

# Processing DNA Storage through Programmable Assembly in a Droplet-Based Fluidics System

Minsang Yu, Doyeon Lim, Jungwoo Kim, and Youngjun Song\*

DNA can be used to store digital data, and synthetic short-sequence DNA pools are developed to store high quantities of digital data. However, synthetic DNA data cannot be actively processed in DNA pools. An active DNA data editing process is developed using splint ligation in a droplet-controlled fluidics (DCF) system. DNA fragments of discrete sizes (100–500 bps) are synthesized for droplet assembly, and programmed sequence information exchange occurred. The encoded DNA sequences are processed in series and parallel to synthesize the determined DNA pools, enabling random access using polymerase chain reaction amplification. The sequencing results of the assembled DNA data pools can be orderly aligned for decoding and have high fidelity through address primer scanning. Furthermore, eight 90 bps DNA pools with pixel information (png: 0.27–0.28 kB), encoded by codons, are synthesized to create eight 270 bps DNA pools with an animation movie chip file (mp4: 12 kB) in the DCF system.

## 1. Introduction


Digital DNA data storage can store sequence information at the molecular level and has been a recent focus of interest because of its capacity for storing extremely dense information<sup>[1,2]</sup> and unique computing ability.<sup>[3]</sup> Digital data is encoded into DNA as a packet of information, represented as size (structure)<sup>[4–6]</sup> or sequence-based DNA, and contains a combination of binary digits, similar to electronic data.<sup>[4,7,8]</sup> Furthermore, to solve mathematical problems in Boolean logic operation and cell computing, single-strand DNA (ssDNA) sequences, which can undergo hybridization and displacement reactions, have been developed as nanocomputing materials for diverse computational research, such as in neural networks.<sup>[3,9–12]</sup> Despite the functional duality

of DNA for computing and memory storage using sophisticated biological technology, processing digital data stored in DNA is not well studied. Most studies focused on data capacity,<sup>[13]</sup> storage applications,<sup>[14]</sup> and sequence encoding methods.<sup>[1]</sup> For DNA editing, a DNA nanoswitch system has been developed with fluorescent monitoring or size (structure)-based detection, which can be used as computing and storage system.<sup>[15]</sup> These systems can process stored DNA data but have limited data capacity due to display limitations, such as electrophoresis lanes and many fluorescent probes. For high-capacity storage, synthetic oligos are encoded with two bits of binary information in each sequence. For parallel storage, the synthetic oligo sequences are tethered with index sequences.<sup>[16]</sup> Common forward and reverse primers are allocated to the

synthetic oligo pool to retrieve the stored DNA data using biological methods such as polymerase chain reaction (PCR), a tool that exponentially amplifies the original DNA molecule. Additionally, next-generation sequencing (NGS) is widely used to detect digital data stored in DNA.<sup>[17–19]</sup> Furthermore, diverse encoding methods have been developed to reduce sequencing errors, such as homopolymers and high GC contents, to correct NGS sequencing results, and to ensure high accuracy.<sup>[20,21]</sup>

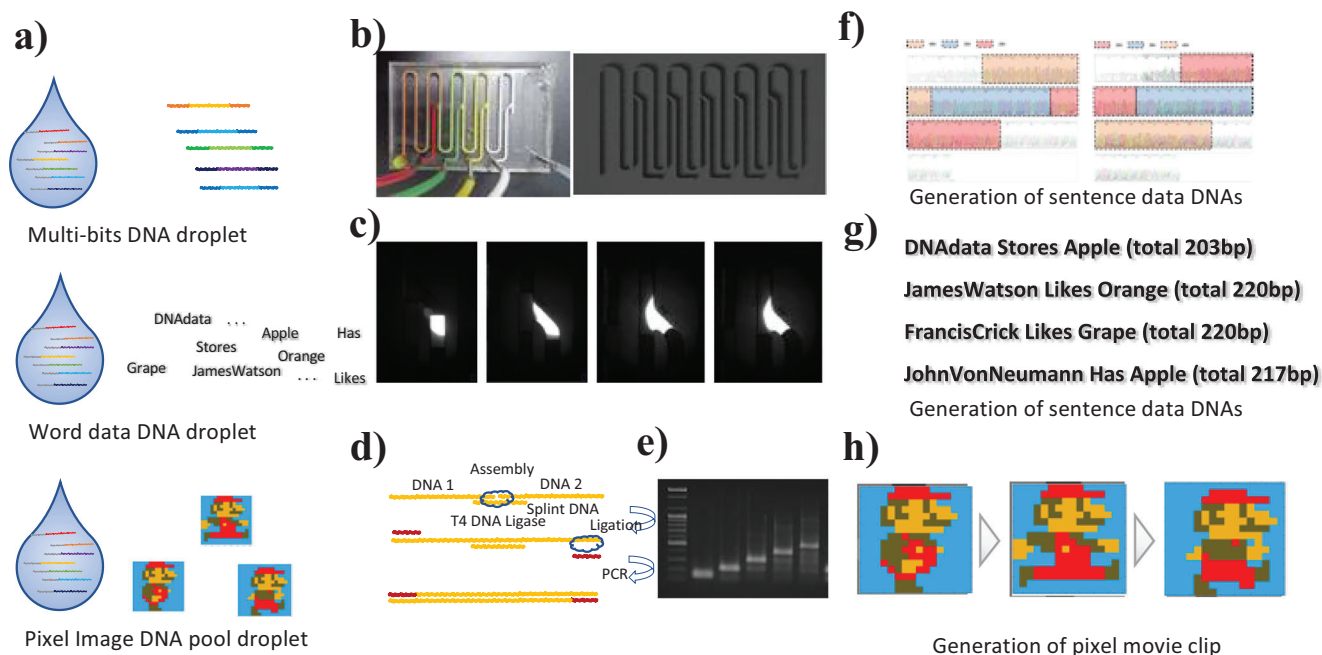
Primer sequences have limitations in reducing DNA data density; however, they help improve data by enabling precise and targeted amplification and ensuring efficient stored information retrieval. Similar to digital data signal amplification using electronic circuits, data molecules with standard forward and reverse primers are amplified using PCR. Moreover, using orthogonal primers, PCR can select a set of DNA data sequences from pools.<sup>[22,23]</sup> Independent primers can perform computational functions such as Boolean logic operations for file selection,<sup>[24]</sup> and primers can also be used for data deletion using index marker DNA hybridization for data security.<sup>[25]</sup> However, high-capacity DNA data storage has not been actively developed for data processing, such as editing or ordering data. Most studies on DNA data storage have focused on cold data storage, where files cannot be frequently accessed.<sup>[26]</sup> Primers and sophisticated biotechnological methods can demonstrate that digital DNA has a functional duality that can be used for computing and data storage. Similar to processing in memory (PIM), recently introduced in electronic device systems to combat data processing limitations, the DNA data molecules, which have data stored in primer sequences, can be used for computing. Specialized devices may be helpful for DNA molecules to handle and manipulate

M. Yu, Y. Song  
Standard Bioelectronics. Co.  
511 Michuhol Tower, Gaetbeol-ro 12, Incheon 21999, South Korea  
D. Lim, J. Kim, Y. Song  
Department of Nano-Bioengineering  
Incheon National University  
Academy-ro 119, Incheon 22012, South Korea  
E-mail: yjunsong@inu.ac.kr

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/advs.202303197>

© 2023 The Authors. Advanced Science published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/advs.202303197



**Figure 1.** Programmable DNA data storage assembly in a droplet controlling fluidics (DCF) system. a) DNA data droplet of the diverse type format (top: size-based bit data, center: sequence-based encoded-word data, bottom: sequence-based encoded pixel art). b) The image and diagram of the DCF chip with multiple y-junction channels. c) Images of droplet association in y-junction. d,e) Diagram and gel image of DNA data assembly. f) Sanger sequencing results for the DNA bits programmable assembly. g,h) The NGS decoding results of the sentences DNA data (g), and the NGS decoding results of the DNA pool movie clip (h).

individual DNA sequences. An electric field-assisted DNA microarray has been developed for hybridization and displacement for short ssDNA sequence processing, allowing random access for three-bit storage and computing.<sup>[27]</sup> However, its ability to handle high-capacity DNA and discrete data molecules is limited. A droplet-based DNA handling system could be helpful for discrete data molecules. Regarding developing DNA storage systems, an electrowetting device can produce a DNA data storage system with 127 sites that can manipulate DNA data into a single pool.<sup>[28]</sup>

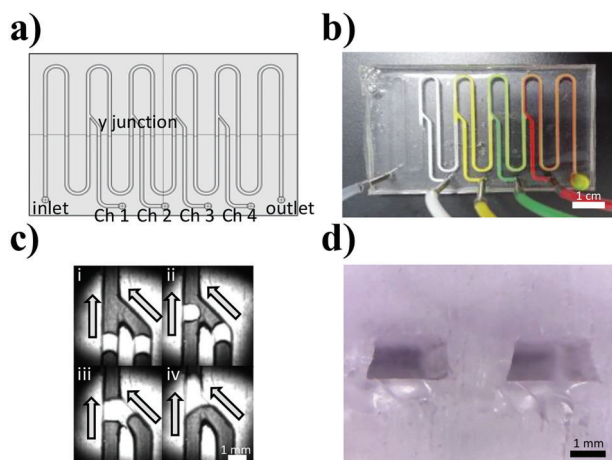
However, this system uses a limited electrode array to focus on the DNA data storage spot and droplet path. Thus, owing to the lack of DNA data editing technology, DNA data operating devices might be helpful to allow DNA data editing, such as DNA data connection. Here, we demonstrate a droplet device for processing DNA data (Figure 1). The digitized droplets were associated with a fluidic system constructed using a 3D-printed mold.<sup>[11]</sup> The DNA data were ligated using splint DNA ligation in the associated droplet. For size and structure-based DNA storage, 100 bps ssDNAs were assembled with splint DNA in the fluidic chip, creating DNA molecules of varying lengths (100–500 bps), each able to display up to five bits of information. Furthermore, we demonstrate that the assembled DNA sequence order was altered using splint DNAs (Figure 1f). The encoded-word DNA data were in series and parallel processed to produce diverse DNA sequences with sentence information to show the PIM storage for high-density DNA data storage (Figure 1g). Finally, we demonstrate that the three DNA pools of the pixel image data, encoded by codon sequences, were created using animation DNA pools. Unlike word DNA data sequences, using an American Standard

Code for Information Interchange (ASCII) code with error correction code (ECC), which applies  $2^7$  Galois field reed-solomon (RS) code,<sup>[29]</sup> the codon-based sequence provided 1.33 bits per bp logical capacity to prevent homopolymers for repeated pixel information. Furthermore, splint address sequences are used for DNA assembly to process DNA data, scanning to search for flexible DNA payloads (unknown data sizes and positions), and aligning them to synchronize the sequences. Due to the capability of the programmable assembly of DNA, the droplet DNA ligation system can be used for DNA digital data storage. In addition, DNA data pool ligation can increase storage ability.<sup>[30]</sup>

## 2. Results

### 2.1. Digitized Droplet Controlling Fluidics System

Digital data, a sequential combination of discrete “0” and “1” numbers, store information, including documents, pictures, and movie files. For example, letter information codes such as ASCII and Universal Transformation Format (UTF-8)<sup>[2,3,31–33]</sup> operate based on a set of sequentially digitized data. This data follows the bits string data format, including the most significant bit (MSB) and a lower significant bit (LSB). The droplet carrier and DNA molecules can be beneficial to store discrete data in DNA as discrete DNA information can be manipulated. A fluidics system operated by droplet carriers can be provided to ensure the programmable DNA droplet association to store DNA droplet data. We fabricated a droplet controlling polydimethylsiloxane (PDMS) fluidics system, casted using a 3D printed mold to control discrete DNA data molecules.<sup>[15,34]</sup> A detailed discussion regarding



**Figure 2.** Digitized DCF system. a) Schematic diagram representing the four y-junction channels in the DCF chip. b) Image of the DCF chip. c) Microscopy images of droplet association. d) Cross-sectional image of the channels.

the process is provided in Experimental Section and Supporting Information 1. For 2  $\mu\text{l}$  of DNA molecular droplets, a droplet controlling fluidics (DCF) channel was designed, which was 1 mm width and 1 mm height (Figure 2a,b). For multiple DNA data droplet manipulation using the DCF chip, we designed a main channel with four y-junctions, in which the droplet association can be operated. DNA droplets in the main DCF channel were  $\approx 2$  mm long with a 1 mm gap. Figure 2c represents the microscopy images of the droplet association at the DCF y-junction. In DCF systems, DNA droplets storing discrete data can be used for programmable assembly of information through droplet association. The associations of the discrete DNA droplets can provide more intelligent DNA information systems, such as DNA computing<sup>[35]</sup> and digital data storage<sup>[36]</sup> using DNA hybridization and enzymatic reactions. DNA information storage through hybridization and enzymatic reactions in the DCF can serve as a multi-bit DNA storage system<sup>[37,38]</sup> and as digital DNA molecular data storage.<sup>[39]</sup>

## 2.2. DNA Droplet Assembly using Splint Ligation

DNA ligation helps manipulate DNA molecular size and structure, and gel electrophoresis can display digitized DNA information, such as letter codes.<sup>[15,40,41]</sup> In particular, splint DNA ligation can aid in programmable assembly using sequence information. Furthermore, DNA droplet carriers can be used as a discrete handling tool.<sup>[31,42,43]</sup>

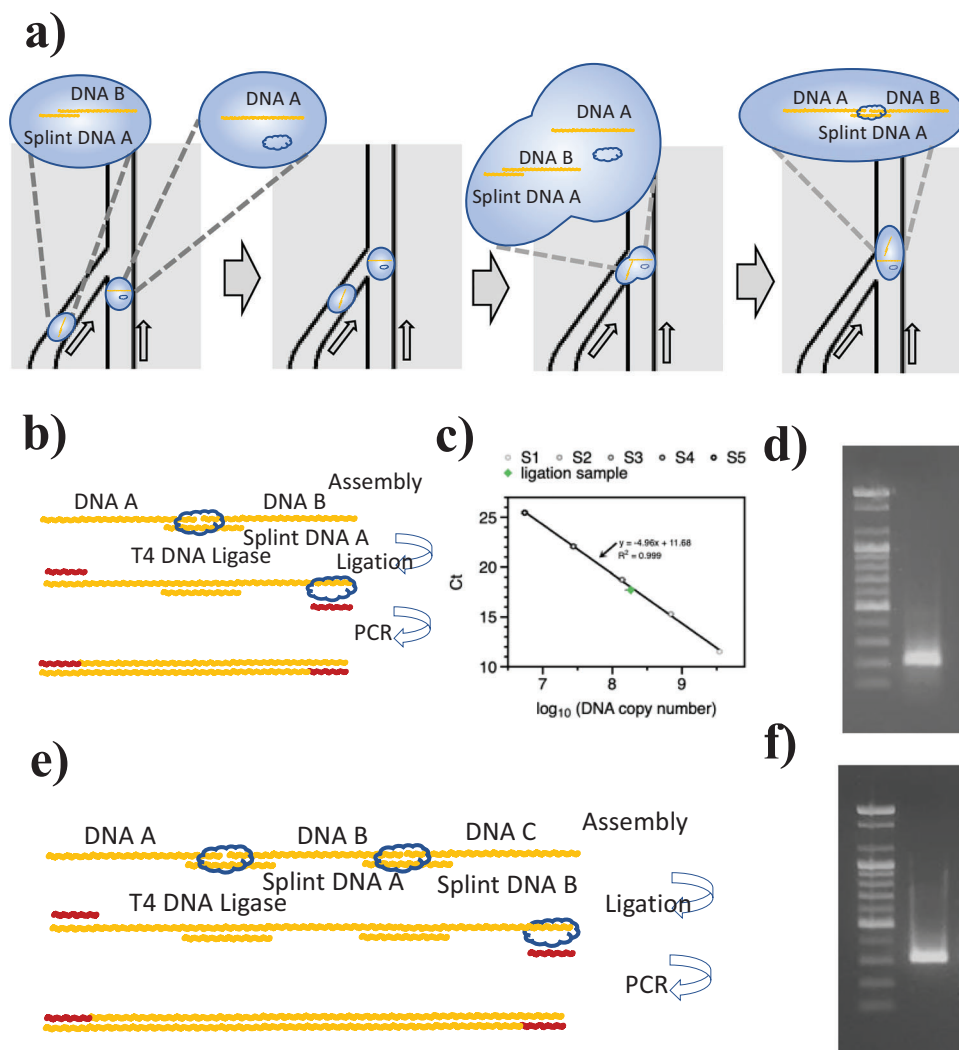
We demonstrated DNA droplet assembly with splint ligation in DCF for discrete programmable DNA information (Figure 3). A single droplet with DNA A (100 bps ssDNA) and T4 DNA ligase in the main channel moved into the y-junction. Another droplet with polynucleotide kinase-treated DNA B (100 bps ssDNA), which included a pre-assembled splint DNA sequence (30 bps ssDNA), was associated with the droplet in the main channel (Figure 3a). The two DNA sequences in the associated droplet were hybridized with splint DNA and ligated using T4 DNA ligase, which can be activated at room temperature<sup>[32]</sup> (25 °C) for

20 min. The linked DNA molecules are thermostable and permanently connected through ligation (Figure 3b). The 200 bps DNA fragment ligation was confirmed and amplified using PCR. The copy number of the ligated 200 bp DNA fragment was quantified using quantitative PCR (qPCR) to measure the efficiency of splint ligation (Figure 3c). Due to using a high amount of T4 DNA ligase (400 units), the ligation efficiency was 18.5%. This result was calculated using a copy number of  $1.67 \times 10^{12}$ , a decrease from the original copy number of  $9.04 \times 10^{12}$  (Figure S2, Supporting Information). Figure 3d shows the gel electrophoresis results of the ligated 200 bps DNA fragment. In the series process, three DNA fragments (A, B, and C) were ligated in the DCF system and verified through PCR. Figure 3f shows the gel electrophoresis results of the 300 bps DNA sequence, which has information tethered to DNA A, B, and C.

## 2.3. Multi-Bit Droplets DNA Ligation

For multi-bit DNA information, the discrete DNA bands in gel electrophoresis can conveniently display DNA information, such as alphanumeric characters.<sup>[15]</sup> The gel electrophoresis results of five bits of DNA information displayed different DNA gel bands from 100 to 500 bps induced by programmable ligation (Figure 4a). By combining ssDNA with splint DNA, which can be informationally tethered, stored digital DNA data can be displayed as binary data of 100 bps in length. Digital DNA data, such as alphanumeric characters encoded as the Baudot code,<sup>[15]</sup> can be represented by the ligated DNA, which are 100 bps fragments incorporated into LSB and 500 bps into MSB.<sup>[34–36,42]</sup> The ligation is because the address sequence of DNA is assembled with another DNA fragment by the associated splint DNA, which contains the address information. For four serial ligations, the efficiency is expected to be at 0.1% without calculating the increasing ligation time for the DNA in the first droplet due to the low efficiency (18.5%) of a single splint ligation process. We have provided gel electrophoresis results with and without ligation and PCR to assess the feasibility of DNA association (Figure S3, Supporting Information). Lane 2 shows the assembled DNA bands with ligation, lane 3 represents the assembled DNA bands without ligation, lane 4 has a strong ligated DNA band, and lane 5 shows DNA bands of assembled DNA without ligation. The gel electrophoresis results show that PCR amplification can help assess whether the recovered ligated DNA fragments are high-quality.

Moreover, due to the addressed splint DNA, the alignment order of the DNA sequence can be changed in DCF. Figure 4b shows the 500 bps DNA gel bands from different DNA ligation combinations. Figure 4c,d shows Sanger sequencing results for DNA fragments in the series order of DNA A, B, C, D, and E (lane 2 in Figure 4b) and DNA fragments in the series order of DNA A, D, C, B, and E (lane 3 in Figure 4b). Furthermore, the NGS read results were analyzed and displayed in Figures S4 and S5 (Supporting Information). The sequencing data corresponded to the original DNA information, maintaining its fidelity and accuracy. More detailed information is provided in Supporting Information 5. The splint DNA ligation method created serial information-tethered DNA representing discrete DNA information to perform multi-bit assembly. Furthermore, droplet



**Figure 3.** DNA droplet assembly using splint ligation. a) The DNA droplet in the DCF system. b) The splint ligation process with DNA A and B. c) Standard qPCR curve for the 200-bps ligation sample (aliquoted to 1/40 000 for the droplet,  $n = 3$ ). d) The gel electrophoresis results of the splint DNA-ligated 200 bps fragment. e) The two series splint ligation experiments (DNA A, B, and C), and f) The gel electrophoresis results of the splint DNA-ligated 300 bps fragment.

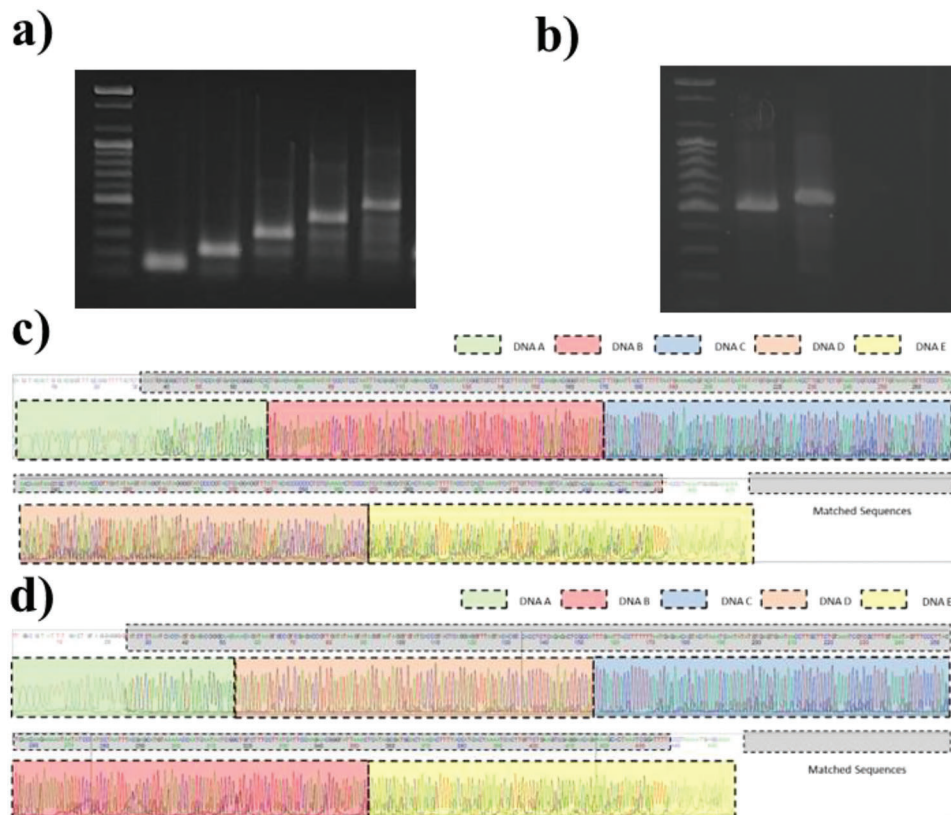
assembly in the DCF system can provide sequence-encoded DNA data storage.<sup>[28]</sup>

## 2.4. Programmable Assembly for DNA Word Data

The DNA word data sequences, encoded by ASCII,<sup>[38]</sup> were designed in orthogonal address primer sequences to exhibit programmable word information combinations. For error data correction (one character or seven bits), the word data sequences were encoded with a seven-bps ECC redundancy using the shortened RS code.<sup>[29]</sup> The Galois field was defined as  $2^7$  due to the symbol of the seven bits ASCII encoding method (0–127). Similar to previous multi-bit droplet DNA ligation processes, the DNA word data in the droplet can be assembled with splint DNA in DCF (Figure 5a). Figure 5b and Table S2 (Supporting Information) provide the word list and sequence information. The ECC

code and the encoding/decoding method are presented in more detail in Supporting Information 7.

Figure 5c shows the gel electrophoresis results to produce various DNA sentence data. The 203 bps DNA sentence data (lane A) fragment, synthesized using two splint ligation processes in DCF, represented “DNAdata Stores Apple.” The synthesized diverse DNA sentence data fragments represented “JamesWatson Likes Orange,” shown in lane B (220 bps), “FrancisCrick Likes Grape,” shown in lane C (220 bps), and “JohnVonNeumann Has Apple,” shown in lane D (217 bps). The DNA sentence data fragments were analyzed through NGS. The resultant sequences, trimmed with forward and reverse address sequences, were decoded to digital data using consensus sequences (Figure 5d,h; Figure S7, Supporting Information). The data sizes were not defined (the sequence length of the DNA data can be modulated, depending on data size); therefore, we searched for the data sequence positions and sizes using the address sequence



**Figure 4.** DNA ligation of multi-bits droplets. a) 100–500 bps discrete DNA band results. b) Two different 500 bps DNA fragments after droplet DNA assembly. c,d) Sanger sequencing results for DNA fragments in the series order of DNA A, B, C, D, and E (c), and DNA A, D, C, B, and E (d).

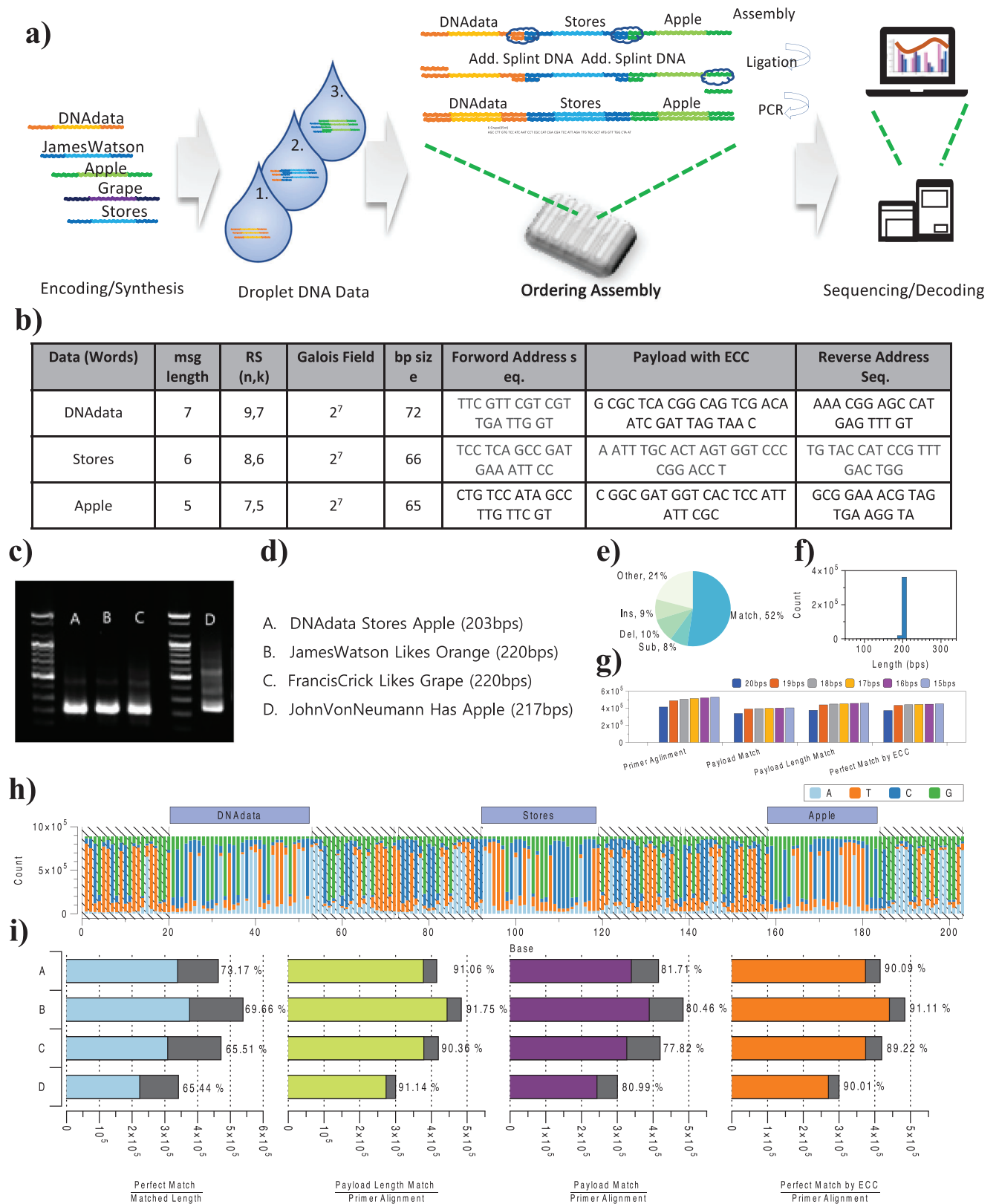
information to retrieve the data stored in DNA sequences. After the first 20 bps of the forward was scanned until 20th bases, the reverse address sequences were scanned from the 21st base to the end of the NGS sequences until the corresponding reverse address sequences were determined to identify the address sequences in NGS raw data. The protocol is detailed in “Alignment Using Address Sequences.” Figure 5e shows the analysis of the sequencing results for the DNA sentence data fragment in lane A. The degree of perfectly matched sequences was 52.4%, and the substitution error was 7.6%. One and two deletion errors were 10.2%, and the insertion error was 8.8% in the NGS raw data. Figure S6 (Supporting Information) provides the sequencing results of other DNA sentence data fragments. The most frequently read sequences were 203 bps (59.2%) in length (Figure 5f). Filters can be used to ensure highly reliable results, designed to match the exact address sequences (20 bps), albeit at the cost of reduced coverage. Figure 5g provides the increased number of counts for address sequence alignment, perfect matches, payload data length matches, and the results of applying ECC to the matched address sequence, which was 15–20 bps. Figures S7 and S9 (Supporting Information) show the consensus sequencing results, including payload data with and without address sequence alignment. A more detailed explanation and count results of other DNA sentence data are provided in Supporting Information 10.

The consensus sequencing result for the DNA sentence data “DNadata Stores Apple” is provided in Figure 5h. Owing to the

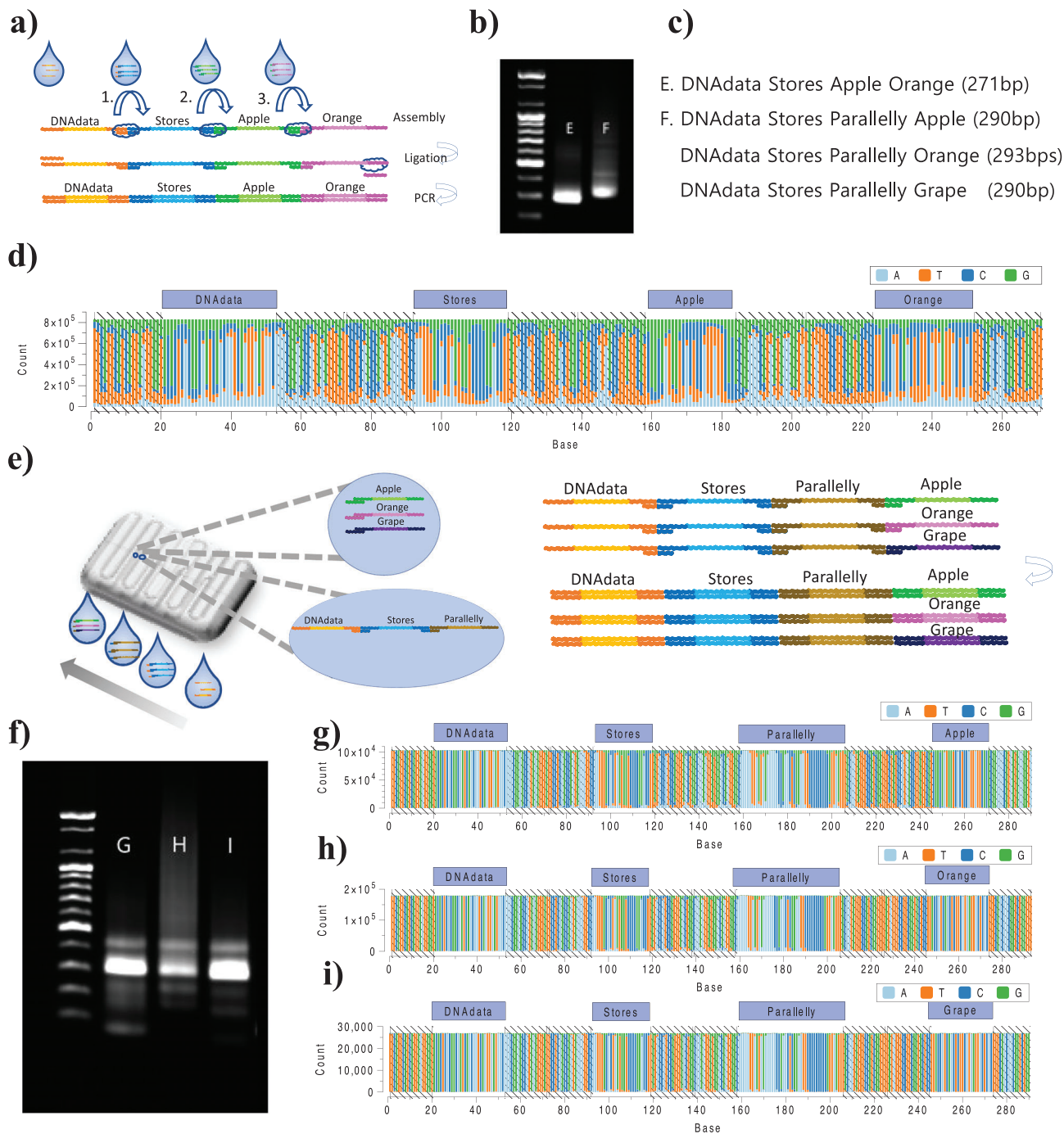
high decoding yield, including the high percentage of perfect match sequences, applying the ECC was unnecessary to recover the ACSII data. The sequencing results with ECC are analyzed in more detail in Figure 5i. The perfectly-matched DNA sequences comprise 73.17%, 69.66%, 65.51%, and 65.44% of lanes A–B, respectively. The payload matching the length in the DNA sentence data, resorted by address alignment, comprises 91.06%, 91.75%, 90.36%, and 91.14% of lanes A–D, respectively. Furthermore, we provided the number of matched sequence data encoded with an ECC (Figure S10, Supporting Information). The percentages of match sequences with ECC applied were increased from 81.71% to 90.09% in lane A, 80.46%–91.11% in lane B, 77.82%–89.22% in lane C, and 80.99%–90.01% in lane D. We conducted DNA synthesis of various sentence structures, using the DNA words. Their information conveyed the intended meaning even if the sentences were not grammatically correct.

## 2.5. Synthesizing Multiple DNA Data Pools

To process DNA data storage, the DNA droplet assembled using ligation in DCF has the benefits of programmable, serial, and parallel data assembly. The DNA word fragments were synthesized using a serial DNA data assembly to form the DNA sentence fragment (lane E of Figure 6b) for “DNadata Stores Apple Orange” (Figure 6a–c). The consensus NGS results of the DNA sentence data are shown in Figure 6d.



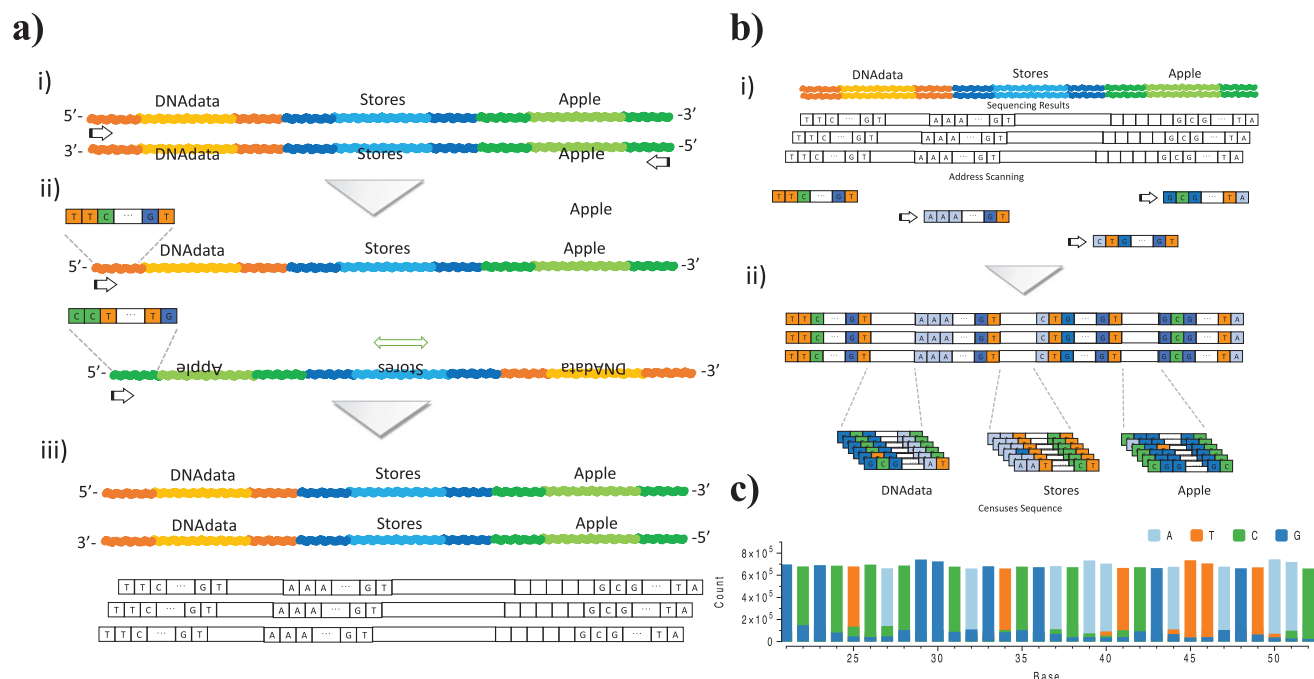
**Figure 5.** Programmable ordering assembly for sequence-encoded word DNA data. a) The synthesis and decoding process of droplet assembly. b) The sequence and encoded information of DNA words. c) Gel electrophoresis results. d) The lengths of DNA sentence data fragments. e–g) Sequencing analysis of lane A. h, i) The consensus sequencing results of lane A h), and the sequencing analysis results (i).



**Figure 6.** Serial and parallel DNA data storage. a) A diagram of the ligation of multiple serial word data DNA fragments. b) The gel electrophoresis results of serial and parallel DNA data storage. c) The length and data of sentence DNA fragments. d) The NGS results of serial DNA data storage. e) A diagram of parallel processing in DCF. f–i) The gel electrophoresis results for PCR-based random access of parallel DNA data storage (f), and the random-access NGS results using primer files (g–i).

The serially synthesized DNA “DNAdata Store Parallely” fragment was assembled with three different word data DNAs, “Apple,” “Orange,” and “Grape,” in a single droplet to split the synthesis for parallel DNA data storage processing (Figure 6e). The reverse address of the serially synthesized DNA fragment was assembled in parallel with three different splint DNAs, which were

hybridized DNA word data of “Apple,” “Orange,” and “Grape” (Tables S2 and S4, Supporting Information). The DNA pool, assembled in series and parallel in DCF, had three different sentence DNA data fragments, “DNAdata Stores Parallely Apple” (290 bps), “DNAdata Stores Parallely Orange” (293 bps), and “DNAdata Stores Parallely Grape” (290 bps) (lane F in Figure 6c).



**Figure 7.** The decoding process for programmable droplet DNA data storage assembly. a) The decoding process with re-alignment to the original information using the reverse primer information. b,c) Address sequence scanning (b), and the consensus sequencing of the DNA data (c).

We used two strategies to read the multiple DNA data pools: (1) scanning the address sequences to access the data pool *in silico* and (2) using a primer-based random-access approach. In the *in silico* process, the NGS reads of the multiple DNA data pools were scanned and sorted using all address sequences to recognize file information such as size and position and the number of entries (Figure S11, Supporting Information). Three split DNA sequences were decoded in series and parallel to recover the digital data. PCR was performed on the DNA data pools using address primers to select the DNA files. The NGS results were then decoded to recover the digital data. Figure 6f–i presents the gel electrophoresis results, which show random access results with address primers through PCR and NGS. The programmable ordering assembly of DNA data in DCF can be conveniently manipulated to represent diverse DNA data, owing to the serial and parallel synthesis, due to the assembly address sequences. Moreover, the address sequences can be widely used as alignment markers in the decoding process.

## 2.6. Alignment using Address Sequences

In digital DNA data storage, primer sequences are pre-defined information and can access DNA data in pools through PCR. Additionally, the primer address sequences, used for programmable data connections through splint ligation, can align with the DNA data file as a synchronization marker<sup>[39]</sup> to improve the data fidelity and define the data in series and parallel. The data size and position can easily recover original digital data.

Figure 7 shows the decoding protocol for digital DNA data synthesized through droplet DNA data assembly using multiple address information. The reverse data sequences in the syn-

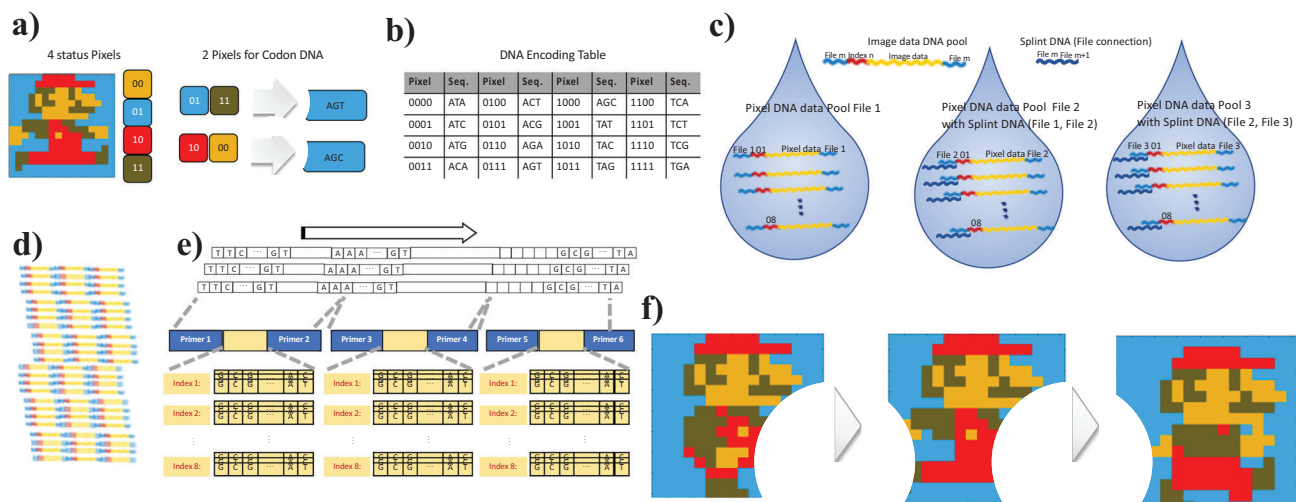
thesized DNA sentence, read by paired-end sequencing, were searched using reverse primer information and flipped original sequence information. The flipped data provided approximately twice the coverage with high fidelity.<sup>[44]</sup> Owing to the original ss-DNA assembly with the splint DNA direction, the flipped DNA sentence data was serially converted from the 3' to 5' direction (Figure 7a<sup>iii</sup>). The next step was to scan the address sequences (Figure 7b). After the DNA data size was assigned by scanning the last address sequence, the file information (positions and size of the data sequence) was searched and defined by scanning other address sequences. As the positions, data size, and order of DNA data were assigned by scanning the address sequences, the data sequences were statistically analyzed using consensus sequence and decoded original word data (Figure 7c), and the sentence data were represented (Figure 7b<sup>ii</sup>). Moreover, the decoding protocol can search the parallel assembled DNA data by scanning multiple address sequences.

The address sequence assembly and searching method can efficiently optimize the data length and increase the sequence fidelity by aligning the address sequences. Moreover, decoding *in de novo* sequencing can be performed rapidly owing to assigning the data position and flipping the original sequence using reverse sequences. Address alignment is a suitable method for droplet DNA assembly to analyze serial and parallel assembled DNA data information.

## 2.7. Droplet Assembly of the DNA Data Pool

Due to synthetic lengths, DNA data storage molecules are stored in parallel in a DNA pool with index sequences. The DNA pools with common address sequences can be processed using droplet





**Figure 8.** Programmable ordering assembly for DNA pools. a) The encoding process of four colors of pixel data. b) The two pixels one codon-based encoding table. c) DNA pool with splint DNA in the droplet. d) The diagram of the assembled movie chip data DNA. e, f) The schematic diagram of the decoding process (e), and the movie clip images (Video S1, Supporting Information) (f).

assembly in the DCF to manipulate the DNA pool files and create new information.

**Figure 8** shows the processing of parallel DNA data storage pools, in which a movie clip was created using three image files (0.27–0.28 kB) of  $16 \times 16$  pixels with four colors. A pixel image file was encoded to DNA sequences through a two-pixel per one codon code translation to prevent homopolymers, frequently repeated when the same pixel color is shown in a specific or background area.<sup>[25]</sup> Three sequences were allocated based on two pixels (four bits of data) (Figure 8a). The DNA sequences produced by the two-pixel-based one codon code (1.3 bits per bps) did not create more than two repeated homopolymers (Figure S12, Supporting Information). The encoding protocol is provided in more detail in Supporting Information 13. The encoded pixel image file was stored in parallel in the synthesized 90 bps DNA pool with a two-bps index and two 20 bps address sequences. Each DNA pool for the three “Super Mario” image files with splint DNAs in the droplet (Figure 8c) connected to create the movie clip file (12 kB). Figure 8e shows the parallel decoding process for aligning the data with the index sequences and constructing the image frame in each address sequence. After consensus sequencing, the DNA sequence reads were sorted using primers for the frames and by the index for the pixel order. The sequence data were decoded using binary information and arranged in  $16 \times 16$  matrixes. Although the four-color pixel images and movie clip files require pixel and color information for digital file conversion, the NGS result of the associated DNA pool was successfully decoded to produce the movie clip (Video S1, Supporting Information).

In our DCF system, three droplets containing 0.27–0.28 kB data in eight 90 bps oligos (33.7–35 B per oligo) created the movie clip file (12 kB) using linear droplet ligation. However, the droplet, which can contain  $9.04 \times 10^{12}$  copy oligos, has 303–315 GB storage capacity in the single droplet ( $\times 1000$  copies for ligation efficiency). The system could be expanded to logical droplet assembly, using logic circuit fluidics for DNA association<sup>[11]</sup> instead of linear droplet ligation.

### 3. Discussion

Digital DNA data storage is a unique research focus on next-generation information storage of materials because of its capacity for high-density information. DNA molecules have been developed with limited functions, such as allowing random access using address primer information, but methods for processing DNA data are lacking. In this study, we demonstrate the programmable assembly of DNA data storage using address information in the DCF system. DNA sequences of 100 bps were assembled into five discrete size-based DNA sequences using splint ligation to display the multi-bit DNA information and programmable order assembly according to the splint DNA information. For DNA processing, DNA word data encoded in the sequence were combined in series and parallel, and single and multiple sentence data DNA pools to produce new information. The synthesized DNA pool contains sentence information accessed through PCR or decoded using the address primer information. During decoding, the data size and position in the sequences were defined in series and parallel to recover the original digital data using the address information easily. The sequences were aligned and sorted to improve fidelity. Finally, we demonstrate the programmable assembly of DNA pools that encode pixel images. The DNA pools of image frames assembled in series were sequenced and decoded in parallel to produce movie clips.

DNA storage is a valuable medium that can provide beneficial tools for high-density<sup>[37,45]</sup> and long-term data storage<sup>[46,47]</sup> to replace traditional solid-state storage materials such as compact disks, hard disk drives, and solid-state drives. Furthermore, DNA, functioning as data-bearing molecules in suspension, can be manipulated to process digital DNA data storage using programmable DNA assembly in the DCF system. Specifically, 303–315 GB droplets can provide a vesicle for discrete DNA data with programmable associations in the system. The synthetic ssDNA pool can allow a more reasonable data-producing price and capacity.<sup>[37,48]</sup> Splint ligation enables sequence assembly,

including synthesizing multiple DNA data pools. Including our data capacity, the system could be expanded to logical droplet assembly, using logic circuit fluidics for DNA association<sup>[11]</sup> instead of linear droplet association. Even though we only demonstrated a few to tens kBs DNA data processing system using a 3 × 2-inch droplet fluidics chip system with individual syringe pumps, the system can be extended to a scalable data processing machine with optimized processes using a room-scale machine-based system, such as a CatalogDNA machine<sup>[49]</sup> and droplet fluidics system.<sup>[28,50]</sup>

In addition, address sequence searching can provide high-fidelity sequences and fast de novo sequence decoding and optimizes data density for DNA synthesis. We focused only on short-sequence DNA for NGS. The processes in the droplet chip system are not fully automated and include NGS protocols, such as PCR and A-tailing.<sup>[51]</sup> However, the serial assembly of parallel DNA pools in droplets could be expanded for long-sequence DNA data storage for Oxford nanopore sequencing, a real-time sequencing technology.<sup>[30]</sup> Furthermore, processing DNA using the droplet system can be extended to efficiently analyze large datasets in DNA data storage<sup>[1,52]</sup> and DNA computing.<sup>[3,17,24,53]</sup> Programmable DNA data assembly techniques in chips can potentially be used as valuable tools in memory computing and data storage for DNA through the enhanced development of DNA algorithms and discrete droplet systems.

## 4. Experimental Section

**The DCF System:** The fluidic mold was designed by AutoCAD 2020 (Autodesk, Inc., San Francisco, CA, USA) and printed using Form 2 (Formlabs, Inc., Somerville, MA, USA) at the Device Lab (Incheon, Republic of Korea). The PDMS was cast in the mold to create the fluidics chip. The DCF system was constructed manually using NE-1000 syringe pumps (New Era Pump Systems, Inc., Farmingdale, NY, USA) with a 1 ml syringe (BD, Inc., Franklin Lakes, NJ, USA). For the DNA droplet assembly process, DNA in the droplets of the main channel was mixed with T4 DNA ligase and ligase buffer (New England BioLabs, Inc., Notting Hill, Australia). The DNA molecular droplets were prepared by serially adding 2 μl of DNA data droplets with a 5 μl air gap in the capillary tube using a syringe pump. The droplet in the main channel was carefully moved and placed at the y-junction of the channel. The injected droplet at the junction was associated with the droplet in the main channel. The associated droplet was then moved to the next y-junction in the main channel to assemble the programmable discrete DNA droplet. The droplet assembly process was monitored using a Leica DMI8 inverted microscope (Leica Microsystems, Inc., Wetzlar, Germany). For DNA sequence ligation, the assembled droplets were incubated at room temperature for 20 min and moved to the next y-junction. After assembly, the T4 DNA ligase enzyme was deactivated at 65 °C for 10 min to provide a consistent processing time and prevent unknown enzyme reactions during ligation and PCR.

**DNA Preparation:** DNA sequences were designed using nupack.com and obtained from Integrated DNA Technologies, Inc. (Coralville, IA, USA) and Macrogen, Inc. (Seoul, Korea). All DNA sequences are listed in Supporting Information and DNA list file. For ligation, the 5' ends of DNA sequences were phosphorylated with T4 polynucleotide kinase (New England BioLabs, Inc.) The phosphorylated DNA sequences were hybridized with splint DNA in the annealing process, during which the temperature was decreased from 95 to 30 °C by 1 °C min<sup>-1</sup> in a SimpliAmp thermal cycler (ThermoFisher Scientific, Inc., Waltham, MA, USA). All DNA droplets were prepared to achieve a final concentration of 4 μM in 50 mM NaCl buffer (Samchun Pure Chemical Co., Seoul, Korea). In the 20 μl main droplet, the DNA samples were mixed with 400 U T4 DNA ligase (New

England BioLabs, Inc., Notting Hill, Australia) in 1× reaction buffer (New England BioLabs, Inc., Notting Hill, Australia).

**The Encoding and Decoding Processes:** The DNA data encoding and decoding processes were developed using Matlab 2021a (Mathworks, Inc., Natick, MA, USA). For the ASCII letter encoding of words, the text information could be represented as the ASCII code converted into binary information and stacked with RS ECC encoding for the seven bits error corrections. The binary information of the words was translated into the four DNA bases (A, T, C, and G). The encoded DNA letters were aligned with primer information for amplification and ligation. For two pixels, which were codon-based encoding of the pixel images, 16 × 16-pixel image frames, which have four color states per pixel, were decoded and linearly aligned. Each image frame was decoded using the same process. The four state colors were represented as 00, 01, 10, and 11. The binary information was translated by the codon DNA bases per two pixels. The encoded DNA sequences were fragmented with index sequences and aligned using primer information for DNA data oligo pools. The PIM DNA data, processed using NGS, were read and decoded by each decoding protocol. All DNA data and source codes are available in the GitHub (<https://github.com/bioelectronicsbiooptics/Droplet-DNA.git>) and NCBI (PRJNA955202) databases.

**PCR and Detection:** For conventional PCR, the samples with 300 nm of each primer (Integrated DNA Technologies, Inc., Coralville, IA, USA) were amplified in 0.2 mm dNTP (Ezymomics, Inc., Seoul, South Korea) and 1 U Taq polymerase (ThermoFisher Scientific, Inc., Waltham, MA, USA), using a SimpliAmp thermal cycler. For qPCR, the DNA samples were amplified with PowerUp SYBR Green Master Mix (Thermo Fisher Scientific, Inc., Waltham, MA, USA), using a QuantStudio 5 Real-Time PCR System (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The DNA and ligated DNA fragments were aliquoted to 1/40 000 to measure the fluorescent intensity using five standard samples diluted to 1/5 from 15.4 pg μl<sup>-1</sup> to 24.2 fg μl<sup>-1</sup>. For gel electrophoresis, the DNA samples were visualized with 5 × SYBR green (Thermo Fisher Scientific, Inc., Waltham, MA, USA) and run on 1.5% (w/v) agarose gel (Sigma–Aldrich Company Ltd., St. Louis, MO, USA) in 1 × Tris/borate/EDTA buffer at 100 V for 40 min. The gels were imaged using a LED-Transilluminator (Korea Ace Scientific Company Ltd., Seoul, South Korea) and an Amersham Typhoon 5 (Cytiva, Inc., Marlborough, MA, USA). Sanger sequencing and NGS were performed using an ABI PRISM 3730XL Analyzer and Illumina Miseq (Illumina, Inc., San Diego, CA, USA) from Macrogen, Inc.

**Statistical Analyses:** For qPCR analysis, five standard samples (n = 3) of double-stranded DNA were calculated using a DNA copy number calculator (<https://www.thermofisher.com/>). For the DNA quantity in the droplet, the copy numbers were multiplied by 40 000 aliquoted amounts.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (NRF-2018R1D1A1A02085457, NRF-2021R1C1C1009300, and NRF-2021R1A4A3031992).

## Conflict of Interest

Y.S. is the C.E.O. of Standard Bioelectronics. Co.

## Author Contributions

M.Y. and D.L. contributed equally to this work. M.Y. performed the experiment. D.L. analyzed and plotted data and revised the manuscript. J.K. encoded the data DNA sequences and decoded DNA sequencing files. Y.S. designed and supervised the experiment.

## Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Keywords

address primer scanning, DNA data storage, droplet-controlled fluidics system, processing of DNA storage, splint ligation

Received: May 18, 2023

Revised: July 11, 2023

Published online: September 26, 2023

- [1] Z. Ping, S. Chen, G. Zhou, X. Huang, S. J. Zhu, H. Zhang, H. H. Lee, Z. Lan, J. Cui, T. Chen, W. Zhang, H. Yang, X. Xu, G. M. Church, Y. Shen, *Nat. Comput. Sci.* **2022**, 2, 234.
- [2] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, E. Birney, *Nature* **2013**, 494, 77.
- [3] L. M. Adleman, *Science* **1994**, 266, 1021.
- [4] K. Chen, J. Kong, J. Zhu, N. Ermann, P. Predki, U. F. Keyser, *Nano Lett.* **2019**, 19, 1210.
- [5] G. D. Dickinson, G. M. Mortuza, W. Clay, L. Piantanida, C. M. Green, C. Watson, E. J. Hayden, T. Andersen, W. Kuang, E. Graugnard, R. Zadegan, W. L. Hughes, *Nat. Commun.* **2021**, 12, 2371.
- [6] K. Chen, J. Zhu, F. Bošković, U. F. Keyser, *Nano Lett.* **2020**, 20, 3754.
- [7] L. Ceze, J. Nivala, K. Strauss, *Nat. Rev. Genet.* **2019**, 20, 456.
- [8] A. Doricchi, C. M. Platnich, A. Gimpel, F. Horn, M. Earle, G. Lanzavecchia, A. L. Cortajarena, L. M. Liz-Marzán, N. Liu, R. Heckel, R. N. Grass, R. Krahe, U. F. Keyser, D. Garoli, *ACS Nano* **2022**, 16, 17552.
- [9] L. Qian, E. Winfree, J. Bruck, *Nature* **2011**, 475, 368.
- [10] S. A. Salehi, X. Liu, M. D. Riedel, K. K. Parhi, *Sci. Rep.* **2018**, 8, 8312.
- [11] W. Lee, M. Yu, D. Lim, T. Kang, Y. Song, *ACS Nano* **2021**, 15, 11644.
- [12] Y. Du, P. Peng, T. Li, *ACS Nano* **2019**, 13, 5778.
- [13] M. Castillo, *Am. J. Neuroradiol.* **2014**, 35, 1.
- [14] J. Koch, S. Gantenbein, K. Masania, W. J. Stark, Y. Erlich, R. N. Grass, *Nat. Biotechnol.* **2020**, 38, 39.
- [15] A. R. Chandrasekaran, O. Levchenko, D. S. Patel, M. MacIsaac, K. Halvorsen, *Nucleic Acids Res.* **2017**, 45, 11459.
- [16] L. Organick, B. H. Nguyen, R. McAmis, W. D. Chen, A. X. Kohll, S. D. Ang, R. N. Grass, L. Ceze, K. Strauss, *Small Methods* **2021**, 5, 2001094.
- [17] D. Deamer, M. Akeson, D. Branton, *Nat. Biotechnol.* **2016**, 34, 518.
- [18] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, K. Strauss, presented at Proc. of the 21st Int. Conf. on Architectural Support for Programming Languages and Operating Systems, ACM, Atlanta, GA, USA **2016**, pp. 637–649.
- [19] S. M. H. T. Yazdi, R. Gabrys, O. Milenkovic, *Sci. Rep.* **2017**, 7, 5011.
- [20] T. T. Nguyen, K. Cai, K. A. Schouhamer Immink, H. M. Kiah, presented at 2020 IEEE Int. Sympos. on Inform. Theory (ISIT), IEEE, Los Angeles, CA, USA **2020**, p. 694.
- [21] X. Li, M. Chen, H. Wu, *Briefings Bioinf.* **2023**, 24, bbac484.
- [22] P. M. Stanley, L. M. Strittmatter, A. M. Vickers, K. C. K. Lee, *Biotechnol. Adv.* **2020**, 45, 107639.
- [23] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, C. N. Takahashi, S. Newman, H.-Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze, K. Strauss, *Nat. Biotechnol.* **2018**, 36, 242.
- [24] J. L. Banal, T. R. Shepherd, J. Berleant, H. Huang, M. Reyes, C. M. Ackerman, P. C. Blainey, M. Bathe, *Nat. Mater.* **2021**, 20, 1272.
- [25] J. Kim, J. H. Bae, M. Baym, D. Y. Zhang, *Nat. Commun.* **2020**, 11, 5008.
- [26] K. Matange, J. M. Tuck, A. J. Keung, *Nat. Commun.* **2021**, 12, 1358.
- [27] Y. Song, S. Kim, M. J. Heller, X. Huang, *Nat. Commun.* **2018**, 9, 281.
- [28] S. Newman, A. P. Stephenson, M. Willsey, B. H. Nguyen, C. N. Takahashi, K. Strauss, L. Ceze, *Nat. Commun.* **2019**, 10, 1706.
- [29] I. S. Reed, G. Solomon, *SIAM* **1960**, 8, 300.
- [30] R. Lopez, Y.-J. Chen, S. Dumas Ang, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Seelig, K. Strauss, L. Ceze, *Nat. Commun.* **2019**, 10, 2933.
- [31] X. Zhang, T. Li, F. Liu, Y. Chen, J. Yao, Z. Li, Y. Huang, J. Wang, *Mol. Cell* **2019**, 73, 130.
- [32] J. Greene, in *Recombinant DNA Principles and Methodologies*, CRC Press, Boca Raton, FL, USA **1998**.
- [33] Y. Zhang, A. McMullen, L.-L. Pontani, X. He, R. Sha, N. C. Seeman, J. Brujic, P. M. Chaikin, *Nat. Commun.* **2017**, 8, 21.
- [34] A. Kulsoom, D. Xiao, A.-u. Rehman, S. A. Abbas, *Multimed. Tools Appl.* **2016**, 75, 1.
- [35] A. Rehman, D. Xiao, A. Kulsoom, M. A. Hashmi, S. A. Abbas, *Multimed. Tools Appl.* **2019**, 78, 9355.
- [36] A. Rehman, X. Liao, A. Kulsoom, S. A. Abbas, *Multimed. Tools Appl.* **2015**, 74, 4655.
- [37] Y. Erlich, D. Zielinski, *Science* **2017**, 355, 950.
- [38] D. Salomon, in *Data Compression: The Complete Reference*, Springer, London, **2007**.
- [39] S. Chandak, K. Tatwawadi, B. Lau, J. Mardia, M. Kubit, J. Neu, P. Griffin, M. Wootters, T. Weissman, H. Ji, presented at 2019 57th Ann. Allerton Conf. on Commun., Control, and Computing (Allerton), IEEE, Monticello, IL, USA **2019**, p. 147.
- [40] K. Halvorsen, W. P. Wong, *PLoS One* **2012**, 7, e44212.
- [41] N. G. Portney, Y. Wu, L. K. Quezada, S. Lonardi, M. Ozkan, *Langmuir* **2008**, 24, 1613.
- [42] Y. Wang, T. Cao, J. Ko, Y. Shen, W. Zong, K. Sheng, W. Cao, S. Sun, L. Cai, Y.-L. Zhou, X.-X. Zhang, C. Zong, R. Weissleder, D. Weitz, *Adv. Sci.* **2020**, 7, 1903463.
- [43] B. Zhu, Y. Zhao, J. Dai, J. Wang, S. Xing, L. Guo, N. Chen, X. Qu, L. Li, J. Shen, J. Shi, J. Li, L. Wang, *ACS Appl. Mater. Interfaces* **2017**, 9, 18434.
- [44] C. Wang, S. Krishnakumar, J. Wilhelmy, F. Babrzadeh, L. Stepanyan, L. F. Su, D. Levinson, M. A. Fernandez-Viña, R. W. Davis, M. M. Davis, M. Mindrinos, *Proc. Natl. Acad. Sci. USA* **2012**, 109, 8676.
- [45] T. Jeewandara, **2021**, <https://phys.org/news/2021-12-dna-storage-nanoscale-electrode-wells.html>.
- [46] C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland, *Science* **2001**, 293, 1763.
- [47] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspina, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, et al., *Science* **2010**, 328, 710.
- [48] M. Hao, H. Qiao, Y. Gao, Z. Wang, X. Qiao, X. Chen, H. Qi, *Commun. Biol.* **2020**, 3, 416.
- [49] <https://www.cambridgeconsultants.com/us/case-studies/catalog-dna-data-encoding-machine>.
- [50] C. N. Takahashi, B. H. Nguyen, K. Strauss, L. Ceze, *Sci. Rep.* **2019**, 9, 4998.
- [51] S. R. Head, H. K. Komori, S. A. LaMere, T. Whisenant, F. Van Nieuwerburgh, D. R. Salomon, P. Ordoukhanian, *BioTechniques* **2014**, 56, 61.
- [52] C. Bee, Y.-J. Chen, M. Queen, D. Ward, X. Liu, L. Organick, G. Seelig, K. Strauss, L. Ceze, *Nat. Commun.* **2021**, 12, 4764.
- [53] K. M. Cherry, L. Qian, *Nature* **2018**, 559, 370.