# Filtering high-dimensional methylation marks with extremely small sample size: an application to gastric cancer data

**Xin Chen**
University of Calgary
**Qingrun Zhang**
University of Calgary
**Thierry Chekouo** ( thierry.chekouotekou@ucalgary.ca )
University of Calgary

**Research Article**

**Keywords:** DNA methylation, BACkPAy, LIMMA, Bayesian model

**Posted Date:** March 13th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-284773/v1

RESEARCH

# Filtering high-dimensional methylation marks with extremely small sample size: an application to gastric cancer data

Xin Chen[1], Qingrun Zhang[1,2,3] and Thierry Chekouo[1,2,3*]

*Correspondence: thierry.chekouotekou@ucalgary.ca
[1]Department of Mathematics and statistics, University of Calgary, 2500 University Dr NW, T2N 1N4 Calgary, Canada
Full list of author information is available at the end of the article

**Abstract**

**Background:** DNA methylations in critical regions are highly involved in cancer pathogenesis and drug response. However, to identify causal methylations out of a large number of potential polymorphic DNA methylation sites is challenging. This high-dimensional data brings two obstacles: first, many established statistical models are not scalable to so many features; second, multiple-test and overfitting become serious. To this end, a method to quickly filter candidate sites to narrow down targets for downstream analyses is urgently needed.

**Methods:** BACkPAy is a pre-screening Bayesian approach to detect biological meaningful clusters of potential differential methylation levels with small sample size. BACkPAy prioritizes potentially important biomarkers by the Bayesian false discovery rate (FDR) approach. It filters non-informative sites (i.e. non-differential) with flat methylation pattern levels accross experimental conditions. In this work, we applied BACkPAy to a genome-wide methylation dataset with 3 tissue types and each type contains 3 gastric cancer samples. We also applied LIMMA (Linear Models for Microarray and RNA-Seq Data) to compare its results with what we achieved by BACkPAy. Then, Cox proportional hazards regression models were utilized to visualize prognostics significant markers with The Cancer Genome Atlas (TCGA) data for survival analysis.

**Results:** Using BACkPAy, we identified 8 biological meaningful clusters/groups of differential probes from the DNA methylation dataset. Using TCGA data, we also identified five prognostic genes (i.e. predictive to the progression of gastric cancer) that contain some differential methylation probes, whereas no significant results was identified using the Benjamin-Hochberg FDR in LIMMA.

**Conclusions:** We showed the importance of using BACkPAy for the analysis of DNA methylation data with extremely small sample size in gastric cancer. We revealed that RDH13, CLDN11, TMTC1, UCHL1 and FOXP2 can serve as predictive biomarkers for gastric cancer treatment and the promoter methylation level of these five genes in serum could have prognostic and diagnostic functions in gastric cancer patients.

**Keywords:** DNA methylation; BACkPAy; LIMMA; Bayesian model

## Background

DNA methylation is a biochemical process of adding a methyl group at the 5' carbon of the cytosine ring in a nucleotide. It is an epigenetic modification in which chemicals tag DNA and regulate gene expressions. Promoter DNA methylation is associated with genes silencing, which contributes to the development of diseases,

especially cancers. An active research field is to detect probes associated with differential methylation levels under contrasting conditions (e.g., sex and tissue types). However, the dimensionality of the problem makes it much harder. The number of features (probes) in methylation dataset is typically at least on the order of several thousand, whereas the number of samples may be few, presenting challenges in multiple hypothesis testing as well as overfitting.

In DNA methylation analysis, $\beta$ value is a quantitative indicator of the methylation level. The formula is shown below

$$\beta = \frac{Max(M, 0)}{Max(M, 0) + Max(U, 0) + 100}, \tag{1}$$

where $Max(M, 0)$ is the intensity of methylated allele, while $Max(U, 0)$ is the intensity of unmethylated allele[1]. $\beta$ value varies between 0 and 1, which represents the degree of DNA methylation in a sample. Generally, "zero" indicates there is no DNA methylation in CpG sites of the sample; "one" means that the focal CpG site in all the cells of the sample is methylated. Additionally, we used 0.2 and 0.8 as the thresholds of hypomethylation and hypermethylation. Alternatively, a $\beta$ value could be transformed to a M-value by the following formula

$$M = log_2 \frac{\beta}{1 - \beta}. \tag{2}$$

We can see the M-value and $\beta$-value are related through a log2 ratio transformation. However, the range of M-value is from -inf to +inf, which is larger than $\beta$ value. In this case, "zero" indicates the sample is half-methylated. And positive values mean a methylation rate greater than 50% while negative values suggest a methylation rate less than 50%.

It is quite hard to obtain meaningful results by applying traditional statistical methods like t-test to compare methylation levels among different groups when we have an extremely small sample size in each group. For instance, the ATM group defined in [4] only contains one female and two male samples. To this end, our novel method BACkPAy ([1]) is a suitable choice, which is developed to identify features (probes) that are differentially expressed in varying conditions for downstream analyses[1]. By applying this method, we could filter potentially significant probes by Bayesian FDR method and also obtain several biologically meaningful clusters/groups (eg. UpUp-UpDown, UpDown-UpUp) of probes. Clusters of probes represent the different patterns of methylation levels among groups. For instance, according to UpUp-UpDown cluster, UpUp means there is a significant increase in methylation levels from ATM to CAM and from CAM to NTM for male. On the other hand, UpDown represents a significant increase from ATM to CAM while a significant decrease from CAM to NTM for female.

In Figure 1, we provide an example of significant probes with UpUp-UpDown cluster we aim to detect. The figure shows that five potentially significant probes with differential methylation levels among three groups based on different tissue types have been filtered. On the one hand, we could indicate that the methylation levels between male and female samples in NTM group are different for selected

five probes. On the other hand, for both male and female samples, these five probes have differential methylation levels among three groups.

In this paper, we filtered significant CpG sites with p-values less than 0.05 by ANOVA at first. Then, 15,504 probes passing this critical value were used to detect the methylation differences among three groups comparing male to female samples by two statistical methods: LIMMA and our innovative Bayesian method (BACkPAy). After filtering potentially significant probes combined with TCGA data by BACkPAy, we further fitted a Cox model in order to obtain several significant genes as biomarkers for cancer identification.

In Section 2, we introduced statistical methods in details. In Section 3, we compared the results of two methods through methylation data of gastric cancer. The filtered significant probes by BACkPAy were separated into clusters with differential methylation levels. We also detected significant genes with prognostic function in gastric cancer by Cox model based on TCGA data relevant to potential significant probes filtered by BACkPAy and summarized some methylation biomarkers in different gastric groups (ATM, CAM and NTM). By the analysis of changes in DNA methylation and corresponding RNA gene expression, the effects of hypomethylation or hypermethylation among different gastric groups in terms of these biomarkers have been discussed. Finally, we concluded with a brief discussion in Section 4.

## Methods

### Datasets

In this work, we used a dataset with very small sample size and a large number of features, which is available from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) public functional genomics data repository with GEO number GSE97686 and GSE107161 [4]. They are originated from the Illumina Infinium HumanMethylation450 BeadChip arrays with 424,383 probes and obtained from 9 blood samples with gastric stromal myofibroblasts. Sex and tissue type are treated as experimental and independent variables respectively. There are three different tissue types: primary gastric cancer-associated myofibroblasts (CAM), patient-matched adjacent tissue myofibroblasts (ATM) and unrelated normal tissue myofibroblasts (NTM). Each tissue type contains 3 samples. Before analyzing methylation data, we pre-processed the original $\beta$ values in GSE97686 by logit transformation to get M-values that are more statistically valid for the differential analysis of methylation levels [3]. Similarly, GEO GSE107161 gene expression data contains total RNA with 47312 genes obtained from gastric stromal myofibroblasts, including 3 CAMs, 3 ATMs and 3 NTMs, and hybridised to theIllumina HumanHT-12v4 Expression BeadChip.

Some preliminary processing steps have been applied to the two datasets: (1) we remove genes with missing expression values for GSE107161; (2) and applied ANOVA to filter probes/genes with significant differences among tissue types (P-value$<$0.05). BACkPAy and LIMMA were applied to the remaining 15504 probes and 738 genes for the DNA methylation and gene expression datasets respectively.

In this manuscript, we also used the survival data of gastric cancer from TCGA-STAD dataset, The Cancer Genome Atlas Stomach Adenocarcinoma. TCGA-STAD

level-3 mRNA expression data contains 238 sample subjects with 26,540 mRNA markers.

### BACkPAy

BACkPAy [1] is a recently released software that implements a Monte Carlo Markov chain (MCMC) algorithm for the detection of omics patterns. BACkPAy is one of the few statistical methods that is able to detect omics patterns while identifying differential omics features between experimental conditions when sample sizes are extremely small. Here, we provide a general overview of the method. Additional details on the BACkPAy architecture as well as on the software can be found in [1].

The method relies on a constrained Bayesian mixture model for clustering that groups omics features (e.g., proteins) into biologically meaningful clusters. Patterns of differentially features in varying conditions are obtained as combinations of these clusters. In particular, by applying BACkPAy to the dataset, our aim is to group methylation probes based on their methylation profile over the three different tissue types for male and female samples.

Overall, the methylation data is encoded as $\mathbf{Y} = (\mathbf{y}_s, j = 1, ..., p; s = 1, 2)$, where $\mathbf{y_{js}} = (y_{js1}, y_{js2}, ..., y_{jsn_s})$ is a vector in which each element $y_{jsi}$ represents the M-value of probe $j$ in sample $i = 1, ..., n_s$ of type $s$, $s = 1$ if the sample is male and $s = 2$ if female. Then, we assume that $y_{js}$ comes from a mixture of a finite number $H$ of components. Given the $h$'th component (cluster) of the mixture, the model is written as

$$y_{jsi} = a_{jh} + x_{si1}\beta_{h1} + x_{si2}\beta_{h2} + \epsilon_{jsih}, \quad \epsilon_{jsih} \sim N(0, \sigma_h^2), \tag{3}$$

where $a_{jh}$ is the probe-specific random effect of probe $j$ within cluster $h$, $x_{si1} = 1$ (i.e. $k = 1$) if sample $i$ of type $s$ (male or female) is from ATM group, and 0 otherwise; and $x_{si2} = 1$ (i.e. $k = 2$) if sample $i$ of type $s$ is from NTM group, and 0 otherwise. Finally, $(\beta_{h1}, \beta_{h2})$ is the vector of slopes that measures the tissue type group effects. Through prior distributions of the slopes, we defined $H = 9$ clusters with respect to the signs of the coefficients $\beta_{sh1}$ and $\beta_{sh2}$. For instance, cluster 1 (DownUp) is characterized by features with decreasing methylation level from ATM group to CAM group and increasing methylation level from CAM group to NTM group ($\beta_{11} > 0$ and $\beta_{12} > 0$), and cluster 2 (FlatUp) is characterized by features with constant methylation level from ATM to CAM group and increasing methylation level from CAM group to NTM ($\beta_{21} = 0$ and $\beta_{22} > 0$), etc. Patterns of probes can be obtained by combining those clusters with respect to variable Sex. For instance, pattern of features DownUp-FlatUp is a group of probes that belong to clusters DownUp and FlatUp for male and female samples respectively. We would be interested in 8 patterns illustrated in Figure 2 to find out significant probes showing different methylation levels in each group and in each sex. The 8 patterns show differential methylation for 3 tissue types between males and females. Among these pattern plots, N represents the number of significant probes filtered with threshold=0.5 in each kind of pattern. We can analyze differential methylation from two angles: (i) the differential methylation profile among three groups for both male and female samples. (ii) DNA methylation based on the change or no change in levels between male and female samples at a specific group.

We studied the performance of BACkPAy on both simulated and proteomic cell line data with extremely sample sizes in each experimental condition (e.g. 1 or 2 samples per condition). We also shown that BACkPAy outperforms other methods designed to detect proteins differentially expressed such as LIMMA (linear model approach with robust hyperparameter estimation,[10]), EDGE (Extraction and analysis of Differential Gene Expression,[11]) and MB-statistic (multivariate empirical Bayes statistic, [12]). LIMMA is one of the most popular method for detecting differential features in particular when the sample size is small. However, in contrary to BACkPAy, the method cannot be applied to a dataset with only one sample in each experimental condition. We also applied LIMMA to our GEO data, whose details are presented below.

### LIMMA

LIMMA is an R/Bioconductor software package using linear models with robust hyperparameter estimation to assess differential methylation levels in several (more than 2) groups. It is a popular package. We fitted a linear model using `lmFit` function in R to estimate $\beta_{\mathbf{j}}$, the differences among three groups for the $j$'th probe. Let $\boldsymbol{y}_j^T = (y_{j_1}, ..., y_{j_n})$ define the DNA methylation level (M-value) for three groups with $n = n_1 + n_2 + n_3$ for the $j$'th probe. The expected value of $\boldsymbol{y}_j$ is defined as $E(\boldsymbol{y}_j) = \boldsymbol{X}\boldsymbol{\alpha}_j$, where $\boldsymbol{X}$ is a design matrix providing a representation of the different DNA methylation targets that have been hybridized to the arrays, $\boldsymbol{\alpha}_j$ is a vector of coefficients. $\boldsymbol{\beta}_j = \boldsymbol{C}^T\alpha_j$, where $\boldsymbol{C}$ is a contrast matrix. Thus, the null hypothesis for testing the DNA methylation differences between male and female in each group is $H_0 : \beta_{jt} = 0$ for probe $j = 1, ..., n$ and $t = 1, 2, 3$ representing three groups of tissue type. The test statistic for testing $H_0$ is the moderated t-statistic, based on a Bayesian approach, defined by

$$\tilde{t}_{jt} = \frac{\hat{\beta}_{jt}}{\tilde{s}_j\sqrt{v_{jt}}}. \tag{4}$$

The p-value for testing $H_0$ is calculated from the $t$ distribution with $d_j + d_0$ degrees of freedom. More information on $\tilde{s_j}, v_{jt}, d_j$ and $d_0$ could be found in the reference[1]. If the p-value of $\tilde{t}_{jt}$ is less than 0.05, we could reject the null hypothesis $H_0$, i.e. there is a significant difference between male and female for probe $j$ in group $t$. Conversely, If p-value is larger than 0.05, that means there is no difference between male and female for probe $j$ in group $t$.

## Results

### Gene expression profiling with TCGA data

From the result of BACkPAy shown in Figure 2, we have selected 181 significant probes with 8 patterns. These significant probes were selected by q-value with a threshold of 0.05, which indicates that these probes are likely to have differential methylation levels between males and females (95%). To further investigate probes of interest found using BACkPAy, we analysed TCGA-STAD data that contains approximately 238 cancer patients. We focused on mRNA expression and the overall survival time of those patients. From TCGA data, we identified the mRNA

expression of genes relative to the identified probes. We then fitted univariate Cox proportional hazard models for each gene [13]. Table 1 presents five genes that were found to be associated with survival time (adjusted P-value<0.05). Those genes are RDH13, CLDN11, TMTC1, UCHL1 and FOXP2. Complete results for all genes are presented in the Supplementary material.

For each of the 5 genes, i) we created two groups of samples that have low and high mRNA expressions as described in Figures 4-8; and ii) related those two groups to the overall survival time of gastric cancer patients. The corresponding Kaplan Meir survival curves of the two groups for each gene were generated in Figures 4-8. For gene RDH13, high expression is associated with decreased survival time while for the other genes, CLDN11, TMTC1, UCHL1 AND FOXP2, high expression is associated with increased survival time.

Figure 3 describes the circular plot of DNA methylation of significant probes identified by BACkPAy and corresponding gene expression profiling. In Figure 3, genes with the same color belong to the same pattern (e.g., genes written in blue indicate that important methylation probes from those genes belong to pattern UpUp-UpDown). As these genome regions show distinct DNA methylation patterns in myofibroblast populations, it is possible that these methylation patterns have significant influence on the growth or discovery of gastric tumors.

From Figure 3, we can observe that genes in chromosomes 6, 15 and 20 have similar methylation patterns, DownUp-UpUp or DownUp-DownDown. That means genes in these chromosomes show identical methylation pattern (DownUp; decrease from ATM to CAM and increase from CAM to NTM) for males while they have two different patterns for females (UpUp and DownDown). On the other hand, the two genes, CHMP4C and TRAPPC9 in chromosome 8, show the same methylation pattern (DownDown) for females while the methylation pattern for males is totally different between the two genes (DownUp pattern in CHMP4C, UpDown pattern in TRAPPC9). This indicates that, for gastric cancer, Sex shows a strong effect on the methylation pattern in chromosomes 6, 8, 15 and 20.

Additionally, in the third track from Figure 3, we illustrate that ARHGAP29, GNA13, TRPPAC9, EGFR, CDH23, ANKRD11, ARID3A and MYH9 are all hypomethylated (i.e., an increase in the proportion of unmethylated cytosines in a specific sequence that is normally methylated) in 9 samples with gastric cancer. Among them, gene ARHGAP29 and ARID3A are from UpDown-UpUp cluster, gene GNA13 is from UpUp-UpDown, MYH9 is from DownDown-UpDown and the remaining four genes (TRPPAC9, EGFR, CDH23, ANKRD11) are all from UpDown-DownDown pattern. It implies that hypomethylation of these genes is likely to have a predictive function for gastric cancer and different type of patients (male or female with different tissue type) would have differential hypomethylated probes in details.

### Differentially probes/genes using BACkPAy and LIMMA

In addition of detection patterns, BACkPAy is also able to identify non-differentially probes i.e probes that do not have a significant change between tissue types for male and female samples. Those probes actually belong to group FlatFlat-FlatFlat.

Using our GEO methylation data, BACkPAy identified 11834 (out of 15504) potential differential probes (Bayesian q-value<0.05) while we got only 1080 differential probes using LIMMA (p-value with F distribution<0.05). We note that in the presence of extremely small sample sizes (one or two samples per experimental condition), BACkPAy can be considered as a "pre-screen" method that screens out non-differential probes and keeps potential differential probes but not necessary important ones. On the other hand, for gene expression data (GSE107161), the number of potential differential genes filtered by BACkPAy is 10 out of 738, which is smaller than genes we got by LIMMA (34 out of 738). For these 10 genes, they have no overlap with those filtered obtained from methylation dataset. For both DNA methylation and gene expression datasets, after adjusting the p-values obtained by LIMMA with the Benjamin-Hochberg approach [14], we were not able to find any significant probes/genes (p-value<0.05).

### Methylation profiling of gastric tumors purified from different tissue type and different sex in each group

Table 1 lists prognostics probes with their corresponding groups/patterns and adjusted p-values while relating genes to survival time. For gene RDH13 (or probe cg18743287), there is no significant difference between male and female in both CAM and ATM groups (Up-Up) while the difference can be seen in NTM group (Up-Down). Similarly, we can get the methylation levels between male and female in NTM group are significantly different in gene UCHL1. In addition, the methylation levels between different sex in CAM group have significant differences in CLDN11, TMTC1 and FOXP2. And the last column in Table 1 implies the effect of genes on gastric cancer survival time. Positive represents high expression of this gene associated with better survival time while negative represents high gene expression combined with poorer survival time for gastric cancer.

### Promoter hypomethylation induces RDH13 expression in gastric cancer groups

Figure 9 shows that the RDH13 promoter is hypomethylated in ATM tissue group, which means that its methylation level is lower in ATM than the other two tissue types (CAM and NTM). However, the gene expression among tissue types confirms that RDH13 is significantly upregulated in ATMs compared to the other two tissue types, especially in NTM group. On the other hand, using TCGA data, Figure 4 shows that gastric cancer patients with low RDH13 expression have a better survival than those with high expression. Collectively, these results provide a strong indication that RDH13 expression may be induced by cancer-induced reprogramming, resulting in RDH13 promoter hypomethylation within gastric ATM group. In summary, the hypomethylation of gene RDH13 may provide a proxy or biomarker for gastric identification in ATMs.

### Promoter hypermethylation represses CLDN11 expression in gastric ATM and CAM group

For gene CLDN11, Figure 10 shows DNA methylation levels in each tissue type are NTM>ATM≈CAM, whereas the gene expression plot indicates that CLDN11 is upregulated in ATM and CAM tissue types compared with NTMs. But all of

the methylation levels in three tissue types are commonly hypomethylation. Interestingly, gastric cancer patients with low CLDN11 expression level are associated with poorer overall survival time (Figure 5). Hence, we can conclude that the hypomethylation can enhance CLDN11 expression levels and further inhibit the development of gastric tumors. On the contrary, the hypermethylation of CLDN11 are more likely to accelerate tumor growth. Collectively, all of the results above illustrate that CLDN11 expression may be repressed by hypermethylation in promoter region, which induces the gastric cancer.

### Promoter hypermethylation induces TMTC1 expression in non-cancer group

In order to investigate the cancer-induced change in TMTC1 expression, as the previous two genes, we extracted the survival data of TMTC1 using TCGA data, and the gene expression data of 9 samples in GSE107161 across the three tissue types. These data indicate that the level of TMTC1 promoter DNA methylation gradually changes in gastric cancer with low level in ATM and CAM tissue types and high level in NTM group which its $\beta$ value is about 0.8, hypermethylation, in Figure 11. In addition, the gene expression showed that TMTC1 expression is almost the same among three tissue types. Figure 6 also shows that gastric cancer patients with high TMCTC1 expression have better survival than low expression. In conclusion, these data might provide strong evidence that TMTC1 expression have a repressive influence on gastric cancer, further proving TMTC1 promoter hypermethylation in NTMs will repress gene expression in gastric cancer.

### Promoter hypermethylation represses UCHL1 expression in gastric cancer group

Based on 9 samples, the plots of means of beta-value and gene expression among three groups confirm that these methylation trends represent a negative correlation with UCHL1 expression patterns (Figure 12). Especially, UCHL1 is significantly downregulated in ATMs, the tissue group of gastric cancers. Moreover, patients with high UCHL1 expression have longer survival time than patients with low expression (Figure 7). We can conclude that promoter UCHL1 hypermethylation is associated with the repression of UCHL1 expression in gastric cancer group.

### Promoter hypomethylation induces FOXP2 expression in non-cancer group

For the gene FOXP2, we found that the methylation levels in ATM and CAM samples are quite similar while the level in NTMs is lower. However, FOXP2 expression level is upregulated in NTMs compared with other two tissue types (Figure 13). Moreover, patients with high gene expression have better survival time as shown in Figure 8. Thus, the upregulation of FOXP2 expression in NTMs and the better survival time in high-expressed group confirm that hypomethylation have a significantly positive effect on FOXP2 expression in non-cancer group.

### Association of genes RDH13, CLDN11, TMTC1, UCHL1 and FOXP2 with previous cancer studies

Here, we identify and report the relationship between the five genes and previous cancer disease studies. It is known that RDH13 shares the greatest sequence similarity with RDH11, RDH12 and RDH14, which are integral membrane proteins of the

endoplasmic reticulum and its subcellular localization in prostate cancer LNCaP cells, which express endogenous RDH13 at high levels. The protein expression pattern of RDH13 is in agreement with the presence of RDH13 transcripts in at least 32 adult tissues, as well as in embryonic and cancer tissues, as reported in the Expressed Sequence Tag GenBank database [15]. Gene CLDN11 (claudin-11) has been shown to be silenced in gastric cancer via hypermethylation of its promoter region, and this hypermethylation is significantly correlated with downregulation of CLDN11 expression versus normal tissues [16]. It has also be shown that during the treatment of gastric cancer, differentially downregulated TMTC1 protein is identified in human gastric carcinoma cells [17], and TMTC1 is also found to be correlated with breast cancer at the functional level [18].

Gene UCHL1, Ubiquitin C-terminal hydrolase-L1 (UCHL1) is a de-ubiquitinating enzyme. As pointed out in [19], its function is controversial in different types of cancer diseases as it can be an oncogene (i.e. causes cancer) or a tumor suppressor. It's an oncogene in cancer disesases such as non-small lung cancer, lymphoma, etc. On the other hand, it is also reported that the expression of gene UCHL1 inhibits the growth in silenced cell lines, which indicates that UCHL1 is a kind of tumor suppressor in gastric and other digestive cancers [19].

For gene FOXP2, several researches have also reported the roles of FOXP2 as a tumor suppressor in gastric cancer and other diseases like osteosarcoma and hepatocellular carcinoma. It is revealed that FOXP2 expression was associated with the regulation of microRNAs in cancer cells. With the similar function of UCHL1 as a tumor suppressor in some cancers, FOXP2 could inhibit the growth of cancer cells by suppressing a series of cancer stem cell associated factors [21, 22].

## Conclusion

Gastric cancer is the third leading cause of cancer death worldwide and it is generally accepted that promoter methylation is an epigenetic mechanism playing an important role in regulating gene expression. Broadly, it has been frequently observed in almost all types of human malignancies such as gastric cancer. As a result, the detection of methylated DNA in the serum could provide a promising method for the prognosis and noninvasive diagnosis of gastric cancer. Additionally, we investigated novel serum methylation markers which have diagnostic or prognostic value in patients with gastric cancer.

To demonstrate the utility of BACkPAy, we chose a dataset with very small sample size. It was shown that the Bayesian hierarchical clustering approach in BACkPAy is advantageous for data with high-dimensional but very small sample sizes. In this paper, we identified 181 differential probes that belong to 8 distinct patterns. On the other hand, comparing male to female samples, we conclude that the probes within UpUp-UpDown, UpDown-UpUp, DownUp-DownDown and DownDown-DownUp clusters have significant difference between female and male samples in NTM group, whereas, the probes from UpUp-DownUp, UpDown-DownDown, DownUp-UpUp and DownDown-UpDown clusters have differential methylation level comparing male to female samples in CAM group. The result of patterns was visualized in 3D plots (Figure 2).

Further, we identified 5 prognostic genes (RDH13, CLDN11, TMTC1, UCHL1 and FOXP2) in gastric cancer. Except for RDH13, genes CLDN11, TMTC1, UCHL1 and

FOXP2, have been reported as tumor suppressors due to the inactivation of gene expression in gastric cancer ([4]). Conversely, our analysis implies that the overexpression of gene RDH13 raises the probability of gastric cancer. Specifically, DNA hypomethylation may lead to induce the RDH13 expression in gastric ATM group, whereas DNA hypermethylation is able to cause decreasing expression of UCHL1 in ATMs. Promoter hypermethylation will repress the expression of CLDN11 and FOXP2 in both ATMs and CAMs while DNA hypermethylation have repressive effect on the TMTC1 expression in gastric NTM group.

In summary, we indicated that DNA methylation level in the serum has the potential to serve as prognostic and diagnostic biomarkers for cancers. In this manuscript, we revealed that RDH13, CLDN11, TMTC1, UCHL1 and FOXP2 can serve as predictive biomarkers for gastric cancer treatment and the promoter methylation level of these five genes in serum could have prognostic and diagnostic functions in gastric cancer patients.

**Acknowledgements**
Thierry Chekouo was partially supported by NSERC Discovery Grant number RGPIN-2019-04810. Qingrun Zhang was supported by an NSERC Discovery Grant (RGPIN-2018-05147), a New Frontier in Research Fund (NFRFE-2018-00748) and a Startup Grant offered by the University of Calgary.

**Funding**
NSERC discovering grant, New Frontier in Research Fund and University of Calgary Startup Grant.

**Abbreviations**
Not applicable

**Availability of data and materials**
The datasets used and/or analysed during the current study are available in GEO repository with accession number GSE97686 and GSE107161 and in Genomic Data Commons (GDC) Data Portal with project ID TCGA-STAD.

**Ethics approval and consent to participate**
Not Applicable

**Competing interests**
The authors declare that they have no competing interests.

**Consent for publication**
Not Applicable

**Author's contributions**
Conceptualization: TC. Methodology: TC and XC. Data extraction and implementation of the methods: XC. Formal analysis and investigation: TC, XC and QZ. Writing of the first draft: XC. Writing and editing: TC, XC and QZ. All authors read and approved the final manuscript.

**Authors' information**
Not Applicable

**Author details**
[1]Department of Mathematics and statistics, University of Calgary, 2500 University Dr NW, T2N 1N4 Calgary, Canada. [2]Alberta Children's Hospital Research Institute, University of Calgary, 2500 University Dr NW, T2N 1N4 Calgary, Canada. [3]Department of Biochemistry and Molecular Biology, University of Calgary, 2500 University Dr NW, Calgary, Canada.

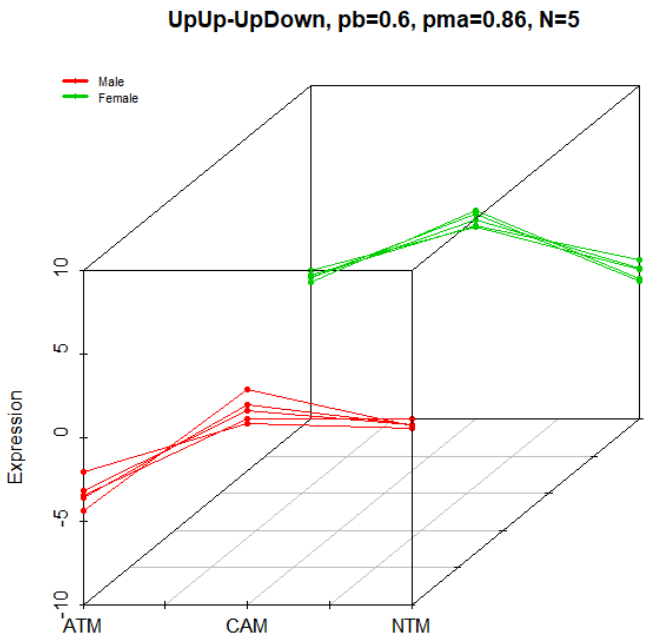
**References**

1. Thierry Chekouo, Francesco C Stingo, Caleb A Class, Yuanqing Yan, et al. Investigating protein patterns in human leukemia cell line experiments: A Bayesian approach for extremely small sample sizes. *Statistical Methods in Medical Research* 2019; 0(0) 1-16.
2. Dongmei Li, Zidian Xie, Marc Le Pape and Timothy Dye. An evaluation of statistical methods for DNA methylation microarray data analysis. *BMC Bioinformatics* 2015; 16:217.
3. Pan Du, Xiao Zhang, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 2010; 11:578.
4. Hanna Najgebauer, Triantafillos Liloglou, et al. Integrated omics profiling reveals novel patterns of epigenetic programming in cancer-associated myofibroblasts. *Carcinogenesis* 2019; Vol. 40, No. 4, 500-512.
5. Gongping Wang, Wei Zhang, Bo Zhou, et al. The diagnosis value of promoter methylation of UCHL1 in the serum for progression of gastric cancer. *BioMed Research International* 2015;

6. Fei Teng, Zhiyuan Xu, Jiahui Chen, et al. DUSP1 induces apatinib resistance by activating the MAPK pathway in gastric cancer. *Oncology Reports* 2018; Vol. 40, Issue 3.
7. Fang Liu, A. Jesse Gore, Julie L. Wilson and Murray Korc. DUSP1 Is a Novel Target for Enhancing Pancreatic Cancer Cell Sensitivity to Gemcitabine. *PLoS ONE* 2014; 9(1): e84982.
8. Xiaotu Ma, Yi-Wei Wang, Michael Q Zhang and Adi F Gazdar. DNA methylation data analysis and its application to cancer research. *Epigenomics* 2013; 5(3): 301–316.
9. John D. Storey. The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *The Annals of Statistics* Vol. 31, No. 6 (Dec., 2003), pp. 2013-2035.
10. Phipson, B, Lee, S, Majewski, IJ, et al. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *Ann Appl Stat* 2016; 10: 946–963.
11. Storey, JD, Xiao, W, Leek, JT, et al. Significance analysis of time course microarray experiments. Proc Natl Acad Sci USA 2005; 102: 12837–1284.
12. Tai, YC, Speed, TP. A multivariate empirical Bayes statistic for replicated microarray time course data. Ann Statist 2006; 34: 2387–2412.
13. MJ Bradburn, TG Clark, SB Love and DG Altman. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. British Journal of Cancer (2003) 89, 431 – 436
14. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. J R Stat Soc B 57:289–300
15. Olga V. Belyaeva, Olga V. Korkina, Anton V. Stetsenko, and Natalia Y. Kedishvili. Human retinol dehydrogenase 13 (RDH13) is a mitochondrial short-chain dehydrogenase/reductase with a retinaldehyde reductase activity. FEBS J. 2008 Jan; 275(1): 138–147.
16. Rachana Agarwal, Yuriko Mori, 1 Yulan Cheng, et al. Silencing of Claudin-11 Is Associated with Increased Invasiveness of Gastric Cancer Cells. PLoS One. 2009; 4(11): e8002.
17. Qian Mao, Pin-Hu Zhang, Jie Yang, et al. iTRAQ-Based Proteomic Analysis of Ginsenoside F2 on Human Gastric Carcinoma Cells SGC7901. Evid Based Complement Alternat Med. 2016; 2016: 2635483.
18. Francesco Moccia, Vittoria Fotia, Richard Tancredi, et al. Breast and renal cancer—Derived endothelial colony forming cells share a common gene signature. ORIGINAL RESEARCH. VOLUME 77, P155-164, MAY 01, 2017
19. Yu-yu Gu, Mei Yang, Mei Zhao, et al. The de-ubiquitinase UCHL1 promotes gastric cancer metastasis via the Akt and Erk1/2 pathways. Tumor Biol. (2015) 36:8379–8387
20. Gongping Wang, Wei Zhang, Bo Zhou, Canhui Jin, Zengfang Wang, Yantong Yang, Zhenzhen Wang, Ye Chen, and Xiaoshan Feng. The Diagnosis Value of Promoter Methylation of UCHL1 in the Serum for Progression of Gastric Cancer. Biomed Res Int. 2015; 2015: 741030.
21. Meng-Ting Chen He-Fen Sun Liang-Dong Li Yang Zhao Li-Peng Yang Shui-Ping Gao Wei Jin. Downregulation of FOXP2 promotes breast cancer migration and invasion through TGF$\beta$/SMAD signaling pathway. Oncology Letters. June-2018 Volume 15 Issue 6
22. Wen-Zhuo Jia, Tao Yu, Qi An, Hua Yang, Zhu Zhang, Xiao Liu, and Gang Xiao. MicroRNA-190 regulates FOXP2 genes in human gastric cancer. Onco Targets Ther. 2016; 9: 3643–3651.


**Figures**

**Figure 1 Example of probes with differential expression pattern, UpUp-UpDown, among three groups when comparing male (red – UpUp trajectory in this example) to female (green – UpDown trajectory in this example) samples.**




**Tables**

**Table 1** Significant genes for diagnosis of gastric cancer by Cox model.The first four columns list prognostics probes filtered by BACkPAy, their related, corresponding groups/patterns of methylation levels and adjusted p-values calculated by Benjamin-Hochbergapproach, respectively. The last column shows the effect of genes on gastric cancer survival time. + indicates high gene expression associated with better survival time while - indicates high gene expression with poorer survival time

| | probe | gene name | pattern | adjusted p-value | sign |
|---|---|---|---|---|---|
| 1 | cg18743287 | RDH13 | UpUp-UpDown | 0.0325 | - |
| 2 | cg17078427 | CLDN11 | UpUp-DownUp | 0.0325 | + |
| 3 | cg05471616 | TMTC1 | DownUp-UpUp | 0.0325 | + |
| 4 | cg09921610 | UCHL1 | DownUp-DownDown | 0.0325 | + |
| 5 | cg20050108 | FOXP2 | UpDown-DownDown | 0.0338 | + |

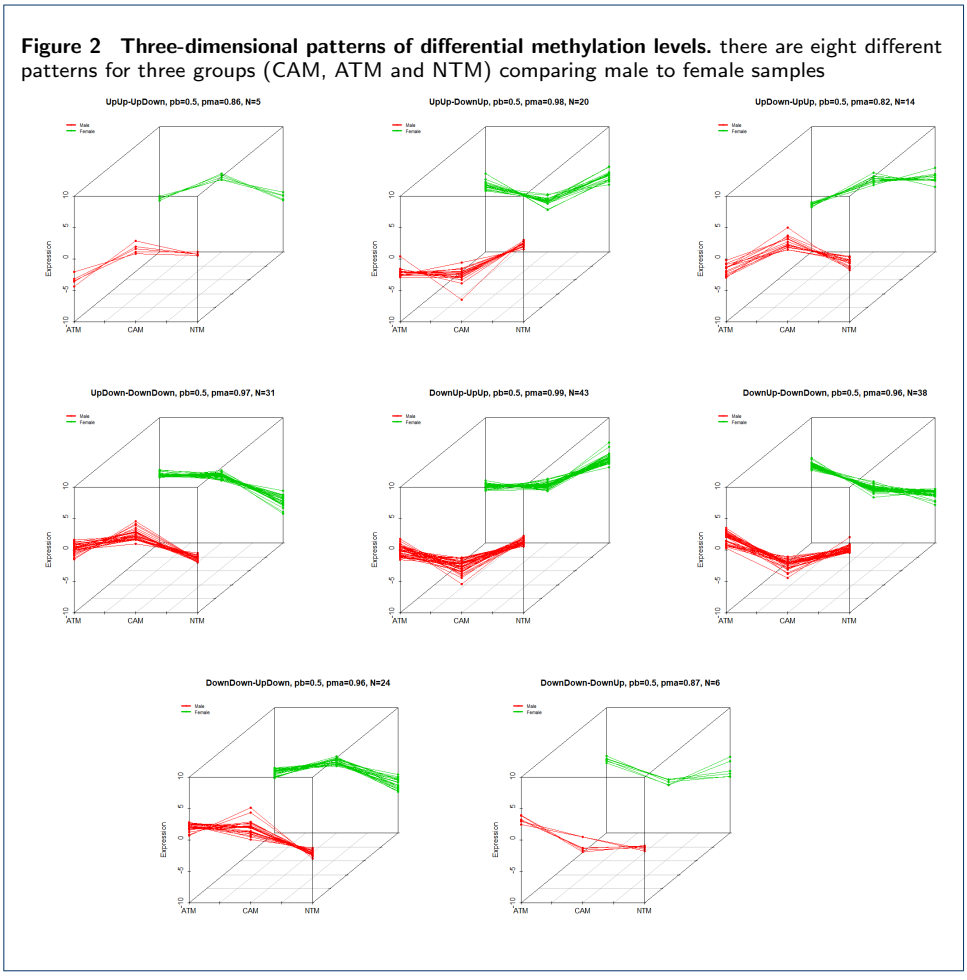



**Figure 2 Three-dimensional patterns of differential methylation levels.** there are eight different patterns for three groups (CAM, ATM and NTM) comparing male to female samples

**Figure 3 Circular plot of DNA methylation levels of significant genes in stromal myofibroblasts.** The outermost ring represents human ideograms, i.e. genome positions by chromosomes. The black lines are cytobands. The labels outside the ring are the names of significant genes selected by BACkPAy. Different colors of these gene labels represent the differential methylation patterns significant genes belong to among ATM (adjacent tissue-derived myofibroblast), CAM (cancer-associated myofibroblast) and NTM (normal tissue-derived myofibroblast) groups comparing male to female samples. The first track shows the adjusted p-value of these significant genes with the survival data in TCGA package. (red: pvalue¡0.05, blue: ¿0.05). The second track represents the DNA methylation level of corresponding CpG sites from significant genes in nine samples. (red: M value¿2, hypermethylated loci, blue: M value¡2, hypomethylated loci).

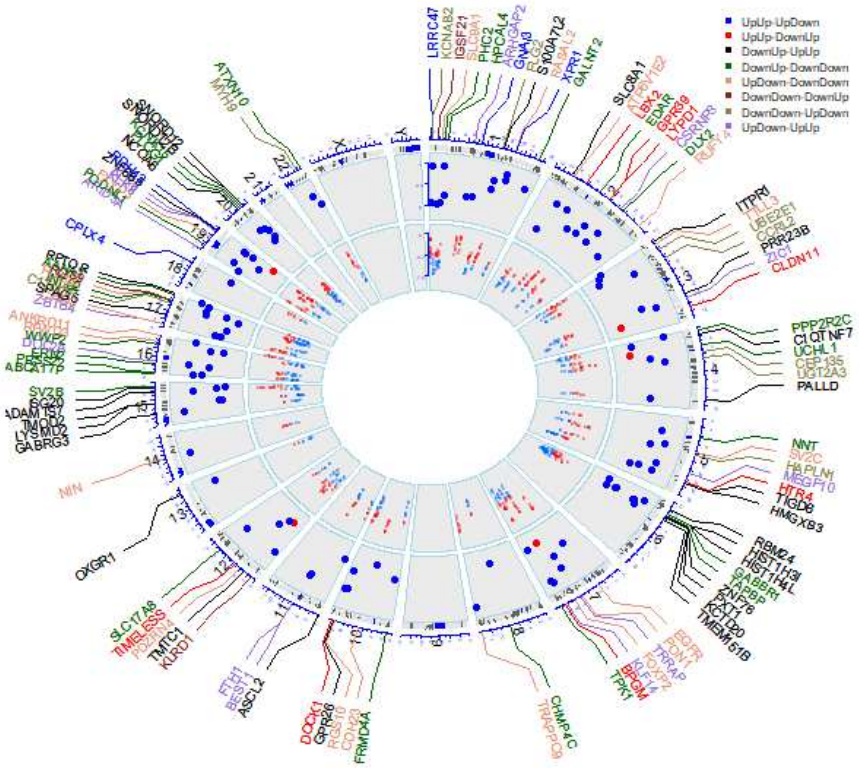



**Figure 4 Survival analysis in RDH13.** The left panel depicts the histogram of gene expression, we set 6 as the threshold. The right panel is Cox proportional hazard regression analysis of overall survival in gastric cancer patients according to the degree of gene expression

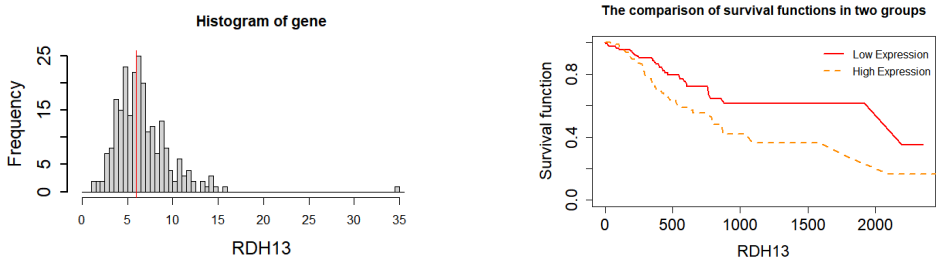


**Figure 5 Survival analysis in CLDN11.** The left panel describes the histogram of gene expression, we set 0.5 as the threshold. The right panel is Cox proportional hazard regression analysis of overall survival in patients with gastric cancer according to the degree of gene expression

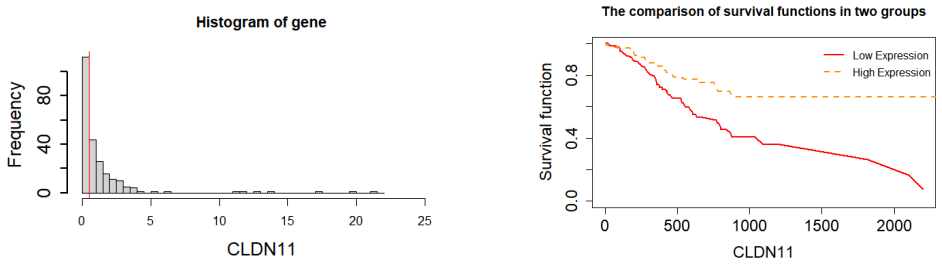



**Figure 6 Survival analysis in TMTC1.** The left panel depicts the histogram of gene expression, we set 2 as the threshold to divide overall expression into two groups. The right plot is the overall survival plot in gastric cancer patients according to the degree of gene expression by Cox model

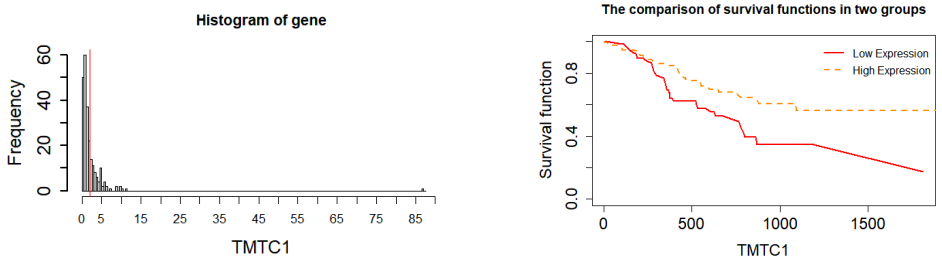



**Figure 7 Survival analysis in UCHL1.** The left panel depicts the histogram of gene expression, we set 2 as the threshold to divide overall expression into two groups. The right plot is the overall survival plot in gastric cancer patients according to the degree of gene expression by Cox model

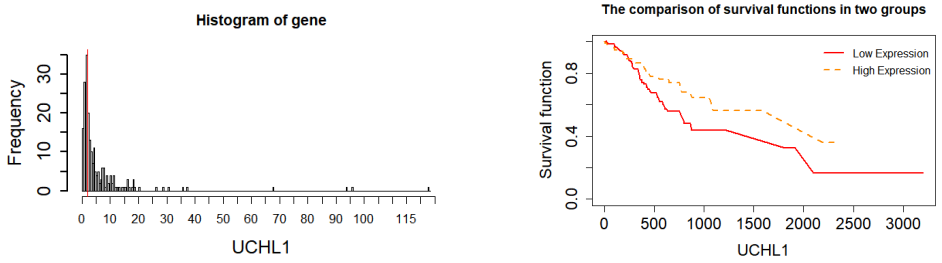



**Figure 8 Survival analysis in FOXP2.** The left panel depicts the histogram of gene expression, we set 1 as the threshold. The right panel is Cox proportional hazard regression analysis of overall survival in gastric cancer patients according to the degree of gene expression

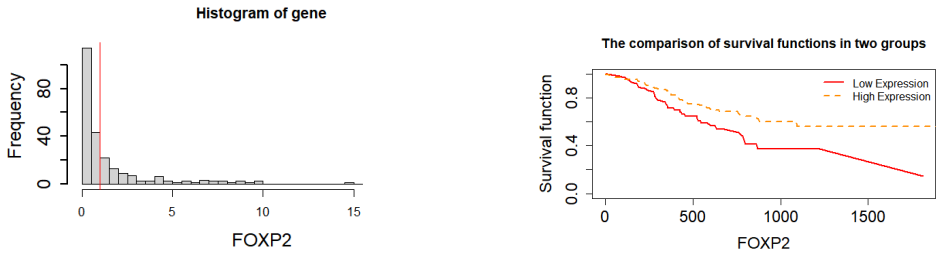


**Figure 9 The overall methylation and gene expression levels of the RDH13 promoter region.**

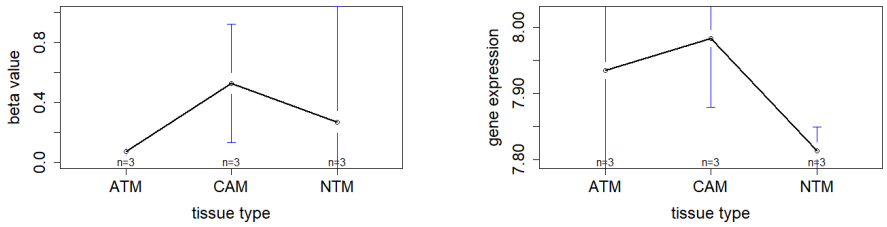



**Figure 10 The overall methylation and gene expression levels of the CLDN11 promoter region.**

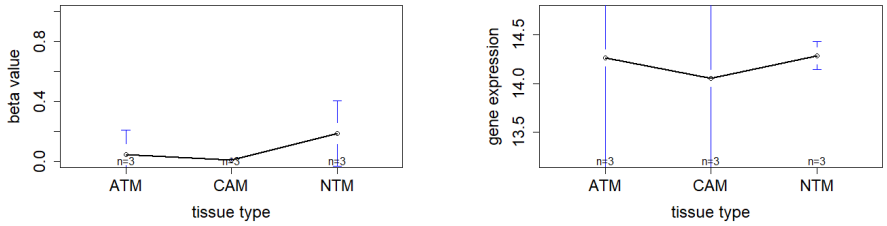



**Figure 11 The overall methylation and gene expression levels of the TMTC1 promoter region.**

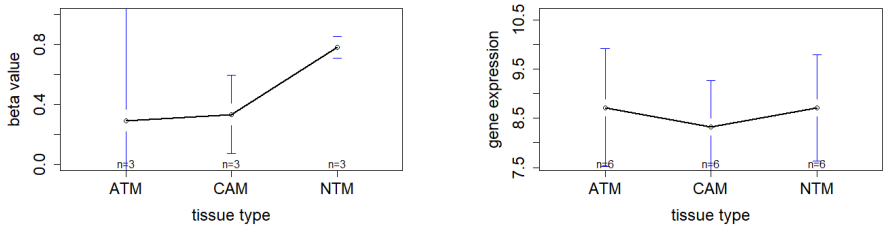



**Figure 12 The overall methylation and gene expression levels of the UCHL1 promoter region.**

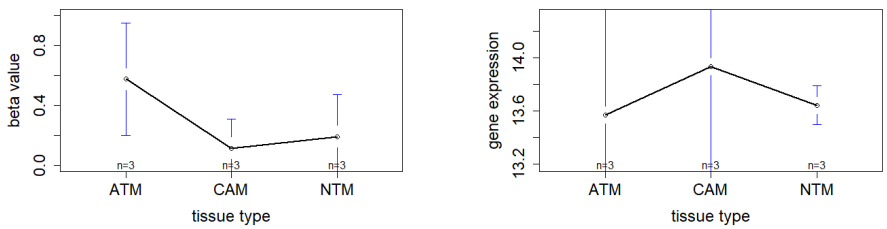
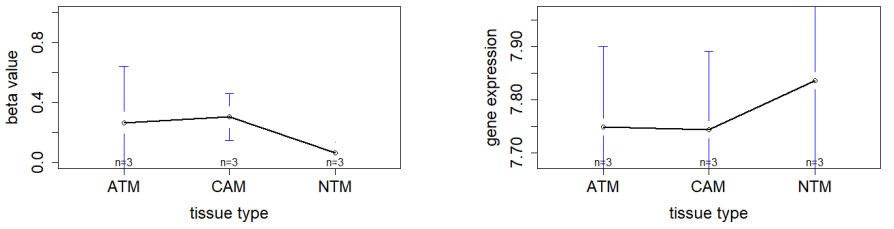


**Figure 13 The overall methylation and gene expression levels of the FOXP2 promoter region.**



# Figures

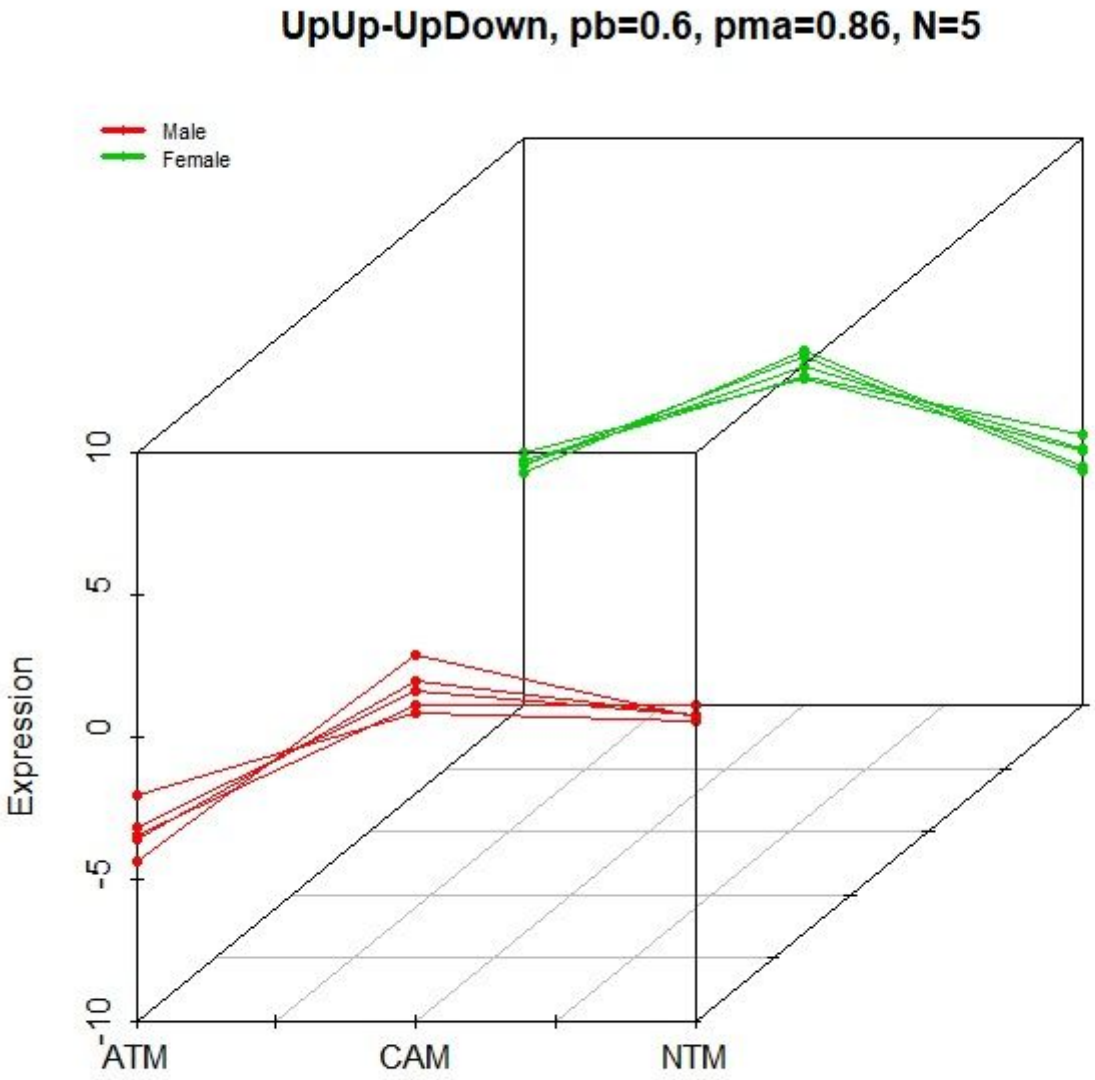
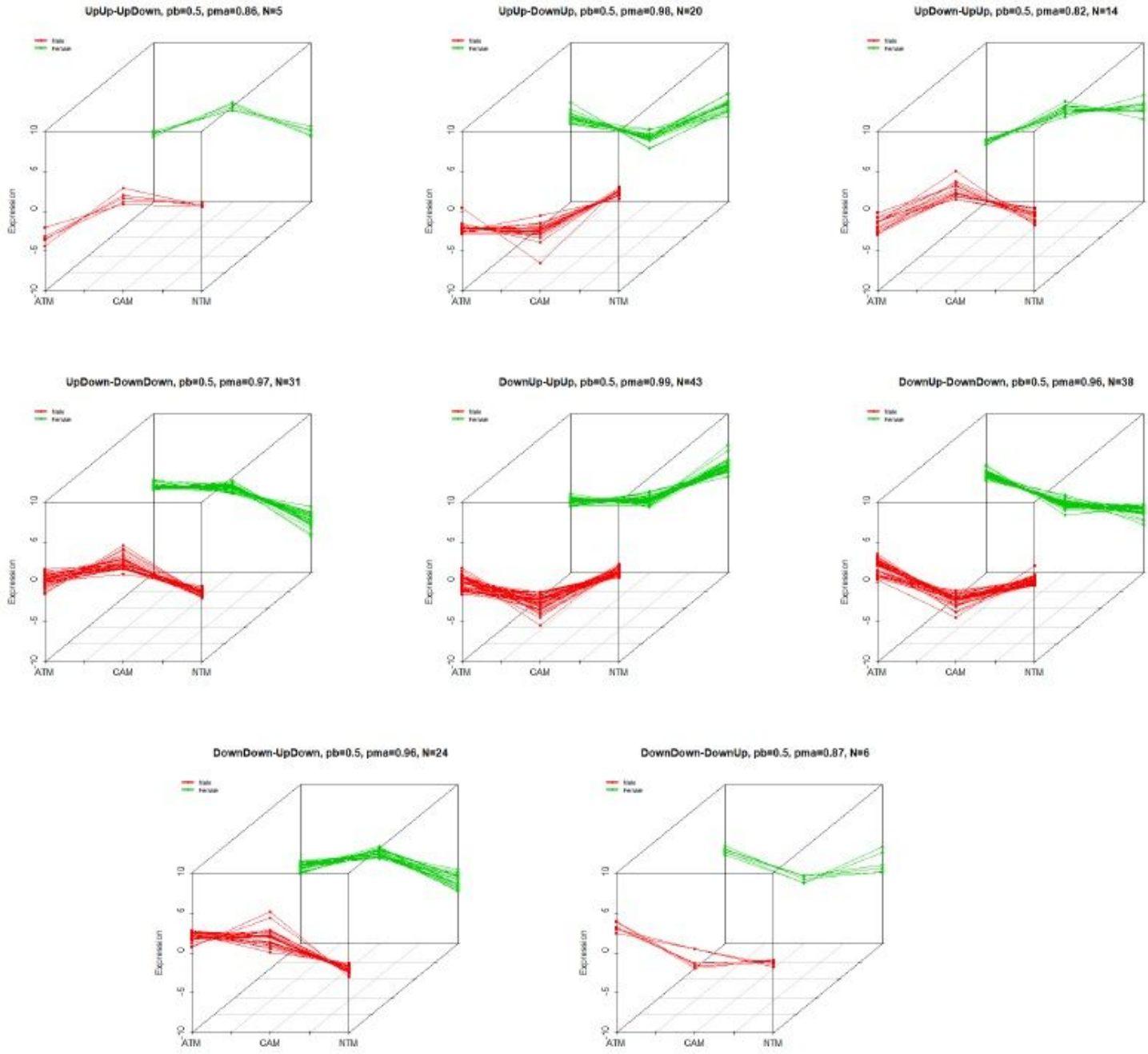


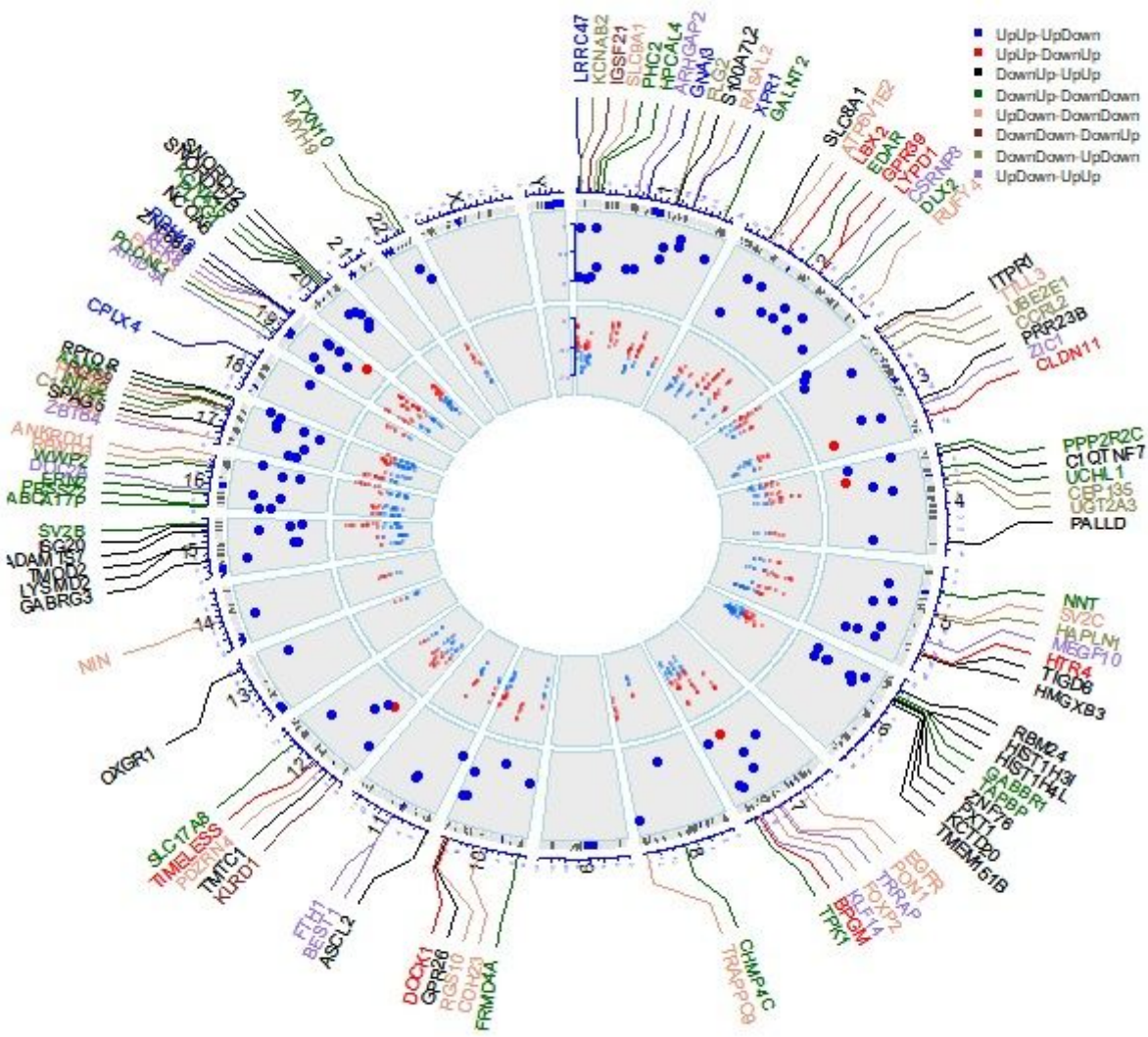

**Figure 1**

Example of probes with differential expression pattern, UpUp-UpDown, among three groups when comparing male (red { UpUp trajectory in this example) to female (green { UpDown trajectory in this example) samples.



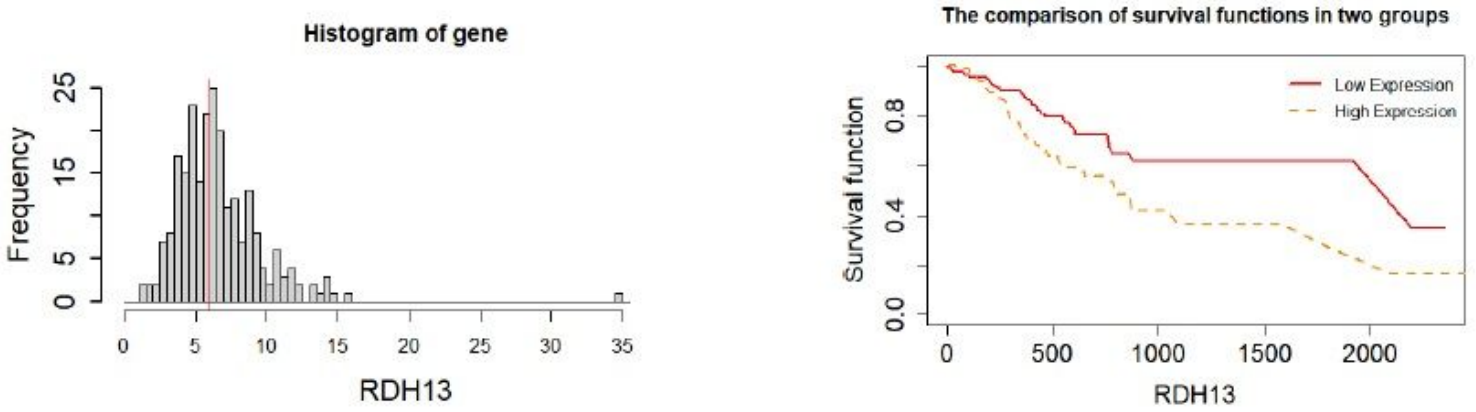

## Figure 2

Three-dimensional patterns of differential methylation levels. there are eight different patterns for three groups (CAM, ATM and NTM) comparing male to female samples



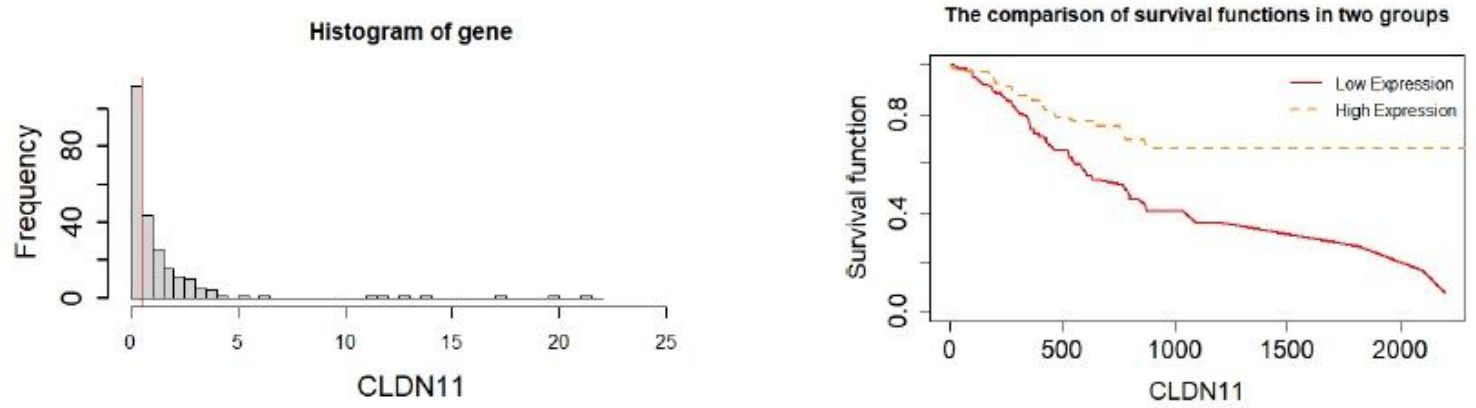

## Figure 3

Circular plot of DNA methylation levels of significant genes in stromal myobroblasts. The outermost ring represents human ideograms, i.e. genome positions by chromosomes. The black lines are cytobands. The labels outside the ring are the names of significant genes selected by BACkPAy. Different colors of these gene labels represent the differential methylation patterns signicant genes belong to among ATM (adjacent tissue-derived myobroblast), CAM (cancer-associated myobroblast) and NTM (normal tissue-derived myobroblast) groups comparing male to female samples. The rst track shows the adjusted p-value of these significant genes with the survival data in TCGA package. (red: pvalue<0.05, blue: >0.05). The second track represents the DNA methylation level of corresponding CpG sites from significant genes in nine samples. (red: M value>2, hypermethylated loci, blue: M value<2, hypomethylated loci).



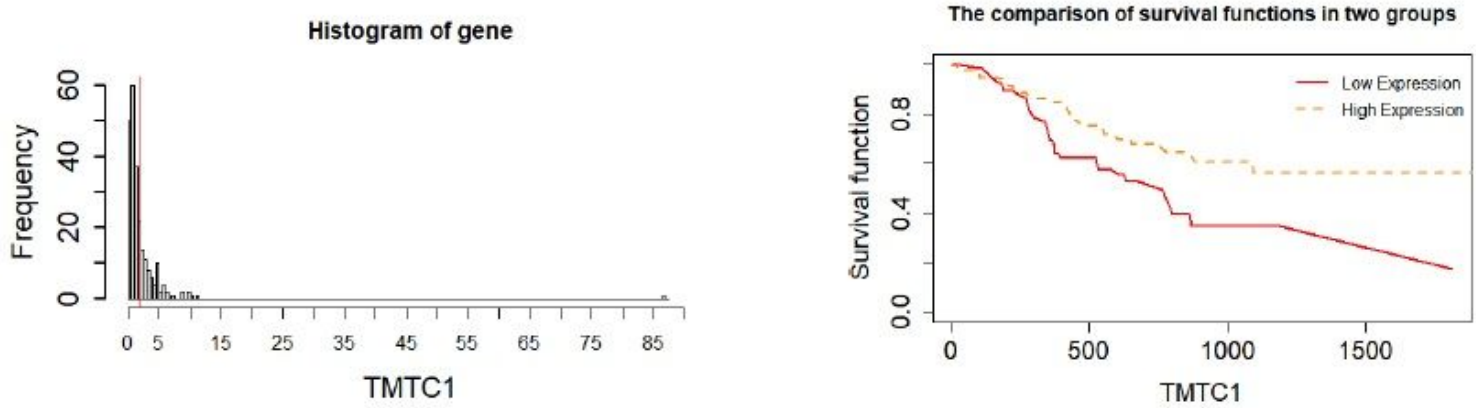

## Figure 4

Survival analysis in RDH13. The left panel depicts the histogram of gene expression, we set 6 as the threshold. The right panel is Cox proportional hazard regression analysis of overall survival in gastric cancer patients according to the degree of gene expression



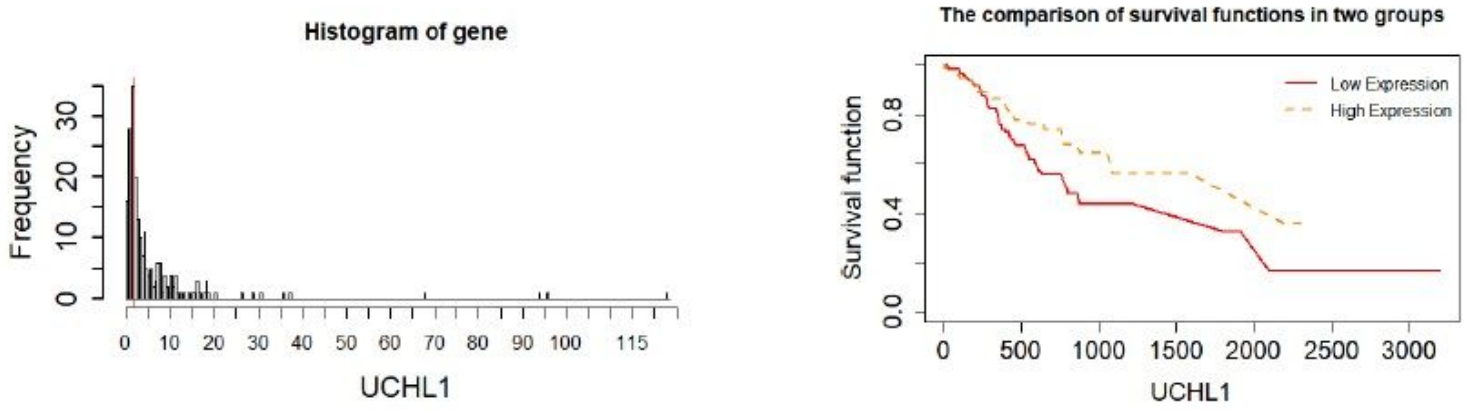

## Figure 5

Survival analysis in CLDN11. The left panel describes the histogram of gene expression, we set 0.5 as the threshold. The right panel is Cox proportional hazard regression analysis of overall survival in patients with gastric cancer according to the degree of gene expression



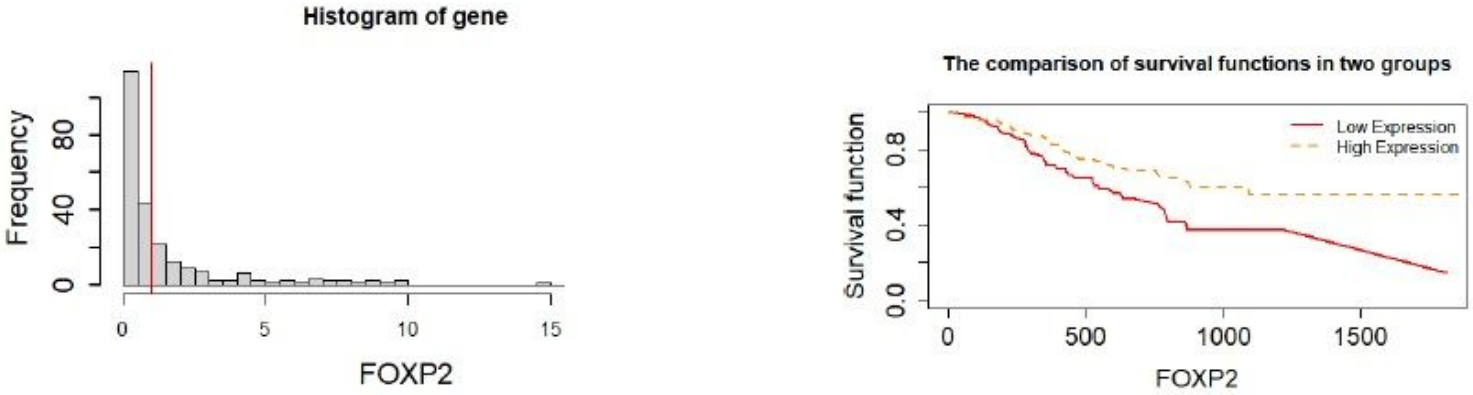
## Figure 6

Survival analysis in TMTC1. The left panel depicts the histogram of gene expression, we set 2 as the threshold to divide overall expression into two groups. The right plot is the overall survival plot in gastric cancer patients according to the degree of gene expression by Cox model




## Figure 7

Survival analysis in UCHL1. The left panel depicts the histogram of gene expression, we set 2 as the threshold to divide overall expression into two groups. The right plot is the overall survival plot in gastric cancer patients according to the degree of gene expression by Cox model




## Figure 8

Survival analysis in FOXP2. The left panel depicts the histogram of gene expression, we set 1 as the threshold. The right panel is Cox proportional hazard regression analysis of overall survival in gastric cancer patients according to the degree of gene expression

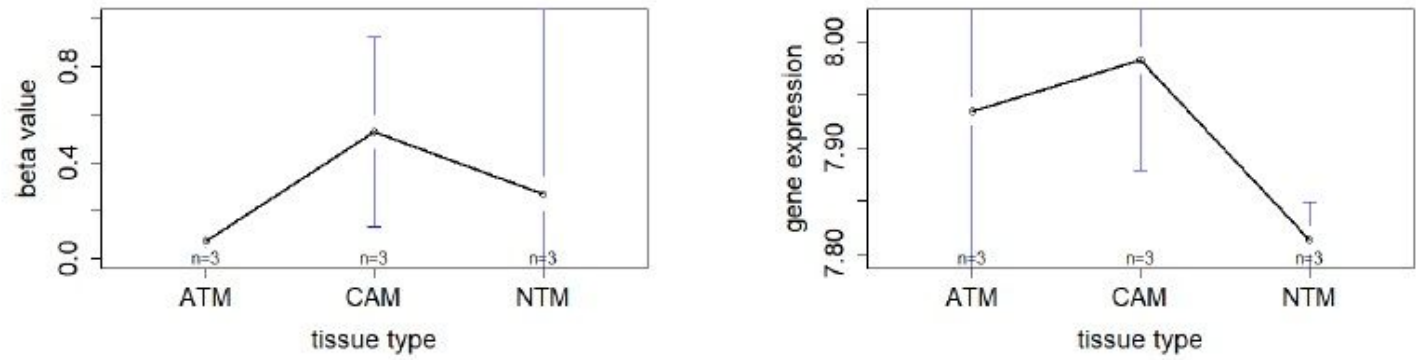



**Figure 9**

The overall methylation and gene expression levels of the RDH13 promoter region.

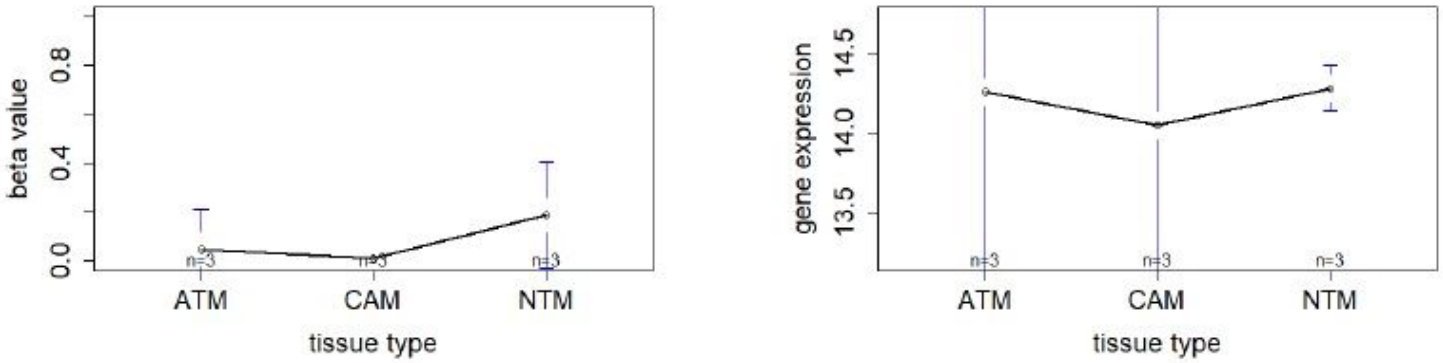



**Figure 10**

The overall methylation and gene expression levels of the CLDN11 promoter region.

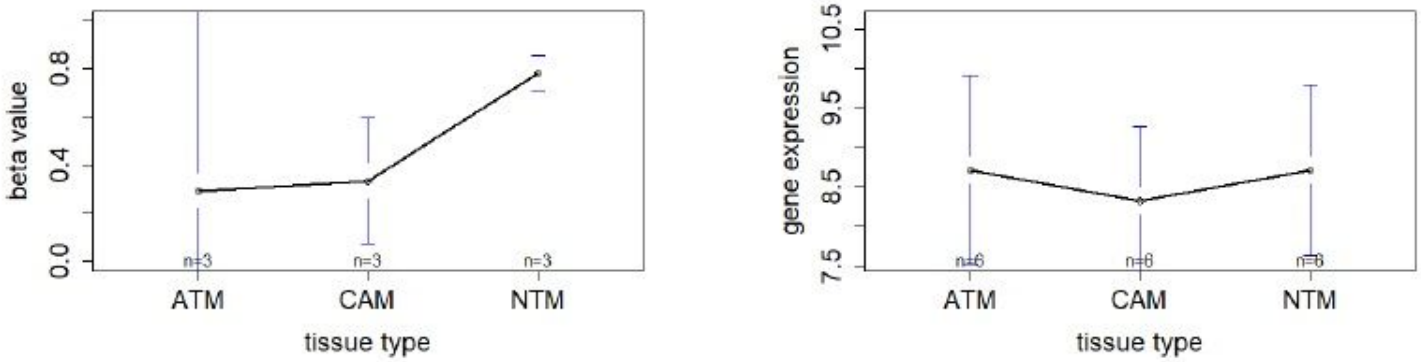



**Figure 11**

The overall methylation and gene expression levels of the TMTC1 promoter region.

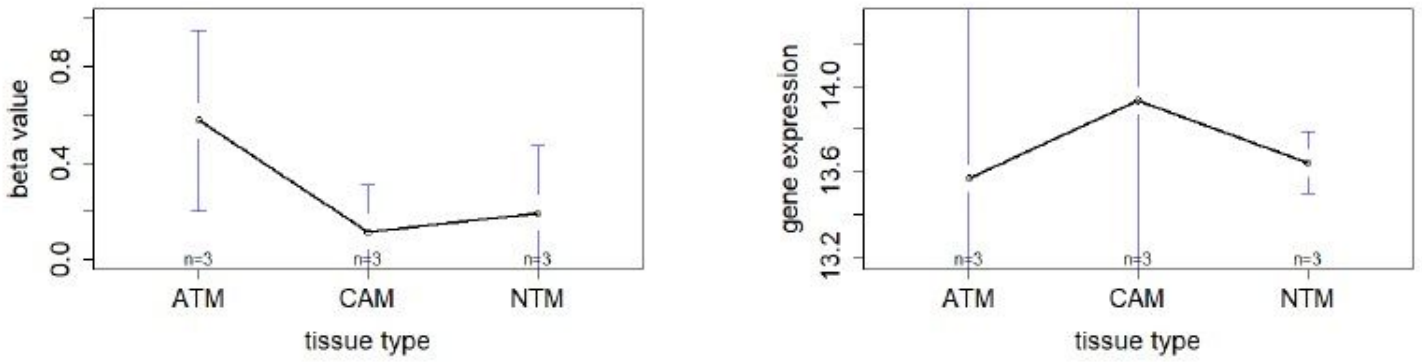



**Figure 12**

The overall methylation and gene expression levels of the UCHL1 promoter region.

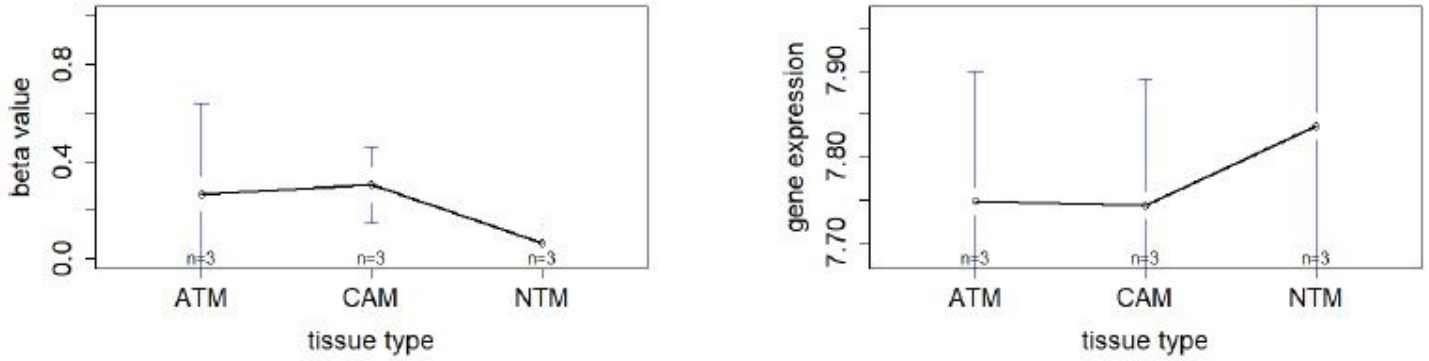



**Figure 13**

The overall methylation and gene expression levels of the FOXP2 promoter region.

# Supplementary Files

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

- BACkPAySupplement.pdf