

國立臺灣大學公共衛生學院流行病學與預防醫學研究所

碩士論文

Institute of Epidemiology and Preventive Medicine

College of Public Health

National Taiwan University

Master Thesis

以貝氏模式利用條件自迴歸分布尋找甲基化誘導之變異基因

Identification of Methylation-driven Genes with Bayesian
Conditional Autoregressive Model

鄭璽容

Shi-Jung Cheng

指導教授：蕭朱杏 博士

盧子彬 博士

Advisor: Chuhsing Kate Hsiao, Ph.D.

Tzu-Pin Lu, Ph.D.

中華民國 104 年 7 月

July, 2015



國立臺灣大學碩士學位論文
口試委員會審定書

論文中文題目

以貝氏模式利用條件自迴歸分布尋找甲基化
誘導之變異基因

論文英文題目

Identification of Methylation-driven Genes with
Bayesian Conditional Autoregressive Model

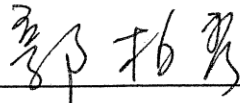
本論文係 鄭璽容 君 (學號 R02849050) 在國立臺灣大
學流行病學與預防醫學研究所完成之碩士學位論文，於民國
104 年 7 月 24 日承下列考試委員審查通過及口試及格，特此
證明。

口試委員：

  (簽名)

(指導教授)





中文摘要



近年來許多相關性研究(association studies)多專注在多基因平台資料(multi-platform genetic data)的整合式分析(integrative analysis)，此類型的研究除了可以包含不同平台資料(如 DNA 與 gene expression)代表的不同生物意義，還可以避免只利用單一平台進行分析的一些缺點，如遺傳力的解釋不佳、模型包含的資訊不足、以及研究成果難以重現等。除此之外，多平台資料的分析還可以讓科學家有機會探討不同平台的基因標記彼此之間互動的情形。

再者，單一平台的基因資料通常都是高維度數據，因此在統計分析上多為單一標記基因的檢定(single-marker test)，這類方法不但忽略同一平台內基因之間的交互作用，也可能面臨多重檢定(multiple tests)所導致的檢力的不足。因此，有些學者發展出同時考慮一組基因的方法，例如以基因集合為主的分析(gene set-based analysis)或以生物路徑(pathway analysis)為主的分析，以降低資料的維度。使用生物路徑的好處是，如此可以瞭解哪些基因參與了特定的細胞功能並且如何相互影響。換句話說，藉由生物路徑，我們可以建立基因之間的交互作用同時保留生物上的解釋意義。

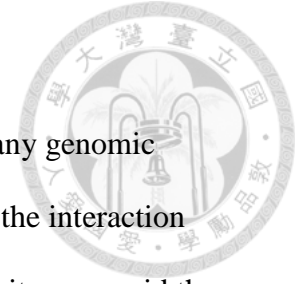
本論文為了在相關性研究中考慮基因之間的關係，並且不侷限於單平台資料，提出了一個貝氏模式，並且以條件自迴歸分佈(conditional autoregressive model)來處理基因間在生物路徑中的關係。這個自迴歸分佈能同時整合受 DNA 甲基化影響的基因、與 RNA 表現的微陣列基因資料，進而偵測對疾病狀態有影響的基因。最後，我們利用卵巢癌的存活資料來示範這個統計模式。實際資料分析的結果顯示，這個模型可以偵測在一個生物路徑中對疾病的存活有影響的基因，其中有些基因與疾病的相關性已經被其他研究學者報導過，其他基因則可能成為未來其他生物

實驗室研究的候選基因。

關鍵字：基因表現、DNA 甲基化、生物路徑、微陣列



ABSTRACT



Multiple-platform analysis has recently become the focus of many genomic research projects. Such analysis offers an opportunity to account for the interaction between genetic observations from different platforms. Additionally, it may avoid the problems encountered in the analysis with single platform genetic markers, such as low heritability, limited information and failure in reproducing findings.

Another problem faced in association studies is the fact that genetic data are often high-dimensional, and thus the most common approaches are single-marker tests. These tests cannot consider gene-gene interaction, and can lead to low statistical power due to corrections for multiple tests. An alternative is to consider sets of genes such as gene set-based analysis or pathway analysis. Through pathways, the knowledge as which genes participate in certain functions and how these genes interact with each other can then be used to construct the relations between genes in statistical analysis, while reserving the biological meaning at the same time.

In this thesis, we propose a Bayesian model with a conditional autoregressive distribution to address the relations among genes in a given pathway. This model also integrates DNA methylation and RNA expression microarray data to detect influential genes. We next illustrate this Bayesian model with an ovarian cancer study. Several influential genes are identified, where some of them have been reported earlier. Finally, we discuss issues and applicability of this proposed model for genetic association studies.

Keywords : gene expression, DNA methylation, pathways, microarray

CONTENTS



中文摘要	i
ABSTRACT	iii
CONTENTS	iv
LIST OF TABLES	v
LIST OF FIGURES	vi
Introduction	1
Methods	5
Real application	10
Results	15
Discussion	20
REFERENCE	24
APPENDIX	37

LIST OF TABLES

Table 1. Background information of clinical data	27
Table 2. Summary statistics of age and survival time	27
Table 3. Selected posterior probabilities of individual gene effects	28
Table 4. Identified influential genes	29



LIST OF FIGURES

Figure 1. Examples of pathways	30
Figure 2. Examples of nodes in KEGG pathways	31
Figure 3. The 95% credible intervals of the coefficients beta	33
Figure 4. Flowchart of the data management procedure	33
Figure 5. Selected genes in the cell cycle pathway	34
Figure 6. Density plots of coefficients γ^+ , γ^- , δ^+ , δ^- , η	35
Figure 7. The 95% credible intervals for the coefficients beta derived from 30000 (red) or 50000 (blue) iterations.....	36




Introduction



The development of gene technology has progressed rapidly since 1990s. Since then, the association studies have been one of the major research focuses. These studies have great contributions to predictions or progression for certain inherited diseases. For instant, the gene *BRCA1* and gene *BRCA2* are notorious in increasing the risk of breast cancers. However, most genetic studies concentrate on constructing a relevance between the target disease and data merely from a single platform (such as RNA expression, SNPs, or copy number variation), leading to problems including low heritability, limited information and difficulties in replicating research results. Especially in cancer studies, the etiology and pathophysiology of cancer are so complicated that it is hard to explain the mechanism through only the information from one metabolic stage of genome. Consequently, many researchers have turned their attentions to multiple-platform analysis.

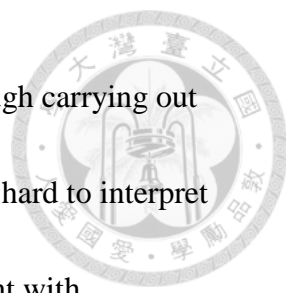
Multiple-platform analysis has many advantages. First, it contains more information with more possible biological interpretations. Next, it offers an opportunity to clarify the interaction between genetic markers from different platforms, which may play a critical role in the occurrence of diseases. Current integrative analysis can be divided into three categories (Wang *et al.*, 2013), sequential integration, model



integration and biological integration. Sequential integration studies, like eQTL, focus on sequentially screening out relevant genetic markers from different platforms. This integrative approach may lose important information when conducting filtering at different stages.

Also, it ignores the interactions between multiple platforms. The second group of integrative analysis, called model integration, aims at building a statistical model to combine information from different platforms. Ray *et al.* (2014) applied joint Bayesian factor analysis to integrate data from different platforms to detect significant disease-related genetic markers. But such methods could encounter difficulties in interpretations, if the biological relationship between different platforms was not considered when establishing the analytic model. The third group, biological integration takes the biological pathways and mechanism into account, while including data from different platforms into analysis. The results of these approaches are more biological interpretable.


Genetic data are known to be high-dimensional, which also contributes to difficulties in analysis. Some analysis for association studies considered single marker tests to detect disease-association genes. As describe earlier, this kind of approaches not only discarded gene-gene integrations, but were also of low power because of multiple testings. Another choice is the approach of dimension reduction (here means extract



information from thousands of variables to fewer components). Though carrying out dimension reduction could decrease the complexity for analysis, it is hard to interpret the results from biological viewpoint. In addition, we need to confront with time-consuming computations. Alternative approach is to consider sets of genes such as pathways. Pathways can be considered as a map of biology mechanism that has particular functions in an organism. Through pathways we could understand which genes participate in this biological activity and how these genes interact with other genes. With this knowledge, we are able to construct appropriate relations between genes in analysis and reserve its biology meanings in the meantime. Furthermore, it reduces largely the number of variables when thousands of genes are classified into different pathways, which brings the convenience in statistical computation.

Pathway Topology (PT)-Based Approaches (Khatri, 2012), one approach of pathway analysis, try to incorporate information of pathway topology to detect disease-related pathways. In the respect, Chang (2014) proposed a model which is able to consider gene effects in pathways by giving each gene a different weight related to the number of its neighboring genes. However, this approach treated every gene in the pathways as equal, and did not consider the interaction between this gene and its neighbors.

In this thesis, following the spirits to contain the information of pathway topology



and to consider multi-platform genetic markers, we propose a Bayesian model to integrate DNA methylation and RNA expression microarray data to detect important genes. Previous research has showed that the change of DNA methylation serves as a good biomarker for disease diagnoses and disease progressions in different cancers (Heyn and Esteller, 2010). Additionally, DNA methylation and RNA expression are regarded as adjustments in different stages. DNA methylation located on the upstream of gene performances is the DNA level of epigenetic regulations, while RNA expression is the downstream of gene performances, indicating different biological functions. In this model, we use the conditional autoregressive model to describe the interaction between genes and the effects from DNA methylation on gene expression. In addition, we illustrate this model with a study of ovarian cancer from The Cancer Genome Atlas (TCGA). The data contain cases with DNA methylations and RNA expression levels. The identified influential genes are discussed and compared with earlier findings. Finally, we discuss issues in this approach and applicability for other genetic association studies.



Method

For each pathway, we construct a regression model for genes in a given pathway.

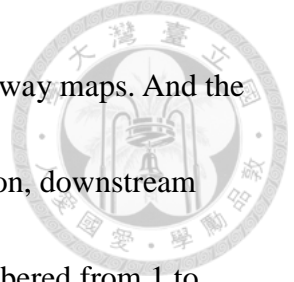
Suppose the total number of patients is n and the total number of genes is p . Let Y_i be the clinical outcome of interest for the i^{th} subject and X_i denote the gene expressions of the i^{th} subject. The expected value of Y_i conditioned on X_i can be written as

$$g(E[Y_i | x_i]) = \sum_{j=1}^p \beta_j x_{ij} + a_0, \quad x_i \sim MVN(\Delta, \Sigma_x)$$

$$X = \begin{bmatrix} x_1^t \\ x_2^t \\ \vdots \\ x_n^t \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & & x_{1,p} \\ x_{2,1} & x_{2,2} & & \\ & & x_{i,j} & \\ x_{n,1} & & & x_{n,p} \end{bmatrix}_{n \times p}$$

where X_i is composed of x_{ij} , indicating the gene expression of the j^{th} genes for the i^{th} subject. It follows a multivariate normal distribution which mean vector Δ and covariance matrix Σ_x . Its setting will be described in details later. First, we separate genes into two groups. One is genes significantly regulated by DNA methylation (genes whose DNA methylation and gene expression are negative), and the other group contains genes without direct regulation effects from DNA methylation (genes whose DNA methylation and gene expression are not negative).

For the first group, we construct its x_{ij} through the steps below: Given a gene j , its gene expression level is influenced by its adjacent genes in the pathway. The definition



of adjacent genes is genes which have one branch with gene j in pathway maps. And the regulation can be divided into upstream activation, upstream inhibition, downstream activation and downstream inhibition. Among all the neighbors (numbered from 1 to $N_A+N_B+N_C+N_D$) suppose N_A neighbors are upstream activation, associating with an effect γ^+ ; N_B neighbors are upstream inhibition, associating with an effect γ^- ; N_C neighbors are downstream activation, associating with an effect δ^+ ; N_D neighbors are downstream inhibition, associating with an effect δ^- . In this group, genes are significantly modulated by their DNA methylation levels. So we add m_j , the DNA methylation level of gene j , with a parameter η to model the effect from DNA methylation. Therefore, x_{ij} would follow a normal distribution with the mean μ equaling

$$\frac{\gamma^+ (gene_{N_{A1}} + \dots + gene_{N_A}) + \gamma^- (gene_{N_{B1}} + \dots + gene_{N_B})}{N_A + N_B + N_C + N_D + 1} + \frac{\delta^+ (gene_{N_{C1}} + \dots + gene_{N_C}) + \delta^- (gene_{N_{D1}} + \dots + gene_{N_D}) + \eta m_j}{N_A + N_B + N_C + N_D + 1}$$

and variance σ^2 equaling $G^2 / (N_A + N_B + N_C + N_D + 1)$.

For the second group, the rule for model construction is similar, but this group includes genes that are not significantly modulated by their DNA methylation levels. Therefore, the parameter η with the effects m_j becomes null. And x_{ij} would follow a normal distribution with μ equaling



$$\frac{\gamma^+ (gene_{N_A} + \dots + gene_{N_{A1}}) + \gamma^- (gene_{N_B} + \dots + gene_{N_{B1}})}{N_A + N_B + N_C + N_D} + \frac{\delta^+ (gene_{N_C} + \dots + gene_{N_{C1}}) + \delta^- (gene_{N_D} + \dots + gene_{N_{D1}})}{N_A + N_B + N_C + N_D}$$

and σ^2 is $G^2 / (N_A + N_B + N_C + N_D)$.

The complete model can be expressed as

$$T_i | x_i \sim Weibull(\text{shape}, \lambda_i)$$

$$\lambda_i = \exp\left(\sum \beta_j x_{ij}\right)$$

$$X_i = \begin{bmatrix} x_1^i \\ x_2^i \\ \vdots \\ x_n^i \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1,p} \\ x_{21} & x_{22} & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & & & x_{n,p} \end{bmatrix}_{n \times p}$$

$$x_i \sim MVN(\Delta, \Sigma x)$$

$$x_{ij} | x_{i,-j} \sim N\left[A + B + \eta m_{ij} \times M(x_j)\right] / num_j, G^2 / num_j$$

$$A = \gamma^+ \sum_{j \in C_{us.a}(j)} x_{ij} + \gamma^- \sum_{j \in C_{us.i}(j)} x_{ij}$$

$$B = \delta^+ \sum_{j \in C_{ds.a}(j)} x_{ij} + \delta^- \sum_{j \in C_{ds.i}(j)} x_{ij}$$

$$num_j = N_{us.a} + N_{us.i} + N_{ds.a} + N_{ds.i} + M(x_j)$$

where

T_i is the survival time of the i^{th} patient

C_i is the censoring time of the i^{th} patient

Y_i is the minimum of T_i and C_i

$Z_i = I(T_i \leq C_i), i = 1, \dots, n$

x_i is the i^{th} person's gene expressions

n is the number of total cases (patients), p is the number of total genes



β is a vector with dimension p by 1

x_{ij} means the gene expression of j^{th} gene in i^{th} patient

$C_{\text{us.a}}(j)$ denotes the set of upstream activation genes of j^{th} gene

$C_{\text{us.i}}(j)$ denotes the set of upstream inhibition genes of j^{th} gene

$C_{\text{ds.a}}(j)$ denotes the set of downstream activation genes of j^{th} gene

$C_{\text{ds.i}}(j)$ denotes the set of downstream inhibition genes of j^{th} gene

$N_{\text{us.a}}$ denotes the number of upstream activation genes of j^{th} gene

$N_{\text{us.i}}$ denotes the number of upstream inhibition genes of j^{th} gene

$N_{\text{ds.a}}$ denotes the number of downstream activation genes of j^{th} gene

$N_{\text{ds.i}}$ denotes the number of downstream inhibition genes of j^{th} gene

$M(x_j) = I(x_j \text{ belongs to genes that significantly regulated by DNA methylation})$

The prior distributions are

$$shape \sim gamma(3, 2)$$

$$\beta_j \sim N(0, 1), j \text{ from } 1 \text{ to } P$$

$$\gamma^+ \sim N(0, 1),$$

$$\gamma^- \sim N(0, 1),$$

$$\delta^+ \sim N(0, 1),$$

$$\delta^- \sim N(0, 1),$$

$$\eta \sim N(-1, 100),$$

$$G^2 \sim gamma(2, 5)$$

Toy example

Following we provide an example to demonstrate the procedures when

constructing our model:

Suppose there are A, B,.....H genes and 10 patients. Define the gene expression

levels x_{ij} for these patients as

$$\mathbf{X} = \begin{bmatrix} x_{1,A} & x_{1,B} & & x_{1,H} \\ x_{2,A} & x_{2,B} & & \\ & & x_{i,j} & \\ x_{10,A} & & & x_{10,H} \end{bmatrix}_{10 \times 8} \quad j = A, B, C, \dots, H, \text{ for the } i^{\text{th}} (i=1,2,\dots,10)$$

patients

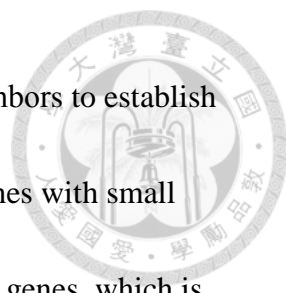
Given the pathway in Figure 1 (a), the model is written as:

$$\begin{aligned} x_{iA} | x_{i,-A} &\sim N([\gamma^+(x_{iB} + x_{iC} + x_{iD}) + \gamma^-(x_{iE}) + \delta^+(x_{iF}) + \delta^-(x_{iG}) + \eta m_{iA}] / 7, G^2 / (3+1+1+1+1)) \\ x_{iB} | x_{i,-B} &\sim N([\delta^+(x_{iA}) + \eta m_{iB}] / 2, G^2 / (0+0+0+1+1)) \\ x_{iC} | x_{i,-C} &\sim N([\delta^+(x_{iA}) + \eta m_{iC}] / 2, G^2 / (0+0+0+1+1)) \\ x_{iD} | x_{i,-D} &\sim N(\delta^+(x_{iA}), G^2 / (0+0+0+1+0)) \\ x_{iE} | x_{i,-E} &\sim N(\delta^-(x_{iA}), G^2 / (0+0+0+1+0)) \\ x_{iF} | x_{i,-F} &\sim N(\gamma^+(x_{iA}), G^2 / (1+0+0+0+0)) \\ x_{iG} | x_{i,-G} &\sim N([\gamma^-(x_{iA}) + \delta^+(x_{iH})] / 2, G^2 / (0+1+1+0+0)) \\ x_{iH} | x_{i,-H} &\sim N(\gamma^+(x_{iG}), G^2 / (1+0+0+0+0)) \end{aligned}$$

Given the pathway in Figure 1 (b), the model is written as:

$$\begin{aligned} x_{iA} | x_{i,-A} &\sim N([\delta^+(x_{iD}) + \delta^-(x_{iF}) + \eta m_{iA}] / 3, G^2 / (0+0+1+1+1)) \\ x_{iB} | x_{i,-B} &\sim N([\delta^-(x_{iD}) + \eta m_{iB}] / 2, G^2 / (0+0+0+1+1)) \\ x_{iC} | x_{i,-C} &\sim N([\delta^+(x_{iE}) + \eta m_{iC}] / 2, G^2 / (0+0+0+1+1)) \\ x_{iD} | x_{i,-D} &\sim N([\gamma^+(x_{iA} + x_{iE}) + \gamma^-(x_{iB}) + \delta^-(x_{iG})] / 4, G^2 / (2+1+0+1+0)) \\ x_{iE} | x_{i,-E} &\sim N([\gamma^-(x_{iC}) + \delta^+(x_{iD})] / 2, G^2 / (0+1+1+0+0)) \\ x_{iF} | x_{i,-F} &\sim N([\gamma^-(x_{iA}) + \delta^+(x_{iG})] / 2, G^2 / (0+1+1+0+0)) \\ x_{iG} | x_{i,-G} &\sim N([\gamma^+(x_{iD} + x_{iF}) + \delta^+(x_{iH})] / 3, G^2 / (2+0+1+0+0)) \\ x_{iH} | x_{i,-H} &\sim N(\gamma^+(x_{iG}), G^2 / (1+0+0+0+0)) \end{aligned}$$





In this model, we take the gene expression of the j^{th} gene's neighbors to establish the distribution of the j^{th} gene. There exist some highly expressed genes with small variation across samples. These genes often belong to house-keeping genes, which is not likely to be our target genes (disease-related genes). They may enlarge the mean of the distribution, and lead to misleading influence of genes. To solve the problem, we consider the coefficient of variation (CV) of each gene at first. If the CV of gene expression is larger than a certain value, then we directly utilize its value of gene expression. Here we use the 10th quantile of CV from all samples as the threshold. Detailed descriptions are stated in the result section. If the CV is smaller than the threshold and the values of gene expression are comparably high across all genes (here we use the 90th quantile of mean of all samples as a threshold), we substitute its value with the median value across all genes and all samples.

Real application

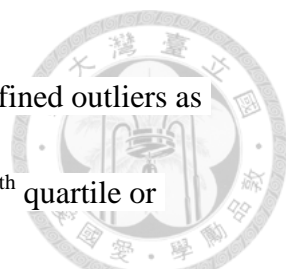
The data of ovarian cancer were downloaded from The Cancer Genome Atlas (TCGA) (<http://cancergenome.nih.gov/>), a NIH website that contains genomics datasets for over 20 types of cancers. In our analysis, we adopt the gene expression (BI

HT_HG-U133A, Level 2) and DNA methylation data (JHU-USC HumanMetylation27, Level 3) of ovarian cancer. Only samples that have both gene expression and methylation data are included in our study. There were 585 cases in total. The steps for data management are described in the following and displayed in Figure 5:

(1) we removed samples with Recurrent Solid Tumor (17 cases) and samples with Solid Tissue Normal (8 cases). The remaining 560 cases were samples with Primary Solid Tumor. A case with duplicate ID was found in the 560 cases, which might indicate that this person has her tissue scanned by microarray twice. However, these two results of scanning were generally consistent, so we deleted one of these cases. It led to 559 cases.

(2) Outliers detection

The signals of gene expression Level 2 data downloaded from TCGA had been already normalized per probe or probe set for each participant's tumor sample. Therefore, our quality control step only aimed at DNA methylation data. First, outliers in each probe were detected and tagged with “1” with boxplot. We defined outliers as data points that were larger than 1.5 times IQR above the 75th quartile or smaller than 1.5 times IQR below the 25th quartile. Next, we calculated how many tags there were for each case. If the case had an extremely large number of tags, say larger than 1.5 times IQR beyond the third



quartiles, the case would be removed (Chang, 2014). (we defined outliers as data points that were larger than 1.5 times IQR above the 75th quartile or smaller than 1.5 times IQR below the 25th quartile.) This excluded 76 samples with extreme “outlying numbers”. At this step, there were 483 cases left.

(3) Quantile normalization (batch effect), clinical information check

In order to eliminate batch effects in DNA methylation, we used quantile normalization aimed at DNA methylation by *normalize.quantiles* function in *preprocessCore* package in R. Next, we matched remaining cases to clinical data, where the clinical data were updated on March 9, 2015. We removed 1 cases without clinical data, 4 cases whose tumor_tissue_site were not Ovary and 32 cases that missed either information of race, vital_status or clinical_stage. At this step, there were 446 cases left.

Because the races and the clinical stage might have impacts on gene levels, we tabulate the frequency tables of race and clinical stages in the 446 cases. We found that a high percentage of cases was white and in high stage. We next narrowed down our analysis to cases that were white and whose stages are between IIIA to VI. This leads to 377 cases for further analysis.

(4) Match gene and gene expression, gene and DNA methylation

In the downloaded DNA methylation data, probes were arranged by their gene



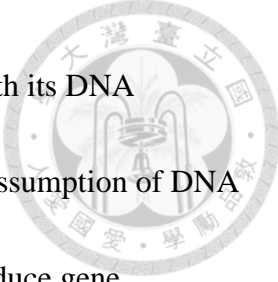
codings. If a gene corresponds to multiple probes, we took the average of all listed probes to be the DNA methylation level of this gene. For gene expression data, the gene coding was referred to the index provided by Affymetrix

(<http://www.affymetrix.com/support/technical/byproduct.affx?product=hgu133>

). Similarly, if a gene corresponded to multiple probes, we took the average of all related probes to be the RNA level of this genes. After arrangement, DNA methylation dataset contained 14310 genes; RNA expression dataset contained 14117 genes. We excluded the genes that only appear in either dataset. Finally, there were 10282 identical genes that could be used in further analysis.

The pathway maps can be downloaded from KEGG. Details about how we arrange pathway information are described in Appendix 2. Here we choose the cell cycle pathway (hsa04110) to demonstrate our model. The reason why we select this pathway is that the pathway has been reported to be associated with ovarian cancer in Fu's study (Fu and Wang, 2013). Also, it has significant pathway effects in Chang's model (Chang, 2014).

Noticed that previously we separate genes into two groups, genes having strong negative association with its DNA methylation, and genes not. The cutoff point is whether the correlation between DNA methylation level and gene expression for this gene is smaller than -0.1. If the correlation is smaller than -0.1, then we classify this




gene to the group in which genes have strong negative association with its DNA methylation. The aim of the classification is to model the biological assumption of DNA methylation mechanism. Increasing DNA methylation level would reduce gene expression owing to the difficulties of RNA polymerase's binding due to DNA methylation (Chen and Pikaard, 1997). That is, the correlation between DNA methylation level and gene expression is supposed to be negative. However, when we examine the correlations of each gene, about 40 percent is positive. This may indicate that there exists other regulation effect that cannot be explained only by DNA methylation effects. Since we consider only pure effects resulting from DNA methylation, only those with correlation smaller than -0.1, their DNA methylation are included into our model.

Computation

The final data included 377 ovarian cancer cases and 41 nodes (can be regarded as 41 genes or gene complex) in the cell cycle pathway. Our dependent variables were survival times of each patient, and correspondingly our link function was survival functions. If the variable "Death_days_to" of the i^{th} person is available, then the survival time (T_i) would be the value of "Death_days_to"; on the contrary, if the variable "Death_days_to" of the i^{th} person is not available, then we define the case as censored and take the value of the variable "Last_contact_days_to" to be the censoring time (C_i) of

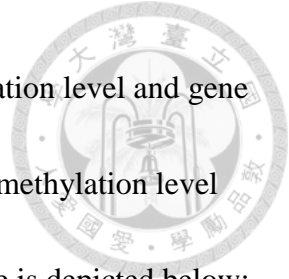
the i^{th} person. Additionally, we use Weibull distribution to fit likelihood functions.



The computation process was carried out by R package version 3.1.2. The posterior samples of parameters were derived from Markov chain Monte Carlo (MCMC) algorithm using R2OpenBUGS package in R. We simulate one chain, 30000 iterations. After burn-in 5000 samples, every 1 in 20 posterior sample was filtered for further analysis. We have compared the results of simulating 30000 iterations and 50000 iterations (Figure 7.) and found the coefficients of beta are very similar. Thus, the consequences of 30000 iterations are presented in this article. The computation time of 30000 iterations was around three to 11 hours for one pathway. The R codes are in Appendix 3. The convergence was checked by MC errors and trace plots.

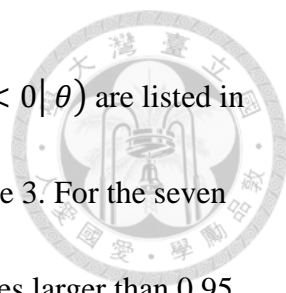
Results

The density plots of posterior samples of γ^+ , γ^- , δ^+ , δ^- , η are displayed in Figure 6. It shows that γ^+ , γ^- , δ^+ , δ^- are all positive but with different scales and centers, indicating that the data support the four categories of relations in pathways. The coefficient η is supposed to be negative because it represents the relations between methylation levels and gene expression levels. However, the posterior samples of η ranges from -0.2 to 0.4.




As previously described, when the correlation between DNA methylation level and gene expression for a specific gene is smaller than -0.1 , we would put the methylation level of the gene into our model. But if the node is a complex, the principle is depicted below: assumed that a gene complex contains three genes A, B and C. If the correlation in gene A and in gene B are smaller than -0.1 , while that in gene C is not, then we would take the mean of the DNA methylation levels of gene A and gene B as the methylation level of this complex; if the correlation only in gene A is smaller than -0.1 , then we would take the DNA methylation levels of gene A as the methylation level of this complex.

After the arrangement, we once again check the correlations between the gene expression level and DNA methylation level of every gene (or gene complex). For all genes that we consider with methylation effects, only one correlation is positive, the others are negative, as expected. Therefore, the arrangement may not be the reason why η ranges from -0.2 to 0.4 , and not in the negative domain. Another explanation may be that when we are calculating the correlations, we consider the marginal distribution of gene expression and DNA methylation level. In contrast, in the model the DNA methylation effects are added in the conditional distributions of x (the gene expressions). The approach to average methylation effects is intuitive and easy for analysis. Such calculation, however, may lose focus of any individual methylation levels. How to consider the DNA methylation level in a complex needs more discussions.



The estimated posterior probabilities $P(\beta_j > 0 | \theta)$ and $P(\beta_j < 0 | \theta)$ are listed in Table 3. And the 95% credible intervals of each β_j are drawn in Figure 3. For the seven genes (or gene complex) with the corresponding posterior probabilities larger than 0.95, the strength of evidence that these genes are likely to associate with patients' survival time is strong. We have tried -0.1 and -0.05 as the cutoff values when determining the inclusion of DNA methylation effects, and the results of the posterior probabilities β_j do not alter drastically. Here we only display the results with -0.1, other results are in Appendix 4.

The significant genes include *CDKN1A*, *MDM2*, gene complex *APC/C /CDC20*, *ATM/ATR*, complex *E2F4/E2F5/RBL1/TFDP1/TFDP2*, *RBI* and *ZBTB17*. Among them, many are consistent with reports from previous literatures. *CDKN1A* has been proved to associate significantly with the increase the hazard of breast cancer patients in Györffy's study (2010), which is a strong evidence because the breast cancer is categorized to gynecological diseases, the same as our target ovarian cancers. And by Ma *et al.* (2011), *CDKN1A* also has been detected to associate with the survival of non-small cell lung cancer. It has been reported that *CDC20* predicts poor prognosis in non-small cell lung cancer patients (Kato *et al.*, 2002), colorectal cancer patients (Wu *et al.*, 2013) and patients with breast cancer (Karra *et al.*, 2014); the expression of *ATM/ATR* will increase after DNA damage, which is an important checkpoint in cell cycle pathway



(Reinhardt *et al.*, 2007). Furthermore, Ye *et al.* (2007) found that expression patterns of gene *ATM* associate with breast cancer survival, and Grabsch *et al.* (2006) reported that expression of gene *ATM* predicts patient survival in colorectal cancer. Speaking to the complex *E2F4/E2F5/RBL1/TFDP1/TFDP2*, the *E2F* family plays a crucial role in the control of cell cycle G2 phase and repress the expression of gene *MYC*, an important regulator in cycle progression and having proved to associate with patients survival in breast cancers (Xu *et al.*, 2010) and lung cancer (Borczuk *et al.*, 2004). Additionally, pervious literatures have shown that over-expressed *TFDP1* associates with progression of hepatocellular carcinomas (Yasui *et al.*, 2003). *RBI* is a famous tumor suppressor gene and functions as a negative regulator in cell cycle pathway. In addition, it has been shown to associate with poor prognosis in patients with non-small cell lung cancer (Zhao *et al.*, 2012).

On the other hand, there is no research report about the association between the expression of *ZBTB17* and complex *MDM2* and cancer survival time. However, it has been reported that *ZBTB17* is involved in the regulation gene *MYC* (Staller *et al.*, 2001); and *MDM2* promotes tumor formation by targeting tumor suppressor proteins like *TP53* (Haupt *et al.*, 1997).

The position of significant genes in the pathway is shown in Figure 5. Generally, they are located at upstream or midstream of the target pathway.



We constructed a survival regression with the same data, as expressed below:

$$h(t) = h_0(t) \exp(\sum \beta_j x_j)$$

$$h_0(t) = \lambda \times shape \times t^{shape-1}$$

where

t is the survival time

β_j is the coefficient of x_j

x_j , means the gene expression of j^{th} genes

The coefficients of the four genes (*CDKN1A*, *GSK3B*, *RBI* and *MYC*) were significant. In this and the above model, *RBI* and *CDKN1A* were significant. Other significant genes in multiple regression model included *MYC* and *GSK3B*. The coefficient of *MYC* in our model also has high posterior probability (0.94). However, as we examine the position of *GSK3B* in the cell cycle pathways, we found *GSK3B* inhibits the complex *CCND/CDK4,6*, where this complex is regulated by many other genes, such as *CDKN2A*, *CDKN2B*, *CDKN2C*, *CDKN2D* and *PCNA*. It indicates that multiple regression approach may detect significant genes, but cannot detect important genes when we consider the pathway information.

Discussion



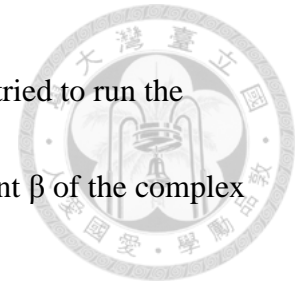
In summary, our proposed model has two advantages. First, it is flexible in in defining functions for the links in pathways. Second, it can be applied to different interested outcomes. In addition to the survival function as demonstrated here, the model can be applied to binary outcomes, such as case control studies or effectiveness studies of medicines, by just changing to logit link or others. Thirds, it efficiently exploits the information in pathways, which may find important disease markers in the biological functions.

Different criteria for trimming data

As previously described, to avoid some house-keeping genes dominating the average value, the gene expression levels may be replaced if its CV is smaller than a certain value. In the model, if the CV of a gene is smaller than the 10th quantile of all samples and the mean gene expression level is larger than the 90th quantile of all samples, its gene expression level is replaced with the median across all genes and all samples. We have checked that both house-keeping genes *ACTB* and *GAPDH* are trimmed with the criteria adopted here.

With this criteria, there is one gene (*SKP1*) being trimmed in cell cycle pathway. We also tried to trim with the 5th quantile of CV of all samples, the results are similar to

that with the 10th quantile as the cutoff point. Additionally, we have tried to run the program with non-trimmed data. The only difference is the coefficient β of the complex *SKP1/SKP2/CUL1/RBX1*. However, its significance is not changed.



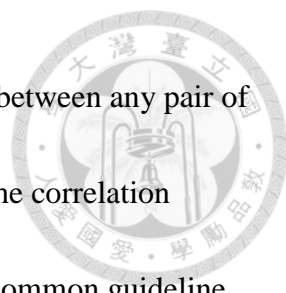
The criteria for outlier detection

For preliminary data process, we choose boxplots to detect outliers. We have also compared the excluded samples by other methods, such as hampel identifier and standardization method (Ben-Gal, 2005; Rousseeuw and Hubert, 2011). The results did not differ much.

Quantile normalization

We have considered the quantile normalization to eliminate batch effects across samples. Some people will trim off five to ten percent data when doing quantile normalization for avoiding using extreme values (Kroll and Wolf, 2002). However, we have excluded the cases with extreme outlier numbers before we did quantile normalization. Therefore, we did not trim data further when doing quantile normalization.

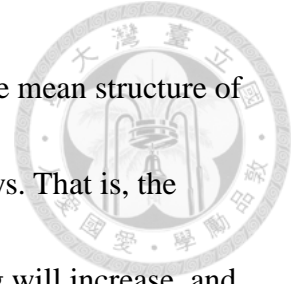
There are issues that need to be addressed. First, other ways to deal with the DNA methylation level of gene complex can be applied. For example, if the gene complex contains three genes, then the determination of whether the complex is DNA-methylated



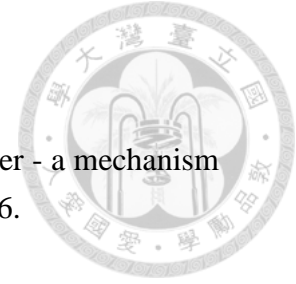
needs clarification. One way is to examine the minimum correlation between any pair of DNA methylation level and gene expression. Other choices may be the correlation between the average methylation level and expressions. There is no common guideline in current practice. Second, in the current model we assume that the DNA methylation of a gene can only influence its own gene expression level. This excludes the case where the DNA methylation may affect other genes. A more general formulation of the autoregressive model to accommodate this phenomenon is possible. However, it may come with the price of computational burden. Third, how to model other regulation effects such as phosphorylation can be considered. In our current model, all activation activities are treated equally. In the future, one can consider whether to distinguish these activation effects with different parameters. Fourth, pathway information is one kind of interactions between genes. To completely understand and model the gene-gene interactions, pathways information may not be enough. If there are different types of data which can convey more interaction information between genes, this should be incorporated into analysis.

The proposed model is designed to be used in the case where the detection of important genes in a given pathway is of interest. This pathway can be selected based on previous analysis, such as the online pathway analysis DAVID or KEGG, or from previous knowledge. If one is interested in incorporating several pathways at the same

time, the current model needs further modifications. For example, the mean structure of each gene expression can contain information from multiple pathways. That is, the pathway effects can be additive. However, the computational loading will increase, and a more efficient algorithm would be worth pursuing.



RERERENCE



Baylin, S.B. and Ohm, J.E. (2006) Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction. *Nat. Rev. Cancer*, 6: 107–116.

Berns, EMJJ. et al. (1996) TP53, and MYC gene alterations independently predict poor prognosis in breast cancer patients. *Genes Chromosomes Cancer*, 16: 170–179.

Ben-Gal, I. E. (2005) Outlier Detection. *The Data Mining and Knowledge Discovery Handbook*, 131–146.

Borczuk, A.C. et al. (2004) Molecular signatures in biopsy specimens of lung cancer. *Am J Respir Crit Care Med*, 170: 167–174.

Broggini, M. et al. (2000) Cell cycle-related phosphatases CDC25A and B expression correlates with survival in ovarian cancer patients. *Anticancer Res*, 20: 4835–4840.

Chang, C.W. (2014) Integrating CGI and Pathway Information in Analysis of Differential DNA Methylation Profiling in Ovarian Cancer. M.S. thesis, University of Taiwan.

Chen, Z.J. and Pikaard, C.S. (1997) Epigenetic silencing of RNA polymerase I transcription: a role for DNA methylation and histone modification in nucleolar dominance. *Genes Dev*, 11: 2124–2136.

Chuang, H.Y. et al. (2010) Pathway-based modeling and diagnosis of cancer development and progression. Ph.D. dissertation, University of California, San diego.

Coleman, T.R., Carpenter, P.B., and Dunphy, W.G. (1996) The Xenopus Cdc6 protein is essential for the initiation of a single round of DNA replication in cell-free extracts. *Cell*, 87: 53–63.

Fu, L.J. and Wang, B. (2013) Investigation of the hub genes and related mechanism in ovarian cancer via bioinformatics analysis. *Journal of Ovarian Research*, 6: 92.

Grabsch, H. et al. Expression of DBA doublestrand break repair proteins ATM and BRCA1 predicts survival in colorectal cancer. *Clin Cancer Research*, 12: 1494–1500.

Godlewski, J. et al. (2010) microRNA-451: A conditional switch controlling glioma cell proliferation and migration. *Cell Cycle*, 9: 2742–2748.

Györfy, B. et al. (2010) An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Research and Treatment*, 123: 725–731.

Haupt, Y. et al. (1997) Mdm2 promotes the rapid degradation of p53. *Nature*, 387: 296–299.

Heyn, H. and Esteller, M. (2010) DNA methylation profiling in the clinic: applications and challenges. *Nature Reviews Genetics*, 13: 679–692.

Karra, H. et al. (2014) Cdc20 and securin overexpression predict short-term breast cancer survival. *Br J Cancer*, 110: 2905–2913.

Kato, T. et al. (2012) Overexpression of CDC20 predicts poor prognosis in primary non-small cell lung cancer patients. *J Surg Oncol*, 106: 423–430.

Khatri, P. (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS computational biology*, 8: 15–22.

Kroll, T. and Wolf, S. (2002) Ranking: a closer look on globalization methods for normalization of gene expression arrays. *Nucleic Acids Res*, 30: 50.

Leone, P.E. et al. (2008) Deletions of CDKN2C in multiple myeloma: biological and clinical implications. *Clin Cancer Research*, 14: 6033–6041.

Ma, H. et al. (2011) Potentially functional polymorphisms in cell cycle genes and the survival of non-small cell lung cancer in a Chinese population. *Lung Cancer*, 73: 32–37.

Mihara, M. et al. (2001) Overexpression of CDK2 is a prognostic indicator of oral cancer progression. *Jpn J Cancer Res*, 92: 352–360.

Momand, J., Wu, H.H. and Dasgupta, G. (2000) MDM2—master regulator of the p53 tumor suppressor protein. *Gene*, 242: 15–29.

Ray, P. et al. (2014) Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30: 1370–1376.

Reinhardt, H.C. et al. (2007) p53-deficient cells rely on ATM- and ATR-mediated checkpoint signaling through the p38MAPK/MK2 pathway for survival after DNA damage. *Cancer Cell*, 11: 175–189.

Rousseeuw, P.J. and Hubert, M. (2011) Robust statistics for outlier detection. *WIREs Data Mining Knowl, Discov*, 1: 73–79.

Staller, P. et al. (2001) Repression of p15INK4b expression by Myc through association with Miz-1. *Nat. Cell Biol.*, 3: 392–399.

Takemasa, I. *et al.* (2000) Overexpression of CDC25B phosphatase as a novel marker of poor prognosis of human colorectal carcinoma. *Cancer Res.*, 60: 3043–3050.

Wang, G., et al. (2009) Osteoblast-derived factors induce an expression signature that identifies prostate cancer metastasis and hormonal progression, *Cancer Res.*, 69: 3433–3442.

Wang, W. et al. (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29, 149–159.

Wu, W.J. et al. (2013) CDC20 overexpression predicts a poor prognosis for patients with colorectal cancer. *J Transl Med.* 11: 142.

Xu, J. et al. (2010) MYC and Breast Cancer. *Genes Cancer*, 1:629–640.

Yasui, K. et al. (2003) Association of over-expressed TFDP1 with progression of hepatocellular carcinomas. *J. Hum. Genet.*, 48: 609–613.

Ye, C. et al. (2007) Expression patterns of the ATM gene in mammary tissues and their associations with breast cancer survival. *Cancer*, 109: 1729–1735.

Zhao, W. et al. (2012) Altered p16(INK4) and RB1 expressions are associated with poor prognosis in patients with nonsmall cell lung cancer. *J Oncol*, 2012: 957437.

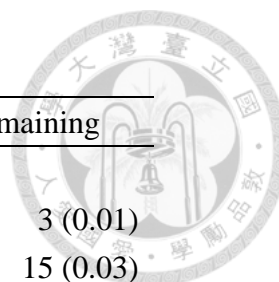


Table1. Background information of clinical data

	excluded	Remaining
Race		
American Indian or Alaska Native	0	3 (0.01)
Asian	4 (0.05)	15 (0.03)
Black or African America	4 (0.05)	20 (0.04)
Native Hawaiian or other Islander	1 (0.01)	0 (0.00)
White	67 (0.88)	414(0.92)
Missing	0	30
Vital_status		
Alive	32 (0.42)	238 (0.50)
Dead	44 (0.58)	242 (0.50)
Missing	0	2
Clinical stage		
Stage IA	0 (0.00)	3 (0.01)
Stage IB	1 (0.01)	2 (0.01)
Stage IC	0 (0.00)	10 (0.02)
Stage IIA	0 (0.00)	3 (0.01)
Stage IIB	0 (0.00)	4 (0.01)
Stage IIC	1 (0.01)	19 (0.04)
Stage IIIA	0 (0.00)	8 (0.02)
Stage IIIB	0 (0.00)	24 (0.05)
Stage IIIC	61 (0.80)	336 (0.70)
Stage IV	13 (0.17)	69 (0.14)
Missing	0	4

Note: 1 case does not contain no clinical info; values in parentheses are percentages by column (no content missing)

Table2. Summary statistics of age and survival time

	Excluded	Remaining
Age	60.7 (11.2)	59.6 (11.6)
Last_contact_days_to	897.7 (757.6)	991.6 (825.0)
Death_days_to	1002.3 (774.5)	1097.6 (737.4)

Note: Values in parentheses are standard deviations

Table 3. Selected posterior probabilities of individual gene effects

Gene	$P(\beta_j > 0 \theta)$	$P(\beta_j < 0 \theta)$	HR (p-value) ²
<i>CDC25A</i>	0.74	0.23	1.22 (0.342)
<i>CDC6</i>	0.37	0.58	0.92 (0.542)
<i>CDKN1A</i>	0.99	0.01	1.20 (0.031)
<i>CDKN2A</i>	0.26	0.65	0.99 (0.905)
<i>CDKN2B</i>	0.18	0.81	1.11 (0.657)
<i>CDKN2C</i>	0.45	0.46	1.00 (0.983)
<i>CDKN2D</i>	0.95	0.05	1.35 (0.231)
<i>EP300</i>	0.10	0.88	0.92 (0.505)
<i>ESPL1</i>	0.74	0.23	0.92 (0.476)
<i>GSK3B</i>	0.77	0.20	1.34 (0.030)
<i>MAD1L1</i>	0.59	0.34	1.02 (0.780)
<i>MDM2</i>	0.03	0.97	1.43 (0.401)
<i>MYC</i>	0.94	0.04	1.18 (0.005)
<i>N.APCCDC20¹</i>	0.01	0.99	1.22 (0.820)
<i>N.APCCFZRI</i>	0.29	0.69	0.91 (0.926)
<i>N.ATMR</i>	0.03	0.96	0.87 (0.387)
<i>N.CCNACDK</i>	0.66	0.30	1.00 (0.992)
<i>N.CCNDCDK</i>	0.36	0.58	0.98 (0.841)
<i>N.CCNECDK</i>	0.21	0.78	0.98 (0.907)
<i>N.CCNHCDK</i>	0.16	0.81	0.91 (0.441)
<i>N.CDC14</i>	0.69	0.29	1.32 (0.234)
<i>N.CDC25BC</i>	0.67	0.31	1.19 (0.270)
<i>N.CDKN1B1C</i>	0.92	0.05	1.24 (0.055)
<i>N.CHEK</i>	0.56	0.39	1.01 (0.968)
<i>N.E2F45</i>	0.98	0.01	1.33 (0.233)
<i>N.GADD</i>	0.06	0.92	0.78 (0.127)
<i>N.MADBUB</i>	0.32	0.60	0.99 (0.975)
<i>N.PPTG12</i>	0.44	0.53	0.96 (0.856)
<i>N.RBL12</i>	0.21	0.78	0.95 (0.774)
<i>N.SKIP</i>	0.06	0.93	0.99 (0.967)
<i>N.SMAD</i>	0.36	0.62	1.07 (0.728)
<i>N.SMC</i>	0.09	0.90	0.87 (0.561)
<i>N.TGFB</i>	0.22	0.76	1.23 (0.384)
<i>PCNA</i>	0.58	0.38	0.97 (0.824)
<i>PLK1</i>	0.06	0.93	0.89 (0.369)

<i>PRKDC</i>	0.93	0.06	1.04	(0.797)
<i>RB1</i>	0.98	0.01	1.44	(0.020)
<i>SFN</i>	0.56	0.35	1.11	(0.113)
<i>TP53</i>	0.52	0.36	1.01	(0.816)
<i>TTK</i>	0.35	0.56	0.97	(0.722)
<i>ZBTB17</i>	0.04	0.96	0.79	(0.374)

Note:

1. If the node in the pathway is a complex, then the gene name starts with “N.”
2. HR(hazard ratio) were derived from the multiple survival regression model.

Table 4. Identified influential genes

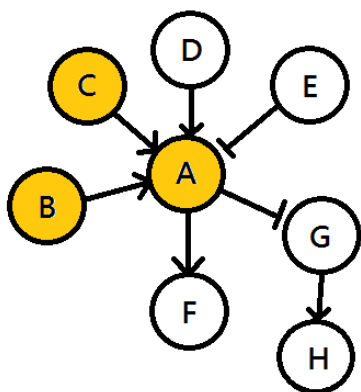
Gene or complex	Posterior probabilities	Coefficient β (p-value)	hazard ratio
<i>CDKN1A</i>	0.97	-0.183 (0.031)	1.20
<i>MDM2</i>	0.97	-0.356 (0.401)	1.43
<i>APC/C /CDC20</i>	0.99	-0.202 (0.820)	1.22
<i>ATM/ATR</i>	0.96	0.142 (0.387)	0.87
<i>E2F4/E2F5/RBL1/TFDP1/TFDP2</i>	0.98	-0.289 (0.233)	1.33
<i>RB1</i>	0.98	-0.363 (0.020)	1.44
<i>ZBTB17</i>	0.96	0.240 (0.375)	0.79

Note: The coefficient β (p-value) and HR (hazard ratio) were derived from multiple survival regression model.

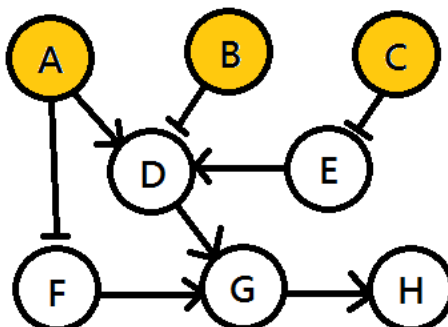


Figure 1. Examples of pathways.

Yellow circle refer to genes whose correlation between DNA methylation and gene expression is negative



(a)



(b)

Figure 2. Examples of nodes in KEGG pathways

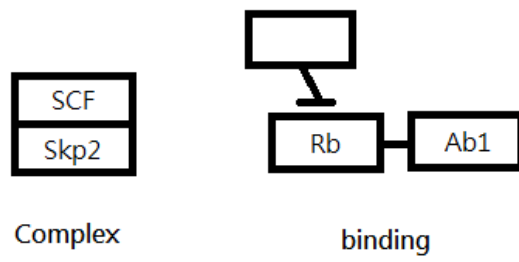


Figure 3. The 95% credible intervals of the coefficients beta

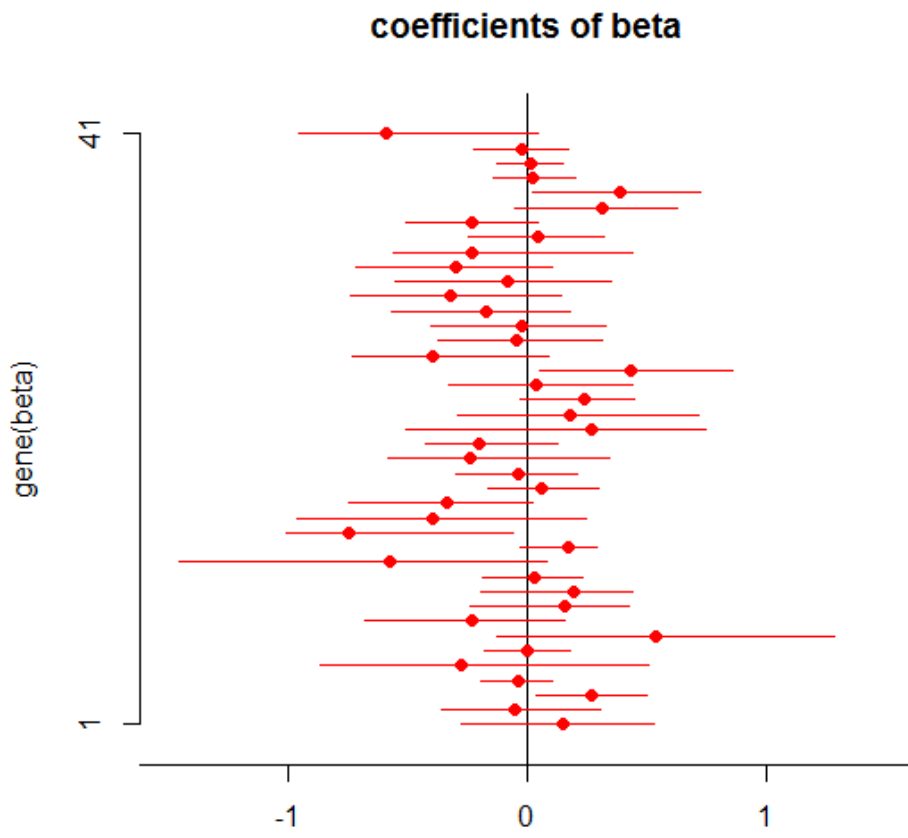
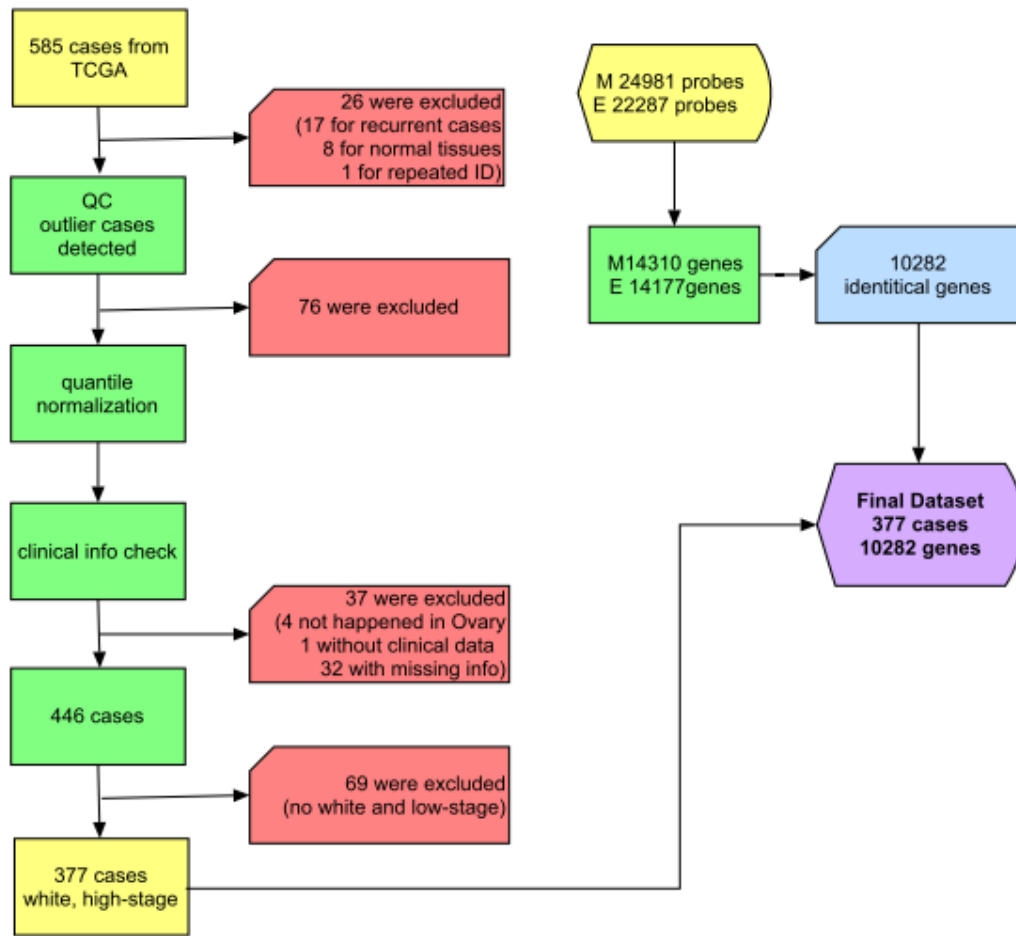




Figure 4. Flowchart of the data management procedure



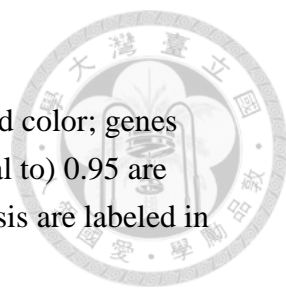


Figure 5. Selected genes in the cell cycle pathway

Genes whose posterior probabilities larger than 0.95 are labeled in red color; genes whose posterior probabilities larger than 0.94 but lower than (or equal to) 0.95 are labeled in yellow color; and genes not incorporated in our data analysis are labeled in gray color.

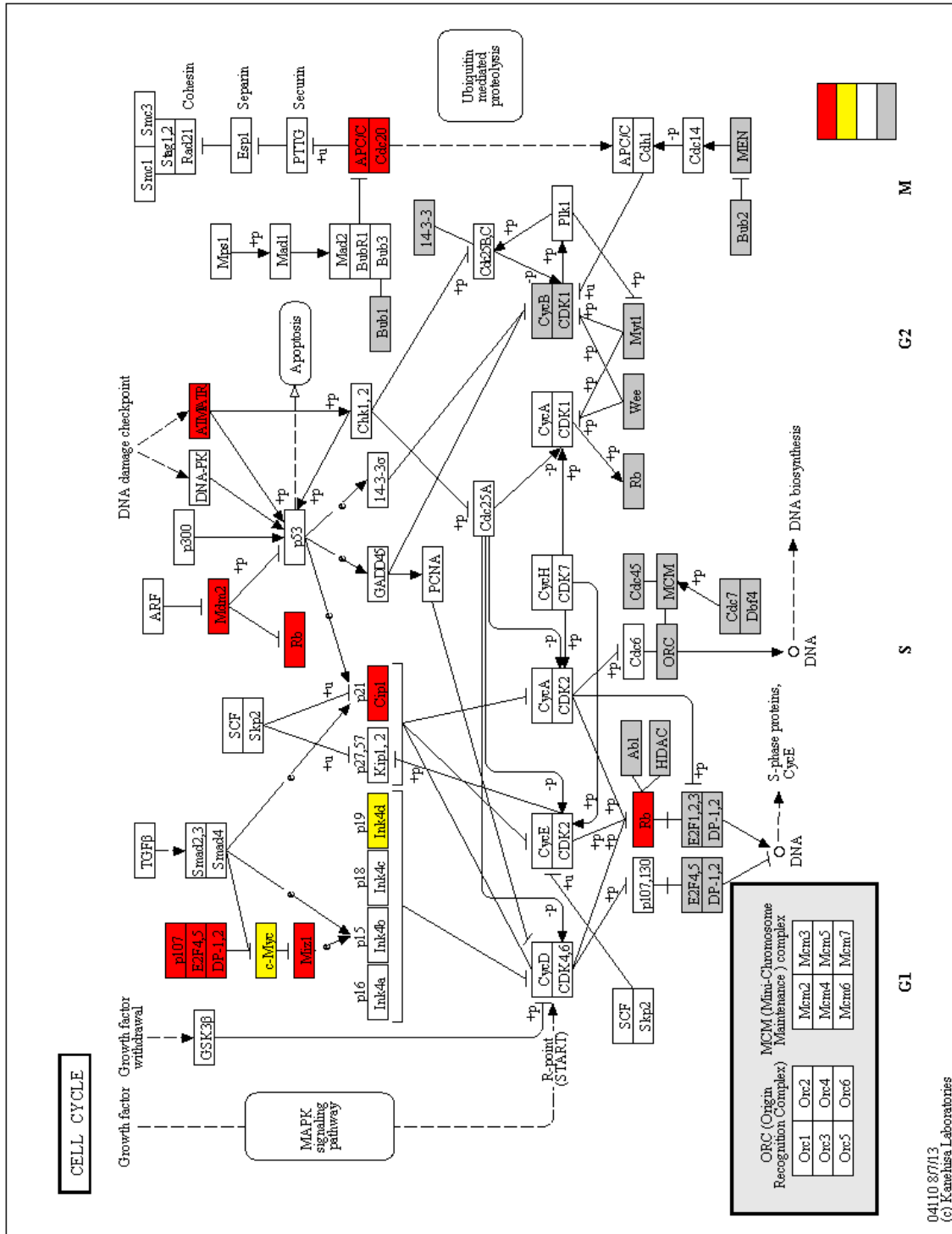




Figure 6. Density plots of coefficients γ^+ , γ^- , δ^+ , δ^- , η

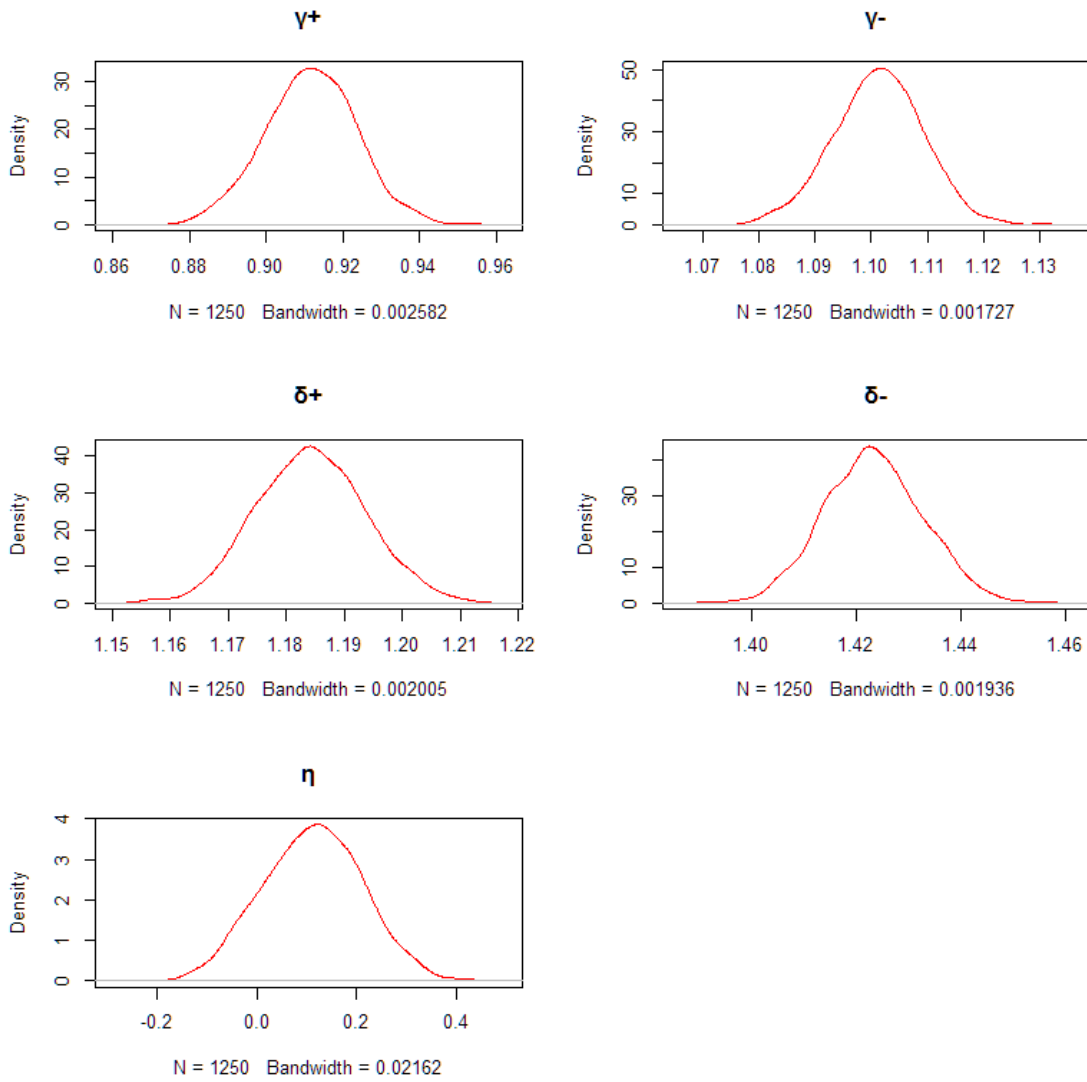
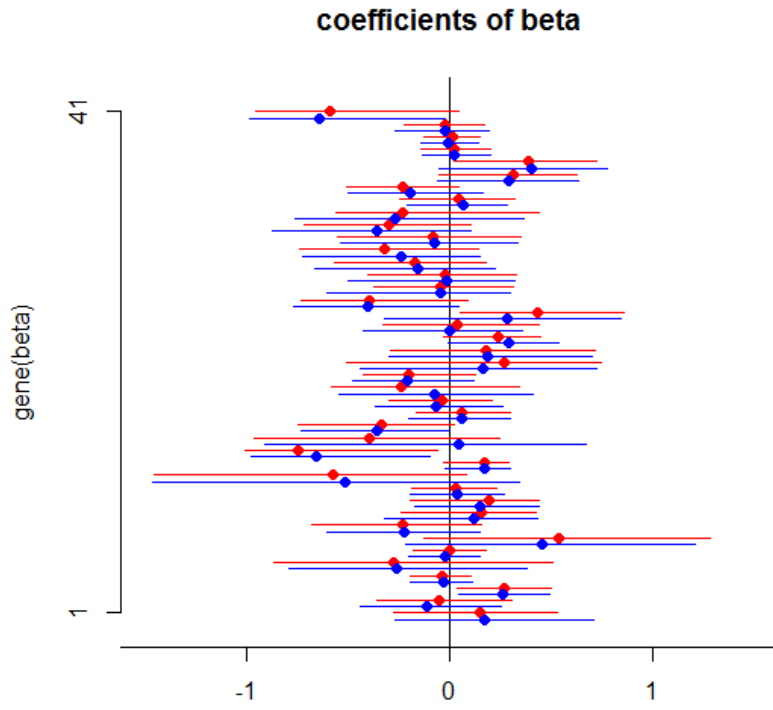


Figure 7. The 95% credible intervals for the coefficients beta derived from 30000 (red) or 50000 (blue) iterations





Appendix.1

Comparison between the remaining samples and excluded samples

Table 1 indicates the frequency of clinical information between remaining samples and excluded samples. First, excluded outliers a little tend to be high-stage patients. Especially in stage IIIC, it differed by 10 percent (89.49-70.13) between excluded outliers and remaining cases.

Second, the remaining cases of five categories in races were 0.01, 0.03, 0.04, 0.00, 0.92, and on the other hand, those of excluded outliers were correspondently 0.00, 0.05, 0.05, 0.01, 0.88, which indicates the distribution of races between excluded outliers and remaining cases are very similar.

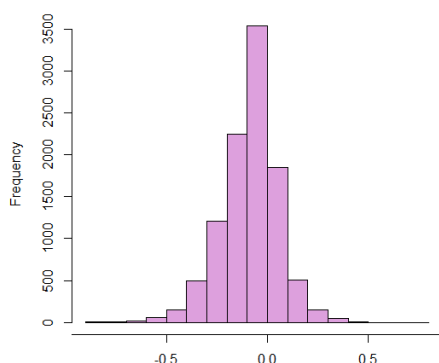
Third, the percentage of vital_status of remaining cases were 0.50, 0.50, and on the other hand, those of excluded outliers were correspondently 0.42, 0.58, which indicates the distribution of vital_status between excluded outliers and remaining cases are very similar.

The summary statistics of continuous variables are listed in table 2. The mean and standard deviation of age in remaining cases were 59.6 and 11.6, on the other hand, those of excluded samples were 60.7, 11.2, which indicates the distribution of ages between two groups are very similar. The mean and standard deviation of death_days_to between two groups are very similar. Also, the mean and standard deviation of last_contact_days_to between two groups are very similar. In conclusion, there are no specific attributes in excluded samples.

Preliminary analysis: Relation between gene expression and DNA methylation

We calculated Pearson correlations between gene expression and DNA methylation level in each gene. The result was shown in Figure S1. Majority of correlations were negative values, which fitted the image that increasing DNA methylation levels will decrease the performance of RNA expression. Conversely, for those genes that had positive relations between gene expression and DNA methylation, we expected there are other mechanisms to regulate the gene's RNA expression.

Figure S1. Correlations between gene expression and DNA methylation level





Appendix.2

Assignment of nodes and lines from pathway maps

Arrangement of nodes

The pathway maps were downloaded from KEGG. Following we will describe the principles when we arrange the relations of gene in pathways. In KEGG pathway map, rectangle figures represent a gene product of a specific gene or a gene set which contents several genes with similar functions. If the rectangle figure represents a gene set, the expression of this node will be replaced with the average of gene expression levels of this gene set. Notice that we will skip the genes in the gene set which are no information in our data when calculate the average value of this gene set.

Second, if more than one rectangle figures are attached to each other, it is on behalf of a complex, which means these gene products have strong association and they only work when these gene products combine to each other. Thus, in our analysis, the complex will be regarded as a single node, and its expression value will be represented as the average of gene expression levels of these genes.

Third, if the data of the specific gene in pathways are missing, we would exclude the gene in our analysis. And the branches (connection to other genes) of these nodes will be ignored. Fourth, we only focus on the gene that is regulated by other genes and the genes binding with other genes (like Ab1 in Figure S2) and not regulating other genes or not regulated by other genes will be also ignored.

Arrangement of lines

The branches (connection to other genes) are divided into two categories, one group is activation (symbol black array) and the other is inhibition (symbol \perp). Except symbol black array and symbol perpendicular, other relations between genes (such as pure straight line, cross shaped, and pure dotted line) in pathways are ignored. Notice that all kind of black array (including activation, phosphorylation, expression, indirect effects (symbol dotted line with black array), and so on) are seen as activation. For every node, we will record that which genes activate it, which genes inhibit it, which genes are activated by it and which genes are inhibited by it. Please be attention that if a gene A is directed to a molecule, and the molecule is directed to other gene B, we all consider that A is directed to B. When finish the arrangement, a pathway map can be organized to one table, like the Table S1. below.

Finally we can use the information of this table to construct the conditional distributions of X. The detail notations of KEGG pathway map can be found at http://www.genome.jp/kegg/document/help_pathway.html.

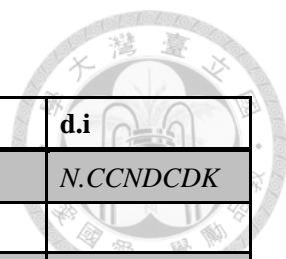
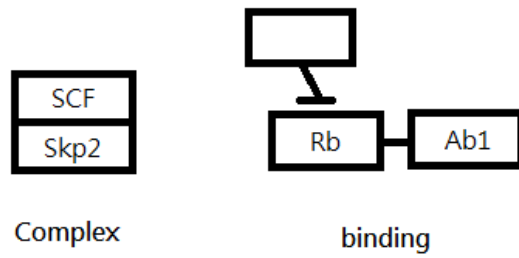


Table S1.

Node	u.a	u.i.	d.a	d.i
<i>PCNA</i>	<i>N.GADD</i>			<i>N.CCNDCK</i>
<i>PLK1</i>			<i>N.CDC25BC</i>	
<i>PRKDC</i>			<i>TP53</i>	
<i>SFN</i>	<i>TP53</i>			
<i>TP53</i>	<i>EP300,PRKDC,N.ATMR, N.CHEK</i>	<i>MDM2</i>	<i>CDKN1A,N.GADD, SFN</i>	
<i>TTK</i>			<i>MAD1L1</i>	
<i>ZBTB17</i>		<i>MYC</i>	<i>CDKN2B</i>	

Note: u.a. means upstream activate gene; u.i. means upstream inhibition gene; d.a. means downstream activate gene; d.i. means downstream inhibition gene.

Figure S2. Examples of nodes in KEGG pathways



The expression level of complex SCF, skp2 would be the average value of their gene expression level; Gene Ab1 binds to Rb, but don't have other regulation connection with other genes. So gene Ab1 would be ignored in our analysis.



Appendix.3

R code

```
#R code for KEGG has04110 pathway
#data
# N = 377; the number of cases
# P = 41; the number of genes in a specific pathway
#t.obs[i] ; time period from death to initial pathologic diagnosis of ith person
#t.cen[i] ; time period from censor to initial pathologic diagnosis of ith person
#num[1:P] ; the number of neighbors and whether has strong correlation with DNA
                methylation of a specific gene
#x[i,1:P] ; the gene expression of ith person
#M[i,1:P] ; the DNA methylation level of ith person

#parameter
#beta[1:P] ; effect of genes in a specific pathway                #β in our model
#tau[1:P] ; variance of 1 to P genes
#a1 ; the effect coming from upstream activation genes                #γ+ in our model
#a2 ; the effect coming from upstream inhibition genes                #γ- in our model
#a3 ; the effect coming from downstream activation genes                #δ+ in our model
#a4 ; the effect coming from downstream inhibition genes                #δ- in our model
#a5 ; the effect DNA methylation                #η in our model
#shape; the shape of Weibull distribution

Model <- function()
{
shape ~ dgamma(3,2)
for (j in 1:P) {
beta[j] ~ dnorm(0,1)
tau[j] <- num[j]/G
}

for (i in 1:N){
HRx[i] <- exp(inprod(x[i,1:P],beta[]))
lamda[i] <- HRx[i]
t.obs[i] ~ dweib(shape,lamda[i])%_%l(t.cen[i],)

mu1[i,1]<- (a2*(x[i,24])+a3*(x[i,19]+x[i,17]+x[i,18]))/num[1]
mu2[i,2]<- (a2*(x[i,17]))/num[2]
```



mu3[i,3]<- (a1*(x[i,31]+x[i,39])+a2*(x[i,30])+a4*(x[i,19]+x[i,17])+a5*M[i,3])/num[3]
mu4[i,4]<- (a2*(x[i,12])+a4*(x[i,18]))/num[4]
mu5[i,5]<- (a1*(x[i,41]+x[i,31])+a4*(x[i,18]))/num[5]
mu6[i,6]<- (a4*(x[i,18])+a5*M[i,6])/num[6]
mu7[i,7]<- (a4*(x[i,18])+a5*M[i,7])/num[7]
mu8[i,8]<- (a3*(x[i,39])+a5*M[i,8])/num[8]
mu9[i,9]<- (a2*(x[i,28])+a4*(x[i,32])+a5*M[i,9])/num[9]
mu10[i,10]<- (a4*(x[i,18]))/num[10]
mu11[i,11]<- (a1*(x[i,40])+a3*(x[i,27]))/num[11]
mu12[i,12]<- (a2*(x[i,4])+a4*(x[i,39]+x[i,37]))/num[12]
mu13[i,13]<- (a2*(x[i,25]+x[i,31])+a4*(x[i,41]))/num[13]
mu14[i,14]<- (a2*(x[i,27])+a4*(x[i,28])+a5*M[i,14])/num[14]
mu15[i,15]<- (a1*(x[i,21])+a5*M[i,15])/num[15]
mu16[i,16]<- (a3*(x[i,39]+x[i,24])+a5*M[i,16])/num[16]
mu17[i,17]<-
(a1*(x[i,1]+x[i,20])+a2*(x[i,23]+x[i,3])+a4*(x[i,37]+x[i,2])+a5*M[i,17])/num[17]
mu18[i,18]<-
(a1*(x[i,1])+a2*(x[i,10]+x[i,4]+x[i,5]+x[i,6]+x[i,7]+x[i,34])+a4*(x[i,37]+x[i,29])+a5*M[i,18])/num[18]
mu19[i,19]<- (a1*(x[i,1]+x[i,20])+a2*(x[i,23]+x[i,3]+x[i,30])+a4*(x[i,37]))/num[19]
mu20[i,20]<- (a3*(x[i,17]+x[i,19])+a5*M[i,20])/num[20]
mu21[i,21]<- (a3*(x[i,15])+a5*M[i,21])/num[21]
mu22[i,22]<- (a1*(x[i,35])+a2*(x[i,24]))/num[22]
mu23[i,23]<- (a2*(x[i,30]+x[i,19])+a4*(x[i,17]+x[i,19])+a5*M[i,23])/num[23]
mu24[i,24]<- (a1*(x[i,16])+a3*(x[i,39])+a4*(x[i,22]+x[i,1])+a5*M[i,24])/num[24]
mu25[i,25]<- (a4*(x[i,13])+a5*M[i,25])/num[25]
mu26[i,26]<- (a1*(x[i,39])+a3*(x[i,34])+a5*M[i,26])/num[26]
mu27[i,27]<- (a1*(x[i,40])+a4*(x[i,14]))/num[27]
mu28[i,28]<- (a2*(x[i,14])+a4*(x[i,9])+a5*M[i,28])/num[28]
mu29[i,29]<- (a2*(x[i,18])+a5*M[i,29])/num[29]
mu30[i,30]<- (a4*(x[i,3]+x[i,23]+x[i,19])+a5*M[i,30])/num[30]
mu31[i,31]<- (a1*(x[i,33])+a3*(x[i,5]+x[i,3])+a4*(x[i,13])+a5*M[i,31])/num[31]
mu32[i,32]<- (a2*(x[i,9])+a5*M[i,32])/num[32]
mu33[i,33]<- (a3*(x[i,31])+a5*M[i,33])/num[33]
mu34[i,34]<- (a1*(x[i,26])+a4*(x[i,18]))/num[34]
mu35[i,35]<- (a3*(x[i,22]))/num[35]
mu36[i,36]<- (a3*(x[i,39]))/num[36]
mu37[i,37]<- (a2*(x[i,12]+x[i,18]+x[i,19]+x[i,17]))/num[37]



mu38[i,38]<- (a1*(x[i,39])+a5*M[i,38])/num[38]

mu39[i,39]<-

(a1*(x[i,8]+x[i,36]+x[i,16]+x[i,24])+a2*(x[i,12])+a3*(x[i,3]+x[i,26]+x[i,38])+a5*M[i,39])/num[39]

mu40[i,40]<- (a3*(x[i,11]))/num[40]

mu41[i,41]<- (a2*(x[i,13])+a3*(x[i,5]))/num[41]

x[i,1]~dnorm(mu1[i,1],tau[1])

x[i,2]~dnorm(mu2[i,2],tau[2])

x[i,3]~dnorm(mu3[i,3],tau[3])

x[i,4]~dnorm(mu4[i,4],tau[4])

x[i,5]~dnorm(mu5[i,5],tau[5])

x[i,6]~dnorm(mu6[i,6],tau[6])

x[i,7]~dnorm(mu7[i,7],tau[7])

x[i,8]~dnorm(mu8[i,8],tau[8])

x[i,9]~dnorm(mu9[i,9],tau[9])

x[i,10]~dnorm(mu10[i,10],tau[10])

x[i,11]~dnorm(mu11[i,11],tau[11])

x[i,12]~dnorm(mu12[i,12],tau[12])

x[i,13]~dnorm(mu13[i,13],tau[13])

x[i,14]~dnorm(mu14[i,14],tau[14])

x[i,15]~dnorm(mu15[i,15],tau[15])

x[i,16]~dnorm(mu16[i,16],tau[16])

x[i,17]~dnorm(mu17[i,17],tau[17])

x[i,18]~dnorm(mu18[i,18],tau[18])

x[i,19]~dnorm(mu19[i,19],tau[19])

x[i,20]~dnorm(mu20[i,20],tau[20])

x[i,21]~dnorm(mu21[i,21],tau[21])

x[i,22]~dnorm(mu22[i,22],tau[22])

x[i,23]~dnorm(mu23[i,23],tau[23])

x[i,24]~dnorm(mu24[i,24],tau[24])

x[i,25]~dnorm(mu25[i,25],tau[25])

x[i,26]~dnorm(mu26[i,26],tau[26])

x[i,27]~dnorm(mu27[i,27],tau[27])

x[i,28]~dnorm(mu28[i,28],tau[28])

x[i,29]~dnorm(mu29[i,29],tau[29])

x[i,30]~dnorm(mu30[i,30],tau[30])

x[i,31]~dnorm(mu31[i,31],tau[31])

x[i,32]~dnorm(mu32[i,32],tau[32])

```
x[i,33]~dnorm(mu33[i,33],tau[33])
x[i,34]~dnorm(mu34[i,34],tau[34])
x[i,35]~dnorm(mu35[i,35],tau[35])
x[i,36]~dnorm(mu36[i,36],tau[36])
x[i,37]~dnorm(mu37[i,37],tau[37])
x[i,38]~dnorm(mu38[i,38],tau[38])
x[i,39]~dnorm(mu39[i,39],tau[39])
x[i,40]~dnorm(mu40[i,40],tau[40])
x[i,41]~dnorm(mu41[i,41],tau[41])
}
```

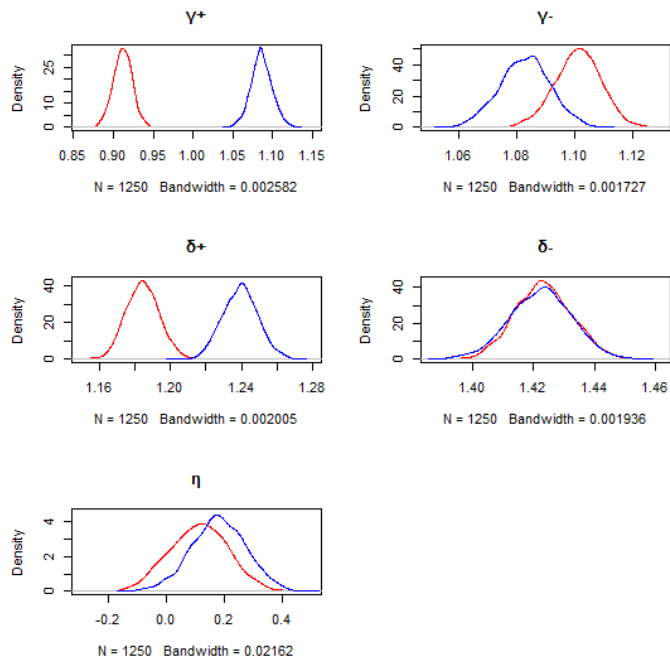
```
a1 ~dnorm(0,1)
a2 ~dnorm(0,1)
a3 ~dnorm(0,1)
a4 ~dnorm(0,1)
a5 ~dnorm(-1,100)
G~dgamma(2,5)
}
```



Appendix.4 Different cutoff values in methylation effect



Figure S3. Density plot of coefficients γ^+ , γ^- , δ^+ , δ^- , η



Note:

Red lines represent the result using cutoff point -0.1 (present model in article)

Blue lines represent the result using cutoff point -0.05

Table S2. Selected posterior probabilities of individual gene effect using cutoff point -0.05

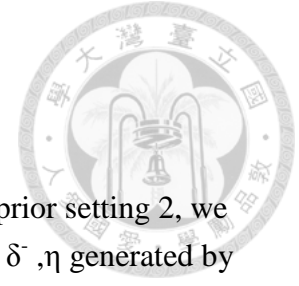
Gene	$P(\beta_j > 0 \theta)$	$P(\beta_j < 0 \theta)$	Gene	$P(\beta_j > 0 \theta)$	$P(\beta_j < 0 \theta)$
<i>CDC25A</i>	0.69	0.28	<i>N.CDC25BC</i>	0.98	0.02
<i>CDC6</i>	0.26	0.70	<i>N.CDKN1B1C</i>	0.94	0.06
<i>CDKN1A</i>	1.00	0.00	<i>N.CHEK</i>	0.37	0.58
<i>CDKN2A</i>	0.25	0.67	<i>N.E2F45</i>	0.68	0.30
<i>CDKN2B</i>	0.27	0.71	<i>N.GADD</i>	0.07	0.93
<i>CDKN2C</i>	0.44	0.46	<i>N.MADBUB</i>	0.61	0.35
<i>CDKN2D</i>	0.55	0.42	<i>N.PPTG12</i>	0.52	0.46
<i>EP300</i>	0.16	0.81	<i>N.RBL12</i>	0.58	0.38
<i>ESPL1</i>	0.38	0.58	<i>N.SKP</i>	0.36	0.60
<i>GSK3B</i>	0.94	0.05	<i>N.SMAD</i>	0.18	0.79
<i>MAD1L1</i>	0.55	0.38	<i>N.SMC</i>	0.04	0.96
<i>MDM2</i>	0.26	0.73	<i>N.TGFB</i>	0.20	0.79
<i>MYC</i>	0.96	0.03	<i>PCNA</i>	0.10	0.87
<i>N.APCCDC20^l</i>	0.00	0.99	<i>PLK1</i>	0.10	0.87
<i>N.APCCFZR1</i>	0.42	0.54	<i>PRKDC</i>	0.65	0.33
<i>N.ATMR</i>	0.01	0.99	<i>RB1</i>	0.91	0.08
<i>N.CCNACDK</i>	0.91	0.06	<i>SFN</i>	0.62	0.30
<i>N.CCNDCDK</i>	0.40	0.56	<i>TP53</i>	0.37	0.49
<i>N.CCNECDK</i>	0.41	0.56	<i>TTK</i>	0.20	0.73
<i>N.CCNHCDK</i>	0.03	0.95	<i>ZBTB17</i>	0.11	0.87
<i>N.CDC14</i>	0.59	0.37			

Note:

If the node in pathway is complex, the front of gene's name would be "N."

Appendix.5

Sensitivity analysis



The prior setting 1 is what we present in the article. And in the prior setting 2, we changed the mean of γ^+ , γ^- , δ^+ , δ^- , η by using the median of γ^+ , γ^- , δ^+ , δ^- , η generated by using the prior setting 1. Detail of setting is shown in the formula below:

Prior setting 1

$$shape \sim gamma(3, 2)$$

$$\beta_j \sim N(0, 1), j \text{ from } 1 \text{ to } P$$

$$\gamma^+ \sim N(0, 1),$$

$$\gamma^- \sim N(0, 1),$$

$$\delta^+ \sim N(0, 1),$$

$$\delta^- \sim N(0, 1),$$

$$\eta \sim N(-1, 100),$$

$$G^2 \sim gamma(2, 5)$$

Prior setting 2 (we take the median of γ^+ , γ^- , δ^+ , δ^- , η in prior setting 1)

$$shape \sim gamma(3, 2)$$

$$\beta_j \sim N(0, 1), j \text{ from } 1 \text{ to } P$$

$$\gamma^+ \sim N(0.9, 1),$$

$$\gamma^- \sim N(1.1, 1),$$

$$\delta^+ \sim N(1.2, 1),$$

$$\delta^- \sim N(1.2, 1),$$

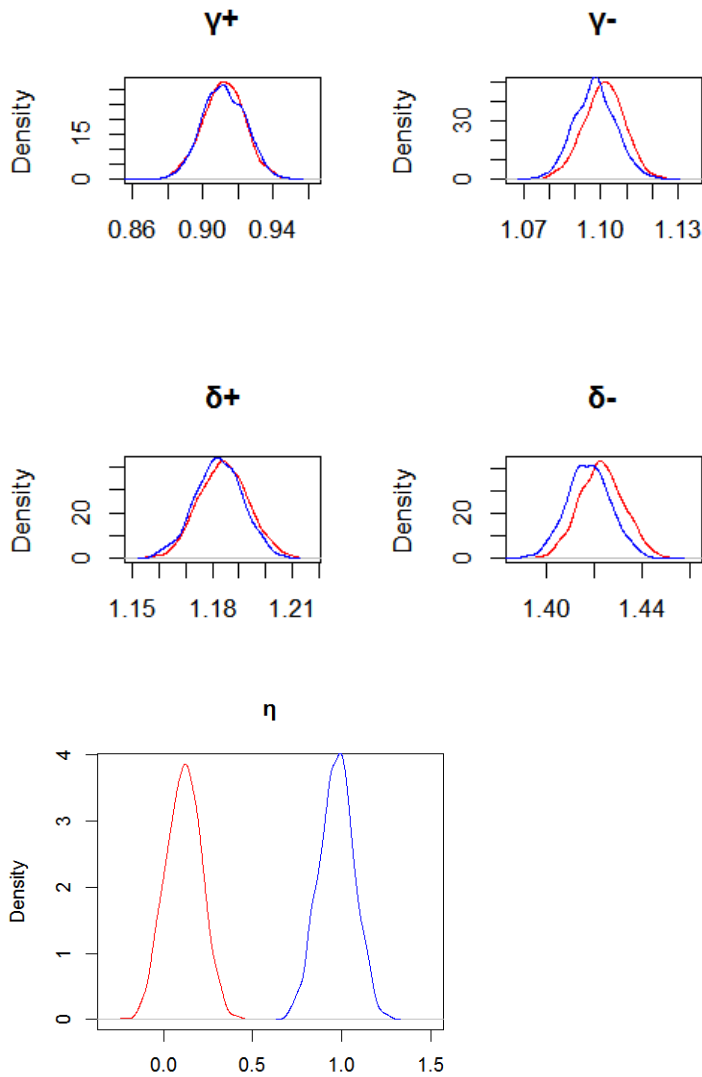
$$\eta \sim N(0.1, 100),$$

$$G^2 \sim gamma(2, 5)$$

The results are shown in following tables and figures.



Figure S4. Density plot of coefficients γ^+ , γ^- , δ^+ , δ^- , η



Note:

Red lines represent the result using prior setting 1 (present model in article)

Blue lines represent the result using prior setting 2

The distribution of γ^+ , γ^- , δ^+ , δ^- are all similar to the density of prior setting 1, which indicates γ^+ , γ^- , δ^+ , δ^- are not sensitive to different priors; on the other hand, in prior setting 1, we have given η a comparably high-information prior. Therefore, it's not surprising that the distribution of η will move after changing the prior.

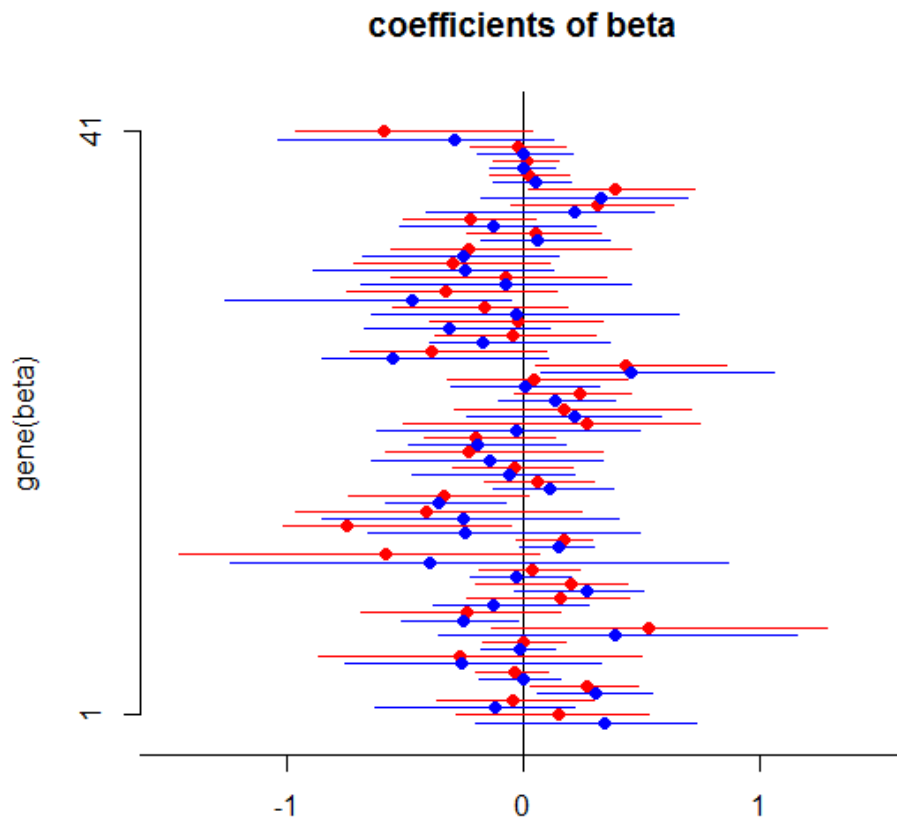
Table S3. Coefficient of beta

Gene	$P(\beta_j > 0 \theta)$	$P(\beta_j < 0 \theta)$	Gene	$P(\beta_j > 0 \theta)$	$P(\beta_j < 0 \theta)$
<i>CDC25A</i>	0.92	0.07	<i>N.CDC25BC</i>	0.77	0.22
<i>CDC6</i>	0.24	0.73	<i>N.CDKN1B1C</i>	0.82	0.14
<i>CDKN1A</i>	0.99(0.99)	0.01	<i>N.CHEK</i>	0.49	0.46
<i>CDKN2A</i>	0.43	0.47	<i>N.E2F45</i>	0.99(0.95)	0.01
<i>CDKN2B</i>	0.15	0.83	<i>N.GADD</i>	0.06	0.93
<i>CDKN2C</i>	0.38	0.52	<i>N.MADBUB</i>	0.27	0.71
<i>CDKN2D</i>	0.81	0.17	<i>N.PPTG12</i>	0.08	0.91
<i>EP300</i>	0.01	0.98(0.88)	<i>N.RBL12</i>	0.48	0.51
<i>ESPL1</i>	0.22	0.74	<i>N.SKP</i>	0.00	1.00(0.93)
<i>GSK3B</i>	0.94	0.05	<i>N.SMAD</i>	0.38	0.59
<i>MAD1L1</i>	0.37	0.58	<i>N.SMC</i>	0.12	0.84
<i>MDM2</i>	0.39	0.61	<i>N.TGFB</i>	0.10	0.88
<i>MYC</i>	0.94	0.04	<i>PCNA</i>	0.62	0.31
<i>N.APCCDC20^l</i>	0.29	0.70	<i>PLK1</i>	0.25	0.72
<i>N.APCCFZRI</i>	0.24	0.74	<i>PRKDC</i>	0.81	0.17
<i>N.ATMR</i>	0.01	0.99(0.96)	<i>RB1</i>	0.91	0.08
<i>N.CCNACDK</i>	0.82	0.14	<i>SFN</i>	0.64	0.28
<i>N.CCNDCDK</i>	0.33	0.61	<i>TP53</i>	0.41	0.47
<i>N.CCNECDK</i>	0.28	0.69	<i>TTK</i>	0.46	0.48
<i>N.CCNHCDK</i>	0.14	0.84	<i>ZBTB17</i>	0.13	0.85
<i>N.CDC14</i>	0.44	0.53			

Note:

Values in the parentheses is the posterior probability derived from prior setting 1

Figure S5. Coefficients of beta



Note:

Red lines represent the result using prior setting 1 (present model in article)

Blue lines represent the result using prior setting 2

Despite the significant genes ($\mathbf{P}(\beta_j > \mathbf{0} | \theta) > 0.95$ or $\mathbf{P}(\beta_j < \mathbf{0} | \theta) > 0.95$) differ in prior setting 1 and 2, their posterior probabilities and 95% credible interval are very similar. And the average differences of posterior probabilities between prior setting 1 and 2 are smaller than five percent.