JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS
Communication

Unnatural Cytosine Analogs Potentiate a Customizable, Enzymatic Method for Integrated Epigenetic and Four-Base Genetic Sequencing

Christian E. Loo, Jared B. Parker, Aleksia Barka, Ruiyao Zhu, Matthew R. Schnell, Rosella G. Cuomo, Heqiao Zhu, Laura Liu, Ashley Chen, Jasleen Gill, Wanding Zhou,* and Rahul M. Kohli*

Cite This: https://doi.org/10.1021/jacs.5c18450
Read Online

ABSTRACT: The interplay of genetic and epigenetic information shapes cell identity, development, and disease. However, standard methods for profiling DNA modifications (e.g., bisulfite sequencing) rely on selective C-to-T conversions, hindering the simultaneous examination of both genetic and epigenetic information. Here, we introduce Integrated Sequencing, which provides high-fidelity mapping of DNA modifications while preserving the native four-base genetic code in single DNA molecules. Integrated-Seq leverages the synthesis of a tethered copy strand with unnatural cytosine analogs that resist enzymatic conversion, combined with a novel DNA deaminase-helicase fusion that drives selective C-to-T conversion of natural cytosines in the original template strand. We demonstrate that Integrated-Seq is compatible with customizable enzymatic readouts to parse 5-methylcytosine and 5-hydroxymethylcytosine, and that preservation of the original four-base genetic code markedly improved enrichment, facilitating analysis of targeted genomic regions. Integrated-Seq thus provides a platform for simultaneous genetic and epigenetic analyses, paving the way for deep insights into fundamental biology and next-generation diagnostics.

Epigenetic DNA modifications to cytosine bases, such as 5-methylcytosine (5mC) and its Ten-Eleven Translocation enzyme (TET)-oxidized derivative, 5-hydroxymethylcytosine (5hmC), play crucial roles in regulating gene expression and cellular identity.[1−3] These modifications are established during differentiation and dynamically altered by biological cues. Consequently, mapping these modifications at base resolution offers powerful insights into development, health, and disease.

The longest-standing epigenetic sequencing approach is bisulfite sequencing (BS-Seq), whereby unmodified C bases are chemically deaminated to uracil and read as C-to-T transitions in sequencing. Although widely used, traditional BS-Seq significantly damages DNA, thereby raising input requirements.[4] As only unmodified C deaminates, BS-Seq also fails to resolve between functionally opposed 5mC and 5hmC marks.[5] Newer methods with BS-Seq variants or enzyme-based deamination have emerged to address some of these limitations.[6,7] APOBEC-coupled epigenetic sequencing (ACE-Seq) was the first method to use a DNA deaminase (APOBEC3A, A3A), selectively deaminating C and 5mC to specifically localize 5hmC alone.[8] Enzymatic methyl sequencing (EM-Seq) extended this approach by protecting 5mC from deamination via TET oxidation, providing a BS-Seq−like readout of 5mC/5hmC together (5modC).[9] While emerging methods such as nanopore can also localize epigenetic marks, enzymatic approaches maintain distinct advantages, given that third-generation sequencing requires higher DNA input and shows limitations resolving 5mC from 5hmC.[10,11]

Many of these established methods, however, also face critical limitations due to their reliance on selective C-to-T conversion. Most notably, this conversion largely changes the four-base genome into three bases, with reduced sequence complexity hindering read mapping against reference genomes and impeding probe-based enrichment of target regions, as three-base probes lose specificity.[12,13] Methods have recently been developed to address these limitations, preserving genetic information in a "copy strand", while still relying on selective C-to-T conversion of the original DNA template (see Figure 1a, Figure S1).[14−16] In both Methyl-SNP-Seq and Five-Letter-Seq, a hairpin adaptor templates complementary strand synthesis before C-to-T conversion, thereby linking two strands within a single DNA molecule. Directional paired-end sequencing then provides epigenetic information from the original strand (read1) and genetic information from the copy strand (read2). Despite this conceptual advance, both methods remain constrained. Methyl-SNP-Seq employs 5mC during copying followed by bisulfite treatment, thus inheriting drawbacks associated with BS-Seq. Five-Letter-Seq instead copies with unmodified C, followed by enzymatic modification and deamination. Although the method is non-destructive, restoring the genetic code requires complex decoding as both strands are deaminated; moreover, loss of the four-base sequence could impair enrichment. Most recently, methyl-CODEC explored an alternative duplex sequencing workflow

Received: October 19, 2025
Revised: January 25, 2026
Accepted: February 26, 2026

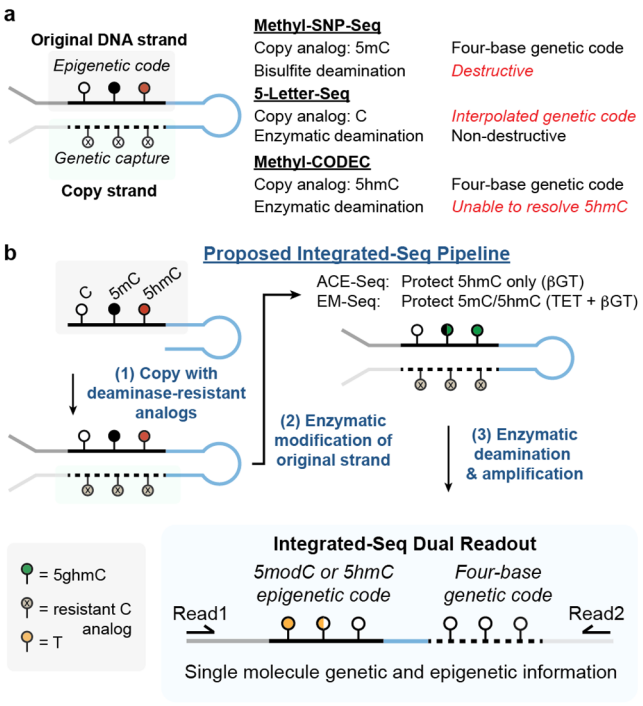

Figure 1. Recent and proposed methods for simultaneous genetic and epigenetic sequencing. (a) A hairpin adaptor and a synthetic copy strand have been exploited for dual sequencing readouts, with noted limitations. (b) Integrated-Seq employs a deaminase-resistant copy strand and distinct enzymatic steps to yield single DNA molecules with linked epigenetic and four-base genetic sequences.

that uses 5hmC for strand-displacing extension. However, this method still relies on reference mapping of converted reads as part of decoding pipeline, and it remains unable to distinguish 5hmC from 5mC.[16]

To overcome these substantial limitations, we sought to develop Integrated Sequencing as a platform for precise, non-destructive mapping of epigenetic modifications, while preserving native four-base genetic information (Figure 1b). In conceiving Integrated-Seq, we drew upon the copy strand strategy but recognized that significant improvements could be achieved by addressing two unmet needs. First, deriving a deaminase-resistant copy strand could potentially preserve the *unaltered* four-base genetic code, akin to Methyl-SNP-Seq, while avoiding destructive chemical deamination. Second, if the copy strand could resist enzymatic modifications, we could enable compatibility with both ACE-Seq and EM-Seq to newly facilitate mapping of *either* 5hmC or 5modC. To realize Integrated-Seq, we thus pursued two goals: (Objective 1) deriving a copy strand strategy to faithfully maintain the four-base genetic code in read2; and (Objective 2) establishing an adaptable pipeline for the selective detection of either 5hmC or 5modC with high fidelity in read1.

To establish our methodology, we employed a mixture of mouse fibroblast genomic DNA (gDNA) combined with controls containing known modifications: unmodified C-containing λ-phage gDNA and a 5mCpG-modified plasmid. Our first task was to develop a robust library construction protocol, integrating precedents from Methyl-SNP-Seq and Five-Letter-Seq.[14,15] This protocol involves hairpin ligation, copy-strand generation, ligation of a deaminase-resistant Y-shaped adaptor, enzymatic deamination, and amplification to generate single molecules with linked strand information. Notably, as A3A is single-stranded DNA specific, we also initially adopted the Five-Letter-Seq strategy of helicase co-incubation, PcrA in our case, henceforth referred to as PcrA + A3A *in trans*.

With a robust library generation protocol established, we turned to Objective 1. Prior work had established potential competing requirements to navigate: structure-function studies suggested that steric bulk at the C5-position of cytosine can help resist A3A deamination;[17,18] however, not all analogs would likely support the overall pipeline.[19,20] We thus broadly considered analogs with different attributes (Figure 2a), including the natural bases C, 5mC, 5hmC, and 5-carboxylcytosine (5caC). For 5hmC, we also tested modification by β-glucosyltransferase (βGT) after copying to make glucosylated-5hmC (g-5hmC). Since bulky unnatural analogs could confer distinct advantages, we also considered 5-vinylcytosine (5vC), 5-ethynylcytosine (5eyC), 5-propynylcytosine (5pyC), and pyrrolocytosine (pyrC) as candidates. Using homogeneously-modified amplicons, we verified by qPCR that all four unnatural analogues did not impact amplification efficiency relative to unmodified C nor result in significant misincorporation events (Figure S2). When amplicons were subjected to A3A treatment, however, we observed that the smaller 5vC and 5eyC are partially deaminated, while 5pyC and pyrC desirably resist deamination. After further validating deamination-resistance of 5pyC and pyrC using mass spectrometry on synthetic oligonucleotides, we nominated these two unnatural analogs for further exploration (Figure S3).

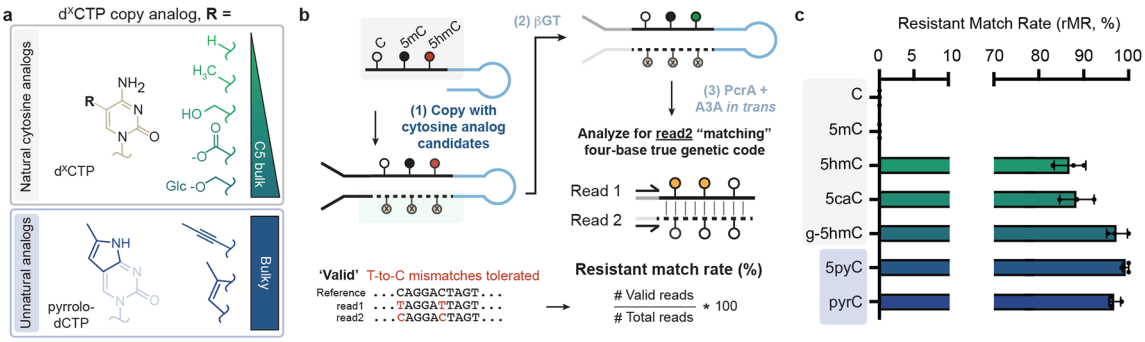

Figure 2. Resistant match rates (rMRs) for accurate genetic readout. (a) Natural and unnatural cytosine analogs examined are depicted. (b) Integrated-Seq libraries were prepared with different copy analogs and parsed for resistant matches using the provided scheme and criteria. (c) rMRs for each analog. rMR % were normalized to the maximum across samples (n = 3, mean, standard deviation shown).

https://doi.org/10.1021/jacs.5c18450
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

Returning to the hairpin-based pipeline, when unmodified C was replaced with any of the four natural or two nominated unnatural analogues in the copy strand, library generation was supported; however, 5caC resulted in lower yields, consistent with a known impact on polymerases (Figure S4).[19] To characterize the resulting libraries, we defined a "resistant match rate" (rMR) metric. Reads were classified as "valid" if read1 and read2 aligned, allowing anticipated T-to-C mismatches, and if read2 contained an unconverted four-base sequence (Figure 2b). Beginning with either C or 5mC copying, we observed undetectable matched reads (rMR = 0%, Figure 2c), consistent with deamination of both strands, leading to a converted read2. With bulkier natural analogues, we observed higher rMRs, consistent with prior studies on A3A reactivity.[8,17] Most interestingly, in addition to the natural g-5hmC, unnatural 5pyC and pyrC showed the highest rMRs, suggesting that these analogues can support strand copying, resist enzymatic deamination, and accurately capture the four-base genetic sequence in read2.

Having identified analogs that retain the genetic code, we turned to Objective 2. We initially focused on Integrated-ACE-Seq, which requires deamination of both unmodified C and 5mC, as no prior method has localized 5hmC while maintaining the intact genetic sequence in read2. To this end, we prepared Integrated-ACE-Seq libraries using the most promising copy analogues, g-5hmC, 5pyC, or pyrC, and compared read1 conversion efficiencies to standard ACE-Seq (Figure 3). For unmodified λ-phage gDNA, we observed near-complete deamination (~99.9%) independent of the copy analogue, consistent with traditional ACE-Seq; however, with 5mCpGs, lower conversion efficiencies (~93–96%) were evident. We therefore undertook a systematic reaction parameter optimization campaign, testing different enzyme preparations, helicase:A3A ratios, and buffer conditions, but only achieved marginal improvements. This represented a major limitation, given the stringent requirements for a low false discovery rate (FDR) for 5hmC detection.

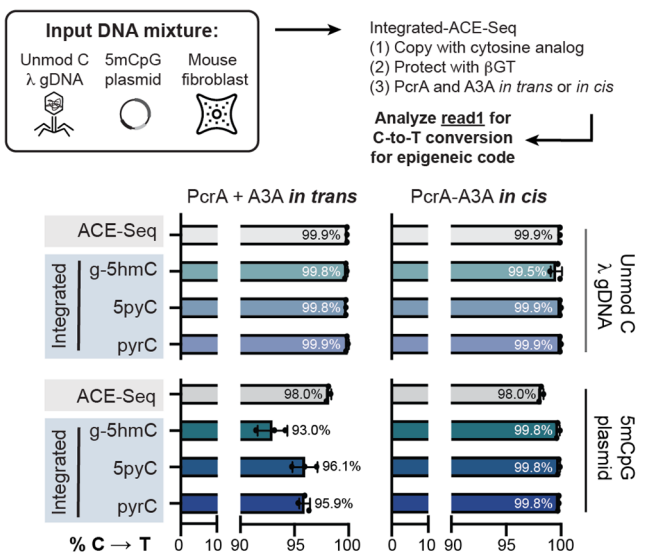

Figure 3. Epigenetic accuracy with Integrated-ACE-Seq. Input DNA was processed using either standard or Integrated-ACE-Seq. Shown are the C-to-T conversion rates for unmodified C or 5mCpG-containing control DNAs with either *in trans* or *in cis* helicase and deaminase (n = 3, mean, standard deviation shown).

We postulated that suboptimal 5mC conversion could arise from kinetic uncoupling of PcrA-mediated DNA unwinding and A3A-mediated deamination. This hypothesis led us to draw on engineering strategies from the genome-editor field where fusions of DNA modifying enzymes are commonly employed.[21] However, we recognized that a helicase-deaminase fusion would be prohibitive for overexpression due to inherent mutagenicity. We thus employed a novel intein strategy whereby PcrA and A3A are expressed separately and linked together *in vitro* using the ultrafast intein pair Cfa^N/Cfa^C, creating a fusion we term PcrA-A3A *in cis*.[22] Using model duplex substrates with different overhangs, we first validated that PcrA-A3A *in cis* showed higher activity relative to that *in trans* (Figure S5). Given these promising results, we employed the fusion in Integrated-ACE-Seq and found significantly improved 5mC deamination (99.8%) compared to PcrA + A3A *in trans* (Figure 3). Interestingly, while rMR values remained consistently high for 5pyC and pyrC, they dropped for g-5hmC (Figure S6), potentially due to impacts on helicase unwinding with bulky g-5hmC.

Advancing toward Objective 2, we next moved from Integrated-ACE-Seq to Integrated-EM-Seq, incorporating a TET oxidation step to localize 5modC. We first assessed the fidelity of 5modC calling by comparing Integrated-EM-Seq to traditional EM-Seq. As with Integrated-ACE-Seq, we observed near-complete deamination (~99.9%) of unmodified λ-phage gDNA (Figure S7). With 5mCpG-control DNA, we observed that, as expected, 5mCs were now efficiently protected from A3A-deamination, allowing accurate 5modC detection.

Our systematic approach thus highlighted two key innovations, deaminase-resistant unnatural analogues and a helicase-deaminase fusion, that together achieved our objectives. Having established the experimental foundation, we developed a custom computational pipeline, Epimatcher, to facilitate the downstream analysis. Unlike BS-Seq which relies on reference genome mapping for modification extraction, Epimatcher reconstructs the four-base sequence by performing a consensus alignment between read1 and the deaminase-resistant copy in read2, using base-quality-based rules and C-to-T-aware scoring, intentionally performed prior to reference alignment. The resulting four-base consensus sequences with cytosine modification states are then mapped using standard four-letter alignment, thereby improving both accuracy and mappability (Table S1).

With the platform established (Scheme S1), we next applied Integrated-Seq to profile diverse tissues. We isolated gDNA from mouse cerebellum, heart, kidney, and blood tissues, where 5hmC levels are known to span a wide range, and prepared matched standard ACE- and EM-Seq libraries for comparison to those from Integrated-Seq. We performed low-coverage sequencing,[23] which accurately estimates global modification levels, and used Epimatcher to quantify 5hmC or 5modC. In the raw reads, the distinctive advantage of Integrated-Seq was evident, as the original four-base genome and epigenetic status were directly observable off the sequencing instrument (Figure S8), while maintaining similar read retention and library yield across methods (Figure S9). Across tissues, detection of 5hmC by Integrated-ACE-Seq and 5modC by Integrated-EM-Seq strongly correlated with both their standard counterparts (R^2 = 0.99, Figure 4a) as well as previously reported measurements using orthogonal methods (Figure S10).[24,25] Subsequent high-depth sequencing of cerebellar samples showed strong correlation between tradi-

https://doi.org/10.1021/jacs.5c18450
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

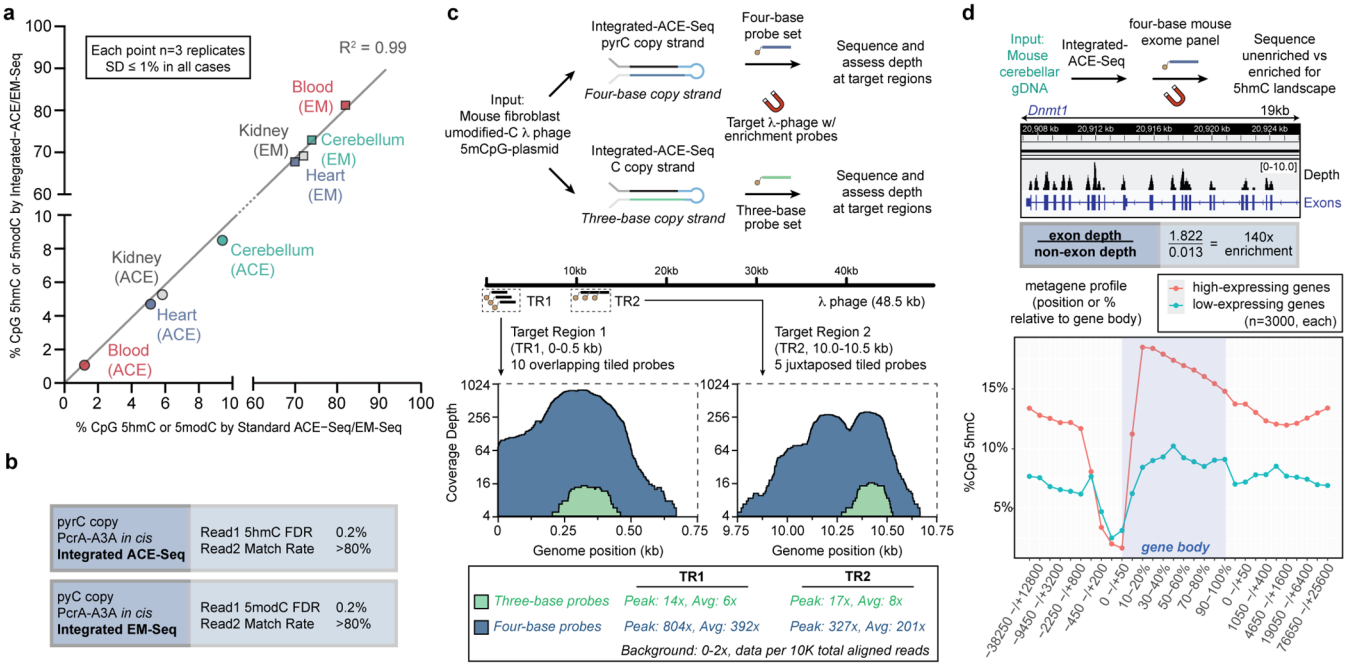

Figure 4. Global and targeted performance of Integrated-Seq. (a) Global 5hmC detection by Integrated-ACE-Seq and 5modC detection by Integrated-EM-Seq correlates highly with their standard ACE-Seq and EM-Seq counterparts across different tissues (n = 3, mean shown, standard deviation smaller than symbol). (b) Both Integrated-ACE-Seq and Integrated-EM-Seq achieve low FDR rates with high match rates. (c) (Top) Workflow schematic for comparison of three-base or four-base probe set targeting regions (TR1, TR2) of the λ-phage genome. (Bottom) Plotted are the post-enrichment per base coverage depth in TR1/TR2 with either three-base or four-base probe sets. (d) (Top) Schematic for exonic enrichment of Integrated-ACE-Seq cerebellar library using a commercial mouse four-base probe panel. (Middle) Representative IGV genome coverage track with read depth and exon structure of genes shown. (Bottom) Metagene plot showing 5hmC content relative to gene body for 3000 highest- and lowest-expressing genes.

tional and Integrated-ACE-Seq, with similar genome-wide coverage patterns and greater read retention (Figure S11 and Table S2). Notably, across all samples, Integrated-Seq consistently achieved high match rates and fidelity, as reflected in the processing metrics (>80% match rate, 0.2% FDR; see Figure 4b, as well as Tables S1 and S2), highlighting versatility and accuracy.

Having applied Integrated-Seq in genome-wide profiling, we next explored whether its four-base genome preservation enables improved targeted enrichment, as many biological questions center on studying epigenetic states of specific loci. Typically, targeted enrichment after C-to-T conversion requires specialized probes with degenerate bases to accommodate the largely three-base genome. To explore this feature, we generated a four-base Integrated-ACE-Seq library using pyrC copying and a three-base version using unmodified C copying and then performed targeted enrichment of two 500-bp regions in λ-phage gDNA using either a four-base or a three-base probe set, respectively (see Figure 4c, Figure S12). Focusing on the two tiled regions of interest, when overall reads were matched, the three-base probe resulted in an average 6- and 8-fold coverage depth, while the four-base probes yielded 392- and 201-fold coverage, a 60- and 25-fold increase attributable to the Integrated-Seq exact four-base genetic copy. Given the opportunities offered by four-base enrichment, we returned to the mouse cerebellum Integrated-ACE-Seq library and enriched using a commercially available four-base mouse exome panel, finding a 140-fold enrichment of exonic regions over the background (Figure 4d). Separating high- and low-expressing genes (n = 3000), metagene plots revealed higher 5hmC in highly expressed genes, with a conserved pattern of 5hmC depletion at promoters, followed by intragenic 5hmC peaking in 5′-coding regions. Thus, Integrated-Seq is fully compatible with four-base enrichment panels, simplifying probe design and synthesis for scalable epigenetic profiling of targeted genomic regions.

Taken together, we establish Integrated-Seq as a non-destructive method for combined epigenetic and four-base genetic profiling. This fully enzymatic workflow represents a notable advance due to its accommodation of low-input samples (unlike Methyl-SNP-Seq), modularity in resolving either 5hmC or 5modC (unlike Methyl-CODEC), and ability to support efficient four-letter–based enrichment (unlike Five-Letter-Seq). Moreover, in principle, Integrated-Seq could be performed on longer DNA molecules and analyzed with third-generation platforms to overcome the current challenges with low input compatibility and sparse DNA modification detection. These attributes position Integrated-Seq for high-value applications such as linking allele-specific methylation to genetic polymorphisms (SNPs), coupling analysis of genetic and epigenetic alterations in CRISPR screens, and, most significantly, advancing cell-free DNA diagnostics by decoding nascent genetic and epigenetic marks detectable early in cancer development.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.5c18450.

Detailed methods, schematics and sequencing library metrics (PDF)

https://doi.org/10.1021/jacs.5c18450
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Authors

Wanding Zhou − *Department of Pathology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States; Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, United States*; Email: zhouw3@chop.edu

Rahul M. Kohli − *Department of Medicine and Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*; orcid.org/0000-0002-7689-5678; Email: rkohli@pennmedicine.upenn.edu

Authors

Christian E. Loo − *Graduate Group in Biochemistry, Biophysics, and Chemical Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Jared B. Parker − *Department of Medicine and Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Aleksia Barka − *Graduate Group in Biochemistry, Biophysics, and Chemical Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Ruiyao Zhu − *Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Matthew R. Schnell − *Graduate Group in Biochemistry, Biophysics, and Chemical Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Rosella G. Cuomo − *Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Heqiao Zhu − *Department of Pathology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States; Center for Computational and Genomic Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, United States*

Laura Liu − *Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*; orcid.org/0000-0003-3120-149X

Ashley Chen − *Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*

Jasleen Gill − *Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, United States*; orcid.org/0000-0003-3091-3102

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.5c18450

Notes

The authors declare the following competing financial interest(s): The University of Pennsylvania has patents pending for some of the materials and protocols with C.E.L., R.M.K., R.G.C., and J.B.P. as inventors. As of November 2025, C.E.L. has served as a scientific advisor to biomodal.

ACKNOWLEDGMENTS

We thank Thomas Muir for the gift of the engineered intein backbone constructs (Addgene plasmids 122949 and 126018). We are grateful to Jenna N. Beyer, George M. Burslem, and Mark S. Dillingham for helpful discussions; and Brooke E. Adams, Eleanor T. Daniels, Alexander J. Krosky, Jacob M. Seiple, and Sol Moe Lee for technical assistance. This work was funded in part by the National Institute of Health grants R35-GM146978 to W.Z., R01-HG010646 to R.M.K., the Trans-disciplinary Program in Translational Medicine and Therapeutics (UL1-TR001878), and F31-HG012892 to C.E.L.

REFERENCES

(1) Schubeler, D. Function and information content of DNA methylation. *Nature* **2015**, *517*, 321−326.

(2) Tahiliani, M.; Koh, K. P.; Shen, Y.; et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **2009**, *324*, 930−935.

(3) Kohli, R. M.; Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **2013**, *502*, 472−479.

(4) Tanaka, K.; Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 1912−1915.

(5) Huang, Y.; Pastor, W. A.; Shen, Y.; Tahiliani, M.; Liu, D. R.; Rao, A. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* **2010**, *5*, No. e8888.

(6) Wang, T.; Loo, C. E.; Kohli, R. M. Enzymatic approaches for profiling cytosine methylation and hydroxymethylation. *Mol. Metab.* **2022**, *57*, 101314.

(7) Dai, Q.; Ye, C.; Irkliyenko, I.; Wang, Y.; Sun, H.-L.; Gao, Y.; Liu, Y.; Beadell, A.; Perea, J.; Goel, A.; He, C. Ultrafast bisulfite sequencing detection of 5-methylcytosine in DNA and RNA. *Nat. Biotechnol.* **2024**, *42*, 1559.

(8) Schutsky, E. K.; DeNizio, J. E.; Hu, P.; et al. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat. Biotechnol.* **2018**, *36*, 1083−1090.

(9) Vaisvila, R.; Ponnaluri, V. K. C.; Sun, Z.; et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* **2021**, *31*, 1280−1289.

(10) Searle, B.; Müller, M.; Carell, T.; Kellett, A. Third-Generation Sequencing of Epigenetic DNA. *Angew. Chem., Int. Ed. Engl.* **2023**, *62*, No. e202215704.

(11) Kong, Y.; Zhang, Y.; Mead, E. A.; et al. Critical assessment of nanopore sequencing for the detection of multiple forms of DNA modifications. *bioRxiv* **2024**, *11* (19), 624260.

(12) Wulfridge, P.; Langmead, B.; Feinberg, A. P.; Hansen, K. D. Analyzing whole genome bisulfite sequencing data from highly divergent genotypes. *Nucleic Acids Res.* **2019**, *47*, No. e117.

(13) Zhou, W.; Johnson, B. K.; Morrison, J.; et al. BISCUIT: an efficient, standards-compliant tool suite for simultaneous genetic and epigenetic inference in bulk and single-cell studies. *Nucleic Acids Res.* **2024**, *52*, No. e32.

(14) Fullgrabe, J.; Gosal, W. S.; Creed, P.; Liu, S.; Lumby, C. K.; Morley, D. J.; Ost, T. W. B.; Vilella, A. J.; Yu, S.; Bignell, H., et al. Simultaneous sequencing of genetic and epigenetic bases in DNA. *Nat. Biotechnol.* **2023**, *41*, 1457.

(15) Yan, B.; Wang, D.; Vaisvila, R.; Sun, Z.; Ettwiller, L. Methyl-SNP-seq reveals dual readouts of methylome and variome at molecule resolution while enabling target enrichment. *Genome Res.* **2022**, *32*, 2079−2091.

(16) Liu, R.; Darbeheshti, F.; Walsh, L.; et al. Methyl-CODEC enables simultaneous methylation and duplex sequencing. *Nucleic Acids Res.* **2025**, *53*, gkaf482.

(17) Schutsky, E. K.; Nabel, C. S.; Davis, A. K. F.; DeNizio, J. E.; Kohli, R. M. APOBEC3A efficiently deaminates methylated, but not TET-oxidized, cytosine bases in DNA. *Nucleic Acids Res.* **2017**, *45*, 7655−7665.

(18) Wang, T.; Fowler, J. M.; Liu, L.; et al. Direct enzymatic sequencing of 5-methylcytosine at single-base resolution. *Nat. Chem. Biol.* **2023**, *19*, 1004−1012.

(19) Kellinger, M. W.; Song, C. X.; Chong, J.; Lu, X. Y.; He, C.; Wang, D. 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **2012**, *19*, 831−833.

(20) Ghanty, U.; DeNizio, J. E.; Liu, M. Y.; Kohli, R. M. Exploiting Substrate Promiscuity to Develop Activity-Based Probes for TET Family Enzymes. *J. Am. Chem. Soc.* **2018**, *140*, 17329−17332.
(21) Anzalone, A. V.; Koblan, L. W.; Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime. *Nat. Biotechnol.* **2020**, *38*, 824−844.
(22) Stevens, A. J.; Brown, Z. Z.; Shah, N. H.; Sekar, G.; Cowburn, D.; Muir, T. W. Design of a Split Intein with Exceptional Protein Splicing Activity. *J. Am. Chem. Soc.* **2016**, *138*, 2162−2165.
(23) Loo, C. E.; Fowler, J. M.; Zhu, H.; et al. Sparse Sequencing permits accurate and efficient quantification of genome-wide cytosine modification levels. *bioRxiv* **2025**, *01* (08), 631958.
(24) Lee, S. M.; Goldberg, D. C.; Cloud, C.; Parker, J. B.; Krapp, C.; Loo, C. E.; Kim, E.; Zhao, I.; Jin, C.; Porecha, R.; et al. A ternary-code DNA methylome atlas of mouse tissues. *Genome Biol.* **2025**, *26*, 343.
(25) Globisch, D.; Munzel, M.; Muller, M.; et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **2010**, *5*, No. e15367.

https://doi.org/10.1021/jacs.5c18450
J. Am. Chem. Soc. XXXX, XXX, XXX−XXX