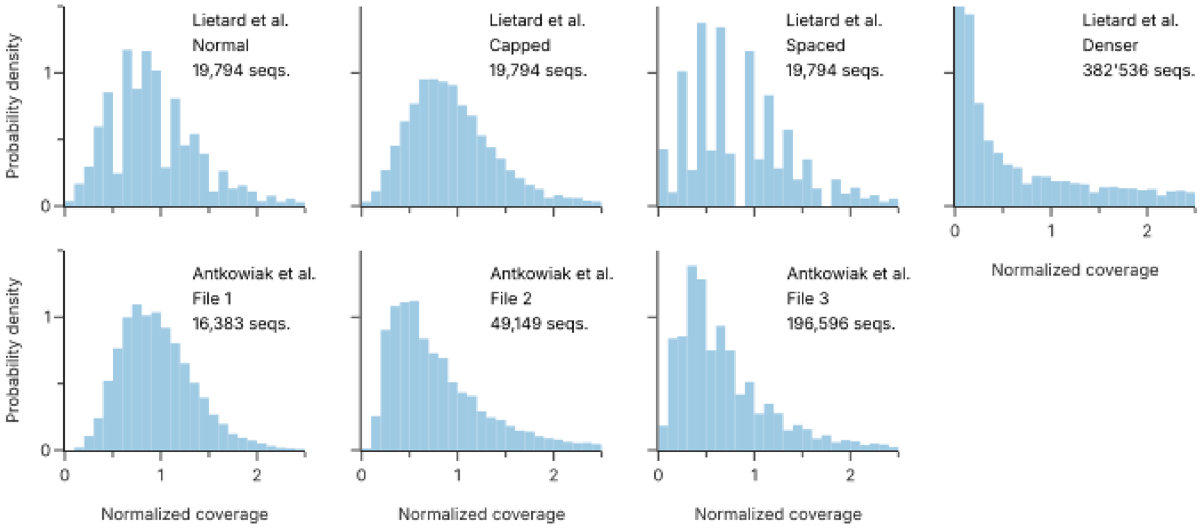
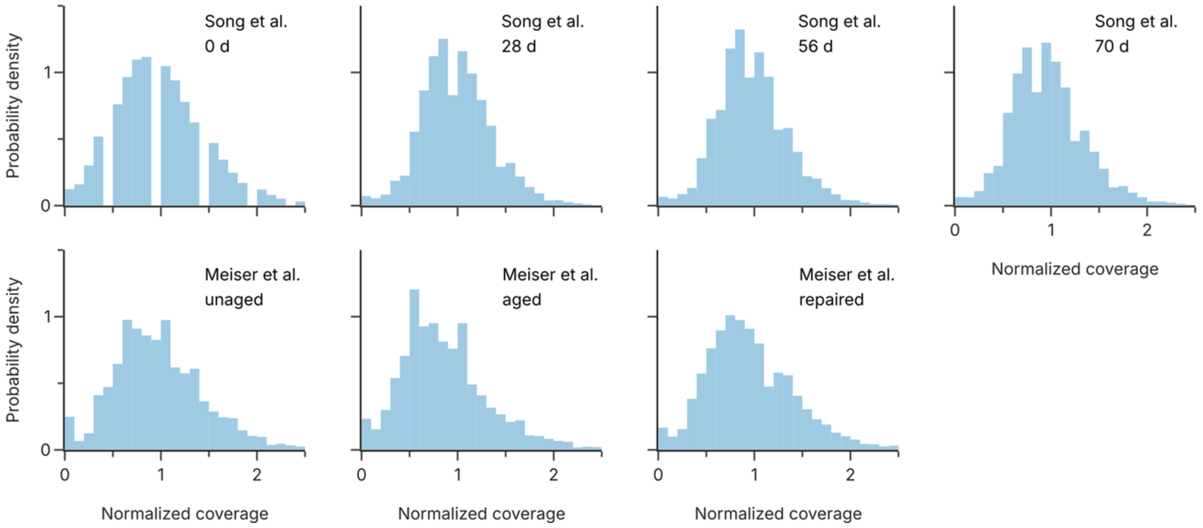
Supplementary Notes

Supplementary Note 1: Breakage patterns in unaged datasets by Meiser et al.[1] and Song et al.[2] The comparison in Fig. 4b and the main text between the aged datasets by Meiser et al.[1] and Song et al.[2], and the reference hydrolysis data by Shapiro[3] is based on the assumption that hydrolysis-induced decay – if it has occurred – will lead to similar base bias patterns. In the case of the unaged datasets however, no hydrolysis-induced decay should have occurred yet, therefore the base bias observed for these datasets is likely due to minor breakage during synthesis or sequencing preparation. This is supported for the unaged dataset by Song et al.[2] by the presence of only 11% broken reads (see Fig. 3c), whereas 89% of reads are full-length, indicating no significant decay has taken place. A similar pattern would also be expected from the unaged data by Meiser et al.[1], however a base bias that is similar to the distribution expected from hydrolysis rates is observed instead. We believe this to have occurred because the sample used by Meiser et al.[1] experienced (hydrolysis-induced) decay prior to the reported experiments on decay. This is supported by the fact that the unaged dataset only contains 48% full-length reads (vs. 89% for Song et al.[2], see Fig. 3c) and has higher error rates to begin with (see Fig. 3a). Possible reasons for this could be insufficient cooling during transport and/or extended storage at elevated temperatures prior to the start of experiments.

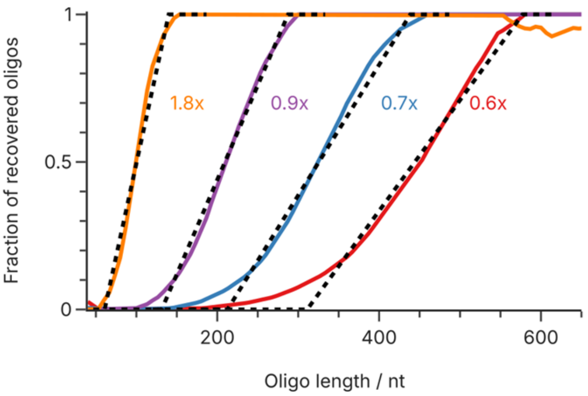
Supplementary Figures



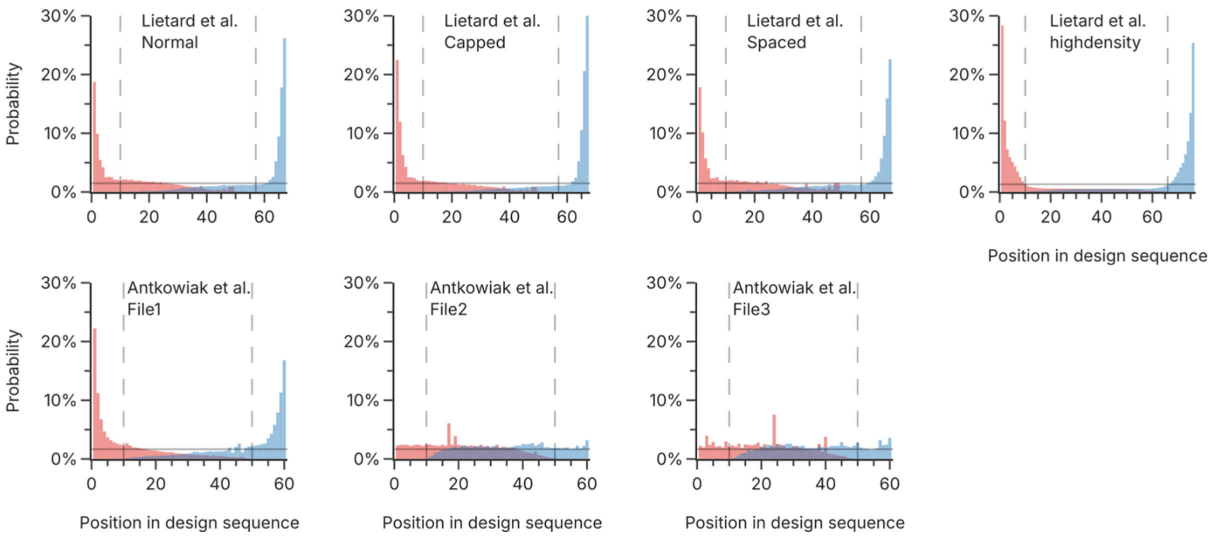
Supplementary Fig. 1: Normalized sequences coverage for photolithographic syntheses. The homogeneity of the oligo pools synthesized by Lietard et al.[4] (top row) and Antkowiak et al.[5] (bottom row) shows an increase in skewness and bias with the number of sequences synthesized in parallel. Due to the low sequencing depth, especially in the datasets by Lietard et al.[4], the data is non-continuous and therefore the histograms appear non-uniform.



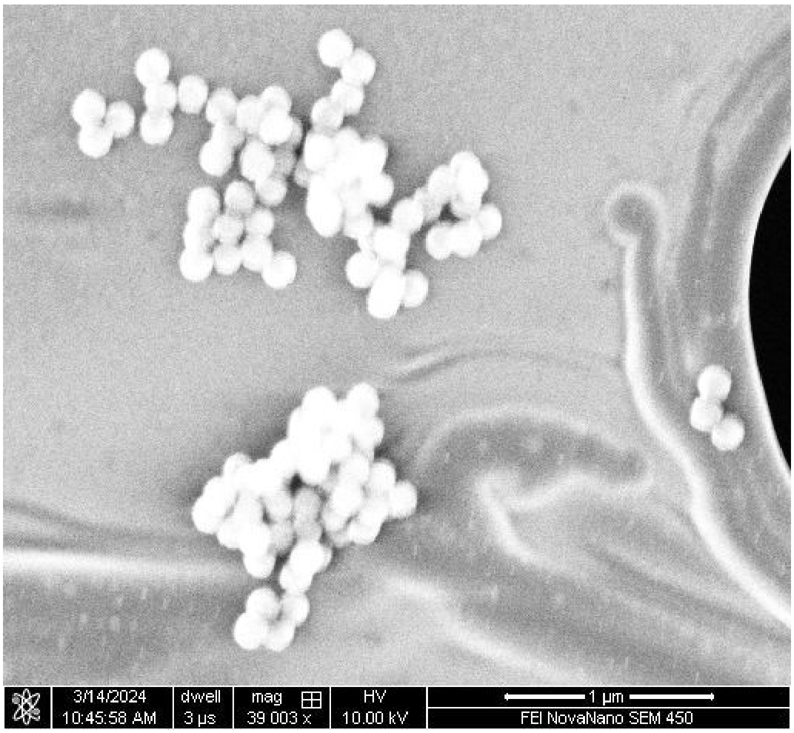
Supplementary Fig. 2: Normalized sequences coverage for DNA decay. The homogeneity of the oligo pools synthesized by Song et al.[2] (top row) and Meiser et al.[1] (bottom row) shows no increase in skewness and bias with aging.



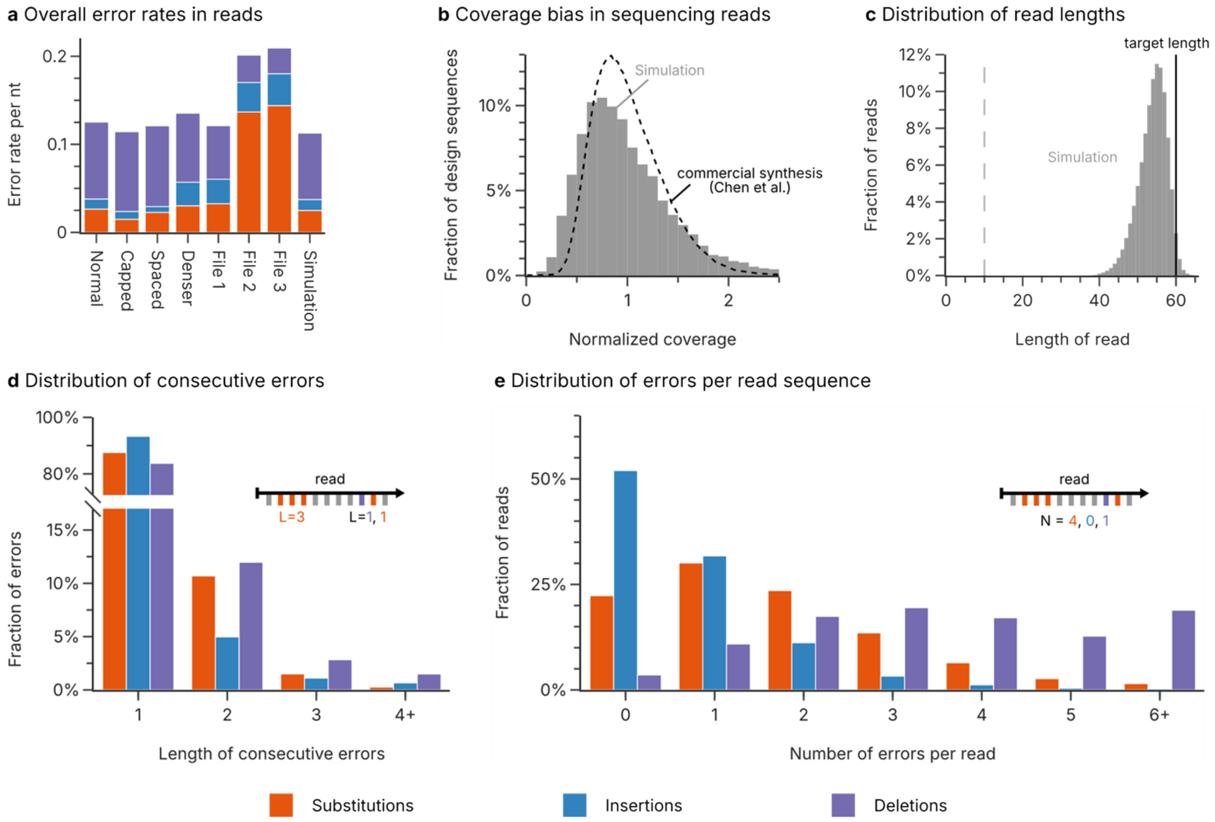
Supplementary Fig. 3: Recovery rate during bead-based clean-up as a function of oligo length and bead ratio. The oligo recovery workflows by Song et al.[2] and Meiser et al.[1] use bead-based clean-up to remove excess adapters after ligation. At different bead ratios (1.8x to 0.6x), the proportion of oligos recovered depends strongly on their individual lengths. This relationship is reasonably well approximated by a piecewise linear function (dotted lines) with a lower cut-off and upper threshold length. The raw data is based on the experimental data provided by Beckman Coulter for AMPure XP beads.[6]



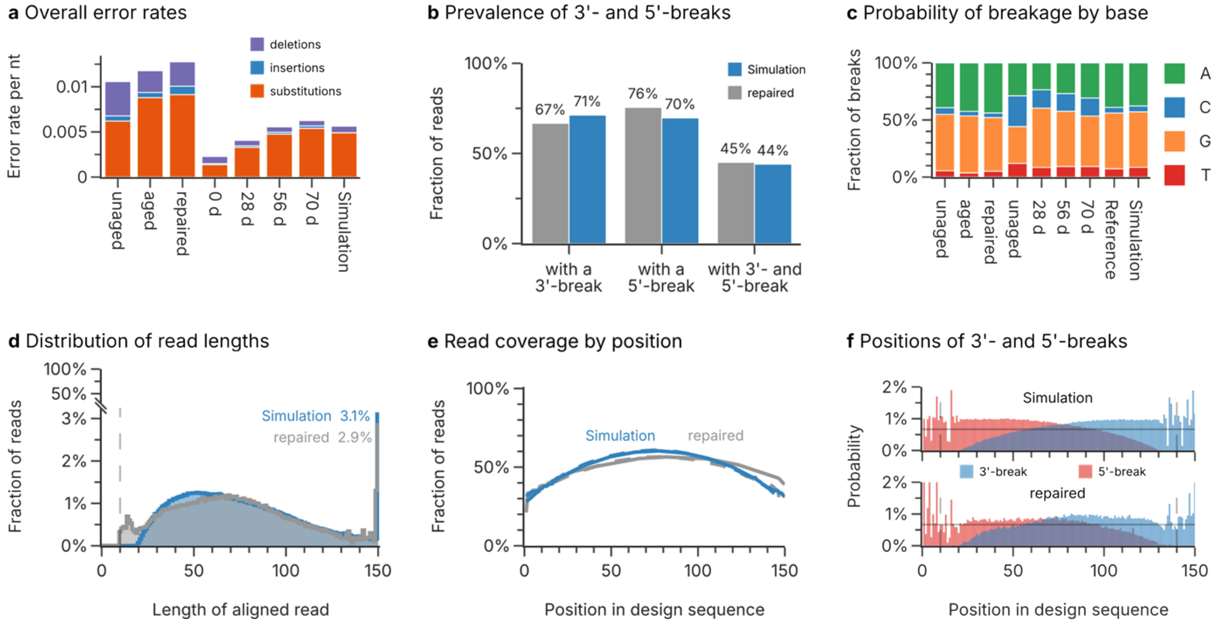
Supplementary Fig. 4: Positional distribution of 3'- (blue) and 5'-ends (red) for photolithographic syntheses. The fragments in the sequencing data by Lietard et al.[4] (top row) and Antkowiak et al.[5] (bottom row) show a clear trend towards truncation at both the 3'- and 5'-ends, as long as 10 nt at each end. The datasets of File 2 and File 3 of Antkowiak et al.[5] do not follow this trend, as the positional distribution is even along the design sequence. Dotted lines indicate the filtering of short sequences by the analysis pipeline (i.e., reads shorter than 10 nt are removed). The solid horizontal line indicates an equal probability along the design sequence.



Supplementary Fig. 5: SEM micrograph of silica nanoparticles encapsulating DNA. This image, compressed into a zip file of 51,632 bytes, was used as input for the codecs benchmarked for the two challenges described in the main text.



Supplementary Fig. 6: Analysis of photolithographic synthesis as simulated by Challenge 1. (a) Overall rate of substitution (orange), insertion (blue), and deletion (purple) errors in the sequencing datasets for photolithographic syntheses by Lietard et al.[4], Antkowiak et al.[5] and the simulation. The error rates represent the median error rate across the length of the sequence, to minimize the effect of the low-diversity regions at the start and end of the sequences. **(b)** Distribution of the sequence coverage in the simulated dataset compared to the coverage distribution for a commercial synthesis by material deposition (Twist Biosciences, data by Chen et al.[7]). **(c)** Length distribution of the reads in the simulated dataset compared to the length of the design sequences (solid line). Only the segment of the read which aligned to the reference sequence is considered. Reads smaller than 10 nucleotides were discarded during mapping of the sequencing data (dotted line). **(d)** The length of consecutive of substitution (orange), insertion (blue), and deletion (purple) errors in the simulated sequencing dataset. **(e)** The number of substitution (orange), insertion (blue), and deletion (purple) errors in the sequencing reads in the simulated dataset. The data in this figure can be compared to Fig. 2 in the main text to validate the accuracy of the digital model of photolithographic synthesis.



Supplementary Fig. 7: Analysis of DNA decay during aging as simulated by Challenge 2. (a) Overall rate of substitution (orange), insertion (blue), and deletion (purple) errors in the sequencing datasets by Meiser et al.[1], Song et al.[2], and the simulation. **(b)** Fraction of reads with different break patterns in the repaired dataset by Meiser et al.[1] (gray) and the simulated dataset (blue). **(c)** Distribution of nucleobases (green: A, blue: C, orange: G, red: T) involved in strand cleavage in the sequencing datasets by Meiser et al.[1], Song et al.[2], the simulation, as well as the theoretical distribution expected from reaction rates.[3] Experimental distributions are estimated from the type of nucleobase preceding a 5'-break in the sequencing data. **(d)** Histograms of the length distribution of aligned sequencing reads in the repaired dataset by Meiser et al.[1] (gray) and the simulated dataset (blue). The fraction of reads with full length in each dataset are given as a percentage. **(e)** Read coverage by position in the design sequence of the repaired dataset by Meiser et al.[1] (gray) and the simulation (blue). The reads are separated by their direction into sense (solid lines) and antisense (dashed lines) directions. A value of 100% corresponds to every read containing a specific position. **(f)** Positional distributions of 3'- (blue) and 5'-breaks (red) in the repaired dataset by Meiser et al.[1] (bottom) and the simulated dataset(top). The horizontal solid line denotes the expected breakage probability if decay occurs uniformly. The data in this figure can be compared to Figs. 3+4 in the main text to validate the accuracy of the digital model of DNA decay during long-term storage.

Supplementary Tables

Supplementary Table 1: Settings for the DNA-RS codec by Antkowiak et al.[5] used in benchmarking.

Challenge	Code rate	Length	Redundancy	# Sequences	Success
Photolithographic Synthesis	0.80 bit nt^{-1}	72	6	7200	0/3
	0.70 bit nt^{-1}	72	6	8200	0/3
	0.60 bit nt^{-1}	72	6	9500	3/3
	0.50 bit nt^{-1}	72	6	11500	3/3

Supplementary Table 2: Settings for the DBGPS codec by Song et al.[2] used in benchmarking.

Challenge	Code rate	Chunk size	# Droplets	Success
DNA Decay	1.40 bit nt^{-1}	35	1800	0/3
	1.30 bit nt^{-1}	35	1930	0/3
	1.20 bit nt^{-1}	35	2100	1/3
	1.10 bit nt^{-1}	35	2300	3/3

Supplementary Table 3: Settings for the DNA-Fountain codec by Erlich et al.[8] used in benchmarking.

Challenge	Code rate	Chunk size	Redundancy factor	Success
Photolithographic synthesis	0.50 bit nt^{-1}	14	1.8	0/3
	0.40 bit nt^{-1}	14	2.5	0/3
	0.30 bit nt^{-1}	14	3.6	0/3
	0.20 bit nt^{-1}	14	5.9	0/3
DNA Decay	0.50 bit nt^{-1}	32	2.4	0/3
	0.40 bit nt^{-1}	32	3.2	0/3
	0.30 bit nt^{-1}	32	4.6	0/3
	0.20 bit nt^{-1}	32	7.5	0/3

References

1 L. C. Meiser, A. L. Gimpel, T. Deshpande, G. Libort, W. D. Chen, R. Heckel, B. H. Nguyen, K. Strauss, W. J. Stark and R. N. Grass, *Commun. Biol.*, 2022, **5**, 1–9.

2 L. Song, F. Geng, Z.-Y. Gong, X. Chen, J. Tang, C. Gong, L. Zhou, R. Xia, M.-Z. Han, J.-Y. Xu, B.-Z. Li and Y.-J. Yuan, *Nat. Commun.*, 2022, **13**, 1–9.

3 R. Shapiro, in *Chromosome Damage and Repair*, eds. E. Seeberg and K. Kleppe, Springer US, New York, NY, 1981, pp. 3–18.

4 J. Lietard, A. Leger, Y. Erlich, N. Sadowski, W. Timp and M. M. Somoza, *Nucleic Acids Res.*, 2021, **49**, 6687–6701.

5 P. L. Antkowiak, J. Lietard, M. Z. Darestani, M. M. Somoza, W. J. Stark, R. Heckel and R. N. Grass, *Nat. Commun.*, 2020, **11**, 5345.

6 Beckman Coulter, AMPure XP: Manual or Automated Purification and Clean-up, Document #AAG-4464DS12.18 (2019).

7 Y.-J. Chen, C. N. Takahashi, L. Organick, C. Bee, S. D. Ang, P. Weiss, B. Peck, G. Seelig, L. Ceze and K. Strauss, *Nat. Commun.*, 2020, **11**, 1–9.

8 Y. Erlich and D. Zielinski, *Science*, 2017, **355**, 950–954.