Published in final edited form as:

Nat Rev Genet. 2023 June ; 24(6): 363–381. doi:10.1038/s41576-022-00559-5.

Navigating the pitfalls of mapping DNA and RNA modifications

Yimeng Kong[1], **Edward A. Mead**[1], **Gang Fang**[1,†]

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

Abstract

Chemical modifications to nucleic acids occur across the Kingdoms of life and carry important regulatory information. Reliable high-resolution mapping of these modifications is the foundation of functional and mechanistic studies. However, mapping technologies may have limitations that sometimes lead to inconsistent results. Some of these limitations are technical in nature and specific to certain types of technology. Here, however, we focus on common (yet not always widely-recognized) pitfalls that are shared among frequently-used mapping technologies and discuss strategies to help technology developers and users to mitigate their effects. While this review is focused primarily on DNA modifications, the pitfalls and navigation strategies we discussed are also applicable to the mapping of RNA modifications.

Introduction

The enzymatic deposition of covalent chemical modifications on nucleic acids mediate versatile and dynamic regulation of both DNA and RNA across the Kingdoms of life[1–4], although the type and abundance of modifications varies among organisms. In mammalian genomes, 5-methylcytosine (5mC) is the most abundant form of DNA modification and, along with its less abundant derivatives that result from its active demethylation (5-hydroxymethylcytosine (5hmC), 5-carboxylcytosine (5caC), 5-formylcytosine (5fC)), it has important roles in development and human diseases[5–7]. In addition to 5mC, bacteria have two other forms of DNA methylation, N6-methyldeoxyadenosine (6mA) and N4-methylcytosine (4mC), all of which have important functions in prokaryotic restriction-modification systems **[G]** and cellular regulation[4,8–12]. Compared to DNA modifications, RNA modifications have greater diversity, with more than 170 distinct forms identified to date[13–22]. The functions of most RNA modifications have yet to be determined, but several, including N6-methyladenosine (m6A) and C5-methylcytidine (m5C), have been shown to have important roles in multiple biological processes in health and diseases[3,14,23].

Although liquid chromatography with tandem mass spectrometry (LC-MS/MS) and antibody-based dot blotting are frequently used as quantification methods (Table 1), reliable

† gang.fang@mssm.edu .

Author contributions

All authors researched data for the article, made substantial contributions to discussions of the content and reviewed and edited the manuscript before submission. Y.K. and G.F. wrote the article.

Competing interests

The authors declare no competing interests.

high-resolution mapping of modified sites is critical for understanding the mechanisms and determining their biological functions[24,25]. Many mapping methods have been developed based on either short-read next-generation sequencing (NGS) technologies or long-read sequencing (LRS) platforms (that is, single-molecule real time (SMRT)[26,27] and nanopore sequencing[28–31]) (Table 1). NGS-based methods require pre-treatment or pre-labelling of the nucleic acid with antibodies, chemicals or enzymes before sequencing to enable modified and unmodified bases to be distinguished (Fig. 1a), whereas LRS-based methods can directly detect modified bases (Fig. 1b).

However, mapping technologies have limitations that reduce their specificity **[G]**, sensitivity **[G]**, or general applicability. These limitations can result in false positive **[G]** (FP) or false negative **[G]** (FN) modification calls and increase the false positive rate **[G]** (FPR) or false negative rate **[G]** (FNR), which are closely related to specificity (specificity=1-FPR) and sensitivity (sensitivity=1-FNR). Some of these limitations are shared among technologies. For example, methods that rely on chemical or enzymatic conversion of non-modified bases are generally prone to generate false positive calls[32,33]: the conversion process is not 100% efficient and non-converted, non-modified bases will be called as modified bases. Despite the promise of LRS-based approaches for direct mapping of a diversity of DNA and RNA modifications[8,13–19,34–36,37], they too have limitations in terms of sensitivity and specificity. These limitations are particularly problematic when mapping modifications with low abundance[38,39] (Box 1) and can confound results and create confusion if they are not resolved or recognized during methods development, evaluation or application. A notable example is the conflicting results obtained for 5mC in mitochondrial DNA (mtDNA) using multiple different techniques. Several studies using whole genome bisulfite sequencing **[G]** (WGBS or BS-seq) reported extensive 5mC methylation in both CpG and non-CpG contexts in mtDNA across multiple species and various conditions[40–42], but other studies suggested that mtDNA methylation levels had been overestimated owing to insufficient bisulfite conversion, read alignment biases and secondary structure of mtDNA[43–46]. A few recent studies performed rigorous method evaluation and reported extremely low or below background levels of 5mC in mtDNA[45–48], calling into question the previously described functional roles of 5mC in mtDNA metabolism[49–51]. Likewise, functionally important 6mA was reported to be present in a few higher eukaryotes[52–60], but other studies have cast doubt on these findings, suggesting that several confounding factors of the 6mA detection technologies might create a significant level of false positive calls, including potential antibody biases in antibody-based DNA immunoprecipitation sequencing **[G]** (DIP-seq) and bacterial contamination in LC-MS/MS[38,61–63]. A recent study took a quantitative metagenomic approach and reported that the vast majority of 6mA in multicellular eukaryotes might be due to bacterial contamination[39]. Similar concerns about false positive calls have been raised for other modifications, including 5hmC in different human cell types[64–66] and 5mC at non-CpG sites[67–71].

In this Review, we discuss issues that commonly arise with widely-used technologies for mapping nucleic acid modifications, focusing on those pitfalls that are less well recognized within the epigenetics and epigenomics research fields. We discuss problems relating to false positive calls, false negative calls, specificity and cross validation among different

technologies. We also suggest strategies to help method developers and users to navigate these pitfalls. Although this review is primarily focused on DNA modifications, the pitfalls and navigation strategies we discuss are also generally applicable to the mapping of RNA modifications. We do not attempt to comprehensively review the biological functions of DNA and RNA modifications or the technical details of specific technologies, for which we refer readers to a number of comprehensive Reviews[8,10,13–20,22,32,34–37].

Sources of false positive mapping calls

For a given genome, because a vast number of nucleotide bases are analyzed, a certain number of false positive calls are hard to avoid due to background noise intrinsic to a technology and multiple hypothesis testing **[G]**. Without necessary adjustments, false positive detections of DNA and RNA modifications can greatly confound data interpretation and downstream functional studies, especially when the modification of interest is of low abundance (Box 1). In this section, we discuss pitfalls that can lead to false positive calls and provide a few strategies for navigation.

Experimental pitfalls

Chemical or enzymatic treatment.—For mapping technologies that involve chemical or enzymatic conversion, the pre-treatment of samples before sequencing can create false positives. For instance, insufficient bisulfite treatment in BS-seq can leave a small percentage of non-modified cytosines unconverted, which are then falsely detected as 5mC in downstream BS-seq analysis[32,33] (Fig. 2a). Over decades of optimization, the rate of incomplete conversion in BS-seq has been dramatically reduced. Morrison *et al.* benchmarked commonly used protocols for BS-seq library preparation protocols and demonstrated incomplete conversion can be as low as 0.5%[72]. This indicates the established BS-seq methods for mapping 5mC in mammalian genomes are generally reliable with relatively low false discovery rates **[G]** (FDR) because CpG methylation is highly abundant[73,74]. However, caution is required when mapping 5mC in regions of the mammalian genome containing less 5mC (such as non-CpG sites or mtDNA) or in species with very low abundance of 5mC in general[67–70] (Box 1). Bisulfite sequencing has also been applied to the quantification and detection of m5C in RNA[75–78]. However, because RNA is single stranded and prone to phosphodiester backbone hydrolysis at high pH, reaction conditions for bisulfite treatment on RNA are less stringent[75,79]. Thus, unmodified cytosines on RNA have a tendency to be left non-converted, adding to false positive calls[75,77,78]. Another example is the detection of the RNA modification pseudouridine (Ψ). RNA labelling with N-cyclohexyl-N'-(2-morpholinoethyl)-carbodiimide metho-p-toluenesulfonate (CMC) can induce reverse transcriptase stalling, which facilitates Ψ detection[80,81]. However, recent studies raised caution that false positives may occur due to non-specific CMC binding to the non-Ψ sites, resulting in the low reproducibility between different studies[82].

Similar caution also applies to restriction enzyme (RE) based mapping methods (RE-seq), which has been used in modification analysis of 6mA[58,83], 5mC[84], 5hmC[85] and others[86]. For example, the 5hmC-sensitive restriction enzyme PvuRts1I has been reported to also cut

at sites of other cytosine modifications when present at a high concentration[85], leading to false positives calls. For RNA modifications, enzymatic methods have been developed to detect m6A with the endoribonuclease MazF recognizing the ACA motif[87,88]. However, a recent study used modification-free RNA from *in vitro* transcription (IVT) as negative controls and assessed the specificity of MazF, highlighting its bias on uncharacterized motifs and RNA secondary structures[89]. Without solid calibration, the false positives caused by the non-specific activity of the enzyme can result in unreliable conclusions of m6A deposition and dynamics.

Antibody bias.—For DIP-seq and RNA immunoprecipitation sequencing **[G]** (RIP-seq), the non-specificity of antibodies is a source of confounding factors that can result in systematic false positives (Fig. 2b). Although commonly-used antibodies typically have hundreds (or even thousands) of fold higher affinity to the intended modification of interest than to unmodified bases, they can still have significant non-specific binding activities across the billions of bases in a large genome or transcriptome. For example, two studies reported that many antibodies for detecting DNA modifications (5mC, 5hmC, 5caC, 5fC and 6mA) tend to bind short tandem repeats due to the intrinsic and non-specific affinity of IgG-based antibodies, which can result in false positive peaks in DIP-seq analysis[61,62]. Similar caution for false positives is needed for antibody-based RIP-seq analysis, which is widely used in the mapping of N1-methyladenosine (m1A)[90–93], m5C[77,94,95] and m6A[93,96–99] and several other RNA modifications[15,18,37,100–103]. Zhang *et al.* reported thousands of m6A-irrelevant peaks in unmethylated adenine-rich regions by RIP-seq with modification-free RNA from IVT[89]. Without proper negative controls, those false positive peaks can be seemingly reproducible between independent RNA samples or studies[89]. For N4-acetylcytidine (ac4C) RNA modification detection on human messenger RNA (mRNA), while an antibody-based method[104] suggested broad genome-wide distribution, chemical conversion methods reported very low abundance[105,106]. This inconsistency raised a concern of antibody artifacts for this rare modification. Off-target antibody binding is not expected to have a big impact for DNA or RNA modifications with high abundance, but can result in a high FDR when the modification of interest has very low abundance (Box 1). For example, because 6mA has been reported to have very low levels (as low as ~0.0001% or undetectable) in most multicellular eukaryotes[53,107–111], DIP-seq based 6mA calling is expected to have a high FDR. Indeed, it has been reported that the non-specific binding of anti-6mA antibodies to repetitive DNA can account for 50–99% of binding among the 'enriched' regions by DIP-seq[61], highlighting the importance of using matched IgG controls. On RNA, the antibody-dependent artifacts were reported to create false positive calls of m1A in 5′ untranslated regions (UTRs) because anti-m1A antibody can also bind mRNA cap structures[112].

Contamination.—False positives can arise from different sources of contamination. For example, DNA extracted using standard protocols can contain residual RNAs[113], which can confound NGS-based DNA sequencing and lead to false positive peaks in DIP-seq experiments, as shown in studies using anti-6mA[62] and anti-5mC antibodies[114,115]. RNA contamination is remarkable for 6mA analysis because mRNA contains abundant m6A and also has a high affinity to anti-6mA antibodies[62] (Fig. 2c). Another source of

contamination comes from exogenous contamination. For example, 5mC on nuclear DNA can confound studies that aim to specifically examine 5mC on mtDNA[46,47]. Another example is, residual levels of bacterial DNA contamination can also confound eukaryotic 6mA studies[39,61,62]. Although the high levels of 6mA in bacterial genomic DNA (gDNA) do not directly contribute to 6mA mapping in a eukaryotic genome of interest, they can lead to overestimation of the global 6mA:A ratio, which can then confound analytical design and data interpretation (see *Cross-validating modification calls*). Similar cautions also apply to the RNA modification analysis. Because DNA is more stable than RNA, it is crucial to exclude DNA contamination at the beginning, otherwise the contaminated DNA, which may also be recognized by the same antibodies[62] or chemicals[116], can introduce false positives.

Sequencing depth.—False positives also arise during library preparation and sequencing. NGS-based approaches can introduce biases during PCR amplification[117,118] that result in non-random sequencing depth throughout the genome and lead to potential false positive calls in methods that rely on accurate read mapping[58,83], such as DIP-seq[61], RIP-seq[119], BS-seq[67] and RE-based[58,83] methods. When a modification of interest has low abundance, the FDR is expected to be high (Box 1). Although LRS of native DNA can avoid this PCR bias, some existing tools use arbitrary cut-offs that depend on sequencing depth to call modification events. For example, in SMRT sequencing, inter-pulse duration **[G]** (IPD) ratios have large variance at low sequencing depth[11,107,120,121], and modification quality value **[G]** (QV, -log10 transformed p value) tends to get overestimated at high sequencing depth[39,121] (Fig. 3a). In DIP-seq data, false positive peaks tend to be called at genomic regions with systematic bias (such as repetitive sequences)[61], especially at high sequencing depth. Without rigorous FDR correction, an arbitrary cut-off on these metrics can cause systematic false positives. Similarly, the analysis of RIP-seq should also pay attention to the bias from sequencing depth and evaluate FDR.

Analytical pitfalls

Sequence variation.—Sequence variation in organisms with polyploid genomes and heteroplasmic **[G]** mtDNA poses a challenge to standard IPD-based analysis of SMRT sequencing data, which compares IPDs observed in the sequenced native DNA to their expected IPD values estimated from a reference genome using an *in silico* model[11,27,120,121]. Specifically, the expected IPD value for a base of interest is estimated based on the genome context around the base (from −10bp to +4bp, with which the SMRT DNA polymerase physically interacts) according to the input reference genome[26,27], and does not capture sequence variation arising from the presence of multiple copies of the genome or mtDNA. Thus, the expected IPD value will be incorrect for any DNA molecules containing sequence variants (for example, a single nucleotide polymorphism, SNP), causing inaccurate IPD deviations not only for the nucleotide itself but also its flanking bases[39,121] (Fig. 3b), leading to systematic false positives. The SMRT platform has also been tailored for direct RNA sequencing **[G]** by switching the standard DNA polymerase to a reverse transcriptase[122]. Although this approach has not been widely used, a slower kinetics (longer IPD) was observed for m6A containing template than a negative control sample. In this context, sequence variations such as RNA polymorphisms[123], is also expected to create false

positives if *in silico* model based on the RNA sequences is developed for RNA modification detection in the future.

Base-calling errors in nanopore sequencing have been utilized to detect RNA modifications building on the general concept that increased error rates during base calling reflects the existence of certain RNA modifications[124–129]. To exclude sequencing errors that are independent from RNA modifications (for example, RNA polymorphisms[123] or repetitive sequence errors[130,131]), these methods usually need matched controls free of RNA modification using IVT RNA. For RNA modifications with relatively high abundance and well characterized motifs, such as m6A in motif RRACH (R=A/G; H=A/C/U)[20,132] or DRACH (D=A/G/U)[21,89,97], in mRNAs, false positives can be minimized by a focused modification analysis at the motif sites. However, for modifications with low abundance identification and a lack of well-characterized motifs, additional care to evaluate false positive rate and FDR is needed (Box 1).

Confounding modifications.—Even when a technology can reliably detect a nucleic acid modification of interest in ground truth datasets, it does not necessarily mean that all the detected events in a real application are specifically this type of modification (Fig. 3c). For example, bisulfite conversion does not distinguish between 5mC and 5hmC, meaning the 'candidates' detected by BS-seq will be a mixture of these two modifications[133,134]. The DNA 6mA events captured by 6mA antibody-based DIP-seq also include reads of RNA origin, which could be cross-contaminated by highly abundant RNA m6A modifications[62]. The detection of 6mA and 4mC by SMRT-seq are usually based on a high IPD ratio signal on the base to be analyzed[27]. However, high IPD ratios can also be caused by 5mC or by marks caused by DNA damages (such as 8-oxoguanine (8-oxoG) and 8-oxo-7,8-dihydroadenine (8-oxoA))[27,135].

The great diversity of RNA modifications poses significantly more challenges to mapping technologies in terms of distinguishing between different forms of RNA modifications. Among all RNA species, transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) are most abundant, and have a high diversity of chemical modifications[17,136–140]. Although mRNAs have multiple forms of modification such as m1A, m5C and m5hC, they are usually of low abundance compared to tRNA and rRNA. m6A, the most abundant form of mRNA modification accounts for 0.1%–0.6% of all adenosines in mammalian mRNA[13,17,141,142]. Some of the other forms of modifications have much lower abundance[13,15,37,82], which pose great challenges for reliable mapping and highlight the need for methods evaluation across a wide range of abundances of a modification of interest (Box 1). For example, the m6A antibodies do not discriminate m6A and N6,2′-O-dimethyl-adenosine (m6Am). Earlier methods distinguished the two RNA modifications based on the unique deposition of m6Am specifically at the first encoded nucleotide adjacent to the 5' cap in transcripts[21,143]. To overcome the mis-annotation between these two structurally similar modifications especially around 5′-UTR, a recent study used an *in vitro* demethylation reaction to selectively remove m6Am before antibody binding to more reliably detect m6A events[144]. New methods are needed to better separate the two structurally similar modifications.

Secondary structure.—LRS can better address read mapping at certain repetitive regions than short read sequencing, however DNA secondary structure **[G]** can affect SMRT DNA polymerase kinetics leading to increased IPD values independent of DNA modifications[107,145] (Fig. 3d). For example, increased IPD ratios have been reported to exist for all four DNA bases in young L1 elements **[G]**, a common repetitive DNA sequences in human, even in the whole genome amplified **[G]** (WGA, modification-free) samples, suggesting possible confounding secondary structures that influence the modification analysis in SMRT sequencing[107]. Nanopore sequencing is expected to be less affected by these factors because it does not rely on polymerase dynamics while the single-strand DNA (ssDNA) or RNA are ratcheted through the nanopore[8,28–31]. However, future research is needed to evaluate whether complex DNA structures such as G-quadraplex can be ratcheted through the commonly used nanopore. As discussed above, DIP-seq can also be confounded by DNA secondary structure in repetitive regions due to non-specific binding of antibody to DNA. In addition, single-stranded RNA is prone to form complex RNA secondary structures[146–150], which can affect the affinity of antibodies, or the reactivity of chemicals or enzymes, in NGS-based approaches for detecting RNA modifications[112,151], and can lead to systematic false positive or false negative calls. For example, RNA secondary structure has also been shown to affect the cleavage efficiency of RNA endoribonuclease MazF used for m6A mapping[89]. RNA secondary structure can also influence the detection of Ψ modifications because it can induce natural reverse transcriptase stalling, confounding the real termination induced by CMC-Ψ labelling[82,152].

Variation in RNA abundance.—The abundance of different types of RNAs varies greatly, with rRNA, tRNA and mRNA accounting for 80–90%, 10–15% and 3–7% of the total RNA, respectively[153]. In addition, mRNA transcripts from different genes or different isoforms of the same genes can have a wide range of expression levels[154,155]. These abundance variations can influence the reliability of RNA modification detection due to differences in statistical power. For example, independent methods for mapping m5C usually have a high reproducibility for tRNA and rRNA, but not for other RNA species[156]. Another study reported that FDR for mapping Ψ on rRNA ranges from ~5% for highly expressed genes but increases to ~12.5% for lowly expressed genes[80].

Minimizing false positive mapping calls

To help technology developers and users navigate the pitfalls reviewed above, we review and discuss a few strategies for rigorous experimental design and methods evaluation.

Matched negative controls

To identify and rule out various confounding factors, it is important to use matched negative control samples to evaluate false positives and FDR. These are critical for both commonly-used technologies and new methods development. For DIP-seq and RIP-seq, an IgG immunoprecipitated control, rather than input DNA or RNA, can help adjust for antibody non-specificity[61]. This is because non-specific IgG control can capture off-target binding due to the intrinsic affinity of IgG for short unmodified DNA repeats and reduce false positives[61] (Fig. 4a). For biochemical and enzyme-based modifications, WGA DNA

and IVT RNA can help estimate non-conversion rates and FDR[45,67,89,157]. For example, unmethylated DNA, either WGA or spike in controls can help to adjust the bias in the BS-seq[72] (incomplete cytosine conversion by sodium bisulfite), *i.e.* BS incubation length and temperatures during the sample preparation. This is particular important for the species with low abundance of 5mC or studies to detect 5mC at non-CpG sites[67–70]. Similarly, for SMRT sequencing and nanopore sequencing, it is critical to evaluate existing and new methods with WGA[39,107,158] (Fig. 4b). It is also important to match the sequencing throughput of negative controls with samples of interest to avoid biases from sequencing depth and the statistical power for modification calling[107]. However, it should be noted that a limitation of WGA is that it does not capture the influences from confounding DNA modifications at the base under investigation or flanking bases. For RNA modification mapping, the negative controls, e.g. IVT RNA, also needs sophisticated design, in terms of sequence and expression diversity[89,159,160]. Another limitation of IVT RNAs is that they are completely modification free, hence they do not capture the confounding effect between different types of RNA modifications, and may not reliably calibrate the background noise that reflects real applications to native RNAs. To better support the specificity of modification detection, it is helpful to examine both wild type cells and cells that have genetic (knockout, knockdown, and/or over expression) or chemical perturbations affecting positive or negative regulators of the modification of interest (such as methyltransferases, demethylases, deaminases, synthases, indirect regulators, among others). Specifically, if the number of called modification events differs significantly upon genetic/chemical perturbation, it strengthens the argument for specificity of modification calling.

FDR evaluation

An important use of negative controls is to help estimate FDR associated with a set of called modification events. Whereas an FPR represents the probability that a false positive call is made among all the sites tested, an FDR informs on the probability of false positive calls among the detected modification events[39,107,157,158]. FDR, in contrast to FPR, depends on the true abundance of a modification of interest in a specific sample. For modifications with low abundance (such as 6mA in most eukaryotes[62], 5hmC in most mammalian cell types[64–66], and 5mC at non-CpG sites in most mammalian cell types[67–70]), the use of FDR can effectively capture the reliability of called modification events considering the very low number of true positive events existing in a sample (Box 1). For each cutoff on p-value (or fold change or other metric), an FDR can be calculated by comparing the number of detected modification events in a sample of interest versus a negative control[107]. The choice of a threshold to report detected modification events can then be reported along with an FDR[39,107,157,158] (Fig. 4c). This strategy is more reliable than the use of an arbitrary cutoff on p-value that might seem to be ‘consistent’ with LC-MS/MS estimation, because DNA modification mapping and LC-MS/MS are not always directly comparable (see *Cross-validating modification calls*). For RNA modifications with relatively high abundance and well characterized motifs, such as m6A (motif RRACH or DRACH), false positives can be minimized by aggregated modification analysis at the motif sites. However, for modifications with low abundance and a lack of well-characterized motifs, extra caution is needed to avoid false positives (Box 1). A helpful approach is to combine the use of FDR evaluation and the examination of mutant cells with knockout of positive (e.g.

RNA methyltransferase[3,20,132,142]) or negative regulators (e.g. RNA demethylase[3,20,132,142]) of a certain RNA modification as critical controls.

Quantification model

As an alternative to FDR, a quantification model can be used to estimate the abundance of a modification of interest. This type of method (recently used to evaluate 5mC in mtDNA[45] and 6mA in eukaryotes[39]) defines a set of features that captures the modified events in a genomic sample, and then trains a machine learning model across a number of positive and negative controls containing the modification at a wide range of abundance (Fig. 4d). The advantage of this strategy is that it can account for background noise within model training; because the model is trained to specifically differentiate between a large number of samples with different levels of modification, false positives are automatically considered as background noise[39,45]. For this strategy, it is important to use independent cross validation (such as LC-MS/MS) along with biological and technical replicates to ensure reliable evaluation and to avoid possible batch effects.

Sources of false negative mapping calls

While any technology is expected to miss a small percentage of modification events due to random chance, herein we focus on the discussion of false negative detections due to systematic biases. Sometimes this pitfall arises because certain existing technologies are powerful for detecting only a subset of DNA or RNA modifications. In other cases, false negative detections are made if there is systematic bias in sample preparation or data analyses.

Technology-intrinsic biases

Some techniques are more effective for detecting certain DNA and RNA modifications than others because of their intrinsic characteristics. For example, SMRT sequencing has stronger signal-to-noise ratios **[G]** for detecting 6mA and 4mC events than 5mC and 5hmC[27,121,161]. The biophysical explanation is that 6mA and 4mC favor a *cis* conformation (methyl group facing into the DNA helix)[162–164], which creates greater steric hindrance for DNA polymerase translocation, significantly influencing the IPD ratio, which is mostly 5–7 for 6mA events and mostly 2–4 for the 4mC events[27,39,107,121,164]. By comparison, the *trans* confirmation favored by 5mC and 5hmC (with the methyl group facing outside the DNA helix) induces moderate IPD changes across multiple flanking bases around 5mC or 5hmC events[27,164,165] (Fig. 5a). Thus, the ability of SMRT sequencing to reliably detect 6mA and 4mC has empowered the rapidly developing field of bacterial epigenomics, it is not effective for the discovery of 5mC in bacterial genomes[27,121,161]. By contrast, nanopore sequencing has a better signal-to-noise ratio for 5mC than 6mA and 4mC[158,166,167], while the ratios for 5mC events vary dramatically across different sequence contexts[158] (Fig. 5b), although active development is underway to improve 5mC calling across more diverse sequence contexts[158,168]. Encouragingly, these technology-specific limitations can be addressed by biochemically pre-treating the DNA. For example, the conversion of 5mC and 5hmC to 5fC and 5caC[164,165] by Ten-Eleven Translocation (TET) enzymes (e.g. TET-assisted pyridine borane sequencing (TAPS[169,170])) (Fig. 5a), or selective labeling of 5mC or 5hmC

with specific chemicals (e.g. Biotin-S-S-5-hydroxymethyl-cytosine[169–171]), was shown to significantly enhance the signal-to-noise ratio in SMRT sequencing. For RNA sequencing, although the first attempt on SMRT platform had observed longer IPD ratio association with m6A by switching the standard DNA polymerase to a reverse transcriptase[122], it has not been widely used for RNA modification detection. In contrast, the nanopore platform has been used increasingly for direct RNA sequencing. Several methods have been developed for the direct detection of various RNA modifications[172,173] either based on sequencing error induced by certain RNA modifications[124–128,174] or changes in electrical signal (ion current or dwell time) that reflect the existence of RNA modification events[175–178].

Experimental biases

Systematic false negative detection can also occur when a sample is inappropriately treated during chemical or enzymatic modification before sequencing. For example, prolonged bisulfite treatment can lead to conversion of 5mC to U (read as T) and increased DNA degradation[32,33], creating subsequent false negative 5mC events (Fig. 5c). For RNA m5C detection, in addition to biases from chemical or enzymatic treatment[75,77,78], the tendency of RNA to naturally degrade[179,180] may further contribute to false negative calls due to lost material and missing reads from low abundant transcripts. A similar concern for false negatives was raised in the detection of m1A RNA modification using the Dimroth reaction, which converts m1A to m6A to eliminate mis-incorporation and truncation patterns[91,181]. The severe alkaline condition in this process can cause significant RNA degradation and result in underestimation of m1A on RNA[90,91,181,182]. Restriction enzyme-based sequencing methods for detecting certain DNA modifications might also be prone to false negatives (Fig. 5d). For instance, the G6mATC-sensitive restriction enzyme DpnI has been reported to recognize the 6mA site in fully methylated C6mATC/G6mATG contexts after 12 hour incubation (in addition to G6mATC, the canonical DpnI motif), but this rarely happens within 30 minutes reaction time[83]. Hence, false negatives may occur at these methylation motifs **[G]** if the incubation time with DpnI is insufficient[83,183,184]. Similarly, other restriction enzymes may require careful analysis to minimize false negatives[9,84,85,185].

Computational biases

A common strategy for mapping DNA modifications using SMRT and nanopore sequencing involves the training of a machine learning model across modified and unmodified bases (Fig. 5b). A series of methods have demonstrated the effective detection of 5mC by nanopore sequencing[166,167,186–190]. However, the models developed in these works were specifically based on training datasets with 5mC at CpG sites. Because of the diverse signal/signatures of 5mC across different sequence contexts in nanopore sequencing[158,168,178], these models specifically detect 5mCpG or a limited diversity of sequence contexts rather than 5mC generally (Fig. 5e). This largely explains why nanopore sequencing studies had not been used for *de novo* 5mC discovery in bacterial genomes, in which 5mC occurs across diverse sequence motifs[8,158]. Recognizing this limitation, a training dataset was generated from an assortment of bacterial species and a method developed that can systematically *de novo* detect these three primary forms of DNA methylation in bacterial genomes[158]. A couple of recent works specifically designed training datasets with 5mC at non-CpG sites, aimed at detecting 5mC in plants, where most 5mC events occur at non-CpG sites[168,191,192].

For SMRT-seq, two recent studies reported new methods for 5mC detection using a convolutional neural network. Because the training datasets are 5mC events specifically from CpG sites[193,194], it is unclear yet if they are generally applicable for detecting 5mC in broader sequence contexts.

Similarly, for RNA modification detection, machine learning models that can directly detect modifications without control samples[177] have broader applicability because they are not limited to modifications that induce base-calling errors, but a general pitfall is that the training datasets may lack diversity and may not be broadly applicable for the detection of a certain form of modification. For example, the tool Nanom6A was trained with direct RNA-Seq data that contains 130 'RRACH' motif sites for 6mA calling[176,177]. Considering the drastic variation of nanopore signal across different sequence contexts of the same modification type[158,186,187], Nanom6A might be reliable specifically for the detection of m6A at 'RRACH' motif sites. More generally, machine learning models trained with datasets with a limited diversity of RNA contexts can also result in both false positive and negative calls[195]. It is encouraging that several recent studies have reported improved performance with machine learning models using extended training data covering more diverse combinations or positive training data extracted from *in vivo* generated datasets[177].

Minimizing false negative mapping calls

Experimental workflow

For experiments that need prior antibody or chemical treatment, it is important to perform appropriate pre-treatments or labelling. Although commercial antibodies and chemicals are thoroughly evaluated with benchmark controls, careful biological and technical replicates are essential to minimize the risk of mistreatment of the real samples. One practical strategy is to include well-characterized positive controls as spike-ins along with DNA or RNA samples of interest[157,169].

Model training

To ensure the broad applicability of modification detection by LRS-based approaches and to mitigate false negatives, more sophisticated training data is required that better represents the modifications of interest in a range of motifs and genome contexts[158,191,192]. The design of model training data also needs to consider practical application. For example, DNA with a primary nucleotide completely replaced by a modified base (for example, all cytosines replaced by 5mC via PCR) serve as great positive controls that represent highly diverse sequence contexts, and have been used in the development of enzyme-based mapping of DNA modifications[157,183,184].

However, the close distances between the modified bases will result in increased and composite signal changes in SMRT and nanopore platforms that do not represent real-world applications in which the modified bases are farther apart from each other. For example, although the 169-kb modified T4 phage DNA with all cytosines replaced by 5hmC served as an excellent positive control in the development of ACE-seq[157], its use in training SMRT or nanopore models should be avoided because the densely–packed 5hmC events does not

represent the real distribution of 5hmC in mammalian genomes and therefore generates an over-estimated signal-to-noise ratio[34,101,157,196,197].

Recent studies have started to use large scale synthesis of modified bases in oligos with random sequence contexts[160]. Although this approach aims to directly address the generally applicability of model training, the importance of modification abundance needs to be considered when using these oligos. A trained model that works reliably on test data with highly abundant modifications can be associated with a high FDR when the model is applied to a real genome in which the modification has much lower abundance (Box 1). Therefore, predictive models trained using synthetic oligos need to be evaluated (especially for FDR) across a wide range of abundance of the modification of interest.

Cross-validating modification calls

While a single technology may have bias, data can be validated by comparing it with calls generated by an independent orthogonal approach[15,37]. For example, BS-seq and nanopore sequencing have been used to cross-validate accurate 5mC profiles[64,134,189]. To cross validate 6mA events across *C. reinhardtii* genome, Fu et al. used both antibody-based DIP-seq and RE-seq (based on 6mA sensitive restriction enzymes) and highlight genomic regions enriched for 6mA events consistent between the two technologies[58,83]. Given the high sensitivity and specificity of LC-MS/MS, it is often used alongside mapping methods as the gold standard for quantifying DNA and RNA modifications[63,108,109,198]. However, various confounding factors mean that not all methods reliably cross-validate each other.

Navigating confounding factors

As a first step in determining appropriate cross-validating techniques, each individual technology should be evaluated for sources and frequency of false positives and false negatives, which should be minimized using matched negative controls and thorough FDR evaluation of the selected cutoff, along with other relevant strategies discussed above. Technologies that depend on the same reagents (for example, antibodies) or materials should be avoided for cross-validation purposes, if possible. For example, dot blotting, DIP-seq and immunohistochemistry staining all rely on antibodies that target the modification of interest, and data analysis will be confounded by shared biases intrinsic to the antibodies: non-specific binding to unmodified bases (for example, in AT rich regions for anti-6mA or anti-m6A antibodies)[61,89,199] (Fig. 2b); inability to distinguish between distinct but similar modification types (for example, DNA 6mA and RNA m6A for anti-6mA antibodies)[62] (Fig. 2c); and the high affinity of IgG to repetitive DNA sequences that tend to form non-canonical DNA secondary structures[61] (Fig. 2b). Thus, it is more reliable to cross-validate results using methods that do not rely on the same reagent. If this is unavoidable, consider using reagents from different sources, for example, antibodies from different brands and vendors[199,200].

Even if two technologies do not use the same reagents or materials, they can sometimes be confounded by factors that affect the results of both technologies. For example, DNA secondary structure can independently cause false positive calls for 5mC and 6mA in both DIP-seq and SMRT-seq. In DIP-seq, the IgG-based 5mC or 6mA antibody has non-specific

affinity to the DNA secondary structures associated with repetitive genomic regions[61] (Fig. 2b); in SMRT-seq, non-specific increases in IPD signals are associated with slower DNA synthesis through genomic regions with complex secondary structures[107,145] (Fig. 3d). Therefore, although certain 5mC or 6mA calls made in a repetitive region may be shared between two seemingly independent technologies, they may still be false positives and an additional technique will be necessary to cross-validate the results.

Similarly, cross validation of results between quantification methods should account for the possibility of confounding owing to exogenous contaminants[38,39,61]. For example, if a gDNA sample is contaminated by exogenous DNA containing a modification of interest, both dot blotting and LC-MS/MS are expected to overestimate its abundance because neither technology distinguishes the source of the modification[39,62] (Fig. 4c). For example, it was recently demonstrated for 6mA in multicellular eukaryotes that even a residual amount of bacterial contamination in the gDNA sample can contribute the vast majority of the detected 6mA[39,62]. Key to this result was a recently developed tool called 6mASCOPE, which took a quantitative metagenomic approach to deconvolve the 6mA results from a genomic DNA sample into different sources, including potential bacterial contamination. Thus, LC/MS-MS and 6mASCOPE together provide more reliable cross-validation of quantification data, which is specific to the genome of interest and robust to contaminants[39].

Finally, sequencing-based 6mA mapping data for many eukaryotic organisms has been widely cross-validated with LC-MS/MS[52–54,56,59,201]. However, this cross-validation approach can have important consequences that are not readily apparent. Sometimes, mapping methods use an arbitrary cut-off to call modifications; for example, IPD ratio and/or QV for SMRT sequencing, or an adjusted p-value for DIP-seq. These cutoffs represent a certain confidence in the set of modification events detected[39,121]. However, the total number of called events will depend on the cut-offs used, which can be very subjective depending on the goal of the study (Fig. 3a; 4c). Importantly, for this reason, these called events are often not directly comparable with absolute quantification of a modification of interest by LC-MS/MS, particularly when partial modifications (that is, a given site is not modified in all cases) are not directly factored into the comparison. Thus, it is usually not reliable to adjust the cutoff of the mapping methods to make the number of called modifications ‘consistent’ with the LC-MS/MS estimation.

Implementing navigation strategies

In practice, different strategies need to be properly integrated depending on the characteristics of a specific form of DNA modification and the goal of a specific application. In this section, we will provide a perspective on how to implement these strategies to map a list of DNA and RNA modifications (Table 2).

Caution for early adaption of emerging technologies.—Most technologies have intrinsic biases that result in false positive and/or false negative calls, especially before comprehensively thorough evaluation has been performed for mapping a certain modification. For example, SMRT-seq was initially designed specifically for mapping prokaryotic DNA methylation (especially 6mA and 4mC), which are not only highly

abundant, but also strongly associated with well-defined motifs[8,11]. Because these two unique characteristics are not shared by most eukaryotes, previous informatics tools for SMRT sequencing were fundamentally not ready yet for mapping 6mA or 4mC in eukaryotes[8,107]. Therefore, premature adaption of SMRT-seq for 6mA mapping in eukaryotes may have confounded a few early studies that reported highly abundant 6mA in eukaryotes[10,38,54,107]. Similarly, nanopore sequencing was initially developed for mapping 5mC specifically at CpG sites in humans[167,189]. Although a few studies adapted it for mapping 5mC and 6mA in bacteria[166,186–188,191], the previous methods were not able to reliably resolve the bacterial methylomes with very diverse methylation motifs[4,9,86]. This challenge was later addressed by development of a new method that can effectively handle the drastic differences in nanopore signal across different sequence contexts even for the same modification type[158]. Without comprehensive evaluation prior to cross validation, early adaptation of emerging technologies that are still under active development can be exciting but comes with risk.

Modifications with high abundance.—If a modification of interest is of high abundance in a genome, mapping technologies are expected to have low FDR in general (Box1). If the modification is also enriched at certain sequence contexts, confirmative motif enrichment analysis along with focused interpretation of individual modification events at motif sites can help enhance the specificity of mapped events[8,39,120,158,161]. This is applicable to the mapping 6mA, 4mC and 5mC in most prokaryotes[8,158], 6mA mapping in certain protozoans[57,58,83,202203], 5mC mapping in mammalian genomes[73,74] as well as mapping the abundant m6A in mammalian mRNAs[20,132]. Several mapping methods have been developed both for NGS[32,33,196] and LRS platforms[120,158,166,167]. In particular, SMRT and nanopore sequencing platforms can support de novo motif discovery from prokaryotes, even with moderate sequencing depth, as discussed in recent studies[8,11,12,158]. However, mapping individual 6mA, 4mC and 5mC events in bacteria is currently challenging using nanopore sequencing, due to the drastic variations of signals across different sequence contexts[158].

Besides the golden standard BS-seq for 5mC mapping in eukaryotes, other methods are seeking to use bisulfite-free treatment to avoid the shortcomings of bisulfite treatment. For example, enzymatic Methyl-seq (EM-seq) uses ten-eleven translocation dioxygenase 2 (TET2) and T4 phage β-glucosyltranferase (T4-βGT) to protect 5mC and 5hmC, and deaminates the unmodified C into U (read as T) by APOBEC3A[204]. Compared to BS-seq, the non-destructive EM-seq has better yield, longer reads, and more evenly distributed genomic coverage when applying to both NGS[72,204] and LRS[205]. In *Arabidopsis*, EM-seq showed more accurate non-CpG estimation due to its lower background noise than BS-seq[206].

Multiple tools for nanopore-seq based 5mC mapping at CpG sites were compared in a study by Liu *et al.*[134], which provide a great roadmap in the choice and integration of different methods for mapping. This study also reported that the discrepancy between 5mC mapping by nanopore sequencing and BS-seq are partially explained by 5hmC events, which motivates the further development of new methods that may help distinguish 5hmC from 5mC by nanopore sequencing[134]. The machine learning models in most existing tools for

nanopore based mapping of 5mC were mainly trained with 5mC events in CpG sites or a limited number of other motifs[166,167,186–190]. Because 5mC events at non-CpG sites usually have much lower abundance than at CpG sites, it is important to use negative controls and perform rigorous FDR evaluation specifically for non-CpG sites (Box 1). Also, given the large variation of signal-to-noise ratio in nanopore sequencing across different sequence contexts, further model training at diverse sequence contexts is necessary to increase accuracy and avoid false negatives. Encouragingly, some recent studies have developed methods for specifically mapping 5mC at non-CpG sites in bacteria and plants[158,168].

Similarly, for SMRT-seq, the current tools for 5mC mapping in mammalian genomes are specifically trained for CpG sites[193,194], and additional methods development is needed for 5mC mapping at more diverse sequence contexts. In addition, because the polymerase kinetics in SMRT sequencing not only depend on chemical modifications to DNA, but also DNA secondary structure[107,145], which tend to cause false positive calls (*e.g.* in genomic regions with complex repeats) (Fig. 4b). Negative controls and rigorous FDR analysis specifically for repetitive regions can help avoid false positives (Fig. 4). It is still unclear yet if DNA secondary structure also affects ion current **[G]** in nanopore sequencing[207]; if so, similar caution may also be helpful for nanopore sequencing based mapping of 5mC at highly repetitive regions.

Modifications with low abundance.—If a modification of interest has low abundance in a genome, great cautions are needed for false positives (Box 1). The abundance of 6mA in multicellular eukaryotes is mostly very low, from 0.1% to 0.0001%, or undetectable[53,107–111,208]. LC-MS/MS or dot blotting needs to be interpreted with caution for possible bacterial contamination[39,61,62]. LC-MS/MS coupled with 6mASCOPE is recommended as the first step to quantitatively deconvolve the total 6mA events into different species of interest and sources of contamination[39]. SMRT-seq has high signal-to-noise ratios for 6mA events, robust IPD signatures across different sequence contexts and independent calling of the four primary DNA nucleotides[10,12,27]. Although nanopore sequencing has been used to systematically detect 6mA motifs in bacteria[158], it is unclear if it is reliable for mapping individual 6mA events in eukaryotes with very low levels of 6mA. To minimize false positives, matched negative controls along with FDR analysis across a wide range of 6mA abundance are critical to adjust for various confounding factors (Fig. 4). Before functional investigation of individual 6mA events, cross validation by independent technologies (*e.g.* restriction enzyme based[58,83] and DIP-seq[61]) should be used while recognizing the possibly hidden confounding factors such as DNA secondary structure[61,145] (Fig. 3d), and other types of DNA modifications[38,164] (Fig. 3c). Finally, it is better to combine the above sequencing-based strategies with *in vitro* (treatment by exogenous methyltransferases) and *in vivo* (genetic manipulation of putative endogenous methyltransferases) experimental validation to enhance the specificity of 6mA mapping in multicellular eukaryotes[10,209,210].

4mC has been largely considered as absent in eukaryotes[25,38,111]. Only recently, 4mC was reported to be present in eukaryotic bdelloid rotifers, in which a 4mC methyltransferase was also identified[211]. Because 4mC is abundant in certain bacterial species[9], further validation

by additional studies and independent technologies are needed to assess the impact of bacterial contamination[38,39], and evaluate FDR as described above for 6mA (Box 1).

5hmC/5fC/5caC generally have much lower abundance than 5mC in most mammalian cell types[6,34,66,196,197,212,213]. As a stable epigenetic modification, 5hmC is present at 1–10% of the level of 5mC depending on the cell types: abundant in early embryo development and brain cells, but much lower in other cell types[6,34,66,196,197]. The abundance of 5fC and 5caC is orders of magnitude lower than 5hmC[6,34,66,196,197,213,214]. As a result, the same cautions should be taken as the analysis of low-abundant 6mA or 4mC in eukaryotes to assess FDRs in methods development and applications (Box 1). A number of methods have been developed for mapping 5hmC/5fC/5caC with NGS methods based on antibody, chemistry or restriction enzymes treatment[85,212–219 220]. For in-depth review of these methods, we refer the readers to previous comprehensive reviews[37,103,196,221]. For SMRT-seq, 5hmC has a slightly better signal-to-noise ratio than 5mC, while signal-to-noise ratios for 5fC and 5caC events are much higher than 5mC and 5hmC[121,164,165]. However, it is more challenging to detect 5hmC, 5fC and 5caC due to their much lower abundance than 5mC (Box 1), and it is important to use negative controls and perform FDR evaluation across a wide range of 5hmC abundance. For nanopore sequencing, Laszlo et al. first showed 5hmC generally has a decreased current signal relative to C, which suggest that 5mC and 5hmC might be discriminated[30]. Wescoe et al. further showed that nanopore sequencing was able to discriminate among five cytosine variants in DNA with different ionic current traces[222]. Although the discrimination accuracies ranged from 92 to 98%, these two studies only examined signals associated with 5mC and 5hmC events in very few specific sequence context[30,222]. Considering the influence of sequence contexts on current signal of the same modification types, it is unclear if 5mC/5hmC/5fC/5caC are generally distinguishable across the complex mammalian genomes. Similar to SMRT-seq, 5hmC/5fC/5caC mapping using nanopore sequencing is more challenging due to their much lower abundance than 5mC (Box 1), and it is critical to evaluate existing and future methods with matched negative controls and FDRs across a wide range of 5hmC/5fC/5caC abundances.

The great diversity of RNA modifications, especially those of low abundance, poses significant challenges to mapping technologies. Compared to tRNA (~20% nucleotides modified, >50 unique forms of modifications[17,136,137]) and rRNA (~2% nucleotides modified; dominant by 2'-OMe and Ψ[138–140]), mRNA has more diverse forms of RNA modification but most modifications have less abundance than those on tRNA and rRNA[13,15,37,82]. The most abundant form of mammalian mRNA modification, m6A, only accounts for 0.1%–0.6% of all adenosines[13,17,141,142]. Some other forms of mRNA modification have even lower abundance[13,15,37,82]. To study these RNA modifications with low abundance, it is critical to evaluate existing and future methods with proper negative controls and FDRs across a wide range of modification abundances. In addition, for specific study of mRNA modifications, it is a good practice to use thorough rRNA and tRNA removal, given the lower abundance of mRNA among total RNA (3–7%)[153]. For example, rRNA contamination was thought to reduce the data covered in mRNA and bring false negative when mapping m1A in human mRNA[91,182].

Although the nanopore platform has been used increasingly for direct mapping of RNA modifications[172,173], most methods are focused on the mapping of m6A[124–128,174–178]. Some recent studies attempted to map less abundant RNA modifications, such as Ψ[174] and m5C[223], but comprehensive assessment of FDR across a wide range of abundance is needed. Considering the high diversity of RNA modifications and drastic variation of nanopore signal across different sequence contexts[158,186,187], it is worth noting that sequencing errors[124–128,174] or changes in electrical signal[175–178] should not be directly interpreted as a specific form of RNA modification without comprehensive assessment of sensitivity and specificity.

For DNA or RNA modifications that are very rare in their abundance, even with the best experimental approach, errors may still arise. One example is the early published 6mA studies, which, despite comprehensive cross-validations, still reported largely false positives as demonstrated in later work[38,61,62]. Essentially, cross validation between independent technologies and reproducibility between independent studies are not substitutes for well-controlled experiments. Each new study should ensure the use of well-controlled experiments before cross validation and comparative analyses with previous studies. Science is difficult and sometimes mistakes need to be made along the path to discovering the right answers. Among the lessons learned from the pursuit of accurate modification detection, rigorous cross validation design can minimize the chance for errors and increase the chance for making more reliable findings.

Conclusions

The study of epigenomes and epitranscriptomes has been revolutionized by the introduction of technologies capable of detecting DNA and RNA modifications at a genome-wide scale. Application of these new technologies has led to a greater appreciation of the diversity and functional importance of dynamic regulation at the DNA and RNA level beyond the primary nucleotides. Broad applications of these technologies in both basic science and biomedical research have highlighted very promising applications that have translational impacts. We have reviewed several pitfalls in the development and use of different technologies for mapping DNA and RNA modifications and discussed strategies to mitigate their effects. We have focused on nucleic acid modifications that are catalyzed by enzymes, but the strategies we discussed also apply to other modifications such as damages, which are caused by endogenous or exogenous stresses[224–226]. In addition, while we have focused on natural and endogenous DNA and RNA modifications **[G]**, it is worth highlighting that these mapping technologies are also applicable for detecting exogenous DNA and RNA modifications **[G]** that are increasingly used as markers to probe functional elements such as chromatin accessibility[227–230], protein-DNA binding[231] and RNA structures[232]. With active method development for mapping both endogenous and exogenous DNA and RNA modifications, researchers will be better equipped to probe the previously hidden epigenetic mechanisms at the DNA and RNA level across the Kingdoms of life.

Acknowledgements

The work was funded by R01 HG011095 (G.F.) and R35 GM139655 (G.F.) from the National Institutes of Health.

Glossary

Restriction-modification system
(R-M system). Rudimentary immune system found in bacteria and other prokaryotic organisms, which provides a defense against foreign DNA. It includes restriction enzyme (R), which cuts specific unmethylated DNA sequences, and the methyltransferase (M), which protects the same DNA sequences

Sensitivity
A mathematical concept describing probability of a positive test conditioned on truly being positive. Also known as recall or true positive rate. In the context of mapping DNA or RNA modifications, it refers to the probability of truly modified events successfully detected as modification by a mapping method

Specificity
A mathematical concept describing probability of a negative test conditioned on truly being negative. Also known as true negative rate. Specificity =1-FPR. In the context of mapping DNA or RNA modifications, it refers to the probability that a modified event detected by a mapping method truly belongs to the modified type of interest

False positive
(FP). A mathematical concept that a test result incorrectly indicates the presence of a condition. In the context of mapping DNA or RNA modifications, FP refers to the case when base is called as modified even though it is not, or a specific modification type of interest is called from a different type of modification

False negative
(FN). A mathematical concept that a test result incorrectly indicates the absence of a condition. In the context of mapping DNA or RNA modifications, it means an authentic modification event of interest classified as either unmodified or a different type of modification

False positive rate
(FPR). A mathematical concept indicating the probability of making false positive calls with a particular test. In the context of mapping DNA or RNA modifications, it means the proportion of false positive modification calls among unmodified bases (or modified bases of other types) by a mapping method

False negative rate
(FNR). A mathematical concept indicating the probability of false negative for a particular test. In the context of mapping DNA or RNA modifications, it means the proportion of false negative calls among all the authentic modifications of interest by a mapping method

Bisulfite sequencing
(BS-seq). The treatment of DNA with bisulfite chemically converts unmethylated cytosines (C) to uracils (U), which is sequenced as thymine (T), while leaving methylated cytosines intact. The methylated base can then be identified by sequencing the bisulfite treated DNA

DNA immunoprecipitation sequencing
(DIP). A method to enrich and sequence DNA fragments containing specific methylation via immunoprecipitation. Specific antibody targeting DNA modifications of interest is incubated with fragmented genomic DNA and precipitated, followed by DNA purification and sequencing

Multiple hypothesis testing
In statistics, the multiple testing problem occurs when one considers a set of statistical inferences simultaneously based on the observed values. The more inferences are made, the more likely erroneous inferences become

False discovery rate
(FDR). A mathematical concept indicating the expected ratio of the number of false positive classifications to the total number of positive classifications. In the context of mapping DNA or RNA modifications, it refers to the probability of false positive calls among the detected modification events by a mapping method

RNA immunoprecipitation sequencing
(RIP). A method to enrich and sequence RNA fragments containing specific methylation via immunoprecipitation. Specific antibody targeting RNA modification of interest is incubated with RNA and precipitated, followed by RNA purification, reverse transcription and sequencing

Inter-pulse duration ratio
(IPD ratio). The deviation of an observed IPD (the time length between emission pulses associated with base incorporation events) from the expected IPD associated with modification-free DNA with the same flanking sequence context. The IPD ratio reflects the presence of a chemical modification of a nucleotide or its neighboring nucleotides

Modification quality value
(QV). −log10 transformed p value. In SMRT sequencing, QV describes the significance of the observed IPDs deviation from the expected level (free of modification)

Heteroplasmic
The presence of more than one type of organellar genome (mitochondrial DNA or plastid DNA) within a cell or individual

Direct RNA sequencing
The technology to sequence RNA nucleotides via direct interrogation of the original RNA strands, without reverse transcription, on the sequencer

DNA secondary structure
In most cases, DNA secondary structures consists of two polynucleotide chains wrapped around one another to form a double helix in a way referred as the canonical B-form. However, across different conditions, some sequence contexts tend to deviate from the B-form via 3D rearrangement of the two polynucleotide chains. Some examples are Z-DNA, cruciform, triplex, G-quadraplex, etc

L1 elements
Class I transposable elements of the long interspersed nuclear elements (LINEs). Also known as LINE-1 or L1 elements. L1s comprise approximately 17% of the human genome

Whole genome amplification
(WGA). The method to amplify the entire genome by random primers. Alternatively, it can be achieved with PCR using specific primers after transposon-based insertion of defined sequence. It usually starts with nanograms of DNA and results in micrograms of amplified products. The amplified DNA is essentially free of DNA modifications and can be used as a negative control

Signal-to-noise ratio
(SNR) is a measure that compares the level of a desired signal to the level of background noise. In the context of mapping DNA/RNA modifications, SNR refers to the signal of certain type of modification in a sequencing technology compared to background signal variation across unmodified bases or other modification types

Methylation motif
A short sequence pattern (usually 2bp~10bp) that are enriched for a certain type of DNA or RNA methylation events in an organism, which is often driven by the recognition preference of DNA or RNA methyltransferases. For example, nearly >95% of adenines at GATC sites are methylated (6mA) in gamma-proteobacteria, >80% of cytosines at CpG sites are methylated (5mC) in the human genome

Current (pA)
The ionic current flowing through a nanopore during nanopore sequencing. pA depends on the nucleotides occupying the constriction point. Chemical modifications to certain nucleotides can create variations in pA values, which is the foundation of modification detection in nanopore sequencing

Endogenous DNA and RNA modifications
DNA and RNA modifications generated during the endogenous metabolic processes in living organisms, usually catalyzed by certain enzymes within the cells

Exogenous DNA and RNA modifications
DNA and RNA modifications formed by exogenous factors that originated from outside the cells. For example, exogenous DNA modifications can be external modified nucleotides randomly incorporated during DNA replication, or directly catalyzed by exogenous DNA methyltransferases in vitro or in vivo

References

1. Greenberg MVC & Bourc'his D The diverse roles of DNA methylation in mammalian development and disease. Nat. Rev. Mol. Cell Biol 20, 590–607 (2019). [PubMed: 31399642]
2. Michalak EM, Burr ML, Bannister AJ & Dawson MA The roles of DNA, RNA and histone methylation in ageing and cancer. Nat. Rev. Mol. Cell Biol 20, 573–589 (2019). [PubMed: 31270442]

3. Jiang X et al. The role of m6A modification in the biological functions and diseases. Signal Transduct. Target. Ther 6, (2021).
4. Sánchez-Romero MA & Casadesús J The bacterial epigenome. Nat. Rev. Microbiol 18, 7–20 (2020). [PubMed: 31728064] This review summarizes the epigenetic regulation by bacterial DNA methylation and its contribution to the phenotypic heterogeneity in bacterial populations.
5. Luo C, Hajkova P & Ecker JR Dynamic DNA methylation: In the right place at the right time. Science (80-.). 361, 1336–1340 (2018).
6. Wu X & Zhang Y TET-mediated active DNA demethylation: Mechanism, function and beyond. Nat. Rev. Genet 18, 517–534 (2017). [PubMed: 28555658]
7. Horvath S & Raj K DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat. Rev. Genet 19, 371–384 (2018). [PubMed: 29643443]
8. Beaulaurier J, Schadt EE & Fang G Deciphering bacterial epigenomes using modern sequencing technologies. Nat. Rev. Genet 20, 157–172 (2019). [PubMed: 30546107] This review discusses the potential of currently available methods, especially LRS technologies for mapping and characterizing bacterial methylomes.
9. Blow MJ et al. The Epigenomic Landscape of Prokaryotes. PLoS Genet. 12, 1–28 (2016).
10. Boulias K & Greer EL Means, mechanisms and consequences of adenine methylation in DNA. Nat. Rev. Genet (2022). doi:10.1038/s41576-022-00456-x
11. 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). [PubMed: 23138224] This paper applies SMRT sequencing to map the 6mA in a bacterium at genome-wide scale and highlights the importance of FDR evaluation.
12. Beaulaurier J et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. Nat Biotechnol 36, 61–69 (2018). [PubMed: 29227468]
13. Kumar S & Mohapatra T Deciphering Epitranscriptome: Modification of mRNA Bases Provides a New Perspective for Post-transcriptional Regulation of Gene Expression. Front. Cell Dev. Biol 9, 1–22 (2021).
14. Barbieri I & Kouzarides T Role of RNA modifications in cancer. Nat. Rev. Cancer 20, 303–322 (2020). [PubMed: 32300195]
15. Helm M & Motorin Y Detecting RNA modifications in the epitranscriptome: Predict and validate. Nat. Rev. Genet 18, 275–291 (2017). [PubMed: 28216634] This review discusses the principles, advantages and drawbacks of new high-throughput methods for of RNA modifications.
16. Frye M, Jaffrey SR, Pan T, Rechavi G & Suzuki T RNA modifications: What have we learned and where are we headed? Nat. Rev. Genet 17, 365–372 (2016). [PubMed: 27140282]
17. Roundtree IA, Evans ME, Pan T & He C Dynamic RNA Modifications in Gene Expression Regulation. Cell 169, 1187–1200 (2017). [PubMed: 28622506]
18. Grozhik AV & Jaffrey SR Distinguishing RNA modifications from noise in epitranscriptome maps. Nat. Chem. Biol 14, 215–225 (2018). [PubMed: 29443978]
19. Saletore Y et al. The birth of the Epitranscriptome: deciphering the function of RNA modifications. Genome Biol. 13, 175 (2012). [PubMed: 23113984]
20. Zaccara S, Ries RJ & Jaffrey SR Reading, writing and erasing mRNA methylation. Nat. Rev. Mol. Cell Biol 20, 608–624 (2019). [PubMed: 31520073]
21. Linder B et al. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat. Methods 12, 767–772 (2015). [PubMed: 26121403]
22. He C Grand Challenge Commentary: RNA epigenetics? Nat. Chem. Biol 6, 863–865 (2010). [PubMed: 21079590]
23. Xue C et al. Role of main RNA modifications in cancer: N6-methyladenosine, 5-methylcytosine, and pseudouridine. Signal Transduct. Target. Ther 7, (2022).
24. Tretyakova N, Villalta PW & Kotapati S Mass spectrometry of structurally modified DNA. Chem. Rev 113, 2395–2436 (2013). [PubMed: 23441727]
25. Boulias K & Greer EL Detection of DNA Methylation in Genomic DNA by UHPLC-MS/MS BT - DNA Modifications: Methods and Protocols. in (eds. Ruzov A & Gering M) 79–90 (Springer US, 2021). doi:10.1007/978-1-0716-0876-0_7

26. Eid J et al. Real-time DNA sequencing from single polymerase molecules. Science (80-.). 323, 133–138 (2009).
27. Flusberg BA et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat. Methods 7, 461–465 (2010). [PubMed: 20453866] This paper provides an early description of direct mapping of 5mC, 5hmC and 6mA using SMRT sequencing.
28. Deamer D, Akeson M & Branton D Three decades of nanopore sequencing. Nat. Biotechnol 34, 518–524 (2016). [PubMed: 27153285]
29. Wang Y, Zhao Y, Bollas A, Wang Y & Au KF Nanopore sequencing technology, bioinformatics and applications. Nat. Biotechnol 39, 1348–1365 (2021). [PubMed: 34750572]
30. Laszlo AH et al. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. Proc. Natl. Acad. Sci. U. S. A 110, 18904–18909 (2013). [PubMed: 24167255]
31. Amarasinghe SL et al. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 1–16 (2020).
32. Krueger F, Kreck B, Franke A & Andrews SR DNA methylome analysis using short bisulfite sequencing data. Nat. Methods 9, 145–151 (2012). [PubMed: 22290186]
33. Darst RP, Pardo CE, Ai L, Brown KD & Kladde MP Bisulfite sequencing of DNA. Curr. Protoc. Mol. Biol 1–17 (2010). doi:10.1002/0471142727.mb0709s91
34. Shi DQ, Ali I, Tang J & Yang WC New insights into 5hmC DNA modification: Generation, distribution and function. Front. Genet 8, 1–11 (2017). [PubMed: 28179914]
35. Amente S et al. Genome-wide mapping of genomic DNA damage: methods and implications. Cell. Mol. Life Sci 78, 6745–6762 (2021). [PubMed: 34463773]
36. Rybin MJ et al. Emerging Technologies for Genome-Wide Profiling of DNA Breakage. Front. Genet 11, (2021).
37. Zhao LY, Song J, Liu Y, Song CX & Yi C Mapping the epigenetic modifications of DNA and RNA. Protein Cell 11, 792–808 (2020). [PubMed: 32440736] This review summarizes the high-throughput methods for mapping five forms of DNA modifications and eight forms of RNA modifications based on NGS platforms, and also summarizes biological discoveries made using these methods.
38. O'Brown ZK et al. Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. BMC Genomics 20, 1–15 (2019). [PubMed: 30606130] This article reports sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA, including both quantification methods and mapping methods.
39. Kong Y et al. Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. Science (80-.). 375, 515–522 (2022).This paper describes a machine learning method that can quantitatively deconvolve 6mA events into eukaryotic species of interest and other sources, and caution for the bacterial contaminations in the study of 6mA in eukaryotic samples.
40. Patil V et al. Human mitochondrial DNA is extensively methylated in a non-CpG context. Nucleic Acids Res. 47, 10072–10085 (2019). [PubMed: 31665742]
41. Bellizzi D et al. The control region of mitochondrial DNA shows an unusual CpG and non-CpG methylation pattern. DNA Res. 20, 537–547 (2013). [PubMed: 23804556]
42. Dou X et al. The strand-biased mitochondrial DNA methylome and its regulation by DNMT3A. Genome Res. 29, 1622–1634 (2019). [PubMed: 31537639]
43. Sharma N, Pasala MS & Prakash A Mitochondrial DNA: Epigenetics and environment. Environ. Mol. Mutagen 60, 668–682 (2019). [PubMed: 31335990]
44. 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).
45. Bicci I, Calabrese C, Golder ZJ, Gomez-Duran A & Chinnery PF Single-molecule mitochondrial DNA sequencing shows no evidence of CpG methylation in human cells and tissues. Nucleic Acids Res. 49, 12757–12768 (2021). [PubMed: 34850165] This paper describes the bias and the technical concerns of BS-seq in mapping 5mC in mtDNA and reports more reliable 5mC levels estimated by machine learning modeling with nanopore data.
46. Mechta M, Ingerslev LR, Fabre O, Picard M & Barrès R Evidence suggesting absence of mitochondrial DNA methylation. Front. Genet 8, 1–9 (2017). [PubMed: 28179914]

47. Matsuda S et al. Accurate estimation of 5-methylcytosine in mammalian mitochondrial DNA. Sci. Rep 8, 1–13 (2018). [PubMed: 29311619]
48. Hong EE, Okitsu CY, Smith AD & Hsieh C-L Regionally Specific and Genome-Wide Analyses Conclusively Demonstrate the Absence of CpG Methylation in Human Mitochondrial DNA. Mol. Cell. Biol 33, 2683–2690 (2013). [PubMed: 23671186]
49. Cao K, Feng Z, Gao F, Zang W & Liu J Mitoepigenetics: An intriguing regulatory layer in aging and metabolic-related diseases. Free Radic. Biol. Med 177, 337–346 (2021). [PubMed: 34715295]
50. Stoccoro A & Coppedè F Mitochondrial dna methylation and human diseases. Int. J. Mol. Sci 22, 1–27 (2021).
51. Lopes AFC Mitochondrial metabolism and DNA methylation: a review of the interaction between two genomes. Clin. Epigenetics 12, 1–13 (2020).
52. Zhang G et al. N6-methyladenine DNA modification in Drosophila. Cell 161, 893–906 (2015). [PubMed: 25936838]
53. Wu TP et al. DNA methylation on N6-adenine in mammalian embryonic stem cells. Nature 532, 329–333 (2016). [PubMed: 27027282]
54. Greer EL et al. DNA Methylation on N6-Adenine in C. elegans. Cell 161, 868–878 (2015). [PubMed: 25936839]
55. Zhou C et al. Identification and analysis of adenine N 6-methylation sites in the rice genome. Nature Plants 4, (2018).
56. Hao Z et al. N6-Deoxyadenosine Methylation in Mammalian Mitochondrial DNA. Mol. Cell 1–14 (2020). doi:10.1016/j.molcel.2020.02.018
57. Wang Y, Chen X, Sheng Y, Liu Y & Gao S N6-adenine DNA methylation is associated with the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in Tetrahymena. Nucleic Acids Res. 45, (2017).
58. Fu Y et al. N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. Cell 161, 879–892 (2015). [PubMed: 25936837]
59. Xie Q et al. N6-methyladenine DNA Modification in Glioblastoma. Cell 175, 1228–1243.e20 (2018). [PubMed: 30392959]
60. Le Xiao C et al. N 6 -Methyladenine DNA Modification in the Human Genome. Mol. Cell 71, 306–318.e7 (2018). [PubMed: 30017583]
61. Lentini A et al. A reassessment of DNA-immunoprecipitation-based genomic profiling. Nat. Methods 15, (2018).This paper demonstrates that systematic off-target binding of antibodies to unmodified short tandem repeats in DIP-seq studies and highlights the importance of using matched IgG negative controls.
62. Douvlataniotis K, Bensberg M, Lentini A, Gylemo B & Nestor CE No evidence for DNA N6-methyladenine in mammals. Sci. Adv 6, 1–10 (2020).This article reports DIP-seq libraries can be contaminated with mammalian mRNA, and that high abundance of m6A in mRNA can contribute to false positive 6mA peaks in DIP-seq.
63. Musheev MU, Baumgärtner A, Krebs L & Niehrs C The origin of genomic N 6-methyl-deoxyadenosine in mammalian cells. Nat. Chem. Biol 16, 630–634 (2020). [PubMed: 32203414]
64. Foox J et al. The SEQC2 epigenomics quality control (EpiQC) study. Genome Biol. 22, 1–30 (2021). [PubMed: 33397451]
65. Skvortsova K et al. Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. Epigenetics and Chromatin 10, 1–20 (2017). [PubMed: 28149326]
66. Branco MR, Ficz G & Reik W Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nat. Rev. Genet 13, 7–13 (2012).
67. Olova N et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 19, 1–19 (2018). [PubMed: 29301551]
68. Ramsahoye BH et al. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc. Natl. Acad. Sci. U. S. A 97, 5237–5242 (2000). [PubMed: 10805783]

69. Patil V, Ward RL & Hesson LB The evidence for functional non-CpG methylation in mammalian cells. Epigenetics 9, 823–828 (2014). [PubMed: 24717538]
70. Jang HS, Shin WJ, Lee JE & Do JT CpG and non-CpG methylation in epigenetic gene regulation and brain function. Genes (Basel). 8, 2–20 (2017).
71. Zhang Y et al. Large-scale comparative epigenomics reveals hierarchical regulation of non-CG methylation in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A 115, E1069–E1074 (2018). [PubMed: 29339507]
72. Morrison J et al. Evaluation of whole-genome DNA methylation sequencing library preparation protocols. Epigenetics and Chromatin 14, 1–15 (2021). [PubMed: 33407878]
73. Robertson KD DNA methylation and human disease. Nat. Rev. Genet 6, 597–610 (2005). [PubMed: 16136652]
74. Suzuki MM & Bird A DNA methylation landscapes: Provocative insights from epigenomics. Nat. Rev. Genet 9, 465–476 (2008). [PubMed: 18463664]
75. Schaefer M, Pollex T, Hanna K & Lyko F RNA cytosine methylation analysis by bisulfite sequencing. Nucleic Acids Res. 37, (2009).
76. Legrand C et al. Statistically robust methylation calling for wholetranscriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. Genome Res. 27, 1589–1596 (2017). [PubMed: 28684555]
77. Huang T, Chen W, Liu J, Gu N & Zhang R Genome-wide identification of mRNA 5-methylcytosine in mammals. Nat. Struct. Mol. Biol 26, 380–388 (2019). [PubMed: 31061524]
78. Schaefer M Chapter Fourteen - RNA 5-Methylcytosine Analysis by Bisulfite Sequencing. in RNA Modification (ed. He CBT-M in E.) 560, 297–329 (Academic Press, 2015).
79. Trixl L, Rieder D, Amort T & Lusser A Bisulfite Sequencing of RNA for Transcriptome-Wide Detection of 5-Methylcytosine BT - Epitranscriptomics: Methods and Protocols. in (eds. Wajapeyee N & Gupta R) 1–21 (Springer New York, 2019). doi:10.1007/978-1-4939-8808-2_1
80. Carlile TM et al. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature 515, 143–146 (2014). [PubMed: 25192136]
81. Li X et al. Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. Nat. Chem. Biol 11, 592–597 (2015). [PubMed: 26075521]
82. Wiener D & Schwartz S The epitranscriptome beyond m6A. Nat. Rev. Genet 22, 119–131 (2021). [PubMed: 33188361] This perspective discusses the technical and analytical challenges that led to inconsistent conclusions regarding the abundance and distribution of six forms of RNA modifications beyond m6A.
83. Luo GZ et al. Characterization of eukaryotic DNA N6-methyladenine by a highly sensitive restriction enzyme-assisted sequencing. Nat. Commun 7, 2–7 (2016).
84. Cohen-Karni D et al. The MspJI family of modification-dependent restriction endonucleases for epigenetic studies. Proc. Natl. Acad. Sci. U. S. A 108, 11040–11045 (2011). [PubMed: 21690366]
85. Sun Z et al. A Sensitive approach to map genome-wide 5-Hydroxymethylcytosine and 5-Formylcytosine at single-base resolution. Mol. Cell 57, 750–761 (2015). [PubMed: 25639471]
86. Roberts RJ, Vincze T, Posfai J & Macelis D REBASE-a database for DNA restriction and modification: Enzymes, genes and genomes. Nucleic Acids Res. 43, D298–D299 (2015). [PubMed: 25378308]
87. Zhang Z et al. Single-base mapping of m6A by an antibody-independent method. Sci. Adv 5, 1–12 (2019).
88. Garcia-Campos MA et al. Deciphering the “m6A Code” via Antibody-Independent Quantitative Profiling. Cell 178, 731–747.e16 (2019). [PubMed: 31257032]
89. Zhang Z et al. Systematic calibration of epitranscriptomic maps using a synthetic modification-free RNA library. Nat. Methods 18, 1213–1222 (2021). [PubMed: 34594034]
90. Dominissini D et al. The dynamic N1 -methyladenosine methylome in eukaryotic messenger RNA. Nature 530, 441–446 (2016). [PubMed: 26863196]
91. Safra M et al. The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. Nature 551, 251–255 (2017). [PubMed: 29072297]

92. Jin H, Huo C, Zhou T & Xie S m1A RNA Modification in Gene Expression Regulation. Genes (Basel). 13, 910 (2022). [PubMed: 35627295]

93. Wei J et al. Differential m 6 A, m 6 A m, and m 1 A Demethylation Mediated by FTO in the Cell Nucleus and Cytoplasm. Mol. Cell 71, 973–985.e5 (2018). [PubMed: 30197295]

94. Cui X et al. 5-Methylcytosine RNA Methylation in Arabidopsis Thaliana. Mol. Plant 10, 1387–1399 (2017). [PubMed: 28965832]

95. Ma J et al. M5C-Atlas: A comprehensive database for decoding and annotating the 5-methylcytosine (m5C) epitranscriptome. Nucleic Acids Res. 50, D196–D203 (2022). [PubMed: 34986603]

96. Lin S, Choe J, Du P, Triboulet R & Gregory RI The m6A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. Mol. Cell 62, 335–345 (2016). [PubMed: 27117702]

97. Dominissini D et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature 485, 201–206 (2012). [PubMed: 22575960]

98. Meyer KD et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. Cell 149, 1635–1646 (2012). [PubMed: 22608085]

99. Gokhale NS et al. N6-Methyladenosine in Flaviviridae Viral RNA Genomes Regulates Infection. Cell Host Microbe 20, 654–665 (2016). [PubMed: 27773535]

100. Delatte B et al. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. Science (80-.). 351, 282–285 (2016).

101. Wu H et al. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev. 25, 679–684 (2011). [PubMed: 21460036]

102. Li W, Li X, Ma X, Xiao W & Zhang J Mapping the m1A, m5C, m6A and m7G methylation atlas in zebrafish brain under hypoxic conditions by MeRIP-seq. BMC Genomics 23, 1–19 (2022). [PubMed: 34979896]

103. Song CX, Yi C & He C Mapping recently identified nucleotide variants in the genome and transcriptome. Nat. Biotechnol 30, 1107–1116 (2012). [PubMed: 23138310]

104. Arango D et al. Acetylation of Cytidine in mRNA Promotes Translation Efficiency. Cell 175, 1872–1886.e24 (2018). [PubMed: 30449621]

105. Sas-Chen A et al. Dynamic RNA acetylation revealed by quantitative cross-evolutionary mapping. Nature 583, 638–643 (2020). [PubMed: 32555463]

106. Thalalla Gamage S, Sas-Chen A, Schwartz S & Meier JL Quantitative nucleotide resolution profiling of RNA cytidine acetylation by ac4C-seq. Nat. Protoc 16, 2286–2307 (2021). [PubMed: 33772246]

107. Zhu S et al. Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. Genome Res 28, 1067–1078 (2018). [PubMed: 29764913] This article demonstrates that sequencing technologies tend to make false positive 6mA calls in eukaryotes, and highlights the importance of FDR evaluation for low-abundant modifications.

108. Liu X et al. N6-methyladenine is incorporated into mammalian genome by DNA polymerase. Cell Res. 31, 94–97 (2021). [PubMed: 32355286]

109. Liu B, Liu X, Lai W & Wang H Metabolically Generated Stable Isotope-Labeled Deoxynucleoside Code for Tracing DNA N6-Methyladenine in Human Cells. Anal. Chem 89, 6202–6209 (2017). [PubMed: 28471639]

110. O'Brown ZK & Greer EL N6-Methyladenine: A Conserved and Dynamic DNA Mark BT - DNA Methyltransferases - Role and Function. in (eds. Jeltsch A & Jurkowska RZ) 213–246 (Springer International Publishing, 2016). doi:10.1007/978-3-319-43624-1_10

111. Schiffers S et al. Quantitative LC–MS Provides No Evidence for m 6 dA or m 4 dC in the Genome of Mouse Embryonic Stem Cells and Tissues. Angew. Chemie - Int. Ed 56, 11268–11271 (2017).

112. Grozhik AV et al. Antibody cross-reactivity accounts for widespread appearance of m1A in 5'UTRs. Nat. Commun 10, 1–13 (2019). [PubMed: 30602773]

113. Tan SC & Yiap BC DNA, RNA, and protein extraction: The past and the present. J. Biomed. Biotechnol 2009, (2009).

114. Edelheit S, Schwartz S, Mumbach MR, Wurtzel O & Sorek R Transcriptome-Wide Mapping of 5-methylcytidine RNA Modifications in Bacteria, Archaea, and Yeast Reveals m5C within Archaeal mRNAs. PLoS Genet. 9, (2013).
115. Thu KL et al. Methylated DNA immunoprecipitation. J. Vis. Exp 23–26 (2009). doi:10.3791/935
116. Amort T & Lusser A Detection of 5-Methylcytosine in Specific Poly(A) RNAs by Bisulfite Sequencing BT - RNA Methylation: Methods and Protocols. in (ed. Lusser A) 107–121 (Springer New York, 2017). doi:10.1007/978-1-4939-6807-7_8
117. Meyer CA & Liu XS Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nat. Rev. Genet 15, 709–721 (2014). [PubMed: 25223782]
118. Sims D, Sudbery I, Ilott NE, Heger A & Ponting CP Sequencing depth and coverage: Key considerations in genomic analyses. Nat. Rev. Genet 15, 121–132 (2014). [PubMed: 24434847]
119. McIntyre ABR et al. Limits in the detection of m6A changes using MeRIP/m6A-seq. Sci. Rep 10, 1–15 (2020). [PubMed: 31913322]
120. Beaulaurier J et al. Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. Nat. Commun 6, (2015).
121. Schadt EE et al. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res. 23, 129–141 (2013). [PubMed: 23093720]
122. Vilfan ID et al. Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. J. Nanobiotechnology 11, 1 (2013). [PubMed: 23343139]
123. Abdi E, Latifi-Navid S & Latifi-Navid H Long noncoding RNA polymorphisms and colorectal cancer risk: Progression and future perspectives. Environ. Mol. Mutagen 63, 98–112 (2022). [PubMed: 35275417]
124. Pratanwanich PN et al. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat. Biotechnol 39, 1394–1402 (2021). [PubMed: 34282325]
125. Jenjaroenpun P et al. Decoding the epitranscriptional landscape from native RNA sequences. Nucleic Acids Res. 49, 1–13 (2021). [PubMed: 33275144]
126. Price AM et al. Direct RNA sequencing reveals m6A modifications on adenovirus RNA are necessary for efficient splicing. Nat. Commun 11, (2020).
127. Parker MT et al. Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6A modification. Elife 9, 1–35 (2020).
128. Liu H et al. Accurate detection of m6A RNA modifications in native RNA sequences. Nat. Commun 10, 1–9 (2019). [PubMed: 30602773]
129. Kimura S, Dedon PC & Waldor MK Comparative tRNA sequencing and RNA mass spectrometry for surveying tRNA modifications. Nat. Chem. Biol 16, 964–972 (2020). [PubMed: 32514182]
130. Ameur A, Kloosterman WP & Hestand MS Single-Molecule Sequencing: Towards Clinical Applications. Trends Biotechnol. 37, 72–85 (2019). [PubMed: 30115375]
131. Fang L et al. DeepRepeat: direct quantification of short tandem repeats on signal data from nanopore sequencing. Genome Biol. 23, 1–27 (2022). [PubMed: 34980209]
132. Shi H, Wei J & He C Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. Mol. Cell 74, 640–650 (2019). [PubMed: 31100245]
133. Huang Y et al. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS One 5, 1–9 (2010).
134. Liu Y et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biol. 22, (2021).This paper benchmarks the computational methods for detecting 5mC in mammalian genome using nanopore sequencing, and demonstrates that 5hmC levels may contribute to the discrepancy between bisulfite and nanopore sequencing.
135. Clark TA, Spittle KE, Turner SW & Korlach J Direct Detection and Sequencing of Damaged DNA Bases. Genome Integr. 2, 10 (2011). [PubMed: 22185597]
136. Suzuki T The expanding world of tRNA modifications and their disease relevance. Nat. Rev. Mol. Cell Biol 22, 375–392 (2021). [PubMed: 33658722]

137. Krutyhołowa R, Zakrzewski K & Glatt S Charging the code — tRNA modification complexes. Curr. Opin. Struct. Biol 55, 138–146 (2019). [PubMed: 31102979]
138. Sloan KE et al. Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. RNA Biol. 14, 1138–1152 (2017). [PubMed: 27911188]
139. Decatur WA & Fournier MJ rRNA modifications and ribosome function. Trends Biochem. Sci 27, 344–351 (2002). [PubMed: 12114023]
140. Sergiev PV, Aleksashin NA, Chugunova AA, Polikanov YS & Dontsova OA Structural and evolutionary insights into ribosomal RNA methylation. Nat. Chem. Biol 14, 226–235 (2018). [PubMed: 29443970]
141. Michaela F, T., H. B., Mikaela, B. & Chuan, H. RNA modifications modulate gene expression during development. Science (80-.). 361, 1346–1349 (2018).
142. Zhang W, Qian Y & Jia G The detection and functions of RNA modification m6A based on m6A writers and erasers. J. Biol. Chem 297, 100973 (2021). [PubMed: 34280435]
143. Mauer J et al. Reversible methylation of m6 Am in the 5′ cap controls mRNA stability. Nature 541, 371–375 (2017). [PubMed: 28002401]
144. Sun H et al. m6Am-seq reveals the dynamic m6Am methylation in the human transcriptome. Nat. Commun 12, 1–12 (2021). [PubMed: 33397941]
145. Guiblet WM et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. Genome Res. 28, 1767–1778 (2018). [PubMed: 30401733]
146. Kwok CK, Tang Y, Assmann SM & Bevilacqua PC The RNA structurome: Transcriptome-wide structure probing with next-generation sequencing. Trends Biochem. Sci 40, 221–232 (2015). [PubMed: 25797096]
147. Mortimer SA, Kidwell MA & Doudna JA Insights into RNA structure and function from genome-wide studies. Nat. Rev. Genet 15, 469–479 (2014). [PubMed: 24821474]
148. Wan Y, Kertesz M, Spitale RC, Segal E & Chang HY Understanding the transcriptome through RNA structure. Nat. Rev. Genet 12, 641–655 (2011). [PubMed: 21850044]
149. Talkish J, May G, Lin Y, Woolford JL & McManus CJ Mod-seq: High-throughput sequencing for chemical probing of RNA structure. Rna 20, 713–720 (2014). [PubMed: 24664469]
150. Weeks KM Advances in RNA structure analysis by chemical probing. Curr. Opin. Struct. Biol 20, 295–304 (2010). [PubMed: 20447823]
151. Helm M, Lyko F & Motorin Y Limited antibody specificity compromises epitranscriptomic analyses. Nat. Commun 10, 9–11 (2019). [PubMed: 30602780]
152. Zaringhalam M & Papavasiliou FN Pseudouridylation meets next-generation sequencing. Methods 107, 63–72 (2016). [PubMed: 26968262]
153. Palazzo AF & Lee ES Non-coding RNA: What is functional and what is junk? Front. Genet 5, 1–11 (2015).
154. Stark R, Grzelak M & Hadfield J RNA sequencing: the teenage years. Nat. Rev. Genet 20, 631–656 (2019). [PubMed: 31341269]
155. Ozsolak F & Milos PM RNA sequencing: Advances, challenges and opportunities. Nat. Rev. Genet 12, 87–98 (2011). [PubMed: 21191423]
156. Hussain S, Aleksic J, Blanco S, Dietmann S & Frye M Characterizing 5-methylcytosine in the mammalian epitranscriptome. Genome Biol. 14, 1–10 (2013).
157. Schutsky EK et al. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. Nat. Biotechnol 36, 1083–1090 (2018).
158. Tourancheau A, Mead EA, Zhang XS & Fang G Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. Nat. Methods 18, 491–498 (2021). [PubMed: 33820988] This paper reports that nanopore sequencing signal displays complex heterogeneity across methylation events of the same type, and develops a method for de novo detection of three forms of bacterial DNA methylation across diverse sequence contexts.
159. Tavakoli S et al. Detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct, long-read sequencing. bioRxiv 2021.11.03.467190 (2021). doi:10.1101/2021.11.03.467190

160. Makhamreh A et al. Messenger-RNA Modification Standards and Machine Learning Models Facilitate Absolute Site-Specific Pseudouridine Quantification. bioRxiv 2022.05.06.490948 (2022). doi:10.1101/2022.05.06.490948

161. Oliveira PH et al. Epigenomic characterization of Clostridioides difficile finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis. Nat. Microbiol 5, 166–180 (2020). [PubMed: 31768029]

162. Hong T et al. Selective detection of N6-methyladenine in DNA via metal ion-mediated replication and rolling circle amplification. Chem. Sci 8, 200–205 (2016). [PubMed: 28451166]

163. Engel JD & Von Hippel PH Effects of methylation on the stability of nucleic acid conformations. Studies at the polymer level. J. Biol. Chem 253, 927–934 (1978). [PubMed: 621212]

164. Clark TA et al. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. BMC Biol. 11, 4 (2013). [PubMed: 23339471]

165. Chaveza L et al. Simultaneous sequencing of oxidized methylcytosines produced by TET/JBP dioxygenases in Coprinopsis cinerea. Proc. Natl. Acad. Sci. U. S. A 111, E5149–E5158 (2014). [PubMed: 25406324]

166. Simpson JT et al. Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods 14, 407–410 (2017). [PubMed: 28218898] This paper describes the use of synthetically methylated DNA to train a hidden Markov model for distinguishing 5mC from unmethylated cytosine in mammalian genome using nanopore data.

167. Rand AC et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat. Methods 14, 411–413 (2017). [PubMed: 28218897]

168. Ni P et al. Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning. Nat. Commun 12, 1–11 (2021). [PubMed: 33397941]

169. Liu Y et al. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat. Biotechnol 37, 424–429 (2019). [PubMed: 30804537]

170. Liu Y et al. Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. Genome Biol. 21, 1–9 (2020).

171. Song CX et al. Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat. Methods 9, 75–77 (2012).

172. Garalde DR et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15, 201–206 (2018). [PubMed: 29334379] This paper provides an early description of direct RNA sequencing and detecting RNA modifications using the nanopore platform.

173. Xu L & Seki M Recent advances in the detection of base modifications using the Nanopore sequencer. J. Hum. Genet 65, 25–33 (2020). [PubMed: 31602005]

174. Begik O et al. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. Nat. Biotechnol 39, 1278–1291 (2021). [PubMed: 33986546]

175. Lorenz DA, Sathe S, Einstein JM & Yeo GW Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base-specific resolution. Rna 26, 19–28 (2020). [PubMed: 31624092]

176. Gao Y et al. Quantitative profiling of N 6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. Genome Biol. 22, 1–17 (2021). [PubMed: 33397451]

177. Qin H et al. DENA: training an authentic neural network model using Nanopore sequencing data of Arabidopsis transcripts for detection and quantification of N 6-methyladenosine on RNA. Genome Biol. 23, 1–23 (2022). [PubMed: 34980209]

178. Leger A et al. RNA modifications detection by comparative Nanopore direct RNA sequencing. Nat. Commun 12, 1–17 (2021). [PubMed: 33397941]

179. Gallego Romero I, Pai AA, Tung J & Gilad Y RNA-seq: Impact of RNA degradation on transcript quantification. BMC Biol. 12, 1–13 (2014). [PubMed: 24417977]

180. Schuierer S et al. A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. BMC Genomics 18, 1–13 (2017). [PubMed: 28049423]

181. Li X et al. Transcriptome-wide mapping reveals reversible and dynamic N1-methyladenosine methylome. Nat. Chem. Biol 12, 311–316 (2016). [PubMed: 26863410]

182. Xiong X, Li X, Wang K & Yi C Perspectives on topology of the human m 1 A methylome at single nucleotide resolution. Rna 24, 1437–1442 (2018). [PubMed: 30131401]
183. Siwek W, Czapinska H, Bochtler M, Bujnicki JM & Skowronek K Crystal structure and mechanism of action of the N6-methyladenine-dependent type IIM restriction endonuclease R.DpnI. Nucleic Acids Res. 40, 7563–7572 (2012). [PubMed: 22610857]
184. Mierzejewska K et al. Structural basis of the methylation specificity of R.DpnI. Nucleic Acids Res. 42, 8745–8754 (2014). [PubMed: 24966351]
185. Kazrani AA, Kowalska M, Czapinska H & Bochtler M Crystal structure of the 5hmC specific endonuclease PvuRts1I. Nucleic Acids Res. 42, 5929–5936 (2014). [PubMed: 24634440]
186. Ni P et al. DeepSignal: Detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Bioinformatics 35, 4586–4595 (2019). [PubMed: 30994904]
187. Liu Q et al. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nat. Commun 10, (2019).
188. McIntyre ABR et al. Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat. Commun 10, 1–11 (2019). [PubMed: 30602773]
189. Yuen ZWS et al. Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. Nat. Commun 12, 1–12 (2021). [PubMed: 33397941]
190. Zhang YZ et al. On the application of BERT models for nanopore methylation detection. Proc. - 2021 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2021 320–327 (2021). doi:10.1109/BIBM52615.2021.9669841
191. Liu Q, Georgieva DC, Egli D & Wang K NanoMod: A computational tool to detect DNA modifications using Nanopore long-read sequencing data. BMC Genomics 20, (2019).
192. Stoiber MH et al. De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. bioRxiv 094672 (2016).
193. Tse OYO et al. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. Proc. Natl. Acad. Sci. U. S. A 118, 1–11 (2021).
194. Ni P et al. DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. bioRxiv 2022.02.26.482074 (2022).
195. Whalen S, Schreiber J, Noble WS & Pollard KS Navigating the pitfalls of applying machine learning in genomics. Nat. Rev. Genet 23, 169–181 (2022). [PubMed: 34837041]
196. Plongthongkum N, Diep DH & Zhang K Advances in the profiling of DNA modifications: Cytosine methylation and beyond. Nat. Rev. Genet 15, 647–661 (2014). [PubMed: 25159599] This review discusses the principles as well as experimental and analytical challenges in mapping 5mC and its oxidation derivatives with BS-seq and other NGS-based methods.
197. Lio CWJ & Rao A TET enzymes and 5hMC in adaptive and innate immune systems. Front. Immunol 10, 1–13 (2019). [PubMed: 30723466]
198. Lai W, Lyu C & Wang H Vertical Ultrafiltration-Facilitated DNA Digestion for Rapid and Sensitive UHPLC-MS/MS Detection of DNA Modifications. Anal. Chem 90, 6859–6866 (2018). [PubMed: 29792685]
199. Liu X, Lai W, Zhang N & Wang H Predominance of N6-Methyladenine-Specific DNA Fragments Enriched by Multiple Immunoprecipitation. Anal. Chem 90, 5546–5551 (2018). [PubMed: 29652489]
200. Debo BM, Mallory B & Stergachis AB Evaluation of N^6-adenine DNA-immunoprecipitation-based genomic profiling in eukaryotes. bioRxiv 2022.03.02.482749 (2022). doi:10.1101/2022.03.02.482749
201. Liang Z et al. DNA N6-Adenine Methylation in Arabidopsis thaliana. Dev. Cell 45, 406–416.e3 (2018). [PubMed: 29656930]
202. Luo GZ et al. N 6 -methyldeoxyadenosine directs nucleosome positioning in Tetrahymena DNA. Genome Biol. 19, (2018).
203. Mondo SJ et al. Widespread adenine N6-methylation of active genes in fungi. Nat Genet (2017). doi:10.1038/ng.3859
204. Vaisvila R et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. Genome Res. 31, 1280–1289 (2021). [PubMed: 34140313]

205. 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). [PubMed: 33468551]
206. Feng S, Zhong Z, Wang M & Jacobsen SE Efficient and accurate determination of genome-wide DNA methylation patterns in Arabidopsis thaliana with enzymatic methyl sequencing. Epigenetics and Chromatin 13, 1–17 (2020). [PubMed: 31918747]
207. Spealman P, Burrell J & Gresham D Inverted duplicate DNA sequences increase translocation rates through sequencing nanopores resulting in reduced base calling accuracy. Nucleic Acids Res. 48, 4940–4945 (2020). [PubMed: 32255181]
208. Shen C, Wang K, Deng X & Chen J DNA N6-methyldeoxyadenosine in mammals and human disease. Trends Genet. 38, 454–467 (2022). [PubMed: 34991904]
209. Wang Y et al. A distinct class of eukaryotic MT-A70 methyltransferases maintain symmetric DNA N6-adenine methylation at the ApT dinucleotides as an epigenetic mark associated with transcription. Nucleic Acids Res. 47, 11771–11789 (2019). [PubMed: 31722409]
210. Beh LY et al. Identification of a DNA N6-Adenine Methyltransferase Complex and Its Impact on Chromatin Organization. Cell 177, 1781–1796.e25 (2019). [PubMed: 31104845]
211. Rodriguez F, Yushenova IA, DiCorpo D & Arkhipova IR Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA. Nat. Commun 13, 1–17 (2022). [PubMed: 34983933]
212. Dietzsch J, Feineis D & Höbartner C Chemoselective labeling and site-specific mapping of 5-formylcytosine as a cellular nucleic acid modification. FEBS Lett. 592, 2032–2047 (2018). [PubMed: 29683490]
213. Song CX et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. Cell 153, 678–691 (2013). [PubMed: 23602153]
214. Zhu C et al. Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. Cell Stem Cell 20, 720–731.e5 (2017). [PubMed: 28343982]
215. Shen L et al. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell 153, 692–706 (2013). [PubMed: 23602152]
216. Neri F, Incarnato D, Krepelova A, Parlato C & Oliviero S Methylation-Assisted bisulfite sequencing to simultaneously map 5fC and 5caC on a genome-wide scale for DNA demethylation analysis. Nat. Protoc 11, 1191–1205 (2016). [PubMed: 27281647]
217. Yu M et al. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell 149, 1368–1380 (2012). [PubMed: 22608086]
218. Booth MJ, Marsico G, Bachman M, Beraldi D & Balasubramanian S Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. Nat. Chem 6, 435–440 (2014). [PubMed: 24755596]
219. B. JM et al. Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. Science (80-.). 336, 934–937 (2012).
220. Füllgrabe J et al. Accurate simultaneous sequencing of genetic and epigenetic bases in DNA. bioRxiv 2022.07.08.499285 (2022). doi:10.1101/2022.07.08.499285
221. Shen L, Song CX, He C & Zhang Y Mechanism and function of oxidative reversal of DNA and RNA methylation. Annu. Rev. Biochem 83, 585–614 (2014). [PubMed: 24905787]
222. Wescoe ZL, Schreiber J & Akeson M Nanopores discriminate among five C5-cytosine variants in DNA. J. Am. Chem. Soc 136, 16582–16587 (2014). [PubMed: 25347819]
223. Stephenson W et al. Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. Cell Genomics 2, 100097 (2022). [PubMed: 35252946]
224. Poetsch AR The genomics of oxidative DNA damage, repair, and resulting mutagenesis. Comput. Struct. Biotechnol. J 18, 207–219 (2020). [PubMed: 31993111]
225. Chatterjee N & Walker GC Mechanisms of DNA damage, repair, and mutagenesis. Environ. Mol. Mutagen 58, 235–263 (2017). [PubMed: 28485537]
226. Huang R & Zhou PK DNA damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy. Signal Transduction and Targeted Therapy 6, (Springer US, 2021).

227. Abdulhay NJ et al. Massively multiplex single-molecule oligonucleosome footprinting. Elife 9, 1–23 (2020).
228. Shipony Z et al. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. Nat. Methods 17, 319–327 (2020). [PubMed: 32042188]
229. Wang Y et al. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. Genome Res. 29, 1329–1342 (2019). [PubMed: 31201211]
230. Stergachis AB, Debo BM, Haugen E, Churchman LS & Stamatoyannopoulos JA Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science (80-.). 368, 1449–1454 (2020).
231. Altemose N et al. DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide. Nat. Methods 19, (2022).
232. Ghut J et al. Determination of isoform-specific RNA structure with nanopore long reads. Nature Biotechnology 39, (2020).
233. Weber M et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet 37, 853–862 (2005). [PubMed: 16007088]
234. Pomraning KR, Smith KM & Freitag M Genome-wide high throughput analysis of DNA methylation in eukaryotes. Methods 47, 142–150 (2009). [PubMed: 18950712]
235. Ficz G et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature 473, 398–404 (2011). [PubMed: 21460836]
236. Xu Y et al. Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. Mol. Cell 42, 451–464 (2011). [PubMed: 21514197]
237. Liu J et al. Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. Nat. Commun 7, 1–7 (2016).
238. Koziol MJ et al. Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. Nat. Struct. Mol. Biol 23, 24–30 (2016). [PubMed: 26689968]
239. Ando Y & Hayashizaki Y Restriction landmark genomic scanning. Nat. Protoc 1, 2774–2783 (2007).
240. Koike K, Matsuyama T & Ebisuzaki T Epigenetics: Application of virtual image restriction landmark genomic scanning (Vi-RLGS). FEBS J. 275, 1608–1616 (2008). [PubMed: 18331348]
241. Suzuki M & Greally JM DNA methylation profiling using HpaII tiny fragment enrichment by ligation-mediated PCR (HELP). Methods 52, 218–222 (2010). [PubMed: 20434563]
242. Oda M & Greally JM The Help Assay BT - DNA Methylation: Methods and Protocols. in (ed. Tost J) 77–87 (Humana Press, 2009). doi:10.1007/978-1-59745-522-0_7
243. Sun Z et al. High-Resolution Enzymatic Mapping of Genomic 5-Hydroxymethylcytosine in Mouse Embryonic Stem Cells. Cell Rep. 3, 567–576 (2013). [PubMed: 23352666]
244. Jia Z et al. A 5-mC dot blot assay quantifying the DNA methylation level of chondrocyte dedifferentiation in vitro. J. Vis. Exp 2017, 5–9 (2017).
245. Jin SG, Wu X, Li AX & Pfeifer GP Genomic mapping of 5-hydroxymethylcytosine in the human brain. Nucleic Acids Res. 39, 5015–5024 (2011). [PubMed: 21378125]
246. Inoue A, Shen L, Dai Q, He C & Zhang Y Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. Cell Res. 21, 1670–1676 (2011). [PubMed: 22124233]
247. Wang Y et al. Naphthalimide derivatives as multifunctional molecules for detecting 5-formylpyrimidine by both PAGE analysis and dot-blot assays. Chem. Commun 54, 1497–1500 (2018).
248. Wheldon LM et al. Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. Cell Rep. 7, 1353–1361 (2014). [PubMed: 24882006]
249. Shen L, Liang Z & Yu H Dot Blot Analysis of N6-methyladenosine RNA Modification Levels. Bio-Protocol 7, 4–8 (2017).
250. Nagarajan A, Janostiak R & Wajapeyee N Dot Blot Analysis for Measuring Global N6-Methyladenosine Modification of RNA BT - Epitranscriptomics: Methods and

Protocols. in (eds. Wajapeyee N & Gupta R) 263–271 (Springer New York, 2019). doi:10.1007/978-1-4939-8808-2_20

251. Mishima E et al. Immuno-northern blotting: Detection of RNA modifications by using antibodies against modified nucleosides. PLoS One 10, 1–17 (2015).

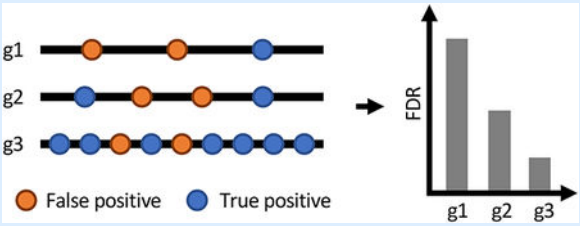
252. Yu M et al. Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfitesequencing. Nucleic Acids Res. 43, 1–10 (2015). [PubMed: 25505162]

253. Faulk C Implications of DNA methylation in toxicology. Toxicoepigenetics: Core Principles and Applications (Elsevier Inc., 2018). doi:10.1016/B978-0-12-812433-8.00006-X

Box 1:

The importance of modification abundance.

The abundance of a DNA or RNA modification in a sample of interest is an essential consideration for experimental design and data interpretation, both during method development and for evaluation of the techniques reviewed here. According to the formula FDR = $N_{fp}/(N_{tp} + N_{fp})$, where N_{tp}, is the number of truly modified events in a sample of interest and N_{fp}, is the number of false positive calls made from a sample of interest (intrinsic to a mapping method), then when the modification of interest is highly abundant in a genome of interest a relatively small number of false positives is associated with a low false discovery rate as $N_{tp} >> N_{fp}$. In such cases (represented by g3 in the figure), false positive calls are not expected to have a major influence in downstream data interpretation. However, when the modification of interest is of low abundance in the genome ($N_{tp} \sim N_{fp}$ or $N_{tp} << N_{fp}$), the same number of false positive calls will result in a much higher false discovery rate, FDR = $N_{fp}/(N_{tp} + N_{fp})$. In such cases (represented by g1 and g2 in the figure), without rigorous evaluation for FDR, false positive calls can greatly confound data interpretation and downstream functional studies. This raises cautions in the mapping of DNA or RNA modifications that are of very low abundance.



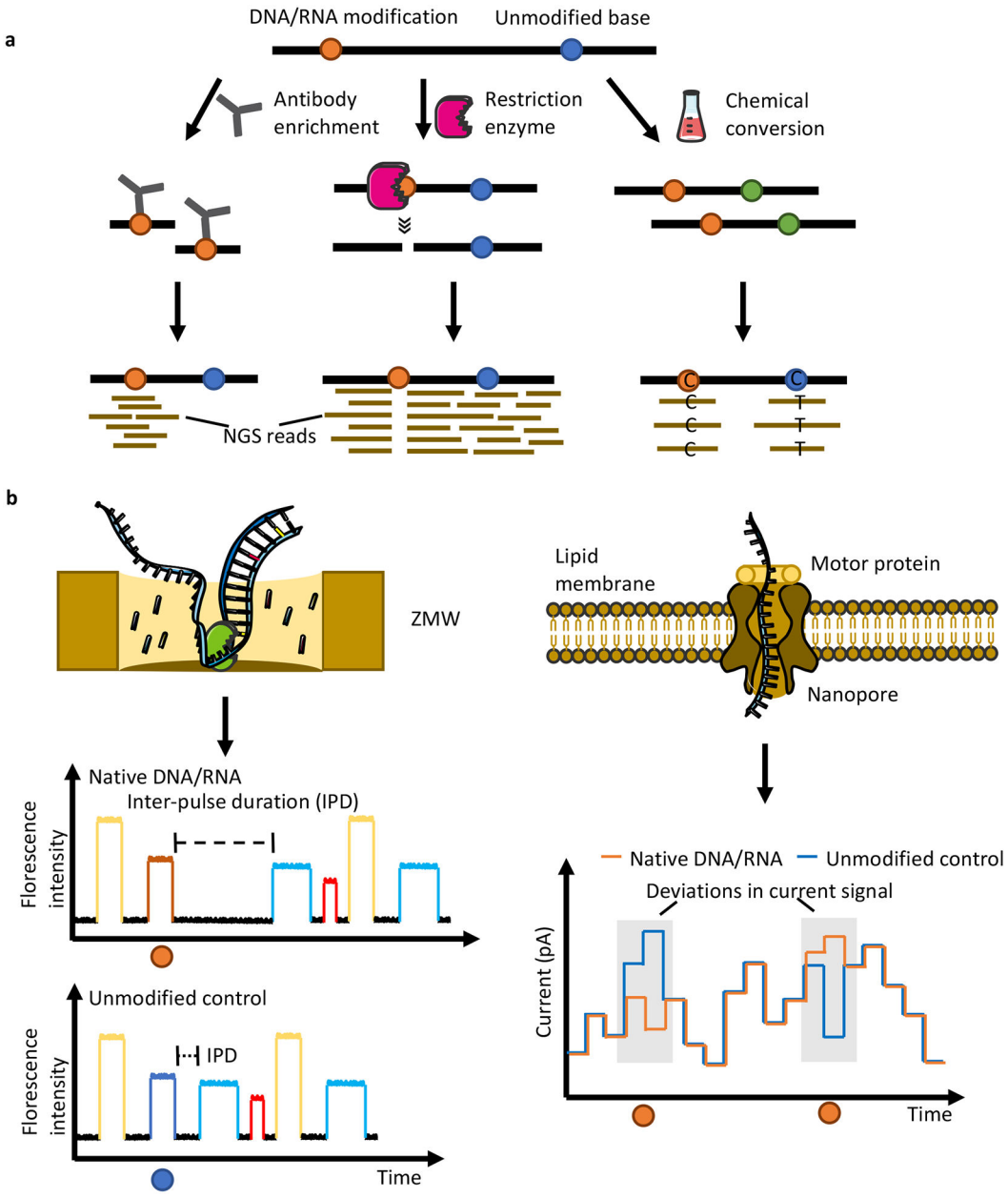


Figure 1. DNA/RNA modification mapping methods based on next generation sequencing and long read sequencing technologies.

a. Next generation sequencing (NGS)-based methods require pre-treatment or pre-labelling of the nucleic acid with antibodies (left), restriction enzymes (middle) or chemicals (right) before sequencing, so that modified and unmodified bases to be distinguished during the NGS sequencing.

b. Long read sequencing (LRS)-based methods can directly detect modified bases. Left, for SMRT sequencing, a DNA polymerase (or reverse transcriptase) is bound within the zero-mode waveguide (ZMW). When dNTP is incorporated at the polymerase active site, it will emit a fluorescent pulse in the corresponding color channel. The order of pulses provides the read sequence and inter-pulse duration between base incorporation events

indicate the presence of a covalent modification in the template DNA/RNA. Right, for nanopore sequencing, it relies on engineered biological nanopores embedded in a lipid membrane to sequence single-stranded DNA (ssDNA) or RNA. The ionic current measured as DNA or RNA gets sequenced through the nanopore depends on the precise set of nucleotides occupying the constriction point. Modified nucleotides in the ssDNA or RNA introduce distinct current patterns, making it possible to detect the existence of modified bases relative to non-modified nucleotides.

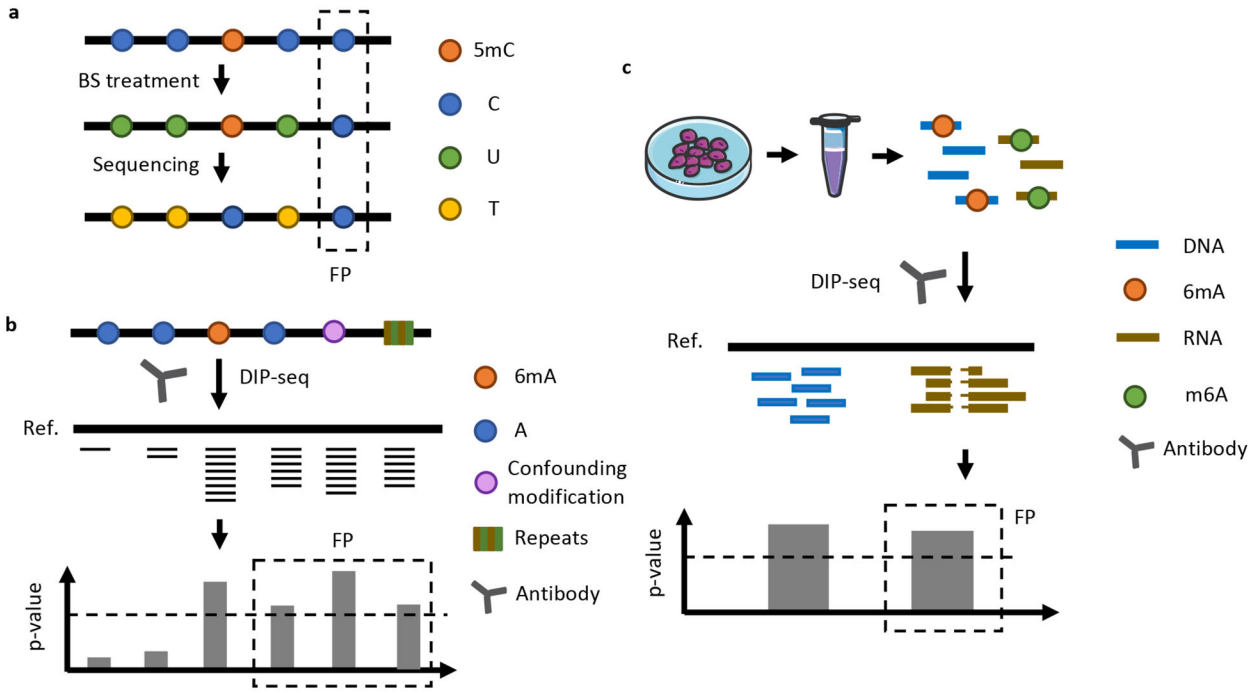


Figure 2. Overview of experimental pitfalls that can lead to false positive calls of DNA/RNA modifications.

a. Insufficient bisulfite (BS) treatment in BS-seq can leave a small percentage of non-modified cytosines unconverted, which are then falsely called as 5-methylcytosine (5mC) in downstream BS-seq data analysis. FP, false positive.

b. The non-specificity of antibodies in DNA immunoprecipitation sequencing (DIP-seq) or RNA immunoprecipitation sequencing (RIP-seq) can result in systematic false positive calls at unmodified bases, modified bases that are not the form of interest, or repetitive sequences with DNA secondary structure. 6mA, N6-methyldeoxyadenosine. Ref., Reference genome.

c. Certain mRNAs contamination through standard DNA extraction protocols may confound next generation sequencing (NGS) DNA sequencing and lead to false positive peaks in DIP-seq.

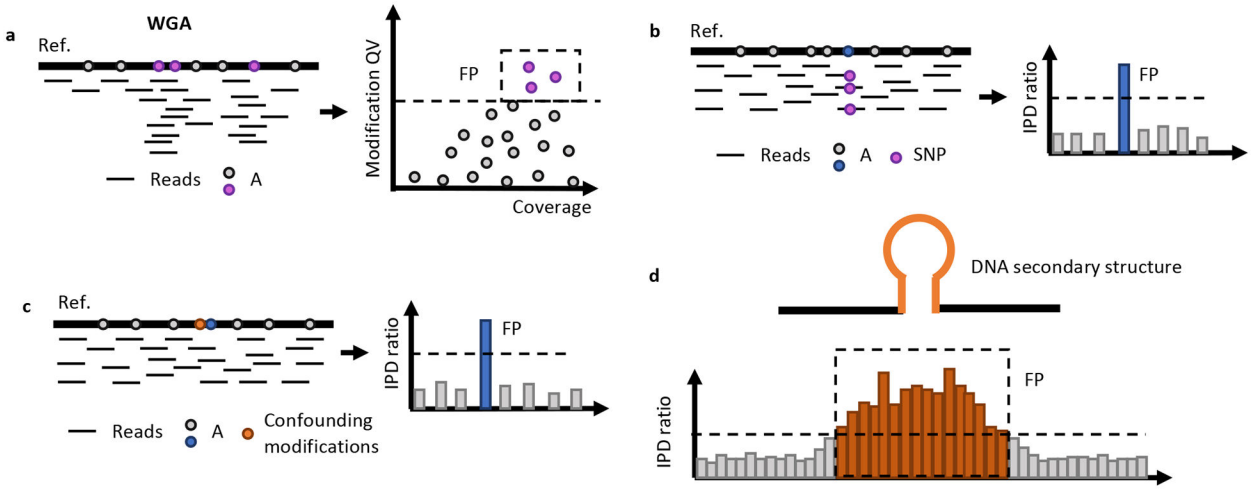


Figure 3. Overview of analytical pitfalls that can lead to false positive calls of DNA/RNA modifications.

a. For single-molecule, real-time sequencing (SMRT-seq), false positives (FP) can arise in methylation free whole genome amplification (WGA) sample, especially at high sequencing depth, because standard tools are based on fixed threshold on modification quality value (QV, −log10 transformed p value). Ref., Reference genome

b. Reference heterogeneity, such as single nucleotide polymorphisms (SNPs), can lead to overestimation of inter-pulse duration (IPD) ratios, resulting in false positives in SMRT-seq.

c. In SMRT-seq, modifications other than the one of interest (such as DNA damage) can affect IPD ratio on neighboring bases (in this case, adenine) and result in false positives. Other sequencing platforms and mapping methods also face similar challenges of confounding modifications.

d. DNA secondary structure may affect DNA polymerase kinetics and create false positive modifications in the flanking neighborhood by SMRT-seq. NGS and nanopore sequencing may also face similar challenges. In addition, single-stranded RNA is prone to form complex RNA secondary structures, which can confound both NGS- and LRS-based methods for detecting RNA modification.

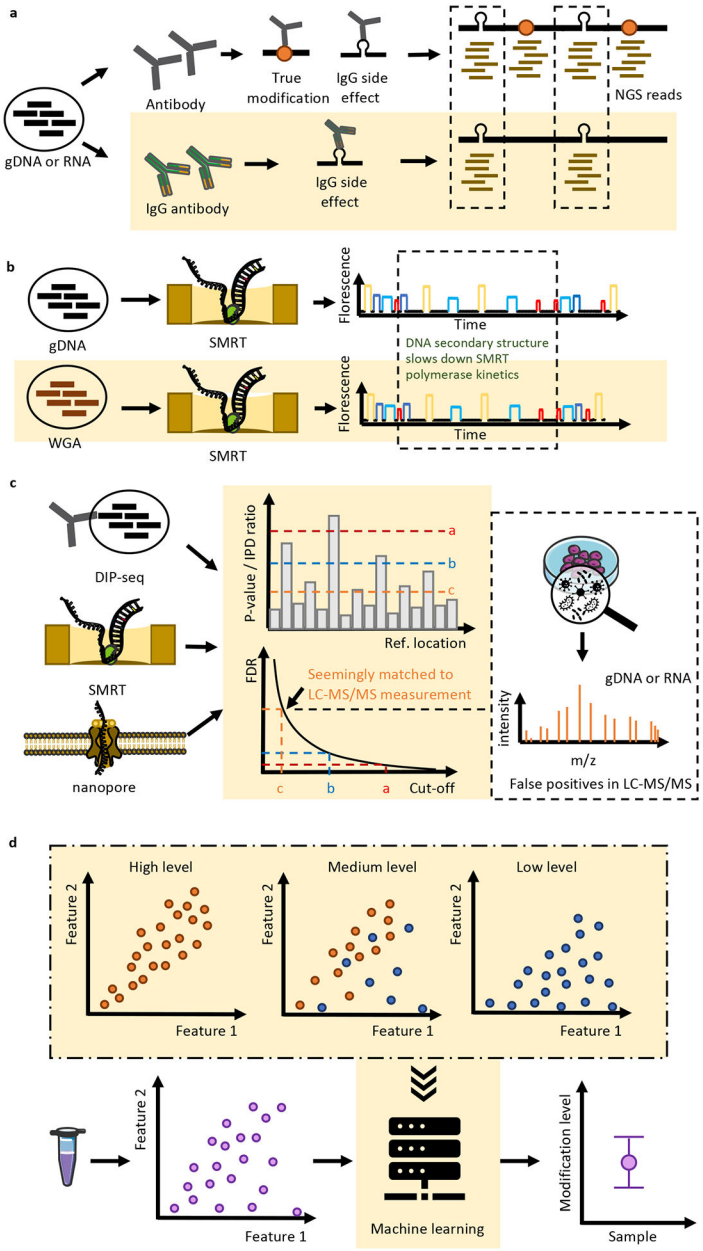


Figure 4. Mitigating false positive mapping calls of DNA and RNA modifications

a. For DIP-seq and RIP-seq, an IgG immunoprecipitated control can help adjust for non-specificity of antibodies and reduce false positive calls.

b. For SMRT sequencing, a whole genome amplification (WGA) control help evaluate the false positive calls due to the abnormal DNA polymerase (or RNA reverse transcriptase) kinetics. For example, systematic reduction in kinetics can be due to the secondary structures that can confound the detection of DNA modifications.

c. For most sequencing methods, it is more reliable to use FDR than the use of an arbitrary cutoff (e.g. p-value or IPD ratio, etc), even though a cutoff might seem to be ‘consistent’ with LC-MS/MS estimation.

d. A quantification model can be used to estimate the abundance of a DNA or RNA modification of interest. The machine learning model is trained with features across a number of positive and negative controls containing the modification at a wide range of abundance. For prediction, the machine learning model can predict modification level along with a confidence interval.

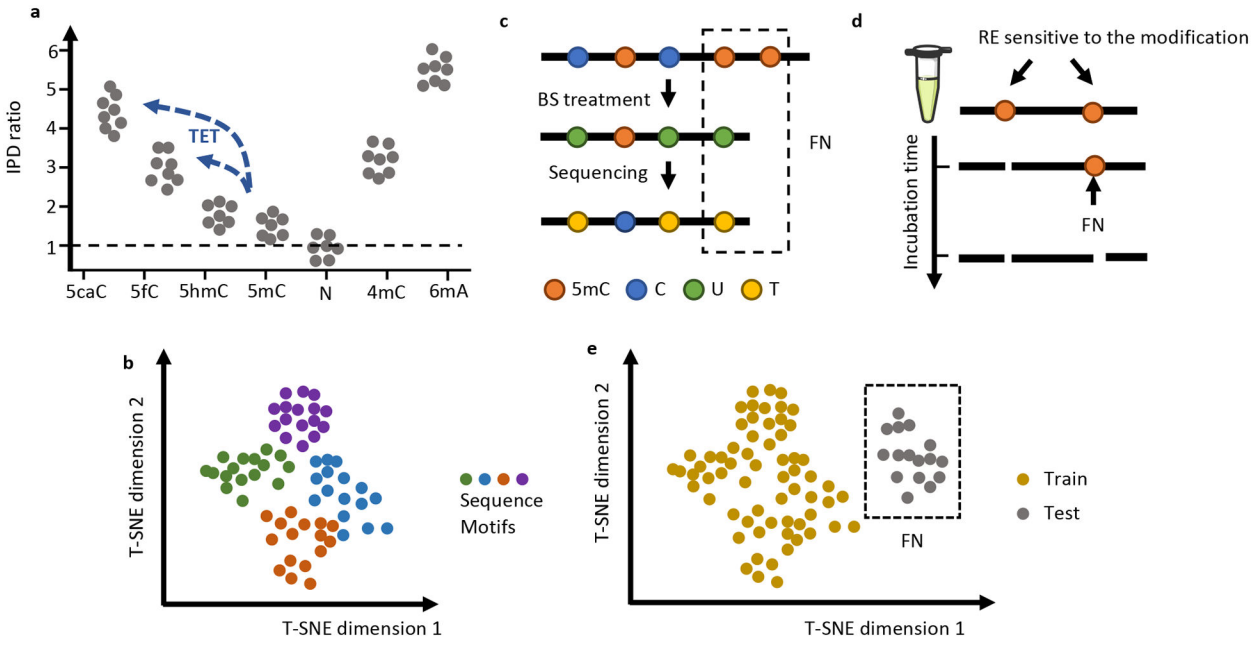


Figure 5. Overview of pitfalls that can lead to false negative calls of DNA modifications.

a. An individual technique is often more effective for detecting certain forms of DNA modifications than others. For example, single-molecule, real-time sequencing (SMRT) sequencing has stronger signal-to-noise ratios for 6mA and 4mC events than 5mC and 5hmC. The signal-to-noise ratios of 5mC and 5hmC can be enhanced by converting 5mC and 5hmC to 5fC and 5caC using the Ten-Eleven Translocation (TET) enzyme. 5CaC, 5-carboxylcytosine; 5fC, 5-formylcytosine; 5mC, 5-methylcytosine; 5hmC, 5-hydroxymethylcytosine; 4mC, N4-methylcytosine; 6mA, N6-methyladenine; N, unmodified bases.

b. In nanopore sequencing, the signal-to-noise ratio can have drastic variations across different sequence contexts (or motifs), even for the same form of DNA modification, as shown with schematic t-distributed stochastic neighbor embedding (t-SNE) map.

c. Prolonged bisulfite treatment can lead to increased conversion of 5mC to Uracil (U, which will be read as T in sequencing) and increased DNA degradation. Both processes can result in false negatives (FNs).

d. False negatives can arise when certain genomic sequence motifs targeted by a restriction enzyme (RE) are not adequately digested, for example, owing to insufficient incubation time.

e. False negatives can arise due to the use of training datasets that do not represent test datasets. For example, machine learning models trained with a limited set of sequence motifs are not generally applicable for mapping the same form of DNA or RNA modifications in other sequence contexts, here shown with a schematic t-SNE map.

Table 1.

Main techniques used for mapping DNA and RNA modifications.

Technique	Underlying principle	DNA modifications detected	RNA modifications detected
Next generation sequencing (NGS)-based methods			
BS-seq	Chemical conversion of non-modified cytosines (read as thymines) upon bisulfite treatment, while 5mCs or 5hmCs resist the treatment and are still read as cytosines.	5mC and 5hmC[32,33,196]	m5C and hm5C[15,75–78]
DIP-seq/ RIP-seq	Antibody enrichment of DNA or RNA fragments containing modifications of interest via immunoprecipitation.	5mC[233,234], 5hmC[235,236], 5fC[215], 5caC[215], 6mA[53,56,237,238] and 4mC[211]	m6A[93,96–99], m1A[90–93], m5C[77,94,95], Ac4C[104] and several other RNA modifications[15,18,37,100–103]
RE-seq	Restriction enzyme digestion (either sensitive or dependent) at specific sequence motif sites to map modification events.	6mA[58,83], 5mC[239–242], 5hmC[85,243]	m6A[87,88]
Long read sequencing (LRS)-based methods			
SMRT-seq	Direct detection of DNA/RNA modifications by monitoring changes in the kinetics of the DNA polymerase or RNA reverse transcriptase kinetics during sequencing.	6mA[8,11,12,27,39,121], 5mC[166,186–190,193], 5fC[164,165], 5caC[164,165], 4mC[8,86]	m6A[122] (proof of concept, not widely used)
Nanopore	Direct detection of DNA and RNA modifications by monitoring the changes in ionic current as ssDNA or RNA is sequenced through the nanopore.	6mA[158,167], 4mC[158], 5mC[134,158,166–168], 5hmC[30,222]	m6A[124–128,172,173], Ψ[125,159,160,174,223] and several other forms of RNA modifications[82,125,178,223]
Methods for quantitation			
LC-MS/MS	Physical separation of different nucleotides and their modified forms by liquid chromatography (LC) coupled with mass analysis by mass spectrometry (MS).	In general, all DNA modifications with internal standards available.	In general, all RNA modifications with internal standards available.
Dot-blotting	Immuno-detection through antibody binding to target DNA or RNA on the membrane.	5mC[244], 5hmC[245], 5fC[246,247], 5caC[246,248], 6mA[52,54,59,201], 4mC[211]	m6A[249,250], m5C[94], m5hC[100] and several other forms of RNA modifications[251]

DIP-seq, DNA immunoprecipitation sequencing; RIP-seq, RNA immunoprecipitation sequencing; RE-seq, restriction enzyme (RE) based sequencing; SMRT-seq, single-molecule, real-time sequencing; BS-seq: bisulfite sequencing

6mA, N6-methyladenine; 4mC, N4-methylcytosine; 5mC, 5-methylcytosine; 5hmC, 5-hydroxymethylcytosine; 5fC, 5-formylcytosine; 5CaC, 5-carboxylcytosine. Ψ, pseudouridine; m5C, 5-methylcytidine; m6A, N6-methyladenosine; m5hC, 5-hydroxymethylcytidine; ac4C, N4-acetylcytidine.

Table 2.

Key considerations for mapping DNA and RNA modifications.

	Prevalence	Chemical-, enzyme- or antibody-based methods with NGS	Direct detection by SMRT sequencing	Direct detection by nanopore sequencing	Cross-validating methods
6mA	• Prokaryotes: prevalent, abundant and motif-driven[4,8–10]. • Some protozoa: prevalent, abundant and motif-driven[4,8–10]. • Most protozoa and multicellular eukaryotes: mostly very low, from 0.1% to 0.0001%, or undetectable[53,107–111,208].	• Be mindful of possible contamination from RNA m6A. • Use matched negative controls for FDR evaluation • Use IgG to avoid FP in DIP-seq. • Ensure optimal digestion conditions in RE-seq to minimize FP and FN.	• High signal-to-noise ratio, reliable for both *de novo* motif discovery and single base resolution mapping in most prokaryotes. • For genomes with low 6mA abundance, proper negative controls are necessary for FDR evaluation.	• *De novo* discovery of 6mA motifs well validated in prokaryotes. Reliable mapping at single base resolution under development. • Applications to eukaryotes still under development. • Matched negative controls are necessary for FDR evaluation.	• LC-MS/MS • SMRT-seq • RE-seq • DIP-seq
4mC	• Prokaryotes: prevalent (especially in thermophilic bacteria and archaea), abundant and motif-driven[4,8–10]. • Eukaryotes: considered as absent in eukaryotes. Only recently, 4mC was reported to be present in eukaryotic bdelloid rotifers[211].	• Use matched negative controls for FDR evaluation • Use IgG to avoid FP in DIP-seq. • Proper treatment in RE-seq to minimize FP and FN.	• Moderate signal-to-noise ratio, reliable for both de novo motif discovery and single base resolution mapping in most prokaryotes. • For genomes with low 4mC abundance, proper negative controls are necessary for FDR evaluation.	• *De novo* discovery of 4mC motifs well validated in prokaryotes. Reliable mapping at single base resolution under development.	• LC-MS/MS • SMRT-seq • DIP-seq • 4mC-TAB-seq[252]

	Prevalence	Chemical-, enzyme- or antibody-based methods with NGS	Direct detection by SMRT sequencing	Direct detection by nanopore sequencing	Cross-validating methods
5mC	• Prokaryotes: prevalent, abundant and motif-driven[4,8–10]. • Eukaryotes: The most abundant form of modified nucleotide in eukaryotic DNA, accounting for up to 5% of cytosines in vertebrates[253]. Mostly in CpG. Also identified in CHG, CHH and other motifs[67–70].	• Ensure optimal bisulfite treatment to avoid FP and FN in BS-seq. • Ensure optimal digestion conditions for RE-seq to minimize FP and FN.	• Methods recently developed to map 5mC at CpG sites at single nucleotide and single molecule resolution[166,186–190,193]. • Be mindful of FP and FN when detecting 5mC at non-CpG sites, because of the low abundance and relatively low signal noise ratio of 5mC.	• Methods widely used to map 5mC at CpG sites at single nucleotide and single molecule resolution. • Be mindful of FP and FN for low abundant 5mC at non-CpG sites.	• LC-MS/MS • BS-seq • EM-seq[204,205] • Nanopore sequencing • SMRT-seq • RE-seq • DIP-seq • TAPS-seq[169,170] • 6-Letter sequencing[220]
5hmC/5fC/ 5caC	• Mainly identified in mammalian cells[6,34,66,196,197]. • 5hmC is present at 1–10% of the level of 5mC depending on the cell types: abundant in early embryo development and brain cells, but much lower in other cell types[6,34,66,196,197]. The abundance of 5fC and 5caC is orders of magnitude lower than 5hmC[6,34,66,196,197,213,214].	• Use matched negative control for FDR evaluation • Use IgG to avoid FP in DIP-seq. • Ensure treatment with chemicals or enzymes is optimized to minimize FP and FN.	• Moderate signal-to-noise ratio for 5hmC and high signal-to-noise ratio for 5fC and 5Cac. • Analysis of synthetic DNA oligos suggested different signatures across 5mC/5hmC/5fC/5caC[164,165]. However, it is still challenging to reliably differentiate among these modifications in real applications. • Be mindful of FP due to low abundance of these modifications.	• Analysis of synthetic DNA oligos suggested different signatures across 5mC/5hmC/5fC/5caC[30,222]. However, it is still challenging to reliably differentiate among these modifications in real applications. • Be mindful of FP due to their low abundance.	• LC-MS/MS • BS-seq • RE-seq • Nanopore sequencing • SMRT-seq • DIP-seq • ACE-seq[157] • 6-Letter sequencing[220] • A number of NGS-based methods based on antibody, chemistry or restriction enzymes treatment[85,157,212–220]
RNA modifications	• tRNA: 20% nucleotides are modified in tRNA and >50 unique modifications have been identified[17,136,137] • rRNA: rRNA also has substantial modifications with ~2% nucleotides being modified but	• Be mindful of nonconversion rate that can add to FP in BS-seq of RNA m5C. • Due to loss of input RNA material,	• The use of reverse transcriptase instead of DNA polymerase was reported to adapt the SMRT platform to directly sequence RNA. However, this	• Multiple methods available based on either base-calling errors or variation in electronic signals[124–128,174–178]. • Sophisticated positive and negative controls	• LC-MS/MS • RE-seq • RIP-seq • Nanopore

	Prevalence	Chemical-, enzyme- or antibody-based methods with NGS	Direct detection by SMRT sequencing	Direct detection by nanopore sequencing	Cross-validating methods
	with smaller diversity: dominance by 2'-OMe and ψ[138–140]. • mRNA: m6A is the most abundant form of mRNA modification, accounting for 0.1% −0.6% of all adenosines in mammalian mRNA[13,17,141,142]. Some of the other forms of modifications have much lower abundance[13,15,37,82].	RNAs with low relative abundance may be missing (FN). • Biases of antibody-, chemistry- and enzyme-based method across different sequence contexts and RNA structure. • Be mindful of FP for RNA modifications at low abundance.	strategy has not been widely followed up[122].	are required with diverse sequence contexts, expression and flanking RNA modifications. Otherwise, machine learning models based on electronical signal may result in FP or FN calls. • Use matched negative controls for FDR evaluation • Challenging to reliably differentiate among different modifications in real applications. • Be mindful of FP for RNA modifications at low abundance.	• A number of other NGS-based methods[15,37,82]

FDR, false discovery rate; FP, false positive; FN, false negative.

DIP-seq, DNA immunoprecipitation sequencing; RIP-seq, RNA immunoprecipitation sequencing; RE-seq, restriction enzyme (RE) based sequencing; SMRT-seq, single-molecule, real-time sequencing; BS-seq, bisulfite sequencing; EM-seq, enzymatic Methyl-seq; 4mC-TAB-seq, 4mC-TET-assisted bisulfite sequencing. TAPS-seq, TET-assisted pyridine borane sequencing; ACE-seq, APOBEC-coupled epigenetic sequencing.

6mA, N6-methyladenine; 4mC, N4-methylcytosine; 5mC, 5-methylcytosine; 5hmC, 5-hydroxymethylcytosine; 5fC, 5-formylcytosine; 5CaC, 5-carboxylcytosine.