Communications Biology

https://doi.org/10.1038/s42003-026-09849-8

Article in Press

Z-Calling: a tool for A/Z (2,6-diaminopurine) base calling and dZ-DNA detection using PacBio HiFi reads

Received: 3 June 2025

Accepted: 2 March 2026

Cite this article as: Wu, B., Chen, Y., Zhou, Y. *et al.* Z-Calling: a tool for A/Z (2,6-diaminopurine) base calling and dZ-DNA detection using PacBio HiFi reads. *Commun Biol* (2026). https://doi.org/10.1038/s42003-026-09849-8

Bo Wu, Ying Chen, Yan Zhou, Longjian Niu, He-Xu Chen, Yating Li, Jia-Yong Zhong, Suwen Zhao, Wei Chi, Yan Zhang & Chuan-Le Xiao

Z-Calling: a tool for A/Z (2,6-diaminopurine) base calling and dZ-DNA detection using PacBio HiFi reads

Bo Wu[1,2,#], Ying Chen[2,#], Yan Zhou[3,#], Longjian Niu[2,#], He-Xu Chen[4,#], Yating Li[3], Jia-Yong Zhong[2], Suwen Zhao[5-7,*], Wei Chi[2,*], Yan Zhang[8-10,*], and Chuan-Le Xiao[1,*]

[1] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou 510060, China

[2] Shenzhen Eye Hospital, Shenzhen Eye Medical Center, Southern Medical University, 18 Zetian Road, Futian District, Shenzhen 518040, China

[3] Jiangsu Key Laboratory of Zoonosis, Yangzhou University, Yangzhou 225009, China

[4] School of Artificial Intelligence, Sun Yat-Sen University, Zhuhai 519000, China

[5] iHuman Institute and School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China

[6] Shanghai Key Laboratory of High-resolution Electron Microscopy, ShanghaiTech University, Shanghai 201210, China

[7] Shanghai Clinical Research and Trial Center, Shanghai 201210, China.

[8] Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), School of Chemical Engineering and Technology, Tianjin University, Tianjin, China

[9] New Cornerstone Science Laboratory, School of Pharmaceutical Science and Technology, Tianjin University, Tianjin 300072, China

[10] Tianjin Key Laboratory for Modern Drug Delivery & High-Efficiency, Collaborative Innovation Center of Chemical Science and Engineering, School of Pharmaceutical Science and Technology, Faculty of Medicine, Frontiers Science Center for Synthetic Biology (Ministry of Education), and Key Laboratory of Systems Bioengineering (Ministry of Education), School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China

*To whom correspondence should be addressed:

Suwen Zhao. Email: zhaosw@shanghaitech.edu.cn

Wei Chi. Email: dr_chiwei@163.com;

Yan Zhang. Email: yan.zhang@tju.edu.cn;

Chuan-Le Xiao. Email:xiaochuanle@126.com;

#These authors contributed equally.

Abstract

The natural occurrence of 2,6-diaminopurine (Z) as a substitute for adenine (A) in certain bacteriophage genomes has profound evolutionary implications and promising biotechnological potential. Progress in this field, however, has been stymied by the absence of reliable methods to detect dZ-DNA, particularly in mixed samples, and to distinguish A from Z at the single-nucleotide level. Here, we introduce Z-Calling, a machine learning-based tool designed to identify dZ-DNA and discriminate A/Z bases directly from PacBio Circular Consensus Sequencing (CCS) reads without additional processing. By analyzing sequence context-dependent kinetic signal changes induced by Z/A substitution, Z-Calling achieves exceptional sensitivity, reliably detecting dZ-DNA even in samples with as little as ~1% dZ-DNA content. Its A/Z base-calling module demonstrates robust performance, with AUC scores of 0.942–0.952 across diverse DNA sequence contexts. Z-Calling represents a significant advancement in accessible and accurate dZ-DNA sequencing, paving the way for its broader application in biotechnology. Z-Calling is freely available at https://github.com/xiaochuanle/Z-Calling .

Introduction

The Watson-Crick base pairing framework maintains DNA's iconic double-helix structure[1] and governs genetic inheritance in nearly all lifeforms. Yet nature defies simplicity: certain viruses, such as the cyanophage S-2L, have evolved to completely replace adenine with 2,6-diaminopurine (Z)—a base that violates conventional pairing rules[2]. Z is distinguished from adenine by an additional amino group at the 2-position of its purine ring, enabling three hydrogen bonds with thymine (T). This enhanced bonding confers greater duplex stability and distinct biophysical properties, positioning dZ-DNA as a functionally unique genetic polymer[3-8]. Collective efforts have characterized the biosynthetic pathways responsible for dZ-DNA synthesis, revealing its widespread distribution and evolutionary role in phage-host dynamics[9-12]. Z incorporation has been shown to enhance performance in diverse applications by leveraging its unique pairing stability and structural properties. For instance, Z-modified toehold probes have been engineered to significantly improve the sensitivity of single-nucleotide variant detection in genotyping assays[6]. In the field of gene editing, Z bases have been utilized to modulate the cleavage activity of CRISPR effectors through non-Watson-Crick base pairing mechanisms[13]. Additionally, substituting adenine with Z in mRNA therapeutics has demonstrated the ability to minimize immunogenicity while enhancing translational capacity, offering alternative strategies for synthetic vaccine design [14]. Beyond therapeutics, the increased stability of dZ-DNA could also be explored for molecular data storage solutions[15] in the future. However, the absence of methods to unambiguously detect dZ-DNA or discriminate A and Z at single-nucleotide resolution remains a major barrier to elucidating its biological functions and exploiting its technological potential.

Accurately detecting dZ-DNA is critical for understanding its evolutionary role and unlocking its biotechnological potential. Natural dZ-DNAs have been reported exclusively in phage genomes that employ specialized enzymes[9,10,12,16] to ensure complete substitution of A with Z. Although over 2,000 viruses have been predicted to harbor dZ-DNA through PurZ homology analyses[9,11,17], conclusive proof of dZ-DNA genomes remains absent for most of them. Current verification methods for phage dZ-DNA rely heavily on techniques like high-performance liquid chromatography (HPLC), which measures nucleotide composition—including the presence of dZ—in fragmented DNA[9-12]. While HPLC quantifies dZ abundance, it fails to pinpoint its exact locations within sequences, a critical gap given that many uncultured viral genomes reside in

complex environmental mixtures. Recent studies reveal that nanopore sequencing, a widely used third-generation sequencing technique, exhibits markedly reduced accuracy across A/T/C/G bases when analyzing dZ-DNA (Z treated as A) compared to canonical DNA[9]. However, such signals remain circumstantial, as similar inaccuracies could arise from unrelated factors like chemical modifications, unknown nucleotides, or sample impurities. Crucially, no robust, user-friendly method exists to definitively identify dZ-DNA within mixed samples containing canonical DNA. This technological void hinders efforts to explore dZ-DNA's biological roles or leverage its unique properties for applied science.

The capacity to resolve Z bases within hybrid dZ-DNA molecules (where A and Z coexist) could unlock groundbreaking applications in synthetic biology, precision medicine, advanced materials[9,18], and molecular computing[15]. Yet this promise remains shackled by technological limitations. Current sequencing platforms fail to reliably distinguish Z from A. First- and next-generation methods, reliant on PCR amplification and fluorescent dNTPs, fundamentally misread Z as A during fluorescence signal interpretation[19,20]. Emerging third-generation technologies offer glimmers of hope. Nanopore sequencing developed by Oxford Nanopore Technologies (ONT) bypasses amplification and instead detects nucleotide-specific perturbations in ionic currents[21]. While theoretically capable of distinguishing Z from A, this approach may require adapted nanopore proteins for optimized signal/noise ratio on dZ-DNA, alongside specialized base-calling models trained on dZ-DNA signatures[22]. PacBio CCS presents a paradox: though PCR-based workflows misidentify Z as A in fluorescence readouts, the platform uniquely records kinetic signals including pulse width (PW) and inter-pulse duration (IPD)[23]. These signals have proven effective for detecting modified bases like 5mC[24-26] and 6mA[27-29], yet PacBio's potential to decode dZ-DNAs—particularly Z bases that form non-Watson-Crick pairing—remains unexplored.

To overcome these limitations, we developed Z-Calling, a machine learning-based tool that leverages PacBio CCS to detect dZ-DNA molecules and to discriminate Z and A bases at the single-nucleotide level. By analyzing polymerase-derived signals—pulse width (PW) and inter-pulse duration (IPD)—across matched dZ-DNA and unmodified DNA sequences, we revealed that Z incorporation generates reproducible, context-specific kinetic signatures. These signatures enable two-tier discrimination: first, classifying dZ-DNA versus canonical DNA, and second, distinguishing Z from A bases within hybrid or complex metagenomic samples. Z-

Calling achieves high sensitivity and precision in dZ-DNA detection, with A/Z base-calling accuracy rivaling state-of-the-art tools for canonical epigenetic modifications (e.g., 5mC or 6mA)[26,27]. This method establishes a generalizable platform for decoding dZ-DNA in diverse contexts, unlocking its potential for engineered biosystems, molecular tagging, and beyond.

Results

Z base alters polymerase kinetics in a context-dependent manner

The structural distortion caused by Z bases (**Figure 1a**)—deviating from canonical Watson-Crick pairing—potentially affects polymerase activity during DNA amplification, creating kinetic signatures that could enable dZ-DNA identification and base discrimination. To test this hypothesis, we constructed comparative PacBio CCS datasets using both Revio and Sequel II platforms. These included canonical DNA and dZ-DNA synthesized by replacing dATP with dZTP during PCR amplification of fruit fly DNA templates. To be noticed, there are three types of DNAs mentioned in this study: canonical DNAs, hybrid DNAs, and dZ-DNAs, representing DNAs with no, partial, and complete dA/dZ substitution, respectively.

The incorporation of Z bases raises questions about potential sequencing biases. Since Z bases pair with thymine (T) through hydrogen bonding (**Figure 1a**), PacBio CCS technology—which uses fluorescent dNTP labeling for base identification (**Figure 1b**)—naturally interprets Z as adenine (A) during sequencing. To quantify whether this substitution compromises sequencing fidelity, we analyzed error rates in dZ-DNA versus canonical DNA using a k-mer-based alignment strategy[30], treating Z as A and benchmarking against native DNA controls. The results showed that dZ-DNA reads with ≥3 sequencing passes achieved >99% base accuracy (QV=20), showing only marginal reductions compared to canonical DNA across both Revio and Sequel II platforms (**Supplementary Note 1**). This robustness stems from CCS's iterative design: repeated bidirectional sequencing of individual molecules progressively corrects errors (**Figure 1c**; **Supplementary Figure 1**). Accuracy plateaued at ~99.9% (QV=30) after ~20 passes (ten full cycles) across both Revio and Sequel II platforms (**Figure 1c**; **Supplementary Data 1**). Notably, hybrid and fully substituted dZ-DNAs exhibited comparable precision (**Supplementary Figure 2; Supplementary Data 2**), demonstrating that A-to-Z replacement minimally impacts sequencing reliability under these conditions.

To evaluate how dA/dZ substitution influences polymerase kinetics, we compared PacBio CCS data from dZ-DNA and canonical DNA (**Supplementary Note 2**). The platform measures two kinetic parameters: pulse width (PW), reflecting nucleotide incorporation time, and inter-pulse duration (IPD), representing pauses between incorporations (**Figure 1d**). By analyzing paired 21-mers differing solely at their central A/Z positions, we minimized interference from neighboring bases. Unlike epigenetic modifications such as 5mC/6mA, which primarily alter IPD[23,28], Z substitution predominantly affects PW within an asymmetric 8-base window spanning one base upstream to seven bases downstream (positions 10–17, **Figure 1e**; **Supplementary Figure 3a-d**; **Supplementary Data 3**). Notably, these PW shifts are highly sequence-dependent, with elongation or shortening varying by context. For instance, at base position 12 in **Figure 1e**, PW was significantly elongated (1.46 ± 0.12-fold, p=2.75e-306 by Wilcoxon test) in the TTZTT context compared to TTATT, whereas significantly shortened (0.84 ± 0.08-fold, p=3.95e-48) in TTZGG compared to TTAGG. Though less frequent, significant IPD changes observed in a small proportion of contexts also followed context-dependent trends (**Supplementary Figure 3a, b**; **Supplementary Data 3**).

While these advances provide critical insights, reliably distinguishing A and Z bases using kinetic signals continues to present challenges. Our analysis demonstrates that kinetic profiles at identical base positions show substantial overlap across experimental conditions (**Supplementary Figure 4**), with differences between A and Z contexts often limited to modest 0.5–1.5-fold shifts in signal magnitude (**Supplementary Figure 5**). Furthermore, dimensionality reduction techniques, such as PCA and t-SNE, failed to resolve clear boundaries between canonical DNA and dZ-DNA samples (**Supplementary Figure 6**). To address these limitations, we pivoted to supervised machine learning frameworks to improve classification accuracy within CCS datasets.

Design of Z-Calling

To assess the contribution of kinetic signals in distinguishing A and Z bases, we trained Random Forest Classification (RFC) models on 21-mer sequence contexts with fixed central A/Z motifs (3 bp). These models achieved an average accuracy of 0.842 ± 0.018 (**Supplementary Data 4**), with pulse width (PW) emerging as the dominant discriminatory feature. Hierarchical

clustering of the feature importance matrix revealed that the influence of positional kinetic signals depends on sequence context (**Figure 2a**). Notably, the nucleotide immediately downstream of the target A/Z base in the template strand exerted the strongest influence, as 3-mer contexts sharing the same downstream base clustered together.

Building on these insights and the need for practical applications, we developed Z-Calling, a computational tool with two primary functions (**Figure 2b**): (1) single-nucleotide resolution base calling of A and Z bases in PacBio CCS reads, and (2) classification of dZ-DNA and canonical DNA reads. Z-Calling employs Multilayer Perceptron (MLP) neural networks for A/Z base calling. Features for each target base include one-hot encoded context k-mers, where each nucleotide is represented by a 4-element vector, along with normalized IPD and PW signals. The MLP architecture consists of an input layer, two hidden layers, and an output layer that calculates a Z probability score for each input feature (**Figure 2c**). We selected the MLP architecture because it consistently outperformed Random Forest baselines and matched the accuracy of more complex deep learning networks in our benchmarks (**Supplemenatry Note 3**; **Supplementary Figure 7**). Furthermore, optimization of the network topology revealed a consistent performance plateau after two hidden layers, confirming that this compact architecture is sufficient to capture the discriminative kinetic signals (**Supplementary Figures 8,9**). To enable accurate A/Z base calling across diverse contexts, we trained the MLP model using a mixed dataset of canonical DNAs, dZ-DNAs, and hybrid DNAs (described in the following section); notably, SHapley Additive exPlanations (SHAP)-based feature importance analysis[31] on this model confirmed that Pulse Width (PW) at specific positions, particularly the base immediately downstream of the target site (PW_12), serves as the dominant predictor for Z-base discrimination (**Supplementary Figure 10a,b**; **Supplementary Note 2**). Additionally, since CCS captures kinetic signals from both DNA strands, Z-Calling distinguishes between T-A and T-Z pairs, generating output in a six-alphabet FASTA format, where 'O' represents a T paired with a Z (**Figure 2b**).

To date, no natural genomes exhibiting A/Z co-occurrence have been identified. We therefore designed the dZ-DNA/canonical DNA classification module of Z-Calling primarily to discriminate between non-hybrid dZ-DNA and canonical DNA CCS reads. This classification is achieved through a two-step approach combining Z base probability scoring via MLP neural networks and classification using a Support Vector Machine (SVM) (**Figure 2d**). First, reads are filtered based

on sequencing pass number, abnormal kinetic signals, and read length (A/Z base count). Next, Z probabilities for all A bases (from both forward and reverse strands) in a read are calculated using the MLP neural network with a model specifically trained for this purpose (described in the following section). To account for variability in CCS read lengths and A/Z counts, Z probability scores are aggregated as frequency distributions across a fixed number (50 by default) of intervals spanning from 0 to 1 (**Figure 2d**). Finally, the SVM classifier categorizes each read as either dZ-DNA or canonical DNA based on the Z probability frequency vector. The SVM model used for Z-Calling read classification was trained on a dZ-DNA/canonical DNA dataset generated on the PacBio Revio platform.

Evaluation of A/Z base calling models

Different feature k-mer (context) length might influence model strength since A/Z substitution impacts kinetic signals on multiple base positions. For dZ-DNA/canonical DNA read classification, we trained MLP models using Revio dZ-DNA/canonical amplicon CCS data. Besides the above fruit fly datasets, we also generated *E. coli* dZ-DNA/canonical DNA dataset except that the sequenced canonical DNA was native genomic DNA. We trained multiple models using k-mer lengths from 7 to 25 (odd numbers only) on a partial of the *E. coli* dZ-DNA/canonical DNA dataset. For testing, we assessed them on three randomly sampled testing datasets from the rest. Assessments showed that the strength of the model, evaluated via area under the receiver operating characteristic curve (AUC), grew with the k-mer size, and its increase rate slowed down as the k-mer size exceeded 19 bp, while the time consumption grew linearly with the k-mer length (**Figure 3a**). For balancing between model strength and computational efficiency, we selected 21-mer for dZ-DNA/canonical DNA read classification. The 21-mer model finally implemented in the read classification module (*k21-full-ZA*) demonstrated effective binary classification of Z and A in dZ-DNA/canonical DNAs, achieving AUCs of 0.9646-0.9796 and F1 scores of 0.9045-0.9384 across fruit fly (both Revio and Sequel II) and *E. coli* (Revio) datasets (**Supplementary Data 5**).

Hybrid DNAs offer significant application potential, therefore distinguishing between A and Z within a single DNA molecule is particularly valuable. Since Z-related kinetic signal alterations are highly context-dependent, we included hybrid DNA data in our training. However, obtaining hybrid DNA datasets suiData for supervised learning is challenging, as techniques for

incorporating A and Z at specific positions within the same molecule have not been established. To address this, we generated HgaI datasets using fruit fly, yeast, rice, and arabidopsis DNAs, featuring dA-dZ in-context islands. These were produced by HgaI digestion followed by selective incorporation of A and/or Z within the 10 random nucleotides flanking the HgaI recognition motif (**Supplementary Figure 11**). We then trained and tested Z-Calling's A/Z calling module using a combination of partial *E. coli* dZ-DNA/canonical DNA amplicon CCS data and the fruit fly HgaI dataset, while the yeast, rice, and arabidopsis HgaI datasets were used for model evaluation.

For selecting the optimal feature k-mer size, we trained A/Z calling MLP models using mixed dZ-DNA/canonical DNA (partial *E. coli* dataset) and HGAI (fruit fly) data with various k-mer lengths (7 to 15). To avoid overfitting to either the dZ-DNA/canonical DNA dataset or HGAI datasets, we chose a k-mer length of 11 bp, which yielded the most consistent performance (AUC and F1) across both datasets (**Supplementary Figure 12**). We initially trained an MLP model (*k11-full-ZA*) using only dZ-DNA/canonical data, but it performed poorly on the HgaI datasets, with AUCs ranging from 0.609 to 0.622 and F1 scores from 0.569 to 0.586 (**Supplementary Data 5**). Increasing the k-mer length worsened performance. In contrast, our mixed training strategy (k11-mixed-AZ) resulted in significant improvements across the four HGAI datasets, achieving AUCs from 0.942 to 0.949 and F1 scores between 0.868 and 0.878 (**Figure 3b**; **Supplementary Data 6**). The k11-mixed-AZ model also demonstrated similar robustness on both the Revio and Sequel II dZ-DNA/canonical DNA datasets (**Figure 3c**), with AUCs ranging from 0.951 to 0.952 and F1 scores between 0.864 and 0.894 (**Supplementary Data 6**). This model, which operates at single-nucleotide resolution, achieved 87.3-89.6% accuracy in dZ-DNA/canonical DNA contexts and 87.3-88.9% accuracy in A-Z-coexisting contexts, reaching performance levels comparable to a leading PacBio 5mC detection tool[26].

Building on this, Z-Calling can be applied to profile Z-base frequencies across an entire genome. For evaluation, we used an engineered yeast (*Saccharomyces cerevisiae*) strain expressing the phage dZ-DNA biosynthetic machinery, with ~24.7% of adenines replaced by Z bases, as confirmed by HPLC-UV analysis (**Supplementary Figure 13**). Applying Z-Calling to the yeast hybrid DNA dataset, 26.4% of the bases paired with thymine were recognized as Z bases, close to the 24.7% detected via HPLC-UV, across both genomes and plasmids (**Supplementary Data 7**). The results were used to profile the aggregated frequencies of Z bases across both native and engineered plasmids in the transformed yeast, unraveling similar random distributions

(**Figure 3d**).

Performance of Z-Calling on detection of dZ-DNAs

We trained an SVM DNA read classifier model based on the k21-full-AZ MLP model using a partial of the fruit fly Revio dZ-DNA/canonical DNA dataset. To evaluate the model, we collected native DNA datasets of various organisms: *Acinetobacter* phage SH-Ab 15497 (Phage 15497), which is natural dZ-DNA; canonical DNAs from *E. coli*, fruit fly, arabidopsis, rice, zebrafish, and human cell lines for checking the false positive discovery rate of the model. Results showed that the SVM model achieved high dZ-DNA/canonical DNA detection accuracy, with true positive rates (TPR, equal to recall) ranging from 93.44% to 99.72% across the four dZ-DNA datasets (including both the amplicons and the SH-Ab 15497 phage) and false positive rates (FPR) ranging from 0.02% to 0.19% across seven canonical DNA datasets (**Figure 4a**; **Supplementary Data 8**). Though both MLP and SVM models were trained on Revio datasets, they performed similarly on tested Sequel II (Chemistry 2.2) datasets (**Figure 4a**). Z-Calling uses the metric positive likelihood ratio (LR+) to assess the confidence of a DNA source containing dZ-DNAs (detailed in Methods). Based on all above-tested dZ-DNA/canonical DNA datasets, the read classification model of Z-Calling had a minimum LR+ of 490 (minimum TPR / maximum FPR among all tested datasets) among the dZ-DNA datasets, suggesting high sensitivity in detecting dZ-DNAs.

We also tested the read classification module of Z-Calling on A/Z-coexisting DNA reads from the transformed yeast with ~24.7% A/Z substitution. In this context, Z-Calling identified 9.06% of the total reads as dZ-DNA, significantly higher (≥45-fold) than any of the canonical DNA controls (**Figure 4a**; **Supplementary Data 8**). This result shows to some extent the validity of our read-classification module despite the coexistence of A and Z in a single molecule.

Phages containing the PurZ gene along with other critical components for dZ-DNA biosynthesis likely harbor genomes in the dZ-DNA format while pending direct evidence[9,11]. A methodology that allows reliable discrimination between dZ-DNAs and canonical DNAs, applicable to mixed datasets from metagenomes would aid in solving these issues (**Figure 2b**). To simulate the scenario on this purpose, we produced 12 artifical metagenomes datasets through mixing different datasets for testing, including 6 from the Sequel II platform (**Figure 4b**) and 6 from the Revio platform (**Figure 4c**). These datasets encompassed canonical, hybrid, and dZ-DNAs, and

originated from Phage 15497, *E. coli*, yeast, arabidopsis, rice, fruit fly, zebrafish, and human (**Supplementary Note 4; Supplementary Data 9**). The reads were classified as canonical or dZ- reads by Z-Calling, and could be easily assigned to species via mapping to the reference genomes due to their high accuracy ($\geqslant$ 99%). Z-Calling worked effectively despite the percentage of dZ-DNA (including hybrid DNAs) being low (in metagenomes 5, 6, 9, 11, 12), or the sources of dZ-DNAs in these mixtures being diverse in species (in metagenomes 7-12 in **Figure 4c**, **Supplementary Data 9**). We observed the only false negative for yeast hybrid DNA reads in the artificial metagenome 7, as only 10 of them were included in the dataset, while the five metagenomes containing no less than 25 yeast hybrid DNA reads were detected as dZ-DNA-positive (**Supplementary Data 9**). In conclusion, Z-Calling has demonstrated its great potential to become a powerful tool for taxonomic annotation of naturally occurring dZ-DNAs and artificial hybrid DNAs that are in the metagenome form.

Z-Calling has primarily been encoded using C++ (except for training scripts) and is computationally efficient with multi-processing capacity (**Supplementary Data 10**; **Supplementary Note 5**). While the training process of Z-Calling requires GPU, its application is GPU-free, making it suitable for deployment on ordinary computers with low computational resource requirements. The computational benchmark of Z-Calling's A/Z calling module on four datasets showed that the A/Z base calling (including BAM filtering + MLP with k11-mixed-AZ) on average utilized 37.48 ± 24.58s per 10k reads or 313.20 ± 98.72s per Gb data, with peak memory usage of 30.92 Gb (**Supplementary Data 10**). The read classification module of Z-Calling (including BAM filtering + MLP with k21-full-dZ + SVM) on average took 65.21 ± 29.62s per 10k reads or 605.30 ± 111.31s per Gb data, with peak memory usage of approximately 29.87 GB (**Figure 4d**; **Supplementary Data 10**), which could be further accelerated through multi-processing of chunked input files.

Discussion

Z-Calling is a machine learning-based tool capable of accurately identifying Z bases at single-nucleotide resolution in DNAs. By leveraging PacBio CCS kinetic signals, it not only distinguishes Z bases from canonical bases but also classifies dZ-DNA and hybrid DNA molecules with high precision. This capability faciliates exploring naturally occurring Z-genome organisms, such as phages harboring dZ-DNAs, whose prevalence in diverse environments

remains to be fully characterized[9,11,17]. Beyond its biological implications, Z-Calling facilitates the broader application of dZ-DNAs in synthetic biology, medicine, and data storage, where precise base-level identification is crucial for validating incorporation and functionality. The robust performance of Z-Calling on both dZ-DNAs and hybrid DNAs underscores its versatility, making it a valuable tool for annotating dZ-DNA-containing sequences in both engineered and natural metagenomic samples.

The distinct impact of Z bases on kinetic signals, compared to modifications such as 5mC and 6mA, likely stems from the unique nature of Z incorporation and its direct involvement in base pairing. While 5mC and 6mA are epigenetic modifications that chemically alter existing canonical bases without disrupting Watson-Crick pairing[32,33], dZTP directly replaces dATP as PCR substrate and forms non-canonical base pairs with thymine via three hydrogen bonds. This altered pairing introduces structural deviations that may affect the dynamics of DNA polymerase during sequencing. Notably, the pulse width (PW) signal alterations observed with Z incorporation exhibit high sequence dependence, suggesting that Z bases influence polymerase kinetics in a manner linked to sequence context and probably local DNA conformation. Unlike 5mC and 6mA, which primarily extend inter-pulse duration (IPD) by affecting polymerase pausing[23,24,27,28], Z bases may induce subtle conformational changes in the polymerase active site, particularly after Z addition, potentially influencing nucleotide binding and incorporation efficiency. Further investigation into the molecular mechanism underlying these changes, including structural studies of polymerase-Z base interactions, could provide valuable insights into the complexities of dZ-DNA processing and its broader implications in DNA synthesis and sequencing technologies.

Beyond binary classification, the coexistence of Z-bases with epigenetic modifications (such as 5mC and 6mA) presents both a challenge and an opportunity for genomic analysis. Developing a unified tool capable of simultaneously detecting all modification types is highly desirable but has been hindered by the lack of 'ground truth' datasets containing multiple, verified modifications at defined loci. However, our analysis suggests that Z-Calling is highly robust to this complexity. Because Z-incorporation primarily alters polymerase Pulse Width (PW) via non-canonical pairing, its kinetic signature is distinct from the Inter-Pulse Duration (IPD) shifts typical of 5mC and 6mA. Empirical testing on native genomes rich in epigenetic marks—including E. coli (6mA) and human/plant genomes (5mC)—showed no increase in false positive rates,

confirming that Z-Calling effectively disentangles Z-signals from the background epigenetic landscape.

While robust for fully substituted genomes, quantifying hybrid samples presents unique hurdles. The observation that the read-level classifier identified a lower proportion of reads in the hybrid yeast sample (~9% detection vs. ~25% Z-content) stems primarily from the model's optimization for fully substituted dZ-genomes. To robustly identify organisms like phages while minimizing false positives, the SVM enforces a high confidence threshold that often excludes hybrid reads with 'diluted' kinetic signatures. Beyond model calibration, the fundamental barrier remains the inherent noise of PacBio kinetic measurements, where the subtle Z-induced shift (0.5–1.5-fold) overlaps substantially with canonical Adenine variance. Distinguishing sporadic Z-signals from stochastic polymerase noise in mixed contexts is significantly more challenging than in fully substituted genomes. Future improvements will likely require a two-pronged approach: developing enzymatic methods to generate 'ground truth' hybrid training data with defined Z-positions, and advancing sequencing technologies — such as optimized polymerase chemistries—to generate more distinct kinetic profiles for non-canonical bases.

In summary, our study demonstrated that Z incorporation primarily affects polymerase pulse width (PW) signals in a sequence-dependent manner, with minimal impact on overall sequencing accuracy. Using these kinetic signals, we developed Z-Calling, which effectively distinguishes A/Z bases at single-nucleotide resolution and classifies dZ-DNA versus canonical DNA reads with high accuracy. Importantly, it performed robustly across datasets from different organisms generated on both Revio and Sequel II (Chemistry 2.2) platforms. Analyses of multiple independent sequencing runs confirmed that Z-base kinetic signatures are fundamental and sData, with the MLP model maintaining robust AUC scores (0.942–0.952) regardless of the biological source or flow cell. Similarly, while the SVM classifier exhibited minor sensitivity fluctuations across runs, it consistently maintained a high ratio of true-to-false positive rates, ensuring reliable dZ-DNA identification. Additionally, Z-Calling can detect dZ-DNAs and Z bases in complex metagenomic mixtures or hybrid DNAs, paving the way for future applications in diversified biotechnologies.

Methods

Dataset Collection and Composition. We generated and acquired 18 independent datasets to train and evaluate machine learning models for dZ-DNA identification and A/Z base calling. These datasets are categorized into three types: (1) Canonical DNAs: Molecules containing only standard bases (A/T/C/G) and natural modifications (e.g., 5mC, 6mA). These were sourced from *Drosophila melanogaster* (S2 cell line), *Saccharomyces cerevisiae* (S288C), *Arabidopsis thaliana* (Col0), *Oryza sativa* (Nipponbare), *Danio rerio* (Tuebingen), and *Escherichia coli* (BL21). *Drosophila melanogaster* (fruit fly) Schneider 2 (S2, ATCC: CRL-1963) cell line, *Saccharomyces cerevisiae* (yeast) strain S288C (ATCC: 204508), and *E. coli* strain B21 (ATCC: BAA-1025) were originally obtained from Genetimes ExCell Technology (Shanghai, China). Human cell line data of HG002 (RRID: CVCL_1C78) and HG00106 (RRID: CVCL_P686) were acquired from the Human Pangenome Reference Consortium[34]. (2) dZ-DNAs: Molecules where adenine is completely replaced by 2,6-diaminopurine. These include synthetic amplicons and naturally occurring dZ-DNA from *Acinetobacter* phage SH-Ab 15497. (3) Hybrid DNAs: Molecules containing both A and Z bases. These were generated via transformed yeast expressing ZTP biosynthesis genes and a restriction-fill-in-ligation strategy (HGAI datasets). For all datasets, identity was verified via BLASTN against the NCBI GenBank database using a minimum of ten reads per sample. Detailed methods applied have been described below.

DNA Extraction and Purification. Genomic DNA was extracted using specialized kits tailored to the source organism. *D. melanogaster* (S2) and *E. coli* DNAs were extracted using the FastPure Blood/Cell/Tissue/Bacteria DNA Isolation Mini Kit (#DC103-01, Vazyme Biotech, Nanjing, China). Yeast DNA was extracted using the Tiangen Yeast Genomic DNA Extraction Kit (#DP307, Tiangen Biotech, Beijing, China). SH-Ab 15497 phages were propagated in *A. baumannii* 15497. Lysates were collected in SM buffer and DNA was extracted using the Lambda Phage Genomic DNA Kit (#ZP317-1， Zoman Biotech, Beijing, China). DNA quantity and purity (A260/A280 ratio of 1.8–2.0) were confirmed using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, USA). Native DNAs of *A. thaliana* accession Col0*, O. Sativa* L. *cv.* Nipponbare*,* and *Danio rerio* (zebrafish) strain Tuebingen were collected in our previous studies[35,36].

PCR Amplification. dZ-DNA amplicons were synthesized by replacing dATPs with dZTPs (TriLink BioTechnologies, San Diego, USA) in the dNTP mix. Each 50 µL reaction contained 50

ng of template DNA (from D. melanogaster S2 or E. coli), 0.4 µM random primers, 0.2 mM each of dZTP, dCTP, dGTP, and dTTP, and 1 unit of Taq DNA polymerase in $1\times$ PCR buffer (Thermo Fisher Scientific, Waltham, USA). Canonical amplicons were generated under identical conditions using dATP instead of dZTP.The thermal cycling profile consisted of an initial denaturation at 95°C for 3 minutes, followed by 20 cycles of: denaturation at 95°C for 30 seconds, annealing at 58°C for 30 seconds, and extension at 72°C for 3 minutes. A final extension was performed at 72°C for 7 minutes. All PCR products were purified using the Qiagen PCR Purification Kit (Qiagen, Hilden, Germany) and verified by gel electrophoresis before library preparation.

HGAI-Digested DNA and dZTP Incoporation. To generate DNA fragments with defined Z-incorporation "islands," genomic DNAs from fruit fly, *A. thaliana*, *O. sativa*, and *E. coli* were processed using a restriction-fill-in-ligation strategy: (1) Approximately 1 µg of genomic DNA was digested with 10 units of HgaI restriction endonuclease (ER1901, Thermo Fisher Scientific, Waltham, US) in a 50 µL reaction volume containing 1× Buffer R. The reaction was incubated at 37°C for 2 hours, followed by heat inactivation at 65°C for 20 minutes. HgaI digestion generates 5-base 5' overhangs (5'-NNNNN-3') downstream of the recognition motif; (2) The 5' overhangs were filled using the Klenow Fragment (3'→5' exo–) (#M0212, New England Biolabs, Ipswich, USA) to prevent degradation of the Z-containing strand. The fill-in reaction included: Digested DNA (~1 µg), 5 units Klenow Fragment (3'→5' exo–), and dN'TP mix: 0.2 mM each of dZTP (N-2003-1, TriLink, San Diego, US), dCTP, dGTP, and dTTP (dATP was excluded). The mixture was incubated at 37°C for 30 minutes. This step selectively incorporated dZTPs into the 5-nucleotide overhangs complementary to the template strand. (3) Following dZ-incorporation, the DNA fragments were purified using 0.6× AMPure PB beads (Pacific Biosciences) to remove excess dZTPs and enzymes. Custom hairpin adapters (13-bp) were ligated to the blunt-ended fragments using T4 DNA Ligase (#M0202, New England Biolabs, Ipswich, USA). The ligation reaction was performed at 25°C for 1 hour. (4) The resulting library was purified again with AMPure PB beads and subjected to PacBio Circular Consensus Sequencing (CCS) on the Revio platform.

Yeast Transformation and HPLC Analysis. The genomic DNAs from transformed yeasts (including plasmids pRS426-ApPurZ-ApdATPase and pRS425-ApDUF550 for dZTP biosynthesis) [9] were enzymatically digested by Nucleoside Digestion Mix (#M0649S, New England Biolabs, Ipswich, USA) and separated on Amplicon Ultra-0.5 mL 3 K centrifugal filters (MilliporeSigma, Burlington, USA). The flow-through was analyzed using a 1260 infinity II instrument (Agilent Technologies, Santa Clara, USA). LC separation was performed using a Syncronis aQ column (Thermo Fisher Scientific, Waltham, USA) with a flow rate of 0.5 mL/min, employing a gradient of solvent A (10 mM NH4AC pH 4.6 in water) and solvent B (methanol). The sample (a gradient of 0-12 min 20-32% B) was injected at a volume of 10 μL with UV detection at 260 nm. Standard deoxynucleosides (dA, dT, dC, dG, and dZ) were used as controls.

Library Preparation and PacBio Sequencing. Libraries were sequenced on PacBio Sequel II (Chemistry 2.2) and Revio platforms (Pacific Biosciences, Menlo Park, USA). For Revio, DNAs were fragmented via ultrasonication to 5–10 kb. Circular Consensus Sequencing (CCS) BAM files were generated using the '--keep-kinetics' option to preserve pulse width (PW) and inter-pulse duration (IPD) signals. We maintained a data quality threshold of at least 60,000 reads per dataset with forward and reverse pass numbers no less than 3.

Sequencing accuracy comparison between dZ-DNA and canonical DNA on PacBio platforms. We assessed the CCS base sequencing accuracy and error distribution of dZ-DNA and canonical DNA amplicons with *D. melanogaster S2 cell line* DNA as templates on both PacBio's Revio and Sequel II platforms. CCS reads with sufficient passes (BAM np tag ≥3) and high similarity (≥95% for excluding contamination) with fruit fly reference genome[37] were retained. Base Phred quality values (QVs) were evaluated for each dataset using Merqury v1.3[30] with native S2 DNA CCS data as control. For each pass number (np tag in BAM) from 3 to 20, QV distributions were statistically compared between dZ-DNA and canonical DNA reads using Mann-Whitney U tests and p values were adjusted using the Benjamini-Hochberg method[38]. Base error types were analyzed using minimap2 v2.28[39] and Bcftools v1.18[40], and we statistically assessed differences in error types (insertion, deletion, and nucleotide substitution) with Z-tests and Cohen's h effect size estimates to quantify practical significance. Detailed command lines and custom scripts used have been detailed in **Supplementary Note 1**.

Analysis of Z base incorporation related kinetic signal alterations. To assess the impact of

Z base incorporation on kinetic signals during PacBio CCS sequencing, we extracted and analyzed pulse width (PW) and inter-pulse duration (IPD) values from CCS reads. Using custom Python scripts , we processed BAM files to retrieve kinetic data for 21 bp k-mers surrounding target A/Z bases in template strand. To avoid interference among adjacent A/Zs, only context k-mers containing one A/Z (target base) were analyzed. The PW and IPD signals were compared between pairs of 21 bp k-mer only differing at the target A/Z base for unraveling alterations due to Z base incorporation. Dimensionality reduction methods, including Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE), were applied to visualize potential patterns between A and Z bases using python codes. Detailed command lines and scripts used have been described in **Supplementary Note 2**.

To evaluate the differences in kinetics signals between dZ-DNA and Canonical DNA CCS reads, we performed a series of statistical analyses using a custom Python script named ZA_signal_ttest_and_medianratio.py. The script processes input files containing kinetics signals for specific context k-mers. Each file includes columns for nucleotide position, base, context k-mer, and corresponding lists of IPD and PW values for each base position within the k-mer. The data was filtered to include only those context k-mers that had at least 10 recorded observations in both conditions ("Canonical A" and "Z"). For each base position within the context k-mers, two-tailed t-tests were conducted to compare the IPD and PW signals between the two conditions. To account for multiple comparisons across the different base positions and context k-mers, the Benjamini-Hochberg (BH) procedure was applied to adjust the p-values.

Random-Forest Classification-based A/Z discrimination. We investigated the possibility of applying a machine learning approach for discrimination between A and Z bases using Random Forest Classification. To identify and evaluate the importance of kinetics signals (IPD and PW) for distinguishing between dZ-DNA and Canonical DNA CCS reads, we also conducted a feature importance analysis using a custom Python script named ZA_signal_feature_importance.py. The script processes input data files containing nucleotide positions, bases, context k-mers, and lists of IPD and PW values for each base position within the k-mer. For each k-mer, the IPD and PW values corresponding to each base position were extracted and treated as features. These features were then used to construct a dataset for subsequent classification analysis. A Random Forest Classifier was employed to model the relationship between the extracted features (IPD and PW values) and the condition labels ("Z" and "Canonical A"). The dataset was split into

training and testing sets to evaluate the model's performance. The classifier was trained on the training set, and its accuracy was assessed using the test set. After training the Random Forest model, the importance of each feature (i.e., IPD and PW values at specific base positions) was quantified. The feature importance scores were calculated based on how much each feature contributed to reducing the impurity in the decision trees of the Random Forest model. The most important features are those that provide the most information for distinguishing between the "Z" and "Canonical A" conditions.

Evaluation of More Complex Model Architectures for Z/A discrimination. To rigorously evaluate the optimal architecture for the Z-Calling framework, we conducted a standardized benchmark of five distinct machine learning models: Random Forest (RF), Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and Graph Convolutional Network (GCN) using our python script 'benchmark_models_gpu.py'. To ensure a fair and unbiased comparison between architectures, all models were trained and tested on identical datasets derived from fruit fly Revio sequencing reads. We generated balanced datasets containing 10,000 samples per class (canonical Adenine vs. 2,6-diaminopurine) to prevent class imbalance artifacts. Input features were standardized across all models using a fixed window size of $k = 11$ (target base $\pm$ 5 bp) with Z-free contexts. The feature vector consisted exclusively of Inter-Pulse Duration (IPD) and Pulse Width (PW) kinetics, ensuring performance differences were attributable solely to model architecture. We isolated nine distinct 3-mer sequence contexts (CAC, CAG, CAT, GAC, GAG, GAT, TAC, TAG, TAT) to evaluate performance stability across diverse genomic backgrounds.

For each replicate, the training data was randomly partitioned into a training set and a validation set (80:20 split). We implemented an Early Stopping mechanism that monitored the Validation Loss at the end of each epoch. And training was automatically halted if the validation metric did not improve for 15 consecutive epochs. Upon training termination, the model weights corresponding to the epoch with the best validation performance were restored. This ensured that the reported AUC metrics reflected each model's peak generalization capability, preventing bias from overfitting (in complex models like GCN) or under-fitting (in slower-converging models).

Architecture of the Z-Calling's single-nucleotide Z/A classification module. We implemented a Multi-Layer Perceptron (MLP) architecture to achieve an optimal balance between classification accuracy and computational efficiency. Kinetic signal-containing BAM

files were parsed using a custom C++ script to retrieve sequence information, quality scores, and kinetic signals, specifically Inter-Pulse Duration (IPD) and Pulse Width (PW). Both forward and reverse strands of each read were processed. Each read was segmented into overlapping k-mers, with the target base was positioned at the center (e.g., the 11th base in a 21-mer). For each nucleotide within the k-mer, we extracted one-hot encoded sequence data (4-element vector) and normalized IPD and PW signals (scaled by 1/255). The resulting features were structured as a two-dimensional tensor of shape $k \times 6$. To optimize throughput, the extraction pipeline utilized multithreading to process reads concurrently.

We evaluated MLP depths ranging from one to five hidden layers; empirical testing indicated that depth beyond two layers yielded no significant performance gains (**Supplementary Note 3; Supplementary Figures 8, 9**). The final architecture consists of an input layer, two hidden layers, and an output layer. The $k \times 6$ input tensor is flattened into a 1D vector prior to entering the network. Both hidden layers comprise a linear transformation with 128 neurons, followed by batch normalization and ReLU activation. The output layer consists of two neurons representing the Z and A classes, with a softmax activation function applied to generate class probabilities. To address potential class imbalances, we employed a custom Focal Loss function and applied gradient clipping during training to ensure numerical stability.

SHAP-Based Feature Importance Analysis. To interpret the underlying features driving the identification of Z-bases, we trained a multi-layer perceptron (MLP) neural network using PyTorch and employed SHAP (SHapley Additive exPlanations), a game-theoretic approach to explain the output of machine learning models, which are carried out using the python script ‘torch_mlp_shap_gpu.py’. We utilized a balanced dataset comprising 1,000,000 "Canonical A" samples and 1,000,000 "Z-base" samples derived from fruit fly dZ-DNA and canonical DNA amplicon sequencing data (sequenced on the Revio platform). Each sample consisted of a 21-mer sequence context centered on the target base, along with the corresponding Inter-Pulse Duration (IPD) and Pulse Width (PW) kinetic signals for each position. The 21-mer context sequence was one-hot encoded, resulting in $21 \times 4 = 84$ binary features (e.g., Seq_1_A, Seq_1_C, etc.). The IPD and PW signals for the 21 positions were extracted as continuous variables, yielding $21 \times 2 = 42$ features (e.g., IPD_1...IPD_21, PW_1...PW_21). The final feature vector for each sample had a dimensionality of 84+42=126. All features were standardized using StandardScaler to have zero mean and unit variance. The MLP architecture comprised two

hidden layers of 128 neurons each, utilizing Batch Normalization, ReLU activation, and Dropout (rate = 0.2) to prevent overfitting, followed by a final linear output layer. The model was trained using the Adam optimizer (learning rate = 0.001) with a binary cross-entropy loss function (BCEWithLogitsLoss) and early stopping based on validation loss. Post-training, we employed SHAP (SHapley Additive exPlanations) to quantify feature importance. Specifically, we used the GradientExplainer to approximate SHAP values by computing the gradient of model outputs with respect to input features. We utilized 500 background samples from the training set to estimate expected values and computed feature attributions for 500 test samples.

Training and evaluation of single-nucleotide MLP model. MLP models for dZ-DNA/canonical DNA classification and A/Z calling were trained to identify Z bases in PacBio CCS reads using different training datasets with different feature k-mer lengths. For read classification, models (*kN-full-AZ* in which N indicates k-mer length and are odd numbers between 7 to 25 bp) was trained using features extracted from *E. coli* dZ-DNA amplicons and native *E. coli* DNA data. The dZ-DNA amplicon features were labeled as the Z class, while native *E. coli* DNA features were labeled as the A class. For training, 20 million A-class features, and 20 million Z-class features were randomly selected. The model was trained over 3 epochs, iterating through batches of training data. All models were evaluated via areas under recursive operating characteristic curves (AUCs) and confusion matrix-derived metrics. The following functions were used to calculate metrics from the confusion matrix:

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall (True Positive Rate)} = TP / (TP + FN)$$

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{False Positive Rate} = FP / (FP + TN)$$

Where TP = True Positives (Z base calls), FP = False Positives, TN = True Negatives (A base calls), and FN = False Negatives.

We also evaluated the models trained using merely full-dZ/dA dataset on the four HGAI datasets that had dA/dZ coexisting contexts within 10 bp regions downstream of HgaI recognized motif (5'-GACGC-3'). The results showed that the models had significantly lower AUCs on the HGAI datasets compared to the full-dZ/dA datasets. For instance, the 11 bp model trained on full-dA/dZ datasets achieved ~0.94 AUCs in identifying DNA molecules containing Z bases, but it

only had 0.609-0.622 AUCs on the HGAI datasets (**Supplementary Data 5**). To address this, we specifically trained models (7 to 15 bp feature k-mer lengths) for A/Z calling using combined data of the fruit fly HGAI dataset (1 million features for either A or Z located in the 10 random bases neighboring HgaI-recognized motif) and the *E. coli* dZ-DNA/native DNA datasets (20 million per class). The models were evaluated on both the four HGAI datasets from arabidopsis, rice, and yeast using the same evaluation metrics (**Supplementary Data 6**).

Training and assessment of SVM read classification model. Support Vector Machines (SVM) were chosen for read-level classification. While simpler models like Logistic Regression underperformed, the SVM effectively captured the decision boundary using the aggregated Z-probability distributions. More complex deep learning architectures were deemed unnecessary for this module, as the SVM achieved >99% accuracy with significantly lower computational overhead. Z-probability distributions calculated by the MLP model were used as features for the SVM. A total of 100,000 dZ-DNA and 100,000 canonical DNA reads of fruit fly sequenced on the Revio platform were used for training. The classifier was trained on read-level aggregated scores generated from 21-bp sequence contexts, and filtering criteria were applied to ensure high-quality data. Kinetic signal filters were applied to remove reads whose median PW/median IPD ratio falls within the top 1% of the distribution or is below 0.3, as these were shown to be significantly enriched for false positives (**Supplementary Data 11**). Reads shorter than 500 bp or containing fewer than 100 A+Z bases were also excluded due to the limited number of A/Z bases.

Development of dZ-DNA source detection pipeline. A source detection pipeline was developed to identify dZ-DNA reads from PacBio CCS data and assign them to reference genomes or contigs. The pipeline applied the same filtering steps used in the SVM classification process, including kinetic signal filtering, pass coverage thresholds, and length-based filtering. Reads classified as dZ-DNA were aligned to a merged reference genome using minimap2[39], and primary alignments were used for taxonomic or contig assignment for reads. If an optional two-column TSV file containing sequence names and taxonomic origins was provided, the reads were assigned to their respective organisms. The pipeline calculated a positive likelihood ratio (LR+) for each organism or contig based on the dZ-DNA read count relative to the total reads assigned to it. The LR+ for each organism or contig is calculated as follows:

$$\mathrm{LR+} = \frac{\text{dZ-DNA read count}}{\text{classified read count}} \div 0.002$$

Here, 0.002 represents the maximum false positive rate observed in all tested dZ-DNA-negative datasets.

To interpret LR+:

LR+ > 1: The test result is more likely to occur in sources containing dZ-DNAs than those without it.

Higher LR+ Values: The higher the LR+, the more confident the test is at identifying sources containing dZ-DNAs. Generally:

- LR+ ≥6: Strong evidence supporting that dZ-DNAs exist and account for >1% of tested reads.
- LR+ >=2 and <6: Weak evidence supporting that dZ-DNAs exist and account for ⩽1% of tested reads.
- LR+ >1 and <2: Cautions should be taken to make any judgments. dZ-DNAs potentially exist at a low content (<0.4%) and more evidence is required to confirm the result.

LR+<=1: No evidence supporting dZ-DNA in source.

And the relative abundance of dZ-DNA of the total DNA of an organism can be roughly estimated as (LR+ / 5)% when LR+ ⩾ 2, which only applies for mixtures between dZ-DNAs and canonical DNAs.

Evaluation of dZ-DNA source detection on artificial metagenome datasets. The dZ-DNA source detection pipeline was tested on mixed datasets, which included various proportions of dZ-DNA and canonical DNA reads (**Supplementary Note 4; Supplemenatry Data 9**). These simulations were designed to mimic different combinations of dZ-DNAs including dZ-DNAs and those containing dA/dZ coexisting contexts (transformed yeast) and canonical DNAs (amplicons and native DNAs) for both PacBio Sequel II (Chemistry 2.2) and Revio platforms. The performance of the pipeline was evaluated based on the positive likelihood ratio (LR+), and dZ-DNA-positive sources were identified using pre-defined thresholds for LR+ (**Supplementary Note 4**) and read counts.

Computational benchmark tests. Computational benchmarks of Z-Calling modules were carried out tests were performed on a computer equipped with an AMD EPYC 7402 24-core

Processor and 64 GB × 4 of DDR4 memory (2667 MHz), running a Linux environment without GPU requirement. The command lines and tested datasets have been detailed in **Supplementary Data 10**.

Statistics and Reproducibility

The performance of MLP and SVM models was evaluated using 18 independent datasets, including 4 dZ-DNA, 9 canonical DNA, and 5 mix-A/Z-DNA datasets. Sample sizes were determined based on empirical standards to ensure sufficient testing power. For each dataset, data were randomly partitioned into training, validation, and test sets. No data were excluded from the analyses. All computational experiments were performed at least 3 times to ensure reproducibility.

Code Availability

All codes written and used by this study have been deposited in our github repository (https://github.com/xiaochuanle/Z-Calling) and in Zenodo (DOI:10.5281/zenodo.17840213)[41]. Partial command lines used in data analysis are described in Supplementary Notes.

Data Availability

Pacbio CCS BAMs (containing kinetics signals) sequenced by our lab have been deposited in Genome Sequence Archive of China National Center for Bioinformation in project ID PRJCA031439 under GSA accessions CRA020168 (https://ngdc.cncb.ac.cn/gsa/browse/CRA020168), CRA019888 (https://ngdc.cncb.ac.cn/gsa/browse/CRA019888), and CRA020191 (https://ngdc.cncb.ac.cn/gsa/browse/CRA020191). Human Sequel Ⅱ (HG002) and Revio (HG00106) datasets were acquired from the study of Baid et al. (2023)[42] and Human Pangenome Reference Consortium[34], which are available at https://console.cloud.google.com/storage/browser/_details/brain-genomics-public/research/deepconsensus/publication/sequencing/hg002_15kb/m64008_201124_002822.subreads.bam?pageState=(%22StorageObjectListData%22:(%22f%22:%22%255B%255D%22))&walkthrough_id=panels--storage--bucket and https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/HG00106/raw_data/PacBio_HiFi/m84081_231112_034048_s4.hifi_reads.bc2070.bam. The plasmids pRS426-ApPurZ-ApdATPase and pRS425-ApDUF550 generated during the current study are available from the corresponding author on reasonable request under a standard Material Transfer Agreement. Source data for the graphs and charts in this study are available in the Figshare repository (DOI: 10.6084/m9.figshare.31281748)[43].

Acknowledgements

We acknowledge financial support from the National Key R&D Program of China (2022YFF1201900 to C.-L.X.), the National Natural Science Foundation of China (no. 32270713, 62350004 to C.-L.X. and no. 32522004 and 32200051 to Y.Zhou); Guangdong Basic and Applied Basic Research Foundation (2020B1515020057 to C.-L.X.); Distinguished Young Scholars of China (no. 32125002 to Y.Zhang); the New Cornerstone Science Foundation (NCI2002321 to Y.Zhang); Natural Science Foundation of Jiangsu Province (BK20220591 to Y.Zhou); Key Project Fund of National Natural Science Foundation (no. 82230031 to W.C.); the Regional Innovation and Development Joint Fund of the National Natural Science Foundation of China (U24A20706 to W.C.); the Key Special Project of 'Cutting-Edge Biotechnology' in the National Key Research and Development Program of China (2024YFC3406200 to W.C.); Sanming Project of Medicine in Shenzhen (No. SZSM202411007 to W.C.); Guangdong Basic and Applied Basic Research Foundation Regional Joint Fund Key Program (2023B1515120051).

Author Contribution

C.-L.X., Y.Zhang. W.C., and S.Z. conceived the study. B.W., Y.C., C.-L.X., and H.-X.C. implemented the algorithms of Z-Calling. H.-X.C., Y.C., and B.W. wrote the codes of Z-Calling. L.N., Y.Zhou., and Y.L. carried out experiments. B.W., Y.Zhou., and J.-Y.Z. carried out data analysis. B.W., Y.Zhang, Y.C., Y.Zhou, L.N., and H.-X.C. wrote the manuscript. S.Z., W.C, C.-L.X., J.-Y.Z., and Y.L. modified and improved the manuscript. All authors read and approved the final version of the manuscript.

Competing Interests

The authors declare no competing interests.

References

1 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953). https://doi.org:10.1038/171737a0

2 Kirnos, M. D., Khudyakov, I. Y., Alexandrushkina, N. I. & Vanyushin, B. F. 2-aminoadenine is an adenine substituting for a base in S-2L cyanophage DNA. *Nature* **270**, 369-370 (1977). https://doi.org:10.1038/270369a0

3 Cheong, C., Tinoco, I., Jr. & Chollet, A. Thermodynamic studies of base pairing involving 2,6-diaminopurine. *Nucleic Acids Res.* **16**, 5115-5122 (1988). https://doi.org:10.1093/nar/16.11.5115

4 Cristofalo, M. *et al.* Nanomechanics of Diaminopurine-Substituted DNA. *Biophys. J.* **116**, 760-771 (2019). https://doi.org:10.1016/j.bpj.2019.01.027

5 Chollet, A. & Kawashima, E. DNA containing the base analogue 2-aminoadenine: preparation, use as hybridization probes and cleavage by restriction endonucleases. *Nucleic Acids Res.* **16 1**, 305-317

(1988).

6 Kang, S., Liu, Q., Zhang, J., Zhang, Y. & Qi, H. 2,6-diaminopurine (Z)-containing toehold probes improve genotyping sensitivity. *Biotechnol. Bioeng.* **121**, 1383-1392 (2024). https://doi.org:https://doi.org/10.1002/bit.28642

7 Haaima, G., Hansen, H. F., Christensen, L., Dahl, O. & Nielsen, P. E. Increased DNA binding and sequence discrimination of PNA oligomers containing 2,6-diaminopurine. *Nucleic Acids Res.* **25**, 4639-4643 (1997). https://doi.org:10.1093/nar/25.22.4639

8 Bailly, C. & Waring, M. J. The use of diaminopurine to investigate structural properties of nucleic acids and molecular recognition between ligands and DNA. *Nucleic Acids Res.* **26**, 4309-4314 (1998). https://doi.org:10.1093/nar/26.19.4309

9 Zhou, Y. *et al.* A widespread pathway for substitution of adenine by diaminopurine in phage genomes. *Science* **372**, 512-516 (2021). https://doi.org:10.1126/science.abe4882

10 Czernecki, D., Bonhomme, F., Kaminski, P.-A. & Delarue, M. Characterization of a triad of genes in cyanophage S-2L sufficient to replace adenine by 2-aminoadenine in bacterial DNA. *Nat. Commun.* **12**, 4710 (2021). https://doi.org:10.1038/s41467-021-25064-x

11 Sleiman, D. *et al.* A third purine biosynthetic pathway encoded by aminoadenine-based viral DNA genomes. *Science* **372**, 516-520 (2021). https://doi.org:10.1126/science.abe6494

12 Pezo, V. *et al.* Noncanonical DNA polymerization by aminoadenine-based siphoviruses. *Science* **372**, 520-524 (2021). https://doi.org:10.1126/science.abe6542

13 Gao, S. *et al.* Harnessing non-Watson–Crick's base pairing to enhance CRISPR effectors cleavage activities and enable gene editing in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2308415120 (2024). https://doi.org:doi:10.1073/pnas.2308415120

14 Zhang, M., Singh, N., Ehmann, M. E., Zheng, L. & Zhao, H. Incorporation of noncanonical base Z yields modified mRNA with minimal immunogenicity and improved translational capacity in mammalian cells. *iScience* **26**, 107739 (2023). https://doi.org:10.1016/j.isci.2023.107739

15 Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nat. Rev. Genet.* **20**, 456-466 (2019). https://doi.org:10.1038/s41576-019-0125-3

16 Czernecki, D. *et al.* How cyanophage S-2L rejects adenine and incorporates 2-aminoadenine to saturate hydrogen bonding in its DNA. *Nat. Commun.* **12**, 2420 (2021). https://doi.org:10.1038/s41467-021-22626-x

17 Tong, Y. *et al.* Alternative Z-genome biosynthesis pathway shows evolutionary progression from Archaea to phage. *Nat. Microbiol.* **8**, 1330-1338 (2023). https://doi.org:10.1038/s41564-023-01410-1

18 Grome, M. W. & Isaacs, F. J. ZTCG: Viruses expand the genetic alphabet. *Science* **372**, 460 - 461 (2021).

19 Rhoads, A. & Au, K. F. PacBio Sequencing and its Applications. *Genom. Proteom. Bioinform.* **13**, 278-289 (2015). https://doi.org:10.1016/j.gpb.2015.08.002

20 Fuller, C. W. *et al.* The challenges of sequencing by synthesis. *Nat. Biotechnol.* **27**, 1013-1023 (2009). https://doi.org:10.1038/nbt.1585

21 Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411-413 (2017). https://doi.org:10.1038/nmeth.4189

22 Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348-1365 (2021). https://doi.org:10.1038/s41587-021-01108-x

23 Flusberg, B. A. *et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461-465 (2010). https://doi.org:10.1038/nmeth.1459

24 Feng, Z. *et al.* Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. *PLoS Comput. Biol.* **9**, e1002935 (2013). https://doi.org:10.1371/journal.pcbi.1002935

25 Tse, O. Y. O. *et al.* Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021). https://doi.org:10.1073/pnas.2019768118

26 Ni, P. *et al.* DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *Nat. Commun.* **14**, 4054 (2023). https://doi.org:10.1038/s41467-023-39784-9

27 Zhang, J. *et al.* 6mA-Sniper: Quantifying 6mA sites in eukaryotes at single-nucleotide resolution. *Sci. adv.* **9**, eadh7912 (2023). https://doi.org:doi:10.1126/sciadv.adh7912

28 Kong, Y. *et al.* Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. *Science* **375**, 515-522 (2022). https://doi.org:doi:10.1126/science.abe7489

29 Jha, A. *et al.* DNA-m6A calling and integrated long-read epigenetic and genetic analysis with fibertools. *Genome Res.* **34**, 1976-1986 (2024). https://doi.org:10.1101/gr.279095.124

30 Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020). https://doi.org:10.1186/s13059-020-02134-9

31 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions in *Proceedings of the 31st International Conference on Neural Information Processing Systems.* 4768–4777 (Curran Associates Inc.).

32 Ehrlich, M. & Wang, R. Y. 5-Methylcytosine in eukaryotic DNA. *Science* **212**, 1350-1357 (1981). https://doi.org:10.1126/science.6262918

33 Luo, G.-Z., Blanco, M. A., Greer, E. L., He, C. & Shi, Y. DNA N6-methyladenine: a new epigenetic mark in eukaryotes? *Nat. Rev. Mol. Cell Biol.* **16**, 705-710 (2015). https://doi.org:10.1038/nrm4076

34 Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437-446 (2022). https://doi.org:10.1038/s41586-022-04601-8

35 Chen, Y. *et al.* High accuracy methylation identification tools on single molecular level for PacBio HiFi data. *bioRxiv*, 2024.2008.2014.607879 (2024). https://doi.org:10.1101/2024.08.14.607879

36 Chen, H. X. *et al.* Accurate cross-species 5mC detection for Oxford Nanopore sequencing in plants with DeepPlant. *Nat. Commun.* **16**, 3227 (2025). https://doi.org:10.1038/s41467-025-58576-x

37 dos Santos, G. *et al.* FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* **43**, D690-D697 (2014). https://doi.org:10.1093/nar/gku1099

38 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc., B: Stat. Methodol.* **57**, 289-300 (1995). https://doi.org:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

39 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018). https://doi.org:10.1093/bioinformatics/bty191

40 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10** (2021). https://doi.org:10.1093/gigascience/giab008

41 Wu, B. *Z-Calling Release v1.0.0*, <https://doi.org/10.5281/zenodo.17840213> (2025).

42 Baid, G. *et al.* DeepConsensus improves the accuracy of sequences with a gap-aware sequence

transformer. *Nat. Biotechnol.* **41**, 232-238 (2023). https://doi.org:10.1038/s41587-022-01435-7

43 Wu, B. *Figure source data of Z-Calling manuscript*, <https://doi.org/10.6084/m9.figshare.31281748> (2026).

Figure 1. PacBio circular consensus sequencing (CCS) of dZ-DNAs.

(**a**) Base pairing between thymine (T) and adenine (A, left) or 2,6-diaminopurine (Z, right). (**b**) Sequencing of dZ-DNAs on the PacBio platform. Bases are recognized via fluorescence-labeled dNTPs, and Z is indistinguishable from A through fluorescence signals. (**c**) Phred quality values (QVs) of dZ-DNA and canonical DNA amplicon CCS reads from the Revio platform. QV is calculated as $-\log_{10}$(base error rate). Z was treated as A in QV calculation for dZ-DNAs. Fruit fly DNAs were amplified by 20 PCR cycles under the same conditions except for the dATP/dZTP substitution (see Methods) to produce the canonical DNA and dZ-DNA amplicons. The horizontal axis indicates the number of read sequencing passes, shaded areas indicate one standard deviation from the mean QVs, and a minimum of 6,920 reads was subjected to statistics for each pass number. (**d**) Schematic diagram explaining kinetic signals and potential alterations due to dA/dZ substitution. (**e**) Kinetic signal ratios (Z/A) across 21 bp contexts for two 5-bp motifs, TT[Z/A]TT and TT[Z/A]GG. The diagram (each square represents a base) on top shows the base positions in the template and nascent strands relative to the central A/Z base in a 21 bp context. For each pair of k-mers differing only at the target A/Z base (11th position), median PW and IPD signals among CCS reads (at least 10 under either condition) were calculated across the 21 positions for both conditions, and the ratios between them were calculated by dividing corresponding values in Z context by those in the A context; violin plots depict the signal ratio distribution across all 21-mer template contexts (1,859 for TT[Z/A]TT and 290 for TT[Z/A]GG) centered with the 5-bp motifs and without other A/Z in context in the template. Boxplots: the center line/ dot, median; boxes, first and third quartiles; whiskers, 5th and 95th percentiles.

Figure 2. Design of Z-Calling.

(**a**) Random Forest Classification (RFC)-based feature importance scores of kinetic signals across A/Z context positions. Kinetic signals (PW and IPD) across the 21 bp context positions (1 to 21, the same as Figure 1e) were arranged by descending order of average importance scores on the horizontal axis. RFC analysis was carried out separately for nine different types of contexts with differential 3-mer center motifs surrounding the target A/Z bases (right vertical axis) using their kinetic signals. The center motifs have been hierarchically clustered via the feature importance scores. The top boxplots depict the distribution of the feature importance scores of the corresponding features on the horizontal axis. In the boxplots: the center line/ dot, median; boxes, first and third quartiles; whiskers, 5th and 95th percentiles. (**b**) Diagram illustrating the Z-Calling framework, training datasets, and application scenarios for dZ-DNA detection (gray arrows) and single-nucleotide A/Z discrimination (green arrows). MLP, Multilayer Perceptron; SVM, Support Vector Machine. In read classification (gray arrows), the positive likelihood ratio (LR+) serves as an indicator for confidence of dZ-DNA presence in tested reads. For base discrimination (green arrows), a six-base alphabet (A, T, C, G, Z, and O) was used to represent double-stranded dZ-DNAs, where O signifies T pairing with Z, and Z-Calling also has provided tools for profiling Z frequencies on A bases across reference genome. (**c**) and (**d**) Architectures of the MLP and SVM networks implemented in Z-Calling. In panel **c**, context sequences and nascent (reverse) strand kinetic signals of target A/Z bases (on the template strand) were extracted as feature input for the MLP networks. Z probability profiles of all A bases including those on the reverse strand were used as input of the SVM read classifier.

Figure 3. Evaluation of the A/Z base discrimination models.

(**a**) Relationship between Multilayer Perceptron (MLP) model performance and feature k-mer size. The models were trained using the same set of randomly sampled 20M A context features and 20M Z context features extracted from the fruit fly Revio dZ-DNA/canonical DNA dataset with different context k-mer sizes, and tested on three replicates of 10M features per A/Z context randomly sampled from the same dataset (training set excluded). CPU time consumption benchmark was performed on the same computer using equal amount of CPU cores for the different models (the same as in **Supplementary Data 10**). A bar's height indicates the mean value of the three replicates, and SDs are shown as error bars. (**b**) and (**c**) ROC (Receiver Operating Characteristic) curves of the k11-mixed-AZ MLP model in discriminating against A/Z-bases in two different types of datasets: dZ-DNA/canonical DNA datasets (left) and A/Z coexisting contexts (right). For the dZ-DNA/canonical DNA datasets, tagged Z context features were extracted from dZ-DNA amplicons, while tagged A context features were extracted from canonical amplicon reads (two fruit fly datasets) or the native *E. coli* DNA reads. HGAI datasets from the four species were subjected to evaluation, with A and Z bases sampled from 10 random nucleotides flanking the HgaI-recognized motif with A/Z coexisting contexts. The AUC (area under the ROC curve) values ranged from 0.9513 to 0.9550 (left panel) and 0.9422 to 0.9488 (right panel). (**d**) Z frequency profiles across three plasmid genomes in the transformed yeast with ~24.7% Z/(A+Z). Each dot depicts the Z base frequency (vertical axis) at an A/T base position (horizontal axis) in the genome, and the lines (10-point moving average) indicate the average of ten A/T bases (local base and left 9) slipping across the genomes. The 2-micron plasmid is native to the original yeast strain. pRS426-ApPurZ-ApdATPase and pRS425-ApDUF550 were genetically engineered to express *PurZ/dATPase* and *DUF550* from *Acinetobacter* phage SH-Ab 15497, respectively.

Figure 4. Assessments of dZ-DNA/canonical read classification by Z-Calling.

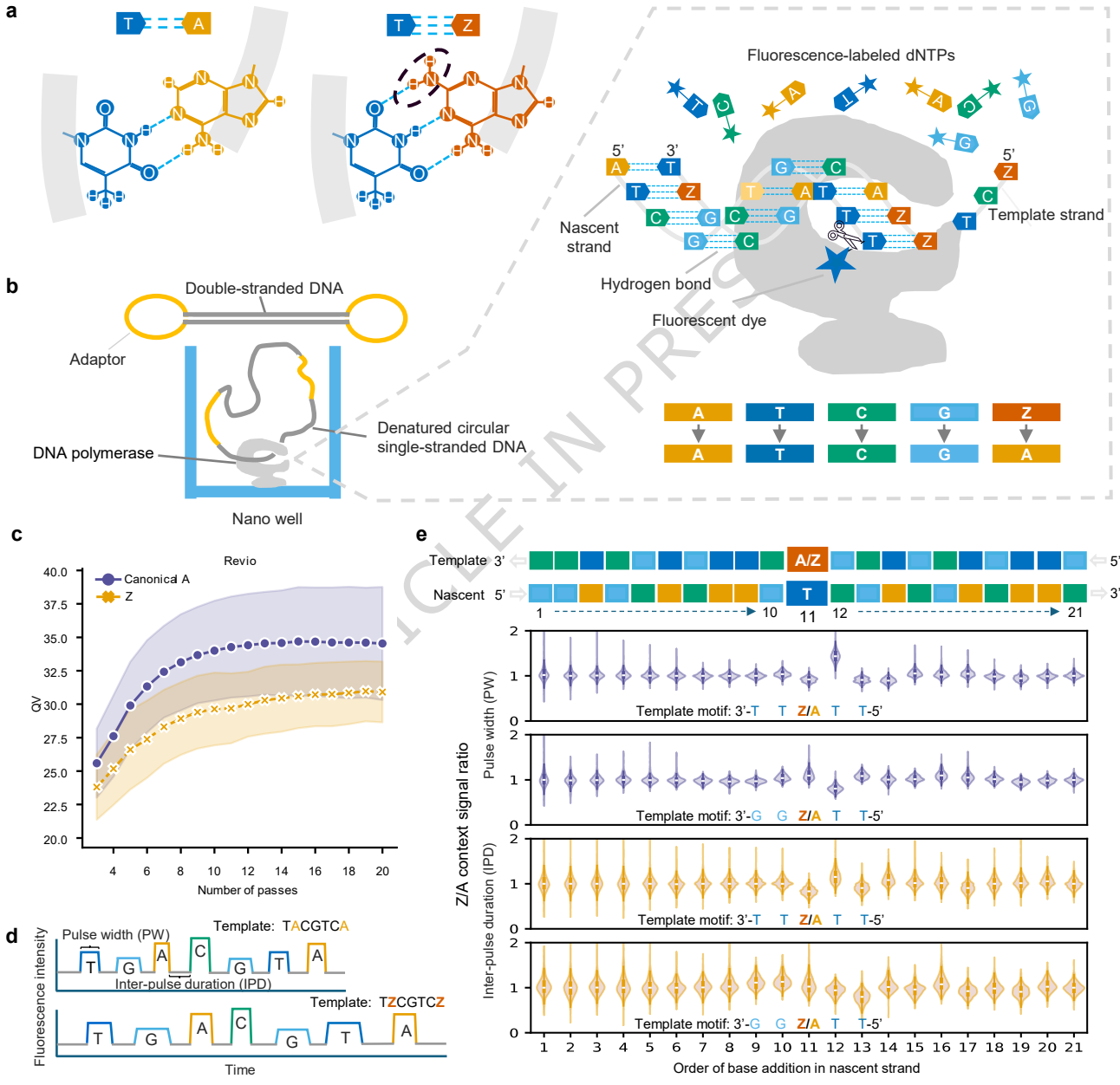
(**a**) Proportion of reads classified as dZ-DNA by the SVM classifier across dZ-negative datasets (blue background) and dZ-positive datasets (red background). The dZ-DNA-negative datasets included native DNA datasets (from arabidopsis, rice, human, *E. coli*, and zebrafish) and canonical amplicon datasets (from fruit fly). The dZ-positive datasets included datasets of native dZ-DNAs (Phage 15497), dZ-DNA amplicons (from *E. coli* and fruit fly), and hybrid DNAs [transformed yeast with ~24.7% Z/(A+Z)]. For each dataset (except for Phage 15497 that has only 6,337 reads in total), three sets of 20,000 reads were randomly sampled for replication. Bar heights represent the mean dZ-DNA ratios detected across replicates, and error bars indicate SDs. (**b**) and (**c**) Detection of dZ-DNA source in artificial Sequel II (**b**) and Revio (**c**) metagenome datasets. In (**b**), fruit fly reads served as the dZ-positive source, while human and zebrafish reads served as dZ-negative controls. In (**c**), dZ-positive sources included fruit fly, *E. coli*, Phage 15497, and transformed yeast. LR+ denotes positive likelihood ratio. The metagenome datasets were generated through mixing dZ-negative and dZ-positive reads whose counts are indicated by the stacked bar heights [proportional to log10(Read count + 0.1) as denoted by the labels] in the bottom panel, in which 0.1 was added to avoid null return due to 0 read count. Detailed statistics are supplied in **Supplementary Data 9**. (**d**) Computational benchmark of read classification time efficiency by Z-Calling. Time consumption of three steps (read filter, MLP, and SVM) in read classification was benchmarked using four different datasets of different read lengths and data amounts. The benchmark tests were carried out on a computer configured with AMD EPYC 7402 24-Core Processor and 256 GB memory. One core was used in Read filter and SVM classification steps, and 16 cores were used in the MLP (*k21-full-AZ*) calling step.

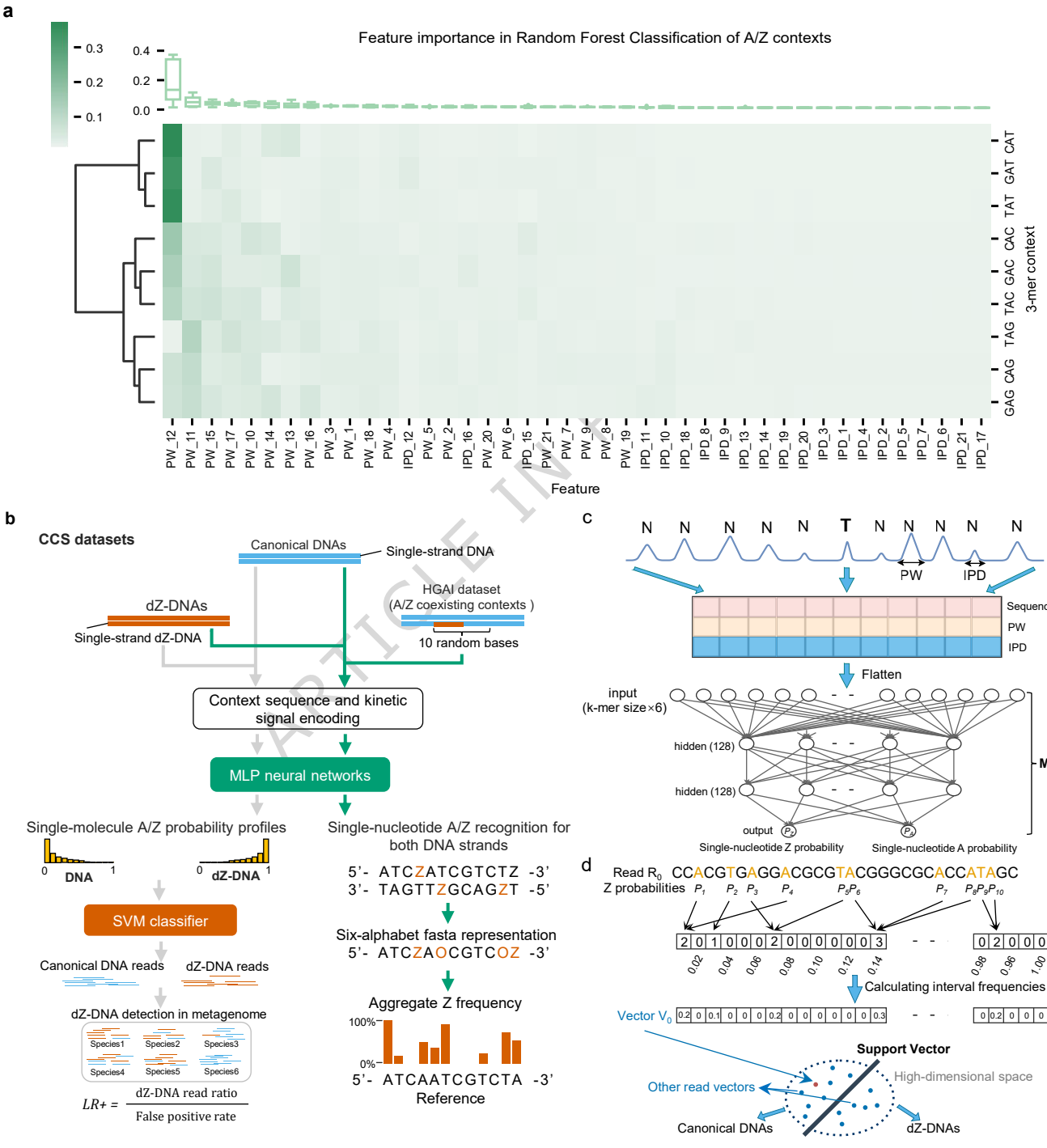
Editorial Summary:

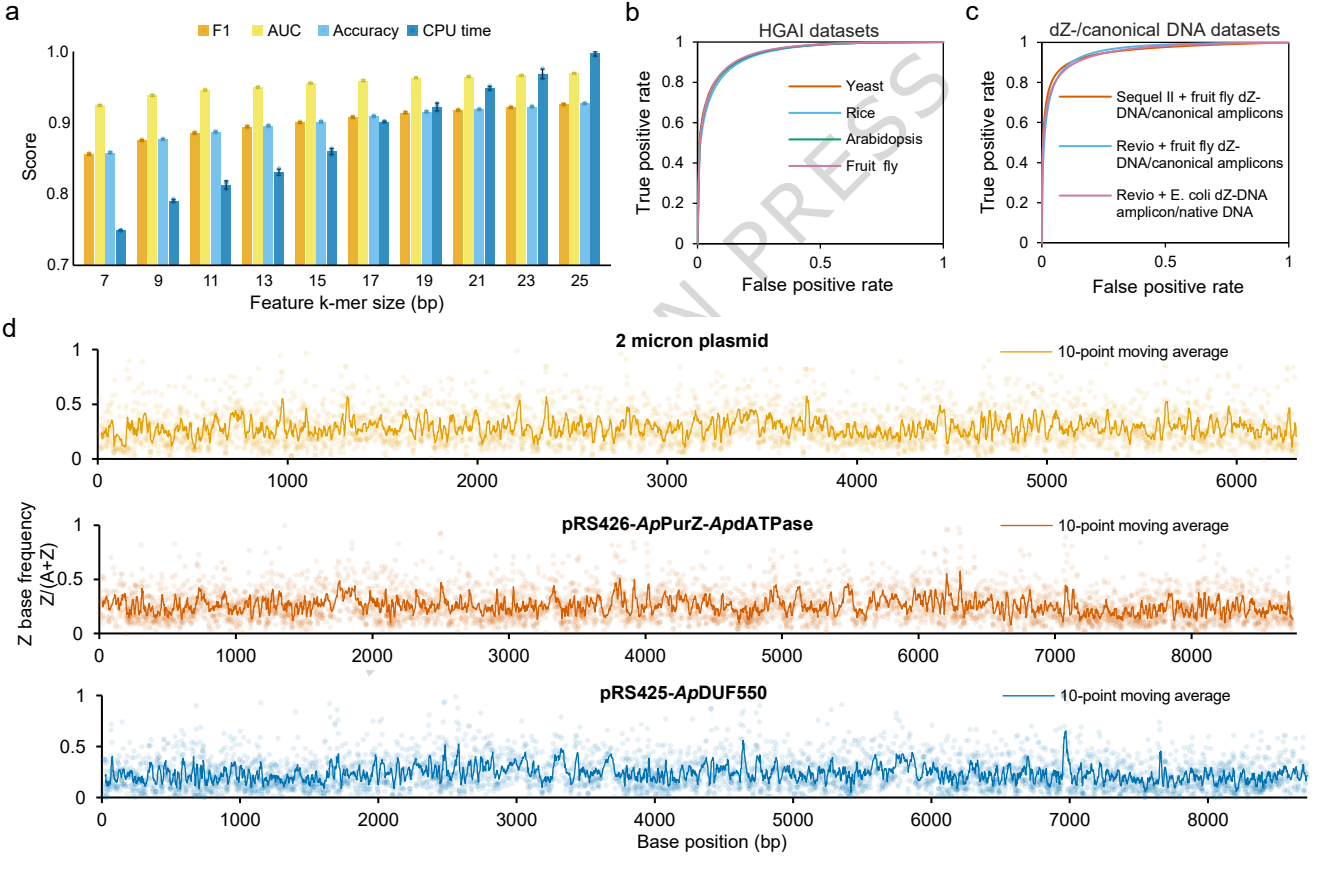
Machine learning models using PacBio kinetic signals enable high-sensitivity detection of dZ-DNA and single-nucleotide A/Z base calling to explore the biological roles and biotechnological potential of 2,6-diaminopurine.

Peer Review Information:

Communications Biology thanks Shanmuga Sozhamannan who co-reviewed with Rachael Sparklin; Osman Doluca and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Rosie Bunton-Stasyshyn. A peer review file is available.




a
Feature importance in Random Forest Classification of A/Z contexts
0.3
0.2
0.1
0.4
0.2
0.0
GAG CAG TAG TAC GAC CAC TAT GAT CAT
3-mer context
PW_12 PW_11 PW_15 PW_17 PW_10 PW_14 PW_13 PW_16 PW_3 PW_1 PW_18 PW_4 IPD_12 PW_5 PW_2 IPD_16 PW_20 PW_6 IPD_15 PW_21 PW_7 PW_9 PW_8 PW_19 IPD_11 IPD_10 IPD_18 IPD_8 IPD_9 IPD_13 IPD_14 IPD_19 IPD_20 IPD_3 IPD_1 IPD_4 IPD_2 IPD_5 IPD_7 IPD_6 IPD_21 IPD_17
Feature
b
CCS datasets
Canonical DNAs
Single-strand DNA
dZ-DNAs
Single-strand dZ-DNA
HGAI dataset
(A/Z coexisting contexts)
10 random bases
Context sequence and kinetic signal encoding
MLP neural networks
Single-molecule A/Z probability profiles
DNA
dZ-DNA
SVM classifier
Canonical DNA reads
dZ-DNA reads
dZ-DNA detection in metagenome
Species1
Species2
Species3
Species4
Species5
Species6
LR+ = dZ-DNA read ratio / False positive rate
Single-nucleotide A/Z recognition for both DNA strands
5'- ATCZATCGTCTZ -3'
3'- TAGTTZGCAGZT -5'
Six-alphabet fasta representation
5'- ATCZAOCGTCOZ -3'
Aggregate Z frequency
100%
0%
5'- ATCAATCGTCTA -3'
Reference
c
N N N N N T N N N N N
PW
IPD
Sequence
PW
IPD
Flatten
input
(k-mer size×6)
hidden (128)
hidden (128)
MLP
output
Single-nucleotide Z probability
Single-nucleotide A probability
d
Read R0 CCACGTGAGGACGCGTACGGGCGCACCATAGC
Z probabilities
2 0 1 0 0 0 2 0 0 0 0 0 0 3
0 2 0 0 0
0.02 0.04 0.06 0.08 0.10 0.12 0.14
0.98 0.96 1.00
Calculating interval frequencies
Vector V0
0.2 0 0.1 0 0 0 0.2 0 0 0 0 0 0 0.3
0 0.2 0 0 0
Support Vector
Other read vectors
High-dimensional space
Canonical DNAs
dZ-DNAs


a
F1
AUC
Accuracy
CPU time
Score
Feature k-mer size (bp)
b
HGAI datasets
Yeast
Rice
Arabidopsis
Fruit fly
True positive rate
False positive rate
c
dZ-/canonical DNA datasets
Sequel II + fruit fly dZ-DNA/canonical amplicons
Revio + fruit fly dZ-DNA/canonical amplicons
Revio + E. coli dZ-DNA amplicon/native DNA
True positive rate
False positive rate
d
2 micron plasmid
10-point moving average
pRS426-ApPurZ-ApdATPase
10-point moving average
pRS425-ApDUF550
10-point moving average
Z base frequency
Z/(A+Z)
Base position (bp)