Mining of Subtype Markers for the Prognosis of Ovarian Cancer based on Methylation Data

Lili Yin

Shengjing Hospital of China Medical University

Ningning Zhang

Shengjing Hospital of China Medical University

Qing Yang (✉ yangqing_sj@126.com)

China Medical University

Research

Keywords: Ovarian cancer; TCGA; methylation; molecular subtype; prognosis marker

Posted Date: March 11th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-16749/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: Ovarian cancer is one of three major malignancies involving the female reproductive system, and its morbidity and mortality are ranked number 3 and number 1 among gynecological tumors, respectively. DNA methylation (MET), as one of the main epigenetic modes, is closely related to the occurrence and development of ovarian cancer. To guide individualized treatment and improve the prognosis in ovarian cancer patients, it is of great significance to elucidate effective MET subtype markers. **Methods:** A total of 571 ovarian cancer MET samples were downloaded from the Cancer Genome Atlas (TCGA), and a COX proportional hazards model was established using the MET spectrum and clinically pathological parameters. Subsequently, the consensus clustering of CpG loci with a significant difference in both univariate and multivariate analyses was performed to screen the molecular subtypes, and these CpG loci were subjected to gene function annotation. Finally, CpG MET loci associated with poor prognosis in ovarian cancer patients were further screened by constructing a weighted gene co-expression network analysis (WGCNA). **Results:** A total of 250 prognosis-related MET loci were obtained by COX regression and 6 molecular subtypes were screened by clustering. There was a remarkable MET difference between most subtypes, of which Cluster 2 had the highest MET level and demonstrated the best prognosis in patients, while Cluster 4 and Cluster 5 had a MET level significantly lower than that of the other subtypes and demonstrated a very poor prognosis. All Cluster 5 samples were at a high grade, while the percentage of Stage IV samples in Cluster 4 was evidently greater than that in the other subtypes. Using the co-expression network, 5 CpG loci were eventually obtained: cg27625732, cg00431050, cg22197830, cg03152385, and cg22809047. The clustering analysis shows that the prognosis in patients with hypomethylation was significantly worse than that in patients with hypermethylation. **Conclusions:** These MET molecular subtypes can be used not only to evaluate the prognosis in ovarian cancer patients but also to fully distinguish the tumor stage and histological grade in these patients. Prognosis-related CpG loci can be applied as biomarkers for individualized treatment in ovarian cancer patients.

Background

Ovarian cancer is a disease of high heterogeneity with varying molecular phenotypes, pathogeneses, and prognoses, and its morbidity is ranked number 3 among malignant gynecological tumors. However, ovarian cancer is not easily detected at an early stage since the ovaries are located deep in the pelvis, and at diagnosis, the tumor is advanced with distant metastases in 70% of cases. Most patients will experience recurrence within 2 years, and there is a lack of effective therapies for recurrent ovarian cancer, which ranks the mortality of ovarian cancer as number 1 among gynecological tumors. Therefore, precision therapy is an urgent demand.

Epithelial ovarian cancer is the most common pathological type and accounts for 80–95% of ovarian malignancies. It is classified into 5 histological subtypes: high-grade serous adenocarcinoma, endometrioid adenocarcinoma, clear cell adenocarcinoma, mucous adenocarcinoma, and low-grade serous adenocarcinoma¹⁻⁵. In a large, prospective, phase-III clinical study, Kommoss et al., showed that

histological grading is necessary in early ovarian cancer but has no significance in prognostic evaluation in advanced ovarian cancer ⁶.

By combining the clinical pathological and molecular biological characteristics of ovarian cancer, [Shih leM et al.](#), classified epithelial ovarian cancer into types I and II. Type I epithelial ovarian cancer mainly includes low-grade serous carcinoma and low-grade endometrioid carcinoma, and most cases have early onset and good prognosis. Type II epithelial ovarian cancer mainly includes high-grade serous carcinoma, high-grade endometrioid carcinoma, and undifferentiated carcinoma, and most cases have rapid onset and poor prognosis. Type I epithelial ovarian cancer is significantly correlated with mutations in BRAF, KRAS, and PTEN, while type II epithelial ovarian cancer is associated with p53 mutations and also very frequently with BRCA1/2 mutations. The dualistic theory reflects different biological behaviors and clinical prognoses of tumors, and such differences are especially remarkable between low-grade and high-grade serous ovarian carcinomas. However, application of the dualistic theory in non-serous ovarian carcinoma is limited; for instance, clear cell carcinoma has many biological behaviors similar to type II epithelial ovarian cancer, despite belonging to type I epithelial ovarian cancer. It is of profound significance to realize precision molecular typing of ovarian cancer for clinical treatment and prognosis monitoring ⁷.

As a result of the continuous improvement in human genome sequencing technology and the advancement of biomedical analysis technology, a new trend has been formed that involves molecular targeted therapy and prognosis evaluation based on the molecular typing of malignant tumors. Molecular-targeted therapy has been successfully applied in several tumors, including ER (+) or HER2(+) breast cancer and EGFR-mutated lung cancer ⁸⁻¹¹, presenting great progress in precision medical treatment. By K-means clustering, Tothill et al. ¹² detected the gene expression spectrum of 285 cases of endometrioid and serous tumors originating from the ovary, peritoneum, and tuba uterine, finally identified 6 molecular subtypes, of which 4 (high interstitium-reactive, high immunity, hypomethylation trix-reactive, and interstitial low immunity) are features of high-grade serous ovarian carcinoma and can be used for predicting prognosis in these patients.

Based on gene expression, TCGA and Tothill et al., divided high-grade serous carcinoma into 4 subtypes: immunoreactive, differentiated, proliferative, and interstitial ¹³. Moreover, the clinical trial completed by Kommoss et al. showed that in high-grade serous ovarian carcinoma patients with proliferative and interstitial molecular subtypes, bevacizumab can improve the progression-free survival to different degrees; however, this study was limited to the molecular typing of the gene expression spectrum ¹⁴.

Despite the impact of gene changes on cancer occurrence, epigenetic changes such as DNA MET also play an important role. Epigenetic inheritance refers to the hereditary changes that occur under the precondition of no changes in DNA sequence, such as histone modification, DNA MET, RNA editing, and gene silencing. The occurrence and growth of ovarian cancer involve several pathways including DNA repair, cell apoptosis, cell cycle regulation, and changes in protooncogenes tumor suppressor genes. Studies have suggested that epigenetic changes in these pathways play essential roles in the

development of ovarian cancer, and the detection of MET signals is helpful for early diagnosis^{15,16}. DNA MET mainly occurs in CpG islands; CpG expression can be inhibited by the hypermethylation of tumor suppressor gene promoters and enhanced by the decreased demethylation probability of protooncogenes. Different regulatory effects of protooncogenes and tumor suppressor genes contribute to the occurrence of cancer¹⁷⁻²⁰. According to relevant studies, the tumor suppressor gene involved in ovarian cancer exists in a hypermethylated state, and changes in its MET level is an important molecular foundation for cancer occurrence^{21,22}. Hu WL et al.,²³ built a DNA MET interaction network for ovarian cancer, breast cancer, and glioma, confirming that the number of DNA MET loci was associated with prognosis in cancer patients; however, no DNA MET molecular typing of ovarian cancer was performed.

In the present study, the univariate and multivariate COX proportional hazards models were established by analyzing Illumina Infinium® Human Methylation 27K data in the TCGA database and combining the MET level and clinical data of the samples. Subsequently, 6 molecular subtypes associated with the prognosis in ovarian cancer patients were screened by consensus clustering of MET spectra with a significant difference in both models, and the ovarian cancer patients were classified using these subtypes. Further, 5 CpG hypomethylation loci related to poor prognosis in ovarian cancer patients were obtained by constructing a WGCNA co-expression network. These loci are of great significance for the clarification of the pathogenesis of ovarian cancer, and can be used as effective tumor markers to provide a reference for clinical prognosis and individualized treatment.

Materials And Methods

1. Preprocessing of ovarian cancer expression datasets and preliminary screening of DNA MET loci

TCGA²⁴ GDC API was utilized to download the latest clinical follow-up information and RNA-Seq data. Illumina Infinium® Human Methylation 27 BeadChip (*Illumina 27K*) microarray was acquired from UCSC Cancer Browser. Samples with complete clinical data and methylation spectrum data were selected. CpG loci with NA (Not Available) > 70% in all samples were deleted. The *impute-KNN* of R package was used to fill the missing value of methylation spectrum. The unstable genomic methylation loci were further removed, involving CpGs and single nucleotide loci on sex chromosomes as well as CpG loci that were not annotated to the gene promoter region²⁵. We divided the datasets into two queues: a training set coupled with a test set. The standards for the subgroups included: 1) Samples were assigned to the training set and the test set randomly; and 2) the data of the two groups should be similar, including age distribution, clinical stage, follow-up time, and mortality ratio.

2. Univariate survival analysis of MET loci in the training set

The research objective was to determine the molecular subtypes of ovarian carcinoma as prognostic determinants. Therefore, CpG loci, which had an important impact on survival, were utilized as a classification feature. First and foremost, a univariate COX proportional hazard model was established based on the methylation level of each CpG loci, age, tumor grade and stage, coupled with survival data

by coxph function of the R-package **survival**. Subsequently, we introduced the significant CpG loci obtained from the univariate model into the multivariate COX proportional hazards model, and took the significant age and clinical attributes in the univariate model as covariables. Ultimately, the CpG loci, which were still significant, were employed as classification features [25]. For each CpG island, the multivariate COX proportional hazards model formula was described below:

$$h(t, x) = h_0(t) \exp(\beta_{\text{methy}} \text{methy}_i + \beta_{\text{age}} \text{age} + \beta_{\text{stage}} \text{stage}) \quad (1)$$

In the formula, “methy_i” is the carrier of the CpG locus methylation level in the sample. “Age” and “stage” describe the age and clinical characteristics of the patients, respectively. “β_{methy}”, “β_{age}”, and “β_{stage}” are regression coefficients. The *P*-values of the COX regression coefficient was adjusted by Benjamini–Hochberg error detection rate. Various comparing processes were carried out.

3. Screening of molecular subtypes by the consensus clustering of methylation profile with a significant difference in both univariate and multivariate analyses

Consensus ClusterPlus in the R package²⁶ was utilized for consensus clustering according to the method described by Zhang et al.²⁵ The subgroups of epithelial ovarian tumors were identified based on the most variable CpG loci. The algorithm is described as follows. First, double sampling of some items and features from the data matrix was conducted, in which each sub-sample was divided into several groups (max.) using a user-specific clustering algorithm (*k-means*, hierarchical clustering, or custom algorithms). The paired consensus value (defined as the proportion of clustering running for the combination of two items) was calculated and stored in the k_i consensus matrix. Second, the final coherent sheaf clustering for each k_i was completed using the distance of 1- consensus value and pruned into k_i group through cutting, which is known as consensus clustering. The algorithm determined the “consensus” clustering by measuring the stability of the clustering results applied to random data subsets from given clustering methods. In each iteration, 80% of the tumors were sampled, and a *k-means* algorithm with the Euclidean squared distance measures were utilized:

$$d = \sqrt{\sum_{k=1}^N (x_k - y_k)^2} \quad (2)$$

There was *k* = 2–10 groups, and these results were compiled for 100 times. The cluster consensus as well as item consensus results were obtained with **Consensus ClusterPlus**.R package. The graphical output results included the heat map of consensus matrix, cumulative distribution function (CDF) diagram, and Δ region diagram. The criteria of clustering number included relatively high consistency within the cluster, relatively low coefficient of variation, and insignificant increase of the area under the CDF curve (AUC). The CV (%) was calculated based on the formula below:

$$CV (\%) = (SD/MN) \times 100 (3)$$

Where SD is the standard deviation, while MN is the average value of the samples. We selected category number as the area under the CDF curve, and there was no significant change. The consensus clustering heatmap was generated using the R package *pheatmap*.

4. Clustering analysis of the methylation expression profile and analysis of the clinical characteristics of screened molecular subtypes

The stable clustering results were selected and the methylation profile was analyzed by clustering analysis. The distance between the MET loci was calculated using the Euclidean distance. Furthermore, the distribution of various molecular subtype samples was analyzed with respect to prognosis, stage, grade, and age.

5. Gene annotation of MET loci

As for the genes corresponding to the gene promoter regions annotated by the selected CpG loci, the transcription factor enrichment analysis was performed by the online tool g:profiler²⁷.

6. WGCNA co-expression analysis of CpG loci

Based on the modification beta value of selected CpG loci, the co-expressed CpG loci were mined by WGCNA co-expression algorithm. The distance between CpG loci was calculated using Pearson Correlation Coefficient. The R package *WGCNA* was used to construct weighted co-expression network, and select a soft threshold of 4 to filter the CpG co-expression modules. The results showed that the co-expression network conformed to the scale-free network. That is to say, the $\log(k)$ of node k presented in the connection is negatively correlated with the $\log(P(k))$ of the probability of node k , with a correlation coefficient larger than 0.8. In order to ensure a scale-free network, we chose $\beta = 4$. The next step was to convert the expression matrix into an adjacency matrix, and then transform the adjacency matrix into a topological matrix. Based on TOM, we utilized average-linkage hierarchical clustering method to cluster genes. The minimum number of genes in each lncRNA network module was set at 30 according to the standard merged dynamic tree cutting. After determining the gene modules with the dynamic cutting method, we calculated eigengenes of each module in turn. The modules were clustered, and the adjacent modules were merged into new modules.

7. Construction of prognosis models and data validation of independent test set

Unsupervised clustering analysis was conducted on the CpG methylation profile selected in the previous step. The similarity between samples was calculated by using the Euclidean distance. The samples were then divided into two groups according to the methylation level of CpG loci. The prognosis differences between the two groups were further analyzed. The methylation profile of 286 samples in the test set were used for validation.

Results

1. Selection of 250 characteristic MET loci

The Illumina Infinium® Human Methylation 27 BeadChip microarray contained 613 samples in total, with 571 samples being screened using MET detection. The missing data imputation of the MET spectrum was performed using Impute in the R software package, and 25,154 MET loci were selected following exclusion of the unstable genomic MET loci. All 571 samples were assigned to either a training set ($n = 285$) or a validation set ($n = 286$). The clinical pathological information in the training set and validation test is shown in **Table 1**.

The MET loci and survival data were analyzed using a univariate COX proportional hazards regression model with $p < 0.05$ as the threshold. A total of 967 loci demonstrated a significant difference in prognosis (**Table S1**), of which the top 20 loci with the most significant difference are shown in **Table 2**.

The prognostic significance of age had a logrank P value of $5.93e-06$, while that of stage was 0.0379 . The significant MET loci were selected using a univariate COX model followed by multivariate COX proportional hazards regression model analysis, with stage and age as covariates. Finally 250 significant MET loci were obtained (**Table S2**).

2. Screening 6 molecular subtypes by consensus clustering of the MET loci

Consensus clustering of MET loci with a significant difference in both univariate and multivariate analyses was performed using *Consensus ClusterPlus* in the R package to screen the molecular subtypes. The similarity between samples was calculated using the Euclidean distance, the clustering was performed with *k-means*, and 80% sampling was conducted 100 times using a double-sampling method. The optimal cluster number was determined by CDF; different colors in the CDF curve represent different cluster numbers (**Figure 1A**), the AUC was larger at 6 and 7 clusters and the clustering effect was better. Further observation of the CDF delta area curve (**Figure 1B**) shows that at 6 clusters, the AUC demonstrates stable clustering results. Following consideration, $k = 6$ was selected and 6 molecular subtypes were obtained.

3. Clustering analysis of the MET expression spectra of 6 molecular subtypes

The composition and number of samples in the 6 clusters were evaluated using the consensus matrix. The color gradient was from white to blue, indicating the consensus of progression. In the matrix permutation, the same clusters were made mutually adjacent; and eventually, a color-coded heat map was displayed and featured by the arrangement of dark blue blocks on a diagonal white background (**Figure 2A**), showing that 285 tumor samples were assigned to these 6 clusters. Furthermore, clustering analysis was performed on 250 MET spectra; the distance between MET loci was calculated using the Euclidean distance, and the heat map was generated by pheatmap using clinical pathological stage and histological type as notes (**Figure 2B**).

The pairwise comparison of various subtypes was performed using a t-test, and the results revealed that most MET loci had a low beta value. There was a significant difference in MET level among the majority of subtypes; the MET level in Cluster 2 was remarkably higher than that in the other 5 subtypes, while the MET levels in Cluster 4 and Cluster 5 were evidently lower than those in the other subtypes (**Table S3**).

4. Analysis of the clinical characteristics of the 6 molecular subtypes

We further analyzed the distribution of the 6 molecular subtypes with respect to prognosis, stage, grade, and age (**Figure 3**). There was a significant difference in the prognosis among the 6 subtypes; the prognosis was best in Cluster 2, worst in Cluster 5, and poor in Cluster 4 (**Figure 3A**), indicating that the prognosis of hypomethylation subtypes was inferior to that of hypermethylation subtypes. The samples in Cluster 5 were all stage III, and the percentage of stage IV samples in Cluster 4 was significantly higher than that in the other subtypes (**Figure 3B**). All samples in Cluster 5 were G3 (**Figure 3C, Table S4**), suggesting that the hypomethylation subtypes were mostly high-grade in clinical pathology. The age of the patients of samples in Cluster 5 was remarkably greater than that of the patients of the other subtypes, and the onset age was 70–80 years old (**Table S5**), while the mean age of the patients of samples in Cluster 2 was the lowest (**Figure 3D**), indicating that the age of patients with hypomethylation subtypes was generally higher than that of patients with hypermethylation subtypes. The above findings suggest, to a certain degree, that these DNA MET subtypes could be used to predict the prognosis, tumor stage, and pathological grade in ovarian cancer patients.

5. Gene annotation and function analysis of the 250 MET loci

There was a total of 285 genes corresponding to the gene promoter regions annotated by 250 CpG loci, and these genes were subjected to transcription factor enrichment analysis using the online tool g:profiler. It was found that 42 genes were significantly enriched to Transcription Factor EC (TFEC) (logrank P value = 0.0107) (**Figure 4A**). At present, the role of TFEC in cancer progression has been studied to a limited extent; thus, to further explore the biological functions in which TFEC may be involved, TFEC-co-expressed molecules in the cBioPortal database were elucidated. Subsequently, the top 300 molecules with the most positive and negative correlations according to Spearman's correlation were selected. Functional enrichment analysis was performed using the DAVID 6.7 database and visualized using “GOplot” in the R software package. Finally, the top 5 biological processes with the most significant functions were chosen: GO:0006955 – immune response, GO:0050776 – regulation of immune response, GO:0006954 – inflammatory response, GO:0045087 – innate immune response, and GO:0007165 – signal transduction. TFEC may promote the occurrence and progression of ovarian cancer by influencing these biological functions(**Figure 4B**).

6. Screening 5 CpG loci by WGCNA co-expression analysis

Using the WGCNA co-expression algorithm, 250 significant CpG loci were mined. Evaluation of the scale-free model was performed at different soft thresholds; a larger value and lower mean connectivity both indicated better compliance with the scale-free distribution. Finally, $\beta = 4$ (**Figure 5A, B**) was selected, and

the settings of height = 0.25, deepSplit = 3, minModuleSize = 10 were chosen. A total of 7 modules were obtained (**Figure 5C**), of which the grey module is the set of genes that could not be clustered to other modules. The statistics of genes in various modules are shown in **Table 3**. The 250 CpG loci were assigned to 7 modules. Pearson's correlation coefficient between the ME of each module and the characteristics of the samples was calculated; a higher correlation coefficient indicates that the module was more important. In **Figure 5D**, the row represents the eigengenes of each module and the column represents the feature information of the samples. The greatest correlation can be seen between the yellow module and Cluster 2 ($R = 0.68$, logrank $P = 7e-40$), the brown module and Cluster 3 ($R = 0.51$, logrank $P = 1e-20$), and the black module and Cluster 5 ($R = 0.61$, logrank $P = 6e-30$).

Since Cluster 2 demonstrated the best prognosis among all the clusters, all CpG loci in the yellow module, mostly correlated with Cluster 2, were selected, and the interaction network was constructed according to their weighted relationship (**Figure 6A**). In this network, the CpG loci with a network centrality > 10 were cg27625732, cg00431050, cg22197830, cg03152385, and cg22809047. Furthermore, the expression relationships among the 22 CpG loci were calculated, and a significantly higher correlation was found among 8 (cg27625732, cg00431050, cg22197830, cg03152385, cg22809047, cg00328227, cg06851207, and cg01777397) (**Figure 6B**). Finally, a total of 5 CpG loci in the intersection were chosen, which had a strong correlation between each other and a centrality > 10 in the weighted network, as the characteristic MET loci of Cluster 2 samples (**Table 4**).

7. Clustering analysis of the 5 CpG loci

Unsupervised clustering analysis was further performed on the MET spectra of the 5 selected CpG loci, and the similarity between samples was calculated using the Euclidean distance. **Figure 7A** shows that the samples were divided into two groups according to the MET level of the 5 CpG loci: Cluster 1 (hypomethylation group) and Cluster 2 (hypermethylation group). The prognosis difference between these two groups was further analyzed (**Figure 7B**), indicating that the prognosis in the hypomethylation group was significantly poorer than that in the hypermethylation group.

8. Model validation using the test dataset

The MET spectra of the 5 CpG loci from the 286 samples in the test dataset were extracted and analyzed by hierarchical clustering (**Figure 8B**). The results show that the MET spectra of the 5 CpG loci were also obviously clustered into two groups: Cluster 1 and Cluster 2. The MET level of Cluster 1 samples was significantly higher than that of Cluster 2 samples. The prognosis difference between Cluster 1 and Cluster 2 (**Figure 8B**) was further analyzed, and it was found that the prognosis in the hypermethylation group was remarkably better than that in the hypomethylation group, which is consistent with the results of the training set.

9. Flow chart of all the analysis

The flow chart of mining of subtype markers for the prognosis of ovarian cancer based on methylation data is shown in **Figure 9**. All the R package covered in this article is listed in **Table S6**.

Discussion

In recent years, an increasing number of studies have focused on exploring the molecular typing of epithelial ovarian cancer to promote the realization of personalized treatment and improve the survival rate in patients; however, the achievements of molecular typing remain in the initial phase. Studies have shown that the occurrence of cancer is associated with genetic changes, and epigenetic abnormalities are also contributors. DNA MET is the major epigenetic modification mode of genomic DNA; it is an important means in regulating the functions of the genome²⁸ and is closely associated with the occurrence, progression, treatment, and prognosis of ovarian cancer. DNA MET-based molecular typing and subtype markers are of great significance for guiding personalized treatment and prognosis evaluation in ovarian cancer patients.

In the present study, 571 ovarian cancer MET samples were downloaded from the TCGA database, 250 MET loci related to the prognosis of ovarian cancer patients were screened by COX regression analysis, and 6 molecular subtypes were selected by clustering with k-means. There was a significant difference in MET loci among most subtypes; the highest MET level and the best prognosis were observed in Cluster 2, and the MET level in Cluster 4 and Cluster 5 was remarkably lower than that in the other subtypes, accompanied by a very poor prognosis. This suggests, to a certain degree, that the prognosis of patients with a hypomethylation subtype was worse than that of patients with a hypermethylation subtype. All samples in Cluster 5 were high-grade, and the mean age of patients in Cluster 5 was higher than that in the other subtypes. The percentage of stage IV samples in Cluster 4 was significantly greater than that in the other subtypes. The above findings suggest that these molecular subtypes can be used not only to evaluate the prognosis in ovarian cancer patients, but also to fully distinguish the tumor stage, histological grade, and age of these patients to guide subsequent treatment.

DNA MET molecular typing also plays a very important role in the diagnosis, treatment, and prognosis of other tumors. Zhang et al.²⁵ screened 9 molecular subtypes by clustering analysis on DNA MET data in 669 breast cancer patients, and the DNA MET mode was reflected in varying races, ages, tumor stages, subject states, histological types, metastatic states, and prognoses. In comparison with PAM50 subtypes using gene expression clustering, DNA MET subtypes are more precise and can be used for the precision treatment of specific histological subtypes of breast cancer.

[Jurmeister](#) constructed a DNA MET map using the whole genome MET data from 600 cases of primary pulmonary, colorectal, and upper gastrointestinal adenocarcinoma, and successfully distinguished between pulmonary enteric adenocarcinoma and metastatic colorectal cancer²⁹.

[Williams](#) et al.³⁰ measured the MET level in different histological subtypes of 154 cases of child germ cell tumors using the Illumina Infinium® Human Methylation 450K chip, identifying 4 molecular subtypes.

The MET level in the germ cell tumors was low, and these molecular subtypes provided information regarding their etiology.

Again using the Illumina Infinium® Human Methylation 450K chip, [Wu SP1 et al.](#)³¹ detected the DNA MET state in 482 and 421 CpG loci in 10 samples of Ewing's sarcoma, 11 samples of synovial sarcoma, and 15 samples of osteosarcoma. Moreover, they developed and validated a whole-genome DNA MET classifier to identify osteosarcoma, Ewing's sarcoma, and synovial sarcoma. MET-based molecular typing is of great significance for diagnosing, recognizing, and treating morphologically overlapping solid tumors.

[Taskesen E et al.](#) integrated the gene expression and DNA MET spectra of 344 samples of acute myeloid leukemia (AML) and established a regression model using Lasso. The results indicated that the subtype prediction of AML cytogenetics and molecular abnormalities could be significantly improved³².

A study by [Rodríguez-Rodero et al.](#), demonstrated that thyroid carcinoma subtypes have promoter-differentiated MET features, and the molecular typing could be realized using abnormal DNA MET expression. Undifferentiated thyroid carcinoma was characterized by abnormal promoter hypomethylation, while differentiated papillary and follicular thyroid carcinoma was characterized by promoter hypermethylation³³.

To further explore the functions of the 250 screened MET loci, gene function annotation of the loci was performed and 42 genes were found to be significantly enriched to TFEC. The *TFEC* gene is located at 7q31.2 and encodes a polypeptide with a length of 347 amino acids, which is mainly localized in the nucleus and cytoplasm. According to a study by [Chung et al.](#), TFEC plays a role as an activating transcription factor (ATF) for the non-myosin heavy chain II-a gene³⁴. At present, evidence for the involvement of TFEC in cancer progression is limited; however, TFEC, MITF, TFEB, and TFE3 are important members of the MIT (microphthalmia-associated transcription factor) family, and recent studies have proven that changes in these transcription factors are related to melanoma, sarcoma, and renal cell carcinoma. With a similar structure to TFEB (another member of the MIT family), TFEC may play an important role in regulating genes related to autophagy and lysosomes³⁵.

The regulation of genes is a complex network; to investigate the effects of TFEC and its relevant factors on the occurrence and progression of tumors, function enrichment analysis was performed and these genes were found to be remarkably enriched to the following biological functions: GO:0006955 – immune response, GO:0050776 – regulation of the immune response, GO:0006954 – inflammatory response, GO:0045087 – innate immune response, and GO:0007165 – signal transduction. Currently, there are no reports of TFEC in ovarian cancer; thus, further investigation is needed.

Finally, 5 CpG loci were screened via the WGCNA co-expression network: cg27625732, cg00431050, cg22197830, cg03152385, and cg22809047. The results show that hypomethylation of these 5 CpG loci was associated with poor prognosis in ovarian cancer patients. The gene annotated by the cg22809047

locus was RPL31, and Maruyama et al.³⁶ have previously shown that in comparison with benign prostate tissues, RPL31 is overexpressed in prostate cancer. In RPL31 siRNA-treated LNCaP and BicR cells, there is an increase in the protein expression levels of the tumor suppressor p53 and its targets, p21 and MDM2. In addition, the inhibition of cell growth and the cell cycle by RPL31 could be recovered by p53 siRNA treatment. RPL31 could be used as the target of molecular treatment for advanced prostate cancer, and we presume that RPL31 could also be used as a target for the treatment of ovarian cancer. ELOVL3 was the gene corresponding to the gene promoter region annotated by the cg00431050 locus. ELOVL3 is a member of the ELOVL (elongase of very long-chain fatty acids) family, which contains a total of 7 members (ELOVL1–7). The proteins encoded by the *ELOVL1–7* genes are involved in the elongation of fatty acid chains of different lengths, and play an important role in regulating the biological synthesis of lipids, fatty acid metabolism, and certain metabolic diseases. There exist only limited studies of the involvement of ELOVL3 in tumors, while ELOVL2 has been widely described in tumors. A study by Kang et al., revealed that breast cancer patients with low ELOVL2 expression have a poor prognosis. ELOVL2 expression has been correlated with the malignant phenotype of breast cancer, and its downregulation induced lipid metabolism reprogramming; thus, ELOVL2 is a novel prognostic biomarker³⁷. We suggest that ELOVL3 expression may also be involved in the occurrence and progression of ovarian cancer by inducing lipid metabolism reprogramming.

Zhang et al.,³⁸ investigated the molecular typing of serous ovarian cancer using the multi-omics data of DNA MET and protein, miRNA, and gene expression, mainly discussing the relationship between molecular typing based on RNA-Seq data and that based on other omics data. They finally screened 9 molecular subtypes based on RNA-Seq data; these subtypes had significant overlap with the molecular subtypes of other omics, but the function analysis results showed that the subtypes based on an omics dataset could not be completely substituted by other omics data.

In the present study, the significance of MET in the molecular typing of ovarian cancer was analyzed using MET data, and the markers of subtypes closely related to the prognosis prediction of ovarian cancer were further screened. A MET data-based ovarian cancer prognosis prediction model was subsequently developed to provide a reference for clinical trials and researchers. In summary, the study by Zhang et al., and our study have different focal points, despite both involving molecular typing.

Subtyping of ovarian carcinomas based on methylation profiles has been reported in a TCGA seminal article¹³, in which 4 subtypes were identified to be significantly associated with differences in age, BRCA inactivation events, and survival based on consensus clustering of variable DNA methylation. The cluster associated with the worst prognosis is characterized by hypomethylation and is associated with old age, which is in accordance with the present findings; however, our approach is different from that in the aforementioned TCGA paper.

Firstly, the samples included in the TCGA paper were 489 cases of high-grade serous ovarian cancer, while the present paper included 571 cases of methylated ovarian cancer, including different clinical stages and grades. Our sample size is larger, and the results are more abundant. Secondly, a multivariate

COX proportional hazards model was performed to elucidate that 250 CpG loci were significant predictors of prognosis, and 6 molecular subtypes were clustered based on the methylation level at these 250 CpG loci. The cluster that was characterized by hypomethylation was associated with a worse prognosis, stage, and grade, and an older patient age. Thirdly, weighted gene co-expression network analysis was further applied to identify the 5 most significant CpG loci, and hypomethylation of these 5 loci was demonstrated to be associated with a worse outcome.

Conclusion

We identified 6 different molecular subtypes using ovarian cancer MET data in the TCGA database and showed that DNA MET molecular typing could accurately support the distinction of tumor stage and pathological grade in ovarian cancer. The specific CpG loci and genes can be used as biomarkers for individualized treatment, which can be carried out in clinical practice according to these prognostic markers. Nevertheless, the present study has certain limitations: only internal validation was performed on MET prognostic loci and no suitable external datasets were obtained; thus, our study results need to be further validated with a larger sample size.

Abbreviations

TCGA: the Cancer Genome Atlas

OS: overall survival

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data during this study are included within this published article and additional files. Any material described in the article can be requested directly from corresponding author on reasonable request. All data in the current study are based on public data available in The Cancer Genome Atlas (TCGA) datasets.

Declaration of Interests

The authors declare no competing financial interests

Non-Financial Conflict of Interest

The authors declare no non-financial conflict of interest.

Funding Sources

No

Author Contributions

Qing Yang designed and conceptualized this study; Lili Yin and Ningning Zhang were major contributors in experiment; Lili Yin analyzed the data; Lili Yin, and Qing Yang were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2006. *CA: a cancer journal for clinicians*. 2006;56(2):106-130.
2. Herzog TJ. Recurrent ovarian cancer: how important is it to treat to disease progression? *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2004;10(22):7439-7449.
3. Pearce CL, Templeman C, Rossing MA, et al. Association between endometriosis and risk of histological subtypes of ovarian cancer: a pooled analysis of case-control studies. *The Lancet Oncology*. 2012;13(4):385-394.
4. Sung P-L, Chang Y-H, Chao K-C, Chuang C-M. Global distribution pattern of histological subtypes of epithelial ovarian cancer: a database analysis and systematic review. *Gynecologic oncology*. 2014;133(2):147-154.
5. Böcker W. [WHO classification of breast tumors and tumors of the female genital organs: pathology and genetics]. *Verh Dtsch Ges Pathol*. 2002;86:116-119.
6. Kommoss S, Schmidt D, Kommoss F, et al. Histological grading in a large series of advanced stage ovarian carcinomas by three widely used grading systems: consistent lack of prognostic significance. A translational research subprotocol of a prospective randomized phase III study (AGO-OVAR 3 protocol). *Virchows Archiv : an international journal of pathology*. 2009;454(3):249-256.

7. Shih I-M, Kurman RJ. Ovarian tumorigenesis: a proposed model based on morphological and molecular genetic analysis. *The American journal of pathology*. 2004;164(5):1511-1518.
8. Miller JA, Kotecha R, Ahluwalia MS, et al. Overall survival and the response to radiotherapy among molecular subtypes of breast cancer brain metastases treated with targeted therapies. *Cancer*. 2017;123(12):2283-2293.
9. Roode SC, Rotroff D, Avery AC, et al. Genome-wide assessment of recurrent genomic imbalances in canine leukemia identifies evolutionarily conserved regions for subtype differentiation. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*. 2015;23(4):681-708.
10. Dolly SO, Collins DC, Sundar R, Popat S, Yap TA. Advances in the Development of Molecularly Targeted Agents in Non-Small-Cell Lung Cancer. *Drugs*. 2017;77(8):813-827.
11. Xu Y, Chen M, Liu C, et al. Association Study Confirmed Three Breast Cancer-Specific Molecular Subtype-Associated Susceptibility Loci in Chinese Han Women. *The oncologist*. 2017;22(8):890-894.
12. Tothill RW, Tinker AV, George J, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2008;14(16):5198-5208.
13. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;474(7353):609-615.
14. Kommoss S, Winterhoff B, Oberg AL, et al. Bevacizumab May Differentially Improve Ovarian Cancer Outcome in Patients with Proliferative and Mesenchymal Molecular Subtypes. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2017;23(14):3794-3801.
15. Ushijima T, Asada K. Aberrant DNA methylation in contrast with mutations. *Cancer science*. 2010;101(2):300-305.
16. Wilting RH, Dannenberg J-H. Epigenetic mechanisms in tumorigenesis, tumor cell heterogeneity and drug resistance. *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy*. 2012;15(1-2):21-38.
17. Esteller M, Silva JM, Dominguez G, et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *Journal of the National Cancer Institute*. 2000;92(7):564-569.
18. Baylin SB, Ohm JE. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? *Nature reviews Cancer*. 2006;6(2):107-116.
19. Esteller M. Epigenetics in cancer. *The New England journal of medicine*. 2008;358(11):1148-1159.
20. Baylin SB, Chen WY. Aberrant gene silencing in tumor progression: implications for control of cancer. *Cold Spring Harb Symp Quant Biol*. 2005;70:427-433.
21. Petrocca F, Iliopoulos D, Qin HR, et al. Alterations of the tumor suppressor gene ARLTS1 in ovarian cancer. *Cancer research*. 2006;66(21):10287-10291.
22. Horak P, Pils D, Haller G, et al. Contribution of epigenetic silencing of tumor necrosis factor-related apoptosis inducing ligand receptor 1 (DR4) to TRAIL resistance and ovarian cancer. *Molecular cancer*

research : *MCR*. 2005;3(6):335-343.

23. Hu W-L, Zhou X-H. Identification of prognostic signature in cancer based on DNA methylation interaction network. *BMC medical genomics*. 2017;10(Suppl 4):63.
24. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;19(1A):A68-A77.
25. Zhang S, Wang Y, Gu Y, et al. Specific breast cancer prognosis-subtype distinctions based on DNA methylation patterns. *Molecular oncology*. 2018;12(7):1047-1060.
26. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics (Oxford, England)*. 2010;26(12):1572-1573.
27. Reimand J, Arak T, Adler P, et al. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic acids research*. 2016;44(W1):W83-W89.
28. Aran D, Hellman A. DNA methylation of transcriptional enhancers and cancer predisposition. *Cell*. 2013;154(1):11-13.
29. Jurmeister P, Schöler A, Arnold A, et al. DNA methylation profiling reliably distinguishes pulmonary enteric adenocarcinoma from metastatic colorectal cancer. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*. 2019;32(6):855-865.
30. Williams LA, Mills L, Hooten AJ, et al. Differences in DNA methylation profiles by histologic subtype of paediatric germ cell tumours: a report from the Children's Oncology Group. *British journal of cancer*. 2018;119(7):864-872.
31. Wu SP, Cooper BT, Bu F, et al. DNA Methylation-Based Classifier for Accurate Molecular Diagnosis of Bone Sarcomas. *JCO Precis Oncol*. 2017;2017.
32. Taskesen E, Babaei S, Reinders MMJ, de Ridder J. Integration of gene expression and DNA-methylation profiles improves molecular subtype classification in acute myeloid leukemia. *BMC bioinformatics*. 2015;16 Suppl 4:S5.
33. Rodríguez-Rodero S, Fernández AF, Fernández-Morera JL, et al. DNA methylation signatures identify biologically distinct thyroid cancer subtypes. *J Clin Endocrinol Metab*. 2013;98(7):2811-2821.
34. Chung MC, Kim HK, Kawamoto S. TFEC can function as a transcriptional activator of the nonmuscle myosin II heavy chain-A gene in transfected cells. *Biochemistry*. 2001;40(30):8887-8897.
35. Haq R, Fisher DE. Biology and clinical relevance of the microphthalmia family of transcription factors in human cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2011;29(25):3474-3482.
36. Maruyama Y, Miyazaki T, Ikeda K, et al. Short hairpin RNA library-based functional screening identified ribosomal protein L31 that modulates prostate cancer cell growth via p53 pathway. *PloS one*. 2014;9(10):e108743.
37. Kang YP, Yoon J-H, Long NP, et al. Spheroid-Induced Epithelial-Mesenchymal Transition Provokes Global Alterations of Breast Cancer Lipidome: A Multi-Layered Omics Analysis. *Frontiers in oncology*. 2019;9:145.

38. Zhang Z, Huang K, Gu C, et al. Molecular Subtyping of Serous Ovarian Cancer Based on Multi-omics Data. *Scientific reports*. 2016;6:26001.

Tables

Table 1 The clinical pathological information in the training set and validation test.

	Validation set	Traning set
Stage		
Stage I	5	11
Stage II	15	12
Stage III	222	215
Stage IV	41	45
Grade		
G1	3	3
G2	38	31
G3	238	243
G4	0	1
Age		
<=60	171	147
>60	115	138

Table 2: Top 20 methylation loci

CpGs	p.value	HR	Low 95%CI	High 95%CI
cg25781123	1.42E-05	144.7952	15.31263	1369.173
cg01278291	1.68E-05	0.013765	0.001955	0.096903
cg21291896	2.93E-05	1.24E+13	8962882	1.72E+19
cg08946332	5.46E-05	0.21193	0.099747	0.450282
cg13804316	8.51E-05	213979.4	469.6279	97496703
cg16179125	9.40E-05	8.255374	2.862145	23.81123
cg13060646	0.000201	4.08882	1.945923	8.591523
cg03750606	0.000282	31.81681	4.916032	205.92
cg15341340	0.000317	3559149	966.7441	1.31E+10
cg08013810	0.00033	0.087427	0.023115	0.330674
cg06797533	0.000383	4.351053	1.932616	9.795869
cg21022435	0.000396	0.00496	0.000263	0.093416
cg22916109	0.000475	1.07E+11	69831.7	1.65E+17
cg10415235	0.00048	6.56E+08	7370.126	5.83E+13
cg05955301	0.000514	8.98125	2.60262	30.99294
cg16016036	0.00056	0.243586	0.109202	0.543343
cg23486067	0.000569	20.19834	3.654797	111.6267
cg25634666	0.00057	0.255673	0.117688	0.555439
cg17332016	0.000602	104162.1	141.5735	76636837
cg03190825	0.000654	15.75275	3.227124	76.89485

Table 2. CpG loci in modules

Module	Count
black	12
blue	34
brown	22
green	16
red	14
turquoise	49
yellow	22

Table 3. The annotation of 5 CpG loci

CpG	Chrom	Start	End	GeneSymbol	Feature_Type
cg03152385	chr16	15094739	15094740	RP11-72I8.1	S_Shore
cg27625732	chr9	1.29E+08	1.29E+08	TBC1D13	N_Shore
cg22197830	chr5	1.35E+08	1.35E+08	TXNDC15	N_Shore
cg22809047	chr2	1.01E+08	1.01E+08	AC016738.4	Island
cg03152385	chr16	15094739	15094740	RRN3	S_Shore
cg00431050	chr10	1.02E+08	1.02E+08	ELOVL3	N_Shore
cg22809047	chr2	1.01E+08	1.01E+08	RPL31	Island

Additional File Legends

Table S1: A total of 967 loci demonstrated a significant difference in prognosis

Table S2: 250 significant MET loci

Table S3: Distribution of methylation levels in each subtype

Table S4: The distribution of grade and stage in each subtype

Table S5: Age distribution of samples in each subtype

Table S6: The R package covered in this article

Figures

Figure 1

A: CDF curve; different colors reflect different cluster numbers; the horizontal axis represents the consensus index; the vertical axis represents the cumulative distribution function (CDF); and a larger AUC indicates better clustering. B: CDF delta area curve of consensus clustering indicating the relative change in the area under the CDF curve for each category number k as compared with k - 1; the horizontal axis represents the category number k and the vertical axis represents the relative change in the area under the CDF curve.

Figure 2

A: Clustering heat map of samples at consensus $k = 6$. Different colors reflect different cluster numbers; the color gradient is from white to blue, indicating the consensus of progression. B: Clustering results of 250 MET loci, clinical stage, and histological grade in 6 clusters of samples; red represents high expression, blue represents low expression.

Figure 3

A: Prognosis differences among the 6 subtypes of samples; different colors represent different molecular subtypes; the horizontal axis reflects the survival time and the vertical axis represents the survival rate. B: Percentages of samples of different clinical stages in the 6 subtypes. C. Percentages of samples of different grades in the 6 subtypes; the horizontal axis represents different molecular subtypes and the vertical axis represents the percentage. D. Age distribution of the patients of the samples in the 6 subtypes; the horizontal axis represents different molecular subtypes and the vertical axis represents age.

Figure 4

A: Transcription factor enrichment results of genes corresponding to the gene promoter regions annotated by 250 CpG loci; green represents the transcription factors, pink represents the genes annotated by MET loci, and purple represents MET loci. B: Chord diagram showing the top 5 enriched GO clusters for genes associated with TFEF. In each chord diagram, enriched GO clusters are shown on the right and genes contributing to enrichment are shown on the left. Positively correlated molecules are

displayed in red and negatively correlated molecules are displayed in blue. Each GO term is represented by one colored line.

Figure 5

A: Evaluation of the scale-free model at different soft thresholds; a larger value indicates better compliance with the features of the biological network. B: Mean connectivity at different soft thresholds; the horizontal axis represents the soft threshold and the vertical axis represents the mean connectivity. C: Gene dendrogram and module colors; different colors represent the genes in different modules. D: Module-feature correlation; the row represents the eigengenes of each module and the column represents the feature information of the samples. Red to green represents a high to low correlation coefficient. The digit in each grid indicates the correlation coefficient between gene modules and the corresponding features, and the digit in the bracket represents the P value.

Figure 6

A: Weighted interaction network of 22 CpG loci in the yellow module; the circle represents CPG loci, the connection line between two circles represents the interaction relationship, and a redder color indicates a larger node degree. B: Correlation coefficient clustering of 22 CpG loci; a more purple color indicates a smaller correlation coefficient and a redder color indicates a larger correlation coefficient.

Figure 7

A: MET spectrum clustering results of the 5 CpG loci. B: Prognosis difference between the hypermethylation and hypomethylation groups formed by clustering. The horizontal axis represents the survival time (months), the vertical axis represents the survival rate, red indicates the hypomethylation group, and blue indicates the hypermethylation group.

Figure 8

A: Expression spectrum clustering results of the 5 CpG loci. B: Prognosis difference between the hypermethylation and hypomethylation groups formed by clustering. The horizontal axis represents the survival time (months), the vertical axis represents the survival rate, red indicates the hypomethylation group, and blue indicates the hypermethylation group.

Figure 9

Flow chart of all the analysis

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS2.docx](#)
- [TableS1.docx](#)
- [TableS6.docx](#)
- [TableS4.docx](#)
- [TableS5.docx](#)
- [TableS3.docx](#)