# Genomic Analysis Using Regularized Regression in High-Grade Serous Ovarian Cancer

Cancer Informatics
Volume 17: 1–4

DOI: 10.1177/1176935118755341

Yanina Natanzon[1], Madalene Earp[1], Julie M Cunningham[2], Kimberly R Kalli[3], Chen Wang[1], Sebastian M Armasu[1], Melissa C Larson[1], David DL Bowtell[4,5], Dale W Garsed[5,6], Brooke L Fridley[7], Stacey J Winham[1] and Ellen L Goode[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. [2]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA. [3]Department of Medicine, Mayo Clinic, Rochester, MN, USA. [4]Garvan Institute of Medical Research, The Kinghorn Cancer Centre, Darlinghurst, NSW, Australia. [5]Peter MacCallum Cancer Centre, Melbourne, Australia. [6]The Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, VIC, Australia. [7]Biostatistics and Informatics Shared Resource, University of Kansas Medical Center, Kansas City, KS, USA.

**ABSTRACT:** High-grade serous ovarian cancer (HGSOC) is a complex disease in which initiation and progression have been associated with copy number alterations, epigenetic processes, and, to a lesser extent, germline variation. We hypothesized that, when summarized at the gene level, tumor methylation and germline genetic variation, alone or in combination, influence tumor gene expression in HGSOC. We used Elastic Net (ENET) penalized regression method to evaluate these associations and adjust for somatic copy number in 3 independent data sets comprising tumors from more than 470 patients. Penalized regression models of germline variation, with or without methylation, did not reveal a role in HGSOC gene expression. However, we observed significant association between regional methylation and expression of 5 genes (*WDPCP, KRT6C, BRCA2, EFCAB13*, and *ZNF283*). CpGs retained in ENET model for BRCA2 and ZNF283 appeared enriched in several regulatory elements, suggesting that regularized regression may provide a novel utility for integrative genomic analysis.

**KEYWORDS:** Elastic Net penalized regression, high-grade serous ovarian cancer, tumor DNA methylation

**RECEIVED:** August 28, 2017. **ACCEPTED:** January 4, 2017.

**TYPE:** Short Report

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is supported by the National Institutes of Health grant, R25 CA92049 (Mayo Cancer Genetic Epidemiology Training Program).

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Ellen L Goode, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. Email egoode@mayo.edu

## Introduction

Epithelial ovarian cancer remains a disease with high mortality[1] due in part to late stage at diagnosis and a high frequency of resistance to chemotherapeutic agents.[2–4] In high-grade serous ovarian cancer (HGSOC), the most common histotype (~70% of cases), genetic variation, aberrant gene expression, and changes in methylation in certain genes have been implicated in etiology.[5–13] Integrating multiple layers of genomic information offers a potential means by which to clarify the genomic architecture of HGSOC. Agnostic, genome-wide, gene-based *omics* integration methods foster hypothesis generation unachievable with single data type candidate gene approaches. In 2015, Pineda et al[14,15] described an innovative agnostic approach to combine genetic variation, gene expression, copy number variation, and methylation using a flexible penalized regression method which allows for simultaneous agnostic dimension reduction and effect estimation. This methodology can help elucidate the genomic complexities characteristic of HGSOC by selecting only genomic features predictive of gene expression. To uncover genetic and epigenomic regulation of expression in HGSOC with the hope of improving HGSOC risk prediction models and identifying potential therapeutic targets, we applied this form of integrated analysis to 3 independent data sets totaling more than 470 patients.

## Materials and Methods

Eligible patients consisted of women with a primary diagnosis of HGSOC in 3 previously described studies: The Cancer Genome Atlas (TCGA) project (N = 339),[16] the Australian Ovarian Cancer Study (AOCS, N = 78),[17] and the Mayo Clinic Ovarian Cancer Case-Control Study (N = 54).[13] The Cancer Genome Atlas cases from the Mayo Clinic were analyzed only as part of the Mayo Clinic set. Fresh frozen primary tumors were used to derive gene expression, DNA methylation, and copy number variation data, and blood was used as a source of DNA for germline genotype. Data for each type, as described in Supplemental Table 1, were processed separately for each study.[16–20]

From among 57 773 genes (Ensembl GTF 75, human genome build 19 [hg19], GRCh37), we restricted analysis to 22 275 protein-coding genes. Quality control steps excluded genes with low gene expression, low gene expression variance, and extremely high expression values, resulting in 9727 genes available in all 3 data sets (Supplemental Figure 2). For each gene, we defined regional gene CpG and single-nucleotide polymorphism (SNP) sites as those residing within 500 kb upstream and downstream of the annotated start and stop positions, respectively. We analyzed data sets sequentially by sample size starting with TCGA, followed by AOCS and then

**Table 1.** Penalized regression (ENET) model[a] of gene-based DNA methylation association with gene expression in TCGA, AOCS, and Mayo Clinic data sets of high-grade serous ovarian cancer.

| | | | | ENET METHYLATION MODEL *P* VALUE | | |
|---|---|---|---|---|---|---|
| GENE NAME | ENSEMBL ID | CHROMOSOME | TRANSCRIPT LENGTH, (kb) | TCGA (N=339)[b] | AOCS (N=78)[c] | MAYO CLINIC (N=54)[d] |
| *WDPCP* | ENSG00000143951 | 2 | 706.5 | .047 | .079 | .008 |
| *KRT6C* | ENSG00000170465 | 12 | 5.3 | .005 | .026 | .066 |
| *BRCA2* | ENSG00000139618 | 13 | 80.8 | .047 | .099 | .052 |
| *EFCAB13* | ENSG00000178852 | 17 | 118 | .010 | .078 | .007 |
| *ZNF283* | ENSG00000167637 | 19 | 24.7 | .007 | .075 | <.001 |
| *DMRTA1* | ENSG00000176399 | 9 | 8.9 | .047 | .014 | >.10 |
| *HCAR3* | ENSG00000255398 | 12 | 2.1 | .043 | .081 | >.10 |
| *GSDMA* | ENSG00000167914 | 17 | 24.5 | .020 | .088 | >.10 |
| *SLC7A5P1* | ENSG00000260727 | 16 | 0.5 | .047 | >.10 | — |
| *ABCC11* | ENSG00000121270 | 16 | 80.7 | <.001 | >.10 | — |
| *ABCA5* | ENSG00000154265 | 17 | 82.9 | .018 | >.10 | — |

Abbreviations: AOCS, Australian Ovarian Cancer Study; ENET, Elastic Net; TCGA, The Cancer Genome Atlas.
[a]Adjusted for gene copy number.
[b]*P* values adjusted for multiple testing. Only results with <.05 *P* value tested in AOCS data set.
[c]*P* values not adjusted for multiple testing. Only results with <.10 *P* value tested in Mayo Clinic data set.
[d]*P* values not adjusted for multiple testing.
— indicates not tested.

Mayo Clinic, only evaluating significant genes/models in subsequent data sets. In TCGA, we analyzed 3 models using ENET penalized regression methods implemented in the R package "glmnet" (v2.0-4)[21]: methylation-only, germline genotype only, and methylation combined with germline genotype (Supplemental Figure 1). Gene expression was the outcome variable in all 3 models; all analyses adjusted for copy number, which was estimated at an Ensembl coordinate midpoint of each gene. Detailed description of ENET parameters, cross-validation, and derivation of unadjusted and adjusted *P* values is shown in Supplemental Materials.

## Results

In TCGA analyses, results from the methylation-only model were significant after multiple test correction, and germline variation (alone or in combination with DNA methylation) was not associated with tumor gene expression. In particular, methylation at 11 genes in TCGA data was associated with gene expression at *P* value of <.05 after multiple testing correction (Table 1). In the AOCS data set, methylation at 8 of these 11 genes was associated with expression (uncorrected *P* value of <.1; Table 1), and, in the Mayo Clinic set, methylation at a subset of 5 genes was associated with expression (*WDPCP, KRT6C, BRCA2, EFCAB13, ZNF283*; uncorrected *P* value of <.1; Table 1; Supplemental Figure). The specific CpGs retained in the methylation model differed between AOCS and Mayo Clinic data sets, likely due to robust correlation of CpGs within each gene.

For the 5 genes showing association between regional methylation and gene expression in all 3 data sets, we examined regional regulatory features of retained CpGs in the AOCS and Mayo Clinic data sets, both of which used the Illumina 450k methylation array. Methylation enrichment analysis comparing distribution of CpGs retained to CpGs not retained in each model was assessed using a Fisher exact test for 4 genomic regulatory features: predicted enhancer elements and experimentally determined DNase I hypersensitivity sites both determined by the ENCODE project, UCSC-defined CpG island features, and UCSC-defined gene regions features. We found no striking patterns observed regarding the 4 genomic regulatory features. However, CpGs retained in the *ZNF283* methylation model were more likely to be located in the north shelf of CpG islands than unretained CpGs (11% vs 4%) in the Mayo Clinic data set (uncorrected *P*=.01, Supplemental Table 2). Also of note, CpGs retained in the *BRCA2* methylation model were more likely to be located in the body of the gene than unretained CpGs (49% vs 30%, uncorrected *P*=.04) in the AOCS data set (Figure 1, Supplemental Table 2); this was supported in the Mayo Clinic data set (40% gene body vs 35% other gene locations), although not statistically significant.

## Discussion

As variation in gene expression is affected by a complex network of correlated genetic and methylation variants, the ENET methodological approach to high-dimensional data expands our

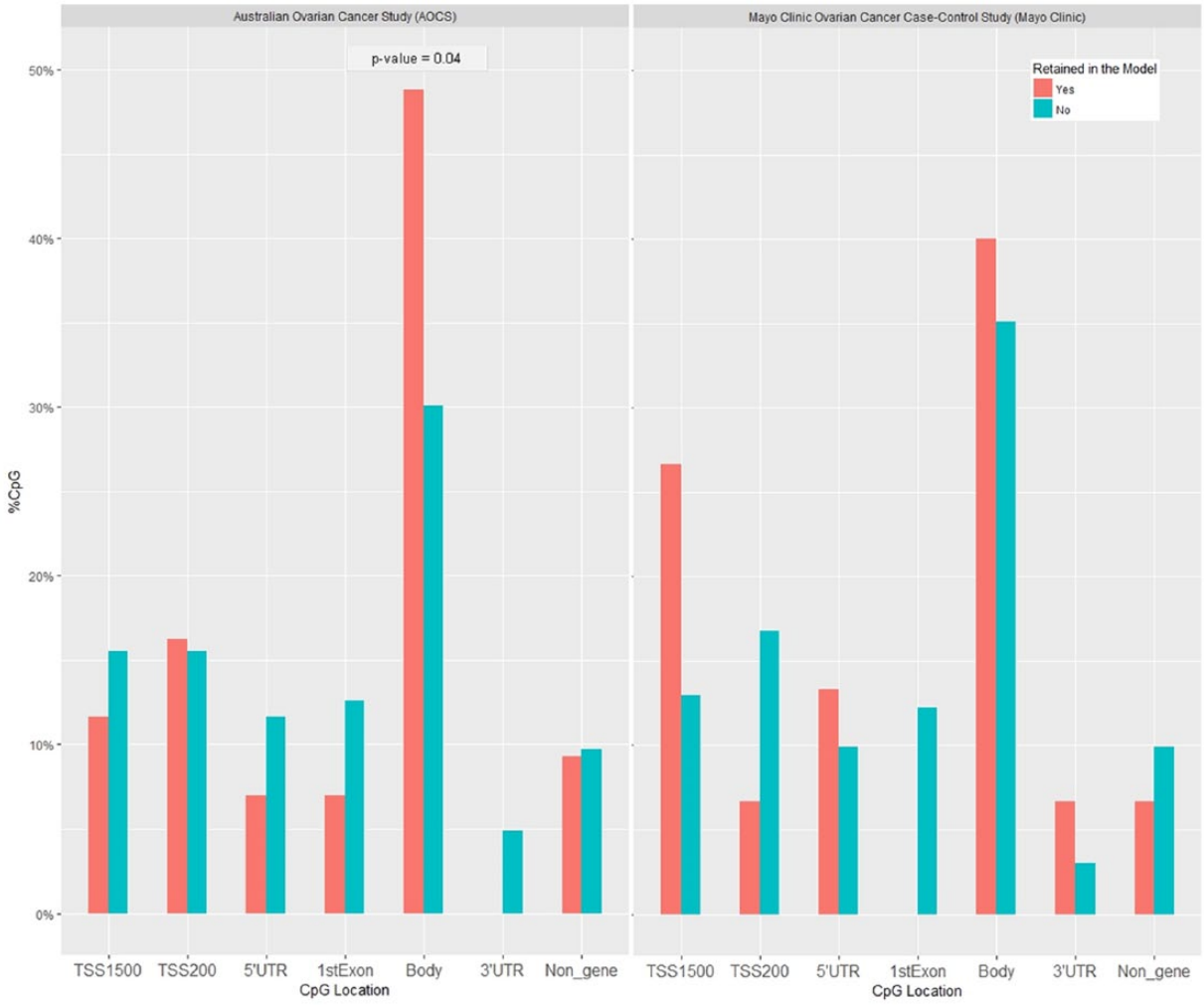


**Figure 1.** Distribution of *BRCA2* CpG locations by retention status in ENET methylation models.
CpG location represent gene region feature category describing the CpG position from UCSC: TSS1500 = 200 to 1500 bases upstream of transcription start site (TSS);
TSS200 = 0 to 200 bases upstream of TSS;
5′UTR = within the 5′ untranslated region, between the TSS and ATG start site;
1stExon = within first exon of the canonical transcript; body = between the ATG and stop codon, irrespective of the presence of introns, exons, TSS, or promoters;
3′UTR = between the stop codon and polyA signal;
Non_gene = region outside of all region listed above;
Annotation acquired from Illumina Infinium HumanMethylation450 v1.2 manifest file, column "UCSC_RefGene_Group";
The *P* value presented represents results of Fisher exact test of CpG location (gene body vs other location);
CpG retention in model (retained vs not retained). CpG counts are provided in Supplemental Table 1.

understanding of gene expression regulation in HGSOC. Concurrently with a gene-centric, genome-wide approach summarizing the effect of multiple CpGs on individual gene expression to reduce multiple testing burden, ENET modeling agnostically selects most predictive CpGs and SNPs while accounting for their correlation structure. Although variation in genetics and in gene expression at several genes detected in our study (BRCA2, *KRT6C, ZNF283*) has been associated with risk of onset, recurrence, and chemoresistance in ovarian and breast cancers,[17,22–24] agnostically evaluated gene-level methylation in these genes has not been previously reported to affect expression in HGSOC. We did not reveal a role for germline genetic variation alone or jointly with DNA methylation in altering gene expression in HGSOC, contrasting the application of ENET in other cancers[25] and the results of other single-variant expression quantitative trait loci methods in HGSOC.[20,26] We cannot rule out potential small associations that could not be detected with our modest-sized data sets. To our knowledge, this is the first study to interrogate all genome-wide protein-coding genes for the impact of methylation on gene expression in HGSOC using Elastic Net regularized regression method. We showed that DNA methylation in 5 genes was associated with gene expression in HGSOC. This method of genome-wide data integration has the potential to improve clinical risk prediction models and reveal novel therapeutic targets in HGSOC.

## Acknowledgements

The authors thank Drs Jonathon Tyrer and Paul PD Pharoah for data sharing.

## Author Contributions

YN, ME, and CW conceived and designed the experiments. YN, ME, CW, SMA, and MCL analyzed the data. YN, ME, and JMC wrote the first draft of the manuscript. YN, ME,

JMC, SMA, and ELG contributed to the writing of the manuscript. YN, JMC, KRK, CW, SMA, MCL, SJW, and ELG agree with manuscript results and conclusions. YN, ME, JMC, SMA, SJW, and ELG jointly developed the structure and arguments for the paper. YN, ME, JMC, SMA, DWG, BLF, SJW, and ELG made critical revisions and approved final version. All authors reviewed and approved of the final manuscript.

## REFERENCES


1. Siegal RL, Miller KD, Jemal A. Cancer statistics. *CA Cancer J Clin*. 2016;66:7–30.
2. Liu J, Cristea MC, Frankel P, et al. Clinical characteristics and outcomes of BRCA-associated ovarian cancer: genotype and survival. *Cancer Genet*. 2012; 205:34–41.
3. Bolton KL, Chenevix-Trench G, Goh C, et al; kConFab Investigators; Cancer Genome Atlas Research Network. Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer. *JAMA*. 2012;307:382–389.
4. Alsop K, Fereday S, Meldrum C, et al. BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group. *J Clin Oncol*. 2012;30:2654–2663.
5. Cunningham JM, Cicek MS, Larson NB, et al. Clinical characteristics of ovarian cancer classified by BRCA1, BRCA2, and RAD51C status. *Sci Rep*. 2014;4:4026.
6. Earp MA, Cunningham JM. DNA methylation changes in epithelial ovarian cancer histotypes. *Genomics*. 2015;106:311–321.
7. Fridley BL, Armasu SM, Cicek MS, et al. Methylation of leukocyte DNA and ovarian cancer: relationships with disease status and outcome. *BMC Med Genom*. 2014;7:21.
8. Hedditch EL, Gao B, Russell AJ, et al. ABCA transporter gene expression and poor outcome in epithelial ovarian cancer. *J Natl Cancer Inst*. 2014;106:dju149.
9. Jim HS, Lin HY, Tyrer JP, et al. Common genetic variation in circadian rhythm genes and risk of Epithelial Ovarian Cancer (EOC). *J Genet Genome Res*. 2015;2:017.
10. Johnatty SE, Tyrer JP, Kar S, et al. Genome-wide analysis identifies novel loci associated with ovarian cancer outcomes: findings from the Ovarian Cancer Association Consortium. *Clin Cancer Res*. 2015;21:5264–5276.
11. Kelemen LE, Terry KL, Goodman MT, et al. Consortium analysis of gene and gene-folate interactions in purine and pyrimidine metabolism pathways with ovarian carcinoma risk. *Mol Nutr Food Res*. 2014;54:2023–2035.
12. Kuchenbaecker KB, Ramus SJ, Tyrer J, et al. Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nat Genet*. 2015;47:164–171.
13. Wang C, Cicek MS, Charbonneau B, et al. Tumor hypomethylation at 6p21.3 associates with longer time to recurrence of high-grade serous epithelial ovarian cancer. *Cancer Res*. 2014;74:3084–3091.
14. Pineda S, Gomez-Rubio P, Picornell A, et al. Framework for the integration of genomics, epigenomics and transcriptomics in complex diseases. *Hum Hered*. 2015;79:124–136.
15. Pineda S, Real FX, Kogevinas M, et al. Integration analysis of three omics data using penalized regression methods: an application to bladder cancer. *PLoS Genet*. 2015;11:e1005689.
16. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474:609–615.
17. Patch AM, Christie EL, Etemadmoghadam D, et al. Whole-genome characterization of chemoresistant ovarian cancer. *Nature*. 2015;521:489–494.
18. Phelan CM, Kuchenbaecker KB, Tyrer JP, et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nature Genet*. 2017;49:680–691.
19. Kalari KR, Nair AA, Bhavsar JD, et al. MAP-RSeq: Mayo analysis pipeline for RNA sequencing. *BMC Bioinform*. 2014;15:224.
20. Amos CI, Dennis J, Wang Z, et al. The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev*. 2017;26:126–135.
21. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1–22.
22. Johnson CA, Collis SJ. Ciliogenesis and the DNA damage response: a stressful relationship. *Cilia*. 2016;5:19.
23. Ricciardelli C, Lokman NA, Pyragius CE, et al. Keratin 5 overexpression is associated with serous ovarian cancer recurrence and chemotherapy resistance. *Oncotarget*. 2017;8:17819–17832.
24. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genet*. 2013;45: 353–361.
25. Barath A, Endreffy E, Bereczki C, et al. Endothelin-1 gene and endothelial nitric oxide synthase gene polymorphisms in adolescents with juvenile and obesity-associated hypertension. *Acta Physiologica Hungarica*. 2007;94:49–66.
26. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*. 2006;22:773–774.