nature biotechnology

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41587-019-0240-x

In the format provided by the authors and unedited.

Data storage in DNA with fewer synthesis cycles using composite DNA letters

Leon Anavy[1*], Inbal Vaknin[2], Orna Atar[2], Roee Amit[2] and Zohar Yakhini[1,3*]

[1]Computer Science Department, Technion – Israel Institute of Technology, Haifa, Israel. [2]Faculty of Biotechnology and Food Engineering, Technion – Israel Institute of Technology, Haifa, Israel. [3]School of Computer Science, Herzliya Interdisciplinary Center, Herzliya, Israel. *e-mail: leon.anavy@gmail.com; zohar.yakhini@gmail.com

Supplementary Information

Supplementary Note

KL divergence is equivalent to maximum likelihood mapping

Define the likelihood of a letter σ, given the observed vector $x^{(N)}$ and the error parameters, as

$$L\left(\sigma; x^{(N)}\right) = P\left(X^{(N)} = x^{(N)} \middle| \sigma\right) = \prod_{i \in \{A,C,G,T\}} \left(p_i(\sigma)\right)^{n_i}$$

Where $n_i = \#of\ occurences\ of\ i\ in\ x^{(N)}$

The log likelihood is defined accordingly:

$$l\left(\sigma; x^{(N)}\right) = \sum_{i \in \{A,C,G,T\}} n_i \log\left(p_i(\sigma)\right)$$

The KL divergence of the observed vector $x^{(N)}$ and the frequency vector of the original letter σ is defined as:

$$KL\left(\pi\left(x^{(N)}\right), p(\sigma)\right) = \sum_{i \in \{A,C,G,T\}} \pi_i\left(x^{(N)}\right) \log\left(\frac{\pi_i\left(x^{(N)}\right)}{p_i(\sigma)}\right) =$$

$$\sum_{i \in \{A,C,G,T\}} \left\{\pi\left(x^{(N)}\right)\right\}_i \log\left(\left\{\pi\left(x^{(N)}\right)\right\}_i\right) - \sum_{i \in \{A,C,G,T\}} \left\{\pi\left(x^{(N)}\right)\right\}_i \log\left(p_i(\sigma)\right) =$$

$$-H\left(\pi\left(x^{(N)}\right)\right) - \sum_{i \in \{A,C,G,T\}} \left\{\pi\left(x^{(N)}\right)\right\}_i \log\left(p_i(\sigma)\right) =$$

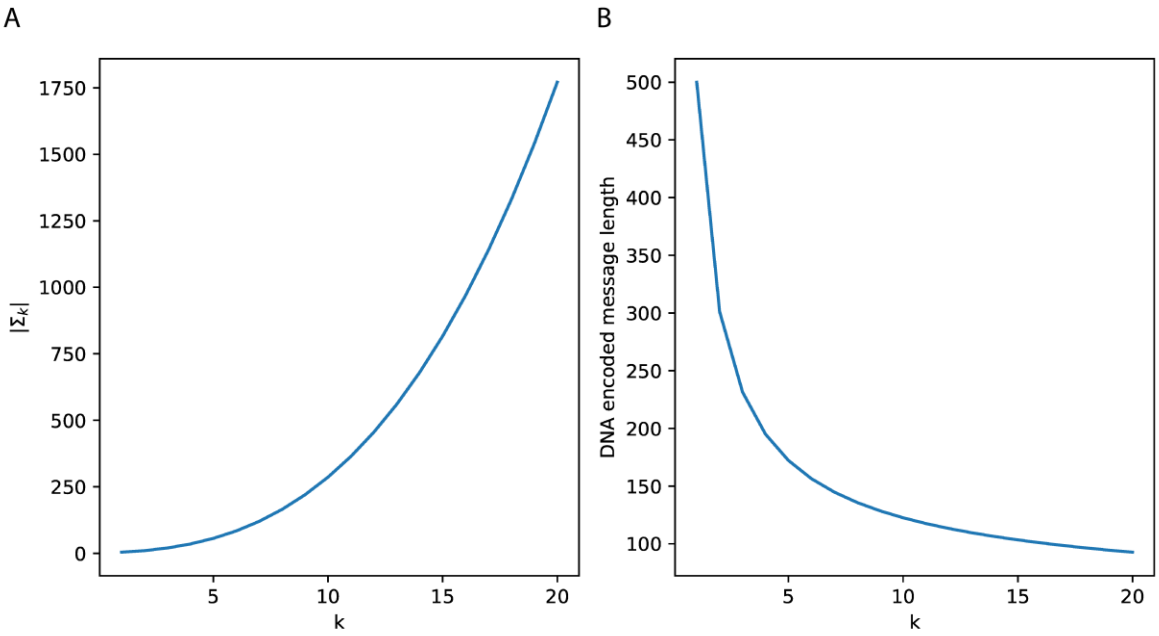
$$-H\left(\pi\left(x^{(N)}\right)\right) - \frac{l\left(\sigma; x^{(N)}\right)}{n}$$

Where $H(X)$ is the Shannon Entropy of X.

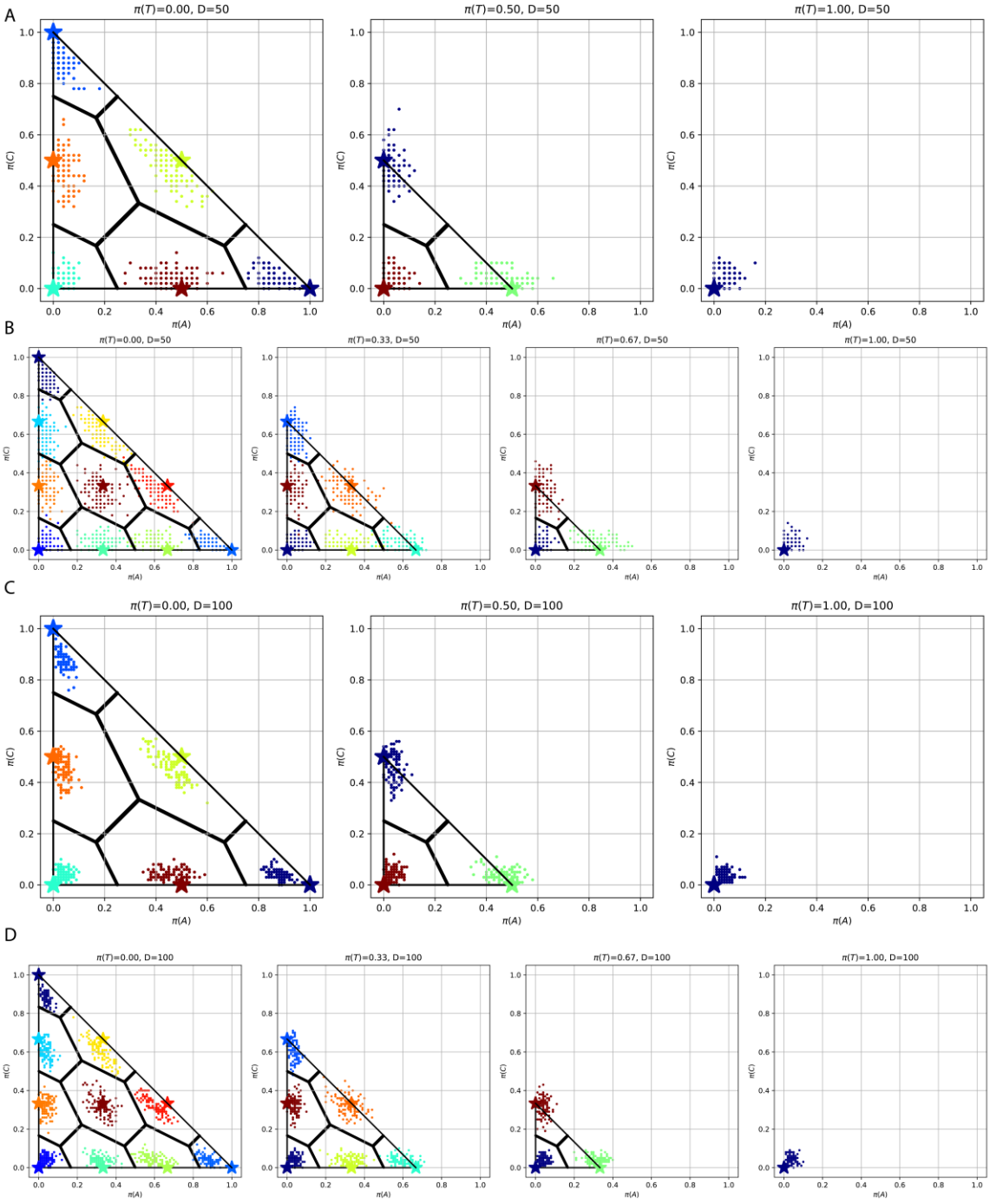
Clearly,

$$argmin_{\sigma \in \Sigma_k}\left(KL\left(\pi\left(x^{(N)}\right), p(\sigma)\right)\right) = argmax_{\sigma \in \Sigma_k}\left(l\left(\sigma; x^{(N)}\right)\right)$$

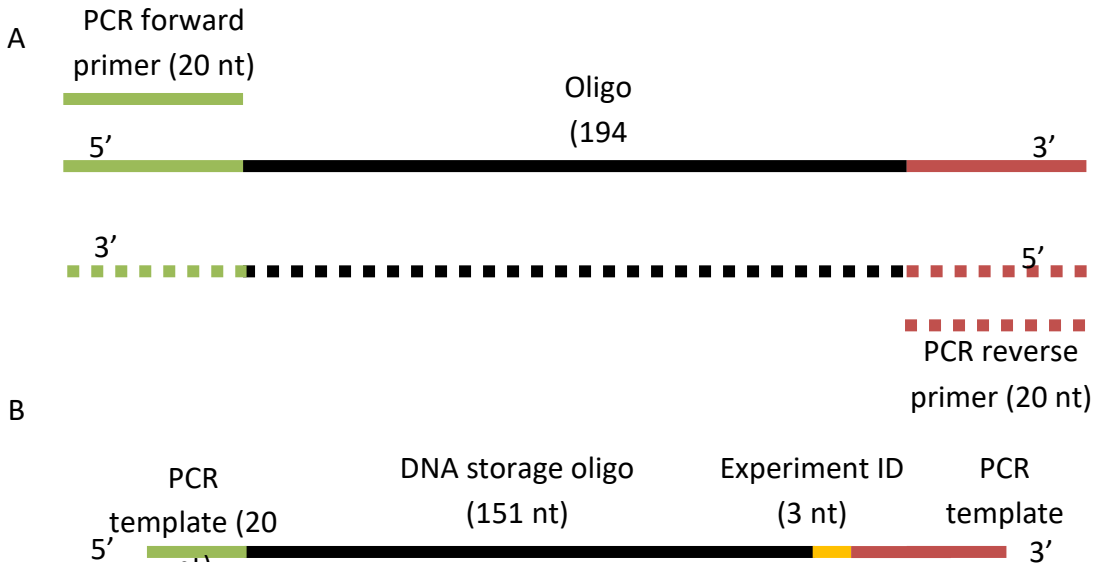
Extended data and figures



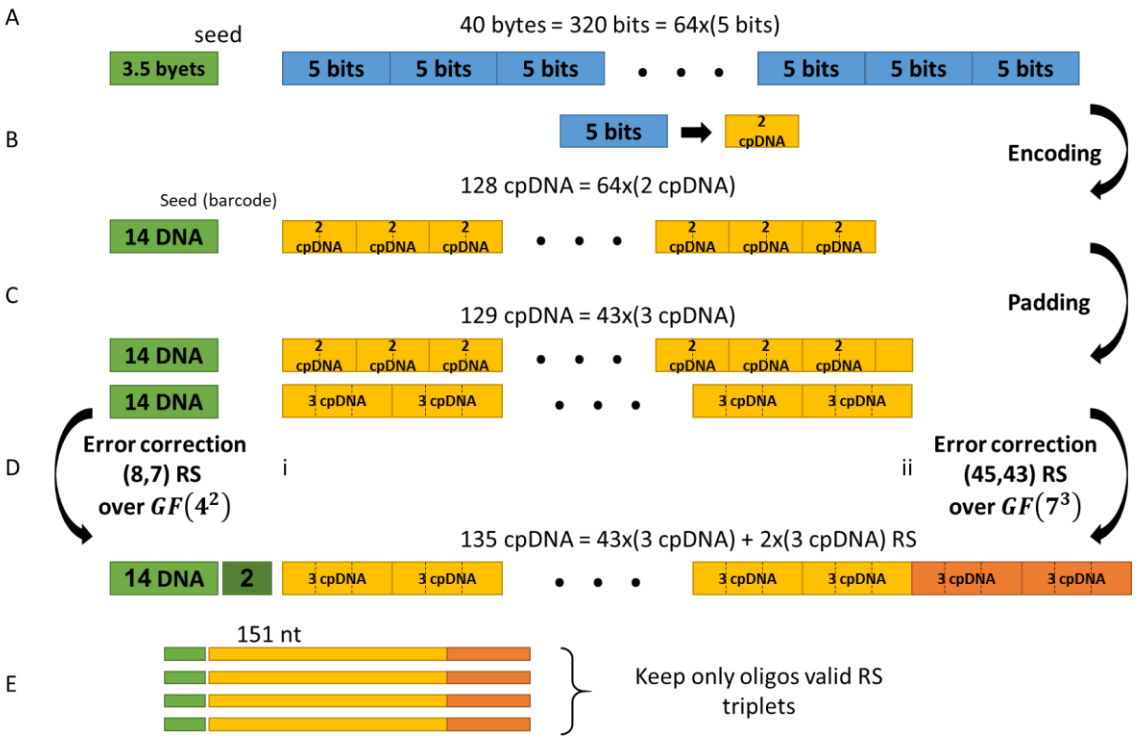
Supplementary Figure 1: Composite alphabet size and information density grow with the resolution parameter *k*. A. Alphabet size as a function of the resolution ***k***. B. required oligo length for naïve encoding of 1000bits as a function of the resolution ***k***.



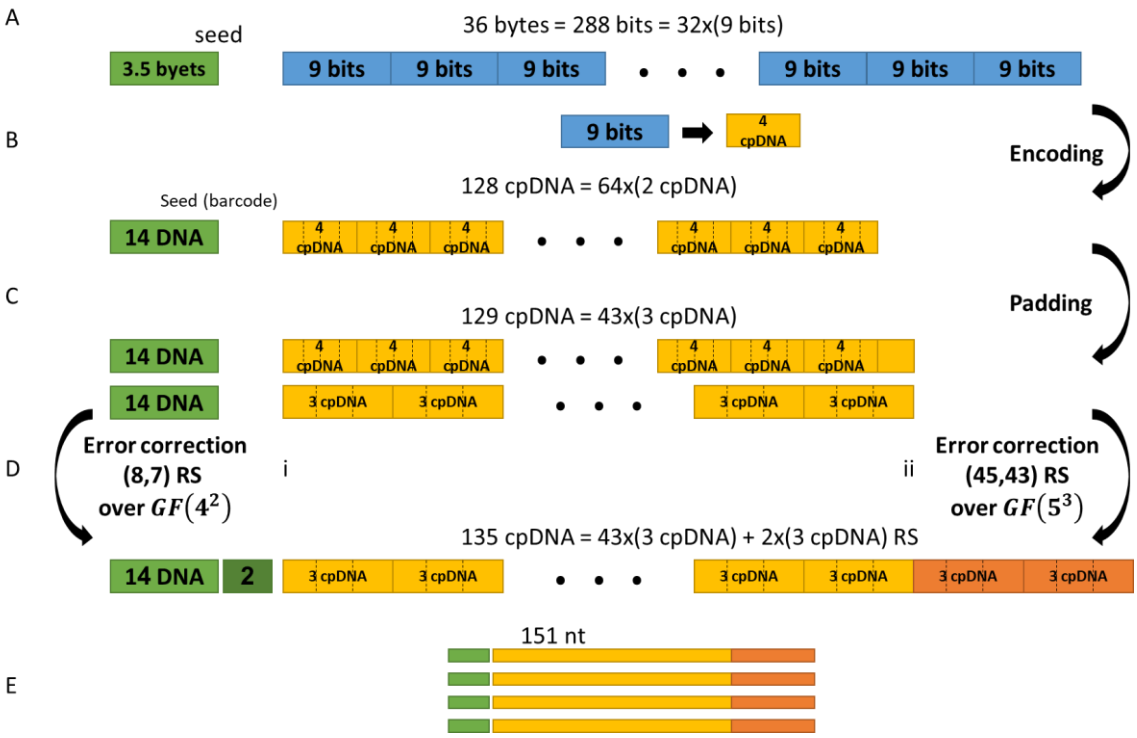
Supplementary Figure 2: Visualization of composite DNA letters inference simulation. Simulation of the letters in two composite DNA alphabets and their inference. Stars represent the base ratio of the designed letter and the small dots represents $\boldsymbol{R = 100}$ realizations. The black lines correspond to the inference areas according to L1 norm inference. A. The 10 letters of the $\boldsymbol{k = 2}$ composite alphabet with simulated sequencing depth of $\boldsymbol{D = 50}$. B. The same as A for the 20 letters of the composite alphabet of $\boldsymbol{k = 3}$. C+D. the same as A+B using a simulated sequencing depth of $\boldsymbol{D = 100}$.



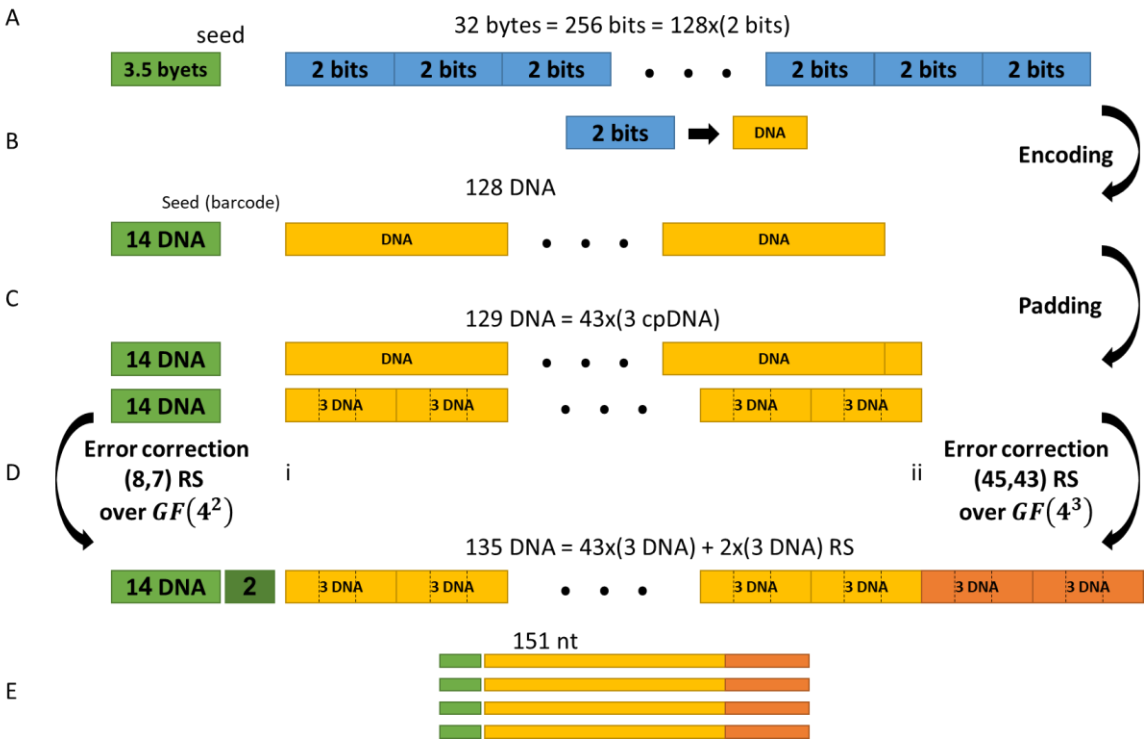
Supplementary Figure 3: Design of the oligos in the large-scale composite DNA experiments. A. The synthesized oligo contains PCR amplification templates. B. The oligo contains 151nt of the encoded message, a barcode of 3nts to identify the experiment ID.



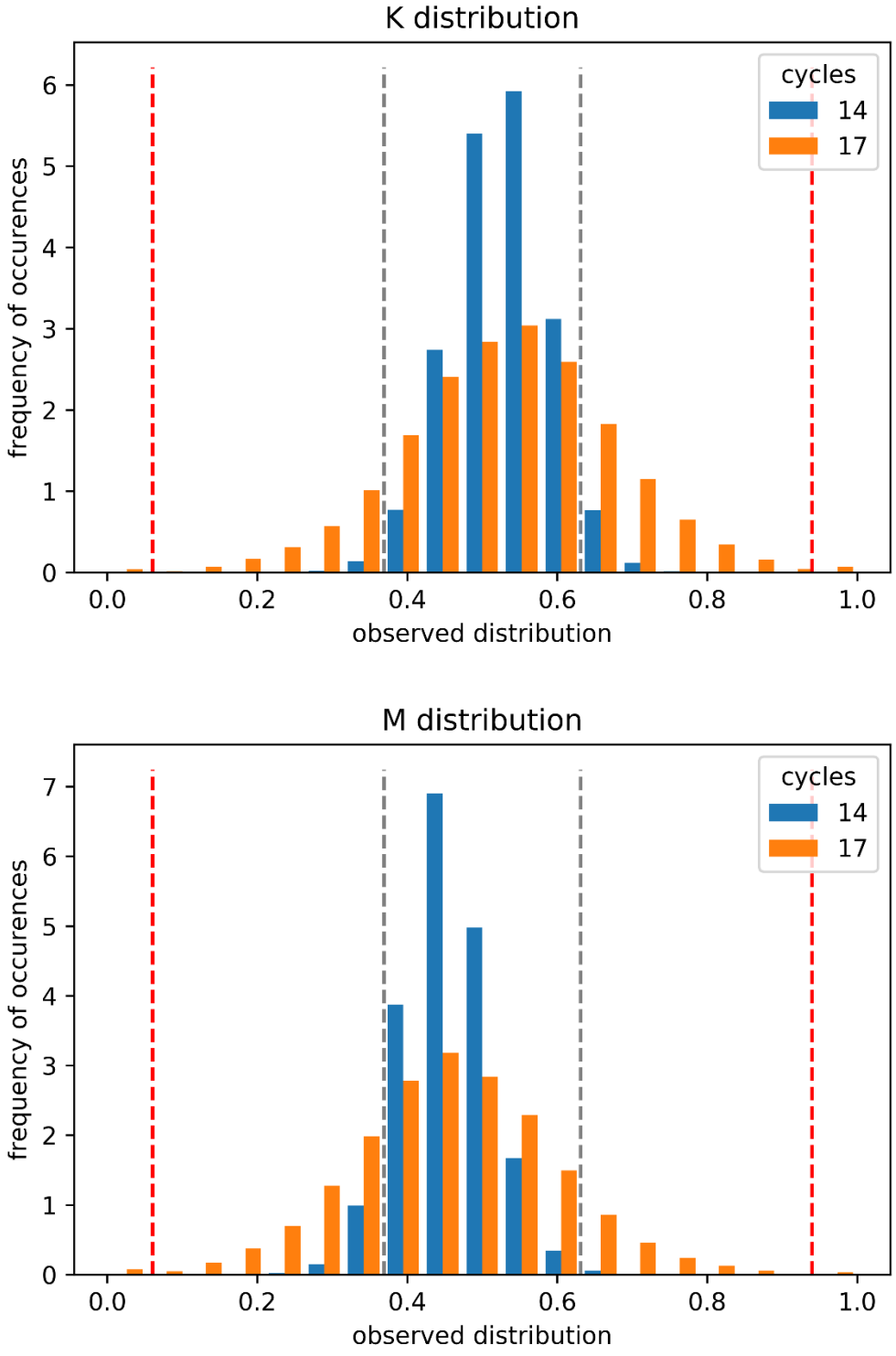
Supplementary Figure 4: Detailed encoding scheme of a six-letter composite DNA based storage system. A. The message was cut to chunks of 40 bytes each and encoded using standard fountain coding with no Reed-Solomon error correction. The fountain coding was altered so that seeds are limited to 3.5 bytes. B. The 3.5 bytes seed was converted to 14 standard DNA nucleotides and the 40 bytes payload of each fountain drop was converted into a 128nt of 6-letter composite alphabet by converting every 5 bits to 2 composite nucleotides. C. The composite message was padded to 129nt by adding a K at the 5' end. D. Reed-Solomon error correction was appended by using a systematic RS encoders to generate a 151nt oligo. (i) The 14nt seed is encoded to 16nt by using a $(\mathbf{8},\mathbf{7})$ RS over $\boldsymbol{GF}(\mathbf{4^2})$. (ii) The 129nt payload is incoded to 135nt by using a $(\mathbf{45},\mathbf{43})$ RS over $\boldsymbol{GF}(\mathbf{7^3})$. E. The encoded oligos are filtered so that only oligos in which all 6 redundancy letters are within $\boldsymbol{\Sigma_6}$ are kept. The desired number of oligos are kept and sent to synthesis.



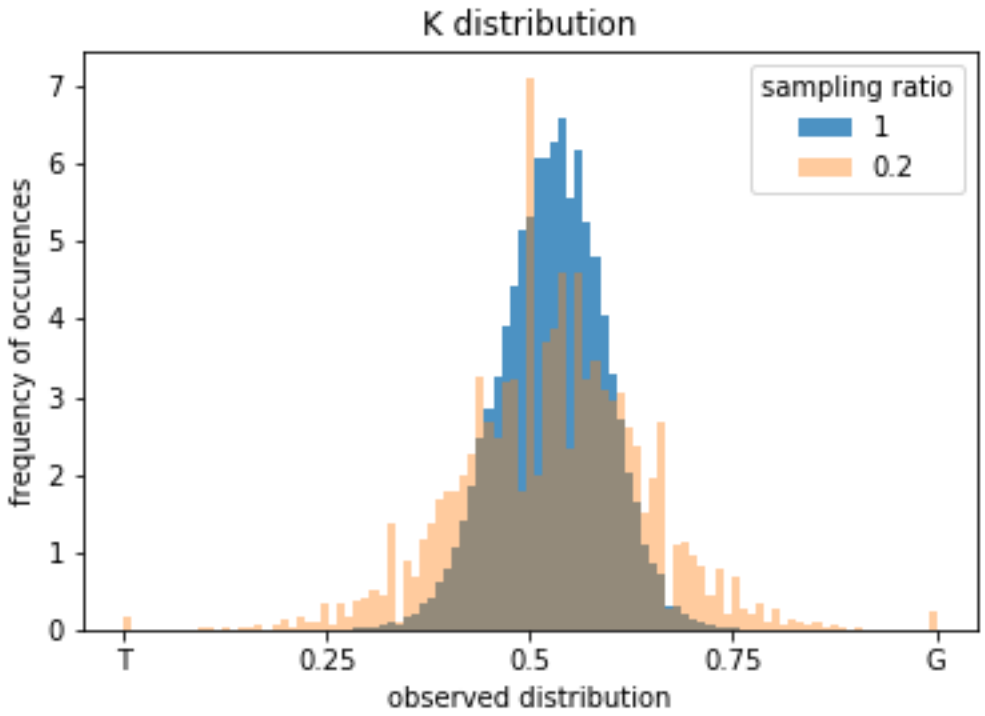
Supplementary Figure 5: Detailed encoding scheme of a five-letter composite DNA based storage system. Similar to Supplementary Figure 4 with the following differences: A. each chunk is 36 bytes. B. The conversion is done by converting every 9 bits to 4 composite DNA nucleotides. D. The payload is encoded using a $(\mathbf{45}, \mathbf{43})$ RS over $\boldsymbol{GF}(\mathbf{5^3})$. E. No filtration is needed since the encoding alphabet is identical to $\boldsymbol{\Sigma_5}$.



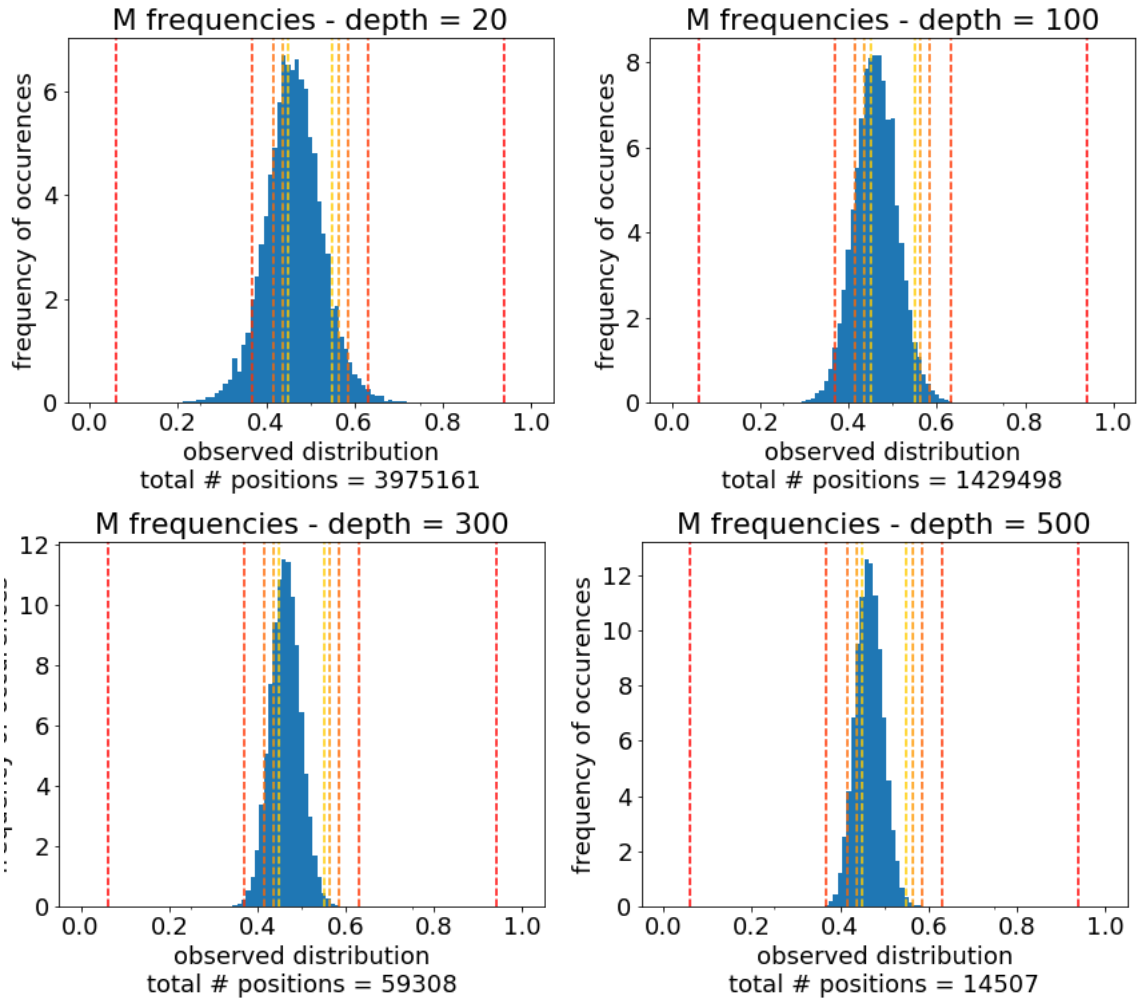
Supplementary Figure 6: Detailed encoding scheme of a four-letter composite DNA based storage system. Similar to Supplementary Figure 4 with the following differences: A. each chunk is 32 bytes. B. The conversion is done by converting every 2 bits to single DNA nucleotides. D. The payload is encoded using a $(\mathbf{45}, \mathbf{43})$ RS over $\boldsymbol{GF}(\mathbf{4^3})$. E. No filtration is needed since the encoding alphabet is identical to $\boldsymbol{\Sigma_4}$.



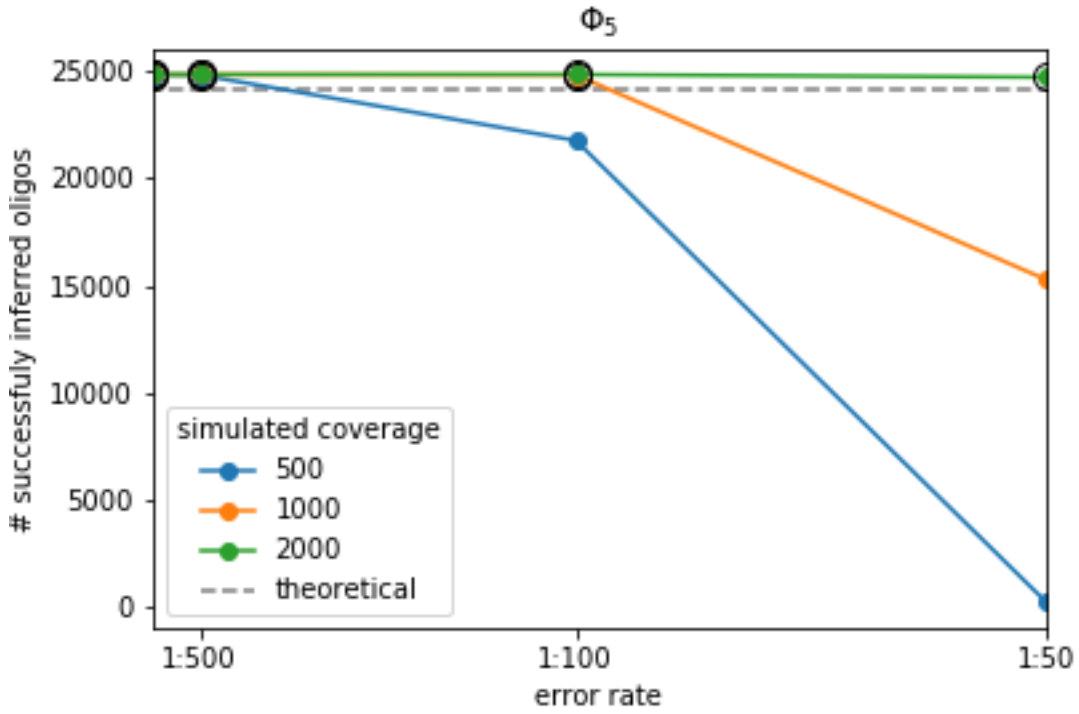
Supplementary Figure 7: Effect of PCR on the composite DNA sequences. Base frequencies for the first two dilutions representing different number of PCR cycles.



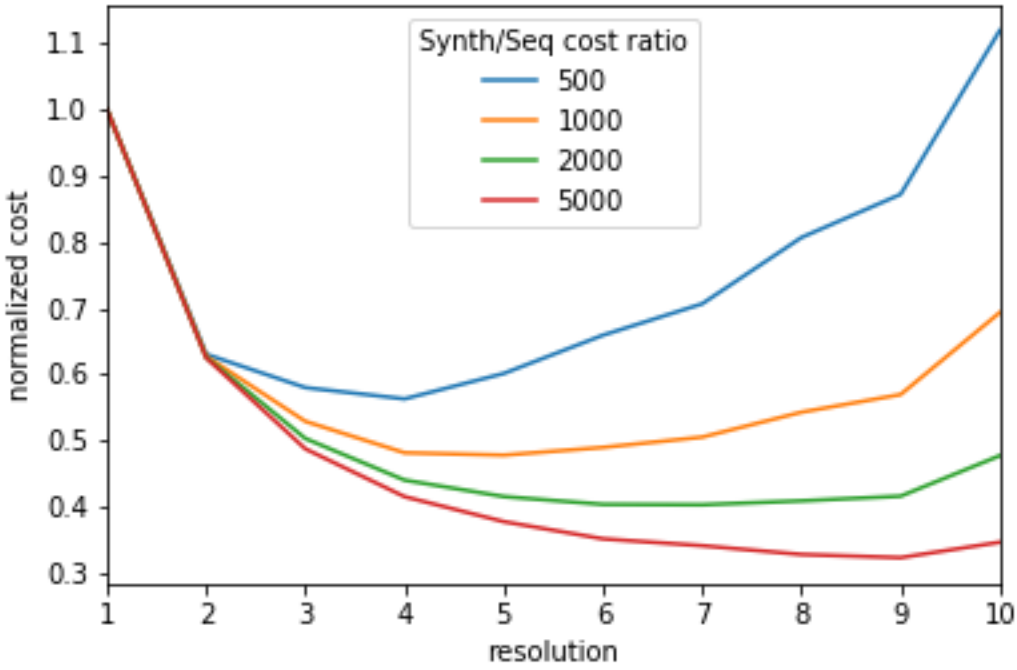
Supplementary Figure 8: Base frequency distribution of the composite letter K. Distribution of base frequencies per synthesized position. The results are presented for the entire sequencing output and for a subsample of 20% of the reads.



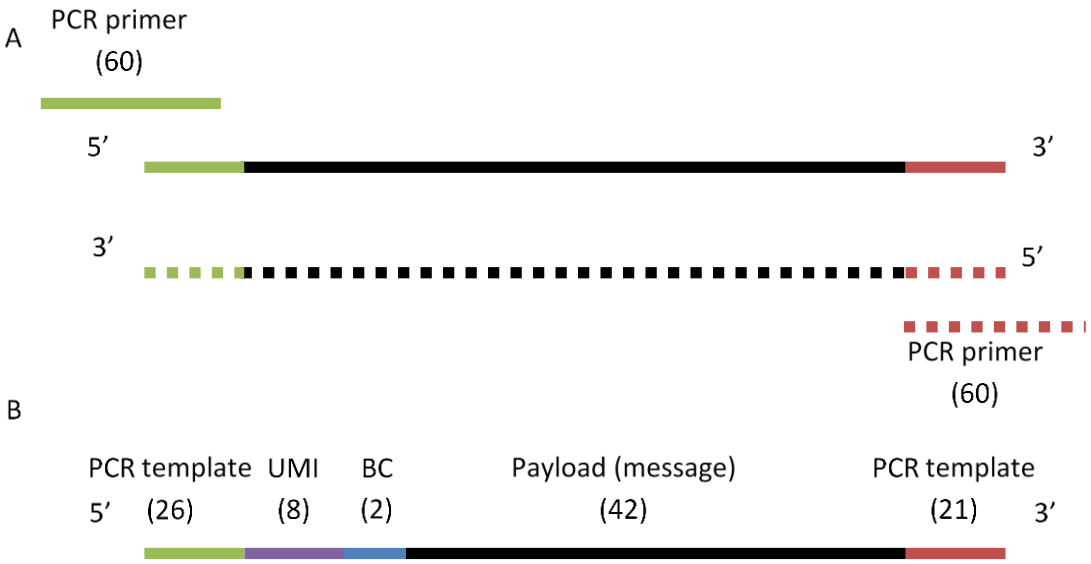
Supplementary Figure 9: Example distribution of the A frequencies for the composite letter M in different sequencing depths. Similar to figure 3d-e. Decision boundaries depicted for $\Phi_k, k = 2,4,6,8,10$.



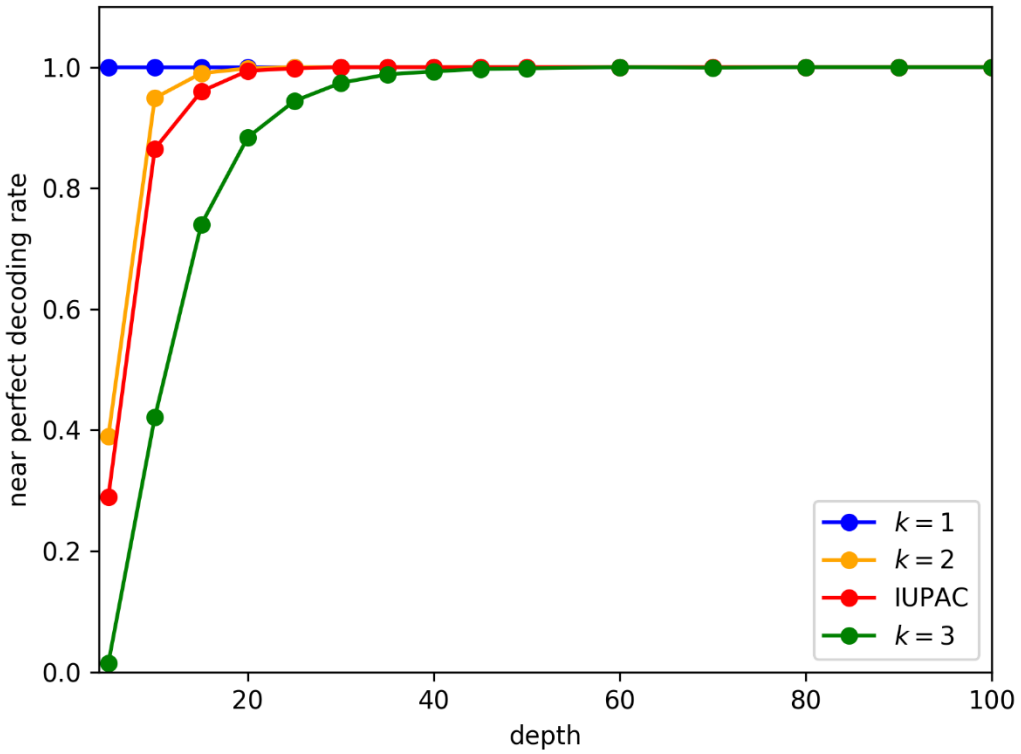
Supplementary Figure 10: Simulation results of Φ_5. Successful inference of Φ_5 composite storage system storing the message from Erlich and Zielinski[7] based on simulations (N=5). The number of correctly inferred oligos is shown as a function of the simulated error rates for 3 sequencing depths (See Online Methods). The theoretical limit of the fountain code (with 0.001 failure probability) is shown as a dashed gray line. Instances for which a successful decoding was achieved are marked with a black circle.



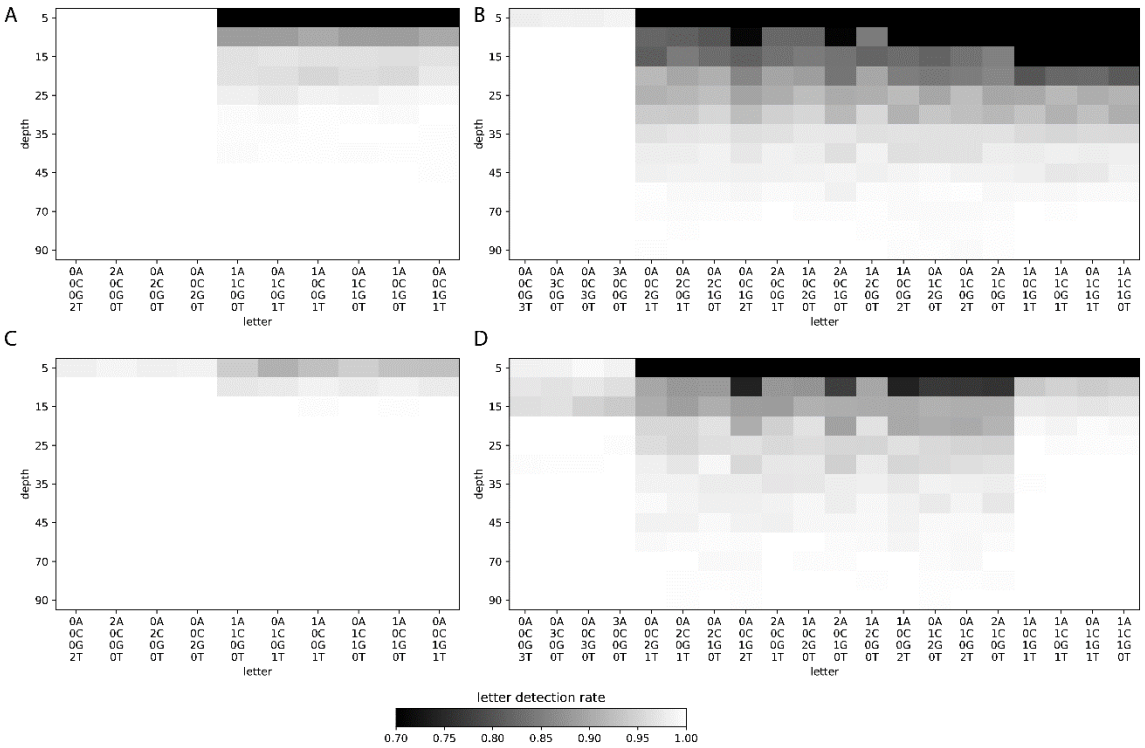
Supplementary Figure 11: Cost analyses of large-scale composite DNA storage systems. The normalized overall cost for a composite DNA based storage system using different alphabets is presented for different synthesis to sequencing cost ratios.



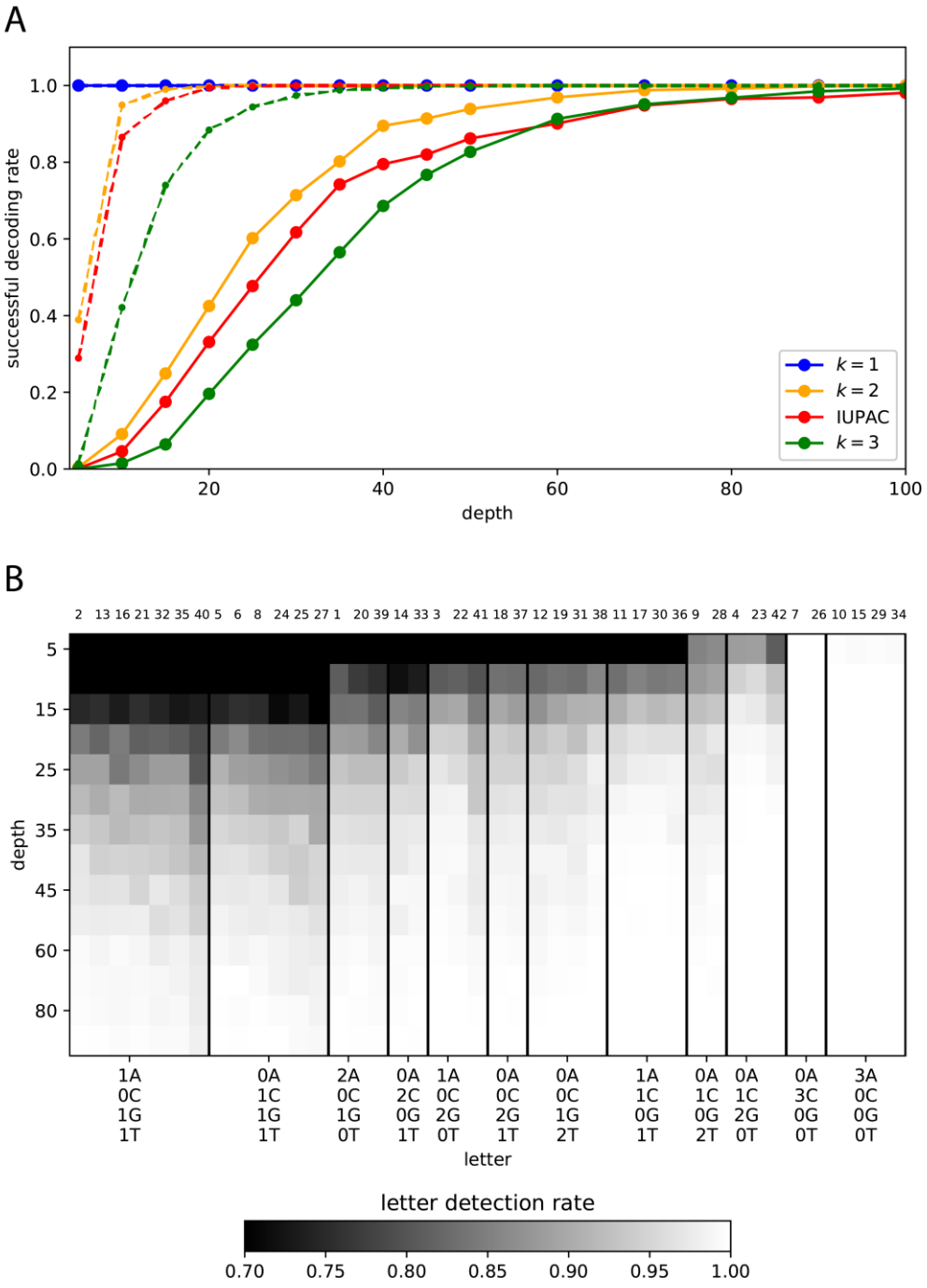
Supplementary Figure 12: Design of the oligos in the large composite alphabets experiments. A. The synthesized oligo contains PCR amplification templates to be used with Illumina Tru-Seq primers to construct a ready-for-sequencing library. B. The oligo contains a UMI region and a barcode of standard (non composite) DNA bases together with a 42 bases long composite DNA sequence.



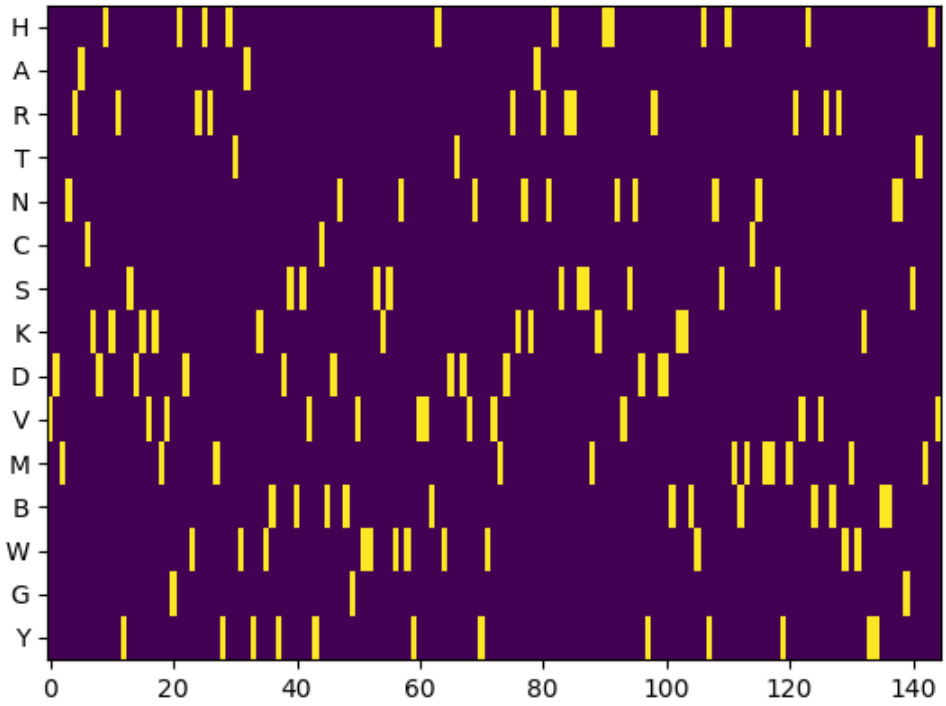
Supplementary Figure 13: Results of the proof of concept experiment. Success rate of near-perfect decoding (allowing a single error) as a function of sequencing depth for the four examined composite alphabets.



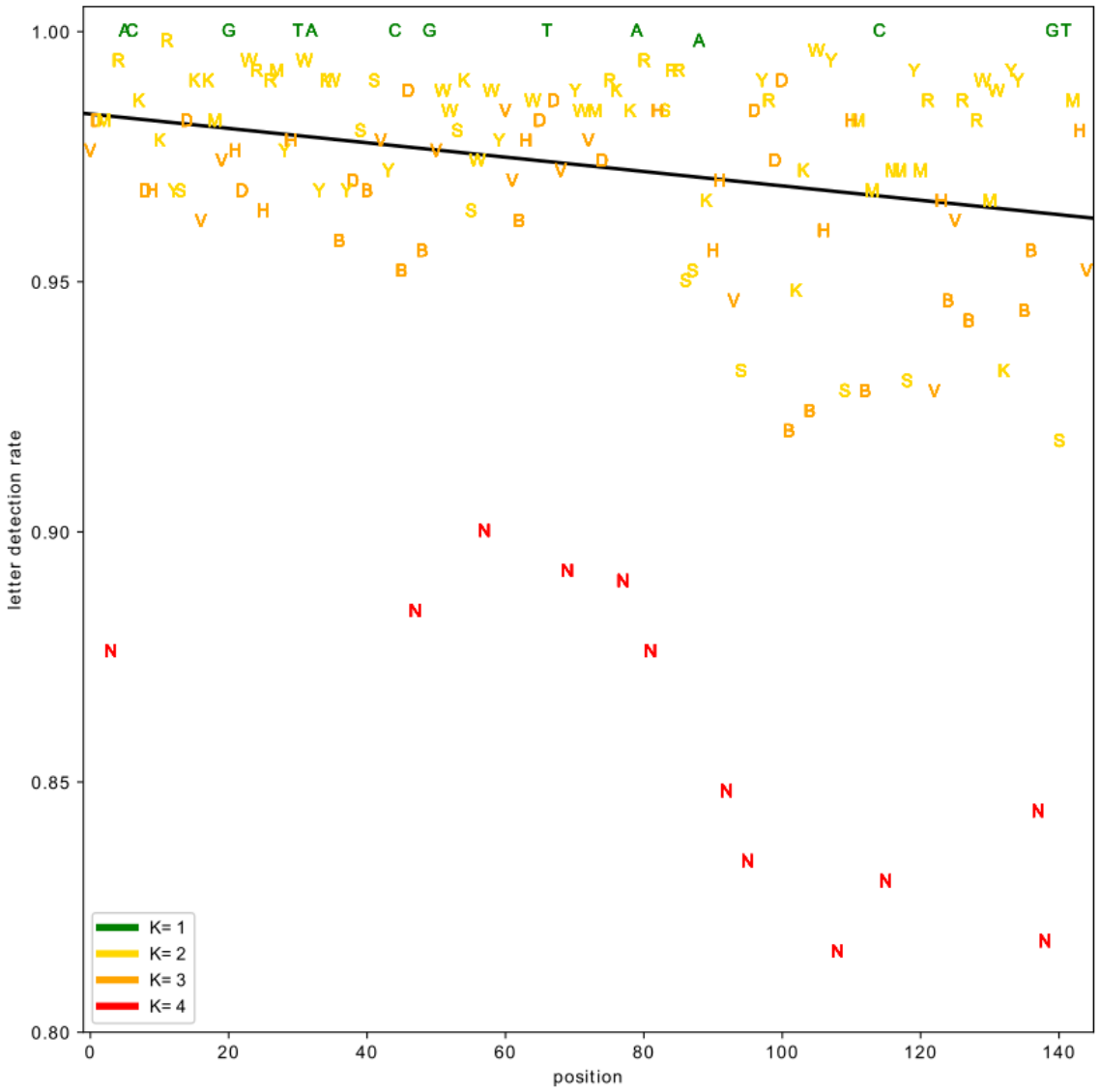
Supplementary Figure 14: Inference of composite DNA letters. Inference rates for the composite letters in two composite DNA alphabets as a function of sequencing depth and inference mechanism. A. The 10 letter composite alphabet of $\boldsymbol{k = 2}$ using L1 norm inference. B. The 20 letter composite alphabet of $\boldsymbol{k = 3}$ using L1 norm inference. C+D. The same as A+B using KL divergence inference.



Supplementary Figure 15: The L1 norm inference for composite DNA letters. A. Successful decoding rate using the KL inference as a function of sequencing depth for the four composite DNA alphabets. The dashed line depicts the results using KL divergence. B. Inference rates for the different composite letters of resolution $\boldsymbol{k = 3}$ as a function of sequencing depth. The positions of the letter in the composite DNA oligo (starting from the 5' end) is stated at the top and the data for each letter is ordered by positioned.



Supplementary Figure 16: Design of the composite DNA oligo for the error analysis experiment. The locations of each of the 15 IUPAC letters (Y axis) along the oligo (X axis) is marked in yellow.



Supplementary Figure 17: Error analysis of larger composite alphabets. Inference rates for the different letters of the "IUPAC" alphabet as function of position in the composite DNA sequence (starting from the 5' end). The letters are colored according to their "native" alphabet. The black line represents a linear trend, excluding the four standard DNA letters and the single letter "N".

Supplementary Table 1 (separate file)

Physical density calculations of composite DNA storage. This includes the large scale experiment and the dilution experiment.

Supplementary Table 2 (separate file)

Logical density calculations of composite DNA storage. This include all the experiments, theoretical encodings and simulation experiments.

Supplementary Table 3 (separate file)

Oligo design for large alphabets experiment and error analysis.

Supplementary Table 4 (separate file)

Oligo design for the large scale composite DNA storage.

Supplementary Table 5 (separate file)

Oligo design for the simulations of large composite alphabet DNA storage.

Supplementary Table 6 (separate file)

Simulation results of large composite alphabet DNA storage.