

<https://doi.org/10.1038/s42003-025-08171-z>

Self-authenticating genomic materials in *Escherichia coli* via advanced genome signatures



Zhaoguan Wang^{1,2,7}, Jingsong Cui^{3,7}✉, Qian Liu^{1,2}, Jiawei Li^{1,2,3}, Chi Guo^{4,5}, Yiyang Liu³, Changyue Jiang³, Hui Xue³, Jiaojiao Li^{1,2}, Yonggang Ke⁶✉ & Hao Qi^{1,2}✉

The authenticity and integrity of synthetic genomic materials containing valuable intellectual property are essential for advancing scientific knowledge and enhancing biosafety. Nevertheless, existing DNA tags and watermarks have limited efficacy due to low mutation tolerance and inadequate digital encoding capacity. Here, we present “Genome Signature”, a biochemically stable and tamper-resistant DNA labeling system that enables the creation of self-authenticating genomes. Central to this system is a Golomb-ruler-derived Genome-Comb, which efficiently maps extensive nucleotide sequences onto limited codons within endogenous genes, significantly improving error correction and data encoding across millions of nucleotides. Using our labeling system, we successfully recorded a 4.5-million-nucleotide genome in living *E. coli*. The Genome Signature effectively encodes data within codon orders, and autonomously identifies and corrects mutations in our computing test, ensuring genome integrity and authenticity. Furthermore, it allows precise tracking of coded sequences across different cells, potentially advancing the development of reliable genomic materials in synthetic biology.

Synthetic genetic materials are gaining increasing significance in scientific research and industrial applications, particularly in fields, such as biomedicine¹, biomanufacturing², environmental bioremediation³, and agriculture⁴. Advances in genome synthesis⁵ and editing technologies⁶ have enabled a gradual expansion in the scale of synthetic sequences with diverse gain-of-function designs in engineered organisms (Supplementary Fig. 1a–c). However, alongside this excitement, societal concerns about the use of these deeply engineered organisms have also emerged⁷. The unauthorized use of cells and their genomic materials⁸, as well as some unpublished research from industrial and government laboratories⁹, often with high commercial value, remains a significant issue that challenged biomanufacturing industries. Additionally, the widespread use of genetically modified organisms (GMOs) with synthetic sequences places significant pressure on natural biological systems through horizontal gene transfer^{10–12}, leading to genomic contamination. In these misuses, abuse, and biocontamination scenarios, genomic materials lack no definitive references for accurately determining their identity and authenticity. As products with both high industrial value and biological risks, synthetic genomes should be

placed under close administration¹³, which demands rapid and accurate determination of the authenticity and provenance of genomic materials found in products.

Synthetic DNA tags are emerging as effective tools for labeling genomic material, thereby facilitating the identification, specification, and protection of intellectual property rights. For instance, PCRtag in synthetic *Saccharomyces cerevisiae* chromosomes¹⁴, watermark in synthetic genome of *M. mycoides*¹⁵, and DNA barcoding in genome of strain engineering¹⁶. However, conventional DNA tags face several significant limitations. While they are useful for genome identification, the introduction of extensive tag sequences into synthetic genomes can result in undesirable alterations to genome size and function. Furthermore, because synthetic DNA tags lack biological function, they are highly vulnerable to mutations during cell replication (Supplementary Table 1)^{17–20}. Continuous and unpredictable mutations can lead to the loss of information encoded in DNA tags.

The genetic code is redundant, with 18 of the 20 protein-producing amino acids encoded by multiple synonymous codons. Synonymous codon substitutions are the major genetic polymorphisms affecting bacterial gene

¹School of Chemical Engineering and Technology, Tianjin University, Tianjin, China. ²State Key Laboratory of Synthetic Biology, Tianjin University, Tianjin, China.

³School of Cyber Science and Engineering, Wuhan University, Wuhan, China. ⁴Artificial Intelligence Institute, Wuhan University, Wuhan, China. ⁵Yangtze River

Delta Technology Innovation Center of Wuhan University, Jiangyin, China. ⁶Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of

Technology and Emory University, Atlanta, GA, USA. ⁷These authors contributed equally: Zhaoguan Wang, Jingsong Cui. ✉e-mail: jscui@whu.edu.cn;

yonggang.ke@emory.edu; haqi@tju.edu.cn

expression in nature and are usually not lethal to the cell²¹. Thus, in addition to nucleotide sequences, codon order is another dimension of information hidden in the genome. Codon redundancy has been used as an artificial marker for microbial genomes, e.g., Liss et al.²² used the gene T7 RNAP tore 498 bits (83 characters), which is the most reported case, but up to 70% of the information space is unused because the binary information is embedded in amino acids with only four or six synonymous codons (Supplementary Table 2). And Wang et al.²³ performed a splitting and fusion studies of the *E. coli* genome by introducing 2521 synonymous codon changes (TCG, TCA, and TAG codons) were replaced with their synonymous codons (AGC, AGT, and TAA) in a specific region of the genome of about 0.54 Mb to achieve genomic tagging. However, despite more studies conducted on codon coding, such as overlaying error correcting codes^{24,25}, or coding with all sense codons²⁶, these labeling tools do not support the validation of detecting subtle changes in a wide range of genome sequences, and lack accurate as well as effective validation of the authenticity and integrity of the entire genome.

Given the critical need to safeguard and track synthetic genomic materials²⁷, establishing systematic paradigms for the protection of synthetic genomes is essential. To meet this challenge, we propose developing an internal genomic signature (Fig. 1a) analogous to the digital signatures²⁸ used in modern electronic documents and transactions. Digital signatures allow for independent verification of electronic documents and transactions without the need for centralized authority, a feature that is vital for optimizing the efficiency of large-scale operations. Similar principles of efficiency and security have driven the success of several groundbreaking technologies, such as quantum digital signature²⁹, quick response codes³⁰. These technologies have achieved widespread adoption by leveraging their core strengths in efficiency and security, enabling them to scale to monumental levels.

In this study, we address the challenge of authenticating synthetic genomic materials by developing the genome-comb hash mapping (Genome-CHM) algorithm, a specialized algorithm based on a Golomb-ruler-derived Genome-Comb, which allows for quick and accurate verification of the integrity and authenticity of individual genomic materials without the need for comparison to reference sequences (Fig. 1b). Using this algorithm, producers of synthetic genomic material can generate strict mapping

associations between the large genomic sequence and the codon sequences of target genes, thus seamlessly encoding the target genes in the genome as genome signatures. By verifying and repairing the mapping associations between genomic sequences and signature genes, producers, users, or any third-party administrators can use our program to validate the encoded genome. Unlike conventional DNA tags that merely label genomes, this method encodes enormous genome sequences into the limited codon order of endogenous genes, preserving the essential and inseparable nature of the genome without altering its overall size. In our biochemical tests, *E. coli* BL21(DE3) with ten genome signatures can grow stably, verification of the encoded genome could be achieved by decoding the genomic signatures; and the genome signature in *E. coli* Nissle 1917 remain recognizable even if the fragment with the signature was transferred into heterologous genomes. In our simulation tests, the genome signature enables the identification and correction of hundreds of random nucleotide mutations across *E. coli* genome, allowing for the verification of its integrity and authenticity. Overall, The Genome-CHM algorithm generates genomic signatures that are robustly scalable and flexible, provide potential self-quantization capabilities for genome, supporting secure authentication-key protection, and rigorous certification protocols. Surpassing current DNA tags, this approach enables the direct documentation and identification of the uniqueness of synthetic genomic materials.

Results

Genome-CHM algorithm for genome signature

For devising a genomic signature with advanced certification capabilities, we face the challenge of creating an algorithm that can efficiently handle large amounts of information and provide error correction at the scale of the entire genome sequence. First, instead of nucleotide, we encoded the codons of the open reading frames (ORFs) of intrinsic genes in a genome, so that the biological function of the gene can effectively protect it from being lost during cell replication. However, to use codon order as an information-carrying medium, one hurdle is the non-uniform storage space due to the uneven number and frequency of synonymous codons for each amino acid. Another issue is that the number of codons in the genome is several orders of magnitude smaller than the number of nucleotides, making codon order comparably limited in terms of storage space. Beyond data encoding, we face

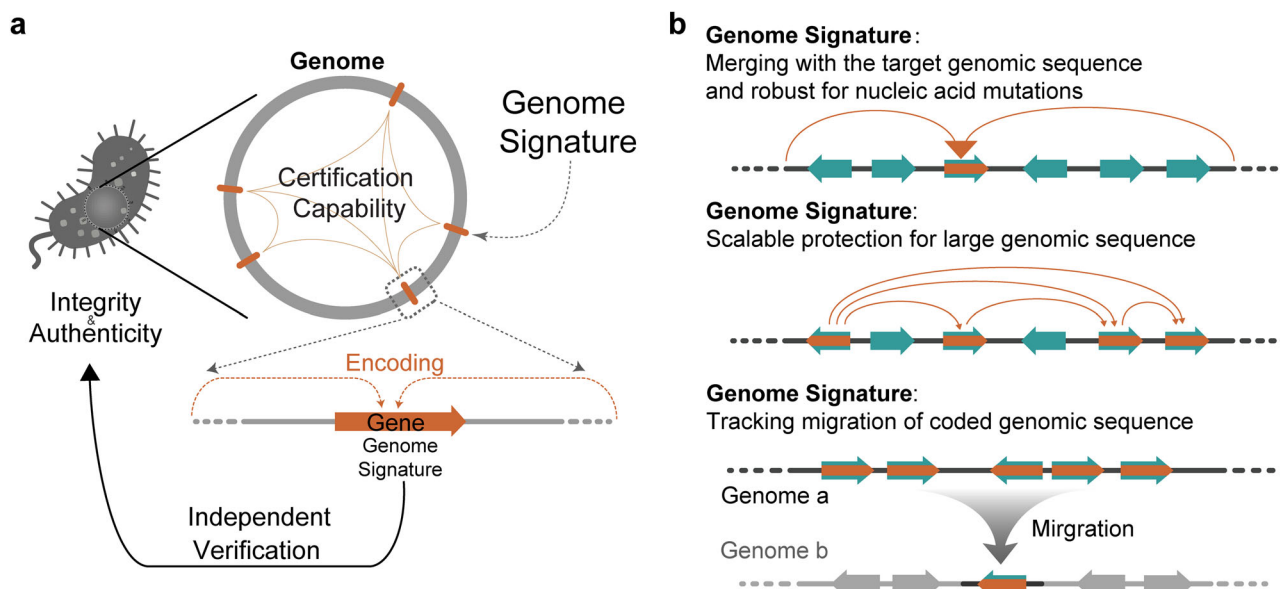


Fig. 1 | Genome signature. Diagram illustrating the implementation and functionality of the genome signature in synthetic genomic materials. **a** Overview of the genome signature's encoding process within a gene. The genome signature integrates into the target genome, providing certification capabilities that ensure integrity and authenticity. This enables independent verification of genomic

sequences. **b** Functionalities provided by the genome signature. Integration with the target genomic sequence enhances robustness against nucleic acid mutations; scalable protection mechanisms applicable to extensive genomic sequences; Tracking and verification of genomic sequence migration from genome a to genome b, ensuring accurate monitoring and protection of coded genomic sequences.

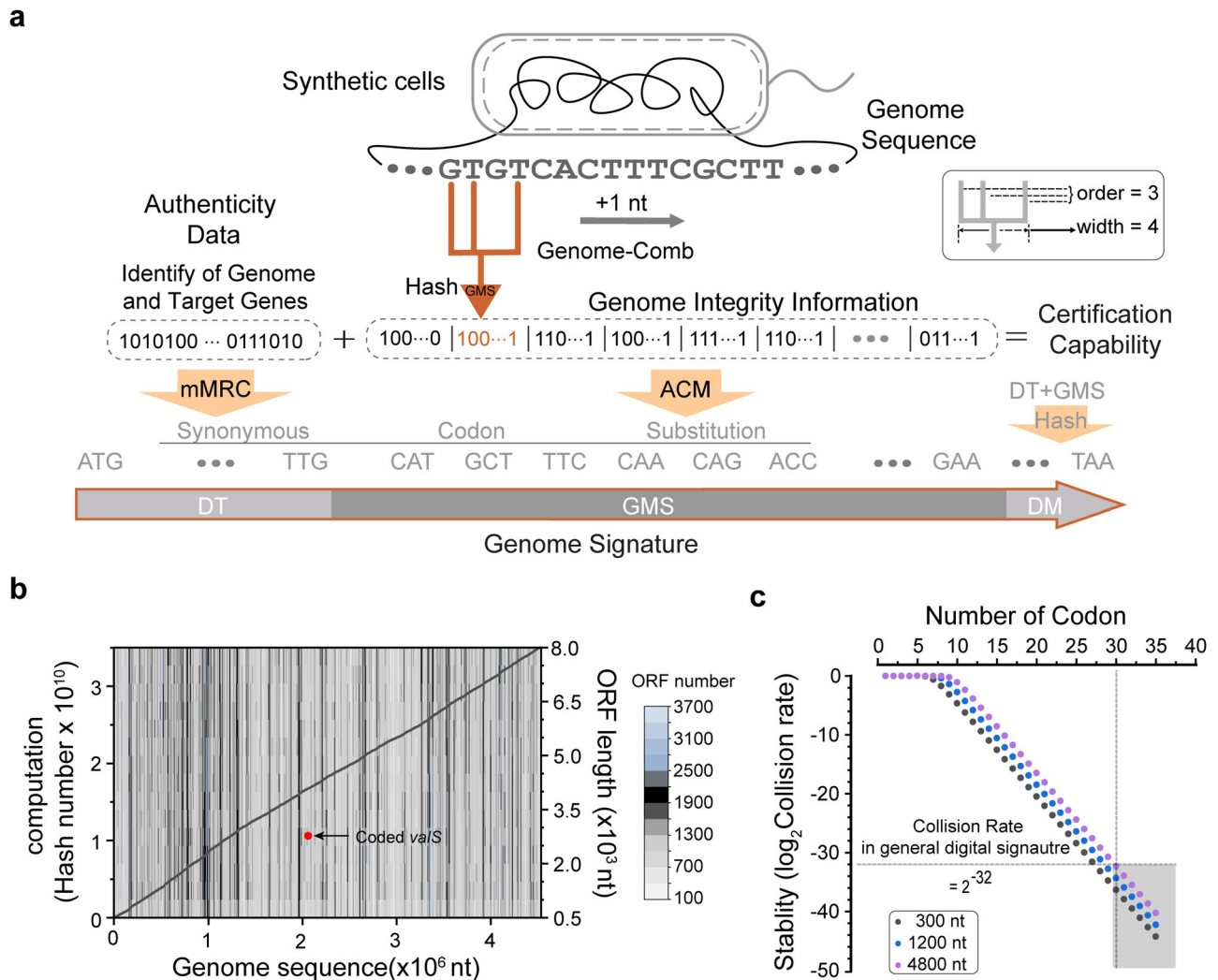


Fig. 2 | Genome-CHM algorithm. **a** Schematic representation of the Genome-CHM algorithm for encoding genome signature into synthetic genome sequence. The Genome-Comb mechanism generates a unique hash from scanning the genomic sequence, integrating it into the genome to include information on authenticity, identity, and integrity (Supplementary Note 1–4). This enables certification capabilities by mapping genome integrity information and identifying any genetic modifications. **b** Computational assessment showcasing the relationship between yielded hash numbers and coded genome sequences. Identification of the coded *valS*

gene (red dot) takes place over 16 million ORF-like sequences (Heat map), each consisting of over 180 codons, found within the 4,521,562 nt *E. coli* genome sequences. This process involves approximately 3.5×10^{10} hash calculations (black line), and the coded *valS* location on the genome indicated by an arrowhead. **c** The stability of the Genome Signature is conducted by evaluating the collision rate of DM decoding computations in relation to the number of codons in the DM of an encoded gene. The collision rate of 2^{-32} (dashed line), which is the accepted level for digital signatures in electronic documents, is achieved for DM with 30 codons.

an even greater challenge: embedding error correction codes into codons to ensure strong tolerance to mutations across the entire genome sequence, which comprises millions of nucleotides.

We developed a highly integrated algorithm Genome-CHM, incorporating modified mixed radix coding (mMRC) and directed hash-adaptive codon mapping (ACM) (Supplementary Note 1), to encode both authenticity data and mutation correction codes for genome integrity information into gene codons (Fig. 2a). First, the authenticity data, including identity information of the genomic materials as well as the identity and length of the target gene, are encoded into the codons at the 5' terminus of the gene, referred to as the data tag (DT), using mMRC (Supplementary Note 2). Generally, each codon is assigned with one integer range of [0,60) according to the codon mapping table (Supplementary Table 3) and the total number of all redundant codons for the same amino acid is fixed at 60. For instance, [41,60) is assigned to GAG (0.32 codon frequency) and [0,41) to GAA (0.68 codon frequency) in *E. coli*. Therefore, codon frequency in encoded genes can be controlled by adjusting the integer range assigned to each codon.

The relevant data are encoded into corresponding integers, which are then mapped to synonymous codons using the mMRC algorithm, rather than simply being converted into a nucleotide sequence. During encoding, the binary data is segmented, and each segment is mapped to a codon at a specific position within the target nucleotide sequence. The selection of synonymous codons at each position is constrained by factors, such as optimizing codon usage and avoiding the formation of undesirable DNA fragments. At each step, a context-specific mapping table is generated to dynamically determine the appropriate codons, excluding those that may violate biochemical or preset constraints. This position-by-position encoding process continues until the entire binary sequence is converted into a synonymous codon sequence. The flexibility in adjusting the synonymous codon (radix) at each position enables high-density encoding while ensuring sequence stability and compatibility with degenerate or composite base schemes. In the current setting, the number of codons within a single gene, as well as the number of genes that can be encoded either together or separately, is effectively limitless. This allows for the flexible expansion of the information-carrying capacity (Supplementary

Table 4). Running on our desktop computer with a medium configuration (Windows 10, 8 Gbyte of memory, 12 cores, Intel(R) Core (TM) i7-8700 CPU), coding DT in essential gene *valS* with 952 codons was completed within seconds.

Next, we developed a reliable algorithm to embed mutation correction codes onto gene codons for protecting genome integrity. This enables independent identification of mutations across the entire genome sequence and verification of integrity without the need for comparison to a reference sequence. However, the algorithm must encrypt large sequences, reaching millions of nucleotides, and its complexity significantly exceeds the capabilities of current methods. We achieve our ambitious goal by reimagining the Golomb ruler, an optimal mathematical structure that has previously been utilized in the fields of radio astronomy and nuclear magnetic resonance spectroscopy³¹. Currently, the main applications of the Golomb ruler include its use in astrophysics for image generation, in coding theory to lower information transmission rate, in crystallography to correctly locate X-ray sensors, and in sonar signal work³². Golomb ruler is a specialized ruler in which the distances between any two markings are unique. For example, a Golomb ruler of length 3 has markings at 0, 1, and 3, corresponding to distinct distances of 1 (between 0 and 1), 2 (between 1 and 3), and 3 (between 0 and 3). Leveraging this variable-length encoding scheme, a specialized Genome-Comb can be constructed with the order of 3 and the width of 4, where order denotes the number of Golomb ruler markings (0, 1, 3), and width represents the number of evenly spaced conventional markings (0, 1, 2, 3) (Fig. 2a). The Genome-Comb is capable of generating quantitative information for sequences of nearly unlimited length. This information records the relative position and content of each nucleotide, which is then mapped to codon order in a specific region within the gene. This mapping is achieved through a combined directed Hash function, commonly referred to as the genome mapping storage (GMS) (Supplementary Note 3). The GMS encompasses hundreds of codons located behind DT in the middle of the gene. Genome-Comb is designed to efficiently identify mutations across extensive sequences, including those that may span millions of nucleotides.

The Genome-Comb scans the genome sequence step-by-step (Supplementary Fig. 2), with the adjustable parameters of order and width. Within each step, a single strand of nucleotides, equivalent to the order of the DNA comb, is scanned into the Hash function $Hash_{GMS}(N_{C,i})$ to produce a monomial, $GP_i(x)$. In this monomial, the degree number represents one codon position in the GMS, and the coefficient corresponds to the value stored at that position (Supplementary Note 3). The mathematical link between a nucleotide strand from Genome-Comb and a codon in the GMS is defined as the comb-hash mapping. Finally, all the monomials are added up via the polynomial operations $GP_{\Sigma}(x)$, and the total summed coefficient is converted to one [0,60) number by modulo 60 and then mapped to one synonymous codon by ACM to generate the GMS sequence.

In the final coding stage, the last 30 codons before the stop codon are encoded as decoding marker (DM) using ACM (Supplementary Fig. 3a). Generally, the nucleotides sequence in DT, GMS, and the last 30 amino acid sequences are firstly encoded to a polynomial $DP_i(x)$ one by one by the function of $Hash_{DM}$ (Supplementary Note 4). Next, all of them are summed up via polynomial operations of $DP_{\Sigma}(x)$ and converted to one [0,60) integer by modulo 60 and then mapped to synonymous codons by ACM to generate the final codon sequence of DM. Thus, DM can not only verify the authenticity and integrity of the coded gene sequence but also effectively correct nucleotide mutations within the sequence. The termination of DM coding signifies the generation of complete genome signatures. In this study, unless explicitly indicated otherwise, the codon usage frequency in coded genes approximates the natural frequency (Supplementary Fig. 3b), as delineated in the codon mapping table. Unlike a simple genome tag, the genome signature of the coded gene is algorithmically merged with the entire genome sequence. Its essential biological function leads to robust stability, making it resistant to being easily lost or rapidly changed during evolution.

Distinct from current frameworks, genome signature offers a certification capability akin to a digital signature, ensuring independent

verifiability and enhanced security. First, the coded gene can be independently identified from the raw sequencing data of the genome, eliminating the need for reference comparison. Generally, the verification of whether an ORF sequence is a coded gene is done by examining its DM (Supplementary Fig. 4a, b, and Supplementary Note 4). The ORF sequence is subjected to generate one new 30-codon string as DM'. If DM' matched with the current last 30-codon sequence, then this ORF is one coded gene. Furthermore, the DM can tolerate the mutation within the ORF sequence. If the DM' is not matched, each nucleotide in the ORF sequence is revised to all potential substitutions, deletions, or insertions one by one. Each revised sequence is then subject to DM decoding. The cost of DM mutation correction, measured in the number of Hash calculations, increases with the number of codons. Theoretically, the DM algorithm can correct multiple random mutations, but at the cost of increased computing resources (Supplementary Note 5).

Validation of genome signature encoded in the *E. coli* genome

The unique mathematical properties of the Golomb ruler-derived Genome-Comb enable the independent verification of the integrity and authenticity of the entire genome sequence. Without requiring sequence comparison, mutations—along with their precise locations and original nucleotide contexts—can be directly identified within the nucleotide sequence from raw genome sequencing data. The procedure for the independent identification of genome signature from the entire genome sequencing data is illustrated in Supplementary Fig. 4. Each ORF-like sequence in the genome, consisting of over 180 codons from ATG to TAA, is individually subjected to DM decoding. As a result, over 16 million genome signature candidate sequences have been decoded in this *E. coli* genome (Fig. 2b). It took a few minutes to identify the genome signature *valS* after approximately $7E + 10$ hash calculations of all the candidate sequences (Supplementary Note 6), running on our desktop computer with a medium configuration mentioned before. Although, the computation cost of mutation correction in DM decoding increases with sequence length, an $O(n^2)$ time-complexity (Supplementary Note 6, Supplementary Table 5), the function of essential genes confers high sequence stability, generally making DM decoding with correction of single nucleotide mutations sufficient for identifying coded genes from sequences comprising millions of nucleotides or more.

The security of genome signature was accessed by quantifying the collision rate in DM decoding (Fig. 2c). For genome signature, a collision can be considered as a sequence with random mutations that generates the same hash value when encoding DM compared to the original sequence (Supplementary Note 7). The collision rates are relatively low, and they decline sharply with an increase in the number of codons in DM. The collision rate for DM with 30 codons can be as low as 2^{-32} , which is the widely accepted standard level for digital signatures in electronic documents.

Once a candidate sequence is confirmed as a genome signature through DM decoding, its corresponding coverage information within the genome, as recorded in the DT, is simultaneously interpreted (Supplementary Note 2). For mutation identification at the genome level, the sequences within this coverage region are then re-scanned by Genome-comb and mapped to the GMS position of the genome signature, generating a new codon order, denoted as GMS'. If GMS' is identical to GMS, it indicates that no mutations have occurred within the coverage region, confirming the sequence integrity within the genome. However, if mutations have occurred, discrepancies between GMS' and GMS will be observed (Fig. 3a). Any nucleotide mutation leads to multiple codon-changes in GMS'. Each codon-change in GMS' can be traced back to multiple nucleotides using the corresponding Comb-Hash mappings. Each nucleotide can then be assigned a QA-score (questionable authenticity score) by counting all Comb-Hash mappings that point to it (Supplementary Note 8). Theoretically, based on the special mathematical feature of the Golomb ruler³², a mutated nucleotide in a sequence will be scanned "order" times by Genome-Comb and causes no more than $2 \times$ order codon-changes in GMS. In GMS, one nucleotide mutation leads to the new Comb-Hash mappings (error-in) and the loss of existing comb-hash mappings (error-out) (Supplementary Fig. 5a).

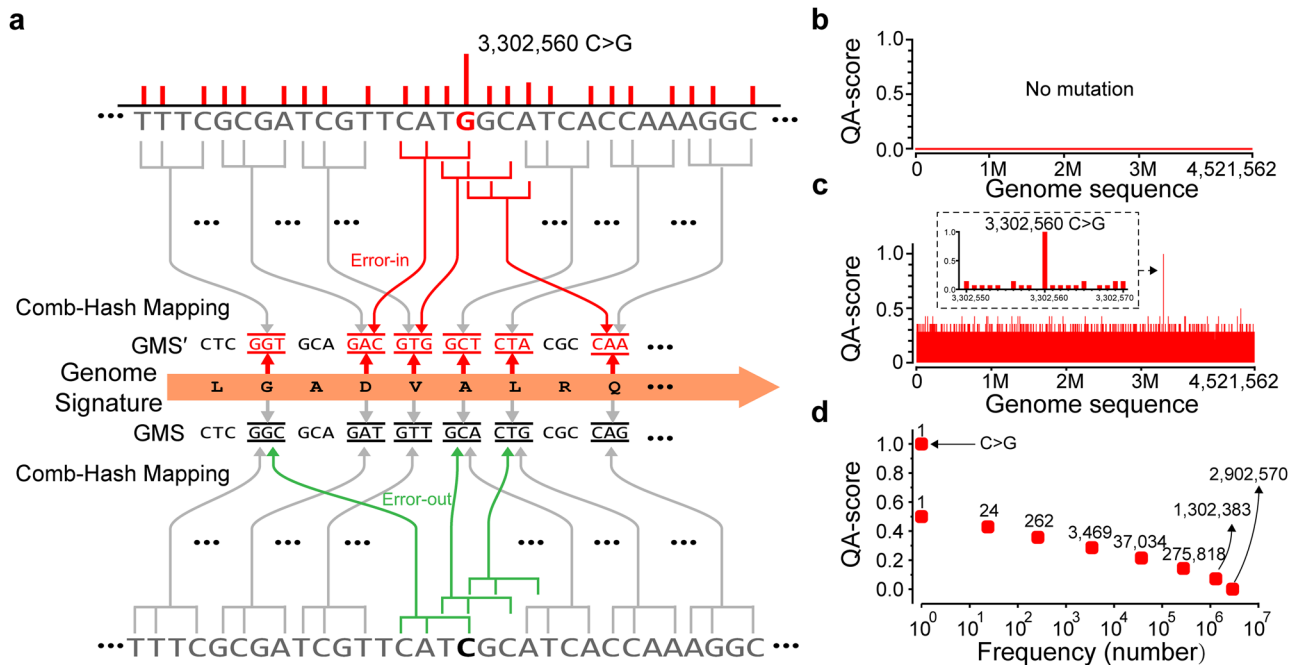


Fig. 3 | QA-score for mutation correction. **a** Diagram of the Comb-Hash Mapping process in GMS decoding illustrating the detection of a specific mutation (3,302,560 C > G). Substitution of 3,302,560 C > G causes changes in GMS. Comb-Hash mappings of both the origin nucleotide C related (error-out, green line) and the substitution G related (error-in, red line) are quantified as a value, termed as QA-score (questionable authenticity score). Non-changed Comb-Hash mappings are

indicated with gray solid lines. **b** Quality Assurance (QA) score profile across the genomic sequence showing baseline values in the absence of mutations. **c** QA score profile revealing a significant spike at the mutation position of substitution 3,302,560 (C > G) within the genome sequence, the inset displays the 20 nucleotides surrounding mutation. **d** The mutated C > G 3,302,560 nucleotide scored 1. All other nucleotides scored ≤ 0.5 , and over 99.99% of them scored below 0.3.

The error-in comb-hash mappings exclusively point to the mutated nucleotide, resulting in a higher QA-score. Conversely, the error-out Comb-Hash mappings lead to an increased number of codon-changes within GMS. However, the strong hash collision results in all codon-changes related to Comb-Hash mappings, except for error-in mappings, pointing toward nucleotides dispersed throughout the entire genome sequence. This leads to a relatively low QA-score noise. As shown in Fig. 3b, all the correct nucleotides in the genome with a single coded gene, *valS* (located at 3,900,299–3,903,154), have a score of zero. Conversely, a nucleotide with a high QA-score either indicates a mutation or is located near a mutation. As shown in Fig. 3c, a nucleotide substitution (3,302,560 C > G) obtained a significantly higher QA-score compared to all other 4,521,561 nucleotides. ~64.19% of all nucleotides scored zero, while over 99.99% of nucleotides had QA-scores lower than 0.3 (Fig. 3d).

Once the coded gene has been identified, the integrity of the entire genome sequence is evaluated using the QA score. This assessment facilitates the independent correction of nucleotide mutations throughout the genome. Specifically, all nucleotides in the genome are ranked by QA score in descending order and then subjected to revision—by substitution with one of the other three bases, deletion, or insertion of A, T, C, or G—until the GMS' aligns with the GMS (Supplementary Fig. 5b). To demonstrate the accuracy of the QA-score, every nucleotide in the *E. coli* genome was mutated one-by-one. Over 99.9% of the mutated nucleotides were correctly identified by the GMS with high QA-scores, ranked in the top 1000 of QA-scores (substitution mutation nucleotides: 4,514,962, insertion mutation nucleotides: 4,515,469, and deletion mutation nucleotides: 4,514,867). The remaining mutated nucleotides were still ranked within the top 1% of high QA-score (Supplementary Fig. 5c). In the GMS algorithm, the sensitivity to substitution is related to the number of Genome-Comb teeth, and the width for indels (Supplementary Note 3). In the GMS algorithm, the sensitivity to substitution is determined by the number of Genome-Comb teeth, which refers to the pattern-matching capability of the algorithm. A higher number of Genome-Comb teeth leads to increased sensitivity and accuracy in detecting substitutions. Additionally, the width of Genome-Comb

influences the sensitivity to indel mutations. A broader width for indels allows the algorithm to detect larger insertions or deletions more effectively. Therefore, GMS is efficient for identifying all types of nucleotide mutations (Supplementary Fig. 6), and it further demonstrates that the Genome-Comb based QA-score is highly efficient in locating and correcting mutations in the whole genome comprising of over millions of nucleotides. The decoding cost of the GMS algorithm can be simply considered as a function of the total sequence length. This means that as the length of the input sequence increases, the decoding cost will also increase in a linear manner. This linear time complexity is referred to as $O(n)$, where n represents the length of the input sequence. This aspect of the GMS algorithm makes it computationally efficient and scalable for large-scale mutation detection tasks (Supplementary Note 9). A total of 27.14 million hash calculations were required to correct one nucleotide substitution in a genome containing a single coded gene of *valS*. In long-term bacterial evolution, it has been observed that single nucleotide mutations account for over 85% of genetic changes evolution¹⁹ (Supplementary Fig. 1d). Specifically, substitution mutations were found to occur almost eight times more frequently than indels. Notably, the majority of these mutations tend to occur outside of essential genes.

Genome signature safeguarded synthetic *E. coli* genome

The effectiveness of the genome signature is influenced by the number of codons present in the coded gene. Although a single gene is constrained by the number of its codons, the stability and data storage capacity of genome recording can be significantly enhanced by incorporating multiple genes (Supplementary Table 4). For example, in the general genomes of *E. coli*, over 300 essential genes can be used as capable candidates. By dividing the genome into fragments and encoding each fragment with multiple genes—and allowing a single gene to encode multiple fragments—the same target sequence is marked by multiple genes with high QA-scores, resulting in greater stability and accuracy of genome recording. The mutated nucleotide is assigned an even higher QA-score by summing up the QA-scores of all genes that encode the same sequence (Supplementary Fig. 7). As a result of

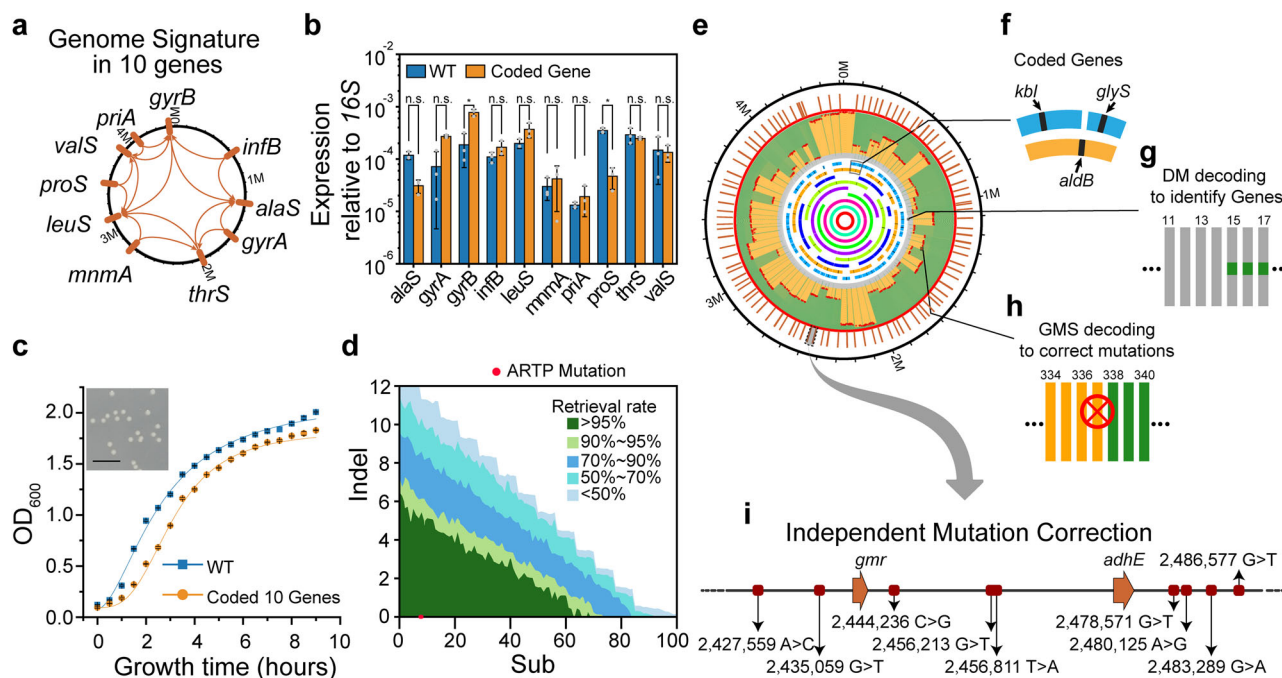


Fig. 4 | Genome signature protected synthetic *E. coli* genome. **a** *E. coli* genome of 4,521,562 nucleotides in length with a genome signature that is coded in ten genes (Supplementary Data 2). Each gene encodes a genome segment ranging from 622,018 to 4,521,562 nucleotides. **b** A genome with ten coded genes is constructed by precise genome editing (“Methods,”) and the expression levels of the ten encoded genes were quantified using quantitative PCR. Error bars represent the means \pm SD from $n = 3$ biologically independent samples. * represent $p < 0.05$; two-tailed t test. **c** The cell growth is measured in standard LB medium, the data represent the means \pm SD from $n = 3$ biologically independent samples, scale bar = 1 cm. The orange bars represent the coded genes, while the blue bars represent the corresponding wild-type genes. **d** Various combinations of random substitutions (X-axis), and indels (Y-axis) were identified in computation experiments. The success rate of correction for each mutation combination was obtained from $n = 100$ computations. The mutations (red dot) were successfully identified in the sequenced genome of cells that were treated with ARTP (“Methods”). **e** The genome of *E. coli*

BL21(DE3) contains the genome signature encoded in a total of 100 genes, their position on the genome was marked as outer red (mutation correction) bars. A nine-layer nested structure is depicted in the center as curved color lines, representing the 100 segments that are coded within these 100 genes (Supplementary Data 2). **f** Gene *kbl* codes a segment of 76,795 nucleotides (right blue curve line), *glyS* codes the downstream segment of 64,937 nucleotides (left blue curve line) and *aldB* codes segment spanning both *kbl* and *glyS* segments of 143,271 nucleotides (orange curve line). **g** All of the 100 coded genes were identified in iterative DM decoding cycles depicted as numbered gray lines. **h** A total of 700 substitution mutations were identified and corrected. A Red cross shows a successfully corrected mutation at round 337 of the GMS decoding. Orange lines indicate sequences with under-terminated authenticity, while green lines represent sequences with GMS-validated authenticity. **i** Distribution of decoded genome signature and mutations in the local genome.

collaboration between multiple joint genes, the accuracy of decoding and locating the mutation position is significantly improved. This collaborative effort allows for more precise identification of the mutated position in the genome (Supplementary Note 10).

E. coli cells with Genome Signature encoded in ten essential genes (total of 7552 codons encoded) were constructed using standard genome editing methods (Fig. 4a). The Genome Signature in this *E. coli* strain consists of five unique genes, each encoding a distinct fragment, while the remaining five genes encode multiple fragments (Supplementary Fig. 7). The genome sequence is structured in five layers, with each fragment being encoded by 4–5 genes. The expression of coding genes was quantified relative to 16S, and statistical analysis showed that most of the genes had no significant differences compared with wild-type sequences. Only *gyrB* and *proS* showed a statistical difference ($P < 0.05$), indicating that the overall effect was still small. (Fig. 4b). Furthermore, the *E. coli* cells containing the genome signature in multiple genes exhibited normal growth rates (Fig. 4c). In the computational experiments, the decoding capability was quantified by performing over a million calculations involving random mutations, including multiple substitutions, insertions, and deletions across the genome (Fig. 4d). Furthermore, mutations were successfully decoded from *E. coli* cells that were sequenced after undergoing treatment with atmospheric and room temperature plasma (ARTP) for random mutations generation.

Moreover, we also tested the correction of fragment mutations that involved multiple consecutive substitutions or indels. In contrast to robust decoding of up to 20 consecutive substitutions, the successful decoding of

indels was limited to only 2–3. As described above, indel causes more codon-changes than substitution (Supplementary Fig. 5a), but indel correction can be further improved, e.g., recording more auxiliary information, such as the fragment size, in the genome signature. In theory, identifying fragment indels is a highly complex problem, and there are currently no practical algorithms available³³. However, Genome Signature has the ability to locate the positions of fragment indels on the genome (Supplementary Fig. 7). Since fragments with severe mutations do not impact the decoding of other genome fragments, the encoding of the genome signature in multiple genes efficiently reduces the risk of collapse during decoding. Long-term bacterial evolution studies^{18–20} have shown that bacterial genomes accumulate only a few mutant nucleotides each year (Supplementary Table 1). Therefore, by encoding the genome signature in ten joint genes, the genome demonstrated improved robustness.

Furthermore, a more effective genome signature was encoded in 100 jointly coded genes (totaling 73,557 codons), which included 11 essential and 89 nonessential genes. The 100 genes formed a nine-layer nested structure (Fig. 4e) to achieve a high capability for mutation correction. The shortest segment of the genome coded in the gene *frwA* (834 codons) is 24,260 nucleotides, while *fecA* (775 codons) codes the entire genome consisting of 4,521,562 nucleotides. Consequently, on average, every segment of the genome sequence is coded by seven genes. In a computational experiment, the genome sequences with randomly mutated nucleotides were subjected to decoding. All 100 coded genes were identified in 100 repeated DM decoding cycles (Fig. 4f, g, and Supplementary Table 6) after correcting

mutations in the sequence of all coded genes in about 2 s. Next, hundreds of mutations outside the coded gene sequences were identified and corrected through 703 iterative GMS decoding cycles, which took approximately 136.82 h (Fig. 4h, i and Supplementary Movie 1). After the correction of 700 substitution mutations, the integrity and authenticity of the genome were successfully determined without querying any references for comparison. In another computational experiment (Supplementary Fig. 8), a total of 97 substitutions and ten indels were identified and corrected, resulting in the recovery of approximately 97.39% of the whole genome sequence. However, there was a failure to correct 53 substitutions and 50 indels on a single fragment of 118,126 nt. Severe mutations may overload the computation and cause a failure in decoding one fragment, but the damage can be controlled without affecting other fragments.

Tracking the genome signature across distinct genomes

A partial sequence of a synthetic genome, particularly one designed with gain-of-function features, has the potential to be horizontally transferred into other cells³⁴, thereby raising significant biosafety concerns within the rapidly advancing field of synthetic biology. Concurrently, the unauthorized use of genomic sequences featuring unique intelligent designs poses a major issue regarding intellectual property rights in the GMO-related industry³⁵. The genome signature, with its high stability and distinct mechanism, can precisely trace an encoded sequence embedded within various genomic contexts. This capability allows it to determine the origin and authenticity of the sequence without the need for sequence comparison against references, which is particularly valuable for confirming its ownership in legal disputes.

To demonstrate the feasibility, a 7035 nt fragment of the *E. coli* Nissle 1917 genome was encoded with the genome signature of 2202 nt in the gene *iutA* (Fig. 5a). *E. coli* Nissle 1917 is a probiotic strain that is widely used in biomedicines and food production³⁶. The Nissle cell was constructed using precise genome editing techniques and grew normally (Fig. 5b). To mimic the horizontal transfer of genomic material, this Nissle genomic fragment containing the genome signature was inserted into three different locations within the *E. coli* BL21 genome. The genetically invaded BL21 cells were cultured, and total genomic DNA was collected. Subsequently, the collected DNA samples were subjected to high-accuracy long-read sequencing using the PacBio HiFi platform, which has a sequencing error rate lower than 0.1% (“Methods”). A total of 323,846 PacBio long reads, each with a length of over 2 Kbp, were collected and subjected to individual genome signature decoding (Supplementary Table 7). All 312,631 long reads were processed, and nucleotide mutation correction and authenticity determination were carried out using approximately $7.3E + 14$ hash calculations. Ultimately, 40 sequences carrying the full 7035-nt invaded sequence were successfully determined (Fig. 5c), while only a partial invaded sequence (coded *iutA*) could be determined in 148 sequences. To assess the limitations of genome signature, an Australian lungfish genome, which is regarded as one of the largest sequenced genomes to date³⁷, was utilized to track migrated genomic sequences. Initially, a single, comprehensive genome sequence was constructed by concatenating all 14 chromosomes (GCA_016271365.1) of the lungfish, resulting in a final size of 34,557,647,948 nt (Fig. 5d). The Nissle fragment was inserted at base position 18,144,212,270 within this composite genome. Remarkably, even in the context of such high sequence complexity, the full invaded fragment was successfully determined through approximately 11.7 billion Hash calculations.

Unlike traditional DNA tag or watermark frameworks, genome signatures possess the capability to both identify the presence of their own and directly assess the integrity and authenticity of the entire genomic sequence. Consequently, genome signatures could serve as a powerful tool for monitoring the contamination of synthetic sequences in natural organisms and detecting the unauthorized use of intelligently designed genomic materials.

Discussion

Today, genomic sequence information is indeed considered a form of Big Data, with global databases storing dozens of exabytes³⁸. Furthermore, the field is anticipated to experience even more rapid growth in the future,

presenting additional hurdles for existing genome characterization methods. The high-integrity storage of genomic data in centralized public databases, essential for genomic analysis, has been hindered by various technical, ethical, and legal concerns. This powerful genome signature platform has the potential to help improve the centralized storage of crucial genetic information. Instead of relying on a centralized database, the platform stores the information for integrity, identification, specification, and intellectual property rights protection directly in the genome itself. With the multitude of challenges our society faces due to the rapid development of GMOs, genome signature promises viable solutions and can play a significant role in advancing synthetic biology for both scientific research and real-world applications.

Genome signature can be encrypted into a genome in living cells while also preserving the integrity of the genome structure and causing no changes to the size of the genome sequence. The inherent connection between coded genes and the genome sequence diminishes the likelihood of damage or loss in genome-wide genome signatures. In this study, we developed living bacteria *E. coli* cells with a robust genome signature in as many as ten genes, but the method can easily be scaled up to include more genes. Microorganisms that are fully annotated and have applicable gene editing tools have the opportunity to be able to realize genome signatures, according to the selection principle of genes (Supplementary Note 11). Genome signatures represent an alternative platform for directly and accurately determining genome integrity and verifying the authenticity of genomic sequences, without the need for reference-based sequence comparisons. Unlike conventional DNA tag frameworks, genome signatures provide an alternative approach the administration of synthetic genomic materials by offering a reference-free approach. Just like the stability and security achieved in digital signature technology, once encrypted, the algorithmic connection and digital data encrypted inside essential genes cannot be tampered with. Furthermore, the combination of directed hashing in the genome-CHM algorithm gives genome signatures the ability to provide certification. The use of authentication keys is able to ensure high security. As a result, genome signatures have the potential to achieve a high level of biosafety and can be a powerful tool for protecting intellectual property and safeguarding GMOs. The exponential expansion of synthetic genomes has spurred widespread concerns regarding their potential impact on multiple facets of society, with a special emphasis on environmental safety. Genome signatures can address these concerns by enabling independent monitoring, without the need for reference-based sequence comparisons, of potential contamination of synthetic genome sequences in natural systems.

However, genome signatures still face certain limitations, such as different tolerance for substitutions and indels. This highlights the need for continued algorithmic optimization, such as exploring different Genome-Comb combinations, to improve indel processing efficiency and reduce computational overhead. Furthermore, further research is required to expand the applicability of genome signatures. In particular, their potential use in widely studied organisms, such as *Bacillus subtilis*, *Lactococcus lactis*, and *Saccharomyces cerevisiae*, need to be further explored. The stability and robustness of genome signatures under more complex environmental conditions also need to be systematically evaluated. Nevertheless, with continued in-depth research, genome signatures are expected to contribute significantly to the establishment of a standardized platform for monitoring and managing synthetic genomic materials.

Methods

Biological resources

E. coli BL21(DE3) S-ABP (our lab) was used for construct the strain containing the Genome Signature collectively referred to as *E. coli* BL21(DE3)³⁹. Sequences of ten coded genes for constructing Genome Signature contained strain were synthesized by Liuhe BGI (Beijing, China). All primers were synthesized by Azenta (Suzhou, China). Original sequence for simulating the migration was from *E. coli* Nissle 1917. The genome data of the lungfish *Neoceratodus forsteri* (Australian lungfish) is from Genbank (GCA_016271365.1). The original genome sequence and the information

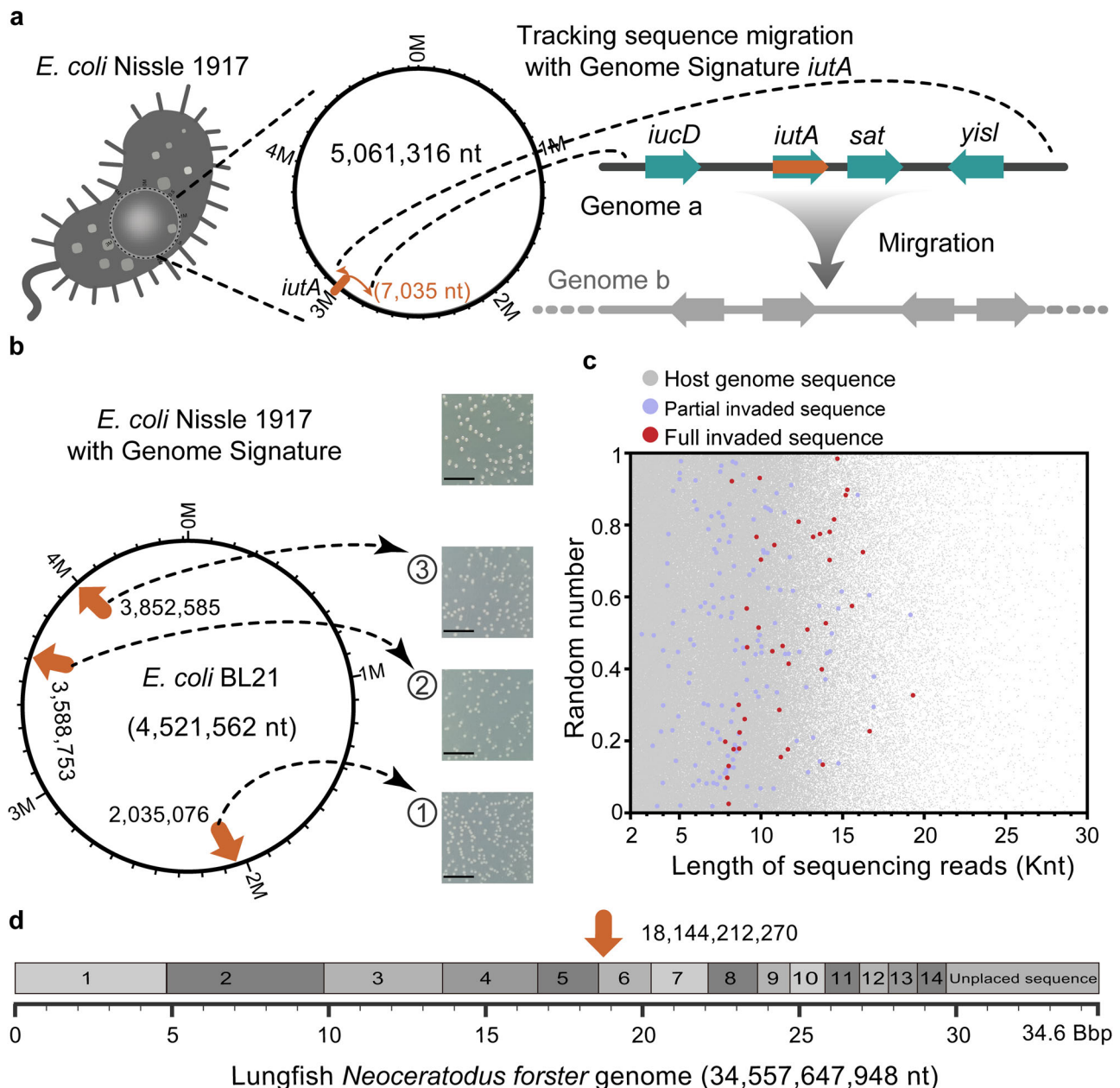


Fig. 5 | Tracking genome signature across distinct genomes. a One fragment of 7035 nt of *E. coli* Nissle 1917 with genome signature in one gene of *iutA*. **b** Nissle *iutA* fragment was inserted in the genome of *E. coli* BL21 at different locations (1: 2,035,076; 2: 3,588,753; 3: 3,852,585). *E. coli* cells cultured on standard LB solid medium scale bar = 1 cm. **c** These *E. coli* BL21 strains were sequenced using the PacBio platform, and reads exceeding 2 Kbp in length were directly decoded. Reads

with a determined full *Nissle* fragment are indicated by red dots, while reads with only the determined coded *iutA* gene but without the full length of the fragment are represented by purple dots. The y-axis indicates the random number of the sequencing read. **d** Nissle coded *iutA* fragment (inserting at 18,144,212,270) was successfully determined in a giant genome sequence constructed by jointing all 14 lungfish chromosomes (GCA_016271365.1), a total of 34,557,647,948 nt in length.

about all coded genes used in this study are listed in Supplementary Data 1 and 2, respectively. Primers for the Sanger sequencing and quantitative PCR (qPCR) are listed in Supplementary Data 3.

Strain construction

To construct a strain containing the genome signature, we prepared ten linear recombinant engineering templates to replace the corresponding wild-type gene sequence with their respective genome signatures. The DNA sequence of the linear recombinant template includes a 500 bp upstream sequence adjacent to the gene in the genome (the upstream homology arm), the genome signature sequence, the ampicillin resistance gene sequence as a screening marker, and a 500 bp downstream sequence adjacent to the gene in the genome (the downstream homology arm). The upstream and

downstream homology arms were amplified from the genome in a 50 μ l PCR reaction using PrimeSTAR Max DNA Polymerase (R045Q, Takara, China), respectively. Overlapping PCR was used to prepare the linear recombination template for gene editing. These fragments were then electroporated into competent cells for genome editing, and *L*-arabinose was used to induce the removal of the selection marker, following the method proposed by Kuhlman et al.⁴⁰. In order to implement *thrS* coding, a copy of *infC* was made on plasmid following the method proposed by Ayyub et al.⁴¹.

We used the genome editing methods based on pECas/pEcgRNA⁴² for simulating the migration of a fragment of 7035 nt watermarked by Genome Signature *iutA* from *E. coli* Nissle 1917 into *E. coli* BL21. gRNA target for each of these positions was designed on <https://chopchop.cbu.uib.no/> to construct pECgRNA-target plasmid. The linear recombinant template for

gene editing includes the 500 bp upstream homology arm, the migration fragment, and the 500 bp downstream homology arm. The sequence of the homology arm was amplified from the genome of *E. coli* BL21 in a 50 µl PCR reaction using PrimeSTAR Max DNA Polymerase, respectively. The migration fragment was amplified from the genome of *E. coli* Nissle 1917 in a 50 µl PCR reaction too. The homology arm DNA fragment and migration fragment were assembled by overlapping PCR to prepare the linear recombinant template for gene editing. Competent cells for 100 µl were mixed with 1000 ng of linear recombinant template and 100 ng pEcgRNA-target plasmids. The resulting mixture was then electroporated. Cells were resuspended in 1 ml SOB medium after electroporation. The cells were allowed to recover at 37 °C for 1 h and then plated on LB agar plates with 50 µg ml⁻¹ Kanomycin, and 100 µg ml⁻¹ spectinomycin incubated overnight at 37 °C.

To verify the results of genome editing, colony PCR was performed using primers designed approximately 150 ~ 200 bp upstream and downstream of the editing site. PCR reactions were carried out following the manufacturer's instructions using EasyTaq DNA Polymerase (AP111-11, TransGen, China) and a specific primer set (F/R) corresponding to each genome editing site. The PCR products were then subjected to Sanger sequencing for validation.

Growth curve monitoring

The growth of *E. coli* BL21(DE3) with ten joint genome signatures was tested in LB medium (10 g l⁻¹ tryptone, 5 g l⁻¹ yeast extract, 10 g l⁻¹ NaCl). First, the cells were cultured in LB medium overnight at 37 °C with shaking at 220 r.p.m. The overnight cultured cells were added to LB medium in the ratio of 1:100, respectively, and cultured at 37 °C with shaking at 220 r.p.m. The optical density (OD) was measured at 565 nm every 0.5 h using a DEN-1B Densitometer (Keison products, Germany).

Quantitative PCR

qPCR experiment was performed to assess the expression of ten genes (*infB*, *alaS*, *gyrA*, *thrS*, *mmmA*, *leuS*, *proS*, *valS*, *gyrB*, and *priA*) in *E. coli* BL21(DE3) cell and cell with ten joint genome signatures. The cells were cultured in LB medium overnight at 37 °C with shaking at 220 r.p.m. Total RNA was extracted from growth-saturated cells using TRIzol reagent (15596026, Invitrogen, USA) according to the manufacturer's protocol. HiScript II Q RT SuperMix for qPCR (+ gDNA wiper) (R222-01, Vazyme, China) was used to remove genomic DNA contamination in the sample and reverse transcript the sample. qPCR reaction was prepared with AceQ qPCR SYBR Green Master Mix (Vazyme, China) and performed in 20 µl reaction mixtures using a QuantStudio6 Flex Real-Time PCR System (Thermo Fisher Scientific, USA) following the manufacturer's protocol. We selected 16S RNA as a reference gene. $\Delta C_T = C_T^{\text{gene}} - C_T^{16S}$, so gene expression relative to 16S RNA is $2^{-\Delta C_T}$.

Mutagenesis using ARTP

Fresh single colony of *E. coli* BL21(DE3) pTKRED was cultured in 3 ml of LB medium with 100 µg ml⁻¹ spectinomycin overnight at 30 °C. The next day, the overnight cells were transferred to LB medium with 100 µg ml⁻¹ spectinomycin at 1:100. Once the OD ($\lambda = 565$ nm) of the culture reached 0.6, cells were diluted 10 times by LB medium, 20 µl diluted culture was treated by ARTP using an ARTP mutation system (ARTP-IIS, Tmxtree Biotechnology, Wuxi, China) with the following parameters: (1) the radio frequency power input was 120 W; (2) the flow of pure helium was ten standard liters per min; (3) the distance between the plasma torch nozzle exit and the slide was 2 mm; and (4) the treatment times were 120 s. After treatment, the slide was washed with 1 ml of LB medium in a 1.5 mL falcon tube, and cultivated at 30 °C and 220 r.p.m. for 1 h. Then the culture medium was gradient diluted and coated onto an LB plate. A single colony was picked to 3 ml of LB medium with 100 µg ml⁻¹ spectinomycin and culture at 30 °C overnight. The next day, the cells were transferred to LB medium with 100 µg ml⁻¹ spectinomycin at 1:100. Cells were collected by centrifugation once the OD ($\lambda = 565$ nm) of the culture reached 0.6 and

washed twice with 1 ml of PBS buffer (PH7.4 slightly), then the bacteria were frozen by liquid nitrogen and send for whole genome sequencing. The sequencing reads were aligned using BWA (version 4.2.0.0) and analyzed with GATK (version 0.7.17).

Genome long-read sequencing

Single colonies of *E. coli* BL21 with migrated fragments at different locations were selected and cultured in 3 mL LB medium overnight respectively. The next day, the cells were mixed and transferred to three 1 L LB medium at 1:1000. Cells were collected by centrifugation once the OD ($\lambda = 565$ nm) of the culture reached 0.6 and washed twice by 50 mL of PBS buffer (PH7.4 slightly), then, the bacteria were frozen in liquid nitrogen and sent for sequencing on the PacBio Sequel IIe platform to obtain HiFi reads. Reads with mass scores below Q5 were filtered out using fastp (version 0.23.1) and divided into three independent sets, each containing over 90,000 reads. Each read was individually processed by the decoder to identify potential signatures.

The giant genome of lungfish

The genome data of the lungfish is 34,557,647,948 nt after removing 129 N. The final linear genome is obtained by connecting 15 fragments one after the other according to the following sequence: chromosome 1–14 and unplaced sequence fragment. Linear genome carrying a migration fragment containing the genome signature *iutA* serves as the decoding input.

Computing resources

The simulation data were obtained from the following machines: 32Gbyte of memory, 64 cores, 2.50 GHz Intel Xeon CPU E5-2682 v4; 16Gbyte of memory, eight cores, Intel Xeon Processor (Skylake, IBRS); 48Gbyte of memory, 12 cores, 2.67 GHz Intel Xeon CPU × 5640. A list of computational details and statistical tests can be found in Supplementary Table 6.

Statistics and reproducibility

qPCR assays and growth curve experiments included $n = 3$ biological replicates for each sample. Decoding statistics for invaded sequence were conducted $n = 3$ independent analysis. Simulation test for the decoding of genome with ten joint genome signatures were performed $n = 100$ for each mutation combination. Simulation test for DM decoding were performed $n = 50$ for each gene. Data visualizations were created using Graphpad Prism v9, Origin v2021 and Python v3.11. The figure legends provide detailed information on the specific data points for each experiment, including sample sizes (n values), p -values, and other relevant metrics.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data generated in this study, including data from *E. coli* mutated by ARTP and *E. coli* with integrated migration sequences, are available at <https://figshare.com/s/fc3e05395b5dadca2bd5>. The source data for Fig. 4e, Figs. S6, S8 are available at <https://figshare.com/s/a9f0746ecab51c700445>. Other source data behind the graphs in the paper are provided as Supplementary Data 4. Additional information is available from the corresponding author upon reasonable request.

Code availability

Demo for the genome signature encoding and decoding is available from the following links: <https://genomemory.org> or <https://doi.org/10.5281/zenodo.15314176>⁴³.

Received: 4 November 2024; Accepted: 5 May 2025;

Published online: 16 May 2025

References

- Weber, W. & Fussenegger, M. Emerging biomedical applications of synthetic biology. *Nat. Rev. Genet.* **13**, 21–35 (2011).
- Moe-Behrens, G. G., Davis, R. & Haynes, K. A. Preparing synthetic biology for the world. *Front. Microbiol.* **4**, 5 (2013).
- Rylott, E. L. & Bruce, N. C. How synthetic biology can help bioremediation. *Curr. Opin. Chem. Biol.* **58**, 86–95 (2020).
- Ke, J., Wang, B. & Yoshikuni, Y. Microbiome engineering: synthetic biology of plant-associated microbiomes in sustainable agriculture. *Trends Biotechnol.* **39**, 244–261 (2021).
- Zhang, W., Mitchell, L. A., Bader, J. S. & Boeke, J. D. Synthetic genomes. *Annu. Rev. Biochem.* **89**, 77–101 (2020).
- Chari, R. & Church, G. M. Beyond editing to writing large genomes. *Nat. Rev. Genet.* **18**, 749–760 (2017).
- Schmidt, M. & de Lorenzo, V. Synthetic bugs on the loose: containment options for deeply engineered (micro)organisms. *Curr. Opin. Biotechnol.* **38**, 90–96 (2016).
- Paracchini, V. et al. Molecular characterization of an unauthorized genetically modified *Bacillus subtilis* production strain identified in a vitamin B 2 feed additive. *Food Chem.* **230**, 681–689 (2017).
- Crook, O. M. et al. Analysis of the first genetic engineering attribution challenge. *Nat. Commun.* **13**, 7374 (2022).
- Lee, J. W., Chan, C. T. Y., Slomovic, S. & Collins, J. J. Next-generation biocontainment systems for engineered organisms. *Nat. Chem. Biol.* **14**, 530–537 (2018).
- Ghaly, T. M. et al. Discovery of integrons in archaea: platforms for cross-domain gene transfer. *Sci. Adv.* **8**, eabq6376 (2022).
- Hall, R. J., Whelan, F. J., McInerney, J. O., Ou, Y. & Domingo-Sananes, M. R. Horizontal gene transfer as a source of conflict and cooperation in prokaryotes. *Front. Microbiol.* **11**, 1569 (2020).
- Bruetschy, C. The EU regulatory framework on genetically modified organisms (GMOs). *Transgenic Res.* **28**, 169–174 (2019).
- Mitchell, L. A. et al. Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. *Science* **355**, eaaf4831 (2017).
- Gibson, D. G. et al. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**, 52–56 (2010).
- Tellechea-Luzardo, J. et al. Linking engineered cells to their digital twins: a version control system for strain engineering. *ACS Synth. Biol.* **9**, 536–545 (2020).
- Foster, P. L., Lee, H., Popodi, E., Townes, J. P. & Tang, H. Determinants of spontaneous mutation in the bacterium *Escherichia coli* as revealed by whole-genome sequencing. *Proc. Natl. Acad. Sci.* **112**, e5990–e5999 (2015).
- Lee, H., Popodi, E., Tang, H. X. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci.* **109**, e2774–e2783 (2012).
- Leon, D., D’Alton, S., Quandt, E. M. & Barrick, J. E. Innovation in an *E. coli* evolution experiment is contingent on maintaining adaptive potential until competition subsides. *PLoS Genet.* **14**, e1007348 (2018).
- Reeves, P. R. et al. Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. *PLoS One* **6**, e26907 (2011).
- Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).
- Liss, M. et al. Embedding permanent watermarks in synthetic genes. *PLoS One* **7**, e42465 (2012).
- Wang, K., de la Torre, D., Robertson, W. E. & Chin, J. W. Programmed chromosome fission and fusion enable precise large-scale genome rearrangement and assembly. *Science* **365**, 922–926 (2019).
- Heider, D. & Bamekow, A. DNA-based watermarks using the DNA-Crypt algorithm. *BMC Bioinforma.* **8**, 176 (2007).
- Hafeez, I., Khan, A. & Qadir, A. DNA-LCEB: a high-capacity and mutation-resistant DNA data-hiding approach by employing encryption, error correcting codes, and hybrid twofold and fourfold codon-based strategy for synonymous substitution in amino acids. *Med. Biol. Eng. Comput.* **52**, 945–961 (2014).
- Houghton, D. & Balado, F. BioCode: two biologically compatible Algorithms for embedding data in non-coding and coding regions of DNA. *BMC Bioinforma.* **14**, 121 (2013).
- Mueller, S. & On, D. N. A. Signatures, Their dual-use potential for GMO counterfeiting, and a cyber-based security solution. *Front. Bioeng. Biotechnol.* **7**, 189 (2019).
- Lin, W. *Digital Signature* (ed. Mulder, V., Mermoud, A., Lenders, V., & Tellenbach, B.) 77–81 (Springer, 2023).
- Yin, H. L. et al. Experimental quantum secure network with digital signatures and encryption. *Natl. Sci. Rev.* **10**, 11 (2023).
- Yuan, T. L., Wang, Y. L., Xu, K., Martin, R. R. & Hu, S. M. Two-layer QR codes. *IEEE Trans. Image Process.* **28**, 4413–4428 (2019).
- Polash, M. M. A., Newton, M. A. H. & Sattar, A. Constraint-based search for optimal Golomb rulers. *J. Heuristics.* **23**, 501–532 (2017).
- Memarsadeghi, N. NASA computational case study: golomb rulers and their applications. *Comput. Sci. Eng.* **18**, 58–62 (2016).
- Spee, W. J. P. & Weber, J. H. Bounds on the maximum cardinality of indel and substitution correcting codes. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* **10**, 349–358 (2024).
- Edelaar, P. & Bolnick, D. I. Non-random gene flow: an underappreciated force in evolution and ecology. *Trends Ecol. Evol.* **27**, 659–665 (2012).
- Mueller, S., Jafari, F. & Roth, D. A covert authentication and security solution for GMOs. *BMC Bioinf.* **17**, 389 (2016).
- Grozdanov, L. et al. Analysis of the genome structure of the nonpathogenic probiotic *Escherichia coli* strain Nissle 1917. *J. Bacteriol.* **186**, 5432–5441 (2004).
- Meyer, A. et al. Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* **590**, 284–289 (2021).
- Stephens, Z. D. et al. Big Data: astronomical or genomics? *PLoS Biol.* **13**, e1002195 (2015).
- Wu, Y. et al. Efficient In vitro full-sense-codons protein synthesis. *Adv. Biol.* **6**, 2200023 (2022).
- Kuhlman, T. E. & Cox, E. C. Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Res.* **38**, e92 (2010).
- Ayyub, S. A., Dobriyal, D. & Varshney, U. Contributions of the N- and C-terminal domains of initiation factor 3 to its functions in the fidelity of initiation and antiassociation of the ribosomal subunits. *J. Bacteriol.* **199**, e00051 (2017).
- Li, Q. et al. A modified pCas/pTargetF system for CRISPR-Cas9-assisted genome editing in *Escherichia coli*. *Acta Biochim. Biophys. Sin.* **53**, 620–627 (2021).
- Wang, Z. G. et al. Genome signature for self-authenticating genomic materials in *Escherichia coli*. *Zenodo* <https://doi.org/10.5281/zenodo.15314176> (2025).

Acknowledgements

We would like to express our gratitude to Y.Z. for insightful discussions. Additionally, thanks to X.W. and M.Q. for their unwavering support throughout the project. This work was supported by grants from the National Key R&D Program of China (Grant No. 2020YFA0712100, 2020YFA0712104).

Author contributions

Z.G.W. and J.S.C. contributed equally to this work. Z.G.W. and Q.L. conducted biochemical experiments and data analysis, prepared the figures, and contributed to the manuscript preparation. Z.G.W. and J.J.L. contributed to the genome editing experiments; J.S.C., H.X., J.W.L., C.G., C.Y.J., and Y.Y.L. wrote the code and conducted all the computation experiments. Y.G.K. contributed to the population genetics study design, figure preparation, and interpretation of the results; H.Q., J.S.C., Y.G.K., and Z.G.W. prepared the manuscript together. H.Q. and J.S.C. supervised the whole project, led the development of the algorithm and designed all experiments.

Competing interests

The authors declare the following competing interests: J.W.L. participated in this work while studying at Wuhan University and interning at Jinuochuangwu (Wuhan) technology Ltd., and then continued his involvement at Tianjin University. J.S.C., H.Q., H.X., Y.Y.L., and J.W.L. are named as inventors on a granted patent held by Wuhan University, serial no. CN202110102256.3. J.S.C., H.X., J.W.L., and Y.Y.L. are named as inventors on a patent application held by Wuhan University, serial no. CN202210938867.6. J.S.C., H.Q., C.Y.J., H.X., Y.Y.L., and J.W.L. are named as inventors on a patent application held by Jinuochuangwu (Wuhan) technology Ltd., serial no. CN202410420104.1. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at

<https://doi.org/10.1038/s42003-025-08171-z>.

Correspondence and requests for materials should be addressed to Jingsong Cui, Yonggang Ke or Hao Qi.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary handling editor: T.G.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025