Experimental Assessment of PCR Specificity and Copy Number for Reliable Data Retrieval in DNA Storage

Lee Organick[1*], Yuan-Jyue Chen[2], Siena Dumas Ang[2], Randolph Lopez[3], Karin Strauss[2,*], and Luis Ceze[1,*]

[1]University of Washington, Paul G. Allen School of Computer Science and Engineering, Seattle, 98195, USA
[2]Microsoft Research, Redmond, 98052, USA
[3]University of Washington, Department of Bioengineering, Seattle, 98195, USA
[*]leeorg@cs.washington.edu, kstrauss@microsoft.com , luisceze@cs.washington.edu

ABSTRACT

Synthetic DNA has been gaining momentum as a potential storage medium for archival data storage[1–9]. Digital information is translated into sequences of nucleotides and the resulting synthetic DNA strands are then stored for later individual file retrieval via PCR[7–9] (**Fig. 1a**). Using a previously presented encoding scheme[9] and new experiments, we demonstrate reliable file recovery when as few as 10 copies per sequence are stored, on average. This results in density of about 17 exabytes/g, nearly two orders of magnitude greater than prior work has shown[6]. Further, no prior work has experimentally demonstrated access to specific files in a pool more complex than approximately 10^6 unique DNA sequences[9], leaving the issue of accurate file retrieval at high data density and complexity unexamined. Here, we demonstrate successful PCR random access using three files of varying sizes in a complex pool of over 10^{10} unique sequences, with no evidence that we have begun to approach complexity limits. We further investigate the role of file size on successful data recovery, the effect of increasing sequencing coverage to aid file recovery, and whether DNA strands drop out of solution in a systematic manner. These findings substantiate the robustness of PCR as a random access mechanism in complex settings, and that the number of copies needed for data retrieval does not compromise density significantly.

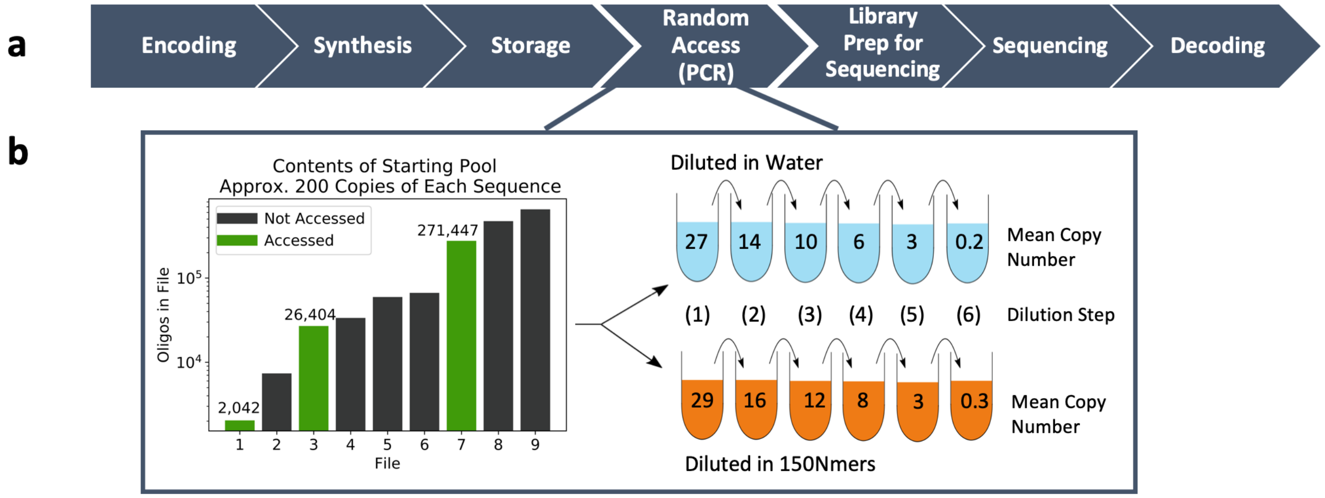
Main

For synthetic DNA data storage to become a viable alternative to electronic archiving, many unique DNA sequences must be physically storable in a single pool and then randomly and reliably accessed. Random access requires far fewer resources to recover data since only relevant files are sequenced and analyzed. Theoretically, for maximum density, only one copy of each sequence would be necessary to perform the Polymerase Chain Reaction (PCR) random access reaction. In practice, however, this is not the case for two reasons: stochastic variations in copy numbers that arise from sub-sampling the pool during random access, and copy number variations that arise from synthesis. Knowing the minimum copy number for each PCR reaction is crucial for storing DNA data; without it, one might store too few copies to access the data or too many, wasting orders of magnitude of density.

Previous work[1,2,6,9] recognized the importance of storage density for DNA to become a practical archival storage, but did not explore PCR random access accuracy when accessing extremely small subsets of data from a dense, complex pool, with the most complex random access performed on a pool comprised of just over 10^7 unique sequences[9]. Prior work also showed that thousands of unique files can be stored in a single pool with successful retrieval of desired files[9]; however, that pool contained approximately 10^4 unique sequences. This paper examines the ability of PCR to recover files from pools of far greater complexity, ranging from over 10^6 to over 10^{10} unique sequences per microliter (**Fig. 1b, 1c**). If PCR file retrieval fails in complex settings, we would observe a marked difference between sequences recovered in complex versus less complex settings. In addition, we might observe an inability to recover small files in complex conditions. Encouragingly, we observed neither of these symptoms.

In this work, we randomly accessed three files from a large pool of DNA at varying copy numbers (**Fig. 1b**). The small file, comprised of 2,042 sequences, represents approximately 0.1KB of digital data. The medium file consists of 26,404 sequences and 1.7KB of digital data, and the large file has 271,447 sequences and 18KB.

We then sequenced all three files at all stages of dilution to measure the rate of sequences lost (**Fig. 2a, 2b**). A copy number of 10, for example, means that, on average, each sequence was present 10 times in solution. The original, undiluted pool of



c

Dilution Step	Initial Pool	1	2	3	4	5	6
uL Prior Sample Used	n/a	11.0	57.9	63.8	40.0	19.0	5.0
uL Diluent Used	n/a	99.0	52.1	21.3	20.0	31.0	45.0
Water Diluent: Unique Sequences/µL	1.60E+06	1.60E+06	1.60E+06	1.60E+06	1.60E+06	1.60E+06	1.60E+06
150N Diluent: Unique Sequences/µL	1.60E+06	1.11E+10	1.17E+10	1.19E+10	1.20E+10	1.22E+10	1.23E+10
Smallest percent of pool accessed in a single PCR reaction	0.13	1.84E-05	1.74E-05	1.72E-05	1.70E-05	1.67E-05	1.65E-05

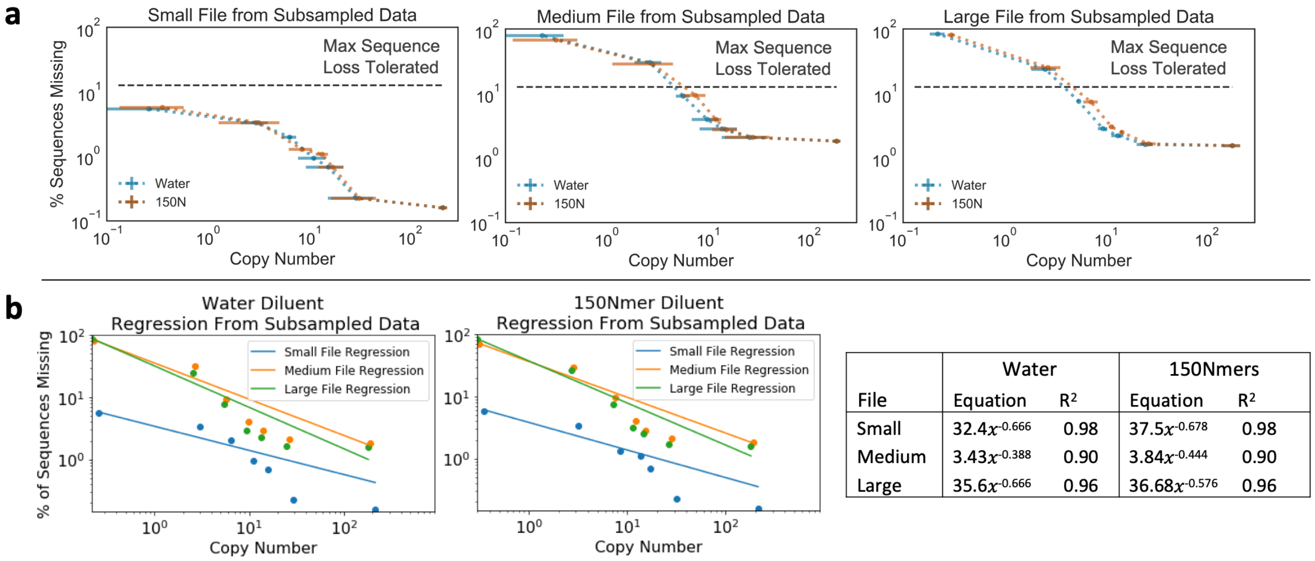
Figure 1. (**a**) A high-level representation of the DNA data storage pipeline. (**b**) (Left) The bar chart depicts contents of the initial, undiluted pool. (Right) The illustration shows the serial nature of subsequent dilutions. *Mean copy number* refers to the mean number of copies of each file's unique sequences as determined by qPCR (**Supplemental Section 1**). One serial dilution used water as the diluent in each step; the other used a solution of 150Nmers to dilute the pool to much greater complexity. (**c**) Details of how the samples were diluted. Note that the dilution steps were identical regardless of diluent. The smallest percent of pool accessed is calculated by dividing the size of the smallest file by the number of unique sequences in the 1 μL of solution used for PCR random access. This percentage refers to the 150N diluent pool since the small file in the water diluent pool is a constant 0.13%.

DNA encoded nine files (1.6M total oligonucleotides (oligos)) and was subsequently serially diluted with water to result in copy numbers ranging three orders of magnitude (**Fig. 1b**). To dramatically increase pool complexity, we repeated this process, this time diluting the samples with 1ng/μL 150Nmers, random sequences of DNA the same length as our original pool (**Fig. 1c**). Theoretically, we would expect to encounter a duplicate sequence only after 4^{150} ($2*10^{90}$) sequences had been synthesized; thus, we would expect each strand to be unique. By diluting samples with 1ng/μL 150Nmers, pool complexity increased from $1.6*10^6$ unique sequences in the initial pool (**Fig. 1b**) to over 10^{10} unique sequences per microliter (see **Supplemental Section 2**). At approximately 15.6 bytes of data encoded per strand[9], this encoding scheme and experimental protocol emulate approximately 200 GB of digital data per microliter, and nearly 9 TB in the final 50μL solution.

We found no distinguishable difference when comparing sequence loss between water and complex 150Nmer dilution conditions (**Fig. 2a**). For the large file, only three samples distinguishably differed for the two complexity conditions; this difference was negligible, with a mean difference of sequences missing of 0.98% and standard deviation of 1.82%.

Regardless of file accessed, decreasing the copy number yielded similar behavior (**Fig. 2**). The loss of sequences for all three files was best modeled with a power regression with relatively high accuracy (**Fig. 2b**). However, though the medium and large files behaved almost indistinguishably, the small file did not lose sequences at a similar rate after the copy number fell below approximately 10 (see **Fig. 2a**). Instead, it lost fewer sequences than the larger two files. This is likely due to a combination of copy number being slightly higher than calculated, fewer sequences initially missing due to variation in synthesis, and the distribution of sequences being slightly more uniform (see **Supplemental Section 3** for detailed analysis).

Encouragingly, the size of the file being recovered also does not impact the copy number required in complex pool conditions (**Fig. 2, Fig. 3a**). Regardless of pool complexity or size of file accessed, only 10 copies of each strand on average, with a standard deviation of 3, are required for successful recovery with no bit errors. A pool complexity emulating nearly 200 GB of digital data per PCR reaction did not hinder file recovery, and no evidence suggests that we are approaching the limit of



File	Water Equation	Water R²	150Nmers Equation	150Nmers R²
Small	$32.4x^{-0.666}$	0.98	$37.5x^{-0.678}$	0.98
Medium	$3.43x^{-0.388}$	0.90	$3.84x^{-0.444}$	0.90
Large	$35.6x^{-0.666}$	0.96	$36.68x^{-0.576}$	0.96

Figure 2. (**a**) Each plot illustrates a file's loss of sequences recovered at 20x coverage, directly comparing the samples diluted in water to those diluted in 150Nmers (150nt sequences comprised of random nucleotides). The threshold of the maximum number of sequences that can be lost while still permitting file recovery is plotted for reference, as determined by previous work[9]. Error bars represent 95% confidence intervals. X-axis errors are taken from triplicate qPCR data (see Methods), and y-axis errors are the result of 100 simulations of the original sequencing data sub-sampled to 20x sequencing coverage (see Methods). (**b**) Each plot illustrates behavioral similarities for each file in each diluent condition, with a power regression overlayed (see **Supplemental Section 4**). The data used here are also sub-sampled to a sequencing coverage of 20x.

pool complexity. Storing many unique sequences in one pool reduces the need for physical isolation, one of the largest density overheads facing this technology.

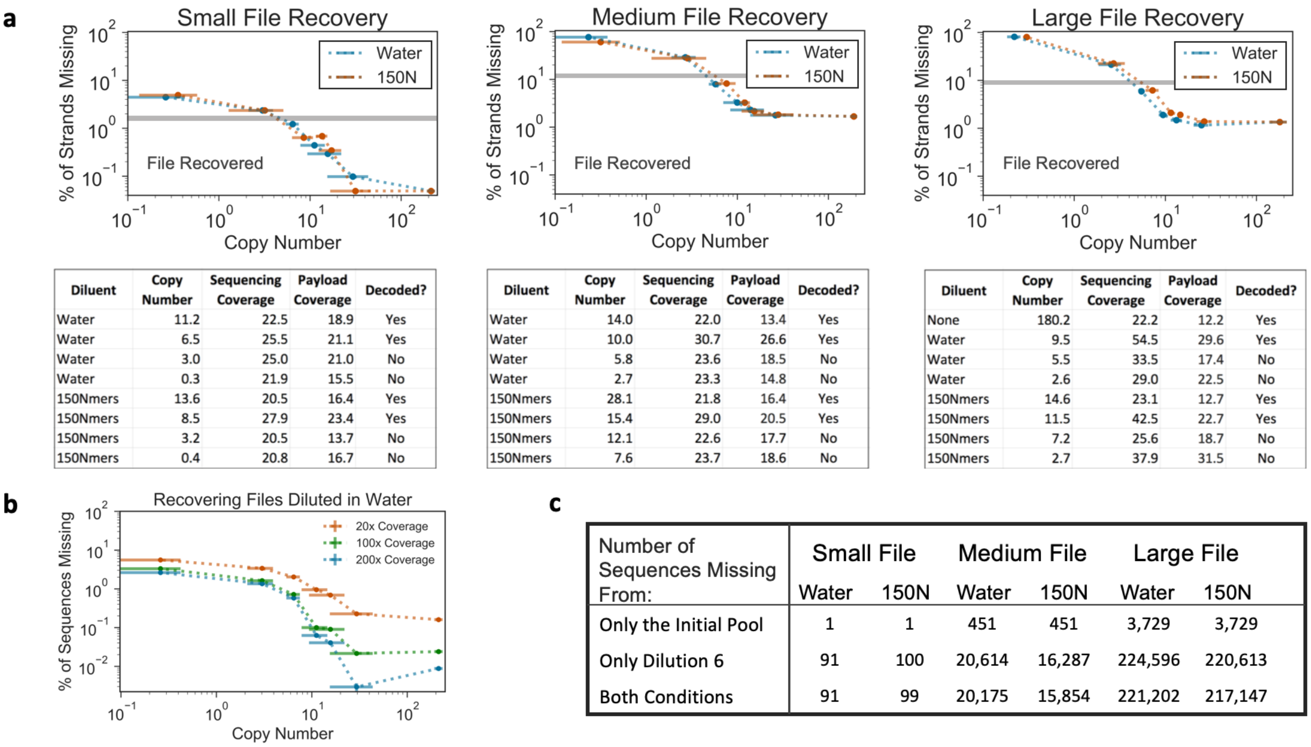
Determining the need for a minimum of approximately 10 copies of each DNA sequence in the PCR reaction to successfully recover a file enables us to calculate a density of 17 exabytes/gram (see **Supplemental Section 5**), nearly two orders of magnitude denser than prior work[6] and closer to the maximum theoretical density predicted for DNA data storage[1]. This is due to this particular encoding and decoding scheme's error tolerance, trading information density per DNA sequence for more efficient sequencing coverage[9] and increased physical density as a result of lower required copy number.

To further investigate the role of complex pools on sequence recovery, the behavior of each individual sequence was compared for both dilution conditions. We measured sequence behavior by examining the proportion of each sequence present in each dilution. Having a percentage that consistently changes for a subset of sequences would indicate that strands are being disproportionately accessed and amplified. We observed no difference in sequence behavior between the two different dilution conditions. Thus, sequences may be changing proportions or are observed to be missing due to stochastic variation in sub-sampling. Further supporting this hypothesis is the fact that, although most sequences absent from the initial pool are also absent from the most "stressed," low copy number pools, this does not happen to some of the sequences; this indicates that these sequences disappeared and reappeared due to stochastic effects rather than systematic interactions that made certain sequences irretrievable. See **Supplemental Section 6** for more detail and analysis.

This significant finding demonstrates the robustness of PCR itself and primer design methodology presented in prior work[9]. It assures users of this encoding scheme that we have not yet reached the limit of how complex a pool can be before it inhibits the ability to recover desired information.

While storage density is a crucial component of DNA data storage, sequencing efficiency is also critical due to its significant time and cost. We found that the samples successfully recovered at the lowest possible copy number had a mean sequencing coverage of 35x with a standard deviation 11x, with a mean copy number of 10 and a standard deviation of 3 when accounting for all files and all diluent conditions (**Fig. 3a**). Thus, we show that it is possible to successfully recover files with no bit errors using the encoding scheme presented in prior work[9] at high physical density and with excellent sequencing efficiency.

We next examined the degree to which simply sequencing more of the prepared sample aids file recovery. To do so, all water-diluted samples from the small file were sequenced a second time from the exact same post-library-preparation material, this time with a much higher sequencing coverage. Previously, all samples had a mean sequencing coverage of 24.5x with a standard deviation of 1.6x and were sub-sampled randomly with replacement to 20x coverage. Here, samples had a mean



Diluent	Copy Number	Sequencing Coverage	Payload Coverage	Decoded?
Water	11.2	22.5	18.9	Yes
Water	6.5	25.5	21.1	Yes
Water	3.0	25.0	21.0	No
Water	0.3	21.9	15.5	No
150Nmers	13.6	20.5	16.4	Yes
150Nmers	8.5	27.9	23.4	Yes
150Nmers	3.2	20.5	13.7	No
150Nmers	0.4	20.8	16.7	No

Diluent	Copy Number	Sequencing Coverage	Payload Coverage	Decoded?
Water	14.0	22.0	13.4	Yes
Water	10.0	30.7	26.6	Yes
Water	5.8	23.6	18.5	No
Water	2.7	23.3	14.8	No
150Nmers	28.1	21.8	16.4	Yes
150Nmers	15.4	29.0	20.5	Yes
150Nmers	12.1	22.6	17.7	No
150Nmers	7.6	23.7	18.6	No

Diluent	Copy Number	Sequencing Coverage	Payload Coverage	Decoded?
None	180.2	22.2	12.2	Yes
Water	9.5	54.5	29.6	Yes
Water	5.5	33.5	17.4	No
Water	2.6	29.0	22.5	No
150Nmers	14.6	23.1	12.7	Yes
150Nmers	11.5	42.5	22.7	Yes
150Nmers	7.2	25.6	18.7	No
150Nmers	2.7	37.9	31.5	No

Number of Sequences Missing From:	Small File		Medium File		Large File	
	Water	150N	Water	150N	Water	150N
Only the Initial Pool	1	1	451	451	3,729	3,729
Only Dilution 6	91	100	20,614	16,287	224,596	220,613
Both Conditions	91	99	20,175	15,854	221,202	217,147

Figure 3. **(a)** The results of decoding for each sample near the threshold for file recovery. Done post-sequencing, decoding involves clustering sequences, finding consensus, then correcting errors[9]. Sequencing coverage is determined by comparing alignment of actual sequences recovered to sequences expected by the reference template; it is the result of total alignments divided by sequences in the file. Decoding is blind to any references and is thus more vulnerable to losing sequences. **(b)** For the small file, when each sample diluted in water was sequenced at greater depth, minimal improvement on the proportion of missing sequences occurred. The 100x coverage data was found by sub-sampling the data used to create the 200x coverage data. **(c)** Missing sequences are compared between the initial pool (prior to any dilutions) and dilution 6 (the last dilution). The fact that some sequences are missing only from the initial pool but not the final dilution suggests that the lost sequences are a result of stochastic variation that occurs during sub-sampling for file recovery.

sequencing coverage of 549x with a standard deviation of 148x and were sub-sampled randomly to 200x coverage. This resulted in a mean of 1.8% fewer missing sequences for 200x coverage, with a standard deviation of 1.1% (**Fig. 3b**). This method of increasing sequencing coverage has the benefit of not requiring additional material from the original pool because there is extra material post library preparation to sequence many times over. However, to significantly improve levels of recovery, over an order of magnitude more sequencing resources must be used. Since increasing sequencing coverage significantly increases the cost of recovery, this process is useful only as a last resort.

In summary, as DNA data storage becomes an increasingly viable alternative to mainstream data storage methods, the ability to individually access very small portions of data from complex pools of DNA takes on great importance to ensure competitive data density. By showing that we can reliably access files that encode 0.1 KB, 1.7KB, and 18KB from a complex pool emulating nearly 200 GB of digital data per PCR random access reaction, we demonstrate the ability of this storage option to enable efficient, reliable data retrieval. We also provide the most accurate estimates to date of the physical redundancy necessary to ensure successful recovery of data encoded in DNA. We have no reason to think that we have reached the limit of how complex the pool could be or how small a file we can recover from such a pool, further supporting the robustness and unprecedented storage potential of DNA.

Methods

Dilution

The process of diluting the starting pool, once in water and once in 150Nmers (strands of DNA 150nt in length, with ‘N’ as the input when ordering from IDT to result in random sequences) was repeated twice to confirm consistency in qPCR behavior;

however, only one of the dilutions was sequenced, and those are the copy numbers reported throughout this paper. **Fig. 1c** details the volume of diluent and sample used for each dilution step. To minimize variation in copy number due to pipetting error, the pipette used to perform the dilutions for the two samples was not adjusted between uses.

qPCR Protocol

From all dilution samples, the same three files were amplified in triplicates via qPCR. Each qPCR reaction used an arbitrary ultramer from the relevant file to yield a standard curve. We ordered all ultramers from IDT. See **Supplemental Section 1** for primer and ultramer sequences as well as amplification efficiencies.

Each file was amplified using the following qPCR recipe: 1 μL of diluted pool, 0.5 μL of the appropriate forward primer, 0.5 μL of the appropriate reverse primer, 10 μL of 2x Kapa HiFi enzyme mix, 7 μL of molecular grade water, and 1 μL 20x Eva Green. The following qPCR protocol was used: (1) 95°C for 3 min, (2) 98°C for 20 s, (3) 62°C for 20 s, (4) 72°C for 15 s, (5) repeat steps 2-4 as needed.

Calculating Copy Numbers

For the large file, qPCR measurements were taken in triplicate, with an arbitrary ultramer from the file used as a custom qPCR standard (also measured in triplicate) to measure the number of copies of each oligo for every sample. The file's mean amplification efficiency and difference from its standard curve are detailed in **Supplemental Section 1**. This identical method of qPCR with a custom standard was used to calculate the first (undiluted) copy number for the medium and small files. However, only the first, undiluted sample could be quantified in this way; this was due to the difference in amplification efficiency between the small and medium files and their respective standards, likely the result of the low number of target strands leading to non-specific amplification (see **Supplemental Section 1**). A different method was thus used to determine copy numbers for the remaining samples associated with the small and medium files.

For the small and medium files, the diluted samples' copy numbers were calculated using the large file's dilution factors. Here, because diluting the pool dilutes all three files simultaneously, the dilution factor between subsequent large file samples was the same as it was between subsequent samples for the small and medium files. Thus, the dilution factor (DF) between each subsequent dilution was found for the large file's samples. The initial, undiluted sample's copy number for the small or medium file (CN_0) was then multiplied by the first (DF_1) to yield the copy number of the first dilution, $CN_1 = CN_0 * DF$, with the general formula:

$$CN_n = CN_{n-1} * DF_n$$

Calculating Margin of Error for Copy Numbers

For the large file, error is presented as the 95% confidence interval found from the variation of the triplicate qPCR reactions. This was calculated from the standard error for the triplicate qPCR, translated from cycle standard error to copy number standard error using the standard curve. Next, to find the 95% CI for each copy number, the standard error was multiplied by the appropriate t-table value with 2 degrees of freedom (4.303).

However, for the small and medium files, calculating the 95% CI entails incorporating both the error caused by initial qPCR measurement and that from the large file's observed dilution factors used to calculate subsequent copy numbers. This is because the equation used to calculate copy numbers is $CN_n = CN_{n-1} * DF_n$, where CN_{n-1} and DF_n both have a previously calculated 95% confidence interval.

To calculate the qPCR measurement error for each sample (δCN_n), the first undiluted sample's 95% CI was calculated in the same way as it was for the large file (detailed above). For each remaining sample, because qPCR data was not used and copy number was calculated with the large file's dilution factor, the prior dilution sample's copy number variation (δCN_{n-1}) was multiplied by the observed dilution factor (DF_n) as measured from the large file. Thus:

$$\delta CN_n = \delta CN_{n-1} * DF_n$$

To calculate the error that results from the large file's dilution factor measurements (δDF_n), we first found the greatest dilution factor that could have been calculated within the 95% confidence interval for calculated copy numbers by the following:

$$DF_{max} = \frac{CN_n + \delta CN_n}{CN_{n-1} - \delta CN_{n-1}}$$

Thus, the variation in copy number due to the dilution factor (δDF) was:

$$\delta DF_n = |CN_{n-1} * DF_{max} - CN_{n-1} * DF_n|$$

Finally, to determine the final 95% CI for small and medium files' diluted samples' copy numbers (δCN_n), the propagation equation incorporated both δCN_{n-1} and δDF_n and was represented by the following, easily solvable, standard error propagation equation:

$$\frac{\delta CN_n}{|CN_n|} = \sqrt{\left(\frac{\delta CN_{n-1}}{CN_{n-1}}\right)^2 + \left(\frac{\delta DF_{n-1}}{DF_{n-1}}\right)^2}$$

Library Preparation and Sequencing

All files were amplified using the following recipe: 1 μL of sample, 0.5 μL of the appropriate forward primer, 0.5 μL of the appropriate reverse primer, 10 μL of 2x Kapa HiFi enzyme mix, and 8 μL molecular grade water. The following PCR protocol was used: (1) 95°C for 3 min, (2) 98°C for 20 s, (3) 62°C for 20 s, (4) 72°C for 15 s, (5) repeat steps 2-4 as needed according to prior qPCR.

See methods for NGS library preparation via ligation and sequencing presented by Organick et al[9].

Sub-sampling Data to Calculate Percent of Sequences Missing

To remove the effect of varying sequencing coverage, aligned sequences were randomly sub-sampled with replacement down to 20x coverage (*20x coverage = number of oligos in file * 20*), and the resulting number of missing strands was recorded. This was performed 100 times per file per sample to yield mean percent sequences missing and the 95% confidence interval.

References

1. Church, G. M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in DNA. *Sci.* **337**, 1628 (2012). URL http://science.sciencemag.org/content/337/6102/1628. DOI 10.1126/science.1226355.
2. Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nat.* **494**, 77–80 (2013). URL http://dx.doi.org/10.1038/nature11875https://www.nature.com/articles/nature11875. DOI 10.1038/nature11875.
3. Yazdi, S. M. H. T., Yuan, Y., Ma, J. & Zhao, H. A Rewritable , Random-Access DNA-Based Storage System. *Nat. Publ. Group* 1–10 (2015). URL http://dx.doi.org/10.1038/srep14138https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4585656/{#}S1. DOI 10.1038/srep14138.
4. Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie - Int. Ed.* **54**, 2552–2555 (2015). DOI 10.1002/anie.201411378.
5. Blawat, M. *et al.* Forward Error Correction for DNA Data Storage. *Procedia Comput. Sci.* **80**, 1011–1022 (2016). URL http://www.sciencedirect.com/science/article/pii/S1877050916308742. DOI https://doi.org/10.1016/j.procs.2016.05.398.
6. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Sci.* **355**, 950–954 (2017). DOI 10.1126/science.aaj2038.
7. Bornholt, J. *et al.* Toward a Dna-Based Archival Storage System Storing Data in Dna Molecules Offers Extreme Density and Durability. *ASPLOS* (2016).
8. Yazdi, S. M. H. T., Gabrys, R. & Milenkovic, O. Portable and Error-Free DNA-Based Data Storage. *Sci. Reports* **7**, 5011 (2017). URL https://doi.org/10.1038/s41598-017-05188-1. DOI 10.1038/s41598-017-05188-1.
9. Organick, L. *et al.* Random access in large-scale DNA data storage. *Nat. Biotechnol.* (2018). URL http://www.nature.com/doifinder/10.1038/nbt.4079. DOI 10.1038/nbt.4079.

Acknowledgements

The authors would like to thank Leila Zelnick for her help reviewing statistics, Jennifer Rogers and Nicholas Nuechterlein for their insight on modeling data, Christopher Takahashi for his help finding bugs in code, and Xiaomen Liu for his assistance in the lab.

Author Contributions Statement

L.O. conceived of, designed, performed, and analyzed experiments. Y-J.C. and R.L. helped designed experiments and analyze data. S.D.A. helped design sequencing error analysis methods and performed sequencing analysis. K.S. and L.C. directed and supervised the work.

Additional information

Competing financial interests Authors Y-J.C., S.D.A., and K.S. are currently employed by Microsoft.