

Tuning the DNA methylation measurement resolution to optimize biological knowledge extraction.

Ricky Lim, Touati Benoukraf^a,

^aCancer Science Institute of Singapore, National University of Singapore

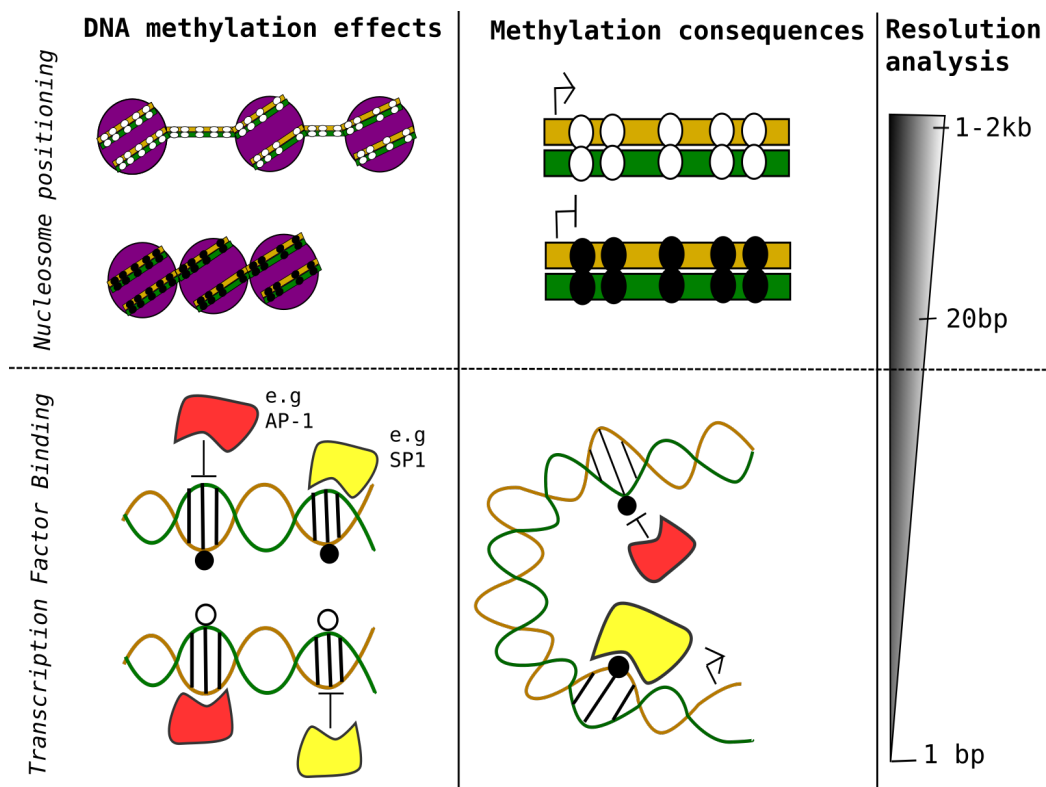
Abstract

Keywords: DNA Methylation, bisulfite sequencing, gene regulation, Chromatin modelling,

1. Introduction

Here we mainly discuss DNA methylation in the context of CpG dinucleotides.

2. Nucleotide Resolutions Required to Analyze Methylation Data in Mammals



Biological Insight	Resolution	References
Gene activity	20-2000bp	[4], [8], [1]
Nucleosome positioning	20 bp	[12], [16], [5]
Transcription factor binding	1 bp	[13], [9], [19]

2.1. DNA methylation and gene activity

Early perception about the function of DNA methylation has been linked to gene activity. Such perception described a global negative correlation between DNA methylation and gene expression. Hypermethylation at promoter regions has been linked to decreased expression levels of their downstream genes. In opposite, reduced global methylation that is triggered by administering 5-Azacytidine could reactive gene expression [11], [6].

In 1983, gene expression affected by DNA methylation was studied on globin gene using southern blot analysis by Busslinger *et al.* This study at the resolution of ~ 1 kb reported that the methylation in the gene region of γ -globin prevents its expression [4]. This study has contributed to the perception of the negative correlation between DNA methylation to gene expression.

Hodges, *et al* reported DNA methylation changes during hematopoietic development. In their report they applied a sliding-window of 50 bp to average methylation levels across promoters of genes. They reported that the greatest differential methylation among different cell stages during hematopoietic development is observed at ~ 1 -2 kb downstream of TSSs within the promoter regions. The observed trend is that the overall methylation is selectively reduced in the transcribed genes associated with lineage-specific development . Furthermore, they correlated the methylation and expression levels in 100 bp bins. In all cell types, they observed a negative correlation between methylation and expressions peaking at ~ 1 kb downstream of TSS [8]

In 2014 with more systematic approach the preception was challenged

by Lou, *et al.* They developed a general model capturing the relationship between DNA methylation and gene expression using whole-genome bisulfite sequencing and RNA sequencing in human cells. The DNA methylation and RNA levels of genes were averaged at the resolution of gene length and 2 kb upstream. At this resolution, they could systematically deduce a global trend of their relationship. They observed a ‘L-Shape’ scatterplot between DNA methylation and RNA gene expression. ‘L-Shape’ scatterplot highlights two extreme methylation levels, i.e, high and low levels. The high level of methylation is correlated with low level gene expression. Whereas the low level of methylation is correlated with high level gene expression. This deduced global trend supported the early perception on the function of DNA methylation on gene expression [14].

In 2015, Amabile *et al* studied the function of DNA methylation during leukaemia development. In this study, they compared the DNA methylation profiles between the cancer and iPS (induced pluripotent stem cell)-state cell lines. iPS-state cell lines were generated using reprogramming approach by overexpression of OCT4, SOX2, KLF4, and c-MYC. At iPS-state, the methylation profile of cancer cell line was defaulted into embryonic-like methylation profile [15]. Although, iPS-specific methylation profile was also still present, as reported by Huang *et al* [10]. They compared the mean promoter methylation profiles between the cancer and iPS-state cell lines in 2000bp-window of methylation scores derived from RRBS. From such comparison, they identified 500 gene promoters that are hypermethylated in cancer cell line. These hypermethylated genes were associated with leukaemia-specific DNA methylation. From this study in 2015, by averaging methylation scores

in 2000bp-window they could identify differentially methylated promoter regions of genes which are hypermethylated in leukaemia.

2.2. DNA methylation and nucleosome positioning

Links between DNA methylation and nucleosomal positioning have been investigated from several studies.

A genome-wide study by Kelly, *et al* in 2012 at the scale of 20 bp-resolution in scoring the methylation level, has provided insight on the links between DNA methylation and nucleosomal positioning. In their study, they reported a correlation of DNA methylation and nucleosomal occupancy at promoter regions. Active promoters are characterized by parsed nucleosomes and unmethylated DNA. Whereas inactive promoters were marked by dense nucleosomes and unmethylated DNA (poised/repressed gene expression) or methylated DNA (silent gene expression). Furthermore, they also observed an anticorrelation between DNA methylation and nucleosomal occupancy at CTCF sites. At these sites, DNA methylation level is at its peak in the linker region between nucleosomes. CTCF is being flanked by well-positioned nucleosomes with high level of DNA methylation in between. [12].

2.3. DNA methylation and transcription factor binding

DNA methylation may act as a differential binding site for a transcription factor (TF). This role has been reported in several studies.

A study by Fujimoto, *et al* in 2005, analyzed at the resolution of single CpG dinucleotides from bisulphite genomic sequencing. From this analysis, they reported the effect of adjacent methylation that blocks the AP-1 site,

leads to the reactivation of TrkA gene during cancer progression [7]. At similar single CpG resolution, a study performed by Zhu, *et al* in 2003 could reveal the effect of methylation on adjacent CpG sites on the binding of Sp1/Sp3 in the p21 promoter. The methylation on p21 promoter leads to the inactivation of p21 in human lung cancer, H719 cell line. This is due to the inhibition promoted by methylation of Sp1 bindings [19].

Methylation studies by Kim, *et al* in 2003 at single-base resolution reported that the presence of a methyl mark may change the binding of TF YY1. YY1 binding sites are differentially methylated between two parental alleles. Within YY1 binding site, one CpG dinucleotide is harboured. The methylation of this CpG occurs maternally, but not paternally, abolishes the binding of YY1. Therefore, a single base change, promoted by the presence of a methyl group could disrupt the binding of YY1 [13].

A study by Hu, *et al* in 2013 using systematic protein microarray, has identified 1300 TF in their binding preference to 150 distinct methylated DNA sequences. They carried out bioinformatics analysis at the scale of a single-base. With this single-base resolution, they can reveal the preference of the Transcription factors, among others is KLF4, to the methylated motif sequences [9].

2.4. DNA methylation and genomic imprinting

Epigenetic machinery that can mark parental alleles of genes genome-wide is required to establish genomic imprinting. One of the machinery that has been widely under investigation is DNA methylation. DNA methylation can establish differential parental methylation during gametogenesis by *de novo* testis or ovary-specific methyltransferase. Following gametogenesis, DNA

methylation is maintained stably during embryonic division. The maintenance is achieved by DNA methyltransferase. These DNA methylations acting as a mark are inherited to the next generation independent of sex. For this reason, in the germline the sex-specific DNA marks are being erased by active or passive demethylation [2].

Imprinted genes work in clusters. A single cluster may harbour 3 up to 12 imprinted genes in the range of 80 - 3700 kb. The example of cluster is *igf2* clusters. This cluster is involved in the regulation of fetal growth. In mouse, *H19* is located 90 kb downstream from *igf2*. The differentially methylated region (DMR) controlling *H19* was identified at 2 kb upstream of *H19*. This DMR is paternally-derived identified with RFLP (Restriction Fragment Length Polymorphism) within the size of ~ 4 kb [3].

Maternally-inherited DMR has also been identified with the *igf2r* clusters using RFLP at the resolution ~ 3 kb. This DMR is required for the maternal expression of *igf2r* [17].

With higher resolution at base resolution, Xie, *et al* in 2012 could systematically recover the known DMRs and identify more candidate DMRs (additional 23 DMRs) in mouse brain cells using methylC-seq. Additionally, they also reported the occurrences of non-CG methylation. The methylation scores computed for non-CG methylation, however, at lower resolution. They apply average methylation score in 1kb-windows. This is due to lower methylation signal for non-CG sites. Therefore with lowering the resolution, they could aggregate the methylation score by averaging signals within 1kb-windows. [18]

References

- [1] Amabile, G., Di Ruscio, A., Müller, F., Welner, R. S., Yang, H., Ebralidze, A. K., Zhang, H., Levantini, E., Qi, L., Martinelli, G. et al. (2015). Dissecting the role of aberrant DNA methylation in human leukaemia. *Nature communications* 6.
- [2] Barlow, D. P. and Bartolomei, M. S. (2014). Genomic imprinting in mammals. *Cold Spring Harbor perspectives in biology* 6, a018382.
- [3] Bartolomei, M. S., Webber, A. L., Brunkow, M. E. and Tilghman, S. M. (1993). Epigenetic mechanisms underlying the imprinting of the mouse H19 gene. *Genes & Development* 7, 1663–1673.
- [4] Busslinger, M., Hurst, J. and Flavell, R. (1983). DNA methylation and the regulation of globin gene expression. *Cell* 34, 197–206.
- [5] Collings, C. K., Waddell, P. J. and Anderson, J. N. (2013). Effects of DNA methylation on nucleosome stability. *Nucleic acids research* 41, 2918–2931.
- [6] Feinberg, A. P. and Tycko, B. (2004). The history of cancer epigenetics. *Nature Reviews Cancer* 4, 143–153.
- [7] Fujimoto, M., Kitazawa, R., Maeda, S. and Kitazawa, S. (2005). Methylation adjacent to negatively regulating AP-1 site reactivates TrkA gene expression during cancer progression. *Oncogene* 24, 5108–5118.
- [8] Hodges, E., Molaro, A., Dos Santos, C. O., Thekkat, P., Song, Q., Uren, P. J., Park, J., Butler, J., Rafii, S., McCombie, W. R. et al. (2011). Direc-

- tional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Molecular cell* 44, 17–28.
- [9] Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., Shin, J., Cox, E., Rho, H. S., Woodard, C. et al. (2013). DNA methylation presents distinct binding sites for human transcription factors. *Elife* 2, e00726.
- [10] Huang, K., Shen, Y., Xue, Z., Bibikova, M., April, C., Liu, Z., Cheng, L., Nagy, A., Pellegrini, M., Fan, J.-B. et al. (2014). A panel of CpG methylation sites distinguishes human embryonic stem cells and induced pluripotent stem cells. *Stem cell reports* 2, 36–43.
- [11] Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* 13, 484–492.
- [12] Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P. and Jones, P. A. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome research* 22, 2497–2506.
- [13] Kim, J., Kollhoff, A., Bergmann, A. and Stubbs, L. (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, *Peg3*. *Human molecular genetics* 12, 233–245.
- [14] Lou, S., Lee, H.-M., Qin, H., Li, J.-W., Gao, Z., Liu, X., Chan, L. L., Lam, V., So, W.-Y., Wang, Y. et al. (2014). Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter

- and gene body methylation in transcriptional regulation. *Genome Biol* *15*, 10–1186.
- [15] Mikkelsen, T. S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B. E., Jaenisch, R., Lander, E. S. and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. *Nature* *454*, 49–55.
- [16] Statham, A. L., Taberlay, P. C., Kelly, T. K., Jones, P. A. and Clark, S. J. (2015). Genome-wide nucleosome occupancy and DNA methylation profiling of four human cell lines. *Genomics Data* *3*, 94–96.
- [17] Stöger, R., Kubička, P., Liu, C.-G., Kafri, T., Razin, A., Cedar, H. and Barlow, D. (1993). Maternal-specific methylation of the imprinted mouse *Igf2r* locus identifies the expressed locus as carrying the imprinting signal. *Cell* *73*, 61–71.
- [18] Xie, W., Barr, C. L., Kim, A., Yue, F., Lee, A. Y., Eubanks, J., Dempster, E. L. and Ren, B. (2012). Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* *148*, 816–831.
- [19] Zhu, W.-G., Srinivasan, K., Dai, Z., Duan, W., Druhan, L. J., Ding, H., Yee, L., Villalona-Calero, M. A., Plass, C. and Otterson, G. A. (2003). Methylation of adjacent CpG sites affects Sp1/Sp3 binding and activity in the p21Cip1 promoter. *Molecular and cellular biology* *23*, 4056–4065.