Original Manuscript

Common variants at the CHEK2 gene locus and risk of epithelial ovarian cancer

Nuremberg, Erlangen 91054, Germany, 102Department of Medicine, Division of Hematology and Oncology, University of California at Los Angeles, David Geffen School of Medicine, Los Angeles CA 90095, USA, 103Biostatistics and Informatics Shared Resource, University of Kansas Medical Center, Kansas City, KS 66160, USA, 104Shanghai Cancer Institute, Shanghai 200032, China, 105Department of Women’s Cancer, Institute for Women’s Health, University College London, London W1T 7DN, UK, 106Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Victoria 3010, Australia, 107Department of Health Science Research, Mayo Clinic, Rochester, MN 55905, USA, 108Cancer Prevention and Control, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA, 109Department of Biomedical Sciences, Community and Population Health Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA, 110Cancer Research Initiatives Foundation, Sime Darby Medical Centre, Subang Jaya 47500, Malaysia, 111Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, 112Departments of Obstetrics and Gynaecology, Hannover Medical School, Hannover 30625, Germany, 113Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen DK-2100, Denmark, 114Molecular Unit, Department of Pathology, Herlev Hospital, University of Copenhagen, Copenhagen DK-2100, Denmark, 115Gyn Clinic, Rigshospitalet, University of Copenhagen DK-2100, Denmark, 116Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Aichi 464-8681, Japan, 117Department of Virus, Lifestyle and Genes, Danish Cancer Society Research Center, Copenhagen DK-2100, Denmark, 118Women’s Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA, 119Department of Gynecology, University of Calgary, Calgary T2N 2N4, Canada, 120Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY 14263, USA, 121College of Pharmacy and Health Sciences, Texas Southern University, Houston, TX 77004, USA, 122Department of Gynecologic Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, 123Department of Gynaecology, Radboud University Medical Centre, Nijmegen 6500 HB, The Netherlands, 124Vesalius Research Center, VIB, Leuven 3000, Belgium, 125Laboratory for Translational Genetics, Department of Oncology, University of Leuven 3000, Belgium, 126Division of Gynecological Oncology, Department of Obstetrics, University Hospitals Leuven, 3000, Belgium, 127Cancer Control Research, BC Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada, 128Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY 14263, USA, 129College of Pharmacy and Health Sciences, Texas Southern University, Houston, TX 77004, USA, 130Department of Gynecologic Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, 131Department of Preventive Medicine, Kyushu University Faculty of Medical Sciences, Fukuoka 812-8582, Japan, 132Department of Health Research and Policy - Epidemiology, Stanford School of Medicine, Stanford, CA 94305, USA, 133Prosserman Centre for Health Research, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario MSG 1X5, Canada, 134Institute of Cancer Sciences, University of Glasgow, Wolfson Wohl Cancer Research Centre, Beatson Institute for Cancer Research, Glasgow G61 1QH, UK, 135Women’s Cancer Research Program, Magee-Women’s Research Institute and University of Pittsburgh Cancer Institute, Pittsburgh, PA 15222, USA, 136Department of Obstetrics, Gynecology and Reproductive Sciences, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA 15222, USA, 137Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA 15222, USA, 138Women’s College Research Institute, Toronto, Ontario M5G 1N8, Canada, 139Department of Pathology, Rigshospitalet, University of Copenhagen DK-2100, Denmark, 140The University of Texas School of Public Health, Houston, TX 77270, USA, 141Department of Obstetrics and Gynaecology, University Malaya Medical Centre, University Malaya, Kuala Lumpur 50603, Malaysia, 142Department of Gynecological Oncology, Roswell Park Cancer Institute, Buffalo, NY 14263, USA, 143Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA, 144Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI 48109, USA, 145Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center, Tampa, FL 33612, USA, 146Channing Division of Network Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA 02115, USA, 147Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA, 148Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT 06520, USA, 149Department of Gynecologic-Oncology, Princess Margaret Hospital, and Department of Obstetrics and Gynecology, Faculty of Medicine, University of Toronto, Toronto, Ontario M5G 1X6, Canada, 150Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA, 151Department of Epidemiology, University of Washington, Seattle, WA 98109, USA, 152Department of Gynecology, Jena University Hospital - Friedrich Schiller University, Jena D-07743, Germany, 153Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37203, USA, 154Cancer Epidemiology Program, University of Hawaii Cancer Center, Hawaii 96826, USA, 155Department of Gynaecological Oncology, Glasgow Royal Infirmary, Glasgow G4 0SF, UK, 156University Malaya Cancer Research Institute, Faculty of Medicine, University Malaya Medical Centre, University Malaya, Kuala Lumpur 50603, Malaysia, 157Department of Obstetrics, Gynecology and Oncology, IInd Faculty of Medicine, Warsaw Medical University and Brodnowski University, Warsaw 03-242, Poland, 158Department of Obstetrics and Gynecology, University of Ulm, Ulm 89075, Germany, 159Department of Medical Oncology, 02115, Dana-Farber Cancer Institute, Boston, MA, USA, 160Cancer Epidemiology Program, Division of Population Sciences, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA and 161Cancer Control and Population Sciences, Duke Cancer Institute, Durham, NC 27710, USA, 162Present address: Women’s Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA and 163Present address: Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
Abstract

Genome-wide association studies have identified 20 genomic regions associated with risk of epithelial ovarian cancer (EOC), but many additional risk variants may exist. Here, we evaluated associations between common genetic variants [single nucleotide polymorphisms (SNPs) and indels] in DNA repair genes and EOC risk. We genotyped 2896 common variants at 143 gene loci in DNA samples from 15 397 patients with invasive EOC and controls. We found evidence of associations with EOC risk for variants at FANCA, EXO1, E2F4, E2F2, CREB5 and CHEK2 genes (P ≤ 0.001). The strongest risk association was for CHEK2 SNP rs17507066 with serous EOC (P = 4.74 x 10⁻⁸). Additional genotyping and imputation of genotypes from the 1000 genomes project identified a slightly more significant association for CHEK2 SNP rs6005807 (P with rs17507066 = 0.84, odds ratio (OR) 1.17, 95% CI 1.11–1.24, P = 1.1 x 10⁻⁴). We identified 293 variants in the region with likelihood ratios of less than 1:100 for representing the causal variant. Functional annotation identified 25 candidate SNPs that alter transcription factor binding sites within regulatory elements active in EOC precursor tissues. In The Cancer Genome Atlas dataset, CHEK2 gene expression was significantly higher in primary EOCs compared to normal fallopian tube tissues (2.70 × 10⁻⁸). We also identified an association between genotypes of the candidate causal SNP rs12166475 (r² = 0.99 with rs6005807) and CHEK2 expression (P = 2.70 × 10⁻⁸). These data suggest that common variants at 22q12.1 are associated with risk of serous EOC and CHEK2 as a plausible target susceptibility gene.

Introduction

Epithelial ovarian cancer (EOC) is the leading cause of death from cancers of the female reproductive tract (1). Established protective risk factors for the disease include pregnancy, oral contraceptives and breast-feeding, factors which reduce the number of lifetime ovulatory cycles. The biological mechanisms underlying these associations are not well understood but may involve proliferation, inflammation, oxidative stress, mutation and DNA repair in the ovarian and/or fallopian tube epithelia (2-4). Several studies have highlighted the critical role in ovarian tumorigenesis of maintaining genomic integrity, either through DNA repair or apoptotic pathways (5).

The importance of highly penetrant genetic risk factors for EOC has become evident from studies of familial clustering of breast and ovarian cancers, also caused by mutations in the BRCA1 and BRCA2 (6-8). Several additional clustering genes exist that confer more moderate risks of EOC (e.g. MSH6, MSH2, MLH1, RAD51C, RAD51D and BRIP1 (9-12)). Finally, several common genetic variants conferring relatively mild risks of EOC have been identified in genome-wide association studies (GWAS) (13-21).

Whereas GWAS are highly successful at identifying common variant susceptibility alleles for multiple complex traits and diseases using case-control study designs, only a fraction of the total number of risk variants for any one disease have been identified to date. For example, Michailidou et al. (22) estimated that many common risk variants for breast cancer await discovery; and the same is likely true of other cancers. One approach to identify additional susceptibility alleles is to integrate knowledge of disease biology with germline genetic datasets (i.e. candidate gene or pathway studies). This approach has recently been successful in identifying the Sp15 (TERT-CLPTM1L) susceptibility locus for breast, prostate and ovarian cancers (19).

CHEK2 is a putative tumor suppressor gene that encodes a protein kinase activated in response to DNA damage and has also been shown to interact with and phosphorylate BRCA1, promoting cellular survival after DNA damage (23). A deletion variant in CHEK2 (CHEK2*1100delC) is associated with a 2–3-fold increased risk of breast cancer but is not associated with ovarian cancer risk (24). However, other CHEK2 coding variants may confer susceptibility to ovarian cancer, and in one study, five CHEK2 missense variants were identified in an analysis of 360 EOC cases, two of which were predicted to be damaging in functional assays (25). Common low penetrance variants may also exist in and around the CHEK2 gene. We previously evaluated associations for germline genetic variants spanning 53 DNA damage response and repair genes and risk of invasive serous EOC in 364 EOC cases and 761 controls from the North Carolina Ovarian Cancer Study (NCOCS), and identified borderline evidence of an association for two variants (rs5762746 and rs6005835) in CHEK2 (26).

In this study, we expanded the examination of risks associated with DNA repair genes by genotyping 2896 single nucleotide polymorphisms (SNPs) and indel variants spanning 143 gene loci in DNA samples of 15 397 EOC cases and 30 816 controls as part of the Collaborative Oncological Gene-environment Study (COGS). Imputation of additional SNPs in COGS further defined risk associations at the CHEK2 locus. Finally, we integrated germline genetic data with functional annotation of the region to identify the most likely disease-causing susceptibility variants and target genes at this locus.

Materials and methods

Genetic association analyses

Study datasets

Genetic association analyses were carried out using data from several Ovarian Cancer Association Consortium (OCAC) genotyping projects. Study subjects were of European ancestry (determined using principal components analysis of genotype data); 2162 cases and 2564 controls from a GWAS from North America (“US GWAS”), 1763 cases and 6118 controls from a UK-based GWAS (“UK GWAS”) and 441 cases and 442 controls from a second GWAS from North America (“Mayo GWAS”) (13-19,21). In total, 11 030 cases and 21 693 controls from 41 OCAC studies were genotyped using the iCOGS array (“OCAC-iCOGS” stage 1 data). The USA and UK GWAS were comprised of several independent case-control studies, and samples from...
some of these studies were also subsequently genotyped using the iCOGS array. Combined, these studies comprised 15 396 independent cases and 30 817 controls. All duplicates were removed. Further details of the component studies are in Supplementary Table 1, available at Carcinogenesis Online. Details of genotyping platform are shown in Supplementary Table 2, available at Carcinogenesis Online.

Variant selection
To select genes, we expanded upon the 53 genes included in our earlier investigation (26) using the gene sets described in Wood et al. (27) plus literature and gene ontology database searches, ultimately identifying 143 genes whose functions relate to DNA damage recognition and repair processes for inclusion in this study. For each gene, we identified all SNPs within genome windows ranging from 10 kb upstream of the transcription start sites to 10 kb downstream of the transcription end sites of the 143 genes DNA repair genes. These variants were included in the US ovarian cancer GWAS database of SNPs imputed by the MACH software package against Hapmap Phase II genotypes (Release 22, NCBI build 36) for 60 CEU founders. We used these data to conduct a preliminary association analysis to identify tag sets for each region, tagging polymorphisms with minor allele frequencies (MAF) of at least 0.025 to an r-squared of 0.80 or above. We ranked the genes on basis of the most significant variants within each gene. We tagged SNPs at TPS3, CHEK2 and the top five ranked genes to an r-squared of 0.975, the genes ranked 6 through 124 to an r-squared of 0.90 and the 20 lowest ranked genes to an r-squared of 0.80. We then chose the SNP with the highest Illumina design score as the tag in each r-squared bin, choosing the most highly significant SNP when there were ties. This yielded 3651 variants that were included on the COGS Illumina custom iSelect chip (iCOGS). After quality control analysis, 3252 variants passed QC of which 53 had MAF < 0.02 (Supplementary Table 3 and Figure 1, available at Carcinogenesis Online).

Imputation
OCAC-iCOGS samples and each of the GWAS sets were imputed separately. Variants were imputed from the 1000 Genomes Project data using the v3 April 2012 release as the reference panel. We used a two-step procedure, which involved prephasing in the first step (using the SHAPEIT software and imputation, of the phased data in the second step using the IMPUTE version 2 software (28). To perform the imputation, we divided the data into segments of ~5 Mb and excluded variants from the association analysis if their imputation accuracy was r2 < 0.25 or their MAF was < 0.005. The number of successfully imputed SNPs by MAF is shown in Supplementary Table 4, available at Carcinogenesis Online.

Data analysis
Analyses were restricted to women of European intercontinental ancestry. We performed principal components analysis using a set of ~37 000 unlinked markers to control for population substructure, an in-house program was utilized (available at http://cge.medschl.cam.ac.uk/software). Unconditional logistic regression treating the number of alternate alleles carried as an ordinal variable (log-additive, codominant model) was used to evaluate the association between each variant and EOC risk. A likelihood ratio statistic was used to examine significance of association, and per-allele log odds ratios (OR) and 95% confidence interval (CI) were estimated. The logistic regression model was adjusted for study and population substructure by including study-specific indicators and a variable number of eigenvalues from the principal components analyses. The number of principal components was chosen based on the position of the inflexion of the principal components screen plot. Two principal components were included in the analysis of the UK and US GWAS data sets, one was used for the Mayo GWAS and five were used for the COGS-OCAC dataset. Results from the three GWAS and COGS were combined using fixed-effects, inverse variance weighted meta-analysis.

Functional analyses
Public databases
To perform functional annotation of 293 candidate variants and expression quantitative trait locus (eQTL) analysis of the CHEK2 locus, we mined the following databases: ENCODE (http://genome.ucsc.edu); Haploreg (http://www.broadinstitute.org/mammals/haploreg/haploreg.php); and The Blood eQTL Browser (http://genenetwork.nl/bloodeqtlbrowser) (29).

Profiling of epigenetic marks in ovarian cancer and ovarian cancer precursor cells
FAIREseq and ChiPseq profiles were generated for two immortalized normal ovarian surface epithelial cell lines (IOE4 and IOE11, generated in-house) and two fallopian serous epithelial cell lines (FTT3 and FT246, from Dr R Drapkin) as described previously (30). The CaOV3 and UWB1.289 (31) ovarian cancer cell lines (from CRUK and ATCC, respectively) were also profiled. Prior to performing ChiPseq/FAIREseq, cell lines were authenticated using the Promega PowerPlex16HS Assay (performed at the University of Arizona Genetic Core), and mycoplasma-specific PCR was performed to ensure cell lines were not contaminated with mycoplasma infections. ChiPseq was performed using antibodies that recognized histone 3 lysine 27 acetylation (H3K27ac) and histone 3 lysine 4 mono-methylation (H3K4me) (Abcam). Identification of variants predicted significantly to alter transcription factor motifs was performed as described by Hazelett et al. (32).

Collection of normal epithelial samples
Early passage primary normal ovarian surface epithelial cells and fallopian tube secretory epithelial cells were obtained from disease-free ovaries and fallopian tubes collected during gynecological surgical procedures taking place at the University of Southern California, Los Angeles (the Gynecological Tissue and Fluid Repository), University College Hospital, London and Oregon Health and Science University. All samples were collected with informed patient consent. Methods for the collection have been described previously (33,34). RNA extraction was performed from cells at 80% confluence using standard protocols. Cell lines were confirmed to be free of mycoplasma, but were not authenticated as they were novel cell lines used at an early passage.

eQTL data analysis
For each sample, 500 ng RNA were reverse transcribed using the Superscript III kit (Life Technologies). TaqMan® was used to quantify CHEK2 gene expression using a TaqMan® gene expression probe (Hs00200485_m1, Life Technologies). Four control genes were also included: ACTB, Hs00357333_g1; GAPDH, Hs02758991_g1; HMBS, Hs00609293_g1; and HPRT1 Hs00280069_m1 (all Life Technologies). Relative expression levels were calculated using the ΔΔCt method. Correlations between genotype and gene expression were calculated in R using a Jonckheere-Terpstra trend test (for three groups) or the Wilcoxon rank sum statistic (for 2 groups).

eQTL analysis using TCGA data
Publicly available microarray and germline genotyping data for high-grade serous EOCs was downloaded from TCGA (Lawrence et al, Nat Comm, accepted). For each case, germline genotyping data were used to determine ancestry using principal components through EIGENSTRAT software (HapMap profiles were used as a control set). Only cases with complete Northern or Western European ancestry were included. cis-eQTL analyses were performed for all genes in a 1 MB region spanning the top variant, using a method we have described previously (35). Associations between risk variant genotypes and mRNA expression for 339 cases were evaluated using a linear regression model adjusting for the effects of copy number and methylation. The Benjamini-Hochberg method was used to adjust for
multiple testing. A significant association was defined by a false discovery rate of less than 0.1.

Results

Genetic association analyses of DNA repair gene loci

For 143 DNA repair genes, we identified 2896 tagging SNPs (minor allele frequencies (MAFs) > 2%) that lie within 10 kb upstream and downstream of the transcription start and end sites of each gene; these variants were genotyped as part of iCOGS. Of these, 2 621 were successfully genotyped in 46 213 subjects from 43 studies. This sample included 15 397 women diagnosed with invasive EOC, of whom 9608 cases had serous ovarian cancer, and 30 816 controls. Details of the study populations, genotyping platforms used for each data set and quality control analysis are given in Supplementary Tables 1–3 and Figure 1, available at Carcinogenesis Online.

Supplementary Table 5, available at Carcinogenesis Online, lists the DNA repair genes evaluated in this study, including the number of tag SNPs at each locus, and the significance of association (P value) with serous ovarian cancer for the most significant risk-associated variant for each gene. The data for all SNPs and genes are illustrated in Figure 1A. SNPs at 6 different genes were associated with serous ovarian cancer risk at a P value threshold of 0.001: FANCA on chromosome 16q24.3 (P = 0.001); EXO1 on chromosome 1q43 (P = 0.0005); E2F4 on chromosome 16q22.1 (P = 0.0005); E2F2 on chromosome 1p36.12 (P = 0.0004); CREB5 on chromosome 7p15.1 (P = 0.0002) and CHEK2 on chromosome 22q12.1 (P = 4.7 × 10⁻⁷).

The genomic inflation factor λ for the combined meta-analysis was 1.15 (adjusted value to 1000 cases and controls λ₁₀₀₀ = 1.01). This may be due to cryptic population structure not accounted for by adjusting for principal components. However, there was no residual inflation observed for association with clear cell and mucinous ovarian cancers (λ = 1.01 and 0.93, respectively) and minimal inflation for the larger set of SNPs on the iCOGS array that had not been selected as candidates for EOC susceptibility (λ = 1.07, λ₁₀₀₀ = 1.004).

Genetic association analyses of the CHEK2 gene locus

CHEK2 showed the strongest evidence of association with serous ovarian cancer risk, for SNP rs17507066 (odds ratio (OR) 0.86; 95% confidence interval (95% CI) 0.81–0.91). Because CHEK2 is a known moderately penetrant susceptibility gene for breast cancer, additional common variants in the region spanning this gene had been included on the iCOGS array at providing a greater density within the region than for other DNA repair genes (16). Genotype data were available for a further 176 variants in this sample set in addition to the 24 tagging SNPs originally evaluated. Further genotyping identified rs9625477 with a marginally more significant association with serous ovarian cancer risk (P = 2.4 × 10⁻⁷). The data for the association analysis of all 200 genotyped variants in the region in serous ovarian cancer are given in Supplementary Table 6, available at Carcinogenesis Online.

We further evaluated this region after imputing genotypes for variants identified through the 1000 Genomes Project for all participants of European ancestry (Supplementary Table 7, available at Carcinogenesis Online). After excluding poorly imputed SNPs, a total of 4785 SNPs with an imputation r² > 0.3 and an estimated MAF > 0.02% spanning a 2 Mb region on 22q12.1 (nucleotide position 28 000 000 to 30 000 000) were analyzed for their associations with high-grade serous ovarian cancer (HGSC) risk. This analysis identified multiple additional variants highly correlated with rs9625477, several of which were more significantly associated with disease risk. All imputed risk-associated variants with P value less than threshold 10⁻⁶ are given in Supplementary Table 7, available at Carcinogenesis Online. The most significant risk association was for SNP rs6005807 (OR 1.17, 95% CI 1.11–1.24, P = 1.1 × 10⁻³), which is correlated with rs9625477 (r² = 0.99). Risk associations for genotyped and imputed SNPs in HGSC across the regions are illustrated in Figure 1B.

We stratified risk associations by histological subtype. The most significant SNP for serous ovarian cancer (rs6005807) was more weakly associated with all invasive subtypes of EOC combined (P = 2.9 × 10⁻⁵) and showed no evidence of association for clear cell (P = 0.65), endometrioid (P = 0.55) or mucinous (P = 0.33) subtypes. However, other variants in the region (all imputed) showed subtype-specific associations: rs73871015 (r² = 0.97 with rs6005807) was the strongest risk allele for the clear cell subtype (P = 0.0002); rs34051361 (r² = 0.56 with rs6005807) was associated with the endometrioid subtype (P = 0.0001); and the variant 22:29126347:D (r² = 0.92 with rs6005807) was associated with the mucinous subtype (P = 0.0001). Summary results are given in Table 1.

Functional annotation of risk-associated variants

Two hundred and ninety-three variants (genotyped or imputed), representing the most likely candidate causal variants at the locus, had likelihood ratios greater than 1:100 compared with the most significant SNP in serous ovarian cancer. The majority of candidate causal variants were SNPs (267/293, 91.1%); the remaining 26/293 (8.9%) were indel polymorphisms. We annotated these SNPs with respect to protein coding genes, predicted functional motifs and regulatory elements cataloged in ENCODE and Haploreg. Two SNPs were located in protein coding regions of the TTC28 gene; both were synonymous and therefore unlikely to be of functional importance. The remaining variants were located in non-coding DNA regions: 274 (93.5%) were located within introns of the TTC28 gene; and 12 (4.1%) were located within introns of the CHEK2 gene (Figure 2). Fifty-four SNPs (18.4%) coincide with enhancer or promoter elements annotated in ENCODE, 51 SNPs (17.4%) are located in DNase hypersensitivity domains and 241 SNPs (82.3%) are predicted to alter transcription factor binding motifs in Haploreg (Supplementary Table 8, available at Carcinogenesis Online). To define further the overlaps between risk variants and putative functional features, we identified those variants predicted significantly to alter transcription factor binding sites (TFBSs) identified using data from FactorBook (32,36). We only considered TFBS variants that lie within regulatory DNA regions active in EOC precursor tissues. Active regulatory elements in normal ovarian and fallopian epithelial cells were profiled using formaldehyde assisted isolation of regulatory element sequencing (FAIRE-seq) to identify regions of open chromatin, and chromatin immunoprecipitation sequencing (ChIP-seq) for histone modification marks H3K4me1 and H3K27ac (37). We identified 25 instances where candidate SNPs altered TFBSs within active regulatory sequences (Supplementary Table 9, available at Carcinogenesis Online), suggesting that these SNPs may be the most likely candidate causal variants at this locus. HOCCOMOCO (38) was also used to identify transcription factors that may bind to the risk associated SNPs. We found that rs12166475 is predicted to affect binding of WT1, and that rs9620817 and rs16986509 are predicted to alter TFBSs for BRCA1; both transcription factors are known to be important in risk of development of HGSC. These data are summarized in Supplementary Table 10, available at Carcinogenesis Online and Figure 2C.
Functional analyses of candidate genes

We used somatic data to evaluate the role in EOC development of all protein-coding genes within a 1Mb region spanning the most significant risk-associated SNP (rs6005807) to identify the most likely susceptibility target gene. Six genes lie in the region: rs6005807 is located in an intron of TTC28 (tetratricopeptide repeat domain 28); 5' prime of TTC28 are CHEK2, HSCB (HscB mitochondrial iron-sulfur cluster co-chaperon), CCDC117 (coiled-coil domain containing 117), XBP1 (X-box binding protein 1), and ZNRF3 (zinc and ring finger 3). We evaluated somatic genetic alterations in primary ovarian tumors for these six genes using The Cancer Genome Atlas (TCGA) data and other public databases (summarized in Table 2).

These analyses revealed that 11% of high-grade serous ovarian cancer (HGSOC) cases showed copy number gain or amplification in the region spanning the six candidate genes, whereas homozygous deletions were rare (<1% cases). We identified a somatic coding sequence mutation in both the CHEK2 and ZNRF3 genes out of 316 sequenced HGSOC cases. The mutation in CHEK2 is a missense (R346H) predicted to be of ‘high impact’ (mutationassesor.org). The ZNRF3 mutation is also missense (P805H), but predicted to have little functional impact. Using another database of somatic mutation frequencies (COSMIC), which includes data for over 8000 tumors, we observed that CHEK2 was the most frequently mutated (in 2.5% of cases) of all the genes in the region (42).

We examined differences in gene expression between normal and cancer tissues (Figure 2D). CHEK2 gene expression was significantly higher in HGSOCs (n = 489) compared to normal fallopian tube tissues (P = 3.7×10−8). TTC28 was the only other
gene in the region that was differentially expressed in ovarian cancers with lower expression in tumors compared to normal control tissues \( (P = 0.01) \). We also evaluated the expression of these genes in a stepwise model of early-stage ovarian epithelial cell transformation driven by overexpression of the CMYC gene and mutant KRAS \((Figure 2E)\) \((43)\). In this model, CHEK2, HSBC,

**Table 1.** Risk associations for the top SNPs at the CHEK2 locus, by histological subtype

<table>
<thead>
<tr>
<th>Top SNP</th>
<th>EOC subtype</th>
<th>Position</th>
<th>Number of cases</th>
<th>SNP location (nearest gene)</th>
<th>( r^2 ) COGS</th>
<th>Odds ratio</th>
<th>95% CI</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs6005807</td>
<td>Serous</td>
<td>28934313</td>
<td>9608</td>
<td>Intron (TTC28)</td>
<td>1</td>
<td>1.17</td>
<td>1.11–1.24</td>
<td>( 1.1 \times 10^{-7} )</td>
</tr>
<tr>
<td>All invasive</td>
<td>15 397</td>
<td></td>
<td></td>
<td></td>
<td>1.12</td>
<td>1.07–1.18</td>
<td>( 2.9 \times 10^{-6} )</td>
<td></td>
</tr>
<tr>
<td>Clear cell</td>
<td>1172</td>
<td></td>
<td></td>
<td></td>
<td>1.04</td>
<td>0.89–1.20</td>
<td>( 0.65 )</td>
<td></td>
</tr>
<tr>
<td>Endometrioid</td>
<td>2385</td>
<td></td>
<td></td>
<td></td>
<td>1.03</td>
<td>0.93–1.15</td>
<td>( 0.55 )</td>
<td></td>
</tr>
<tr>
<td>Mucinous</td>
<td>1112</td>
<td></td>
<td></td>
<td></td>
<td>1.07</td>
<td>0.92–1.24</td>
<td>( 0.33 )</td>
<td></td>
</tr>
<tr>
<td>rs5752754</td>
<td>All invasive</td>
<td>28925542</td>
<td>15 397</td>
<td>Intron (TTC28)</td>
<td>0.94</td>
<td>1.11</td>
<td>1.06–1.16</td>
<td>( 2.4 \times 10^{-6} )</td>
</tr>
<tr>
<td>rs78371015</td>
<td>Clear cell</td>
<td>29245611</td>
<td>1172</td>
<td>Intergenic (ZNRF3)</td>
<td>0.97</td>
<td>1.40</td>
<td>1.17–1.67</td>
<td>( 0.0002 )</td>
</tr>
<tr>
<td>rs34051361</td>
<td>Endometrioid</td>
<td>29582557</td>
<td>2385</td>
<td>Intergenic (KREMEN1)</td>
<td>0.56</td>
<td>1.57</td>
<td>1.24–1.98</td>
<td>( 0.0001 )</td>
</tr>
<tr>
<td>chr22:29126347:D</td>
<td>Mucinous</td>
<td>29126347</td>
<td>1112</td>
<td>Intergenic (ZNRF3)</td>
<td>0.92</td>
<td>1.56</td>
<td>1.25–1.98</td>
<td>( 0.0001 )</td>
</tr>
</tbody>
</table>

**Figure 2.** Functional annotation of candidate genes and SNPs at the 22q12 locus. We analysed the genes and ovarian cancer risk associated variants in a 1MB region spanning the top ranked risk SNP at this locus: (A) The location of all protein coding genes in the 1 MB region spanning the ranked risk SNP; and the 298 SNPs in the region with a 100:1 likelihood of being causal; (B) Illustration of the regulatory elements identified by RNAseq and ChIPseq analysis (H3K4me1, H3K4me3 and H3K27ac) catalogued by ENCODE, illustrates the extent of overlap between candidate SNPs and epigenetic marks; (C) The three most significant candidate SNPs identified through a combination of functional annotation, eQTL analyses and transcription factor binding site prediction. Position weight matrices for each factor are shown with SNP position indicated. Ref, reference allele; Alt, alternative allele; PCT, percent of maximum position weight matrix score for each allele; (D) Gene expression in 489 HGSOCs and 8 normal fallopian tube tissue specimens performed by TCGA indicates CHEK2 is the most significantly differentially expressed gene in the region; (E) mRNA expression of each gene in an in vitro early stage transformation model of ovarian cancer. Expression in CMYC and KRAS transformed cells compared to untransformed immortalized ovarian epithelial (IOE) cells \( (P < 0.05) \).
Table 2. Candidate gene analyses and eQTL analyses

<table>
<thead>
<tr>
<th>Gene</th>
<th>Functiona</th>
<th>TCGA gene expression tumor (n = 489) versus normal FT (n = 8)</th>
<th>COSMIC somatic mutations</th>
<th>Expression quantitative trait locus (eQTL) analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P value</td>
<td>Mean Normal FT</td>
<td>Mean HGSOC</td>
</tr>
<tr>
<td>TTC28</td>
<td>Involved in mitosis and cytokinesis (39)</td>
<td>0.010</td>
<td>0.347</td>
<td>0.001</td>
</tr>
<tr>
<td>CHEK2</td>
<td>Involved in DNA damage response</td>
<td>3.7 \times 10^{-4}</td>
<td>−1.29</td>
<td>0.011</td>
</tr>
<tr>
<td>HSCB</td>
<td>Involved in iron-sulfur cluster biogenesis (40)</td>
<td>0.368</td>
<td>0.065</td>
<td>−0.020</td>
</tr>
<tr>
<td>CCDC117</td>
<td>Not characterized</td>
<td>0.098</td>
<td>−0.151</td>
<td>−0.005</td>
</tr>
<tr>
<td>XBP1</td>
<td>Transcription factor, involved in the immune system</td>
<td>0.130</td>
<td>0.268</td>
<td>−0.015</td>
</tr>
<tr>
<td>ZNRF3</td>
<td>Ubiquitin ligase, involved in Wnt signaling (41)</td>
<td>0.975</td>
<td>0.006</td>
<td>0.004</td>
</tr>
</tbody>
</table>

NS, no significant eQTLs detected; ND, not done; OSEC, ovarian surface epithelial cell; N/A, not available; in TCGA level 3 data there were no data available for HSCB, CCDC117 and ZNRF3.

*Gene functions were identified using the NCBI ‘Gene’ pages unless alternative references are provided. Blood eQTL analyses were performed using data from Ref (29), the top disease-associated eQTL SNP is shown for each gene. For CHEK2, HSCB, CCDC117 and XBP1, the top eQTL SNP was also disease associated. *Minor alleles were associated with increased expression of CHEK2, HSCB, CCDC117 and XBP1.
CCDC117, XBP1 and ZNF3 all showed significantly increased expression in ovarian epithelial cells (P < 0.05) that had undergone early stage neoplastic transformation, suggesting that the upregulation of these genes may be an early event in EOC development.

eQTL analyses

We used eQTL analysis to evaluate associations between risk genotypes and the levels of mRNA expression for candidate genes in the region (29). We looked for cis-eQTL associations in both normal and tumor tissues. We did not detect an eQTL for CHEK2 in normal ovarian/fallopian epithelial cultures. In peripheral blood samples (N > 5300) we observed a particularly strong association between rs12165715 and XBP1 expression (P = 1.16 × 10⁻⁶⁵) (29). We also detected associations between rs12166475 and CHEK2 expression (P = 2.70 × 10⁻¹⁴), rs9620817 and HSCB expression (P = 1.21 × 10⁻⁷), and rs16986509 and CCDC117 expression (P = 1.29 × 10⁻⁴). The variant most significantly associated with ovarian cancer risk at this locus (rs6005807) showed significant eQTL associations for CHEK2 (P = 9.22 × 10⁻⁸) and HSCB (P = 1.85 × 10⁻⁷). Importantly, for CHEK2, HSCB, CCDC117 and XBP1, the most significant eQTL SNPs were also disease associated.

These data are summarized in Table 2. Three of the top 25 candidate variants from in silico functional annotation were amongst the most significant eQTL associations: rs12166475, associated with expression of CHEK2 coincides with a TFBS for EGR1, a transcription factor previously implicated in EOC development; rs9620817, associated with HSCB expression, is predicted to alter CEBPB and ETS1 motifs; and rs16986509 associated with CCDC117 expression is predicted to alter TFBSs for TAL1 and UA2 (Figure 2C).

Finally, we evaluated eQTL associations for all six protein-coding genes in the region in 339 primary HGSO tissues using publicly available data from TCGA. Variations in gene expression were adjusted for changes in DNA copy number and methylation variation in each tumor. We observed no significant associations between rs6005807 and CHEK2, TTC28 or XBP1 at a P value threshold of 0.05, and false discovery rate threshold of 0.1. Details of the eQTL analyses are provided in Table 2.

Discussion

DNA repair mechanisms are important in the initiation and development of EOC and the current study represents the most comprehensive analysis of common genetic variation at DNA repair genes and EOC risk to date. We evaluated 2621 candidate variants spanning 143 gene regions at several different DNA repair pathways and found strong evidence of risk associations for SNPs at the CHEK2 gene locus that were just below the threshold for genome-wide significance. This is consistent with a smaller previous study in which we showed borderline evidence of association with clear cell, endometrioid and mucinous EOC at P values of 0.0002. Because these three EOC subtypes are less common than serous EOC, the weakness of these associations may simply reflect the smaller sample sizes available for their genotyping. One caveat to the association analyses lies in the minor residual inflation observed in the test statistics. This may be due to population structure but given that this inflation was greater than for other sets of SNPs this seems an unlikely explanation. An alternative explanation is an overall burden of weak susceptibility signals within this set of SNPs.

The most significant associations were identified using imputed genotypes, based on an estimated imputation r² that indicates a very high correlation between imputed genotypes and actual genotypes. It is therefore unlikely that there are other common variants within the region that may represent more highly associated SNPs than those already identified, and so these alleles represent the candidate causal variants for this locus. We identified 293 candidate causal polymorphisms that are virtually indistinguishable from each other with respect to their risk associations and any one (or even several) could be the causal SNP(s) influencing expression of the target susceptibility gene. Only two of the variants were in protein-coding regions and both were synonymous changes, suggesting that the causal SNP(s) likely reside in non-coding DNA. As such we neither know the functional basis for the genetic susceptibility, nor the target susceptibility gene(s). Our in silico analysis of these variants with respect to non-coding regulatory biofeatures profiled in multiple different cell lines by ENCODE, and in EOC precursor tissues, identified 25 risk SNPs that intersect with annotated functional elements. While this represents a relatively small number of candidate causal polymorphisms and functional targets at this locus, several other candidate causal SNPs may exist. None of the cell lines we evaluated have been comprehensively analysed for the full catalogue of non-coding regulatory elements and is possible that additional variants overlap regulatory marks that were not profiled; for example CTCF repressor marks and non-coding RNAs.

We identified three risk SNPs located within regulatory elements active in ovarian cells. These SNPs were also the most significant SNPs from eQTL analysis in the region: rs12166475 associated with CHEK2; rs9620187 associated with HSCB; and rs16986509 associated with CCDC117. In each case the minor allele was associated with increased gene expression; consequently higher CHEK2, HSCB and CCDC117 expression was associated with reduced cancer risk, whereas higher XBP1 expression was associated with higher cancer risk. The strongest candidate SNP is rs12166475, which alters the binding site for EGR1, a transcription factor involved in epithelial-to-mesenchymal
transition in EOC (46). The alternative allele of this variant is also predicted to increase the binding affinity of WT1, a biomarker commonly expressed in HGSOCs. WT1 can have both repressive and activating effects on gene expression (47). Additional functional analysis of the possible interaction between rs12166475 and CHEK2 will be required to validate these findings and to elucidate the transcriptional consequences of allele-dependent EGR1/WT1 binding at the site of this SNP. The SNP rs9620817, which is most significantly associated with HSCB expression, is also a strong candidate. This SNP is predicted to alter CEBPB and ETS1 transcription factor binding sites (TFBSs), although the difference in predicted binding affinity between the two SNP alleles is much greater for ETS1. Both rs16986509 and rs9620817 also alter binding sites for BRCA1, which may be of significance given that BRCA1-associated pathways are deregulated in approximately half of all HGSOCs (48). BRCA1 can function as a co-repressor or co-activator, and regulates gene expression by interacting with a myriad of different transcription factors, including TP53 and CMYC (49).

Although we conditioned our analysis at 22q12.1 on tagging variants spanning the CHEK2 gene, five other genes lie within a 500-kb region at either side of the most risk associated SNP that could be the target of risk-associated variants at this locus. However, somatic analysis of ovarian tumors from TCGA suggests that CHEK2 is the most likely target. It is the only gene in the region that is differentially expressed in ovarian tumors compared to normal fallopian tube tissues, suggesting that it may play a role in EOC development. While CHEK2 was overexpressed in ovarian tumors compared to normal fallopian tubes, in our eQTL analyses reduced CHEK2 expression was associated with increased cancer risk, which may suggest that overexpression occurs at later stages of tumorigenesis but lower CHEK2 expression is involved in early cancer development. This hypothesis is consistent with the moderate risk of breast cancer conferred by CHEK2 loss-of-function variants, where large population-based studies report estimated odds ratios for rare protein-truncating and splice-junction variants on the order of 6.18 (95% CI: 1.76–21.8) and 8.75 (95% CI: 1.06–72.2) for missense substitutions (50). Breast and EOC have shared genetic etiology for both high and low penetrance susceptibility genes, providing a rationale for why germline genetic variants in or around CHEK2 may be associated with EOC risk. No similar rationale applies for other genes in the region but eQTL analyses identified several genotype-gene expression associations that indicate alternative candidate target genes. The strongest association was with the XBPI gene in peripheral lymphocytes, although the most significant eQTL SNP for this gene was not predicted to alter a TFBS within an active regulatory element in EOC precursor cells. We also identified highly statistically significant cis-eQTL associations between risk SNPs and expression of both CHEK2 and HSCB. However, the true importance of these eQTL associations is unclear given that they were identified in tissues that are not associated with EOC development. Several previous studies have highlighted the importance of tissue-specific gene expression when evaluating eQTLs, and stressed the need to perform eQTL analysis in tissues relevant to disease development (51). We did not identify eQTL associations for any of these genes in primary EOCs, although these studies were underpowered.

In summary, we provide evidence that common genetic variants in a region on chromosome 22q12.1 are associated with risk of serous ovarian cancer. The most likely target susceptibility gene at this locus is CHEK2 based on a combination of its known role in DNA damage response pathways, somatic variation in gene expression suggesting a role in EOC development, and a significant eQTL association with risk-associated variants. Future studies will be needed to increase the power of our genetic association studies by increasing sample size to confirm that this region is a true susceptibility locus for ovarian cancer. Finally, detailed functional characterization of this locus will be needed to confirm the functional impact of these candidate SNPs on their regulatory elements and establish their interactions and influence on the target susceptibility gene, as has been shown for other common variant susceptibility genes and alleles.

**Supplementary material**

Supplementary Tables 1–10 and Figure 1 can be found at [http://carcin.oxfordjournals.org/](http://carcin.oxfordjournals.org/)

**Funding**

The scientific development and funding for this project were supported by the following: the Genetic Associations and Mechanisms in Oncology (GAME-ON): a NCI Cancer Post-GWAS Initiative (U19-CA148112); the COGS project is funded through a European Commission’s Seventh Framework Programme grant (agreement number 223175-HEALTH-F2-2009-223175); the Ovarian Cancer Association Consortium is supported by a grant from the Ovarian Cancer Research Fund thanks to donations by the family and friends of Kathryn Sladek Smith (PPD/RPCI07); and Department of Defense Award (W81XWH-12-1-0561, W81XWH-07-0449). F.M. was supported by National Institutes of Health K07-CA080668, S.S.T. and E.M.P. were supported in part by Department of Defense Award W81XWH-12-1-0561. Funding of the constituent studies was provided by: California Cancer Research Program (00-01389V-20170, 210200); Cancer Prevention Institute of California; Department of Defense (DAMD17-02-1-0666, DAMD17-02-1-0669, W81XWH-10-1-02802); Fred C. and Katherine B. Andersen Foundation; Lon V Smith Foundation grant LVS-39420; Mayo Foundation; Minnesota Ovarian Cancer Alliance; National Institutes of Health (K07-CA095666, K07-CA143047, K22-CA138563); National Center for Research Resources/General Clinical Research Center grant M01-RR000506, N01-CN025403, N01-CN55424, N01-PC67001, N01-PC67010, P01-CA17054, P50-CA14089, R01-CA15083, P30-CA072720, P50-CA105009, F50-CA136393, P50-CA159981, R01-CA058860, R01-CA074850, R01-CA080742, R01-CA092044, R01-CA112523, R01-CA122443, R01-CA126841, R01-CA160565, R01-CA54419, R01-CA58598, R01-CA61132, R01-CA67016, R01-CA83918, R01-CA87538, R01-CA95023, R01-CA96378, R01-CA114343, R03-CA113148, R03-CA115195, U01-CA69417, U01-CA17966), U1M1-CA182910; Rutgers Cancer Institute of New Jersey; and the US Public Health Service (PSA-042205). Personal support: K.L is supported by a K99/R00 grant from the National Cancer Institute (Grant number 1K99CA184415-01). D.F.E. is a Principal Research Fellow of Cancer Research UK. G.C.-T. and P.M.W. are supported by the National Health and Medical Research Council. B.K. holds an American Cancer Society Early Detection Professorship (SIOP-06-258-01- COUN) and the National Center for Advancing Translational Sciences (NCATS), Grant UL1TR000124.

**Acknowledgements**

This study would not have been possible without the contributions of the following: A. M. Dunning, P. Hall (COGS); D. C. Tessier, F. Bacot, D. Vincent, S. LaBoissière and F. Robidoux and the staff of the genotyping unit (Genome Quebec); D. C. Whitman, P. M.
References


