A Fuzzy sequencer for rapid DNA fragment counting and genotyping

Wenxiong Zhou[1,#], Li Kang[1,2,#], Shuo Qiao[1,2,#], Haifeng Duan[3,#], Chenghong Yin[4,#] , Zhizhao Liao[1], Mingchuan Tang[1], Ruiying Zhang[3], Lei Li[3], Lei Shi[3], Meijie Du[5,6], Yipeng Wang[4], Wentao Yue[4], Yan Xiao[7], Lin Di[1,8], Xiannian Zhang[1], Yuhong Pang[1], Mingkun Li[9], Lili Ren[7], Jianbin Wang[5,*], Zitian Chen[3,*], and Yanyi Huang[1,2,8,*]

Affiliations:

[1] Biomedical Pioneering Innovation Center (BIOPIC), Beijing Advanced Innovation Center for Genomics (ICG), Peking-Tsinghua Center for Life Sciences, School of Life Sciences, Peking University, Beijing 100871, China

[2] College of Chemistry and Molecular Engineering, and Beijing National Laboratory for Molecular Sciences, Peking University, Beijing 100871, China

[3] Cygnus Biosciences, No.29 Kechuang 7th Street, BDA, Beijing 100176, China

[4] Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Beijing 100026, China

[5] School of Life Sciences, Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing 100871, China

[6] College of Life Science, Henan Normal University, Henan 453007, China

[7] Institute of Pathogen Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[8] Institute for Cell Analysis, Shenzhen Bay Laboratory, Guangdong 518132, China

[9] Beijing Institute of Genomics, Chinese Academy of Sciences, and China National Center for Bioinformation, Beijing, China

*Correspondence to: jianbinwang@tsinghua.edu.cn (J.W.), czitian@gmail.com (Z.C.), yanyi@pku.edu.cn (Y.H.).

# These authors contributed equally to this work.

Abstract: High-throughput sequencing technologies generate a vast number of DNA sequence reads simultaneously, which are subsequently analyzed using the information contained within these fragmented reads. The assessment of sequencing technology relies on information efficiency, which measures the amount of information entropy produced per sequencing reaction cycle. In this study, we propose a fuzzy sequencing strategy that exhibits information efficiency more than twice of currently prevailing cyclic reversible terminator sequencing methods. To validate our approach, we developed a fully functional and high-throughput fuzzy sequencer. This sequencer implements a highly efficient fluorogenic sequencing-by-synthesis chemistry and underwent testing across various application scenarios, including copy-number variation detection, noninvasive prenatal testing, transcriptome profiling, mutation genotyping, and metagenomic profling. Our findings unequivocally demonstrate that the fuzzy sequencing strategy outperforms existing methods in terms of information efficiency and delivers accurate resequencing results with faster turnaround times.

One Sentence Summary: A fuzzy sequencer can exceed current limit of information efficiency of DNA sequencers for resequencing applications.

High-throughput 'next-generation' sequencing (NGS)[1-2] has been extensively employed for precise quantification of nucleic acids at a whole genome, exome, or transcriptome scale[3-4]. Among the primary applications of NGS is resequencing, which encompasses copy-number variation (CNV) assessment[5-6], transcriptome profiling (RNA-seq)[7-8], and noninvasive prenatal testing (NIPT)[9] that rely on counting sequence-based DNA fragments. The accuracy of counting largely depends on the effectiveness of extracting information from DNA fragments and accurately mapping it onto reference sequences. Although mapping algorithms continue to evolve, the sequencing steps themselves have remained largely unchanged for several years.

Next-generation sequencing relies on the incorporation of nucleotides to encode information, and the efficiency of information encoding varies across different systems. The majority of high-throughput DNA sequencers utilize sequencing-by-synthesis (SBS) chemistries, with signal generation schemes falling into two major variations: single-nucleotide addition (SNA) or cyclic reversible terminator (CRT)[10-11]. SNA chemistry produces natural DNA duplex but suffers from relatively low intrinsic information efficiency, averaging at 0.67 nt/cycle or 1.33 bit/cycle for encoding random sequences (Supplementary Fig. 1a). In contrast, multi-step CRT chemistry, although producing scar-containing duplex, offers 2 bit/cycle intrinsic information efficiency with an exact 1 nt/cycle extension rate (Supplementary Fig. 1b). Enhancing the information entropy encoding efficiency becomes a clear avenue for improving sequencing performance.

While the information entropy embedded in the nucleotide sequence plays a decisive role in mapping a read to a reference, mapping accuracy is also influenced by the corresponding region of the reference. One example is the mate-pair reads which, compared with conventional paired-end reads, are more mappable because of larger reference target size instead of higher information entropy. This effect becomes particularly evident in genomes with a high proportion of repeat sequences. From the perspective of the sequencer, the reference-side or extrinsic information entropy in the reads is equally significant but often overlooked. This oversight is primarily due to prevalent DNA sequencers providing explicit sequences of four bases, where the intrinsic information efficiency is equal to the extrinsic information efficiency, despite the differences in efficiency between SNA and CRT. Enhancing the encoding

efficiency of extrinsic information entropy could also lead to improvements in sequencing technology.

Decoupling the intrinsic and extrinsic information entropy within the same read suggests a change in sequencing read formatting. When aligning a DNA fragment sequence to reference sequences, the alignment process relies not on the string format, but rather on the information entropy contained within it. The reference genome is typically represented by the conventional and natural four base-letters (A, C, G, and T), but the read itself does not necessarily need to adhere to the same formatting. As long as the information is sufficiently rich for unique alignment, alternative formatting can be employed. Although sequencing by oligonucleotide ligation and detection (SOLiD) is being phased out in the market due to low-efficiency chemistry and very short read lengths, it demonstrated the possibility of sequencing with alternate formatting through a color-space strategy[12].

From a sequencing strategy standpoint, achieving higher information efficiency (> 2 bit/cycle) for both intrinsic and extrinsic information entropy could revolutionize the entire sequencing field. In this report, we introduce a fuzzy SBS strategy that sacrifices explicit sequence information in exchange for improved information encoding performance. The fuzzy SBS sequencing strategy can be implemented by incorporating more than one type of nucleotide per reaction cycle (Fig. 1a-c, Supplementary Fig. 1c). As a proof of concept, we developed a high-throughput sequencer to execute a dual-base flowgram fuzzy SBS strategy. We demonstrated that SNA chemistry can be re-engineered to accommodate the fuzzy decoding approach, thereby surpassing the limitations of information efficiency and enabling rapid resequencing and counting.

Fuzzy SBS can be implemented using five different flowgrams, allowing the introduction of nucleotides beyond the limitation of a single type per cycle (see Supplementary Text 2.2). Briefly, *BitSeq* employs a dual-base flowgram, where a combination of two nucleotides is alternately added in each cycle. For instance, K (a mix of G/T) is added during odd cycles, while M (a mix of A/C) is added during even cycles (see Fig. 1a). BitSeq generates fuzzy sequences since it's not possible to distinguish the signal associated with the incorporation of the two nucleotides in the mix. However, if the two nucleotides in the mix can be distinguished, such as being

labeled by different fluorophores, then the information from each cycle can be elevated. This modified approach is referred to as *SuperBitSeq* (see Fig. 1b,c). Similarly, by employing three-base flowgrams, we can have *TritSeq*, *SuperTritSeq*, and *HyperTritSeq*, depending on the distinguishability of the three nucleotides in the mix introduced in each cycle. All of these fuzzy SBS approaches can achieve intrinsic and/or extrinsic information efficiency beyond 2 bits per cycle. In this paper, the focus will primarily be on BitSeq and SuperBitSeq as they strike a reasonable balance between information efficiency and signal dynamic range per reaction cycle.

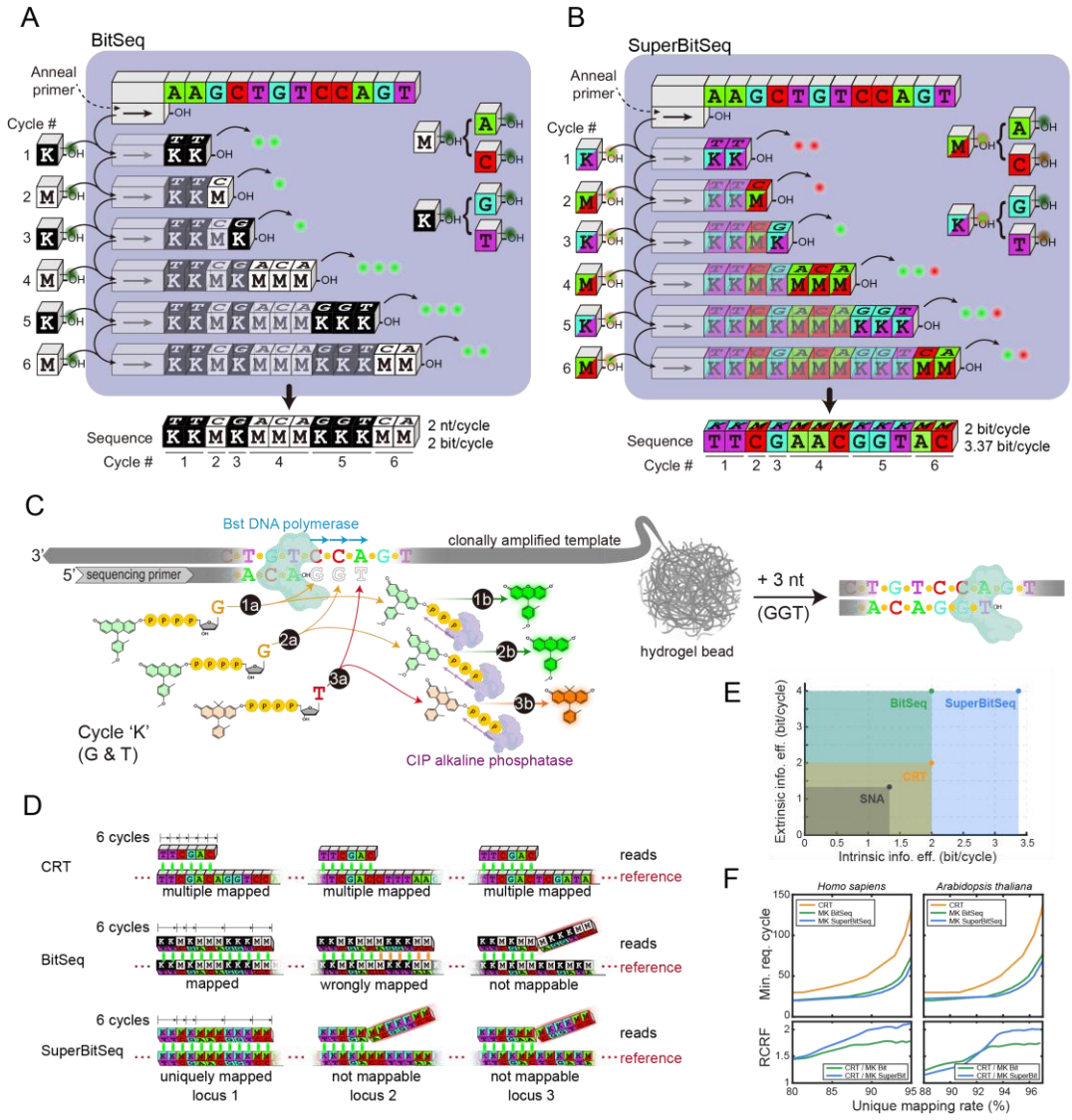

Fig. 1. Fuzzy sequencing is efficient in information and read length. (A) Schematic of BitSeq. (B) Schematic of SuperBitSeq. (C) Schematic of fluorogenic sequencing chemistry for BitSeq and SuperBitSeq. (D) Longer read length or more information enable more accurate genome mapping. (E) The information efficiency of different sequencing technologies. Compared with SNA or CRT, BitSeq and SuperBitSeq show higher information efficiency that can be illustrated as the product of intrinsic and extrinsic components. (F) Simulated unique mapping rate of DNA reads by fuzzy sequencing and CRT for different genomes. For mapping the specific proportion of the genome, both BitSeq and

SuperBitSeq need much shorter minimally required sequencing cycles, and this advantage can also be reflected by reaction cycle reduction factor (RCRF) that defined by the ratio of minimally required cycle numbers between technologies.

We conducted *in silico* analysis to validate the alignment accuracy of the fuzzy SBS encoding strategy. In the BitSeq, DNA sequences consist of six degenerate bases: M (A/C), K (G/T), R (A/G), Y (C/T), W (A/T), and S (C/G). Each experimental run utilizes only two complementary degenerate bases, resulting in three possible forms of bit sequences (M-K, R-Y, or W-S). To map the BitSeq and SuperBitSeq signals, we encode them as sequence reads that can be aligned using conventional mappers such as Burrows-Wheeler Aligner (BWA)[13-14] and Bowtie2 (Supplementary Text 2.1, Supplementary Table 1-4)[15-16]. For our analysis, we randomly selected 1×10^7 positions from the human reference genome and 1×10^6 positions from the *Arabidopsis* reference genome. We simulated the BitSeq and CRT sequencing from these positions by extending the nucleotide strings with certain cycles (Supplementary Fig. 2), and then aligned the simulated reads back to their respective reference genomes. We observed that reads generated with more sequencing cycles provided greater information, resulting in a higher unique mapping rate (UMR) for both CRT and BitSeq approaches. Notably, SuperBitSeq consistently exhibited the highest UMR (Fig. 1d-f, Supplementary Table 5-56). Additionally, the UMR of BitSeq with the M-K form was consistently higher than that of CRT and BitSeq with R-Y form. This preference for the M-K form may be attributed to the occurrence of transition base substitutions following gene duplication during evolution[17]. Consequently, we selected the M-K form for the rest of our study.

We constructed a laboratory prototype sequencer for conducting high-throughput BitSeq (Fig. 2a,b). The bit sequencer comprises a microfluidic flow cell for implanting clonally amplified DNA fragments for sequencing and an imaging system for signal acquisition (Supplementary Fig. 3-5). We implemented fluorogenic SBS chemistry to naturally align with the BitSeq flowgram[18-20]. Bst polymerase is capable of recognizing and incorporating Tokyo Green (TG) terminal phosphate-labeled fluorogenic nucleotides (TG-TPLFNs, or TG-dN4P) into nascent strands in the presence of Mn^{2+} [21]. We constructed DNA fragments into a library with two specific end adapters for clonal amplification on hydrogel microbeads carrying amplification primers. The DNA fragments hybridized on the beads were then amplified into clones

through emulsion PCR[22] or surface-tethered recombinase polymerase amplification (RPA). Each clone typically contained approximately 30,000 copies of DNA for the SBS reaction.

To load the clonally amplified hydrogel microbeads for sequencing, we created a microfluidic flow cell (Fig. 2c) containing 28 million 30-femtoliter microwells (Fig. 2d,e). Each cycle of the SBS reaction can extend one or more bases and produce a corresponding amount of fluorophore. The microwells can be sealed to prevent signal cross-talk due to fluorophore diffusion. These microwells were fabricated by selective wet-etching of fiber-optic plates (FOP), allowing the fluorogenic signal to propagate through the fiber waveguide to the outer surface of the flow cell for imaging (Fig. 2f). The inner surface of the microwells was made hydrophilic, while the adjoining surface between microwells was topologically coated to be hydrophobic, enabling the highly parallel sealing of the SBS reactions using oil (Supplementary Fig. 6,7).

The oil-based microwell sealing is robust, reversible, and compatible with fluorogenic SBS reaction conditions. In each cycle of the fluorogenic SBS reaction, we first filled the microwells with reaction buffer containing polymerase and two fluorogenic unnatural nucleotides at low temperature of 4 ℃. We then sealed the microwells by steadily flowing in fluorinated oil (Fig. 2f). Next, we quickly raised the temperature of the flowcell to 65℃ to initiate the synthesis reaction, followed by cooling down to 25℃ for fluorescence image acquisition. After that,we removed the sealing by flowing in isopropanol and performed an aqueous wash to reset the reaction conditions for the next cycle. The acquired images were aligned to identify each microwell, which was then registered for tracking throughout all reaction cycles (Fig. 2g).

Defected regions resulting from local sealing or unsealing failures were eliminated before further processing (Fig. 2h). Empty microwells and microwells with multiple beads were identified and filtered out for further analysis. The fluorescence signal of each monoclonal microwell was extracted from the images after noise reduction and background correction. Due to the asynchronization of the molecular reaction within each clone, the intensity-cycle series could not be directly converted into bit sequences without a meticulous dephasing process. To supervise the dephasing algorithm, we spiked a standard DNA library, constructed from the lambda

phage genome, into the samples to be sequenced together. As a result, the sequenced reads could be easily classified into two categories: standard dots (SD) and library dots (LD), based on the characteristic signal profiles of the first few cycles. We employed the dual-base flowgram dephasing model to fit the SD signals and obtain the parameters[23], which were then applied to LD signals to construct a flux matrix for dephasing correction and convert the signals into bit sequences.

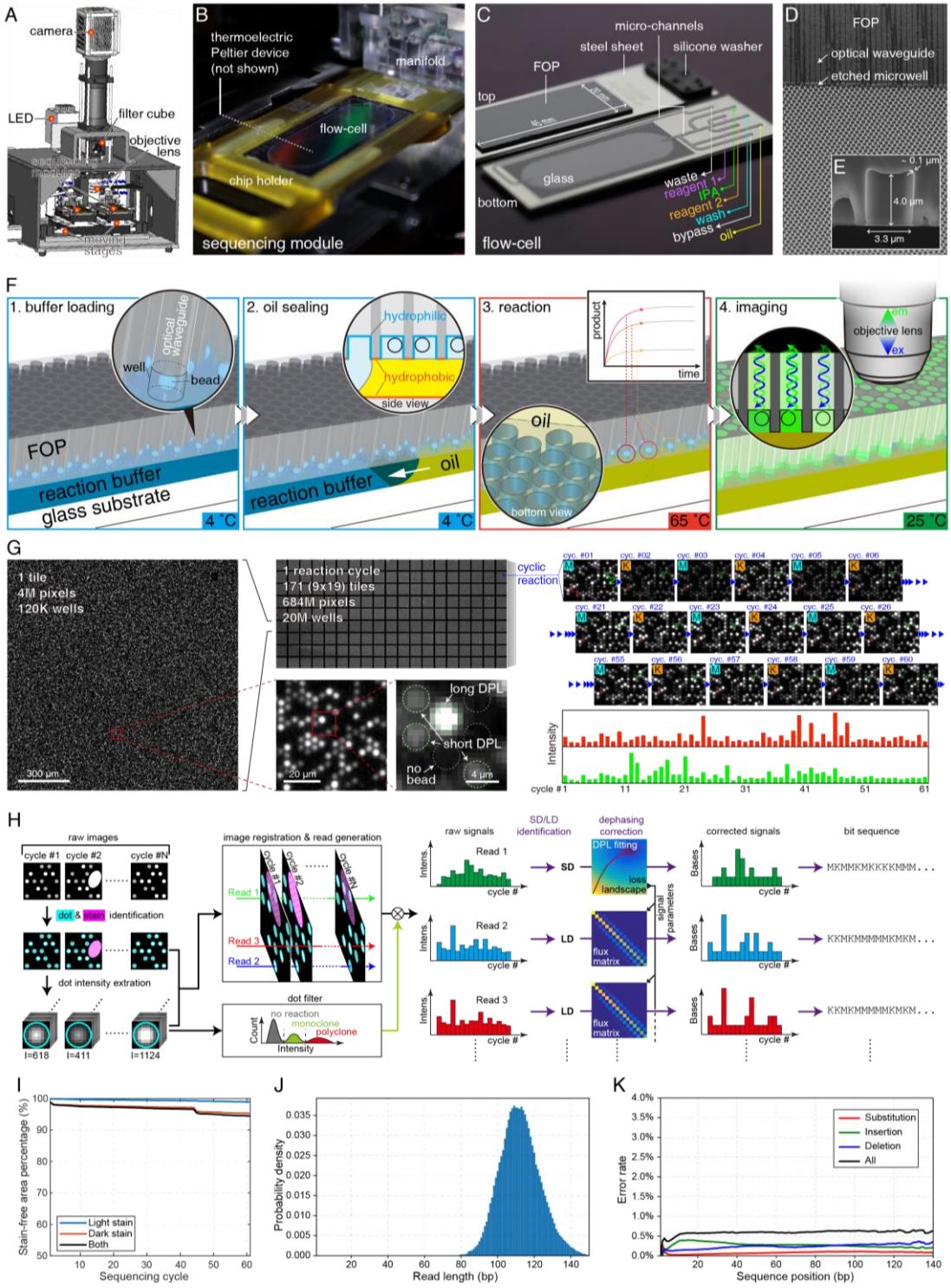

Fig. 2. Implementation of BitSeq. (A) The design of a prototype for high-throughput fluorogenic fuzzy sequencing. (B) The photo of a sequencing module, which consists of a flow-cell connected with a manifold for liquid routing, and the flow-cell is placed on a thermoelectric Peltier device for temperature control. (C) The photo of a sequencing flow-cell, which is made of a glass slide and a fiber optical plate (FOP) and a flow channel in between with shape defined by laser-cut double side adhesive. (D) Scanning electron microscopy image of the microwell array made of selective etching of FOP. These femtoliter microwells are placed inward the flow chamber. (E) The size of a microwell. (F) The running procedure of one sequencing cycle. Firstly, the micro-beads with clonal amplified templates for sequencing are introduced into the microwells and tethered on the inner surface of the microwells. Then, the reaction buffer is primed into the flow chamber and all microwells are sealed by oil flowing into the chamber to separate each microwell from cross-talk. The SBS reaction is triggered by elevation of the temperature. The fluorogenic product is proportional to the bases that incorporated during each reaction cycle. When the reaction is finished, the flow-cell is cooled to take fluorescent images. (G) One typical field of view from 171 tiles a cycle and the procedure of fluorescence intensity extraction. The microwells that contained beads will generate fluorescence after reaction of each cycle, and the higher intensity indicates longer degenerate polymer length (DPL) in that reaction cycle. Each microwell is addressable and indexed to produce an intensity series that can be later deduced into bit sequences. (H) Schematic of the data processing pipeline. Defects are filtered and the intensity of each microwell at each reaction cycle is extracted and further filtered to keep monoclones for dephasing correction. The correction parameters are determined by fitting the signal to the standard dots (SD) which have been spiked into the sequencing library. (I) >90% of the area remains stain-free after 61 cycles of sequencing. (J) The read length distribution of bit sequences. (K) The error rate of BitSeq.

We initially tested our prototype sequencer on a lambda phage genomic library (Supplementary Fig. 8, Supplementary Table 57). After 61 cycles of repeated sealing and sequencing, more than 90% of the chip area remained unaffected by stains (Fig. 2i, Supplementary Fig. 9). In a single run, we generated 6,578,364 reads with an average length of 112.2 bp (Fig. 2j, Supplementary Fig. 10-15). By mapping the bit sequences back to the reference genome, we successfully aligned 99.11% of the reads, which had an average length of 111.6 bp. The error rate of BitSeq was found to be 0.39% in the first 10 bp and 0.62% in the 130-140 bp region, which were acceptably low. This indicates the feasibility of accurate read mapping and DNA fragment counting (Fig. 2k).

Using the same aligning strategy, we then evaluated the performance of the BitSeq approach in four widely used resequencing applications: copy number variation (CNV) identification, noninvasive prenatal testing (NIPT), transcriptomic analysis (RNA-seq), and metagenomics sequencing (mNGS).

We sequenced two genomic DNA samples collected from a normal male and a Down syndrome male patient at low coverage (0.1X). The sequencing coverage across the whole genome was mostly uniform for both samples (Supplementary Fig. 16,17) except for the sex chromosomes which were half of that of autosomes. The additional copy of chromosome 21 in the Down syndrome patient was accurately

counted through BitSeq, as validated by Ion Torrent sequencer as well. When down-sampled to the same number of reads, data from BitSeq and ion torrent showed approximate variations as indicated by the median of the absolute values of all pairwise differences (MAPD).

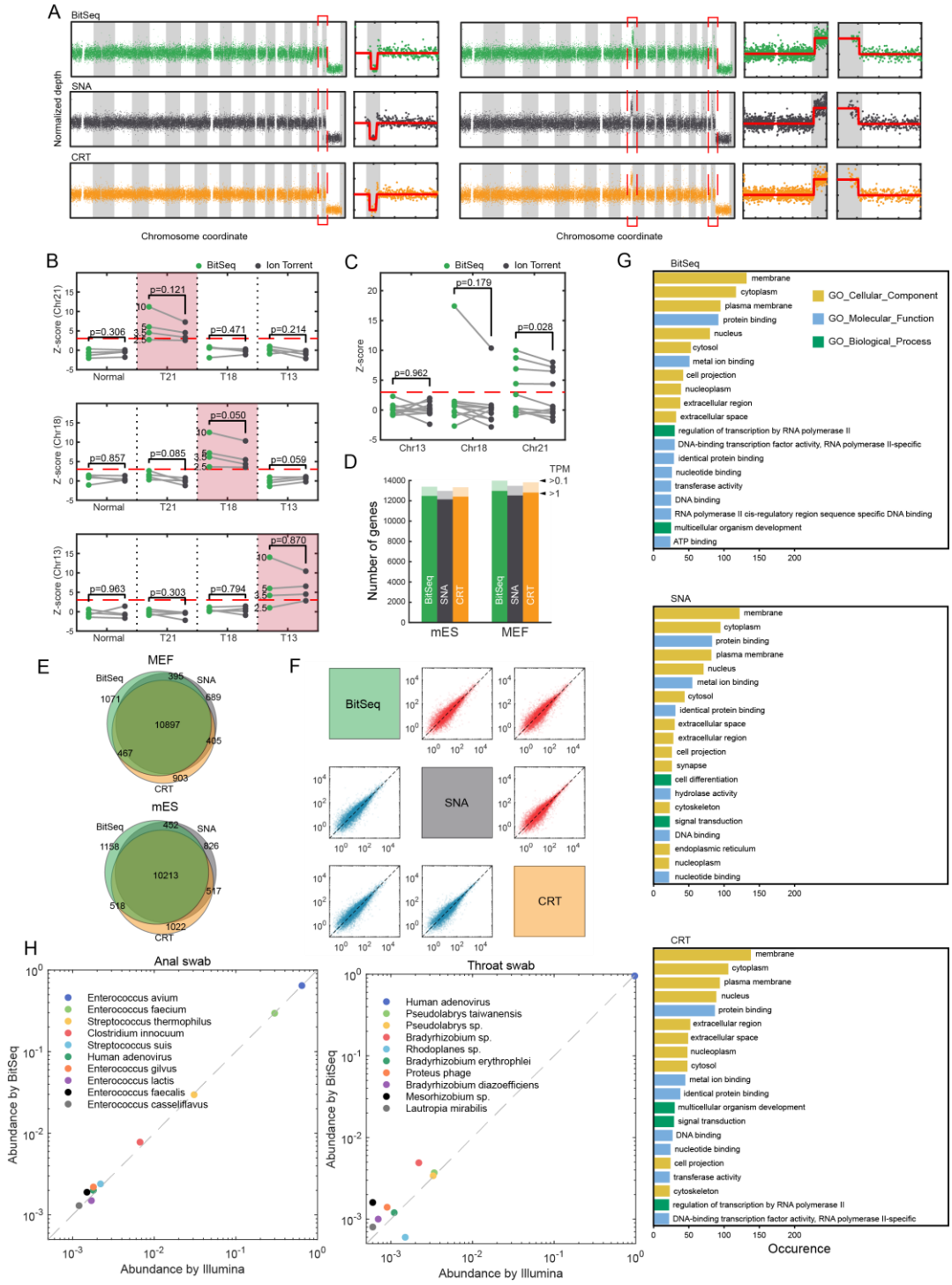

Fig. 3. Comparison of BitSeq with commercial sequencers in resequencing. (A) CNV of a DiGeorge syndrome patient (left) and a development-delayed patient (right). (B) Z-scores of normal and mock trisomy NIPT ccfDNA samples. (C) Z-scores of single-blinded NIPT ccfDNA samples. The red dash in B-C indicates the threshold of 3. (D) Number of genes detected using different thresholds

for RNA-seq data of mES and MEF cells. (E) Venn diagram of TPM>1 genes. (F) TPM comparison of MEF (blue) and mES (red) cell line between BitSeq and commercial sequencers. (G) Gene Ontology analysis of genes that were only detected by each sequencer in MEF. (H) Microbe abundance comparison between BitSeq and Illumina.

Besides aneuploidies, BitSeq can also faithfully detect the clinically relevant small size CNVs. We used BitSeq to identify the 2.9-Mb deletion at Chr22.q11.21 that was associated with DiGeorge syndrome from a patient, and two CNVs (20.8-Mb gain at Chr 11.q and 10.0-Mb gain at Chr 22.q11) from another patient with development delay (Fig. 3a) with high concordance with other conventional sequencing approaches. Besides, BitSeq also showed similar MAPD around 0.03 compared with other approaches (Supplementary Table 58).

We then tested the effectiveness of BitSeq on NIPT, a widely used resequencing application in prenatal diagnostics that requires high accuracy of DNA counting. We started the test with mock samples prepared by spiking fragmented genomic DNA from trisomy patients into circulating cell-free DNA (ccfDNA) of an unpregnant woman with different ratios. These 12 mock NIPT samples had trisomy in Chr21, Chr18 and Chr13, respectively, and the mixing proportion was 10%, 5%, 3.5% and 2.5%. We also collected 28 NIPT ccfDNA samples from pregnant women with normal fetuses. We applied BitSeq to all the samples and used 24 out of the 28 normal samples as the control set (the other 4 used as the negative test set) to calculate the Z-score for each NIPT sample (Fig. 3b). Using Z=3 as the cut-off, BitSeq shows comparable power to identify the positive trisomy NIPT samples when the mock fetal DNA fraction is not less than 3.5%. We further tested 10 single-blinded true NIPT samples, 5 of which were tested positive based on the same cut-off using BitSeq (Fig. 3c). These results were confirmed by parallel tests by Ion Torrent sequencers as well as the clinical records (Supplementary Fig. 18,19, Supplementary Table 58).

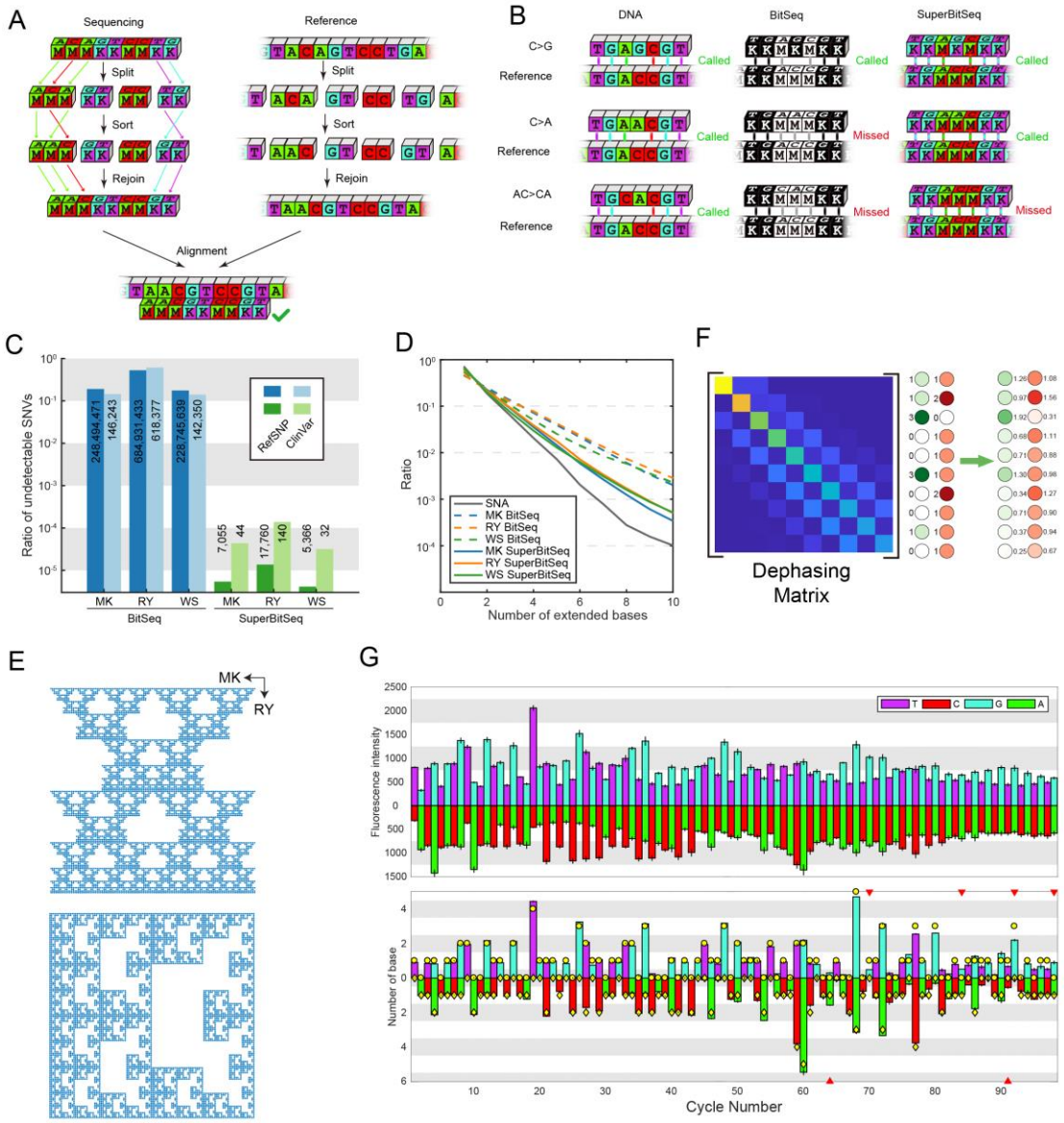

Fig. 4. Properties of SuperBitSeq. (A) Encoding strategy of SuperBitSeq in M-K form as an example. (B) Distinguishable and indistinguishable SNV types of BitSeq and SuperBitSeq. (C) Ratio of indistinguishable SNV by BitSeq and SuperBitSeq. (D) Frequencies of extended base number of different flowgrams. (E) Fractal of encoded SuperBitSeq signals. (F) Schematic of dephasing correction of SuperBitSeq signals. (G) The fluorescence intensities (top) and their dephasing corrected signals (bottom) of a single-template SuperBitSeq experiment.

We next verified that BitSeq could be used for RNA-seq. Using mouse embryonic fibroblasts (MEF) and embryonic stem (mES) cells, BitSeq could provide almost identical results as Ion Torrent or Illumina sequencers did. The number of genes detected was comparable, and the gene expression levels were correlated with high linearity between these approaches (Fig. 3d-f). We also checked the Gene Ontology (GO) terms of genes only detected by one sequencer and found no specific preferences in between (Fig. 3g, Supplementary Fig. 19b-d).

Combined with metagenomic sequencing, BitSeq may identify pathogens in acute infections timely. We tested the capability of BitSeq in metagenomics by sequencing one anal swab and one throat swab from a previously reported pediatric patient suffered from multi-organ abscesses[24]. The 10 species types with the highest abundance identified by BitSeq and Illumina sequncer are identical. And their abundances, ranging from less than 0.1% to over 95%, are also close (Supplementary Fig. 3h).

All these experiments show that BitSeq is fully capable for those applications based on DNA fragment counting. We expect that other re-sequencing approaches that do not require single-base resolution but focus on pattern alterations such as bisulfite sequencing for DNA methylation[25-27] or Hi-C for chromatin topology analysis[28-29] are also possible with BitSeq. Thanks to its high information and long extension length per cycle as well as the fast fluorogenic sequencing chemistry, BitSeq would enable more efficient DNA fragment counting that is valuable in both basic science research and clinical testing.

We further considered the possibility to extend fuzzy sequencing strategy from BitSeq to SuperBitSeq by labeling the two different nucleotides in each reaction cycle with distinguishable fluorophores. Like BitSeq, we also designed an encoding strategy for SuperBitSeq that is compatible with prevailing short reads mappers like BWA (Fig. 4a). The differential nucleotide labeling in SuperBitSeq will elevate the information efficiency embedded in the SBS process, enabling the detection of significantly more SNVs that cannot be identified through BitSeq (Fig. 4b). For example, in BitSeq the A>C/G>T mutations cannot be identified in the M-K form, nor the A>G/C>T mutations in the R-Y form. In SuperBitSeq they can all be well captured, except for only specific indistinguishable base swap events such as the AC>CA in the M-K form. To evaluate the potential of SuperBitSeq for SNP/SNV detection, we filtered dbSNP (build 155)[30] and ClinVar (20210908)[31] to exclude items with N, and obtained 1,297,977,577 and 1,009,591 items respectively. About 17-52% of known human SNP/SNVs cannot be distinguished by BitSeq, whereas the undetection rates are ~0.01% or even lower for SuperBitSeq (Fig. 4c). This means that SuperBitSeq can *de facto* detect almost all known SNP/SNVs. More importantly, in each reaction cycle the fluorescence signal released by extended bases can be split

into two channels, resulting in reduced requirement for signal detection dynamic range and enhanced accuracy in detecting long copolymers (Fig. 4d, Supplementary Fig. 20). Interestingly, we also found that the encoded superbit sequences can form a fractal when mapped the points in a square (Fig. 4e, Supplementary Text 2.4).

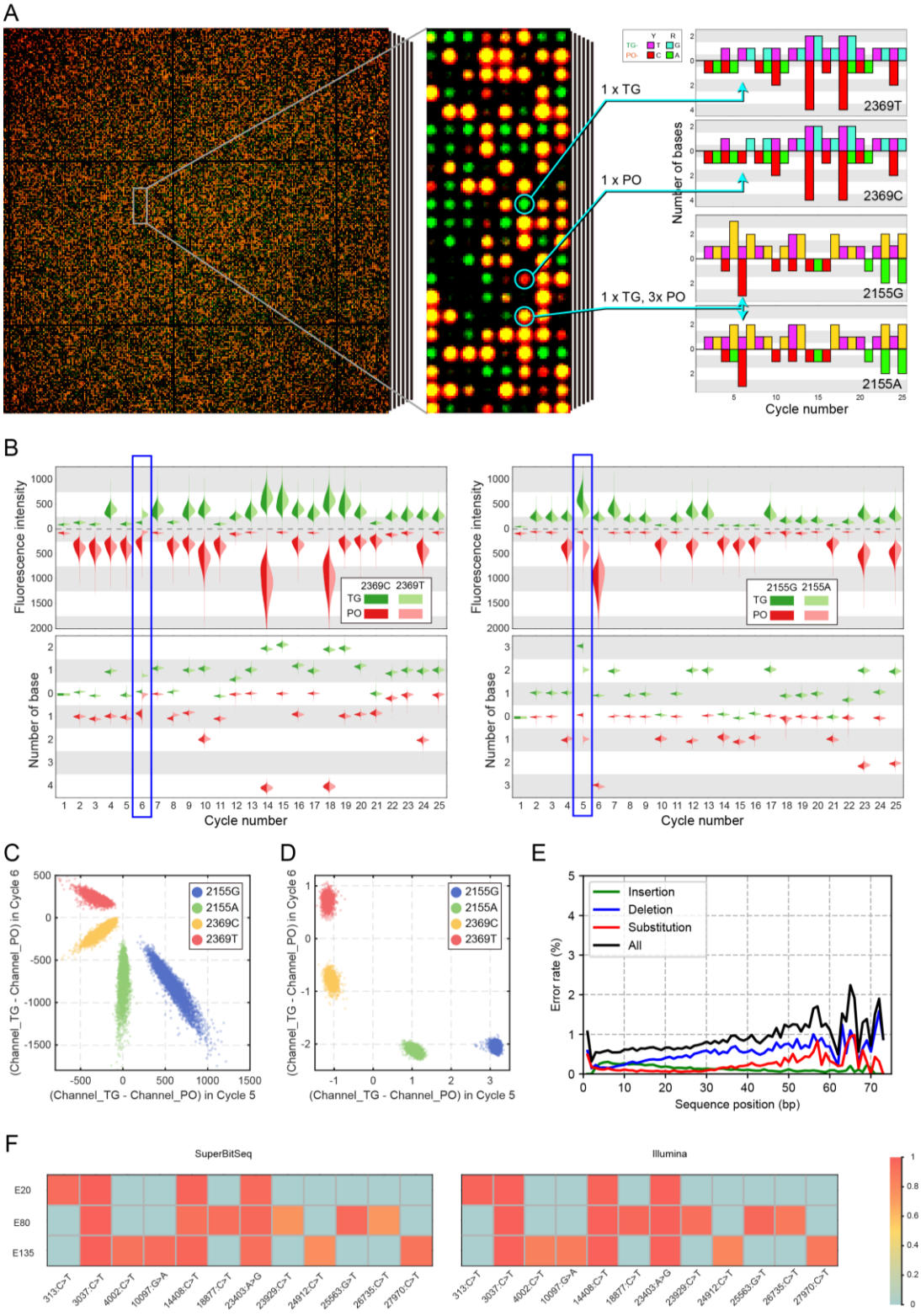

Fig. 5. High-throughput SuperBitSeq identifies single-nucleotide variants (SNVs). (A) A merged fluorescence image of green (TG-labeled T or G) and red (PO labeled C or A) channels, from which the signal intensities are extracted. The image was associated with Cycle 6 in this specific experiment and yellow dots represent the co-existing of green and red-signal. The signal series of four different DNA templates are shown in the right. (B) Violin plot of fluorescence intensity and dephasing corrected signals of four different DNA templates. The blue boxes show the distinguishable signals that differentiate the SNVs using SuperBitSeq. (C-D) The four different DNA templates can be distinguished by the difference of fluorescence intensities (C) or dephasing-corrected signals (D) in Cycle 5-6. (E) Error rate of SuperBitSeq in SARS-CoV-2 samples. (F) Allele frequencies in SARS-CoV-2 samples detected by SuperBitSeq and an Illumina sequencer.

To implement SuperBitSeq experimentally, we synthesized a novel fluorophore, Peking Orange (PO), which was designed to exhibit excellent fluorogenic properties (on/off ratio ~ 2000) when terminally labeled on the tetraphosphate nucleotides (PO-TPLFNs, or PO-dN4P, Supplementary Fig. 21,22,23a). PO was used to pair with TG for labeling two different nucleotides, respectively, in each reaction cycle. PO and TG have similarly high fluorescence quantum efficiency but distinct excitation and emission spectra to avoid signal cross-talk (Supplementary Fig. 23b). PO-dN4P has great photo and thermal stability, and showed excellent signal-to-base linearity upon nucleotide extension reaction (Supplementary Fig. 23c-e). The same dephasing algorithm we optimized for TG-dN4P substrates could be seamlessly adopted to PO-dN4P when sequencing a single template (Supplementary Fig. 24). According to our previous *in silico* simulation, dephasing correction for SuperBitSeq could be built upon the algorithm of BitSeq by separating two channels for independent correction with the same dephasing parameters (Fig. 4f). We also experimentally confirmed this strategy using the same single template for SuperBitSeq (Fig. 4g, Supplementary Fig. 25).

We added a fluorescence channel to the bit sequencer and converted it to the superbit sequencer (Fig. 5a, Supplementary Fig. 26), while the image processing and signal extraction pipeline was largely adapted from before. We demonstrated the SNV detectability of SuperBitSeq by sequencing the G719S and T790M mutations of EGFR gene (Fig. 5b). The signal acquired from 9 imaging tiles (0.67 x 0.67 mm^2 each, 212,874 reads in total) could be clearly clustered into 4 groups (Fig. 5c,d), and each group represented one of the DNA templates with single nucleotide difference. With dephasing correction, the signals of these clusters clearly unveiled the significant difference at the specific reaction cycle, which is associated with the SNVs we targeted.

While BitSeq can only determine the pathogen species and abundances in metagenomic sequencing, SuperBitSeq allows fast pathogen genotyping and origin tracing, which is valuable in epidemic outbreaks. We sequenced 3 SARS-CoV-2 samples from a previously reported epidemiology study[32] for 26 cycles using SuperBitSeq. The total reaction time is only 42.6 min and we got about 200k reads. After mapping the SuperBitSeq reads to the reference genome, we restored the SNV information by separating the read and the reference into two semi-sequences, aligning individually, and then merging together (Supplementary Fig. 28). The experimental data demonstrated low error rate ($\approx 1\%$, Fig. 5e) and an average read length of 43 bp (Supplementary Fig. 29a). The SNVs called by SuperBitSeq are consistent with those called by an Illumina sequencer (Fig. 5f).

One reasonable extension of our information-richer SuperBitSeq is to combine three rounds of SuperBitSeq with orthogonal dual-base mixes for error correction code (ECC) sequencing, which may surpass the previously demonstrated high accuracy ECC. Our *in silico* simulation showed that compared to traditional ECC, error-correction using three rounds of superbit sequences has lower false negative rate in detecting potential sequencing errors under the same noise level (Supplementary Fig. 27). Both BitSeq and SuperBitSeq are specific forms of the fuzzy SBS strategy that encode DNA with high information efficiency for sequencing. More advanced forms such as TritSeq or SuperTritSeq are definitely worth of more expectation. However, the signal dynamic range of TritSeq would be too large to practice for fluorogenic chemistry, despite TritSeq would generate longer extension length and higher information efficiency than BitSeq. There are two major challenges that have not been well addressed in this work. First, the fundamental mathematical structure of the fuzzy coding of DNA is yet fully understood. For example, after transforming the bit sequences into binary faction number and then converting into decimal fraction numbers, every infinite long DNA sequence can be mapped to a point in the square region $[0,1) \times [0,1)$ (Fig. 4e), and formed fractal patterns for both SuperBitSeq with M-K and W-S forms. These two fractal patterns have identical Hausdorff dimension of ≈ 1.7716, although the nature of it is still unknown (Supplementary Text 2.4). Second, currently available fragment mapping algorithms and softwares, such as BWA or Bowtie2, were designed for recognizing natural base format but not the

fuzzy sequences. Although we have circumvented this problem by renaming the bit strings to be compatible with current softwares, a more general mapping tool that directly handles the nature of fuzzy sequence is still highly desirable.

References and Notes

1. Levy, S. E. & Myers, R. M. Advancements in next-generation sequencing. *Annu. Rev. Genomics. Hum. Genet.* **17**, 95-115 (2016).
2. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature*. **550**, 345-353 (2017).
3. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med*. **363**, 166-176 (2010).
4. Li, R. et al. A body map of somatic mutagenesis in morphologically normal human tissues. *Nature*. **597**, 398-403 (2021).
5. Li, X. et al. Improvement in the risk assessment of oral leukoplakia through morphology-related copy number analysis. *Sci. China. Life. Sci.* **64**, 1379-1391 (2021).
6. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet*. **16**, 172-183 (2015).
7. Wu, A. R., Wang, J., Streets, A. M. & Huang, Y. Single-cell transcriptional analysis. *Annu. Rev. Anal. Chem.* **10**, 439-462 (2017).
8. Di, L. et al. RNA sequencing by direct tagmentation of RNA/DNA hybrids. *Proc. Natl. Acad. Sci. U. S. A*. **117**, 2886-2893 (2020).
9. Wong, F. C. K. & Lo, Y. M. D. Prenatal diagnosis innovation: genome sequencing of maternal plasma. *Annu. Rev. Med*. **67**, 2.1-2.14 (2016).
10. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet*. **17**, 333-351 (2016).
11. Metzker, M. L. Sequencing technologies – the next generation. *Nat. Rev. Genet*. **11**, 31-46 (2010).
12. Valouev, A. et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome. Res.* **18**, 1051-1063 (2008).
13. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. **26**, 589-595 (2010).
14. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997 [q-bio.GN] (2013).
15. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie2. *Nat. Methods*. **9**, 357-359 (2012).
16. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*. **35**, 421-432 (2019).
17. Gaudelli, N. M. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA

cleavage. *Nature*. **551**, 464-471 (2017).

18. Sood, A. et al. Terminal phosphate-labeled nucleotides with improved substrate properties for homogeneous nucleic acid assays. *J. Am. Chem. Soc*. **127**, 2394-2395 (2005).
19. Sims, P. A., Greenleaf, W. J., Duan, H. & Xie, X. S. Fluorogenic DNA sequencing in PDMS microreactors. *Nat. Methods*. **8**, 575-580 (2011).
20. Chen, Z. et al. Fluorogenic sequencing using halogen-fluorescein-labeled nucleotides. *ChemBioChem*. **16**, 1153-1157 (2015).
21. Chen, Z. et al. Highly accurate fluorogenic DNA sequencing with information theory-based error correction. *Nat. Biotechnol*. **35**, 1170-1178 (2017).
22. Li, M. et al. BEAMing up for detection and quantification of rare sequence variants. *Nat. Methods.* **3**, 95-97 (2006).
23. Zhou, W. et al. A virtual sequencer reveals the dephasing patterns in error-correction code DNA sequencing. *Natl. Sci. Rev*. **8**, nwaa227 (2021).
24. Wu, M. et al. Pathogenesis diagnosis of a pediatric patient suffering from multi-organ abscesses. *Diagn. Pathol*. **18**, 85 (2023).
25. Guo, H. et al. The DNA methylation landscape of human early embryos. *Nature*. **511**, 606-610 (2014).
26. Li, X. et al. The comprehensive DNA methylation landscape of hematopoietic stem cell development. *Cell. Discov*. **7**, 86 (2021).
27. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology*. **31**, 745-759 (2020).
28. Luo, X. et al. 3D genome of macaque fetal brain reveals evolutionary innovations during primate corticogenesis. *Cell*. **184**, 723-740 (2021).
29. Ding, B. et al. Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability. *Sci. Adv*. **7**, eabi6020 (2021).
30. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic. Acids. Res*. **29**, 308-311 (2001).
31. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic. Acids. Res*. **46**, D1062-D1067 (2017).
32. Song, S. et al. Genomic Epidemiology of SARS-CoV-2 in Pakistan. *Genom. Proteom. Bioinf.* **19**, 727-740 (2021).
33. Loman, N. J. et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434-439 (2012).

34. Seshan, V. E. & Olshen, A. DNAcopy: DNA copy number data analysis. R package version 1.70.0 (2022).
35. Patro, R. et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. **14**, 417-419 (2017).
36. Pu, D. et al. A real-time decoding sequencing based on dual mononucleotide addition for cyclic synthesis. *Anal. Chim. Acta*. **852**, 274-283 (2014).
37. Pu, D. & Xiao, P. A real-time decoding sequencing technology – new possibility for high throughput sequencing. *RSC. Adv*. **7**, 40141-40151 (2017).
38. Scheres, S. H. W. A Bayesian view on cryo-EM structure determination. *J. Mol. Biol*. **415**, 406-418 (2012).

Acknowledgments

The authors thank Ying Shang, Wentao Li, Lingguo Du, Xiaomeng Wu, Bingxiao Feng, Jiawei Huang and Runxin Du for experimental and analytical assistance. **Funding**: Funding was partially provided by National Natural Science Foundation of China Grant 21927802 (Y.H., J.W.), 22050002 (Y.H.) and T2188102 (Y.H.), Beijing Municipal Science and Technology Commission Grant Z201100005320016 (Y.H.) and Z211100003321006 (Y.H.), and Beijing Advanced Innovation Center for Genomics. **Author contributions**: Conceptualization: Y.H., Z.C., and W.Z. Experiment: W.Z., L.K., S.Q., H.D., R.Z., L.S., Y.X., W.Y., Y.W., L.D., Y.P., M.L., L.R., and C.Y. Data Analysis: W.Z., L.K., S.Q., Z.L., M.T., L.L., M.D., X.Z., J.W., Z.C., and Y.H. Writing: W.Z., L.K., J.W., Z.C., and Y.H. **Competing interests**: W.Z., L,K., S.Q., H.D., Z.C. and Y.H. are inventors on the patent applications of fuzzy sequencing technologies. H.D., Z.C. and Y.H. are founders, and R.Z., L.L. and L.S. are employees, of Cygnus Biosciences. Other authors declare no competing interests.