOX4 overexpressed in endometrial tumors compared with normal tissue from controls without endometrial cancer.	74	N/A	N/A	P < 0.005	N/A	Y	19887623

* Over 150 articles on SOX4 expression and cancer have been published since 1995. Articles published in 2015 and those related specifically to endometrial cancer are presented here.

NR: Not Reported, OS: Overall Survival, RFS: Recurrence Free Survival

Table S1.4. Description of studies included in meta-analysis.

Cohort	Cases	Controls	Subtype	Age Criteria	Platform
ANECS	606	3083	Endometrioid	18-79	Illumina Infinium 610K
SEARCH	681	5190	Endometrioid	18-69	Illumina Infinium 610K, Illumina Infinium 1.2M
NSECG	925	895	All	≤70	Illumina 660K, Illumina Hap550
E2C2	2695	2777	All	>18	Illumina Human OmniExpress, Illumina Human 660W
Total	4907	11945			

Chapter 2

Exome-wide association study of endometrial cancer in a multiethnic population.

Maxine M. Chen[1], Marta Crous-Bou[1,2], Veronica W. Setiawan[3], Jennifer Prescott[2], Sara H. Olson[4], Nicolas Wentzensen[5], Amanda Black[5], Louise Brinton[5], Chu Chen[6], Constance Chen[1], Linda S. Cook[7,8], Jennifer Doherty[9], Christine M. Friedenreich[8], Susan E. Hankinson[2,10], Patricia Hartge[5], Brian E. Henderson[3], David J. Hunter[1], Loic Le Marchand[11], Xiaolin Liang[4], Jolanta Lissowska[12], Lingeng Lu[13], Irene Orlow[4], Stacey Petruzella[4], Silvia Polidoro[14,15], Loreall Pooler[3], Timothy R. Rebbeck[16], Harvey Risch[13], Carlotta Sacerdote[14,15], Frederick Schumacher[17], Xin Sheng[3], Xiao-ou Shu[18], Noel S. Weiss[19], Lucy Xia[3], David Van Den Berg[3], Hannah P. Yang[5], Herbert Yu[11], Stephen Chanock[5], Christopher Haiman[3], Peter Kraft[1], Immaculata De Vivo[1,2]*

[1] Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, Massachusetts, United States of America

[2] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America

[3] University of Southern California, Los Angeles, California, United States of America

[4] Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America

[5] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America

[6] Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

[7] University of New Mexico, Albuquerque, New Mexico, United States of America

[8] Alberta Health Services – CancerControl Alberta, Calgary, Alberta, Canada

[9] Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, United States of America

[10] Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts, United States of America

[11] University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America

[12] Department of Cancer Epidemiology and Prevention, M Sklodowska-Curie Cancer Center and Institute of Oncology, Warsaw, Poland

[13] Yale University School of Public Health, New Haven, Connecticut, United States of America

[14] Center for Cancer Prevention (CPO-Piemonte), Turin, Italy

[15] Human Genetic Foundation (HuGeF), Turin, Italy

[16] Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America

[17] Cancer Prevention Institute of California, Fremont, California, United States of America

[18] Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

[19] University of Washington, Seattle, Seattle, Washington, United States of America

Published in *PLoS One*; 2014. doi: 10.1371/journal.pone.0097045

Exome-Wide Association Study of Endometrial Cancer in a Multiethnic Population

Maxine M. Chen[1], Marta Crous-Bou[1,2], Veronica W. Setiawan[3], Jennifer Prescott[2], Sara H. Olson[4], Nicolas Wentzensen[5], Amanda Black[5], Louise Brinton[5], Chu Chen[6], Constance Chen[1], Linda S. Cook[7,8], Jennifer Doherty[9], Christine M. Friedenreich[8], Susan E. Hankinson[2,10], Patricia Hartge[5], Brian E. Henderson[3], David J. Hunter[1], Loic Le Marchand[11], Xiaolin Liang[4], Jolanta Lissowska[12], Lingeng Lu[13], Irene Orlow[4], Stacey Petruzella[4], Silvia Polidoro[14,15], Loreall Pooler[3], Timothy R. Rebbeck[16], Harvey Risch[13], Carlotta Sacerdote[14,15], Frederick Schumacher[17], Xin Sheng[3], Xiao-ou Shu[18], Noel S. Weiss[19], Lucy Xia[3], David Van Den Berg[3], Hannah P. Yang[5], Herbert Yu[11], Stephen Chanock[5], Christopher Haiman[3], Peter Kraft[1], Immaculata De Vivo[1,2]*

1 Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, Massachusetts, United States of America, **2** Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States of America, **3** University of Southern California, Los Angeles, California, United States of America, **4** Memorial Sloan-Kettering Cancer Center, New York, New York, United States of America, **5** Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, United States of America, **6** Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, **7** University of New Mexico, Albuquerque, New Mexico, United States of America, **8** Alberta Health Services – CancerControl Alberta, Calgary, Alberta, Canada, **9** Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, United States of America, **10** Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, Massachusetts, United States of America, **11** University of Hawaii Cancer Center, Honolulu, Hawaii, United States of America, **12** Department of Cancer Epidemiology and Prevention, M Sklodowska-Curie Cancer Center and Institute of Oncology, Warsaw, Poland, **13** Yale University School of Public Health, New Haven, Connecticut, United States of America, **14** Center for Cancer Prevention (CPO-Piemonte), Turin, Italy, **15** Human Genetic Foundation (HuGeF), Turin, Italy, **16** Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, United States of America, **17** Cancer Prevention Institute of California, Fremont, California, United States of America, **18** Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **19** University of Washington, Seattle, Seattle, Washington, United States of America

Abstract

Endometrial cancer (EC) contributes substantially to total burden of cancer morbidity and mortality in the United States. Family history is a known risk factor for EC, thus genetic factors may play a role in EC pathogenesis. Three previous genome-wide association studies (GWAS) have found only one locus associated with EC, suggesting that common variants with large effects may not contribute greatly to EC risk. Alternatively, we hypothesize that rare variants may contribute to EC risk. We conducted an exome-wide association study (EXWAS) of EC using the Infinium HumanExome BeadChip in order to identify rare variants associated with EC risk. We successfully genotyped 177,139 variants in a multiethnic population of 1,055 cases and 1,778 controls from four studies that were part of the Epidemiology of Endometrial Cancer Consortium (E2C2). No variants reached global significance in the study, suggesting that more power is needed to detect modest associations between rare genetic variants and risk of EC.

Citation: Chen MM, Crous-Bou M, Setiawan VW, Prescott J, Olson SH, et al. (2014) Exome-Wide Association Study of Endometrial Cancer in a Multiethnic Population. PLoS ONE 9(5): e97045. doi:10.1371/journal.pone.0097045

Editor: Paolo Peterlongo, IFOM, Fondazione Istituto FIRC di Oncologia Molecolare, Italy

Received February 24, 2014; **Accepted** April 15, 2014; **Published** May 8, 2014

Funding: The Nurses' Health Study (NHS) investigators are supported by the NCI, NIH Grants Number 1R01 CA134958, 2R01 CA082838, P01 CA087969, and R01 CA49449. MMC is supported by training grant 5T32CA009001-38 from the NCI. MCB is also supported by a Sara Borrell postdoctoral fellowship from the Spanish Ministry of Health, Carlos III Health Institute. The Multiethnic Cohort Study (MEC) is supported by the NCI, NHI Grants Number CA54281, CA128008, and 2R01 CA082838. The Fred Hutchinson Cancer Research Center (FHCRC) is supported by NCI, NIH Grant Number 2R01 CA082838, NIH RO1 CA105212, RO1 CA 87538, RO1 CA75977, RO3 CA80636, NO1 HD23166, R35 CA39779, KO5 CA92002 and funds from the Fred Hutchinson Cancer Research Center. The EDGE Study (Estrogen, Diet, Genetics, and Endometrial Cancer) is supported by NCI, NIH Grants Number 2R01 CA082838, R01CA83918 (Olson PI) and P30-CA008748 (Cancer Center Support Grant). The Alberta Health Services Study (AHS) was supported by operating grants obtained from the National Cancer Institute of Canada with funds from the Canadian Cancer Society and the Canadian Institute for Health Research. The AHS is also supported by NCI, NIH Grant Number 2R01 CA082838. Dr Christine Friedenreich is supported by career awards from Alberta Innovates-Health Solutions and the Alberta Cancer Foundation through the Weekend to End Women's Cancers Breast Cancer Chair. Dr Linda Cook was supported through the Canada Research Chairs program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nhidv@channing.harvard.edu

Introduction

Endometrial cancer (EC), a cancer of the uterine epithelial lining that typically occurs near or after menopause, is the most common cancer of the female reproductive organs and the 10th leading cause of cancer death in women in the developed world [1–3]. EC is strongly associated with estrogen-only post-menopausal hormone therapy [4,5] and excess body weight [6] due to increased aromatization of C-19 steroids by excess adipose tissue [7]. These risk factors support the "unopposed estrogen"

Table 1. Studies participating in the exome-wide association study of endometrial cancer.*

Study	Study Acronym	Study Design	Cases	Controls	Location	Mean BMI at diagnosis (cases)	Mean age at diagnosis (cases)	Total
Alberta Health Services	AHS	Case-Control	517	937	CANADA	32.3	58.5	1454
Estrogen, Diet, Genetics and Endometrial Cancer	EDGE	Case-Control	271	244	USA (NJ)	32.3	60.7	515
Fred Hutchinson Cancer Research Center	FHCRC	Case-Control	55	58	USA (WA)	31.0	60.5	113
Multiethnic Cohort Study	MEC	COHORT	326	659	USA (CA, HI)	28.8	65.5	985
								3067

*Sample size before quality control.
doi:10.1371/journal.pone.0097045.t001

hypothesis in which EC may develop because of the unchecked mitogenic effects of estrogen in the absence of sufficient progesterone [8]. Some studies have shown that family history increases risk two to three-fold in younger women who have a first-degree female relative with EC [9,10], while among older women the association is less strong. In addition, there is an increased risk of EC in women with Lynch syndrome [11], a hereditary autosomal dominant condition that confers a high risk of colorectal cancer as well. These observations suggest that germline genetics may contribute to EC susceptibility.

Genome-wide association studies (GWAS) have successfully identified more than a hundred susceptibility loci for a variety of cancer types [12]. Three GWAS studies of EC have been conducted to date with only one identifying a novel genome-wide significant locus, rs4430796, ($p = 7.1 \times 10^{-10}$) associated with EC [13] at the *HNF1B* gene region on chromosome 17q12. Two independent studies subsequently replicated the association with rs4450796 [14,15]. However, two other GWAS studies of EC [14,16] were not able to identify additional genome-wide significant loci, suggesting that common variants with large effects may not highly contribute to the familial risk of EC.

Most risk alleles discovered through GWAS have modest effect sizes that do not account for much heritability of common diseases [17]. Moreover, GWAS studies have focused on common variants (>5%) in the general population. Low frequency variants make up a large fraction of genetic variation in humans and may explain a substantial portion of the heritability in cancer etiology. Recent exome-sequencing studies have found rare variants in candidate susceptibility genes for familial colorectal cancer [18], breast cancer [19], and prostate cancer [20], suggesting that analysis of rare variants may also provide insight into the etiology of EC. However, exome-sequencing studies require samples sizes that are not amenable to large epidemiological studies due to the high cost currently needed to achieve sufficient statistical power.

There has been a push to develop statistically powerful, yet relatively inexpensive, methods to detect associations for rare variants with larger effect sizes. Illumina has recently developed the Infinium HumanExome BeadChip (exome array) from non-synonymous variants found at least 3 times on more than 2 data sets from the whole-exome sequencing of more than 12,000 individuals. This array provides a platform from which we can begin to survey the landscape of rare variation in a large number of samples.

We genotyped rare variants in a multiethnic population of 3,067 women (1,169 EC cases and 1,898 controls) from the Epidemiology of Endometrial Cancer Consortium (E2C2) [21] in order to test the hypothesis that rare variants in coding regions may be associated with EC risk.

Methods

Ethics committee from each participating study (Alberta Health Services; Estrogen, Diet, Genetics and Endometrial Cancer Study; Multiethnic Cohort Study) obtained written informed consent from all study participants. All written consent was approved from the Institutional Review Board (IRB) from each institution (Alberta Health Services, Canada; Memorial Sloan Kettering, USA; University of Hawaii Cancer Center, USA; Keck School of Medicine-University of Southern California, USA).

Alberta Health Services, Memorial Sloan Kettering, University of Hawaii Cancer Center, and University of Southern California institutional review boards specifically approved the present study (Exome-Wide Association Study of Endometrial Cancer), as well as the written consent obtained from participants.

Table 2. Cases and Controls by Reported Ethnicity and Study.

	Alberta		EDGE		FHCRC		MEC	Total		
	Case	Control	Case	Control	Case	Control	Case	Control	Case	Control
Caucasian	446	866	196	177	1	0	0	0	643	1043
Latina	0	0	8	8	17	26	98	203	123	237
Asian	0	0	2	0	0	0	117	228	119	228
African American	0	0	18	8	8	15	68	137	94	160
Hawaiian	0	0	0	0	0	0	26	53	26	53
Unknown	27	40	2	4	21	13	0	0	50	57
Total	473	906	226	197	47	54	309	621	1055	1778

doi:10.1371/journal.pone.0097045.t002

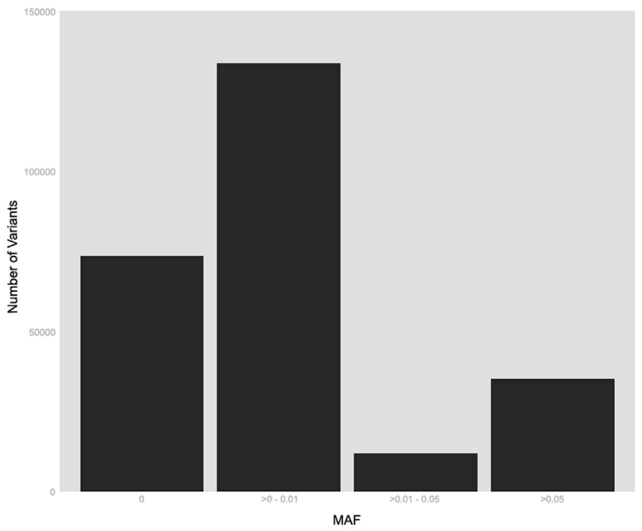

Figure 1. Minor allele frequency for all variants successfully genotyped over all ethnicities. The number of variants is plotted by the minor allele frequency over all ethnicities. These variants include those that are monomorphic in all ethnicities.
doi:10.1371/journal.pone.0097045.g001

Participating studies also obtained IRB certification, permitting data sharing according to the NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome- Wide Association studies (GWAS).

Study Population

Exome array genotyping was performed on 3,067 samples from 3 retrospective case-control studies: the Alberta Health Services Study (AHS) [22], the Estrogen, Diet, Genetics and Endometrial Cancer study (EDGE) [23], and the Fred Hutchinson Cancer Research Center (FHCRC) study and 1 case-control study nested within the prospective Multiethnic Cohort Study (MEC) [24]. Studies participating in this analysis are described in Table 1 and in our previous GWAS[14]. Of the women included in the study, 1,169 were EC cases and 1,898 were controls. Cases were restricted to those diagnosed with the most common subtype of EC (type I) while controls were cancer free and had an intact uterus. Controls were matched to cases by age and study site.

Genotyping and Quality Control

DNA was extracted at each study site from buffy coat or cheek-cell samples following the manufacturer's protocol and genotyped at the University of Southern California using the Infinium Human Exome BeadChip (Illumina Inc., San Diego, CA) as part of the Stage II replication of the E2C2 GWAS. The BeadChip included 9,232 custom markers, 2,211 of which are specifically relevant to EC, in addition to the 247,870 markers coding primarily for protein-altering variants already included in the BeadChip's default design.

Genotype calling was performed with Illumina GenCall on all samples (n = 3,067) using the MEC cluster file (16,000 multiethnic samples) for the non-custom markers and autoclustering for the custom markers. Variants were excluded from analyses if call rates were < 90% (n = 115), the variant was monomorphic (n = 77,521), the loci had no observed founders and missing all genotypes (n = 1,962), the variant was an insertion or deletion allele (n = 117), or the variant deviated from Hardy-Weinberg equilibrium at p-value < 0.0001 in any ethnic group (n = 248).

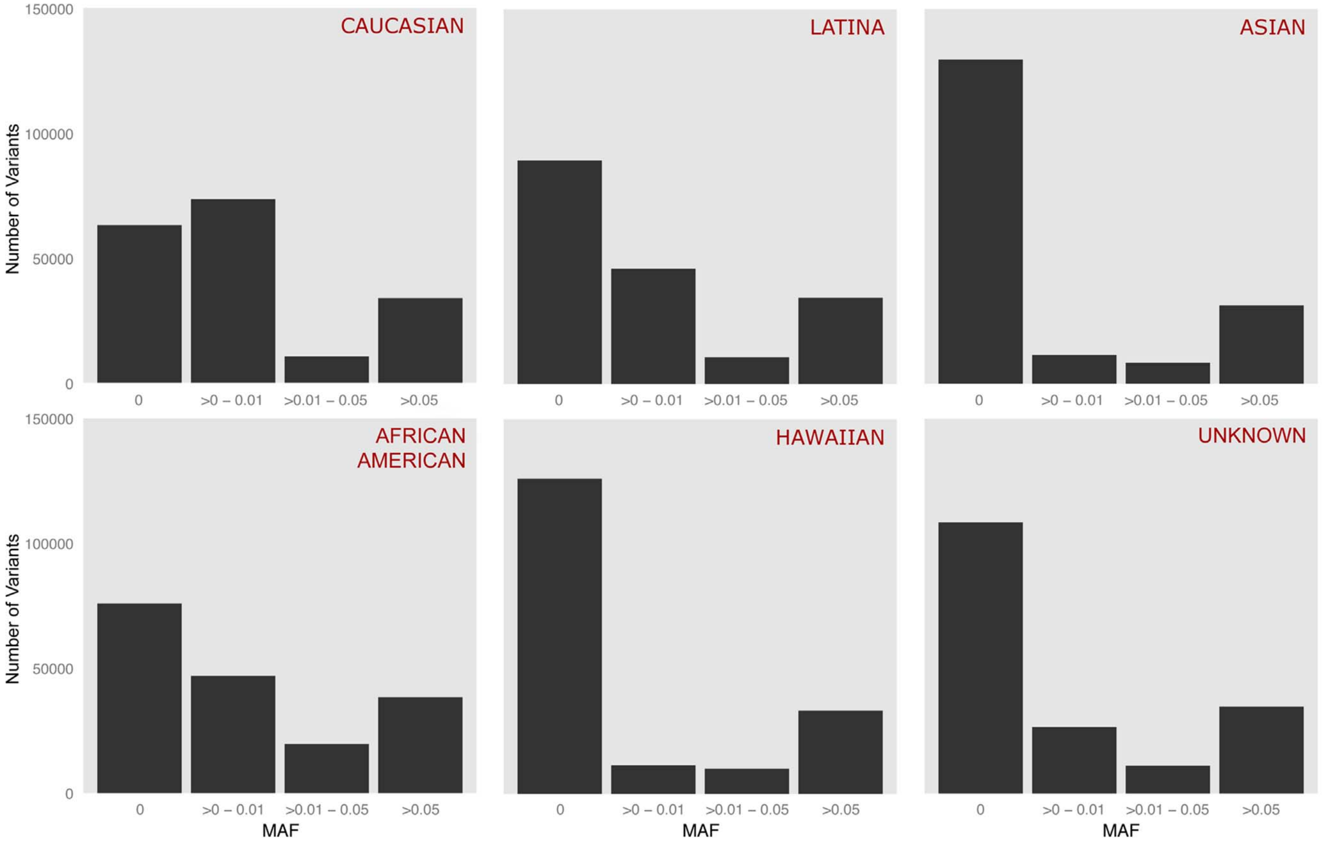

Figure 2. Minor allele frequency for all variants successfully genotyped by reported ethnicity. The number of variants is plotted by the minor allele frequency for each ethnicity. All these variants are polymorphic in at least one reported ethnicity.
doi:10.1371/journal.pone.0097045.g002

The final disease trait analysis data set contained 177,139 successfully genotyped variants.

In total, 3,031 out of 3,067 samples were successfully genotyped with call rates ≥ 90%. Of these, we removed 40 duplicate samples (genotype concordance rate > 99.9%) used for assay quality control and 15 samples for other quality control reasons. We conducted principal components analysis (PCA) to identify self-reported ethnicity outliers and infer ancestry with EIGENSOFT v 4.2 [25] using 47,097 custom and non-custom SNPs with genotyping rates > 90% and MAF > 1%. The HapMap phase II (build 37) CEU, YRI, and JPT-CHB samples were used as reference populations. Using the first 5 principal components, we determined 7 individuals that were ethnicity outliers and excluded them from analyses. After further removal of 136 outliers (more than 3.5 standard deviations from the mean) of sample heterozygosity by ethnicity, 2,833 women (1,055 EC cases and 1,778 controls) remained for disease trait analysis.

Statistical Analysis

Single variant association analysis. Single variant analyses were performed overall and stratified by self-reported ethnic group. For each SNP, we estimated odds ratios (OR) and 95% confidence intervals (CI) using unconditional logistic regression, assuming an additive genetic model (0, 1, 2 copies of the minor allele) and adjusting for body mass index (BMI in kg/m^2), age, study site, plate, and the first 4 principal components to account for population stratification. All single variant analyses were performed using PLINK v 1.07 [26].

Gene-based analysis. As an additional method to discover rare variants associated with EC, gene-based testing was performed using SKAT-O [27] over all ethnicities. SKAT-O combines gene-burden tests and SKAT, a SNPset level test for association using kernel machine methods, in special cases for an optimized approach that maximizes power. These analyses were also adjusted for BMI, age, study site, plate and the first 4 principal components. In total, 16,245 genes with at least one variant were tested.

Statistical significance. We determined single variant association to reach global significance if the unadjusted p-value was $<2.82 \times 10^{-7}$, corresponding to a Bonferroni correction for 177,139 tests. Gene-based associations were considered significant for unadjusted p-values $<3.08 \times 10^{-6}$, corresponding to a Bonferroni correction for 16,245 tests.

In accordance to NIH/NCI policy all data will be submitted to the database of Genotypes and Phenotypes (dbGaP, http://www.ncbi.nlm.nih.gov/gap).

Results

Association analyses included 177,139 successfully genotyped variants with MAF > 0 from a total of 257,102 variants included in the array. Population characteristics of the four participating studies (AHS, EDGE, FHCRC, and MEC) are described in Table 1. Mean age at diagnosis for cases ranged from 58.5 years in AHS to 65.5 years in MEC and mean BMI at diagnosis for cases ranged from 28.8 kg/m^2 in MEC to 32.3 kg/m^2 in AHS and EDGE. Of the 3,067 samples genotyped, 2,833 were included in

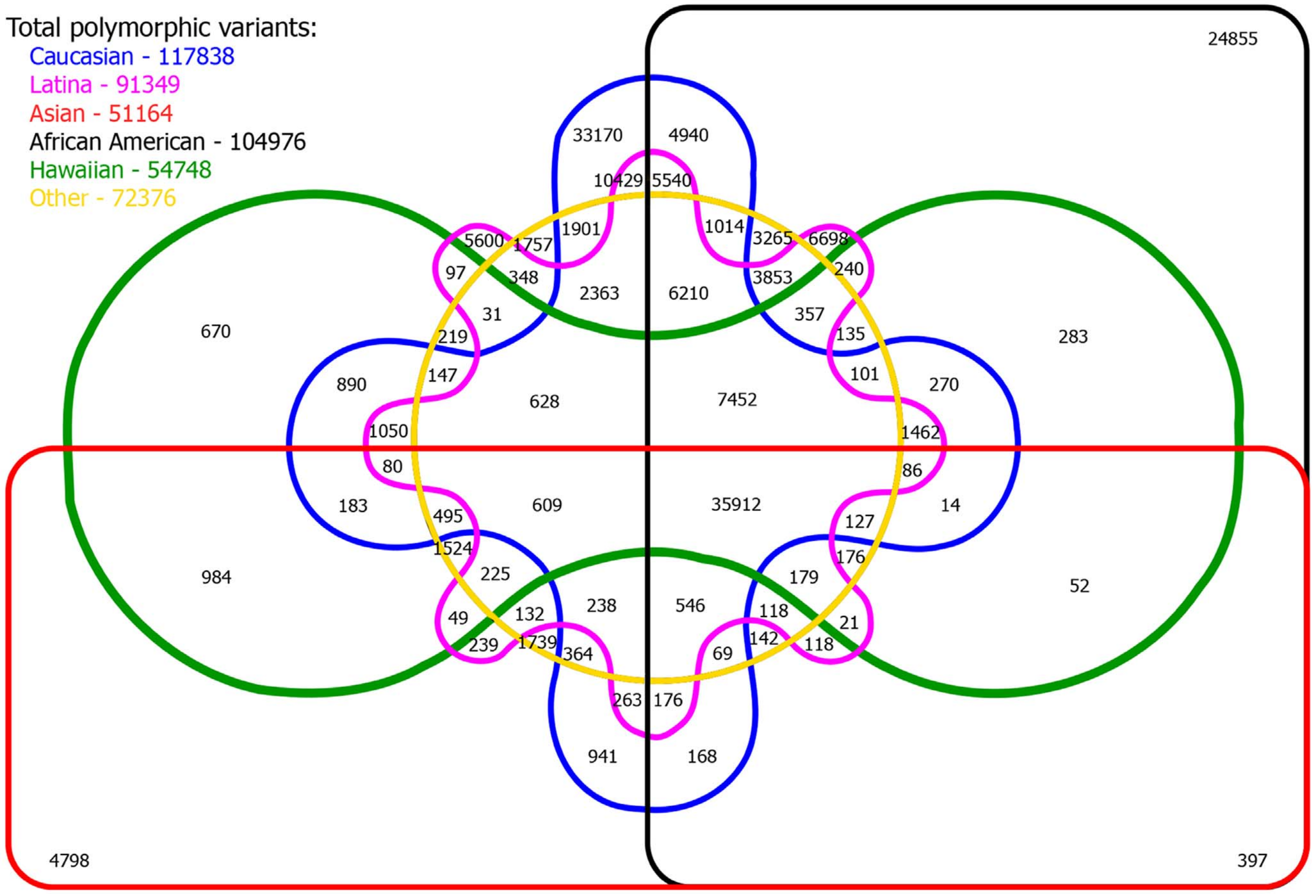

Figure 3. Six-way Venn diagram showing polymorphic putative functional variants shared by reported ethnicities. Numbers of shared variants are shown at intersections. The total numbers of polymorphic variants by ethnicity are listed in the upper-left hand corner.
doi:10.1371/journal.pone.0097045.g003

the analysis. There were no differences in age, BMI, and ethnicity between excluded cases and those included in the analysis (results not shown). Of these 2,833 individuals, there were 254 self-reported African-Americans, 347 self-reported Asians, 1,686 self-reported Caucasians, 79 self-reported Hawaiians, 360 self-reported Latinas, and 107 who did not report a specific ethnicity (Table 2).

Variant Distribution among Reported Ethnicities

In this study population, 77,521 variants (30.4%) were found to be monomorphic across all reported ethnicities and 177,139 variants (69.6%) were polymorphic in at least one ethnic population with 74.0% of polymorphic alleles having MAF ≤ 1% (Figure 1). Of the variants that were polymorphic in at least one ethnic population, 42.0% in African Americans, 71.7% in Asians, 34.9% in Caucasians, 69.7% in Hawaiians, 49.5% in Latinas, and 60.0% in those of unknown ethnicity were monomorphic (Figure 2). The MAF distributions were fairly similar among Asians, Hawaiians, and those who did not report a specific ethnicity while African Americans, Caucasians, and Latinas shared more similarities in MAF with each other than with Asians, Hawaiians, and those of unknown ethnicity. About 20.2% (n = 35,912) of variants were shared by all 5 reported ethnicities while Caucasians and Latinas had the most variants in common at 41.1% (n = 72,878) (Figure 3). Caucasians had the most unique polymorphic variants (18.7%), followed by African-Americans (14.0%), Latinas (3.2%), Asians (2.7%), those who did not report ethnicity (1.0%), and Hawaiians (0.4%).

Single Variant Association for Endometrial Cancer

No variants reached global significance in single variant association of EC for all ethnicities combined (Figure 4a, Table 3) when correcting for multiple comparisons using the Bonferroni adjustment ($p < 2.82 \times 10^{-7}$). The strongest associations were for variants with >0.05 MAF (Table 3) located within 50 kb of the long non-protein coding intergenic RNA, *LINC00520* (rs1953358, OR = 1.36, $p = 4.76 \times 10^{-7}$) and in the intron region of *PROS1* (rs8178648, OR = 1.71, $p = 1.53 \times 10^{-6}$), which codes for protein S, a cofactor to protein C in the anti-coagulation pathway. In Caucasians, who make up the majority of the overall analysis, only rs8178648 remained suggestively associated with OR = 1.98 and $p = 3.35 \times 10^{-6}$ (Figure 4b, Table 3). There were no globally significant or suggestive variants in African Americans, Asians, Hawaiians, Latinas, and those who did not report ethnicity (Table S1).

Gene-based Analysis of Endometrial Cancer

None of the gene-based tests of association were globally significant ($p < 3.08 \times 10^{-6}$) after adjusting for multiple comparisons (Table S2). Of the 16,245 genes tested, the most significant EC association was with *KRT81* ($p = 2.21 \times 10^{-5}$), a member of the keratin gene family located on 12q13. *PROS1*, where rs8178648 is located, was not significantly associated with EC (p = 0.6789) when testing over all ethnicities neither when testing only in Causasians (results not shown).

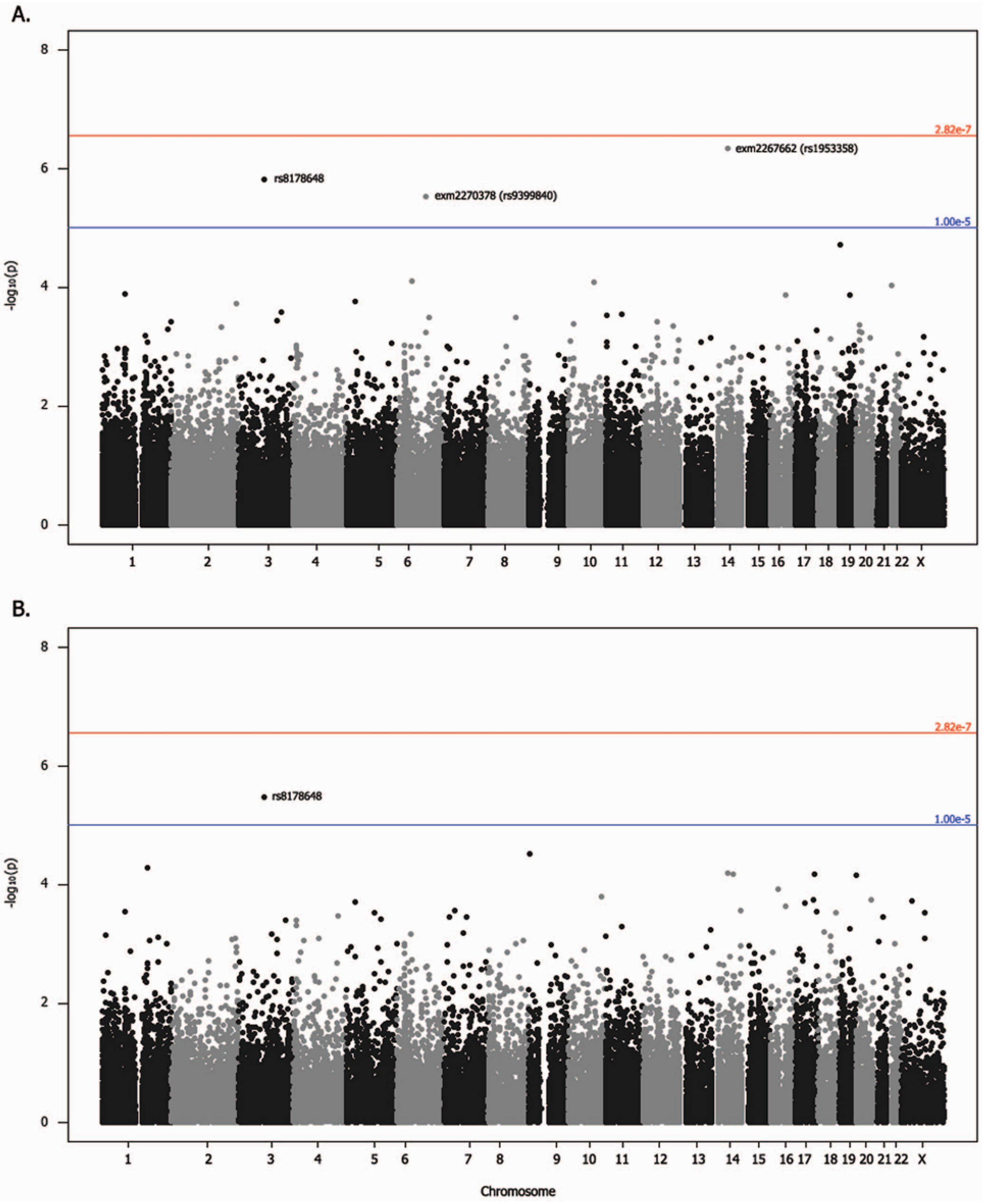

Figure 4. Manhattan plots for the endometrial cancer association analysis. Results of single variant analyses ($-\log_{10}p$) are plotted against chromosome position (NCBI build 37) for association over all ethnicities (A) and for associations within Caucasians (B). Suggestive variants are labeled above. Results were adjusted for age at diagnosis, BMI, study site, plate, and the first four principal components.
doi:10.1371/journal.pone.0097045.g004

Table 3. Top five most significant associations of single coding variants with endometrial cancer risk.*

All Cases (n = 1055) vs. Controls (n = 1778)

Variant	Chr	Position (bp)	Gene/Locus	A1	A2	MAF (all)	OR (95% CI)	P-value
exm2267662 (rs1953358)	14	56295580	LINC00520	G	A	0.49	1.36 (1.20, 1.53)	4.76E-07
rs8178648	3	93605739	PROS1	G	A	0.09	1.71 (1.37, 2.12)	1.53E-06
exm2270378 (rs9399840)	6	104076463	n/a	C	T	0.47	0.75 (0.67, 0.85)	3.01E-06
exm1401784	19	1796166	ATP8B3	T	C	0.23	0.72 (0.61, 0.83)	1.92E-05
exm558041 (rs6926980)	6	56917538	KIAA1586	A	G	0.23	0.75 (0.65, 0.87)	7.95E-05

Caucasian Cases (n = 639) vs. Caucasian Controls (n = 1042)

Variant	Chr	Position (bp)	Gene/Locus	A1	A2	MAF (all)	OR (95% CI)	P-value
rs8178648	3	93605739	PROS1	G	A	0.09	1.98 (1.49, 2.65)	3.35E-06
exm736725 (rs10974657)	9	4622453	SPATA6L	C	T	0.09	2.34 (1.57, 3.50)	3.00E-05
rs10753688	1	165666448	ALDH9A1	C	T	0.41	1.43 (1.20, 1.70)	5.18E-05
exm2267662 (rs1953358)	14	56295580	LINC00520	G	A	0.49	0.71 (0.60 0.84)	6.49E-05
exm1113971 (rs141549345)	14	74401030	LOC283922	A	G	0.03	0.36 (0.22, 0.59)	6.56E-05

*Adjusted for age at diagnosis, BMI at diagnosis, study site, plate, and the first four principal components.
doi:10.1371/journal.pone.0097045.t003

Discussion

We present an initial exploration into whether rare variants are associated with EC risk in a multiethnic population from the E2C2. No variants reached global significance ($p < 2.82 \times 10^{-7}$) in the single variant association analyses of EC in all ethnicities combined or when stratified by reported ethnicity. Additionally, no gene-based test of association reached global significance ($p < 3.08 \times 10^{-6}$).

Among all ethnicities, rs8178648 on chromosome 3 maintained a suggestive association with EC (OR = 1.707, 95% CI: 1.363–2.123, $p = 1.53 \times 10^{-6}$). The variant lies within the intron region of *PROS1*, a gene coding for protein S, a cofactor in the anticoagulant pathway that causes autosomal dominant hereditary thrombophilia when mutated [28]. *PROS1* expression has been reported to be elevated in aggressive prostate cancer tissue [29] and thyroid cancer tissue [30], suggesting it may have a role in cancer etiology or progression. *PROS1* has been found to be directly upregulated by progestins [31] and downregulated by 17β-Estradiol, an estrogen that regulates gene expression via the estrogen receptor [32], making it susceptible to imbalances in the sex hormone metabolic pathway, which is implicated in EC etiology. However, *PROS1* was not significantly associated with EC ($p = 0.6789$) when using SKAT-O and no other GWAS have found significant or suggestive variants in this gene.

One weakness of this study is our limited sample size, which was not sufficiently powered to detect rare variants with modest effects associated with EC. Additionally, the exome array content is predominantly based on European ancestry whereas our study included a substantial number of samples with other ancestries. Incomplete exome array coverage of all functional variants and indels that may impact EC risk may also have limited the scope of our study. However, our analysis is one of only two studies [33] using the exome array to examine associations between rare variants and complex diseases in large multiethnic populations. Our study is also the first to utilize the exome array with EC and serves as an extension to our previous examination of common variants on EC risk.

A previous GWAS [13] identified one novel locus near *HNF1B*, rs4430796, inversely associated with EC risk. We replicated the findings in our GWAS [14], but no other common variants associated with EC have been determined. Exome arrays that focus on rare variants, which are hypothesized to have larger effect sizes than common variants, have been used to successfully identify new loci influencing insulin processing and secretion in type 2 diabetics [34]. To date, analyses of cancer sites using exome arrays have failed to find strong evidence that rare variants are highly associated with cancer, revealing only one variant significantly associated with breast cancer and none with prostate cancer [33]. Similarly, we have not identified any loci significantly associated with EC. Due to our limited sample size, our study was estimated to be sufficiently powered to detect ORs > 2.53 for low frequency variants (MAF = 0.02). An OR of 2.00 (MAF = 0.01) would also need around 4,250 cases and 7,250 controls to be sufficiently powered. Even for variants with higher MAFs similar to what was observed for rs8178648, a study detecting a per-allele OR of 1.70 would require at least 1,107 cases and 1,871 controls to be considered sufficiently powered ($\beta = 0.80$). Therefore, larger studies need to be conducted in order to detect novel associations with rare variants.

In conclusion, our study found no evidence that rare variants with large effect sizes are associated with EC risk. Though we were able to identify a few suggestive associations, as with rs8178648, much larger studies would be needed to identify a more modest influence of rare variants on the risk of EC.

Supporting Information

Table S1 1–5. Single variant association results. Top 100 most significant single variant associations with endometrial cancer by ethnic groups.
(XLSX)

Table S2 SKAT-O gene based association results. SKAT-O gene based associations with endometrial cancer for all ethnicities combined.
(XLSX)

Author Contributions

Conceived and designed the experiments: IDV PK. Analyzed the data: MMC MCB. Wrote the paper: MMC MCB IDV. Reviewed the analysis and results and provided insightful comments on the manuscript: VWS JP SHO NW AB LB CC CC LSC JD CMF SEH PH BEH DJH LLM XL JL LL IO SP SP LP TRR HR CS FS XS X-oS NSW LX DVDB HPY HY SC CH.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E, et al. (2011) Global cancer statistics. CA Cancer J Clin 61: 69–90. doi: 10.3322/caac.20107.
2. Bray F, Ren JS, Masuyer E, Ferlay J (2013) Global estimates of cancer prevalence for 27 sites in the adult population in 2008. Int J Cancer 132: 1133–1145. doi: 10.1002/ijc.27711.
3. American Cancer Society. (2013) Cancer Facts & Figures 2013. American Cancer Society, Atlanta.
4. Persson I, Adami HO, Bergkvist L, Lindgren A, Pettersson B, et al. (1989) Risk of endometrial cancer after treatment with oestrogens alone or in conjunction with progestogens: results of a prospective study. BMJ 298: 147–151.
5. Grady D, Gebretsadik T, Kerlikowske K, Ernster V, Petitti D (1995) Hormone replacement therapy and endometrial cancer risk: a meta-analysis. Obstet Gynecol 85: 304–313. doi: 10.1016/0029-7844(94)00383-o.
6. Bergstrom A, Pisani P, Tenet V, Wolk A, Adami HO (2001) Overweight as an avoidable cause of cancer in Europe. Int J Cancer 91: 421–430.
7. Schmandt RE, Iglesias DA, Co NN, Lu KH (2011) Understanding obesity and endometrial cancer risk: opportunities for prevention. Am J Obstet Gynecol 205: 518–525. doi: 10.1016/j.ajog.2011.05.042.
8. Kaaks R, Lukanova A, Kurzer MS (2002) Obesity, endogenous hormones, and endometrial cancer risk: a synthetic review. Cancer Epidemiol Biomarkers Prev 11: 1531–43.
9. Gruber SB, Thompson WD (1996) A population-based study of endometrial cancer and familial risk in younger women. Cancer and Steroid Hormone Study Group. Cancer Epidemiol Biomarkers Prev 5: 411–7.
10. Lucenteforte E, Talamini R, Montella M, Dal Maso L, Pelucchi C, et al. (2009) Family history of cancer and the risk of endometrial cancer. Eur J Cancer Prev 18: 95–99. doi: 10.1097/CEJ.0b013e328305a0c9.
11. Vasen HF, Watson P, Mecklin JP, Jass JR, Green JS, et al. (1994) The epidemiology of endometrial cancer in hereditary nonpolyposis colorectal cancer. Anticancer Res 14: 1675–1678.
12. Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, et al. (2009) A Catalog of Published Genome-Wide Association Studies. NHGRI. http://www.genome.gov/gwastudies. Accessed 14 May 2013.
13. Spurdle AB, Thompson DJ, Ahmed S, Ferguson K, Healey CS, et al. (2011) Genome-wide association study identifies a common variant associated with risk of endometrial cancer. Nat Genet 43: 451–454. doi: 10.1038/ng.812.
14. De Vivo I, Prescott J, Setiawan VW, Olson SH, Wentzensen N, et al. (2014) Genome-wide association study of endometrial cancer in E2C2. Hum Genet 133: 211–224. doi: 10.1007/s00439-013-1369-1.
15. Setiawan VW, Haessler J, Schumacher F, Cote ML, Deelman E, et al. (2012) HNF1B and endometrial cancer risk: results from the PAGE study. PLoS One 7: e30390. doi: 10.1371/journal.pone.0030390.
16. Long J, Zheng W, Xiang YB, Lose F, Thompson D, et al. (2012) Genome-wide association study identifies a possible susceptibility locus for endometrial cancer. Cancer Epidemiol Biomarkers Prev 21: 980–987. doi: 10.1158/1055-9965.epi-11-1160.
17. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109: 1193–1198. doi: 10.1073/pnas.1119675109.

18. Gylfe AE, Katainen R, Kondelin J, Tanskanen T, Cajuso T, et al. (2013) Eleven candidate susceptibility genes for common familial colorectal cancer. PLoS Genet 9: e1003876. doi: 10.1371/journal.pgen.1003876.
19. Thompson ER, Doyle MA, Ryland GL, Rowley SM, Choong DY, et al. (2012) Exome sequencing identifies rare deleterious mutations in DNA repair genes FANCC and BLM as potential breast cancer susceptibility alleles. PLoS Genet 8: e1002894. doi: 10.1371/journal.pgen.1002894.
20. Fitzgerald LM, Kumar A, Boyle EA, Zhang Y, McIntosh LM, et al. (2013) Germline missense variants in the BTNL2 gene are associated with prostate cancer susceptibility. Cancer Epidemiol Biomarkers Prev 22: 1520–1528. doi: 10.1158/1055-9965.epi-13-0345.
21. Olson SH, Chen C, De Vivo I, Doherty JA, Hartmuller V, et al. (2009) Maximizing resources to study an uncommon cancer: E2C2—Epidemiology of Endometrial Cancer Consortium. Cancer Causes Control 20: 491–496. doi: 10.1007/s10552-008-9290-y.
22. Friedenreich CM, Cook LS, Magliocco AM, Duggan MA, Courneya KS (2010) Case-control study of lifetime total physical activity and endometrial cancer risk. Cancer Causes Control 21: 1105–1116. doi: 10.1007/s10552-010-9538-1.
23. Fortuny J, Sima C, Bayuga S, Wilcox H, Pulick K, et al. (2009) Risk of endometrial cancer in relation to medical conditions and medication use. Cancer Epidemiol Biomarkers Prev 18: 1448–1456. doi: 10.1158/1055-9965.epi-08-0936.
24. Kolonel LN, Henderson BE, Hankin JH, Nomura AM, Wilkens LR, et al. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol 151: 346–357.
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909. doi: 10.1038/ng1847.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575. doi: 10.1086/519795.
27. Lee S, Wu MC, Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13: 762–775. doi: 10.1093/biostatistics/kxs014
28. Andersen BD, Bisgaard ML, Lind B, Philips M, Villoutreix B, et al. (2001) Characterization and structural impact of five novel PROS1 mutations in eleven protein S-deficient families. Thromb Haemost 86: 1392–1399.
29. Saraon P, Musrap N, Cretu D, Karagiannis GS, Batruch I, et al. (2012) Proteomic profiling of androgen-independent prostate cancer cell lines reveals a role for protein S during the development of high grade and castration-resistant prostate cancer. J Biol Chem 287: 34019–34031. doi: 10.1074/jbc.M112.384438.
30. Griffith OL, Melck A, Jones SJ, Wiseman SM (2006) Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers. J Clin Oncol 24: 5043–5051. doi: 10.1200/jco.2006.06.7330.
31. Hughes Q, Watson M, Cole V, Sayer M, Baker R, et al. (2007) Upregulation of protein S by progestins. J Thromb Haemost 5: 2243–2249. doi: 10.1111/j.1538-7836.2007.02730.x.
32. Suzuki A, Sanda N, Miyawaki Y, Fujimori Y, Yamada T, et al. (2010) Down-regulation of PROS1 gene expression by 17beta-estradiol via estrogen receptor alpha (ERalpha)-Sp1 interaction recruiting receptor-interacting protein 140 and the corepressor-HDAC3 complex. J Biol Chem 285: 13444–13453. doi: 10.1074/jbc.M109.062430.
33. Haiman CA, Han Y, Feng Y, Xia L, Hsu C, et al. (2013) Genome-wide testing of putative functional exonic variants in relationship with breast and prostate cancer risk in a multiethnic population. PLoS Genet 9: e1003419. doi: 10.1371/journal.pgen.1003419.
34. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stancakova A, et al. (2013) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet 45: 197–201. doi: 10.1038/ng.2507.

Table S2.1.1: Single variant association results for Asians

Chromosome #	SNP	Unadjusted P-value	Bonferroni Adjusted P-Value
11	exm885609	5.92E-05	1
12	exm1045438	7.83E-05	1
12	exm1045327	8.69E-05	1
9	exm765224	8.84E-05	1
16	exm1213609	0.0001298	1
12	exm1045314	0.0001671	1
11	exm-rs716274	0.0001908	1
15	exm1148780	0.0002289	1
3	exm292415	0.0003752	1
8	rs17716313	0.0005914	1
6	exm558041	0.000607	1
22	exm-rs5751614	0.0006081	1
9	exm776342	0.0006112	1
2	exm2261151	0.0006663	1
15	exm1148781	0.0008642	1
16	exm1272937	0.0008675	1
4	exm398469	0.0009138	1
10	exm2267042	0.0009394	1
3	exm350059	0.001004	1
23	exm2268576	0.001141	1
15	exm1154888	0.001147	1
12	exm2267431	0.001212	1
16	exm2267874	0.001213	1
1	exm-rs1192415	0.001217	1
17	rs12951993	0.001233	1
16	exm2267916	0.001236	1
16	exm2260592	0.00124	1
3	exm350067	0.001256	1
11	exm2277008	0.001269	1
16	exm-rs153782	0.001401	1
13	exm2267587	0.001425	1
12	exm1010442	0.001449	1
14	exm2272159	0.001502	1
9	exm2262604	0.001519	1
2	exm2269090	0.001523	1
9	exm-rs11243897	0.001529	1
1	exm118456	0.001564	1
2	exm200258	0.001738	1
11	rs7126796	0.001862	1
3	rs12493155	0.001879	1
11	rs11022755	0.001911	1
5	exm510285	0.00194	1
12	exm2260080	0.001949	1
7	rs11767887	0.001958	1
4	exm2256634	0.001977	1
3	exm2255690	0.001985	1
11	exm2250161	0.002114	1
11	rs4146388	0.002204	1
20	exm1517960	0.002218	1
11	exm-rs7941030	0.002259	1
1	exm87958	0.002298	1

Table S2.1.1 (continued): Single variant association results for Asians

2	exm2265342	0.002316	1
9	exm-rs4366181	0.002325	1
11	rs12290622	0.002335	1
11	rs1481891	0.002335	1
11	rs10832017	0.002335	1
1	exm-rs11805303	0.002378	1
5	exm449239	0.002378	1
19	exm1421052	0.002402	1
9	exm-rs10983238	0.002419	1
9	rs11141494	0.002421	1
10	exm-rs704010	0.002451	1
17	rs8182331	0.002453	1
13	exm1059306	0.002473	1
8	exm684306	0.002478	1
1	exm45228	0.002499	1
10	exm853566	0.002539	1
17	exm1352677	0.002565	1
11	exm965604	0.002618	1
18	exm1372005	0.002674	1
11	rs7107287	0.002688	1
7	exm636394	0.002708	1
18	exm1388919	0.002733	1
4	exm435282	0.00274	1
11	exm961573	0.002746	1
8	exm730169	0.002768	1
3	exm2255751	0.00279	1
1	exm2254080	0.002801	1
15	exm2252136	0.00285	1
15	exm1150082	0.00285	1
2	exm2254461	0.002861	1
9	exm2271104	0.002864	1
16	exm2252452	0.002936	1
11	exm2273473	0.002974	1
5	exm2265961	0.002985	1
5	exm-rs11249661	0.003038	1
2	exm220598	0.003092	1
2	exm-rs12477314	0.003168	1
2	exm2269055	0.003221	1
3	exm340449	0.003224	1
7	exm654195	0.003247	1
8	exm2262480	0.003406	1
19	exm1456338	0.003457	1
1	exm38236	0.003461	1
3	rs2005618	0.003527	1
1	exm153094	0.003564	1
11	exm965464	0.003601	1
11	exm965436	0.003601	1
2	exm-rs2268363	0.003613	1

Table S2.1.2: Single variant association results for African-Americans

Chromosome #	SNP	Unadjusted P-value	Bonferroni Adjusted P-Value
5	exm2266031	1.63E-05	1
5	exm463076	5.43E-05	1
5	exm463057	7.19E-05	1
18	custom_18-46372096	0.0001135	1
6	exm518563	0.0001463	1
8	exm690789	0.0002483	1
4	exm2269837	0.0002537	1
18	custom_18-46385146	0.0002857	1
11	exm-rs10437653	0.0003017	1
11	exm930530	0.0003107	1
11	exm-rs1387153	0.0003639	1
3	exm2269446	0.000448	1
4	exm2269905	0.0004591	1
18	custom_18-46367489	0.0004801	1
11	exm2250173	0.0004848	1
11	exm930634	0.0004848	1
2	exm-rs6738825	0.0005052	1
3	exm364452	0.0005141	1
17	exm1275605	0.0005188	1
7	exm643860	0.0005368	1
2	exm2269382	0.000546	1
10	exm2259513	0.0005543	1
11	exm2267254	0.0005847	1
18	custom_18-46382325	0.0006624	1
1	exm166750	0.0006862	1
3	exm2269482	0.0007456	1
21	exm1564210	0.0007881	1
10	exm816552	0.0008464	1
10	exm816522	0.0008464	1
10	exm816567	0.0008464	1
2	exm2273361	0.0009571	1
13	exm-rs17369571	0.0009591	1
6	exm584901	0.0009821	1
2	exm228898	0.0009932	1
18	custom_18-46376821	0.001014	1
20	exm1523944	0.001027	1
2	exm2265350	0.001046	1
18	custom_18-46381502	0.001079	1
18	custom_18-46381891	0.001079	1
18	custom_18-46381543	0.001079	1
12	exm2271867	0.001088	1
11	exm930441	0.001109	1
1	rs6541017	0.001121	1
11	exm920401	0.001122	1
10	exm-rs2281880	0.001173	1
1	exm2253243	0.001314	1
15	exm1149977	0.001315	1
11	exm913057	0.001321	1
6	exm2262096	0.001386	1
18	rs2078131	0.001404	1
9	exm-rs7025486	0.001513	1

Table S2.1.2 (continued): Single variant association results for African Americans

23	exm2263187	0.001521	1
14	exm1102109	0.001575	1
7	exm-rs10253361	0.00159	1
9	exm-rs4366181	0.001597	1
2	exm2265254	0.001639	1
20	exm1534109	0.001679	1
6	exm573837	0.00168	1
20	exm1525055	0.001718	1
12	exm1005604	0.00175	1
4	exm390487	0.001792	1
1	rs10800956	0.001839	1
11	exm955691	0.001844	1
10	exm811848	0.001851	1
18	exm2268119	0.001863	1
14	rs2693694	0.001904	1
2	exm-rs10172646	0.001907	1
9	exm770191	0.001915	1
9	exm770208	0.001915	1
14	exm2251950	0.001921	1
3	rs874151	0.001927	1
3	rs73224955	0.001927	1
3	exm360527	0.001956	1
1	exm2232887	0.002061	1
18	exm1387188	0.002079	1
18	exm2268085	0.002163	1
12	exm975791	0.002165	1
8	exm695233	0.002187	1
11	rs9645657	0.00219	1
5	exm468680	0.002222	1
3	exm365274	0.002225	1
17	rs7407003	0.002238	1
17	rs9898816	0.002238	1
2	exm-rs13031237	0.002249	1
10	exm807357	0.002263	1
6	exm529518	0.002301	1
14	exm2251898	0.002315	1
11	exm930937	0.002326	1
10	rs2884127	0.002361	1
6	exm568967	0.002363	1
17	exm1285270	0.002372	1
3	exm-rs1435703	0.002373	1
12	exm-rs7965445	0.002377	1
19	exm1452707	0.002379	1
6	exm-rs406238	0.002399	1
12	exm-rs7315621	0.002429	1
3	rs56374105	0.00245	1
1	exm-rs2274910	0.002517	1
6	exm-rs1077394	0.002535	1

Table S2.1.3: Single variant association results for Hawaiians

Chromosome #	SNP	Unadjusted P-value	Bonferroni Adjusted P-Value
15	exm1185372	0.0009239	1
15	exm1185366	0.0009239	1
15	exm1185518	0.0009349	1
15	exm1185392	0.0009349	1
15	exm1185487	0.0009349	1
15	exm1185480	0.0009349	1
15	exm1185460	0.0009349	1
15	exm1185450	0.0009349	1
11	exm876605	0.0009683	1
22	exm-rs242076	0.001072	1
3	exm-rs991258	0.001161	1
3	exm-rs3773643	0.001189	1
1	rs11589267	0.001371	1
11	exm875622	0.001375	1
14	exm-rs12431733	0.001404	1
15	exm1146681	0.001429	1
3	exm353478	0.001617	1
6	exm587799	0.001711	1
9	exm-rs842304	0.001783	1
7	exm624358	0.001834	1
4	exm2269873	0.001859	1
2	exm173753	0.001944	1
2	exm173893	0.001944	1
2	exm173743	0.001944	1
21	rs2826487	0.002069	1
15	exm1146708	0.002095	1
18	exm2268117	0.002162	1
9	rs1307279	0.0022	1
9	exm-rs6474694	0.002248	1
1	exm165415	0.002324	1
1	exm-rs10493340	0.002399	1
8	exm734237	0.002401	1
2	exm2269153	0.00241	1
9	exm753957	0.002446	1
2	exm2254604	0.002479	1
2	rs4331558	0.002505	1
14	exm1084268	0.002546	1
5	exm2264133	0.002616	1
11	exm-rs1393350	0.002713	1
7	exm2258484	0.002731	1
22	exm2268419	0.002733	1
22	exm-rs240343	0.002738	1
7	rs3918181	0.002809	1
8	exm-rs1835740	0.002827	1
22	exm2268423	0.002833	1
2	exm265772	0.002852	1
2	exm2254527	0.002927	1
4	exm398022	0.002945	1
6	exm512358	0.00308	1
11	exm958649	0.003139	1
12	exm2251234	0.003149	1

Table S2.1.3 (continued): Single variant association results for Hawaiians

8	exm732117	0.003195	1
11	exm946837	0.003283	1
9	exm2266901	0.003493	1
13	exm-rs17086609	0.00353	1
3	rs1523060	0.003536	1
7	exm-rs10224002	0.00357	1
22	exm-rs132628	0.003679	1
23	exm-rs2430212	0.003727	1
17	rs4968857	0.003861	1
15	exm-rs3212335	0.003864	1
18	exm2268076	0.004062	1
16	exm1263301	0.004103	1
23	exm2268470	0.004223	1
4	exm397941	0.00425	1
11	custom_11-111059568	0.004282	1
4	exm403814	0.004286	1
11	rs10502135	0.004324	1
7	exm2258303	0.004355	1
7	exm596878	0.004355	1
23	exm2268512	0.004372	1
6	exm2257795	0.004399	1
21	exm-rs2836754	0.004403	1
11	rs9645657	0.004423	1
11	custom_11-111064723	0.004423	1
11	custom_11-111055744	0.004423	1
10	exm862777	0.004446	1
19	exm-rs8099917	0.004453	1
11	rs12274451	0.004497	1
16	exm2272428	0.004509	1
6	exm-rs793834	0.004518	1
6	exm-rs10455248	0.004548	1
18	exm2268086	0.00462	1
1	rs2100516	0.004661	1
12	exm1022813	0.004902	1
9	rs7846809	0.004965	1
14	exm2272066	0.004973	1
11	custom_11-111080246	0.004976	1
11	custom_11-111081590	0.004976	1
11	custom_11-111082062	0.004976	1
11	custom_11-111087889	0.004976	1
11	custom_11-111081140	0.004976	1
2	exm269902	0.004991	1
20	exm1550771	0.005016	1
8	exm732084	0.00505	1
3	exm366819	0.005055	1
18	exm1378750	0.00509	1
18	exm1378823	0.00509	1
3	exm372903	0.005101	1

Table S2.1.4: Single variant association results for Latinas

Chromosome #	SNP	Unadjusted P-value	Bonferroni Adjusted P-Value
19	exm1490555	1.59E-05	1
6	exm2266273	5.02E-05	1
19	exm1408667	8.63E-05	1
12	exm1042018	0.0001501	1
7	exm-rs2429582	0.0001872	1
9	rs10971520	0.0002028	1
22	exm2255082	0.000326	1
8	exm719372	0.0003627	1
7	exm2270585	0.0003924	1
3	rs7616008	0.0004201	1
20	rs13041173	0.0004583	1
9	rs3119692	0.0004881	1
6	exm554218	0.0004963	1
3	rs62270253	0.0004976	1
19	exm1408655	0.0005272	1
21	exm1563904	0.0006965	1
3	rs77356594	0.0007369	1
3	rs62270252	0.0007369	1
3	rs12632496	0.0007369	1
3	rs12633064	0.0007373	1
10	exm-rs2631681	0.0007833	1
13	rs1414318	0.0008057	1
1	exm149177	0.0008281	1
6	exm2264184	0.0008305	1
8	exm723247	0.0009351	1
14	exm2260295	0.0009738	1
19	exm-rs3865444	0.0009963	1
6	exm594521	0.001028	1
3	exm305550	0.001091	1
11	exm2250364	0.001152	1
5	exm456072	0.001153	1
1	exm146787	0.001156	1
6	exm574754	0.001165	1
11	exm883596	0.001182	1
19	exm1435303	0.001192	1
12	exm1014783	0.0012	1
23	exm1654811	0.001217	1
3	rs9852437	0.001234	1
3	rs12632239	0.001304	1
11	exm-rs297325	0.001359	1
3	rs17393618	0.001447	1
8	exm2262499	0.001457	1
10	exm807539	0.001496	1
18	exm2253431	0.001506	1
4	exm2265752	0.001537	1
1	exm-rs11264625	0.001558	1
3	rs1868172	0.001604	1
6	exm-rs12210887	0.00167	1
1	exm-rs6427356	0.001676	1
7	exm2266378	0.001681	1
18	rs4798367	0.001843	1

Table S2.1.4 (continued): Single variant association results for Latinas

9	rs4879707	0.002063	1
11	exm2250302	0.002071	1
17	exm2253161	0.002076	1
12	exm2251021	0.002101	1
1	exm116693	0.002182	1
2	rs849511	0.002197	1
4	exm2265768	0.002226	1
7	exm2273436	0.002288	1
14	exm-rs8017161	0.002333	1
17	exm1322546	0.002338	1
3	exm2269583	0.002406	1
20	exm1521580	0.002418	1
5	exm2266069	0.002437	1
3	exm305462	0.002458	1
3	exm339992	0.002462	1
3	exm-rs6803290	0.002477	1
1	exm32146	0.00252	1
9	rs10813982	0.00259	1
1	rs10874888	0.002598	1
16	exm2272437	0.002622	1
5	exm452656	0.002673	1
1	exm2249926	0.002678	1
9	rs11103218	0.002771	1
11	exm-rs1357339	0.002812	1
6	exm554324	0.002844	1
10	exm-rs2893923	0.003011	1
6	exm2266355	0.003042	1
15	1kg_15-61137875	0.003136	1
10	exm812431	0.003222	1
2	exm195234	0.00334	1
9	exm-rs1980889	0.003347	1
10	exm807345	0.003378	1
1	exm32337	0.003407	1
5	exm2270197	0.003487	1
16	exm2272449	0.003564	1
18	exm-rs7236477	0.003671	1
17	exm1301695	0.003681	1
6	exm-rs4715166	0.003685	1
7	exm634822	0.003751	1
5	exm-rs35391	0.003782	1
19	exm-rs4072910	0.003819	1
19	exm2268219	0.003864	1
8	rs4243863	0.003907	1
18	exm2272696	0.003908	1
6	exm-rs9491140	0.003921	1
9	exm771317	0.003937	1
19	exm2268189	0.003957	1
16	exm1271100	0.004037	1

Table S2.1.5: Single variant association results for those with unknown ethnicity

Chromosome #	SNP	Unadjusted P-value	Bonferroni Adjusted P-Value
15	exm2267766	0.0001504	1
9	exm802933	0.0004482	1
2	rs3770244	0.0004595	1
6	exm-rs1592404	0.0005002	1
6	exm-rs9405124	0.0005002	1
7	exm2270665	0.0005839	1
1	exm48643	0.0006707	1
19	exm1440681	0.0007402	1
16	exm1262021	0.0007678	1
3	exm2255974	0.0008015	1
9	rs706134	0.0008338	1
9	exm2271182	0.0009934	1
14	rs1022714	0.001008	1
6	exm-rs2856717	0.001027	1
6	exm-rs2647012	0.001027	1
8	rs11136727	0.001077	1
14	rs12435927	0.001109	1
14	rs1804799	0.001109	1
14	rs10483802	0.001109	1
6	exm-rs9275141	0.001157	1
8	exm721520	0.001168	1
8	rs60806454	0.001174	1
5	exm2256958	0.0012	1
2	exm2269147	0.001234	1
17	exm-rs2138852	0.001242	1
18	exm-rs8084703	0.001253	1
8	rs10087922	0.001298	1
5	exm2270175	0.0013	1
15	exm1162598	0.001356	1
2	exm2265151	0.001379	1
21	exm-rs2825388	0.0014	1
19	exm1398517	0.001428	1
6	exm-rs206018	0.001445	1
6	exm-rs3130171	0.00146	1
20	exm-rs3790268	0.001474	1
4	exm2256417	0.001523	1
4	rs3774937	0.001568	1
8	exm2258600	0.001645	1
11	rs11022759	0.001748	1
5	exm461047	0.001852	1
13	exm2271916	0.001855	1
1	rs16829304	0.001883	1
6	exm-rs1265048	0.001916	1
4	exm-rs12651106	0.001916	1
14	exm2272057	0.001923	1
6	exm-rs3094549	0.001947	1
13	exm1076135	0.001956	1
9	exm2266812	0.001957	1
2	exm263537	0.002011	1
14	exm1100483	0.002078	1
14	exm1100436	0.002078	1

Table S2.1.5 (continued): Single variant association results for those with unknown ethnicity

9	exm779121	0.002097	1
9	exm2271122	0.00213	1
1	exm70355	0.002207	1
5	exm2265917	0.002225	1
6	exm-rs3869115	0.00228	1
2	exm253142	0.002388	1
19	exm1483213	0.00241	1
23	exm1649561	0.002412	1
1	exm2268742	0.002499	1
1	exm75532	0.002637	1
7	exm-rs4510766	0.002649	1
5	exm501284	0.00267	1
11	exm910171	0.002691	1
12	exm1014631	0.002718	1
11	exm2267271	0.002719	1
11	exm907378	0.002719	1
11	exm907410	0.002719	1
3	exm291505	0.002746	1
9	exm743762	0.002782	1
12	exm1014783	0.00286	1
19	exm1415151	0.002882	1
1	exm-rs4845552	0.002885	1
15	exm2267784	0.002924	1
11	exm2250464	0.002956	1
11	exm2271608	0.00297	1
8	rs10108639	0.002974	1
2	rs993598	0.002984	1
6	exm-rs1612904	0.003053	1
6	exm-rs9275596	0.003053	1
1	exm-rs1935881	0.003068	1
1	rs6679342	0.00308	1
4	exm423281	0.003095	1
11	exm2249573	0.003108	1
21	exm2273010	0.003108	1
3	rs11915310	0.003126	1
9	rs706127	0.003139	1
6	exm536445	0.003157	1
2	exm2269209	0.003182	1
7	exm2270615	0.003244	1
11	exm2218005	0.003355	1
16	exm1263247	0.003537	1
6	exm-rs7744001	0.003549	1
2	exm262503	0.003554	1
17	exm1350119	0.003589	1
1	exm773	0.003628	1
17	rs4357980	0.00366	1
10	exm865498	0.003663	1
11	exm909579	0.003768	1

Table S2.2. Top 50 SKAT-O gene based association results by unadjusted p-value (all ethnicities)

Gene ID	# of Variants in Gene Set	Unadjusted P-value	Bonferroni Adjusted P-Value
KRT81	4	2.21E-05	0.36
C3orf33	4	1.40E-04	1
C17orf81	4	1.95E-04	1
COL5A3	22	3.03E-04	1
BCL2L12	4	3.07E-04	1
FAM161B	9	3.96E-04	1
ATR	18	4.83E-04	1
SULF1	9	5.79E-04	1
FAM38B	7	6.85E-04	1
MAT2B	2	6.89E-04	1
EIF5A	1	7.81E-04	1
EIF5AL1	1	7.81E-04	1
C1orf95	1	1.14E-03	1
KRT7	5	1.15E-03	1
RUSC1	8	1.15E-03	1
NAA20	2	1.20E-03	1
TTLL9	6	1.24E-03	1
TDRD12	2	1.27E-03	1
PJA2	6	1.28E-03	1
DEFB112	3	1.29E-03	1
WDR72	14	1.31E-03	1
C1orf104	7	1.43E-03	1
PDCD5	1	1.44E-03	1
BCL7A	3	1.49E-03	1
GFOD2	3	1.66E-03	1
PRDM1	9	1.66E-03	1
WDR7	9	1.69E-03	1
ARMC2	7	1.77E-03	1
CCDC155	7	1.93E-03	1
ALLC	6	1.97E-03	1
SNAPC2	10	2.00E-03	1
TM4SF4	1	2.00E-03	1
CCDC39	23	2.04E-03	1
C16orf3	3	2.21E-03	1
TAS2R5	9	2.22E-03	1
ZFAND2A	4	2.28E-03	1
MARVELD3	4	2.33E-03	1
ZNF780B	5	2.42E-03	1
NFATC3	7	2.45E-03	1
SLC4A1	13	2.46E-03	1
EPHB4	6	2.48E-03	1
CCDC89	2	2.48E-03	1
BCAS1	8	2.54E-03	1
ASZ1	5	2.61E-03	1
EZR	4	2.61E-03	1
UNC45A	9	2.63E-03	1
SLC2A7	9	2.63E-03	1
PRKCB	5	2.66E-03	1
ASMTL	5	2.68E-03	1
COLEC12	9	2.73E-03	1

Chapter 3

Mutation analysis of endometrial cancer in a population-based study by targeted next-generation sequencing

Maxine M. Chen[1], Marta Crous-Bou[1,2], Michael J. Downing[3], Jennifer Prescott[2], George L. Mutter[3], Immaculata De Vivo[1,2]

[1] Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA 02115, USA

[2] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

[3] Department of Pathology, Division of Women's and Perinatal Pathology, Harvard Medical School, Brigham and Women's Hospital, Boston, MA 02115, USA.

Abstract

Endometrial carcinoma (EC), a malignancy that arises from the epithelial lining of the uterus, is heterogeneous at histologic and molecular levels. Risk factors and outcomes also differ by type. Though studies have characterized the genomic landscape of endometrial carcinoma, few have integrated histologic, clinical, and prospectively collected epidemiologic data in to the analysis. We have collected formalin-fixed paraffin embedded tumor tissue from women diagnosed with EC between 1976 and 2012 who were enrolled in the Nurses' Health Study, a large ongoing prospective cohort study. Through targeted next-generation sequencing, we interrogated 50 cancer related genes to identify genetic variants in 37 ECs and correlate findings with immunohistochemical, histologic, and epidemiologic data. Mutations most frequently occurred in *TP53* (57%), *PTEN* (46%), and *PIK3CA* (38%). *TP53* mutations were seen in 83% of ECs that immunostained positive for mutant p53, with the most frequent *TP53* mutations occurring in R248. Well-differentiated tumors had an elevated ($p < 0.05$) frequency of *PTEN* and *PIK3CA* mutation. The mutation profiles of these samples are consistent with previous studies, supporting the viability of archival paraffin embedded tissue in mutation detection. This study's interdisciplinary approach to tumor characterization may help inform future development of personalized models for EC.

Introduction

Endometrial carcinoma (EC) is a heterogeneous disease that originates from the epithelial lining of the uterus. It is the most common gynecological malignancy among females in the developed world[1]. Between 2005 and 2011, 18.3% of women diagnosed with EC in the United States did not survive more than five years, a case fatality percentage greater than that of breast cancer (10.6%)[2]. Established risk factors for EC include excess body weight (2.5-fold increased risk)[3] and estrogen-only post-menopausal hormone therapy (9.5-fold increased risk with 10 or more years of use)[4]. Cigarette smoking is known to reduce EC risk by about 20%[5].

EC includes endometrioid and non-endometrioid histotypes. Endometrioid tumors may have squamous or mucinous differentiation, whereas non-endometrioid tumors encompass the disparate pathogenetic subtypes of serous, clear cell, and carcinosarcomas [6]. Serous carcinomas are the most common non-endometrioid tumors, which are generally more aggressive and have poorer prognosis than endometrioid tumors. Development of endometrioid-type EC has been traditionally attributed to the mitogenic effects of excess estrogen from established environmental exposures in the absence of sufficient progesterone[7,8]. Non-endometrioid tumors are historically considered estrogen-independent.

New evidence has called into question the broad categorization of EC into two subtypes with different estrogen dependencies. A pooled analysis out of the Epidemiology of Endometrial Cancer Consortium observed that the risk factor patterns of high-grade endometrioid tumors and non-endometrioid tumors were similar[9], suggesting that unopposed estrogen may increase the risk of non-endometrioid tumors as well. Additionally, the Cancer Genome Atlas (TCGA) performed a comprehensive genomic analysis of endometrial cancer and identified four subtypes based on genomic alterations that revealed the similarity between high-grade endometrioid cancers and serous carcinomas based on mutation presence[10]. Thus, to improve our understanding of EC heterogeneity, studies must be undertaken to integrate genomic alterations, established risk factors, and histological characteristics.

Investigating the molecular profile of EC tumors may also refine the diagnostic tools used to assess histological subtype and prognosis. *TP53* is a tumor suppressor that is frequently mutated in cancers, often stabilizing p53 protein in a way that can be detected by immunostaining[11]. P53 mutation, which is indicated by positive immunostaining, is associated with poor prognosis[12–14] and often distinguishes serous carcinomas from endometrioid carcinomas[15]. However, p53 positive immunostaining is also found in some endometrioid tumors and not all serous carcinomas stain positive for p53 mutation. Insight into the mutational landscape of p53 and other diagnostic markers may improve marker classification and have future clinical implications.

In this study, we performed targeted next-generation sequencing of 50 cancer-related genes in 37 EC cases with from the Nurses' Health Study (NHS) that were immunostained for p53 protein and have detailed histologic information. The NHS is a population-based cohort that has prospectively collected environmental exposure information dating back to 1976, providing a unique opportunity to correlate the mutations that arise in the tumor not only with p53 immunohistochemistry and tumor histology, but also with the lifestyle exposure history of these cases. The purpose of this study is to characterize mutations from cancer-related genes that arise in EC and incorporate clinical data, histologic information, and exposure history for a comprehensive analysis of EC cases from a population-based cohort.

Methods

Study population and sample collection.

In 1976, the Nurses' Health Study prospective cohort enrolled 121,700 female resident nurses from the United States between the ages of 30 and 55. Self-administered questionnaires on lifestyle exposures and medical histories were obtained at baseline and every two years thereafter. Greater than 90% response rates have been achieved for each follow-up cycle.

Self-reported cases of incident endometrial carcinoma with no prior history of cancer were confirmed by medical record review. Participants were asked for permission to collect formalin-fixed paraffin embedded tissue blocks containing representative samples of the endometrial carcinoma. After consent, specimens were obtained from the participant's pathology department. Hematoxylin and eosin-stained slides were centrally reviewed by a gynecologic pathologist (GLM) based on published criteria[16,17].

Immunohistochemistry.

From each surgical specimen, three representative 0.6mm diameter cores from one paraffin block of the primary endometrial tumor were planted in a tissue microarray. Serial sections (4um) of each tissue microarray were stained for the following hematoxylin and eosin, p53 (Leica Biosystems clone PAb 1801; 1:300), cytokeratin (Dako clone AE1/AE3; 1:200) in replicate independent staining runs. Following incubation with primary antibody, slides were washed and incubated with an appropriate biotinylated secondary antibody, and signal was detected by addition of avidin peroxidase in a chromogenic reaction carried out with 3-3' diaminobenzadine to yield a brown reaction product. Whole stained slides were digitally scanned at 40x optical magnification by a Hamamatsu Nanozoomer scanner. Images were visually scored for tumor nuclear staining positivity across a 90% threshold. Cases with >90% of nuclei staining were considered "p53 protein abnormal", a staining pattern indicating mutational stabilization and accumulation of the p53 protein, and lesser staining proportions considered "p53 protein wild type". Duplicate stains were independently scored for marker specific signal within tumor cells, and discordant replicates resolved by re-review. Because of the limited tissue quantity available in the tissue cores, we were unable to reliably score p53 protein null mutants. The expected density of p53 nuclear staining cells in wild type cell populations was insufficient to be robustly represented amongst the three 0.6mm cores.

Case selection.

A total of 40 primary endometrial carcinomas from hysterectomy specimens (20 p53 protein abnormal and 20 p53 protein wild type) were selected to maximize power to assess genomic differences between p53 abnormal and wild type cases. Mixed type endometrial tumors, carcinosarcomas, poorly stained specimens, and specimens with inadequate amount of remaining tumor were excluded. Cases were further selected to represent balanced numbers of Stage 1 (n=23) vs Stage 2 or higher (n=17) disease.

DNA extraction, library preparation and next generation sequencing.

Paraffin tissue cores from representative tumor areas were treated with deparaffinization solution and digested with proteinase K. DNA was isolated by automated extraction with the QIAamp DNA FFPE Tissue kit (QIAGEN, Hilden, Germany) on the QIAcube system. Genomic DNA (gDNA) was quantified using the TaqMan RNaseP Detection Reagents kit (Thermo Fisher Scientific, Foster City, CA). One sample was excluded for insufficient gDNA, leaving 39 samples for library preparation and sequencing. Barcoded libraries were prepared using Life Technologies' (Carlsbad, CA) Ion AmpliSeq™ Library Kit v2.0 according to manufacturer's protocol. Briefly, 10ng of gDNA from each sample was amplified using the Ion AmpliSeq Cancer Hotspot Panel v2 primer pool, which covers 50 oncogenes and tumor suppressor genes (Table S3.1). Primer sequences were partially digested to facilitate ligation of IonXpress™ Barcode Adapters to the amplicons. Barcoded libraries were quantified using the the Ion Library Quantitation Kit. After quantitation, barcoded libraries were combined to create multiplexed libraries with a final concentration of 8pM. Emulsion PCR was performed using the Ion OneTouch™ 2 instrument and template-positive Ion Sphere Particles were enriched using the Ion OneTouch™ ES system according to manufacturer's protocol. Thirty-nine samples were loaded onto Ion 318v2 chips and sequenced on the IonTorrent Personal Genome Machine (PGM™) at 500 flows using the Ion PGM™ Sequencing 200 Kit v2.

Quality control and data analysis.

Sequencing reads were aligned to hg19 using the in-house Torrent Suite software (version 5.0.3) and variants were called using the Variant Caller plug-in. Variants were excluded from further analysis if quality score <= 20, base coverage <= 500x, and global minor allele frequency >= 0.01. Two samples had identical mutation profiles and were excluded from the analysis, leaving 37 samples for variant annotation. Variants were annotated with ANNOVAR (2016Feb01). Only nonsynonymous mutations with minor allele frequencies < 1% and hotspot variants found in the Catalogue of Somatic Mutations in Cancer (COSMIC)[18] were included in further analysis. Statistical analyses and data visualization was performed using R v3.0.2, GenVisR package for R, or cBioPortal (OncoPrinter and MutationMapper)[19,20]. Statistical correlations between genes with hotspot variants and tumor characteristics or exposure history were assessed using Fisher's exact test or the Kruskal-Wallis test when appropriate. Significance was assessed at $p < 0.05$.

Results

Case Characteristics

Clinicopathological characteristics and exposure history of cases sequenced are summarized in Table 3.1. The mean age at diagnosis for all cases was 70.4 years. The mean body mass index (BMI) at diagnosis for all cases was 27.9. Those who were p53 protein abnormal by immunohistochemistry had a lower mean BMI of 26.5 compared to those who were p53 protein wild type (BMI: 29.3). Twenty-five women had a BMI of 25 or greater. Ever smokers comprised 38% of all cases.

There were 23 cases with endometrioid carcinoma, 11 cases with serous carcinoma, and 3 cases with clear cell carcinoma. About 90% of p53 protein wild type tumors were endometrioid whereas 56% of the p53 protein abnormal tumors were serous. Of all cases, 57% had stage I tumors and 43% had tumors of stage II or greater. Stage I tumors accounted for 72% of those who were p53 abnormal and 42% of those who were p53 wild type. There were eleven endometrioid grade I cases, all of whom were p53 wild type, five endometrioid grade II cases, and seven endometrioid grade III cases. Serous and clear cell cases were not graded and account for 67% of p53 abnormal tumors. In practice, they can be considered high grade tumors.

Immunohistochemistry

Immunostaining for p53 was performed on slides cut from paraffin blocks with identifiable tumor tissue for all cases. Nineteen cases were p53 protein wild type. Figure 3.1a is representative of a p53 wild type immunostain. Eighteen cases were p53 protein abnormal. P53 abnormal immunostains were characterized by strong nuclear staining in >90% of tumor cells (Figure 3.1b), indicating excess p53 accumulation.

Hotspot Mutation Profile Over All Cases

The lowest average depth of sequencing achieved was 1500x and all samples had at least 92% of target amplicons covered at >500x. Genes that were mutated in at least one sample are summarized in Figure 3.2. Three samples, all endometrioid subtype, did not have any non-synonymous hotspot mutations after filtering. The median number of mutations identified per sample was 3 (range 0-8). Out of 111 total mutations across all samples, there were 97 missense mutations, 9 nonsense mutations, and 5 frameshift mutations. The most frequently mutated genes among our samples were *TP53* (21 cases, 57%)*, PTEN* (17 cases, 46%), and *PIK3CA* (14 cases, 38%). *PTEN* had the most nonsense mutations (6/9, 67%) out of all the genes in our panel.

Hotspot Mutation Profile by p53 Staining Outcome

Characteristics of the hotspot mutation profiles by p53 immunohistochemistry are summarized in Figure 3.2 and Figure S3.1. *TP53* hotspot mutations were found in 15 out of 18 (83%) p53 protein abnormal tumors and in 6 out of 19 (32%) p53 protein wild type tumors. All five *TP53* mutated endometrioid tumors that were p53 wild type were grade 1 (Table 3.2). The remaining p53 wild type tumor with mutated *TP53* was of clear cell histology. *PTEN* hotspot mutations were more prevalent in p53 wild type tumors (12/19, 63%) than in p53 abnormal tumors (5/18, 28%). *PIK3CA* was fairly evenly distributed across p53 immunohistochemical profiles, with 6 out of 18 samples mutated among p53 abnormal tumors and 8 out of 19 samples mutated among p53 wild type tumors. Additionally, *KRAS* mutations occurred more often in p53 wild type tumors (6/19, 32%) than in p53 abnormal tumors (2/18, 11%) whereas

FBXW7 mutations occurred more often in p53 abnormal tumors (4/18, 22%) than in p53 wild type tumors (1/19, 5%).

Closer examination of our *TP53* results revealed differences in mutation position between p53 protein abnormal and p53 protein wild type tumors. All but two mutations in p53 wild type tumors arose in exon 6 through exon 8. The most mutated amino acids in p53 wild type tumors were R213 and R248 (Figure 3.3, Table 3.2). In p53 abnormal tumors, mutations were identified throughout exons 4-8. Amino acid R248 was mutated in almost half of p53 abnormal tumors (7/15, 47%). All other mutations in p53 abnormal tumors only appear once. The most common *TP53* mutations in p53 wild type tumors were R213*/Q and R248Q, while R248W is the most common *TP53* mutation in p53 abnormal tumors.

Hotspot Mutation Profile by Histology

To determine whether hotspot mutation profiles differed by histology, we stratified mutations by tumor stage, histologic subtype, and histologic grade. There were no statistically significant differences in hotspot mutations by stage in our study (Figure S3.2). *PIK3CA*, *PTEN*, *KRAS*, and *ATM* mutations were more frequent in stage I tumors than in stage II or greater tumors. This pattern is consistent with TCGA[10] when stratified by tumor stage (Table S3.2). *FGFR2* and *KIT* mutations were somewhat more frequent in stage II or greater tumors than in stage I tumors.

The percent of endometrioid tumors with *TP53* mutations was significantly lower than the percent of non-endometrioid tumors with *TP53* mutations (p = 0.0475, Fisher's exact test). No other mutations were significantly different by type. Endometrioid tumors had a higher frequency of *PTEN*, *PIK3CA*, and *KRAS* mutations. Non-endometrioid tumors have more *FBXW7* mutations. This pattern is consistent with TCGA when stratified by histologic subtype (Table S3.3).

To stratify our sample set by histologic grade, grade 3 endometrioid tumors were grouped with non-endometrioid tumors, which are considered high grade. Stratification by histologic grade revealed that *PTEN* is significantly mutated (p = 0.0067, Kruskal-Wallis test) in grade 1 endometrioid tumors compared to other grades (Figure 3.4) Similarly, *PIK3CA* is significantly mutated in grade 1 endometrioid tumors (p = 0.0150, Kruskal-Wallis test). TP53 is mutated more frequently in grade 3 tumors, though this result was not significantly different from other grades. *KRAS* was most frequently mutated in grade 1 endometrioid tumors.

Hotspot Mutation Profile by Exposure History

Being overweight or obese is a major risk factor for endometrial cancer; therefore we dichotomized BMI into overweight (BMI ≥ 25) and normal weight (< 25) to assess whether hotspot mutation profiles were correlated with weight. All fourteen women who had *PIK3CA* mutated tumors were in the overweight category (Figure 3.5). This was significant at p = 0.0009, though this did not replicate in TCGA samples (p = 0.6198). There were no other significant differences in mutation frequency by BMI. *TP53*, *PTEN*, and *KRAS* were more frequently mutated in overweight women. *FGFR2* and *APC* were more frequently mutated in normal weight women.

Smoking is associated with a decrease risk of EC. We assessed the mutation profiles of those who have a history of smoking (ever smokers) to those who have not smoked (never smokers). Mutation frequencies did not significantly differ by smoking status. Overall, hotspot mutations were less frequent in smokers, though *RET*, *EGFR*, and *CTNNB1* mutations only occurred among those who smoked.

Discussion

Though the genomic landscape of EC has been comprehensively characterized in fresh frozen tissue by TCGA[10], our study shows that targeted next-generation sequencing panels performed on archival formalin-fixed paraffin embedded tissue can similarly characterize important aspects of the mutational profile of EC. Previous studies have shown that *PIK3CA* mutations occur in about 30% of endometrioid tumors and 20% of non-endometrioid tumors[21]. Our study similarly revealed *PIK3CA* mutations in about 40% of endometrioid tumors and 30% of non-endometrioid tumors. *TP53* mutations appeared in about 80% of our non-endometrioid tumors, similar to the 88% of non-endometrioid EC samples in TCGA with *TP53* mutations. However, the frequency of *TP53* mutations in our endometrioid samples was about three times as high compared to TCGA (44% vs. 13.5%). This may be due to differences in the distribution of grade or stage between our studies. Among sequenced endometrioid tumors, our study has a higher proportion of stage II or greater tumors (39.1% vs. 19%), histologic grade I tumors (47.8% vs. 37.5%) and histologic grade III tumors (30.4% vs. 24.5%) than TCGA. Otherwise, similar patterns of mutation frequency such as more *PTEN* mutations, less *FBXW7* mutations, and more *KRAS* mutations in endometrioid tumors compared to non-endometrioid tumors were found.

Our analysis of p53 immunostaining in conjunction with hotspot mutation data reveals interesting avenues to explore regarding the molecular basis behind p53 staining in EC. As expected, *TP53* mutations occurred much more frequently in p53 protein abnormal tumors than in p53 protein wild type tumors. However, presence of *TP53* mutations and p53 wild type staining were not mutually exclusive. Our integration of immunohistochemical and sequencing data was able to provide potential leads as to why this may be the case. *TP53* mutations in p53 wild type tumors were more concentrated in exons 6-8, whereas mutations in p53 abnormal tumors were evenly spread out among exons 4-8. R248W was the most frequent *TP53* mutation in p53 abnormal tumors, but was not present at all in p53 wild type tumors. Perhaps these differences in amino acid changes affect the stability of the p53 protein and the ability of immunostaining to detect these mutant proteins.

Mutations in R248 are frequent in cancer[22–25] and are known for their oncogenic properties[26]. Specifically, R248Q has been seen to promote cell invasion in endometrial cancer cell lines[27], while studies in other cell lines have shown that R248W mutations result in increased migration, cell cycle propagation, drug resistance, increased colony formation, and genomic instability[28]. Given that R248W may be frequently mutated among p53 mutants, the highly oncogenic nature of the mutation may contribute to the poorer prognosis of those who are p53 abnormal.

Our study was also one of few to correlate the mutation spectrum of EC with histologic data and exposure history. We see that mutational profiles do differ by histologic grade, but not necessarily by tumor stage. Like TCGA, we also see that traditional endometrioid and non-endometrioid subtyping is not nuanced enough to capture distinct mutational profiles. Furthermore, analyzing exposure history and sequencing data revealed correlation between presence of *PIK3CA* mutation and overweight status. Though this correlation was not replicated in TCGA data, this provides a starting point for exploring how environmental exposures may influence tumor development.

The small sample size limits the study to descriptive and exploratory analyses. The wealth of exposure information from our original cohort would be better utilized with many more cases. However, our results can aid in generating hypotheses and potential leads for larger studies to explore. Sequencing was only performed on tumor tissue and there is no straightforward way to rule out germline variants. To mitigate this issue, we filtered conservatively by ruling out any variants found in dbSNP with MAF ≥ 1% and excluding mutations not recorded in COSMIC. This filtering may result in false negatives, though most of our interest was in the mutational profile rather than identifying novel mutations. Targeted sequencing allowed us to have high base coverage, leading to more accurate base calling for potentially heterogeneous cell populations such as the tumor tissue in this study.

To our knowledge, our study is the first to incorporate sequencing, immunohistochemistry, histologic and epidemiologic data, providing high-dimensional annotation of EC tumors. As a proof-of-principle study, we demonstrate that commercial targeted sequencing panels on formalin-fixed paraffin-embedded tissue can produce comparable results to larger sequencing studies and that combining sequencing data with immunostaining can provide insight into a marker's diagnostic utility. Similar annotation of EC tumors integrating other diagnostic markers and epidemiologic data in larger sample sets from population-based studies are needed to develop personalized models that improve prediction, diagnosis, and treatment.

References

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136,** E359–386 (2015).
2. Howlader, N. *et al.* SEER Cancer Statistics Review, 1975-2012, National Cancer Institute. Bethesda, MD. (2014). Available at: http://seer.cancer.gov/csr/1975_2012/. (Accessed: 10th January 2016)

3. Zhang, Y. *et al.* Overweight, obesity and endometrial cancer risk: results from a systematic review and meta-analysis. *Int. J. Biol. Markers* **29,** e21–29 (2014).

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

5. Zhou, B. *et al.* Cigarette smoking and the risk of endometrial cancer: a meta-analysis. *Am. J. Med.* **121,** 501–508.e3 (2008).

6. Mendivil, A., Schuler, K. M. & Gehrig, P. A. Non-endometrioid adenocarcinoma of the uterine corpus: a review of selected histological subtypes. *Cancer Control* **16,** 46–52 (2009).

7. Bokhman, J. V. Two pathogenetic types of endometrial carcinoma. *Gynecol. Oncol.* **15,** 10–17 (1983).

8. Key, T. J. & Pike, M. C. The dose-effect relationship between 'unopposed' oestrogens and endometrial mitotic rate: its central role in explaining and predicting endometrial cancer risk. *Br. J. Cancer* **57,** 205–212 (1988).

9. Setiawan, V. W. *et al.* Type I and II endometrial cancers: have they different risk factors? *J. Clin. Oncol.* **31,** 2607–2618 (2013).

10. Cancer Genome Atlas Research Network *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497,** 67–73 (2013).

11. Harris, C. C. & Hollstein, M. Clinical Implications of the p53 Tumor-Suppressor Gene. *N. Engl. J. Med.* **329,** 1318–1327 (1993).

12. Pisani, A. L. *et al.* HER-2/neu, P53, and DNA analyses as prognosticators for survival in endometrial carcinoma. *Obstet. Gynecol.* **85,** 729–734 (1995).

13. Hamel, N. W. *et al.* Prognostic Value of p53 and Proliferating Cell Nuclear Antigen Expression in Endometrial Carcinoma. *Gynecol. Oncol.* **62,** 192–198 (1996).

14. Powell, B. *et al.* p53 protein overexpression is a prognostic indicator of poor survival in stage I endometrial carcinoma. *Int. J. Oncol.* **14,** 175–179 (1999).

15. Lomo, L. *et al.* Histologic and immunohistochemical decision-making in endometrial adenocarcinoma. *Mod. Pathol.* **21,** 937–942 (2008).

16. Creasman, W. Revised FIGO staging for carcinoma of the endometrium. *Int. J. Gynaecol. Obstet.* **105,** 109 (2009).

17. *World Health Organization Classification of Tumours. Pathology and Genetics of Tumours of the Breast and Female Genital Organs*. (IARC Press, 2013).

18. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805–D811 (2015).

19. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discov.* **2,** 401–404 (2012).

20. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6,** pl1 (2013).

21. Dedes, K. J., Wetterskog, D., Ashworth, A., Kaye, S. B. & Reis-Filho, J. S. Emerging therapeutic targets in endometrial cancer. *Nat. Rev. Clin. Oncol.* **8,** 261–271 (2011).

22. Yokoyama, N. *et al.* Mutations of p53 in gallbladder carcinomas in high-incidence areas of Japan and Chile. *Cancer Epidemiol. Biomark. Prev.* **7,** 297–301 (1998).

23. Powell, B., Soong, R., Iacopetta, B., Seshadri, R. & Smith, D. R. Prognostic significance of mutations to different structural and functional regions of the p53 gene in breast cancer. *Clin. Cancer Res.* **6,** 443–451 (2000).

24. Smith, G. *et al.* Mutations in APC, Kirsten-ras, and p53--alternative genetic pathways to colorectal cancer. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 9433–9438 (2002).

25. McBride, S. M. *et al.* Mutation frequency in 15 common cancer genes in high-risk head and neck squamous cell carcinoma. *Head Neck* **36,** 1181–1188 (2014).

26. Brachova, P., Thiel, K. W. & Leslie, K. K. The Consequence of Oncomorphic TP53 Mutations in Ovarian Cancer. *Int. J. Mol. Sci.* **14,** 19257–19275 (2013).

27. Dong, P. *et al.* Mutant p53 gain-of-function induces epithelial–mesenchymal transition through modulation of the miR-130b–ZEB1 axis. *Oncogene* **32,** 3286–3295 (2013).

28. Muller, P. A. J. & Vousden, K. H. Mutant p53 in Cancer: New Functions and Therapeutic Opportunities. *Cancer Cell* **25,** 304–317 (2014).

Table 3.1. Sample characteristics by p53 staining status

	All cases (n = 37)	*p53 abnormal* (n = 18)	*p53 wild type* (n = 19)
Mean age at diagnosis	70.4	69.4	71.4
Mean BMI	27.9	26.5	29.3
BMI >=25 (%)	25 (68%)	11 (61%)	14 (74%)
Smoking Status (%)			
Ever	14 (38%)	8 (44%)	6 (32%)
N/A	3 (8%)	1 (6%)	2 (11%)
Stage (%)			
I	21 (57%)	13 (72%)	8 (42%)
II+	16 (43%)	5 (28%)	11 (58%)
Grade[a] (%)			
I	11 (30%)	0 (0%)	11 (58%)
II	5 (14%)	3 (17%)	2 (11%)
III	7 (19%)	3 (17%)	4 (21%)
None	14 (38%)	12 (67%)	2 (11%)
Histology (%)			
Endometrioid	23 (62%)	6 (33%)	17 (90%)
Serous	11 (30%)	10 (56%)	1 (5%)
Clear	3 (8%)	2 (11%)	1 (5%)

[a] Non-endometrioid tumors were not graded.

Table 3.2. *TP53* hotspot mutations in endometrial carcinoma patients

ID	**Type**	**Grade**[a]	**Stage**	**P53 Status**	***TP53* Mutations**
101	Clear Cell	.	1	abn	R248W, M169I
662	Clear Cell	.	1	abn	R273H
038	Endometrioid	2	1	abn	Y220C
379	Endometrioid	2	1	abn	F109V
177	Endometrioid	3	1	abn	A276G
215	Endometrioid	3	1	abn	R248W
529	Endometrioid	3	1	abn	R248W, K132T
029	Serous	.	3	abn	R248W
129	Serous	.	3	abn	I195T, F134L, T102I, P72A
138	Serous	.	1	abn	N239S
239	Serous	.	4	abn	R248Q
274	Serous	.	1	abn	R175H
286	Serous	.	1	abn	R248W
673	Serous	.	2	abn	V157F
714	Serous	.	3	abn	R248L
310	Clear Cell	.	2	wt	R248Q
012	Endometrioid	1	1	wt	R213Q
025	Endometrioid	1	1	wt	R213fs, R213X
142	Endometrioid	1	2	wt	P72A
309	Endometrioid	1	2	wt	P278H, C277Y, P151S
736	Endometrioid	1	1	wt	R248Q

Abbreviations : abn, abnormal; wt, wild type

[a] Non-endometrioid tumors were not graded (indicated by . symbol).

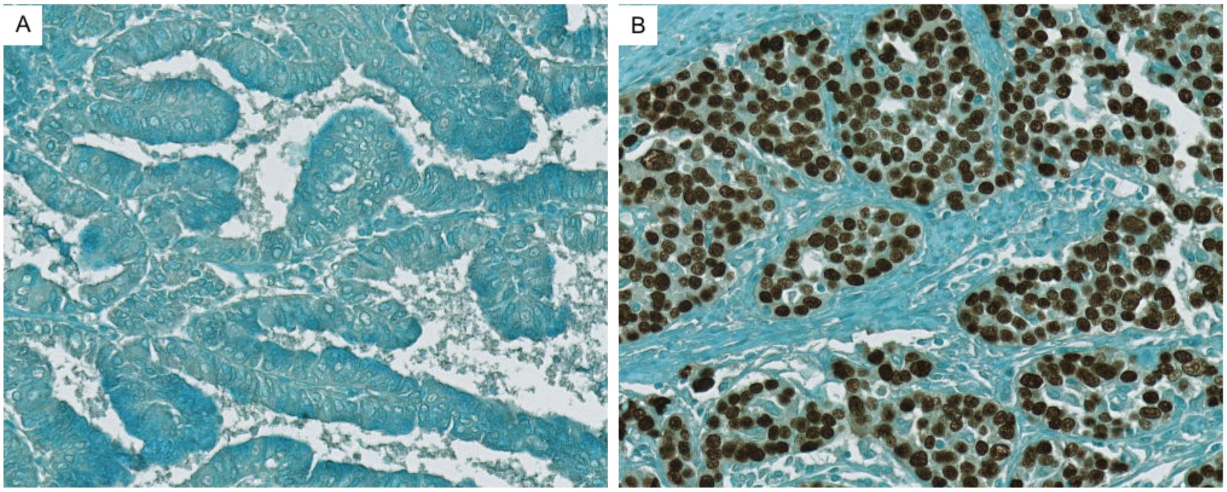

Figure 3.1. P53 immunostaining of endometrial carcinoma. (A) Example of p53 protein wild type endometrial tumor tissue. (B) Example of p53 protein abnormal endometrial tumor tissue.

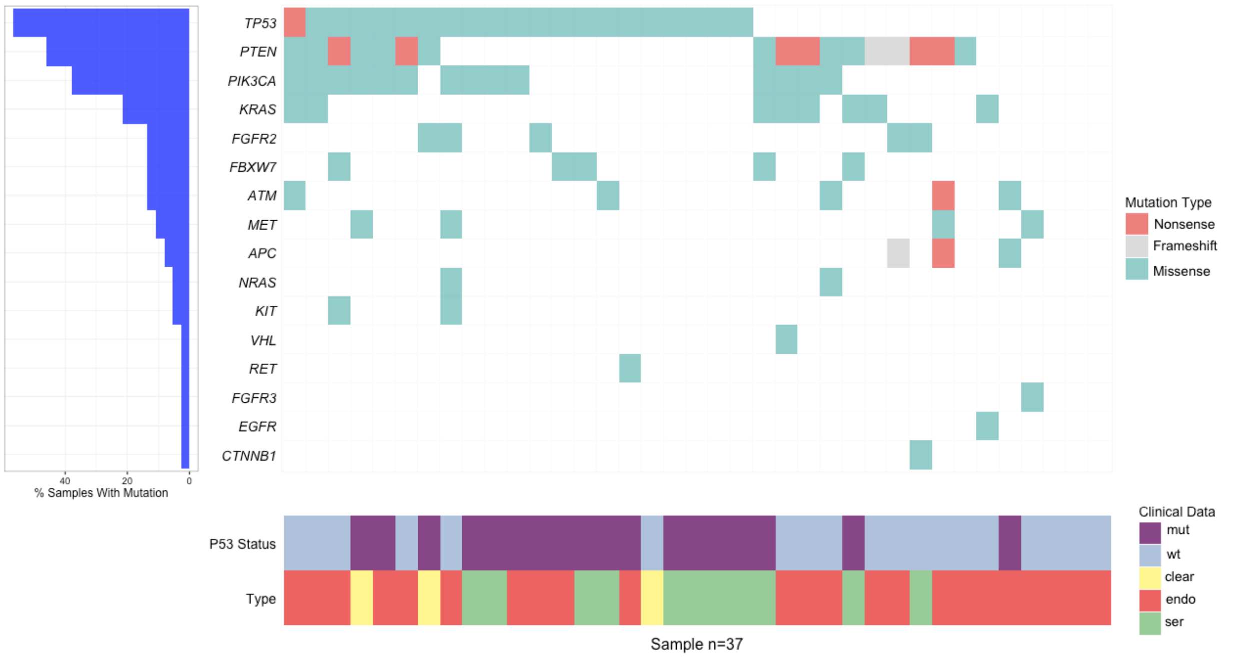

Figure 3.2. The mutation profile of endometrial carcinoma in 37 samples from a population-based study. Main panel: Each column represents the mutation presence of one sample. Red rectangles are nonsense mutations, Blue rectangles are missense mutations, and gray rectangles are frameshift mutations. A sample can have more than one type of mutation in each gene, but the most deleterious mutation is represented. Left panel: The percent of all samples with at least one mutation in the corresponding gene in the main panel. Bottom panel: P53 status and tumor type of the corresponding sample in the main panel. Abbreviations: mut, p53 mutant (abnormal); wt, p53 wild type; clear, clear cell carcinoma; endo, endometrioid carcinoma; ser, serous carcinoma.

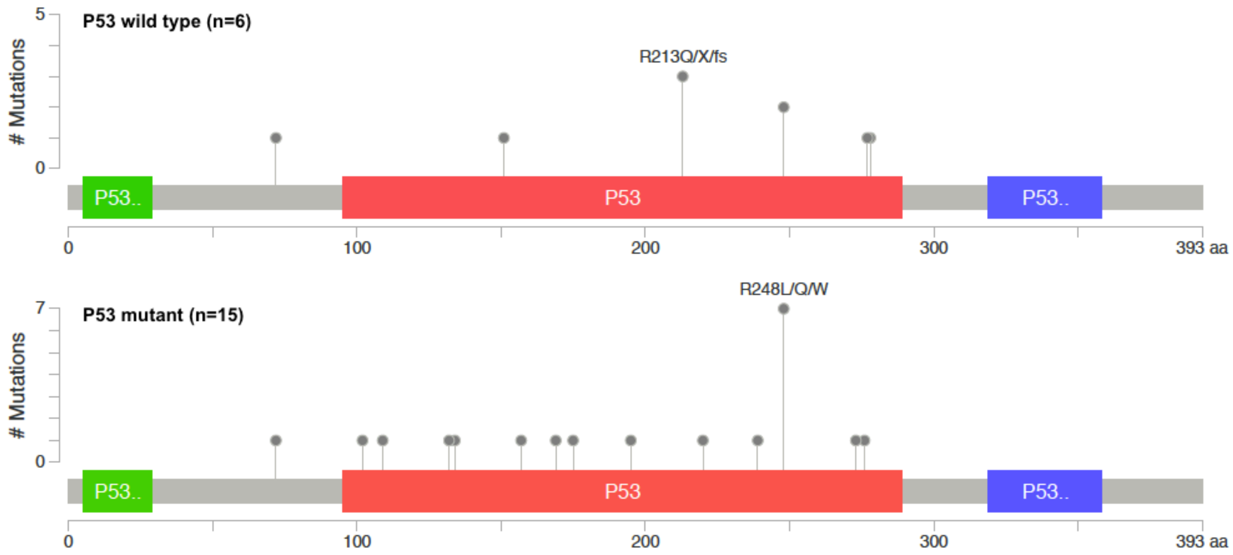

Figure 3.3. Locations of *TP53* mutations identified among the 37 endometrial cancer cases by p53 status. Each lollipop represents the number of mutations that occurred at that amino acid position. Top panel: *TP53* R213 was most frequently mutated in p53 protein wild type tumors. Bottom panel: P53 protein mutant (abnormal) tumors had *TP53* mutations that were spread throughout exons 4-8. The most frequent amino acid changes occurred on R248.

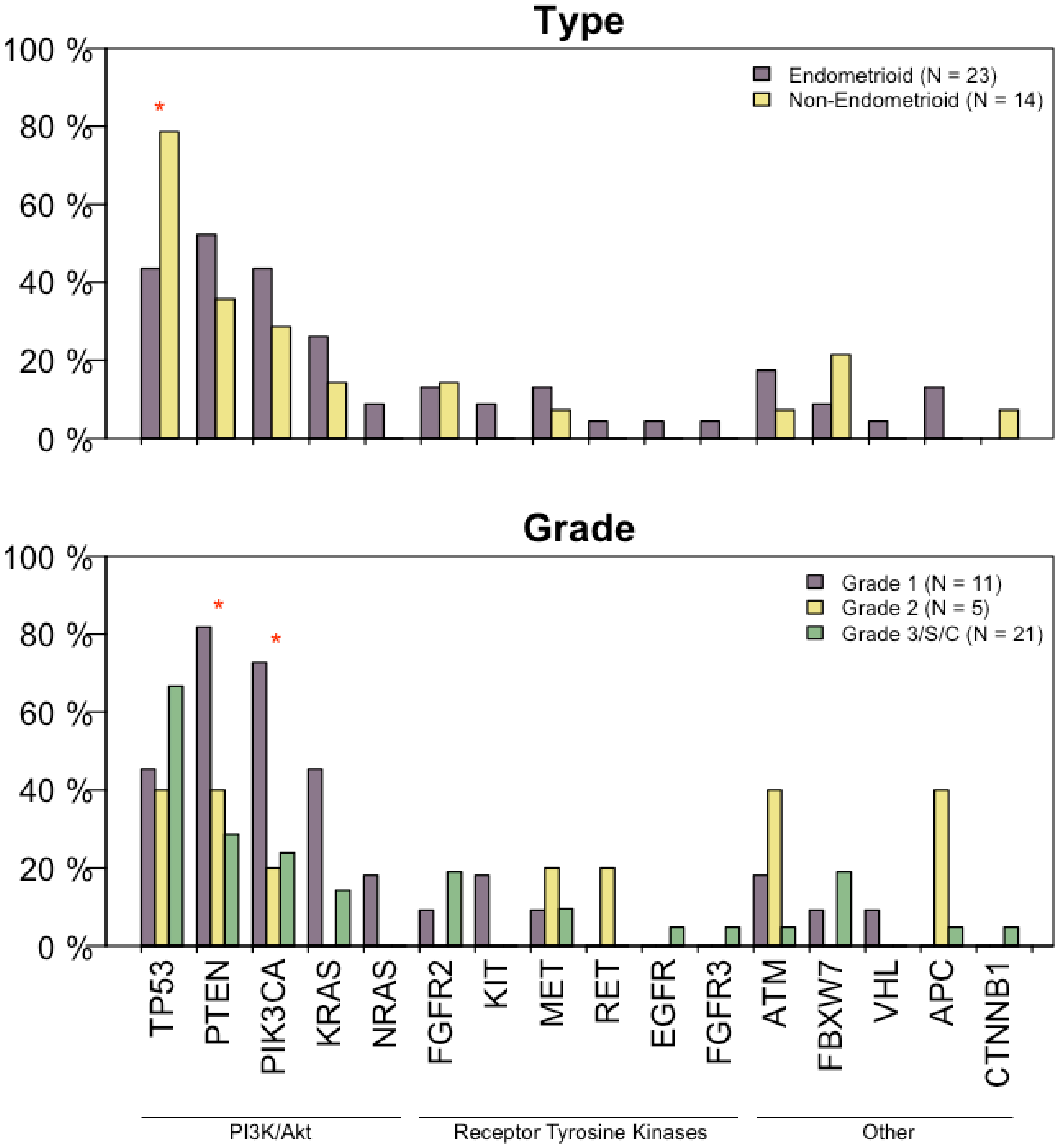

Figure 3.4. Percentage of samples with at least one hotspot mutation stratified by histological characteristics. Genes are grouped by pathway features. Red asterisks indicate significant correlation between frequency of gene mutation and morphological feature at $P < 0.05$. Top panel: Presence of *TP53* mutation in our study is significantly correlated with type. Bottom panel: Presence of *PTEN* and *PIK3CA* mutations is significantly correlated with endometrioid grade.

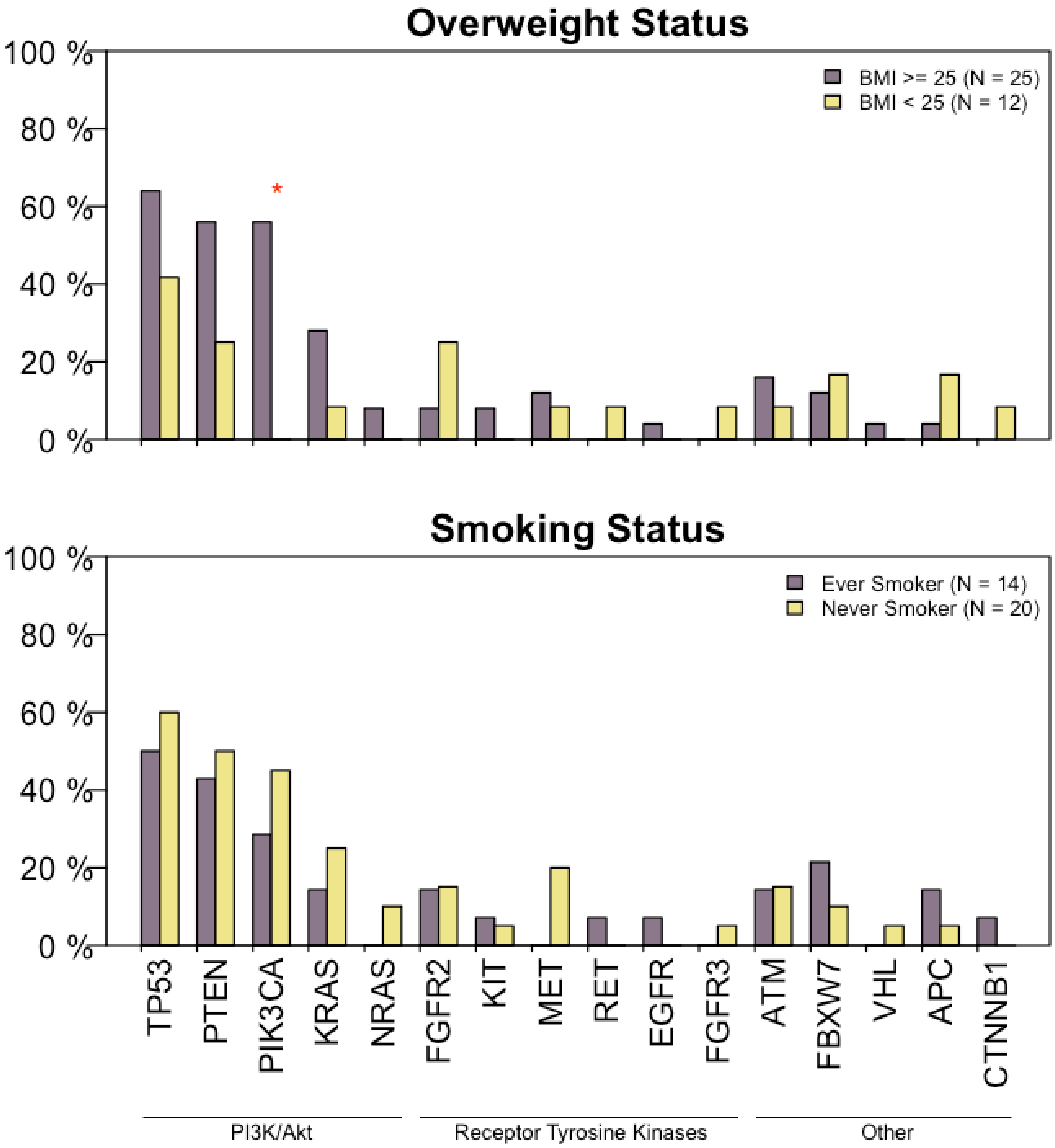

Figure 3.5. Percentage of samples with at least one hotspot mutation stratified by exposure history. Genes are grouped by pathway features. Red asterisks indicate significant correlation between frequency of gene mutation and morphological feature at $P < 0.05$. Top panel: Presence of *PIK3CA* mutation in our study is significantly correlated with overweight (BMI ≥ 25) status. This result did not replicate in TCGA. Bottom panel: The frequency of gene mutation did not differ by smoking status.

Table S3.1. Genes captured in the Cancer Hotspot Panel v.2

Ion Ampliseq Cancer Hotspot Panel v2				
ABL1	EGFR	GNAQ	KRAS	PTPN11
AKT1	ERBB2	GNAS	MET	RB1
ALK	ERBB4	HNF1A	MLH1	RET
APC	EZH2	HRAS	MPL	SMAD4
ATM	FBXW7	IDH1	NOTCH1	SMARCB1
BRAF	FGFR1	IDH2	NPM1	SMO
CDH1	FGFR2	JAK2	NRAS	SRC
CDKN2A	FGFR3	JAK3	PDGFRA	STK11
CSF1R	FLT3	KDR	PIK3CA	TP53
CTNNB1	GNA11	KIT	PTEN	VHL

Table S3.2. Frequency of Mutations by Tumor Stage within TCGA and Our Study

	TCGA (n = 246)		Our Study (n = 37)	
Gene	Stage I (n = 176)	Stage II+ (n = 70)	Stage I (n = 21)	Stage II+ (n = 16)
APC	20 (11.4%)	9 (12.9%)	2 (9.5%)	1 (6.3%)
ATM	22 (12.5%)	7 (10.0%)	4 (19.0%)	1 (6.3%)
CTNNB1	58 (33.0%)	18 (25.7%)	0 (0.0%)	1 (6.3%)
EGFR	4 (2.3%)	4 (5.7%)	0 (0.0%)	1 (6.3%)
FBXW7	22 (12.5%)	16 (22.9%)	3 (14.3%)	2 (12.5%)
FGFR2	23 (13.1%)	8 (11.4%)	2 (9.5%)	3 (18.8%)
FGFR3	2 (1.1%)	1 (1.4%)	1 (4.8%)	0 (0.0%)
KIT	11 (6.3%)	6 (8.6%)	0 (0.0%)	2 (12.5%)
KRAS	43 (24.4%)	9 (12.9%)	6 (28.6%)	2 (12.5%)
MET	7 (4.0%)	6 (8.6%)	3 (14.3%)	1 (6.3%)
NRAS	8 (4.5%)	1 (1.4%)	1 (4.8%)	1 (6.3%)
PIK3CA	100 (56.8%)	31 (44.3%)	10 (47.6%)	4 (25.0%)
PTEN	130 (73.9%)	30 (42.9%)	12 (57.1%)	5 (31.3%)
RET	7 (4.0%)	4 (5.7%)	1 (4.8%)	0 (0.0%)
TP53	32 (18.2%)	37 (52.9%)	13 (61.9%)	8 (50.0%)
VHL	3 (1.7%)	0 (0.0%)	1 (4.8%)	0 (0.0%)

Table S3.3. Frequency of Mutations by Subtype within TCGA and Our Study

	TCGA (n = 248)		Our Study (n = 37)	
Gene	Endometrioid (n = 200)	Non-Endometrioid (n = 48)	Endometrioid (n = 23)	Non-Endometrioid (n = 14)
APC	27 (13.5%)	2 (4.2%)	3 (13.0%)	0 (0.0%)
ATM	28 (14.0%)	1 (2.1%)	4 (17.4%)	1 (7.1%)
CTNNB1	74 (37.0%)	0 (0.0%)	0 (0.0%)	1 (7.1%)
EGFR	8 (4.0%)	0 (0.0%)	1 (4.3%)	0 (0.0%)
FBXW7	23 (11.5%)	15 (31.3%)	2 (8.7%)	3 (21.4%)
FGFR2	27 (13.5%)	4 (8.3%)	3 (13.0%)	2 (14.3%)
FGFR3	3 (1.5%)	0 (0.0%)	1 (4.3%)	0 (0.0%)
KIT	17 (8.5%)	0 (0.0%)	2 (8.7%)	0 (0.0%)
KRAS	51 (25.5%)	1 (2.1%)	6 (26.1%)	2 (14.3%)
MET	12 (6.0%)	1 (2.1%)	3 (13.0%)	1 (7.1%)
NRAS	8 (4.0%)	1 (2.1%)	2 (8.7%)	0 (0.0%)
PIK3CA	110 (55.0%)	22 (45.8%)	10 (43.5%)	4 (28.6%)
PTEN	159 (79.5%)	2 (4.2%)	12 (52.2%)	5 (35.7%)
RET	11 (5.5%)	0 (0.0%)	1 (4.3%)	0 (0.0%)
TP53	27 (13.5%)	42 (87.5%)	10 (43.5%)	11 (78.6%)
VHL	3 (1.5%)	0 (0.0%)	1 (4.3%)	0 (0.0%)

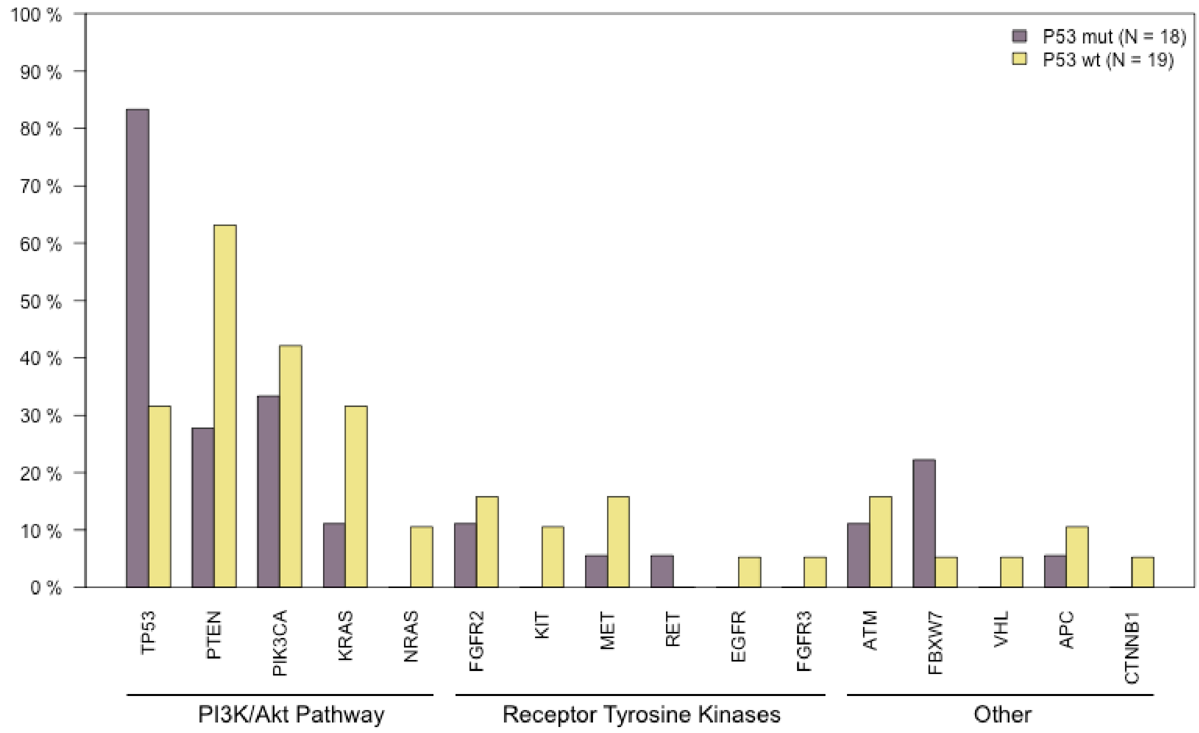

Figure S3.1. Percentage of hotspot mutated samples by p53 status.

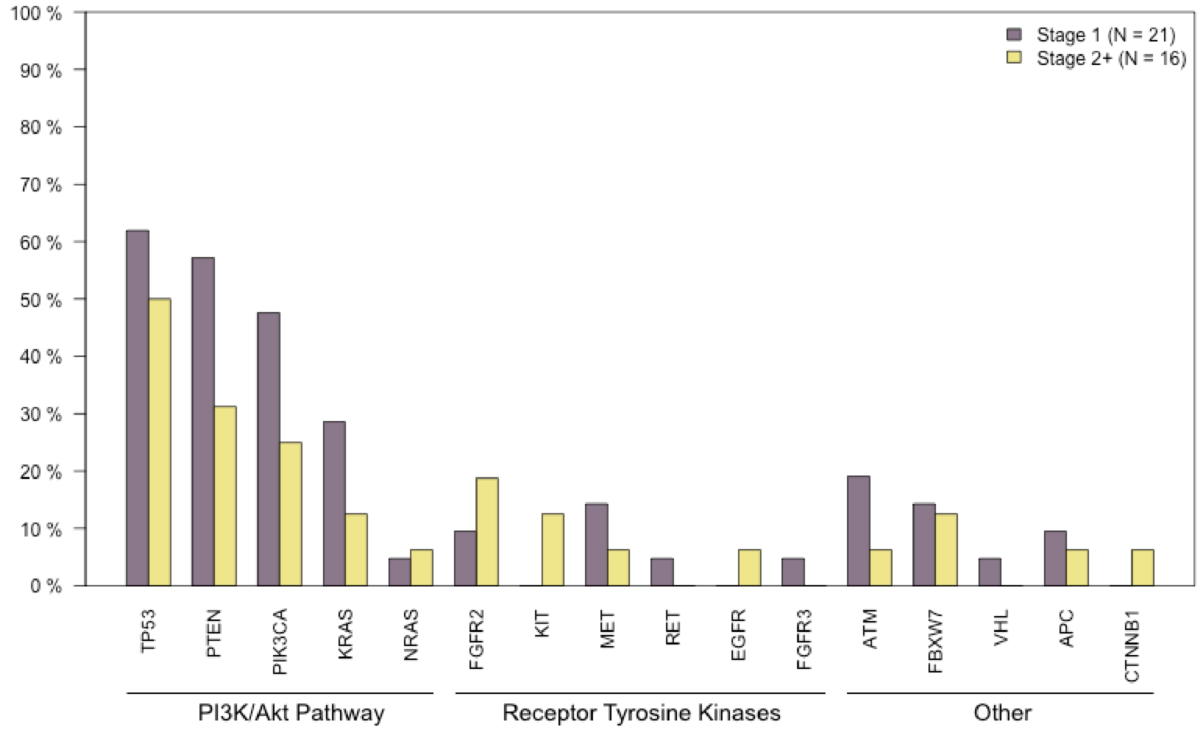

Figure S3.2. Percentage of hotspot mutated samples by tumor stage.