

MIT Open Access Articles

*Multiplexed Sequence Encoding: A
Framework for DNA Communication*

The MIT Faculty has made this article openly available. ***Please share***
how this access benefits you. Your story matters.

Citation: Zakeri Bijan, Peter A. Carr, and Timothy K. Lu. "Multiplexed Sequence Encoding: A Framework for DNA Communication." PLoS ONE 11:4 (2016): e0152774.

As Published: <http://dx.doi.org/10.1371/journal.pone.0152774>

Publisher: Public Library of Science

Persistent URL: <http://hdl.handle.net/1721.1/103080>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution



RESEARCH ARTICLE

Multiplexed Sequence Encoding: A Framework for DNA Communication

Bijan Zakeri^{1,2*}, Peter A. Carr^{2,3*}, Timothy K. Lu^{1,2}

1 Department of Electrical Engineering and Computer Science, Department of Biological Engineering, Research Laboratory of Electronics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, United States of America, **2** MIT Synthetic Biology Center, 500 Technology Square, Cambridge, MA 02139, United States of America, **3** MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420, United States of America

* bjan.zakeri@oxfordalumni.org (BZ); carr@ll.mit.edu (PAC)



CrossMark
click for updates

OPEN ACCESS

Citation: Zakeri B, Carr PA, Lu TK (2016) Multiplexed Sequence Encoding: A Framework for DNA Communication. PLoS ONE 11(4): e0152774. doi:10.1371/journal.pone.0152774

Editor: Albert Jeltsch, Universität Stuttgart, GERMANY

Received: December 3, 2015

Accepted: February 28, 2016

Published: April 6, 2016

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Data Availability Statement: All relevant data are within the paper.

Funding: This work is sponsored by the Defense Advanced Research Projects Agency under Air Force Contract #FA8721-05-C-0002, and NIH grants 1R01EB017755, 1DP2OD008435, and 1P50GM098792.

Competing Interests: The authors declare competing financial interests. B.Z., P.A.C., and T.K.L. are authors on a US patent application (Name: DNA Encryption Technologies, Application No.: 62069994). This does not alter the authors' adherence to PLOS ONE policies on sharing data and materials. There

Abstract

Synthetic DNA has great propensity for efficiently and stably storing non-biological information. With DNA writing and reading technologies rapidly advancing, new applications for synthetic DNA are emerging in data storage and communication. Traditionally, DNA communication has focused on the encoding and transfer of complete sets of information. Here, we explore the use of DNA for the communication of short messages that are fragmented across multiple distinct DNA molecules. We identified three pivotal points in a communication—data encoding, data transfer & data extraction—and developed novel tools to enable communication via molecules of DNA. To address data encoding, we designed DNA-based individualized keyboards (iKeys) to convert plaintext into DNA, while reducing the occurrence of DNA homopolymers to improve synthesis and sequencing processes. To address data transfer, we implemented a secret-sharing system—Multiplexed Sequence Encoding (MuSE)—that conceals messages between multiple distinct DNA molecules, requiring a combination key to reveal messages. To address data extraction, we achieved the first instance of chromatogram patterning through multiplexed sequencing, thereby enabling a new method for data extraction. We envision these approaches will enable more widespread communication of information via DNA.

Introduction

Communication has many faces. While the general objective of transferring information between different parties remains constant, the medium for information transfer continues to evolve. In nature, DNA has been used for billions of years as the chemical of choice for transferring information across cells, species, and generations [1]. Now, advances in biotechnology are enabling the use of DNA for the transfer and storage of non-biological information [2–9].

Rapid advances in digital technologies over the past decades have enabled efficient and facile communication regardless of whether our messages are short tweets of 140 characters or long communiqués of thousands of pages. However, our ever-increasing reliance on digital

are no other declarations relating to employment, consultancy, patents, products in development or modified products, etc.

technologies may make it worthwhile to explore new methods of information storage and transfer with alternative characteristics. Key attributes of DNA including high-density data storage, static and stable data maintenance, efficient reproducibility, and the lack of technological obsolescence, mean that DNA has distinct advantages over current magnetic and optical data storage platforms [1], thus warranting further exploration into DNA-specific writing and reading technologies.

With synthesis and sequencing speeds rising, and costs rapidly declining [10,11], DNA is an intriguing option for the transfer and storage of digital information [1]. DNA molecules have been used for hiding messages [2] and storing digital data [7–9]. In these studies, an encoding algorithm was used to convert digital data into nucleotide sequences that were then written, transferred, and read using DNA as the storage medium. The encoding algorithms were also programmed to reduce homopolymeric stretches as bits and trits were converted to bases. These methods allow users without prior biological or computer programming skills to encode any computer file into DNA sequences that can then be synthesized by commercial vendors. Therefore, we identified these three pivotal points of a communication—data encoding, data transfer & data extraction—to develop new methods for DNA-based communications (Fig 1a). 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 the information stored in m , DNA-specific cryptography and steganography methods may be implemented [12–15], akin to conventional digital data transfer that incorporate encryption algorithms such as AES, RSA, Twofish, and others. However, in these early days of DNA data storage and communication, we believe it to be also useful to explore different methods of data encoding, data transfer, and data extraction to find additional opportunities afforded by DNA compared to conventional magnetic and optical platforms.

Here we present a new framework for the communication of short messages in DNA that explores the fragmentation of a message across multiple distinct DNA molecules (Fig 1b). To encode a message (m) in DNA, we used an encoding key (k) to convert plaintext into bases, while at the same time reducing the occurrence of homopolymers. k was designed as a substitution cipher that assigned ‘codons’ to characters based on the frequency of occurrence of characters in English text. Additionally, decoy messages (d) were also encoded using k and incorporated within a communication. To decode the information, a combination key is required to discern the message m from the decoy information d , where subsequently k can be used to decode m . To transfer data, we established a secret-sharing system where m can be fragmented across 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, we investigated a new method that allows for the multiplexed sequencing of multiple DNA molecules with a common primer, where regions within distinct DNA molecules that have matching information can be identified from a single sequencing reaction via chromatogram patterning.

Materials and Methods

Plasmids

Constructs were cloned using standard molecular biology techniques. KOD Hot Start DNA Polymerase (VWR) was used to PCR amplify a p15A origin and a chloramphenicol resistance gene, which were then fused to gBlocks from IDT (Table 1) using Gibson assembly. Random DNA sequences were generated at http://www.bioinformatics.org/sms2/random_dna.html. Constructs were sequence verified by Genewiz Inc. (Cambridge, MA).

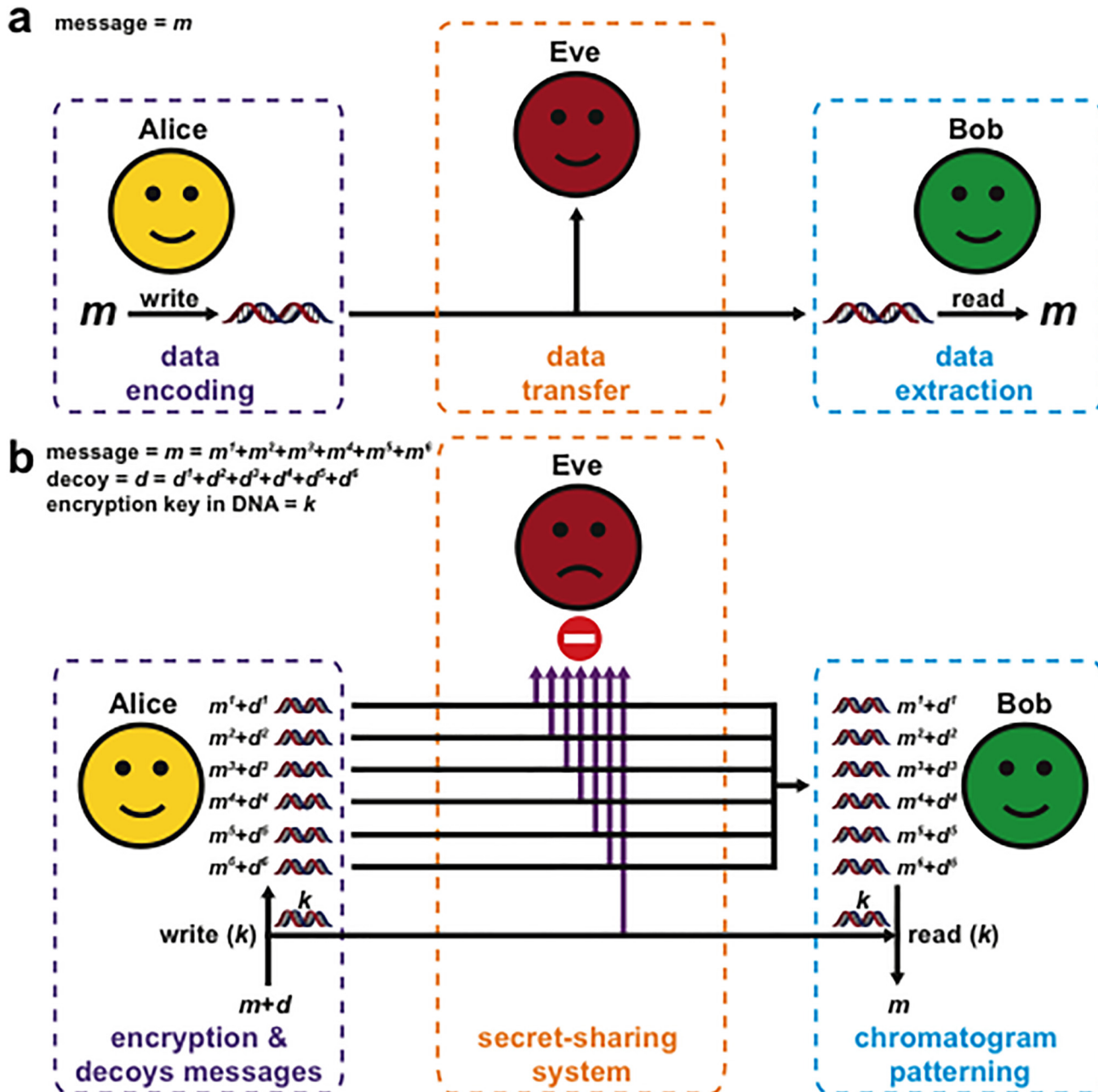


Fig 1. DNA communication. (a) For Alice to send a message (m) to Bob, she must first write the data into DNA and then physically send the DNA to Bob, who can read the DNA and extract the data. Eve, who is eavesdropping, can physically intercept and read m . Here we have identified three areas to explore within the communication channel between Alice and Bob: data encoding, data transfer, and data extraction. (b) Fragmented DNA communication. Data encoding: m can be mixed with decoy (d) data and fragmented, then written into DNA, where the key (k) is used to encode the data and 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. Data extraction: chromatogram patterning can be used by Bob to extract data via multiplexed sequencing reactions.

doi:10.1371/journal.pone.0152774.g001

Sanger Sequencing

Constructs were purified using Qiagen kits and stored in cell culture grade water (Cellgro). Constructs were diluted to 30 ng/ μ L and sent for sequencing at indicated ratios. Primer_{ExternalFw} (GACATTAACCTATAAAAAATAGGC), Primer_{ExternalRv} (GCATCTTCCAGGAAATCTC),

Table 1. Sequences of the constructs used in this study.

Construct	Plasmid	Sequence
iKey-64	pBZ38	TTTTTTTTTCGGAGCTGAGACCGAACGTAGGCTTCGGCACTGTTAGAAGATATCAACAATTCACGTATGCGCGTGGTAACTTGTCTTTTGATTCACTGCCATTCTGCGGAGCTCCCATTTCAGATCCACCTGGAGGGGAAAGATAGTTTATGTCACACAGTACTAACAAAAACCCGGTTTAGCTCTAGCGGCTCCTGCCCGTTTTTTTTT
DNA-1	pBZ27	TGGCCACGATCCATGCTAACGCTCTGCGTAGGGATGAATCCCCTTTTGAACCTGTTCTACTGACGGACGAGCTGATAGGTAGCCGAAGTAGTGATACGATCCACACATGCCATCATTGCATACTCGTGCATTCAATGATGCATAGTCACGTAGTCCATATGGTAATGGTGATGTCAAGTCACATGTCAATACTCGTACTAGAACTGAGCGCGATGACTGGCGAGCTGGTGCCTCCGAGGGCTGGTTCGAGCGACTAAGTTGAATGCGCAGACCGATCGAGACGACTCTAGCGCTGGAATAAATCAGAATAAAGA
DNA-2	pBZ28	CCCACCAACTGCAATAGACGGTACTGTACACCCTGTTTTACAGCAACGGGAAAGGAGGATCACTTTCTACAATTGTGTGCTGGACTGACAGTCGCATATCCACACATGCCATCATTGCATACTCGTGCATTCAATGATGCATCTACACGTAGTCCATATGGTAATGGTGATGTCACTACACATGTCAATACTCGTACTAGAACTGAGCGCGATACGACTCGCCATAGGGTTCGCCGGCTCGCACTGACTACCTTACGCTCTAGCCAGATCGGAGCCGCGCATGACCCCTGTGATATAATACCGTTTAC
n1	pBZ29	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGACGTAGTTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTGATCATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCATCCAGTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTAACGTCTCTTCCACCTTTCCAAAAAGTAAACCCGACTGATCGGCATACGGCAACAGTACTCTGACTACCATAGTAGTGAGATGGTGGATTACGATCGCGTGATCTGAGTATCAATTGATCTATAGTGGATTGACTGATGATCGTACTGTGCTGACTGACTCTGACGTGATCTCAGTGCATATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC
n2	pBZ30	GCTAGTTATTGCTCAGCGGCATCGTATGACGATGACTGACTTAGCAACTGTCGAGTAATATGACCTGAGAGCTACTGATCTGACTAGCTAAGCTTGCATGCACGTCATGATCCACTATAGATCAATGACTCAGATCAGCGATATCGACGTTGACTAGTCAAGCTAGATCCACATATGCTGTATGTGCGTAGTCGATGTCATGACTATGTTTTACAGCAACGGGAAAGGAGGACCGTCTATTGGCAGTATTGGTGGGATCTTGTAACTAACGTCAAGATAGGGATGATCTCTGACGCATACACGCATTAGATGCCGTCTGCATATATGGCAACAGTGGATACGACTCGATCATCGAGTTCCGATGCTAGCACTGACTACGTTACGCTCTGATCTCAGACGATAGTCAGATCGGAGTCAGCTGCATGACGACAGTGCATGCTAGCGTTGATCTCATGCATCCTACACTCATGAGACTCGTACTGACTGCTGCACTAGACTGTCCCTATAGTGAGTCGTATTA
n3	pBZ31	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCGACGATATTCGACGTAGTTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTGATCATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCATCCAGTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTAACGTCTCTTCCACCTTTCCAAAAAGTAAACACCACTGACGTATCGACTACGCACATACAGCATATGTGGATGATCACTGACTGACTGAACTACGATCATGGTGTATGTGAGCGTGTATGTGCTCGTACTGGAGAAACGGCAACAGTGGATGATTGACGTACGACTGCTAGCTCAGGTCATATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC
n4	pBZ32	GCTAGTTATTGCTCAGCGGCATCGTATGACGATGACTGACTTAGCAACTGTCGAGTAATATGACCTGAGAGTCAGTGCTCATGATGTCATCCACTGTTGCCGTTTTCCCTACACGAGCACATACACGCTCACATACACCATGATGACTAGCATGATCATCCACTGTTAGTATCAGCAGCGGCATGATCTGATGACGATCATGACTGTTTTACAGCAACGGGAAAGGAGGACCGTCTATTGGCAGTATTGGTGGGATCTTGTAACTAACGTCAAGATAGGGATGATCTCTGACGCATACACGCATTAGATGCCGTCTGCATATATGGCAACAGTGGATACGACTCGATCATCGAGTTCCGATGCTAGCACTGACTACGTTACGCTCTGATCTCGACGATAGTCAGATCGGAGTCAGCTGCATGACGACAGTGCATGCTAGCGTTGATCTCATGCATCCTACACTCATGAGACTCGTACTGACTGCTGCACTAGACTGTCCCTATAGTGAGTCGTATTA
n5	pBZ33	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCACGGATATTCGACGTAGTTCGCTCGACCTGACTCGTGAGTGCAGCTACGTGTGATCATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCATCCAGTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTAACGTCTCTTCCACCTTTCCAAAAAGTAAACACTGACTGTCATTGCTGATCATGATGCCGGCGTGATCTAGATACACGGTGGATTTCAGCTACTAGTGAATCATGACGTGAGAAGCATGAACGATATGAAGAAGTTATGTGGATAGCTGTCGACTGATCGTATCGATGCAGTCCCTCAGGTCATATTACTCGACAGTTGCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC
n6	pBZ37	GCTAGTTATTGCTCAGCGGCATCGTATGACGATGACTGACTTAGCAACTGTCGAGTAATATGACCTGAGAGCTATCGATGACGACTGATGTCATCATGATCCACATAAATTCTTCATATCGTTTCATGCTTCTCACGTCATGATAACGCATCCACCATCTCACTACTATGGTAGTCGAGTACACTGTTGCCGATGCGCGATGTCATGTTTACAGCAACGGGAAAGGAGGACCGTCTATTGGCAGTATTGGTGGGATCTTGTAACTAACGTCAAGATAGGGATGATCTCTGACGCATACACGCATTAGATGCCGTCTGCATATATGGCAACAGTGGATACGACTCGATCATCGAGTTCCGATGCTAGCACTGACTACGTTACGCTCTGATCTCGACGATAGTCAGATCGGAGTCAGCTGCATGACGACAGTGCATGCTAGCGTTGATCTCATGCATCCTACACTCATGAGACTCGTACTGACTGCTGCACTAGACTGTCCCTATAGTGAGTCGTATTA

doi:10.1371/journal.pone.0152774.t001

Primer_{Key} (TAATACGACTCACTATAGGG), and Primer_{Message} (GCTAGTTATTGCTCAGCGG) were used for Sanger sequencing reactions performed in triplicate at Genewiz Inc. under 'Difficult Template' settings. Genewiz Inc. was not consulted prior, during, or after this study and all Sanger sequencing reactions were performed under blind conditions to ensure bias was not introduced in the results.

Next-Generation Sequencing

An outside party (MIT BioMicro Center, Cambridge, MA) performed next-generation sequencing (NGS) sequencing and analysis on a mixture of $n_1+n_2+n_3+n_4+n_5+n_6$. Plasmids were purified using Qiagen kits and stored in cell culture grade water (Cellgro). To confirm purity, plasmids (300ng) were run on a 1% agarose gel. Plasmids were then mixed at equal concentrations of 30 ng/ μ L and 900 ng of the mixture was submitted to the MIT BioMicro Center. Blind experimental conditions were used throughout the sequencing and annotation process.

Briefly, for NGS sequencing a Nextera kit (Epicentre) followed by 1.5% agarose BluePippin (Sage Science) isolation of 450–800 bp inserts was used to generate a library. A MiSeq (Illumina) run on a 600 nt v3 kit was used for pair-end sequencing. Sequence assemblies were then performed using various programs including: SOAP Denovo, Trinity, Mira, Velvet, and RAST annotation.

Results and Discussion

To date, several elegant methods have been proposed for encoding digital information in DNA, each taking a unique approach to convert digital data into bases while at the same time reducing the occurrence of homopolymeric stretches [16]. However, within these early days of the field different encoding methods need to be investigated and the pros and cons of different approaches evaluated until the field converges on a single standardized and DNA-centric encoding method.

To convert plaintext to bases for DNA encoding, we took inspiration from written text. We combined the familiarity of text-based communication—the QWERTY keyboard—and the genetic code to develop individualized keyboards (iKeys) that serve as a facile method 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 $4^3 = 64$ distinct codons. Accordingly, codons are units of nucleotides that encode information that is then translated into function. Here, we abstract the concept of a codon to encode information by mapping the 64 distinct codons onto a modified QWERTY keyboard to produce a personalized code—iKey-64—for translating text into DNA (Fig 2a). This serves as an encoding key (k) for converting a message (m) into a DNA encodable language (Fig 1b), akin to a substitution cipher. Furthermore, any specific version of iKey-64 can itself be encoded in DNA and provided as an additional component of a communication, serving as a unique dictionary for each message (Fig 2b and 2c).

It is known that stretches of homopolymers in DNA often lead to sequencing inaccuracies [16]. To mitigate this problem, we rationally designed iKey-64 to reduce the incidence of homopolymers in DNA messages by basing codon assignment on the frequency of use of letters in the English language [17] (Table 2 and Fig 3). Higher frequency characters were designated to codons containing 3 different nucleotides, lower frequency characters to codons with the same nucleotide in the first and last position, and the least frequent characters were assigned to codons with 2 or more homopolymeric stretches. Here we use English as an example, but a similar approach can be used for other languages. Since the codons AAA, CCC, GGG, and TTT are assigned to function keys—that can encode any user-defined function—no

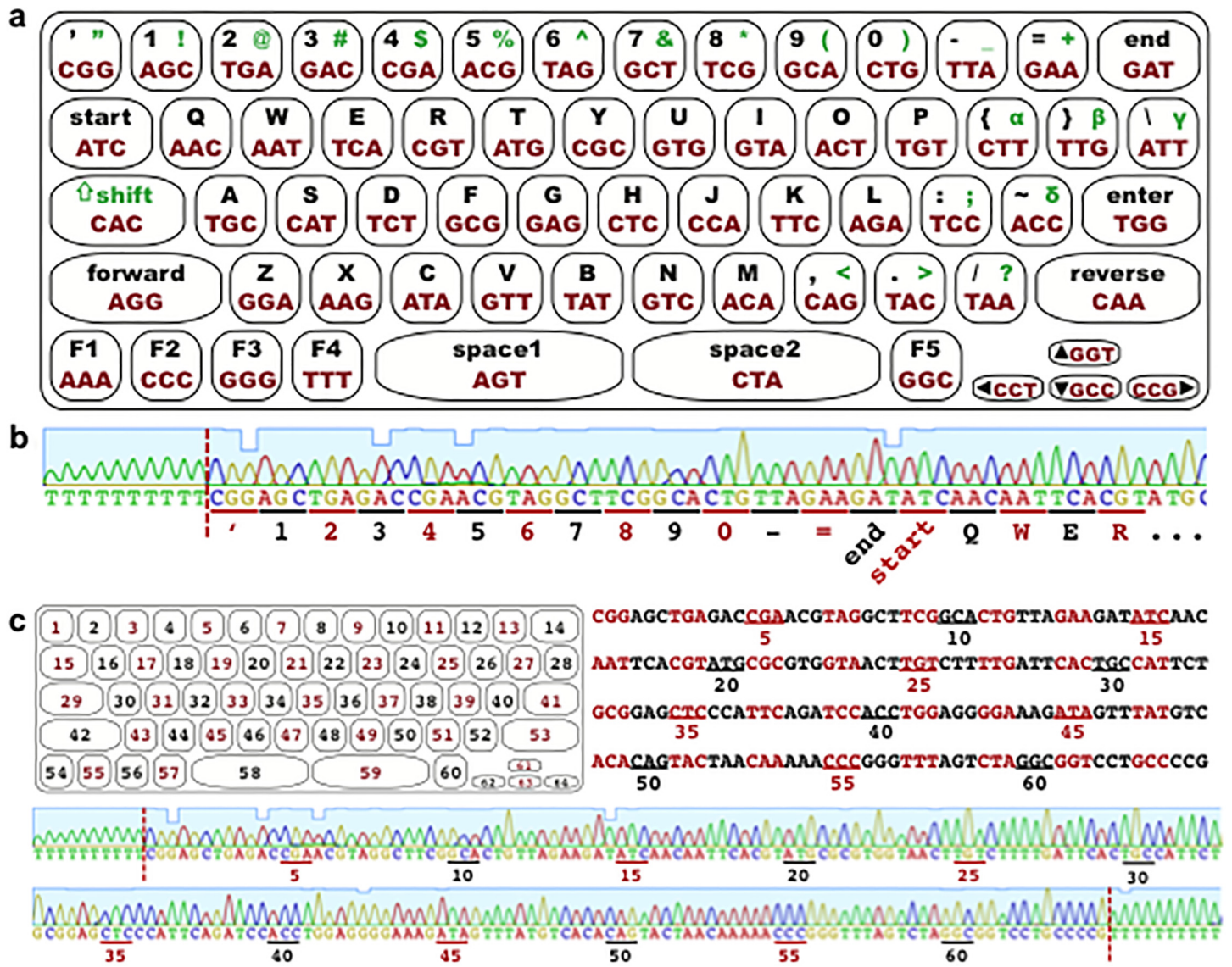


Fig 2. 64 button iKey for chromatogram patterning. (a) iKey-64, used to convert plaintext to codons for writing information in DNA molecules. Messages begin with 'start', finish with 'end', 'forward' and 'reverse' provide information on the strand containing the desired message, and 'space1' and 'space2' may be used to produce troughs in chromatograms. The 'shift' codon precedes capitalized letters or upper characters. Codons can be randomized to produce up to $24!12!28! = 9.1 \times 10^{61}$ iKey-64 variants. (b) The iKey-64 variant from (a) written in synthetic DNA and read by Sanger sequencing. Shown is the first row and part of the second row of the iKey keyboard (flanked by 10 T nucleotides). (c) Top: iKey-64 buttons and codons were numbered to write the keyboard onto a strand of DNA. Bottom: iKey-64 written in DNA. Codons were flanked by 10 Ts to separate the start and end of the keyboard from surrounding DNA for identification, marked by red lines.

doi:10.1371/journal.pone.0152774.g002

homopolymeric stretches longer than 4 bases are possible when encoding regular English text (Fig 2a). For example, the letter VK would be encoded with bases GTTTTC, where the maximum homopolymeric stretch of 4 Ts would be reached. Additionally, since all numerals (0–9) were assigned to codons containing 3 different nucleotides, no homopolymeric stretches longer than 2 bases are possible when encoding numbers, including instances where digital data stored in bits, trits, etc. is converted to bases (Fig 2a). For example, the numerals 110011 would be encoded with bases AGCAGCCTGCTGAGCAGC, where the maximum homopolymeric stretch of 2 Cs would be reached. In the event where multiple consecutive function keys are used,

Table 2. The frequency of letters used in English based on the Concise Oxford Dictionary and adapted from [17].

Letter	Frequency	Letter	Frequency
E	11.16%	M	3.01%
A	8.50%	H	3.00%
R	7.58%	G	2.47%
I	7.54%	B	2.07%
O	7.16%	F	1.81%
T	6.95%	Y	1.78%
N	6.65%	W	1.29%
S	5.74%	K	1.10%
L	5.49%	V	1.01%
C	4.54%	X	0.29%
U	3.63%	Z	0.27%
D	3.38%	J	0.20%
P	3.17%	Q	0.20%

doi:10.1371/journal.pone.0152774.t002

spaces can be used to reduce homopolymeric stretches. In subsequent experiments we investigate new methods for fragmented DNA communication. Therefore, we encode text as an example since the contents of the communication are not our focus, but similar approaches should be applicable for other data formats.

To investigate this DNA platform for information transfer, we sought to disseminate texts across multiple different DNA strands so that the desired message would be revealed only if the correct strand combinations were analyzed. A single communication channel between Alice and Bob can be intercepted by Eve at a single point of contact, thereby compromising the message m (Fig 1a). However, a fragmented communication channel would require multiple points of contact for interception by Eve (Fig 1b). This approach can add an additional layer of protection for a communication and also provide opportunities to explore introducing tiers of complexity within a communication that is afforded by the unique makeup of DNA as a chemical polymer for information storage. Therefore, we created a fragmented communication platform that we call Multiplexed Sequences Encoding (MuSE), a secret-sharing system [18] that allows for communication of a message m across multiple distinct DNA molecules.

To extract information that is fragmented by MuSE across multiple distinct DNA molecules, one would have to sequence the DNA molecules individually then compare the sequences to look for regions of sequence identity to locate encoded messages. However, the distinct nature of DNA as a data storage medium provided us with an opportunity to explore alternative methods of data extraction. Accordingly, we sought to develop a platform that allows for multiple distinct DNA molecules to be sequenced within a single reaction, whereby the encoded data shared among DNA molecules could be easily located via patterns formed in sequencing chromatograms.

In designing MuSE, we 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 homogeneous peak would be observed, and where two bases differ a small heterogeneous peak would be observed, thereby producing a pattern (Fig 4a). Not surprisingly, the naïve sequencing of multiple DNA strands with a common primer is unable to achieve chromatogram patterning, and instead it produces poor readouts (Fig 5). However, the codons in iKey-64 were rationally assigned to characters based on the frequency of use of individual

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 TGG 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}

Fig 3. Rational design of iKey-64 for encoding information into DNA, while reducing the incidence of homopolymers and achieving chromatogram patterning. The buttons of iKey-64 were separated into 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.

doi:10.1371/journal.pone.0152774.g003

characters, thereby serving to reduce the incidence of homopolymers in DNA messages that reduce the accuracy of sequencing reactions. Therefore, we expected the design of iKey-64 to mitigate the problem of base calls moving out of phase when multiple DNA molecules were sequenced simultaneously with a common primer as observed in Fig 5.

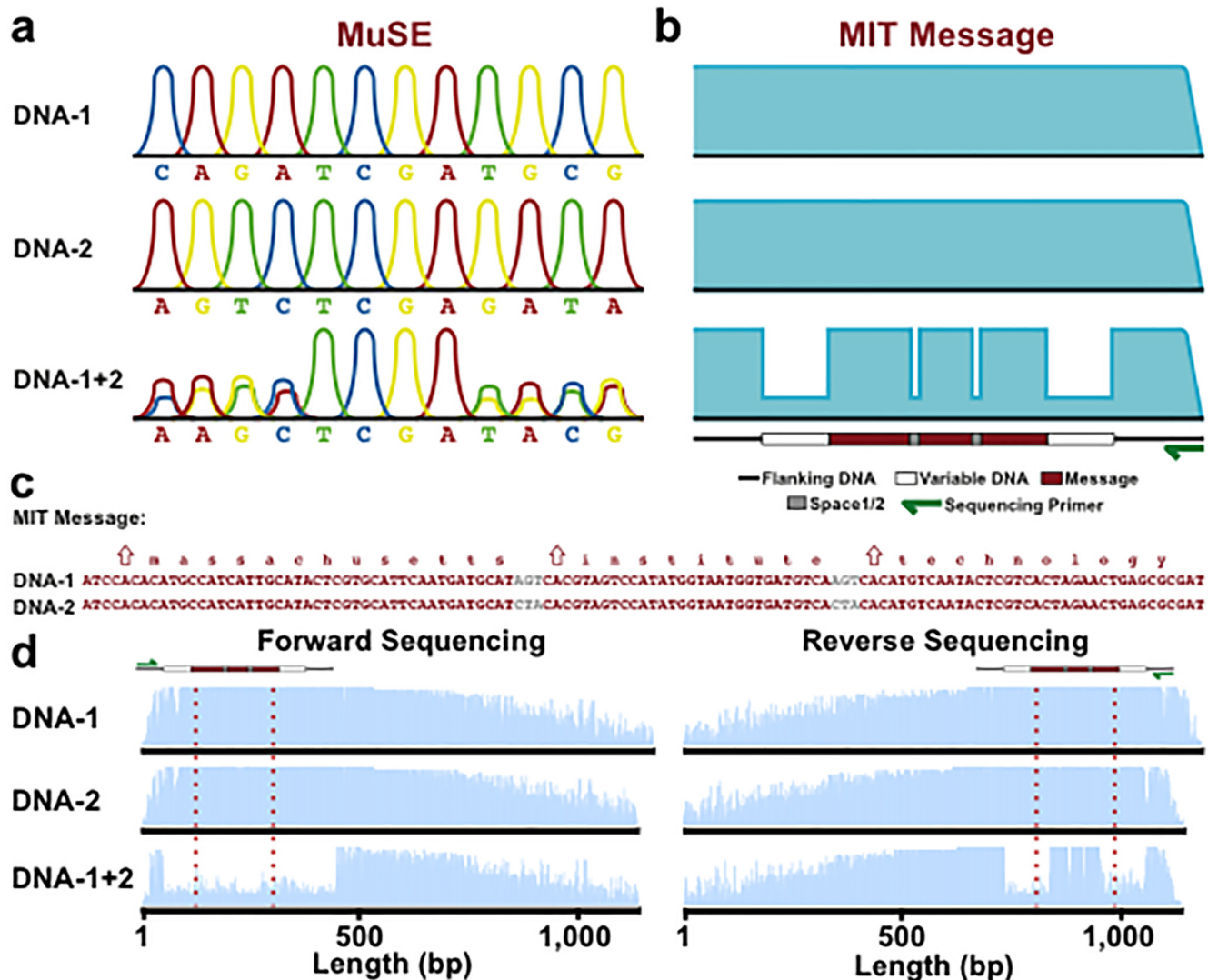


Fig 4. Chromatogram patterning with MuSE. (a) Schematic for chromatogram patterning. When two DNA strands are co-sequenced with a common primer via Sanger sequencing, different overlapping nucleotides produce small heterogeneous peaks while matching nucleotides produce large homogeneous peaks. Peaks are kept in alignment via iKey-64. (b) Schematic of chromatogram patterning for the message ‘Massachusetts Institute Technology’ via MuSE. (c) Sequences for ‘Massachusetts Institute Technology’ used in (b) and encoded with iKey-64. (d) Chromatograms observed from Sanger sequencing of the DNA-encoded message described in (b) and (c). When DNA-1 and DNA-2 are co-sequenced at equal concentrations with a common primer (green arrows), chromatogram patterning is achieved during reverse (Primer_{ExternalRv}) but not forward (Primer_{ExternalFw}) sequencing due to the flanking variable DNA regions. Red lines surround embedded messages.

doi:10.1371/journal.pone.0152774.g004

To test whether chromatogram patterning could be achieved with MuSE, we used iKey-64 to encode the message ‘Massachusetts Institute Technology’ on two DNA strands, where space1 (AGT) was used with the first DNA strand (DNA-1) and space2 (CTA) with the second DNA strand (DNA-2) to demarcate individual words in the sequences (Fig 4b and 4c). In this design, co-sequencing both DNA strands together should introduce troughs around words in the resulting chromatogram, thereby providing a simple method to locate the message from a single sequencing reaction. As expected, individual sequencing of DNA-1 and DNA-2 produced high quality reads, but gave no indication of the presence or location of a message (Fig 4d). However, in a DNA-1+2 mixture, forward sequencing with a common primer did not reveal a message through chromatogram patterning, but rather camouflaged the message (Fig 4d). This was due

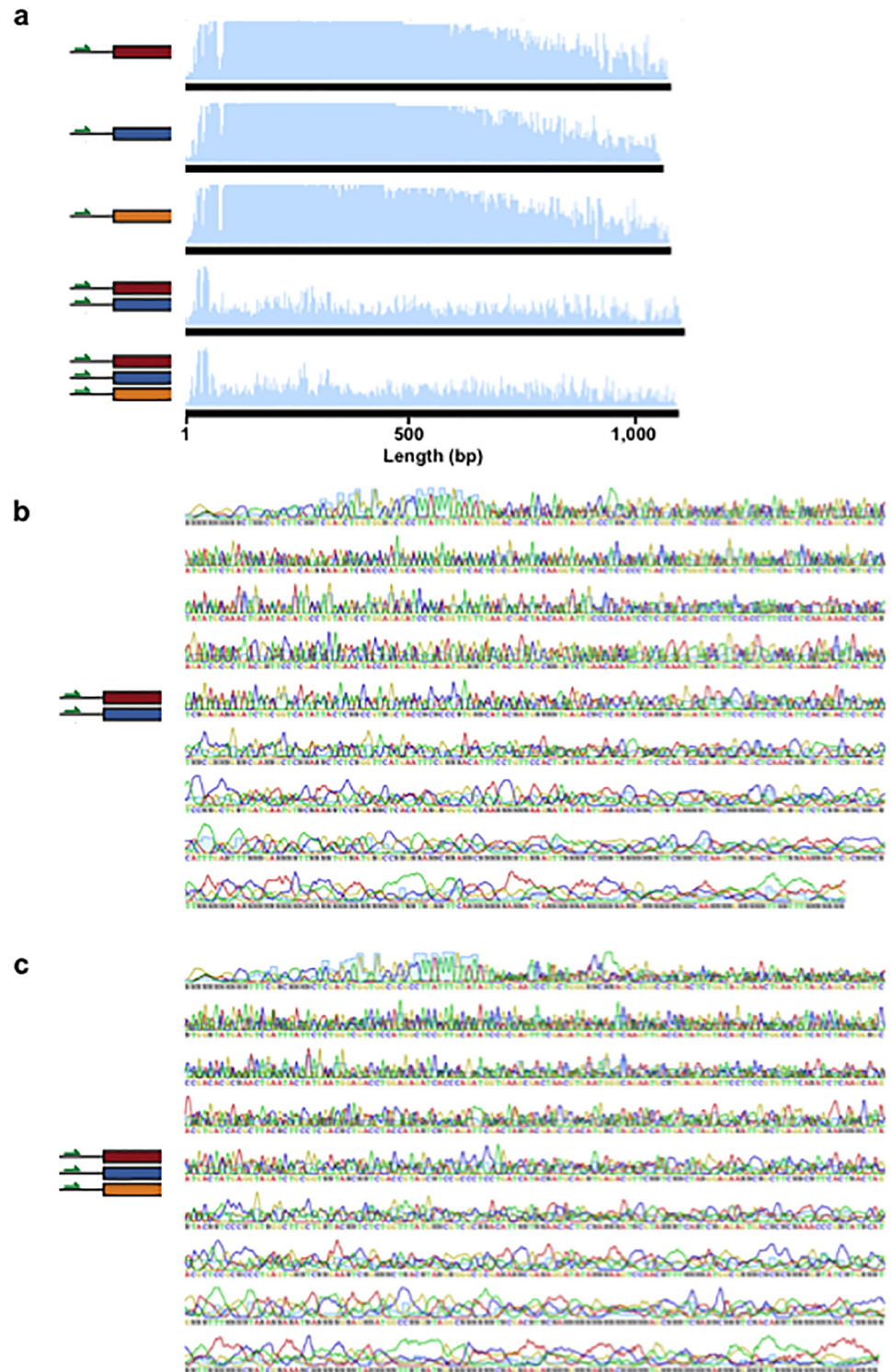


Fig 5. Naïve co-sequencing of multiple DNA strands. (a) Red (DNA-1), blue (n1), and orange (iKey-64) strands have different sequences but they all share a common upstream region and sequencing primer ($\text{Primer}_{\text{ExternalFw}}$). Individual sequencing of each strand produces high quality reads, but the resulting reads are of poor quality when two (red and blue) or three (red, blue, and orange) strands are co-sequenced. (b) Close-up of the chromatogram of red and blue co-sequencing. (c) Close-up of the chromatogram of red, blue, and orange co-sequencing.

doi:10.1371/journal.pone.0152774.g005

to variable DNA sequences placed upstream of the messages, where stretches of C and A homopolymers at the 5' ends interfered with base determination during Sanger sequencing, thus causing intentional misalignment of the recognized bases in the chromatogram (Fig 6a and 6b). Only reverse sequencing of DNA-1+2 with a common primer produced a distinct pattern in the chromatogram, readily identifying the location of the message to be decoded with iKey-64 (Fig 4d). Since there were no interfering stretches of homopolymers in the variable DNA regions, there were no shifts in the base calls during sequencing, thus leading to predictable chromatogram patterning from a multiplexed sequencing reaction (Fig 6c and 6d). Therefore, as a proof-of-concept we demonstrated that information from multiple DNA molecules can be extracted in a single reaction.

While individual sequencing of each strand followed by sequence alignments can be used to extract information from multiple DNA molecules, chromatogram patterning provides opportunities to explore new methods for data extraction and for incorporating information in DNA mixtures. To illustrate, the degree of contrast achieved in the chromatogram patterns can be tuned in a MuSE communication by adjusting the ratio of DNA-1/DNA-2 (Figs 7 and 8). This serves as a method to embed information in chromatograms discreetly so that alignments of DNA sequencing data to known templates cannot be used to identify embedded information (Fig 9). Such an approach provides new opportunities for exploring ways to store information in DNA, where data extraction is dependent on multiplexed DNA sequencing.

Next we wanted to determine whether we could use the MuSE method to disseminate information encoded with iKey-64 across more than 2 DNA molecules. This would enable us to introduce more complexity into a fragmented communication channel (Fig 1b). To demonstrate, we sought to fragment a communication that contained an intended message and a decoy message across 6 distinct DNA molecules. Such a communication would include three components (Fig 10): (1) secret-sharing system: the intended message and the decoy message along with instructions on how to differentiate between the two would be disseminated across 6 DNA molecules, (2) encoding key: the information would be converted from plaintext into bases using iKey-64, and (3) combination key: a puzzle would enable the end-user to identify the strand combinations that need to be analyzed in order to extract the desired message.

Accordingly, iKey-64 was used to encode watermarks, a combination key, a desired message, and a decoy message within 525 bp regions across six synthetically produced DNA strands, recreating a World War II communication made during the establishment of Bletchley Park [19] (Figs 10 and 11a), a significant point in cryptography history. The functions of the elements are: (i) watermarks—an identification tag for each DNA strand that allows the end-user to categorize each strand according to the combination key, (ii) combination key—a riddle whose solution provides the correct combinations of DNA strands required to analyze in order to unlock the desired message, (iii) message—the desired information to be communicated, and (iv) decoy—a false message to be revealed if improper strand combinations are analyzed, for example as a result of an incorrect solution to the combination key.

A workflow of the process for the WWII communication encoded with iKey-64 is shown in Fig 11a to demonstrate how an end-user would extract information from our DNA communication. The first step would be to pool a partial sample of the available 6 DNA molecules ($n_1 + n_2 + n_3 + n_4 + n_5 + n_6$) obtained from the fragmented DNA communication within the secret-sharing system. The next step is to identify the combination key in order to know which strand combinations need to be analyzed to reveal the desired message. Co-sequencing of the pooled DNA molecules with Primer_{Key}, which is common to all 6 DNA molecules, followed by decoding with iKey-64 should reveal the information: “Pascal’s triangle: d2r6-reverse” (Fig 10). Here a simple combination key was chosen to demonstrate the concept, and this riddle means that the desired message is revealed from sequencing DNA pairs in the reverse direction as ordered

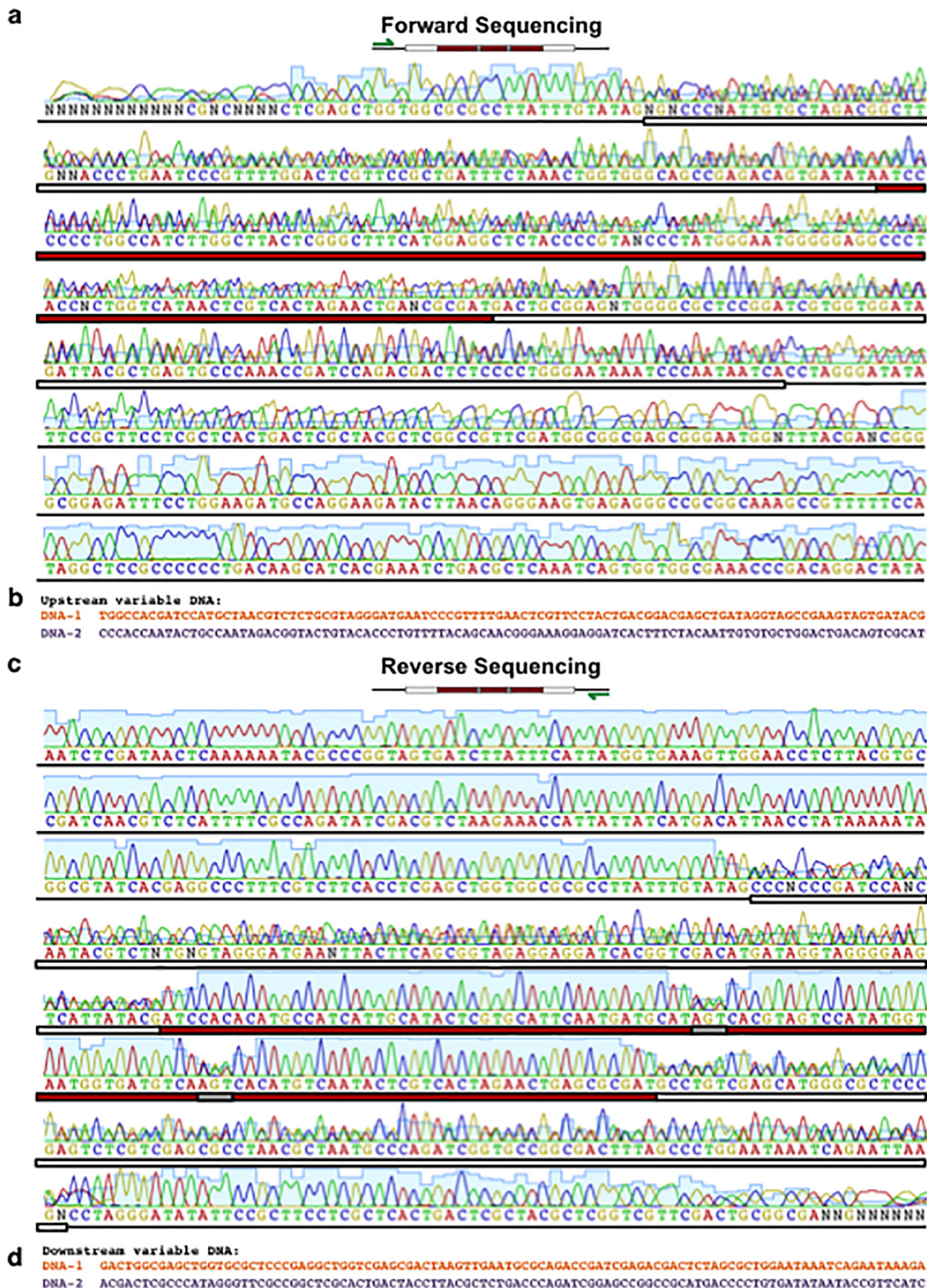


Fig 6. Chromatogram patterning requires the alignment of base calls to be maintained during co-sequencing of DNA strands. (a) Close-up of the chromatogram for forward co-sequencing of DNA-1+2 encoding the MIT message (red box) from Fig 4d. (b) Sequence of the upstream variable DNA regions (Fig 4b), corresponding to the upstream flanking region of the MIT message. (c) Close-up of the chromatogram for reverse co-sequencing of DNA-1+2 encoding the MIT message (red box) from Fig 4d. (d) Sequence of the downstream variable DNA regions (Fig 4b), corresponding to the downstream flanking region of the MIT message. Samples were co-sequenced at equal concentrations and the green arrows depict the sequencing primers (Primer_{ExternalFw} and Primer_{ExternalRv}).

doi:10.1371/journal.pone.0152774.g006

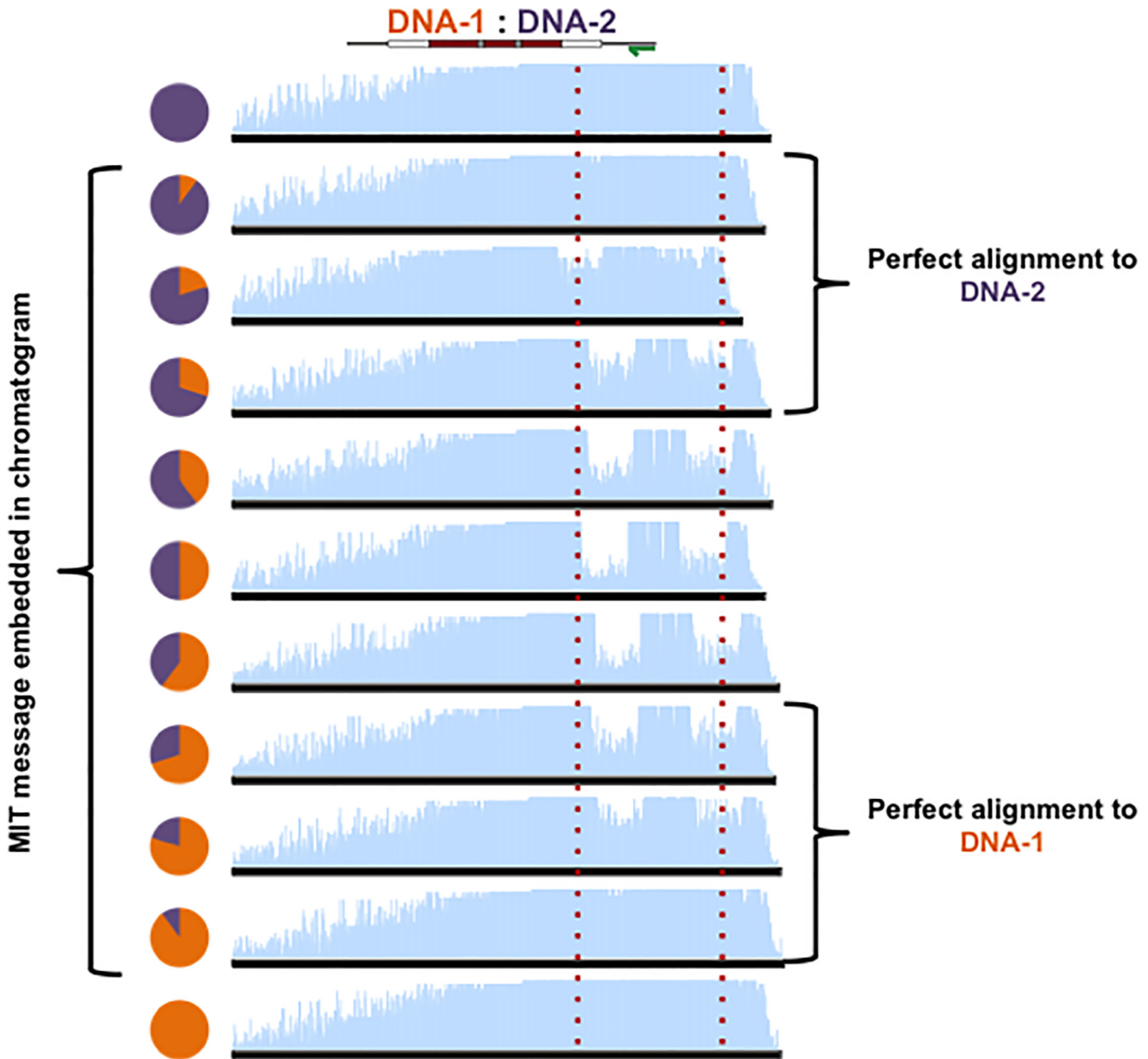


Fig 7. MuSE can be tuned to hide information in DNA communications. Chromatogram patterning can be tuned to discreetly embed information in sequencing data by varying the ratios of DNA-1 (orange) and DNA-2 (purple). Red lines surround embedded messages.

doi:10.1371/journal.pone.0152774.g007

in Pascal’s triangle from diagonal 2 down until row 6. Next, the desired message can be extracted by co-sequencing the correct DNA pairs using the sequencing primer $\text{Primer}_{\text{Message}}$, which is common to all 6 DNA molecules. Thus, if strand pairs n_1+n_2 , n_3+n_4 , and n_5+n_6 were to be co-sequenced using $\text{Primer}_{\text{Message}}$ and decoded with iKey-64, then the embedded message “Bletchley Park: GC&CS Codebreakers” would be revealed. If, for example, one were to misinterpret the key, then a decoy message would be revealed—“Captain Ridley’s Shooting Party”—as a result of co-sequencing DNA pairs n_2+n_3 , n_4+n_5 , and n_6+n_1 , a circular permutation of the correct combination key.

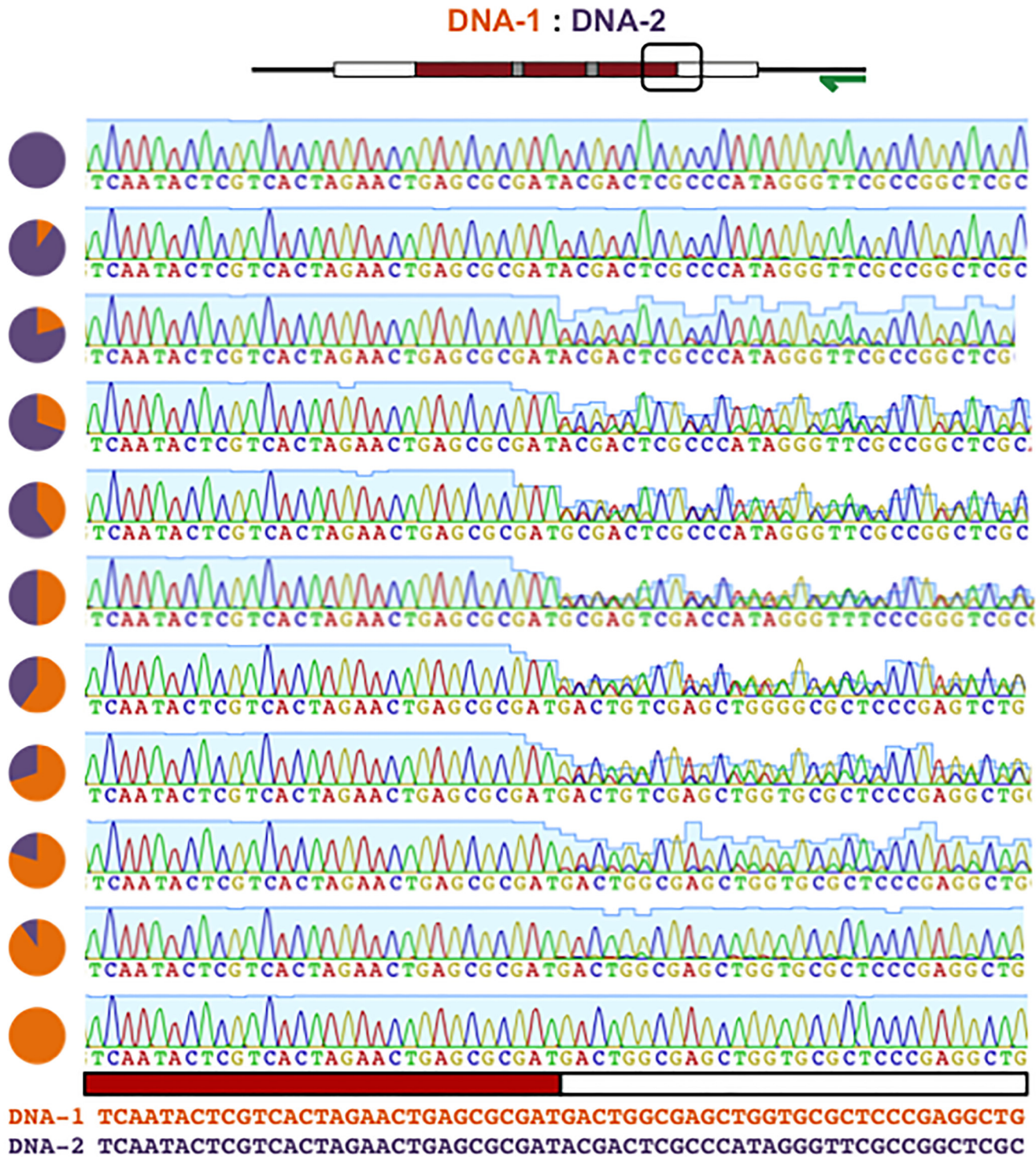


Fig 8. Discreet embedding of information in chromatograms. A close-up of chromatogram patterns formed with MuSE tuning from Fig 7. Message encoding regions (red box) contain single peaks while variable DNA regions (white box) contain two overlapping peaks whose heights can be adjusted by varying the ratios of DNA-1:DNA-2. The chromatogram close-ups correspond to the boxed region in the MIT message schematic shown above.

doi:10.1371/journal.pone.0152774.g008

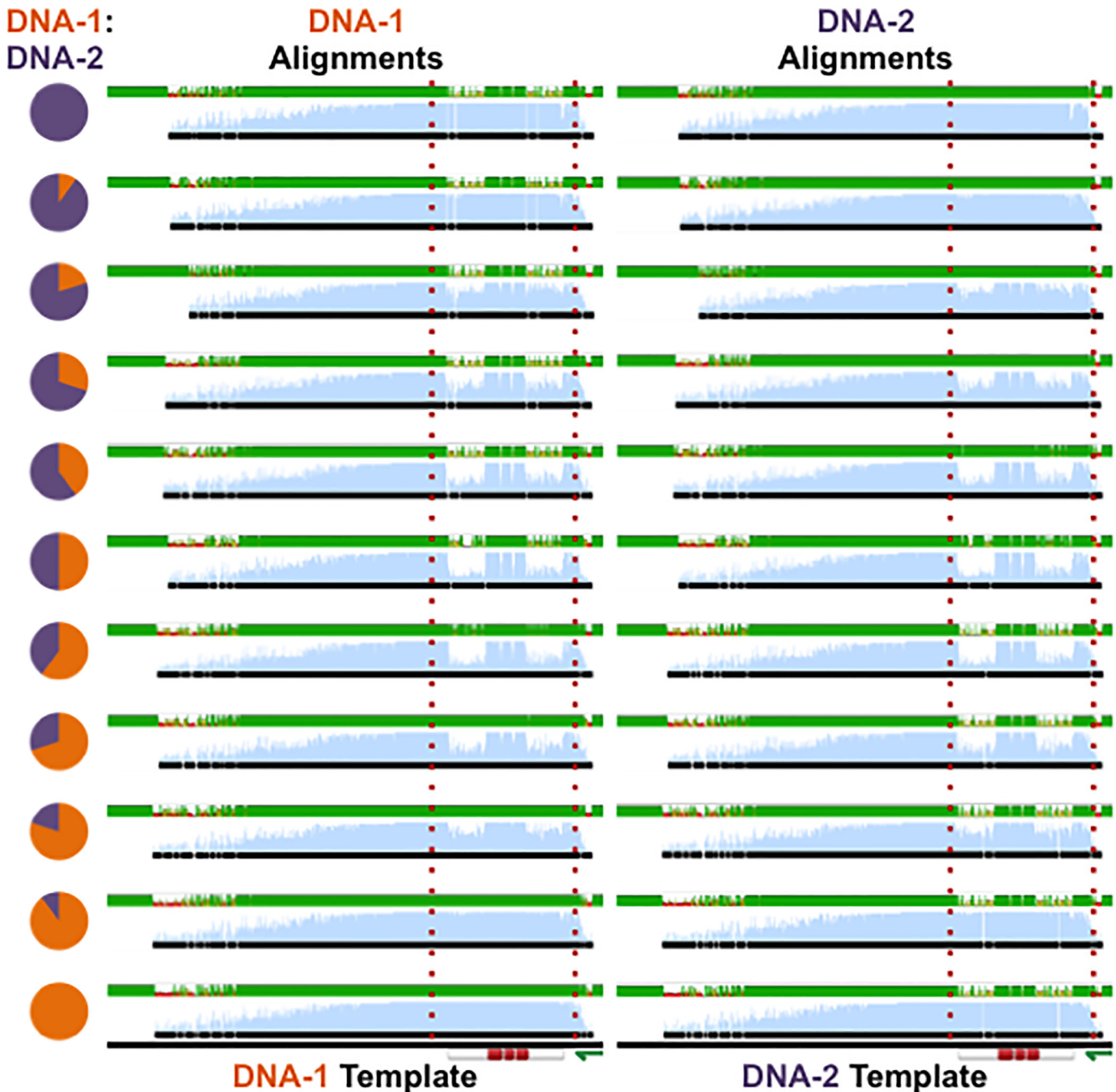


Fig 9. Discretely embedded messages cannot be identified by sequence alignments. By varying the ratios of DNA-1 (orange) and DNA-2 (purple), the degree of chromatogram patterning can be tuned (Fig 7). When one partner is present at a lower concentration, chromatogram patterning is still achieved; however, the resulting chromatogram aligns perfectly with the more concentrated partner. Therefore, messages may be discretely encoded between two DNA strands and revealed in chromatograms, but not identified by sequence alignments. Left: alignment of chromatograms from Fig 7 with DNA-1. Right: alignment of chromatograms from Fig 7 with DNA-2. Red lines surround embedded messages.

doi:10.1371/journal.pone.0152774.g009

In the event that the end-user does not have access to $\text{Primer}_{\text{Key}}$ and $\text{Primer}_{\text{Message}}$ —an unauthorized user such as Eve (Fig 1b)—then random sequencing primers may be used. For example, the sequencing primers $\text{Primer}_{\text{ExternalFw}}$ or $\text{Primer}_{\text{ExternalRv}}$ (Fig 12) may be used instead of $\text{Primer}_{\text{Key}}$ and $\text{Primer}_{\text{Message}}$ to extract messages embedded in DNA fragments. As a way to

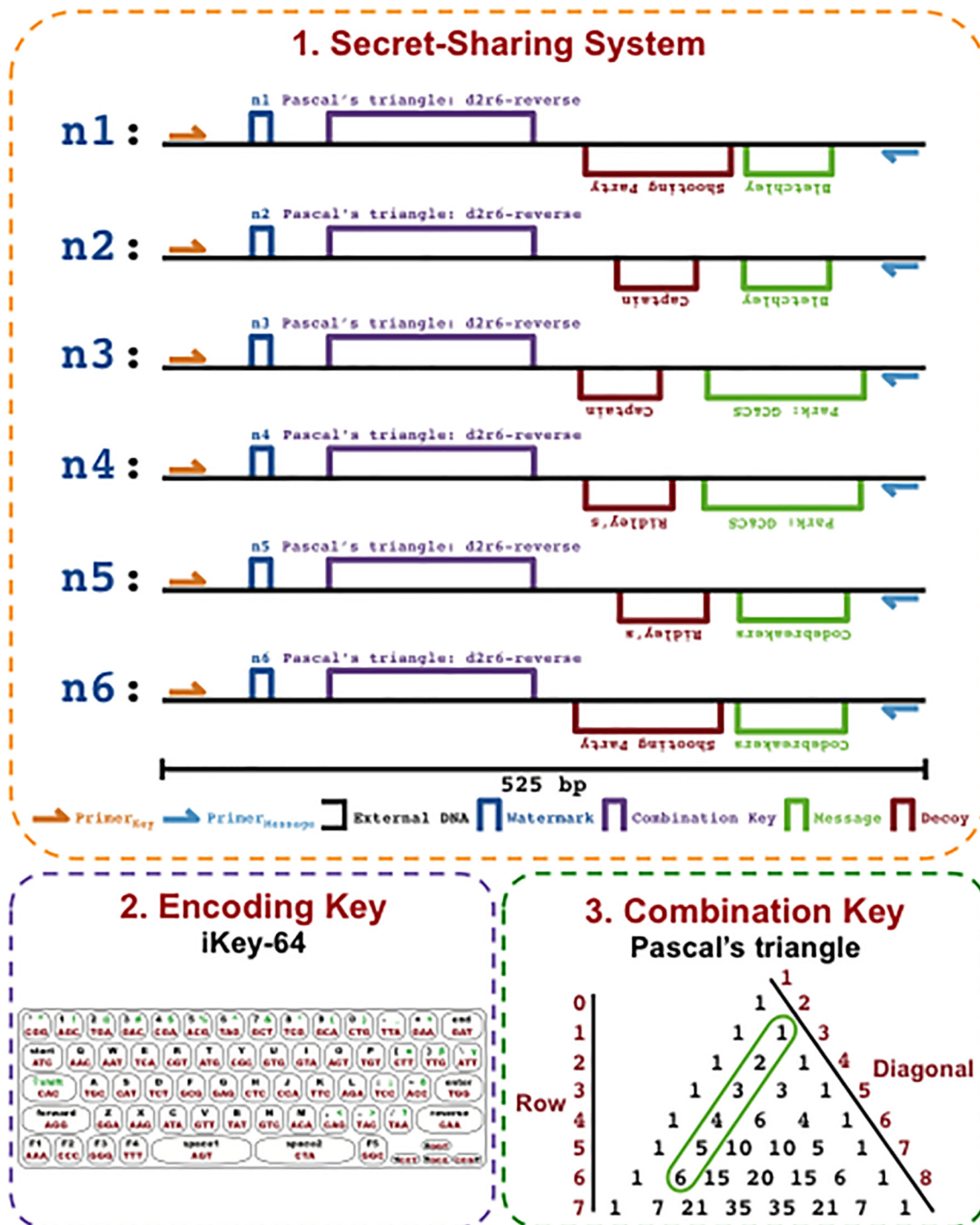


Fig 10. A fragmented WWII communication. 1. Secret-Sharing System: a recreated WWII communication was encoded across six DNA molecules and included watermarks, a combination key, a desired message, and a decoy message. 2. Encoding Key: iKey-64 was used to encode the information included in the WWII communication. 3. Combination Key: identifies which strands contain the desired message, here if strands are sequenced according to the Combination Key—obtained from Pascal's triangle—with the appropriate primers, then the desired message is revealed.

doi:10.1371/journal.pone.0152774.g010

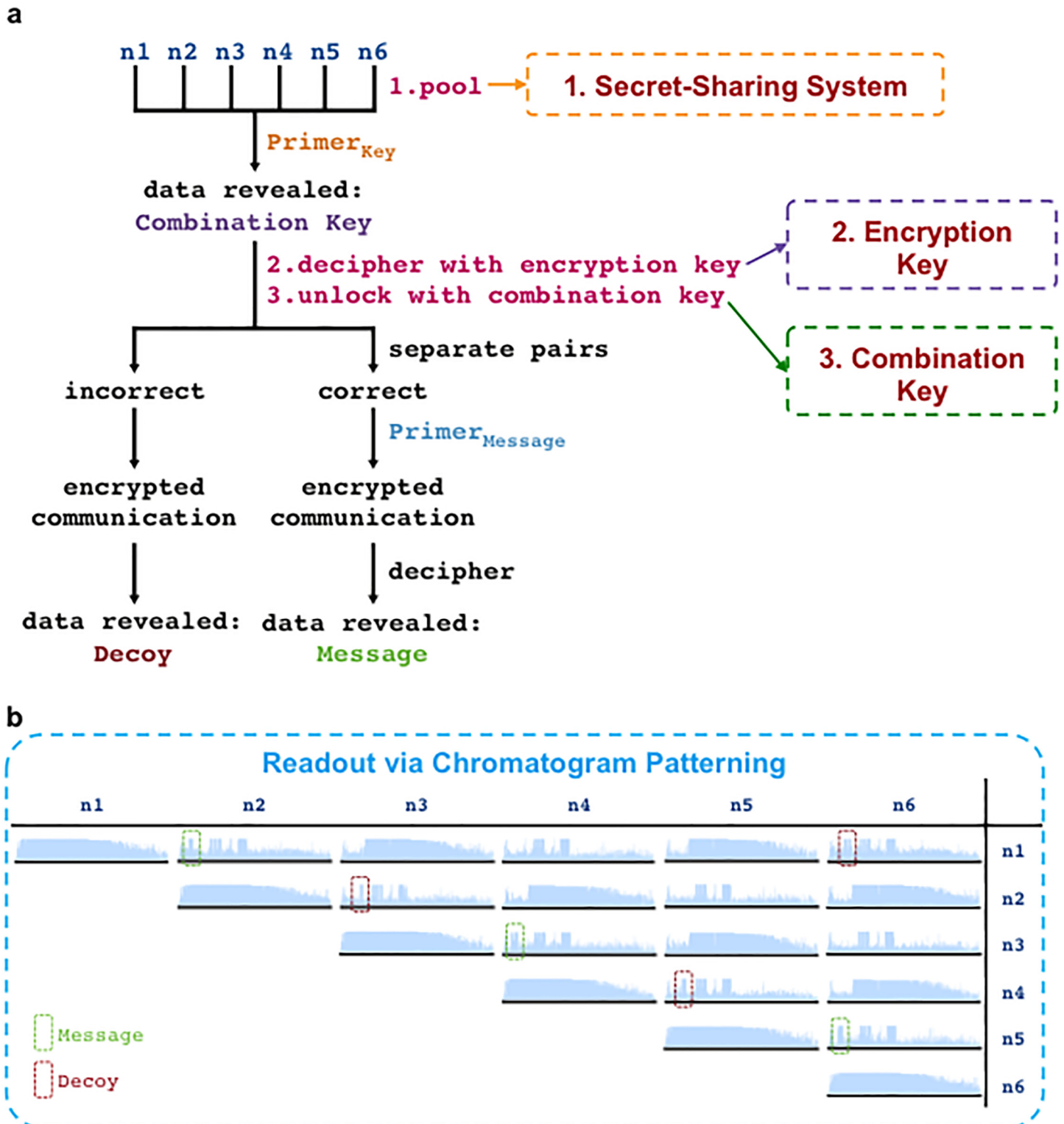


Fig 11. A WWII communication recreated in DNA. (a) Workflow for extracting the desired message from the WWII communication. Workflow steps 1, 2, and 3 are highlighted in pink and can be viewed in detail in Fig 10. Data containing strands are pooled and sequenced with Primer_{Key} to reveal the combination key. Decoding and solving the combination key will reveal the correct strand pairs to analyze with Primer_{Message} to reveal the desired message. Analysis of incorrect strand pairs will reveal a decoy message. (b) Chromatograms of an n1 x n6 matrix of DNA strands from the WWII communication (Fig 10) tuned and co-sequenced with Primer_{Message}. Boxes highlight patterns that communicate either the desired message (green) or the decoy message (red).

doi:10.1371/journal.pone.0152774.g011

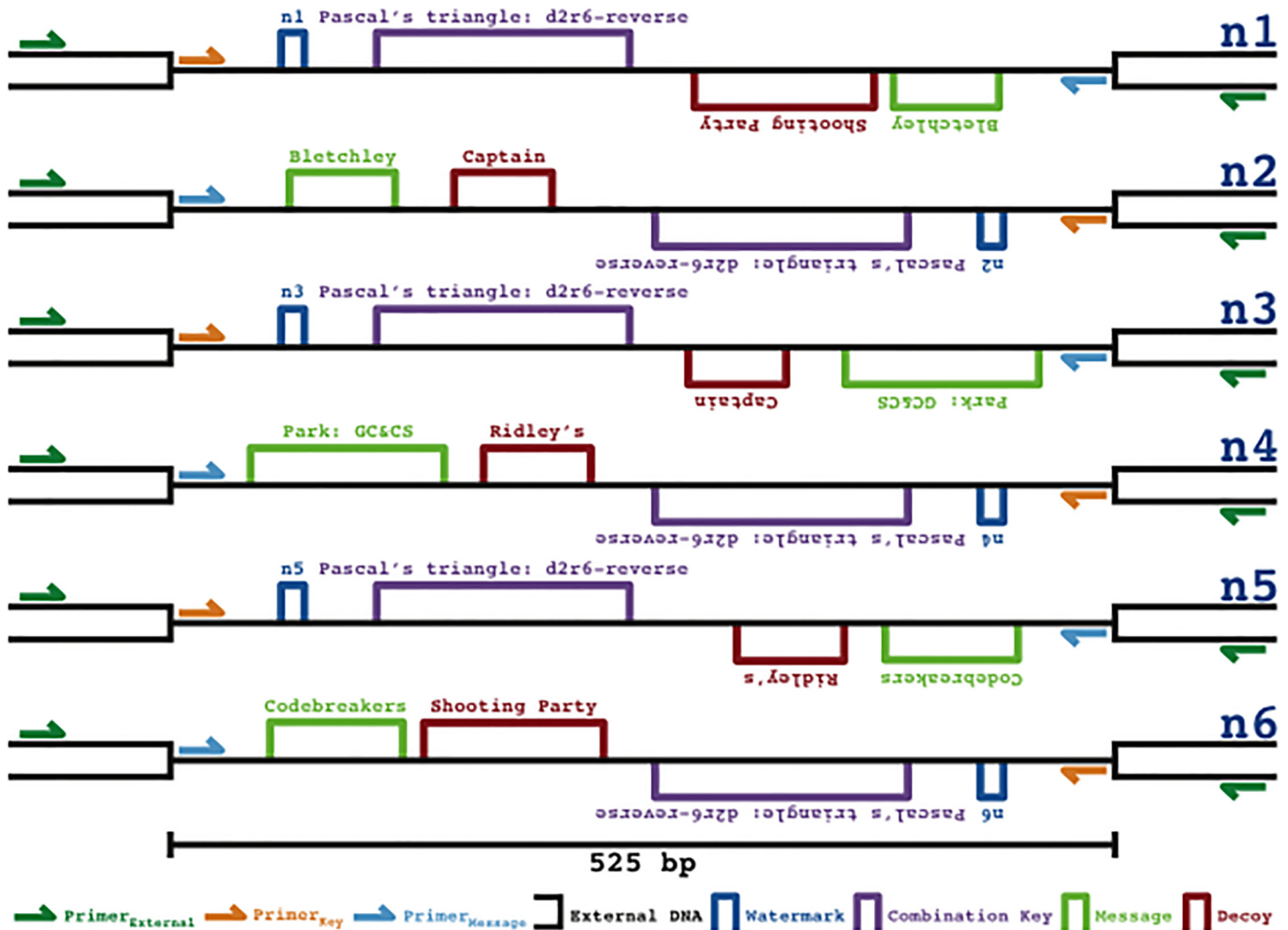


Fig 12. DNA camouflage. The 525 bp information-encoding regions of the WWII communication were flipped between the forward and reverse strands to provide a camouflage effect against sequencing with random primers ($Primer_{ExternalFw}$ and $Primer_{ExternalRv}$). While the external DNA regions surrounding the information containing regions were identical, strands n1/n3/n5 were placed in the forward direction and strands n2/n4/n6 in the reverse direction, with watermarks used to determine the orientation.

doi:10.1371/journal.pone.0152774.g012

obfuscate random sequencing attempts of pooled DNA samples, we flipped the information-containing regions of our WWII communication between the forward and reverse strands. We hypothesized that this would create a camouflage effect, where co-sequencing reactions would not produce chromatogram patterning and instead produce sequencing reads of poor quality that did not provide reliable sequence information (Fig 12). As intended, co-sequencing with $Primer_{ExternalFw}$ or $Primer_{ExternalRv}$ did not produce chromatogram patterning, regardless of whether message or decoy pairs (Fig 13), or all six strands were co-sequenced (Fig 14).

On the other hand, if the appropriate sequencing primers are used as per the data extraction workflow (Fig 11a), then the information from the fragmented DNA communication can be efficiently extracted. To demonstrate, when $Primer_{Key}$ is used to co-sequence a pooled sample of all six DNA molecules from the WWII communication (Fig 10), then the combination key “Pascal’s triangle: d2r6-reverse” is revealed via chromatogram patterning while the other data encoding regions (watermark, message, and decoy data) do not lead to chromatogram

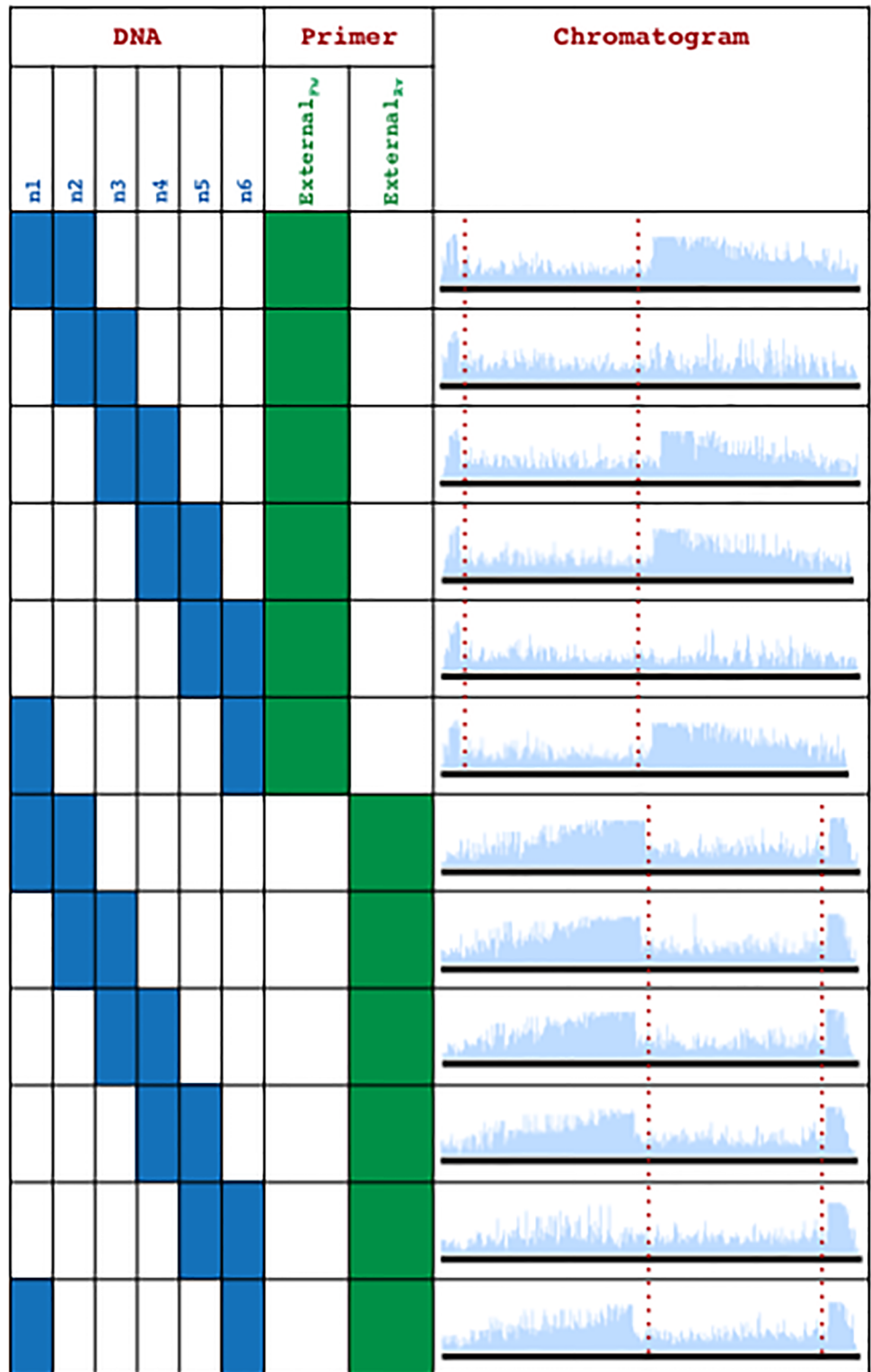


Fig 13. Data extraction from the WWII communication using Primer_{ExternalFw} and Primer_{ExternalRv} produces poor quality sequencing reads (message encoding regions are between the red lines).

doi:10.1371/journal.pone.0152774.g013

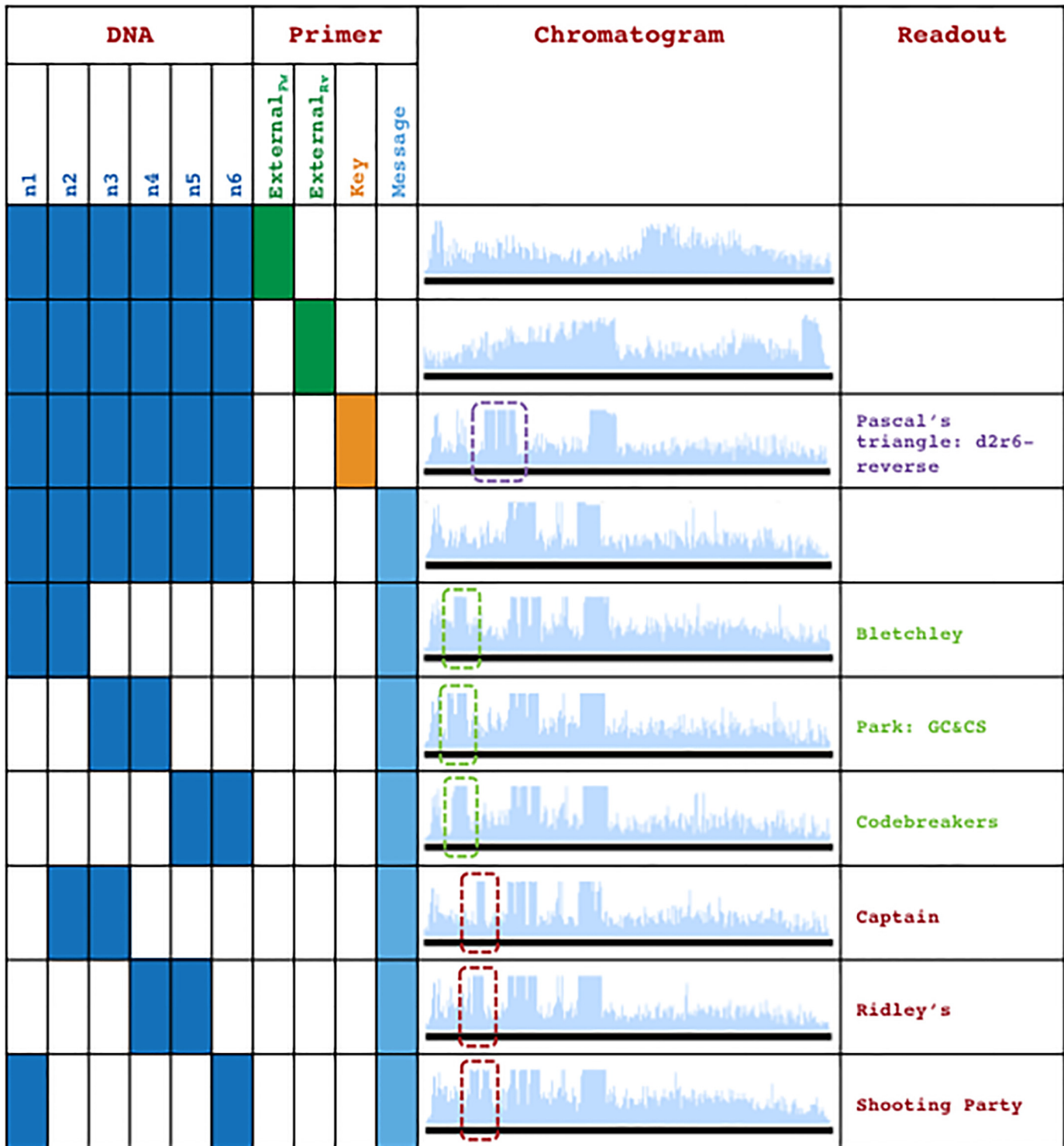


Fig 14. WWII communication readouts of tuned and co-sequenced DNA strands.

doi:10.1371/journal.pone.0152774.g014

patterning (Fig 14). Similarly, chromatogram patterning is not observed as expected when $\text{Primer}_{\text{Message}}$ is used for co-sequencing all six strands, since the proper strand combinations are not being co-sequenced as per the combination key. However, co-sequencing of DNA pairs with $\text{Primer}_{\text{Message}}$ as per the order in Pascal's triangle— $n1+n2$, $n3+n4$, and $n5+n6$ —reveals the message “Bletchley Park: GC&CS Codebreakers” via chromatogram patterning (Fig 15). Alternatively, the co-sequencing of the incorrect pairs— $n2+n3$, $n4+n5$, and $n6+n1$ —reveals the decoy message “Captain Ridley's Shooting Party” (Fig 15). Expectedly, co-sequencing of other pair combinations did not lead to any patterning (Fig 11b).

While co-sequencing of a pooled DNA communication with primers that are not specific to the messages results in poor quality sequencing reads and a camouflage effect with Sanger sequencing (Fig 13), an unauthorized end-user may use other sequencing platforms such as next-generation sequencing (NGS) to gain access to encoded information. To recreate such a scenario, we tested the difficulty associated with NGS analysis of DNA samples, where the end-user has no prior knowledge of the DNA sequences or what is encoded within them. Accordingly, we prepared a pooled and purified DNA sample containing the DNA molecules $n1+n2+n3+n4+n5+n6$ from the WWII communication (Fig 10). We then submitted the sample for NGS analysis to an outside party under blind experimental conditions, asking them to provide us with the assembled contents of the sample (Fig 16a and 16b). While sequencing of the mixture produced ~2 million reads (Table 3), the blind assembly of the reads to reconstruct the contents proved difficult and inconclusive. However, after the initial analysis we informed the outside party that there were 6 plasmids in the sample, each containing 525 bp messages as inserts. We further provided the vector sequence and asked for the exact sequences of the messages in the sample. A second round of analysis identified 6 assembled sequences that represented our encoded information (Table 4). Alignment of the 6 identified sequences with $n1$, $n2$, $n3$, $n4$, $n5$, and $n6$ templates provided most of the information in the six DNA molecules, with $n1$, $n2$, $n3$, and $n5$ providing almost perfect sequence alignments (Fig 16c). Therefore, an end-user should be able to extract data from a fragmented communication using NGS with prior knowledge of the DNA contents and the encoding method.

Conclusions

Rapid advances in DNA synthesis and sequencing technologies are enabling the use of DNA for non-biological applications. One promising application that has emerged is the use of synthetic DNA for data storage, both for communication and long-term data archiving [1]. However, in these early days within the field we advocate exploring different methods—of putting information into, transporting information as, and taking information out of DNA molecules—with the long range goal of attaining a standard that can be accepted by industry and implemented for future applications.

Here we developed a method of encoding information in DNA that reduces the formation of homopolymers by taking into account the frequency of usage of different characters in English text. Our iKey-64 method is designed to convert both plaintext and numerals into a DNA language, while allowing for personalization. Users can shuffle codons assigned to the keyboard or alter the characters within the keyboard to develop a unique layout. With chromatogram patterning and homopolymer reduction, codon shuffling will enable 9.1×10^{61} iKey-64 variants out of a maximum of $64! = 1.3 \times 10^{89}$ iKey-64 variants (Fig 3). Furthermore, we developed a secret-sharing system in MuSE that explores fragmentation of messages across multiple distinct DNA molecules, and enables a new method—chromatogram patterning—to locate messages within DNA molecules. Our encoding, transfer, and data extraction methods are proof-of-concept experiments that are designed to explore new approaches of

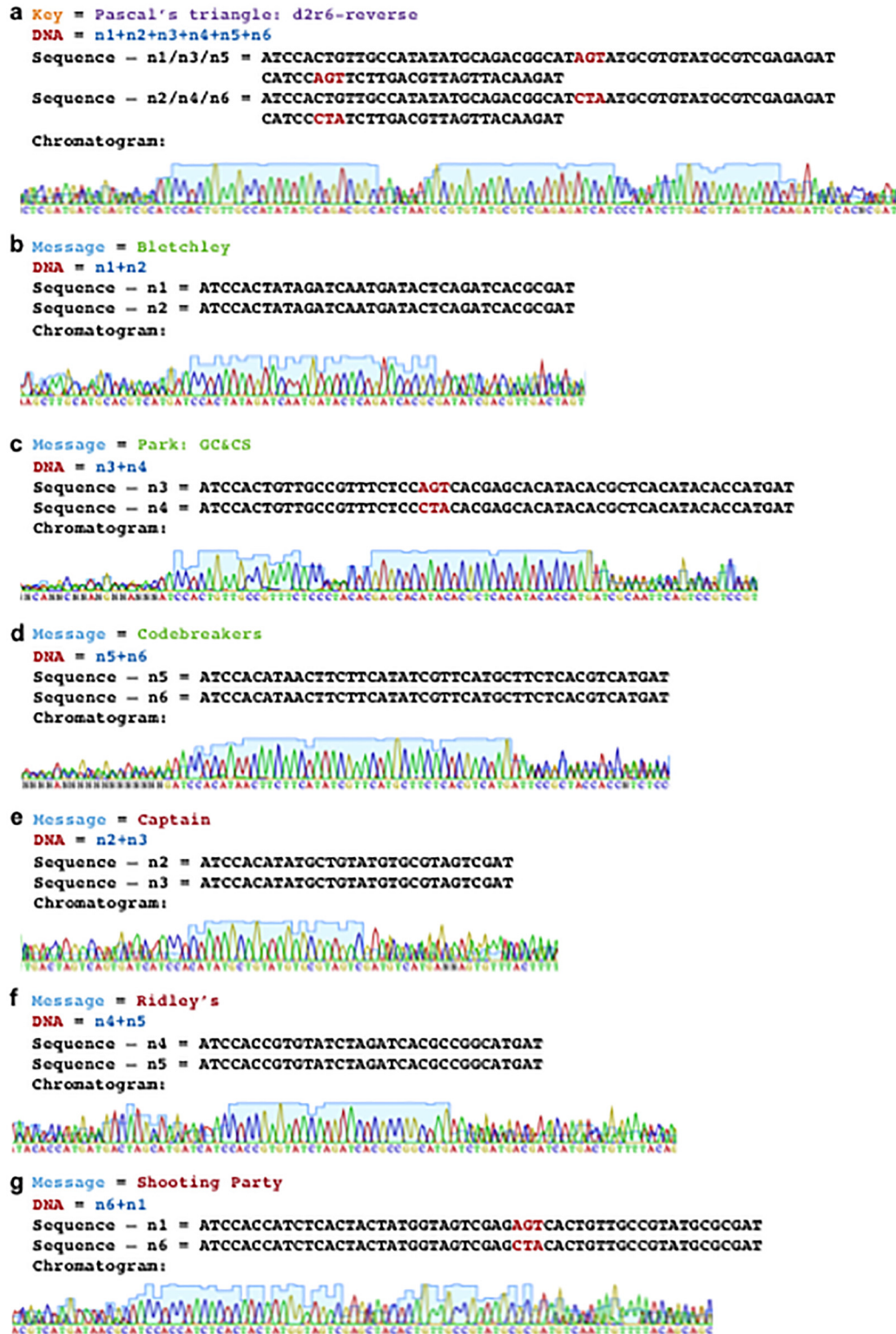


Fig 15. Examination of the peaks produced during co-sequencing of the WWII communication. Details of the encoded information, strand combinations tuned and co-sequenced, DNA sequence of embedded messages, and close-ups of the chromatogram patterns produced are shown for the WWII communication including: (a) the combination key, (b), (c), (d), the desired message, and (e), (f), (g), the decoy message. Space1 was used for all odd numbered strands ($n1, n3, n5$) and space2 was used for all even numbered strands ($n2, n4, n6$) to demarcate words. Space1/2 codons are shown in red.

doi:10.1371/journal.pone.0152774.g015

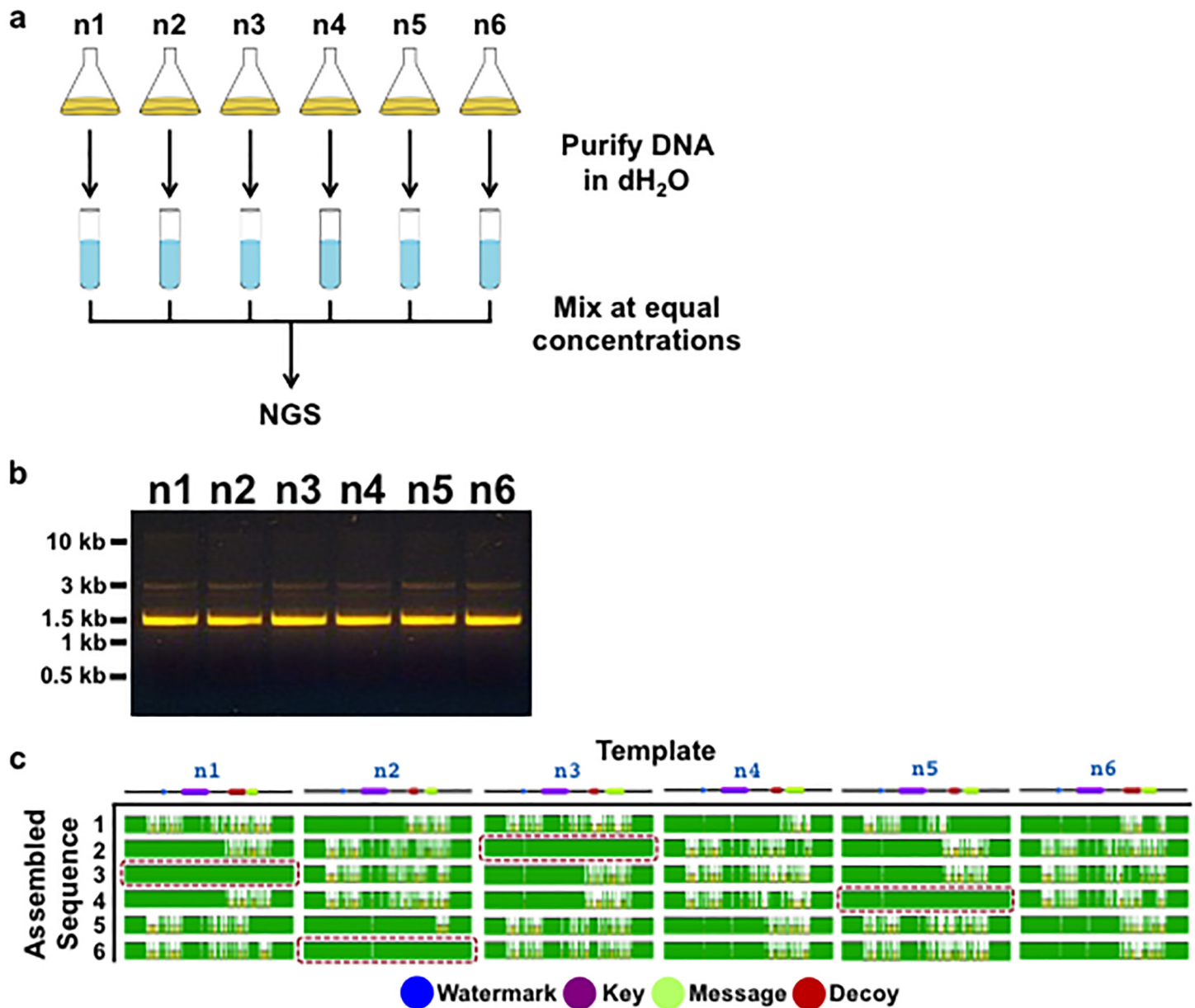


Fig 16. Next-generation sequencing of the WWII communication. (a) Plasmids containing n1, n2, n3, n4, n5, and n6 sequences (Fig 10) were grown and purified in dH₂O, mixed at equal concentrations of 30 ng/μL, and submitted to an outside party (MIT BioMicro Center) for NGS sequencing and assembly under blind experimental conditions. (b) 300 ng of plasmids containing n1, n2, n3, n4, n5, and n6 sequences were run on a 1% agarose gel to demonstrate purity. (c) The outside party (MIT BioMicro Center) 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 4) 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 templates are shown. Shown below is a legend for the color-coding of the templates. Boxes highlight assembled sequences with near perfect alignment to corresponding templates.

doi:10.1371/journal.pone.0152774.g016

communicating via DNA, and to initiate thought and discussion on how to structure DNA communication channels.

Some limitations of our approach are that the iKey-64 method of encoding has not been tested for encoding numerals, and other methods [7–9] may prove more efficient. The iKey-64 method does not incorporate data compression and it would be interesting to explore ways to adapt data compression methods for encoding using this method. Also, the iKey-64 character-

Table 3. 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 16) 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 into the original constructs proved challenging and the results were inconclusive. Information regarding the plasmids containing the WWII communication: 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.

doi:10.1371/journal.pone.0152774.t003

codon assignment was based on the frequenting of usage of characters in the Oxford English Dictionary and with numerals categorized as high frequency characters. While this method was tested for English-based communications, in theory character assignment can be modified for other languages using a similar approach. Additionally, encoding with iKey-64 needs to be combined with other cryptography such as AES, RSA, Twofish and other methods to ensure data security. This will in turn require customized versions of iKey to be designed to allow for the encoding of encrypted data. Moreover, our chromatogram patterning method has been designed for Sanger sequencing. Since we were exploring the fragmented communication of short messages that could be read with a single Sanger read, we chose to use Sanger sequencing for initial proof-of-concept experiments. However, future experiments will need to focus on using NGS sequencing methods as they are more efficient, cost effective, and allow for the investigation of more complex fragmented communications. Additionally, the concept of storing information within overlapping sequencing reads in multiplexed sequencing reactions may also be adaptable for nanopore sequencing and NGS methods.

Thus far our experiments exclusively utilized DNA maintained *in vitro* in the form of plasmid DNA. Maintaining a full set of MuSE plasmids inside the same cell is problematic due to likely segregation loss, as the plasmids share common replication origins and resistance markers [20]. However, DNA maintained *in vivo* can also be used for the communication of digital information, for example for genome watermarking applications [1]. In future experiments, we aim to utilize programmable post-translational protein assembly [21–24] to develop an addition module that will allow for the intracellular dual maintenance of two plasmids with a common origin and selection marker, which would enable *in vivo* MuSE. We intend for these early explorations to stimulate the development of future DNA communication tools that in turn may further broaden adoption of DNA for communication, an increasing possibility with the development of portable sequencing devices [25,26].

Table 4. Identified sequences from NGS analysis.

Assembled Sequence	Sequence
1	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGAGATCAACGCTAGCATCGCACTGTCGTCATGCAGCTGACTCCGATCTGACTATCGTCTGAGATCAGAGCGTAACGTAGTCAGTGCATGATGCGAACTCGATGATCGAGTCGTATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCATCCCTATCTTGACGTTAGTTACAAGATCCCACCAATACTGCCAATAGACGGTCCTCTTTCCCGTTGCTGTAAAACAGTCA TGATCGTCATCAGATCATGCCGGCGTGATCTAGATACACGGTGGATTAGCTACTAGTCAATCATGACGTGAGAAGCAT GAACGATATGAAGAAGTTATGTGGATAGCTGTCGACGTGATCGTATCGATGCAGTCCTCAGGTCATATTACTCGACAGTT GCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC
2	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGACGTAGTCGCTCGACCTGACTCG TGAGTGCAGCTACGTGTCAGTCATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCAT CCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTAACGTCTCTTCCACCTTTCCAAAAAGTAAACACCCA TGACGTATCGACTACGCACATACAGCATATGTGGATGATCACTGACTGACTGAACTACGATCATGGTGTATGTGAGCGTG TATGTGCTCGTGACTGGAGAAACGGCAACAGTGGATGATTGACGTACGACTGCTAGCTCAGGTCATATTACTCGACAGTT GCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC
3	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGACGTAGTCGCTCGACCTGACTCG TGAGTGCAGCTACGTGTCAGTCATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCAT CCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTAACGTCTCTTCCACCTTTCCAAAAAGTAAACCCGAC TGATCGGCATACGGCAACAGTACTCTGACTACCATAGTGTAGTGTAGGATGATTACGATCGCGTGATCTGAGTATCA TTGATCTATAGTGGATTGACTGATGATCGTACTGTGCTGACTGACTGACGTGCGATCTCAGGTCATATTACTCGACAGTT GCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC
4	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGATCATGATTCTGATCTAGTCCAGCAGTAGAGTCGTCTCGATCGATCTGTGCATCGTCAGCGATATTCGACGTAGTCGCTCGACCTGACTCG TGAGTGCAGCTACGTGTCAGTCATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCAT CCAGTTCTTGACGTTAGTTACAAGATTGGCCACGATCCATGCTAACGTCTCTTCCACCTTTCCAAAAAGTAAACTGAC TGCATTGTCGATCATCATGCCGGCGTGATCTAGATACACGGTGGATTGACTACTAGTCGAATCATGACGTGAGAAGCAT GAACGATATGAAGAAGTTATGTGGATAGCTGTCGACGTGATCGTATCGATGCAGTCCTCAGGTCATATTACTCGACAGTT GCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC
5	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGAGATCAACGCTAGCATCGCACTGTCGTCATGCAGCTGACTCCGATCTGACTATCGTCTGAGATCAGAGCGTAACGTAGTCAGTGCATGATGCGAACTCGATGATCGAGTCGTATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCAT CCCTATCTTGACGTTAGTTACAAGATCCCACCAATACTGCCAATAGACGGTCCTCTTTCCCGTTGCTGTAAAACATAGT CATGACATCGACTACGCACATACAGCATATGTGGATCTAGCTTACTAGTCAACGTGATATCGCGTGATCTGAGTATCA TTGATCTATAGTGGATTGACTGATGATCGTACTGTGCTGACTGACTGACGTGCGATCTCAGGTCATATTACTCGACAGTT GCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC
6	TAATACGACTCACTATAGGGACAGTCTAGTGCAGCAGTCAGTACGAGTCTCATGAGTGTAGGATGCATGAGATCAACGCTAGCATCGCACTGTCGTCATGCAGCTGACTCCGATCTGACTATCGTCTGAGATCAGAGCGTAACGTAGTCAGTGCATGATGCGAACTCGATGATCGAGTCGTATCCACTGTTGCCATATATGCAGACGGCATAGTATGCGTGTATGCGTCGAGAGATCAT CCCTATCTTGACGTTAGTTACAAGATCCCACCAATACTGCCAATAGACGGTCCTCTTTCCCGTTGCTGTAAAACATAGT CATGACATCGACTACGCACATACAGCATATGTGGATCTAGCTTACTAGTCAACGTGATATCGCGTGATCTGAGTATCA TTGATCTATAGTGGATCATGACGTGCATGCAAGCTTAGCTAGTCAGATCAGTACTCTCAGGTCATATTACTCGACAGTT GCTAAGTCAGTCATCGTCATACGATGCCGCTGAGCAATAACTAGC

doi:10.1371/journal.pone.0152774.t004

Acknowledgments

This work is sponsored by the Defense Advanced Research Projects Agency under Air Force Contract #FA8721-05-C-0002, and NIH grants 1R01EB017755, 1DP2OD008435, and 1P50GM098792. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.

Author Contributions

Conceived and designed the experiments: BZ PAC. Performed the experiments: BZ. Analyzed the data: BZ PAC TKL. Contributed reagents/materials/analysis tools: BZ PAC TKL. Wrote the paper: BZ PAC TKL.

References

1. Zakeri B, Lu TK (2015) DNA nanotechnology: new adventures for an old warhorse. *Curr Opin Chem Biol* 28: 9–14. doi: [10.1016/j.cbpa.2015.05.020](https://doi.org/10.1016/j.cbpa.2015.05.020) PMID: [26056949](https://pubmed.ncbi.nlm.nih.gov/26056949/)
2. Clelland CT, Risca V, Bancroft C (1999) Hiding messages in DNA microdots. *Nature* 399: 533–534. PMID: [10376592](https://pubmed.ncbi.nlm.nih.gov/10376592/)
3. Bancroft C, Bowler T, Bloom B, Clelland CT (2001) Long-term storage of information in DNA. *Science* 293: 1763–1765.
4. Liss M, Daubert D, Brunner K, Kliche K, Hammes U, Leihner A, et al. (2012) Embedding permanent watermarks in synthetic genes. *PLoS ONE* 7: e42465. doi: [10.1371/journal.pone.0042465](https://doi.org/10.1371/journal.pone.0042465) PMID: [22905136](https://pubmed.ncbi.nlm.nih.gov/22905136/)
5. Cox JP (2001) Long-term data storage in DNA. *Trends Biotechnol* 19: 247–250. PMID: [11412947](https://pubmed.ncbi.nlm.nih.gov/11412947/)
6. Sennels L, Bentin T (2012) To DNA, all information is equal. *Artif DNA, PNA & XNA* 3: 109–111.
7. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, et al. (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* 494: 77–80. doi: [10.1038/nature11875](https://doi.org/10.1038/nature11875) PMID: [23354052](https://pubmed.ncbi.nlm.nih.gov/23354052/)
8. Church GM, Gao Y, Kosuri S (2012) Next-generation digital information storage in DNA. *Science* 337: 1628. PMID: [22903519](https://pubmed.ncbi.nlm.nih.gov/22903519/)
9. Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ (2015) Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew Chem* 54: 2552–2555.
10. Carr PA, Church GM (2009) Genome engineering. *Nat Biotechnol* 27: 1151–1162. doi: [10.1038/nbt.1590](https://doi.org/10.1038/nbt.1590) PMID: [20010598](https://pubmed.ncbi.nlm.nih.gov/20010598/)
11. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* 4: 265–270. doi: [10.1038/nnano.2009.12](https://doi.org/10.1038/nnano.2009.12) PMID: [19350039](https://pubmed.ncbi.nlm.nih.gov/19350039/)
12. Haughton D, Balado F (2013) BioCode: two biologically compatible Algorithms for embedding data in non-coding and coding regions of DNA. *BMC Bioinforma* 14: 121.
13. Heider D, Barnekow A (2007) DNA-based watermarks using the DNA-Crypt algorithm. *BMC Bioinforma* 8: 176.
14. Tulpan D, Regoui C, Durand G, Belliveau L, Leger S (2013) HyDEn: a hybrid steganocryptographic approach for data encryption using randomized error-correcting DNA codes. *BioMed Res Int* 2013: 634832. doi: [10.1155/2013/634832](https://doi.org/10.1155/2013/634832) PMID: [23984392](https://pubmed.ncbi.nlm.nih.gov/23984392/)
15. Kawano T (2013) Run-length encoding graphic rules, biochemically editable designs and steganographical numeric data embedment for DNA-based cryptographical coding system. *Commun & Integr Biol* 6: e23478.
16. Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 55: 641–658. doi: [10.1373/clinchem.2008.112789](https://doi.org/10.1373/clinchem.2008.112789) PMID: [19246620](https://pubmed.ncbi.nlm.nih.gov/19246620/)
17. Oxford University Press. What is the frequency of the letters of the alphabet in English? Available at <http://www.oxforddictionaries.com/us/words/what-is-the-frequency-of-the-letters-of-the-alphabet-in-english>. (2014).
18. Ferguson N, Schneier B, Kohno T (2010) *Cryptography engineering: design principles and practical applications*. Indianapolis: Wiley Publishing, Inc.
19. Bletchley Park Trust. Captain Ridley's Shooting Party. Available at <http://www.bletchleypark.org.uk/content/hist/worldwartwo/captridley.rhtm>. (2014).
20. San Millan A, Peña-Miller R, Toll-Riera M, Halbert ZV, McLean AR, Cooper BS, et al. (2014) Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nat Commun* 5: 5208. doi: [10.1038/ncomms6208](https://doi.org/10.1038/ncomms6208) PMID: [25302567](https://pubmed.ncbi.nlm.nih.gov/25302567/)
21. Zakeri B, Howarth M (2010) Spontaneous intermolecular amide bond formation between side chains for irreversible peptide targeting. *J Am Chem Soc* 132: 4526–4527. doi: [10.1021/ja910795a](https://doi.org/10.1021/ja910795a) PMID: [20235501](https://pubmed.ncbi.nlm.nih.gov/20235501/)
22. Zakeri B, Fierer JO, Celik E, Chittock EC, Schwarz-Linek U, Moy VT, et al. (2012) Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin. *Proc Natl Acad Sci United States Am* 109: E690–E697.
23. Veggiani G, Zakeri B, Howarth M (2014) Superglue from bacteria: unbreakable bridges for protein nanotechnology. *Trends Biotechnol* 32: 506–512. doi: [10.1016/j.tibtech.2014.08.001](https://doi.org/10.1016/j.tibtech.2014.08.001) PMID: [25168413](https://pubmed.ncbi.nlm.nih.gov/25168413/)
24. Zakeri B (2015) *Synthetic Biology: A New Tool for the Trade*. *ChemBiochem* : Eur J Chem Biol 16: 2277–2282.

25. Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proc Natl Acad Sci United States Am* 106: 7702–7707.
26. Blazej RG, Kumaresan P, Mathies RA (2006) Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc Natl Acad Sci United States Am* 103: 7240–7245.