

DNA Base Modification Detection using Single Molecule, Real-Time Sequencing

Introduction

Base modifications are important to the understanding of biological processes such as gene expression, host-pathogen interactions, DNA damage and DNA repair¹. Single Molecule, Real-Time (SMRT[®]) sequencing has the potential to revolutionize the study of base modifications through direct detection of unamplified source material.

Traditionally, it has been a challenge to study the wide variety of base modifications that are seen in nature. Most high-throughput techniques focus on cytosine methylation -- made accessible through bisulfite treatment when sequencing or PCR-based techniques to detect the methylation at a single base resolution². SMRT sequencing, in contrast, does not require genetic alterations to the source material in order to view base modifications. Instead, the kinetics of base addition is measured during the normal course of sequencing. These kinetic measurements present characteristic patterns in response to a wide variety of base modifications.³

As a result of this breakthrough methodology to detect base modifications, it is now possible to sequence modifications other than 5-methylcytosine. Bacterial modifications such as 6-methyladenine, 4-methylcytosine, or more recently identified eukaryotic modifications such as 5-hydroxymethylcytosine⁴, are accessible to study using a single sequencing method on the PacBio[®] RS system. As our understanding of kinetic information grows, the analysis of base modifications using SMRT technology will continue to become easier and faster.

Types of Base Modification in Biology

DNA base modifications have a variety of functional roles which include:

- Epigenetic markers for influencing gene expression such as 5-methylcytosine, 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine.
- Bacterial identity markers for affecting host-pathogen interactions such as 6-methyladenine, 4-methylcytosine, and 5-methylcytosine.
- Bacterial epigenetic markers for regulating DNA replication and repair and transcription regulation such as 6-methyladenine.
- Products of DNA damage such as 8-oxoguanine and 8-oxoadenine.

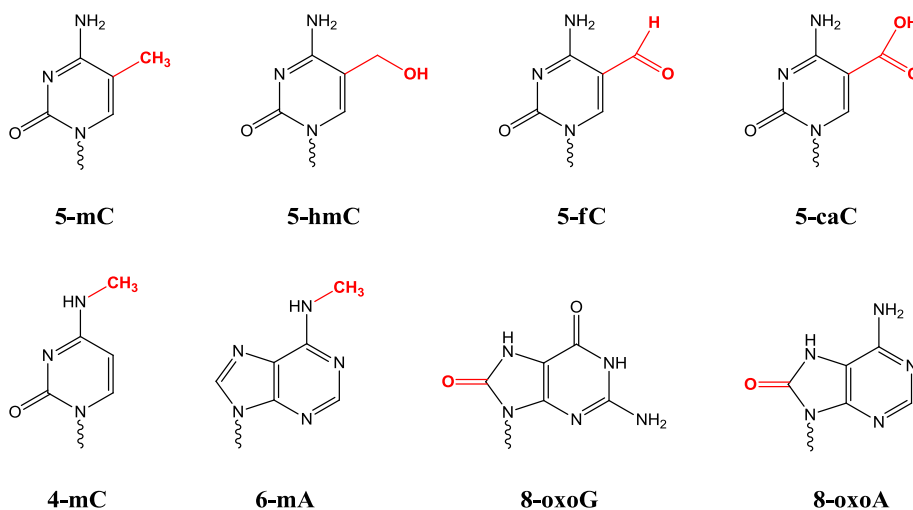


Figure 1. Molecular structures of base modifications including 5-methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC), 5-carboxylcytosine (5-caC), 4-methylcytosine (4-mC), 6-methyladenine (6-mA), 8-oxoguanine (8-oxoG), and 8-oxoadenine (8-oxoA).

Today, the most commonly studied base modification is 5-methylcytosine, commonly referred to as “methylation” despite the fact that other bases are also naturally modified by methyl groups. Methylation has gathered interest from researchers in a variety of disciplines including early developmental biology, cancer biology, and neurological disorders. Methylation and de-methylation help regulate gene expression and have been linked to several human diseases through mechanisms such as deactivating tumor suppressors or activating oncogenes⁵.

Other modifications, such as 6-methyladenine (6-mA) in bacteria, have been studied with lower resolution methods – such as chromatography or through methylation’s protective effect against restriction endonucleases – because they are not easily accessible with standard sequencing techniques. This modification is associated with basic functions such as DNA replication and repair⁶. It is also common in protists and plants, and some studies suggest that it may also be present in mammalian DNA⁷.

SMRT sequencing is capable of detecting 6-mA as well as other common bacterial base modifications. As a result, the technology is expected to increase our understanding of a broad array of biological processes. The potential benefits of detecting base modification, using SMRT sequencing, include:

- Single-base resolution detection of a wide variety of base modifications (including those in Figure 1 and more).
- Single-molecule resolution over long-read distances.
- Unamplified double-stranded input DNA, which means that strand-specific modifications, such as hemimethylation, are detectable.
- Hypothesis-free base detection which allows discovery of unknown or unexpected modifications through the effects on sequencing kinetics (as described below).

Studying Polymerase Kinetics with SMRT Sequencing

SMRT Sequencing allows the observation of single DNA polymerases reading individual molecules of DNA in real time. Therefore, the kinetics of DNA polymerization is observable on a single-molecule basis. The kinetic characteristics, such as the time duration between two successive base incorporations, are altered by the presence of a modified base in the DNA template³. This manifests as an increased space between fluorescence pulses, which is called the interpulse duration (IPD), as shown in Figure 2 below.

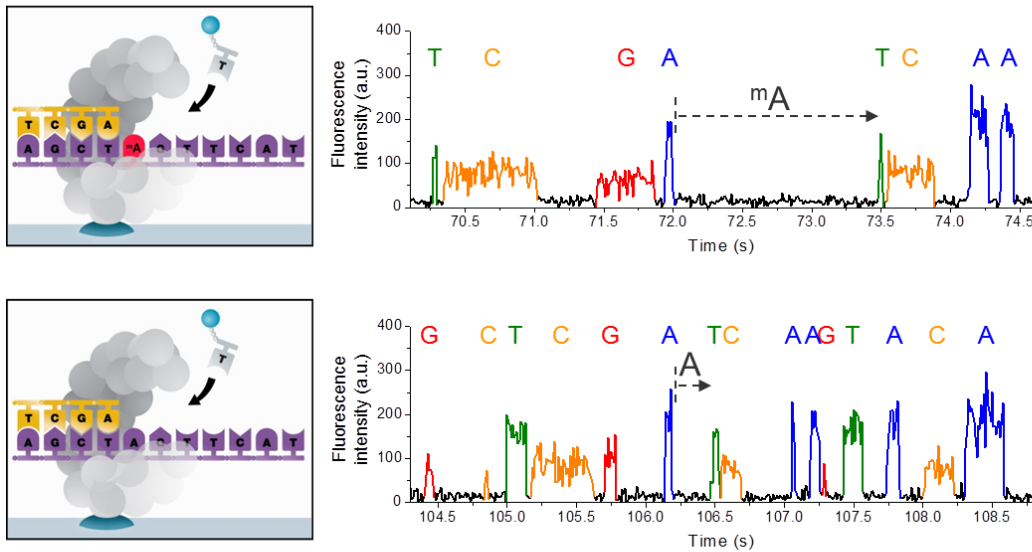


Figure 2. Principle of detecting modified DNA bases during SMRT sequencing. The presence of the modified base in the DNA template (top), shown here for 6-methyladenine, results in a delayed incorporation of the corresponding T nucleotide, i.e. longer interpulse duration (IPD), compared to a control DNA template lacking the modification (bottom).³

These changes in the DNA polymerase speed, relative to an unmodified DNA control template lacking modified bases, can be measured for each template position to indicate the presence of modified bases in the DNA template. In order to quantify the change in IPD distributions between a control sample and a native sample, we define the *IPD ratio* as the ratio of the mean IPD in the native sample to the mean IPD in the control samples.

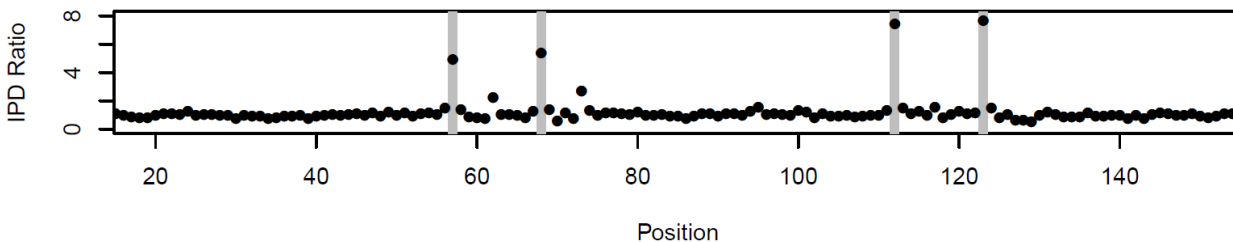


Figure 3. Detection of 6-methyladenine using IPD ratios. For each template position (x-axis), the ratio (y-axis) of average interpulse durations (IPDs) for the native DNA and the unmodified control template is plotted. Excursions from the baseline indicate the presence of a modified base (four 6-methyladenines in this example marked with vertical grey bars), slowing down the polymerase at least five-fold at the position of the modification⁹.

The kinetic effects are not necessarily limited to just the nucleotide incorporation across the modified base position in the DNA template. This is because the DNA polymerase is in intimate contact with the DNA over an extended region of approximately eleven bases, and the modified base can impart effects on the polymerase dynamics at several positions over this region. This results in kinetic “signatures” which can aid in the identification of the type of base modification. For example, for the three common bacterial identity markers⁸, 5-methylcytosine has typical characteristic kinetic signals two and six bases downstream of the methylated position, 6-methyladenine at the modified position and five bases downstream, and 4-methylcytosine just at the position of the modification (Figure 4).

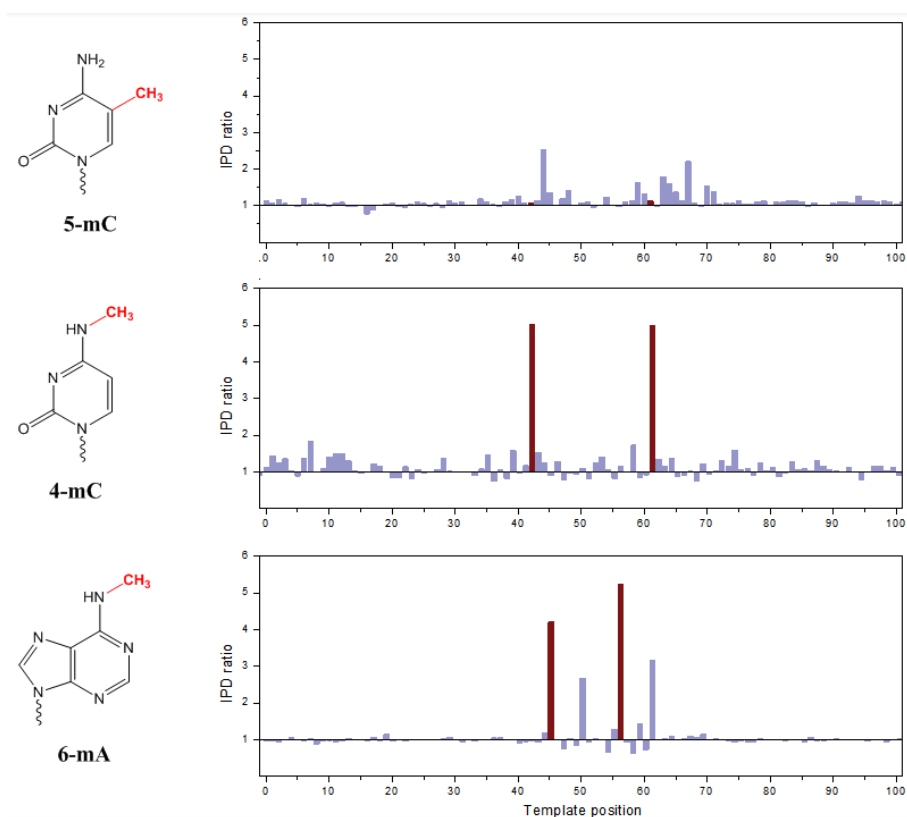


Figure 4. Kinetic signatures for the three common bacterial identity markers 5-methylcytosine (5-mC), 4-methylcytosine (4-mC), and 6-methyladenine (6-mA). The methylated template positions are highlighted in red. Note that the kinetic signatures vary both in magnitude as well as in the length of the region over which the polymerase dynamics are affected.

The different signal magnitudes translate to different amounts of sequencing-fold coverage required to obtain similar confidence levels of detection. In addition, the shape and strength of the kinetic signature depends on the surrounding sequence context. Future commercial products are expected use the sequence context to better interpret the modification signal.

Based on the measurement and analysis principle outlined above, a typical experiment for detecting DNA base modifications is designed as follows:

- Obtain unamplified DNA of interest and prepare a SMRTbell™ library for SMRT sequencing (*see the Pacific Biosciences Sample Preparation and Sequencing Guide*).
- Use a small aliquot of the original DNA sample and perform a whole-genome amplification (WGA) reaction to obtain the control DNA sample lacking any base modifications. Then prepare a SMRTbell library for this sample.
- Perform adequate SMRT sequencing for both samples to obtain the necessary coverage needed for the modifications under study, and sufficient overall coverage to characterize the genome or portion of the genome of interest.
- Use the bioinformatics tools to perform kinetic analysis for base modification⁹.

Examples of Current Applications

Direct DNA sequencing of base modifications requires that the following general considerations be fulfilled during the experimental design:

1. Unamplified DNA is necessary. Amplification of DNA through PCR, whole-genome amplification (WGA), or other techniques, will result in the loss of base modifications. Therefore, enough unamplified DNA has to be available to make a sequencing library. In late 2011, the sample requirement for preparing a ~500 bp SMRT sequencing library is 250 ng of input DNA. As DNA input requirements are reduced over time, an increasingly small or targeted region of a genome will be sufficient for studies without amplification requirements.
2. The different magnitudes of kinetic signal for the different base modifications translate to different amounts of sequencing fold coverage per template position required to obtain adequate and similar confidence levels of detection. As of late 2011, we recommend the minimum sequencing fold coverage *per strand* as follows⁹:

4-methylcytosine	25x
5-methylcytosine	250x
5-hydroxymethylcytosine	250x
glucosylated 5-hydroxymethylcytosine*	25x
hydroxymethylcytosine enriched with the Hydroxymethyl Collector™ Kit†	5x
6-methyladenine	25x
8-oxoguanine	25x

3. The genome, or portion of a genome, to be interrogated also has to be compatible with the current throughput performance of the PacBio RS. As of late 2011, the base throughput of a two hour run on the PacBio RS is 90 Mb. Over time, we expect SMRT technology throughput to increase, making larger genome sizes more practical to sequence for base modifications.

* Generated by T4 Phage β-glucosyltransferase (Josse, J. and Kornberg, A. (1962) J. Biol.Chem., 237, 1968-1976)

† Available from Active Motif (<http://www.activemotif.com/catalog/775/hydroxymethyl-collector-trade>)

Applications particularly suitable for SMRT sequencing, based on the criteria above, include:

- Bacterial whole-genome base modification studies. For example, sequencing of *E. coli* to find the sites of 5-mC, 4-mC, and 6-mA modification when studying virulence, gene expression, or pathogen-host interactions.
- Isolated and purified mitochondrial, chloroplast or other small genomes, containing known or novel DNA base modifications.
- Enrichment of portions of a larger genome, for example isolation of DNA regions modified with hydroxymethylcytosine (5-hmC), and enriched by specific biotinylation chemistries of 5-hmC and subsequent streptavidin pulldown.

Conclusion

SMRT sequencing is the only commercially available technology capable of measuring the kinetics of base incorporation during a sequencing run. This kinetics information can be used to identify sites in the target DNA that have been chemically modified in a variety of ways including methylation, formylation, carboxylation, and more. These base modifications are associated with several biological processes including gene expression and DNA oxidation damage.

Most of these modifications have not been extensively studied due to difficulties in adapting experimental techniques to a higher throughput sequencing method – we expect that these modifications will be possible to study with SMRT sequencing. Examples of novel studies that may be performed on the PacBio *RS* today include full-genome bacterial modification studies and hydroxymethylcytosine studies in mammalian genome regions enriched for that modification.

We anticipate that over time the range of target genomes that will be addressable by SMRT sequencing will improve. We expect to ultimately address regions as small as single genes as well as larger whole genomes. We further expect that the advances afforded by SMRT sequencing will enhance genetic studies into gene regulation, DNA damage and repair, bacterial virulence, and other important biological pathways. Novel DNA modifications could potentially be discovered, expanding the utility of sequencing to even broader areas of study.

Bibliography

1. Trygve Tollefsbol (2010) Handbook of Epigenetics: The New Molecular and Medical Genetics. Academic Press.
2. Clark S.J., Statham A., Stirzaker C., Molloy P.L. & Frommer, M. DNA methylation: bisulphite modification and analysis. Nat. Protocols 1, 2353–2364 (2006).
3. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods 7:461-465.
4. Kriaucionis S. & Heintz N. The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. Science 324, 929–930 (2009); Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324, 930–935 (2009).
5. Esteller M. Epigenetics in cancer. New England Journal of Medicine 358, 1148-1159 (2008).
6. Marinus M.G. & Casadesus, J. Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. FEMS Microbiol. Rev. 33, 488–503 (2009).
7. Ratel D, Ravanat JL, Berger F, Wion D, N6-methyladenine: the other methylated base of DNA. BioEssays 28,309-315 (2006).
8. Roberts R.J., Vincze T., Posfai J. and Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 38, D234-236 (2010).
9. "DNA Modification Detection with SMRT Sequencing using R" found at <https://github.com/PacificBiosciences/R-kinetics>.

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2011, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at <http://www.pacificbiosciences.com/licenses.html>.

Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT and SMRTbell are trademarks of Pacific Biosciences in the United States and/or certain other countries. All other trademarks are the sole property of their respective owners. 11/07/2011 PN 001-640-287-01