Critical assessment of nanopore sequencing for the detection of multiple forms of DNA modifications

Yimeng Kong[1,2], Yanchun Zhang[1†], Edward A. Mead[1†], Hao Chen[1], Christian E. Loo[3], Yu Fan[1], Mi Ni[1], Xue-Song Zhang[4], Rahul M. Kohli[3] and Gang Fang[1#]

[1] *Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA*

[2] *Center of Clinical Laboratory Medicine, Zhongda Hospital, School of Medicine, Advanced Institute for Life and Health, Southeast University, Nanjing, China*

[3] *Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA*

[4] *Center for Advanced Biotechnology and Medicine, Rutgers University, New Brunswick, NJ, USA*

[†] *These authors contributed equally to this work.*

 # Address correspondence to: gang.fang@mssm.edu

Abstract:

While nanopore sequencing is increasingly used for mapping DNA modifications, it is important to recognize false positive calls as they can mislead biological interpretations. To assist biologists and methods developers, we describe a framework for rigorous evaluation that highlights the use of false discovery rate with rationally designed negative controls capturing both general background and confounding modifications. Our critical assessment across multiple forms of DNA modifications highlights that while nanopore sequencing performs reliably for high-abundance modifications, including 5-methylcytosine (5mC) at CpG sites in mammalian cells and 5-hydroxymethylcytosine (5hmC) in mammalian brain cells, it makes a significant proportion of false positive detections for low-abundance modifications, such as 5mC at CpH sites, 5hmC and N6-methyldeoxyadenine (6mA) in most mammal cell types. This study highlights the urgent need to incorporate this framework in future methods development and biological studies, and advocates prioritizing nanopore sequencing for mapping abundant over rare modifications in biomedical applications.

Introduction:

The advent of nanopore sequencing technology has revolutionized the landscape of genomic research[1,2]. Not only does it generate high throughput long reads, but it also has a unique ability for the direct mapping of DNA modifications[3–5]. It circumvents the conventional requirements for chemical, antibody, or enzymatic treatments, enabling the analysis of native DNA[4,5]. The capacity for resolving both primary nucleotide sequence and DNA modifications along individual long reads greatly empowers the study of human epigenomes in health and diseases, as well as a broad spectrum of species[6–9].

DNA modifications, such as 5-methylcytosine (5mC) and its derivative 5-hydroxymethylcytosine (5hmC), play a crucial role in the regulation of genomic functions and are pivotal to understanding the complex mechanisms underlying various biological processes and diseases[10–12]. However, as the adoption of nanopore sequencing expands and the repertoire of computational tools for detecting DNA modifications grows, it has become increasingly important to critically assess the technology's fidelity, particularly regarding its propensity to generate false positives. Misinterpretation of these false positive modification calls can mislead further biological studies and clinical applications[4,13–16]

A notable example was the debate over DNA N6-methyladenine (6mA) in the mammalian genome. In brief, 6mA was thought to be exclusively present in the DNA of bacteria until several studies reported its detection in nematodes, insects, mice, and humans[17–21]. While several studies reported important functions of 6mA in human development and diseases[18,21], subsequent studies then discovered multiple confounding factors and raised caution about the level of 6mA in the human genome[15,16,22,23], leading to an active debate. The debate was well clarified when bacterial contamination was demonstrated to have contributed to the significantly overestimated 6mA abundance in the mammalian genome[4,13]. In another example of confusion caused by different sequencing modalities, several studies using nanopore sequencing and/or whole genome bisulfite sequencing reported extensive 5mC in mitochondrial genomes (mtDNA) across multiple species[24,25], while other studies suggested that mtDNA methylation levels had been overestimated due to confounding factors[26–29]. This debate was addressed when subsequent studies performed rigorous method evaluation[27] and reported the 5mC level is extremely low in mammalian mtDNA, calling into question the previously described functional roles of 5mC in mtDNA metabolism.

These examples of false positive detections highlight the critical need for rigorous development and evaluation of modification mapping methods to prevent erroneous assumptions that could misguide future research. Motivated to address these challenges, we recently wrote a perspective detailing the pitfalls of detecting DNA and RNA modifications and strategies to navigate them[4]. One of the most important pitfalls that requires attention is the risk of false positives when mapping DNA modifications of low abundance. A method may perform very well in detecting DNA modification of high abundance in a genome but can give mostly false positive calls when applied to a sample with low abundance of the same modification. Essentially, this is related to the important statistical concept of false discovery rate (FDR). Fundamentally, the reliability of a modification mapping technology not only depends on the intrinsic properties of the technology itself but also the abundance of the DNA modification of interest in a specific sample. However, existing methods for modification detection are usually trained and evaluated using datasets with predefined ratios of modified to non-modified bases (e.g. 1:1) that do not represent physiologically relevant levels of these modifications (e.g. 5hmC and 6mA are much less abundant than 5mC in most human cell

types). This gap between method development and their ultimate applications underlines the critical need for FDR estimation to safeguard data interpretation.

In this study, we focused on nanopore sequencing because it represents the most promising technology for mapping a variety of DNA (and RNA) modifications. We generated data using the latest R10.4.1 nanopore sequencing kits and critically assessed several versions of the official software for modification calling including a very recent release (as of September 15, 2024), which was reported to have high accuracy for detecting 5mC, 5hmC, and 6mA. Our assessment revealed contrasting performance of these models based on the abundance of the modifications. The technology exhibits high reliability for detecting high-abundance modifications, such as 5mC at CpG sites in mammalian genomes. However, it demonstrates a marked propensity for false positives when detecting low-abundance modifications. For instance, we observed this phenomenon with 5mC at non-CpG sites (namely 5mCpH) and 5hmC, which, although common in mammalian brain cells, are considerably less prevalent in other cell types[30–34]. This disparity in detection reliability is also evident with 6mA, which is abundant in bacteria and certain protozoan species, but occurs at much lower levels in higher eukaryotes[15,16,22,23].

To mitigate the risk of false positive calls and provide guidance for broad users and tool developers, we present a framework that highlights the use of rationally designed negative controls (capturing both general background and confounding modifications) and FDR analysis to evaluate the reliability of detected modification events. This approach aims to help researchers navigate through the pitfalls of nanopore sequencing, particularly in the detection of low-abundance modifications (**Fig. 1a**).

Results:

False positive calls can dominate detection when the DNA modification of interest has low abundance.

To illustrate false positive calls, we first performed a whole genome amplification (WGA) of a mouse genomic DNA (gDNA) sample. The amplified DNA from the WGA process is essentially modification-free[4,35], serving as a negative control. The WGA sample was sequenced with the latest R10.4.1 kit and flow cell (**Supplementary Table 1**), followed by read-level 5mC modification calling using the latest (as of September 15, 2024) official software DORADO (**Methods, Supplementary Table 2**). Although no or very low-level 5mC sites were expected from the WGA sample, 0.16%-3.19% of unmethylated C sites were called as 5mC from nanopore reads, representing false positives (**Fig. 1b; Supplementary Fig. 1**). The 3.19%-0.16% of false positive calls correspond to different thresholds on the read-level classification probability (P_{mod}) in the DORADO output, ranging from low-confident (0.5) to high-confident (0.99) (**Fig. 1b; Supplementary Fig. 1; Methods**). As expected, the percentage of false positive 5mC sites decreased as the threshold increased. While 3.19% false positive 5mC sites were detected with a P_{mod} threshold of 0.5, ~0.16% false positive 5mC sites were detected with a P_{mod} threshold of 0.99 (**Fig. 1b**). Consistent results were observed across various WGA preparation methods (**Methods**), validating that the WGA samples are reliable negative controls (**Supplementary Fig. 2**). This analysis demonstrated that even with the latest software and a high confident threshold, the nanopore read-level 5mC calling can still make millions of false positive calls across a mammalian genome.

Next, we performed the same analysis on the native gDNA samples extracted from a mouse prefrontal cortex (mPFC). 6.92%-3.31% of unmethylated C sites were called as 5mCs across the same ranges (0.5-0.99 of P_{mod} thresholds) (**Fig. 1c; Supplementary Figs. 1 & 2**). While the percentage of unmethylated C sites changed dramatically with increased (more strict) P_{mod} threshold, the percentage of 5mC sites stays fairly stable (**Supplementary Fig. 1**), which is largely consistent with previously reported levels of 5mC/C (4-5% as estimated by LC-MS/MS) in mPFC[36].

The use of FDR to safeguard DNA modification detection across various abundance levels.

Our above analysis demonstrates that the reliability of modification calls using the same calling method (DORADO in this case) depends on the abundance of a modification of interest in a specific sample. While a seemingly small proportion (0.16%) of false positive calls may seem negligible in the mPFC sample (with abundant 5mC), it could significantly mislead any biological interpretations in samples with the low 5mC/C levels, such as the WGA sample. This statistical concept is well established and is nicely captured by the measure FDR, defined as the ratio of N_{fp}/N_p, where N_p is the total number of all positive modification calls detected from the native sample (true positive N_{tp} + false positive calls N_{fp}), and N_{fp} can be inferred from the negative WGA sample (free of modification). FDR decreases from 0.46 to 0.05 for the native mPFC gDNA data, as the P_{mod} threshold increased from 0.5 to 0.99, consistently with the FDR of a biological replicate (mPFC2) (**Fig. 1d**). To apply FDR analysis to the WGA dataset, we generated 2-round WGA dataset on mPFC2 to estimate N_{fp}, and compared it with the 1-round WGA dataset of mPFC2 across different P_{mod} thresholds. Notably, the FDR remained >0.9 for the mouse WGA data across P_{mod} thresholds (**Fig. 1d**). Taking the P_{mod} threshold of 0.99 for example, 90.46% of 5mC events called from the mouse WGA data are false positives, while only 4.9% of the 5mC sites called from the native mPFC data are false positives.

The mouse WGA and native mPFC data described above represent two cases with either very low or high abundant 5mC. Although 5mC is generally abundant in mammalian genomes, much lower 5mC levels have been observed in other eukaryotic species such as yeast[37] and Drosophila[37], as well as in mammalian mtDNA[26–29]. To further illustrate the use of FDR across various abundance levels of 5mC (representing its wide range across different eukaryotic and prokaryotic genomes), we simulated a series of 5mC/C levels across different orders of magnitude. Specifically, we randomly mixed nanopore reads from the native mPFC and mouse WGA samples at different proportions creating 5mC/C levels from 10^{-1} to 10^{-6}. As shown in **Fig. 1e,** FDR increases as 5mC/C level decreases: for 5mC/C levels below 10^{-3}, FDR stays close to 1, meaning the vast majority of the 5mC calls are false positives within a gDNA sample with rare 5mCs. This observation is consistent with the definition of FDR = $N_{fp}/(N_{tp} + N_{fp})$. When the modification of interest is highly abundant in a genome of interest, a relatively small number of false positives are associated with a low FDR as $N_{tp} \gg N_{fp}$ (the native mPFC sample, **Fig. 1d**). However, when the modification of interest is of very low abundance in the genome ($N_{tp} \sim N_{fp}$ or $N_{tp} \ll N_{fp}$), the false positive calls will result in a much higher FDR, meaning the vast majority of the called modifications are false positives (the mouse WGA sample in **Fig. 1d**). This raises caution for the use of nanopore sequencing to call 5mC from species with extremely low levels of 5mC. To avoid this pitfall, FDR analysis should be applied, instead of relying on arbitrary threshold on P_{mod}, to reliably detect and interpret the called DNA modifications.

Critical evaluation of the detection of 5mCpG, 5mCpH, 5hmC, and 6mA in mPFC.

The above analysis was focused on 5mC across all sequence contexts in the mPFC sample. In mammalian genomes, 5mC is known to be mostly prevalent at CpG sites, but much less prevalent at non-CpG sites (i.e. 5mCpH, H=A/T/C): while certain brain cells have a high abundance of 5mCpH, most other cell types have very low levels of 5mCpH[30–32]. As expected, although 5mCpG calling has low FDRs in the mPFC sample, 5mCpH calling generally has much higher FDRs (**Fig. 1f**). This contrast between the reliability of 5mCpG and 5mCpH calling highlights the need for caution in the detection and interpretation of 5mCpH from various mammalian cell types. It also underscores the need to apply FDR evaluation specifically to modifications of interest within particular sequence contexts, rather than assuming equal reliability across all detected DNA modifications of the same type.

We further evaluated the use of nanopore sequencing for the direct detection of 5hmC, which is the oxidized derivative of 5mC and an intermediate in the DNA demethylation pathway. Compared to 5mC, 5hmC is much less abundant in mammal cells: while it is enriched in neurons, most other mammalian cells have very low levels of 5hmC[38,39]. As expected, a much lower 5hmC/C level was observed in mPFC than 5mC across all P_{mod} thresholds (**Supplementary Fig. 3**), consistent with previous estimations by LC-MS/MS[40] (5mC/C: ~4%-5% and 5hmC/C: ~0.6%). 5hmC calling has a much higher FDR (**Fig. 1g-i, Supplementary Figs. 4 & 5**). Consistent results were observed in the mPFC sample with the DORADO model designed specifically for 5mCpG and 5hmCpG calling ("5mCG_5hmCG" model) (**Supplementary Fig. 6**). Notably, while 5hmCpG has a relatively low FDR, 5hmCpH calling showed the highest FDR, consistent with the previous studies using orthogonal enzymatic methods that show 5hmC enrichment in the brain is almost exclusively at CpG sites[33,39] (**Fig. 1i, Supplementary Fig. 5b**).

Next, we assessed the direct detection of 6mA, a DNA modification that is prevalent in bacteria but extremely rare in mammalian cells[13,15,16,22,23,41]. Although the official tools recently enabled 6mA calling, the FDRs estimated on the mPFC sample are ~1 across P_{mod} thresholds, suggesting that the called 6mA events are essentially all false positives (**Supplementary Fig. 7)**. Amid the ongoing debate over 6mA in mammalian genomes[13,15,16], this assessment highlights the critical need for the use of FDR to safeguard against unrecognized false positive calls in DNA modification detection that could mislead biological and biomedical studies, consuming years of research efforts and resources.

Critical evaluation of the detection of 5mCpG, 5mCpH, 5hmC and 6mA in hLCL

Furthermore, we applied the FDR-based framework to a human lymphoblastoid cell line (hLCL, **Methods**). Read-level analysis on native nanopore reads showed 2.4%-4.9% 5mC/C level across the genome (**Supplementary Fig. 8**), consistent with previous characterization (~2.5%-4.0%) as estimated by LC-MS/MS[42]. FDR evaluation showed a similar trend of decrease, from 0.76 to 0.08 as the P_{mod} threshold increased from 0.5 to 0.99 (**Fig. 2a**), suggesting 5mC calling is largely reliable as observed for native mPFC gDNA.

However, FDR analysis of 5mCpH, 5hmC and 6mA calling showed that nanopore-sequencing-based calling of these three types of modifications are dominated by false positives in the hLCL sample, with high

FDRs (~1 across all P_{mod} thresholds for 5hmC and 6mA), consistent across versions of DORADO and base calling models (**Fig. 2a-d, Supplementary Fig. 9-12**). These observations are highly consistent with previous studies that showed: (1) 5mCpH is rare in most mammalian cell types except certain brain cells[30–34]; (2) the 5hmC/C level in hLCL is very low (only ~0.001%) as reported by LC-MS/MS[42]; and (3) 6mA in hLCL is rare, consistent with our previous assessment of 6mA in hLCL using the independent PacBio sequencing[13,14].

By comparing mPFC (low 5hmC enriched at CpG) and hLCL (very rare 5hmC), the FDR evaluation showed that 5hmCpG calls are largely reliable in mPFC, but not in hLCL (**Fig. 2e**). These findings are consistent with previous estimates of 5hmC/C level by LC-MS/MS, which showed that hLCLs have much lower 5hmC/C levels compared to mPFC : ~0.6% in mouse cerebral cortex[36,40] vs. ~0.001% in hLCL[42]. This comparison between different cell types highlights the distinct reliability of the same modification calling tool across samples with different levels of DNA modification.

The use of more comprehensive negative controls to account for confounding DNA modifications

The above FDR analyses were all based on the use of WGA samples as negative controls. While WGA serves as very helpful negative control to capture technology-centric false positives, it does not consider the confounding effect among different forms of DNA modifications (**Fig. 1a)**. On this basis, we hypothesized that the abundant 5mC in CpG sites on the mammalian genome may further confound nanopore based detection of 5hmC which also locates in CpG context, leading to false positive 5hmC calls (5mC being called as 5hmC that are not represented by WGA samples). For example, the FDR of ~0.4, as estimated using WGA as the negative control (**Fig. 2d & e**), appeared to support that 60% of 5hmCpG events called from hLCL are “reliable”. If this were true, it suggests 5hmC/CG levels of 0.047%-2.1%. Considering CpG dinucleotides account for 9.82% of total Cs (**Methods**), this equates to 0.0046%-0.21% 5hmC/C in the native hLCL genome (**Supplementary Fig. 12**), which is much higher than the 5hmC/C level (~0.001%) estimated by LC-MS/MS and BS-seq/oxBS-seq technologies from the same cell type[42]. This alerted us to consider the abundant 5mCpG events in hLCL may have confounded the detection of 5hmC, leading to false positive 5hmCpG calls not captured by WGA.

To test the hypothesis, we sequenced three bacterial strains with well-characterized 5mC motifs that cover all C contexts (CA, CT, CG and CC), but no 5hmC events (given the absence of associated TET enzymes[43] or hydroxymethylases[44]; see **Methods**). These bacterial data serve as helpful negative controls (abundant 5mC yet 5hmC free) that allow us to evaluate false positive 5hmC calls due to the confounding 5mC on the same context. Specifically, the *Neisseria gonorrhoeae* strain has GG5mCC and T5mCACC motifs[8,45]; the *Helicobacter pylori* strain has GG5mCC and G5mCGC motifs[3,45]; and the *Escherichia coli* K12 MG1655 strain has C5mCWGG (W=A/T)[45,46]. All of these motifs were nearly ~100% 5mC methylated across the genome by their own methyltransferases[45,47–49].

We assessed 5hmC calling in these bacterial data (**Methods**) and found that DORADO miscalls ~3.94%-0.11% of 5mC as false positive 5hmC, when P_{mod} increases from 0.5 to 0.99. These false positive 5hmC/5mC levels in native bacteria are significantly higher than 5hmC/C levels in the matched WGA controls (**Fig. 2f, Supplementary Fig. 13**), supporting that these 5hmC calls were indeed false positives

confounded by the abundant 5mC events. Importantly, the false positive 5hmC/5mC level miscalled from 5mC on bacterial genomes are higher than the 5hmCpG/CpG levels called from hLCL (**Fig. 2g, Supplementary Fig. 13**).

Even with a simulation of 70% 5mC/C levels among these motifs (to mimic the ~70% 5mCpG/CpG level on human genome, see **Methods**), the rate of false positive 5hmC calls due to confounding with 5mC remains higher than 5hmCpG/CpG levels detected in the hLCL data (**Fig. 2g**). This result implies that although the WGA-based negative control estimated a seemingly low FDR for 5hmCpG calling from hLCL, the use of additional negative controls, high in confounding 5mC but 5hmC-free, provided critical estimation of high FDR (~1, **Supplementary Fig. 14**), suggesting that the 5hmCpG events called from the hLCL sample are nearly all false positives. This analysis not only highlights that caution is needed when detecting 5hmC on a genome with abundant 5mC, but also the generally applicable need for more comprehensive negative controls in cases where different forms of DNA modifications can confound each other (**Fig. 1a**).

Discussion

In this study, we have critically assessed the reliability of nanopore sequencing for detecting multiple forms of DNA modifications, demonstrating that false positives can significantly influence data interpretation, especially for low-abundant modifications. This raises the need for caution in the current common practice where arbitrary thresholds are used for calling DNA modifications from nanopore data, which can confound data interpretation and mislead downstream biological analysis.

To address these challenges and guide both users and method developers, we have presented a framework that emphasizes the use of rationally designed negative controls (capturing both general background and confounding modifications) and FDR analysis (**Fig. 1a**). Using this framework, we demonstrated that the abundance of modification plays a critical role in the FDR evaluation. Specifically, we applied this framework to assess the calling of multiple modification types (5mC, 5hmC, and 6mA) at different genomic contexts (CpG and CpH), and we made observations consistent across both mouse and human samples using various versions of DORADO (v0.5.3 and the latest v0.7) and basecalling models (v4.2.0, v4.3.0 and the latest v5.0.0) as of Sept 15, 2024.

For 5mC, while detection at CpG sites is generally reliable with a low FDR, the same cannot be said for CpH sites, because 5mCpH is typically rare in most mammalian cells[30–34]. For 5hmC, although high reliability can be achieved in cell types where they are relatively abundant, such as the mPFC, high false positive rates are expected in most mammalian cell types where they are rare. For 6mA, which is prevalent in bacteria but largely absent in mammalian genomes[13,15,16], our FDR analysis showed that it cannot be reliably detected yet using nanopore sequencing even with the latest official model release. The generally high FDRs associated with 5mCpH, 5hmC, and 6mA calling from most mammalian cell types are in great contrast with the high detection accuracy (97.9% and 97.5% for 5mC/5hmC and 6mA all context detection, respectively) for DNA modifications reported by the nanopore sequencing official tool DORADO (https://nanoporetech.com/platform/accuracy/#variant-calling). This is essentially because the DORADO models were all developed and evaluated using training datasets with predefined ratios of modified and non-modified bases, which do not reflect the true biological variability in modification abundance in mammalian cell types.

Our work highlights the urgent need for the community to incorporate the FDR assessment framework along with rationally designed negative controls in their nanopore-based DNA modification mapping studies, especially for method development and evaluation. For biological studies and biomedical applications, our study supports the general reliability of 5mCpG mapping across mammalian cell lines, but limited applicability of existing 5hmCpG mapping methods to brain tissues or other tissues with high 5hmC abundance. In addition, our findings indicate a critical need for further efforts to incorporate data with physiologically relevant levels of modifications in the model training and evaluation, along with matched data generated using gold-standard methods such as ACE-seq[44] and EM-seq[50]. Before new reliable methods are developed and critically assessed for mapping low-abundance modifications, nanopore sequencing should be prioritized for mapping abundant, rather than rare, modifications in biological studies and biomedical applications.

Furthermore, we emphasized the importance of employing comprehensive negative controls that account for the confounding effects of abundant modifications on the detection of rarer ones, by demonstrating a significant level of false positive 5hmC called from abundant 5mCpG methylation in human cells (**Fig. 2f & g**). As the nanopore sequencing community aims to advance toward the simultaneous detection of various DNA modifications, such as DNA methylation, DNA damage, and abasic sites[51,52], the potential for cross-confounding effects cannot be overlooked. This caution also extends to the direct nanopore sequencing of RNA, where the diverse landscape of RNA modifications presents similar challenges[4,53,54].

In conclusion, the implications of our findings are particularly significant given the widespread application of nanopore sequencing in fields ranging from fundamental biological research to clinical diagnostics, including the analysis of cell-free DNA in liquid biopsies[55,56]. Detecting low-abundance modifications remains a considerable challenge, underscoring the need for a rigorous, transparent framework to evaluate the reliability of detected events.

Figure Captions

Fig. 1 A framework for the rigorous evaluation of DNA modification detected by nanopore sequencing, and demonstration of false positives and the use of false discovery rate (FDR) to evaluate multiple forms of DNA modification called from mouse prefrontal cortex (mPFC) samples.

a, The framework for evaluating the reliability of detected modifications using negative controls and FDRs. The use of rationally designed negative controls tailored based on prior knowledge including modification levels, genome contexts and coexistence of abundant confounding modifications.

b, False positive 5mC sites among the total C (%) (*y-axis*) in mPFC gDNA subjected to whole genome amplification (WGA, free of modification) by applying different thresholds on modification probability (P_{mod}) assigned by the official DORADO software (*x-axis*). Unmethylated C levels (%) for native mPFC and mouse WGA samples among all P_{mod} thresholds were shown in **Supplementary Fig. 1**.

c, 5mC sites among the total C (%) (*y-axis*) identified on reads from native mPFC gDNA across different thresholds on P_{mod} (*x-axis*).

d, FDR evaluation of read-level 5mC detection on mPFC sample (*y-axis*) across different thresholds on P_{mod} (*x-axis*). 2-round WGA on mPFC2 was compared the 1-round WGA on mPFC2 to estimate the FDR of the WGA sample (**Methods**).

e, FDR evaluation for a list of samples that have 5mC/C levels with different orders of magnitude, which were simulated by randomly mixing native reads with WGA reads to mimic a number of 5mC/C levels (**Methods**).

f, FDR evaluation for 5mC in CpG sites (left) and CpH sites (right) in the simulated samples with varying 5mC/C levels across different magnitude orders (**e**).

g, FDR evaluation of 5mC and 5hmC calls made from the native mPFC sample using DORADO software (v0.5.3) with the 5mC_5hmC model (v4.3.0). Consistent results were observed on the latest DORADO version (v0.7) and base calling model (v5.0.0) as of Sept 15, 2024 (**Supplementary Fig. 4**).

h, FDR evaluation of 5mC calls made at CpG and CpH sites in the native mPFC sample using DORADO software (v0.5.3) with the 5mC_5hmC model (v4.3.0). Consistent results were observed on the latest DORADO version (v0.7) and base calling model (v5.0.0) as of Sept 15, 2024 (**Supplementary Fig. 5a**).

i, FDR evaluation of 5hmC calls made at CpG and CpH sites in the native mPFC sample using DORADO software (v0.5.3) with the 5mC_5hmC model (v4.3.0). Consistent results were observed on the latest DORADO version (v0.7) and base calling model (v5.0.0) as of Sept 15, 2024 (**Supplementary Fig. 5b**).

Fig. 2. FDR evaluation of multiple forms of DNA modifications called from human LCL cells (hLCL) using DORADO software (v0.5.3) with the model (v4.3.0).

a, FDR evaluation for read-level 5mC and 5hmC calls across different thresholds on P_{mod} (*x-axis*). Consistent results were observed on the latest DORADO version (v0.7) and basecalling model (v5.0.0) as of Sept 15 2024 (**Supplementary Fig. 10**)

b, FDR evaluation for read-level 6mA calls.

c, FDR evaluation of 5mC calls at CpG and CpH sites. Consistent results were observed on the latest DORADO version (v0.7) and basecalling model (v5.0.0) as of Sept 15, 2024 (**Supplementary Fig. 11a**)

d, FDR evaluation of 5hmC calls at CpG and CpH sites. Consistent results were observed on the latest DORADO version (v0.7) and basecalling model (v5.0.0) as of Sept 15, 2024 (**Supplementary Fig. 11b**).

e, Comparison of 5hmC at CpG and CpH sites between mPFC and hLCL samples.

f, False positive (FP) 5hmC/5mC calls from bacterial genomes with well-characterized 5mC motifs (nearly 100% methylated) yet free of 5hmC.

g, The 5hmCG/CG levels detected in hLCL samples vs the FP 5hmC/5mC levels detected from bacterial 5mC motif sites that are free of 5hmC. The mean and standard deviation (represented by bars) of 5hmC levels were calculated from all 5mC motifs. FP 5hmC calls were estimated from 5hmC calls made from native bacterial reads alone, or in a mixture of 70% native bacterial reads and 30% WGA reads (**Methods**).

Acknowledgments

We thank Magdalena Ksiezarek, Yujie Liu and Yangmei Li for their help with nanopore sequencing library preparation and bacterial genomic DNA samples that we used as negative controls in this study. This work was supported by grants no. R35 GM139655 and R01 HG011095 (G.F.) from the National Institutes of Health. C.E.L. was supported by F31 HG012892. This work was supported in part by the staff and resources of Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. Y.K. was an employee at Icahn School of Medicine at Mount Sinai. During this project, she assumed a new role as a Professor at the Advanced Institute for Life and Health, Department of Medicine, Southeast University (SEU) in Nanjing, China, where she was supported by her start-up fund (RF1028623368) and Southeast University Interdisciplinary Research Program for Young Scholars (2024FGC1004) from SEU.

Competing interests

The authors declare no competing interests.

Author contributions

Study design: Y. K and G.F.; Data analysis: Y.K.; Data processing: Y.K., Y. Z. and H.C.; Cell culture and sequencing: E.A.M.; Data interpretation: Y. K., Y. Z., H. C., E.A.M., C.E.L., Y. F., M. N., X-S Z., R. M. K. and G. F.; Supervised research: G.F.; Wrote first draft of paper: Y.K. and G.F.; Approved paper: all authors.

References

1. Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).

2. Cheetham, S. W., Faulkner, G. J. & Dinger, M. E. Overcoming challenges and dogmas to understand the functions of pseudogenes. *Nat. Rev. Genet.* **21**, 191–201 (2020).

3. Tourancheau, A., Mead, E. A., Zhang, X. S. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat. Methods* **18**, 491–498 (2021).

4. Kong, Y., Mead, E. A. & Fang, G. Navigating the pitfalls of mapping DNA and RNA modifications. *Nat. Rev. Genet.* (2023) doi:10.1038/s41576-022-00559-5.

5. Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.* **20**, 157–172 (2019).

6. Kolmogorov, M. *et al.* Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat. Methods* **20**, 1483–1492 (2023).

7. Charalampous, T. *et al.* Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.* **37**, 783–792 (2019).

8. Cao, L. *et al.* mEnrich-seq: methylation-guided enrichment sequencing of bacterial taxa of interest from microbiome. *Nat. Methods* **21**, 236–246 (2024).

9. Ewing, A. D. *et al.* Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol. Cell* **80**, 915-928.e5 (2020).

10. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science (80-.).* **361**, 1336–1340 (2018).

11. Michalak, E. M., Burr, M. L., Bannister, A. J. & Dawson, M. A. The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat. Rev. Mol. Cell Biol.* **20**, 573–589 (2019).

12. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).

13. Kong, Y. *et al.* Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science (80-.).* **375**, 515–522 (2022).

14. Zhu, S. *et al.* Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. *Genome Res* **28**, 1067–1078 (2018).

15. Lentini, A. *et al.* A reassessment of DNA-immunoprecipitation-based genomic profiling. *Nat. Methods* **15**, (2018).

16. Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B. & Nestor, C. E. No evidence for DNA N6-methyladenine in mammals. *Sci. Adv.* **6**, 1–10 (2020).

17. Greer, E. L. *et al.* DNA Methylation on N6-Adenine in C. elegans. *Cell* **161**, 868–878 (2015).

18. Xie, Q. *et al.* N6-methyladenine DNA Modification in Glioblastoma. *Cell* **175**, 1228-1243.e20 (2018).

19. Zhang, G. *et al.* N6-methyladenine DNA modification in Drosophila. *Cell* **161**, 893–906 (2015).

20. Wu, T. P. *et al.* DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature* **532**, 329–333 (2016).

21. Xiao, C. Le *et al.* N 6 -Methyladenine DNA Modification in the Human Genome. *Mol. Cell* **71**, 306-318.e7 (2018).

22. O'Brown, Z. K. *et al.* Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics* **20**, 1–15 (2019).

23. Musheev, M. U., Baumgärtner, A., Krebs, L. & Niehrs, C. The origin of genomic N 6-methyl-deoxyadenosine in mammalian cells. *Nat. Chem. Biol.* **16**, 630–634 (2020).

24. Patil, V. *et al.* Human mitochondrial DNA is extensively methylated in a non-CpG context. *Nucleic Acids Res.* **47**, 10072–10085 (2019).

25. Dou, X. *et al.* The strand-biased mitochondrial DNA methylome and its regulation by DNMT3A. *Genome Res.* **29**, 1622–1634 (2019).

26. Owa, C., Poulin, M., Yan, L. & Shioda, T. Technical adequacy of bisulfite sequencing and pyrosequencing for detection of mitochondrial DNA methylation: Sources and avoidance of false-positive detection. *PLoS One* **13**, 1–19 (2018).

27. Bicci, I., Calabrese, C., Golder, Z. J., Gomez-Duran, A. & Chinnery, P. F. Single-molecule mitochondrial DNA sequencing shows no evidence of CpG methylation in human cells and tissues. *Nucleic Acids Res.* **49**, 12757–12768 (2021).

28. Mechta, M., Ingerslev, L. R., Fabre, O., Picard, M. & Barrès, R. Evidence suggesting absence of mitochondrial DNA methylation. *Front. Genet.* **8**, 1–9 (2017).

29. Shao, Z., Han, Y. & Zhou, D. Optimized bisulfite sequencing analysis reveals the lack of 5-methylcytosine in mammalian mitochondrial DNA. *BMC Genomics* **24**, 1–18 (2023).

30. Kinde, B., Gabel, H. W., Gilbert, C. S., Griffith, E. C. & Greenberg, M. E. Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6800–6806 (2015).

31. He, Y. & Ecker, J. R. Non-CG Methylation in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **16**, 55–77 (2015).

32. Madrid, A., Chopra, P. & Alisch, R. S. Species-specific 5 mC and 5 hmC genomic landscapes indicate epigenetic contribution to human brain evolution. *Front. Mol. Neurosci.* **11**, 1–12 (2018).

33. Jeong, H. *et al.* Evolution of DNA methylation in the human brain. *Nat. Commun.* **12**, (2021).

34. de Mendoza, A. *et al.* The emergence of the brain non-CpG methylation system in vertebrates. *Nat. Ecol. Evol.* **5**, 369–378 (2021).

35. Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat Biotechnol* **30**, 1232–1239 (2012).

36. Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5**, 1–9 (2010).

37. Capuano, F., Mülleder, M., Kok, R., Blom, H. J. & Ralser, M. Cytosine DNA methylation is found in drosophila melanogaster but absent in saccharomyces cerevisiae, schizosaccharomyces pombe, and other yeast species. *Anal. Chem.* **86**, 3697–3702 (2014).

38. Gross, J. A. *et al.* Characterizing 5-hydroxymethylcytosine in human prefrontal cortex at single base resolution. *BMC Genomics* 1–14 (2015) doi:10.1186/s12864-015-1875-8.

39. He, B. *et al.* Tissue-specific 5-hydroxymethylcytosine landscape of the human genome. *Nat. Commun.* **12**, 1–12 (2021).

40. Münzel, M. *et al.* Quantification of the sixth DNA base hydroxymethylcytosine in the brain. *Angew. Chemie - Int. Ed.* **49**, 5375–5377 (2010).

41. Boulias, K. & Greer, E. L. Means, mechanisms and consequences of adenine methylation in DNA. *Nat. Rev. Genet.* (2022) doi:10.1038/s41576-022-00456-x.

42. Foox, J. *et al.* The SEQC2 epigenomics quality control (EpiQC) study. *Genome Biol.* **22**, 1–30 (2021).

43. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: Mechanism, function and beyond. *Nat. Rev. Genet.* **18**, 517–534 (2017).

44. Schutsky, E. K. *et al.* Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotechnol.* **36**, 1083–1090 (2018).

45. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE: a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **51**, D629–D630 (2023).

46. Anton, B. P. *et al.* Complete genome sequence of ER2796, a DNA methyltransferase-deficient strain of Escherichia coli K-12. *PLoS One* **10**, 1–22 (2015).

47. Oliveira, P. H. & Fang, G. Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria. *Trends Microbiol.* **29**, 28–40 (2021).

48. Blow, M. J. *et al.* The Epigenomic Landscape of Prokaryotes. *PLoS Genet.* **12**, 1–28 (2016).

49. Sánchez-Romero, M. A. & Casadesús, J. The bacterial epigenome. *Nat. Rev. Microbiol.* **18**, 7–20 (2020).

50. Sun, Z. *et al.* Nondestructive enzymatic deamination enables single-molecule long-read amplicon sequencing for the determination of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Genome Res.* **31**, 291–300 (2021).

51. Cao, B. *et al.* Nick-seq for single-nucleotide resolution genomic maps of DNA modifications and damage. *Nucleic Acids Res.* **48**, 6715–6725 (2020).

52. Liu, Z. J., Martínez Cuesta, S., van Delft, P. & Balasubramanian, S. Sequencing abasic sites in DNA at single-nucleotide resolution. *Nat. Chem.* **11**, 629–637 (2019).

53. Helm, M. & Motorin, Y. Detecting RNA modifications in the epitranscriptome: Predict and validate. *Nat. Rev. Genet.* **18**, 275–291 (2017).

54. Lucas, M. C. *et al.* Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. *Nat. Biotechnol.* **42**, 72–86 (2024).

55. Luo, H., Wei, W., Ye, Z., Zheng, J. & Xu, R. hua. Liquid Biopsy of Methylation Biomarkers in Cell-Free DNA. *Trends Mol. Med.* **27**, 482–500 (2021).

56. Herrgott, G. A. *et al.* Detection of diagnostic and prognostic methylation-based signatures in liquid biopsy specimens from patients with meningiomas. *Nat. Commun.* **14**, 1–19 (2023).

57. Zhang, X. S. *et al.* Maternal cecal microbiota transfer rescues early-life antibiotic-induced enhancement of type 1 diabetes in mice. *Cell Host Microbe* **29**, 1249-1265.e9 (2021).

58. Ando, T. *et al.* A Helicobacter pylori restriction endonuclease-replacing gene, hrgA, is associated with gastric cancer in Asian strains. *Cancer Res.* **62**, 2385–2389 (2002).

59. Zhang, X. S. & Blaser, M. J. DprB facilitates inter- and intragenomic recombination in helicobacter pylori. *J. Bacteriol.* **194**, 3891–3903 (2012).

60. Fan, Y. *et al.* Long-read metagenomics empowers precise tracking of bacterial strains and their genomic changes after fecal microbiota transplantation. *bioRxiv* 2024.09.30.615906 (2024) doi:10.1101/2024.09.30.615906.

61. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Materials and Methods

Mouse Prefrontal Cortex (mPFC)

10-12 mg PFC tissue was extracted from healthy female C57BL6 mice (Jackson Laboratory, ME, USA), aged 14-16 weeks old. Mice were reared in a specific pathogen-free (SPF) vivarium at Rutgers University's School of Public Health animal facility by methods previously described[57]. Each brain was quickly removed and placed in an ice-cold petri dish, and prefrontal cortex (PFC) tissue was carefully extracted by brain dissection. The mPFC samples were flash-frozen individually in dry ice and stored at -80°C for subsequent DNA extraction. DNA isolation was carried out by a Wizard Genomic DNA Purification Kit (Promega Corporation, WI, USA), following the kit recommendations, including the optional RNase treatment. Another mPFC sample (mPFC2) was collected from a different mouse individual using the same protocol.

Human Lymphoblastoid Cell Lines (hLCL)

Human lymphoblastoid cells samples were originally obtained from Coriell Institute for Medical Research (Coriell; cat# GM 12878, NJ, USA) and the methods utilized for LCL samples were as detailed previously[8].

Neisseria gonorrhoeae FA 1090

N. gonorrhoeae FA1090 was obtained from American Type Culture Collection (ATCC, cat# 700825, VA, USA) and the methods utilized for *N. gonorrhoeae* were detailed as previously described[8].

Escherichia coli K-12 MG1655

E. coli K-12 MG1655 sample and methods were detailed using methods detailed previously[8].

Helicobacter pylori JP26

H. pylori strain JP26 was originally isolated from a gastric cancer patient in Japan[58], maintained as detailed previously[59], and stored at -80°C. To obtain a working stock, the *H. pylori* strain was grown at 37°C in 5% CO_2 on Trypticase soy agar (TSA) plates with 5% sheep blood (ThermoFisher Scientific # R01200, MA, USA) for 3 days, then some of the bacterial lawn was transferred to a new plate and incubated at 37°C in 5% CO_2 for 2 days. Finally, bacteria were collected in a 1ml PBS rinse of the plate for DNA extraction with the Qiagen DNeasy Blood and Tissue Kit (Qiagen# 69504, MD, USA).

Whole genome amplification (WGA)

REPLI-g multiple displacement amplification was carried out for *E.coli*, hLCL, mPFC, mPFC2 gDNA samples by the methodology described previously[60]. For the *E. coli,* mPFC2 samples, an additional sample was created for each (designated as 2-round WGA samples) by subjecting the first round of WGA samples to a second round of REPLI-g amplification. The rationale was to compare the two rounds of WGA to ensure DNA amplification was adequate and the WGA samples serve as reliable negative controls.

ONT sequencing

Sequencing was primarily performed with PromethION P2 solo and MinION Mk1b and Mk1c instruments, R10.4.1 flow cells from Oxford Nanopore Technologies as described previously[60]; Ampure bead incubations were extended to 45 minutes improve elution. The rapid kit generated mPFC2 library was sequenced on a MinION Flow Cell (R10.4.1) until the estimated data yield reached 5GB.

Read-level modification analysis

Modified base calling was performed with DORADO (v0.5.3) or DORADO (v0.7) with the genome reference for the species to be analyzed, including mouse (GCF_000001653.27_GRCm39), human (GCA_000001405_GRCh38), *N. gonorrhoeae* FA1090 (RefSeq accession NC_002946.2), *H. pylori* strain JP26 (RefSeq accession NZ_CP023448.1) and *E. coli* K-12 MG1655 (RefSeq accession NC_000913.3). The counts of CpG dinucleotides and total cytosines of human genome were calculated using seqtk (v1.3-r106). Modified base models were selected depending on the modifications of interest (**Supplementary Table 2**).

For the mPFC and hLCL datasets, counts of modified and unmodified bases were estimated using *modkit (v0.2.6) summary* with -f 0.001 with adjusted P_{mod} (from 0.5 to 0.99) applied on --filter-threshold (https://a.storyblok.com/f/196663/x/a461c7bdec/modkit-epi2me-poster-po_1268-en-_v1_08aug2024.pdf). For mPFC2, counts of modified and unmodified bases were estimated using *modkit (v0.2.6) summary* with --no-sampling with adjusted P_{mod} (from 0.5 to 0.99) applied on --filter-threshold. For mPFC and hLCL datasets analyzed with DORADO (v0.7), random subsamples of the datasets (**Supplementary Table 1**) were used, and the counts of modified and unmodified bases were estimated using *modkit (v0.2.6) summary* with --no-sampling with adjusted P_{mod} (from 0.5 to 0.99) applied on --filter-threshold. Modification levels were calculated with total number of modified calls passing (confidence >= threshold) calls among total number of calls extracted.

False discovery rate (FDR) calculation

The FDR corresponding to a specific P_{mod} threshold was estimated by comparing global distribution of the measure obtained from the native DNA sample with that from a WGA (methylation-free) sample.

Specifically, for a given P_{mod} threshold, the FDR is calculated as follows:

$$\text{FDR} = \frac{f_{WGA}(P_{mod}>thres)}{f_{Native}(P_{mod}>thres)}$$

where P_{mod} denotes the classification probability by DORADO software for the base to be analyzed; $f_{WGA}(P_{mod}>\text{thres})$ denotes the fraction of methylated sites out of the total counts in WGA with P_{mod} > thres and $f_{native}(P_{mod}>\text{thres})$ denotes the fraction of methylated sites out of the total counts in native with P_{mod} > thres. Threshold of P_{mod} from 0.5 to 0.99 was applied. FDR was capped at 1 to keep it conceptually meaningful.

Simulation of 5mC level of different magnitude orders on mPFC sample

To create 5mC levels of different magnitude orders, native mPFC and mouse WGA samples were randomly subsampled and mixed to create 5mC levels of different magnitude orders. Briefly, reads were randomly subsampled from two samples using samtools[61] (v1.17) to create mock samples with 10M reads in total for each. Read counts from two samples were normalized based on the average read lengths to ensure the ratio of yields within the mock samples. Fraction of all modification calls were performed with *modkit (v0.2.6) summary* -f 0.2 with adjusted P_{mod} (from 0.5 to 0.99) applied on --filter-threshold.

CpG and CpH sites analyses on mPFC and hLCL samples

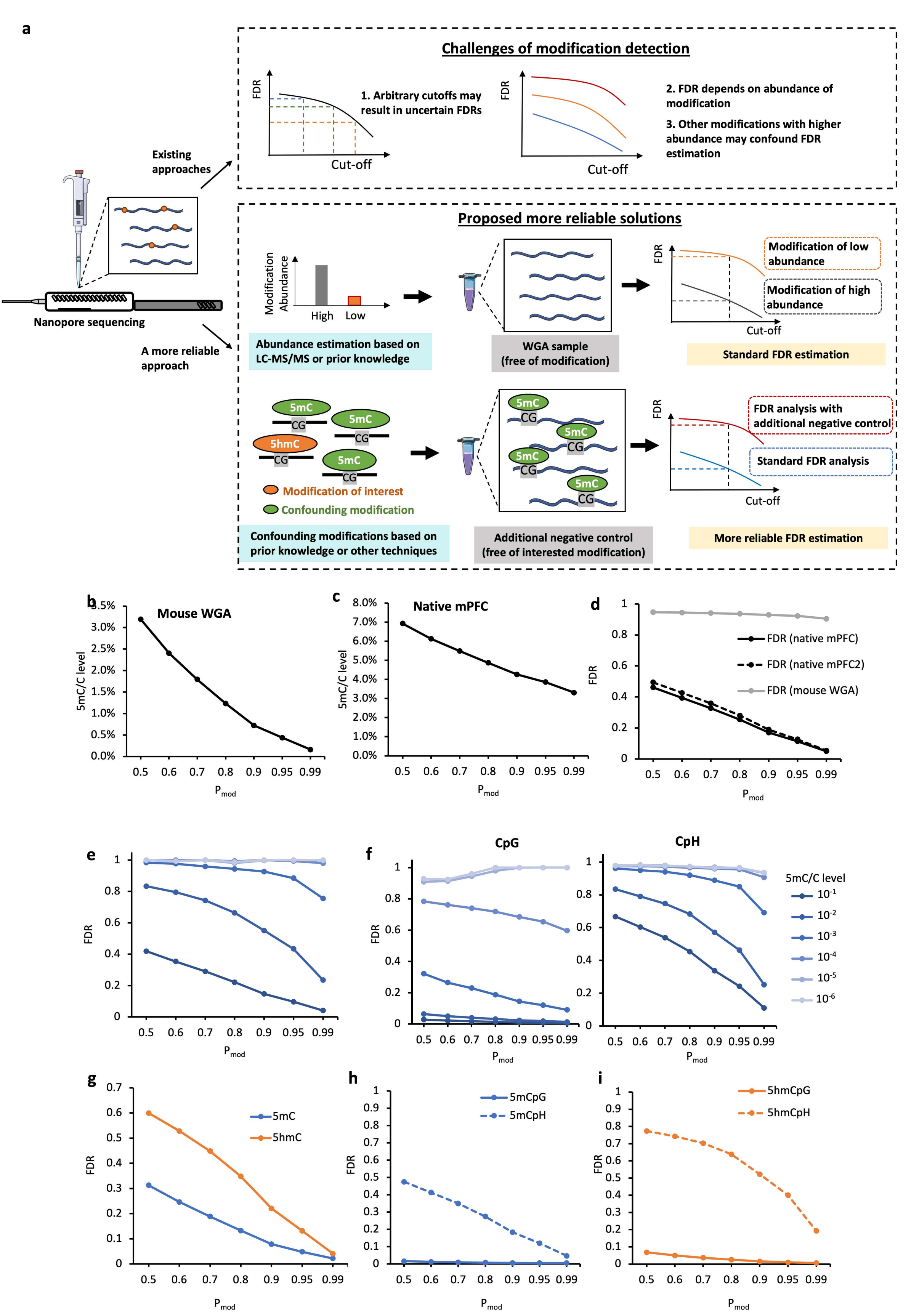
CpG and CpH sites on mouse and human genomes were retrieved using modkit with *modkit motif-bed reference.fasta CG/CH 0*. The pass-threshold was estimated with only base modification probabilities that were aligned to positions overlapping intervals in the CpG or CpH sites using *modkit --include-bed*. The CpG and CpH modification levels and FDR analysis were performed as the methods described above.

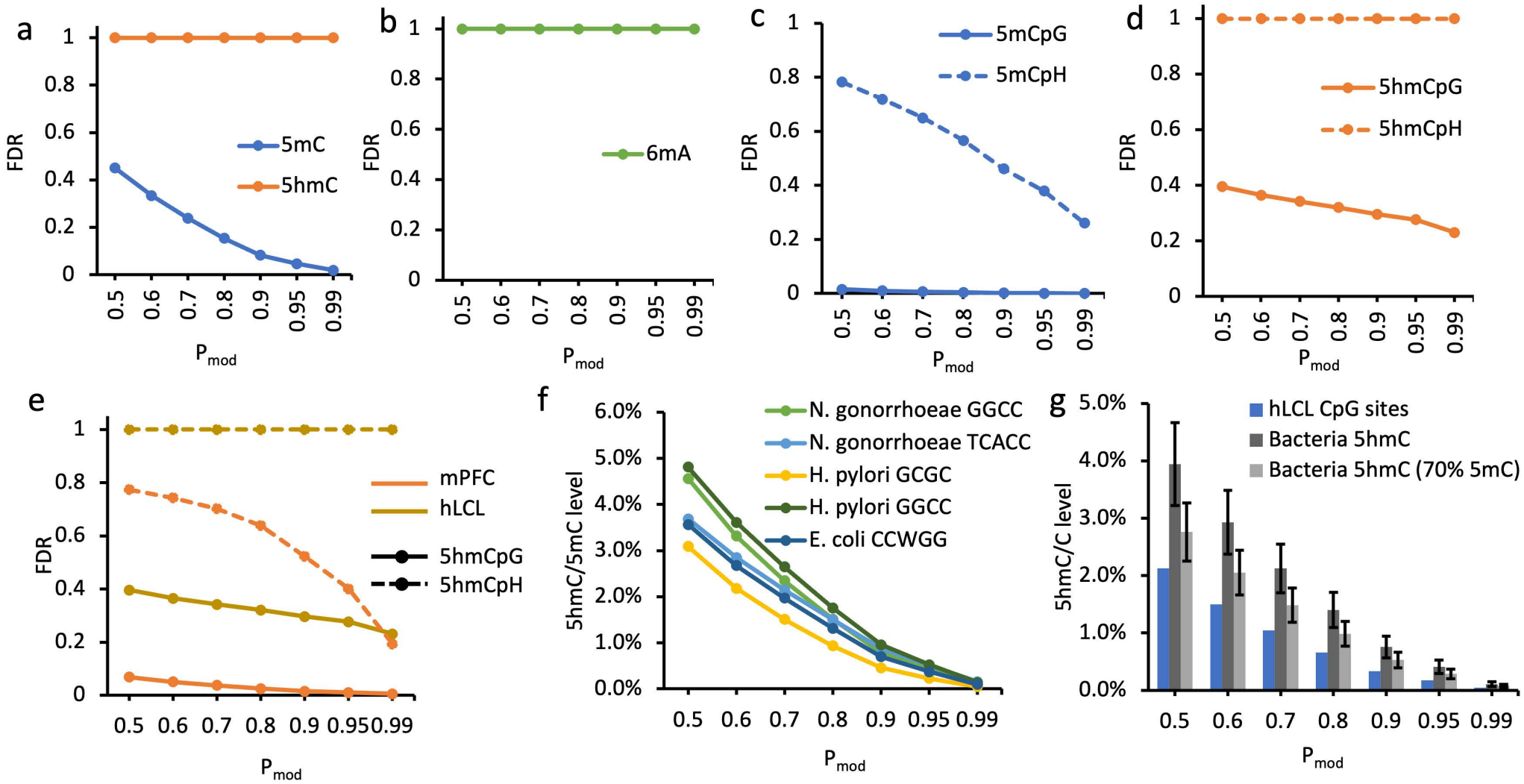
5hmC analysis across bacteria 5mC motifs

Cytosine sites within 5mC motif on bacteria genomes were retrieved using modkit *(v0.2.6)* with *modkit motif-bed* function. The motifs included GGCC and TCACC in *N. gonorrhoeae*, GCGC and GGCC in *H. pylori* and CCWGG in *E. coli*. For each motif, the counts of modified and unmodified bases from native and WGA samples were estimated using *modkit (v0.2.6) summary* with --no-sampling with adjusted P_{mod} (from 0.5 to 0.99) applied on *--filter-threshold*. The mean and standard deviation of 5hmC/C levels in native bacterial samples were then calculated across the five motifs.

Ethics declarations

All animal procedures were conducted following protocols approved by the Rutgers University Institutional Animal Care and Use Committee (IACUC; protocol numbers 201900013 and 201900032).


a
Challenges of modification detection
FDR
Cut-off
1. Arbitrary cutoffs may result in uncertain FDRs
2. FDR depends on abundance of modification
3. Other modifications with higher abundance may confound FDR estimation
Existing approaches
Nanopore sequencing
A more reliable approach
Proposed more reliable solutions
Modification Abundance
High
Low
Abundance estimation based on LC-MS/MS or prior knowledge
WGA sample (free of modification)
Modification of low abundance
Modification of high abundance
Standard FDR estimation
5mC
5hmC
CG
Modification of interest
Confounding modification
Confounding modifications based on prior knowledge or other techniques
Additional negative control (free of interested modification)
FDR analysis with additional negative control
Standard FDR analysis
More reliable FDR estimation
b
Mouse WGA
5mC/C level
P mod
c
Native mPFC
d
FDR (native mPFC)
FDR (native mPFC2)
FDR (mouse WGA)
e
f
CpG
CpH
5mC/C level
10-1
10-2
10-3
10-4
10-5
10-6
g
5mC
5hmC
h
5mCpG
5mCpH
i
5hmCpG
5hmCpH


a
FDR
5mC
5hmC
P mod
b
FDR
6mA
P mod
c
FDR
5mCpG
5mCpH
P mod
d
FDR
5hmCpG
5hmCpH
P mod
e
FDR
mPFC
hLCL
5hmCpG
5hmCpH
P mod
f
5hmC/5mC level
N. gonorrhoeae GGCC
N. gonorrhoeae TCACC
H. pylori GCGC
H. pylori GGCC
E. coli CCWGG
P mod
g
5hmC/C level
hLCL CpG sites
Bacteria 5hmC
Bacteria 5hmC (70% 5mC)
P mod