# Mapping and characterizing N6-methyladenine in eukaryotic genomes using single molecule real-time sequencing

Shijia Zhu, John Beaulaurier, Gintaras Deikus, et al.

*Genome Res.* published online May 15, 2018
Access the most recent version at doi:10.1101/gr.231068.117

# Mapping and Characterizing N6-methyladenine in Eukaryotic Genomes using Single Molecule Real-Time Sequencing

Shijia Zhu[1], John Beaulaurier[1], Gintaras Deikus[1], Tao P. Wu[2], Maya Strahl[1], Ziyang Hao[3], Guanzheng Luo[3], James A. Gregory[4], Andrew Chess[1], Chuan He[3], Andrew Xiao[2], Robert Sebra[1], Eric E. Schadt[1] and Gang Fang[1#]

[1] *Department of Genetics and Genomic Sciences and Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA*

[2] *Department of Genetics and Yale Stem Cell Center, Yale School of Medicine, New Haven, CT 06520, USA*

[3] *Department of Chemistry and Institute for Biophysical Dynamics, Howard Hughes Medical Institute, The University of Chicago, Chicago, IL 60637, USA*

[4] *Center for Genomics of Neurodegenerative Disease, New York Genome Center, New York, NY 10013, USA*

[#]*Address correspondence to: gang.fang@mssm.edu*

**N6-methyladenine ($m^6dA$) has been discovered as a novel form of DNA methylation prevalent in eukaryotes, however, methods for high resolution mapping of $m^6dA$ events are still lacking. Single-molecule real-time (SMRT) sequencing has enabled the detection of $m^6dA$ events at single-nucleotide resolution in prokaryotic genomes, but its application to detecting $m^6dA$ in eukaryotic genomes has not been rigorously examined. Herein, we identified unique characteristics of eukaryotic $m^6dA$ methylomes that fundamentally differ from those of prokaryotes. Based on these differences, we describe the first approach for mapping $m^6dA$ events using SMRT sequencing specifically designed for the study of eukaryotic genomes, and provide appropriate strategies for designing experiments and carrying out sequencing in future studies. We apply the novel approach to study two eukaryotic genomes. For green algae, we construct the first complete genome-wide map of $m^6dA$ at single nucleotide and single molecule resolution. For human lymphoblastoid cells (hLCLs), joint analyses of SMRT sequencing and independent sequencing data suggest that putative $m^6dA$ events are enriched in the promoters of young, full length LINE-1 elements (L1s). These analyses demonstrate a general method for rigorous mapping and characterization of $m^6dA$ events in eukaryotic genomes.**

## Introduction

N6-methyladenine ($m^6dA$) is the most prevalent form of DNA methylation in prokaryotes, most commonly associated with restriction-modification (RM) systems that defend hosts against invading foreign genomes(Casadesús and Low 2006; Wion and Casadesús 2006). In addition, increasing evidence suggests $m^6dA$ also plays important roles in the regulation of bacterial gene expression(Fang et al. 2012), cell cycle(Kozdon et al. 2013), virulence(Heithoff et al. 1999) and antibiotic susceptibility(Jen et al. 2014). The prevalence of $m^6dA$ in eukaryotes was unclear until recent studies demonstrated their existence in algae(Fu et al. 2015), fungi (Mondo et al. 2017), worm(Greer et al. 2015), insect (Zhang et al. 2015), and recently in vertebrates(Koziol et al. 2016) including mammals (Wu et al. 2016). These recent studies have revealed diverse functions impacted by $m^6dA$ events in eukaryotes including the regulation of gene expression(Fu et al. 2015; Wu et al. 2016; Zhou et al. 2016; Mondo et al. 2017), transposons(Zhang et al. 2015; Wu et al. 2016), and cross talk with histone modifications(Fu et al. 2015; Wu et al. 2016). The existence of $m^6dA$ modifications across a diverse set of eukaryotic genomes opens up an exciting paradigm(Luo et al. 2015) in epigenetics and epigenomics regarding the regulation of biological processes in eukaryotic systems, in addition to the widely studied cytosine methylation.

Several methods have been developed to map $m^6dA$ in eukaryotic genomes. DNA immunoprecipitation (DIP) with anti-N6-methyladenine antibodies followed by next-generation sequencing ($m^6dA$-DIP-seq) has identified genomic regions enriched for

$m^6dA$ events in several species (Fu et al. 2015; Greer et al. 2015; Zhang et al. 2015). The combination of $m^6dA$-DIP-seq and exonuclease digestion ($m^6dA$-CLIP-exo-seq) provides increased resolution (Fu et al. 2015). However, due to the nature of antibody-based methods, both $m^6dA$-DIP-seq and $m^6dA$-CLIP-exo-seq, lack the ability to identify $m^6dA$ events at single nucleotide resolution and might be confounded by certain biases (Lentini et al. 2017). Furthermore, the immunoprecipitation process loses information necessary to study cell-to-cell epigenetic heterogeneity (*i.e. partial methylation*) in the cell population of interest (Fu et al. 2015; Greer et al. 2015; Zhang et al. 2015). Thus, these antibody-based approaches are limited with respect to their ability to elucidate the characteristics of $m^6dA$ events at high resolution. A complementary approach was developed using $m^6dA$-sensitive or $m^6dA$-dependent restriction enzymes (REs)(Fu et al. 2015; Luo et al. 2016), both of which provide single-nucleotide resolution and enable estimates of partial methylation at each nucleotide position. However, these restriction enzyme-based methods have a fundamental limitation in that they can only examine a limited set of motif sites (specific to the restriction enzymes used) and therefore provide a largely incomplete view of $m^6dA$ methylome in any given organism.

Single molecule real-time (SMRT) sequencing (Eid et al. 2009) by Pacific Biosciences Inc. enabled the genome wide mapping of $m^6dA$ in prokaryotes at single-nucleotide resolution (Fang et al. 2012) and at single-molecule level (Beaulaurier et al. 2015). SMRT sequencing monitors not only the pulse fluorescence associated with each incorporated nucleotide, but also the time between the incorporation events, termed the inter-pulse duration (IPD). Deviation of an IPD distribution from the expected level, as

reflected by the IPD ratio, is highly correlated with the presence of modifications of the nucleotide corresponding to the IPD deviation or its neighboring nucleotides (Flusberg et al. 2010; Schadt et al. 2013). Given this feature of SMRT sequencing, $m^6dA$ methylomes have been mapped for hundreds of bacterial and archaeal genomes, revealing many novel insights into $m^6dA$ biology in prokaryotes(Sánchez-Romero et al. 2015; Blow et al. 2016). SMRT sequencing has also been used to detect $m^6dA$ in some eukaryotic species (Greer et al. 2015; Wu et al. 2016; Mondo et al. 2017). Although promising, the effective use of SMRT sequencing for studying $m^6dA$ in eukaryotic genomes has not been rigorously examined.

In fact, there are fundamental differences (**Fig. 1a**) between the prokaryotic and eukaryotic $m^6dA$ methylomes that raise a caution in the use of SMRT sequencing, and more generally third generation sequencing (Manrao et al. 2012; Laszlo et al. 2013; Schreiber et al. 2013), for the detection of $m^6dA$ events in eukaryotic genomes. First, $m^6dA$ abundance ($m^6dA/A$) is orders of magnitudes lower in eukaryotes than in prokaryotes (Casadesús and Low 2006; Fang et al. 2012; Fu et al. 2015; Greer et al. 2015; Luo et al. 2015; Zhang et al. 2015) (**Fig. 1a**). Given a certain false positive rate (FPR) associated with IPD-based detection of DNA modifications, $m^6dA$ calls from eukaryotes with low $m^6dA$ abundance are expected to have high false discovery rates (FDR). If the FDR becomes too high, then true $m^6dA$ events would be overwhelmed by the large number of false positive ones. Second, $m^6dA$ events in prokaryotes are highly sequence specific due to their involvement in and the nature of RM systems (Casadesús and Low 2006; Fang et al. 2012). Typically, an active methyltransferase

methylates nearly all occurrences (often >95%) of its target sequence motif in a prokaryotic genome (Fang et al. 2012; Blow et al. 2016) (**Fig. 1a**). In contrast, $m^6dA$ events are much less motif-driven in eukaryotes (Fu et al. 2015; Luo et al. 2016; Wu et al. 2016), likely due to their involvement in functional regulation rather than RM systems (**Fig. 1a**). For example, $m^6dA$ motifs have been identified in *Chlamydomonas reinhardtii* (Fu et al. 2015), *Plasmodium falciparum*(Luo et al. 2016) and mouse embryonic stem cells (mESCs) (Wu et al. 2016), where very few occurrence (often <3%) of the motif across the genome sites are methylated, *i.e. weakly motif-driven*. Complicating matters further, other types of DNA modifications (DNA damages, $m^5C$ and its derivatives in the process of demethylation, etc.) occurring at neighboring bases can disturb the IPD ratios at an adenine site in question (Flusberg et al. 2010; Schadt et al. 2013), leading to false positive $m^6dA$ calls. As a result, the weakly motif-driven nature of $m^6dA$ events in eukaryotic genomes poses a critical challenge in differentiating $m^6dA$ events from other types of DNA modifications. Finally, cell-to-cell epigenetic heterogeneity of $m^6dA$ has been increasingly recognized in prokaryotes (Casadesús and Low 2013; Manso et al. 2014; Beaulaurier et al. 2015) and $m^6dA$ in eukaryotes is expected to be similarly heterogeneous, if not more so, considering the large number of cell types and subpopulations of a given cell type (Huang et al. 2000; Heintzman et al. 2009; Miller et al. 2012; Shulha et al. 2013) (**Fig. 1a**). Thus, the ability to study eukaryotic $m^6dA$ methylation at single-molecule resolution and to characterize cell-to-cell heterogeneity is desired to achieve a better understanding of $m^6dA$ biology in eukaryotes.

Motivated by the above challenges, here we propose the first approach (**Fig. 1b**) for mapping of $m^6dA$ events using SMRT sequencing specifically designed for the study of eukaryotic genomes. Using well-characterized $m^6dA$ methylomes, we systematically investigate the factors affecting the sensitivity and specificity of $m^6dA$ detection at the levels of single nucleotides, single molecules, and individual motifs. This comprehensive evaluation provides a strategic framework that can help the study design, $m^6dA$ detection and interpretation in future studies of eukaryotic $m^6dA$ methylomes using SMRT sequencing, as well as its critical integration with independent and complementary sequencing methods. We applied the approach to examine $m^6dA$ in green algae and human genomes. These applications demonstrate a general method and guideline for mapping and rigorous characterization of $m^6dA$ events in eukaryotic genomes.

## Results

### A novel approach and comprehensive evaluations for detecting and characterizing $m^6dA$ in eukaryotic genomes

We developed a novel approach with three core components to address challenges posed by the three fundamental differences between the prokaryotic and eukaryotic $m^6dA$ methylomes. We will first present each of the core components and associated evaluations, followed by the application of the novel approach to green algae and human genomes (**Fig. 1b**).

Design of SMRT sequencing and rigorous detection of m$^6$dA events. The genome-wide mapping of 5-methylcytosine (m$^5$C) via bisulfite sequencing builds on the accurate base calling of Illumina sequencing (Schuster 2008). In contrast, SMRT sequencing-based detection of DNA modifications is facilitated through statistical tests (Fang et al. 2012; Schadt et al. 2013; Beaulaurier et al. 2015) comparing the observed distribution of IPD values at each nucleotide locus with the expected IPD value of the same base in the same sequence context, but without methylation. The latter IPD value is estimated from whole genome amplified (WGA; methylation free) samples(Flusberg et al. 2010). For a given genome, millions or billions of nucleotides are tested, making false positive calls due to multiple hypotheses testing a serious concern (**Supplemental Text; Supplemental Fig S1**). To account for the multiple hypothesis testing, we use a false discovery rate (FDR) (Reiner et al. 2003; Fang et al. 2012) calculated by comparing the global distribution of IPD ratios (or *p* values from student's *t*-test(2012)) in native versus WGA samples (**Methods**). For a given genome, the FDR of m$^6$dA detection conceptually reflects the fraction of false positives among total m$^6$dA calls and depends on two major factors: the fraction of methylated adenines across the genome, $f(m^6dA/A)$; and per-strand sequencing depth, $coverage$ (*i.e.* average number of IPD values for each strand of the genome reference). $f(m^6dA/A)$ can be estimated from high performance liquid chromatography (HPLC) coupled with m$^6$dA mass spectrometry (MS)(Fu et al. 2015; Greer et al. 2015; Zhang et al. 2015) (**Supplemental Text**), while $coverage$ depends on SMRT sequencing read depth. Using bacteria with well characterized m$^6$dA methylomes (**Supplemental Table S1**), we systematically evaluated the variation in FDR over different levels of $f(m^6dA/A)$ and $coverage$

(**Methods**). As expected, for each level of detection sensitivity, lower FDRs can be achieved with higher levels of $f(m^6dA/A)$ and $coverage$ (**Fig. 2a-b**). We further estimated the expected levels of FDRs for genomes with different levels of m$^6$dA $f(m^6dA/A)$ at different levels of $coverage$ (**Fig. 2c; Methods**). Notably, while moderate $coverage$ (e.g. ~20x) is sufficient to achieve a fairly low FDR (e.g. ~0.03) for species with higher $f(m^6dA/A)$ levels (e.g. ~1%), deep $coverage$ (e.g. ~150x) is necessary for species with low $f(m^6dA/A)$ levels (e.g. ~0.001%) in order to achieve even modest FDRs (e.g. ~0.2). This systematic evaluation provides a rational strategy that can help determine the depth of SMRT sequencing in future studies of eukaryotic genomes based on the $f(m^6dA/A)$ values estimated from HPLC/MS data. It is worth noting that the coverage required to achieve a certain level of FDR estimated in **Fig. 2c** is specifically for calling fully (~100%) methylated m$^6$dA events at single nucleotide resolution. For other types of epigenomic analyses such as motif discovery and consensus analysis across multiple genomic sites (e.g. transcription start sites (Fu et al. 2015) etc.), the requirement on sequencing depth can be lower depending on specific m$^6$dA patterns in different organisms.

Unbiased discovery of m$^6$dA motifs. In contrast to bacteria, m$^6$dA motifs in eukaryotic species are typically only weakly motif-driven, *i.e.* the motif-specific fraction of methylated motif sites across the genome, $f_m(m^6dA/A)$, is often very low (<3%; **Fig. 1a**). When a motif is only methylated at a low fraction across the genome, it becomes difficult to differentiate between the enrichment of the motif due to m$^6$dA events and

methylation-independent enrichment of the motif reflecting the intrinsic sequence composition of a eukaryotic genome or certain regions of interests (Bailey 2011). To address this challenge in SMRT sequencing-based m$^6$dA motif enrichment analysis, we develop a *motif enrichment score* that is calculated as the odds ratio between the frequency of a motif among putative m$^6$dA sites (IPD ratio > $r$) and the frequency of the motif among all adenine sites in the genome (**Supplemental Fig S2**). The reciprocal of this enrichment score approximates the FDR of the motif sites with m$^6$dA events (**Methods**). To illustrate the use of the motif enrichment score, we first examine a m$^6$dA motif (CAAAAA; $f_m(m^6dA/A) > 95\%$) in a strain of the bacterium *Clostridium difficile*, where the motif has an enrichment score of 1.08 x $10^5$ ($r = 4$) (**Fig. 2d**), meaning that CAAAAA is 108,000-fold enriched among A's with IPD ratio > 4 compared to all the A's sites in the genome (**Supplemental Text**). Next, we collected 11 bacterial species/strains that contain a total of 55 confident m$^6$dA motifs (**Fig. 2e, Supplemental Table S1**) to systematically evaluate the use of motif enrichment score for detecting m$^6$dA motifs with low $f_m(m^6dA/A)$ levels as expected in eukaryotic genomes. With the 55 m$^6$dA motifs as background truth, we calibrated motif enrichment scores over different $f_m(m^6dA/A)$ levels of abundance (**Fig. 2f; Methods**).

Single-molecule analysis to estimate partial m$^6$dA methylation. In the above methods and analyses, m$^6$dA calling relies on IPDs pooled from separate molecules for each genomic locus. This *aggregated* analysis works well when each m$^6$dA locus has nearly 100% methylation across all molecules. However, epigenetic heterogeneity is often

observed in both bacteria(Casadesús and Low 2013; Manso et al. 2014; Beaulaurier et al. 2015) and eukaryotic species (Heng et al. 2009; Smallwood et al. 2014), where only a fraction of cells are methylated at each genomic locus, i.e. *partial methylation* (**Fig. 1a**). Partial methylation is characterized by a locus-specific fraction of methylation $f_l(m^6dA/A)$. Using an *E. coli* strain with a well-characterized methylome (Fang et al. 2012) (**Methods**), we found that partial methylation significantly reduces the reliability of $m^6dA$ event calling by aggregate analysis (**Supplemental Fig S3**). To better estimate partial methylation and study cell-to-cell $m^6dA$ heterogeneity in eukaryotic genomes, we developed a method for single molecule-resolution analysis of SMRT sequencing data. In brief, IPDs are grouped by molecules for each genomic site and compared to the expected IPD values (**Fig. 2g**; **Methods**). The IPD ratios of methylated ($m^6dA$) and non-methylated sites (at single-molecule level) follow two normal distributions with means of 1 and ~4 and variances that decrease as per-molecule, per-strand sequencing coverage increases (**Fig. 2g; Methods, Supplemental Text**). Using 4,359 $m^6dA$ sites with different levels of methylation fraction, ranging from 22% to 97%, subsampled from a well-characterized *E. coli* $m^6dA$ methylome (**Methods**), we found the single-molecule level analysis provides a more accurate estimation of partial methylation than existing aggregate methods without single-molecule level analysis (**Fig. 2h**). In addition, while aggregate analysis can hardly detect the GATC motif in *E. coli* methylome with simulated partial methylation (**Fig. 2i**; motif enrichment score = 1.3; FDR >0.75), single-molecule analysis clearly recognizes the GATC motif with a motif enrichment score of 25 (**Fig. 2j**; FDR<0.05).

## A comprehensive characterization of $m^6dA$ in *C. reinhardtii*

The first genome-wide detection of $m^6dA$ in green algae *C. reinhardtii* was achieved recently, revealing that $m^6dA$ has a periodic pattern of deposition around transcriptional start sites (TSSs) that is inversely correlated with nucleosome positioning(Fu et al. 2015). In this previous study, Fu *et al.* developed three complementary sequencing-based methods and found that certain motifs were enriched for $m^6dA$ events, of which GATC and CATG were confirmed by $m^6dA$-RE-seq. In a more recent study, Luo *et al*. developed a more sensitive version of $m^6dA$-RE-seq by using a methylation-dependent restriction enzyme, leading to the discovery of two additional $m^6dA$ motifs (CATC and GATG)(Luo et al. 2016). While these two studies fundamentally enhanced our understanding of the $m^6dA$ methylome of *C. reinhardtii*, the single-nucleotide resolution $m^6dA$ map that they provide remains incomplete, and there are possibly additional $m^6dA$ motifs yet to be discovered.

A complete map of $m^6dA$ and cross validation with five independent methods. In order to construct a complete $m^6dA$ methylome of *C. reinhardtii* at single-nucleotide resolution, we performed high coverage SMRT sequencing of both native and WGA samples of the same *C. reinhardtii* strain used in these recent studies (Fu et al. 2015; Luo et al. 2016) (**Supplemental Table S1**). We used an IPD ratio threshold of > 4.5 (FDR < 0.05; **Fig. 3a; Methods; Supplemental Fig S4a; Supplemental Text**) to calculate motif enrichment scores. Among the sixteen 4-mer motifs centered at AT, nine (VATB, V = A, C or G and B = C, G or T) are significantly enriched for the $m^6dA$ events in native DNA

but not in WGA DNA (**Fig. 3b & Supplemental Fig S4b**). In **Fig. 3b**, each $4 \times 4$ heatmap corresponds to all sixteen 4-mer motifs, for which 2nd and 3rd bases are fixed at the center/title (e.g. AA). The rows and columns in the heatmaps represent the first and last bases of 4-mer motifs. Each cell in the following $4 \times 4$ heatmaps shows the motif enrichment score based on native DNA sample. Take the $4 \times 4$ heatmap with AT on top for example, the upper left corner corresponds to the motif CATG. A red color indicates that CATG has a very high motif enrichment score of ~200. For motifs centered at AA, CA and GA, high motif enrichment scores are also observed when the last base is T (**Fig. 3b**); this is essentially a trivial consequence of the VATB motifs. It is also worth noting that a small number of additional 4-mer motifs have moderate methylation scores (**Fig. 3b**, yellow entries) in the native data, to some extent similar to that seen in the WGA data (**Supplemental Fig S4b**). This observation suggests that some background noise (**Supplemental Text**), which may contribute to spurious motif enrichment, should be removed using WGA data as a negative control. Therefore, to further filter the background, we calculated the ratio of the motif scores between the native DNA and the WGA control, demonstrating an even cleaner motif enrichment (**Supplemental Fig S4c**). At single-nucleotide level, the 117,735 methylated VATB sites (FDR < 0.05) represent 98.3% of total genomic $m^6dA$ calls (**Supplemental Table S2**), and ~0.3% of total A sites in the genome. A cross-check among five independent $m^6dA$ detection methods shows that single-nucleotide $m^6dA$ events called by SMRT sequencing are highly consistent with detections made by $m^6dA$-RE-seq and $m^6dA$-DIP-seq (**Fig. 3c, Supplemental Fig S4d**). It is worth nothing that a $m^6dA$ event called from SMRT-seq data will be missed by $m^6dA$-RE-seq if the event resides in a motif context

not recognized by the RE. m$^6$dA-DIP-seq can miss a m$^6$dA event due to certain bias or lack of sensitivity commonly associated with antibody based approaches (**Supplemental Fig S4d).** Thus, in addition to the four motifs confirmed by RE-based methods (Fu et al. 2015; Luo et al. 2016), SMRT sequencing-based motif analysis revealed five additional m$^6$dA 4-mer motifs and provides a complete motif characterization of the *C. reinhardtii* m$^6$dA methylome.

High-resolution characterization of m$^6$dA deposition. We next performed a comprehensive characterization of the *C. reinhardtii* m$^6$dA methylome. We first checked whether the five additional motifs discovered from SMRT sequencing follow a periodic deposition pattern similar to the four previously known motifs(Fu et al. 2015; Luo et al. 2016). The methylated sites of all the nine 4-mers (VATB), but not the other seven 4-mers (non-VATB), are enriched at TSSs with a similar periodic pattern (**Fig. 3d**) that inversely correlates with nucleosome positioning (**Fig. 3e**). Next, the completeness and single-nucleotide resolution of this m$^6$dA map allows us to examine the frequency of m$^6$dA events in linker DNAs: an average of one m$^6$dA locus occurs in the linker DNAs between the adjacent nucleosomes near TSSs, with some linkers having ~10 m$^6$dA events and some having none (**Fig. 3f; Supplemental Fig S4e**). The depletion of m$^6$dA events in the close proximity of TSSs (**Fig. 3e**) motivated us to check the frequency of VATB motif sites in this region. We found VATB and non-VATB sites have a similar periodic frequency that reaches its peak density near TSSs (**Fig. 3e**), yet the VATB sites close to TSSs are non-methylated. Beyond TSSs, regions with high nucleosome occupancy also have high density of VATB sites, yet low levels of m$^6$dA (**Fig. 3e**).

These discrepancies between VATB density and m$^6$dA methylation density suggest the existence of additional factors in the deposition of m$^6$dA events in *C. reinhardtii* beyond the proximity to TSS and the clearly defined m$^6$dA motif. A further integrative analysis of SMRT sequencing data and the RNA-seq gene expression data from *Fu et al.*(Fu et al. 2015) shows that m$^6$dA events at VATB sites are associated with active gene expression (**Fig. 3g and h; Supplemental Fig S4f**), while there is no such correlation between gene expression and the frequency of VATB motif sites (**Supplemental Fig S4g**).

Single-molecule strand-specific characterization. A unique advantage of SMRT sequencing is the ability to examine methylation states of the two reverse complementary VATB sites at the two strands of each molecule. This allows us to further characterize m$^6$dA events at VATB sites in terms of full-, non- or hemi-methylation states at single-molecule resolution with strand specificity. We examined m$^6$dA calls in GATC (**Fig. 3i**) and CATG sites (**Fig. 3j**) detected by m$^6$dA-RE-seq (Fu et al. 2015) and the methylated VATB sites detected by SMRT sequencing (FDR <0.05; **Fig. 3k; Supplemental Fig S4h**). Consistently, most examined molecules were fully methylated on both strands (**Fig. 3i-k**; top right corners). We also found that some VATB sites were hemi-methylated (**Fig. 3i-k**; top left and bottom right corners), which could be right after DNA replication forks and haven't been fully methylated yet. Some VATB sites were non-methylated on both strands of single molecules (**Fig. 3i-k**; bottom left corners), despite these loci having high consensus m$^6$dA methylation levels. Collectively, the above comprehensive characterizations reveal the first complete m$^6$dA

map of *C. reinhardtii* and motivate future research towards mechanistic understanding of $m^6dA$ deposition.

**Integrative analysis of SMRT sequencing data of human lymphoblastoid cells**

After interrogating the $m^6dA$ distribution in a unicellular eukaryote, we next apply the new method to investigate a more complex genome. The recent discovery of $m^6dA$ in mammalian genomes and the enrichment of $m^6dA$ in young, full-length L1s in mESCs opened new research opportunities (Wu et al. 2016). To date, the deposition patterns of $m^6dA$ in human genomes remain unclear. Human lymphoblastoid cells (hLCLs) are transformed from B cells by Epstein-Barr viruses for immortalization and have been widely used in large-scale studies of human genetics and genomics (Reedman and Klein 1973; Young and Rickinson 2004). Recently, whole genome-wide SMRT sequencing data of hLCLs have been generated to improve human genome assembly (Zook et al. 2016), which also provide a good opportunity to detect putative $m^6dA$ events.

We first used dot blotting to compare hLCLs with negative oligos and mESCs(Wu et al. 2016). The results suggest the existence of $m^6dA$ in hLCLs at a $m^6dA/A$ level lower than what was observed in mESCs (**Supplemental Fig. S5**). It is worth noting that, because EBV genome co-exists with human genome in hLCLs, the $m^6dA$ dot blots reflects $m^6dA$ in both EBV genome and human genome. In such cases, sequencing based study is

necessary for a specific genome of interest. We collected the genome-wide SMRT sequencing data (specifically, the subset with P6-C4 chemistry) publicly available for three hLCL samples (HG002, HG003 and HG004) (Zook et al. 2016). Considering the low level of $f(m^6dA/A)$ and the sequencing coverage (~18x per reference strand for aggregate analysis), a genome-wide analysis of m$^6$dA events with current data would probably be associated with a high FDR (**Fig. 2c**). Therefore, in the current study, we focused on full length L1s with different ages (Castro-Diaz et al. 2014) (**Methods**) to test whether putative m$^6$dA is enriched on young full-length L1s in human genome as in mESCs (Wu et al. 2016). A consensus analysis of IPD ratios on adenine (A) sites in the +/-6,000bp beyond the 5' UTRs of L1s across all the 7,108 full length L1s showed that, consistent among the three hLCLs, there is an enrichment of high IPD ratios at young (age < 10 million years) full-length L1s (**Fig. 4a; Supplemental Fig S6; Methods**), but much less enrichment in old L1s (**Fig. 4a&b; Supplemental Fig S6**). In addition, this consensus analysis shows that the mean IPD ratio on A sites is relatively higher in the promoter and proximal region of young L1s than in the flanking regions (**Fig. 4a)**. In mammalian genome, the majority of CG dinucleotides are methylated (m$^5$C), which can confound the m$^6$dA analysis based on SMRT sequencing data because m$^5$C events can affect IPDs at multiple flanking nucleotides (Schadt et al. 2013). To scrutinize this consensus pattern, we next examine multiple factors that may confound SMRT sequencing-based detection of putative m$^6$dA events (**Supplemental Text; Methods**): effects of m$^5$C events (Hata and Sakaki 1997) on neighboring IPDs (Flusberg et al. 2010; Schadt et al. 2013) (**Supplemental Fig S7**), outlier IPDs (Fang et al. 2012; Beaulaurier et al. 2015), SNP effects (both homozygous and heterozygous genotypes),

and the use of *in silico* IPD estimation. We found that the consensus IPD ratio pattern in young full-length L1 remains after rigorous filtering of these possible confounding factors (**Supplemental Fig S8 & S9; Methods**). However, we found that when the same analysis was applied to the other three types of nucleotides (C, G and T), similar consensus patterns are observed even after the effect of all these confounding factors are filtered out (**Supplemental Fig S10**). This unexpected observation suggests the possible co-enrichment of other DNA modifications (beyond $m^5C$) in the young L1s together with $m^6dA$, or the possible existence of DNA secondary structure in addition to DNA modifications, which are also expected to affect DNA polymerase kinetics in SMRT sequencing. Without orthogonal validation methods, SMRT sequencing data alone is unable to differentiate among these possibilities.

We therefore used $m^6dA$-DIP-seq as an independent method to examine the human LCLs derived from the same cell lines (**Supplemental Table S3**). A consensus analysis of $m^6dA/A$ density across all the 7,108 full-length L1s shows that $m^6dA$ events are enriched in young, but not old, full-length L1s of human LCLs (**Fig. 4c; Supplemental Fig. S11; Methods**). In addition, we performed a further analysis to examine the possibility that the consensus $m^6dA$ pattern across young L1s by $m^6dA$-DIP-seq could be the result of certain biases (Lentini et al. 2017). Specifically, the exact same $m^6dA$-DIP-seq protocol was also performed for hLCL WGA DNA, where essentially no $m^6dA$ events are expected, and used as alternative control to input DNA (**Supplemental Table S3**). We observed consistent pattern when two controls are used to compare with $m^6dA$-DIP-seq of native hLCL DNA (**Fig. 4d; Methods**). These analyses of $m^6dA$-DIP-

seq data suggest that $m^6$dA events are enriched in the young full-length L1 of hLCLs and that the $m^6$dA/A level is relatively higher in the promoter and proximal region of young L1s than the downstream region, similar to the observations made from IPD analysis of SMRT sequencing data (**Fig. 4a&b**). A 2-mer motif analysis of the hLCL SMRT sequencing data (across young L1s) showed that AG is the most enriched for putative $m^6$dA events among all eight 2-mer motifs (**Fig. 4e**), although it is worth noting the WGA sample also showed modest, weaker enrichment for AG (**Supplemental Fig S12a; Methods; Supplemental Text**). In a further analysis of 4-mer motifs, AAGG and CAAG showed the highest motif enrichment scores that are specific to native DNA (**Fig. 4f)** but not WGA DNA (**Supplemental Fig S12b)**. We also estimated single-nucleotide sequence conservation in L1s through multiple alignment of young full-length L1s (L1HS and L1PA2 (Castro-Diaz et al. 2014); **Methods**) and found that the loci with highest frequency of putative $m^6$dA sites (adjusted by the frequency of A's) across young L1s generally occur at the loci that are highly conserved across full length L1s (Darling et al. 2004) (**Fig. 4g**) and the loci with highest relative frequency of AG (AG/A) (**Fig. 4g**). These observations suggest the deposition and function of $m^6$dA may be related to sequence conservations in the promoter and proximal regions of young L1s (Goodier and Kazazian 2008); this, however, need to be validated by future independent methods that have better resolution than $m^6$dA-DIP-seq and less constrained sequence specificity than $m^6$dA-RE-seq.

## Discussion

The recent discovery of $m^6dA$ in eukaryotic genomes opens up a new and promising dimension of epigenetic research, however, methods for high resolution, complete mapping of $m^6dA$ events are still lacking. Here we presented a novel set of methods and an analytical framework for $m^6dA$ characterization in eukaryotic genomes using SMRT sequencing. The key motivation of this study was the characteristics of eukaryotic $m^6dA$ methylomes that fundamentally differ from those of prokaryotes, yet all previous computational methods for SMRT sequencing based detection of $m^6dA$ events were designed specifically for the study of prokaryotic methylomes. In addition, we highlighted the importance of tailoring sequencing design and analytical strategy for an organism considering the $m^6dA/A$ abundance in its genome as determined by mass spectrometry and dot blots. For organisms with high $m^6dA/A$ abundance, confident (low FDR) $m^6dA$ events can be called at both single nucleotide and single molecule resolution, allowing a variety of in-depth characterization, as demonstrated in our analysis of the *C. reinhardtii* $m^6dA$ methylome. For organisms with low $m^6dA/A$ abundance, however, $m^6dA$ events called by SMRT sequencing data are essentially putative events and must be treated with caution, because they are expected to have high FDR as estimated in **Fig. 2c**. In applications that belong to the latter case, consensus analyses, which are more resistant to false positive calls, should be adopted when applicable, as illustrated in the study of young L1s in hLCLs.

Importantly, instead of specifically detecting $m^6dA$ events, SMRT sequencing can detect any form of DNA modifications that significantly affect DNA polymerase kinetics as measured by IPD. Different types of DNA modifications at a site of interest or its neighboring sites can lead to similar IPD ratios at the site (Flusberg et al. 2010; Schadt et al. 2013). In a bacterial genome, the forms of DNA methylation are relatively limited ($m^6dA$, $m^5C$, $m^4C$) and highly motif-driven, which fundamentally ease the detection and differentiation of $m^6dA$ events from other DNA modifications. In contrast, $m^6dA$ events in eukaryotic genomes are much less abundant, weakly motif-driven, and possibly co-existing with other forms of DNA modifications. These differences between bacterial and eukaryotic methylomes call for critical attention in the interpretation of putative $m^6dA$ calls based on SMRT sequencing to avoid mis-interpretation of false positive events. The methods and the overall framework we presented in this study highlight the importance of rational design of experiments and SMRT sequencing, as well as rigorous analysis and interpretation of SMRT sequencing data in combination with independent and complementary techniques.

Finally, it is worth noting that the strengths and challenges associated with SMRT sequencing discussed above also largely apply to other third generation real time sequencing techniques that also hold promise for the detection of DNA methylation, e.g. Oxford Nanopore (Manrao et al. 2012; Laszlo et al. 2013; Schreiber et al. 2013). Essentially, similar to SMRT sequencing, these other third-generation sequencing methods indirectly detect DNA modifications based on features captured during real-

time single molecule sequencing. Therefore, similar cautions are likely needed in the use of these third-generation sequencing technologies in the mapping and characterization of different forms of DNA modifications in eukaryotic genomes.

## Methods

**Pre-processing of SMRT sequencing data for IPD analysis.** We followed the preprocessing steps as implemented in SMRTportal (URL). In brief, an initial filtering step removes all subreads with ambiguous alignments (MapQV<240), low accuracy (<75%) or short-aligned length (<50 bases). Next, an additional filtering step removes the subread IPD values from the mismatched positions with respect to the reference sequence. Subread IPD normalization corrects for any potential slowing of polymerase kinetics over the course of an entire read (which consists of many subreads) and is done by dividing subread IPD values by their mean.

**Estimation of false discovery rate (FDR) for single nucleotide-level m$^6$dA calls**. The FDR corresponding to a specific threshold on a given statistical measure (e.g. IPD ratio, *t*-test *p* value or identificationQv) is estimated by comparing global distribution of the measure obtained from the native DNA sample with that from a whole-genome amplified (WGA, methylation free) sample. Specifically, the FDR is calculated as follows:

$$FDR = \frac{f_{WGA\_As}(m > thres)}{f_{native\_As}(m > thres)}$$

where *m* denotes a given statistical measure; $f_{WGA\_As}(m > thres)$ denotes the fraction of As with $m > thres$ out of all As in WGA, and $f_{native\_As}(m > thres)$ denotes the fraction of As with $m > thres$ out of all As in native DNA sample. There are cases where a WGA sample is not available or some true m$^6$dA motifs are known a priori or discovered based on motif enrichment analysis. In such cases, for each motif, FDRs can be estimated for single nucleotide-level m$^6$dA calls only among the A sites corresponding to the motif across the genome. Specifically, we describe motif-specific FDR estimation for a specific motif using data from native DNA alone:

$$FDR_{motif} = \frac{f_{native\_As}(m > thres)}{f_{native\_motif}(m > thres)}$$

where $f_{native\_motif}(m > thres)$ denotes the frequency of motif sites with $m > thres$ among all sites of that putative m$^6$dA motif in native DNA.

**Expected FDRs for single nucleotide-level m$^6$dA calls over different levels of $f(\boldsymbol{m^6dA/A})$ and $\boldsymbol{coverage}$**. We used an *E. coli* C227 methylome that has been well characterized in the previous study(Fang et al. 2012) (Supplemental Table S1). By subsampling both m$^6$dA motif sites and non-m$^6$dA-motif sites, we generated test datasets with different levels of $f(\boldsymbol{m^6dA/A})$ and $coverage$. For each dataset, we estimated the FDR corresponding to the same cutoff on IPD ratio (>4).

**Methylation enrichment score of a putative m$^6$dA motif**. For a real m$^6$dA methylation motif, it is expected that the fraction of A's with high IPD ratios in that motif, i.e. $f_{native\_motif}(m > thres)$, should be higher than the background in the same native DNA sample, i.e. $f_{native\_As}(m > thres)$. So, we define the *motif enrichment score* as the odds between the two fractions:

$$S_{motif}\{native\} = \frac{f_{native\_motif}(m > thres)}{f_{native\_As}(m > thres)}$$

The denominator can also be defined as the A sites in native DNA excluding a certain motif. Because m$^6$dA/A level is mostly <5% in both bacteria and eukaryotes, the two alternative definitions are practically the same, and the currently defined one is easier to calculate. It is worth noting that, the *motif enrichment score* is mathematically the reciprocal of motif specific FDR:

$$S_{motif}\{native\} = \frac{1}{FDR_{motif}}$$

Certain intrinsic biases in SMRT sequencing (e.g. possible biases associated with the *in silico* control model as described above) can contribute to the small but statistically significant enrichment of certain motifs independent of DNA modifications. These biases can be estimated by calculating methylation enrichment scores for a specific motif using a WGA sample without DNA methylation, i.e. $S_{motif}\{WGA\}$. A motif is enriched for m$^6$dA events if it has a high enrichment score that is specific to the native data but not the WGA data.

**Methylation enrichment scores for motifs with different fraction of m$^6$dA methylation**. We collected 11 bacterial m$^6$dA methylomes (55 m$^6$dA motifs) that have been well characterized in previous studies (Fang et al. 2012; Beaulaurier et al. 2015; Pak et al. 2015). All the m$^6$dA motif sites are pooled together as true m$^6$dA events. By subsampling from these m$^6$dA sites, we generated test motifs with different fraction of methylation: 100%, 50%, 10%, 1% and 0.1%. For each of these fractions, we estimated the methylation motif score, $S_{motif}\{native\}$, corresponding to different thresholds of IPD ratios and *t*-test p values.

**Single molecule, single nucleotide level calculation of IPD ratios**. Considering each molecule separately, the IPD values (post-filtering) are grouped by their strand and mapped genomic position, and the mean value is calculated. At each genomic position of a single strand, the mean IPD values for each molecule follow the Gaussian distribution based on Central Limit Theory.

**Methylation fraction calling for each site at single molecule level**. The Central Limit Theorem (CLT) states that, given a sufficiently large sample size, the average of all samples from the same population tends toward a normal distribution, even if the original variables themselves are not normally distributed. Meanwhile, the mean of a sample approximates the mean of the population. In the context of IPD based DNA modification detection, the IPD values follow an exponential distribution, however, the *mean* of the IPD values that come from the same molecule and at the same genomic location (referred to as *IPD at a single molecule level*) follows normal distributions: either a single normal distribution (fully-methylated or non-methylated sites) or a mixture of normal distributions (partially methylated sites). Accordingly, given a single site, we can use a Gaussian mixture model (GMM) to estimate the extent of partial methylation. The GMM comprises two normal distributions from methylated and non-methylated molecules; the mean of IPD for non-methylated molecules are estimated from either *in silico* control model or WGA; the mean of IPD for methylated molecules are learned from the data, and the estimated proportion of two normal distributions reflects the fraction of methylated and non-methylated molecules. Furthermore, the CLT states that the variance of the sample approximates the variance of the population divided by the sample size. Accordingly, as read coverage increases for each molecule, the variance of normal distributions of *IPDs at the single molecule level* decreases, providing a better power for the separation between the methylated and non-methylated molecules. Therefore, the CLT provides a theoretical foundation to use GMM to call methylation fraction at a single molecule level.

**Simulation of partial m$^6$dA methylation from well-characterized bacterial m$^6$dA methylomes**. We use the SMRT sequencing data for *E. coli* C227 strain (both native and WGA), generated in a recent study (Fang et al. 2012). In *E. coli*, most GATC sites are ~100% m$^6$dA methylation(Fang et al. 2012). To simulate partially methylated GATC sites, we randomly select single molecules from both native and WGA data, and mix them in different proportions to generate GATC sites with different levels of partial methylation. For each GATC site, the true fraction of m$^6$dA methylation is calculated based on the number of unique molecules from the native and WGA data.

***C. reinhardtii* DNA extraction:** The frozen cell pellet is grounded in liquid nitrogen using a plastic pestle and 1.5 mL LoBind Eppendorf micro centrifuge tubes. We used the NucleoSpin® Plant II (Macherey Nagel, Catalog #740770.50) kit and followed the standard protocol for lysis buffer PL1, using ~100mg of tissue/extraction column to extract the DNA. The concentration and quality of the resulting DNA are checked using the Qubit dsDNA High Sense kit and 12k DNA BioAnalyzer chip.

**Human lymphoblastoid cells (LCLs).** Genome-wide SMRT-seq data were from a recent study(Zook et al. 2016). The full human SMRT-seq data contains a mixture of two SMRT-seq chemistries: P5_C3 and P6_C4. Different chemistries are associated with different DNA polymerase kinetics that can significantly impact IPD values, which may lead to false positive calls. To achieve most rigorous data analysis, we chose to use P6_C4 SMRT runs only. gDNA is available from Coriell Biorepository: NA24143, NA24149 and NA24385.

**Genome references**: SMRT-seq data were mapped to the appropriate genomes using BLASR via SMRTportal (URL). Reads from Illumina sequencing data are mapped using BWA 0.7.8 (Li and Durbin 2009). The *C. reinhardtii* data were mapped to *Chlamydomonas* genome (JGI) Version 9.1. For human LCL data, we build a customized reference by extracting the [-10,000nt, +10,000nt] regions surrounding the 5' UTR of full-length L1s from the UCSC hg19, and mapped the human data to the faux-reference.

**Overlap analysis between m$^6$dA calls by different methods**. For m$^6$dA-RE-seq, its overlap with SMRT-seq is defined as the ratio between the CATG/GATC sites detected by both SMRT-seq and m$^6$dA-RE-seq, and the CATC/GATG sites detected by SMRT-seq. For m$^6$dA-DIP-seq and m$^6$dA CLIP-seq, their overlap with SMRT-seq is defined as the ratio between putative m$^6$dA sites that are detected by SMRT-seq and covered by at least one peak called from m$^6$dA-DIP-seq/m$^6$dA CLIP-seq, and all the m$^6$dA sites detected by SMRT-seq. The above overlap in a region of interest mainly depends on the ratio between true positive and false positive SMRT-seq m$^6$dA detections in that region.

**m$^6$dA dot blots**. We followed the same protocol as used in the recent study(Wu et al. 2016). Briefly, first, DNA samples were denatured at 95 degrees for 5 min, cooled down on ice, neutralized with 10% vol of 6.6 M ammonium acetate. Samples were spotted on the membrane (Amersham Hybond-N+, GE) and air dry for 5 min, then UV-crosslink (2× auto-crosslink, 1800 UV Stratalinker, STRATAGENE). Membranes were blocked in blocking buffer (5% milk, 1% BSA, PBST) for 2 h at room temperature, incubated with m$^6$dA antibodies (202-003, Synaptic Systems, 1:1000) overnight at 4 degrees. After 5 washes, membranes were incubated with HRP linked secondary anti-rabbit IgG antibody (1:5,000, Cell Signaling 7074S) for 30 min at room temperature. Signals were detected with ECL Plus Western Blotting Reagent Pack (GE Healthcare).

**Full length L1 elements and their evolutionary ages.** We collected the Human LINE-1 (L1) transposon annotations from the RepeatMasker (Tarailo-Graovac and Chen 2009). Those ~6kb long L1s were treated as full-length L1s (Babushok and Kazazian 2007). The evolutionary age for each L1 subfamily is based on Castro-Diaz et al.(Castro-Diaz et al. 2014).

**Consensus analysis of IPD ratios across different L1s**. The full-length L1s identified as described above were aligned based on their 5' UTR sites. At each aligned position, the IPD ratios of a specific base (A/G/C/T) across different L1s were aggregated, and normalized to the frequency of that corresponding base (A/G/C/T).

**Estimating false positive rate of $m^6dA$ calls for adenines close to $m^5C$ sites**. *E. coli* K-12 has $m^5C$ at the second cytosine at CC(A/T)GG sites(Kahramanoglou et al. 2012). We used SMRT sequencing data for *E. coli* from a recent study(Fang et al. 2012) and examined the IPD ratios for A sites within +/-10bp from CC(A/T)GG sites to estimate the false positive, excluding known $m^6dA$ events at G<u>A</u>TC and A<u>A</u>CNNNNNNGTGC/GCACNNNNNNG<u>A</u>A. Based on these selected A sites, we estimate the false positive rate of $m^6dA$ calls due to neighboring $m^5C$ sites (Supplemental Fig S7a)
.

**$m^6dA$ DIP sequencing**. We followed the same protocol as used in the recent study (Wu et al. 2016). Briefly, genomic DNA from human lymphoblastoid cell lines derived from a family trio were purified with DNeasy kit (QIAGEN, 69504). For each sample, 5 µg DNA was sonicated to 200–500 bp with Bioruptor. Then, adaptors were ligated to genomic DNA fragments following the Illumina protocol. The ligated DNA fragments were denatured at 95 degree for 5 min. Then, the single-stranded DNA fragments were immunoprecipitated with 6 mA antibodies (5 µg for each reaction, 202-003, Synaptic Systems) overnight at 4 degrees. $m^6dA$ enriched DNA fragments were purified according to the Active Motif hMeDIP protocol. IP DNA and input DNA were PCR amplified with Illumina indexing primers, and were then subjected to multiplexed library construction and sequencing with Illumina HiSeq sequencing.

**Analysis of $m^6dA$ DIP sequencing data**. BWA 0.7.8 (Li and Durbin 2009) was used to align the human $m^6dA$-DIP-seq reads to the UCSC hg19. Peaks called from the green algae genome were obtained from the authors of the original study.

**Consensus analysis of $m^6dA$-DIP-seq reads across different L1s**. The putative full-length L1s were aligned based on their 5' UTR. At each aligned position, the $m^6dA$-DIP-seq read coverage for different L1s were aggregated, and normalized to the A/T frequency across all the full length L1s. To further rule out the possibility of biased background distribution, we also normalized the average read coverage to the aggregated read coverage from $m^6dA$-DIP-seq of WGA DNA or input DNA.

## Data Access

The sequencing data are deposited in the Sequence Read Archive (SRA) with the following the accession numbers: SRP102471 (SMRT-seq of *C. innocuum* Native DNA with 1 SMRTcell), SRP102628 (SMRT-seq of *C. difficile* Native DNA with 2 SMRTcells and WGA with 2 SMRTcells), SRP105216 (SMRT-seq of *H. pylori* Native DNA with 2 SMRTcells), SRP102373 (SMRT-seq of *S. aureus* Native DNA with 1 SMRTcell), SRP105217 (SMRT-seq of *C. reinhardtii* Native DNA with 20 SMRTcells and WGA with 18 SMRTcells), and SRP128153 (SRX3538573: $m^6$dA-DIP-seq of HG002 Native DNA, SRX3538574: input DNA of HG002, SRX3538575: $m^6$dA-DIP-seq of HG003 Native DNA, SRX3538576: input DNA of HG003, SRX3538577: $m^6$dA-DIP-seq of HG004, SRX3538578: input DNA of HG004, SRX3538579: $m^6$dA-DIP-seq of GM12878 Native DNA, SRX3538580: $m^6$dA-DIP-seq of GM12878 WGA, and SRX3538581: input DNA of GM12878).

**Code availability**. The novel methods presented in the manuscript are implemented in R (Team 2013), and the codes are available in Supplemental Material and at https://github.com/fanglab/SMRTER

## Acknowledgements

We thank the members of the Fang laboratory for critical discussion, and the people who contributed to the generation of the publically available SMRT sequencing data for the human lymphoblastoid cell lines. The work is partially funded by the seed grant (G.F.) from Icahn Institute for Genomics and Multiscale Biology, R01 GM114472 (G.F.) from National Institutes of Health, and a Nash Family Research Scholar Award (G.F.) from the Friedman Brain Institute. This work was also supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

## Author Contributions

G.F. conceived project and supervised the research. S.Z. and G.F. designed the methods and experiments. S.Z. performed most of the computational analyses. G.D., T.P.W., M.S., Z.H., J.A.G. and R.S. conducted the wet lab experiments. G.D. and R.S. designed and conducted SMRT sequencing. S.Z, J.B., T.P.W., Z.H., G.L., A.C., C.H., A.X., R.P., E.E.S. and G.F. contributed to data analysis and interpretation. S.Z. and G.F. wrote the manuscript with inputs from all co-authors.

## Competing Financial Interests

E.E.S. is on the scientific advisory board of Pacific Biosciences.

## References


2012. Detecting DNA Base Modifications Using Single Molecule, Real-Time Sequencing. *Pacific Biosciences White Book*.

Babushok DV, Kazazian HH. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Human mutation* **28**: 527-539.

Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**: 1653-1659.

Beaulaurier J, Zhu S, Sebra R, Zhang X-S, Rosenbluh C, Deikus G, Shen N, Munera D, Waldor MK, Blaser M et al. 2015. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. *Nature communications* **6:7438** DOI: 10.1038/ncomms8438.

Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, Froula J, Kang DD, Malmstrom RR, Morgan RD. 2016. The epigenomic landscape of prokaryotes. *PLoS Genet* **12**: e1005854.

Casadesús J, Low D. 2006. Epigenetic gene regulation in the bacterial world. *Microbiology and molecular biology reviews* **70**: 830-856.

Casadesús J, Low DA. 2013. Programmed Heterogeneity: Epigenetic Mechanisms in Bacteria. *Journal of Biological Chemistry* **288**: 13929-13935.

Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, Friedli M, Duc J, Jang SM, Turelli P, Trono D. 2014. Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes & development* **28**: 1397-1409.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research* **14**: 1394-1403.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133-138.

Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, Feng Z, Losic B, Mahajan MC, Jabado OJ. 2012. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nature biotechnology*.

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.

Fu Y, Luo G-Z, Chen K, Deng X, Yu M, Han D, Hao Z, Liu J, Lu X, Doré LC. 2015. N 6-Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas. *Cell* **161**: 879-892.

Goodier JL, Kazazian HH. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**: 23-35.

Greer EL, Blanco MA, Gu L, Sendinc E, Liu J, Aristizábal-Corrales D, Hsu C-H, Aravind L, He C, Shi Y. 2015. DNA Methylation on N 6-Adenine in C. elegans. *Cell* **161**: 868-878.

Hata K, Sakaki Y. 1997. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene* **189**: 227-234.

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**: 108-112.

Heithoff DM, Sinsheimer RL, Low DA, Mahan MJ. 1999. An essential role for DNA adenine methylation in bacterial virulence. *Science* **284**: 967-970.

Heng HH, Bremer SW, Stevens JB, Ye KJ, Liu G, Ye CJ. 2009. Genetic and epigenetic heterogeneity in cancer: A genome-centric perspective. *Journal of cellular physiology* **220**: 538-547.

Huang F-P, Platt N, Wykes M, Major JR, Powell TJ, Jenkins CD, MacPherson GG. 2000. A discrete subpopulation of dendritic cells transports apoptotic intestinal epithelial cells to T cell areas of mesenteric lymph nodes. *The Journal of experimental medicine* **191**: 435-444.

Jen FE-C, Seib KL, Jennings MP. 2014. Phasevarions mediate epigenetic regulation of antimicrobial susceptibility in Neisseria meningitidis. *Antimicrobial agents and chemotherapy*: AAC. 00004-00014.

Kahramanoglou C, Prieto AI, Khedkar S, Haase B, Gupta A, Benes V, Fraser GM, Luscombe NM, Seshasayee AS. 2012. Genomics of DNA cytosine methylation in Escherichia coli reveals its role in stationary phase transcription. *Nature communications* **3**: 886.

Kozdon JB, Melfi MD, Luong K, Clark TA, Boitano M, Wang S, Zhou B, Gonzalez D, Collier J, Turner SW. 2013. Global methylation state at base-pair resolution of the Caulobacter genome throughout the cell cycle. *Proceedings of the National Academy of Sciences* **110**: E4658-E4667.

Koziol MJ, Bradshaw CR, Allen GE, Costa AS, Frezza C, Gurdon JB. 2016. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nature structural & molecular biology* **23**: 24-30.

Laszlo AH, Derrington IM, Brinkerhoff H, Langford KW, Nova IC, Samson JM, Bartlett JJ, Pavlenok M, Gundlach JH. 2013. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences* **110**: 18904-18909.

Lentini A, Lagerwall C, Vikingsson S, Mjoseng HK, Douvlataniotis K, Vogt H, Green H, Meehan RR, Benson M, Nestor CE. 2017. A reassessment of DNA immunoprecipitation-based genomic profiling. *bioRxiv*: 224279.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Luo G-Z, Blanco MA, Greer EL, He C, Shi Y. 2015. DNA N6-methyladenine: a new epigenetic mark in eukaryotes? *Nature Reviews Molecular Cell Biology* **16**: 705-710.

Luo G-Z, Wang F, Weng X, Chen K, Hao Z, Yu M, Deng X, Liu J, He C. 2016. Characterization of eukaryotic DNA N6-methyladenine by a highly sensitive restriction enzyme-assisted sequencing. *Nature communications* **7**.

Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, Pavlenok M, Niederweis M, Gundlach JH. 2012. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature biotechnology* **30**: 349-353.

Manso AS, Chai MH, Atack JM, Furi L, Croix MDS, Haigh R, Trappetti C, Ogunniyi AD, Shewell LK, Boitano M. 2014. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nature communications* **5**.

Miller JC, Brown BD, Shay T, Gautier EL, Jojic V, Cohain A, Pandey G, Leboeuf M, Elpek KG, Helft J. 2012. Deciphering the transcriptional network of the dendritic cell lineage. *Nature immunology* **13**: 888-899.

Mondo SJ, Dannebaum RO, Kuo RC, Louie KB, Bewick AJ, LaButti K, Haridas S, Kuo A, Salamov A, Ahrendt SR. 2017. Widespread adenine N6-methylation of active genes in fungi. *Nature genetics*.

Pak TR, Altman DR, Attie O, Sebra R, Hamula CL, Lewis M, Deikus G, Newman LC, Fang G, Hand J. 2015. Whole-genome sequencing identifies emergence of a quinolone resistance mutation in a case of Stenotrophomonas maltophilia bacteremia. *Antimicrobial agents and chemotherapy* **59**: 7117-7120.

Reedman BM, Klein G. 1973. Cellular localization of an Epstein-Barr virus (EBV)-associated complement-fixing antigen in producer and non-producer lymphoblastoid cell lines. *International Journal of Cancer* **11**: 499-520.

Reiner A, Yekutieli D, Benjamini Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**: 368-375.

Sánchez-Romero MA, Cota I, Casadesús J. 2015. DNA methylation in bacteria: from the methyl group to the methylome. *Current opinion in microbiology* **25**: 9-16.

Schadt EE, Banerjee O, Fang G, Feng Z, Wong WH, Zhang X, Kislyuk A, Clark TA, Luong K, Keren-Paz A. 2013. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. *Genome research* **23**: 129-141.

Schreiber J, Wescoe ZL, Abu-Shumays R, Vivian JT, Baatar B, Karplus K, Akeson M. 2013. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences* **110**: 18910-18915.

Schuster SC. 2008. Next-generation sequencing transforms today's biology. *Nature methods* **5**: 16.

Shulha HP, Cheung I, Guo Y, Akbarian S, Weng Z. 2013. Coordinated cell type–specific epigenetic remodeling in prefrontal cortex begins before birth and continues into early adulthood. *PLoS Genet* **9**: e1003433.

Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods* **11**: 817-820.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*: 4.10. 11-14.10. 14.

Team RC. 2013. R: A language and environment for statistical computing.

Wion D, Casadesús J. 2006. N6-methyl-adenine: an epigenetic signal for DNA–protein interactions. *Nature Reviews Microbiology* **4**: 183-192.

Wu TP, Wang T, Seetin MG, Lai Y, Zhu S, Lin K, Liu Y, Byrum SD, Mackintosh SG, Zhong M. 2016. DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature*.

Young LS, Rickinson AB. 2004. Epstein–Barr virus: 40 years on. *Nature Reviews Cancer* **4**: 757-768.

Zhang G, Huang H, Liu D, Cheng Y, Liu X, Zhang W, Yin R, Zhang D, Zhang P, Liu J. 2015. N 6-Methyladenine DNA Modification in Drosophila. *Cell* **161**: 893-906.

Zhou C, Liu Y, Li X, Zou J, Zou S. 2016. DNA N6-methyladenine demethylase ALKBH1 enhances osteogenic differentiation of human MSCs. *Bone Research* **4**: 16033.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data* **3**.

## FIGURE LEGENDS

**Figure 1.** Differences between bacterial and eukaryotic m$^6$dA methylomes and a novel approach for mapping m$^6$dA events in eukaryotic organisms. **(a)** Comparison between bacterial and eukaryotic m$^6$dA methylomes over three aspects. **(b)** A novel approach for mapping and characterizing m$^6$dA events in eukaryotic genomes. The novel approach, including a set of methods as summarized on the left, is comprehensively evaluated using subsampled bacterial m$^6$dA methylome data, and applied to *Chlamydomonas reinhardtii* (green algae) and human lymphoblastoid cells (LCLs).

### Figure 2: Comprehensive evaluation of m$^6$dA detection based on SMRT-seq data

(**a & b**) Sensitivity-FDR curves at different levels of per strand SMRT-seq coverage (**a**) and fraction of methylated A sites in the genome (**b**). Curves are estimated based on either p value or IPD ratio; both are shown. FDR estimation is based on the coverage-matched native (*Escherichia coli* with m$^6$dA at GATC sites; Methods) and WGA samples.

(**c**) FDRs estimated for different combinations of per strand SMRT-seq $coverage$ and fraction of m$^6$dA sites, $f(m^6dA/A)$, in the genome. FDR estimation is based on the coverage-matched native and WGA samples (Methods) at an IPD ratio of 4.

(**d**) Motif specific methylation detection leads to more reliable m$^6$dA calls with lower FDRs.

(**e**) Distribution of *p* values (-$\log_{10}$) and IPD ratios of $m^6$dA events (red) and non-methylated A's (black) from eleven well-characterized bacterial $m^6$dA methylomes.

(**f**) Enrichment score for motifs with different fractions of motif sites methylated across the genome $f_m(m^6dA/A)$, estimated based on *p* value (-$\log_{10}$, left) and IPD ratio (right). SMRT-seq data from eleven bacterial species/strains with well-characterized $m^6$dA methylomes are used for this simulation analysis.

(**g**) Schematic illustrating single molecule-level analysis for the estimation of partial methylation. A single molecule (two DNA strands and two adapters) and the subreads that are produced from the top strand of this molecule in SMRT-seq (top panel). For a given genomic position, when non-single molecule analysis is performed, IPD ratios for the methylated and non-methylated subreads follow two exponential distributions (red and black curves in the second panel). In contrast, when single molecule analysis was performed, IPDs ratios across all molecules follow two normal distributions with smaller variance over increasing coverage per molecule-strand (third and fourth panels).

(**h**) Estimation of partial methylation $f_l(m6dA/A)$ by aggregate analysis (left panel) and single molecule-level analysis (right panel). X-axis: background truth $f_l$ based on simulation; Y-axis: estimated $f_l$ . Dots: 4,359 A's with known fraction of $m^6$dA methylation based on subsampling from a well-characterized *E. coli* $m^6$dA methylome.

(**i & j**) Distribution of IPD ratios for partially methylated $m^6$dA sites and non-methylated A's based on aggregate analysis (**i**) and single molecule level analysis (**j**). The inset provides an enlarged view. The motif enrichment score for the same, known methylation

motif GATC significantly differs between the two types of analyses (1.3 in aggregated analysis vs. 25 in single molecule analysis).

**Figure 3. Characterization of a complete m$^6$dA methylome of *C. reinhardtii* reveals novel biological insights.**

(**a**) FDR estimation by comparing the IPD ratio distribution of *C. reinhardtii* native (red) with WGA (black) samples. The inset provides an enlarged view. (**b**) A rigorous motif enrichment analysis reveals that V<u>A</u>TB (V = A, C or G and B = C, G or T) is the m$^6$dA motif of in *C. reinhardtii*. Each $4 \times 4$ heatmap corresponds to all sixteen 4-mer motifs, for which 2nd and 3rd bases are fixed at the center/title (e.g. <u>A</u>A). The rows and columns in the heatmaps represent the first and last bases of 4-mer motifs. Each cell in the following $4 \times 4$ heatmaps shows the motif enrichment score based on the native DNA sample. (**c**) Putative m$^6$dA sites called by SMRT-seq are highly consistent with those detected by independent techniques: m$^6$dA-DIP-seq (DIP), m$^6$dA-CLIP-exo-seq (CLIP) and m$^6$dA-RE-seq (RE). (**d**) VATB, but not non-VATB (i.e. TATN/NATA), motifs have a periodic pattern of IPD ratio distribution around TSS's. Average IPD ratio (normalized by motif frequency) for each of the nine VATB motifs (top panel) and each of the seven non-VATB motifs (bottom panel) are plotted around TSS's. (**e**) Relationship across four different distributions (top to bottom panels): average IPD ratio of VATB sites, nucleosome positioning, and frequency of V<u>A</u>TB and non-VATB motif sites. Peaks and valleys of the periodic patterns are indicated by red and blue dots, aligned across the four panels. (**f**) Illustrative examples showing m$^6$dA sites near the TSS's of three

genes. This figure is adapted from *Fu et al.*(Fu et al. 2015) where we project $m^6dA$ sites detected by SMRT-seq (red dots; FDR < 0.05; randomly generated heights to ease visualization) on top of GATC and CATG sites detected by $m^6dA$-RE-seq (blue bars; middle panel) and nucleosome occupancy (bottom panel). (**g**) $m^6dA$ events at VATB sites are associated with active gene expression. Average IPD ratios are compared between two groups of genes with high (FPKM>1) and low (FPKM<1) expression levels. (**h**) The correlation between the gene expression level in *C. reinhardtii* and methylated VATB on gene promoters. X-axis represents the number of methylated VATB sites (IPD ratio > 4.5; FDR=0.05) within [0, +2,000bp] of TSS's the number of VATB sites with IPD ratio > 4.5 (FDR=0.05) in the [0,+2,000] of TSS's. Y-axis represents the mean $\log_2$ FPKM of genes. The error bars represent standard errors. (**i-k**) Single molecule, strand-specific analysis of SMRT-seq data to examine full-, non- or hemi-methylation status at $m^6dA$ sites. Three sets of $m^6dA$ sites are analyzed: $m^6dA$ in GATC sites (**h**) and CATG sites (**i**) based on $m^6dA$-RE-seq(Fu et al. 2015) and (**j**) VATB sites with high aggregate IPD ratio (IPD ratio > 4.5; Methods) based on SMRT-seq. The X- and Y- axes denote the single molecule, strand-specific IPD ratio of each pair of reverse-complementary VATB sites at the two strands of each single molecule.

**Figure 4. $m^6dA$ deposition on full length L1s in human LCLs**

(**a**) Mean IPD ratio of A sites (adjusted by the frequency of A's) across 1,274 young (evolutionary age<10 million years), full-length (>6,000bp) L1s for three hLCL lines, respectively. Consistent across the trio, the IPD ratio is relatively higher in the promoter and proximal region than the flanking regions.

(**b**) The mean IPD ratio of A sites at full length L1s is inversely correlated with the L1s' evolutionary ages in Human LCLs. The heatmap shows the mean IPD ratio of A's on each L1, [0, +500] from 5' UTR start site, for each of the trio. As indicated in the sidebar, L1s (rows) are ordered by their evolutionary ages. Consistently across the trio, the IPD ratio of A sites is higher in younger full-length L1s than in older L1s.

(**c**) Average $m^6$dA-DIP-seq read count (adjusted for the read count in the input DNA sample and the A/T content) on Human LCLs young (1,274), middle-aged (4,164), and old L1 elements (1,670), respectively. Consistent with SMRT-seq data, $m^6$dA is enriched at the promoter and proximal region of young full-length L1s.

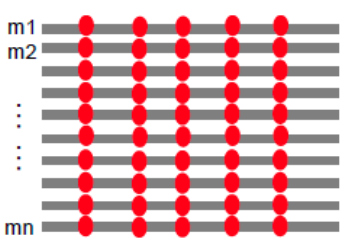
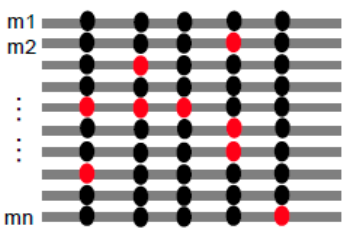
(**d**) Average $m^6$dA-DIP-seq read count adjusted for the A/T content and the read count in two control samples on Human LCL young L1 elements, respectively: input DNA as control (black curve in top panel) and $m^6$dA-DIP-seq on WGA as control (blue curve in bottom panel).

(**e**) Motif AG is enriched for putative $m^6$dA events. The barplot represents the motif enrichment score of all dinucleotide motifs in each of the trio. The putative methylated position is underscored. It suggests that motif AG is enriched for high IPD ratios in clear contrast to all the other dinucleotides.

(**f**) Motif enrichment analysis of human young full-length L1s. Each $4 \times 4$ heatmap corresponds to all sixteen 4-mer motifs, for which 2nd and 3rd bases are fixed at the center/title. The rows and columns in the heatmaps represent the first and last bases of 4-mer motifs. Each cell in the following $4 \times 4$ heatmaps shows the motif enrichment score based on the native DNA sample.

(**g**) Peaks of putative $m^6dA$ events across human young full-length L1s occur at loci with certain sequence features. Top panel: level of sequence conservation across young full-length L1 elements based on multiple alignment by Mauve(Darling et al. 2004); two middle panels: frequency of AG dinucleotides (relative to A's) and A's on young full-length L1s; bottom panel: frequency of putative $m^6dA$ events at each locus across all young full-length L1s (averaged among the trio). The peaks of sequence conservation, AG/A frequency and $m^6dA$ frequency across young full-length L1s are co-localized as indicated by the red, blue and green dots.

**a**

| | Bacterial m⁶dA methylomes | Eukaryotic m⁶dA methylomes |
|---|---|---|
| **m⁶dA/A abundance** | Generally high, although it varies across species (some 0%, some >2%) | Generally orders of magnitude lower than bacteria, although it varies across cell types |
| **Motif-driven nature**<br>Genome<br>Motif sites with methylation<br>Motif sites without methylation |  <br>Highly motif driven: nearly all (>95%) the motif sites are methylated across the genome |  <br>Weakly motif driven: a small fraction (e.g. <3%) of motif sites are methylated |
| **Cell-to-cell heterogeneity**<br>Molecule<br>Methylation events<br>Not methylated | 100% methylation<br>  <br>m1<br>m2<br>mn<br>Generally common | Partial methylation<br>  <br>m1<br>m2<br>mn<br>Common in certain species | Unclear in previous studies; expected to be similarly heterogeneous, if not more so, given the large number of cell types and subpopulations within each cell type |

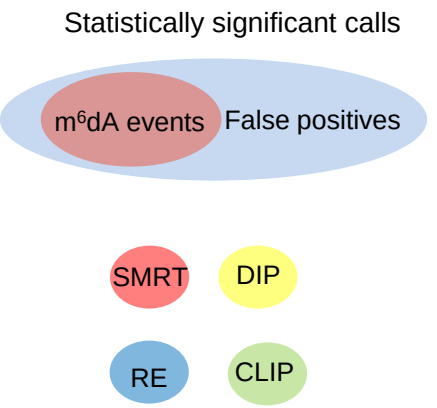


**b**

## A novel framework for mapping of m⁶dA events in eukaryotic genomes

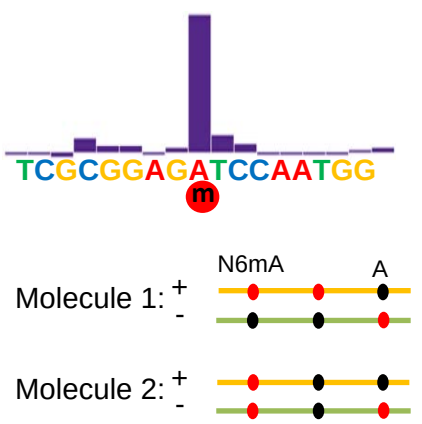
**Effective use of negative controls**

- Control for false positive m⁶dA calls
- Adjust for biases in m⁶dA motif discovery
- Integration of indepdenent, complementary techniques

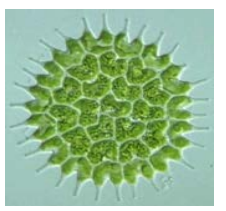



**High resolution analyses**

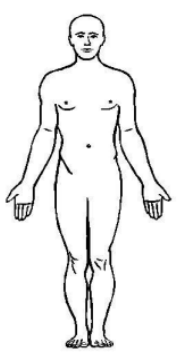
- Single-base resolution mapping
- Single-molecule, strand-specific characterization




**Comprehensive evaluation based on subsampling of well-characterized bacterial m⁶dA methylomes**



**A complete map of m⁶dA events in *C. reinhardtii* at single-base and single-molecule resolution**



**Analysis of putative m⁶dA events in human lymphoblastoid cells**