"Understanding epigenetic regulation in bacteria – one base at a time"

DNA methylation (addition of a $–CH_3$ group to one of the four bases, A, C, G or T) is one way by which the expression of genes is controlled in many organisms. Research has demonstrated additional roles of DNA methylation in bacteria in recognition of foreign DNA, restriction-modification (R-M) systems, and antibiotic resistance [1,2]. Single molecule sequencing can be adapted to identify DNA methylation patterns at single cell resolution. Beaulaurier et al. propose SMALR (Single-molecule Modification Analysis of Long Reads), a technique that combines Single Molecule Real-Time (SMRT) DNA sequencing of long reads with a loop over smaller sub-sequences (100-250 bp) from each long read, improving detection of methylated bases and concurrently overcoming the high error rate present in single-molecule sequencing output [1]. They demonstrate the effectiveness of SMALR by using it to characterize heterogeneity within genome-wide methylation patterns (methylomes) of 6 different bacteria, among other things.

Background description

Current approaches for identification of different methylation patterns (N^6-methyladenine (6mA), 5-methylcytosine (5mC) and N^4-methylcytosine (4mC)) from genomic data require chemical treatment of the DNA prior to sequencing - a process called bisulfite sequencing. The bisulfite treatment converts un-methylated cytosine (C) to a Thymine (T), and leaves the methylated C's (mC) as C's, making it difficult to distinguish 4mC from 5mC [2]. This approach also lacks the resolution necessary to identify genome-wide methylation activity, since smaller reads restrict the reconstruction of complex and repetitive regions of the bacterial genome.

SMRT technology uses differences in speed of integration of new bases (and base modifications) during DNA replication (Inter-Pulse Duration, *IPD*) to identify the different bases being added (**Figure 1**). Hence, the DNA remains in its native form, and the long reads enable us to 'read' longer stretches of DNA in real time, providing base-level resolution of complex and repetitive

sequences. However, this approach is confounded by the need for multiple observations from long DNA molecules in order to identify base-specific IPD values. Furthermore, there is important regulatory potential of 4mC and 6mA in bacterial R-M systems, and a method for accurately identifying these methylation marks will aid research in this direction [2]. In order to address these issues, the authors developed SMALR, a framework that combines a single long-read SMRT sequencing step with multiple SMRT sequencing steps for smaller 'sub-reads'.

Description of research (methods and results)

SMALR identifies specific patterns of change in IPD by pooling information from single-pass long read SMRT sequencing and deep coverage of small read SMRT sequencing on the sub-reads that make up the long DNA strand. Multiple sequencing of the smaller reads (circular-consensus sequencing) provides multiple observations of IPDs for each base in the sequence. These values are pooled together for each base. This is done for both the methylated (native) and corresponding unmethylated (Whole Genome Amplified, *WGA*) sequences. For each base, the mean log WGA IPD value for each base is deducted from the mean log IPD value of the matching native base, to calculate a single molecule, single nucleotide (SM_{SN}) score (**Figure 2**). The same approach, when applied to entire molecules, gives an Agg_{SN} score for each DNA strand/position (**Figure 2**). Methylated sites have a high SM_{SN} score (**Figure 3b**).

The authors used the SM_{SN} approach to identify 6mA in known methylated 5'-CTGCAG sites in a specific *E. coli* strain, using matching WGA samples as unmethylated controls. As expected, they found that SM_{SN} scores improve detection of 6mA signal when present (sensitivity) and discriminate between 6mA and other signals (specificity) with increase in coverage of each small subread (**Figure 3a**). They also attained high sensitivity (98.5%) and specificity (99.5%) using minimum per-molecule coverage of 15. Similar results were obtained for 5mC detection. In addition, they took previously annotated methylation sites (methylation motifs) in different

bacteria and distinguished two distinct groups of motifs (methylated and unmethylated) based on their SM_{SN} scores. An analysis of all bacterium-motif pairs showed that while most motifs were methylated across all bacteria, *C. salexigens and H.pylori* also had a high density of non-methylated motifs (**Figure 4**). This heterogeneity could be result of environment-influenced variation in the methylase activity in different cells (phase variation) [1,3]. Since the SMALR approach allows strand-specific mapping of the methylation events, the authors used the data to test for phase variation in methylation motifs that were targeted by specific phase-variable methylases in *H. pylori* colonies. As expected, they found clear unimodal distributions of the pooled IPD scores (SM_P), indicative of motif methylation in colonies with active phase-variable methylases, and quantifiable minor subpopulations with inactive methylases (**Figure 5**). Furthermore, quantification of SM_{SN} scores at 5 different loci on the *C. crescentus* genome at different time-points during DNA replication revealed a trend towards hemimethylation at initially methylated 5'-GANTC sites, in keeping with the expected methylation patterns around the replication fork. Interestingly, the authors found that the terminus (*Ter*) re-methylated much quicker after the passage of the replication fork than other genomic regions. Research can be done on the replication processes influencing synchronization of bacterial genome methylation.

Discussion

This approach cleverly overcomes major barriers in single-molecule analysis of methylomes, by using a combination of short- and long- libraries to address high error rates seen in SMRT sequencing. In addition to the applications outlined in the paper, the ability to apply this approach in a reference free, *de novo* manner expands its application to methylome sequencing of DNA from bacterial species that have draft (incomplete) reference genomes available (or no reference genome available at all!). It can also be used to accurately identify different strains of bacteria in a mixed cell population (metagenomics). Lastly, it has been reported elsewheres that

R-M systems are a major barrier to DNA transformation (adding foreign DNA to a cell) in bacteria [2]. SMALR can be used to investigate if differences in the sequence context of 4mC, 5mC, and 6mA have any correlation with difficulty in transforming different bacterial species.

Figures and Tables

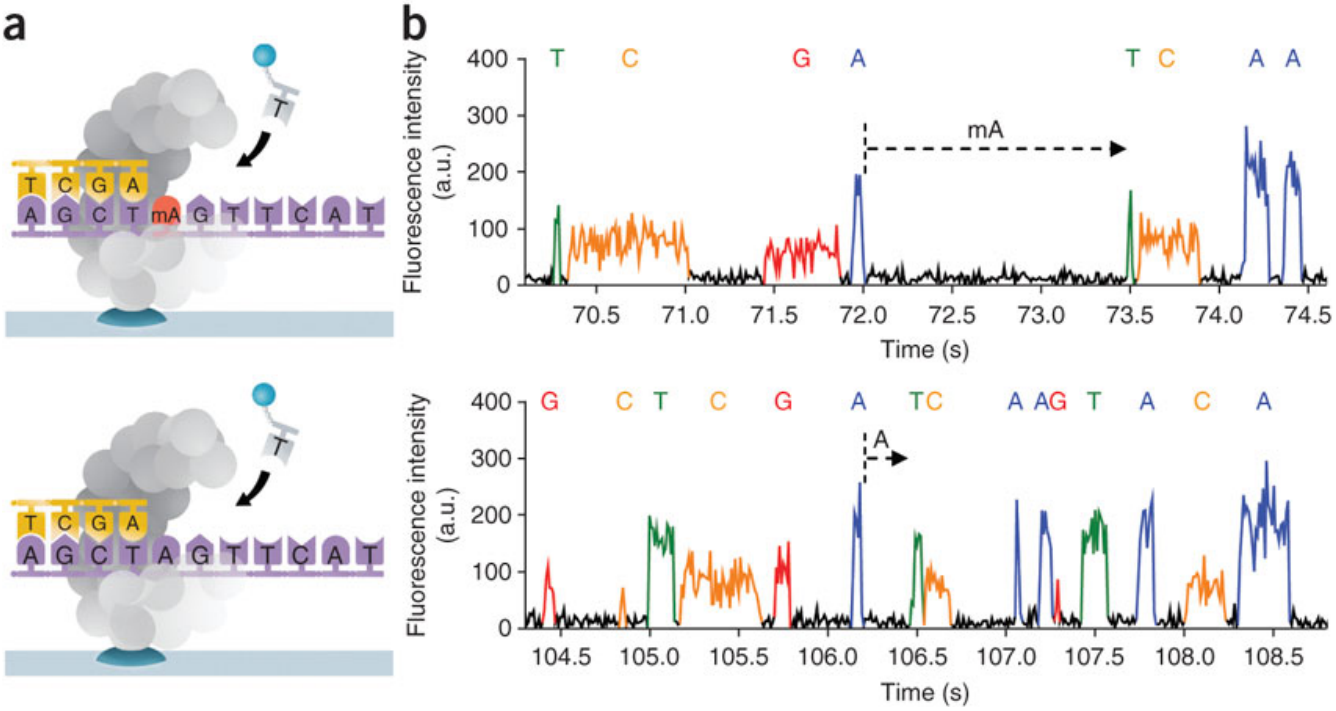

Figure 1. Differences in Inter-Pulse Duration can be used to identify which base has been added by the polymerase during strand extension[4]. The addition of each base by the polymerase generates a fluorescent emission, with different colours indicating the 4 different bases. The time between 2 fluorescent emissions indicates the time taken for the nucleotide to be incorporated. In this particular case, the addition of a Thymine complementing the mA takes longer (upper block), as compared to the addition of a Thymine complementing the un-methylated A (lower block).

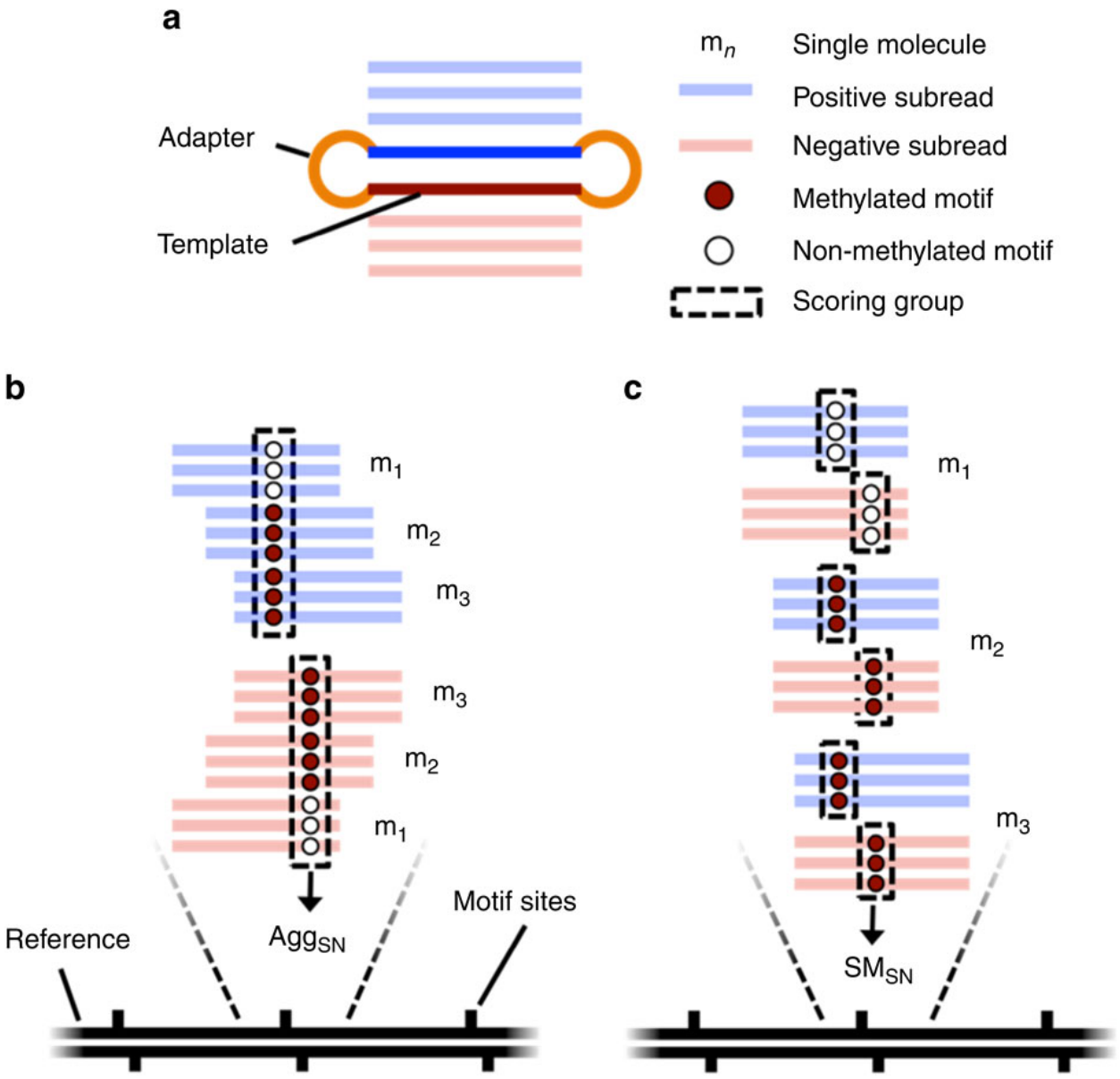

Figure 2. Method for calculation of SM_{SN} score and Agg_{SN} score using short-read data combined with long-read data. (a) A single SMRT sequencing molecule (short DNA sequence + circular adapters) can be used as template to sequence the same molecule multiple times and generate sub-reads. **(b)** The IPD values from all the subreads aligning to a given strand, at a given genomic position, can be aggregated across all molecules in order to identify a consensus methylated base (Agg_{SN} value). **(c)** Alternatively, each subread from each molecule can be considered separately, and the SM_{SN} score can be calculated for each molecule (and then the strand, and genomic position) in order to identify a consensus methylated base.

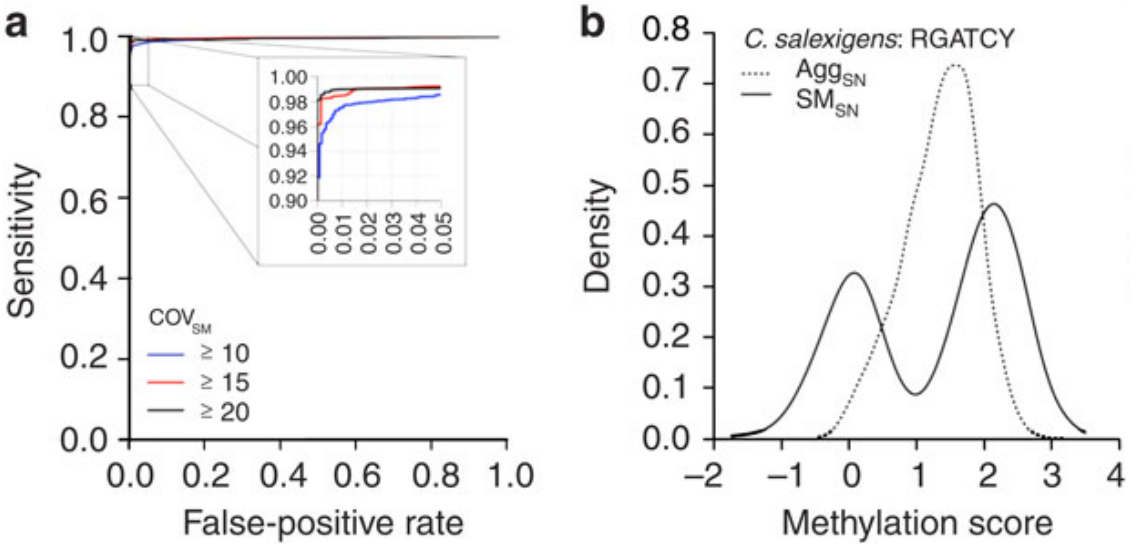

Figure 3. Performance of SM_{SN} scores in detection of DNA methylation in *E. coli* O104:H4 C227-11 strain. (a) The performance of the SM_{SN} score in detecting 6mA DNA methylation at 5'-CTGCAG motif improves as the coverage of each single molecule increases (cov_{SM}). **(b)** The aggregate, single nucleotide score across all molecules spanning a position (Agg_{SN}), as compared to the bimodal distribution evident when calculating molecule-specific scores (SM_{SN}) for partially un-methylated 5'-RGATCY motif in *C. salexigens*. The bimodal distribution enables accurate and objective estimation of these distinct fractions.

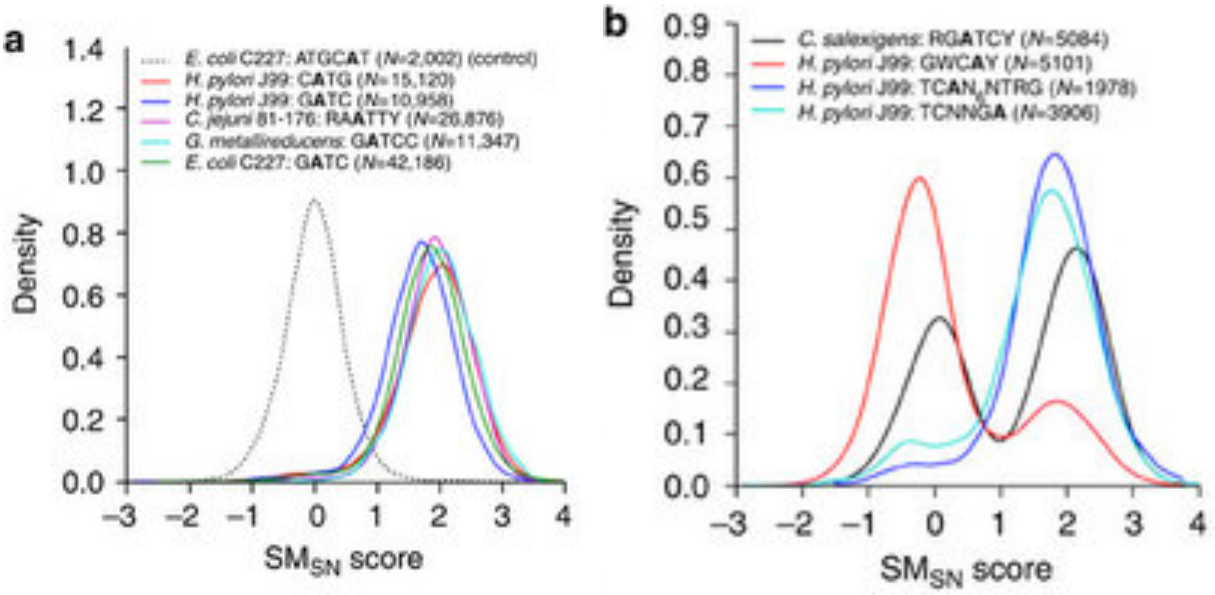

Figure 4. SM_{SN} scores reveal epigenetic heterogeneity in bacterial samples. **(a)** SM_{SN} scores for bacterium-motif pairs, for methylated motifs. A non-methylated motif is shown for comparison. **(b)** Among the bacterial species that showed significant non-methylated motif fractions (*H. pylori* and *C. salexigens*), a minor variation can be seen in the SM_{SN} scores associated with each motif peak. This is due to subtle differences in the chemistry version used for SMRT sequencing of the native and WGA samples.[1]

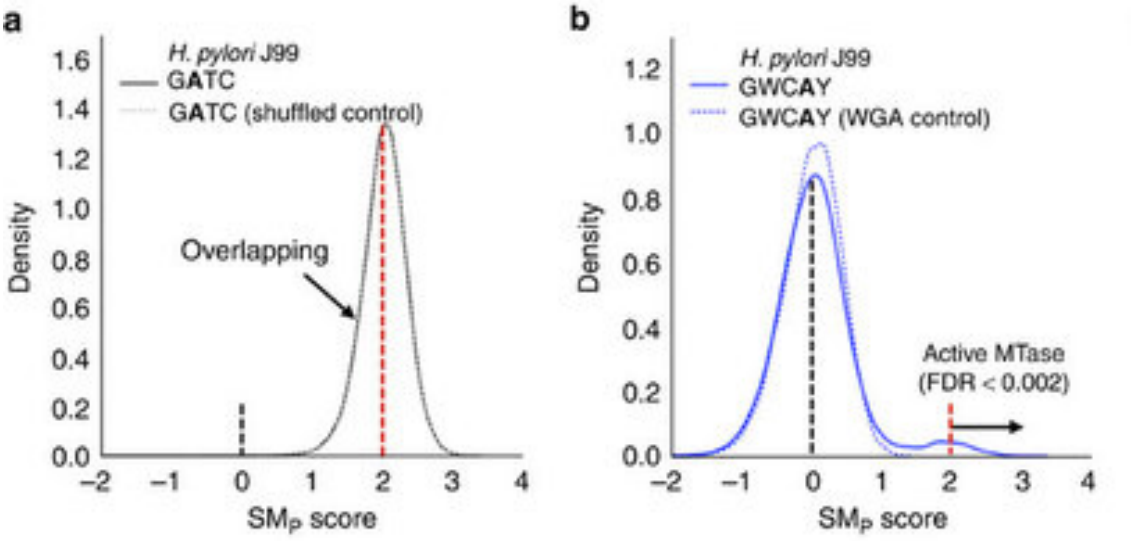

Figure 5. Identification of differences in SM_p score distributions due to phase-variable methylation in *H. pylori* J99. (a) SM_p distribution for *H. pylori* J99 motif 5'-GATC and a shuffled unmethylated control shows a single peak at 2, indicating a fully active methylase acting at the site. **(b)** SM_p distribution for *H. pylori* J99 motif 5'-GWCAY and a WGA unmethylated control shows a major peak around 0, and a minor peak around 2. This is indicative a mostly inactive methylase around 5'-GWCAY in most of the cells.

References

1. Beaulaurier, John; Zhang, Xue-Song; Zhu, Shijia; Sebra, Robert; Rosenbluh, Chaggai; Deikus, Gintaras; Shen, Nan; Munera, Diana; Waldor, Matthew K. (2015-06-15). **"Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes".** *Nature Communications* **6**: 7438. doi:10.1038/ncomms8438. PMC 4490391. PMID 26074426.
2. Yu, Miao; Ji, Lexiang; Neumann, Drexel A.; Chung, Dae-hwan; Groom, Joseph; Westpheling, Janet; He, Chuan; Schmitz, Robert J. (2015-12-02). **"Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite-sequencing".** *Nucleic Acids Research* **43** (21): e148. doi:10.1093/nar/gkv738. ISSN 0305-1048. PMC 4666385. PMID 26184871.
3. van der Woude, Marjan W.; Bäumler, Andreas J. (2004-07-01). **"Phase and Antigenic Variation in Bacteria".** *Clinical Microbiology Reviews* **17** (3): 581–611. doi:10.1128/CMR.17.3.581-611.2004. ISSN 0893-8512. PMC 452554. PMID 15258095.
4. Flusberg, Benjamin A.; Webster, Dale R.; Lee, Jessica H.; Travers, Kevin J.; Olivares, Eric C.; Clark, Tyson A.; Korlach, Jonas; Turner, Stephen W. (2010-06-01). **"Direct detection of DNA methylation during single-molecule, real-time sequencing".** *Nature Methods* **7** (6): 461–465. doi:10.1038/nmeth.1459. ISSN 1548-7105. PMC 2879396. PMID 20453866.