De Novo Non-Canonical Nanopore Basecalling Enables Private Communication using Heavily-modified DNA Data 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[2], 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.)

Abstract

Hidden messages in DNA molecules by employing chemical modifications has been suggested for private data storage and transmission at high information density. However, rapidly decoding these "molecular keys" with corresponding basecallers remains challenging. We present DeepSME, a nanopore sequencing and deep-learning based framework towards single-molecule encryption, demonstrated by using 5-hydroxymethylcytosine (5hmC) substitution for individual nucleotide recognition rather than sequential interactions. This non-natural, motif-insensitive methylation disrupts ion current, resulting in a readout failure of 67.2%-100%, concealing the privacy within the DNAs. We further develop an alignment-free DeepSME

basecaller as a key to reconstitute the digital information. Our three-stage training pipeline, expands k-mer size from 4^6 to 4^9, achieving over 92% precision and recall from scratch. DeepSME deciphers fully 5hmC concealed text and image within 16× coverage depth with an F1-score of 86.4%, surpassing all the state-of-the-art basecallers. Demonstrated on edge computing devices, DeepSME holds supreme potential for DNA-based private communications and broader bioengineering and medical applications.

Introduction

DNA storage has emerged as a promising solution for the requirement of digital memories in the 'zettabyte era' due to its density and durability[1]. Design and realization of molecular protection[2] places ever-increasing demands on the accurate storage, transmission and assessment of DNA data rapidly and privately. Many efforts have been devoted to exploring an accessible approach for information privacy, such as cryptoristors[3], metahologram[4,5], quantum photonic system[6,7], DNA based technology[8–11], synthetic macromolecules[12,13], where concealed information would not be evident to unsuspecting persons' examination[10]. Among them, biomolecular steganography, which utilizes chemical modifications or interactions instead of computational schemes, has been demonstrated using nucleic acids, proteins[14], aptamers[15], and bacteria[16] for information concealment. For instance, DNA-based steganography developed by Clelland et al.[17] was further leveraged by DNA origami cryptography, which creates a key with a size of over 700 bits with oligonucleotides[11]. However, these strategies exploiting molecular interactions with low information densities and time-consuming processing, have not fully explored the potential of chemical modifications to densely packed nucleobases, thus preventing them from being used in sequencing-based approaches.

Sequencing non-canonical DNA holds promise for private communication because current sequencing technologies and basecallers are designed for canonical DNA. When encountering modified nucleobases, these state-of-the-art sequencing technologies introduce various errors, making it challenging to retrieve the correct sequence[18]. Such methylated[19] or oxidized[20] modifications necessitate specialized methods across different sequencing platforms, such as next generation sequencing (NGS)[21] or long-read ones. For NGS techniques, WGBS[22,23], RRBS[24], MeDIP-seq[25], MRE-seq[26] etc. have been developed for single specific types, where the non-canonical bases may not react completely or break the DNA strands into pieces[27]. These methods can also be applied onto the long-read fluorescent sequencing (PacBio), where

62 the drawbacks are the same. Meanwhile, for long-read nanopore sequencing, non-canonical
63 bases can be identified without any changes. The detection algorithms can be categorized into
64 two classes: 1) Identification from alphabet errors. This includes Epinano for m6A[28], differ for
65 m6A[29], DRUMMER for m6A[30], nanoRMS for Ψ U/Nm[31], ELIGOS for m6A[32], NanoCEM for
66 unspecific modifications[33] and Dinopore for Inosine[34]; 2) Identification from current features,
67 containing DeepSignal for 5mC[35], DeepMod for 5mC[36], MINES for m6A[37], nanoDoc for
68 unspecific modifications[38], nanom6A for m6A[39], Nanocompore for m6A[40], Yanocomp for
69 m6A[41], xPore for m6A[42], Penguin for Ψ [43] m6Anet for m6A[44] and exos for 8-oxo-dG[20]; Rerio
70 for 5mC/5hmC and IL-AD for 5mC/5hmC/6mA[45]. Aiming for biological science, these
71 developments of basecallers limit themselves within the identification of light-weighted
72 modification[46] or motifs[47]. None of them - to the best of our knowledge – has been reported to
73 deal with heavily or nearly complete modifications, which remains challenging because of the
74 lack of basecallers and references.

75 To address this limitation, a *de novo* approach to basecalling heavily modified DNA is required
76 for pristine raw data. Nanopore data readouts, maintaining sequential information of any
77 nucleobase or analogues into ion flow blockades, nowadays utilize neural networks (NN)[48]
78 vastly for accurate basecalling. Well-trained NNs can act as the cipher to pair with a specific
79 molecular key, the inherent non-reversibility of neural networks prevents direct determination
80 of the key from the network's weights. However, training NNs is often a joy with tears, since
81 rounds of sophisticated alignments and time-consuming corrections of datasets may take
82 place[49]. Besides, existing prior knowledge (k-mer dictionary) and references are critical to
83 reduce the high error rates[50]. The bottlenecks for nanopore private communication have two
84 aspects: 1) to arrange private messages wisely into non-canonical DNAs and 2) to rapidly and
85 correctly build up correlated nanopore NN with limited correct alignments. To date,
86 constructing a framework for heavily or completely modified DNAs, is therefore highly
87 appreciated.

88 In this work, we propose a private framework utilizing motif-insensitive 5hmC modification
89 combined with a nanopore sequencing-based *de novo* basecaller for privacy. We conceal text
90 and image into fully 5hmC modified DNA sequences and validate our approach by nanopore
91 sequencing and basecaller training, aiming to address the challenges to build the basecaller
92 posed to heavily modified bases. We develop a three-stage alignment-free training framework
93 for Deep-learning framework towards Single-Molecule Encryption (DeepSME), which tackles
94 the basecalling bottleneck of the heavily modified dataset by expanding k-mer dictionary.

Independent k-mer tables are generated from scratch, allowing us to process the modified sequences and their corresponded signal disruptions without prior references at single-molecule level. This framework, together with the training method of private DeepSME basecaller underpins the potential for concealed DNA-based data storage and communication with high information density, addressing the increasing demand for robust information privacy in an era of evolving biotechnological threats.

Results

Fig. 1a illustrates the scenario for digital information stored in DNA for private 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)[51] or nanopore sequencing (NPS)[52] for public communications. Here, Alice acts as the sender who initiates the communication either publicly or privately. For the former case, canonical bases and conventional basecallers are publicly available. Here we proposed a Deep-learning framework towards Single-Molecule Encryptions (DeepSME) for the latter case between Alice and Bob. This allows Alice to hide information within non-canonical DNAs and creates a basecaller that serves as the key for Bob. Moreover, unsuspecting person Eve is not capable of accessing the information using conventional basecallers.

Following the roles for Alice, Bob and Eve, Fig. 1b displays an example of chemical structures of Alice's choice - 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 on the blockage of nanopore signals could be expected with close-to-neutral hydrophilicity. Following synthetic methods[36], all the cytosines are substituted *in vitro* by the chemically stable 5hmCs regardless of any motifs or contexts. This complete replacement of 5hmC surpasses the medication ratio in any biological systems, in particular human genome[23,53,54], where only 1.28% of C (2.62‰ of a given whole microbial genome) is currently modified to 5mC or 5hmC.

We next examined the privacy of full 5hmC modification using nanopore sequencing in Fig. 1c. As DNA molecules pass through the nanopore, they generate characteristic ionic current signals that are recorded as squiggle patterns (shown as black lines in the figure). For canonical

DNA (top panel), the current signal (shown in cyan) is correctly interpreted by conventional basecallers to reveal the Alice's reference sequence (GCTGACCCGCCGCATCGGTG). However, current signals in the bottom panel drastically deviated (shown in orange) when containing full 5hmC modifications. These heavily modified signals confuse conventional basecallers, for example Bonito/Dorado, resulting in severe misinterpretation of the message (TGCTTGCCGTTCGCGTTATA, with 16 out of 20 bases incorrectly called), demonstrating the successful privacy protection away from Eve in this scenario.

For the corresponding key produced by either Alice or Bob, we propose a three-stage alignment-free pipeline for DeepSME basecaller, shown in Fig. 2a. Eleven samples of DNA sequences ranging from 1145 nt to 1341 nt are firstly sequenced to construct the 6-mer quality check dataset without any sequence-to-current alignments[50]. Tolerating these errors caused by the alignment-free segmentation, an initial 6-mer QC 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[55], squigulator[56], and deepsimulator[57], *in silico* dataset can be generated for length variation and expanded the dictionary from 4^6 to 4^9. Not only mitigating the overall 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 5). Last, three types of microbial gDNA are sequenced by nanopore and processed 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 and segmentations for every 10,000 heavily modified nucleobases manually with no prior knowledge. Taking conventional segmentation tool Tombo[50] (Nanopolish[58] also works) on canonical dataset as a control, 3.9% errors will always exist. Considering billions of sequences and bases that are being used, Fig. 2c shows the overall error rate as a function of length variation. Given approximately 450 nt/s as the sequencing speed of a nanopore[59], 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)

159 contribute to the length variations of 40 nt to 85 nt. The longer the segmentation goes, the more
160 errors the preliminary DeepSME holds.

161 The inset of Fig. 2c illustrates the segmentation-induced length variation. A fixed number of
162 bases can be assigned to segmented current blocks each time as a chunk, creating a dataset
163 containing tens of thousands of chunks for the training of the basecaller. The length variation
164 itself is proportional to the segmented time. Our preliminary DeepSME segmented chunks at
165 2.22 s is estimated to have a total error of 11.6%, which seems tolerable for further
166 investigations.

167 We further analyzed the reduced ionic currents with the modification of 5hmC. Fig. 2d shows
168 the normalized current difference for the canonical C and 5hmC in the extracted k-mer model
169 from DeepSME in the second training stage. By substituting either C or 5hmC at each position
170 (x-axis), the mean averaged current difference (MACD) is calculated over all k-mers in the
171 model. In this 6-mer model, the base at index 2 has the highest impact on the current change.
172 This is in a reasonable agreement with reported k-mer model from ONT[60]. Additionally, the
173 weight of the bases at index #0, #1, #2, and #3 on the current significantly decreases when
174 5hmC modification takes place. This indicates that 5hmC modification leads to a reduction in
175 the normalized currents at these corresponding base positions compared to the canonical state.

176 On the other hand, Fig. 2e presents a comparative analysis of normalized ionic current levels
177 between canonical and 5hmC-modified DNA sequences. Each point in the figure represents a
178 specific k-mer (6-mer), where those not containing any C's shown in black, and those
179 containing 1 to 6 C's shown in orange. The intensity of the orange color correlates with the
180 number of cytosines in the k-mer - darker orange indicates more cytosines. The black dashed
181 line with a slope of 0.83 with the offset of 0.58 represents a linear fitting for 6-mers without
182 C's. Notably, this line deviates from the gray diagonal one, aligning with the k-mer for 5hmC
183 modified DNA reported by Kovaka et al[60]. This significant deviation in slope indicates
184 substantial changes in the normalized current signals, making the non-natural heavily
185 modification more effective for private communications. Furthermore, k-mers containing
186 5hmC modifications consistently show lower normalized current levels compared to their
187 canonical counterparts, which matches well with our previous observations in Fig. 2d.

188 The basecalling accuracy and recall are evaluated at the third stage. Fig. 2f illustrates the
189 relationship between recall and precision of DeepSME at three stages for both canonical C and
190 5hmC. First, the preliminary basecallers are located at the bottom left, meaning low precision

and low recall capabilities since only 6-mer features of DNA 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 lacks the ability of generalization for variant conditions. Last, the reinforced basecallers show significant improvements in 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 canonical one. Compared with the commercial basecallers (Bonito FAST, HAC - high accuracy and SUP - super accuracy config), the reinforced 5hmC basecaller is beyond the F1-score of 0.85 (grey dashed line), surpassing Bonito SUP, HAC as well as FAST.

We turn our focus onto the decoding results of 5hmC modified DNA that conceals text file (sustech_introduction.txt, 978 bytes, 55 strands) and image file (sustech_logo.jpg, 7,775 bytes, 432 strands). Without knowing exact types of DNA modifications, Fig. 3a shows the decoding performance of the identical text file using state-of-the-art basecallers on modified DNA from the unsuspecting observer Eve's view with 16× coverage. For the basecallers that could not reach the coverage amount, all basecalled sequences were obliged to use. Guppy 6.0 achieved a decode rate of 5.45% (3/55), Bonito/Dorado reached 3.64% (2/55), Rerio 5mC 5hmC managed 3.64% (2/55), while IL-AD was unable to decode any sequences (0/55) at all.

Encouraged by the high precision and recall of DeepSME, we perform decoding experiments displayed in Fig. 3b from the private observer Bob's view. The recovered texts are shown with the increasing physical redundancy (the coverage depth from 2× to 16×). It is evident that DeepSME produced up to 52.73% (29/55) of the text after decoding with 4×, reached up to 94.5% (52/55) data retrieval with 8× coverage and correctly decoded all of the text with 16× coverage. Fig. 3c compares the text data retrieval performance of the two DeepSMEs as a function of coverage depth. They exhibit almost identical data recover rate, with small variations across experiments. At each coverage level, the data recover rate was measured 10 times with different random seeds.

To gain more insights from the two basecallers, the substitution probabilities of canonical 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[52].

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

Discussion

Our work demonstrates that the challenge of pairing key generation with nanopore basecallers can be coped with the DeepSME framework. This could enable private communication between Alice and Bob with a superior F1-score of 86.42%, shown in Fig. 3f Our DeepSME framework offers several advantages:

1) *Knowledge growing*. First, the initial 6-mer QC DeepSME basecaller does extract 5hmC-modified features. This is supported by the shift of k-mer current in Figure 2e, where 89.63% of C-containing k-mers deviate from the diagonal and only 12.35% of k-mers that don't contain C deviate from the diagonal. Second, the Enhanced DeepSME basecaller could learn more abundant features from 6-mers to 9-mers, which can be ascribed to the substantial precision increase on the modified test dataset from 0% to 82.87% (see Figure 2f and Supplementary Table 7). Third, the reinforced DeepSME basecaller could get polished smoothly along with experimental 9-mer gDNA dataset. This can be accredited by the fact of achieving 92.99% precision and 92.93% recall.

2) *Alignment free.* We used the strategy of fixing chunk size of 1000 bp instead, shown in Fig. 2c. Based on the assumption of constant sequencing velocity at 450 bp, this strategy holds an error rate as high as 11.6%. Compared with 5.63% error as the minimal, this high error rate is no more a trouble as the overall basecalling errors went down to 7.01% in Fig. 2f.

3) *Prior-knowledge free*. Our training of fully modified dataset can be performed without pre-existing k-mer tables nor basecaller weights. This is underpinned by a smart use of the output from the Connectionist Temporal Classification (CTC) layer of our 6-mer QC DeepSME basecaller. Once this basecaller gets trained, one can directly derive a new k-mer table from the CTC layer, which is particularly beneficial to link with the trained weights. To this end, prior k-mer tables and basecaller weights can be detoured.

4) *Affordable*. We show in the Methods that 266.4 minutes of thee-stage basecaller training in 19.9 GB VRAM are sufficient for our open-sourced deep-learning neural network. While lightweight models can reduce computational demands, they led to much lower F1-scores (75.77%-84.36%). Different programming language-based software (Python-based Bonito and

255 C-based Dorado) and hardware platform (NVIDIA RTX 3090 and Jetson Xavier AGX) were
256 thoroughly evaluated (see Supplementary Table 11 and 12), confirming the operational
257 feasibility of the current model size in edge-computing environments.

258 For the generalizability, DeepSME basecallers could be custom tailored and delivered for
259 variant combinations on promise, including diverse modification types, modification ratios as
260 well as nanopore sensors. For instance, our preliminary tests demonstrate successful alignment
261 of the sequence primer current extracted from the 6-mer QC sequence (using an R9.4.1 pore)
262 to its corresponding current in an R10.4.1 pore via Dynamic Time Warping[61] (DTW)
263 (Supplementary Information Figure 6). Although high error rates have been mitigated by
264 Composite Hedges Nanopore (CHN) code, the advances on R10.4.1 flowcells could further
265 manage the error rate below 1%[62]. This cross-pore compatibility suggests the adaptability of
266 the DeepSME framework across different types of nanopores or other hardware such as
267 CycloneSEQ[63] or PolyseqOne[64].

268 Our three-stage training pipeline is essential, as removing any stage would compromise the
269 overall performance and reliability of the framework. First, the initial 6-mer QC DeepSME
270 basecaller lays the foundation of generating a unique 6-mer table for subsequent steps. Without
271 this initial step, fewer than 0.2% of the chunks can be identified for the rest steps, making
272 basecaller training nearly impossible. Second, the *in silico* training of the Enhanced DeepSME
273 basecaller boosts the accuracy to over 80% while expanding the k-mer coverage from 6-mers
274 to 9-mers. Without 81,490 simulated strands, the 6-mer QC DeepSME basecaller would
275 recognize less than 0.5% of chunks of gDNA dataset, hindering basecalling. However, the
276 Enhanced DeepSME basecaller has the recall rate remaining at 2.5%, showing its overfitting
277 of *in silico* data. Last, the reinforced DeepSME basecaller incorporating real-world gDNA data,
278 reached 93.0% precision on DNA storage dataset (Fig. 3). This demonstrates the successful
279 mitigation of overfitting risks.

280 For the unsuspecting observer Eve, it is hardly possible to circumvent DeepSME with public
281 basecallers and their weights. Alternative basecallers have fewer than 5% of the correct
282 sequences for modified sequences in Fig. 3f (see the Supplementary Table 7-11 for recall and
283 precision of these basecallers). We believe this could be attributed to the shifted current
284 deviations (Fig. 1c and Fig. 2e): some of the points are still located close to the black dashed
285 line, which could be recognized by basecaller and aligned to the reference sequence.
286 Meanwhile, other points are far away from the black dashed line, being ill-interpreted.

Beyond enumerating basecalling attempts, the interception of DeepSME could be executed by Eve as the weights lack inherent encryption. Thus, employing traditional[65] or post-quantum cryptographic algorithms[66] becomes essential and be a critical direction to avoid cybersecurity threats between Alice and Bob. Furthermore, reverse engineering of modification patterns represents another potential vulnerability. However, the diversity of existing modifications[19,67,68] and the availability datasets for heavily modified DNA, may hinder the timely training of such reverse-engineering models, providing DeepSME with a valuable, albeit potentially temporary, window for privacy, though not absolute security.

Lastly, privacy may be inferred, assuming an ultimate modification caller that can accurately decode any level of modified data. While theoretically plausible, achieving this in practice presents considerable hurdles. One one hand, the complexity of the "molecular key space" increases with 429 existing types[19,67,68] and upcoming new ones. On the other hand, the parameters and memory consumption will inflate drastically for training an 100-letter basecaller (may need $5.6 + 14.3 \times \frac{100^6}{4^6} = 3.49 \times 10^9$ GB of VRAM). These practical issues offers DeepSME a substantial layer of privacy against current threats and resource-constrained adversaries.

We speculate that DeepSME could evolve into a functional encryption system. The original DNA sequence could stand for the plaintext, the nanopore current readouts from modified DNA for the ciphertext, the specific modification types and their ratios for the key space, while *in vitro* modification process and DeepSME weights governed basecalling for encryption and decryption, respectively. While sharing similarities as a quintuple scheme for the definition of encryption[69], we are at a very infant stage towards encryption. Given the key length is defined as $\log_2(C_N^M)$, where M is the known nucleobase analogues and N is the chosen nucleobase types, our implementation with a single modification type achieves a key length of approximately $\log_2(\mathrm{C}_4^{429})$ = 30.37 bits, supposing the existence of at least 429 known nucleobase modifications[19,67,68]. This is far behind National Institute of Standards and Technology's (NIST) recommended minimum of 128 bits for secure encryption[65]. In this theory, 24 types of nucleobase analogues used in one *in vitro* process could fulfil the key length requirement of $\log_2(C_{24}^{429})$ = 129.89 bits. Alternatively, block encryption offers a promising approach to increase key length. By splitting the data into blocks and employing different combinations of nucleobase analogues for each block, we can linearly increase the overall key length. For instance, using four blocks and focusing on a subset of 140 common base analogues

mentioned by Juan Et. al.[19], we can select four distinct sets of six analogues to encrypt each data block. This approach could theoretically achieve a key length of $4 \times \log_2(C_6^{140}) = 132.51$ bits, thereby meeting NIST's requirements. These methods, while beyond the scope of this work, highlight the potential of DeepSME for future development as a robust encryption system.

We also verified the integration of DeepSME with post-quantum cryptography (PQC) such as frameworks like FrodoKEM[71]. This could also be applied onto the raw texts and images. For tamper-evident DNA storage protocols; we foresee that techniques such as molecular tags, sequence watermarks, and enzymatic protection mechanisms could be explored[72]. Some simplied methods, like monitoring 5hmC/canonical cytosine ratios and basecaller discrepancies which reuse the developed basecallers, may also provide effective tamper-evidence. We further note that 5hmC DeepSME basecaller also performs well on canonical DNA sequences (see Supplementary Table 11), which shows its promise to be utilized as a framework to detect methylations from epigenomics.

For ethical considerations, we have implemented multi-layered biosecurity screening measures and advocate for adherence to guidelines like the International Gene Synthesis Consortium (IGSC) standards and ISO regulations[73]. Concerns regarding illicit data smuggling via microbes are taken into our thoughts. Although they could be mitigated by the likely toxicity of extensive 5hmC modifications[74], proactive biosecurity, ethical guidelines, and dual-use risk assessments remain essential for responsible DeepSME development, as we have solely focused on *in vitro* investigations[75].

In conclusion, we have proposed and experimentally validated a communication framework (DeepSME) with non-naturally existing hydromethylation DNAs for information privacy, paired with a *de novo* trained nanopore basecaller. Avoiding tremendous alignments and corrections, we demonstrate that our three-stage training method is essential for DeepSME to achieve both 92% recall and precision incrementally, exceeding state-of-the-art basecallers on the digital information stored in 5hmC DNA dataset. Within just 16× coverage depth and negligible prior knowledge requirements, our independent DeepSME framework fulfills the readout of heavily-modified single-molecule DNA, showcasing its great promise towards biomolecular-based private communications, unclonable functions and anticounterfeiting systems.

Statistics & Reproducibility

351 Statistics used to analyze DNA storage decoding performance are described in the methods
352 section. Analysis can be reproduced using datasets deposited in the NCBI SRA and Zenodo
353 (see "Data availability" statement) and the code developed for this work (see "Code
354 availability" statement). All sequencing data from 6-mer QC sequences and Bacterial Genome
355 DNA (gDNA) were used to train DeepSME. All sequencing data from DNA storage strands
356 were used to perform the DNA decoding experiments. The decoding rates under different
357 coverages shown in Fig. 3c were determined by repeating the analysis 10 times using random
358 seeds 0-9 when selecting reads from each sequence group.

359 **Methods**

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

366 For each 50 µl reaction system, 15 µl of ddH2O was added along with 25 µl of 2× Phanta Max
367 Buffer. Additionally, 1 µl of each 10 mM dATP, dTTP, dCTP, and dGTP was included. To this
368 mixture, 2 µl of Forward Primer and 2 µl of Reverse Primer were added. Finally, 1 µl of Phanta
369 Max Super-Fidelity DNA Polymerase (1 U/µl) and 1 µl of DNA were incorporated to complete
370 the reaction mixture. Specifically, 5-Hydroxymethyl-dCTP (Cat. No. NU-932s) was purchased
371 from Jena Bioscience as the alternative PCR substrate for dCTP. The PCR conditions were as
372 follows: 1. Pre-denaturation at 95°C for 3 min, 2. Denaturation at 95°C for 15 seconds, 3.
373 Annealing at 60°C for 15 seconds, 4. Extension at 72°C for 1.2 minutes or 1 minute per 1kb
374 sequence, 5. Final Extension at 72°C for 5 minutes. After performing 30 cycles, the resulting
375 products can be considered as a complete conversion from cytosine (C) to methylated 5hmC.
376 To confirm the fidelity of amplification with DNA modification, the 5hmC PCR products were
377 sent to Sanger sequencing showing identical sequences (Supplementary Figure 16).

378 **Electrical Readout via Nanopore Sequencing.** Following PCR amplification, both the
379 canonical (ATCG) and 5hmC-modified ($AT^{5hm}CG$) DNA products were separately subjected
380 to electrophoresis on a 1.5% agarose gel for length validation and purification. The bands
381 corresponding to the reference sequence length were excised, followed by DNA extraction
382 using the E.Z.N.A. Cycle Pure Kit (V-spin) from Omega Bio-Tek. A solution containing 200

383 ng of DNA was prepared using the ND608 kit from Vazyme for DNA damage repair, end
384 preparation and adapter ligation. This includes two steps: 1) Damage repair and end preparation:
385 2 µl Damage Prep Enzyme, 5 µl End Prep Enzyme, and 10 µl End Prep Buffer were added,
386 with ddH_2O supplemented to a total volume of 65 µl. This mixture was incubated in a PCR
387 thermal cycler at 30°C for 20 minutes, followed by 65°C for 15 minutes. 2) Adapter ligation:
388 25 µl Rapid Ligation Buffer 2, 5 µl Rapid DNA Ligase, and 5 µl AMX-F from the LSK-110
389 kit from Oxford Nanopore Technology were added to the reaction solution, and the mixture
390 was incubated in a PCR thermal cycler at 20°C for 15 minutes.

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

407 Furthermore, DNA quantification was performed using a Qubit Fluorometer to measure 50
408 fmol of DNA. For flow cell priming and loading according to the LSK-110 protocol, 30 µl of
409 thawed and mixed Flush Tether (FLT) was added directly to the tube of thawed and mixed
410 Flush Buffer (FB) and mixed by vortexing at room temperature. Then, 800 µl of the priming
411 mix was loaded into the flow cell (FLO-MIN106) via the priming port, avoiding the
412 introduction of air bubbles, and left to sit for 5 minutes. During this time, the library for loading
413 was prepared by mixing 37.5 µl Sequencing Buffer II (SBII), 25.5 µl Loading Beads II (LBII),
414 and 12 µl DNA library including 50 fmol of DNA. An additional 200 µl of the priming mix
415 was loaded into the flow cell via the priming port, again avoiding air bubbles. The prepared

416 library was then mixed gently by pipetting up and down just before loading. Finally, 75 µl of
417 the sample was added to the flow cell via the SpotON sample port in a dropwise fashion,
418 ensuring each drop flowed into the port before adding the next.

419 Finally, the MinION was connected to a computer, and the MinKNOW v24.02.6 software was
420 launched. The LSK-110 protocol and other default settings were selected. We also adjust the
421 "sequence length cutoff options" on MinKNOW (200 bp for DNA Storage experiment, 1000
422 bp for 6-mer QC DeepSME Basecaller experiment). Finally, the sequencing button was clicked
423 to begin collecting nanopore sequencing current data.

424 **Constructing of DeepSME Pipeline.** *Design and Synthesis of 6-mer QC sequence.* The
425 Cate_NAN plasmid (see Supplementary Table S17) of total length 5262 nt was used to
426 constructing ten sequences of our 6-mer quality control (QC) dataset. Ten pairs of forward and
427 reverse primers, ranging from 40 nt to 50 nt (see Supplementary Table S18), were designed
428 and synthesized by Tsingke Biotech to generate corresponding ten samples of DNA sequences
429 with length of 1330 nt, 1340 nt, 1340 nt, 1340 nt, 1340 nt, 1200 nt, 1200 nt, 1230 nt, 1226 nt
430 and 1262 nt containing approximately 20 nt to 30 nt without cytosine (C) in their forward and
431 reverse primers to serve as barcode of the sequence (see Supplementary Table S19). An
432 additional sequence of 1144 nt was derived from a synthesized 1206 nt DNA template. These
433 eleven key sequences contribute to 99.93% physical redundancy of 6-mers (4093/4096) with a
434 median occurrence of three. The non-C segments on the eleven sequences were further used to
435 facilitate subsequent data classification.

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

442 *Preliminary Training of DeepSME with 6-mer datasets.* The sequenced raw data were
443 processed using Guppy 6.0 basecaller (`dna_r9.4.1_450bps_hac.cfg`)
444 (https://nanoporetech.com/document/Guppy-protocol) as a control. The barcode region of
445 basecalled fastq sequences were aligned back to the ionic currents using Tombo[50] 1.5.1
446 (https://github.com/nanoporetech/tombo) to provide primer cutoff point. The current data were
447 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` based on `basecaller_benchmark` repository[48], see supplementary information Table 14). 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 in 151.8 minutes with an RTX 3090 GPU.

Extracting 6-mer statistics (k-mer model). Ionic current data were segmented by the probability matrix obtained from the CTC segmentation[76] of the trained 6-mer QC DeepSME, then enhanced by executing three times of k-mer extraction and alignment by Tombo to get the k-mer table (6-mer). Sequenced average currents, standard deviations as well as averaged dwell times and related standard deviations can be analyzed for either canonical 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[56] v0.3.0 (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 (https://github.com/nanoporetech/bonito). These datasets were used to train 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 Table. 15) over 5 epochs with a learning rate of 5e-4, resulting in the Enhanced DeepSME in 64.6 minutes with an RTX 3090 GPU.

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/),

480 Aeromonas_hydrophila_BJ054
481 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_046708825.1/), and Vibrio_cholera_E1
482 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_026013235.1/) microorganisms were
483 extracted using the Qiagen Miniprep kit. The DNA was fragmented using the TD502 kit
484 (TruePrep DNA Library Prep Kit V2 for Illumina) from Vazyme and modified using 5-
485 Hydroxymethyl-dCTP (NU-932s) from Jena Bioscience. PCR was performed with Phanta Max
486 Super-Fidelity DNA Polymerase (P505-d1) from Vazyme over 18 cycles under the following
487 conditions of six steps: 1) Extension at 72°C for 3 minutes. 2) Pre-denaturation at 95°C for 3
488 min, 3) Denaturation at 95°C for 15 min, 4) Annealing at 60°C for 15 seconds, 5) Extension at
489 72°C for 60 s-72 s per kilobase sequence, 6) Final Extension at 72°C for 5 minutes. The
490 modified DNA samples were sequenced using the Oxford Nanopore Technology R9.4.1
491 flowcell and LSK-110 reagents.

492 *Reinforced DeepSME with Fully5-hmC Bacteria gDNA Datasets.* The Enhanced DeepSME
493 was used to basecall gDNA and map to the corresponding genome. The gDNA datasets from
494 the three microbial samples also lead to 118,096 chunks with a chunk size of 3,600 (0.9 s).
495 These data were then used to reinforce the Encoder-Decoder architecture with CRF decoding
496 layer (`dna_r9.4.1_e8_sup@v3.3` model in Bonito 0.8.1) over 5 epochs with a learning
497 rate of 5e-4, yielding our final reinforced DeepSME, which takes 50.0 minutes with an RTX
498 3090 GPU. The weights of our basecallers could also be compatible with Dorado and be ready
499 for deployed to edge devices like Jetson Xavier AGX for further use. We have also tried using
500 FrodoKEM framework to encrypt our 5hmC-DeepSME basecaller weight with FrodoKEM-
501 640-AES algorithm (see https://github.com/sparkcyf/FrodoKEM_demo)to demo the
502 feasibility for integrating DeepSME with post-quantum cryptography.

503 **Recovering Data Stored in 5-hmC DNA using DeepSME.** Dataset for digital information
504 stored in DNA were prepared using our Composite Hedges Nanopore (CHN) codec[77] in the
505 four-letter configuration. The binary forms of two selected files (219 bytes and 4109 bytes)
506 were extracted and split into payloads with 36 bytes. Next, Reed-Solomon (40,36) code was
507 used to add redundancy to these payloads to generate segments. Then, a 34 nt barcode as
508 addresses for random access followed with three 5-mer anchors are added to each segment.
509 Finally, an oligo pool containing 487 single-stranded 243 nt DNA sequences was sent to Twist
510 Biosciences, including 55 sequences that encode a .txt file and 432 sequence that encode a .jpeg
511 image (see Supplementary Table 20 and Supplementary Table 21).

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

516 The reinforced 5hmC DeepSME was used to process the sequenced raw data. The resulting
517 fastq and aligned bam files were performed using the aboved-mentioned CHN decoder. For
518 performance metrics, "the rate of the number of reads that could be aligned to the reference
519 FASTA in total reads" was used as Recall, "the rate of the number of correct bases in total
520 bases" was used as Precision, and their product was used as the F1-score.

521

522 **Data availability**

523 The sequencing pod5 files from 5hmC-modified 6-mer QC sequences, Bacterial Genome
524 DNA (gDNA) for training the 5hmC DeepSME basecaller and conducting the DNA storage
525 decoding experiments have been deposited to Zenodo
526 (https://doi.org/10.5281/zenodo.12704171). The FASTQ file generated by the 5hmC
527 reinforced DeepSME from Bacterial Genome DNA (gDNA) has been deposited to NCBI
528 Sequence Read Archive (SRA) accession number SRR32782578
529 (https://www.ncbi.nlm.nih.gov/sra/SRX28066745), SRR32782579
530 (https://www.ncbi.nlm.nih.gov/sra/SRX28066744), SRR32782580
531 (https://www.ncbi.nlm.nih.gov/sra/SRX28066743) under NCBI BioProject PRJNA1238011
532 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1238011). The FASTQ file generated by
533 the reinforced 5hmC DeepSME from the 5hmC DNA storage dataset has been deposited to
534 NCBI Sequence Read Archive (SRA) accession number SRR32782578
535 (https://www.ncbi.nlm.nih.gov/sra/SRX28067610) under NCBI BioProject PRJNA1238011
536 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1238011). All other data described in this
537 work are available in the main text, provided in the supplementary materials, or can be
538 reproduced using the deposited datasets and GitHub code (see "Code availavility" section).
539 Source data are provided with this paper.

540 **Code availability**

541 The code package for this study is available in the GitHub repository (https://github.com/spar
542 kcyf/DeepSME) and is also available on Zenodo under https://doi.org/10.5281/zenodo.15064
543 695 .

544

545 **References**

546 1. Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information
547 storage in synthesized DNA. *Nature* **494**, 77–80 (2013).

548 2. Handbook of Information and Communication Security. *Handb. Inf. Commun. Secur.*
549 (2010) doi:10.1007/978-3-642-04117-4.

550 3. Kim, S. Il *et al.* Cryptographic transistor for true random number generator with low
551 power consumption. *Sci. Adv.* **10**, eadk6042 (2024).

552 4. Qu, G. *et al.* Reprogrammable meta-hologram for optical encryption. *Nat. Commun.*

11, 5484 (2020).

5. 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).

6. Paraïso, T. K. *et al.* A photonic integrated quantum secure communication system. *Nat. Photonics* **15**, 850–856 (2021).

7. 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).

8. Dickinson, G. D. *et al.* An alternative approach to nucleic acid memory. *Nat. Commun.* **12**, 2371 (2021).

9. Kuzdraliński, A. *et al.* Unlocking the potential of DNA-based tagging: current market solutions and expanding horizons. *Nat. Commun.* **14**, 6052 (2023).

10. Meiser, L. C. *et al.* Synthetic DNA applications in information technology. *Nat. Commun.* **13**, 352 (2022).

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

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

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

14. 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).

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

16. 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).

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

583 18. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing
584 and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).

585 19. Alfonzo, J. D. *et al.* A call for direct sequencing of full-length RNAs to identify all
586 modifications. *Nat. Genet.* **53**, 1113–1116 (2021).

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

589 21. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of
590 next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

591 22. Irizarry, R. A. *et al.* Comprehensive high-throughput arrays for relative methylation
592 (CHARM). *Genome Res.* **18**, 780–790 (2008).

593 23. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**,
594 355–364 (2023).

595 24. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-
596 resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).

597 25. Taiwo, O. *et al.* Methylome analysis using MeDIP-seq with low DNA concentrations.
598 *Nat. Protoc.* **7**, 617–636 (2012).

599 26. Li, D., Zhang, B., Xing, X. & Wang, T. Combining MeDIP-seq and MRE-seq to
600 investigate genome-wide CpG methylation. *Methods* **72**, 29–40 (2015).

601 27. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation
602 sequencing methods for chromatin biology. *Nat. Rev. Genet.* **15**, 709–721 (2014).

603 28. Liu, H., Begik, O. & Novoa, E. M. EpiNano: Detection of m6A RNA Modifications
604 Using Oxford Nanopore Direct RNA Sequencing. in *Methods in Molecular Biology*
605 vol. 2298 31–52 (2021).

606 29. Parker, M. T. *et al.* Nanopore direct RNA sequencing maps the complexity of
607 arabidopsis mRNA processing and m6A modification. *Elife* **9**, e49658 (2020).

608 30. Price, A. M. *et al.* Direct RNA sequencing reveals m6A modifications on adenovirus
609 RNA are necessary for efficient splicing. *Nat. Commun.* **11**, 6016 (2020).

610 31. Begik, O. *et al.* Quantitative profiling of pseudouridylation dynamics in native RNAs
611 with nanopore sequencing. *Nat. Biotechnol.* **39**, 1278–1291 (2021).

612 32. Jenjaroenpun, P. *et al.* Decoding the epitranscriptional landscape from native RNA
613 sequences. *Nucleic Acids Res.* **49**, e7 (2021).

614 33. Guo, Z. *et al.* Nanopore Current Events Magnifier (nanoCEM): a novel tool for
615 visualizing current events at modification sites of nanopore sequencing. *NAR*
616 *Genomics Bioinforma.* **6**, lqae052 (2024).

617 34. Nguyen, T. A. *et al.* Direct identification of A-to-I editing sites with nanopore native
618 RNA sequencing. *Nat. Methods* **19**, 833–844 (2022).

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

621 36. Ahsan, M. U., Gouru, A., Chan, J., Zhou, W. & Wang, K. A signal processing and
622 deep learning framework for methylation detection using Oxford Nanopore
623 sequencing. *Nat. Commun.* **15**, 1448 (2024).

624 37. Lorenz, D. A., Sathe, S., Einstein, J. M. & Yeo, G. W. Direct RNA sequencing enables
625 m6A detection in endogenous transcript isoforms at base-specific resolution. *RNA* **26**,
626 19–28 (2020).

627 38. Ueda, H. nanoDoc: RNA modification detection using Nanopore raw reads with Deep
628 One-Class Classification. *bioRxiv* 2020.09.13.295089 (2021)
629 doi:10.1101/2020.09.13.295089.

630 39. Gao, Y. *et al.* Quantitative profiling of N 6-methyladenosine at single-base resolution
631 in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA
632 sequencing. *Genome Biol.* **22**, 22 (2021).

633 40. Leger, A. *et al.* RNA modifications detection by comparative Nanopore direct RNA
634 sequencing. *Nat. Commun.* **12**, 7198 (2021).

635 41. Parker, M. T., Barton, G. J. & Simpson, G. G. Yanocomp: robust prediction of m6A
636 modifications in individual nanopore direct RNA reads. *bioRxiv* 2021.06.15.448494
637 (2021).

638 42. Pratanwanich, P. N. *et al.* Identification of differential RNA modifications from
639 nanopore direct RNA sequencing with xPore. *Nat. Biotechnol.* **39**, 1394–1402 (2021).

640 43. Hassan, D., Acevedo, D., Daulatabad, S. V., Mir, Q. & Janga, S. C. Penguin: A tool
641 for predicting pseudouridine sites in direct RNA nanopore sequencing data. *Methods*

642 **203**, 478–487 (2022).

643 44. Hendra, C. *et al.* Detection of m6A from direct RNA sequencing using a multiple
644 instance learning framework. *Nat. Methods* **19**, 1590–1598 (2022).

645 45. Wang, Z. *et al.* Adapting nanopore sequencing basecalling models for modification
646 detection via incremental learning and anomaly detection. *Nat. Commun.* **15**, 7148
647 (2024).

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

650 47. Yuen, Z. W. S. *et al.* Systematic benchmarking of tools for CpG methylation detection
651 from nanopore sequencing. *Nat. Commun.* **12**, 3438 (2021).

652 48. Pagès-Gallego, M. & de Ridder, J. Comprehensive benchmark and architectural
653 analysis of deep learning models for nanopore sequencing basecalling. *Genome Biol.*
654 **24**, 71 (2023).

655 49. Teng, H. *et al.* Chiron: translating nanopore raw signal directly into nucleotide
656 sequence using deep learning. *Gigascience* **7**, giy037–9 (2018).

657 50. Stoiber, M. H. *et al.* De novo Identification of DNA Modifications Enabled by
658 Genome-Guided Nanopore Signal Processing. *bioRxiv* 094672 (2016)
659 doi:10.1101/094672.

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

662 52. Lopez, R. *et al.* DNA assembly for nanopore data storage readout. *Nat. Commun.* **10**,
663 2933 (2019).

664 53. Hood, L. & Rowen, L. The human genome project: Big science transforms biology
665 and medicine. *Genome Med.* **5**, 79 (2013).

666 54. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids*
667 *Res.* **53**, D1243–D1249 (2025).

668 55. GitHub - nanoporetech/scrappie: Scrappie is a technology demonstrator for the Oxford
669 Nanopore Research Algorithms group.

670 56. Gamaarachchi, H., Ferguson, J. M., Samarakoon, H., Liyanage, K. & Deveson, I. W.
671 Simulation of nanopore sequencing signal data with tunable parameters. *Genome Res.*

672 **34**, 778–783 (2024).

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

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

677 59. Tyler, A. D. *et al.* Evaluation of Oxford Nanopore's MinION Sequencing Device for
678 Microbial Whole Genome Sequencing Applications. *Sci. Rep.* **8**, 1–12 (2018).

679 60. Kovaka, S. *et al.* Uncalled4 improves nanopore DNA and RNA modification detection
680 via fast and accurate signal alignment. *bioRxiv* 2024.03.05.583511 (2024).

681 61. Silva, D. F. & Batista, G. E. A. P. A. Speeding up all-pairwise dynamic time warping
682 matrix calculation. in *16th SIAM International Conference on Data Mining 2016, SDM*
683 *2016* 837–845 (2016). doi:10.1137/1.9781611974348.94.

684 62. Sereika, M. *et al.* Oxford Nanopore R10.4 long-read sequencing enables the generation
685 of near-finished bacterial genomes from pure cultures and metagenomes without short-
686 read or reference polishing. *Nat. Methods* **19**, 823–826 (2022).

687 63. Liang, H. *et al.* Efficiently Constructing Complete Genomes with CycloneSEQ to Fill
688 Gaps in Bacterial Draft Assemblies. *bioRxiv* 2024.09.05.611410 (2024)
689 doi:10.1101/2024.09.05.611410.

690 64. Li, Q., Sun, C., Wang, D. & Lou, J. GCRTcall: a Transformer based basecaller for
691 nanopore RNA sequencing enhanced by gated convolution and relative position
692 embedding via joint loss training. 1–8 (2024) doi:10.3389/fgene.2024.1443532.

693 65. Barker, E. Recommendation for Key Management: Part 1 – General. (2020)
694 doi:10.6028/NIST.SP.800-57PT1R5.

695 66. Alagic, G. *et al.* Status Report on the Third Round of the NIST Post-Quantum
696 Cryptography Standardization Process. (2022) doi:10.6028/NIST.IR.8413.

697 67. Sood, A. J., Viner, C. & Hoffman, M. M. DNamod: The DNA modification database.
698 *J. Cheminform.* **11**, 30 (2019).

699 68. Boccaletto, P. *et al.* MODOMICS: A database of RNA modification pathways. 2021
700 update. *Nucleic Acids Res.* **50**, D231–D235 (2022).

701 69. Jonsson, J. & Kaliski, B. Public-Key Cryptography Standards (PKCS) #1: RSA

Cryptography Specifications Version 2.1. (2003) doi:10.17487/RFC3447.

70. Bos, J. *et al.* CRYSTALS - Kyber: A CCA-Secure Module-Lattice-Based KEM. *Proc. - 3rd IEEE Eur. Symp. Secur. Privacy, EURO S P 2018* 353–367 (2018) doi:10.1109/EUROSP.2018.00032.

71. FrodoKEM Learning With Errors Key Encapsulation Algorithm. (2017).

72. Berezin, C. T., Peccoud, S., Kar, D. M. & Peccoud, J. Cryptographic approaches to authenticating synthetic DNA sequences. *Trends Biotechnol.* **42**, 1002–1016 (2024).

73. Biotechnology – Nucleic acid synthesis – Part 2: Requirements for the production and quality control of synthesized gene fragments, genes, and genomes. (2024).

74. Xing, X. W. *et al.* Mutagenic and cytotoxic properties of oxidation products of 5-methylcytosine revealed by next-generation sequencing. *PLoS One* **8**, (2013).

75. Biodefense in the Age of Synthetic Biology. *Biodefense Age Synth. Biol.* (2018) doi:10.17226/24890.

76. 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 267–278 (2020).

77. Zhao, X. *et al.* Composite Hedges Nanopores codec system for rapid and portable DNA data readout with high INDEL-Correction. *Nat. Commun.* **15**, 9395 (2024).

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), NSQKJJ under grant K21799109 and K21799116, Zhejiang Provincial Collaborative Innovation Center for High- end Digital Intelligence Diagnosis and Treatment Equipment and Center for Computational Science and Engineering at Southern University of Science and Technology.

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. provided the plasmid samples 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 construction 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 pertaining to the training framework of DeepSME. The remaining authors declare no competing interests.

748

749 **Fig. 1 | Obfuscating digital information stored in DNA for private communication via**
750 **nanopore sequencing.** (**a**) Schematics of private communication for files and texts stored in
751 DNA. Steganography is achieved by obfuscating data stored inside DNA via modifications by
752 the sender Alice. Canonical DNAs can be transmitted publicly and readout using conventional
753 basecaller by everyone. Non-canonical DNAs in principle cannot be easily revealed by an
754 unsuspecting observer Eve, while the private observer Bob should be able to use DeepSME as
755 the key to decipher the messages from Alice. (**b**) An example of Alice's choice with the
756 chemical structure of deoxycytidine triphosphates (dCTP), which is labelled as cytosine (C)
757 and 5-hydroxymethylcytosine (5hmC), respectively. (**c**) Nanopore sequencing of canonical and
758 fully methylated 5hmC DNA strands using a state-of-the-art basecaller (Bonito/Dorado). The
759 ionic current signals of C (cyan) and 5hmC (orange) are interpreted into DNA sequences and
760 aligned with reference sequences/messages from Alice. The error bases are labelled in red,
761 corresponding to the information received by Eve using the conventional basecaller.

762

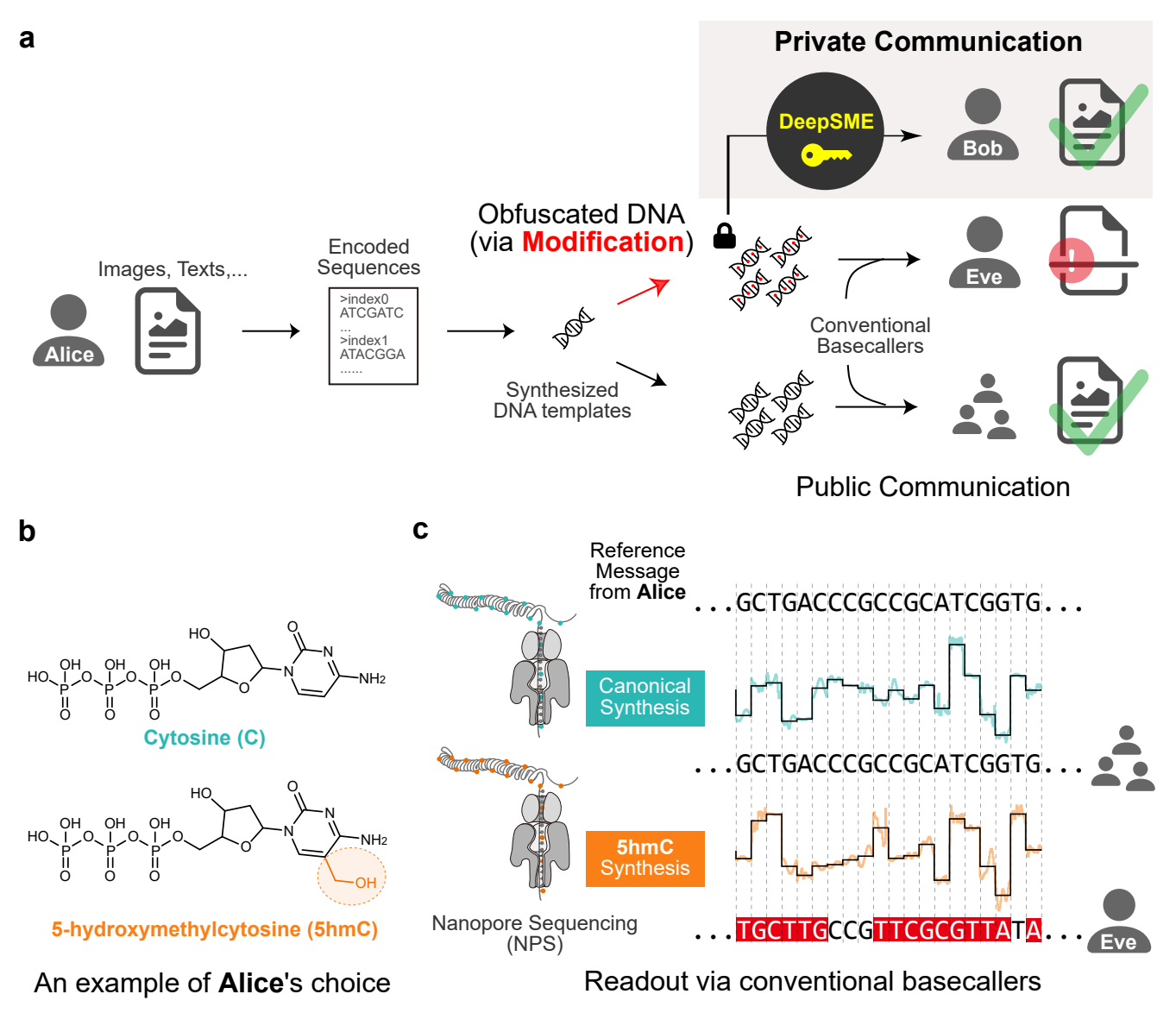
763

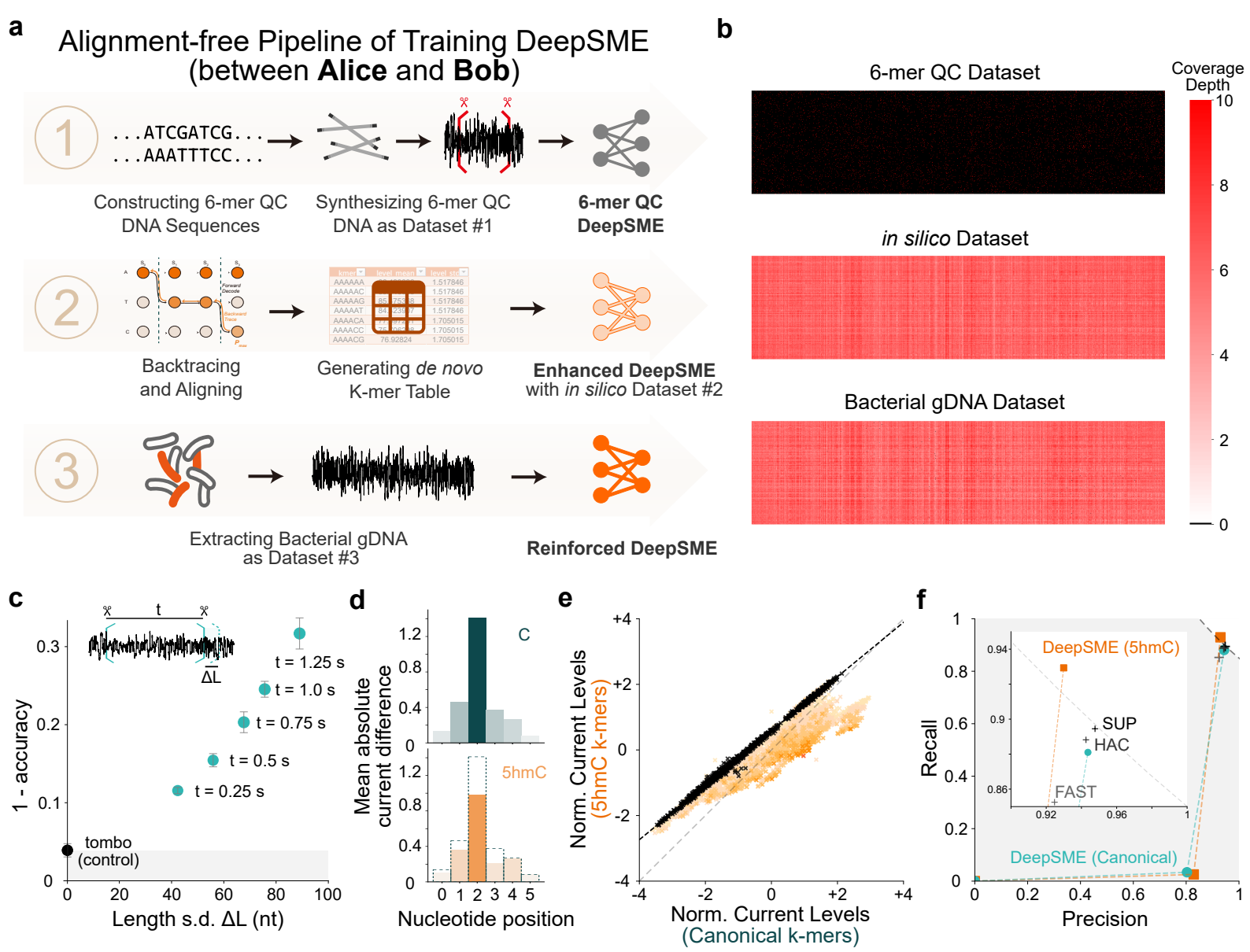
764 **Fig. 2 | *De novo* construction of non-canonical DeepSME basecaller between Alice and**
765 **Bob.** (**a**) Training DeepSME with a three-stage alignment-free pipeline. (**b**) 9-mer coverages
766 of three used training datasets. (**c**) Training error rate as a function of standard deviation of
767 segmented lengths. Inset depicts the uncertainty (standard deviation of a certain segmented
768 length) (**d**) Mean absolute current difference (MACD) of both canonical cytosine (top) and
769 5hmC (bottom) at different nucleotide positions of a k-mer oligonucleotide. (**e**) Methylated
770 5hmC vs. canonical cytosine of normalized current levels. (**f**) Recall vs. precision of DeepSME
771 for both canonical and methylated 5hmC.

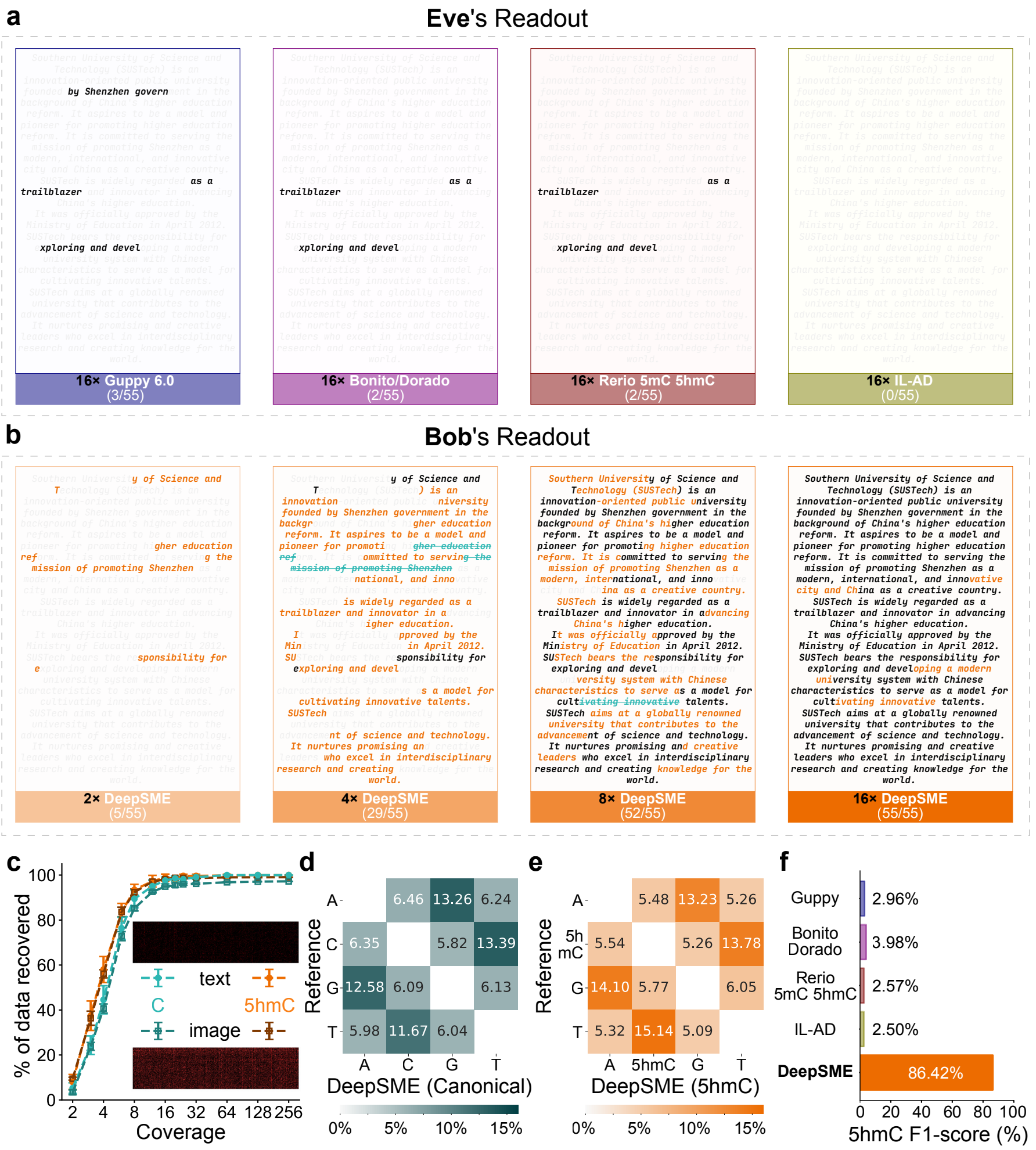
772

773

774 **Fig. 3 | Deobfuscation performance of digital information enabled by DeepSME.** (**a**)
775 Decoding performance of the text file by the unsuspecting observer Eve using Guppy 6.0
776 (`dna_r9.4.1_450bps_hac` config), Bonito/Dorado (`dna_r9.4.1_e8_sup@v3.3`
777 config), Rerio 5mC 5hmC (`res_dna_r941_min_modbases_5mC_5hmC_v001` config),
778 Incremental Learning and Anomaly Detection (IL-AD, 5hmC config) with coverage depths of
779 16×. (**b**) Decoding performance of the text file by the private observer Bob using DeepSME
780 with coverage depths of 2×, 4×, 8× and 16×. The transcoded strand amounts are 5, 33, 52 and
781 55 of total 55. (**c**) The percentage of data recovered as a function of coverage for the text file
782 stored in canonical (cyan) or 5hmC (orange) using corresponding canonical or 5hmC
783 DeepSME, respectively (n=10). (**d**) Base bias of substitutions for canonical DeepSME. (**e**)
784 Base bias of substitutions for 5hmC DeepSME. (**f**) F1-scores for variant basecallers on
785 obfuscated (5hmC) DNA.


a
Private Communication
DeepSME
Bob
Obfuscated DNA
(via Modification)
Images, Texts,...
Alice
Encoded
Sequences
>index0
ATCGATC
...
>index1
ATACGGA
.....
Synthesized
DNA templates
Eve
Conventional
Basecallers
Public Communication
b
Cytosine (C)
5-hydroxymethylcytosine (5hmC)
An example of Alice's choice
c
Reference
Message
from Alice
...GCTGACCCGCCGCATCGGTG...
Canonical
Synthesis
...GCTGACCCGCCGCATCGGTG...
5hmC
Synthesis
Nanopore Sequencing
(NPS)
...TGCTTGCCGTTCGCGTTATA...
Eve
Readout via conventional basecallers


a
Alignment-free Pipeline of Training DeepSME
(between Alice and Bob)
1
...ATCGATCG...
...AAATTTCC...
Constructing 6-mer QC DNA Sequences
Synthesizing 6-mer QC DNA as Dataset #1
6-mer QC DeepSME
2
Backtracing and Aligning
Generating de novo K-mer Table
Enhanced DeepSME with in silico Dataset #2
3
Extracting Bacterial gDNA as Dataset #3
Reinforced DeepSME
b
6-mer QC Dataset
in silico Dataset
Bacterial gDNA Dataset
Coverage Depth
c
1 - accuracy
Length s.d. ΔL (nt)
t = 1.25 s
t = 1.0 s
t = 0.75 s
t = 0.5 s
t = 0.25 s
tombo (control)
d
Mean absolute current difference
Nucleotide position
C
5hmC
e
Norm. Current Levels (5hmC k-mers)
Norm. Current Levels (Canonical k-mers)
f
Recall
Precision
DeepSME (5hmC)
DeepSME (Canonical)
SUP
HAC
FAST


a
Eve's Readout
16× Guppy 6.0
(3/55)
16× Bonito/Dorado
(2/55)
16× Rerio 5mC 5hmC
(2/55)
16× IL-AD
(0/55)
b
Bob's Readout
2× DeepSME
(5/55)
4× DeepSME
(29/55)
8× DeepSME
(52/55)
16× DeepSME
(55/55)
c
% of data recovered
Coverage
text
C
5hmC
image
d
Reference
DeepSME (Canonical)
A 6.46 13.26 6.24
C 6.35 5.82 13.39
G 12.58 6.09 6.13
T 5.98 11.67 6.04
0% 5% 10% 15%
e
Reference
DeepSME (5hmC)
A 5.48 13.23 5.26
5hmC 5.54 5.26 13.78
G 14.10 5.77 6.05
T 5.32 15.14 5.09
0% 5% 10% 15%
f
Guppy 2.96%
Bonito Dorado 3.98%
Rerio 5mC 5hmC 2.57%
IL-AD 2.50%
DeepSME 86.42%
5hmC F1-score (%)