論文の内容の要旨

Observing comprehensive DNA methylomes via single molecule real-time sequencing: application to diploid and centromeric methylation

(一分子 DNA シーケンサによる DNA メチル化情報の網羅的観測手法 ― 二倍体ゲノムとセントロメア領域への応用)

鈴木 裕太

1. CpG methylation detection from kinetics information of SMRT sequencing data

SMRT (single molecule real-time) sequencing, or PacBio sequencing, has been adopted in hundreds of sequencing studies despite its relatively high cost and raw read error rate, because it can produce longer reads than conventional NGS (next generation sequencers) and its random error profile eventually enabled extremely accurate genome assembly. It is also useful in epigenetics studies as it can produce kinetics information that reflects the methylation status of DNA sample. However, methylation analysis using SMRT sequencing was not applied to vertebrate genomes including human genome, as no method had enough power to detect cytosine methylation accurately. I developed an algorithmic strategy, *AgIn*, to extract methylation information from SMRT reads of practical sequencing depth, and I reported the method can achieve good detection accuracy (~93% sensitivity and precision for detecting unmethylated CpGs) with reads of depth ~30x (Figure 1). The method was successfully applied to catalog methylation statuses of repetitive elements in human genome and extremely homologous (>99% similarity over 4.6kbpp) *Tol2* transposon in medaka genome [1]. The method was continuingly adapted to newer version of SMRT sequencing protocol and it currently works well with the latest P6-C4 chemistry for the PacBio RSII instrument.

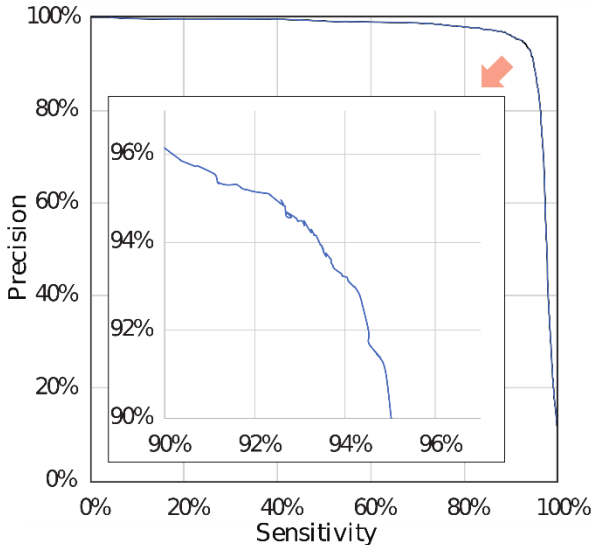

Figure 1. Prediction performance of AgIn method

2. Allele-specific methylation analysis using SMRT sequencing

This study is essentially an extension of AgIn to resolve another difficulty in the epigenetics studies. In diploid genomes, the methylation status of CpG sites in the same region can be different for two homologous chromosomes, and such situation is known as ***allele-specific methylation*** (ASM) events. ASM can regulate gene expression; for example, ***genomic imprinting***, in which the imprinted genes are expressed only from paternal or maternal chromosomes, are largely explained by the existence of ASM at the imprinting center. Also,

disruption of ASM status is known to cause diseases. Several methods were developed to detect ASM events genome-widely, from use of methylation-sensitive restriction enzymes to model-based estimation using short read bisulfite sequencing data, but these methods were unable to observe methylation status of the genome comprehensively for various reason. To overcome the situation, I developed a new strategy to observe directly genome-wide ASM events using SMRT sequencing reads, noting that the primary difference between two homologous chromosomes was nothing but heterozygous SNVs (single nucleotide variants), thus any method to detect ASM should relate heterozygous SNVs and methylation status around them. The proposed method assumes the availability of heterozygous SNV sites and, especially, their phasing information. This may sound demanding at first, but recent advent of linked-read technology (from 10x Genomics) lowered this hurdle. After aligning SMRT reads onto reference genome, the reads were separated according to alleles of heterozygous SNVs they contained, giving two sets of reads each represents single allele. Then, for each set, AgIn was applied to call CpG methylation status of the regions. By applying this strategy to two samples, AK1 (Asian Korean) and HG002 (Ashkenazim), I successfully identified thousands of CpG islands (CGIs) which shown ASM. The CGIs with strongest ASM signal were often located around the promoter regions of imprinted genes such as *TP73*, *ZNF597*, *ZNF331*, *HYMAI*, *MEST*, *PEG3*, *PEG13*. As a result, the list of CGIs with strong ASM signal was significantly (p=0.007, U test) populated with those were associated with imprinted genes. I also found that these ASM CGIs had unique distribution within genome in terms of chromatin state defined in ENCODE project; while most general CGIs were in TSS-like regions, ASM CGIs were rather found in actively transcribed regions or repressed regions. As an individual example, Figure 2 depicts the observed methylation statuses over the *GNAS* complex locus in AK1 genome. The locus is known to show a complex expression patterns regulated by ASM. I confirmed the ASM pattern of the locus was consistent with the known ASE (allele-specific expression) pattern. In collaboration with University of Iowa, I compared ASM calls with ASE data generated from

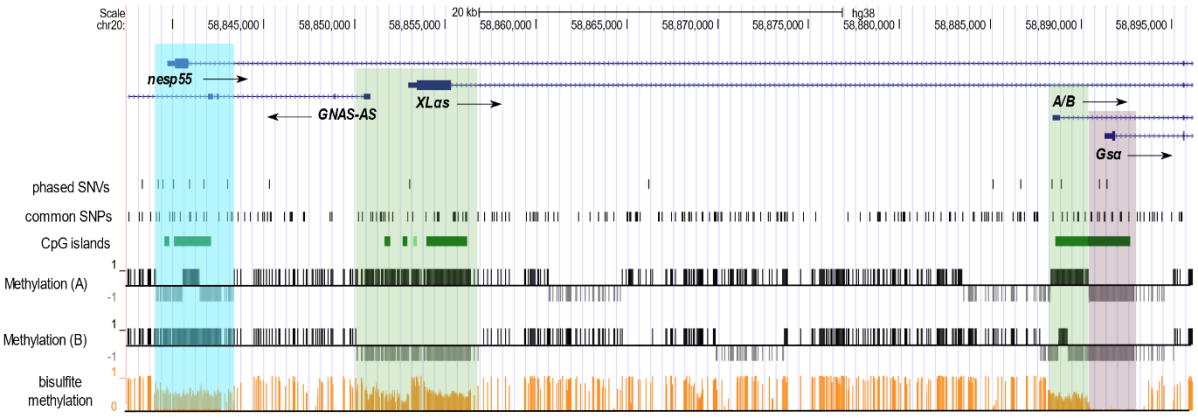

Figure 2. Allele-specific methylation over the GNAS locus. Genes in each of four shaded regions are known to be expressed (from the left) maternally, paternally, paternally, and biallelically.

long reads and short reads RNA-seq. As expected, I found that the expressing allele and the unmethylated allele were coincide for the ASM/ASE regions. Based on this observation, since ASE was much difficult to detect comprehensively due to scarcity of SNVs within exons, I claimed ASM can be a surrogate for ASE status of the gene. I also claimed these findings were possible only by using long reads because the majority of CpG sites in personal human genomes were located distant from any heterozygous SNV, which were sparsely distributed. This work is in preparation for publication [2].

3. CpG methylation analysis of centromeric repeat reagions in medaka genome

Centromeres were possibly the most difficult regions in any genome sequencing study, and epigenetic characterization of centromeres was largely indirect and descriptive as conventional method such as bisulfite sequencing could not observe methylation over its highly repetitive (Mbp-scale arrays of alpha-satellite) structure. I applied AgIn algorithm to medaka centromeres, which were assembled in contigs in the latest version of medaka genome [3], and identified the regions with unmethylated CpGs from a total of 11 chromosomes of two medaka inbred strains, Hd-rR and HSOK, which diverged ~2.5Mya (SNP rate ~ 2.5%) (Figure 3). By analyzing the sequence composition (k-mer distribution) of unmethylated and methylated repeats, I claimed these variations in methylation occurred recently, at least after the divergence of two strains (Figure 4). Therefore, it implied that the change in methylation status in each chromosome or strain could be independent, and I hypothesized that it may precede ultimate alteration in functionality of centromeres. I validated my methylation calls in centromeric regions by comparing them with calls from bisulfite sequencing data wherever they are available, although I found bisulfite sequencing was not sufficient to observe the overall picture of centromeric methylation patterns.

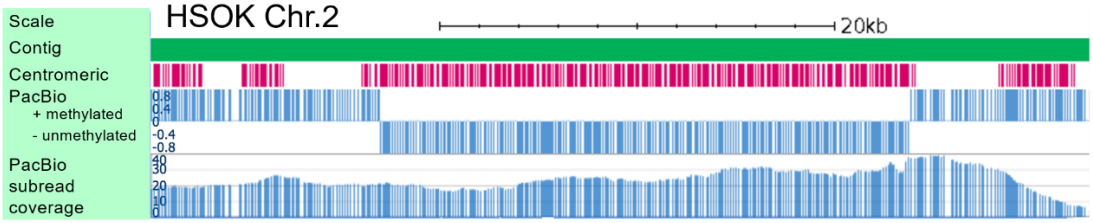

Figure 3. Methylation of centromeric repeats in medaka genome, HSOK chromosome 2.

Conclusion

In these works, I demonstrated the utility of SMRT sequencing for epigenetics study, especially for allele-specific methylation analysis and methylation analysis in complex region such as centromeres. The analyses of ASM in human genomes could retrieve imprinted genes and was consistent with the expression pattern of the ASE genes, which validated the accuracy of the method. By applying the method to centromeric repeat regions, I uncovered the cryptic methylation patterns of the regions, arriving at the hypothesis on an evolutionary drive of centromere sequences.

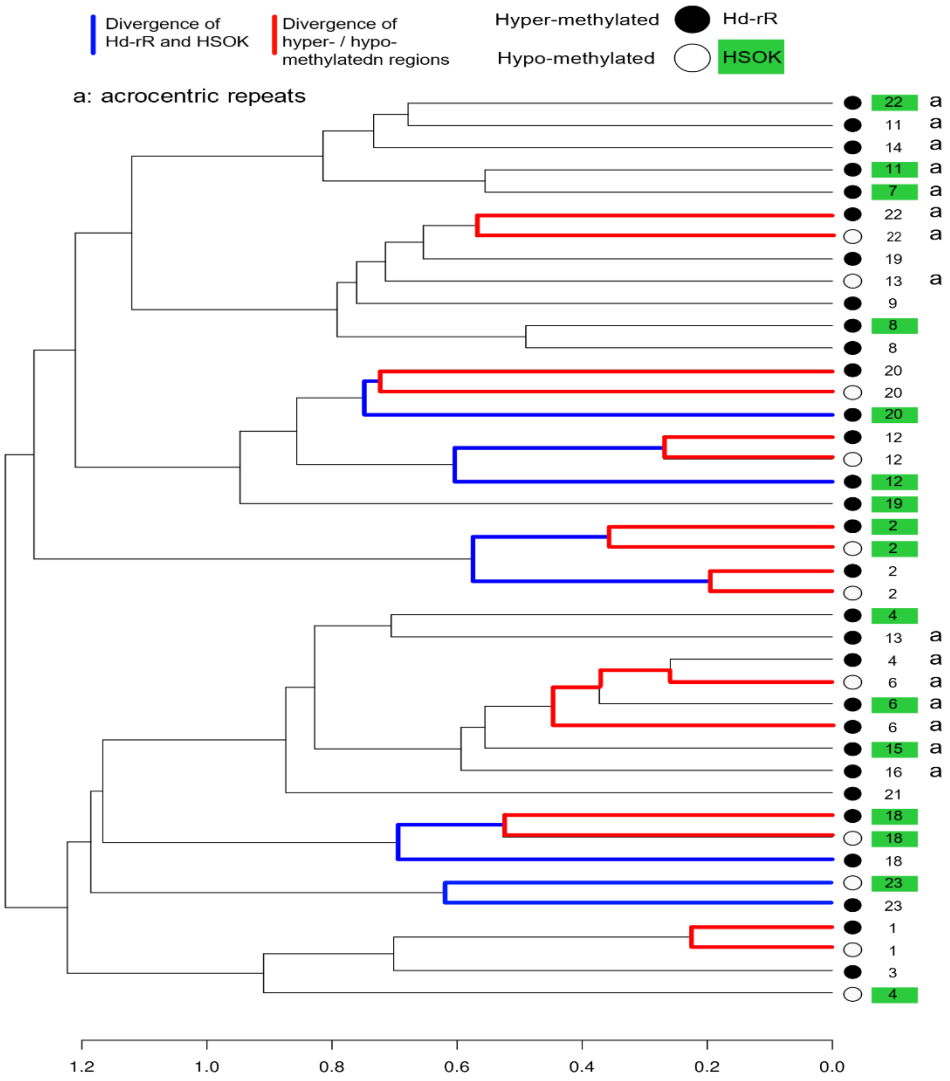

Figure 4. Clustering based on sequence similarity of centromeric repeats with unmethylated or methylated CpGs. Assuming the sequence similarity reflects the evolutionary distances, divergence of two strains precedes diversification of methylation pattern.

References

[1] Suzuki et al. "AgIn: measuring the landscape of CpG methylation of individual repetitive elements." *Bioinformatics*, 32, 19 **(2016)** 2911-2919.

[2] Suzuki et al. "Personal diploid methylomes and transcriptomes via phased heterozygous variants and single-molecule real-time sequencing." *in preparation*.

[3] Ichikawa, Tomioka, Suzuki, Nakamura, Doi et al. "Centromere evolution and CpG methylation during vertebrate speciation." *Nature Communications.* ", 8 **(2017)**: 1833.