Comprehensive Epigenomic Analysis for Identifying DNA Methylation Panel for Ovarian Cancer Detection

Ye Zhang

Biomedical Engineering Department, Johns Hopkins University, Baltimore, MD 21218

Abstract

Aberrant DNA methylation is a well-known contributor to carcinogenesis and the study of DNA methylation as a biomarker for cancer detection has become increasely popular. However, the translation of discoverd DNA methylation biomarkers into commercially available clinical test are limited. Therefore a comprehensive approach to identify highly-specific DNA methylation biomarkers and develop a panel of DNA methylation genes for specific cancer detection is under urgent demand. In this study, DNA methylation data of 96 samples achieved from Illumina MethylationEPIC BeadChip array was analyzed to identify methylation genes that are highly specific to high-grade serous ovarian carcinoma (HGSOC).

Introduction

Over the past decades, biomarkers have been playing an increasely important role in cancer detection and diagnosis with the advancement of genomic profiling technilogies. More specifically, cancer biomarkers have been applied for measureing the risk of cancer development in a specific biological sample (e.g. tissue), or the risk of cancer progression/response to a certain therapeutic strategy [1]. For example, anaplastic lymphoma kinase (ALK) gene mutation has been found in several types of cancer including non-small cell lung cancer (NSCLC) and neuroblastoma, and checking ALK gene rearrangements and overexpression has been used to determine NSCLC prognosis and plan cancer treatment [2,3]. Also, studies have shown that people who inherit certain mutation in BRCA1 and BRCA2 genes have a higher risk of getting breast cancer [4]. Therefore the BRCA gene test can help people with family history of breast cancer to determine whether they carry an inherited BRCA mutation and help unaffected high-risk people for close surveillance or even prevention.

Among the advancement of cancer biomarker studies, DNA methylation is a well-known contributor to all forms of cancer and can occur early in carcinogenesis, sometimes even prior to the development of precursor lesions [5,6]. The application of DNA methylation as a biomarker for cancer detection has emerged in recent years and shows potential in developing powerful tools for cancer prediction and diagnositics. For example, the DNA test of aberrant NDRG4 and BMP3 methylation in stool samples as part of coloresctal cancer screening has received FDA approval in 2014 [7]. Despite these encouraging development in cancer biomarker studies, there is still a large gap between initial biomarker discovery studies and their clinical translation. A recent review article performed a search of the PubMed database and found that around 1,800 DNA methylation biomarkers reported in more than 14,000 scientific articles, only 14 biomarkers have been translated into a commercially available clinical test [8]. Therefore a comprehensive approach to identify highly-specific DNA methylation biomarkers and develop small panels of DNA methylation biomarkers for a specific cancer is still under demand.

High-grade serous ovarian carcinoma (HGSOC) is one of most lethal cancers, with over 60% of women not diagnosed until late stages [9]. Though some studies have reported aberrant gene promoter methylation related to HGSOC, there still remains an urgent need for the identification of biomarkers capable of detecting HGSOC. In this project, we investigated the DNA methylation data resulted from MethylationEPIC BeadChip array, which contains genome-wide methylomic screening of 96 tissue samples including 23 HGSOC, 23 endometrioid endometrial carcinoma (EEC), 10 uterine serous carcinoma (USC), 4 high-grade tumors named "unclassified cohort" and 36 normal ovarian samples [10]. Using a comprehensive epigenomic approach, we

identified novel panels of DNA methylation genes in ovarian cancers, which can potentially used to distinguish HGSOC from healthy tissues with high specificity.

Methods

Data loading and preprocessing

DNA methylation data were processed from Illumina MethylationEpic array iDat files using functions implemented in the `minfi` package [11] from the Bioconductor bioinformatics software project [12]. First a core sample sheet containing the Illumina ExpicArray information of the 96 samples and a sample diagnosis table were loaded using `xlsx` package [13]. After matching Illumina MethylationEpic array sample information and phenodata, methylation data were imported from idat files using `minfi` package. With `ilm10b2.hg19` package [14] from Bioconductor, the updated hg19 annotations for the probe sequences represented on the Illumina Expicarray were remapped to hg38 based on the UCSC liftOver mapping [15].

Differentially Methylated Region Identifying

After methylation data reading, CpG-dense regions including islands, shores, and shelves were filtered within 1,500 bases of the transcription start site. Then the `bumphunter` algorithm [16] was applied, with a bootstrap null, cutoff of 0.2 and 100 iterations, to identify differentially methylated regions (DMRs) between HGSC tumors and normal ovarian samples. After bumphunter, the 1,000 most variable methylated probe were selected for subsequent analysis. At the same time, DMRs containing at least 2 differentially methylated probes were also selected for comparison.

NMF Consensus Clustering

After DMR identification, consensus clustering of DNA methylation profile was performed using the `NMF` package [17, 18]. To estimate the best consensus clustering groups, a rank range of 2 to 5 groups were used for NMF clustering respectively and the cophenetic correlation values were calculated for each rank value. Then candidate genes in each group were ranked according to area under ROC curve (AUC) calculated on HGSC tumors and normal ovarian samples.

Reproducibility

Everything presented in this report are reproduced in the R markdown file `Final_Report.Rmd`. In order to save time for knitting, the imported methylation dataset and data with prelimiary exploration have been saved as `data.RData`, which has been updated into a shared Dropbox folder. The uploaded data can be used in order to reproduce the exact same results in this report within shorter time because the methylation data is large and NMF algorithm takes a long time.

Results and Discussion

This project aims to explore and identify highly specific differentially methylated regions in HGSOC tissues as methylation biomarkers for HGSOC detection. DNA methylation data was achieved by performing Illumina MethylationEpic array analysis on 96 malignant and healthy gynecologic tissues. After preliminary selection of candidate DMRs, NMF algorithm was applied for consensus clustering of DNA methylation profile to generate optimum consensus methylation clusters. Candidate methylation biomarkers were then ranked and selected from each cluster based on their AUCs calculated on HGSC tumors and normal ovarian samples.

Genome-wide methylation analysis

The 96 samples run on Illumina MethylationEpic array include 23 HGSOC, 37 non-HGSOC malignant, and 36 normal ovarian tissues. The 37 non-HGSOC tumor samples comprised 23 endometrioid endometrial (EEC), 10 serous endometrial (ESC), and 4 other poorly differentiated (possibly HGSOC), unclassified tumors. The normal-appearing ovarian tissues included 11 fallopian, 13 endometrial, and 12 endocervical mucosal epithelial samples (**Supplementary Table S1**). Previous study of the methylation data has performed unsupervised hierarchical clustering of the 5,000 most variable differentially methylated probes, which showed that the methylomes of HGSOC and non-HGSOC malignant were significantly distinguishable from normal ovarian samples [10]. Previous study has also demonstrated that the clustering pattern remained largely unchanged if the unsupervised analysis was performed using only probes located within CGIs, shores, and shelves [10], indicating that the exploration of gene candidate can be focued within these CpG-dense regions. **Figure 1** below presents the unsupervised hierarchical clustering of the 5,000 most variable CpG sites located within CpG islands, shores, and shelves of the 96 tissue samples.

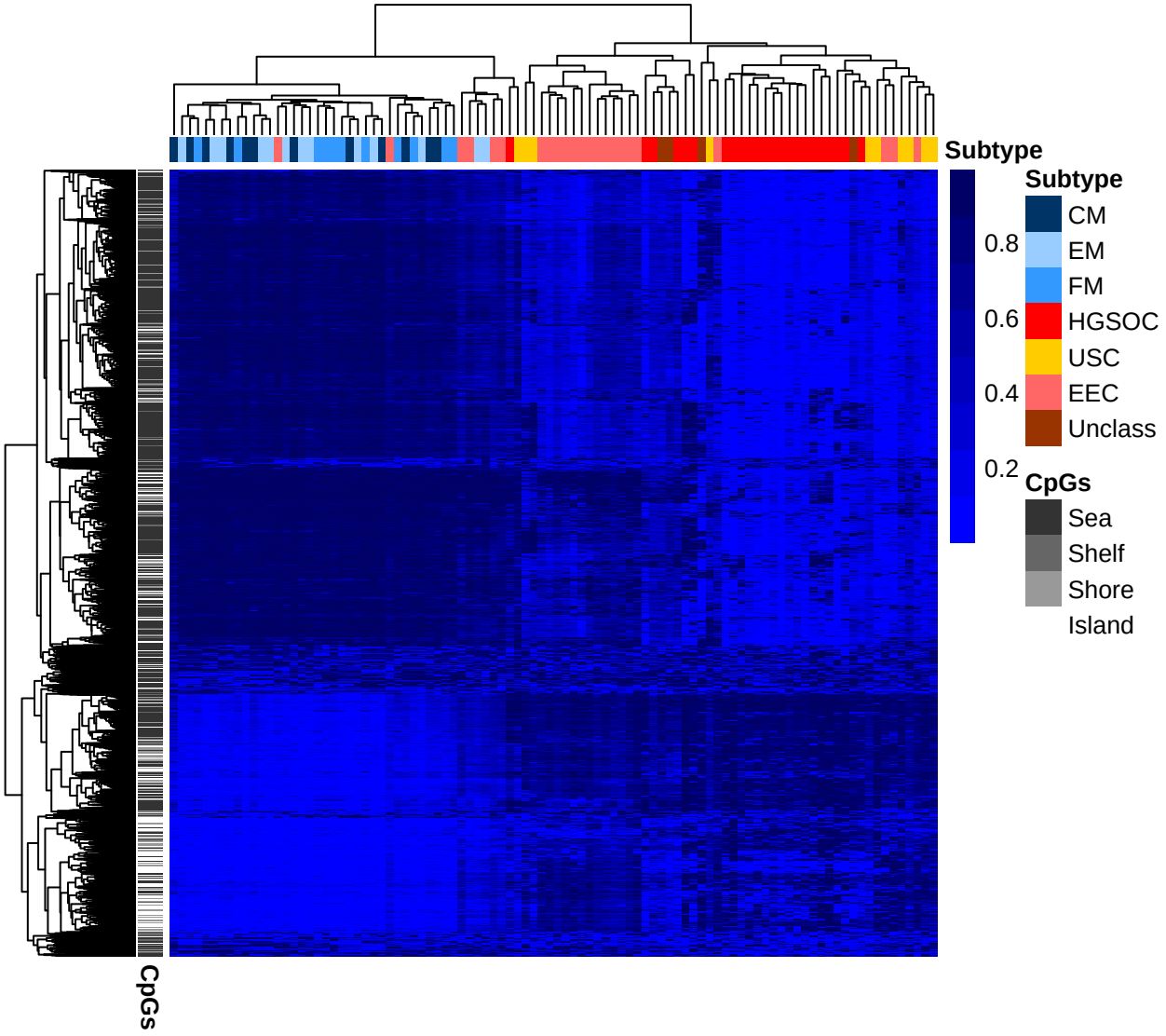

Figure 1. Methylation heatmap of 5000 most differentially methylated probes located within CpG islands, shores and shelves of the 96 samples.

Selection of Candidate Methylation Regions

In order to identify candidate methylation genes for translation into locus-specific methylation biomarker assays, bioinformatic algorithms were applied to smooth methylation levels across adjacent CpG sites [25]. Starting with 866,238 probes used in Illumina MethylationEpic assay for the analysis of 96 tissue samples, CpG sites lyin in CpG-dense regions (islands, shores, and shelves) within 1,500 bp of the transcription start site were first selected, giving 137,670 probes. Then 72,438 probes with high-specificity methylation, whose $\beta < 0.2$ in normal ovarian samples were selected for subsequent consensus clustering using NMF algorithm. Moreover, DMRs containing at least 2 differentially methylated probes within the same region were selected, resulting in 294 probes located in these 91 high-confidence DMRs (**Supplementary Table S2**). Methylation heatmap of these 294 probes in 96 ovarian tissue samples is presented in **Figure 2**.

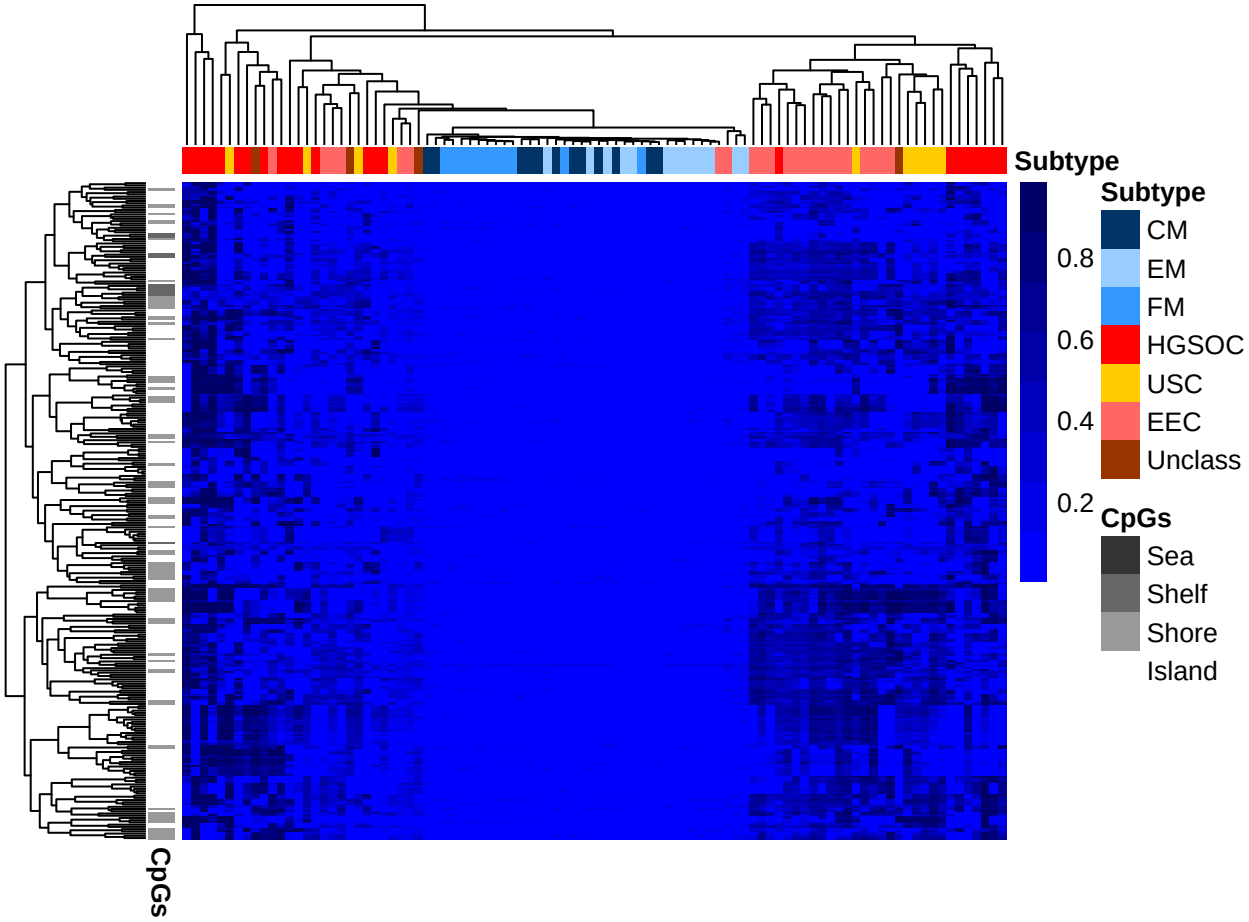

Figure 2. Methylation heatmap of 294 differenrially methylated probes located within 91 high-confidence DMRs of the 96 ovarian tissue samples.

Identification of Candidate Genes Using Consensus Methylation Clustering

After preliminary selection of differentially methylated probes, a nonnegative matrix factorization (NMF) algorithm was applied to analyze the the correlations between methylation patterns to identify complementary biomarkers for screening. First, consensus clustering was performed for the methylation data of selected 294 probes. Cophenetic correlation coefficients for rank ranging from 2 to 5 were calculated and shown in Figure 3, among which cophenetic correlation has the highest value of 0.9917 when rank equals 4. Therefore the selected 274 probes were grouped into 4 (rank = 4) consensus methylation clusters, as shown in Figure 4. There were 104, 78, 53, and 59 probes grouped into clusters 1 to 4, respectively. At the same time, 93

DNA methylation genes represented by these 294 probes were also grouped into four clusters(as listed in **Supplementary Table S3**).

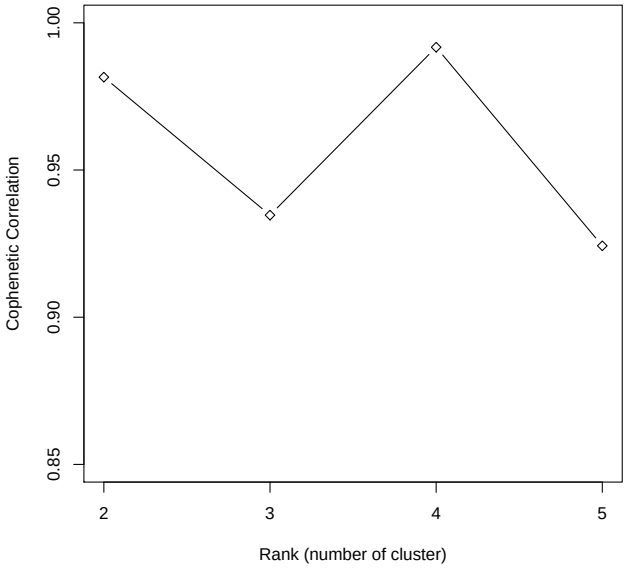

Figure 3. Cophenetic correlation coefficients for consensus clustering with rank ranging from 2 to 5. When rank = 4, cophenetic correlation has the highest value of 0.9917.

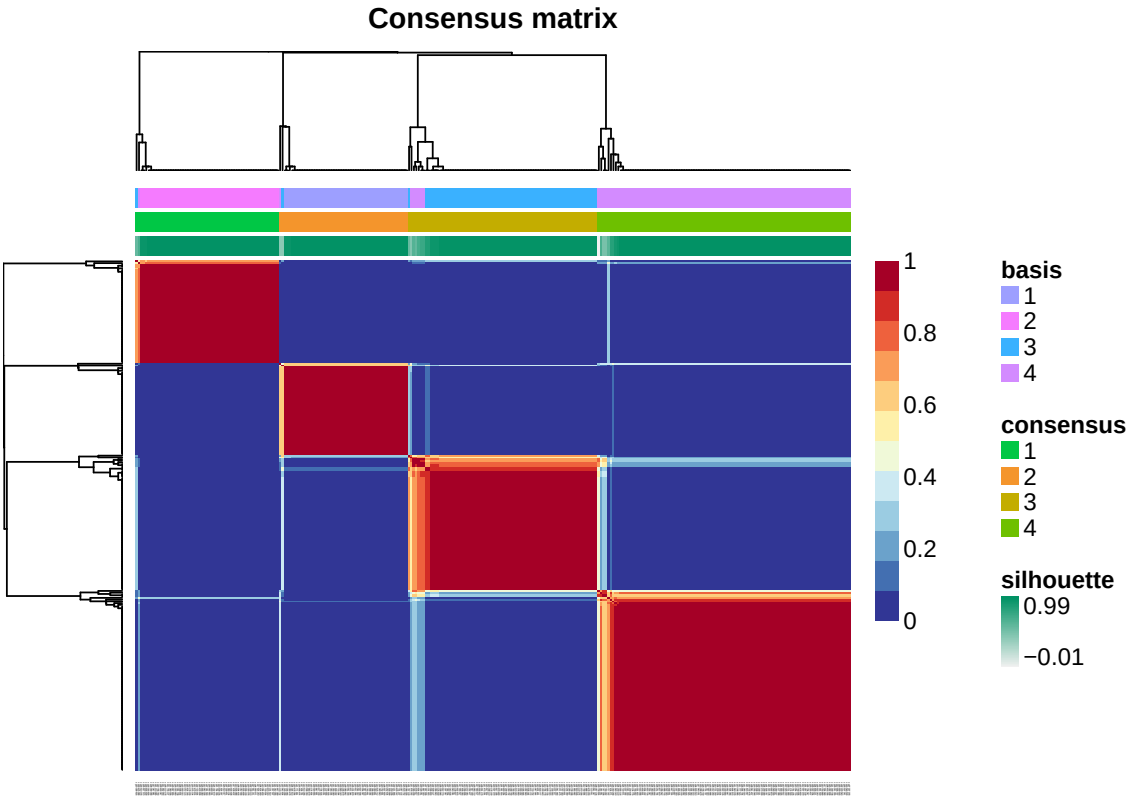

Figure 4. Correlation matrix showing four consensus clustering of methylation patterns with 274 probes in 96 tissue samples. Strong and weak correlations are shown in red and blue respectively.

To identify highly sensitive and specific biomarkers for HGSOC detection, the candidate genes in each cluster group were ranked according to their AUC calculated on HGSC tumors and normal ovarian samples

Table 1: Top Candidate Methylation Genes (within 93 genes) Identified in 4 Clusters

genes	probes	auc	NMF_cluster
HIST1H2BB	cg15387132, cg07701237, cg26426142, cg02221866, cg21250296, cg11503599	0.93	1
PCDHGA6	cg01357507	0.96	2
LOC200726	cg20841047, cg08057136, cg12110911	0.96	3
HIST2H2BF	cg04888113, cg12769994	0.97	3
PTPRN	cg03970036, cg15119274	0.95	3
NEUROD1	cg02819605, cg01863682, cg20709008	0.96	3
SNTG1	cg09246637, cg19254369, cg00230631	0.93	3
IRX2	cg18371475, cg15941948	0.95	3
PCDHB15	cg03572772, cg18606364	0.93	3
TUBB6	cg16546503, cg07307078, cg03507241	0.94	4
C6orf174	cg25929533, cg12695797	0.94	4
KCNK2	cg08453036, cg17934948	0.96	4
SIX6	cg07747306, cg13769906	0.93	4

after consensus clustering, as listed in **Table 1** below. There is one methylation gene *HIST1H2BB* in cluster 1 and one gene *PCDHGA6* in cluster 2 with AUC over 0.92, and there are 7 genes (*LOC200726*, *PTPRN*, *IRX2*, etc.) in cluster 3 and 4 genes (*TUBB6*, *C6orf174*, et al) in cluster 4 over the AUC threshold respectively. This panel is consistent with the results published in previous study [10], and suggesting some new candidates such as *HIST1H2BB* and *PCDHGA6*. Though AUC of these genes may not as high as those picked up in previous study, including genes from all the four clusters can help prevent missing any of the major methylation within the data.

Moreover, since the 294 probes are selected from the 137,670 probes by filtering those probes located in DMRs that contained only one differentially methylated probe within the same region, it is possible that those filtered regions/genes can still serve as good biomarkers for HGSOC detection. Therefore, methylation data of 137,670 probes were utilized instead for consensus clustering and identification of candidate genes. Because running NMF algorithm on 137,670 probes occupied too much computer memory, the most variable 1000 probes within these 137,670 probes were selected and grouped with consensus clustering using NMF algorithm. Similarly, cophenetic correlation coefficients for rank ranging from 2 to 5 were calculated and shown in **Figure 5**, among which cophenetic correlation has the highest value of 0.9755 when rank value equals 2. Then the most variable 1000 probes were grouped into 2 (rank = 2) consensus methylation clusters, as shown in **Figure 6**, with 546 probes in cluster 1 and 454 probes in cluster 2 respectively. At the same time, 466 DNA methylation genes where these 1000 probes located were also grouped into two clusters.

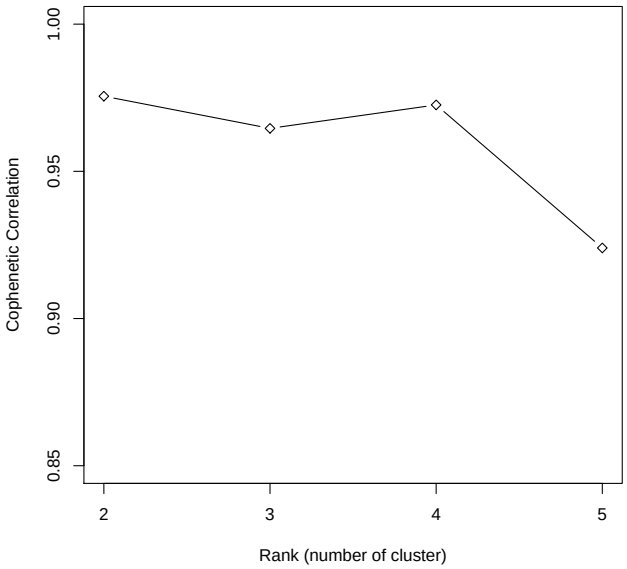

Figure 5. Cophenetic correlation coefficients for consensus clustering with rank ranging from 2 to 5. When k =2, cophenetic correlation has the highest value of 0.9755.

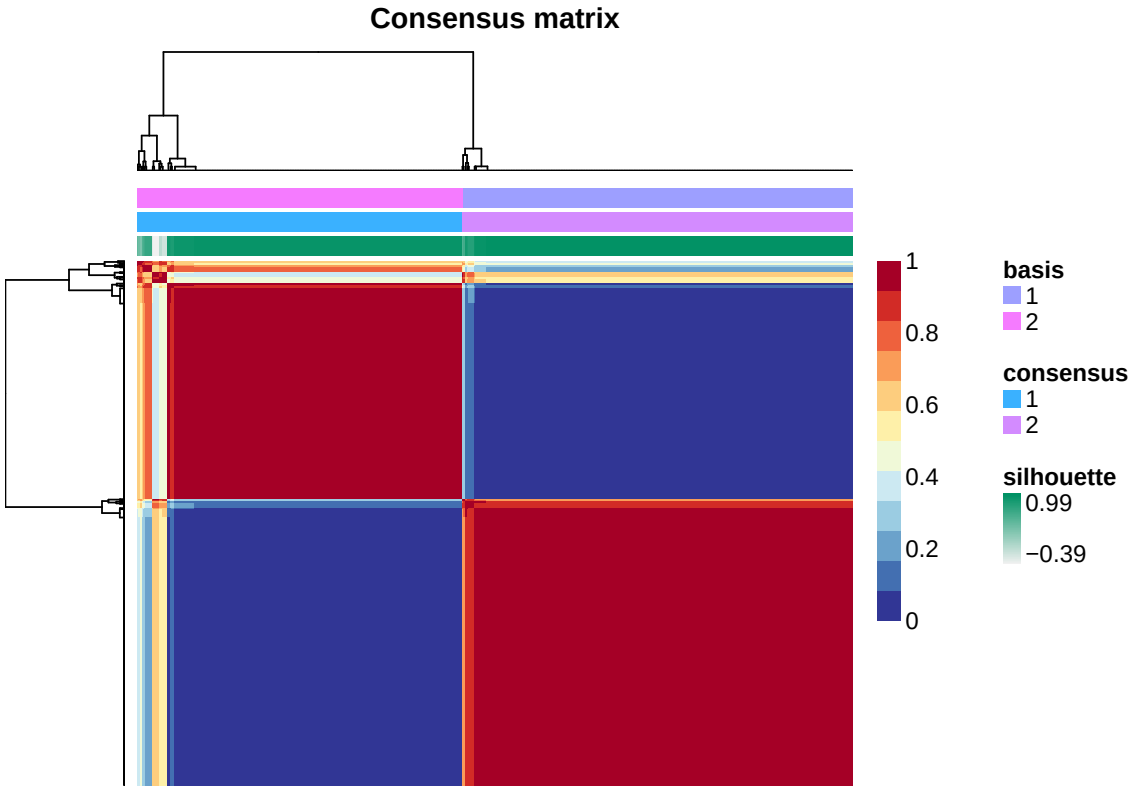

Figure 6. Correlation matrix showing two consensus clustering of methylation patterns using 1000 most variable probes in 59 samples (HGSOC & Normal tissue samples). Strong and weak correlations are shown in red and blue respectively.

Similarly, the candidate genes in each cluster group were ranked according to AUC calculated on HGSC tumors and normal ovarian samples after consensus clustering to select gene panel as candidate biomarkers . With a threshold of 0.95, the top 5 genes in cluster 1 and the top 16 genes in cluster 2 were selected (as listed in **Table 2**). Some genes listed in **Table 1**, such as *HIST2H2BF* and *KCNK2*, were also presented

Table 2: Top Candidate Methylation Genes (within 466 genes) Identified in 2 Clusters

genes	probes	auc	NMF_cluster
ZBTB16	cg07673133	0.97	1
LOC200726	cg20841047, cg08057136, cg12110911	0.96	1
HOXC13	cg25936147	0.96	1
C14orf23	cg09241022, cg13046832	0.96	1
PABPC5	cg16401529	0.96	1
HIST2H2BA	cg13688123	0.99	2
PCDHGA4	cg04246144, cg09066326, cg03835609, cg02741229, cg18617005, cg18507379, cg25353450, cg13933262	0.99	2
ZFP41	cg12680609	0.99	2
HIST2H2BF	cg14497172	0.98	2
ANTXR2	cg12667673	0.98	2
GRM6	cg25423752, cg02229543	0.98	2
TACR2	cg11893281	0.98	2
CD8A	cg12606911	0.97	2
NXPE3	cg17650664	0.97	2
KCNK2	cg08453036, cg17934948	0.96	2
SIX3	cg06622151, cg19186145	0.96	2
HAND2	cg05155840, cg04771946	0.96	2
PCDHGA6	cg01357507	0.96	2
NOVA1	cg22896309	0.96	2
NKX2-8	cg06621744	0.96	2
FBN1	cg10480343, cg22535307, cg08151731	0.96	2

here, while obviously more new candidate methylation gene were selected. These genes, on which there may only be one methylation probe located, can be considered to serve as methylation biomarkers in locus-specific assays for HGSOC detection.

In addition, since the cophenetic correlation coeffient of rank equalling 4 is 0.9726, which is very close to that of rank equalling 2, the consensus clustering of the most variable 1000 probes into 4 (rank = 4) methylation clusters were also taken into consideration, as shown in **Figure 7**. The NMF algorithm grouped 369 probes into cluster 1, 231 probes into cluster 2, 205 probes into cluster 3, 195 probes into cluster 4 respectively. Meanwhile, the 466 DNA methylation genes where these 1000 probes located were also grouped into four clusters. In a similar way as above, the candidate genes in each cluster group were ranked according to AUC calculated on HGSC tumors and normal ovarian samples after consensus clustering. With a threshold of 0.95, there are 2 genes selected in cluster 1 (*CD8A* and *NXPE3*), 3 genes selected in cluster 2 (*ZBTB16*, *HOXC13* and *C14orf23*), 7 genes selected in cluster 3 (*HIST2H2BA*, *PCDHGA4*, etc) and 4 genes in cluster 4 (*ANTXR2*, *TACR2*, etc), as listed in **Table 3**. All the 16 genes picked up here were covered in the gene panel with rank =2, which include 21 gene candidates. Though consensus clustering with higher cophenetic correlation coeffient (rank =2) can provide more gene candidates using the same AUC threshold, more clustering groups can help pick up the combination of gene candidates if even smaller gene panel is desired.

Table 3: Top Candidate Methylation Genes (within 466 genes)Identified in 4 Clusters

genes	probes	auc	NMF_cluster
CD8A	cg12606911	0.97	1
NXPE3	cg17650664	0.97	1
ZBTB16	cg07673133	0.97	2
HOXC13	cg25936147	0.96	2
C14orf23	cg09241022, cg13046832	0.96	2
HIST2H2BA	cg13688123	0.99	3
PCDHGA4	cg04246144, cg09066326, cg03835609, cg02741229, cg18617005, cg18507379, cg25353450, cg13933262	0.99	3
ZFP41	cg12680609	0.99	3
HIST2H2BF	cg14497172	0.98	3
HAND2	cg05155840, cg04771946	0.96	3
PCDHGA6	cg01357507	0.96	3
PABPC5	cg16401529	0.96	3
ANTXR2	cg12667673	0.98	4
TACR2	cg11893281	0.98	4
NOVA1	cg22896309	0.96	4
NKX2-8	cg06621744	0.96	4

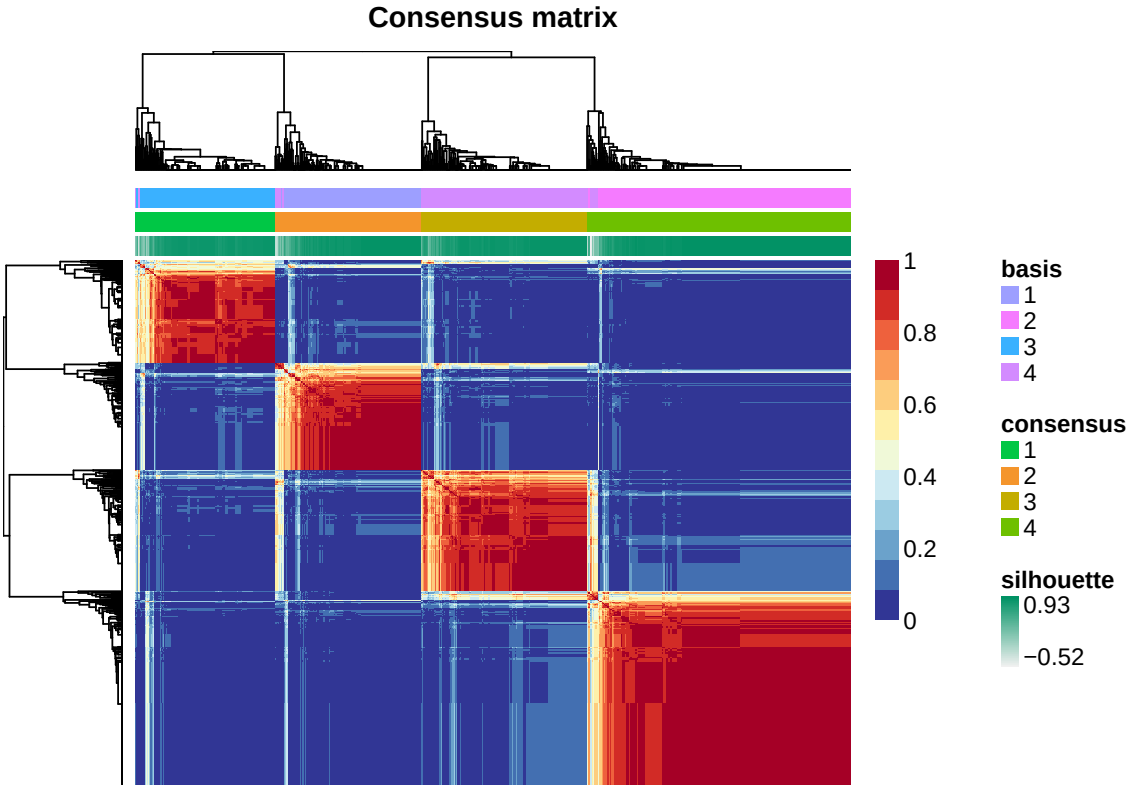

Figure 7. Correlation matrix showing four consensus clustering of methylation patterns using most variable 1000 probes in 59 samples (HGSOC & Normal tissue samples). Strong and weak correlations are shown in red and blue respectively.

Conclusion

In summary, using a comprehensive epigenomics approach, we have identified promising DNA methylation biomarkers for HGSOC screening. These highly specific methylation genes/loci will serve as candidate regions to design methylation-specific primers and probes in locus-specific assays (e.g. methylation-specific PCR) for HGSOC detection. At the same, this approach can be applied for the analysis of methylation data from other cancer types and potentially identify novel panels of methylation biomarkers for detection of other cancer types.

Reference

1. Goossens N, Nakagawa S, Sun XC, Hoshida Y. Cancer biomarker discovery and validation. Transl Cancer Res 2015;4(3):256-69.
2. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448(7153):561-U3.
3. Mosse YP, Laudenslager M, Longo L, Cole KA, Wood A, Attiyeh EF, et al. Identification of ALK as a major familial neuroblastoma predisposition gene. Nature 2008;455(7215):930-U22.
4. Wooster R, Weber BL. Breast and ovarian cancer. New Engl J Med 2003;348(23):2339-47 doi DOI 10.1056/NEJMra012284.
5. Baylin SB, Ohm JE. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? Nat Rev Cancer 2006;6(2):107-16.
6. Bartlett TE, Chindera K, McDermott J, Breeze CE, Cooke WR, Jones A, et al. Epigenetic reprogramming of fallopian tube fimbriae in BRCA mutation carriers defines early ovarian cancer evolution. Nat Commun 2016;7.
7. Koch A, Joosten SC, Feng Z, de Ruijter TC, Draht MX, Melotte V, et al. Analysis of DNA methylation in cancer: location revisited. Nat Rev Clin Oncol 2018;15(7):459-467.
8. Imperiale TF, Ransohoff DF, Itzkowitz SH, Levin TR, Lavin P, Lidgard GP, et al. Multitarget Stool DNA Testing for Colorectal-Cancer Screening. New Engl J Med 2014;370(14):1287-97.
9. Society AC. Cancer facts and figures 2017. Atlanta, GA: American Cancer Society; 2017.
10. Pisanic TR, Cope LM, Lin SF, Yen TT, Athamanolap P, Asaka R, et al. Methylomic Analysis of Ovarian Cancers Identifies Tumor-Specific Alterations Readily Detectable in Early Precursor Lesions. Clin Cancer Res 2018;24(24):6536-47 doi 10.1158/1078-0432.Ccr-18-1199.
11. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. Bioinformatics, 30(10), 1363-1369. doi: 10.1093/bioinformatics/btu049 (URL: http://doi.org/10.1093/bioinformatics/btu049).
12. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods 2015;12:115.
13. Adrian A. Dragulescu and Cole Arendt (2018). xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.1. (URL: https://CRAN.R-project.org/package=xlsx).
14. Hansen KD. IlluminaHumanMethylationEPICanno.ilm10b2.hg19: Annotation for Illumina's EPIC methylation arrays; R package version 0.6.0;2016.
15. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res 2002;12:996-1006.
16. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin DM, Feinberg AP, Irizarry RA (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International journal of epidemiology, 41 (1), 200-209. doi: 10.1093/ije/dyr238 (URL: http://doi.org/10.1093/ije/dyr238).
17. Renaud Gaujoux, Cathal Seoighe (2010). A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010, 11:367. (http://www.biomedcentral.com/1471-2105/11/367).
18. Huang RL, Su PH, Liao YP, Wu TI, Hsu YT, Lin WY, et al. Integrated Epigenomics Analysis Reveals a DNA Methylation Panel for Endometrial Cancer Detection Using Cervical Scrapings. Clin Cancer Res 2017;23(1):263-72.