Network-Based Integration of GWAS and Gene Expression Identifies a HOX-Centric Network Associated with Serous Ovarian Cancer Risk


1Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom. 2Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, United Kingdom. 3Department of Medical Oncology, The Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts. 4Department of Preventive Medicine, Keck School of Medicine, University of Southern California Norris Comprehensive Cancer Center, Los Angeles, California. 5Radboud University Medical Centre, Radboud Institute for Health Sciences, Nijmegen, the Netherlands. 6Division of Epidemiology, Director of Genetic Epidemiology Research Institute, School of Medicine, University of California Irvine, Irvine, California. 7Byelorussian Institute for Oncology and Medical Radiology Aleksandrov N.N., Minsk, Belarus. 8Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. 9Peter MacCallum Cancer Institute, East Melbourne, Victoria, Australia. 10Cancer Prevention and Control.
Abstract

Background: Genome-wide association studies (GWAS) have so far reported 12 loci associated with serous epithelial ovarian cancer (EOC) risk. We hypothesized that some of these loci function through nearby transcription factor (TF) genes and that putative target genes of these TFs as identified by coexpression may also be enriched for additional EOC risk associations.

Methods: We selected TF genes within 1 Mb of the top signal at the 12 genome-wide significant risk loci. Mutual information, a form of correlation, was used to build networks of genes strongly coexpressed with each selected TF gene in the unified microarray dataset of 489 serous EOC tumors from The Cancer Genome Atlas. Genes represented in this dataset were subsequently ranked using a gene-level test based on results for germline SNPs from a serous EOC GWAS meta-analysis (2,196 cases/4,396 controls).

Results: Gene set enrichment analysis identified six networks centered on TF genes (HOXB2, HOXB5, HOXB6, HOXB7 at 17q21.32 and HOXD1, HOXD3 at 2q31) that were significantly enriched for genes from the risk-associated end of the ranked list (P < 0.05 and FDR < 0.05). These results were replicated (P < 0.05) using an independent association study (7,035 cases/21,693 controls). Genes underlying enrichment in the six networks were pooled into a combined network.

Conclusion: We identified a HOX-centric network associated with serous EOC risk containing several genes with known or emerging roles in serous EOC development.

Impact: Network analysis integrating large, context-specific datasets has the potential to offer mechanistic insights into cancer susceptibility and prioritize genes for experimental characterization.

Introduction

The genetic architecture of inherited susceptibility to epithelial ovarian cancer (EOC) appears similar to other hormone-related cancers, but fewer EOC risk loci have been discovered compared with breast and prostate cancer, probably due to restricted sample size (1). Genome-wide association studies (GWAS) have so far identified 12 risk loci associated with serous EOC, the most common subtype of EOC (1–7). These account for 4% of the excess familial risk of EOC, while rare, high-penetrance mutations contribute an additional 5% (8). This suggests that many undiscovered common serous EOC risk variants exist.

Very stringent statistical thresholds are generally used to declare common variant susceptibility alleles at so-called genome-wide significance (P < 5 × 10⁻⁸). However, when there is limited statistical power, hundreds or thousands of single-nucleotide polymorphisms (SNP) with small effect sizes will not reach genome-wide significance (9). A key challenge in genetic epidemiology is to identify these risk SNPs with small effects. One approach is ever-larger studies allied to better coverage of common variation across the genome to increase statistical power. However, even a case–control study with 100,000 samples has just 23% power to detect at genome-wide significance an allele of frequency 5% that confers a per-allele relative risk of 1.1.

GWAS pathway analysis has emerged as a complement to imputation, single-variant testing and meta-analysis for the discovery of true genetic associations in the pool of SNPs that are below genome-wide significance (10). Pathway studies are guided by the hypothesis that true risk associations are more likely to cluster in genes that share a common biologic function potentially dysregulated in disease pathogenesis. However, incomplete annotation and canonical representation of pathways in the literature are major limiting factors (11).

One approach to overcome this limitation is by analyzing GWAS signals within the reduced search space of dynamic networks constructed from pairwise interactions observed in large, independent, tissue-specific transcriptomic datasets (12). Furthermore, GWAS of cancer and other diseases increasingly suggest that at least some genome-wide significant risk loci act through nearby transcription factor (TF) genes (13–15). Target genes of these TFs, in turn, have been found to be enriched for SNPs that fail to reach genome-wide significance but are nominally associated with the disease (16, 17). Therefore, we adopted a risk locus TF gene-centric approach to integrating serous EOC transcriptomic and GWAS datasets. Seven of the 12 known genome-wide significant serous EOC risk loci harbor at least one TF gene in the 2-Mb interval centered on the top SNP at the locus. This includes nine members of the HOXD cluster and 10 members of the HOXB cluster at the 2q31 and 17q21.32 loci, respectively. The target genes of most homeobox (HOX) TFs remain largely unknown due to their promiscuous DNA-binding properties in vitro (18). Because genes highly coexpressed with TF genes are more likely to represent their targets (19), and coexpression has been linked to shared function (20), we used the genes highly coexpressed with each TF in The Cancer Genome Atlas (TCGA) high-grade serous EOC microarray dataset to build hub-and-spoke type TF-target gene networks (21). We then systematically interrogated these networks for overrepresentation of genes containing SNPs ranked high for their association with serous EOC risk in a GWAS meta-analysis. Our aims were to prioritize hub TF genes whose networks demonstrated such overrepresentation as candidates for post-GWAS functional characterization and to use these networks to identify novel pathways and potential sub-genome-wide significant risk loci involved in serous EOC development. Most GWAS SNPs lie outside protein-coding regions of the genome and may affect cancer susceptibility by regulating a gene or genes up to a mega-base away making such integrative genomic approaches to prioritizing genes in these 1-Mb regions imperative (22, 23).

Materials and Methods

Ranking genes based on GWAS results for serous EOC risk

This study used P values for association with serous EOC risk from a published meta-analysis of a North American and UK GWAS for 2,508,744 SNPs that were either genotyped or imputed in a total of 2,196 serous EOC cases and 4,396 controls (1). The meta-analysis was restricted to subjects of European descent and the HapMap II (release 22) CEU panel served as reference for imputation. The North American and UK GWAS together with the replication dataset described below are summarized in Supplementary Table S1. All participants provided written informed
consent and each contributing study was approved by the appro-riate local institutional ethical review board. A complete list of genes annotated with start and end positions of each gene was downloaded via the Bioconductor package TxDb.Hsapiens.UCSC. hg19.knownGene (v2.8.0). After removing genes with ambiguous location, all SNPs were mapped to genes with boundaries defined by the start and end positions. The genes were ranked in descending order of the negative logarithm (base 10) of the minimum P value among all SNPs in each gene after adjusting this P value for the number of SNPs in the gene using a modification of the Sidak correction (24). This accounts for linkage disequilibrium (LD) between SNPs while reducing the effect of gene size on the minimum P value (25). Pearson r between gene size and minimum P value improved from −0.3 to +0.1. Adequacy of correction was further confirmed by quantile–quantile plots (Supplementary Fig. S1). In all, 10,693 genes that were also represented in the TCGA gene expression dataset (described below) were ranked on the basis of the GWAS results for subsequent analysis.

Constructing serous EOC-specific coexpression networks with TF genes at risk loci as hubs

We selected genes with experimentally confirmed TF activity as described by Vazquez and colleagues (26) that were less than 1 Mb from the most significant SNP at each of the 12 loci (1–7) known to be associated with serous EOC risk at genome-wide significance (Table 1). The SNPs listed in Table 1 were obtained from the Collaborative Oncological Gene-environment Study (COGS) Primer (27). Nineteen of the 29 TF genes selected belonged to either the HOXB or the HOXD cluster. We relied on coexpression in serous EOC tissue to define possible context-specific target genes of the Tfs.

Gene set enrichment analysis

Each TF gene and the genes coexpressed with it were treated as a single hub-and-spoke network. Gene set enrichment analysis (GSEA; ref. 30) was used to determine overrepresentation of genes from the serous EOC-associated end of the ranked list generated on the basis of the GWAS meta-analysis, in each network. Before evaluating each network by GSEA, its parent hub TF gene was excluded from the network. Ten thousand permutations were performed with the number increased to 50,000 for networks with P = 0 at 10,000. Genes in the ranked list that were not among the 11,864 genes in the TCGA dataset were excluded before GSEA as their coexpression with the Tfs could not be evaluated. GSEA is optimized for gene sets/networks containing 15 to 500 genes (30). Applying this cutoff led to the exclusion of 11 of the 25 networks from the primary GSEA. We also conducted a secondary GSEA run allowing for networks with more than 10 genes. This enabled evaluation of 24 out of the 25 coexpression networks. Only the network centered on HOXD9 was excluded in the second GSEA run because it contained just six genes. Networks with GSEA P < 0.05 and FDR < 0.05 were considered significant. We also reported fold enrichment as the ratio of the number of genes from the top 5% of the ranked list that were observed in each network to the number expected. Cytoscape (v2.8.3) was used for network visualization (31).

Replication analyses

First, we examined the effect of LD between SNPs in closely spaced genes. The frequent coexpression of genes colocalized on the genome coupled with LD between SNPs in such genes has the potential to inflate the GSEA signal (25). We subjected all SNPs in genes that were in the ranked list input to GSEA to LD-based clumping using PLINK (v1.07; ref. 32) and CEU LD information from HapMap II (release 22). Starting from the most significant SNP, all SNPs with r² > 0.7 within 250 kb of it were removed and the step repeated for the next best available SNP. Using the LD-thinned SNP list, SNP-to-gene mapping, gene ranking and GSEA were repeated as described above.

Second, to replicate significant findings from our primary analysis, we used an independent dataset from the COGS. The

---

Table 1. Summary of 12 genome-wide significant serous EOC risk loci

<table>
<thead>
<tr>
<th>Locus</th>
<th>SNP</th>
<th>Chr</th>
<th>Position</th>
<th>OR (95% CI)</th>
<th>TF genes &lt; 1 Mb away from SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2q31</td>
<td>rs711830</td>
<td>2</td>
<td>177,037,311</td>
<td>1.12 (0.88–1.45)</td>
<td>HOXD3, HOXD6, HOXD9, HOXD10, HOXD13b</td>
</tr>
<tr>
<td>3q25</td>
<td>rs7083979</td>
<td>3</td>
<td>163,908,723</td>
<td>1.10 (0.93–1.31)</td>
<td>IRX4b</td>
</tr>
<tr>
<td>4q32</td>
<td>rs696105</td>
<td>4</td>
<td>1,279,790</td>
<td>1.09 (1.05–1.13)</td>
<td>IRX4b</td>
</tr>
<tr>
<td>5p5</td>
<td>rs10066990</td>
<td>5</td>
<td>1,279,790</td>
<td>1.09 (1.05–1.13)</td>
<td>IRX4b</td>
</tr>
<tr>
<td>8q21</td>
<td>rs17782652</td>
<td>8</td>
<td>82,653,644</td>
<td>1.19 (1.12–1.26)</td>
<td>MYC</td>
</tr>
<tr>
<td>8q24</td>
<td>rs7184937</td>
<td>8</td>
<td>129,541,475</td>
<td>1.11 (1.03–1.21)</td>
<td>HOXB6, HOXB2, HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8, HOXB9, HOXB10, HOXB11, HOXB13, HOXD2</td>
</tr>
<tr>
<td>9p22</td>
<td>rs1084113</td>
<td>9</td>
<td>16,915,021</td>
<td>1.11 (1.03–1.21)</td>
<td>HOXB6, HOXB2, HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8, HOXB9, HOXB10, HOXB11, HOXB13, HOXD2</td>
</tr>
<tr>
<td>10p12</td>
<td>rs7084454</td>
<td>10</td>
<td>21,821,274</td>
<td>1.10 (0.97–1.24)</td>
<td>HOXB6, HOXB2, HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8, HOXB9, HOXB10, HOXB11, HOXB13, HOXD2</td>
</tr>
<tr>
<td>17q21</td>
<td>rs7405776</td>
<td>17</td>
<td>36,093,022</td>
<td>1.12 (1.08–1.17)</td>
<td>HOXB1, HOXB2, HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8, HOXB9, HOXB10, HOXB11, HOXB13, HOXD2</td>
</tr>
<tr>
<td>17q21.31</td>
<td>rs2077606</td>
<td>17</td>
<td>43,529,293</td>
<td>1.15 (1.12–1.19)</td>
<td>HOXB6, HOXB2, HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8, HOXB9, HOXB10, HOXB11, HOXB13, HOXD2</td>
</tr>
<tr>
<td>17q21.32</td>
<td>rs728345</td>
<td>17</td>
<td>46,502,917</td>
<td>1.12 (1.08–1.16)</td>
<td>HOXB6, HOXB2, HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8, HOXB9, HOXB10, HOXB11, HOXB13, HOXD2</td>
</tr>
<tr>
<td>19p13</td>
<td>rs711830</td>
<td>19</td>
<td>17,389,704</td>
<td>1.14 (0.99–1.17)</td>
<td>KLF2, NR2F6</td>
</tr>
</tbody>
</table>

Abbreviations: Chr, chromosome; CI, confidence interval.

*TF gene was not represented in the TCGA dataset.

†Associated with serous EOC risk in BRCA1 mutation carriers only (7).

Correction for 11,864 pairwise tests was set in ARACNE to retain only the genes most strongly coexpressed with each TF gene.
COGS dataset included 7,035 serous EOC cases and 21,693 controls of European descent after exclusion of overlap with the samples used in the primary analysis (Supplementary Table S1). These were genotyped using the iCOGS custom array (1) and imputed into the 1000 Genomes March 2012 EUR reference panel (33). The imputed dataset was filtered to retain 7,768,381 SNPs with minor allele frequency > 0.03 for consistency with the GWAS meta-analysis. SNP-to-gene mapping (without LD-based clumping), gene ranking and GSEA were repeated as described above.

Analysis of combined network

The 'leading-edge' subsets of genes from all coexpression networks that were significant in the primary analysis were combined into a single network. The leading-edge subset is the core of a gene set or network that drives its enrichment signal. It is defined as those members of the network that are ranked higher than the position in the ranked gene list where the network enrichment score is maximum (30). Subsequent analyses were conducted on genes in the combined network.

First, we input the genes to the Disease Association Protein–Protein Link Evaluator (DAPPLE, v1.0; ref. 34). DAPPLE uses a database of 169,810 well-established pairwise interactions between 12,793 proteins to connect input proteins (genes) directly or indirectly via a single protein not in the input. By biological meaningful networks are more connected than random ones and DAPPLE tests the significance of indirect connectivity in the resultant network using permutation.

Second, we measured coexpression between each TF gene that was the hub of a significant network in the primary analysis and 17,255 other genes with expression levels profiled in the 245 sample Australian Ovarian Cancer Study (AOCS) serous EOC microarray dataset (GSE9899; refs. 35, 36) using MI calculated by ARACNE. All 17,255 genes were ranked on the basis of the strength of their coexpression with each hub TF. For each hub TF, we then counted the number of genes coexpressed with it in the combined network that were also observed among the top (strongest) 1% and top 10% of coexpression interactions for the TF in the AOCS dataset as arbitrary estimates of tissue-specific replication of these coexpression-derived interactions.

Third, we used PSCAN (v1.2.2; ref. 37) to test overrepresentation of known TF-binding motifs from TRANSFAC (Human; ref. 38) up to 1 kb upstream of transcription start sites of genes in the combined network. HOXB and HOXD motifs were absent in PSCAN but some homeodomain motifs were represented.

Results

The network-based analytic strategy we used to integrate serous EOC GWAS and gene expression datasets is outlined in Fig. 1. A total of 29 TF genes were located within 1 Mb of the top risk-associated SNP at seven of the 12 known genome-wide significant serous EOC risk loci (Table 1). Of these, four TFs were not represented in the TCGA dataset used in this study and could not be evaluated further (Table 1). Computing pairwise mutual information between the somatic expression levels of each remaining TF gene (fixed as a hub) and the 11,863 other genes in the TCGA dataset using an adaptive partitioning procedure with a Bonferroni-adjusted threshold of \( P < 0.01 \) to retain only the most strongly coexpressed genes yielded 25 hub-and-spoke type networks ranging in size from 6 to 1,953 genes. Fourteen TF genes from five risk loci were hubs of coexpression networks that included between 17 and 368 genes placing them in the gene set size range (15–500 genes) optimal for GSEA (Tables 1 and 2).

A total of 689,882 of the approximately 2.5 million SNPs from the serous EOC GWAS meta-analysis were located within the boundaries of 10,693 of the 11,864 genes represented in the TCGA data. These genes were ranked in descending order using the negative logarithm (base 10) of the modified Sidak-corrected minimum \( P \) value among all SNPs in each gene.

Six of the 14 networks tested by GSEA (with hubs HOXB2, HOXB7, HOXB6, and HOXB5 at 17q21.32, and hubs HOXD3 and HOXD1 at 2q31) were significantly enriched for genes that ranked high for their association with serous EOC risk (GSEA \( P < 0.05 \) and FDR < 0.05; Table 2). The significant networks demonstrated between 1.5- and 8-fold enrichment for genes from the top 5% of the ranked list as compared with expectation (Table 2). Supplementary Table S2 lists the genes in each significant network.

The six networks identified by the primary analysis remained significant in replication analysis using the LD-thinned SNP list, indicating that the primary GSEA signals were not being driven by strongly correlated SNPs in closely spaced genes (Supplementary Table S3). All six networks were also significantly associated (\( P = 0.03 \) to 4 \( \times 10^{-4} \)) with serous EOC risk in replication analysis using the independent COGS dataset (Supplementary Table S4). Next, we observed that 33 genes were shared by more than one of the six networks. Furthermore, 18 genes across the six networks were located less than 1 Mb from the most significantly associated SNPs at the 2q31 and 17q21.32 risk loci. To eliminate potential inflation of the GSEA signal due to genes coexpressed locally with the hub TF genes anchoring the significant networks, we pooled the 249 genes in the six original networks and removed the 18 genes at 2q31 and 17q21.32. We also counted the overlapping genes only once in the pooled network. The result was a set of 174 genes of which 170 were covered by SNPs in the COGS dataset. This filtered set was also significantly associated with serous EOC risk in GSEA using the ranked gene list derived from the COGS data (\( P = 0.007 \)). A secondary GSEA run with less stringent parameters that allowed inclusion of the 24 coexpression networks with more than 10 genes (including the 10 previously excluded networks containing > 500 genes) did not identify any additional significant networks at the \( P < 0.05 \) and FDR < 0.05 threshold (Supplementary Table S5). Supplementary Table S6 shows differences in the number of genes from the TCGA dataset covered by intragenic SNPs in the ranked lists generated from the GWAS meta-analysis, LD-thinned, and COGS datasets.

Given that all networks significant in the primary analysis were centered on hub TFs from the HOX gene family and the observation that some of their targets were shared, we merged the leading edge subsets (i.e., the genes underlying the GSEA signal) from each of the six networks. This yielded a combined network of 50 nonredundant genes most strongly associated with serous EOC risk and 81 interactions between them (Fig. 2). The 50 genes were submitted as input to the web-based tool, DAPPLE, to determine whether they demonstrate significant connectivity at the protein–protein interaction (PPI) level, a frequent characteristic of disease-associated networks (39). DAPPLE connected the proteins encoded by the genes using a database of high-confidence PPIs, allowing for connections between input genes/proteins that were either direct or indirect via a single intermediate protein not in the input group. Two PPI networks were formed (Supplementary Fig. S2), with the larger network involving 29 of

Kar et al.
the 50 input genes/proteins. Compared with 10,000 permuted networks with similar underlying topology, the input proteins were more likely to be connected to each other indirectly (through a single intermediate protein) than expected by chance alone ($P = 0.01$). Next, we evaluated whether the genes coexpressed with the hub TF genes in the combined network were more likely to be among the genes most strongly coexpressed with the same TFs in an independent microarray dataset from the AOCS. All interactions in the combined network between a hub TF gene and another HOX cluster gene and 25 of 44 (57%) hub TF and non-HOX cluster gene interactions were among the top 10% of interactions for the corresponding hub TF in the AOCS dataset (for HOX–non-HOX interactions, binomial test $P = 2 \times 10^{-14}$). In fact, 32 of 37 (86%) HOX–HOX and 13 of 44 (30%) HOX–non-HOX interactions were seen at a more stringent top 1% level (for HOX–non-HOX interactions, binomial test $P = 4 \times 10^{-16}$). HOX–HOX interactions were analyzed separately because HOX cluster members are already known to interact with each other (40). Finally, we assessed enrichment of known TF-binding motifs up to 1 kb upstream of the 40 non-HOX cluster genes in the combined network using the online tool PSCAN to look for possible overrepresentation of the few homeodomain motifs that are well established. The binding motif for the cooperative homeodomain TFs, HOX-A9-MEIS1, was the second most overrepresented among the 40 non-HOX cluster genes in the combined network ($P = 0.008$), second only to the motif for SPZ1 ($P = 0.006$). Thus, the AOCS dataset and the PSCAN tool further supported the TF-target gene interactions observed in the combined network.

**Discussion**

In this study, we used network analysis to integrate TCGA gene expression data with GWAS meta-analysis summary findings to identify six networks significantly associated with serous EOC risk and replicated our results using the independent COGS dataset. Network construction was guided by the premise that TF genes...
Table 2. GSEA results for 14 serous EOC risk locus TF-centric gene coexpression networks

<table>
<thead>
<tr>
<th>Serous EOC risk locus</th>
<th>Network hub TF gene</th>
<th>Coexpression network size</th>
<th>Normalized enrichment</th>
<th>Nominal P</th>
<th>FDR q value</th>
<th>Top 5% fold enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1q21.32</td>
<td>HOXD2</td>
<td>17</td>
<td>1.64</td>
<td>2.0E-05b</td>
<td>4.0E-05b</td>
<td>8.00</td>
</tr>
<tr>
<td>17q21.32</td>
<td>HOXB7</td>
<td>31</td>
<td>1.62</td>
<td>0b</td>
<td>5.0E-05b</td>
<td>4.50</td>
</tr>
<tr>
<td>17q21.32</td>
<td>HOXB6</td>
<td>33</td>
<td>1.43</td>
<td>5.0E-04</td>
<td>3.0E-03</td>
<td>4.50</td>
</tr>
<tr>
<td>2q31</td>
<td>HOXD5</td>
<td>52</td>
<td>1.39</td>
<td>5.0E-04</td>
<td>4.9E-03</td>
<td>3.67</td>
</tr>
<tr>
<td>17q21.32</td>
<td>HOXB5</td>
<td>30</td>
<td>1.37</td>
<td>4.0E-03</td>
<td>5.1E-03</td>
<td>3.00</td>
</tr>
<tr>
<td>2q31</td>
<td>HOXD1</td>
<td>69</td>
<td>1.25</td>
<td>8.1E-03</td>
<td>0.04</td>
<td>1.50</td>
</tr>
<tr>
<td>17q21.32</td>
<td>SP2</td>
<td>134</td>
<td>1.18</td>
<td>7.7E-03</td>
<td>0.09</td>
<td>1.57</td>
</tr>
<tr>
<td>17q21.32</td>
<td>HOXB3</td>
<td>103</td>
<td>1.18</td>
<td>0.02</td>
<td>0.08</td>
<td>2.17</td>
</tr>
<tr>
<td>17q21.32</td>
<td>TBX2</td>
<td>46</td>
<td>1.10</td>
<td>0.19</td>
<td>0.22</td>
<td>1.33</td>
</tr>
<tr>
<td>17q21.32</td>
<td>NFX2L1</td>
<td>315</td>
<td>1.06</td>
<td>0.12</td>
<td>0.35</td>
<td>1.06</td>
</tr>
<tr>
<td>19p13</td>
<td>NR2F6</td>
<td>312</td>
<td>1.04</td>
<td>0.23</td>
<td>0.43</td>
<td>1.50</td>
</tr>
<tr>
<td>8q24</td>
<td>MYC</td>
<td>130</td>
<td>1.02</td>
<td>0.37</td>
<td>0.45</td>
<td>0.71</td>
</tr>
<tr>
<td>3q25</td>
<td>SHOX2</td>
<td>368</td>
<td>1.00</td>
<td>0.53</td>
<td>0.56</td>
<td>0.95</td>
</tr>
<tr>
<td>2q31</td>
<td>HOXD20</td>
<td>85</td>
<td>1.00</td>
<td>0.52</td>
<td>0.52</td>
<td>0.80</td>
</tr>
</tbody>
</table>

*Hub-and-spoke networks containing 15 to 500 genes constructed from serous EOC TCGA data.

A striking feature of the combined network is that it connects genes at five of the 12 genome-wide significant serous EOC risk loci. This is particularly compelling because only TF genes from two of these loci (2q31 and 17q21.32) were used as seed hubs for the six significant networks that underlie the combined network. The other three were picked up by the combination of coexpression and enrichment analysis, which is, without any a priori input. A small number of genes contribute to the combined network from the three non-hub loci in this group of five: BNC2 (HOXD1 network member) at 9p22; HNF1B (HOXB6 and HOXB7 networks) and LHX1 (HOXD1 network) at 17q12; and ABHD8 (HOXB7 network) at 19p13. Serous EOC risk SNPs at 17q12 have previously been associated with significantly higher HNF1B methylation (47). Likewise, the SNPs at the 9p22 risk locus are likely to represent multiple transcriptional regulatory elements acting on BNC2 based on an integrated functional analysis (Buckley and colleagues; unpublished data). Finally, functional work on 19p13 indicates that ABHD8 is the most likely target of serous EOC and breast cancer risk SNPs at this locus (Lawrenson and colleagues; unpublished data; refs. 4, 48). Taken together, these findings for ABHD8, HNF1B, and BNC2 underscore the power of integrating GWAS with tissue-specific gene expression data in a network paradigm to prioritize the genes likely to be regulated by genome-wide significant risk variants for downstream functional characterization (23).

The addition of orthogonal biologic prior to statistical evidence from GWAS may also unravel interesting targets at loci that are nominally significant but fail to reach genome-wide significance. For example, the oncogene WT1 that appears in the combined network is a highly specific serous ovarian lineage marker (49). The top SNP in this gene has P = 0.012 in the GWAS meta-analysis that strengthens to P = 0.0078 in the COGS dataset. However, while it is intriguing to identify such possible functional targets associated with serous EOC susceptibility, it will require extensive experimental validation to confirm their role in disease predisposition. Profiling of the relevant HOXB and HOXD TF binding in fallopian tube and ovarian surface epithelial cells and in ovarian cancer cells may identify binding sites containing serous EOC risk SNPs analogous to what has been done to establish the role of other TFs in breast cancer susceptibility (50, 51). The emergence of genome editing technologies such as the CRISPR-Cas9 system may further enable modification of risk near genome-wide significant serous EOC risk SNPs may be the functional targets of these SNPs. The TFs may regulate pathways of target genes, represented by genes strongly coexpressed with each TF gene that are, in turn, enriched for SNPs nominally associated with serous EOC susceptibility. The product of the multistep analysis presented here is a combined network of 50 genes and 81 interactions (Fig. 2). A significant proportion of these interactions were further supported by coexpression analysis in the independent AOC5 microarray dataset.

Hub TF genes of the six significant networks and the genes in their combined network suggest that specific members of the HOXD and HOXB clusters potentially mediate the effects of the 2q31 and 17q21.32 serous EOC risk loci, respectively. Although this analysis was ongoing, a parallel comprehensive functional follow-up of the 2q31 locus revealed that SNPs at this locus affect serous EOC development through HOXD9 (Lawrenson and colleagues, submitted for publication). HOXD9 could not be evaluated as a hub TF using GSEA because its network contained just six genes (including HOXD1 and HOXD3) at the stringent cutoff we used for identifying only the most strongly coexpressed genes. However, it is worth noting that HOXD9 is a leading edge or core member of two of the significant networks identified (centered on HOXD1 and HOXD3), and therefore appears in the combined network uncovered by this analysis (Fig. 2). Thus, HOXD9 may have a putative master regulatory role upstream of HOXD1 and HOXD3 in serous EOC development. At 17q21.32, HOXB5, HOXB6, and HOXB7 have collectively been implicated in early carcinogenic reprogramming of transcription in epithelial cells (41). HOXB7 is also known to be involved in DNA repair catalyzed by poly(ADP-ribose) polymerase (42). HOXB7 expression is elevated in ovarian cancer and overexpression is associated with marked proliferation of immortalized normal ovarian surface epithelial cells by upregulation of basic fibroblast growth factor (43). Hypermethylation of HOXB5, a possible tumor suppressor, is also a frequent somatic change reported in ovarian cancer and corresponding in vitro models (44). Overall, the identification of a novel HOX-centric pathway is consistent with the emerging role of developmental genes in ovarian carcinogenesis (45). Other members of the HOX gene family have previously been shown to guide both reproductive tract patterning during early development and EOC cellular morphology (46).
Conception and design: applying the modified Sidak correction. SNPs in these HOX-binding sites to modulate and confirm the role of the HOX target genes in cellular models that reflect early stages of ovarian cancer (52).

The observation that 29 of the 50 genes in the combined network were also part of a PPI network with significant indirect connectivity suggests that this gene network is biologically coherent. The failure to capture the remaining 21 coexpressed genes in the PPI network also hints that some interactions are perhaps tissue-specific and/or specific to the transcriptome because DAPPLE, the tool used for the PPI network analysis, relies on protein interactions found in different tissues (34).

There are limitations inherent in this study. GSEA is optimized to detect enrichment in gene sets or networks containing between 15 and 500 genes. Although we applied it to the 10 networks containing more than 500 genes in a secondary analysis, our failure to detect any additional significant networks among the very large networks does not rule out the presence of genuine enrichment in these networks. Coexpression is likely not the perfect proxy for defining TF targets and the risk locus–TF gene and TF–target gene interactions suggested by this analysis warrant follow-up.

In conclusion, by identifying a HOX-centric gene coexpression network associated with serous EOC risk, this report highlights the potential of network analysis to combine GWAS with other molecular data to offer insights into the mechanisms linking population studies with cancer biology.

**Disclosure of Potential Conflicts of Interest**

P.A. Fasching reports receiving commercial research support from Amgen and Novartis and has received speakers bureau honoraria from Amgen, Novartis, Pfizer, Celgene, Roche, and GlaxoSmithKline. U. Menon reports receiving a commercial research grant from Abcodia and has ownership interest (including patents) in Abcodia. No potential conflicts of interest were disclosed by the other authors.

**Authors’ Contributions**


www.aacrjournals.org Cancer Epidemiol Biomarkers Prev; 24(10) October 2015 1581

Figure 2. Cytoscape visualization of the combined network. Hub TF genes of the six significant networks underlying the combined network are outlined in green. Candidate genes at genome-wide significant risk loci other than the input hub loci (2q31 and 17q21.32) are outlined in blue. Nodes are colored according to the P-value of the most significant intragenic SNP in the GWAS meta-analysis (uncorrected for the number of intragenic SNPs). Diamond-shaped genes have P < 0.05 after applying the modified Sidak correction.

The following authors would like to acknowledge specific grant support from the NIH: J.A. Doherty (R01-CA168758), R.P. Edwards (R31-CA156368). K. Odnuni (PS0-CA159981), C.L. Pearse (R03-CA115195), J. Permutt-Wey (R01-CA154343), C.M. Phehan (R01-CA149429), E.M. Plooe (T22-CA209001), S.J. Rams (R01-CA172404), A. Risch (R01-CA074850), J.H. Rothenstein (P30-CA142445), W. Sieh (K07-CA143047), L.E. Sucheston-Campion (P50-CA159981), Rachel Palmieri Weber (R01-CA076616), and from Cancer Research UK: A. Gentry-Maharaj (C315/A2621) and P.D.P. Pharoah (C490/A10119). Funding of the constituent studies of the OCAC-GWAS and COGS are listed by funding agency with each grant number in parentheses: American Cancer Society (CRTG-06-196-01-CCE), the California Cancer Research Program [00-01389-20101, 001525403, 002000]; the Canadian Institutes for Health Research (MOP-66727); Cancer Council Queensland; Cancer Council South New South Wales; Cancer Council South Australia; Cancer Council Tasmania; Cancer Foundation of Canada; Cancer Society of New Zealand; Ovarian Cancer Research Consortium (OCRC); the Ovarian Cancer Research Foundation; the Ovarian Cancer Research Fund; the U.S. National Cancer Institute (1K99CA184415-01); the U.S. National Institutes of Health (R01-CA087538, R01-CA080978, R01-CA106414, R01-CA122443, R01-CA124523, R01-CA134433, R01-CA126841, R01-CA136924, R01-CA149429, R03-CA113148, R05-CA115195, R37-CA070867, R37-CA070867, UI01-CA069417, UI01-CA071966, and Intramural research funds); the U.S. Army Medical Research and Material Command (DAMD17-98-1-8659, DAMD17-99-C-9401); the U.K. National Institute for Health Research Biomedical Research Centres at the University of Cambridge, Imperial College London, University College Hospital “Women’s Health Theme” and the Royal Marsden Hospital, WorkSafeBC.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received November 24, 2014; revised May 26, 2015; accepted June 29, 2015; published OnlineFirst July 24, 2015.


Network-Based Integration of GWAS and Gene Expression Identifies a HOX-Centric Network Associated with Serous Ovarian Cancer Risk

Siddhartha P. Kar, Jonathan P. Tyrer, Qiyuan Li, et al.

Cancer Epidemiol Biomarkers Prev 2015;24:1574-1584. Published OnlineFirst July 24, 2015.

Updated version
Access the most recent version of this article at:
doi:10.1158/1055-9965.EPI-14-1270

Supplementary Material
Access the most recent supplemental material at:
http://cebp.aacrjournals.org/content/suppl/2015/07/29/1055-9965.EPI-14-1270.DC1

Cited articles
This article cites 49 articles, 14 of which you can access for free at:
http://cebp.aacrjournals.org/content/24/10/1574.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.