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,#] , Chao Liu[5,#], Zhizhao Liao[1], Mingchuan Tang[1], Ruiying Zhang[3], Lei Li[3], Lei Shi[3], Meijie Du[6,7], Yipeng Wang[4], Wentao Yue[4], Yan Xiao[8], Lin Di[1,9], Xiannian Zhang[1], Yuhong Pang[1], Mingkun Li[10], Lili Ren[8], Jianbin Wang[6,*], Zitian Chen[3,*], and Yanyi Huang[1,2,9,*]

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]Department of Infectious Diseases, Peking University Third Hospital, Beijing 100871, China

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

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

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

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

[10] 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.

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

57 Next-generation sequencing relies on the incorporation of nucleotides to encode
58 information, and the efficiency of information encoding varies across different
59 systems. The majority of high-throughput DNA sequencers utilize sequencing-by-
60 synthesis (SBS) chemistries, with signal generation schemes falling into two major
61 variations: single-nucleotide addition (SNA) or cyclic reversible terminator (CRT)[10-
62 11]. We define the intrinsic information efficiency of a sequencer as the information
63 entropy per cycle of the sequencing signals. According to this definition, SNA
64 chemistry produces natural DNA duplex but suffers from relatively low intrinsic
65 information efficiency, averaging at 0.67 nt/cycle or 1.33 bit/cycle for encoding
66 random sequences. In contrast, multi-step CRT chemistry, although producing scar-
67 containing duplex, offers 2 bit/cycle intrinsic information efficiency with an exact 1
68 nt/cycle extension rate (Supplementary Fig. 1a). Enhancing the information entropy
69 encoding efficiency becomes a clear avenue for improving sequencing performance.

70 While the information entropy embedded in the nucleotide sequence plays a
71 decisive role in mapping a read to a reference, mapping accuracy is also influenced by
72 the corresponding region of the reference. One example is the mate-pair reads which,
73 compared with conventional paired-end reads, are more mappable because of larger
74 reference target size instead of higher information entropy. This effect becomes
75 particularly evident in genomes with a high proportion of repeat sequences. From the
76 perspective of the sequencer, we define the extrinsic information efficiency as the
77 information entropy per cycle of the sequenced DNA fragment, or two times of the
78 read length per cycle. This reference-side extrinsic information is equally significant
79 but often overlooked. This oversight is primarily due to prevalent DNA sequencers

80 providing explicit sequences of four bases, where the intrinsic information efficiency
81 is equal to the extrinsic information efficiency, despite the differences in efficiency
82 between SNA and CRT. Enhancing the extrinsic information efficiency could also
83 lead to improvements in sequencing technology.

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

95 From a sequencing strategy standpoint, achieving higher information efficiency
96 (> 2 bit/cycle) for both intrinsic and extrinsic information entropy could revolutionize
97 the entire sequencing field. In this report, we introduce a fuzzy SBS strategy that
98 sacrifices explicit sequencing information in exchange for improved information
99 encoding performance. The fuzzy SBS sequencing strategy is based on the
100 fluorogenic sequencing chemistry using the terminal phosphate-labeled fluorogenic
101 nucleotides[13-15] (TPLFN, Fig. 1a) we reported earlier. In this chemistry, the
102 fluorophore Tokyo Green (TG) labeled TPLFN can turn from a dark state to a light
103 state after catalysis by the Bst polymerase and alkaline phosphatase, and its
104 fluorescence intensity indicates the number of the incorporated nucleotides (Fig. 1b).
105 We constructed DNA fragments into a library with two specific end adapters for
106 clonal amplification on hydrogel microbeads carrying amplification primers. The
107 DNA fragments hybridized on the beads were then amplified into clones through
108 emulsion PCR[16] or surface-tethered recombinase polymerase amplification (RPA)
109 (Fig. 1c). Each clone typically contained approximately 30,000 copies of DNA for the
110 SBS reaction. After annealing with the sequencing primer, the DNA fragments can be
111 effective sequenced through a conventional SNA strategy (Fig. 1d).

112 Fuzzy SBS can be implemented using different flowgrams (Fig. 1e-f,
113 Supplementary Fig. 1b), allowing the introduction of nucleotides beyond the
114 limitation of a single type per cycle (see Supplementary Text 2.1). Briefly, *BitSeq*
115 employs a double-nucleotide addition flowgram, where a combination of two
116 nucleotides is alternately added in each cycle. For instance, K (a mix of G/T) is added
117 during odd cycles, while M (a mix of A/C) is added during even cycles (see Fig. 1e).
118 BitSeq generates fuzzy sequences since it's not possible to distinguish the signal
119 associated with the incorporation of the two nucleotides in the mix. Similarly, TritSeq
120 employs a triple-nucleotide addition flowgram, where a combination of three
121 nucleotides is alternately added in each cycle (Supplementary Fig. 1b). All of these
122 fuzzy SBS approaches can achieve intrinsic and/or extrinsic information efficiency
123 beyond 2 bits per cycle, effectively reducing the cycle number required to achieve a
124 specific read length and ultimately leading to faster turn-around speed
125 (Supplementary Table 1). In this paper, the focus will primarily be on BitSeq and its
126 dichromatic form SuperBitSeq as they strike a reasonable balance between
127 information efficiency and signal dynamic range per reaction cycle.

128 We constructed a laboratory prototype sequencer for conducting high-throughput
129 BitSeq (Fig. 2a,b). The bit sequencer comprises a microfluidic flow cell for
130 implanting clonally amplified DNA fragments for sequencing and an imaging system
131 for signal acquisition (Supplementary Fig. 2-4). We implemented fluorogenic SBS
132 chemistry to naturally align with the BitSeq flowgram[13-15,17].

133 The microfluidic flow cell (Fig. 2c) contains 28 million 30-femtoliter microwells
134 (Fig. 2d,e). Each cycle of the SBS reaction can extend one or more bases and produce
135 a corresponding amount of fluorophore. The microwells can be sealed to prevent
136 signal cross-talk due to fluorophore diffusion. These microwells were fabricated by
137 selective wet-etching of fiber-optic plates (FOP), allowing the fluorogenic signal to
138 propagate through the fiber waveguide to the outer surface of the flow cell for
139 imaging (Fig. 2f). The inner surface of the microwells was made hydrophilic, while
140 the adjoining surface between microwells was topologically coated to be
141 hydrophobic, enabling the highly parallel sealing of the SBS reactions using oil
142 (Supplementary Fig. 5,6).

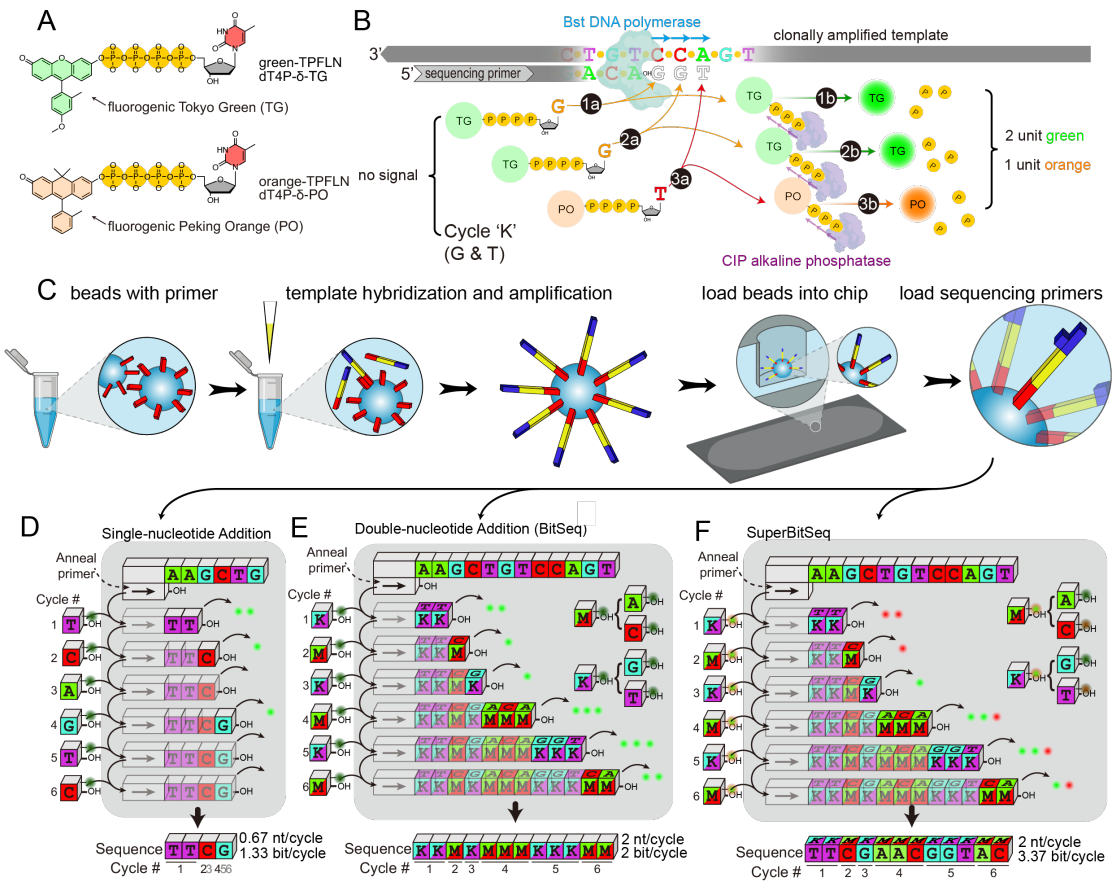

Fig. 1. Schematic of fuzzy sequencing. (A) Chemical structure of the terminal phosphate-labeled fluorogenic nucleotides (TPLFN) used in fuzzy sequencing. (B) Schematic of the fluorogenic sequencing chemistry adopted by fuzzy sequencing. (C) Schematic of the fuzzy sequencing pipeline. The DNA templates are amplified on the beads and loaded into the wells of the sequencing chip. After hybridized by the sequencing primers, the DNA templates are primed for sequencing. (D-F) Schematic of the nucleotide flowgrams used by single-nucleotide addition (D), monochromatic double-nucleotide addition (BitSeq, E) and dichromatic double-nucleotide addition (SuperBitSeq, F).

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 °C. We then sealed the microwells by steadily flowing in fluorinated oil (Fig. 2f). Next, we quickly raised the temperature of the flowcell to 65°C to initiate the synthesis reaction, followed by cooling down to 25°C 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).

162 Defected regions resulting from local sealing or unsealing failures were
163 eliminated before further processing (Supplementary Fig. 7A). Empty microwells and
164 microwells with multiple beads were identified and filtered out of further analysis.
165 The fluorescence signal of each monoclonal microwell was extracted from the images
166 after noise reduction and background correction. Due to the asynchronization of the
167 molecular reaction within each clone, the intensity-cycle series could not be directly
168 converted into bit sequences without a meticulous dephasing process. To supervise
169 the dephasing algorithm, we spiked a standard DNA library, constructed from the
170 lambda phage genome, into the samples to be sequenced together. As a result, the
171 sequenced reads could be easily classified into two categories: standard dots (SD) and
172 library dots (LD), based on the characteristic signal profiles of the first few cycles.
173 We employed the dual-base flowgram dephasing model to fit the SD signals and
174 obtain the parameters[18], which were then applied to LD signals to construct a flux
175 matrix for dephasing correction and to convert the signals into bit sequences.

176 We initially tested our prototype sequencer on a lambda phage genomic library
177 (Supplementary Fig. 8, Supplementary Table 2). After 61 cycles of repeated sealing
178 and sequencing, more than 90% of the chip area remained unaffected by stains
179 (Supplementary Fig. 7B, 11). In a single run, we generated 6,578,364 reads with an
180 average length of 112.2 bp (Fig. 2h, Supplementary Fig. 9-15). To map the BitSeq
181 signals, we encode them as sequence reads that can be aligned using conventional
182 mappers such as Burrows-Wheeler Aligner (BWA)[19-20] and Bowtie2 (Supplementary
183 Text 2.2, Supplementary Table 3-6)[21-22]. By mapping the bit sequences back to the
184 reference genome, we successfully aligned 99.11% of the reads, which had an average
185 length of 111.6 bp. The error rate of BitSeq was found to be 0.39% in the first 10 bp
186 and 0.62% in the 130-140 bp region, which were acceptably low. This indicates the
187 feasibility of accurate read mapping and DNA fragment counting (Fig. 2i).

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

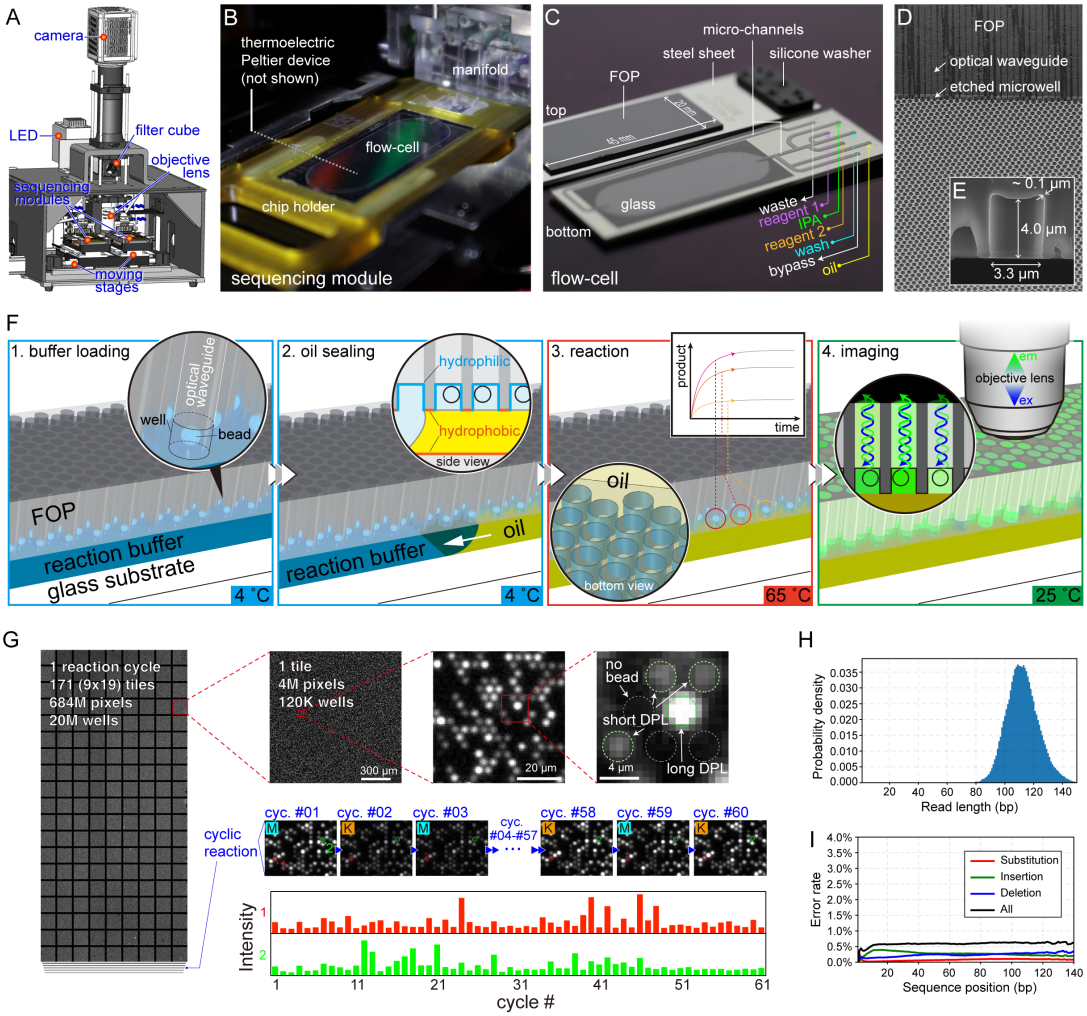

Fig. 2. Implementation of BitSeq. (A) The design of a prototype for high-throughput fluorogenic fuzzy sequencing. (B) The photo of a sequencing module, which consists of a flow-cell connected with a manifold for liquid routing, and the flow-cell is placed on a thermoelectric Peltier device for temperature 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) The read length distribution of bit sequences. (I) The error rate of BitSeq.

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.

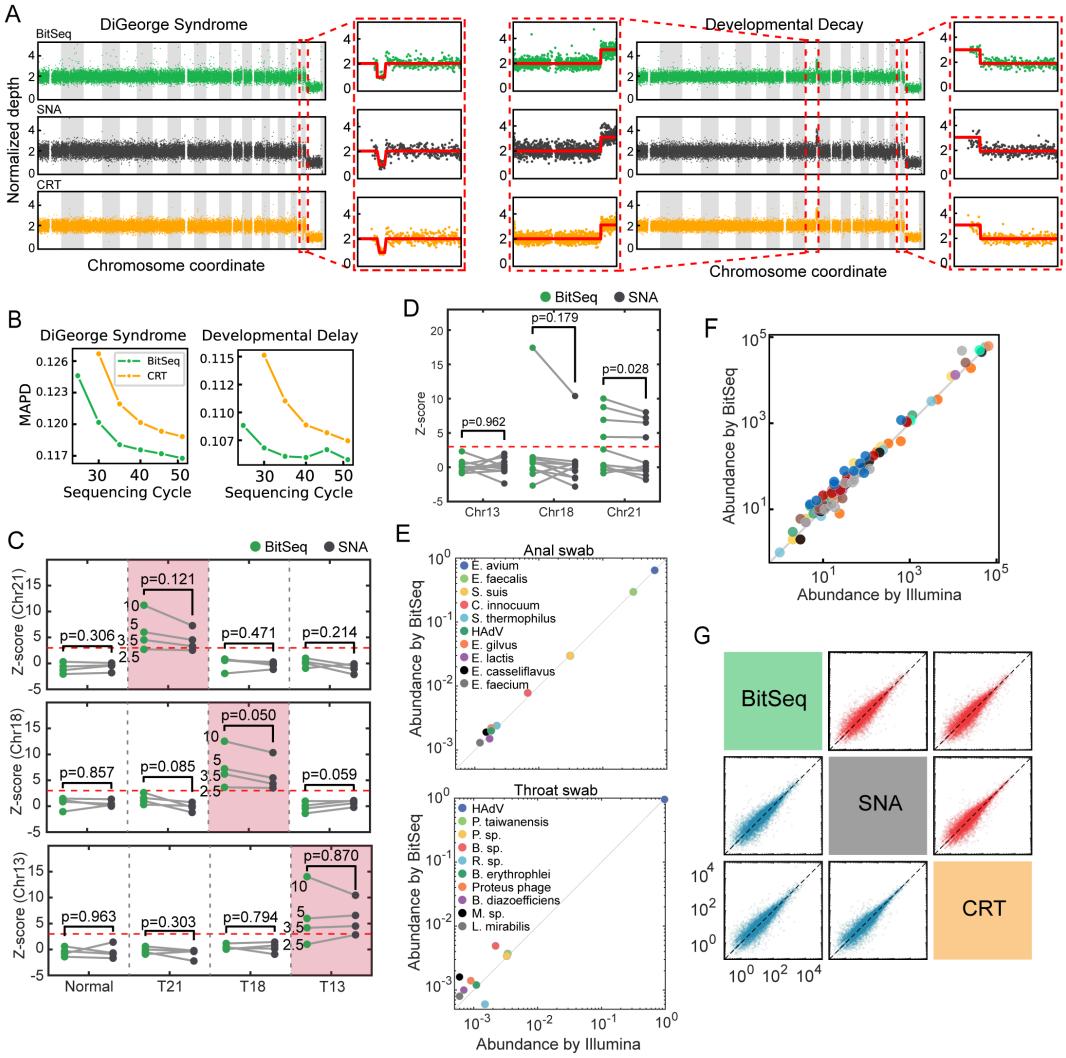

Fig. 3. Comparison of BitSeq with commercial sequencers in resequencing. (A) CNV of a DiGeorge syndrome patient (left) and a development-delayed patient (right). (B) BitSeq shows lower MAPD than CRT under the same sequencing cycles. MAPD for CRT under 25 cycles is missing because such short reads cannot be mapped uniquely to the genome. (C) Z-scores of normal and mock trisomy NIPT ccfDNA samples. (D) Z-scores of single-blinded NIPT ccfDNA samples. The red dash in B-C indicates the threshold of 3. (E-F) Microbe abundance comparison between BitSeq and Illumina by metagenomic sequencing (E) and targeted metagenomic sequencing (F). (G) TPM comparison of MEF (blue) and mES (red) cell line between BitSeq and commercial sequencers.

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

230 development delay (Fig. 3a) with high concordance with other conventional
231 sequencing approaches. Notably, BitSeq showed lower median of the absolute values
232 of all pairwise differences (MAPD) than CRT when truncated to read lengths with the
233 same sequencing cycles, indicating more precise CNV determination (Fig. 3b).

234 We then tested the effectiveness of BitSeq on NIPT, a widely used resequencing
235 application in prenatal diagnostics that requires high accuracy of DNA counting. We
236 started the test with mock samples prepared by spiking fragmented genomic DNA
237 from trisomy patients into circulating cell-free DNA (ccfDNA) of an unpregnant
238 woman with different ratios. These 12 mock NIPT samples had trisomy in Chr21,
239 Chr18 and Chr13, respectively, and the mixing proportion was 10%, 5%, 3.5% and
240 2.5%. We also collected 28 NIPT ccfDNA samples from pregnant women with
241 normal fetuses. We applied BitSeq to all the samples and used 24 out of the 28 normal
242 samples as the control set (the other 4 used as the negative test set) to calculate the Z-
243 score for each NIPT sample (Fig. 3c). Using Z=3 as the cut-off, BitSeq shows
244 comparable power to identify the positive trisomy NIPT samples when the mock fetal
245 DNA fraction is not less than 3.5%. We further tested 10 single-blinded true NIPT
246 samples, 5 of which were tested positive based on the same cut-off using BitSeq (Fig.
247 3d). These results were confirmed by parallel tests by Ion Torrent sequencers as well
248 as the clinical records (Supplementary Fig. 18,19a).

249 Combined with metagenomic sequencing, BitSeq may identify pathogens in acute
250 infections timely. We tested the capability of BitSeq in metagenomics by sequencing
251 one anal swab and one throat swab from a previously reported pediatric patient
252 suffered from multi-organ abscesses[23]. The 10 species types with the highest
253 abundance identified by BitSeq and Illumina sequncer are identical. And their
254 abundances, ranging from less than 0.1% to over 95%, are also close (Fig. 3e). In
255 addition, using a custom panel, we also tested BitSeq on targeted metagenomics by
256 sequencing bronchoalveolar lavage fluid samples from 14 patients with community-
257 acquired pneumonia. The species types as well as their normalized read number
258 reflecting their relative abundance, are also consistent (Fig. 3f).

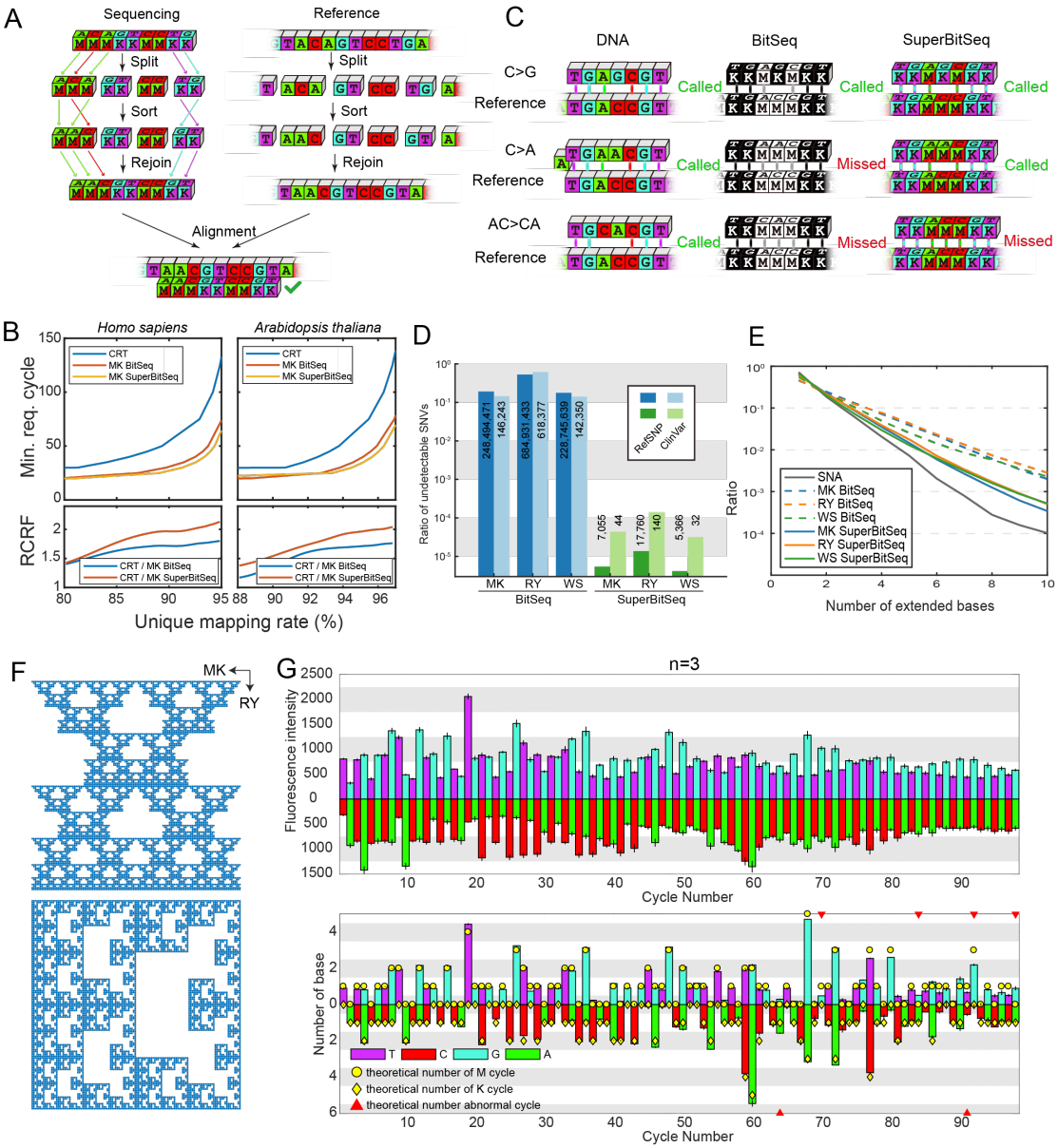

Fig. 4. Properties of SuperBitSeq. (A) Encoding strategy of SuperBitSeq in M-K form as an example. (B) Simulated unique mapping rate of DNA reads by fuzzy sequencing and CRT for different genomes. For mapping the specific proportion of the genome, both BitSeq and SuperBitSeq need much shorter minimally required sequencing cycles, and this advantage can also be reflected by reaction cycle reduction factor (RCRF) that defined by the ratio of minimally required cycles between technologies. (C) Distinguishable and indistinguishable SNV types of BitSeq and SuperBitSeq. (D) Ratio of indistinguishable SNV by BitSeq and SuperBitSeq. (E) Freqencies of extended base number of different flowgrams. (F) Fractal of encoded 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. 3g, Supplementary Fig. 19b-c). We also

274 checked the Gene Ontology (GO) terms of genes only detected by one sequencer and
275 found no specific preferences in between (Supplementary Fig. 19d-f).

276 All these experiments show that BitSeq is fully capable for those applications
277 based on DNA fragment counting. We expect that other re-sequencing approaches
278 that do not require single-base resolution but focus on pattern alterations such as
279 bisulfite sequencing for DNA methylation[24-26] or Hi-C for chromatin topology
280 analysis[27-28] are also possible with BitSeq. Thanks to its high information and long
281 extension length per cycle as well as the fast fluorogenic sequencing chemistry,
282 BitSeq would enable more efficient DNA fragment counting that is valuable in both
283 basic science research and clinical testing.

284 We further considered the possibility to extend fuzzy sequencing strategy from
285 BitSeq to SuperBitSeq by labeling the two different nucleotides in each reaction cycle
286 with distinguishable fluorophores (Fig. 1f). SuperBitSeq possesses the same extrinsic
287 information efficiency as BitSeq, but even higher intrinsic information efficiency
288 (Supplementary Fig. 1d).

289 Like BitSeq, we also designed an encoding strategy for SuperBitSeq that is
290 compatible with prevailing short reads mappers like BWA (Fig. 4a). We conducted *in*
291 *silico* analysis to validate the alignment accuracy of the fuzzy SBS encoding strategy.
292 For our analysis, we randomly selected 1×10^7 positions from the human reference
293 genome and 1×10^6 positions from the *Arabidopsis* reference genome. We simulated
294 the BitSeq, SuperBitSeq and CRT sequencing from these positions by extending the
295 nucleotide strings with certain cycles (Supplementary Fig. 20), and then aligned the
296 simulated reads back to their respective reference genomes. We observe that reads
297 generated with more sequencing cycles provided greater information, resulting in a
298 higher unique mapping rate (UMR) for both fuzzy sequencing and CRT approaches.
299 Notably, SuperBitSeq consistently exhibited the highest UMR (Fig. 4b,
300 Supplementary Fig. 21, Supplementary Table 7-58). Additionally, the UMR of BitSeq
301 with the M-K form was consistently higher than that of CRT and BitSeq with R-Y
302 form. This preference for the M-K form may be attributed to the occurrence of
303 transition base substitutions follow gene duplication during evolution[29].

304 The elevated information efficiency enables the detection of significantly more
305 SNVs that cannot be identified through BitSeq (Fig. 4c). For example, in BitSeq the

306 A>C/G>T mutations cannot be identified in the M-K form, nor the A>G/C>T
307 mutations in the R-Y form. In SuperBitSeq they can all be well captured, except for
308 only specific indistinguishable base swap events such as the AC>CA in the M-K
309 form. To evaluate the potential of SuperBitSeq for SNP/SNV detection, we filtered
310 dbSNP (build 155)[30] and ClinVar (20210908)[31] to exclude items with N, and obtained
311 1,297,977,577 and 1,009,591 items respectively. About 17-52% of known human
312 SNP/SNVs cannot be distinguished by BitSeq, whereas the undetection rates are
313 ~0.01% or even lower for SuperBitSeq (Fig. 4d). This means that SuperBitSeq can *de*
314 *facto* detect almost all known SNP/SNVs. More importantly, in each reaction cycle
315 the fluorescence signal released by extended bases can be split into two channels,
316 resulting in reduced requirement for signal detection dynamic range and enhanced
317 accuracy in detecting long copolymers (Fig. 4e, Supplementary Fig. 22). Interestingly,
318 we also found that the encoded superbit sequences can form a fractal when mapped
319 the points in a square (Fig. 4f, Supplementary Text 2.4).

320 To implement SuperBitSeq experimentally, we synthesized a novel fluorophore,
321 Peking Orange (PO), which was designed to exhibit excellent fluorogenic properties
322 (on/off ratio ~ 2000) when terminally labeled on the tetraphosphate nucleotides (PO-
323 TPLFNs, or PO-dN4P, Fig. 1a, Supplementary Fig. 23,24,25a). PO was used to pair
324 with TG for labeling two different nucleotides, respectively, in each reaction cycle.
325 PO and TG have similarly high fluorescence quantum efficiency but distinct
326 excitation and emission spectra to avoid signal cross-talk (Supplementary Fig. 25b).
327 PO-dN4P has great photo and thermal stability, and showed excellent signal-to-base
328 linearity upon nucleotide extension reaction (Supplementary Fig. 25c-e). The same
329 dephasing algorithm we optimized for TG-dN4P substrates could be seamlessly
330 adopted to PO-dN4P when sequencing a single template (Supplementary Fig. 26).
331 According to our previous *in silico* simulation, dephasing correction for SuperBitSeq
332 could be built upon the algorithm of BitSeq by separating two channels for
333 independent correction with the same dephasing parameters[18]. We also
334 experimentally confirmed this strategy using the same single template for
335 SuperBitSeq (Fig. 4g, Supplementary Fig. 27).

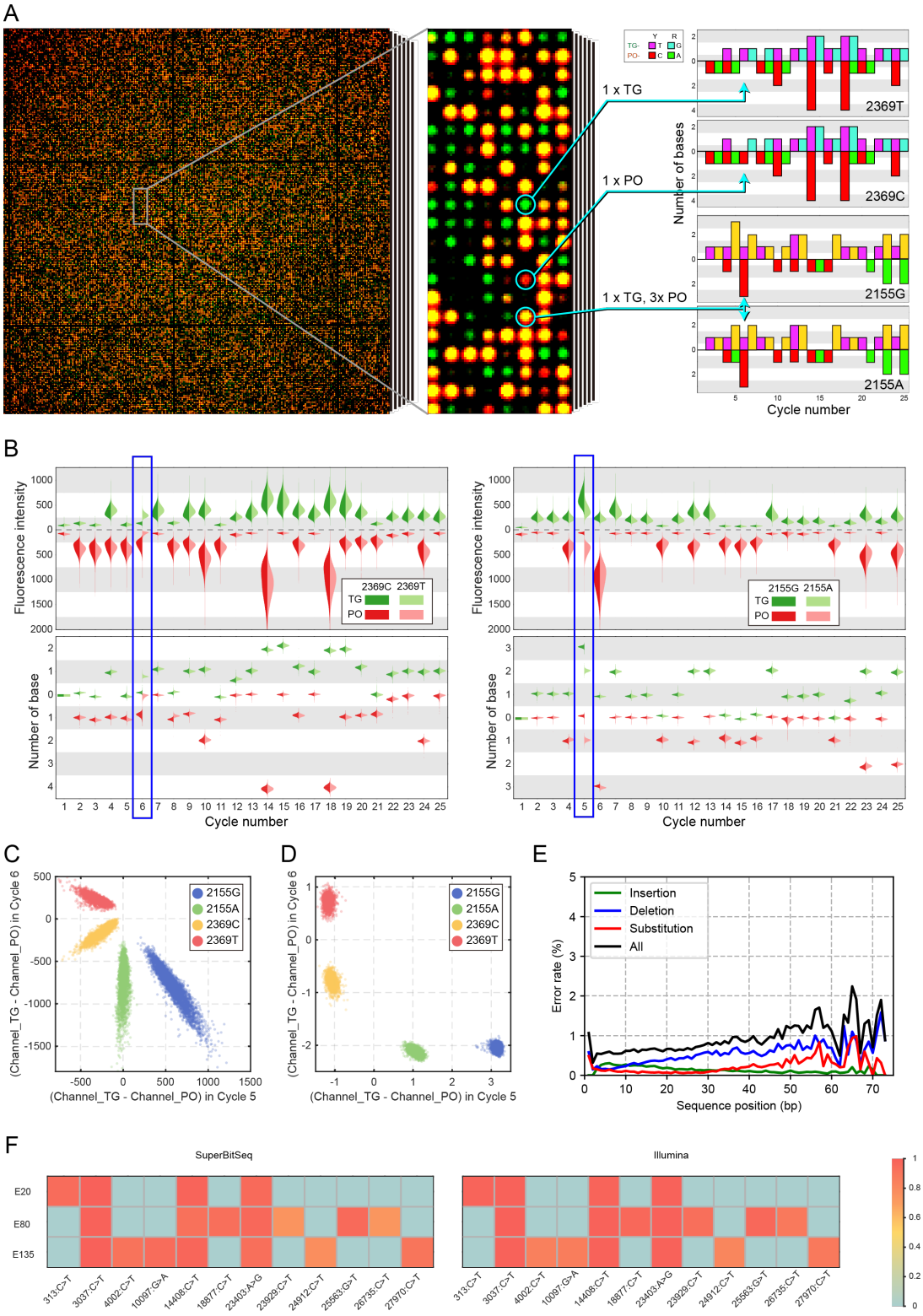

Fig. 5. High-throughput SuperBitSeq identifies single-nucleotide variants (SNVs). (A) A merged fluorescence image of green (TG-labeled T or G) and red (PO labeled C or A) channels, from which the signal intensities are extracted. The image was associated with Cycle 6 in this specific experiment and yellow 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.

We added a fluorescence channel to the bit sequencer and converted it to the superbit sequencer (Fig. 5a, Supplementary Fig. 28), 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. 29). The experimental data demonstrated low error rate ($\approx$1%, Fig. 5e) and an average read length of 43 bp (Supplementary Fig. 30a). 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. 31). 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 three 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. 4f), 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. Last, further optimizations are still required to reduce the overall operational time and cost. Although cycle number reduction by fuzzy sequencing leads to reduced reaction time and reagent cost, multiple steps such as template amplification, fluidics and imaging still take up a large proportion of time, and the throughput should be further increased to reduce the average cost.

397 **References and Notes**

398 1. Levy, S. E. & Myers, R. M. Advancements in next-generation sequencing. *Annu. Rev. Genomics.*
399 *Hum. Genet.* **17**, 95-115 (2016).
400 2. Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature*. **550**, 345-353 (2017).
401 3. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl.*
402 *J. Med*. **363**, 166-176 (2010).
403 4. Li, R. et al. A body map of somatic mutagenesis in morphologically normal human tissues.
404 *Nature*. **597**, 398-403 (2021).
405 5. Li, X. et al. Improvement in the risk assessment of oral leukoplakia through morphology-related
406 copy number analysis. *Sci. China. Life. Sci.* **64**, 1379-1391 (2021).
407 6. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the
408 human genome. *Nat. Rev. Genet*. **16**, 172-183 (2015).
409 7. Wu, A. R., Wang, J., Streets, A. M. & Huang, Y. Single-cell transcriptional analysis. *Annu. Rev.*
410 *Anal. Chem.* **10**, 439-462 (2017).
411 8. Di, L. et al. RNA sequencing by direct tagmentation of RNA/DNA hybrids. *Proc. Natl. Acad.*
412 *Sci. U. S. A*. **117**, 2886-2893 (2020).
413 9. Wong, F. C. K. & Lo, Y. M. D. Prenatal diagnosis innovation: genome sequencing of maternal
414 plasma. *Annu. Rev. Med*. **67**, 2.1-2.14 (2016).
415 10. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-
416 generation sequencing technologies. *Nat. Rev. Genet*. **17**, 333-351 (2016).
417 11. Metzker, M. L. Sequencing technologies – the next generation. *Nat. Rev. Genet*. **11**, 31-46
418 (2010).
419 12. Valouev, A. et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of
420 universal sequence-dictated positioning. *Genome. Res.* **18**, 1051-1063 (2008).
421 13. Sims, P. A., Greenleaf, W. J., Duan, H. & Xie, X. S. Fluorogenic DNA sequencing in PDMS
422 microreactors. *Nat. Methods*. **8**, 575-580 (2011).
423 14. Chen, Z. et al. Fluorogenic sequencing using halogen-fluorescein-labeled nucleotides.
424 *ChemBioChem*. **16**, 1153-1157 (2015).
425 15. Chen, Z. et al. Highly accurate fluorogenic DNA sequencing with information theory-based
426 error correction. *Nat. Biotechnol*. **35**, 1170-1178 (2017).
427 16. Li, M. et al. BEAMing up for detection and quantification of rare sequence variants. *Nat.*
428 *Methods.* **3**, 95-97 (2006).
429 17. Sood, A. et al. Terminal phosphate-labeled nucleotides with improved substrate properties for

430 homogeneous nucleic acid assays. *J. Am. Chem. Soc*. **127**, 2394-2395 (2005).
431 18. Zhou, W. et al. A virtual sequencer reveals the dephasing patterns in error-correction code DNA
432 sequencing. *Natl. Sci. Rev*. **8**, nwaa227 (2021).
433 19. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
434 *Bioinformatics*. **26**, 589-595 (2010).
435 20. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
436 1303.3997 [q-bio.GN] (2013).
437 21. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie2. *Nat. Methods*. **9**,
438 357-359 (2012).
439 22. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of
440 threads on general-purpose processors. *Bioinformatics*. **35**, 421-432 (2019).
441 23. Wu, M. et al. Pathogenesis diagnosis of a pediatric patient suffering from multi-organ abscesses.
442 *Diagn. Pathol*. **18**, 85 (2023).
443 24. Guo, H. et al. The DNA methylation landscape of human early embryos. *Nature*. **511**, 606-610
444 (2014).
445 25. Li, X. et al. The comprehensive DNA methylation landscape of hematopoietic stem cell
446 development. *Cell. Discov*. **7**, 86 (2021).
447 26. Liu, M. C. et al. Sensitive and specific multi-cancer detection and localization using
448 methylation signatures in cell-free DNA. *Annals of Oncology*. **31**, 745-759 (2020).
449 27. Luo, X. et al. 3D genome of macaque fetal brain reveals evolutionary innovations during
450 primate corticogenesis. *Cell*. **184**, 723-740 (2021).
451 28. Ding, B. et al. Noncoding loci without epigenomic signals can be essential for maintaining
452 global chromatin organization and cell viability. *Sci. Adv*. **7**, eabi6020 (2021).
453 29. Gaudelli, N. M. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA
454 cleavage. *Nature*. **551**, 464-471 (2017).
455 30. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic. Acids. Res*. **29**, 308-
456 311 (2001).
457 31. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting
458 evidence. *Nucleic. Acids. Res*. **46**, D1062-D1067 (2017).
459 32. Song, S. et al. Genomic Epidemiology of SARS-CoV-2 in Pakistan. *Genom. Proteom. Bioinf*.
460 **19**, 727-740 (2021).
461 33. Loman, N. J. et al. Performance comparison of benchtop high-throughput sequencing platforms.
462 *Nat. Biotechnol.* **30**, 434-439 (2012).

463 34. Seshan, V. E. & Olshen, A. DNAcopy: DNA copy number data analysis. R package version
464 1.70.0 (2022).
465 35. Patro, R. et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat.*
466 *Methods*. **14**, 417-419 (2017).
467 36. Pu, D. et al. A real-time decoding sequencing based on dual mononucleotide addition for cyclic
468 synthesis. *Anal. Chim. Acta*. **852**, 274-283 (2014).
469 37. Pu, D. & Xiao, P. A real-time decoding sequencing technology – new possibility for high
470 throughput sequencing. *RSC. Adv*. **7**, 40141-40151 (2017).
471 38. Scheres, S. H. W. A Bayesian view on cryo-EM structure determination. *J. Mol. Biol*. **415**, 406-
472 418 (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.), T2225005 (J.W.) and T2188102 (Y.H.), Beijing Municipal Science and Technology Commission Grant Z201100005320016 (Y.H.), Z211100003321006 (Y.H.) and Z221100007022003 (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., C.L., 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.