

# Binary Transition Pattern (BTP) for Erythropoietic DNA Methylation Data

## Najyah Garoot1\* and Byung Guk Kim2

<sup>1</sup>Department of Computer Science, University of Massachusetts, Lowell, University Ave, Lowell, MA 01854, USA <sup>2</sup>Department of Information Systems, College of Computing and Information Technology, King Abdul-Aziz University, Jeddah, Saudi Arabia

#### Abstract

**Objective:** Despite whole-genome data analysis for understanding the implications of differential DNA methylation patterns and the nature of their correlation with gene expression (GE) regulation, the observations remain controversial.

**Methods:** Unlike the usual methods of selecting differentially methylated regions (DMR) by hypothesis statistical testing, we clustered CpG sites by their actual log<sub>2</sub>FC over time. Each cluster is represented by a 5-digit binary code.

**Results:** We identified novel stage-specific TFBS during erythropoiesis. Our stage-specific differential methylation clusters have the ability to determine TFBS of co-regulators such as GATA1 and PU.1 at hypermethylated CpG loci.

**Conclusion:** Hypomethylated CpG loci clusters that are associated with GE upregulation are independent of hypermethylated sites associated with downregulation. The analy-sis confirms that DNA methylation has unidirectional correlation with GE.

Keywords: Binary transition pattern; Erythropoietic DNA; Methylation data

# Introduction

Several recent studies have analyzed DNA methylation as ratio of intensities, methylation percentage,  $\beta$ -value, or M-value to determine possible regulatory regions called differentially methylated regions (DMR) using statistical methods. An alternative approach to differential methylation (DM) analysis is loci-based analysis, where variability in loci methylation is measured instead of testing hypothesis of mean difference [1].

DNA methylation profiling and clustering methods were developed for detecting meaningful patterns. Hierarchical and non-hierarchical clustering methods are the most commonly used methods.

Usually, non-parametric approaches do not consider the time sequence of samples. We developed a time-sensitive clustering method that uses actual methylation measures free of statistical estimates and has no sample-size requirement. The method develops stage-specific binary transition patterns (BTP) to profile methylation data based on its fold change (FC) values.

Instead of using clustering to provide initial profiling, our BTP presents flexibly divided clusters further based on the pattern itself, genomic regions of interest, or methylation status (hypo/hyper). BTP introduced a summarized view of methylation changes over time regardless of sample size.

An initial analysis of raw DNA methylation data of chromosome 16 had defined a list of DMRs with adjusted p value <0.02 and presented CpG clusters that were enriched with biologically significant transcription factor binding site (TFBS) for erythroid development [2]. The analysis relied on the maximum logFC for the five empirical Bayes (EB) tests of DM (start, early, intermediate, late, and last), ignoring the logFC of the other differentiation phases. Our study illustrated how consideration of the methylation status of all stages could form specific and meaningful clusters of CpG loci.

We compared the selection of stage-specific DMRs by statistical threshold alone and by statistical threshold and methylation pattern combined. Then, we showed the power of pattern-specific analysis of all CpG sites without excluding any site below a statistical threshold. The analysis included promoter, gene body, intergenic regions, CpG islands (CGIs), and erythroid-specific enhancers.

### Methods

We studied chromosome 16 DNA methylation dataset of 12 samples from adult bone marrow which were collected at six-time points of erythroid differentiation: day 0, 3, 7, 10, 13, and 16 [GSE44054] [3]. Data was previously processed and normalized for stage-specific DM analysis [2].

In this study, we considered the methylation pattern of each CpG fragment regardless of their DM threshold. Each CpG fragment of a differentiating erythroid cell is assign a 5-digit binary code that represented the five log,FC of development stages.

To define the methylation pattern, the transition summary between six stages of erythroid development was represented in five binary bits. First, maximum  $|-\log_2 FC|$  was selected. If the value of the selected  $-\log_2 FC$  is negative, we assigned 0 for the first binary bit of the BTP bits and 1 if it was positive. Therefore, if the leading bit of a BTP is 0, the loci will be hypomethylated at a certain stage and 1 if it is hypermethylated. Each of the remaining four bits in a BTP was assigned 0 if there is loss of methylation during inter-stage transition, and 1 otherwise (Figure 1). An example of pattern conversion is provided in Table 1.

## Results

Initially, CpG loci were assigned to differential stages based

\*Corresponding author: Najyah Garoot, Department of Computer Science, University of Massachusetts, Lowell, 1 University Ave, Lowell, USA, Tel: (781)308-0614; E-mail: ngaroot@kau.edu.sa

Received May 10, 2018; Accepted May 30, 2018; Published June 11, 2018

Citation: Garoot N, Kim BG (2018) Binary Transition Pattern (BTP) for Erythropoietic DNAMethylation Data. JComputSciSystBiol 11:229-232. doi:10.4172/jcsb.1000277

**Copyright:** © 2018 Garoot N, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

on their maximum absolute -log<sub>2</sub>FC. By assessing distribution of hypomethylated (hypo) DMR based on their BTP, in addition to their differential stage, we observed that CpG loci that peak at a specific stage are likely to lose and gain methylation following a specific pattern (Figure 1). We noted that not all BTPs are found in hypo-DMR. The majority of hypo DMR follows the methylation pattern of single switch between '0' to '1', which indicates stage specificity.

We verified whether the TFBS enrichment results of stagespecific hypo DMR results was due to stage specificity alone or certain methylation patterns by comparing our results with previously published TFBS enrichment results of stage-specific DMR regardless of their pattern of methylation [2].

TFBS enrichment analysis (EA) of each start phase pattern ('01110' and '01111') did not yield any significant results. We thus considered '0111X' where X can be either 0 or 1. The enrichment result of BTP '0111X'. Table 1 showed stronger enrichment of the GATA family TFBS compared to the EA of all hypo-DMR that have maximum  $\log_2$ FC during the start phase regardless of their pattern.

Next, we analyzed the early phase DMRs with pattern '0011X' regardless of the last BTP bit and were able to obtain more significant TFBS enrichment results than those obtained by analyzing all hypo DMR that have maximum  $\log_2$ FC during the early phase, in addition to the enrichment of *Zfp809* (Table 2).

One of the most stage-specific methylation patterns is '00011' as it is the largest cluster and is exclusive to the intermediate phase (day 7 to day 10). More significant EA results were achieved when we ignored the last bit (i.e., pattern "0001X") compared to EA for all intermediate DMRs. Furthermore, the pattern "00011" by itself had three significantly enriched TFBS (Table 2), whereas pattern "0001X" showed five TFBS with q-value < 0.01 (Table 3).

The "00001" BTP DMR clusters of the late stage showed no significant TFBS enrichment. Hence, we assessed BTP "00101", where 31% DMR peaks were in the late stage and 69% DMR peaks were in the early stage (Figure 1), although no significant enrichment was observed.

Generally, when comparing two similar patterns with different last digits, we noticed that the number of hypo DMRs ending with '0' were more than the number of hypo DMRs ending with '1' (i.e., more DMRs tended to lose methylation at the last phase).

In addition, the switch in methylation pattern between 0 and 1 (when  $-\log_2 FC$  increases or decreases) occurs when  $-\log_2 FC$  is maximum. Consequently, patterns with single switch between 0 and 1, such as 01111, 00111, 00011, and 00001 were exclusive to a specific stage (the differentiated stage).



While comparing hypo DMR genomic distribution, we noticed that hypo-DMR erythroid-specific enhancers have higher number of hypo DMR with pattern '0111X' (Figure 2), which is specific to the initial phase of differentiation. Successive hypomethylation of enhancers, promoters, and CGI, followed by gene body regions, is expected during cell development.

## Whole-Chromosome Analysis of Hypomethylated BTP

To evaluate if BTP can improve specificity, we included all CpG loci of chromosome 16 regardless of their statistical significance.

Regions of hypomethylated CpG loci patterns 00001, 00011, 00111, and 01111 (pattern of single switch from 0 to 1) were still specific to a single stage (Figure 2). The rest of the patterns represent CpG loci that have more than one peak of hypomethylation, such as BTP '00101', which loses methylation twice during erythroid differentiation. In other words, the CpG loci of this pattern gained methylation after day 7 and regained it back after day 13. The more delayed is the first switch of methylation, higher are the numbers of hypomethylated regions. In addition, the earlier the switch from 0 to 1 in the pattern, the more is the enrichment of TFBS as shown in Table 4.

Pattern '00001' shows that TFBS of poor cell-type specificity are enriched during the late stage. For example, HRE, and AR-half sites are enriched in the preceding stages as well. Similarly, ZIC1, is a TF known for its role in the development and maturation of many tissues (not specific for hematopoietic development) [4]. This poor specificity of TFBS also may indicate minimum transcription activation of celltype specific genes during the late stage of differentiation, as increase in termination of gene body transcription is expected during this stage.







J Comput Sci Syst Biol, an open access journal ISSN: 0974-7231

## Whole-Chromosome Analysis of Hypermethylated BTP

It is also important to examine the patterns of hyper BTP (CpG loci with maximum positive  $-\log_2 FC$ ) to understand if they have different pattern distribution as they had different  $\log_2 FC$  levels compared to hypomethylated CpG loci. A comparison of Figures 2 and 3 shows that hypo BTPs, but not hyper BTPs, show a gradual increase in CpG loci frequency with progress in differentiation.

In addition, hypo BTPs that end with 0 have more CpG loci (Figure 2) than patterns that end with 1, whereas hyper BTPs show no such consistent behavior (Figure 3). This can emphasize the importance of analyzing hypo- and hyper methylated sites independently.

Unlike hypo BTP, we were not able to detect a pattern-specific TFBS enrichment in majority of hyper BTP. However, by excluding hypermethylated CpG loci at CGI regions, we were able to detect a single BTP with strong pattern-specific TFBS enrichment; 11110 (Table 5).

We noticed that CGI site counts are maximum during the intermediate and late phase BTPs (Figure 3). This trend of hypermethylation is expected toward the maturation phase of erythroid development but was difficult to observe without pattern-specific clusters.

Rank	TF Name	P value	log P value	q value
1	GATA3	1.00E-07	-1.69E+01	0
2	GATA2	1.00E-06	-1.42E+01	0.0001
3	GATA1	1.00E-05	-1.32E+01	0.0002
4	GATA4	1.00E-05	-1.28E+01	0.0002

Table 1: TFBS enrichment analysis results for start DMR of hypo BTP '0111X'.

Rank	TF Name	P value	log P value	q value
1	ZNF322	1.00E-04	-1.11E+01	0.0051
2	ZFP809	1.00E-04	-1.00E+01	0.007
3	RAR:RXR	1.00E-04	-9.89E+00	0.007

Table 2: TFBS enrichment analysis results for early DMR of hypo BTP '0011X'.

Rank	TF Name	P value	log P value	q value
1	NANOG	1.00E-05	-1.11E+01	0.0003
2	BCL6	1.00E-04	-1.00E+01	0.0017
3	SCL	le-4	-1.06E+01	0.0028
4	AR-half site	1.00E-04	-9.89E+00	0.0031
5	NF1-half site	1.00E-03	-8.78E+00	0.0098

 Table 3: TFBS enrichment analysis results for intermediate DMR of hypo BTP '0001X'.

Methylation pattern	Total CpG loci	TFBS Count	TFBS enrichment (q-value < 0.01)
1	10,899	3	HRE, AR-halfsite, Zic
11	7,971	4	HRE, Six1, ZNF322, AR-halfsite
111	4,267	7	RAR:RXR, AR-halfsite, ZNF322, PR, Zfp809, ZFX, SCL
101	4,893	8	ZNF711, HRE, FOXM1, FOXA, ZFX, EBF1, FOXA1, AP-2gamma
0111X <sup>.</sup>	2,296	10	GATA4, GATA3, STAT3+il21, SCL, GATA2, HRE, GATA1, HOXB4, STAT4, EBF1

\*01111 by itself did not show enrichment with q-value <0.01

Table 4: TFBS enrichment results for hypo BTP clusters.

## Discussion

Application of BTP to hypomethylated stage-specific DMRs revealed methylation patterns of high stage specificity. In particular, the major switch of log<sub>2</sub>FC value from hypo- to hypermethylation across differentiation stages was successfully represented by the BTP switch between 0 and 1. Furthermore, BTP analysis of DNA methylation data had the power to consider crucial but not statistically significant CpG sites.

Smaller clusters of pattern-specific DMRs were able to show stronger enrichments of TFBS than stage-specific clusters of DMR regardless of the methylation pattern. We observed that DMRs with patterns 0111X, 0011X, and 0001X have better TFBS enrichment than patterns with a constant fifth digit. A comparison of Figures 1 and 3 revealed that selection of DMR based on its statistical significance was able to filter out non-pattern-specific last phase peaks.

GATA1, GATA2 and, GATA3 are well known erythropoietic factors, whereas GATA4, along with GATA5 and GATA6, are endodermal factors. We observed GATA1, 2, 3, and 4 TFBS specifically enriched with 0111X, but GATA1 and GATA4 had the same q-value (Table 1). Although GATA4 expression in not known to be associated with erythroid development, a study demonstrated GATA4's ability to induce globin gene expression (GE) when the gene was introduced in *GATA1*-null embryonic bodies [5].

We also identified a '00011' cluster that was enriched with *BCL6* and *NANOG* TFBS. Silencing of NANOG was reported to downregulate *BCL6* and other genes in the cell-cycle signaling pathway [6,7]. The hypomethylation of their binding sites during the intermediate phase of erythroid differentiation can further justify their possible co-regulation of other genes.

Our methylation patterns analysis of all CpG loci on chromosome 16 (not only DMRs) revealed a general trend of erythroid differentiationrelated methylation. We observed gradual increase in hypomethylation with progress in differentiation, which might be due to the increase in the number of active regulatory regions (Figure 2) as the cell approaches maturity. This can partially explain the presence of fewer enriched TFBS as differentiation progresses, which might be due to the decreasing specificity of hypomethylated regions (Table 4). On the contrary, hyper BTP (Figure 3) did not show the same gradual change of methylation as hypo BTP (Figure 2).

втр	CpG loci	Count	TFBS enrichment (q-value<0.01)
10110	449	2	SCL, Ascl1
11010	221	2	Ets1-distal, Elk4
11110	275	13	AR-half site, EWS:ERG, SCL, ETV2, ERG, EHF, ETS1, FLI1, ELF5, PU.1, GATA1, EWS:FLI1, ETS1-distal

 $\label{eq:table_transformation} \end{tabular} \end{tabul$ 

Even though, pattern '11110' is not the most frequent BTP (Figures 2 and 3), it has the maximum number of significantly enriched erythroidspecific TFBS of all BTPs after excluding CGI. In the hypo BTP '0111X' cluster, GATA1 binding sites were significantly enriched during the start phase (Table 4), indicating that DNA methylation changes for transcription activation and termination occurs at independent CpG sites.

Hyper BTP showed super significant enrichment to a single late stage pattern, which is '11110'. It was the smallest pattern-specific cluster with the highest TFBS enrichment when CGI were excluded (Table 5).

The significant enrichment of *GATA1* TFBS with hypo BTP '01111' of the start stage and the hyper BTP '11110' of the late stage suggest that DNA methylation is negatively correlated to GE in a single direction. In other words, TFBS hypomethylation activates or/and upregulates GE but they are not responsible for downregulating the same genes.

## Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

#### Acknowledgements

We thank Dr. Robert Haney, University of Massachusetts, Lowell, for his comments and review of the text.

#### References

- Wilhelm BCS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, et al. (2013) Review of processing and analysis methods for DNA methylation array data. Br J Cancer 109: 1394-1402.
- Garoot N, Kim B (2018) Stage-specific DMR during human erythropoiesis in chromosome 16. Genetics Res.
- Yu Y, Mo Y, Ebenezer D, Bhattacharyya S, Liu H, et al. (2013) High resolution methylome analysis reveals widespread functional hypomethylation during adult human erythropoiesis. J Biol Chem 288: 8805-8814.
- O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44: D733-D745.
- Hosoya OS, Mochizuki N, Suzuki M, Ohneda O, Ohneda K, et al. (2006) GATA-4 incompletely substitutes for GATA-1 in promoting both primitive and definitive erythropoiesis in vivo. J Biol Chem 281: 32820-32830.
- Choi SC, Choi JH, Park CY, Ahn CM, Hong SJ, et al. (2012) Nanog regulates molecules involved in stemness and cell cycle-signaling pathway for maintenance of pluripotency of P19 embryonal carcinoma stem cells. J Cell Physiol 227: 3678-3692.
- Jen J, Wang YC (2016) Zinc finger proteins in cancer progression. J Biol Chem 23: 53.