(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 May 2016 (19.05.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/077079 A1

(51) International Patent Classification:
G06F 21/00 (2013.01)

(21) International Application Number:
PCT/US2015/058120

(22) International Filing Date:
29 October 2015 (29.10.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/069,994 29 October 2014 (29.10.2014) US

(71) Applicant: **MASSACHUSETTS INSTITUTE OF TECHNOLOGY** [US/US]; 77 Massachusetts Avenue, Cambridge, MA 02139 (US).

(72) Inventors: **LU, Timothy, Kuan-Ta**; 1 Harvard Place, Charlestown, MA 02129 (US). **CARR, Peter, A.**; 33 Woburn Street, Medford, MA 02122 (US). **ZAKERI, Bijan**; 11 Overlook Ridge Drive, Apt. 309, Revere, MA 02151 (US).

(74) Agent: **VATLAND, Janice, A.**; Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210-2206 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

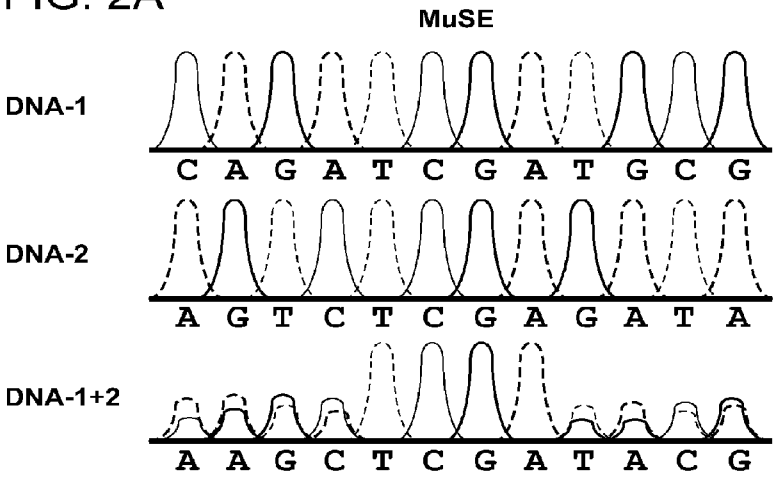
— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

(54) Title: DNA ENCRYPTION TECHNOLOGIES

FIG. 2A



(57) Abstract: In some aspects, the instant disclosure relates to the multiplexed encryption of information on nucleic acid molecules. In some aspects, the instant disclosure relates to a method of secure communication of information disseminated across at least one nucleic acid molecule, the method comprising (a) obtaining a modified keyboard comprising a personalized platform for translating text into a nucleic acid sequence; (b) translating a quantum of information into a nucleic acid message sequence using the modified keyboard of (a); and, (c) obtaining an at least one nucleic acid molecule, each molecule comprising: (i) the complete or a portion of the nucleic acid message sequence, and (ii) at least one contiguous stretch of randomized variable nucleic acid sequence flanking and/or inserted into the message sequence, thereby producing a nucleic acid molecule or a set of nucleic acid molecules containing the entire quantum of information.

WO 2016/077079 A1

DNA ENCRYPTION TECHNOLOGIES

RELATED APPLICATIONS

This application claims the benefit of U.S. provisional application serial number USSN 62/069,994, filed on October 29, 2014, and entitled "DNA Encryption Technologies", the entire content of which is incorporated herein by reference.

FEDERALLY SPONSORED RESEARCH

This invention was made with government support under Contract No. N66001-12-C-4016 awarded by the Space and Naval Warfare Systems Center. The government has certain rights in the invention.

BACKGROUND OF INVENTION

As the costs and time constraints of DNA synthesis and sequencing are rapidly declining, DNA is emerging as a viable medium for information storage. Previously, DNA has been used for hiding messages and storing large texts, however these methods require advanced laboratories with trained scientists to extract information. Simpler writing and reading methods are required for DNA communication to become more adopted.

SUMMARY OF INVENTION

In some aspects, the instant disclosure relates to a method of secure communication of information disseminated across at least one nucleic acid molecule, the method comprising (a) obtaining a modified keyboard comprising a personalized platform for translating text into a nucleic acid sequence; (b) translating a quantum of information into a nucleic acid message sequence using the modified keyboard of (a); and, (c) obtaining an at least one nucleic acid molecule, each molecule comprising: (i) the complete or a portion of the nucleic acid message sequence, and (ii) at least one contiguous stretch of randomized variable nucleic acid sequence flanking and/or inserted into the message sequence, thereby producing a nucleic acid molecule or a set of nucleic acid molecules containing the entire quantum of information. In some embodiments, the nucleic acid molecules are naturally-occurring. In some embodiments, the nucleic acid molecules are synthesized or non-naturally occurring. In some embodiments, the sequences of the nucleic acids are naturally-occurring. In some embodiments, the sequences of the nucleic acid molecules are synthesized or non-naturally occurring. In some embodiments, the

modified keyboard comprises codons. In some embodiments, the codons are designed to normalize frequency of character usage.

In some aspects, the instant disclosure relates to a method of secure communication of information contained on a single nucleic acid molecule, the method comprising (a) obtaining a nucleic acid molecule of known sequence; (b) obtaining a modified keyboard comprising a personalized platform for translating nucleic acid sequence into text; and, (b) generating a quantum of information translated from the nucleic acid sequence using the modified keyboard of (a). In some embodiments, the modified keyboard comprises codons. In some embodiments, the codons are designed to normalize frequency of character usage.

In some embodiments, the method further comprises co-sequencing the set of nucleic acid molecules using one or more common primers. In some embodiments, the co-sequencing produces patterns in a chromatogram. In some embodiments, the method further comprises identifying nucleic acid sequence corresponding to areas of high intensity peaks on the chromatogram. In some embodiments, the method further comprises identifying nucleic acid sequence corresponding to areas of low intensity peaks on the chromatogram. In some embodiments, co-sequencing produces no chromatogram pattern. In some embodiments, the method further comprises identifying nucleic acid sequence using sequence alignments generated by bioinformatics software. In some embodiments, the method further comprises extracting the quantum of information contained within the set of nucleic acid molecules by using the modified keyboard to translate the nucleic acid sequence from the one or more nucleic acid molecules.

In some embodiments, the modified keyboard comprises homopolymer codons. In some embodiments, the keyboard comprises homopolymer codons located on functional keys. In some embodiments, the codons are greater than 3 nucleotides in length. In some embodiments, the codons are 4, or 5, or 6, or 7, or 8, or 9, or 10, or 11, or 12, or 13, or 14, or 15, or 16, or 17, or 18 nucleotide bases in length. In some embodiments, the codons are of mixed lengths. In some embodiments, the variable nucleic acid sequence comprises contiguous homopolymer codons.

In some embodiments, the instant disclosure relates to methods of extracting a quantum of encrypted information from a plurality of nucleic acid molecules. In some embodiments, the encrypted information is extracted by nucleic acid sequencing. In some embodiments, the nucleic acid sequencing is co-sequencing. In some embodiments, the co-sequencing is DNA co-

sequencing. In some embodiments, the DNA co-sequencing is performed by Sanger sequencing. In some embodiments, the plurality of nucleic acid molecules are sequenced with at least one common primer. In some embodiments, data produced from nucleic acid sequencing is analyzed by sequence alignment. In certain embodiments, the nucleic acid molecule(s) are in silico.

In some aspects, the instant disclosure relates to a method of producing an individualized keyboard for the conversion of plaintext into nucleic acid encodable language, the method comprising: (a) producing a library of codons; (b) assigning each member of the library to a different symbol; and, (c) arranging the symbols into an array, thereby producing an individualized keyboard. In some embodiments, the codons of the library are greater than three nucleotide bases in length. In some embodiments, the codons of the library are 4, or 5, or 6, or 7, or 8, or 9, or 10, or 11, or 12, or 13, or 14, or 15, or 16, or 17, or 18 nucleotide bases in length. In some embodiments, the codons of the library are of mixed lengths. In some embodiments, the symbol is selected from the group consisting of letter, number, word, punctuation mark or pictogram, logogram and/or any other relevant references to linguistic principles of different languages.

BRIEF DESCRIPTION OF DRAWINGS

Figures 1A-1C depict one embodiment of the iKey platform. Figure 1A depicts a graphical representation of one embodiment of an iKey-64, used to convert plaintext to codons for DNA transcription. Messages begin with 'start', finish with 'end', 'forward' and 'reverse' provide information on the strand containing the desired message, and 'space1' and 'space2' can be used to produce troughs in chromatograms. Codons can be randomized to produce one-time iKeys. Figure 1B shows that in this embodiment, iKey-64 buttons and codons were numbered to transcribe the keyboard on to a single strand of DNA (SEQ ID NO: 24). Figure 1C depicts this embodiment of iKey-64 transcribed on DNA (SEQ ID NO: 1). Codons were flanked by 10 Ts (SEQ ID NO: 1) to separate the start and end of the keyboard from surrounding DNA for identification.

Figures 2A-2E depict chromatogram patterning with Multiplexed Sequence Encryption (MuSE). Figure 2A depicts a schematic for chromatogram patterning. When two DNA strands are co-

sequenced, different overlapping nucleotides produce small peaks while identical ones produce large peaks. Peaks are kept in alignment via iKey-64. In Figure 2A, SEQ ID NOs: 48 through 50 appear from top to bottom, respectively. Figure 2B depicts a schematic demonstrating 'Massachusetts Institute Technology' being patterned with MuSE and iKey-64. Figure 2C depicts the sequence of 'Massachusetts Institute Technology used in Figure 2B. In Figure 2C, SEQ ID NOs: 51 and 52 appear from top to bottom, respectively. Figure 2D shows DNA-1+2 are co-sequenced at equal concentrations with a common primer (arrows), chromatogram patterning is achieved during reverse ($Primer_{ExternalRv}$) but not forward ($Primer_{ExternalFw}$) sequencing due to the flanking variable DNA regions. Figure 2E shows that chromatogram patterning can be tuned by varying the ratios of DNA-1 (light shading) and DNA-2 (dark shading).

Figures 3A-C show that chromatogram patterning requires the alignment of base calls to be maintained during co-sequencing of DNA strands. Figure 3A shows a close-up of the chromatograms for forward; the consensus sequence listed below the alignment is represented by SEQ ID NO: 25. Figure 3B shows a close-up of the chromatograms for reverse sequencing of DNA-1+2 encoding the MIT cipher shown in Figure 2D; the consensus sequence listed below the alignment is represented by SEQ ID NO: 26. Samples were co-sequenced at equal concentrations and the arrow depicts the sequencing primer. Figure 3C shows the sequence of upstream (SEQ ID NOs: 14-15) and downstream (SEQ ID NOs: 16-17) variable DNA regions from Figure 2B.

Figure 4 shows that MuSE can be tuned to discreetly encode messages in a mixed DNA population. By varying the ratios of DNA-1 (light shading) and DNA-2 (dark shading), the degree of chromatogram patterning can be tuned (Figure 2E). When one partner is present at a lower concentration chromatogram patterning is still achieved, however the resulting chromatogram would align perfectly with the more concentrated partner. Therefore, messages may be discreetly encoded between multiple DNA strands and revealed in chromatograms, but not identified by sequence alignments. Left: alignment of chromatograms from Figure 2E with DNA-1. Right: alignment of chromatograms from Figure 2E with DNA-2.

Figure 5 shows discreetly embedded messages in chromatograms. A close-up of chromatogram patterns formed with MuSE tuning (Figure 2E). Message encoding regions (shaded box) contain single peaks while variable DNA regions (unshaded box) contain two overlapping peaks whose heights can be adjusted by varying the ratios of DNA-1 (SEQ ID NO: 2) and DNA-2 (SEQ ID NO: 3). The portions of DNA-1 and DNA-2 that are shown in the alignment are represented by SEQ ID NO: 53 and SEQ ID NO: 54.

Figures 6A-6B show a combinatorial cipher depicting a WWII communication. Figure 6A shows that one embodiment of iKey-64 was used to transcribe watermarks, a key, a cipher, and a decoy message between 6 DNA strands. If the strands are sequenced according to the key (Pascal's triangle on left) with the appropriate primers, then the correct communication would be revealed. Figure 6B shows the chromatograms of an n1 x n6 matrix of strands tuned and co-sequenced with $Primer_{Cipher}$. Chromatogram patterning is not achieved when incorrect pairs are co-sequenced.

Figure 7 shows combinatorial cipher readouts from the WWII communication of Figures 6A-6B. Tuning and co-sequencing of multiple DNA strands reveals a variety of messages depending on the primers used and the order of strands co-sequenced.

Figure 8 shows that the combinatorial cipher of Figures 6A-6B does not produce chromatogram patterning if non-specific primers are used for co-sequencing. Co-sequencing of cipher and decoy message containing pairs at equal concentrations with non-specific primers that are common to all strands ($Primer_{ExternalFw/Rv}$) that bind outside of the information containing 525-bp region (Figure 6A) does not produce chromatogram patterning.

Figures 9A-9G show an examination of the peaks produced during co-sequencing of the combinatorial WWII cipher of Figures 6A-6B. Figure 9A shows DNA sequencing information (SEQ ID NOs: 27-29) and close-up chromatogram for the Key. Figures 9B-9D show DNA sequencing information (SEQ ID NOs: 30-38) and close-up chromatogram for the Cipher. Figures 9E-9G show DNA sequencing information (SEQ ID NOs: 39-47) and close-up chromatogram for the Decoy message.

Figure 10 shows a 256 button iKey for introducing redundancies for transcribing plaintext in to a DNA encodable format. This is a theoretical design for an iKey-256 based on a four-nucleotide codon. While it is not designed to produce chromatogram patterning, iKey-256 would introduce redundancies in the transcription of plaintext on to DNA by equaling the frequencies of buttons for the letters used in English (Table 2). Increased number of 'start', 'end', 'shift', and 'space' buttons were implemented to reduce the overuse of any individual codon. To highlight the start and end of any message from the surrounding DNA, all 5 'start' and 'end' codons may be used together to identify messages written within even a genome. Furthermore, a '|' button was introduced to replace all punctuation characters as offline communication by DNA need not abide by grammatical rules.

Figures 11A-11B show DNA-based communication. Figure 11A provides an example of NDA communication in which for Alice to send a message (*m*) to Bob, she must first write the data into DNA and then physically send the DNA to Bob, who can read the DNA and extract the data. Eve, who is eavesdropping, can physically intercept and read *m*, making the communication channel unsecure. Three areas that can improve communication between Alice and Bob include data encoding, data transfer, and data extraction. Figure 11B provides an example of improved DNA communication. Data encoding: *m* can be mixed with decoy (*d*) data and fragmented, then written into DNA with one-time pad encryption, where the key (*k*) can itself be written in DNA. Data transfer: DNA encoded *k* and fragmented *m+d* components can be transmitted between Alice and Bob using multiple different channels based on a secret-sharing system. Interception of an incomplete set of DNA communications by Eve will not provide the data in *m*. Data extraction: chromatogram patterning can be used by Bob to rapidly extract data via multiplexed sequencing reactions.

Figures 12A-12C show naïve co-sequencing of multiple DNA strands. Figure 12A shows DNA-1 (top), n1(second from top), and iKey-64 (third from top) strands have different sequences but they all share a common upstream region and sequencing primer ($Primer_{ExternalFw}$). Individual sequencing of each strand produces high quality reads, but the resulting reads are of poor quality when two (*e.g.*, DNA-1 and n1) or three (*e.g.*, DNA-1, n1, and iKey64) strands are co-sequenced. Figure 11B depicts a close-up of the chromatogram of DNA-1 (SEQ ID NO: 2) and

n1 (SEQ ID NO: 4) co-sequencing. Figure 11C depicts a close-up of the chromatogram of DNA-1, n1, and iKey64 co-sequencing (SEQ ID NOs: 2, 4 and 1, respectively).

Figure 13 shows an example of a workflow of extracting the correct message from a DNA communication that incorporates the iKey, MuSE, and chromatogram patterning techniques. Workflow steps 1, 2, and 3 can be viewed in detail in Figures 6A-6B and Figure 14. Data containing strands are pooled and sequenced with $Primer_{Key}$ to reveal the combination key. Deciphering and unlocking of the combination key will reveal the correct strand pairs to analyze with $Primer_{Message}$ to reveal the message. Analysis of incorrect strand pairs will reveal a decoy communication.

Figure 14 shows an example of a combinatorial message depicting a WWII communication. iKey-64 (Encryption Key) was used to write watermarks, a key, a message, and a decoy between six DNA strands (Secret-Sharing System). If strands are sequenced according to the Combination Key—obtained from Pascal's triangle—with the appropriate primers, then the correct communication is revealed.

Figure 15 shows an example of DNA camouflage. The 525 bp information-encoding regions of DNA were flipped between the forward and reverse strands to provide a camouflage effect against sequencing with random primer ($Primer_{ExternalFw/Rv}$). While the external DNA regions surrounding the information containing regions were identical, strands n1/n3/n5 were encoded in the forward direction and strands n2/n4/n6 in the reverse direction, with watermarks used for orientation.

Figures 16A-16C show an example of next-generation sequencing of a communication disseminated across six DNA strands. Figure 16A shows plasmids containing n1, n2, n3, n4, n5, and n6 sequences (Figure 15) were grown and purified in dH_2O, mixed at equal concentrations of 30 ng/μL, and submitted to an outside party for NGS sequencing and assembly under blind experimental conditions. Figure 15B shows 300 ng of plasmids containing n1, n2, n3, n4, n5, and n6 sequences run on a 1% agarose gel to demonstrate purity. Figure 16C shows the outside party was provided with the number of plasmids, vector sequences, and the size of messages inserted into the vectors and asked to assemble the messages encoded in the plasmids. They

assembled 6 sequences (Table 5) that represent the messages n1, n2, n3, n4, n5, and n6. Here the alignment of the 6 assembled sequences with n1, n2, n3, n4, n5, and n6 are shown. Shown below the alignment is a legend for the color-coding of the templates. Boxes highlight assembled sequences with near perfect alignment to corresponding templates.

DETAILED DESCRIPTION OF INVENTION

In some embodiments, methods are provided herein for the storage, transfer and retrieval of encrypted information within at least one nucleic acid molecule In some aspects, the instant disclosure relates to a method of secure communication of information disseminated across at least one nucleic acid molecule, the method comprising (a) obtaining a modified keyboard comprising a personalized platform for translating text into a nucleic acid sequence; (b) translating a quantum of information into a nucleic acid message sequence using the modified keyboard of (a); and, (c) obtaining at least one nucleic acid molecule, each molecule comprising: (i) the complete or a portion of the nucleic acid message sequence, and (ii) at least one contiguous stretch of randomized variable nucleic acid sequence flanking and/or inserted into the message sequence, thereby producing a nucleic acid molecule or a set of nucleic acid molecules containing the entire quantum of information. In some embodiments, the nucleic acid molecules are naturally-occurring. In some embodiments, the nucleic acid molecules are synthesized or non-naturally occurring. In some embodiments, the sequences of the nucleic acids are naturally-occurring. In some embodiments, the sequences of the nucleic acid molecules are synthesized or non-naturally occurring.

In some aspects, the instant disclosure relates to a method of secure communication of information contained on a single nucleic acid molecule, the method comprising (a) obtaining a nucleic acid molecule of known sequence; (b) obtaining a modified keyboard comprising a personalized platform for translating nucleic acid sequence into text; and, (b) generating a quantum of information translated from the nucleic acid sequence using the modified keyboard of (a).

In certain aspects, the instant disclosure relates to the use of a keyboard to encrypt text information into nucleic acid sequence. For example, the keyboard can be a modified keyboard, in which the keys are modified relative to a standard "QWERTY" keyboard such that each key

corresponds to specific combination of nucleotides. In some embodiments, the modified keyboard is used as a "one-time pad". As used herein, a "one-time pad" refers to a device for the encryption of information, wherein each character of a plaintext (*e.g.*, information) is encrypted by combining it with the corresponding bit or character of a single-use, random, secret pad or key (*e.g.*, a modified keyboard) using modular addition. In some embodiments, the keyboard disclosed herein is a physical keyboard comprising a set of keys, wherein each key is associated with a particular codon. In some embodiments, the modified keyboard comprises homopolymer codons. In some embodiments, the keyboard comprises homopolymer codons located on functional keys. In some embodiments, homopolymer codons are associated only with functional keys. As used herein, a "functional key" refers to a key that does not translate a letter, number, word, punctuation mark or pictogram, logogram and/or any other relevant references to linguistic principles of different languages. In some embodiments, the keyboard is a virtual keyboard comprising a set of keys, wherein each key is associated with a particular codon. As used herein, a "virtual keyboard" is a keyboard appearing on a computer screen, the keys of which may be activated by a user clicking a mouse or contacting a touch screen. In some aspects, the instant disclosure relates to a method of producing an individualized keyboard for the conversion of plaintext into nucleic acid encodable language, the method comprising: (a) producing a library of codons; (b) assigning each member of the library to a different symbol; and, (c) arranging the symbols into an array, thereby producing an individualized keyboard. In some embodiments, the codons of the library are three nucleotide bases in length, such as those depicted in Figure 1A. In some embodiments, the codons of the library are greater than three nucleotide bases in length. In some embodiments, the codons of the library are 4, or 5, or 6, or 7, or 8, or 9, or 10, or 11, or 12, or 13, or 14, or 15, or 16, or 17, or 18 nucleotide bases in length. In some embodiments, the codons of the library are of mixed lengths. In some embodiments, the symbol is selected from the group consisting of letter, number, word, punctuation mark or pictogram, logogram and/or any other relevant references to linguistic principles of different languages.

As used herein "nucleic acid" refers to a DNA or RNA molecule. Nucleic acids are polymeric macromolecules comprising a plurality of nucleotides. In some embodiments, the nucleotides are deoxyribonucleotides or ribonucleotides. In some embodiments, the nucleotides comprising the nucleic acid are selected from the group consisting of adenine, guanine, cytosine, thymine, uracil and inosine. In some embodiments, the nucleotides comprising the nucleic acid

are modified nucleotides. Methods of modifying nucleotides are generally known in the art. Non-limiting examples of nucleotide modifications include phosphorothioate backbone modifications, 2'-O-methyl group sugar modifications and the substitution of non-naturally occurring nucleotide bases (for example, nucleotides derivatized at the 5-, 6-, 7- or 8-position). In some embodiments, the nucleotide modification is fusion of DNA terminal ends with at least one protein. In some embodiments, the nucleic acids of the instant disclosure are natural. Non-limiting examples of natural nucleic acids include genomic DNA, and plasmid DNA. In some embodiments, the nucleic acids of the instant disclosure are synthetic. As used herein, the term "synthetic nucleic acid" refers to a nucleic acid molecule that is constructed via the joining nucleotides by a synthetic or non-natural method. One non-limiting example of a synthetic method is solid-phase oligonucleotide synthesis. In some embodiments, the nucleic acids of the instant disclosure are isolated.

Aspects of the instant disclosure relate to the translation of information into nucleic acid sequence. In some embodiments, the amount of information to be translated into nucleic acid sequence may be measured as a quantum. As used herein, a "quantum of information" refers to a pre-determined amount of information that is expressed in the appropriate unit. Non-limiting examples of appropriate units include characters, letters, words, phrases, sentences, numbers and symbols. In some embodiments, nucleic acid sequence that comprises translated information is referred to herein as "nucleic acid message sequence". In some embodiments, information may be translated into nucleic acid sequence using codons. As used herein, "codon" refers to a group of consecutive nucleotides that form a single unit of genetic code. Naturally-occurring codons are three nucleotides in length and represent the 20 common amino acids used to build proteins. In some embodiments, the codons used to translate information into DNA sequence are naturally-occurring codons that comprise three nucleotides. In some embodiments, the codons used to translate information into DNA sequence are greater than 3 nucleotides in length. In some embodiments, the codons are 4, or 5, or 6, or 7, or 8, or 9, or 10, or 11, or 12, or 13, or 14, or 15, or 16, or 17, or 18 nucleotide bases in length. In some embodiments, the codons are of mixed lengths. Also contemplated herein is the use of homopolymer codons. The term "homopolymer" describes a codon consisting essentially of a homogenous population of nucleotides. In some embodiments, homopolymer codons may be represented by the formulae including but not limited to $[A]_n$,$[C]_n$, $[G]_n$, $[T]_n$, $[U]_n$ and $[I]_n$, wherein n is an integer representing the length of the codon. Further non-limiting examples of homopolymer codons

include AAA, GGG, CCC, TTT, GGG, UUU, III, AAAA, GGGG, TTTT, CCCC, UUUU, and IIII. In some embodiments, the modified keyboards disclosed herein comprises homopolymer codons. In some embodiments, the homopolymer codons are located on the functional keys of a modified keyboard.

In some aspects, the instant disclosure relates to methods of secure communication of information by translation of said information into nucleic acid sequence. In some embodiments, the nucleic acid sequence is natural or naturally-occurring. In some embodiments, the nucleic acid sequence is synthetic or synthesized. In order to further obscure the identity of translated information, the translated information may be camouflaged within larger fragments of natural genomic or plasmid nucleic acid sequence, or variable nucleic acid sequence, to produce an encrypted nucleic acid molecule. In some embodiments, the synthesized nucleic acid molecules comprise nucleic acid message sequence and at least one contiguous stretch of randomized variable nucleic acid sequence. In some embodiments, the synthesized nucleic acid molecules comprise nucleic acid message sequence and no randomized variable nucleic acid sequence. As used herein "variable" refers to randomized nucleic acid sequence that does not comprise nucleic acid message sequence. In some embodiments, variable DNA sequence camouflages information translated into nucleic acid sequence by disrupting the fidelity of base calling during nucleic acid sequencing. In some embodiments, the variable nucleic acid sequence of the instant disclosure comprises one or more homopolymer codons. In some aspects, the presence of homopolymer codons in variable nucleic acid sequence causes an intentional misalignment of nucleic acid sequences during sequence analysis. Such misalignment may be useful in disguising the location of the encrypted information.

In some embodiments, the instant disclosure relates to methods of extracting a quantum of encrypted information from a one or more of nucleic acid molecules. In some embodiments, the encrypted information is extracted by nucleic acid sequencing. In some embodiments, the nucleic acid sequencing is co-sequencing. In some embodiments, the co-sequencing is DNA co-sequencing. In some embodiments, the DNA co-sequencing is performed by Sanger sequencing. Other non-limiting methods of DNA co-sequencing include Maxam-Gilbert sequencing, bridge PCR, nanopore sequencing and Next Generation Sequencing (*e.g.*, Single-molecule real-time sequencing, Ion Torrent sequencing, pyrosequencing, Illumina sequencing, sequencing by ligation (SOLiD)). In some embodiments, the plurality of nucleic acid molecules

are sequenced with at least one common primer. In some embodiments, the plurality of nucleic acid molecules are sequenced with 2, or 3, or 4, or 5, or 6, or 7, or 8, or 9, or 10 common primers.

In some embodiments, the method further comprises co-sequencing the set of nucleic acid molecules using one or more common primers to produce a chromatogram. A "chromatogram" refers to a visual representation of a DNA sample produced by a sequencing machine. Chromatograms depict a sequence of nucleic acid base calls as a series of peaks along a histogram. In some embodiments, the method described herein further comprises identifying information translated into nucleic acid sequence corresponding to areas of high intensity peaks on the chromatogram. In some embodiments, the method further comprises identifying nucleic acid sequence corresponding to areas of low intensity peaks on the chromatogram. In some embodiments, nucleic acid sequencing produces no chromatogram pattern. In some embodiments, the method further comprises identifying nucleic acid sequence using sequence alignments generated by bioinformatics software. In some embodiments, the method further comprises extracting the information contained within a single nucleic acid molecule or the set of nucleic acid molecules by using the modified keyboard to translate the nucleic acid sequence from the at least one nucleic acid molecule.

In some embodiments, the nucleic acid sequences and molecules described herein are in silico. As used herein, the term "in silico" refers to nucleic acid sequences or molecules produced by means of computer modeling or computer simulation. Without being bound by any particular theory, the instant disclosure contemplates the utility of in silico nucleic acid sequences and molecules for the nucleic acid encryption methods described herein. In some embodiments, in silico nucleic acid molecules or sequences may be encrypted using the methods described herein. In some embodiments, encrypted in silico nucleic acid molecules or sequences are useful for the archiving and protection of digital data.

EXAMPLES

Example 1: Materials and Methods

Plasmids

Constructs were cloned using standard molecular biology techniques, where KOD Hot Start DNA Polymerase (VWR) was used for all PCRs with primers from IDT. Synthetic DNA sequences were purchased as gBlocks from IDT (Table 1) and assembled with PCR amplified

p15A origin and chloramphenicol resistance gene fusions using Gibson assembly with 25 bp sequence overlaps, either with a commercial kit (NEB) or homemade mixture[24], and transformed in to *E. coli* DH5αPRO (F^- φ80*lac*ZΔM15 Δ(*lac*ZYA-*arg*F)U169 *deo*R *rec*A1 *end*A1 *hsd*R17(rk^-, mk^+) *pho*A *sup*E44 *thi*-1 *gyr*A96 *rel*A1 λ^-, PN25/tet^R, Placiq/lacI, Sp^r). Random DNA sequences were generated at http://www.bioinformatics.org/sms2/random_dna.html. All constructs were sequence verified by Genewiz Inc. (Cambridge, MA).

Sequencing

All constructs (Table 1) were purified using Qiagen kits and stored in cell culture grade water (Cellgro). Constructs were diluted to a final concentration of 30 ng/μL and sent for sequencing at indicated concentrations. $Primer_{ExternalFw}$ (GACATTAACCTATAAAAATAGGC) (SEQ ID NO: 10), $Primer_{ExternalRv}$ (GCATCTTCCAGGAAATCTC) (SEQ ID NO: 11), $Primer_{Key}$ (TAATACGACTCACTATAGGG) (SEQ ID NO: 12), and $Primer_{Cipher}$ (GCTAGTTATTGCTCAGCGG) (SEQ ID NO: 13) were used for all sequencing reactions as indicated. Sequencing reactions were all performed by Genewiz Inc. (Cambridge, MA) under 'Difficult Template' settings to ensure stringent sequencing conditions were employed. All sequencing reactions were performed in triplicate. Genewiz Inc. was not consulted prior, during, or after this study and all Sager sequencing reactions were performed under blind conditions by Genewiz Inc. to ensure bias was not introduced in the results. Geneious Pro 5.5.8 was used to analyze chromatograms, perform ClustalW alignments, and produce figures.

Table 1: DNA Constructs

Construct	Plasmid	Sequence	Seq ID NO:
iKey-64	pBZ38	TTTTTTTTTTCGGAGCTGAGACCGAACGTAGGCTTCGGCACTGTTAGAAGATATCAACAATTCACGTATGC GCGTGGTAACTTGTCTTTTGATTCACTGCCATTCTGCGGAGCTCCCATTCAGATCCACCTGGAGGGGAAAG ATAGTTTATGTCACACAGTACTAACAAAAACCCGGGTTTAGTCTAGGCGGTCCTGCCCCGTTTTTTTTT	1
DNA1	pBZ27	TGGCCACGATCCATGCTAACGTCTCTGCGTAGGGATGAATCCCGTTTTGAACTCGTTCCTACTGAGGGACG AGTCGATAGGTAGCCGAAGTAGTGATACGATCCACACATGCCATCATTGCATACTCGTGCATTCGATGATG CATAGTCACGTAGTCGTATGGTAATGGTGATGTCAAGTCACATCTCAATACTCGTCACTAGAACTGAGCG CGATGACTGGCGAGCTGGTGCGTCCCGAGGCTGGTCGAGCGACTAAGTTGAATCGCGAGACCGATCGAGA CGACTCTAGCGCTGGAATAAATCAGAATAAAGA	2
DNA2	pBZ28	CCCACCAATACTGCCAATAGACGGTACTGTACACCCTGTTTTACAGCAACGGGAAAGGAGGATCACTTTCT ACAATTGTGTGCTGGACTGACAGTCGCATATCCACACATGCCATCATTGCATACTCGTGCATTCGATGATG CATCTACGTAGTCGTATGGTAATGGTGATGTCACTACACATCTCAATACTCGTCACTAGAACTGAGCG CGATACGACTCCGCCCATAGGGTTCCCCGGCTCGCACTGACTACCTTACGCTCTGACCCACATCCGAGCCGG CCGCATGACCCCTGTGATATAATACCGTTCATC	3
n1	pBZ29	TAATACGACTCACTATAGGGACAGTCTAGTCCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGAT CATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGACGTA GTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTCAGTCATCCACTGTTGCCATATATGCAGACGGCATA GTATGCGTGTATGCGTCGAGAGATCATCCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTA ACGTCTCTTCCACCTTTCCCAAAAAGTAACACCGACTGATCGCCATACGGCAACAGTGACTCTCGACTAC CATAGTAGTGAGATGGTGGATTACGATCGCGTGATCTGAGTATCATTGATCTATAGTGGATTGACTGATGA TCGTACTGTCGTACTGACTCTGACGTCGATCTCAGGTCATATTACTCGACAGTTGCTAAGTCACTCATCGT CATACGATGCCGCTGAGCAATAACTAGC	4
n2	pBZ30	GCTAGTTATTGCTCAGCGGCATCGTATGACGATGACTGACTTAGCAACTGTCGAGTAATATGACCTGAGAG CTACTGATCTGACTAGCTAAGCTTGCATGCACGTCATGATCCACTATAGATCAATGATACTCAGATCACGC GATATCGACGTTGACTAGTCAAGCTAGATCACATATGCTGTATGTGCGTAGTCGATGTCATGACTATCTT TTACAGCAACGGGAAAGGAGGACCGTCTATTGGCAGTATTGGTGGGATCTTGTAACTAACGTCAAGATAGG GATGATCTCTCGACGCATACACGCATTAGATGCCGTCTGCATATATGGCAACAGTGGATACGACTCGATCA TCGAGTTCGCATGCTAGCACTGACTACGTTACGCTCTGATCTCAGACGTATAGTCAGATCGGAGTCAGCTGC ATGACGACAGTGCGATGCTAGCGTTGATCTCATGCATCCTACACTCATGAGACTCGTACTGACTGCTGCAC TAGATCTGTCCCTATAGTGAGTCGTATTA	5
n3	pBZ31	TAATACGACTCACTATAGGGACAGTCTAGTCCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGAT CATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGACGTA GTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTCAGTCATCCACTGTTGCCATATATGCAGACGGCATA GTATGCGTGTATGCGTCGAGAGATCATCCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTA ACGTCTCTTCCACCTTTCCCAAAAAGTAACACCGACTGATCGCCATACGGCAACAGTGACTCTCGACTAC CATAGTAGTGAGATGGTGGATTACGATCGCGTGATCTGAGTATCATTGATCTATAGTGGATTGACTGATGA TGATCACTGACTGACTGAACTACGATCATGGTGTATGTGAGCGTGTATGTGCTCGTCGACTGGAGAAACGGC AACAGTGGATGATTGACGTACGACTGCTAGCTCAGGTCATATTACTCGACAGTTGCTAAGTCACTCATCGT CATACGATGCCGCTGAGCAATAACTAGC	6
n4	pBZ32	GCTAGTTATTGCTCAGCGGCATCGTATGACGATGACTGACTTAGCAACTGTCGAGTAATATGACCTGAGAG TCAGTGCTCATGATGTCAATCCACTGTTGCCGTTTCTCCCTACACGAGCACATACACGCTCACATACACCA TGATGACTAGCATGATCATCCGCCGTGTATCTAGATCACGCCGGCATGATCTGATGACGATCATGACTGTT TTACAGCAACGGCAAAGGAGGACCGTCTATTGGCAGTATTGGTGGGATCTTGTAACTAACGTCAAGATAGG GATGATCTCTCGACGCATACACGCATTAGATGCCGTCTGCATATATGGCAACAGTGGATACGACTCGATCA TCGAGTTCGCATGCTAGCACTGACTACGTTACGCTCTGATCTCGACGATAGTCAGATCGGAGTCAGCTGC ATGACGACAGTGCGATGCTAGCGTTGATCTCATGCATCCTACACTCATGAGACTCGTACTGACTGCTGCAC TAGATCTGTCCCTATAGTGAGTCGTATTA	7
n5	pBZ33	TAATACGACTCACTATAGGGACAGTCTAGTCCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGAT CATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGACGTA GTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTCAGTCATCCACTGTTGCCATATATGCAGACGGCATA GTATGCGTGTATGCGTCGAGAGATCATCCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTA ACGTCTCTTCCACCTTTCCCAAAAAGTAACACCGACTGATCGCCATACGGCAACAGTGACTCTCGACTAC CATAGTAGTGAGATGGTGGATTACGATCGCGTGATCTGAGTATCATTGATCTATAGTGGATTGACTGATGA TCGTACTGTCGTACTGACTCTGACGTCGATCTCAGGTCATATTACTCGACAGTTGCTAAGTCACTCATCGT CATACGATGCCGCTGAGCAATAACTAGC	8
n6	pBZ37	GCTAGTTATTGCTCAGCGGCATCGTATGACGATGACTGACTTAGCAACTGTCGAGTAATATGACCTGAGAG CTACTGATCTGACTAGCTAAGCTTGCATGCACGTCATGATCCACTATAGATCAATGATACTCAGATCACGC GATATCGACGTTGACTAGTCAAGCTAGATCACATATGCTGTATGTGCGTAGTCGATGTCATGACTATCTT TTACAGCAACGGGAAAGGAGGACCGTCTATTGGCAGTATTGGTGGGATCTTGTAACTAACGTCAAGATAGG GATGATCTCTCGACGCATACACGCATTAGATGCCGTCTGCATATATGGCAACAGTGGATACGACTCGATCA TCGAGTTCGCATGCTAGCACTGACTACGTTACGCTCTGATCTCAGACGATAGTCAGATCGGAGTCAGCTGC ATGACGACAGTGCGATGCTAGCGTTGATCTCATGCATCCTACACTCATGAGACTCGTACTGACTGCTGCAC TAGACTGTCCCTATAGTGAGTCGTATTA	9

Example 2: Secure Offline Communication via DNA Linguistics

Introduction

The Internet has revolutionized communication with its great speed and volume but remains vulnerable to security breaches. For certain applications where security supersedes speed, the offline transfer of data remains vital. Moving beyond pen and paper, DNA is increasing being used as a medium for information storage and communication[1-6], and DNA cryptography and steganography have emerged as platforms for securing embedded information against unauthorized individuals[7-10].

Three important points of a communication have been investigated—data encoding, data transfer & data extraction—to develop new innovations specifically for DNA-based communications (Figure 11A). To illustrate, if Alice sends a message (*m*) to Bob, she would first write—encode and synthesize—the information in DNA molecules and send it to Bob who would then read—sequence and decode—the message (*m*). However, during the transfer of *m* between Alice and Bob, Eve could intercept the communication and read *m*. To protect *m*, DNA-specific cryptography and steganography methods may be implemented, however many of these methods are experimentally unproven and do not make accommodations for challenges in DNA synthesis and sequencing, such as minimizing homopolymeric stretches.

Here a new framework for the facile and secure communication of short messages in DNA is presented (Figure 11B). To securely encode data, an encryption key (*k*)—that functions as a one-time pad—and decoys (*d*), where *k* is required to decode the message (*m*) and a combination key is required to discern *m* from *d* was implemented. To securely transfer data, a secret-sharing system was established, where *m* can be dispersed throughout a mixture of different DNA molecules, requiring Eve to physically intercept and interrogate multiple separate data transmission lines to gain access to *m*. To facilitate data extraction, chromatogram patterning, a method that allows the bypassing of sequence alignments and instead permits information to be extracted from multiple DNA molecules in a single sequencing reaction was developed.

Taking inspiration from one-time pads, considered to be an unbreakable form of encryption[11-15], described herein is a rationally designed individualized keyboard (iKey) that is amenable to randomization, serves as a facile platform to transfer plaintext on to DNA, and can achieve chromatogram patterning through co-sequencing of multiple DNA strands. Using an iKey, the secret-sharing Multiplexed Sequence Encryption (MuSE) system was developed for

the secure offline communication of information that is disseminated across multiple DNA strands but can be extracted in one step. By recreating a World War II communication from Bletchley Park, it is demonstrated herein that watermarks, a key, a cipher, and a decoy can be written on DNA and the correct information is revealed only if specific strands are co-sequenced.

Development of iKey and MuSE

Here, the familiarity of text-based communication, the QWERTY keyboard, and the genetic code were combined to develop an iKey that serves as a facile platform for DNA communication.

The natural genetic code employs three-letter DNA words (codons) to represent the 20 common amino acids used to build proteins. The four-letter DNA alphabet of adenine (A), cytosine (C), guanine (G) and thymine (T) thus yields $4^3 = 64$ codons. These 64 codons were mapped onto a modified QWERTY keyboard to produce a personalized platform – iKey-64 – for translating text on to DNA (Figure 1A). The codons in iKey-64 can be randomized to produce a unique iKey for every message to provide additional security for communications, akin to a one-time pad[11]. Any specific version of iKey-64 can itself be encoded in DNA and provided as an additional component of a communication, where it can serve as a unique dictionary for each message (Figures 1B-1C).

To increase the security of encoded messages in addition to the substitution cipher of iKey-64, texts were disseminated between multiple DNA strands so that the desired message would be revealed only if the correct strand combinations were analyzed. This multiplexing is at the heart of the MuSE strategy, which is a secret-sharing system where a message can be stored securely by being fragmented and distributed between multiple parties[16]. Analyzing only a single strand would yield either nonsense or incorrect messages designed to mislead unauthorized individuals.

Conventionally, to extract information embedded on multiple DNA strands, one would first have to sequence each strand separately and then perform sequence alignments. In designing MuSE, it was expected that when multiple DNA strands are analyzed together by Sanger sequencing using a common primer, at chromatogram positions where two bases are identical a large peak would be observed and where two bases differ a small peak would be observed, thereby producing a pattern (Figure 2A). However, the simultaneous sequencing of

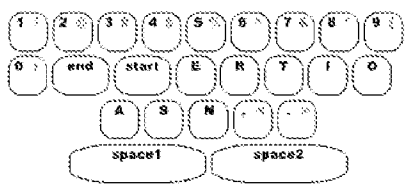
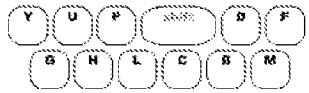
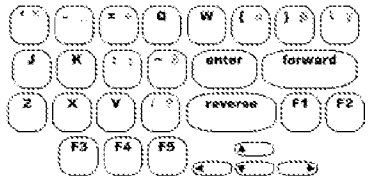
multiple DNA strands with a common primer cannot be used, as it leads to poor chromatograms and non-specific reads (Figures 12A-12C). Chromatogram patterning is based on the rational design of iKey-64 (Tables 2-3), where the aim was to reduce the incidence of homopolymers in DNA messages as long stretches of homopolymers lead to sequencing inaccuracies[17]. The homopolymer codons AAA, CCC, GGG, and TTT are assigned to four function keys, ensuring that in normal text no homopolymer longer than four bases is possible. Even letter combinations yielding four identical bases (such as GTT-TTC representing V-K on the keyboard) are kept quite rare. Therefore, the codon assignment of iKey-64 was based on the frequency of use of letters in the English language[18] to minimize the occurrence of homopolymers and achieve chromatogram patterning.

As shown in Table 3, the buttons of this embodiment of the iKey-64 were separated in to 3 categories based on the frequency of use as judged by qualitative measures. Category 1 is for the most frequently used buttons and is encoded by codons that contain three different nucleotides. Category 2 is for less frequently used buttons and is encoded by codons that contain the same nucleotide in the first and third position. Category 3 is for the least frequently used buttons and is encoded by codons that contain two or more homopolymers. Since iKey-64 is similar in design to a one-time pad, many possible versions exist and the last column provides the number of potential permutations that exist for randomly shuffling the codons between the buttons. The frequency of letters in the English alphabet were based on Table 2. If chromatogram patterning is not desired, then all 64 buttons in iKey-64 can be randomly shuffled for transcription of plaintext on to DNA.

Table 2: Rational Design of iKey-64: Letter Frequency

Letter	Frequency	Letter	Frequency
E	11.1607%	M	3.0129%
A	8.4966%	H	3.0034%
R	7.5809%	G	2.4705%
I	7.5448%	B	2.0720%
O	7.1635%	F	1.8121%
T	6.9509%	Y	1.7779%
N	6.6544%	W	1.2899%
S	5.7351%	K	1.1016%
L	5.4893%	V	1.0074%
C	4.5388%	X	0.2902%
U	3.6308%	Z	0.2722%
D	3.3844%	J	0.1965%
P	3.1671%	Q	0.1962%

Table 3: Rational Design of iKey64: Examples of iKey Permutations

Category	Codons	Buttons	Permutations
1	TCA TCG TAC TAG TGC TGA CTA CTG CAT CAG CGT CGA ATC ATG ACT ACG AGT AGC GTC GTA GCT GCA GAT GAC		6.2×10^{23}
2	TCT TAT TGT CTC CAC CGC ATA ACA AGA GTG GCG GAG		4.8×10^{8}
3	TTT TTC TTA TTG TCC TAA TGC CTT CCT CCC CCA CCG CAA CGG ATT ACC AAT AAC AAA AAG AGG GTT GCC GAA GGT GGC GGA GGG		3.0×10^{29}
iKey-64 Total Permutations with Chromatogram Patterning			9.1×10^{61}
iKey-64 Total Permutations without Chromatogram Patterning			1.3×10^{89}

iKey-64 was tested for MuSE by writing the cipher 'Massachusetts Institute Technology' on two DNA strands, where "space1" (AGT) was used in DNA-1 and "space2" (CTA) with DNA-2 to demarcate individual words in the sequences (Figures 2B-2C). Co-sequencing both DNA samples together would introduce troughs around words in the chromatogram. Individual sequencing of DNA-1 and DNA-2 produced high quality reads, however in a DNA-1+2 mixture forward sequencing with a common primer did not produce chromatogram patterning, but rather camouflaged the cipher (Figure 2D). This was due to the variable DNA sequences placed upstream of the ciphers, where stretches of C and A homopolymers at the 5' ends interfered with base determination during Sanger sequencing causing intentional misalignment of the recognized bases in the chromatogram (Figures 3A-3C). On the other hand, reverse sequencing of DNA-1+2 with a common primer produced a distinct pattern on the chromatogram. Since there were no interfering stretches of homopolymers in the variable DNA regions, there were no

shifts in the base identities during sequencing leading to predictable chromatogram patterning and a single-step extraction of information from the two strands (Figures 3B-3C).

MuSE can be tuned to embed data in chromatograms discreetly so that sequence alignments derived from chromatograms cannot be used to identify embedded information. Adjusting the ratio of DNA-1/DNA-2 allows the degree of contrast achieved in the chromatogram patterns to be varied (Figure 2E). When DNA-1 or DNA-2 is present at 10-30%, chromatogram patterning is still achieved upon close examination of individual peaks, but the resulting sequence produced is only that of the more concentrated partner (Figures 4-5). Therefore, an unauthorized user would be unable to see embedded messages directly in the sequence output or in alignments.

Multiplexed Sequencing of Strand Combinations

For additional security, MuSE can be used to disseminate information across many DNA strands, where multiplexed sequencing of different strand combinations will provide different readouts (Figure 13). To demonstrate this, watermarks, a key, a cipher, and a decoy message were encoded across six strands in a 525 bp region of DNA to recreate a World War II communication made during the establishment of Bletchley Park (Figure 6A and Figure 14)[19]. The functions of the elements are: (i) watermarks – an identification tag for each strand, (ii) key – a riddle whose solution would provide the correct strand combinations required for co-sequencing to reveal the cipher in the secret-sharing system, (iii) cipher – the desired message to be communicated, and (iv) decoy – a false message to be revealed if improper strand combinations were used for co-sequencing.

To extract the information via co-sequencing, two different primers – $Primer_{Key}$ and $Primer_{Cipher}$ – that are common to all six strands are required. As a demonstration for this exercise a simple key was chosen, where co-sequencing of all of the strands with $Primer_{Key}$ revealed the message: Pascal's triangle: d2r6-reverse (Figure 6A). This serves as a combination key and means the cipher is revealed from pairs as ordered is Pascal's triangle diagonal 2 down until row 6 on the reverse strand. If strand pairs n1+2, n3+4, and n5+6 were to be co-sequenced using $Primer_{Cipher}$, then the embedded message 'Bletchley Park: GC&CS Codebreakers' would be revealed. However, if one were to for example misinterpret the key, then a decoy message could be revealed. Here, one decoy message was embedded – 'Captain Ridley's Shooting Party' – that would be revealed if one were to co-sequence pairs n2+3, n4+5, and n6+1, a circular

permutation of the key. Of course, more than one decoy message could be embedded to further introduce complexity in communications. Alternatively, an unauthorized user may use random primers – $Primer_{ExternalFw/Rv}$ – instead of $Primer_{Key}$ and $Primer_{Cipher}$ to extract messages if they were embedded in large DNA regions. To obfuscate this approach, the embedded information was alternated between the forward and reverse strands to provide a camouflage effect (Figure 15. Since any secure communication would have a limited quantity of DNA (enough to extract the desired message once), an unauthorized user would be unable to exhaustively explore primer sequences to extract information without advanced scientific protocols.

As expected, co-sequencing with $Primer_{ExternalFw/Rv}$ did not produce chromatogram patterning, whether cipher/decoy pairs or all six strands were co-sequenced (Figures 7-8). However, co-sequencing of all six strands with $Primer_{Key}$ produced the readout 'Pascal's triangle: d2r6-reverse', while the cipher/decoy containing regions did not produce chromatogram patterning. Similarly, chromatogram patterning was not observed in the cipher/decoy containing regions when $Primer_{Cipher}$ was used for co-sequencing all six strands. On the other hand, sequencing of pairs with $Primer_{Cipher}$ as per the order in Pascal's triangle – n1+2, n3+4, and n5+6 – revealed the cipher via chromatogram patterning (Figures 9A-9G). Similarly, co-sequencing of the incorrect pairs – n2+3, n4+5, and n6+1 – led to a decoy message to be revealed. Expectedly, co-sequencing of other pair combinations did not lead to any patterning (Figure 6B). This demonstrated that in addition to the security afforded by iKey-64 and MuSE, one must also decipher the key accurately to unlock embedded messages.

If unauthorized individuals were to gain access to a DNA communication, next-generation sequencing (NGS) might also be attempted for extracting messages. To recreate such a scenario, the difficulty associated with NGS analysis of unknown DNA samples was tested. A purified mixture of DNA samples n1+n2+n3+n4+n5+n6 was prepared and submitted for NGS analysis to an outside party under blind experimental conditions, with a request to provide the assembled contents of the sample (Figure 16A-16B). While sequencing of the mixture produced ~2 million reads, the blind assembly of the reads to reconstruct the contents proved difficult and inconclusive (Table 4). However, after the initial analysis the outside party was informed that there were 6 plasmids in the sample, each containing 525 bp messages as inserts. The vector sequence was then provided and the outside party asked for the exact sequences of the messages in the sample. A second round of analysis identified 6 assembled sequences that represented our messages (Table 5). Alignment of the 6 identified sequences with n1, n2, n3, n4, n5, and n6

templates provided most of the information in the six messages, with n1, n2, n3, and n5 providing almost perfect alignments (Figure 16C). This demonstrated the difficulty associated with blind sequencing of a MuSE communication without any prior knowledge of DNA contents. Even if the sequences of a DNA communication were identified after considerable time and expense, the contents of a communication would still likely be protected by the iKey, combination key, and decoy/non-coding sequences.

Table 4: Next-generation sequencing statistics of assembled reads under blind experimental conditions.

	n1+n2+n3+n4+n5+n6
Sequence size	1,407,947
Number of scaffolds	2,851
% GC	51.1
Shortest contig size	300
Median sequence size	423
Mean sequence size	493.8
Longest contig size	4,625
Number of subsystems	22
Number of coding sequences	984
Number of RNAs	0

*NGS sequencing of a mixture of samples n1+n2+n3+n4+n5+n6 (Fig S10) produced 1,997,179 reads at 300 bp with 47% GC content. Shown are the statistics of the assembled scaffolds by the MIT BioMicro Center under blind experimental conditions. While the DNA samples produced high quality reads, under blind experimental conditions assembly of the reads in to the original constructs proved challenging and the results were inconclusive. n1 = 2,346 bp/47.4% GC, n2 = 2,346 bp/47.3% GC, n3 = 2,346 bp/47.5% GC, n4 = 2,346 bp/47.6% GC, n5 = 2,346 bp/47.4% GC, n6 = 2,346 bp/47.3% GC.

Table 5: Identified sequences from NGS analysis.

Assembled Sequence	Sequence	SEQ ID NO:
1	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGA GTCTCATGAGTGTAGGATGCATGAGATCAACGCTAGCATCGCACTG TCGTCATGCAGCTGACTCCGATCTGACTATCGTCTGAGATCAGAGC GTAACGTAGTCAGTGCTAGCATGCGAACTCGATGATCGAGTCGTAT CCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTC GAGAGATCATCCCTATCTTGACGTTAGTTACAAGATCCCACCAATA CTGCCAATAGACGGTCCTCCTTTCCCGTTGCTGTAAAACAGTCATGA TCGTCATCAGATCATGCCGGCGTGATCTAGATACACGGTGGATTCA GCTACTAGTCGAATCATGACGTGAGAAGCATGAACGATATGAAGAA GTTATGTGGATAGCTGTCGACGTGATCGTATCGATGCAGTCCTCAG GTCATATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCC GCTGAGCAATAACTAGC	18
2	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGA GTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGC AGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGAC GTAGTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTCAGTCATCC ACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGA GAGATCATCCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCC ATGCTAACGTCTCTTCCACCTTTCCCAAAAAGTAACACCACCATGACG TATCGACTACGCACATACAGCATATGTGGATGATCACTGACTGACT GAACTACGATCATGGTGTATGTGAGCGTGTATGTGCTCGTGACTGG AGAAACGGCAACAGTGGATGATTGACGTACGACTGCTAGCTCAGGT CATATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCCGC TGAGCAATAACTAGC	19
3	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGA GTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGC AGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGAC GTAGTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTCAGTCATCC ACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGA GAGATCATCCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCC ATGCTAACGTCTCTTCCACCTTTCCCAAAAAGTAACACCGACTGATC GCGCATACGGCAACAGTGACTCTCGACTACCATAGTAGTGAGATGG TGGATTACGATCGCGTGATCTGAGTATCATTGATCTATAGTGGATTG ACTGATGATCGTACTGTCGTACTGACTCTGACGTCGATCTCAGGTCA TATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCCGCTG AGCAATAACTAGC	20
4	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGA GTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGC AGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGAC GTAGTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTCAGTCATCC ACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGA GAGATCATCCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCC ATGCTAACGTCTCTTCCACCTTTCCCAAAAAGTAACACTGACTGCAT TCGTGATCATCATGCCGGCGTGATCTAGATACACGGTGGATTCAGC TACTAGTCGAATCATGACGTGAGAAGCATGAACGATATGAAGAAGT TATGTGGATAGCTGTCGACGTGATCGTATCGATGCAGTCCTCAGGTC ATATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCCGCT GAGCAATAACTAGC	21
5	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGA GTCTCATGAGTGTAGGATGCATGAGATCAACGCTAGCATCGCACTG TCGTCATGCAGCTGACTCCGATCTGACTATCGTCTGAGATCAGAGC	22

	GTAACGTAGTCAGTGCTAGCATGCGAACTCGATGATCGAGTCGTAT CCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTC GAGAGATCATCCCTATCTTGACGTTAGTTACAAGATCCCACCAATA CTGCCAATAGACGGTCCTCCTTTCCCGTTGCTGTAAAACATAGTCAT GACATCGACTACGCACATACAGCATATGTGGATCTAGCTTGACTAG TCAACGTCGATATCGCGTGATCTGAGTATCATTGATCTATAGTGGAT TGACTGATGATCGTACTGTCGTACTGACTCTGACGTCGATCTCAGGT CATATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCCGC TGAGCAATAACTAGC	
6	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGA GTCTCATGAGTGTAGGATGCATGAGATCAACGCTAGCATCGCACTG TCGTCATGCAGCTGACTCCGATCTGACTATCGTCTGAGATCAGAGC GTAACGTAGTCAGTGCTAGCATGCGAACTCGATGATCGAGTCGTAT CCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTC GAGAGATCATCCCTATCTTGACGTTAGTTACAAGATCCCACCAATA CTGCCAATAGACGGTCCTCCTTTCCCGTTGCTGTAAAACATAGTCAT GACATCGACTACGCACATACAGCATATGTGGATCTAGCTTGACTAG TCAACGTCGATATCGCGTGATCTGAGTATCATTGATCTATAGTGGAT CATGACGTGCATGCAAGCTTAGCTAGTCAGATCAGTAGCTCTCAGG TCATATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCCG CTGAGCAATAACTAGC	23

*After blind analysis by the MIT BioMicro Center did not provide the contents of the unknown sample submitted for analysis, further information about the plasmids and vector sequences was provided. Shown here are the 6 assembled and identified sequences each 525 bp, representing the messages encoded in n1, n2, n3, n4, n5, and n6 generated by the MIT BioMicro Center after a second round of analysis. Alignments to n1, n2, n3, n4, n5, and n6 are in Figure 16C.

iKey-64 is designed to convert plaintext in to a DNA encodable language. If chromatogram patterning is desired, the codons may potentially be shuffled to enable 9.1×10^{61} variants (Table 3). However, if chromatogram patterning is not desired then a maximum of 1.3×10^{89} variants exist, significantly increasing the security of encoded information. As a communication medium, knowledge of the appropriate primers, combination key, and incorporation of decoy messages would also provide additional data security. Nevertheless, data encoded using iKey-64 would still not be truly random due to the frequency of use for each button, but additional measures may be implemented to increase security: (i) Cryptography – plaintext information may first be subject to advanced cryptographic algorithms, (ii) Linguistics – principles of linguistics may be applied to the layout of iKeys to modify alphabets for DNA communication, introduce new grammar rules or create iKeys in different languages, and (iii) Codons – increasing the number of nucleotides per codon can introduce redundancies in the buttons to adjust for character usage frequency. To illustrate, four nucleotides codons can be

used to create a 256 button keyboards such as iKey-256 (Figure 10). When the number of buttons for each letter is adjusted to reflect its frequency in English text, then the probability of using a button for E would equal Q. Similar redundancies may also be introduced for buttons representing numerals, grammar, and other user-defined functions. For instance, the frequency of numerals may be adjusted according to Benford's Law[20].

To further extend the iKey system, codons can be used to represent words or phrases in addition to characters. It is estimated that the vocabulary of an educated native English speaking adult consists of ~17,000 lemmas, while only 10 lemmas constitute 25% of the words used in English[21, 22]. Using 8-nucleotide codons could generate iKeys with 65,536 buttons, sufficient to include all of the commonly used words in English as well as accommodate individual letters, numerals, grammatical characters, functional characters, and high frequency words. Theoretically, the iKey platform may be designed to incorporate the entire English language. The Oxford English Dictionary (OED), the most comprehensive record of the English language, contains 291,500 entries and a total of 615,100 word forms[23]. To encode all of the entries of the OED on an iKey would require 10-nucleotide codons to generate a 1,048,576 button keyboard. Additionally, the dictionary is composed of 59 million words containing 350 million characters resulting in 5.9 characters/word. This would require 18 nucleotides to encode with an iKey-64 but only 10 nucleotides for an iKey-1,048,576, representing a 44% reduction in DNA requirements.

References

1. Bancroft, C., Bowler, T., Bloom, B. & Clelland, C. T. Long-term storage of information in DNA. Science 293, 1763-1765 (2001).

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

3. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. Science 337, 1628 (2012).

4. Liss, M. et al. Embedding permanent watermarks in synthetic genes. PLoS One 7, e42465 (2012).

5. Cox, J. P. Long-term data storage in DNA. Trends Biotechnol. 19, 247-250 (2001).

6. Sennels, L. & Bentin, T. To DNA, all information is equal. Artif. DNA PNA XNA 3, 109-111 (2012).

7. Haughton, D. & Balado, F. BioCode: two biologically compatible Algorithms for embedding data in non-coding and coding regions of DNA. BMC Bioinformatics 14, 121-2105-14-121 (2013).

8. Heider, D. & Barnekow, A. DNA-based watermarks using the DNA-Crypt algorithm. BMC Bioinformatics 8, 176 (2007).

9. Tulpan, D., Regoui, C., Durand, G., Belliveau, L. & Leger, S. HyDEn: a hybrid steganocryptographic approach for data encryption using randomized error-correcting DNA codes. Biomed. Res. Int. 2013, 634832 (2013).

10. Kawano, T. Run-length encoding graphic rules, biochemically editable designs and steganographical numeric data embedment for DNA-based cryptographical coding system. Commun. Integr. Biol. 6, e23478 (2013).

11. Ekert, A. & Renner, R. The ultimate physical limits of privacy. Nature 507, 443-447 (2014).

12. Gehani, A., LaBean, T. & Reif, J. DNA-based Cryptography. DNA Based Computers V: Dimacs Workshop DNA Based Computers V June 14-15, 1999 Massachusetts Institute of Technology 54, 233 (2000).

13. Mao, C., LaBean, T. H., Relf, J. H. & Seeman, N. C. Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. Nature 407, 493-496 (2000).

14. Hirabayashi, M., Kojima, H. & Oiwa, K. in (eds Peper, F., Umeo, H., Matsui, N. & Isokawa, T.) 174-183 (Springer Japan, 2010).

15. Hirabayashi, M., Kojima, H. & Oiwa, K. Effective algorithm to encrypt information based on self-assembly of DNA tiles. Nucleic Acids Symp. Ser. (Oxf) (53):79-80. doi, 79-80 (2009).

16. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. Clin. Chem. 55, 641-658 (2009).

17. http://www.oxforddictionaries.com/us/words/what-is-the-frequency-of-the-letters-of-the-alphabet-in-english.

18. Ferguson, N., Schneier, B. & Kohno, T. in Cryptography engineering: design principles and practical applications (Wiley Publishing, Inc., Indianapolis, 2010).

19. http://www.bletchleypark.org.uk/.

20. Alves, A. D., Yanasse, H. H. & Soma, N. Y. Benford's Law and articles of scientific journals: comparison of JCR and Scopus data. Scientometrics 98, 173-184 (2014).

21. http://www.oxforddictionaries.com/us/words/the-oec-facts-about-the-language.

22. Goulden, R., Nation, I. S. P. & Read, J. How large can a receptive vocabulary be? Applied Linguistics 11, 341-363 (1990).

23. http://public.oed.com/history-of-the-oed/dictionary-facts/.

24. Gibson, D. G. Enzymatic assembly of overlapping DNA fragments. Methods Enzymol. 498, 349-361 (2011).

What is claimed is:

CLAIMS

1. A method of secure communication of information contained on a single nucleic acid molecule, the method comprising:

(a) obtaining a nucleic acid molecule of known sequence;

(b) obtaining a modified keyboard comprising a personalized platform for translating nucleic acid sequence into text; and,

(b) generating a quantum of information translated from the nucleic acid sequence using the modified keyboard of (a).

2. A method of secure communication of information disseminated across at least one nucleic acid molecule, the method comprising:

(a) obtaining a modified keyboard comprising a personalized platform for translating text into a nucleic acid sequence;

(b) translating a quantum of information into a nucleic acid message sequence using the modified keyboard of (a); and,

(c) obtaining a at least one nucleic acid molecules, each molecule comprising (i) the complete or a portion of the nucleic acid message sequence and (ii) at least one contiguous stretch of randomized variable nucleic acid sequence flanking and/or inserted into the message sequence, thereby producing a nucleic acid molecule or a set of nucleic acid molecules containing the entire quantum of information.

3. The method of claim 1 or claim 2, wherein the modified keyboard comprises codons.

4. The method of claim 3, wherein the codons are designed to normalize frequency of character usage.

5. The method of any one of claims 1 to 4, further comprising sequencing the nucleic acid molecule or set of nucleic acid molecules using one or more common primers.

6. The method of claim 5, wherein the sequencing produces a chromatogram.

7. The method of claim 5, wherein the sequencing produces data that is analyzed by sequence alignment or bioinformatics methods.

8. The method of claim 6, further comprising identifying nucleic acid sequence corresponding to areas of high intensity peaks on the chromatogram.

9. The method of claim 6, further comprising identifying nucleic acid sequence corresponding to areas of low intensity peaks on the chromatogram.

10. The method of any one of claims 6-9, further comprising extracting the quantum of information contained within the set of nucleic acid molecules by using the modified keyboard to translate the nucleic acid sequence identified in any one of claims 6-9.

11. The method of any one of claims 1-10, wherein the modified keyboard comprises homopolymer codons located on functional keys.

12. The method of any one of claims 1-11, wherein the codons are greater than 3 nucleotides in length.

13. The method of claim 12, wherein the codons are 4, or 5, or 6, or 7, or 8, or 9, or 10, or 11, or 12, or 13, or 14, or 15, or 16, or 17, or 18 nucleotide bases in length.

14. The method of any one of claims 1-13, wherein the codons are of mixed lengths.

15. The method of any one of claims 1-14, wherein the variable nucleic acid sequence comprises contiguous homopolymer codons.

16. The method of any one of claims 6-15, wherein the sequencing is performed by Sanger sequencing, bridge PCR, nanopore sequencing, or Next Generation Sequencing.

17. The method of any one of claims 1-16, wherein the at least one nucleic acid molecule is sequenced with at least one common primer.

18. The method of any one of claims 1-17, wherein the nucleic acid molecule(s) are in silico.

19. A method of producing an individualized keyboard for the conversion of plaintext into nucleic acid encodable language, the method comprising:

(a) producing a library of codons;

(b) assigning each member of the library to a different symbol; and

(c) arranging the symbols into an array, thereby producing an individualized keyboard.

20. The method of claim 19, wherein the codons are greater than three nucleotide bases in length.

21. The method of claim 19 or claim 20, wherein the codons are 4, or 5, or 6, or 7, or 8, or 9, or 10, or 11, or 12, or 13, or 14, or 15, or 16, or 17, or 18 nucleotide bases in length.

22. The method of any one of claims 19-21, wherein the codons are of mixed lengths.

23. The method of any one of claims 19-22, wherein the symbol is selected from the group consisting of letter, number, word, punctuation mark, pictogram or logogram.

24. The method of any one of claims 2-18, wherein the variable sequence comprises at least one contiguous stretch of homopolymer codons.

25. The method of any one of claims 19-23, wherein the individualized keyboard comprises homopolymer codons associated only with functional keys.

26. The method of any one of claims 19-23, wherein the codons are designed to normalize frequency of character usage.

FIG. 1A

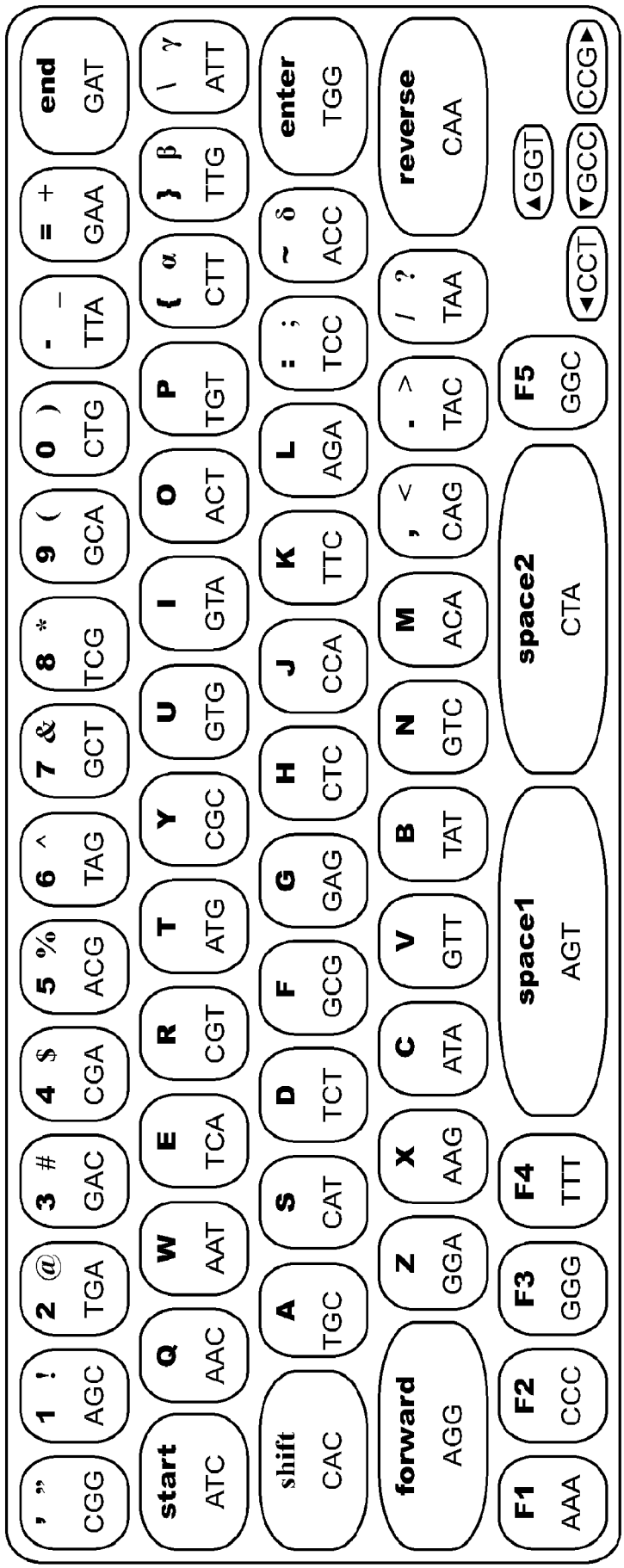
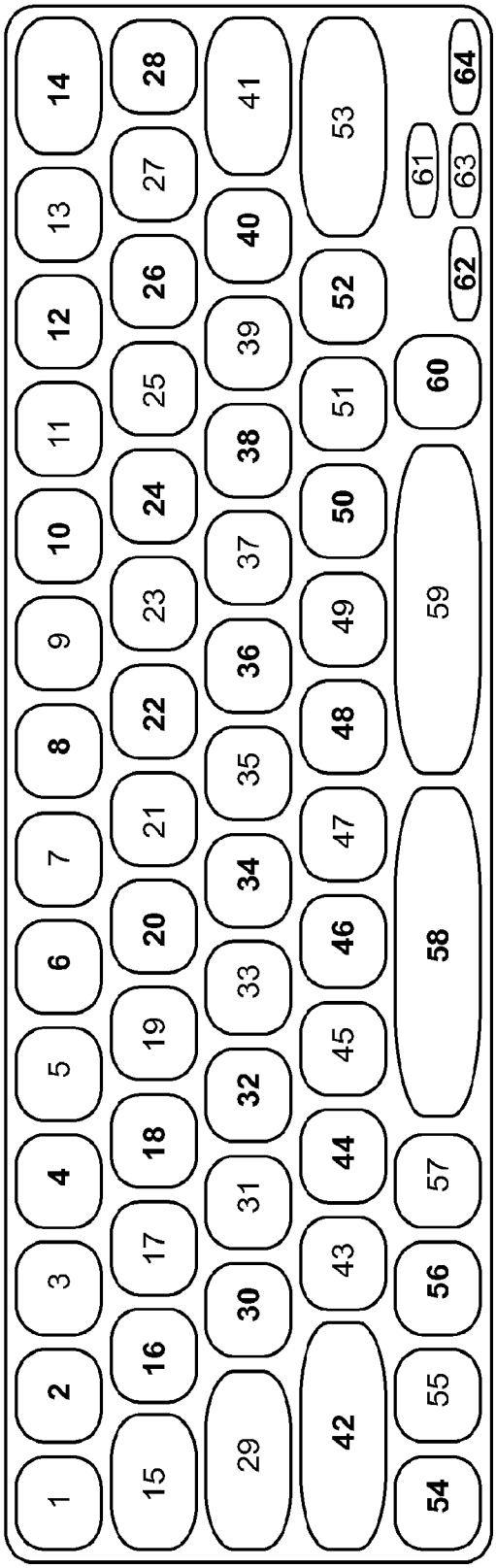

' "
CGG
1 !
AGC
2 @
TGA
3 #
GAC
4 $
CGA
5 %
ACG
6 ^
TAG
7 &
GCT
8 *
TCG
9 (
GCA
0)
CTG
- _
TTA
= +
GAA
end
GAT
start
ATC
Q
AAC
W
AAT
E
TCA
R
CGT
T
ATG
Y
CGC
U
GTG
I
GTA
O
ACT
P
TGT
{ α
CTT
} β
TTG
\ γ
ATT
shift
CAC
A
TGC
S
CAT
D
TCT
F
GCG
G
GAG
H
CTC
J
CCA
K
TTC
L
AGA
: ;
TCC
~ δ
ACC
enter
TGG
forward
AGG
Z
GGA
X
AAG
C
ATA
V
GTT
B
TAT
N
GTC
M
ACA
, <
CAG
. >
TAC
/ ?
TAA
reverse
CAA
F1
AAA
F2
CCC
F3
GGG
F4
TTT
space1
AGT
space2
CTA
F5
GGC
▲GGT
◀CCT
▼GCC
CCG▶

FIG. 1B



CGG**AGC**TGA**GAC**CGA**ACG**TAG**GCT**TCG**GCA**CTG**TTA**GAA**GAT**ATC**AAC**
AAT**TCA**CGT**ATG**CGC**GTG**GTA**ACT**TGT**CTT**TTG**ATT**CAC**TGC**CAT**TCT**
GCG**GAG**CTC**CCA**TTC**AGA**TCC**ACC**TGG**AGG**GGA**AAG**ATA**GTT**TAT**GTC**
ACA**CAG**TAC**TAA**CAA**AAA**CCC**GGG**TTT**AGT**CTA**GGC**GGT**CCT**GCC**CCG**

FIG. 1C

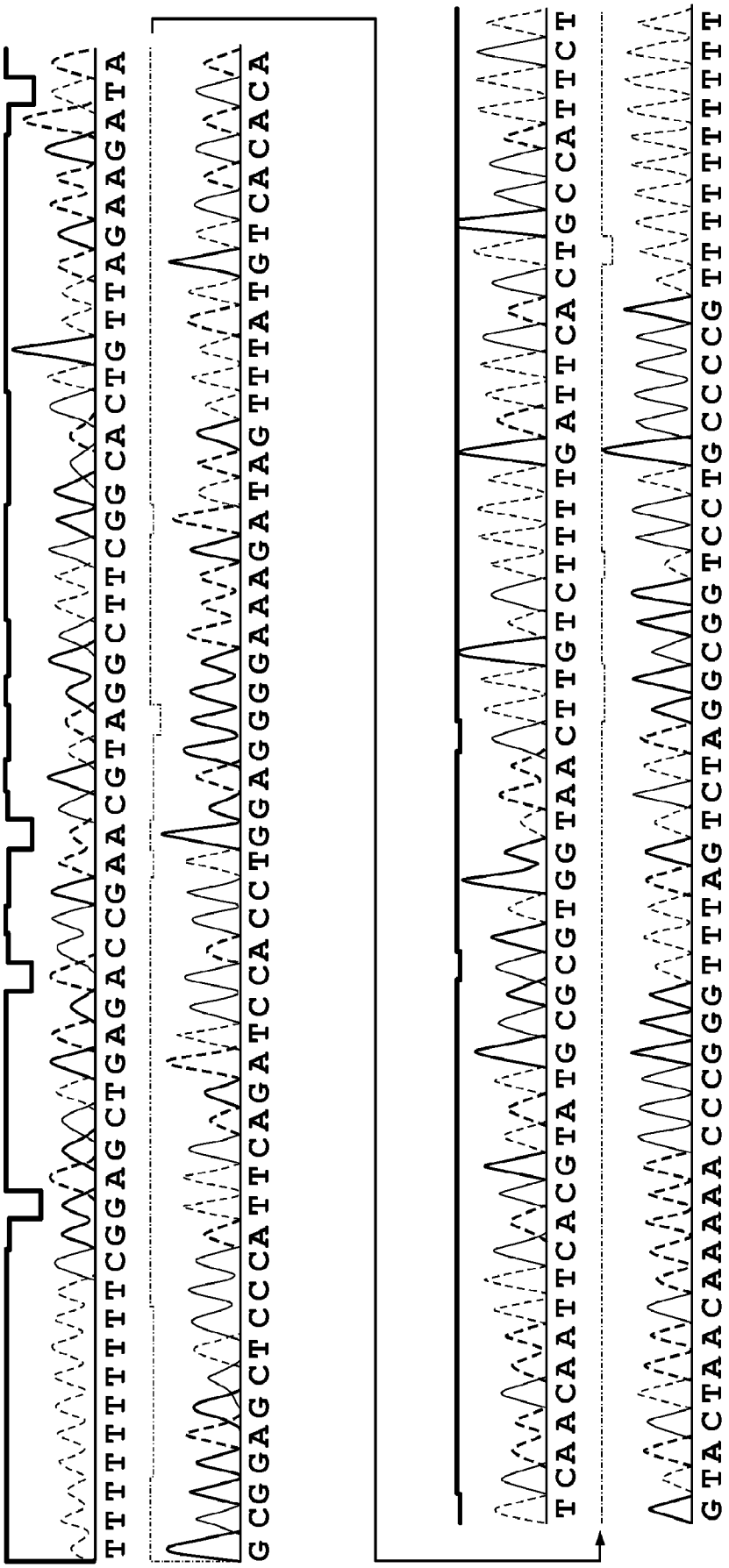

TTTTTTTTTCGGAGCTGAGACCGAACGTAGGCTTCGGCACTGTTAGAAGATA
GCGGAGCTCCCATTCAGATCCACCTGGAGGGGAAAGATAGTTTATGTTCACACA
TCAACAATTCACGTATGCGCGTGGTAACTTGTCTTTTGATTCACTTGCCATTCCT
GTACTAACAAAAACCCCGGGTTTAGTCTAGGCGGTCCCCCGTTTTTTTTTT

FIG. 2A

FIG. 2B

FIG. 2C

MIT Cipher:

DNA 1 ATCCACACATGCCATCATTGCATACTCGTGCATTCAATGATGCAT**AGT**CACGTAGTCCA

DNA 2 ATCCACACATGCCATCATTGCATACTCGTGCATTCAATGATGCAT**CTA**CACGTAGTCCA

Massachusetts Institute Technology

TATGGTAATGGTGATGTCA**AGT**CACATGTCAATACTCGTCACTAGAACTGAGCGCGAT

TATGGTAATGGTGATGTCA**CTA**CACATGTCAATACTCGTCACTAGAACTGAGCGCGAT

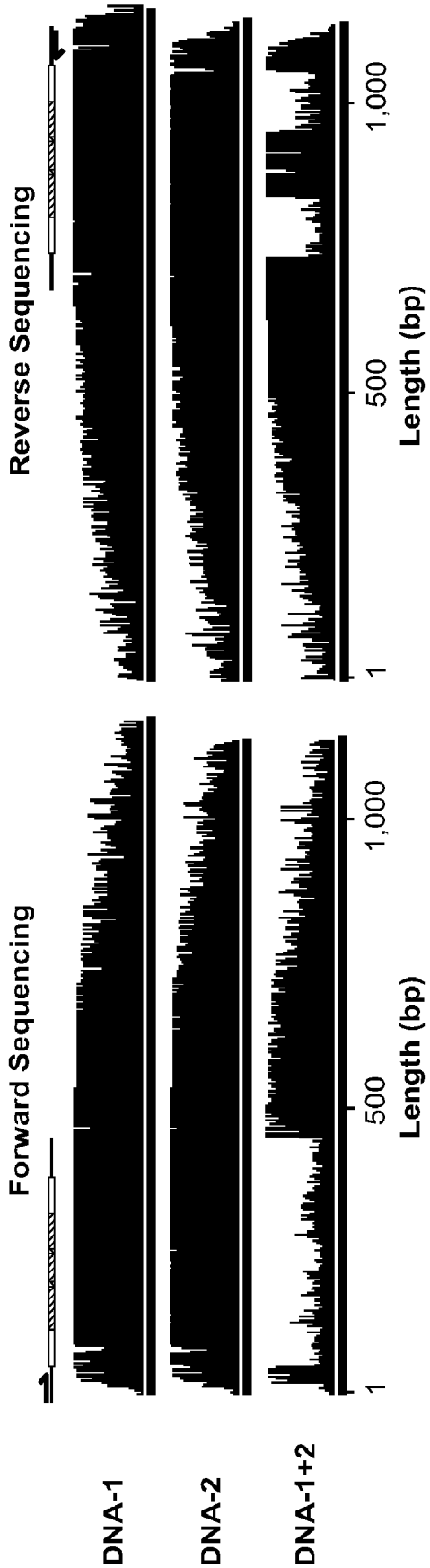

Forward Sequencing
Reverse Sequencing
DNA-1
DNA-2
DNA-1+2
1
500
1,000
Length (bp)

FIG. 2D

FIG. 2E

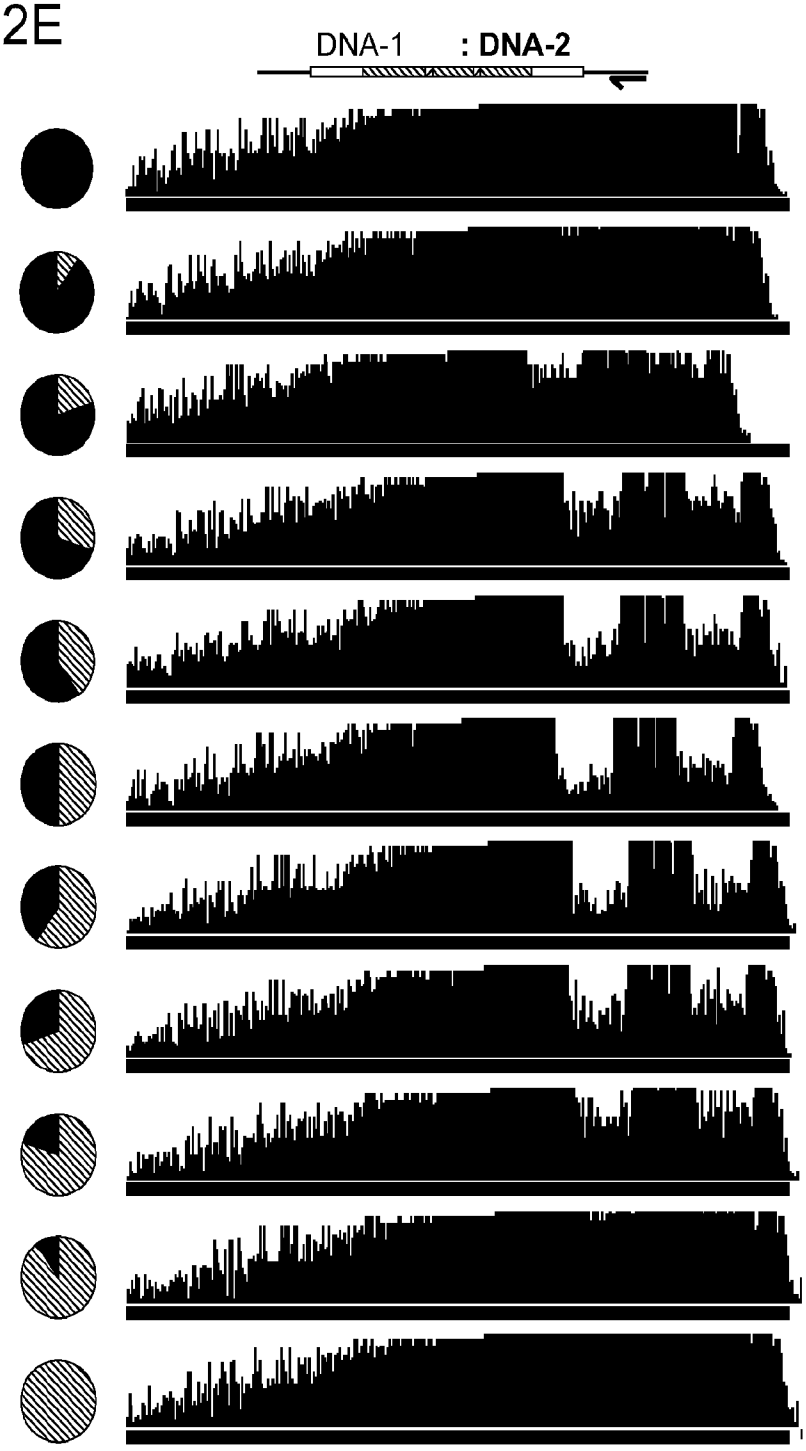


FIG. 3A

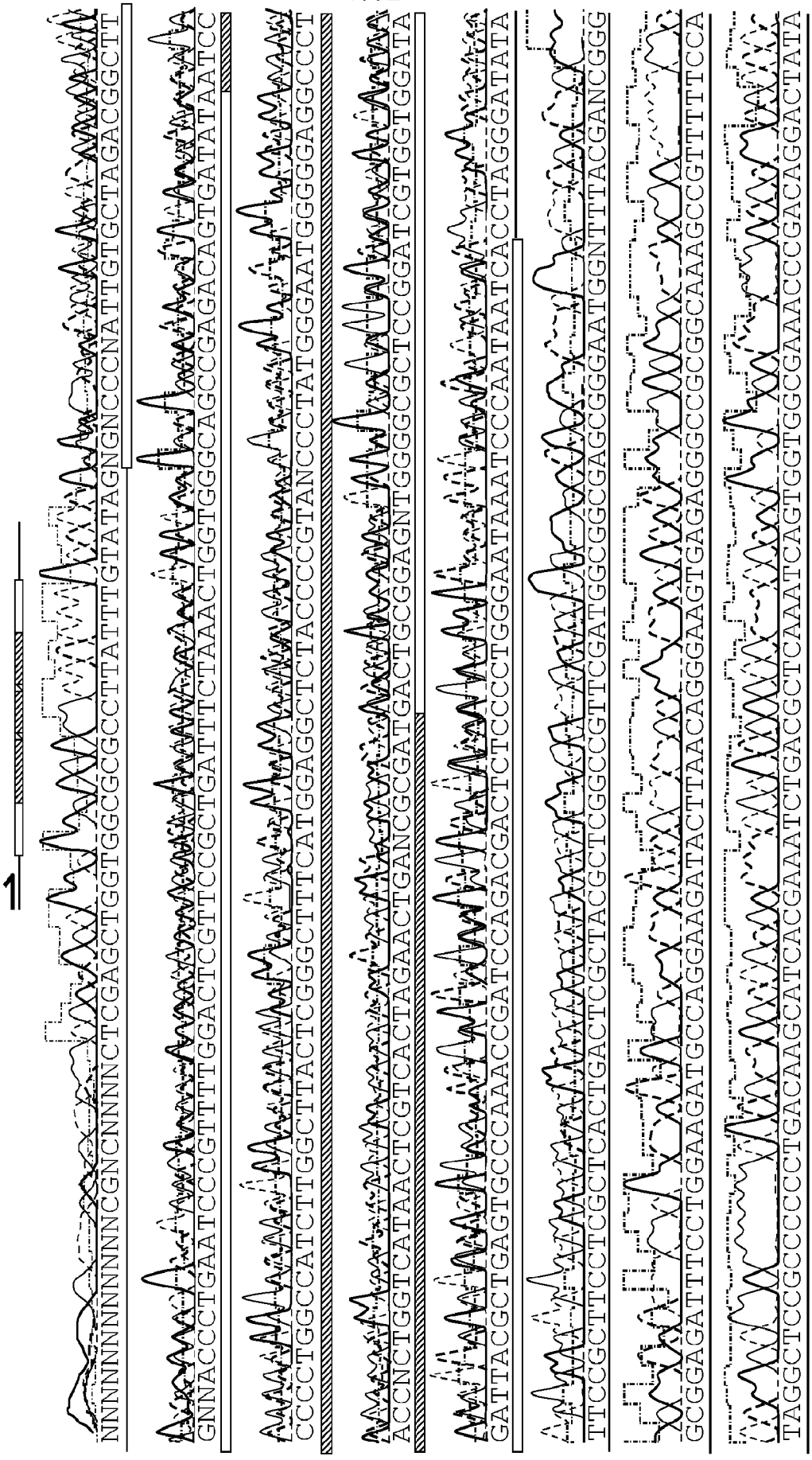


FIG. 3B

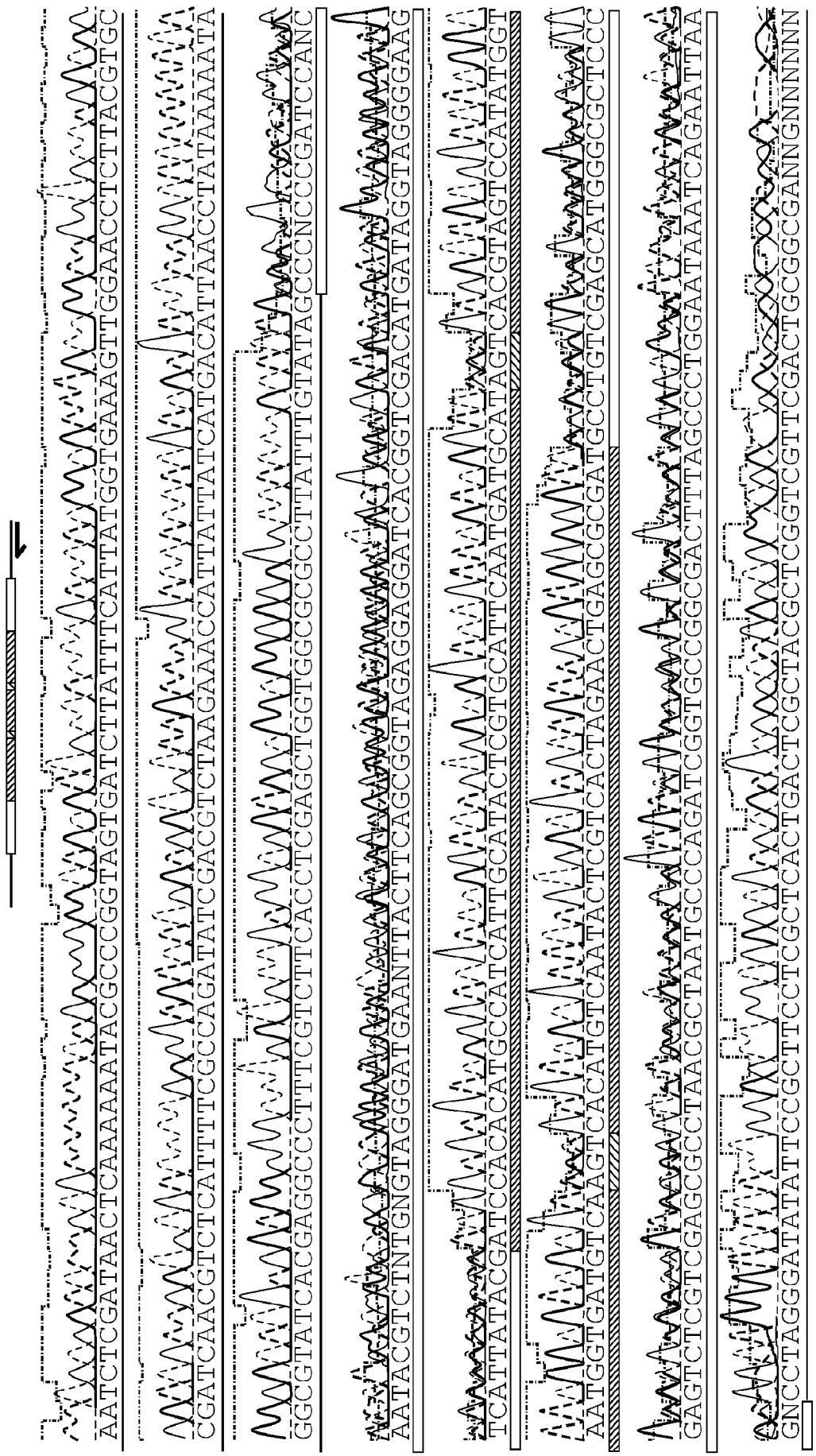


FIG. 3C

Upstream variable DNA: ! SEQ ID NO: 14-15

TGGCCACGATCCATGCTAACGTCTCTGCGTAGGGATGAATCCCGTTTTGAACTCGTTCCTACTGACGGA

CCCACCAATACTGCCAATAGACGGTACTGTACACCCTGTTTTACAGCAACGGGAAAGGAGGATCACTTT

CGAGCTGATAGGTAGCCGAAGTAGTGATACG!

CTACAATTGTGTGCTGGACTGACAGTCGCAT!

Downstream variable DNA: ! SEQ ID NO: 16-17

GACTGGCGAGCTGGTGGCGCTCCCGAGGCTGGTCGAGCGACTAAGTTGAATGCGCAGACCGATCGAGACG

ACGACTCGCCCATAGGGTTCGCCGGCTCGCACTGACCTACCTTACGCTCTGACCCAGATCGGAGCCGGCC

ACTCTAGCGCTGGAATAAATCAGAATAAAGA!

GCATGACCCCCTGTGATATAATACCGTTCATC!

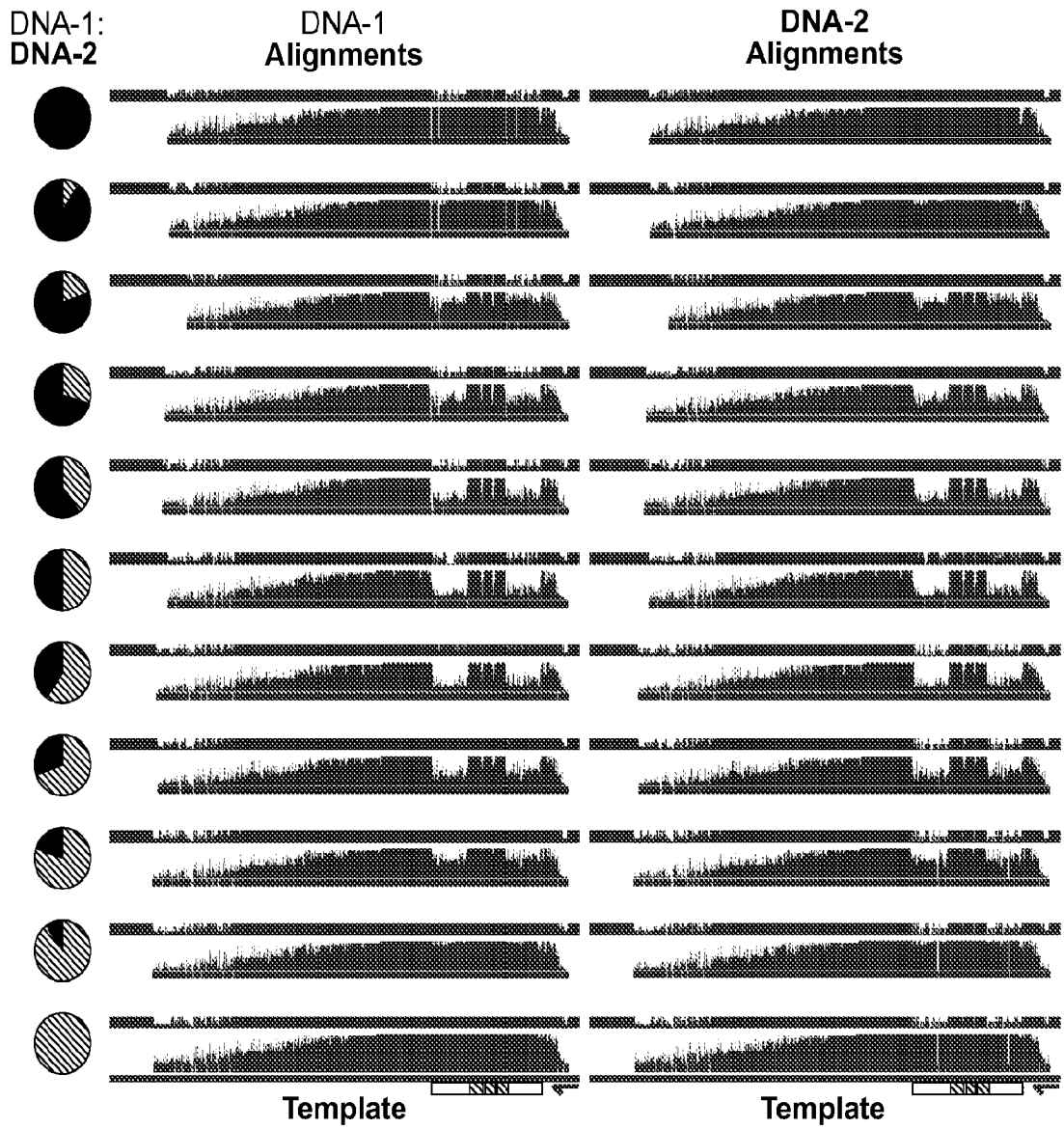
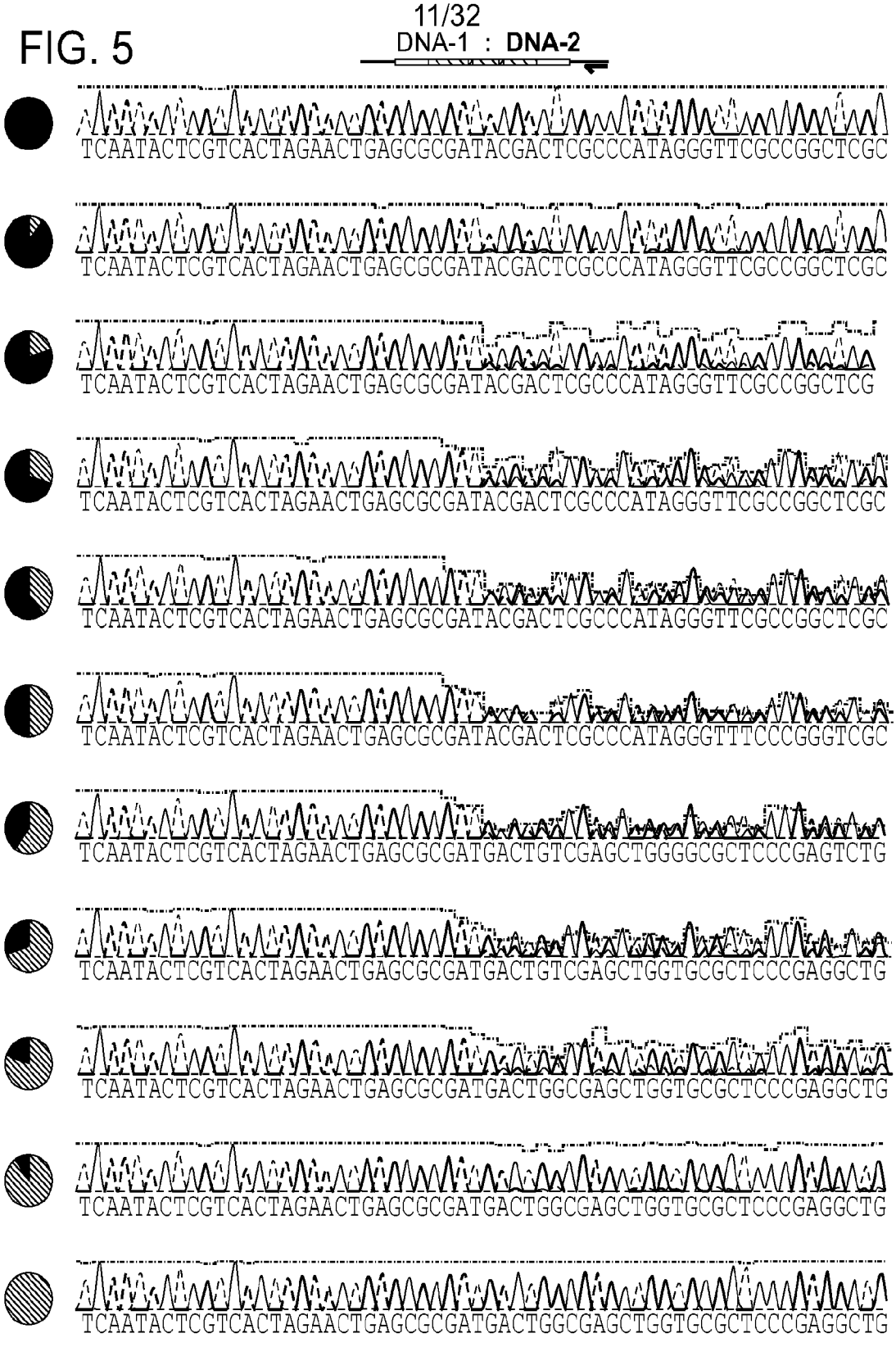

DNA-1:
DNA-2
DNA-1
Alignments
DNA-2
Alignments
Template
Template

FIG. 4

FIG. 5



SEQ ID NO: 53

TCAATACTCGTCACTAGAACTGAGCGCGATACGACTCGCCCATAGGGTTCGCCGGCTCGC!

! SEQ ID NO: 54

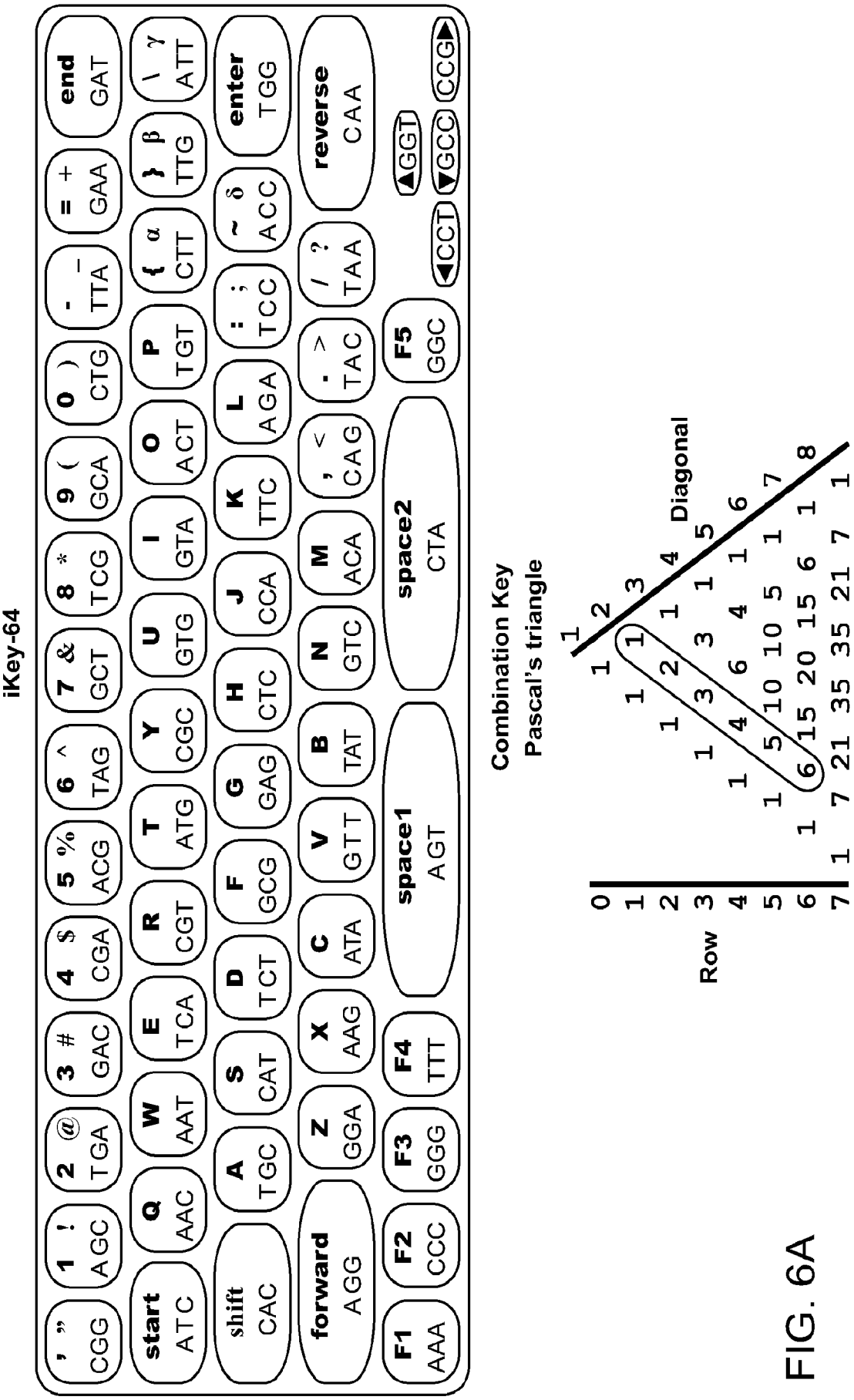

iKey-64
' " CGG
1 ! AGC
2 @ TGA
3 # GAC
4 $ CGA
5 % ACG
6 ^ TAG
7 & GCT
8 * TCG
9 (GCA
0) CTG
- _ TTA
= + GAA
end GAT
start ATC
Q AAC
W AAT
E TCA
R CGT
T ATG
Y CGC
U GTG
I GTA
O ACT
P TGT
{ α CTT
} β TTG
\ γ ATT
shift CAC
A TGC
S CAT
D TCT
F GCG
G GAG
H CTC
J CCA
K TTC
L AGA
: ; TCC
~ δ ACC
enter TGG
forward AGG
Z GGA
X AAG
C ATA
V GTT
B TAT
N GTC
M ACA
, < CAG
. > TAC
/ ? TAA
reverse CAA
F1 AAA
F2 CCC
F3 GGG
F4 TTT
space1 AGT
space2 CTA
F5 GGC
▲GGT
◀CCT
▼GCC
CCG▶
Combination Key
Pascal's triangle
Diagonal
Row
0
1
2
3
4
5
6
7
1
1 2
1 3
1 4
1 5
1 6
1 7
1 8
1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1

FIG. 6A

FIG. 6A (Cont.)

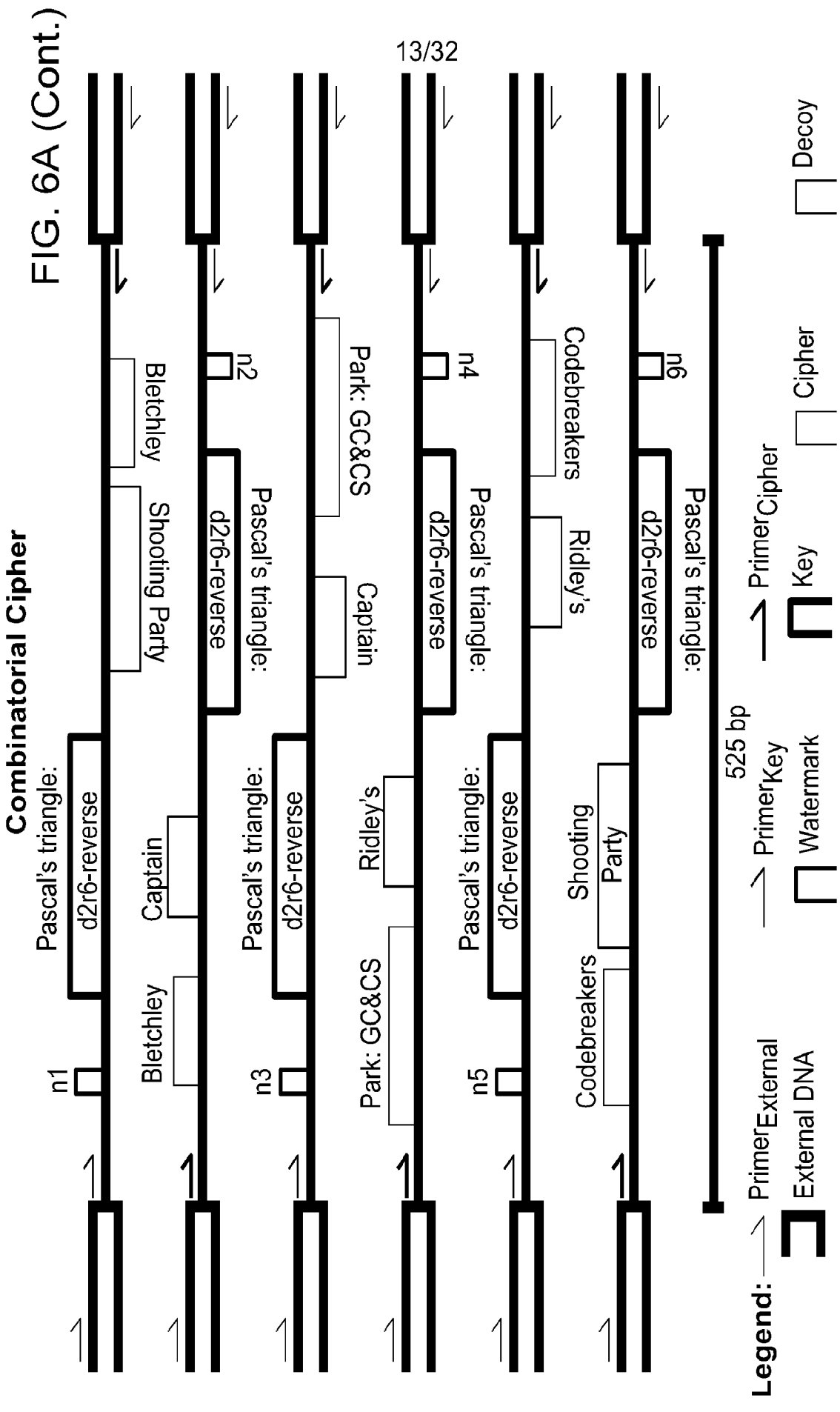

Combinatorial Cipher
Pascal's triangle:
d2r6-reverse
n1
Bletchley
Captain
Shooting Party
n2
Park: GC&CS
Ridley's
Codebreakers
n3
n4
n5
n6
Shooting
Party
525 bp
Legend:
Primer External
External DNA
Primer Key
Watermark
Primer Cipher
Key
Cipher
Decoy

FIG. 6B

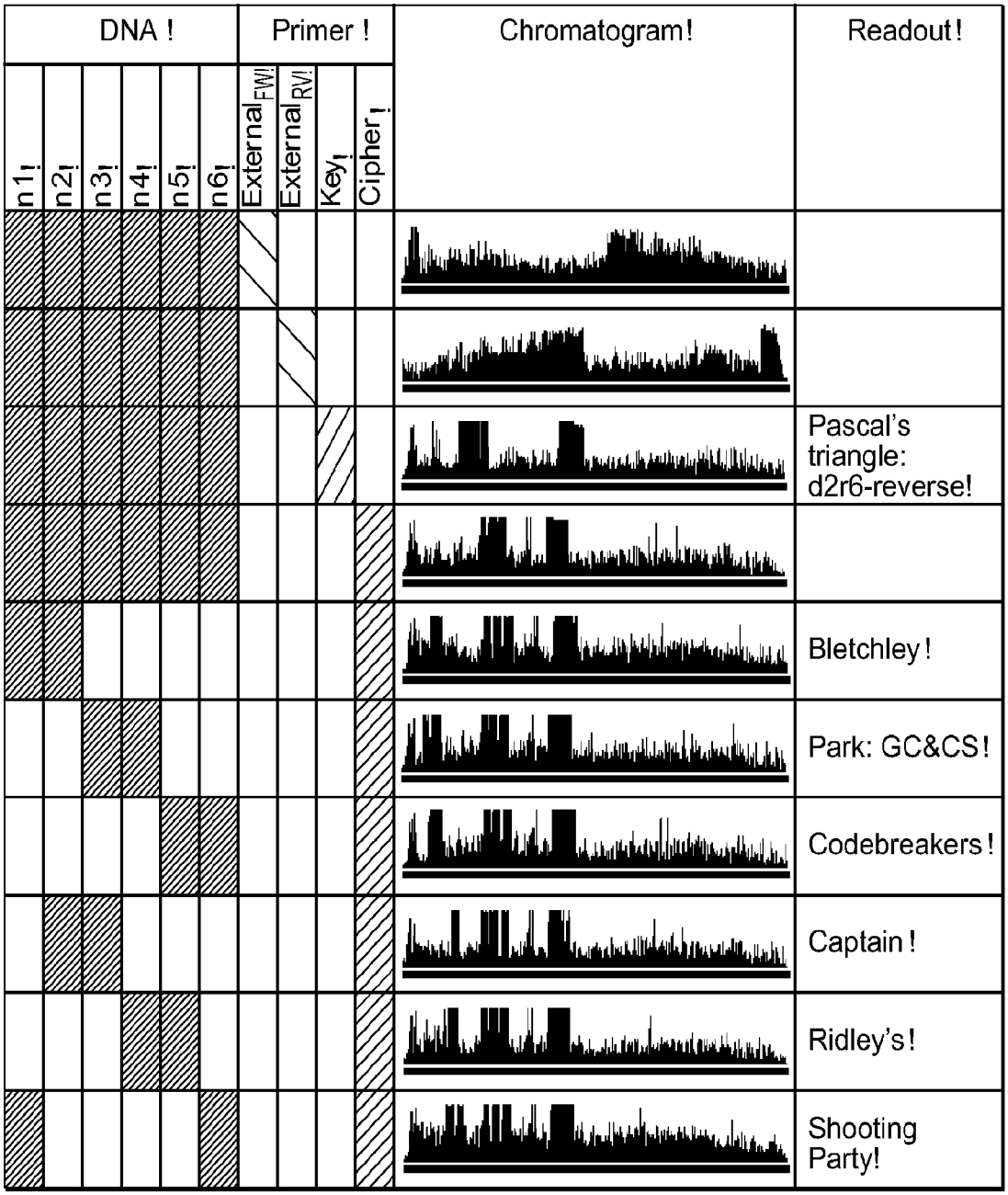

DNA !
Primer !
Chromatogram!
Readout!
n1!
n2!
n3!
n4!
n5!
n6!
External FW!
External RV!
Key!
Cipher !
Pascal's triangle: d2r6-reverse!
Bletchley!
Park: GC&CS!
Codebreakers!
Captain!
Ridley's!
Shooting Party!

FIG. 7

FIG. 8

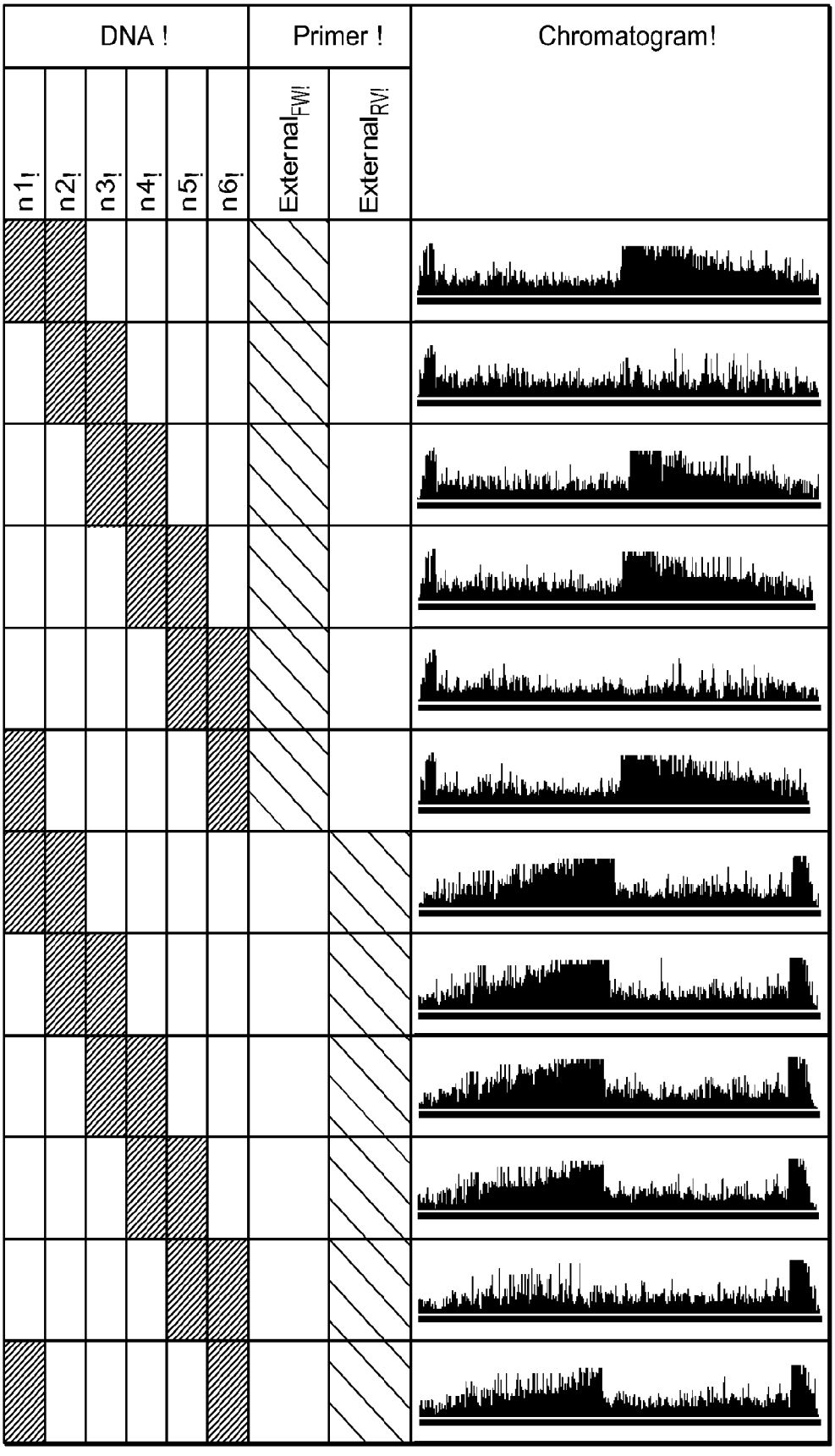

DNA !
Primer !
Chromatogram!
n1!
n2!
n3!
n4!
n5!
n6!
External$_{FW!}$
External$_{RV!}$

FIG. 9A

Key = Pascal's triangle: d2r6-reverse!
DNA = n1+n2+n3+n4+n5+n6 !
Sequence – Odd = ATCCACTGTTGCCATATATGCAGACGGCAT**AGT**ATGCGTGTATGCGTCGAGAGAT ! SEQ ID NO: 27
! ! CATCC**AGT**TCTTGACGTTAGTTACAAGAT!
Sequence – Even =ATCCACTGTTGCCATATATGCAGACGGCAT**CTA**ATGCGTGTATGCGTCGAGAGAT!
! ! CATCC**CTA**TCTTGACGTTAGTTACAAGAT ! SEQ ID NO: 28
Chromatogram: !

CTCGATGATCGAGTCGCATCCACTGTTGCCATATATGCAGACGGCATCTAATGCGTGTATGCGTCGAGAGATCATCCCTATCTTGACGTTAGTTACAAGATTGCACNCGAT SEQ ID NO: 29

FIG. 9B

Cipher = Bletchley !
DNA = n1+n2 !
Sequence – Odd = ATCCACTATAGATCAATGATACTCACGATCACGCGAT ! SEQ ID NO: 30
Sequence – Even = ATCCACTATAGATCAATGATACTCAGATCACGCGAT &!
Chromatogram: ! SEQ ID NO: 31

AAGCTTGCATGCACGTCATGATCCACTATAGATCAATGATACTCAGATCACGCGATATCGACGTTGACTAGT SEQ ID NO: 32

FIG. 9C

Cipher = Park: GC&CS !

DNA = n3+n4 ! SEQ ID NO: 33

Sequence – Odd = ATCCACACTGTTGCCGTTTCTCCC**AGT**CACGAGCACATACACGCTCACATACCATGAT !

Sequence – Even = ATCCACACTGTTGCCGTTTCTCCC**CTA**CACGAGCACATACACGCTCACATACCATGAT !

Chromatogram: ! SEQ ID NO: 34

SEQ ID NO: 35

FIG. 9D

Cipher = Codebreakers !

DNA = n5+n6 !

Sequence – Odd = ATCCACACATAACTTCTTCATATCGTTCATGCTTCTCACGTCATGAT ! SEQ ID NO: 36

Sequence – Even = ATCCACACATAACTTCTTCATATCGTTCATGCTTCTCACGTCATGAT ! SEQ ID NO: 37

Chromatogram: !

SEQ ID NO: 38

FIG. 9E

Cipher = Captain !
DNA = n2+n3 !
Sequence – Odd = ATCCACATATGCTGTATGTGCGTAGTCGAT ! SEQ ID NO: 39
Sequence – Even = ATCCACATATGCTGTATGTGCGTAGTCGAT ! SEQ ID NO: 40
Chromatogram: !

SEQ ID NO: 41

FIG. 9F

Cipher = Ridley's !
DNA = n4+n5 !
Sequence – Odd = ATCCACACCGTGTATCTAGATCACGCCGGCATGAT ! SEQ ID NO: 42
Sequence – Even = ATCCACACCGTGTATCTAGATCACGCCGGCATGAT ! SEQ ID NO: 43
Chromatogram: !

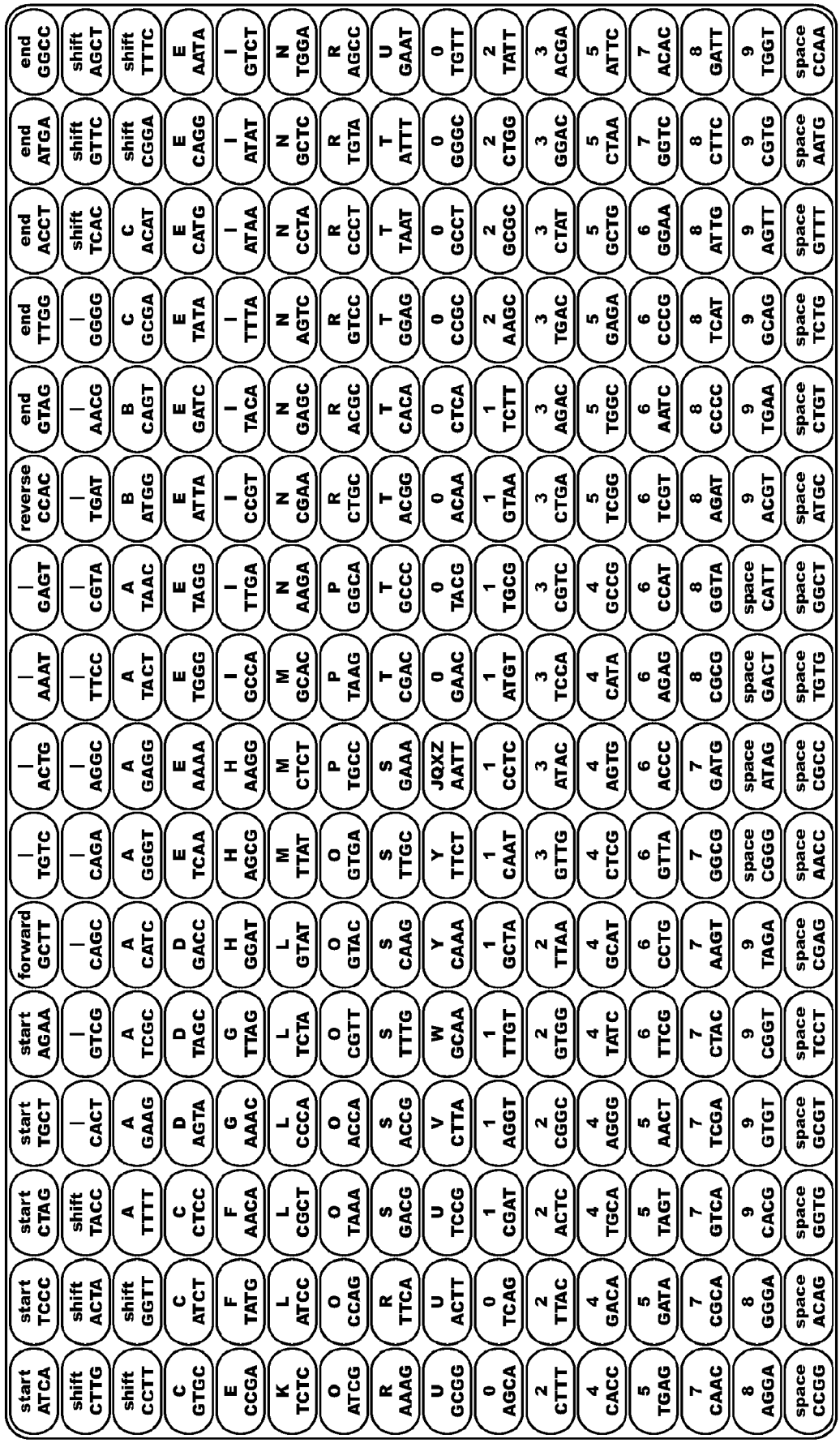
SEQ ID NO: 44

FIG. 9G

Cipher = Shooting Party !
DNA = n6+n1 !
Sequence – Odd = ATCCACCATCTCACTACTATGGTAGTCGAG**AGT**CACTGTTGCCGTATGCGCGAT ! SEQ ID NO: 45
Sequence – Even = ATCCACCATCTCACTACTATGGTAGTCGAG**CTA**CACTGTTGCCGTATGCGCGAT ! SEQ ID NO: 46
Chromatogram: !

SEQ ID NO: 47

FIG. 10


start ATCA | start TCCC | start CTAG | start TGCT | start AGAA | forward GCTT | | TGTC | | ACTG | | AAAT | | GAGT | reverse CCAC | end GTAG | end TTGG | end ACCT | end ATGA | end GGCC
shift CTTG | shift ACTA | shift TACC | | GTCG | | CAGC | | CAGA | | AGGC | | TTCC | | CGTA | | TGAT | | AACG | | GGGG | | shift TCAC | shift GTTC | shift AGCT
shift CCTT | shift GGTT | A TTTT | A GAAG | A TCGC | A CATC | A GGGT | A GAGG | A TACT | A TAAC | A ATGG | B CAGT | C GCGA | C ACAT | shift CGGA | shift TTTC
C GTGC | C ATCT | C CTCC | D AGTA | D TAGC | D GACC | E TCAA | E AAAA | E TGGG | E TAGG | E ATTA | E GATC | E TATA | E CATG | E CAGG | E AATA
E CCGA | F TATG | F AACA | G AAAC | G TTAG | H GGAT | H AGCG | H AAGG | I GCCA | I TTGA | I CCGT | I TACA | I TTTA | I ATAA | I ATAT | I GTCT
K TCTC | L ATCC | L CGCT | L CCCA | L TCTA | L GTAT | M TTAT | M CTCT | M GCAC | N AAGA | N CGAA | N GAGC | N AGTC | N CCTA | N GCTC | N TGGA
O ATCG | O CCAG | O TAAA | O ACCA | O CGTT | O GTAC | O GTGA | P TGCC | P TAAG | P GGCA | R CTGC | R ACGC | R GTCC | R CCCT | R TGTA | R AGCC
R AAAG | R TTCA | S GACG | S ACCG | S TTTG | S CAAG | S TTGC | S GAAA | T CGAC | T GCCC | T ACGG | T CACA | T GGAG | T TAAT | T ATTT | U GAAT
U GCGG | U ACTT | U TCCG | V CTTA | W GCAA | Y CAAA | Y TTCT | JQXZ AATT | 0 GAAC | 0 TACG | 0 ACAA | 0 CTCA | 0 CCGC | 0 GCCT | 0 GGGC | 0 TGTT
0 AGCA | 0 TCAG | 1 CGAT | 1 AGGT | 1 TTGT | 1 GCTA | 1 CAAT | 1 CCTC | 1 ATGT | 1 TGCG | 1 GTAA | 1 TCTT | 2 AAGC | 2 GCGC | 2 CTGG | 2 TATT
2 CTTT | 2 TTAC | 2 ACTC | 2 CGGC | 2 GTGG | 2 TTAA | 3 GTTG | 3 ATAC | 3 TCCA | 3 CGTC | 3 CTGA | 3 AGAC | 3 TGAC | 3 CTAT | 3 GGAC | 3 ACGA
4 CACC | 4 GACA | 4 TGCA | 4 AGGG | 4 TATC | 4 GCAT | 4 CTCG | 4 AGTG | 4 CATA | 4 GCCG | 5 TCGG | 5 TGGC | 5 GAGA | 5 GCTG | 5 CTAA | 5 ATTC
5 TGAG | 5 GATA | 5 TAGT | 5 AACT | 6 TTCG | 6 CCTG | 6 GTTA | 6 ACCC | 6 AGAG | 6 CCAT | 6 TCGT | 6 AATC | 6 CCCG | 6 GGAA | 7 GGTC | 7 ACAC
7 CAAC | 7 CGCA | 7 GTCA | 7 TCGA | 7 CTAC | 7 AAGT | 7 GGCG | 7 GATG | 8 CGCG | 8 GGTA | 8 AGAT | 8 CCCC | 8 TCAT | 8 ATTG | 8 CTTC | 8 GATT
8 AGGA | 8 GGGA | 9 CACG | 9 GTGT | 9 CGGT | 9 TAGA | space CGGG | space ATAG | space GACT | space CATT | 9 ACGT | 9 TGAA | 9 GCAG | 9 AGTT | 9 CGTG | 9 TGGT
space CCGG | space ACAG | space GGTG | space GCGT | space TCCT | space CGAG | space AACC | space CGCC | space TGTG | space GGCT | space ATGC | space CTGT | space TCTG | space GTTT | space AATG | space CCAA

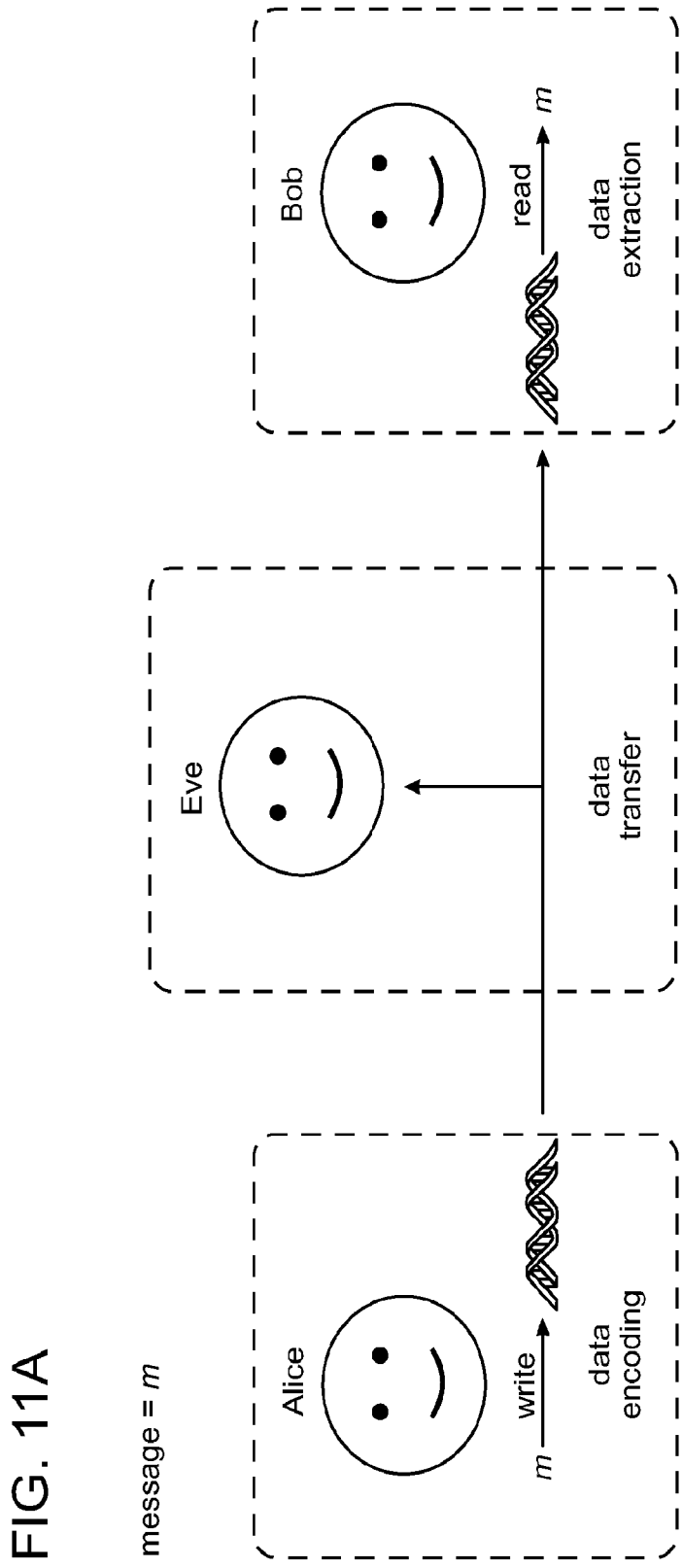
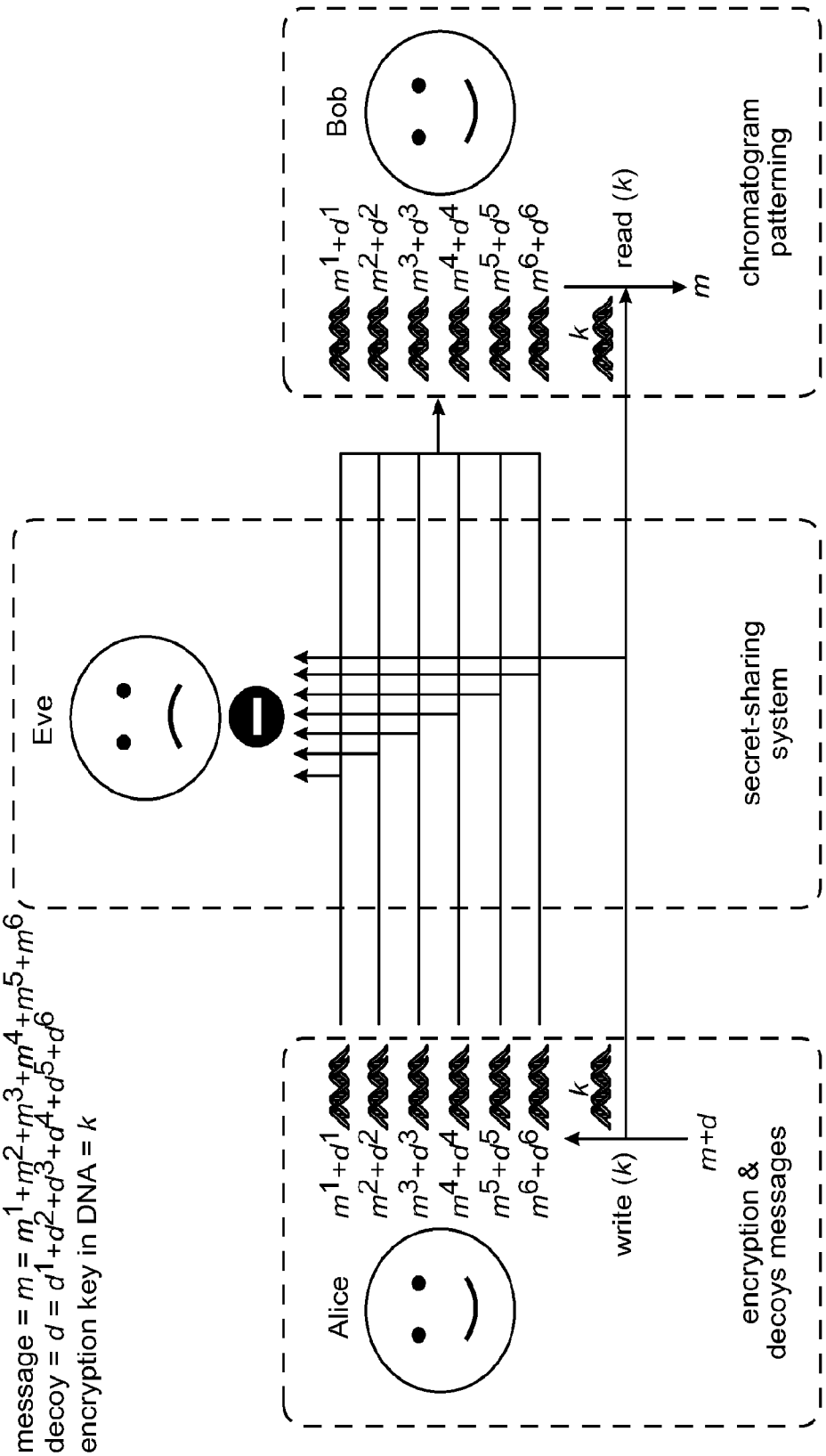

message = m
Alice
write
m
data
encoding
Eve
data
transfer
Bob
read
m
data
extraction

FIG. 11A

FIG. 11B


message = $m = m^1+m^2+m^3+m^4+m^5+m^6$
decoy = $d = d^1+d^2+d^3+d^4+d^5+d^6$
encryption key in DNA = k
Alice
m^1+d^1
m^2+d^2
m^3+d^3
m^4+d^4
m^5+d^5
m^6+d^6
k
write (k)
$m+d$
encryption & decoys messages
Eve
secret-sharing system
Bob
read (k)
m
chromatogram patterning

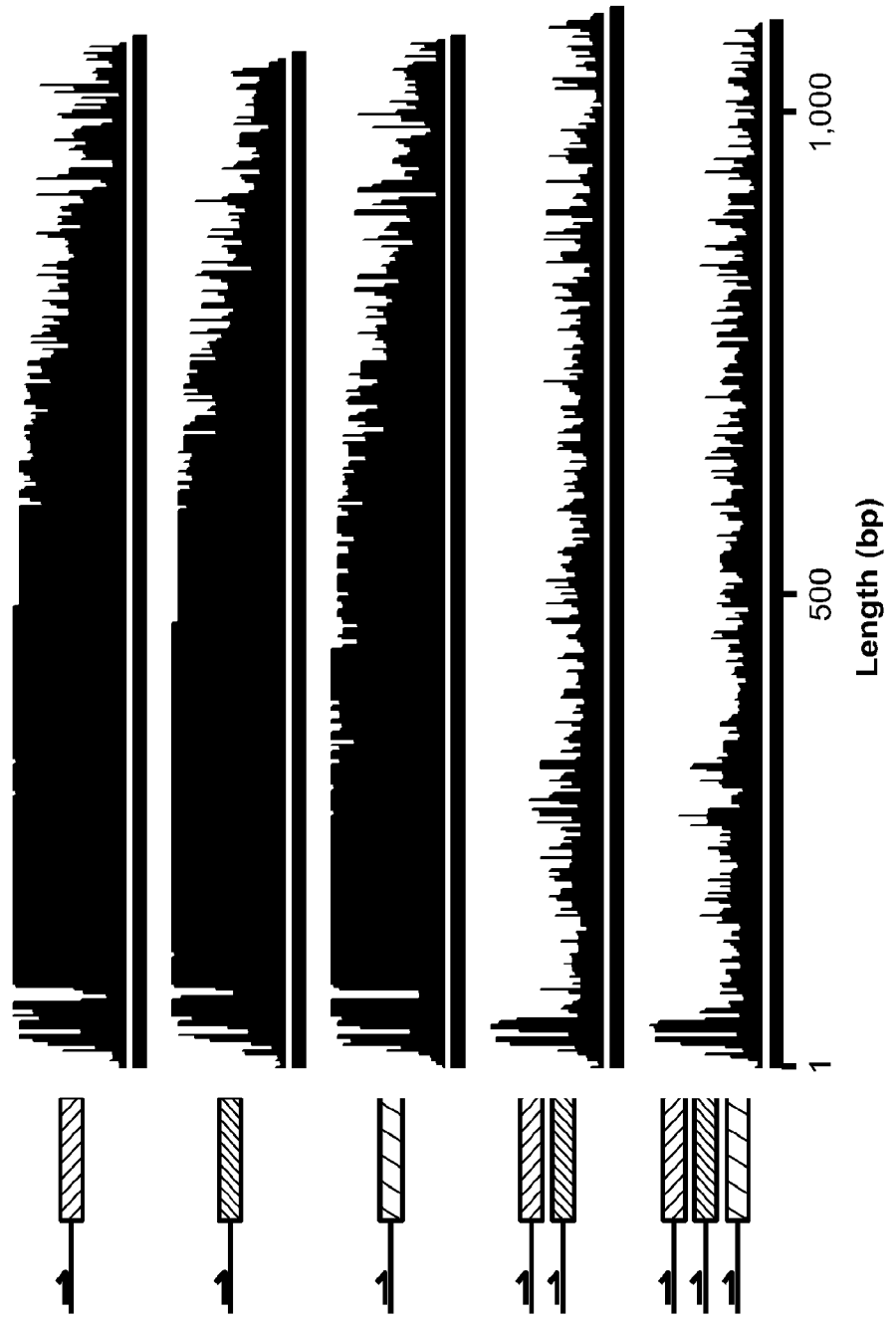
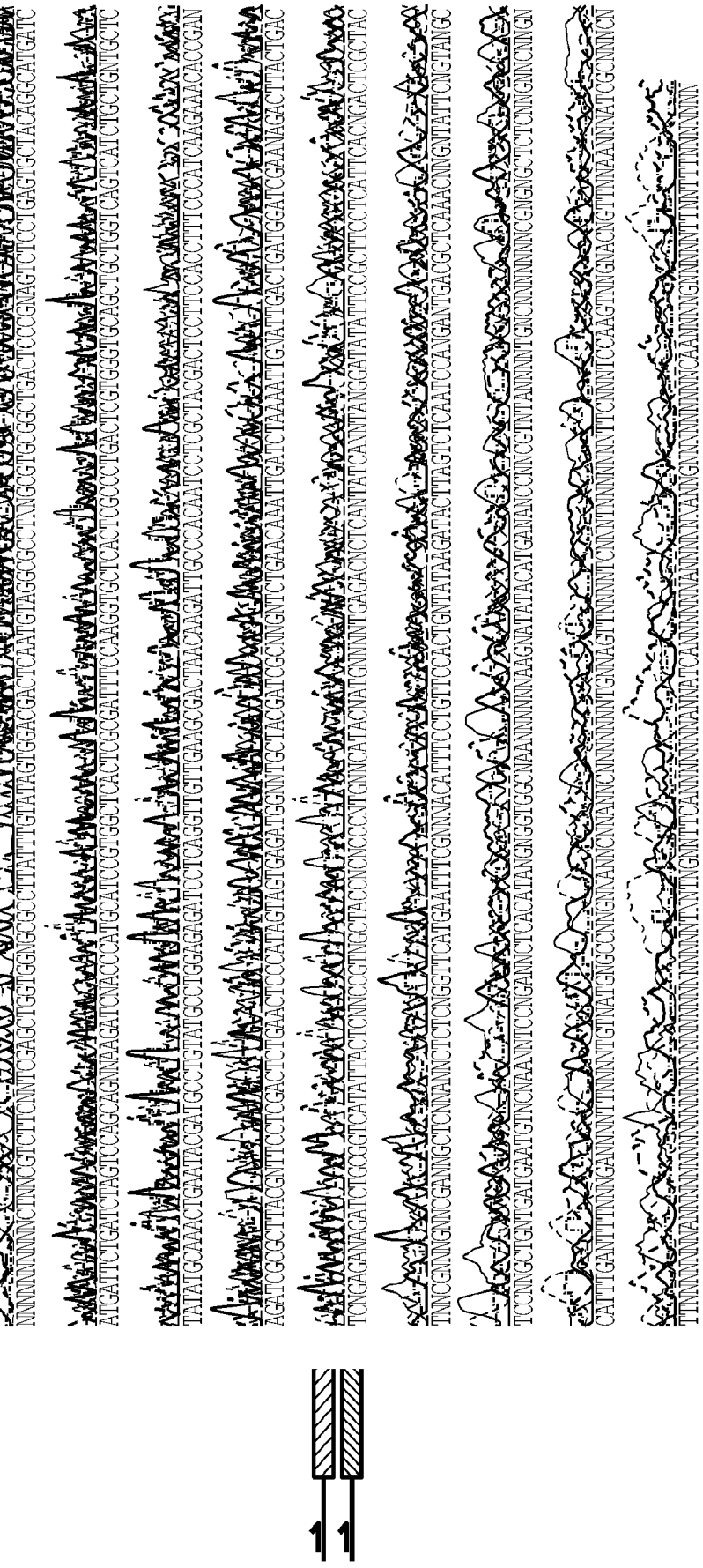


FIG. 12A

FIG. 12B



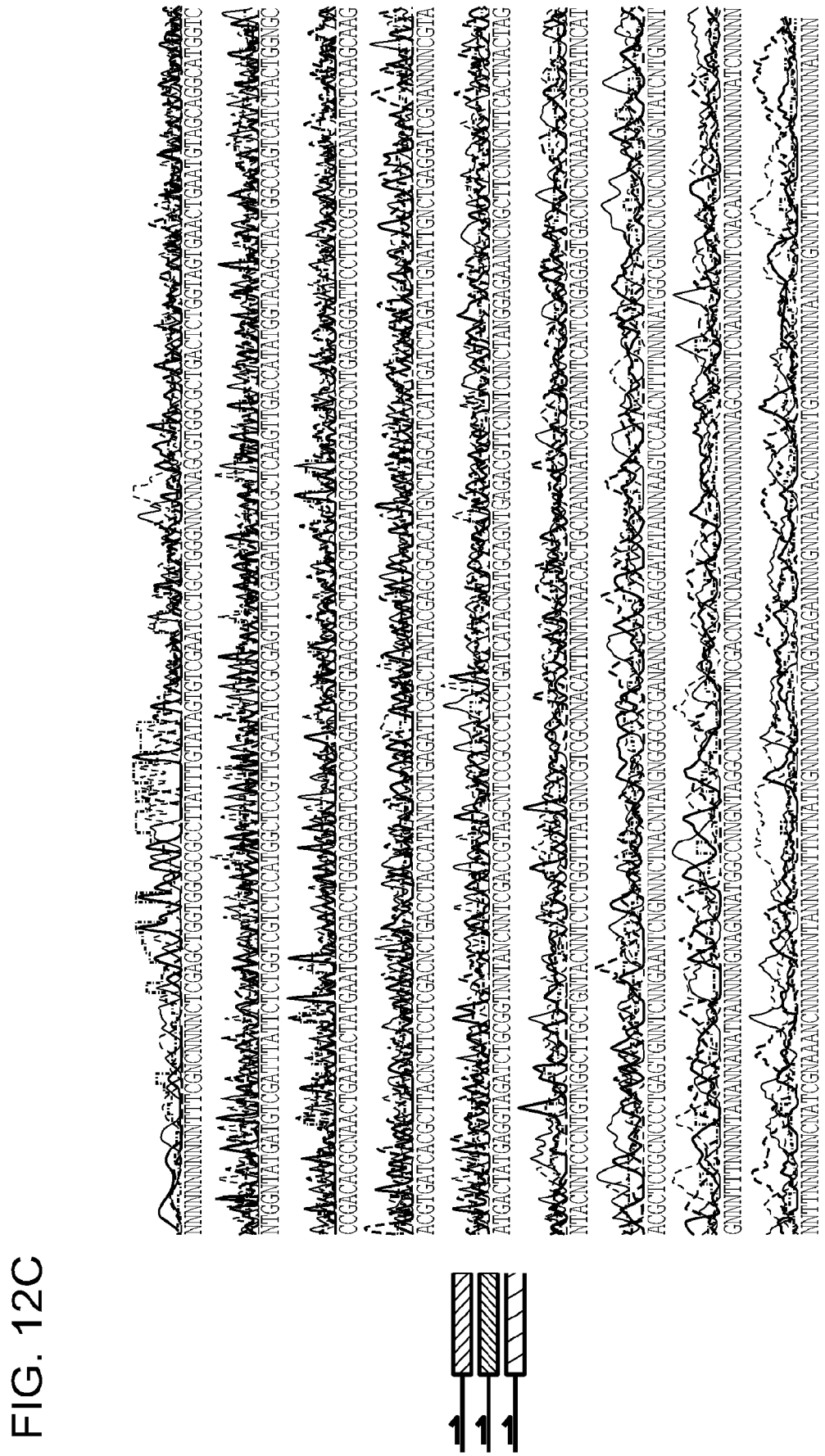


FIG. 12C

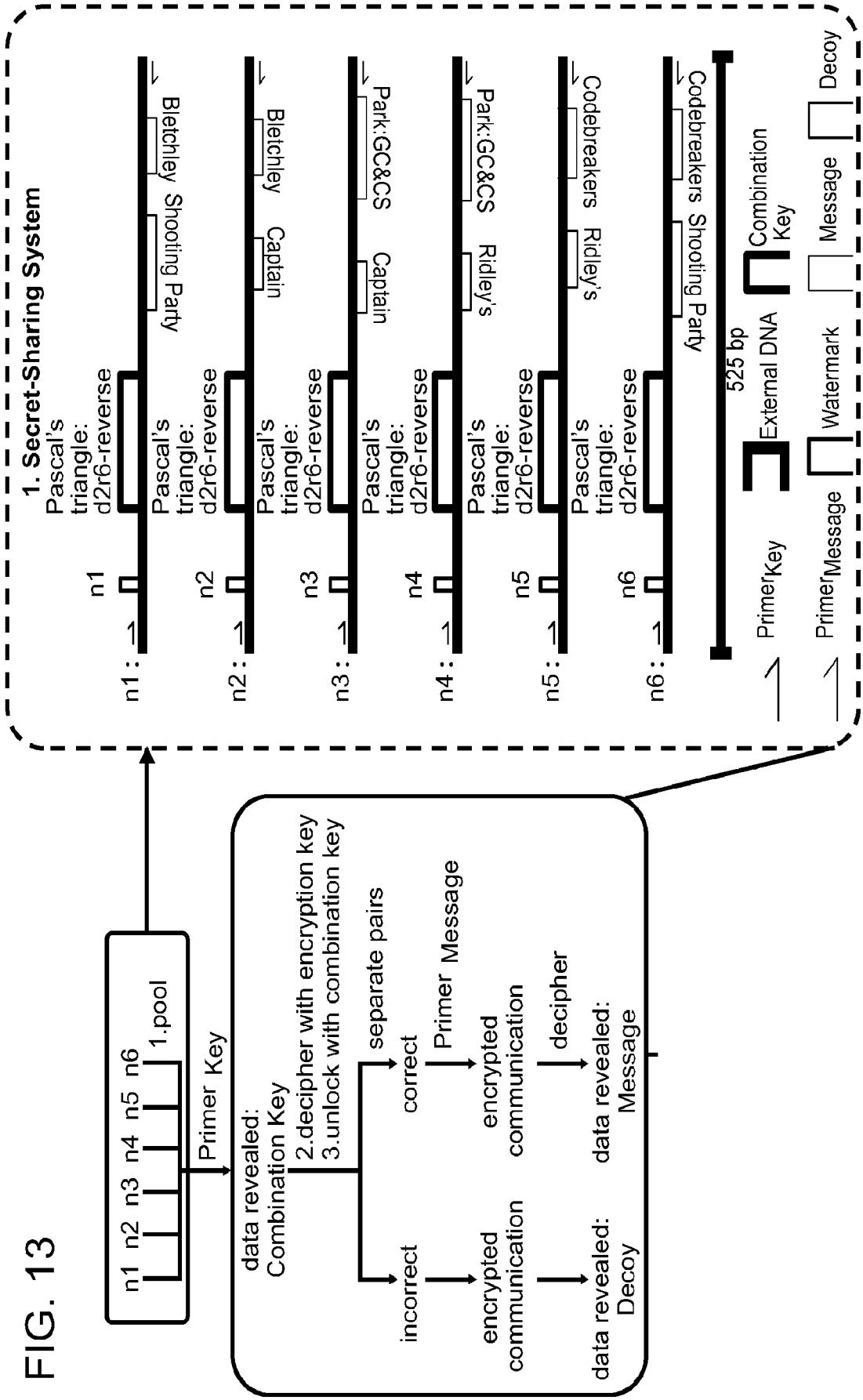

1. Secret-Sharing System
Pascal's triangle: d2r6-reverse
Bletchley Shooting Party
Bletchley Captain
Park:GC&CS Captain
Park:GC&CS Ridley's
Codebreakers Ridley's
Codebreakers Shooting Party
n1
n2
n3
n4
n5
n6
525 bp
External DNA
Combination Key
Message
Watermark
Decoy
Primer Key
Primer Message
1.pool
Primer Key
data revealed: Combination Key
2.decipher with encryption key
3.unlock with combination key
separate pairs
correct
Primer Message
encrypted communication
decipher
data revealed: Message
incorrect
encrypted communication
data revealed: Decoy

FIG. 13

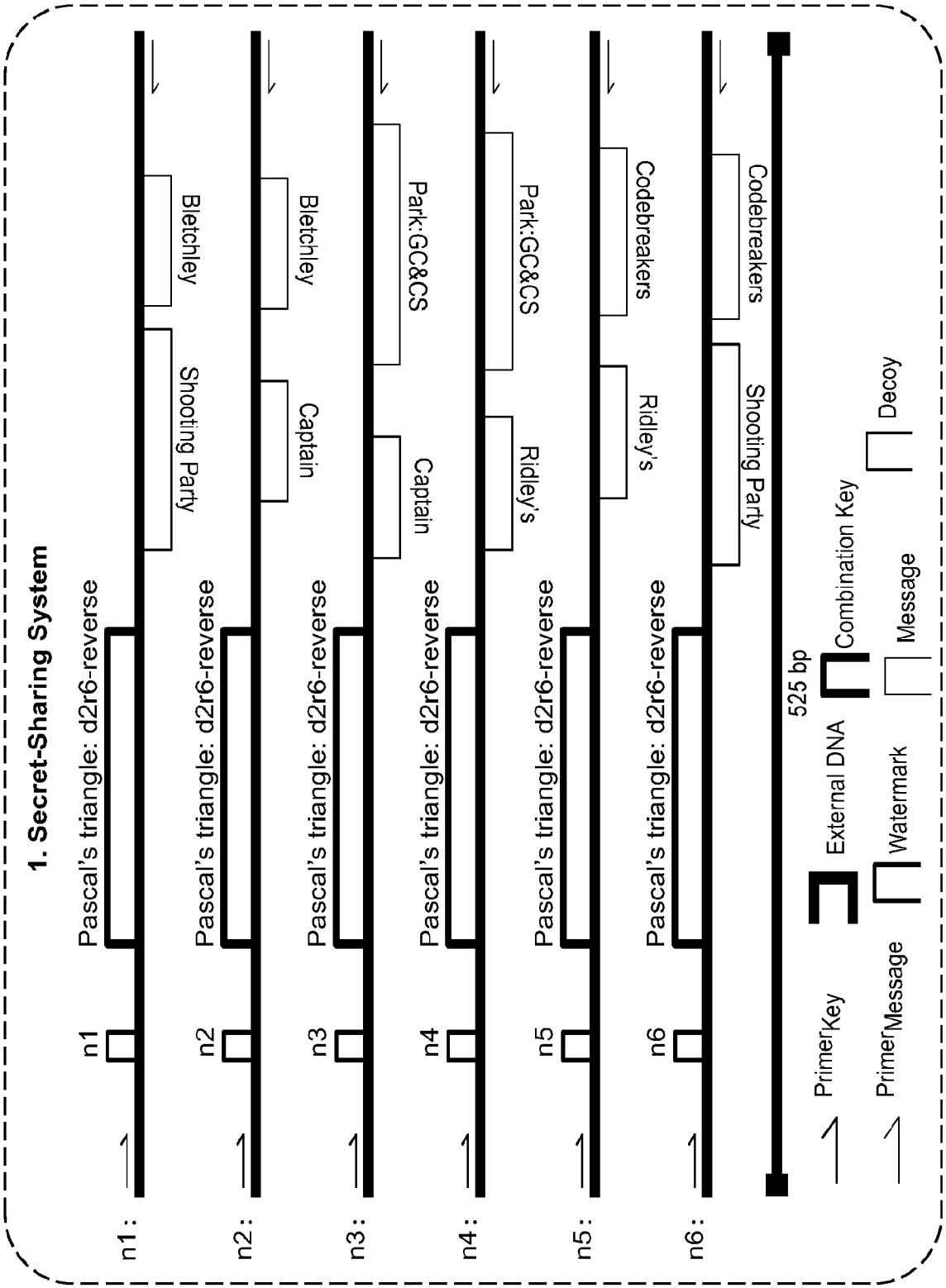

1. Secret-Sharing System
n1:
n2:
n3:
n4:
n5:
n6:
n1
n2
n3
n4
n5
n6
Pascal's triangle: d2r6-reverse
Bletchley
Shooting Party
Captain
Park:GC&CS
Ridley's
Codebreakers
525 bp
Primer Key
Primer Message
External DNA
Watermark
Combination Key
Message
Decoy

FIG. 14A

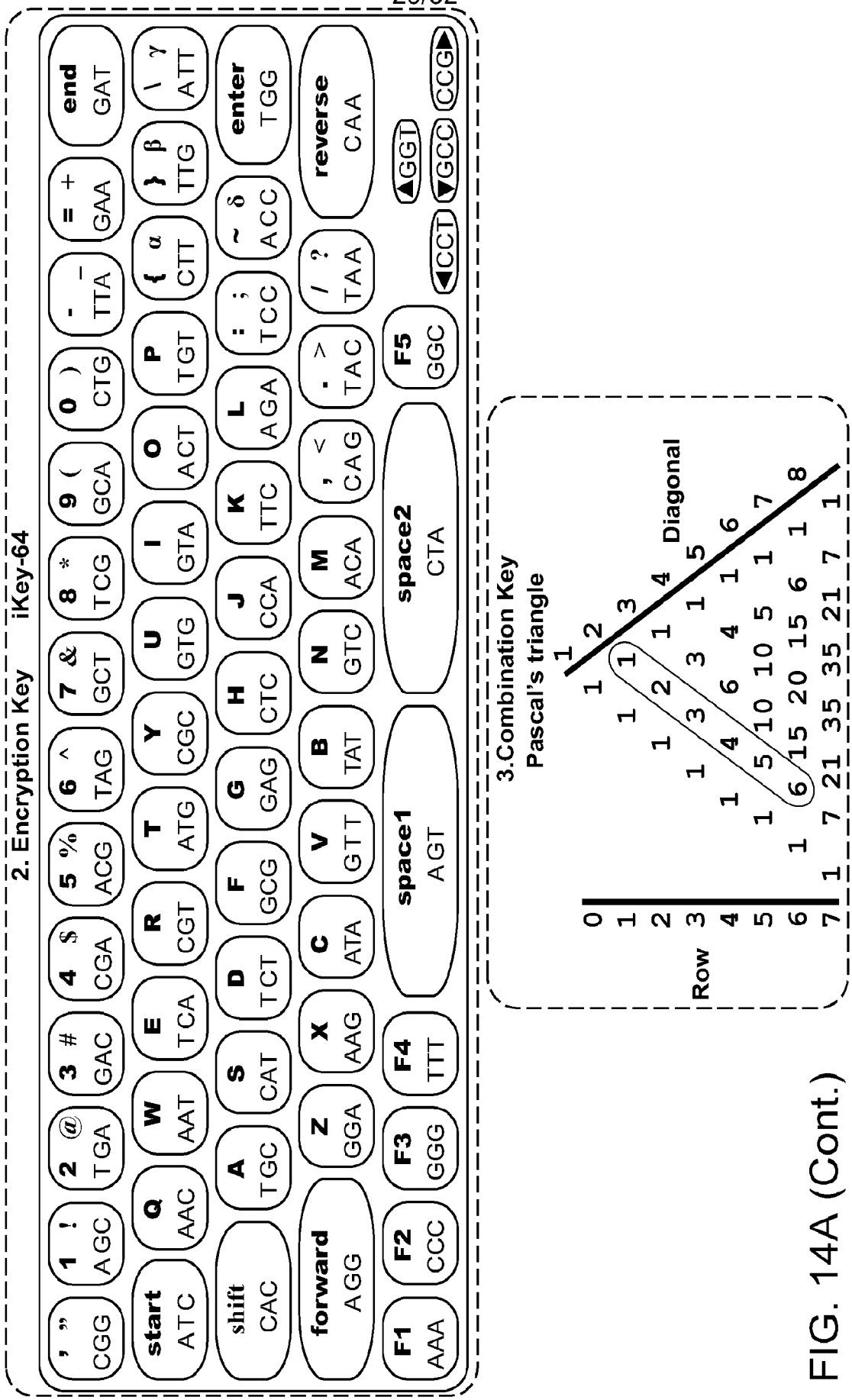

2. Encryption Key iKey-64
' " CGG
1 ! AGC
2 @ TGA
3 # GAC
4 $ CGA
5 % ACG
6 ^ TAG
7 & GCT
8 * TCG
9 (GCA
0) CTG
- _ TTA
= + GAA
end GAT
start ATC
Q AAC
W AAT
E TCA
R CGT
T ATG
Y CGC
U GTG
I GTA
O ACT
P TGT
{ α CTT
} β TTG
\ γ ATT
shift CAC
A TGC
S CAT
D TCT
F GCG
G GAG
H CTC
J CCA
K TTC
L AGA
: ; TCC
~ δ ACC
enter TGG
forward AGG
Z GGA
X AAG
C ATA
V GTT
B TAT
N GTC
M ACA
, < CAG
. > TAC
/ ? TAA
reverse CAA
F1 AAA
F2 CCC
F3 GGG
F4 TTT
space1 AGT
space2 CTA
F5 GGC
◀CCT
▼GCC
CCG▶
▲GGT
3.Combination Key
Pascal's triangle
Row
Diagonal

FIG. 14A (Cont.)

FIG. 15

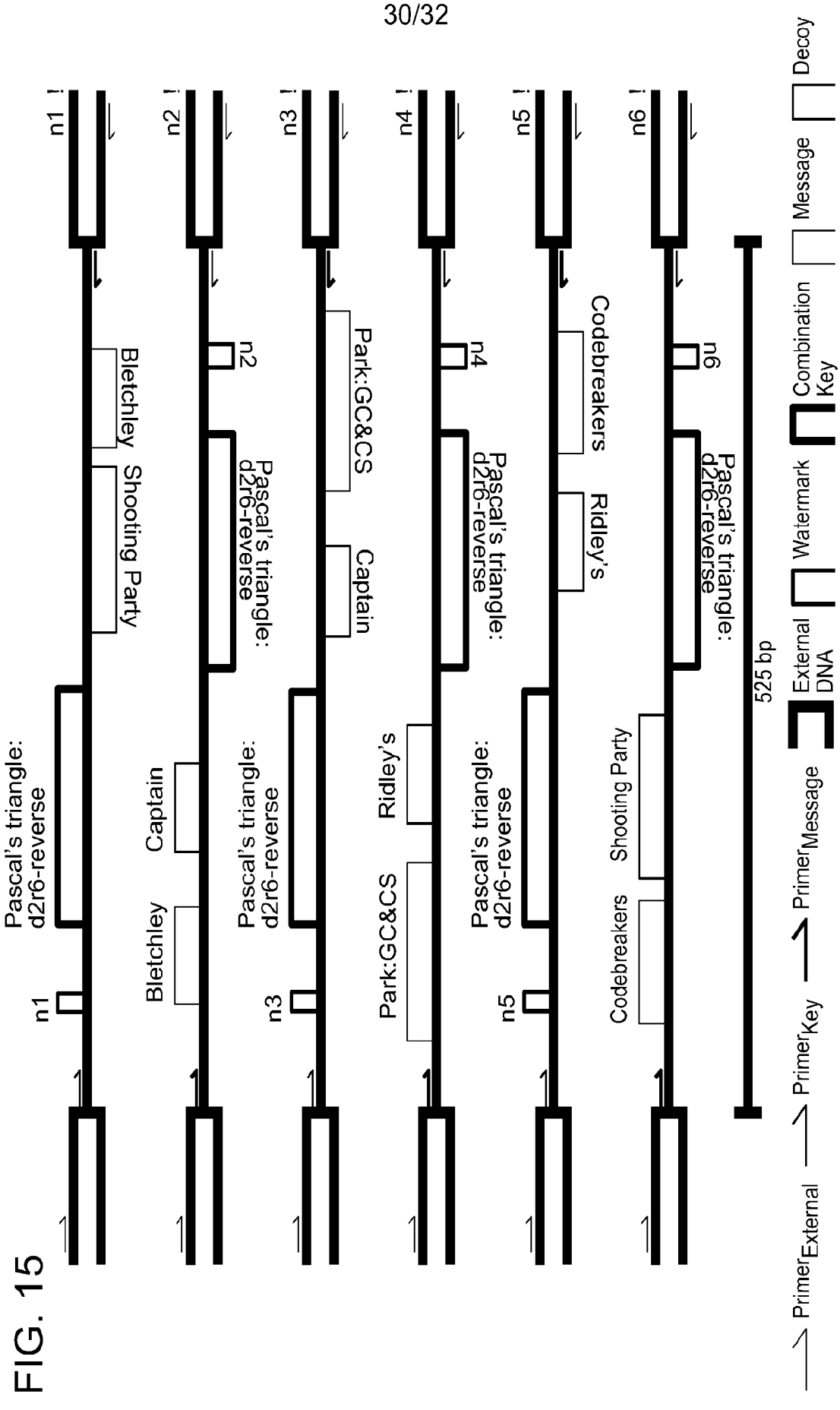

n1
Pascal's triangle: d2r6-reverse
Bletchley
Shooting Party
n2
Bletchley
Captain
Pascal's triangle: d2r6-reverse
n2
n3
n3
Park:GC&CS
Captain
Pascal's triangle: d2r6-reverse
n4
n4
Park:GC&CS
Ridley's
Pascal's triangle: d2r6-reverse
n5
n5
Codebreakers
Ridley's
Pascal's triangle: d2r6-reverse
n6
n6
Codebreakers
Shooting Party
Pascal's triangle: d2r6-reverse
525 bp
Primer External
Primer Key
Primer Message
External DNA
Watermark
Combination Key
Message
Decoy

FIG. 16A

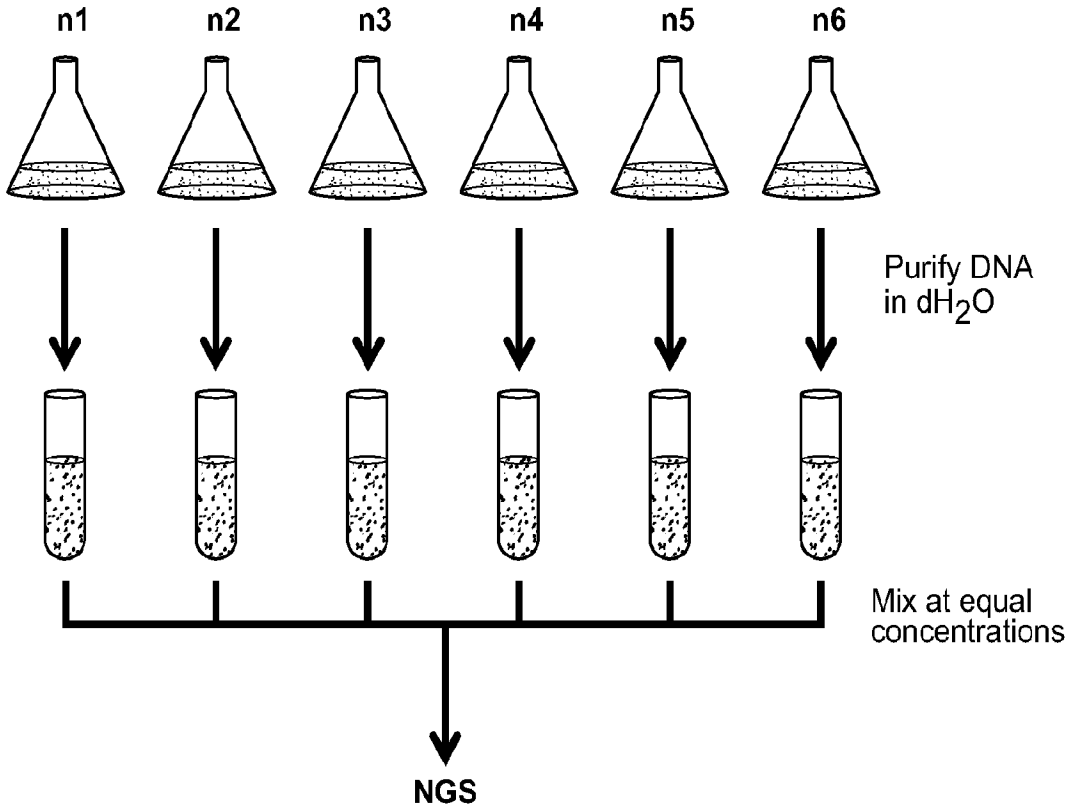


FIG. 16B

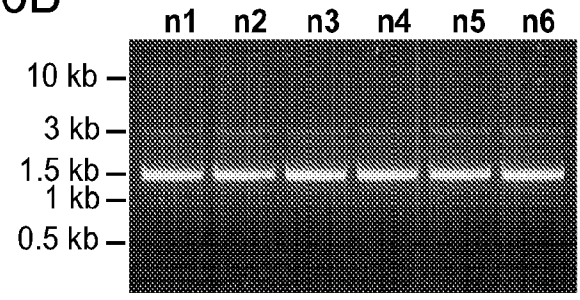
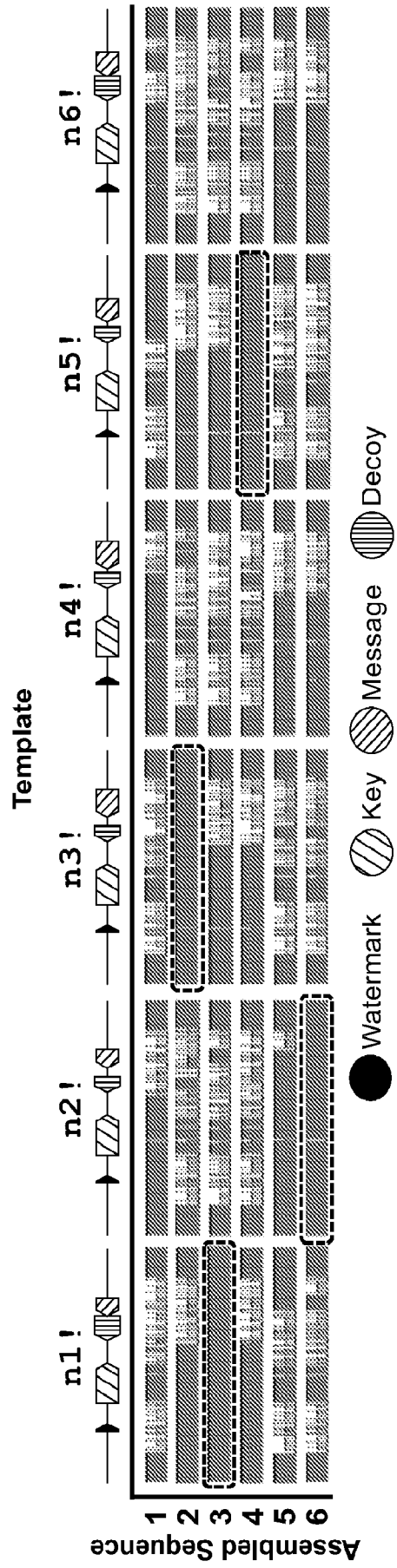


FIG. 16C


Template
n1!
n2!
n3!
n4!
n5!
n6!
Assembled Sequence
1
2
3
4
5
6
Watermark
Key
Message
Decoy

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/058120

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 21/00 (2016.01)
CPC - G06F 21/6254 (2016.02)
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - G06F 12/14, 17/30, 21/00; G09C 1/00; H04L 9/32 (2016.01)
CPC - G06F 21/6254; H04L 2209/60, 2209/88, 9/0822, 9/32 (2016.02)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC 380/283; 506/9; 705/51 (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Patbase, Google Patents , PubMed, Google.
Search terms used: encrypt secure information storage cipher DNA RNA genome nucleic acid keyboard codons

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2013/0046994 A1 (SHAW) 21 February 2013 (21.02.2013) entire document	1-4, 19-21
A	WO 2003/019159 A1 (FIRST GENETIC TRUST INC et al) 06 March 2003 (06.03.2003) entire document	1-4, 19-21
A	US 2012/0230326 A1 (GANESHALINGAM et al) 13 September 2012 (13.09.2012) entire document	1-4, 19-21
A	SHAW et al. "Genomics-based Security Protocols: From Plaintext to Cipherprotein," International Journal on Advances in Security, 2 January 2011(02.01.2011), Vol 4, Pgs. 106-117. entire document	1-4, 19-21
A	WO 2011/053868 A1 (SYNTHETIC GENOMICS, INC. et al) 05 May 2011 (05.05.2011) entire document	1-4, 19-21

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
01 April 2016	21 APR 2016
Name and mailing address of the ISA/ Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, VA 22313-1450 Facsimile No. 571-273-8300	Authorized officer Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/058120

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 5-18, 22-26
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.