Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31

Jennifer Permuth-Wey, Kate Lawrenson, Howard C. Shen et al.$^\dagger$

Epithelial ovarian cancer (EOC) has a heritable component that remains to be fully characterized. Most identified common susceptibility variants lie in non-protein-coding sequences. We hypothesized that variants in the 3' untranslated region at putative microRNA (miRNA)-binding sites represent functional targets that influence EOC susceptibility. Here, we evaluate the association between 767 miRNA-related single-nucleotide polymorphisms (miRSNPs) and EOC risk in 18,174 EOC cases and 26,134 controls from 43 studies genotyped through the Collaborative Oncological Gene–environment Study. We identify several miRSNPs associated with invasive serous EOC risk (odds ratio $= 1.12, P = 10^{-8}$) mapping to an inversion polymorphism at 17q21.31. Additional genotyping of non-miRSNPs at 17q21.31 reveals stronger signals outside the inversion ($P = 10^{-10}$). Variation at 17q21.31 is associated with neurological diseases, and our collaboration is the first to report an association with EOC susceptibility. An integrated molecular analysis in this region provides evidence for ARHGAP27 and PLEKHM1 as candidate EOC susceptibility genes.

$^\dagger$A full list of authors and their affiliations appears at the end of the paper.
Genome-wide association studies (GWAS) have identified hundreds of genetic variants conferring low penetrance susceptibility to cancer. More than 90% of these variants lie in non-protein-encoding sequences including non-coding RNAs and regions containing regulatory elements (that is, enhancers, promoters, untranslated regions (UTRs)). The emerging hypothesis is that common variants within non-coding regulatory regions influence expression of target genes, thereby conferring disease susceptibility.

MicroRNAs (miRNAs) are short non-coding RNAs that regulate gene expression post-transcriptionally by binding primarily to the 3’UTR of target messenger RNA (mRNA), causing translational inhibition and/or mRNA degradation. MiRNAs have been shown to have a key role in the development of epithelial ovarian cancer (EOC). We and others have found evidence that various miRNA-related single-nucleotide polymorphisms (miRSNPs) are associated with EOC risk, suggesting they may be key disruptors of gene function and contributors to disease susceptibility. However, studies of miRSNPs that affect miRNA–mRNA binding have been restricted by small sample sizes, and therefore have limited statistical power to identify associations at genome-wide levels of significance.

Large-scale studies and more systematic approaches are warranted to fully evaluate the role of miRSNPs and their contribution to disease susceptibility.

Here, we use the in silico algorithms, TargetScan and Pictar, to predict miRNA:miRNA-binding regions involving genes and miRNAs relevant to EOC, and align identified regions with SNPs in the Single Nucleotide Polymorphism database (dbSNP). We then genotype 1,003 miRSNPs (or tagging SNPs with dbSNP) (Methods). We then perform on a custom Illumina Infinium iSelect array designed by the Collaborative Oncological Gene–environment Consortium (OCAC) (Supplementary Table S1). Genotyping was performed on a custom Illumina Infinium iSelect array designed as part of the Collaborative Oncological Gene–environment Study (COGS), an international effort that evaluated 211,155 controls from 43 studies from the Ovarian Cancer Association Consortium (Supplementary Table S1). Genotyping was observed across study sites (Supplementary Fig. S2).

The four susceptibility alleles identified here reside on the H2 haplotype that is reported to be rare in Africans and located within the inversion (Fig. 1). Chromosomes with the non-inverted or inverted segments of 17q21.31, respectively, known as haplotype 1 (H1) and haplotype 2 (H2), represent two distinct lineages that diverged ~3 million years ago and have not undergone any recombination event.

The four susceptibility alleles identified here reside on the H2 haplotype that is reported to be rare in Africans and exhibits strong linkage disequilibrium (LD) among Europeans, consistent with our findings. The H2 haplotype has a frequency of 22% among European women in our primary analysis (Table 1) but only 3.2% and 0.3% among Africans (151 invasive cases, 200 controls) and Asians (716 invasive cases, 1573 controls), respectively.

### Table 1 | Tests of association by histological subtype for directly genotyped and imputed SNPs at 17q21.31 most strongly associated with invasive epithelial ovarian cancer risk among Europeans.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Major &gt; minor allele</th>
<th>Coordinate*</th>
<th>MAF</th>
<th>Subtype</th>
<th>Number of cases (versus 23,491 controls)</th>
<th>Per-allele OR (95% CI)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1052587†</td>
<td>(T &gt; C)</td>
<td>44102604</td>
<td>0.22</td>
<td>All invasives</td>
<td>14,533</td>
<td>1.10 (1.06–1.13)</td>
<td>1.9 × 10^-7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Serous</td>
<td>8,371</td>
<td>1.12 (1.08–1.17)</td>
<td>4.6 × 10^-8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Endometrioid</td>
<td>2,068</td>
<td>1.11 (1.04–1.19)</td>
<td>5.2 × 10^-3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Clear cell</td>
<td>1,025</td>
<td>0.98 (0.88–1.09)</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mucinous</td>
<td>944</td>
<td>1.07 (0.96–1.20)</td>
<td>0.22</td>
</tr>
<tr>
<td>rs12942666δ</td>
<td>(A &gt; G)</td>
<td>43499839</td>
<td>0.22</td>
<td>All invasives</td>
<td>14,533</td>
<td>1.11 (1.07–1.15)</td>
<td>3.3 × 10^-8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Serous</td>
<td>8,371</td>
<td>1.13 (1.11–1.20)</td>
<td>1.0 × 10^-9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Endometrioid</td>
<td>2,068</td>
<td>1.10 (1.02–1.18)</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Clear cell</td>
<td>1,025</td>
<td>1.04 (0.92–1.16)</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mucinous</td>
<td>944</td>
<td>1.04 (0.92–1.16)</td>
<td>0.55</td>
</tr>
<tr>
<td>rs29600000</td>
<td></td>
<td>(T &gt; C)</td>
<td>43534353</td>
<td>0.18</td>
<td>All invasives</td>
<td>14,533</td>
<td>1.12 (1.08–1.16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Serous</td>
<td>8,371</td>
<td>1.16 (1.12–1.20)</td>
<td>3.3 × 10^-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Endometrioid</td>
<td>2,068</td>
<td>1.12 (1.03–1.20)</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Clear cell</td>
<td>1,025</td>
<td>1.05 (0.93–1.16)</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mucinous</td>
<td>944</td>
<td>1.03 (0.90–1.15)</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Abbreviations: CI, confidence interval; MAF, minor allele frequency in controls; OR, odds ratio.

*Genome build NCBI B37/human genome build 19 assembly.

†OR and 95% CI per copy of the minor allele, with adjustment for the first five eigenvalues from principal components analysis.

‡rs1052587 is the most statistically significant miRNA-binding site SNP among all invasives and serous; it resides in a putative miRNA-binding site between microtubule-associated protein tau (MAPT) and miR-34a-5p (chr 17:13942325-13944231).

δrs12942666 is a SNP at 17q21.31 that was directly genotyped as part of COGS; it is in strong linkage disequilibrium (r² = 0.99) with two other 17q21.31 SNPs that were directly genotyped but had less optimal clustering: rs2077606 (P = 3.9 × 10^-10 for the serous subtype) and rs17631303 (P = 4.7 × 10^-7 for the serous subtype).

||rs29600000 represents the most statistically significant SNP at 17q21.31 (among all invasives) that was imputed from the 1000 genome Project reference panel with an R-squared quality metric of 95% or greater (http://www.1000genomes.org/page.php).
with SNPs from the 17q21 locus (spanning 46.2–46.5 MB, build 37).

Genotype clustering was poor for rs2077606, but clustering was good for its correlated SNP, rs12942666 ($\chi^2 = 0.99$) and so results for this SNP are presented instead (Supplementary Fig. S3; Table 1). Subgroup analysis revealed marginal evidence of association for rs12942666 with endometrioid ($P = 0.04$), but not mucinous or clear cell EOC subtypes (Table 1), and results were consistent across studies (Supplementary Fig. S4). Rs12942666 is correlated with the top-ranked mirSNP, rs1052587 ($\chi^2 = 0.76$) (Fig. 1). To evaluate whether associations observed for rs12942666 and rs1052587 represented independent signals, stepwise logistic regression was used; only rs12942666 was retained in the model. This suggests that the cluster which includes rs12942666 is driving the association with EOC risk that was initially identified through the candidate mirSNPs.

Functional and molecular analyses. To evaluate functional evidence for candidate genes, risk-associated SNPs, and regulatory regions at 17q21.31, we examined a 1-MB region centred on rs12942666 using a combination of locus-specific and genome-wide assays and in silico analyses of publicly available data sets, including The Cancer Genome Atlas (TCGA) Project18 (see Methods). Rs12942666 and many of its correlated SNPs lie within introns of Rho GTPase activating protein 27 (ARHGAP27) or its neighbouring gene, pleckstrin homology domain containing family M with RUN domain member 1 (PLEKHM1) (Supplementary Table S2). There are another 15 known protein-coding genes within the region: KIF18B, C1QL1, DCAKD, NMT1, PLCD3, ABCB4, HEXIM1, HEXIM2, FMNL1, C17orf46, MAP3K14, C17orf69, CRHR1, IMP5 and MAFT (Fig. 2a).

To evaluate the likelihood that one or more genes within this region represent target susceptibility gene(s), we first analysed expression, copy number variation and methylation involving these genes in EOC tissues and cell lines (Fig. 2b–g; Supplementary Tables S3 and S4). Most genes showed significantly higher expression ($P < 10^{-4}$ in EOC cell lines versus normal ovarian cancer precursor tissues (OCPTs); ARHGAP27 showed the most pronounced difference in gene expression between cancer and normal cells ($P = 10^{-16}$) (Fig. 2b and Supplementary Table S3). For nine genes, we also found overexpression in primary high-grade serous (HGS) EOC tumours versus normal ovarian tissue in at least one of two publicly available data sets, TCGA series of 568 tumours18 and/or the Gene Expression Omnibus series GSE18520 data set consisting of 53 tumors19 (Fig. 2c and Supplementary Table S3). Analysis of DNA copy number variation in TCGA revealed frequent loss of heterozygosity in this region rather than copy number gains (Supplementary Fig. 5a–b; Supplementary Methods). We observed significant hypomethylation ($P < 0.01$) in ovarian tumours compared to normal tissues for DCAKD, PLCD3, ABCB4, FMNL1 and PLEKHM1 (Fig. 2d and Supplementary Table S4), which is consistent with the overexpression observed for DCAKD, PLCD3 and FMNL1. Taken together, these data suggest that the mechanism underlying overexpression may be epigenetic rather than based on copy number alterations.

We evaluated associations between genotypes for the top risk SNP rs12942666 (or a tagSNP) and expression of all genes in the region (expression quantitative trait loci (eQTL) analysis) in normal OCPTs, lymphoblastoid cell lines and primary ovarian tumours from TCGA. The only significant eQTL association observed ($P < 0.05$) in normal OCPTs was for ARHGAP27 ($P = 0.04$) (Fig. 2e; Supplementary Table S3). Because rs12942666 was not genotyped in tissues analysed in TCGA, we used data for its correlated SNP rs2077606 ($\chi^2 = 0.99$) to evaluate eQTLs in tumour tissues. Rs2077606 genotypes were strongly correlated in EOC tissues (Fig. 2e, orange line).
associated with PLEKHM1 expression in primary HGS-EOCs
\(P = 1 \times 10^{-4}\) (Fig. 2f; Supplementary Table S3). We also detected associations between rs12942666 (and rs2077606) genotypes and methylation for PLEKHM1 and CRHR1 in primary ovarian tumours \(P = 0.020\) and 0.001, respectively) using methylation quantitative trait locus analyses (Fig. 2g; Supplementary Table S4). Finally, the Catalogue of Somatic Mutations in Cancer database\(^{20}\) showed that nine genes in the region, including PLEKHM1, have functionally significant mutations in cancer, although for most genes mutations were not reported in ovarian carcinomas (Supplementary Table S3).

Taken together, these data suggest that several genes at the 17q21.31 locus may have a role in EOC development. The risk-associated SNPs we identified fall within non-coding DNA,
Figure 2 | Expression and methylation analyses at the 17q21.31 ovarian cancer susceptibility locus. (a) Genomic map and LD structure. The location and approximate size of 17 known protein-coding genes (grey) and one microRNA (blue) in the region are shown relative to the location of rs12942666. Orange indicates the location of the inversion polymorphism, and green indicates the region outside the inversion. (b) Gene expression (EOC and normal cell lines). Gene expression analysis in epithelial ovarian cancer (EOC) cell lines (T; n = 51) compared with normal ovarian surface epithelial cells (OSECs) and fallopian tube secretory epithelial cells (FTSECs) (N; n = 73). (c) eQTL analysis (primary EOCs). Boxplots of The Cancer Genome Atlas (TCGA) Affymetrix U133A-array-based gene expression in primary high-grade serous ovarian tumours (T; n = 568) and normal fallopian tube tissues (N; n = 8). Where data were not available in TCGA, gene expression data from the Gene Expression Omnibus series GSE18520 data set containing 53 high-grade serous tumours and 10 normal ovarian tissues are shown (indicated by a red asterisk). (d) Methylation (primary tumours and normal tissue). Methylation analysis of 106 high-grade serous ovarian tumours compared with normal ovarian tissues (n = 7). Methylation data were generated for CpG site(s) associated with each gene using the Illumina 450 methylation array. Pairwise analysis of methylation for an individual CpG for each gene is based on the CpG with most significant inverse relationship to gene expression (that is, cis negative), for a subset of 43 tumours having available gene expression data. Statistically significant cis-negative probes are indicated by a red open circle. (e) eQTL analysis (OSecs/FTSECs). eQTL analysis comparing expression of each gene to genotype for the most statistically significant SNP at 17q21.31 (rs12942666), for 73 normal OSEC/FTSEC lines. Data are presented as box plots comparing expression levels in cases carrying rare homozygotes/heterozygotes, with cases homozygous for the common allele. (f) eQTL analysis (primary EOCs). eQTL analysis comparing expression of each gene by genotype using level 3 gene expression profiling data from Agilent 244K custom arrays and level 2 genotype data from the Illumina IM-Duo BeadChip for 568 high-grade serous ovarian cancer patients from TCGA. In all panels *P < 0.05, **P < 0.01, ***P < 0.001. Grey X’s indicate data not available. Here, genotype data for rs2077606 is used (rather than rs12942666) because rs12942666 was not genotyped in the TCGA data set.

suggesting the functional SNP(s) may be located within an enhancer, insulator or other regulatory element that regulates expression of one of the candidate genes we evaluated. One hypothesis emerging from these molecular analyses is that rs12942666 (or a correlated SNP) mediates regulation of PLEKHM1, a gene implicated in osteopetrosis and endo-toxisis and/or ARHGAP27, a gene that may promote carcinoma-genesis through dysregulation of Rho/Rac/Cdc42-like GTPases. To identify the most likely candidate for being the causal variant at 17q21.31, we compared the difference between log-likelihoods generated from un-nested logistic regression models for rs12942666 and each of 198 SNPs in a 1-MB region featured in Supplementary Table 2. As expected, the log-likelihoods were very similar due to the strong LD; no SNPs emerged as having a likelihood ratio > 20 for being the causal variant.

To explore the possible functional significance of rs12942666 and strongly correlated variants (r² > 0.80), we then generated a map of regulatory elements around rs12942666 using ENCyclopedia of DNA Elements (ENCODE) data and formaldehyde-assisted isolation of regulatory elements sequencing analysis of OCPTs (Supplementary Methods). We observed no evidence of putative regulatory elements coinciding with rs12942666 or correlated SNPs (Fig. 3a). A map of regulatory elements in the entire 1-MB region can be seen in Supplementary Fig. 5c–f. We subsequently used in silico tools (ANNOVAR23, SNPInfo24 and SNPnexus25) to evaluate the putative function of possible causal SNPs (Supplementary Methods). Of 50 SNPs with possible functional roles, more than 30 reside in putative transcription factor binding sites (TFBS) within or near PLEKHM1 or ARHGAP27; 12 SNPs may affect methylation or miRNA binding, and two are non-synonymous coding variants predicted to be of no functional significance (Supplementary Table S2).

As most of the top-ranked 17q21.31 SNPs with putative functions (including two of the top directly genotyped SNPs, rs2077606 and rs17631303) are predicted to lie in TFBS (Supplementary Table S2), we used the in silico tool, JASPAR26, to further examine TFBS coinciding with these SNPs. Two SNPs scored high in this analysis (Supplementary Table S5); the first, rs12946900, lies in a GAGGAA motif and canonical binding site for SPIB, an Ets family member27. Ets factors have been implicated in the development of ovarian cancer and other malignancies28, but little evidence supports a specific role for SPIB in EOC aetiology. The second hit was for rs2077606, which lies in an E-box motif CACCTG at the canonical binding site for ZEB1 (chr. 10p11.2), a zinc-finger E-box binding transcription factor that represses E-cadherin29,30 and contributes to epithelial–mesenchymal transition in EOCs31.

We analysed expression of SPIB and ZEB1 in primary ovarian cancers using TCGA data; we found no significant difference in SPIB expression in tumours compared with normal tissues (Fig. 3bi). In contrast, ZEB1 expression was significantly lower in primary HGS-EOCs compared with normal tissues (P = 0.005) (Fig. 3bii). We validated this finding using qPCR analysis in 123 EOC and OCPT cell lines (P = 8.8 × 10⁻⁴) (Fig. 3biii). As rs2077606 lies within an intron of PLEKHM1, this gene is a candidate target for ZEB1 binding at this site. Our eQTL analysis also suggests ARHGAP27 is a strong candidate ZEB1 target at this locus; ARHGAP27 expression is highest in OCPT cell lines carrying the minor allele of rs2077606 (P = 0.034) (Fig. 3ci). Although we observed no eQTL associations between rs2077606 and ZEB1 expression in lymphoblastoid cell lines (Fig. 3cii), we found evidence of eQTL between rs2077606 and ZEB1 expression in HGS-EOCs (P = 0.045) (Fig. 3ciii). ZEB1 binding at the site of the common allele is predicted to repress gene expression whereas loss of ZEB1 binding conferred by the minor allele may enable expression of ARHGAP27, consistent with the eQTL association in OCPTs (Fig. 3cii). Although these data support a repressor role for ZEB1 in EOC development and suggest ARHGAP27 may be a functional target of rs2077606 (or a correlated SNP) in OCPTs through trans-regulatory interactions with ZEB1, it is important to investigate additional hypotheses as we continue to narrow down the list of target susceptibility genes, SNPs, and regulatory mechanisms that contribute to EOC susceptibility at this locus.

Discussion

The present study represents the largest, most comprehensive investigation of the association between putative miRSNPs in the 3’UTR and cancer risk. This and the systematic follow-up to evaluate associations with EOC risk for non-miRSNPs in the region identified 17q21.31 as a new susceptibility locus for EOC. Although the miRSNPs identified here may have some biological significance, our findings suggest that other types of variants in non-coding DNA, especially non-miRSNPs at the 17q21.31 locus, are stronger contributors to EOC risk. It is possible, however, that highly significant miRSNPs exist that were not identified in our study because (a) they were not pre-selected for evaluation (that
is, they do not reside in a binding site involving miRNAs or genes with known relevance to EOC, or they reside in regions other than the 3'UTR\(^{3,4}\) and/or (b) they were very rare and could not be designed or detected with our genotyping platform and sample size, respectively. Despite these limitations, the homogeneity between studies of varying designs and populations in the OCAC

**Figure 3** | The non-coding landscape and eQTL associations for the rs2077606 susceptibility SNP at 17q21.31. (a) Analysis of the chromatin landscape at ARHGAP27 and PLEKHM1 in normal ovarian surface epithelial and fallopian tube secretory epithelial cells (OSECs/FTSECs) by formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq). Alignment with ENCODE FAIRE-seq tracks (shown) and ChIP-seq tracks (not shown) from non-EOC-related cell lines reveals open chromatin peaks corresponding to (a) promoters (b) CTCF insulator binding sites and (c) H3K4me3 signals, suggestive of a dynamic regulatory region. An H3K4me3 signal at a coding ARHGAP27 mRNA variant (c) located between the genes is highly pronounced in OSEC/FTSEC, suggesting tissue-specific expression and function. Several of the top-ranking SNPs fall within TFBS (Supplementary Table S2). rs12942666 did not coincide with TFBS, but tightly linked SNPs, rs12946900 and rs2077606 fell within predicted binding sites for SPIB and ZEB1, respectively. (b) We analysed the expression of SPIB and ZEB1 in primary high-grade serous tumours from TCGA and found (i) no significant change in SPIB expression but (ii) significant downregulation of ZEB1 in tumours compared with normal tissues. (iii) QPCR analysis of ZEB1 expression in 73 OCPT and 50 EOC cell lines supported the finding that ZEB1 expression is lower in cancer cell lines compared with normal precursor tissues. (c) eQTL analysis in OSECs/FTSECs for different alleles of rs2077606. (i) There was a significant eQTL for ARHGAP27, with the minor (A) allele being associated with increased ARHGAP27 expression. (ii) There was no evidence of an association between rs2077606 genotypes and ARHGAP27 expression in lymphoblastoid cell lines, suggesting this association may be tissue-specific. (iii) We observed a borderline significant eQTL association between ZEB1 mRNA and rs2077606 in tumours from TCGA, with the minor risk allele also associated with lower expression.
and the genome-wide levels of statistical significance imply that all detected associations are robust. Furthermore, molecular correlate analyses of genes within the region suggest that cis-acting genetic variants influencing non-coding DNA regulatory elements, miRNAs and/or methylation underlie disease susceptibility at the 17q21.31 locus. Finally, these studies point to a subset of candidate genes (that is, PLEKH1, ARHGAP27) and a transcription factor (that is, ZEB1) that may influence EOC initiation and development.

This novel locus is one of eleven loci now identified that contains common genetic variants conferring high penetrance susceptibility to EOC in the general population. Genetic variants at several of these loci influence risks of more than one cancer type, suggesting that several cancers may share common mechanisms. For example, alleles at Sp15.33 and 19p13.1 are associated with estrogen-receptor-negative breast cancer and serious EOC susceptibility,22,35, and variants at 8q24 are associated with risk of EOC and other cancers.21,25 Genetic variation at 17q21.31 is also associated with a frontotemporal dementia-spectrum disorder, Parkinson’s disease, developmental delay and alopecia.37–42 Through COGS, the CIMBA also recently identified 17q21.31 variants that modify EOC risk in BRCA1 and BRCA2 carriers (P < 10−8 in BRCA1/2 combined).16 In particular, rs17631303, which is perfectly correlated with rs2077606 and rs12942666, was among the top-ranking SNPs detected by CIMBA16. Consistent with our findings, CIMBA also provide data that suggest EOC risk is associated with altered expression of one or more genes in the 17q21.31 region.16. Thus, results from this large-scale collaboration support a role for this locus in both BRCA1-2- and non-BRCA1-2-mediated EOC development. Before these findings can be integrated with variants from other confirmed loci and non-genetic factors to predict women at greatest risk of developing EOC and provide options for medical management of these risks, continued efforts will be needed to fine map the 17q21.31 region and to fully characterize the functional and mechanistic effects of potential causal SNPs in disease aetiology and development.

Methods

Study population. Forty-three individual OCAC studies contributed samples and data to the COGS initiative. Nine of the 43 participating studies were case-only (GRR, HS, LAX, OR, PVD, RMH, SOC, SRO, UKR); cases from these studies were pooled with case-control studies from the same geographic region. The two national Australian case-control studies were combined into a single study to create a case-control set for participating OCAC studies. Details regarding the 43 participating OCAC studies are summarized in Supplementary Table S1. Briefly, cases were women diagnosed with histologically confirmed primary EOC (invasive or low malignant potential), fallopian tube cancer or primary peritoneal cancer ascertained from population- and hospital-based studies and cancer registries. The majority of OCAC cases (>90%) do not have a family history of ovarian or breast cancer in a first-degree relative, and most have not been tested for BRCA1/2 mutations as a part of their parent study. Controls were women without a current or prior history of ovarian cancer with at least one ovary intact at the reference date. All studies had data on disease status, age at diagnosis/interview, self-reported racial group and histologic subtype. Most studies frequency-matched cases and controls on age group and race.

Selection of candidate genes and SNPs. To increase the likelihood of identifying miRSNPs with biological relevance to EOC, we reviewed published literature and consulted public databases to generate two lists of candidate genes: (1) 55 miRNAs with biological relevance to EOC, generated start and end coordinates for regions of miRNA binding, and database SNP version 129 was mined to identify SNPs falling within the designated binding regions. Of 3,246 unique miRSNPs that were identified, 1,102 obtained adequate design scores using Illumina’s Assay Design Tool. The majority (1,108, 98.3%) of the 1,102 SNPs resided in predicted sites of miRNA binding and therefore represent miRSNPs (r2 > 0.80) for miSNPs that were not designable or had poor-to-moderate design scores. Ninety-nine of the 1102 SNPs failed during custom assay development, leaving a total of 1,003 SNPs that were designed and genotyped.

Genotyping and QC. The candidate miRSNPs selected for the current investigation were genotyped using a custom Illumina Infinium Select Array as part of the international COGS, an effort to evaluate 211,155 genetic variants for association with the risk of ovarian, breast and prostate cancer. Samples and data were included from several consortia, including OCAC, the Breast Cancer Association Consortium, the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA) and the Consortium of Investigators of Modifiers of BRCA1/2 in the Genome (PRACTICAL). Although one of the primary goals of COGS was to replicate and fine-map findings from pooled GWAS from each consortium, this effort also aimed to genotype candidate SNPs of interest (such as the miRSNPs). The genotyping and QC process has been described recently in our report of OCAC’s pooled GWAS findings. Briefly, COGS genotyping was conducted at six centres, two of which were used for OCAC samples: McGill University and Genome Quebec Innovation Centre (Montréal, Canada) (n = 19,806) and Mayo Medical Genomics Facility (n = 27,824). Each 96-well plate contained 250 ng genomic DNA (or 500 ng whole genome-amplified DNA). Raw intensity data files were sent to the COGS data coordination centre at the University of Cambridge for genotype calling and QC using the GenCall algorithm.

Sample QC. One thousand two hundred and seventy-three OCAC samples were genotyped in duplicate. Genotypes were discordant for greater than 40 per cent of SNPs for 22 pairs. For the remaining 1,251 pairs, concordance was greater than 99.6 per cent. In addition, we identified 245 pairs of samples that were unexpected genotypic duplicates. Of these, 137 were phenotypic duplicates and judged to be from the same individual. We used identity-by-state to identify 618 pairs of first-degree relatives. Samples were excluded according to the following criteria: (1) 1,133 samples with a conversion rate (the proportion of SNPs successfully called per sample) of less than 95 per cent; (2) 169 samples with heterozygosity > 5 x 10−7 from the intercontinental ancestry-specific mean heterozygosity; (3) 65 samples with ambiguous sex; (4) 269 samples with the lowest call rate from a first-degree relative pair; (5) 1,686 samples that were either duplicate samples that were non-discordant for genotype or genotypic duplicates that were not concordant for phenotype. A total of 44,308 eligible subjects including 18,174 cases and 26,134 controls were available for analysis.

SNP QC. The process of SNP selection by the participating consortia has been summarized previously. In total, 211,155 SNP assays were successfully designed, including 23,239 SNPs nominated by OCAC. Overall, 94.3% of OCAC-nominated SNPs were genotyped. QC SNPs: (1) the call rate was > 95% with MAF > 5% or < 99% with MAF < 5% (n = 5,201); (2) they were monomorphic upon clustering (n = 2,587); (3) P-values of HWE in controls were < 10−7 (n = 2,914); (4) there was greater than 2% discordance in duplicate pairs (n = 22); (5) no genotypes were called (n = 1,311). Of 1,003 candidate miRSNPs genotyped, 767 (76%) were available for analysis. Of the 237 miRSNPs that were excluded were monomorphic (n = 158, 67%). Genotype intensity cluster plots were visually inspected for the most strongly associated SNPs.

Population stratification. HapMap DNA samples for European (CEU, n = 60), African (YRI, n = 53) and Asian (JPT + CHB, n = 88) populations were also genotyped using the COGS Select. We used the program LAMPM3 to estimate intercontinental ancestry based on the HapMap (release no. 23) genotype frequency data for these three populations. Eligible subjects with >90 per cent. European ancestry were defined as European (n = 39,773) and those with greater than 80 per cent. Asian or African ancestry were defined as Asian (n = 2,382) or African, respectively (n = 387). All other subjects were defined as being of mixed ancestry (n = 1,766). We then used a set of 37,090 unlinked SNPs to perform principal components analysis within each major population subgroup. To enable this analysis on very large sample sizes, we used an in-house program written in C++ using the Intel MKL libraries for eigenvectors (available at http://ccge.medschl.cam.ac.uk/software/).

Tests of association. We used unconditional logistic regression treating the number of minor alleles carried as an ordinal variable (log-ordinal model) to evaluate the association between each SNP and EOC risk. Separate analyses were carried out for each ancestry group. The model for European subjects was adjusted for population substructure by including the first five eigenvalues from the principal components analysis. African- and Asian ancestry-specific estimates were obtained after adjustment for the first two components representing each respective ancestry. Due to the heterogeneous nature of EOC, subgroup analysis was conducted to estimate genotype-specific ORs for serous carcinomas (the most predominant histologic subtype) and the three other main histological subtypes of
EOC: endometrioid, mucinous and clear cell. Separate analyses were also carried out for each study site, and site-specific ORs were combined using a fixed-effect meta-analysis. The χ² test of heterogeneity was estimated to quantify the proportion of total variation due to heterogeneity across studies, and the heterogeneity of ORs between studies was tested with Cochran’s Q statistic. The R statistical package ‘rmeta’ was used to generate forest plots. Statistical analysis was conducted in PLINK.42

Imputation of genotypes at 17q21.31. To increase genomic coverage, we imputed genotype data for the 17q21.31 region (chr17: 40,099,001–44,900,000, human genome build 37) with IMPUTE2.2 (ref. 55) using phase 1 haplotype data from the 1000 Genome project data.43 For each imputed genotype the expected number of minor alleles carried was estimated (as weights). IMPUTE provides estimated allele dosage for SNPs that were not genotyped and for samples with missing data for directly genotyped SNPs. Imputation accuracy was estimated using an r² quality metric. We excluded imputed SNPs from analysis where the estimated accuracy of imputation was low (r²<0.3).

Functional studies and in silico analysis of publicly available data sets. We performed the following assays for each gene in the 1-MB region centred on the most significant SNP at the 17q21.31 locus (see Supplementary Methods): gene expression analysis in EOC cell lines (n = 51) compared with normal cell lines from OCPT,44 including ovarian surface epithelial cells and fallopian tube secretory epithelial cells (n = 73) and Cgp island methylation analysis in HGS ovarian cancer (HGS-EOCs) tissues (n = 106) and normal tissues (n = 7). Genes in the region were also evaluated in silico by mining publicly available molecular data generated for primary EOCs and other cancer types, including TCGA analysis of 568 HGS EOCs,45 the Gene Expression Omnibus series GSE1852 data set of 53 HGS EOCs19 and the Catalogue Of Somatic Mutations in Cancer database.20 We used these data to (1) compare gene expression between (a) EOC cell lines and normal cell lines and (b) tumour tissue and normal tissue from TCGA, (2) to compare gene methylation status in HGS-EOCs and normal tissue, (3) to conduct functional studies and (b) tumour tissue and normal tissue from TCGA, (2) to conduct

References

The scientific development and funding for this project were in part supported by the US National Cancer Institute (R01-CA114343 and R01-CA114343-51) and the Genetic Associations and Mechanisms in Oncology (GAME-ON); a NCI Cancer Post-GWAS Initiative (U19-CA148112). This study made use of the data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.uk/. Funding for the project was provided by the Wellcome Trust under award 076113. The results published here are in part generated upon data generated established by the National Cancer Institute and National Human Genome Research Institute. Information about TCGA, and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov/. D.F.E. is a Principal Investigator with the TCGA Research Network. Additional information

Supplementary Information accompanies this article at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/


Additional information

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/
Washington 98109, USA. 103 Department of Epidemiology, University of Washington, Seattle, Washington 98109, USA. 104 Institut für Humangenetik Wiesbaden, 65187 Wiesbaden, Germany. 105 Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Melbourne, Victoria 3053, Australia. 106 Department of Gynecologic Oncology, The Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw 02-781, Poland. 107 Pediatrics Epidemiology Center, College of Medicine, University of South Florida, Tampa, Florida 33612, USA. 108 Cancer Research Initiatives Foundation, Sime Darby Medical Centre, 47500 Selangor, Malaysia. 109 Department of Surgery, Faculty of Medicine, University Malaya, 59100 Kuala Lumpur, Federal Territory of Kuala Lumpur, Malaysia. 110 Channing Division of Network Medicine, Harvard Medical School and Brigham and Women’s Hospital, Boston, Massachusetts 02115, USA. 111 Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA. 112 Department of Obstetrics and Gynecology, University of Ulm, Ulm 89081, Germany. 113 Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina 27708, USA. 114 Department of Obstetrics and Gynecology, Mayo Clinic, Rochester, Minnesota 55905, USA. 115 Shanghai Cancer Institute, Shanghai 2200-25, China. 116 Department of Epidemiology, Center for Cancer Genetics Research and Prevention, School of Medicine, University of California Irvine, Irvine, California 92697, USA. 117 Department of Statistical Science, Duke University, Durham, North Carolina 27708, USA. 118 Cancer Prevention, Detection and Control Research Program, Duke Cancer Institute, Durham, North Carolina 27708-0251, USA. 119 Department of Obstetrics and Gynecology, Duke Cancer Institute, Durham, North Carolina 27708, USA. 120 Department of Biostatistics, University of Kansas Medical Center, Kansas City, Kansas 66160, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.P.-W. or T.A.S. or S.G. (email: Jenny.Wey@moffitt.org or Thomas.Sellers@moffitt.org or gayther@usc.edu)