DeepSME: De Novo Nanopore Basecalling of Motif-insensitive DNA Methylation and Alignment-free Digital Information Decryptions at Single-Molecule Level

Qingyuan Fan[1], Xuyang Zhao[1], Junyao Li[1], Ronghui Liu[1], Ming Liu[2], Qishun Feng[3], Yanping Long[4], Yang Fu[4], Jixian Zhai[4], Qing Pan[5,], Yi Li[1,*]*

1. School of Microelectronics, MOE Engineering Research Center of Integrated Circuits for Next Generation Communications, Southern University of Science and Technology, 518055, Shenzhen, China

2. School of Medicine, Southern University of Science and Technology, 518055, Shenzhen, China

3. National Clinical Research Center for Infectious Diseases, Shenzhen Third People's Hospital, The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen, 518112, China

4. Department of Biology, School of Life Sciences, Southern University of Science and Technology, 518055, Shenzhen, China

5. College of Information Engineering, Zhejiang University of Technology, 310023, Hangzhou, China

Email: liy37@sustech.edu.cn (Y.L.); pqpq@zjut.edu.cn (Q.P.)

Abstract

Biomolecular encryption employing chemical modifications enables secure approaches for information storage and communications. However, constructing high information density pathways for rapid synthesis and readout remains a challenge to guarantee confidentiality, integrity, and availability (CIA). Here we develop a nanopore sequencing based protocol, demonstrated by complete substitution using 5-hydroxymethylcytosine (5hmC) for individual nucleotide recognition rather than sequential interactions. Such motif-insensitive methylation at the single-molecule level does not naturally exist and results in severe ion current disruption and a 67.2%-100% readout failure, which ensure its ability on the encryption of the data encoded inside the DNA. We further propose and establish an alignment-free DeepSME

basecaller, which is a deep learning-based platform independent on prior models and knowledges. DeepSME utilizes a three-stage training pipeline that initiates tolerable for 11.55% errors, expands its neighboring k-mer dictionary model size from 4^6 to 4^9, and mitigates the errors by only three microbial genomes, giving rise to 92% precision with 92% recall. Fully 5hmC encrypted digital information were deciphered by DeepSME within 16× coverage depth. The versatile and transparent DeepSME pipeline and its F1-score performance of 86.4% surpassing all the state-of-the-art basecallers, support its great potential for meeting the rapidly increasing CIA demands of DNA-based secure communications.

Keywords

DNA storage, secure communication, nanopore sequencing, 5-hydroxymethylcytosine, *de novo* private basecaller

Introduction

Storing and delivering sensitive data safely lie at the heart of information technology. Many efforts have been devoted to exploring an accessible approach to achieve information encryption, such as cryptoristors[1], metahologram[2,3], quantum photonic system[4,5], DNA based technology[6–9], synthetic macromolecules[10,11], where encoded information can be recognized only under specific conditions or tools. Amongst them, biomolecular encryption that utilizes chemical modifications or interactions instead of computational schemes has been described as an alternative where nucleic acids, proteins[12], aptamers[13], and bacteria[14] have been exploited to protect messages for secure communication. For instance, DNA-based steganography developed by Clelland et al.[15] was further leveraged by DNA origami cryptography[16], which creates a key with a size of over 700 bits with oligonucleotides. However, these strategies exploit molecular interactions with low information densities, largely ignoring the chemical modification potential of densely packed nucleobases, thus preventing them from being used in sequencing-based DNA storage applications.

Sequencing-based communication requires nucleobase-level encryption to achieve high information density. Unfortunately, current sequencing data ought to be decoded using public codecs, which offers transparency but compromises security. Although bases can be methylated or oxidized[17], encryption methods are currently limited. However, modifications of all four

common nucleobases are feasible and can be achieved through either natural enzymatic processes[18] or synthetic methods[19].

Taking 5-hydroxymethylcytosine (5hmC) modification for example, it was discovered as a relatively abundant form of cytosine modification in Purkinje neurons and mouse embryonic stem cells (mESCs)[20]. In addition to being an intermediate of 5-methylcytosine (5mC) oxidation, evidence shows that 5hmC is a stable epigenetic mark with regulatory functions.[21] 5hmC - often referred as the "sixth base" of DNA - could be a chemically-stable candidate to replace cytosine to provide base-level protections for the messages in single DNA molecules.

Reversibly unlocking biomolecular protections - in most of the cases - requires either optical or electrical readout approaches. Nanopore sequencing, an emerging single-molecule technique, translates the sequence of nucleobases into ion flow blockades when a single strand of DNA passes through a nanopore. Such squiggled ionic current with distinct levels contains the neighboring k-mer sequence information, which can be interpreted (so-called basecall) into conventional sequenced results via neural networks[22]. Furthermore, modified nucleobases provide additional steric hinderances that give rise to the deviation of current levels. Thus, digital information can be embedded into these significant current discrepancies, which avoid being basecalled into correct sequences.

Reading 5hmC modification has been investigated using MspA nanopores[23–25] and solid-state nanopores[26] since early 2000. For commercial nanopore devices, Liu et. al, reviewed nanopore basecallers for methylations including 5hmC modifications[27], all of which were restricted with light-weighted modification[28] or CpG motifs[29]. This is because in biological systems, only 2.63% of C (2.42‰ of a given whole microbial genome) is currently modified to 5mC or 5hmC[30]. These sparse modifications lead to weak "perturbations" of the signals. Thus, state-of-the-art pipelines such as DeepSignal[31], DeepMod[32] and conventional basecallers such as Rerio-5mC-5hmC and Incremental Learning and Anomaly Detection (IL-AD[33]) may fail for nearly complete 5hmC modifications, which is a prerequisite to the success of nanopore decryption. Currently, the 5hmC-based DNA decryption using nanopore sequencing has yet to be investigated due to the challenges on k-mer model extraction, sophisticated alignment and time-consuming correction of datasets, hindering the development of decryptable DNA readout.

In this work, we proposed a highly informative scheme based on motif-insensitive 5hmC modification for encryption and nanopore sequencing based *de novo* basecaller for decryption. We encrypted text and image into fully 5hmC modified DNA sequences and further verified

by nanopore sequencing via conventional basecallers. We developed a three-stage alignment-free training method for Deep-learning based Single-Molecule Encryption (DeepSME), which tackle the basecalling bottleneck of the heavily modified dataset by expanding k-mer dictionary from scratch. Without any prior knowledges, we provided independent k-mer tables and exploit the properties of these unique signal disruptions at single-molecule level. This scheme together with the training method of secured DeepSME basecaller underpinned the potential for secure DNA-based data storage and communication with high information density, addressing the increasing demand for robust information security in an era of evolving biotechnological threats.

Results

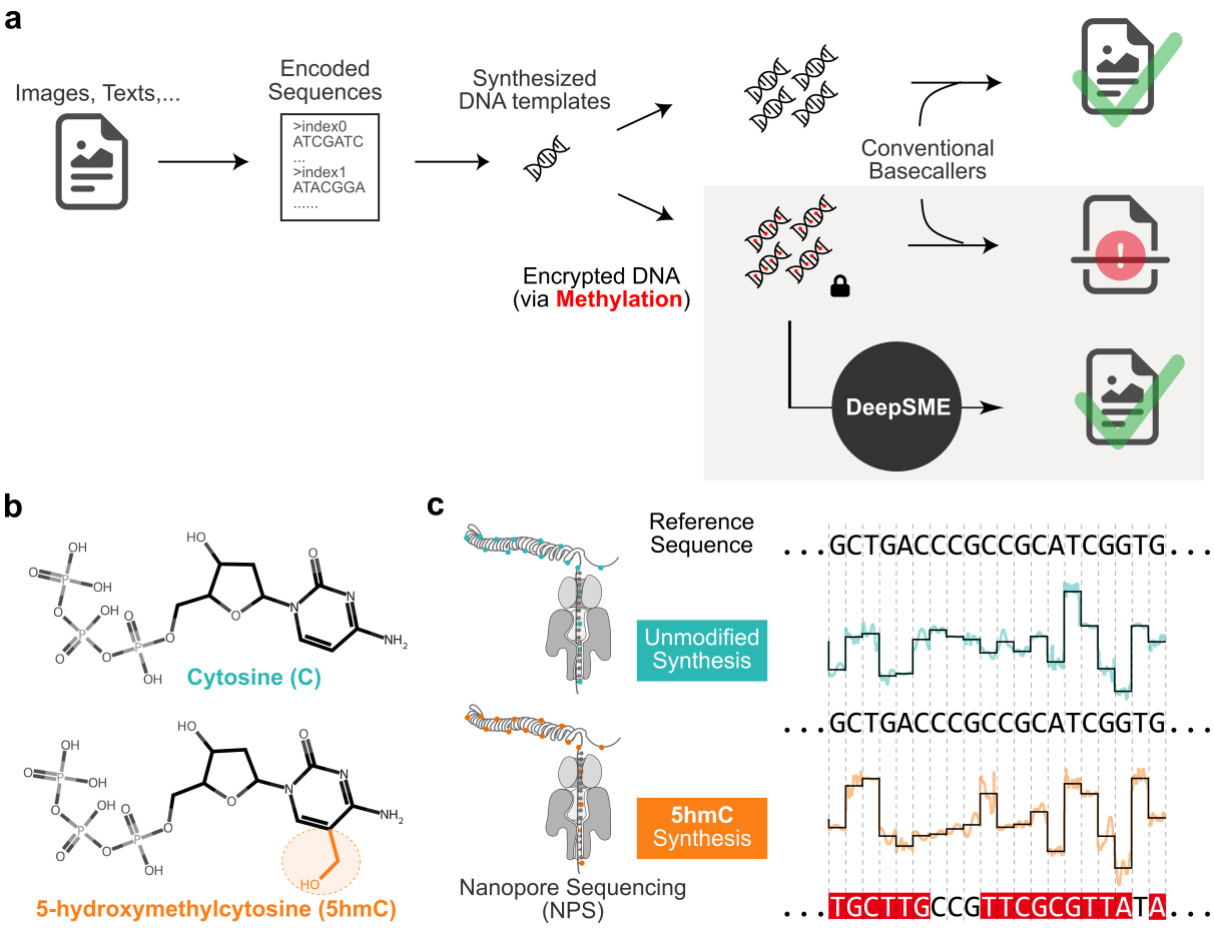

Fig. 1 | Encryption and decryption of digital information stored in DNA using nanopore sequencing. (**a**) Schematics of secure communication for files and texts stored in DNA. (**b**) Chemical structure of deoxycytidine triphosphates (dCTP), which is labelled as cytosine (C) and 5-hydroxymethylcytosine (5hmC), respectively. (**c**) Nanopore sequencing of unmodified and fully methylated 5hmC DNA strands. The ionic current signals of C (cyan) and 5hmC

(orange) are basecalled into DNA sequences, which are aligned with reference sequences. The error bases are labelled in red.

Fig. 1a illustrates the scenario for digital information stored in DNA and proposed encryption and decryption for secure communications. Texts and Images can be transcoded into DNA sequences, followed by the DNA synthesis as templates. Conventional polymerase chain reactions (PCR) amplify the templates into double strands and can be sequenced by either next generation sequencing (NGS)[34] or nanopore sequencing (NPS)[35]. Furthermore, data retrievals can be realized by the combination of basecalling and decoding steps for correcting the errors during public communications. Here secure communication with encrypted DNA strands, shown in Fig. 1a, disturbs the translation process by conventional basecallers such as Guppy 6.0, Bonito as well as Dorado. Thus, digital information is locked and fails to be decoded at single-molecule level of DNA strands. Meanwhile, DeepSME - a Deep-learning method for unlocked Single-Molecule Encryptions that we proposed, is expected to produce correct sequences for decoders.

Next, we are inspired to achieve encryption of stored digital information by DNA methylation**.** Fig. 1b displays the chemical structures of deoxycytidine triphosphate (dCTP), where the two analogies are labelled as cytosine (C) and 5-hydroxymethylcytosine (5hmC). Due to an additional hydroxymethyl group at 5', a distinct steric hinderance could be expected with close-to-neutral hydrophilicity.

Nanopore sequencing is realized to equip with basecallers for information readout. Fig. 1c illustrates the principle of nanopore sequencing for the DNA strands without and with the 5hmC modification. The resulting current signals can be squiggled with guided black lines and further aligned to the combination of neighboring oligonucleotides (k-mers) via basecallers. For instance, the current signal in cyan represents the basecalled reference sequence (GCTGACCCGCCGCATCGGTG). Meanwhile, the identical DNA sequences with 5hmC modification give rise to a distinct current signal in orange, leading to heavily misinterpreted letters (16 out of 20 bases, TGCTTGCCGTTCGCGTTATA) in the state-of-the-art Guppy 6.0. These current differences prevent conventional basecallers from accurately reading the modified sequences, corroborating encryptable 5hmC replacement at single-molecule level.

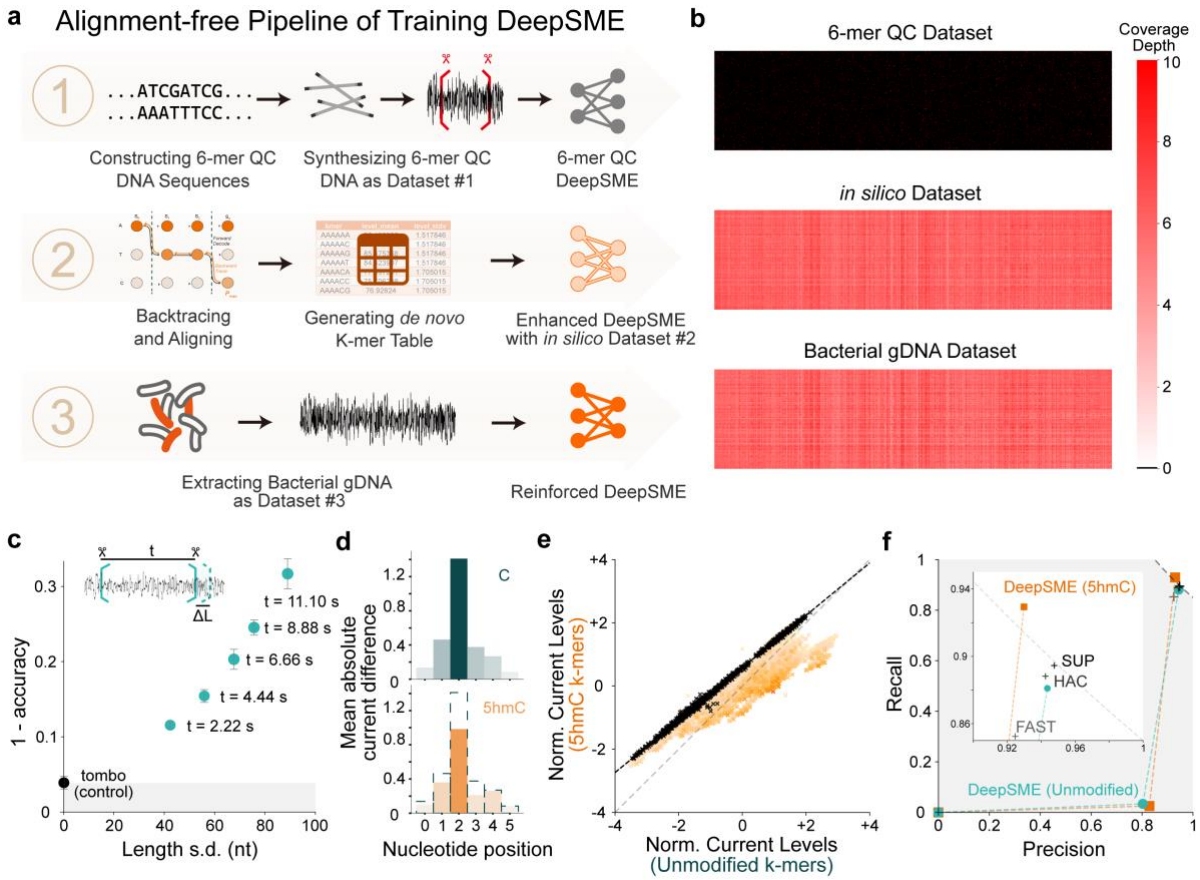

Fig. 2 | *De novo* DeepSME construction. (**a**) Training DeepSME with a three-stage alignment-free pipeline. (**b**) 9-mer coverages of three used training datasets. (**c**) Training error rate as a function of standard deviation of segmented lengths. Inset depicts the uncertainty (standard deviation of a certain segmented length) (**d**) Mean absolute current difference (MACD) of both unmodified cytosine (top) and 5hmC (bottom) at different nucleotide position of a k-mer oligonucleotide. (**e**) Methylated 5hmC vs. unmodified cytosine of normalized current levels. (**f**) Recall vs. precision of DeepSME for both unmodified and methylated 5hmC.

We propose a three-stage alignment-free pipeline for DeepSME basecaller, shown in Fig. 2a. 11 type of DNA sequences ranging from 1145 nt to 1341 nt are firstly sequenced to construct the 6-mer quality check dataset without any alignments. Tolerating these alignment errors, an initial DeepSME is trained to provide a k-mer model as a new dictionary. Benefiting from investigated k-mer dictionary and nanopore simulators such as scrappie squiggle[36], squigulator[37], and deepsimulator[38], *in silico* dataset can be generated for length variation and expanded the dictionary from 4^6 to 4^9. Not only mitigating the alignment errors, simulated current signals can also expand the physical redundancy (also named as coverage depth,

denoting the depth or completeness of DNA sequencing on the reference sequence). Depicted in Fig. 2b, the upper heatmap of full 6-mers shows a sparse occupation (2.60%) in the 9-mer space, while *in silico* dataset can cover almost all the 9-mers, comparable to the experimental microbial whole-genome sequences (gDNA) dataset with slightly different frequency (see Supplementary Table S1). Last, three types of microbial gDNA are involved and automatic aligned by enhanced DeepSME, which leads to the final training of the reinforced DeepSME.

In the first stage, the 6-mer quality check (QC) DNA is sensitive to initiate the DeepSME. It is hardly practical to mark 500 annotations for every 10,000 heavily modified nucleobases manually with no prior knowledges. Taking conventional segmentation tool Tombo[39] (Nanopolish[40] also works) on unmodified dataset as a control, 3.9% alignment errors will always exist. Considering billions of sequences and bases that are being used, Fig. 2c shows that the alignment error rate as a function of length variation. Given approximately 450 nt/s as the sequencing speed of a nanopore[41], individual current signals segmented as 2.22 s (corresponding to roughly 1000 nt), 4.44 s (~ 2000 nt), 6.66 s (~ 3000 nt), 8.88 s (~ 4000 nt) and 11.10 s (~ 5000 nt) contribute to the length variations of 40 nt to 85 nt. The longer the segmentation goes, the more errors the preliminary DeepSME holds.

The inset of Fig. 2c illustrates the segmentation-induced length variation. A fixed number of bases can be assigned to segmented current blocks each time as a chunk, creating a dataset containing tens of thousands of chunks for the training of the basecaller. The length variation itself is proportional to the segmented time. Our preliminary DeepSME segmented chunks at 2.22 s is estimated to have an alignment error of 11.6%, which seems tolerable for further investigations.

DeepSME reveals reduced currents with the presentation of 5hmC in the second stage. Fig. 2d shows the normalized current difference for the unmodified C and 5hmC in the extracted k-mer model from DeepSME. By substituting either C or 5hmC at each position (x-axis), the mean averaged current difference (MACD) is calculated over all k-mers in the model. In this 6-mer model, the base at index 2 has the highest impact on the current change. This is in a reasonable agreement with reported k-mer model from ONT[42]. Additionally, the weight of the bases at index #0, #1, #2, and #3 on the current significantly decreases when 5hmC modification takes place. This indicates that 5hmC modification leads to a reduction in the normalized currents at these corresponding base positions compared to the unmodified state.

On the other hand, Fig. 2e shows a scattered plot of normalized current levels between each unmodified C and 5hmC-modifications. Each point in the figure represents a 6-letter combination, where those not containing any C's shown in black, and those containing 1 to 6 C's shown in orange. The darker the orange color is, the higher the C content goes. The black dashed line with a slope of 0.83 with the offset of 0.58 represents a linear fitting for 6-mers without C's. Surprisingly, this line departs from the gray dashed line with the slope equals to 1 as light-weighted modifications do by Kovaka et al[42]. The slope changes suggest severe disruptions on the averaged normalized current, which benefit the encryption processes. The 6-letter combinations with 5hmC show lower normalized current levels compared with unmodified ones, which is in line with our observations in Fig. 2d.

The basecalling accuracy and recall are overviewed at the third stage. Fig. 2f illustrates the relationship between recall and precision of DeepSME at three stages for both unmodified C and 5hmC. First, the preliminary basecallers locate at the bottom left, meaning low precision and low recall capabilities since only 6-mer conditions (2%) are learnt so far. Second, the enhanced basecallers perform much better on precision but remain low on the recall, which is reasonable since the simulated dataset lack the ability of generalization for variant conditions. Last, the reinforced basecallers show significant improvements on both precision (0%-82.9%-93.0%) and recall (0%-2.5%-92.9%). The reinforced 5hmC basecaller has a slightly lower accuracy but a higher recall than that of unmodified one. Compared with the commercial basecallers (Bonito FAST, HAC - high accuracy and SUP - super accuracy), the reinforced 5hmC basecaller is beyond the F1-score of 0.85 (grey dashed line), surpassing Bonito SUP, HAC as well as FAST.

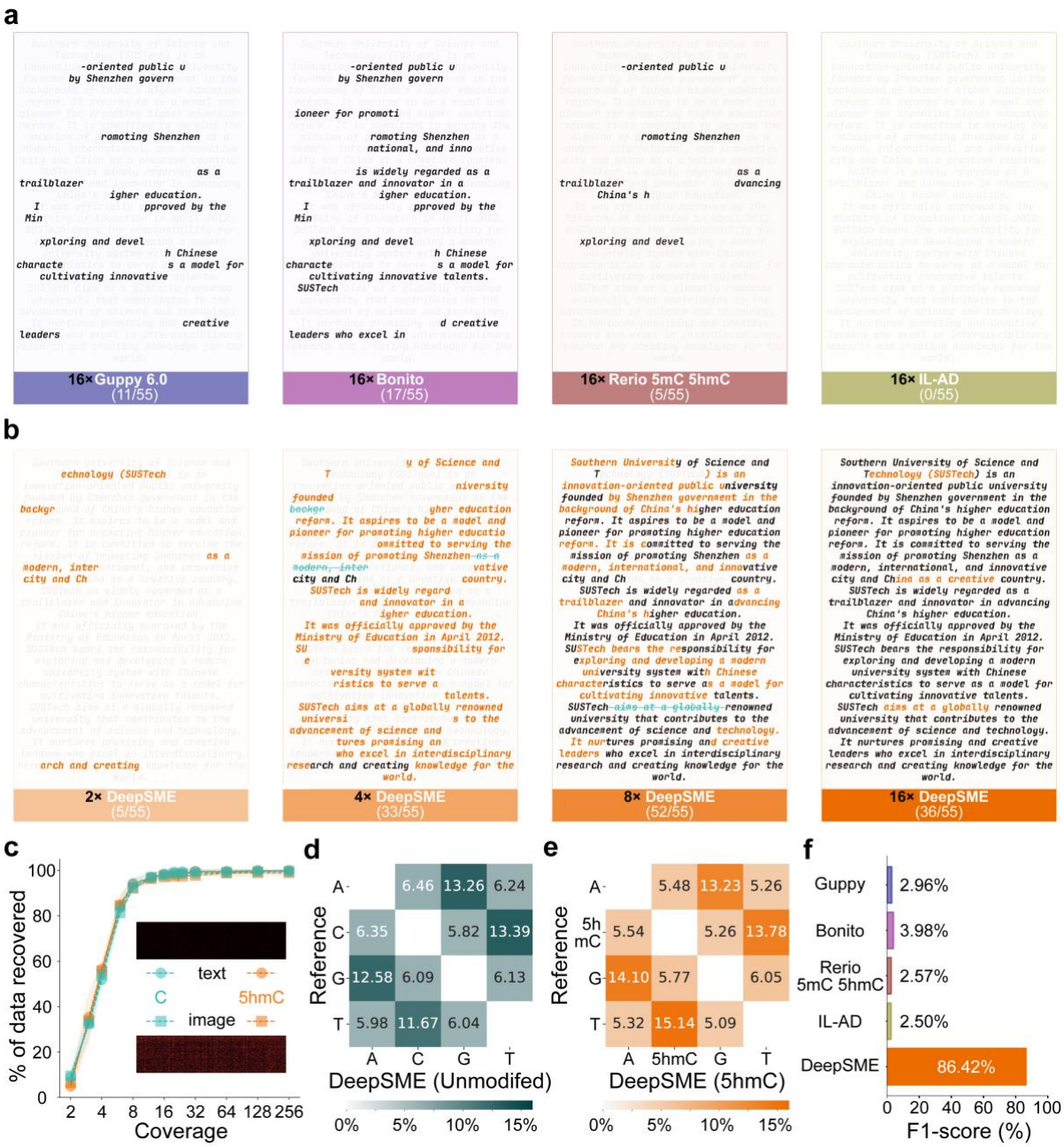

Fig. 3 | Decryption performance of digital information enabled by DeepSME. (**a**) Decryption performance of the text file using Guppy 6.0 (dna_r9.4.1_450bps_hac config), Bonito (dna_r9.4.1_e8_sup@v3.3 config), Rerio 5mC 5hmC (res_dna_r941_min_modbases_5mC_5hmC_v001 config), Incremental Learning and Anomaly Detection (IL-AD, 5hmC config) with coverage depths of 16×. (**b**) Decryption performance of the text file using DeepSME with coverage depths of 2×, 4×, 8× and 16×. The decrypted strand amounts are 5, 33, 52 and 55 of total 55. (**c**) The percentage of data recovered as a function of coverage for both the text file (circles) and the image file (squares) stored in unmodified (cyan) or 5hmC (orange) using corresponding unmodified or 5hmC DeepSME,

respectively. (**d**) Base bias of substitutions for unmodified DeepSME. (**e**) Base bias of substitutions for 5hmC DeepSME. (**f**) F1-scores for variant basecallers on encrypted (5hmC) DNA.

We turn our focus onto the decryption journey of 5hmC modified text file (sustech_introduction.txt, 978 bytes, 55 strands) and image file (sustech_logo.jpg, 7,775 bytes, 432 strands). Fig. 3a shows the decoding performance of the identical text file using state-of-the-art basecallers on encrypted (modified) DNA with 16× coverage. For the basecallers that could not reach the coverage amount, all basecalled sequences were obliged to use. Bonito achieved a decode rate of 30.91%, Guppy 6.0 reached 20.00%, Rerio 5mC 5hmC managed 9.09%, while IL-AD was unable to decode any sequences at all.

Encouraged by the high precision and recall of DeepSME, we perform decryption experiments displayed in Fig. 3b. The recovered texts are shown with the increasing physical redundancy (the coverage depth from 2× to 16×). It is evident that DeepSME produced 60.0% of the text after decoding with 4×, and further reached 94.5% data retrieval with 8× coverage. Fig. 3c compares the data retrieval performance of the two DeepSMEs as a function of coverage depth. They exhibit almost identical characteristics with narrowed variation areas that each point of recovered rate was repeated 10 times.

To gain more insides from the two basecallers, the substitution probabilities of unmodified and 5hmC DeepSME are illustrated in Fig. 3d and 3e. The substitutions from T to C and C to T are 11.67% and 13.39%, while these values go to 15.14% and 13.78% for T-5hmC and 5hmC-T. A-G and G-A are also high for both. The two matrixes are quite comparable, which are also in agreement with the matrix from Lopez et al[35].

To analyze the comparison with alternative basecallers, Fig. 3f presents the F1-scores on our encrypted (modified) DNA dataset. DeepSME achieved a leading F1-score of 0.864 to date, while other basecallers such as Guppy 6.0, Bonito Super, Rerio, and IL-AD had F1-scores ranging only from 0.025 to 0.040.

Discussion

Our work demonstrates a superior F1-score of DeepSME, specifically as high as 86.42% for our synthesized DNA sequences in Fig. 3f as the testing dataset, advancing in the following aspects: 1) *Knowledge extraction*. We used a Connectionist Temporal Classification (CTC) decoding layer as the interface and the criteria for segment the current to obtain a k-mer model as a dictionary to inspect the states of sequences. As an intermediate step, this model twists

experimental 6-mer DeepSME smoothly into a simulated 9-mer one. 2) *Alignment-free and prior knowledge independent*. Conventional basecallers highly reply on the universal k-mer table provided by Oxford Nanopore Technology (ONT). Our method limits the alignment errors stepwise by segmented length (chunk size), simulated and experimental data with high redundancy. 3) *Transparency*. Surprisingly, 8 hours of sequencing and 3 hours of basecaller training of three originated 5hmC-modified microbial genome dataset do fulfill the requirements of digital information and reasonably high performance with an open-sourced deep-learning neural network. Specially tailored DeepSME basecallers could be delivered for variant conditions.

On the other hand, our DeepSME has limitations summarized as followed: 1) *Generalizability*. applying DeepSME directly on biological samples is still doubtful, although it surpasses all the existing basecallers for full 5hmC methylation, to the best of our knowledge. This basecaller has not been optimized for 5hmC with different mixing ratio with C, 5mC and other derivatives. 2) *Insufficiency of datasets*. Depart from Fig. 2b and Fig. 3c, the state-of-the-art basecallers such as Albacore[43], Guppy, Bonito[44], Dorado, to name a few, do have better performances[45] accounting for a wide variety of data sources, including human, animal, microbial, and synthetic DNA. Variant methylations and oxidations exist broadly over them[46], which introduces greater variability and robustness and are not currently included in our dataset. 3) *Cross-talking of primer classification*. Pore models with unknown precision may require more runs of classification experiments to exclude the errors and correctly determine the eleven groups for each 6-mer QC sequence in the sequencing pool. However, our preliminary tests in the Supplementary Figure S5 suggest a smooth transition from R9.4.1 to R10.4.1 use DTW align algorithm[47].

Our three-stage training pipeline is robust for DeepSME where each step is crucial and addresses specific challenges for achieving high F1-score. First, the initial DeepSME lay the fundamental for 6-mers that the second and third 9-mer basecallers reply on. Although the precision and recall are extremely low for the final DNA storage dataset, the second and the third DeepSME cannot converge if this step was skipped. Second, subsequent training with the *in-silico* dataset expands the coverage depth from 6-mers (3%) to 9-mers (100%), mitigating alignment errors by simulated signals which are more predictable and less variations on time scale. The amount of 81,490 simulated strands thus lead to the increase of more than 80% precision, while the recall remains at 2.5%. Directly applying the initial 6-mer QC DeepSME on the third experimental microbial gDNA datasets contains much more alignment errors. Thus,

the second step is important to bridge between the first and the third ones. Finally, the recall rate is rescued by incorporating diverse microbial gDNA datasets. This allows the model to generalize and accurately decode real-world, heavily modified sequences. Omitting any of these steps would hamper the DeepSME's performance, highlighting the vitality of this progressive training approach.

Alternative basecallers other than DeepSME were hardly able to correct fewer than 5% of the sequences for modified sequences in Fig. 3f (see the Supplementary Information Section 6 for recall and precision of these basecallers). We believe this could be attributed to the shifted current deviations (Fig. 1c and Fig. 2e): some of the points still locate on the black dashed line to be recognized and aligned to the reference sequence, while quite a lot of points are well-distanced. This inability to correct 5hmC-modified sequences demonstrates that our base modification approach does enable the encryption method in DNA storage.

It is worth mentioning that 5hmC-modified sequences are likely to "terminate at half a journey"(see Supplementary Figure S3). When sequencing a 5,262 nt long plasmid sequence, 57.9% of the unmodified sequences were longer than 5,100 nt. However, the value for extension exceeding 5,100 nt dropped to only 18.7% for the 5hmC-modified sequences. 4 folds of strands creased at 1,000 nt - 2,000 nt (see Supplementary Figure S1, S2). We speculate that this may be due to alteration in the chemical structure of 5hmC, which reduces the efficiency of helicase (motor protein) binding to the 5hmC methylated strands. Thus, the degradation on the binding strength leads to more sequences terminated much earlier before the finish line. Shorter lengths such as between 1145 nt and 1341 nt that we used facilitate the generation of the datasets, while longer lengths are more beneficial for encryption that less information could be even touched.

In conclusion, we have proposed and verified a scheme for high-informative secure communication combining DNA methylation induced information encryption and nanopore sequencing based decryption, via *de novo* assembly of a private DeepSME basecaller. Avoiding tremendous alignments and corrections, we proof that a three-stage training method that each stage is compulsory for DeepSME to align both recall and precision to 92% incrementally, exceeding state-of-the-art basecallers on the digital information stored 5hmC DNA dataset. Within 16× coverage depth and negligible prior knowledges, our independent DeepSME pipeline fulfills the readout of encryptable single-molecule DNA, showcasing its great promise

for biomolecular-based secure communications, unclonable functions and anticounterfeiting systems.

Methods

Information Encryption via 5-hydroxymethylcytosine (5hmC) based Polymerase Chain Reaction. Synthetic DNA sequences with designed digital information were purchased from Twist Bioscience, while plasmids and primers were obtained from Tsingke Biotech. Polymerase chain reactions (PCR) were performed using 10 mM dATP, dTTP, dGTP and dCTP supplied by Sangon. The PCR process utilized Phanta Max Super-Fidelity DNA Polymerase (P505-d1) from Vazyme.

For each 50 µl reaction system, 15 µl of ddH2O was added along with 25 µl of 2× Phanta Max Buffer. Additionally, 1 µl of each 10 mM dATP, dTTP, dCTP, and dGTP was included. To this mixture, 2 µl of Forward Primer and 2 µl of Reverse Primer were added. Finally, 1 µl of Phanta Max Super-Fidelity DNA Polymerase (1 U/µl) and 1 µl of DNA were incorporated to complete the reaction mixture. Specifically, 5-Hydroxymethyl-dCTP (Cat. No. NU-932s) was purchased from Jena Bioscience as the alternative PCR substrate for dCTP. The PCR conditions were as follows: 1. Pre-denaturation at 95°C for 3 min, 2. Denaturation at 95°C for 15 min, 3. Annealing at 60°C for 15 seconds, 4. Extension at 72°C for 1.2 minutes or 1 minute per 1kb sequence, 5. Final Extension at 72°C for 5 minutes. After performing 30 cycles, the resulting products can be considered as a complete conversion from cytosine (C) to methylated 5hmC.

Electrical Readout via Nanopore Sequencing. Unmodified or 5hmc-modified DNA products after PCR process was subjected to electrophoresis on a 1.5% agarose gel for length validation and purification. The bands corresponding to the reference sequence length were excised, followed by DNA extraction using the E.Z.N.A. Cycle Pure Kit (V-spin) from Omega Bio-Tek. A solution containing 200 ng of DNA was prepared using the ND608 kit from Vazyme for DNA damage repair, end preparation and adapter ligation. This includes two steps: 1) Damage repair and end preparation: 2 µl Damage Prep Enzyme, 5 µl End Prep Enzyme, and 10 µl End Prep Buffer were added, with ddH_2O supplemented to a total volume of 65 µl. This mixture was incubated in a PCR thermal cycler at 30°C for 20 minutes, followed by 65°C for 15 minutes. 2) Adapter ligation: 25 µl Rapid Ligation Buffer 2, 5 µl Rapid DNA Ligase, and 5 µl AMX-F from the LSK-110 kit from Oxford Nanopore Technology were added to the reaction solution, and the mixture was incubated in a PCR thermal cycler at 20°C for 15 minutes.

Next, DNA purification was performed using VAHTS DNA Beads (N411-01) from Vazyme according to the bead purification protocol in the LSK-110 kit. The VAHTS DNA Beads were first resuspended by vortexing. Subsequently, 40 µl of the resuspended beads were added to the reaction in previous step and mixed by flicking the tube. The mixture was then incubated on a Hula mixer (rotator mixer) for 5 minutes at room temperature. After spinning down the sample and pelleting it on a magnet, the supernatant was pipetted off while keeping the tube on the magnet. The beads were then washed by adding 250 µl Short Fragment Buffer (SFB), flicking to resuspend, spinning down, and returning the tube to the magnetic rack to allow the beads to pellet. The supernatant was removed using a pipette and discarded. This washing step was repeated. After spinning down and placing the tube back on the magnet, residual supernatant was pipetted off and the beads were allowed to dry for about 30 seconds, ensuring the pellet did not crack. The tube was removed from the magnetic rack, and the pellet was resuspended in 15 µl Elution Buffer (EB). The mixture was spun down and incubated for 10 minutes at room temperature. The beads were then pelleted on a magnet until the eluate was clear and colorless, for at least 1 minute. Finally, 15 µl of eluate containing the DNA library was removed and retained in a clean 1.5 ml Eppendorf DNA LoBind tube.

Furthermore, DNA quantification was performed using a Qubit Fluorometer to measure 50 fmol of DNA. For flow cell priming and loading according to the LSK-110 protocol, 30 µl of thawed and mixed Flush Tether (FLT) was added directly to the tube of thawed and mixed Flush Buffer (FB) and mixed by vortexing at room temperature. Then, 800 µl of the priming mix was loaded into the flow cell (FLO-MIN106) via the priming port, avoiding the introduction of air bubbles, and left to sit for 5 minutes. During this time, the library for loading was prepared by mixing 37.5 µl Sequencing Buffer II (SBII), 25.5 µl Loading Beads II (LBII), and 12 µl DNA library including 50 fmol of DNA. An additional 200 µl of the priming mix was loaded into the flow cell via the priming port, again avoiding air bubbles. The prepared library was then mixed gently by pipetting up and down just before loading. Finally, 75 µl of the sample was added to the flow cell via the SpotON sample port in a dropwise fashion, ensuring each drop flowed into the port before adding the next.

Finally, the MinION was connected to a computer, and the MinKNOW v24.02.6 software was launched. The LSK-110 protocol and other default settings were selected. We also adjust the sequence length cutoff MinKNOW (20bp, 200bp or 1000bp) according to the expected median length of the sequences. of and the sequencing button was clicked to begin collecting nanopore sequencing current data.

Constructing of DeepSME Pipeline. *Design and Synthesis of 6-mer QC sequence.* The Cate_NAN plasmid (see Supplementary Table S9) of total length 5262 nt was used to constructing ten sequences of our 6-mer quality control (QC) dataset. Ten pairs of forward and reverse primers, ranging from 40 nt to 50 nt (see Supplementary Table S10), were designed and synthesized by Tsingke Biotech to generate corresponding ten groups of DNA sequences with length of 1330 nt, 1340 nt, 1340 nt, 1340 nt, 1340 nt, 1200 nt, 1200 nt, 1230 nt, 1226 nt and 1262 nt containing approximately 20 nt to 30 nt without cytosine (C) in their forward and reverse primers (see Supplementary Table S11). An additional sequence of 1144 nt was derived from a synthesized 1206 nt DNA template. These eleven key sequences contribute to 99.93% physical redundancy of 6-mers (4093/4096) with a median occurrence of three. The non-C segments on the eleven sequences were further used to facilitate subsequent data classification.

Both unmethylated and methylated PCR products were pooled and sequenced using an Oxford Nanopore Technology R9.4.1 flowcell with LSK-110 reagents. Specifically, the sequences were purified using the E.Z.N.A. Cycle Pure Kit (V-spin), followed by end repair and adapter ligation using the DNA Damage Repair Enzyme and End Prep Enzyme from the ND608 kit, and Rapid DNA Ligase, respectively. The final DNA samples were purified using VHATS DNA Beads, quantified with Qubit Flow, and loaded onto the flowcell for sequencing.

Preliminary Training of DeepSME with 6-mer datasets. The sequenced raw data were processed using Guppy 6.0 basecaller (dna_r9.4.1_450bps_hac.cfg) as a control. The basecalled fastq sequences were aligned back to the ionic currents using Tombo[39] 1.5.1 (https://github.com/nanoporetech/tombo). The current data were catalogued into 11 sequences, recognized by the non-C regions of the primers. These segments were also correctly recognized with 5hmC modification, ensuring automated classification of raw data into these 11 types of 6-mer QC sequences. The primer regions were then carefully trimmed, leaving only the 6-mer QC sequence regions.

Ionic current dataset from these 11 sequences were used for a preliminary training of a neural network with Encoder-Decoder architecture with Connectionist Temporal Classification (CTC) decoding layer (modified version of bonito_bonitorev_ctc in basecaller_benchmark repository[45], see supplementary information Table. S7). The stride of the Encoder CNN layer was reconfigured to 1 to achieve high CTC layer resolution of DeepSME, resulting in the training convergence among 6-mer QC dataset.

Extracting 6-mer statistics (k-mer model). Ionic current data were segmented by the probability matrix obtained from the CTC segmentation[48] of the trained 6-mer QC DeepSME. Sequenced average currents, standard deviations as well as averaged dwell times and related standard deviations can be analyzed for either unmodified or fully methylated 5-hmC sequenced by the R9.4.1 nanopore.

Expanding 6-mer to 9-mer with in silico simulations. Using the k-mer model from the previous step, simulated currents were generated for the whole genome fasta sequences of 50 microorganisms (Fasta can be download from https://github.com/marcpaga/nanopore_benchmark/blob/main/download/links_wick_data_train.txt) and simulated using Squigulator[37] 0.3.0 (https://github.com/hasindu2008/squigulator) with a specified k-mer table csv file.

Enhanced training of DeepSME with Simulated Datasets. In silico experiments were performed to generate training datasets of 118,299 chunks with a chunk size of 3,600 (0.9 s) using the data preparation functionality of Bonito 0.8.1. These datasets were used to fine-tune the Encoder-Decoder architecture with Conditional Random Field (CRF) decoding layer (dna_r9.4.1_e8_sup@v3.3 model in Bonito 0.8.1, see supplementary information. S8) over 5 epochs with a learning rate of 5e-4, resulting in the enhanced DeepSME.

Generating and Sequencing of Fully Methylated 5-hmC Bacteria Genome DNA (gDNA). Genome DNA from Pseudomonas_aeruginosa_PAO1 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000006765.1/), Aeromonas_hydrophila_AH54 (unpublished), and Vibrio_cholera_E1 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_026013235.1/) microorganisms were extracted using the Qiagen Miniprep kit. The DNA was fragmented using the TD502 kit (TruePrep DNA Library Prep Kit V2 for Illumina) from Vazyme and modified using 5-Hydroxymethyl-dCTP (NU-932s) from Jena Bioscience. PCR was performed with Phanta Max Super-Fidelity DNA Polymerase (P505-d1) from Vazyme over 18 cycles under the following conditions of six steps: 1) Extension at 72°C for 3 minutes. 2) Pre-denaturation at 95°C for 3 min, 3) Denaturation at 95°C for 15 min, 4) Annealing at 60°C for 15 seconds, 5) Extension at 72°C for 60 s-72 s per 1kb sequence, 6) Final Extension at 72°C for 5 minutes. The modified DNA samples were sequenced using the Oxford Nanopore Technology R9.4.1 flowcell and LSK-110 reagents.

Reinforced DeepSME with Fully5-hmC Bacteria gDNA Datasets. The enhanced DeepSME was used to basecall gDNA and map to the corresponded genome. The gDNA datasets from the three microbial samples also lead to 118,096 chunks with a chunk size of 3,600 (0.9 s). These data were then used to reinforce the Encoder-Decoder architecture with CRF decoding layer (dna_r9.4.1_e8_sup@v3.3 model in Bonito 0.8.1) over 5 epochs with a learning rate of 5e-4, yielding our final reinforced DeepSME.

Decrypting Data Stored in 5-hmC DNA using DeepSME. Dataset for digital information stored in DNA were prepared using our Composite Hedges Nanopore codec[49] in the four-letter configuration. The binary forms of two selected files (219 bytes and 4109 bytes) were extracted and split into payloads with 36 bytes. Next, Reed-Solomon (40,36) code was used to add redundancy to these payloads to generate segments. Then, a 34 nt barcode as addresses for random access followed with three 5-mer anchors are added to each segment. Finally, an oligo pool containing 487 single-stranded 243 nt DNA sequences was sent to Twist Biosciences, including 55 sequences that encode a .txt file and 432 sequence that encode a .jpeg image (see Supplementary Table S12 and Supplementary Table S13).

For the PCR amplification, a pair of carefully designed 23 nt 5'- and 11 nt 3'- flanking sequences was added to both ends of each DNA sequence. Unmodified or modified 5hmC PCR were performed to obtain DNA sequences without or with encryption, followed by above described nanopore sequencing procedure to obtain the current data.

The reinforced DeepSME was used to process the sequenced raw data. The resulting fastq and aligned bam files were performed using the decoder of our Composite Hedges Nanopore codec[49]. For performance metrics, "the rate of the number of bases that could be aligned to the reference FASTA in total bases" was used as Recall, "the rate of the number of correct bases in total bases" was used as Precision, and their product was used as the F1-score.

Data availability

Source data is provided with this paper. The sequencing fastq data that supports the findings o f this study and the weight of the DeepSME model has been deposited in the Zenodo (https://doi.org/10.5281/zenodo.12704171)

Code availability

The code package for this study is available in the GitHub repository (https://github.com/sparkcyf/DeepSME).

References

1. Kim, S. Il *et al.* Cryptographic transistor for true random number generator with low power consumption. *Sci. Adv.* **10**, eadk6042 (2024).
2. Qu, G., Yang, W. & Song, Q. Reprogrammable meta-hologram for optical encryption. *Nat. Commun.* (2020) doi:10.1038/s41467-020-19312-9.
3. Lim, K. T. P., Liu, H., Liu, Y. & Yang, J. K. W. Holographic colour prints for enhanced optical security by combined phase and amplitude control. *Nat. Commun.* **10**, 25 (2019).
4. Paraïso, T. K. *et al.* A photonic integrated quantum secure communication system. *Nat. Photonics* **15**, 850–856 (2021).
5. Babin, C. *et al.* Fabrication and nanophotonic waveguide integration of silicon carbide colour centres with preserved spin-optical coherence. *Nat. Mater.* **21**, 67–73 (2022).
6. Dickinson, G. D. *et al.* An alternative approach to nucleic acid memory. *Nat. Commun.* **12**, 1–10 (2021).
7. Kuzdraliński, A. *et al.* Unlocking the potential of DNA-based tagging: current market solutions and expanding horizons. *Nat. Commun.* **14**, 1–7 (2023).
8. Meiser, L. C. *et al.* Synthetic DNA applications in information technology. *Nat. Commun.* **13**, 352 (2022).
9. Zhang, Y. *et al.* DNA origami cryptography for secure communication. *Nat. Commun. 2019 101* **10**, 1–8 (2019).

10. Soete, M., Mertens, C., Badi, N. & Du Prez, F. E. Reading Information Stored in Synthetic Macromolecules. *J. Am. Chem. Soc.* **144**, 22378–22390 (2022).

11. Arcadia, C. E. *et al.* Multicomponent molecular memory. *Nat. Commun.* **11**, 691 (2020).

12. Wong, N. Y., Xing, H., Tan, L. H. & Lu, Y. Nano-encrypted morse code: A versatile approach to programmable and reversible nanoscale assembly and disassembly. *J. Am. Chem. Soc.* **135**, 2931–2934 (2013).

13. Liu, Y. *et al.* An aptamer-based keypad lock system. *Chem. Commun.* **48**, 802–804 (2011).

14. Palacios, M. A. *et al.* InfoBiology by printed arrays of microorganism colonies for timed and on-demand release of messages. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 16510–16514 (2011).

15. Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. *Nature* **399**, 533–534 (1999).

16. Zhang, Y. *et al.* DNA origami cryptography for secure communication. *Nat. Commun.* **10**, 5469 (2019).

17. Pagès-Gallego, M. *et al.* Direct detection of 8-oxo-dG using nanopore sequencing. *bioRxiv* 2024.05.17.594638 (2024) doi:10.1101/2024.05.17.594638.

18. Jeltsch, A. Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *ChemBioChem* vol. 3 at https://doi.org/10.1002/1439-7633(20020402)3:4<274::aid-cbic274>3.0.co;2-s (2002).

19. Feng, S., Zhong, Z., Wang, M. & Jacobsen, S. E. Efficient and accurate determination of genome-wide DNA methylation patterns in Arabidopsis thaliana with enzymatic methyl sequencing. *Epigenetics and Chromatin* **13**, 1–17 (2020).

20. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. *Science (80-.).* **324**, 929–930 (2009).

21. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nat. Chem. 2014 612* **6**, 1049–1055 (2014).

22. How basecalling works | Oxford Nanopore Technologies. https://nanoporetech.com/platform/technology/basecalling.

23. Laszlo, A. H. *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18904–18909 (2013).

24. Derrington, I. M. *et al.* Nanopore DNA sequencing with MspA. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16060–16065 (2010).

25. Butler, T. Z., Pavlenok, M., Derrington, I. M., Niederweis, M. & Gundlach, J. H. Single-molecule DNA detection with an engineered MspA protein nanopore. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20647–20652 (2008).

26. Wang, F. *et al.* Solid-State Nanopore Analysis of Diverse DNA Base Modifications Using a Modular Enzymatic Labeling Process. *Nano Lett.* **17**, 7110–7116 (2017).

27. Liu, Y. *et al.* DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. *Genome Biol.* **22**, 1–33 (2021).

28. Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* **14**, 411–413 (2017).

29. Yuen, Z. W. S. *et al.* Systematic benchmarking of tools for CpG methylation detection from nanopore sequencing. *Nat. Commun. 2021 121* **12**, 1–12 (2021).

30. Assembly of the Human Genome (GRCh38 Genome Reference Consortium Human Reference 38). https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/.

31. Ni, P. *et al.* DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* **35**, 4586–4595 (2019).

32. Ahsan, M. U., Gouru, A., Chan, J., Zhou, W. & Wang, K. A signal processing and deep learning framework for methylation detection using Oxford Nanopore sequencing. *Nat. Commun. 2024 151* **15**, 1–21 (2024).

33. Wang, Z. *et al.* Adapting Nanopore Sequencing Basecalling Models for Modification Detection via Incremental Learning and Anomaly Detection. at https://doi.org/10.1101/2023.12.19.572431 (2023).

34. Song, L. *et al.* Robust data storage in DNA by de Bruijn graph-based de novo strand assembly. *Nat. Commun.* **13**, 5361 (2022).

35. Lopez, R. *et al.* DNA assembly for nanopore data storage readout. *Nat. Commun.* **10**, 2933 (2019).

36. GitHub - nanoporetech/scrappie: Scrappie is a technology demonstrator for the Oxford Nanopore Research Algorithms group. https://github.com/nanoporetech/scrappie.

37. Gamaarachchi, H. *et al.* Simulation of nanopore sequencing signal data with tunable parameters. *Genome Res.* gr.278730.123 (2024) doi:10.1101/GR.278730.123.

38. Li, Y. *et al.* DeepSimulator1.5: a more powerful, quicker and lighter simulator for Nanopore sequencing. *Bioinformatics* **36**, 2578–2580 (2020).

39. Stoiber, M. *et al.* De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* 094672 (2017) doi:10.1101/094672.

40. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).

41. MinION portable nanopore sequencing device. https://nanoporetech.com/products/sequence/minion.

42. Kovaka, S. *et al.* Uncalled4 improves nanopore DNA and RNA modification detection via fast and accurate signal alignment. *bioRxiv* 2024.03.05.583511 (2024) doi:10.1101/2024.03.05.583511.

43. New basecaller now performs 'raw basecalling', for improved sequencing accuracy. *Oxford Nanopore Technology* https://nanoporetech.com/news/news-new-basecaller-now-performs-raw-basecalling-improved-sequencing-accuracy (2017).

44. Chris Seymour. Bonito: A PyTorch Basecaller for Oxford Nanopore Reads. at https://github.com/nanoporetech/bonito (2019).

45. Pagès-Gallego, M. & de Ridder, J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. *Genome Biol.* **24**, 1–18 (2023).

46. Xu, L. & Seki, M. Recent advances in the detection of base modifications using the Nanopore sequencer. *J. Hum. Genet.* **65**, 25–33 (2020).

47. Silva, D. F. & Batista, G. E. A. P. A. Speeding up all-pairwise dynamic time warping matrix calculation. in *16th SIAM International Conference on Data Mining 2016, SDM 2016* (2016). doi:10.1137/1.9781611974348.94.

48. Kürzinger, L., Winkelbauer, D., Li, L., Watzel, T. & Rigoll, G. CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition. in *Lecture Notes in*

Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) vol. 12335 LNAI (2020).

49. Zhao, X. *et al.* Composite Hedges Nanopores: A High INDEL Correcting Codec System for Rapid and Portable DNA Data Readout. *bioRxiv* (2024) doi:10.1101/2024.07.12.603190.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (no. 2022YFF1203400), National Natural Science Foundation of China (no. 62171211, 32371526, 32100021 and 32371372) and Science and Technology Innovation Commission of Shenzhen (JCYJ20220814170440001, JCYJ20220818100218039, JCYJ20220530113013030 and JCYJ20230807092459028) as well as NSQKJJ under grant K21799109 and K21799116.

Author contributions

Q.Y.F. and Y.L. designed the experiment and the training pipeline. Q.Y.F., R.H.L. and Y.L. conducted the sequence design of the 6-mer QC sequence, Q.Y.F., R.H.L. Y.P.L. and J.X.Z. conducted the primers for the 6-mer QC sequence. M.L., Q.S.F. and Y.F. cultivated the microorganisms needed for sequencing experiments. Q.Y.F. wrote the code for the k-mer extraction and the architecture of the preliminary basecaller. Q.Y.F., J.Y.L and X.Y.Z. conducted the nanopore basecalling and related quality check. Q.Y.F. conducted the constructing of the datasets and training of the basecaller. Q.Y.F., X.Y.Z. and Y.L. prepared the figures and tables. Q.Y.F., Q.P. and Y.L. drafted the manuscript. Q.P. and Y.L. supervised the study. All authors read, revised, and approved the final manuscript.

Competing interests

Q.F., X.Z., J.L., R.L., and Y.L. have a patent filed with application number CN117238360A. The remaining authors declare no competing interests.

Additional information

Supplementary information. The online version contains supplementary material available at XXX

Correspondence and requests for materials should be addressed to Yi Li.