A qualitative transcriptional signature for predicting CpG island methylator phenotype status of the right-sided colon cancer

Tianyi You
Harbin Medical University
wenyuan Zhao (zhaowenyuan@ems.hrbmu.edu.cn)
Harbin Medical University https://orcid.org/0000-0002-6477-9434
Zheng Guo
Harbin Medical University
Kai Song
Harbin Medical University
Wenbing Guo
Harbin Medical University
Yelin Fu
Harbin Medical University
Kai Wang
Harbin Medical University
Hailong Zheng
Harbin Medical University
Jing Yang
Harbin Medical University
Liangliang Jin
Harbin Medical University
Lishuang Qi
Harbin Medical University

Research article

Keywords: Right-sided colon cancer; CpG island methylator phenotype; The qualitative transcriptional signature; Relative expression ordering; Gene pairs

Posted Date: November 21st, 2019

DOI: https://doi.org/10.21203/rs.2.17598/v1

Version of Record: A version of this preprint was published at Frontiers in Genetics on October 29th, 2020.
See the published version at https://doi.org/10.3389/fgene.2020.00971.

Abstract

Background: A part of colorectal cancer which is characterized by simultaneous numerous hypermethylation CpG islands sites is defined as CpG island methylator phenotype (CIMP) status. Stage II and III CIMP-positive (CIMP+) right-sided colon cancer (RCC) patients have a better prognosis than CIMP-negative (CIMP-) RCC treated with surgery alone. However, there is no gold standard available in defining CIMP status. In this work, we developed a transcriptional qualitative signature to individually predict CIMP status for stage II and III RCC. Results: Based on the relative expression orderings (REOs) of gene pairs, a signature composed of 19 gene pairs was developed to predict the CIMP status of RCC through a feature selection process. A sample is predicted as CIMP+ when the gene expression orderings of at least 12 gene pairs vote for CIMP+; otherwise the CIMP-. The difference of prognosis between the predicted CIMP+ and CIMP- groups was more significantly different than the original CIMP status groups. There were more differential methylation and expression characteristics between the two predicted groups. The hierarchical clustering analysis showed that the signature could perform better for predicting CIMP status of RCC than current methods. Conclusions: The qualitative transcriptional signature for classifying CIMP status at the individualized level can predict outcome and guide therapy for RCC patients.

Background

Colorectal cancer (CRC) is the third most commonly diagnosed malignancy and the second leading cause of mortality in the world [1]. The CpG island methylator phenotype-positive (CIMP+) tumor, which is characterized by vast hypermethylation of promoter CpG island sites, accounts for 17%-20% of CRC [2, 3]. Several studies indicated that the stage II and III CRC patients with CIMP+ status are associated with a better prognosis than CIMP-negative (CIMP-) CRC patients, and CIMP+ CRC patients cannot benefit from 5-Fluorouracil (5-FU)-based adjuvant chemotherapy (ACT) [4, 5].

Currently, the CIMP status is commonly detected by methylation-specific polymerase chain reaction (PCR) and methylight techniques. The methylation-specific PCR detects five biomarkers with MINT1, MINT2, MINT31, CDKN2A (p16) and MLH1 [6], and the methylight detects five biomarkers with CACNA1G, IGF2, NEUROG1, RUNX3 and SOCS1[7]. For each panel of CIMP markers, CRC is classified as CIMP+ if three or more CIMP markers are methylated which are also called as CIMP-high (CIMP-H). Besides, the others are classified as CIMP- which are also divided into CIMP-low (CIMP-L) if one or two CIMP markers are methylated and CIMP-0 if none of methylated marker is observed [4, 8]. Because CIMP-L patients have the same prognosis as CIMP-0 patients, and CIMP-L or CIMP-0 patients can benefit from 5-FU-based ACT [9], it is reasonable to group CIMP-L and CIMP-0 as CIMP- in our study. It is worth noting that the technologies commonly used could cause false-positive and false-negative results. The false-positive results arise from the incomplete bisulfite conversion, false priming, and the too low annealing temperature or too many used cycles [10]. The false-negative results are caused by the insufficient amount of input DNA, DNA degradation during bisulfite treatment, low stability of single-strand DNA, and strand-specific PCR amplification [11, 12]. Currently, there is no golden standard with respect to technologies and CIMP

markers for the detection of altered DNA methylation used to define CIMP status [12-14]. Therefore, it is worthwhile to develop a credible signature for predicting CIMP status.

Nowadays, because of the cost-effective of transcriptome analysis and the regulatory relationships between the DNA methylation and gene expression, several quantitative transcriptional signatures have been developed for predicting the CIMP status of CRC patients [15-17]. The quantitative signatures are sensitive to the systematic inter-laboratory biases of microarray or RNA-sequencing experiments, especially batch effects, which are introduced by experimental conditions, regent dosages, microarray technology, and operational procedures [18, 19], resulting in the failures in independent inter-laboratory data. In addition, the quantitative signatures would also be greatly affected by varied proportions of tumor epithelial cell in tumor tissues sampled from different tumor locations of the same patient [20], partial RNA degradation during specimen storage and preparation [21], and amplification bias for minimum specimens even with about 15–25 cancer cells [22], which are common factors that can lead to failures in clinical applications. In contrast, the qualitative signatures based on relative expression orderings (REOs) of gene pairs within a sample are robust against the batch effects, different tumor locations, partial RNA degradation, and amplification bias [23, 24], which could be directly applied to the sample at the individual level in clinical applications [19-22, 25].

Consistent with the differences in anatomy location, the left-sided colon cancer (LCC) and right-sided colon cancer (RCC) have different embryonic developmental sites, genomic patterns and different clinical symptoms [26-28]. Additionally, among the CIMP+ CRC, RCC has a significantly higher prevalence (87%) than LCC (13%) [29]. Thus, in this study, we developed a qualitative transcription signature for predicting CIMP status of stage II and III RCC at the individual levels. The performance of the signature was evaluated in four independent datasets by receiver operating characteristic (ROC) analysis. Meanwhile, based on the patients' relapse-free survival (RFS), we provided evidence that the signature could perform better for identifying CIMP status of RCC than current methods.

Results

Identification of the predictive signature for CIMP status of RCC

Fig. 1 describes the flowchart of this study. The GSE39582 dataset including the largest sample size of stage II and III RCC with CIMP status was used as the training data for selecting an REOs-based signature. Firstly, we identified 2209 DE genes between the 64 CIMP+ RCC samples and the 117 CIMP- RCC samples (limma test, FDR < 0.01). From all gene pairs consisted of at least one DE gene, we extracted 383,591 CIMP-related gene pairs whose specific REOs patterns occurred more frequently in the CIMP+ than in the CIMP- samples (Fisher's exact test, FDR < 0.01). Then, 53 panels of gene pairs were got within different range of FD value. After a redundancy removal process for each panel of gene pairs, we calculated the largest F-score with the optimal vote rule (Fig. 2a, see Methods). Finally, the 19 gene pairs, which obtained the largest F-score within the range of FD more than 0.58, were denoted as 19 gene pairs signatures (19-GPS) for predicting CIMP status of stage II and III RCC (Fig. 2b).

A sample was predicted as CIMP+ if the REOs of at least 12 gene pairs in 19-GPS voted for CIMP+; otherwise the CIMP-. According to the classification rule, the F-score of the signature in the training data was 0.91 (Table 1) with a sensitivity of 0.91 and a specificity of 0.90, and the AUC of ROC curve was 0.95 (95% CI: 92.08%-97.83%) (Fig. 3a).

Table 1. The performance of 19-GPS for right-sided colon cancer (RCC) samples in the training and validation datasets

	pre-CIMP+ (CIMP+:CIMP-)	pre-CIMP- (CIMP+:CIMP-)	sensitivity	specificity	F-score
GSE39582	70(58:12)	111(6:105)	0.91	0.90	0.95
GSE39084	11(6:5)	8(0:8)	1	0.62	0.76
GSE25070	6(5:1)	7(1:6)	0.83	0.86	0.85
E-TABM-328	13(10:3)	9(8:1)	0.91	0.73	0.81
Total RCC	100(79:21)	135(15:120)	0.84	0.85	0.85

Note: The CIMP+ and CIMP- represented the original CIMP status; pre-CIMP+ and pre-CIMP- represented the CIMP status predicted by 19-GPS.

Based on the knowledge that stage II and III CIMP+ RCC patients treated with surgery alone have better prognoses than CIMP- RCC patients [4, 5], we evaluated the reliability of 19-GPS through survival analysis. In the training dataset containing 31 samples of stage III RCC patients treated with surgery alone, one of the 16 original CIMP- samples was reclassified as CIMP+ by 19-GPS (Supplementary Table 1, Additional File 1). The survival analysis showed that the RFS of the 16 predicted CIMP+ patients was significantly longer than the 15 predicted CIMP- patients (log-rank P = 4.90e-3, HR = 0.14, 95% CI = 0.03-0.68, Fig. 4a), which was more significant than the difference between patients with the original CIMP status due to the reclassified sample (log-rank P = 5.24e-3, HR = 0.15, 95% CI = 0.03-0.69, Fig. 4b). It is also known that stage III CIMP- RCC patients treated with 5-Fu-based ACT have better outcomes than patients treated with surgery alone [4]. In the 41 stage III RCC samples of training data for patients receiving 5-Fu-based ACT, 2 of the 29 original CIMP- samples were reclassified as CIMP+ by 19-GPS, and 2 of the 12 original CIMP+ samples were reclassified as CIMP- (Supplementary Table 1, Additional File 1). The survival analysis showed that the RFS of the 29 predicted CIMP- patients receiving 5-Fu-based ACT was significantly longer than the 15 predicted CIMP- patients treated with surgery alone (log-rank P = 5.97e-3, HR = 0.27, 95% CI = 0.10~0.73, Fig. 4c), which was more significant than the different between original CIMP- patients treated with 5-FU-based ACT and surgery alone (log-rank P = 1.69e-2, HR = 0.33,

95% CI = 0.13~0.85, Fig. 4d). The survival analysis validated that 19-GPS could perform better for predicting CIMP status of stage II and III RCC patients than current methods.

There were 12 CIMP- and 6 CIMP+ samples reclassified by 19-GPS in the total of stage II and III RCC of training dataset. We contrasted the gene expression patterns of the 18 signature-disconfirmed samples with the 163 signature-confirmed samples through hierarchical clustering analysis. Firstly, we identified 4685 DE genes between the 58 signature-confirmed CIMP+ samples and the 105 signature-confirmed CIMP- samples in the training dataset (limma test, FDR < 0.01). Secondly, using the expression measurements of the top 100 significant DE genes, the samples were classified into two subgroups using the complete linkage hierarchical clustering analysis based on the Euclidean distance (Fig. 5a). The results showed that all of the samples reclassified as CIMP+ and CIMP- were clustered with the group of signature-confirmed CIMP+ and CIMP- samples, respectively. The gene expression patterns validated the correctness of 19-GPS in training dataset.

Validation of 19-GPS in independent datasets

In three validation datasets (GSE39084, GSE25070 and E-TABM-328) of RCC samples, the CIMP status of samples was predicted based on 19-GPS. In GSE25070, *TMEM150C* and *CCDC170* included in 19-GPS were not detected by Illumina Human Ref-8v3.0 expression beadchip, which resulted in 17 gene pairs available for classification. Then we observed that the classifier of 17 gene pairs achieved the largest F-score when requiring that at least 10 of 17 gene pairs voted for CIMP+ determination in the training dataset, so the vote rule was regarded as the optimal vote rule in GSE25070. Similarly, in E-TABM-328, 18 gene pairs were detected by Whole Human Genome Microarray 4x44K, and CIMP+ determination could be voted by at least 11 of 18 gene pairs as the optimal vote rule. The F-score of the signature were 0.76, 0.85 and 0.81 in GSE39084, GSE25070 and E-TABM-328. The AUC of ROC were 97.44% (95% CI: 91.37%-100%), 91.67% (95% CI: 61. 68%-100%) and 82.23% (95% CI: 70.59%-100%) (Fig. 3b, c, d).

Because the therapeutic and survival information was unavailable in three validation datasets, we compared the gene expression patterns of the signature-disconfirmed samples with the signature-confirmed samples through hierarchical clustering analysis in the validation datasets. Using the expression levels of the top 100 significant DE genes between the signature-confirmed CIMP+ and CIMP- samples (limma test, FDR < 0.01), the samples were classified into two subgroups using the hierarchical clustering analysis (Fig. 5b, c, d). In GSE39084, the result showed 4 of 5 CIMP- samples reclassified as CIMP+ by our signature were clustered with the group of signature-confirmed CIMP+ samples. The similar results were observed in GSE25070 and E-TABM-328 that all of the samples reclassified as CIMP+ and CIMP- were clustered with the group of signature-confirmed CIMP+ and CIMP- samples, respectively. These results provided transcriptional evidence of the correctness of the prediction of 19-GPS.

The differentially methylated CpG sites and expressed genes between CIMP+ and CIMP- samples

The CIMP+ status is characterized by high frequency of promoter hypermethylation whose regions almost locate in tumor suppressor genes [30]. We used the datasets detected both gene expression and

DNA methylation profiles to select the differentially methylated CpG sites between predicted CIMP+ and CIMP- samples (match GSE25070 to GSE25062 and match GSE79793 to GSE79794). The CIMP status predicted by 19-GPS in GSE25070 was used in GSE25062. Then, the 1581 hypermethylated CpG sites were selected between the predicted CIMP+ and CIMP- samples in GSE25062 (limma test, $P < 0.05$, Fig. 6a). The hypermethylated CpG sites located in the regions of 26 tumor suppressor genes which were downloaded from The Cancer Gene Census containing 316 tumor suppressor genes (https://cancer.sanger.ac.uk/census). Meanwhile, the 1147 hypermethylated CpG sites were selected between original CIMP status samples in GSE25062, and they located in the regions of 15 tumor suppressor genes (limma test, $P < 0.05$, Fig. 6b). The results showed that the predicted CIMP+ samples had much more hypermethylated CpG sites and tumor suppressor genes than the original CIMP+ samples.

And then, we calculated the number of hypermethylated CpG sites and tumor suppressor genes of predicted CIMP+ samples based on the same method in GSE79793 and GSE79740. Compared with the predicted CIMP- samples, the predicted CIMP+ samples had 3124 hypermethylated CpG sites which located in the regions of 57 tumor suppressor genes, (limma test, $P < 0.05$, Fig. 6c). Because the samples in above datasets had no original CIMP labels, so we could not compare the difference of the number of hypermethylated CpG sites and tumor suppressor genes between the predicted and original CIMP status. Moreover, the 552 hypermethylated CpG sites between predicted CIMP+ and CIMP- samples were identified in both GSE25062 and GSE79740, which were not randomly distributed among all of the hypermethylated CpG sites ($P < 2.2e\text{-}16$, Hypergeometric test).

Besides, we selected 4771 DE genes between the predicted CIMP+ and CIMP- samples, which were more than 2209 DE genes among the original samples in the training dataset (limma test, FDR <0.05). This indicated that the differences in methylation and gene expression patterns between the predicted CIMP+ and CIMP- samples were more significant than the original samples. In conclusion, the differentially methylated CpG sites and expressed genes analysis provided the evidence that the characteristic of predicted CIMP status of samples conformed to the truly biological property.

The robustness against varied proportions of tumor epithelial cell

Some reports show the qualitative signatures based on REOs of gene pairs are robust against to varied proportions of tumor epithelial cell [20]. To validate the robustness of 19-GPS, our laboratory collected 13 fresh-frozen primary tumor tissue samples through surgical excision. Fresh-frozen primary tumor tissue samples were retrospectively collected at Union Hospital of Fujian Medical University. And the 13 solid tumor tissue samples were from 5 patients whose excisions were from different sampling positions with different information of ‘percentage of tumor cells’ as shown in Table 2. The institutional ethical review boards of Union Hospital of Fujian Medical University approved the protocol, and all patients signed informed consents before sample collection. And we used the fragments per kilobase of exon model per million mapped fragments to quantify the gene expression level from RNA sequencing data. Then, we used 16 gene pairs available for 19-GPS to predict the CIMP status of 13 samples. And the gene

expression levels of 19-GPS were detailed in (Supplementary Table 1, Additional File 2). There were 4 of 5 patients containing samples with different percentage of tumor cells predicted the same CIMP status, and 2 of 3 samples of the one remaining patient were also predicted the same CIMP status (Table 2). The results indicated that the performance of 19-GPS was not effected by varied proportions of tumor epithelial cell.

Table 2. The predicted CIMP status of samples with different percentage of tumor cells

Sample ID	Percentage of tumor cells（%）	Predicted CIMP status
HCF1	40	Negative
HCF2	100	Negative
HCF3	100	Negative
LGL1	50	Negative
LGL2	90	Positive
LGL3	90	Positive
SDL1	100	Negative
SDL2	100	Negative
WCY1	60	Negative
WCY2	100	Negative
WCY3	100	Negative
ZCH1	70	Negative
ZCH3	40	Negative

Discussion

In this study, we developed a robust qualitative transcriptional signature consisting of 19-GPS to individually identify the CIMP status for stage II and III RCC. We also tried to develop a signature to predict CIMP status for stage II and III LCC. However, the prevalence rate of CIMP+ among LCC was only 2.04-6.67% in the training and validation datasets (Supplementary Table 2, Additional File 1), and the statistics showed that the prevalence rate is about 2.67% in several studies [31]. There were so few LCC samples that we could not train or validate a signature to predict the CIMP status for LCC samples.

After identifying DE genes in training dataset, the functional enrichment analysis showed that the 4771 DE genes between the predicted CIMP+ and CIMP- samples were significantly enriched in 55 KEGG pathways (see Methods) (FDR < 0.05, hypergeometric distribution, Supplementary Table 2, Additional File 2). Especially, some cancer-associated pathways for metabolic pathway [32], cell cycle pathway [33], apoptosis pathway [34] were significantly enriched. Among the 55 significantly enriched pathways, mismatch repair pathway plays a critical role in maintaining the integrity and stability of the genome [35].

And the p53 signaling pathway can regulate angiogenesis and metastasis, which is closely related to the progression and outcome of colorectal cancer [36].

The association of CIMP status and the outcome was similar among stage II and III patients, but only stage III patients had a significant difference of survival analysis in the training dataset [5]. This may be due to the fact that the stage II patients had too much censored data to analyze in the training dataset. It is well known that the molecular marker consisting of CIMP and microsatellite instability (MSI) status can more accurately predict the outcome of CRC patients treated with surgery alone, compared with the molecular marker consisted of CIMP or MSI status alone [5, 37]. In the training dataset, we divided stage III RCC patients treated with surgery alone into four groups: CIMP+ with MSI-high (MSI-H) group, CIMP+ with microsatellite stability (MSS) group, CIMP- with MSI-H group and CIMP- with MSS group. We observed that the RFS of predicted CIMP+ with MSI-H group of patients treated with surgery alone was significantly longer than the others (log-rank P = 2.39e-2, Supplementary Fig. 1a, Additional File 3). After dividing samples into four categories, although the sample size was small in four groups, the survival difference between the predicted CIMP patients was more significant than original CIMP patients due to the one reclassified sample (log-rank P = 2.50e-2, Supplementary Fig. 1b, Additional File 3).

Our laboratory proposes the concept of “a sequence for all”, which composed by a series of qualitative transcriptional signatures for the prognostic and predictive biomarkers of CRC, including identifying micro-metastasis after surgery, 5-FU-based ACT benefit of high relapse risk patients, MSI status for CRC patients and so on [23, 24]. The qualitative transcriptional signature for predicting CIMP status in this study could combine with the other panels to predict the prognosis and guide the optimal therapy for CRC patients in clinical application.

Conclusions

In summary, the qualitative transcriptional signature could robustly predict CIMP status of stage II and III RCC at the individualized levels. The CIMP status predicted by 19-GPS can evaluate the outcome and guide the therapy for stage II and III RCC patients treated with surgery alone. The robustness and simplicity of the REO-based signature would make it convenient in clinical settings and worthy to further validate in a prospective clinical trial.

Methods

Data and preprocessing

The gene expression and methylation datasets for colon cancer used in this study were downloaded from the Gene Expression Omnibus database (GEO, http://www.ncbi.nlm.nih.gov/geo/) and the ArrayExpress database (https://www.ebi.ac.uk/arrayexpress/), as described in details in Table 3 and 4.

The training dataset for extracting a REOs-based signature was GSE39582, including 64 CIMP+ and 117 CIMP- stage II and III RCC samples, which recorded the information of relapse-free survival (RFS) of

patients for further survival analyses. Because of the small sample size of RCC in GSE39084, GSE25070 and E-TABM-328, so the three cohorts including a total of 54 RCC samples were combined as the validation cohort to test the predictive signatures. Besides, we used the samples which were detected both gene expression profiles and DNA methylation profiles (match GSE25070 to GSE25062 and match GSE79793 to GSE79794) to select the differentially methylated CpG sites between the CIMP+ and CIMP- samples predicted by the signature.

Table 3. The datasets detected CpG island methylator phenotype (CIMP) status in this study

	GSE39582 (n = 510)	GSE39084 (n = 19)	GSE25070 (n = 22)	E-TABM-328 (n = 47)
Stage				
I	37	-	-	-
II	247	20	-	-
III	167	14	-	-
IV	59	-	-	-
CIMP status				
CIMP+	93	6	6	11
CIMP-	417	13	16	36
Location				
Right	210	19	13	22
Left	300	-	9	25
CIMP Detection				
	Methylight	Methylight	Methylight	Methylation-specific PCR
Adjuvant chemotherapy				
Yes	296	-	-	-
No	201	-	-	-
NA	16	-	-	-
Platform	Affymetrix Human Genome U133 Plus 2.0 Array	Affymetrix Human Genome U133 Plus 2.0 Array	Illumina Human Ref-8v3.0 expression beadchip	Whole Human Genome Microarray 4x44K

Table 4. The datasets detected both gene expression and DNA methylation profiles in this study

	GSE25070 (n = 22)	GSE25062 (n = 22)	GSE79793 (n = 26)	GSE79740 (n = 26)
Data type	Expression profiling	Methylation profiling	Expression profiling	Methylation profiling
CIMP status				
CIMP+	6	6	-	-
CIMP-	16	16		
Platform	Illumina HumanRef-8 v3. 0 expression beadchip	Illumina HumanMethylation27 BeadChip	Illumina HumanHT-12 WG-DASL V4. 0 R2 expression beadchip	Illumina HumanMethylation450 BeadChip

For data measured by the Affymetrix platform, we downloaded the raw mRNA expression data (CEL files) and used the Robust Multi-Array Average algorithm [38] for background adjustment. For data measured by the Illumina and Agilent platform, we directly downloaded the processed data (series matrix files). For each gene expression database, the rule of processing all probes was following: the expression measurements of multiple probes mapping to the same Entrez Gene ID were averaged to obtain a single measurement, and the probes that did not map to any Entrez Gene ID or mapped to multiple Entrez Gene IDs were discarded. For the gene methylation datasets, we only analyzed the 25014 CpG sites detected by both the 27K array and 450K array which were not targeted the X and Y chromosomes. Using methylated signal intensity (M) and unmethylated signal intensity (U), the DNA methylation level of each probe was calculated by M/(U + M + 100) [39].

Differentially methylated CpG sites and expressed genes analysis

For microarray data, we selected differential methylated CpG sites or differentially expressed (DE) genes between two classes of samples using limma algorithm [40]. The P values were adjusted by the Benjamini-Hochberg procedure for multiple testing to control the false discovery rate (FDR) [41].

Signature development for predicting CIMP status of RCC

Firstly, for a gene pair, *i* and *j*, with expression values of E_i and E_j, we used Fisher's exact test [42] to evaluate whether the frequency of a specific REO pattern ($E_i > E_j$ or $E_i < E_j$) was significantly higher in the CIMP+ samples than the frequency in the CIMP- samples. The gene pairs which detected with FDR < 0.01 were defined as CIMP-related gene pairs.

Secondly, because some genes appeared in multiple CIMP-related gene pairs, we narrowed down the number of gene pairs via a redundancy removal method. For a gene that appeared in multiple gene pairs, we only kept the gene pair with the largest frequency difference (FD) value and discarded others. The FD was calculated for each gene pair by the following formula.

, c = 1, 2, the probabilities of observing $E_i > E_j$ in each group.

, the FD value of a gene pair (i, j).

The bigger FD value was, the more stable difference of REOs between two groups of samples was. After that, we obtained a panel of gene pairs with no less than a FD cutoff with 0.01 spacing distance from the maximum to minimum. Finally, we selected the optimal vote rule for each gene panel according to their harmonic mean value (F-score) of sensitivity and specificity in predicted CIMP+ and CIMP- groups. A sample was labeled as CIMP+ if the REOs of at least k gene pairs in the panel of gene pairs were consistent with the specific patterns ($E_i > E_j$) of the training samples, and vice versa. For each k ranging from 1 to the number of gene pairs in the panel of gene pairs, we could compute the corresponding F-score. The F-score was calculated by the following formula.

We selected the k which could reach the largest F-score as the optimal vote rule for each panel of gene pairs. Finally, we selected the panel of gene pairs which reached the largest F-score as the signature.

Sample clustering

Limma algorithm was performed to identify DE genes between the samples with predicted CIMP+ and CIMP- by the signature confirmed with the original CIMP status. Complete linkage hierarchical clustering analysis was performed to stratify RCC samples into two subgroups. The similarity of samples was evaluated by the Euclidean distance based on the expression measurements of DE genes.

Statistical analysis

The RFS is the period from the date of initial surgical resection until the date of the first occurrence of a new tumor event or the final documented data (censored). The Kaplan-Meier method and the log-rank test were used to evaluated the survival curve and compare the difference of survival curves, respectively [43]. Univariable Cox proportional hazards regression model calculated the Hazard Ratio (HR) and the 95% confidence interval (95% CI) [44]. The predictive performance of the signature was calculated by using area under the curve (AUC) of the ROC curve analysis [45]. The functional categories for enrichment analysis were downloaded from KEGG [46]. The hypergeometric distribution model was used to test whether a set of genes observed in a functional term was significantly more than what expected by

random chance. All statistical analyses were performed using the R 3.5.2 software package (http://www.r-project.org/).

Abbreviations

CIMP: CpG island methylator phenotype; CIMP+: CIMP-positive; CIMP-: CIMP-negative; CRC: colorectal cancer; RCC: right-sided colon cancer; LCC: left-sided colon cancer; REOs: relative expression orderings; 19-GPS: 19 gene pairs signatures; F-score: harmonic mean value; FD: frequency difference; 5-FU: 5-Fluorouracil; ACT: adjuvant chemotherapy; PCR: methylation-specific polymerase chain reaction; CIMP-H: CIMP-high; CIMP-L: CIMP-low; ROC: receiver operating characteristic; AUC: area under the curve; RFS: relapse-free survival; HR: hazard ratio; MSI: microsatellite instability; MSI-H: MSI-high; MSS: microsatellite stability; GEO: Gene Expression Omnibus; DE: differentially expressed; FDR: false discovery rate.

Declarations

Ethics approval and consent to participate

The Institutional Ethical Review Boards of Union Hospital of Fujian Medical University approved the protocol, and all patients signed informed consents before sample collection.

Consent for publication

Not applicable.

Availability of data and materials

All training and validation datasets analyzed in this study were downloaded from the public database: GEO and Arrayexpress. The data analyzed during the analysis of robustness against varied proportions of tumor epithelial cell are included in Supplementary Table 1, Additional File 2.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers: 61601151, 81572935, 81872396, 61673143 and 61701143].

Authors' contributions

WZ and ZG conceived the idea, TY conceived and designed the experiments, wrote the manuscript, KS and QL designed the experiments, WG and FY analyzed the data, WK and ZH performed the experiments, YJ and JL helped in writing the manuscript. All authors approved the final version.

Acknowledgements

Not applicable.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: **Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. *CA Cancer J Clin* 2018, **68**(6):394-424.
2. Kudryavtseva AV, Lipatova AV, Zaretsky AR, Moskalev AA, Fedorova MS, Rasskazova AS, Shibukhova GA, Snezhkina AV, Kaprin AD, Alekseev BY *et al*: **Important molecular genetic markers of colorectal cancer**. *Oncotarget* 2016, **7**(33):53959-53983.
3. Jass JR: **Serrated adenoma of the colorectum and the DNA-methylator phenotype**. *Nat Clin Pract Oncol* 2005, **2**(8):398-405.
4. Jover R, Nguyen TP, Perez-Carbonell L, Zapater P, Paya A, Alenda C, Rojas E, Cubiella J, Balaguer F, Morillas JD *et al*: **5-Fluorouracil adjuvant chemotherapy does not increase survival in patients with CpG island methylator phenotype colorectal cancer**. *Gastroenterology* 2011, **140**(4):1174-1181.
5. Ogino S, Nosho K, Kirkner GJ, Kawasaki T, Meyerhardt JA, Loda M, Giovannucci EL, Fuchs CS: **CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer**. *Gut* 2009, **58**(1):90-96.
6. Issa JP: **CpG island methylator phenotype in cancer**. *Nat Rev Cancer* 2004, **4**(12):988-993.
7. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, Kang GH, Widschwendter M, Weener D, Buchanan D *et al*: **CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer**. *Nat Genet* 2006, **38**(7):787-793.
8. Min BH, Bae JM, Lee EJ, Yu HS, Kim YH, Chang DK, Kim HC, Park CK, Lee SH, Kim KM *et al*: **The CpG island methylator phenotype may confer a survival benefit in patients with stage II or III colorectal carcinomas receiving fluoropyrimidine-based adjuvant chemotherapy**. *BMC Cancer* 2011, **11**:344.
9. Juo YY, Johnston FM, Zhang DY, Juo HH, Wang H, Pappou EP, Yu T, Easwaran H, Baylin S, van Engeland M *et al*: **Prognostic value of CpG island methylator phenotype among colorectal cancer patients: a systematic review and meta-analysis**. *Ann Oncol* 2014, **25**(12):2314-2327.
10. Kristensen LS, Mikeska T, Krypuy M, Dobrovic A: **Sensitive Melting Analysis after Real Time-Methylation Specific PCR (SMART-MSP): high-throughput and probe-free quantitative DNA methylation detection**. *Nucleic Acids Res* 2008, **36**(7):e42.
11. Liu Z, Zhou J, Gu L, Deng D: **Significant impact of amount of PCR input templates on various PCR-based DNA methylation analysis and countermeasure**. *Oncotarget* 2016, **7**(35):56447-56455.
12. Advani SM, Advani P, DeSantis SM, Brown D, VonVille HM, Lam M, Loree JM, Mehrvarz Sarshekeh A, Bressler J, Lopez DS *et al*: **Clinical, Pathological, and Molecular Characteristics of CpG Island**

Methylator Phenotype in Colorectal Cancer: A Systematic Review and Meta-analysis. *Transl Oncol* 2018, **11**(5):1188-1201.

13. Jia M, Gao X, Zhang Y, Hoffmeister M, Brenner H: **Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review**. *Clin Epigenetics* 2016, **8**:25.
14. Bae JM, Kim JH, Kwak Y, Lee DW, Cha Y, Wen X, Lee TH, Cho NY, Jeong SY, Park KJ *et al*: **Distinct clinical outcomes of two CIMP-positive colorectal cancer subtypes based on a revised CIMP classification system**. *Br J Cancer* 2017, **116**(8):1012-1020.
15. Xi X, Li T, Huang Y, Sun J, Zhu Y, Yang Y, Lu ZJ: **RNA Biomarkers: Frontier of Precision Medicine for Cancer**. *Noncoding RNA* 2017, **3**(1).
16. Siegfried Z, Simon I: **DNA methylation and gene expression**. *Wiley Interdiscip Rev Syst Biol Med* 2010, **2**(3):362-371.
17. Moarii M, Reyal F, Vert JP: **Integrative DNA methylation and gene expression analysis to assess the universality of the CpG island methylator phenotype**. *Hum Genomics* 2015, **9**:26.
18. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data**. *Nat Rev Genet* 2010, **11**(10):733-739.
19. Qi L, Chen L, Li Y, Qin Y, Pan R, Zhao W, Gu Y, Wang H, Wang R, Chen X *et al*: **Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer**. *Brief Bioinform* 2016, **17**(2):233-242.
20. Cheng J, Guo Y, Gao Q, Li H, Yan H, Li M, Cai H, Zheng W, Li X, Jiang W *et al*: **Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites**. *Oncotarget* 2017, **8**(18):30265-30275.
21. Chen R, Guan Q, Cheng J, He J, Liu H, Cai H, Hong G, Zhang J, Li N, Ao L *et al*: **Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples**. *Oncotarget* 2017, **8**(4):6652-6662.
22. Liu H, Li Y, He J, Guan Q, Chen R, Yan H, Zheng W, Song K, Cai H, Guo Y *et al*: **Robust transcriptional signatures for low-input RNA samples based on relative expression orderings**. *BMC Genomics* 2017, **18**(1):913.
23. Song K, Guo Y, Wang X, Cai H, Zheng W, Li N, Song X, Ao L, Guo Z, Zhao W: **Transcriptional signatures for coupled predictions of stage II and III colorectal cancer metastasis and fluorouracil-based adjuvant chemotherapy benefit**. *FASEB J* 2019, **33**(1):151-162.
24. Zhao W, Chen B, Guo X, Wang R, Chang Z, Dong Y, Song K, Wang W, Qi L, Gu Y *et al*: **A rank-based transcriptional signature for predicting relapse risk of stage II colorectal cancer identified with proper data sources**. *Oncotarget* 2016, **7**(14):19060-19071.
25. Li M, Li H, Hong G, Tang Z, Liu G, Lin X, Lin M, Qi L, Guo Z: **Identifying primary site of lung-limited Cancer of unknown primary based on relative gene expression orderings**. *BMC Cancer* 2019, **19**(1):67.

26. Barton MK: **Primary tumor location found to impact prognosis and response to therapy in patients with metastatic colorectal cancer**. *CA Cancer J Clin* 2017, **67**(4):259-260.
27. Loupakis F, Yang D, Yau L, Feng S, Cremolini C, Zhang W, Maus MK, Antoniotti C, Langer C, Scherer SJ *et al*: **Primary tumor location as a prognostic factor in metastatic colorectal cancer**. *J Natl Cancer Inst* 2015, **107**(3).
28. Shen H, Yang J, Huang Q, Jiang MJ, Tan YN, Fu JF, Zhu LZ, Fang XF, Yuan Y: **Different treatment strategies and molecular features between right-sided and left-sided colon cancers**. *World J Gastroenterol* 2015, **21**(21):6470-6478.
29. Yamauchi M, Morikawa T, Kuchiba A, Imamura Y, Qian ZR, Nishihara R, Liao X, Waldron L, Hoshida Y, Huttenhower C *et al*: **Assessment of colorectal cancer molecular features along bowel subsites challenges the conception of distinct dichotomy of proximal versus distal colorectum**. *Gut* 2012, **61**(6):847-854.
30. Ng JM, Yu J: **Promoter hypermethylation of tumour suppressor genes as potential biomarkers in colorectal cancer**. *Int J Mol Sci* 2015, **16**(2):2472-2496.
31. Natsume S, Yamaguchi T, Takao M, Iijima T, Wakaume R, Takahashi K, Matsumoto H, Nakano D, Horiguchi SI, Koizumi K *et al*: **Clinicopathological and molecular differences between right-sided and left-sided colorectal cancer in Japanese patients**. *Jpn J Clin Oncol* 2018, **48**(7):609-618.
32. La Vecchia S, Sebastian C: **Metabolic pathways regulating colorectal cancer initiation and progression**. *Semin Cell Dev Biol* 2019.
33. Tominaga O, Nita ME, Nagawa H, Fujii S, Tsuruo T, Muto T: **Expressions of cell cycle regulators in human colorectal cancer cell lines**. *Jpn J Cancer Res* 1997, **88**(9):855-860.
34. Stoian M, State N, Stoica V, Radulian G: **Apoptosis in colorectal cancer**. *J Med Life* 2014, **7**(2):160-164.
35. Liu J, Zheng B, Li Y, Yuan Y, Xing C: **Genetic Polymorphisms of DNA Repair Pathways in Sporadic Colorectal Carcinogenesis**. *J Cancer* 2019, **10**(6):1417-1433.
36. Slattery ML, Mullany LE, Wolff RK, Sakoda LC, Samowitz WS, Herrick JS: **The p53-signaling pathway and colorectal cancer: Interactions between downstream p53 target genes and miRNAs**. *Genomics* 2019, **111**(4):762-771.
37. Shiovitz S, Bertagnolli MM, Renfro LA, Nam E, Foster NR, Dzieciatkowski S, Luo Y, Lao VV, Monnat RJ, Jr., Emond MJ *et al*: **CpG island methylator phenotype is associated with response to adjuvant irinotecan-based therapy for stage III colon cancer**. *Gastroenterology* 2014, **147**(3):637-645.
38. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.
39. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F: **Evaluation of the Infinium Methylation 450K technology**. *Epigenomics* 2011, **3**(6):771-784.
40. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Res* 2015, **43**(7):e47.

41. Hochberg Y, Benjamini Y: **More powerful procedures for multiple significance testing**. *Stat Med* 1990, **9**(7):811-818.
42. Crans GG, Shuster JJ: **How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial**. *Stat Med* 2008, **27**(18):3598-3611.
43. Bland JM, Altman DG: **The logrank test**. *BMJ* 2004, **328**(7447):1073.
44. Harrell FE, Jr., Lee KL, Mark DB: **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors**. *Stat Med* 1996, **15**(4):361-387.
45. McClish DK: **Analyzing a portion of the ROC curve**. *Med Decis Making* 1989, **9**(3):190-195.
46. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets**. *Nucleic Acids Res* 2012, **40**(Database issue):D109-114.

Additional File Legends

Additional file 1: The classification of 19-GPS and the number of LCC with CIMP status.

Table S1. The classification of 19-GPS in stage III RCC of GSE39582 in detail. Table S2. The number of CIMP+ and CIMP- LCC in the training and validation datasets.

Additional file 2: The gene expression values of 19-GPS and Enrichment analysis for DE genes.

Table S1. The gene expression values of 19-GPS in CRC samples with different percentage of tumor cells. Table S2. Enrichment analysis for DE genes.

Additional file 3: The Kaplan-Meier curves of RFS of the CIMP with MSI groups.

Fig 1. The Kaplan-Meier curves of RFS of the CIMP with MSI groups identified by 19-GPS and original labels in training database. (a, b) All of stage III RCC of CIMP+ with MSI-H group, CIMP+ with MSS group, CIMP- with MSI-H group and CIMP- with MSS group treated with surgery alone.

Figures

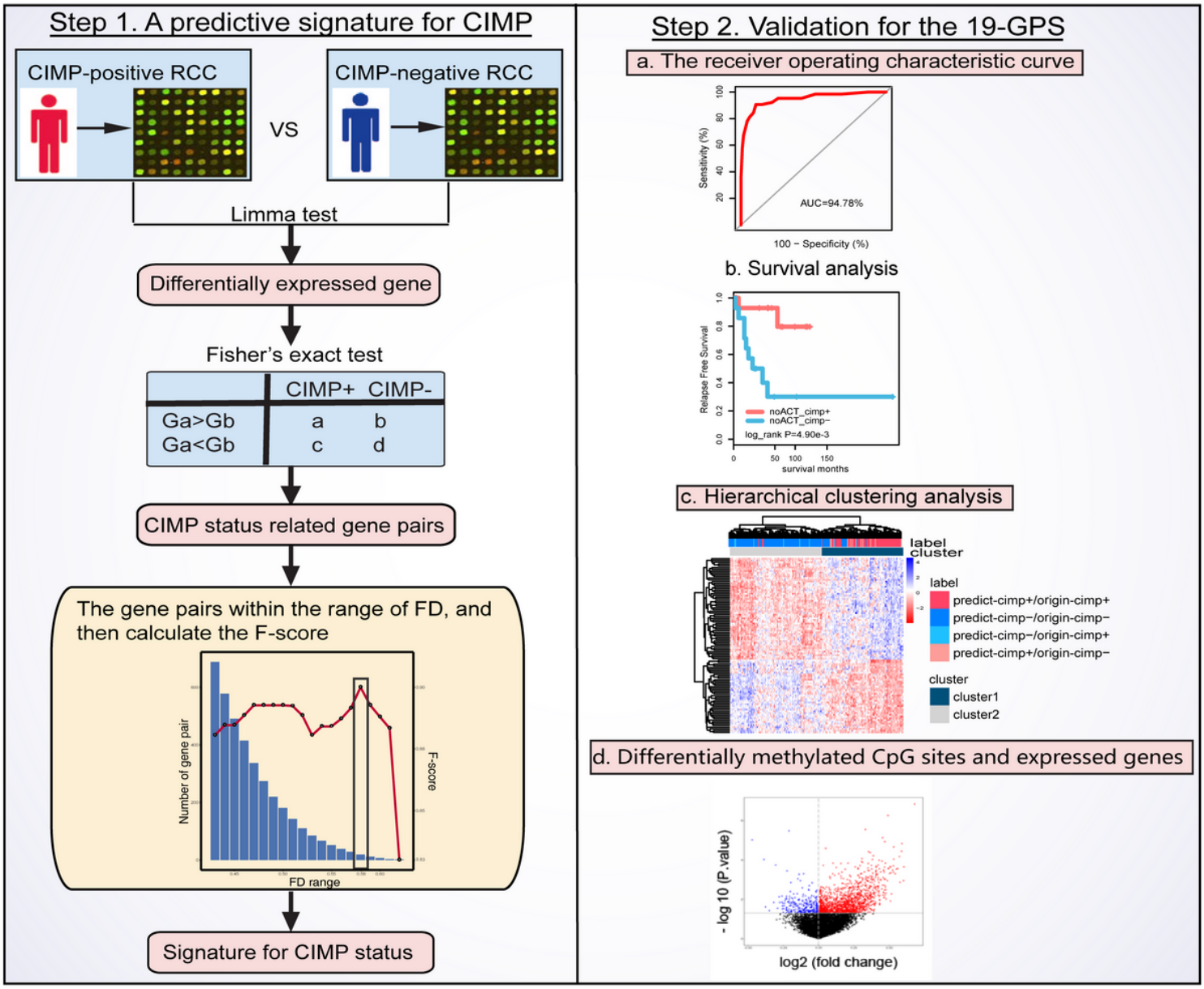

Figure 1

Flowchart of this study. CIMP, CpG island methylator phenotype; RCC, right-sided colon cancer; FD, frequency difference; F-score, harmonic mean value; 19-GPS, 19 gene pairs signatures.

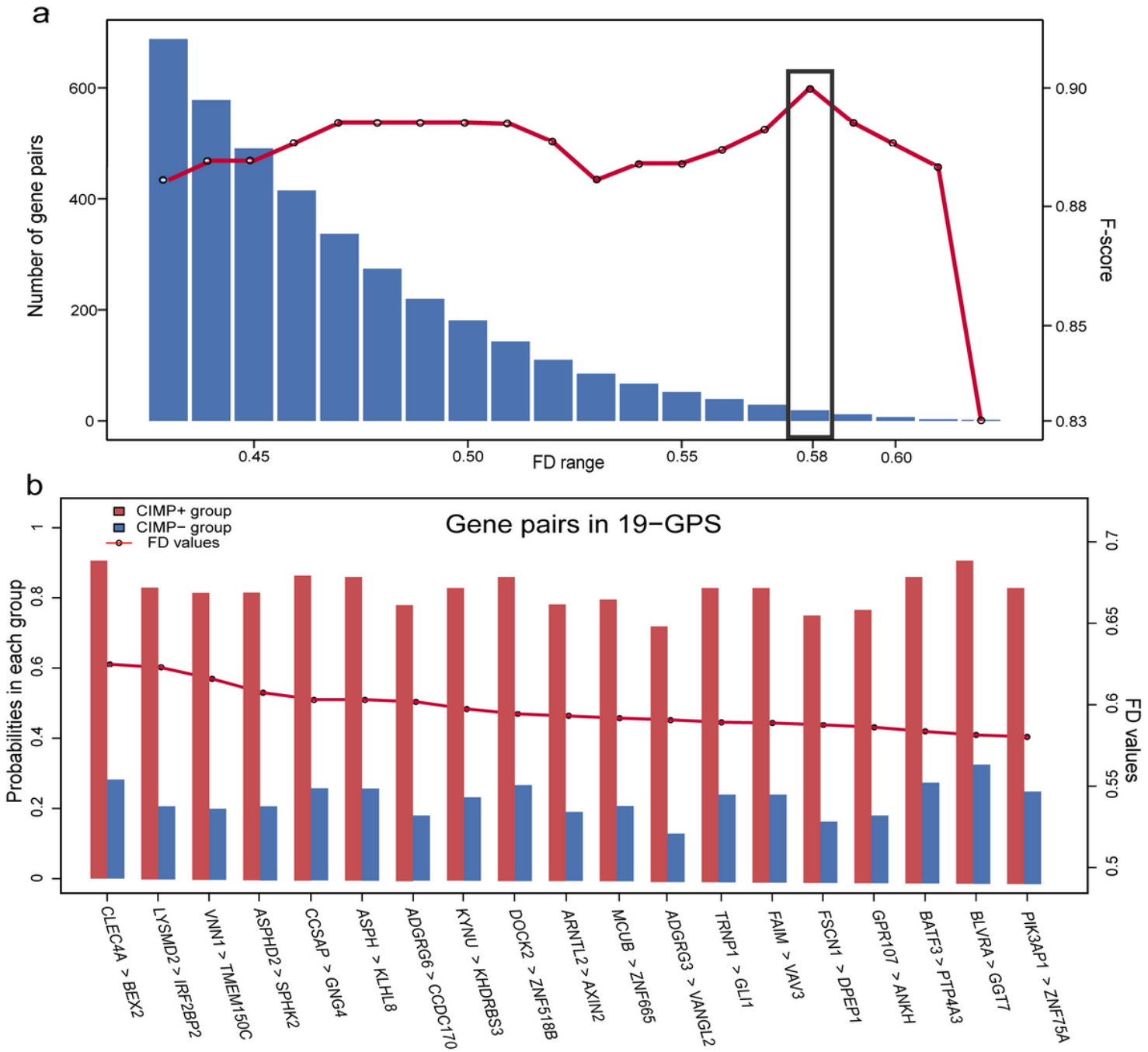

Figure 2

The F-score and number of the gene pairs within different range of FD values (a), and composition of 19-GPS (b). The x-axis represent the range of FD value and the relative expression orderings (REOs) (gene1 > gene2) of 19-GPS, respectively.

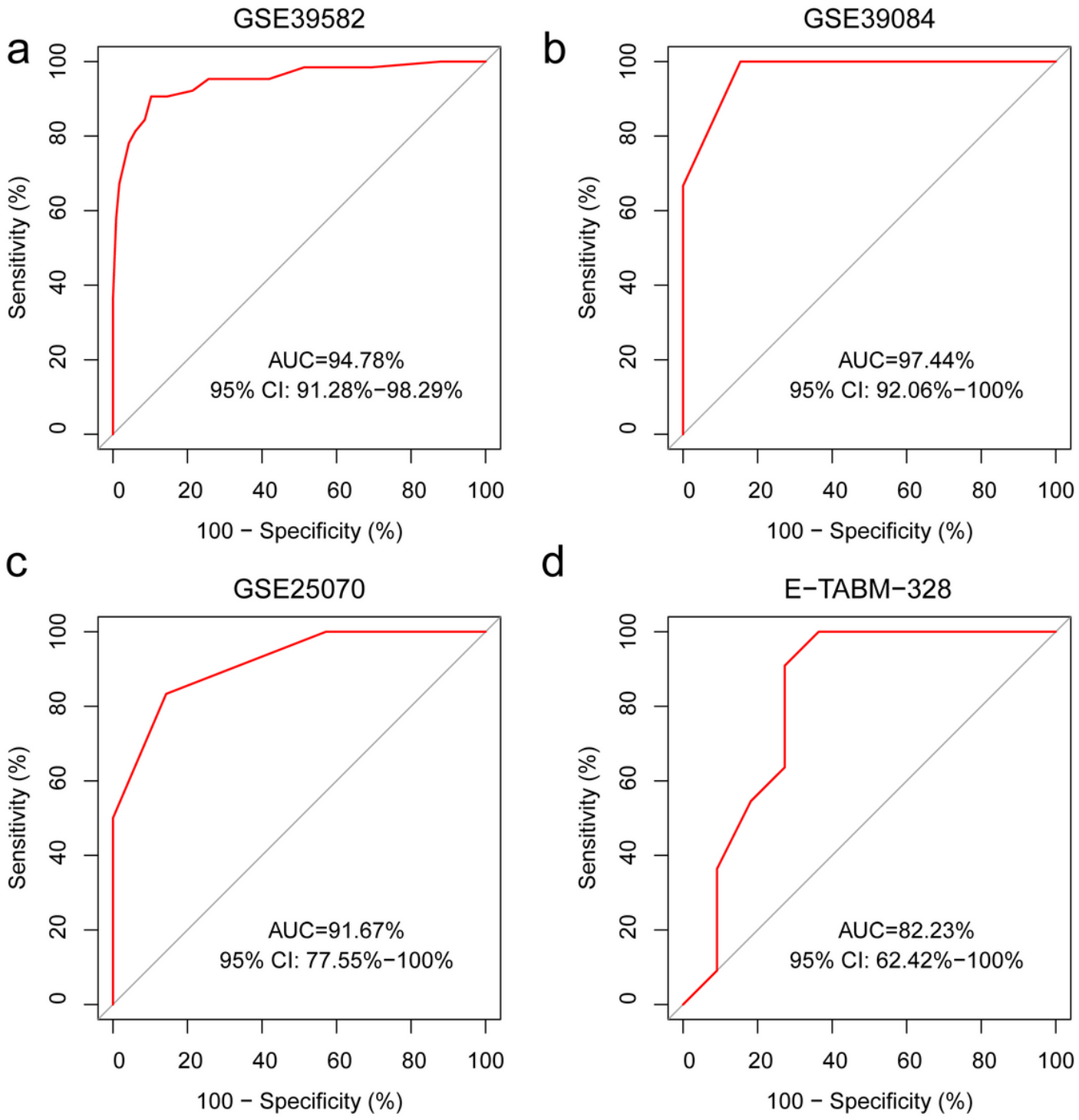

Figure 3

The ROC curves for 19-GPS in four independent datasets. (a) The right-sided colon cancer (RCC) of the training dataset, (b) The RCC of GSE39084, (c) The RCC of GSE25070, (d) The RCC of E-TABM-328.

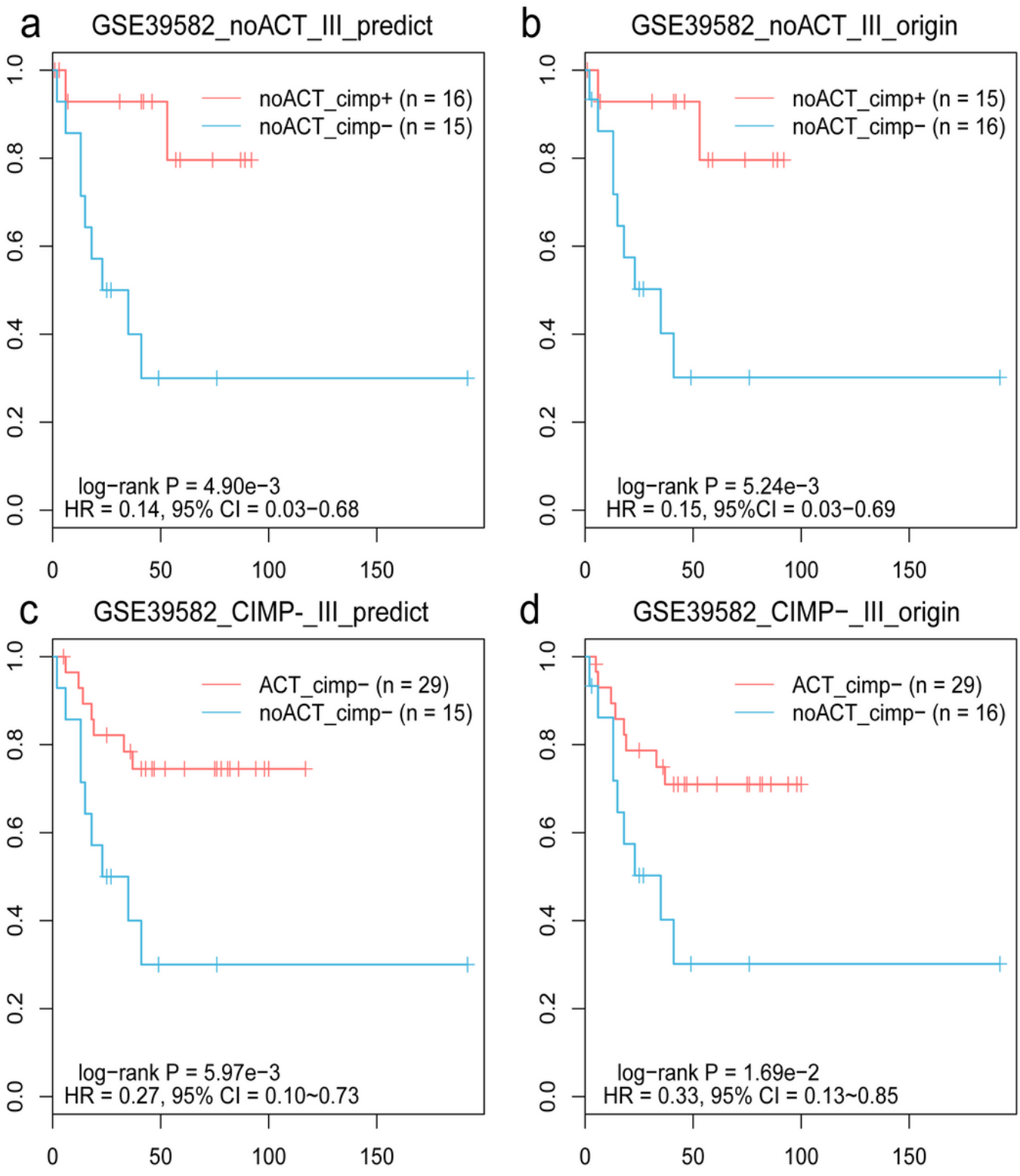

Figure 4

The Kaplan-Meier curves for the prediction of 19-GPS and original CIMP status in training dataset. (a, b) Stage III RCC of the CIMP+ and CIMP- patients treated with surgery alone. (c, d) All of stage III RCC of the CIMP- patients.

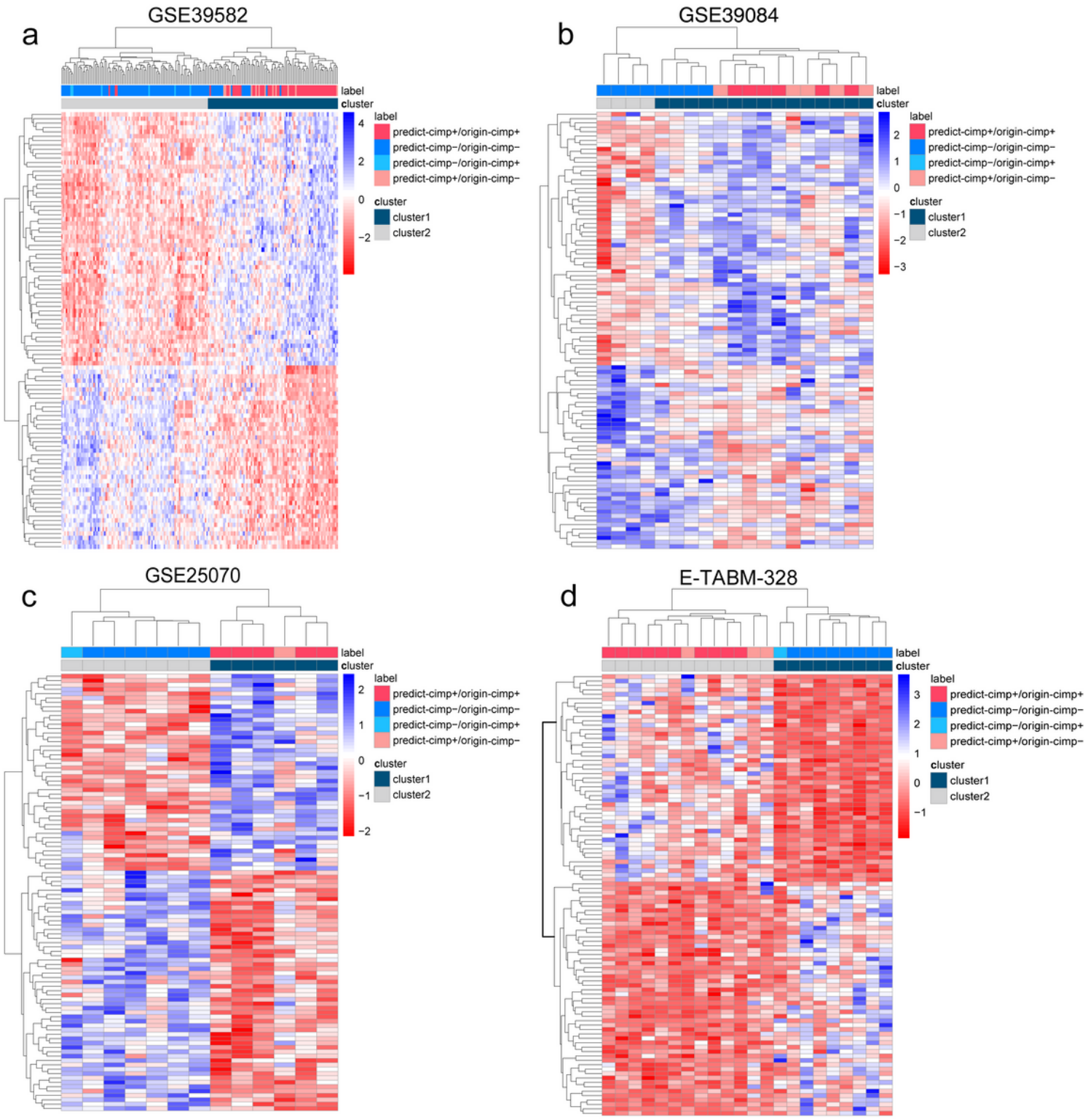

Figure 5

The complete linkage hierarchical clustering analysis of the RCC samples in four independent datasets. (a) GSE39582 (b) GSE39084 (c) GSE25070 (d) E-TABM-328 based on the differentially expressed genes between the signature-confirmed CIMP+ and CIMP- samples. Note: predict-CIMP/origin-CIMP, predict-CIMP represented the predicted CIMP status by 19-GPS and origin-CIMP represented the original CIMP status.

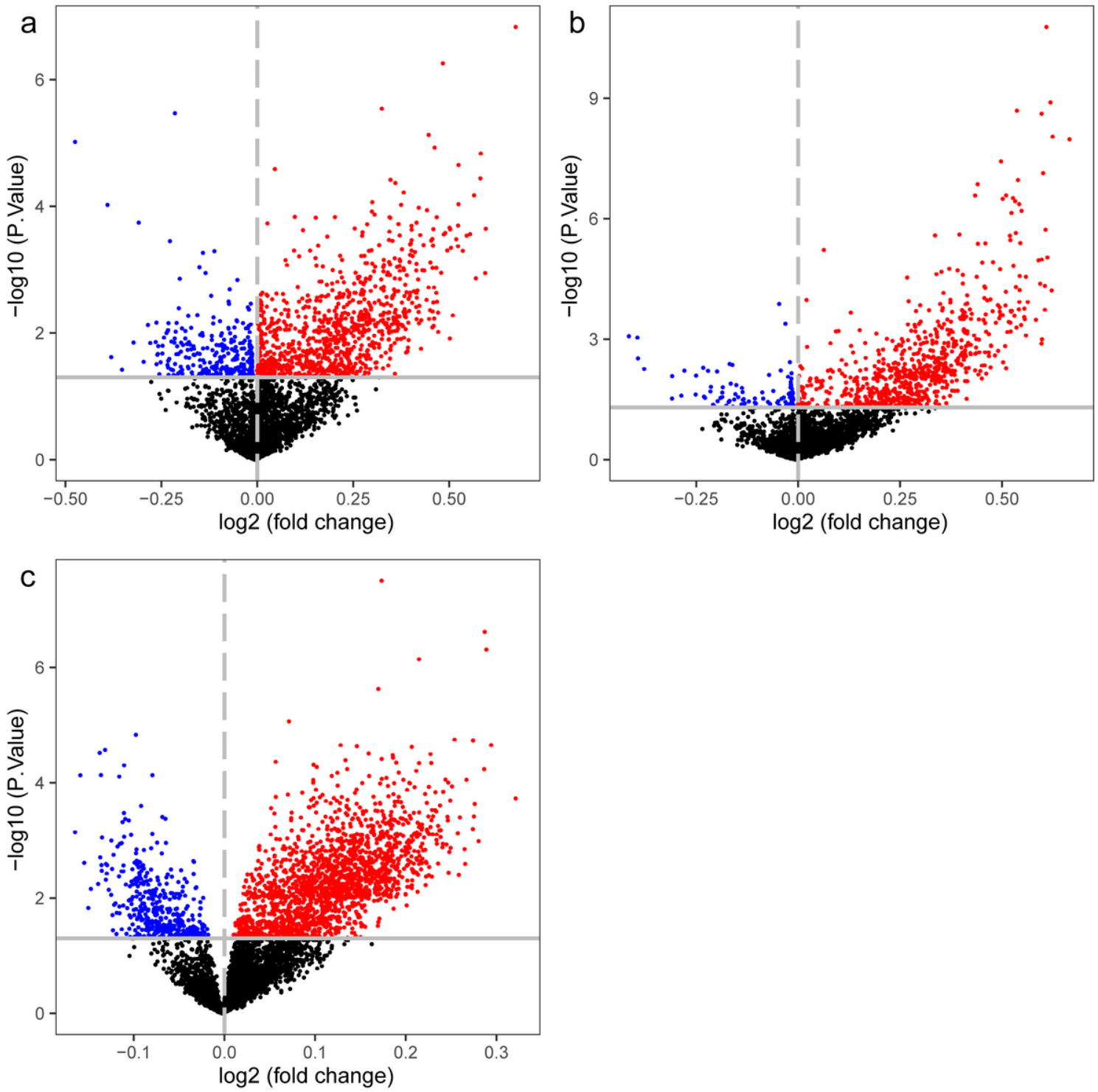

Figure 6

Volcano plots of the differentially methylated CpG sites between CIMP+ and CIMP- samples. (a) The samples with predicted CIMP status by 19-GPS in GSE25062, (b) The samples with original CIMP status in GSE25062, (c) The samples with predicted CIMP status by 19-GPS in GSE79740. The log2 (fold change) beta value difference in DNA methylation between the samples with CIMP+ and CIMP- status is plotted on the x-axis, and the P value (-1* log10 P value) for limma test of differences between the two subtypes is plotted on the y-axis. The CpG sites which are significantly different and log2 (fold change) >

0 between the two subtypes are shown in red, and the CpG sites which are significantly different and log2 (fold change) < 0 are shown in blue.

Supplementary Files

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

- Additionalfile2.xls
- AdditionalFile3.pdf
- Additionalfile1.docx