Information Decay and Enzymatic Information Recovery for DNA Data Storage

Linda C. Meiser[1], Andreas L. Gimpel[1], Tejas Deshpande[1], Gabriela Libort[1], Weida D. Chen[1], Reinhard Heckel[2], Bichlien H. Nguyen[3], Karin Strauss[3], Wendelin J. Stark[1], Robert N. Grass[1,]*

Affiliations:

[1] Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Vladimir-Prelog-Weg 1, CH-8093 Zurich, Switzerland

[2] Department of Electrical and Computer Engineering, Technical University of Munich, Arcistrasse 21, 80333 Munich, Germany

[3] Microsoft Research, Redmond, WA 98052, USA

*to whom correspondence should be addressed rograss@ethz.ch

ABSTRACT

Synthetic DNA has been proposed as a storage medium for digital information due to its high theoretical storage density and anticipated long storage horizons. However, under all ambient storage conditions, DNA undergoes a slow chemical decay process resulting in nicked (broken) DNA strands, and the information stored in these strands is no longer readable. In this work we design an enzymatic repair procedure, which is applicable to the DNA pool prior to readout and can partially reverse the damage. Through a chemical understanding of the decay process, an overhang at the 3' end of the damaged site is identified as obstructive to repair via the base excision-repair (BER) mechanism. The obstruction can be removed via the enzyme apurinic/apyrimidinic endonuclease I (APE1), thereby enabling repair of hydrolytically damaged DNA via Bst polymerase and Taq ligase. Simulations of damage and repair reveal the benefit of the enzymatic repair step for DNA data storage, especially when data is stored in DNA at high storage densities (= low physical redundancy) and for long time durations.

Introduction

Genomic DNA, which is present in all living organisms, can be damaged by extrinsic and intrinsic agents, including factors such as exposure to sunlight, oxidation or hydrolysis and lesions in the DNA strands.[1,2] Although natural proofreading mechanisms are in place, some of these lesions and errors may remain unattended, which, if not repaired, can be mutagenic.[3]

Three basic cellular defense mechanisms, namely direct repair, base excision-repair (BER), and nucleotide excision-repair (NER), are known, in which specific enzymes repair spontaneous DNA lesions. These lesions could be caused by sunlight, oxidation, hydrolysis or exposure to small molecules. In the absence of large quantities of external DNA-damaging agents, most DNA lesions are repaired by the base excision-repair pathway, which has been reconstituted by Dianov and Lindahl in 1994,[2] using purified enzymes, for which Lindahl jointly received the Nobel Prize in Chemistry in 2015[4].

Since the identification of enzymes involved in the base excision repair pathway,[2,3,5–7] repair enzymes restoring the quality of degraded DNA have not only been used for understanding genomic DNA repair, but are used in standard molecular biology routines and have also been proposed for improving the analysis of ancient DNA, the genotyping of forensic samples as well as the tracing of foods, just to mention a few.[8–11] One application of enzymatic repair that has not yet received any attention, however, is the repair of synthetic DNA in DNA data storage applications.

For DNA data storage digital files (bits) are translated into nucleotides, which can then be synthesized and stored for thousands of years.[12–17] Due to constraints in synthesizing and reading very long DNA strands, data is usually stored in a pool of relatively short (ca. 150 nt) oligos.[18] Aside of encoding for actual data, every oligo also comprises an index for information organization and amplification primers, which enable random access[17,19] and the handling of the oligos for readout (sequencing preparation). The technology usually uses error correction codes, which calculate and append redundant information to the original data, so that read/write and storage errors can be compensated for.[14,17,18] Such error correcting codes may thereby enable the correction of individual base errors (e.g. mutations) within a sequence (inner code), as well as the loss of complete sequences (outer code). While such error correcting codes enable the perfect recovery of stored data, they come at the cost of having to write (=synthesize) more nucleotides than would be necessary in the absence of errors. DNA synthesis is currently the bottleneck of the more widespread adoption of DNA data storage and consequently synthesis should be minimized. Our specific interest in terms of minimizing errors and therefore necessary data redundancy lies in the actual DNA storage process itself.

In this work, we analyze the decay process of synthetic DNA, which is expected to occur during a DNA data storage scenario on a molecular level. We identify DNA nicking as the process resulting in the greatest data loss and propose enzymatic repair for a potential solution to reverse this data loss. This solution presents a process for information recovery for synthetic DNA in data storage applications.

During storage DNA may be exposed to several stresses including UV irradiation, oxidation, hydrolysis, alkylation, ionizing radiation or mechanical shear,[20] but independent of the actual storage format, hydrolysis has been identified as a major decay reaction.[14,21] This is due to the near omnipresence of atmospheric humidity, and the relatively high reaction rates of DNA with water as opposed to oxygen. In addition, water is the standard medium for all DNA handling steps, as DNA is not soluble in any other solvent than water.

In terms of DNA decay, two decay modes are of special interest: the mutation of individual bases, and the formation of nicks (=strand breaks) of the DNA backbone. While the nature of mutations has been described in the past,[22] DNA nicks are especially problematic in the view of information readout, as nicked DNA cannot be amplified via polymerase chain reaction (PCR), a procedure typically employed during random access and DNA readout (sequencing preparation) routines (Fig 1c). Consequently, any DNA strand comprising at least one single strand break is not amplified, and is thereby not read. In the context of information integrity, a mutation can result in the loss of up to 2 bits of information, and a nick can result in upwards of 100 bits of data lost (assuming that the sequences are of length more than 50, which they typically are). While these losses can be mitigated by error correction codes, the cost imbalance of mutations vs. nicks remains: The cost for error-correction codes to correct for mutated bases is relatively low: One substitution error requires at least two symbols of redundancy, and one erasure (missing nucleotide) requires at least one symbol of redundancy (one symbol can vary in size, a common choice is 3 nucleotides per symbol).[18,23] The loss of a whole sequence (as a consequence of a single DNA nick), however, is much more expensive, as the whole information of the sequence is lost. From an error correction standpoint this is expensive, as the correction of every nick requires at least one full redundant DNA strand. Specifically, a nick has an expected cost that is about by half of the length of the sequence higher than the cost of a mutation. In practice, the cost of a nick is therefore typically by a factor of 25-100 higher than that of a mutation. As a result, we were interested in investigating the nicking of DNA during storage.

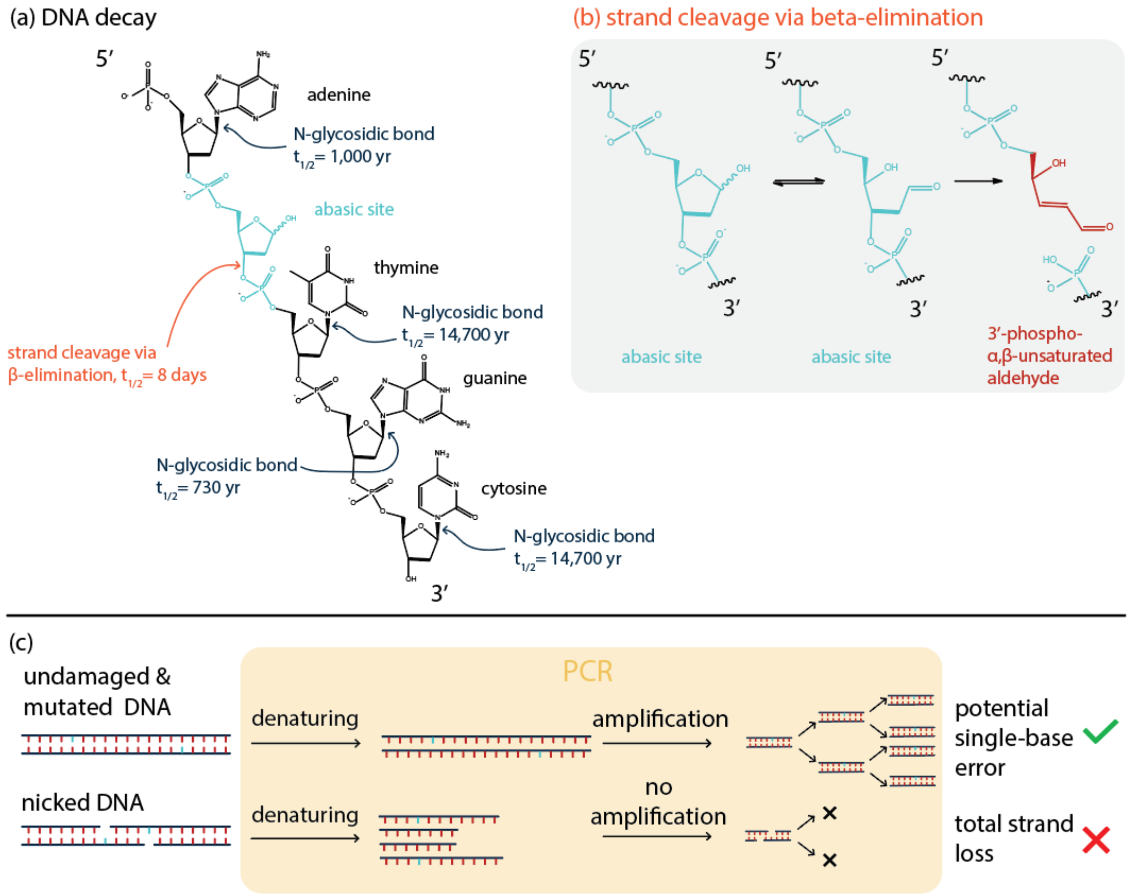

Figure 1: Decay pathways during DNA data storage. *(a) Hydrolysis can lead to the release of bases with rates as indicated. Once an abasic site exists, strand cleavage proceeds via β-elimination. (b) Mechanism of β-elimination resulting in a 3'-phospho-α,β-unsaturated aldehyde.*[24] *(c) Denaturation of undamaged/mutated DNA and nicked DNA resulting in potential single-base loss and total strand loss, respectively.*

Results

We thus investigated the damage of DNA that occurs hydrolytically, by looking at two different sets of DNA with different file sizes (Supplementary Information). We naturally aged the DNA in water, at 25 °C and 30 °C for up to 40 days and quantified the amount of amplifiable DNA via qPCR. While various DNA preservation methods would enable slower hydrolysis due to the partial protection of the DNA from water, it may be expected that the major decay promoting step would still encompass the same hydrolysis reactions, but at a significantly lower rate than in aqueous solution.

To investigate the nicking process, we measured the fragmentation of aged DNA by ssDNA sequencing using Illumina's iSeq100. The slightly unusual sample preparation procedure allowed for analysis of nicked ssDNA fragments.[25,26] In contrast to traditional amplicon sequencing applied in DNA data storage, this preparation procedure does not require PCR as a first step and enables the reading of sequences of incomplete lengths (=fragments). For analysis, the single strand DNA sequences were aligned. This ensures that synthesis or sequencing errors do not affect the fragment analysis, and the nicking process can be directly evaluated by looking at the fragment size distribution. Figure 2 shows the fragment size distribution for a file encoding a 115 kB Jpg image, consisting of 7,373 unique DNA strands for the undamaged file (Fig. 2a) and the aged file (Fig. 2b). Per design, the full length of each sequence is 150 nucleotides. We observe that nearly 23% of alignments are of the correct strand length in the file of undamaged DNA, whereas following aging of the sample, only 0.7% of alignments are of the correct strand length. The fragment-size distribution varies as expected: The aged file contains many small fragments between fragment lengths of 50-70 nucleotides and no prevalence of a specific fragment length. This finding supports the general hypothesis that under well controlled conditions, DNA decay is a random process.[27–30]

Fragmentation results of DNA aged at 30 °C show that nicking the sequence length by one half, approximately. However, no matter how long the resulting fragments are, it may be safely assumed that in a standard sample preparation method involving PCR as initial amplification step,[12–14,17] these fragments would not amplify.

DNA hydrolysis has been firstly investigated by Lindahl[3] showing that guanine and adenine are released from DNA approximately 20 times faster than cytosine and thymine, without a significant effect from whether the DNA is single-stranded or double-stranded.[3] Once the base is lost (leading to an abasic site in the DNA stand), cleavage via β-elimination is expected to proceed with a half-life of about eight days, leaving the phosphate backbone nicked.[21] Figure 1a summarizes literature DNA decay rates including the rates of release of bases to form apurinic and apyrimidinic sites, and rates of strand cleavage via β-elimination.[3,21] Once an abasic site exists, spontaneous β-elimination proceeds under physiological conditions, resulting in a DNA nick. Even in the absence of water, DNA nicking as a result of an acid-catalyzed and water-independent base loss and subsequent strand breakage has been reported.[31]

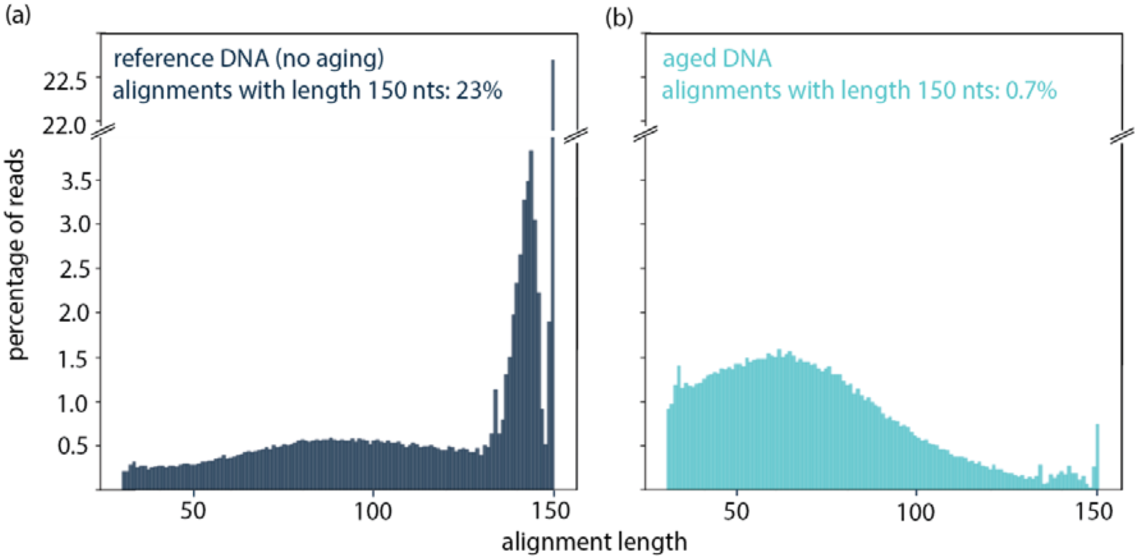

Figure 2: Fragment length distribution analysis. *Fragmentation of DNA file (originally containing 7,373 unique DNA sequences of length 150 nucleotides). Fragmentation analysis has been conducted by dsDNA denaturation, enzymatic single strand library preparation and sequencing using Illumina's iSeq100. Subsequently, alignment was performed to obtain the fragment length distribution, excluding sequences with < 30 nt lenght. (a) reference DNA that was not aged. (b) sample aged at 30 °C for 6 weeks.*

Enzyme selection

As already discussed previously, nature has pathways to repair various types of damage enzymatically. Although damage caused by hydrolysis can be of different chemical morphologies (i.e., resulting in abasic sites, nicks, deaminated cytosine, or fragmentation),[3] our primary interest was to develop a targeted enzymatic repair pathway with the potential to repair the nicks observed in the previous section of this work and thus to ensure that as many DNA strands as possible can be amplified during PCR.

From a most primitive view, one could imagine that a nick comprising a single base loss (see e.g. Fig 1C) could be resolved by the simple use of a polymerase and a ligase. The polymerase would replace the missing base, and the ligase would ligate the two portions of the strand. This works very well in many contexts of molecular biology (e.g. Gibson assembly[32]) However, our experiments using such enzyme mixes were not useful to reassembly the nicks caused by hydrolysis. The reason for this can be found if the chemical mechanism of DNA hydrolysis is more closely investigated: After strand cleavage a 3'-phospho-α,β-unsaturated aldehyde is present in the phosphate backbone[24] (Fig. 1b), which are known to obstruct the base-excision repair pathway.[33,34]

There is, however, an enzyme that allows removing the obstructive 3’-phospho-α,β-unsaturated aldehyde. This enzyme is called apurinic/apyrimidinic endonuclease I (APE1) and is most commonly known for its function to incise the phosphate backbone with its endonuclease activity.[3435] However, APE1 has a functionality to cleave off a 3’phospho-α,β-unsaturated aldehyde (the obstructive group present after DNA nicking) as well, leaving an 3’-OH terminal. This terminal should then be subsequently repaired by the standard base-excision repair pathway using polymerase and ligase.[2,36]

To test the hypothesis that APE1 can remove 3’-phospho-α,β-unsaturated aldehyde groups, we investigated the requirement of the three enzymes (APE1, polymerase, ligase) by individually removing each of the enzymes and quantifying repair using PCR (Fig. 3a), showing that repair mixes without APE1 do not seal the nick in DNA strands. This indicates, that APE1 is indeed required for removing the obstructive end after nicking. We tested various different polymerases and ligases, as well as the addition of other enzymes, combinations of different enzymes of the same type (i.e., a combination of different endonucleases, polymerases or ligases) and observed optimal repair activity with Bst polymerase and Taq ligase in combination with APE1 treatment, as quantified by PCR (Supplementary Information). From our experimental findings, we devised a molecular repair path to seal nicks in DNA strands (Fig. 3b).

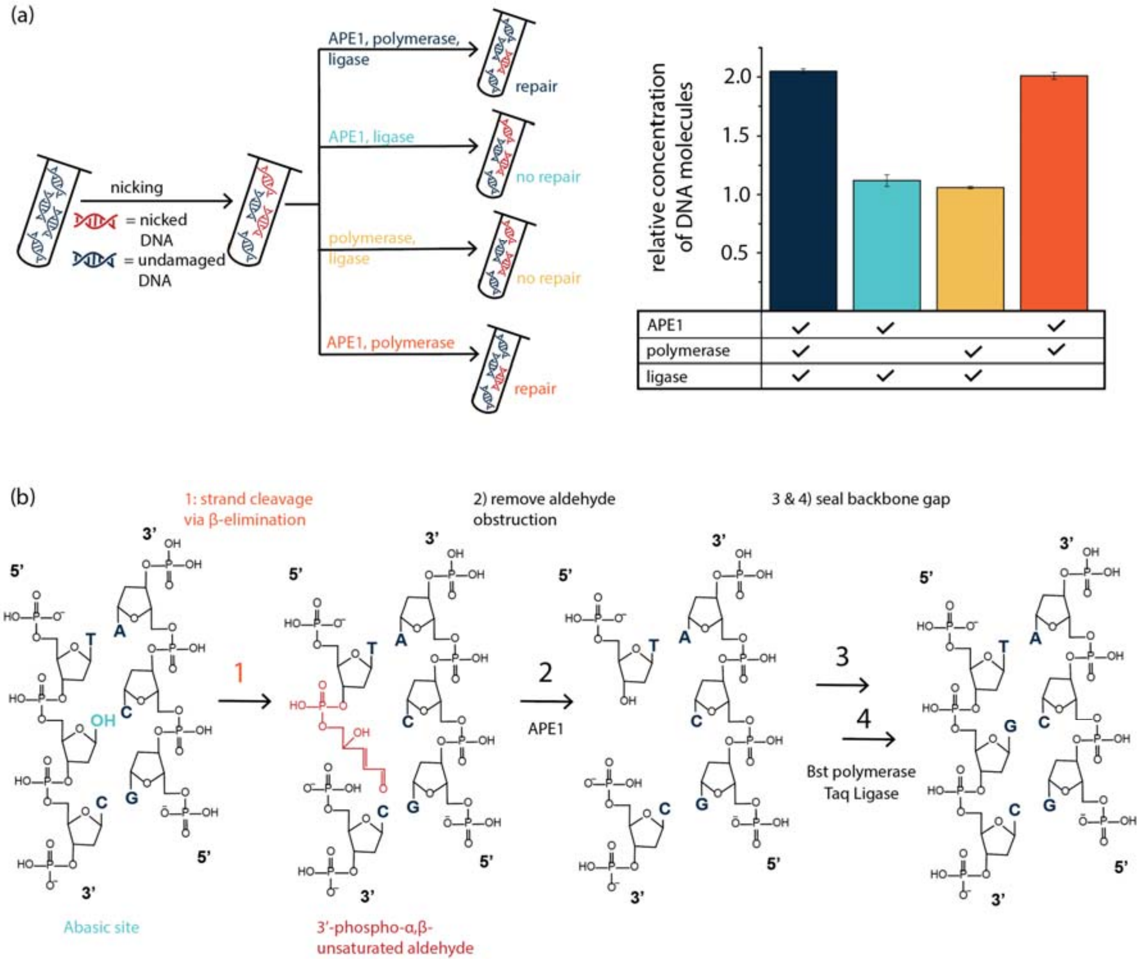

Figure 3: Repair pathway of nicked phosphate backbone. *(a) Quantification of repair potential with three enzymes (APE1, Bst Polymerase, Taq ligase), compared to repair results when leaving out either one of the three enzymes during the repair procedure; s.d. from 3 qPCR replicates. (b) suggested repair pathway of a nick: When the phosphate backbone has been cleaved by β-elimination, APE 1 can ingest the 3'-phospho-α,β-unsaturated aldehyde and render the DNA strand with a 3'-OH terminus such that a polymerase can elongate the backbone and insert the missing base. The taq ligase will then seal the strand.*

We have further investigated the required repair conditions by tuning the relative amounts of these three enzymes as well as the composition of dNTPs (equimolar mixture of dATP, dTTP, dGTP and dCTP), NAD+ and ThermoPol® buffer for optimal repair conditions. We optimized the repair time to be 15 min at a repair temperature of 37 °C. We named this mix the "APL repair mix" based on the initials of **A**PE1, **P**olymerase, **L**igase.

Repair performance

We evaluated the repair performance of the APL repair mix by quantitative PCR, denaturing gel electrophoresis as well as fragmentation analysis by sequencing the damaged and repaired single DNA strands (Fig. 4). Figure 4a shows the percentage of intact DNA per DNA pool investigated (model DNA and file 1 DNA as shown in Supplementary Information) without damage (assumed to be 100 % intact), after damage, and after repair. We observe that self-devised APL mix shows repair of up to 31% of DNA sequences that would have otherwise been lost. Additionally, we observed that the less severe the damage was, the more relative repair was achieved. For example, after damaging model DNA, only 42% of DNA strands were still intact. Repair, however, recovered DNA strands from the pool, increasing the percentage of intact DNA strands to 73% (i.e. 53 % of the damaged sequences could be repaired). On the other hand, after aging file 1 DNA, only 5% of DNA strands were still amplifiable. While with repair, this fraction was improved by more than three fold to 16%, this repair only accounts for 11% of the broken sequences.

To confirm repair of the APL repair mix, we denatured DNA that has been aged at 30 °C for 4 weeks and analyzed the single strand DNA components using gel electrophoresis. (Fig. 4b). Lane M shows a 50 base pair ladder as reference, and L1 shows reference file 1 DNA (no aging, no repair). L2 shows aged file 1 without repair L3 shows aged file 1 DNA that has been repaired using APL repair mix. We see that L2 shows barely any band at 150 nucleotides, indicating that only a few strands of the original sequence length (150 nucleotides) are still intact. L3, on the other hand, shows a clear band at 150 nucleotides, indicating that some of the damaged file must have been repaired by APL repair enzymes to the original length of 150 nucleotides. This confirms our hypothesis that the APL repair mix can repair nicks in the phosphate backbone after hydrolytic damage has occurred.

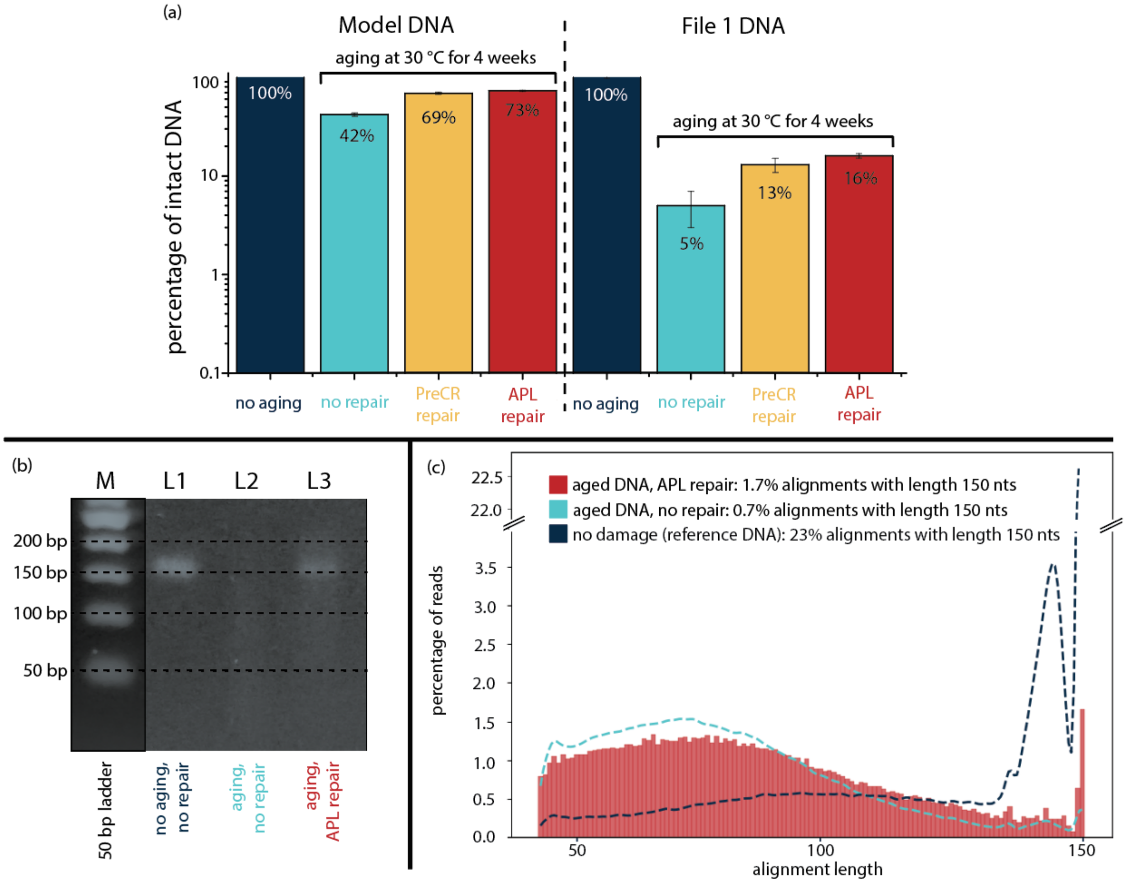

Figure 4: Enzymatic repair of DNA libraries. *(a) repair potential of commercial enzymatic repair mix (PreCR repair) and self-composed enzymatic repair mix (APL repair) shown for damaged files aged at 30 °C for 4 weeks; s.d. from 3 qPCR replicates. (b) denaturing gel of file 1 DNA aged at 30 °C for 4 weeks. Bands show reference DNA (no aging, no repair), damaged DNA without repair (aging, no repair) and damaged DNA that has been treated with APL repair (aging, APL repair). (c) Fragmentation of file 1 DNA after alignment comparing reference and aged DNA to repaired DNA.*

Finally, we analyzed the fragment distribution of the repaired DNA (similarly to the analysis shown in Fig. 2), including dsDNA denaturation, single strand library preparation, sequencing using Illumina's iSeq100 and subsequent alignment (Fig. 4c). Whereas we had observed above that the correct sequence length is found in 23% of alignments in the undamaged file and in 0.7% of alignments in the damaged file, results of enzymatic repair show that fragments got repaired, resulting in the file containing 1.7% of alignments with correct sequence length of 150 nucleotides, more than doubling the number of full-length sequences available for decoding. It may be added, that nicked DNA sequences cannot be decoded simply because they lack PCR primers and would not have been amplified in a standard PCR based dsDNA sequencing preparation workflow or a random access workflow[17,19]. However, broken sequences bring a second problem because during fragmentation, sequence indices get cut off, which are required for decoding.[18] We observe that less shorter fragments (50-70 nucleotides) exist after repair, implying that these sequence lengths were very prevalent for repair.

In combination, PCR, gel electrophoresis and fragmentation analyses show that repair of fragmented synthetic DNA is successful under optimized repair conditions. We observe that the less damage a file has seen, the more repair is possible. This is explainable, as during prolonged damage, a single sequence may be nicked several times, making repair more and more challenging by the decreased melting points of the individual fragments (see discussion in Supplementary Information). Fragment analysis has additionally shown that repair results in an increase in sequences with the correct sequence length, which is an important indicator for whether or not a file can be amplified and sequenced.

As previously mentioned, commercial enzymatic repair mixes exist (i.e., PreCR® repair mix) and can be bought as complete enzymatic kits. Such enzymatic kits have already been investigated for forensic applications, the recovery of ancient DNA or cell-free DNA in maternal plasma for prenatal testing, where compromised DNA samples were repaired using the PreCR® repair kit. Repair results, however, varied across applications: Treating ancient DNA with the PreCR® enzyme kit generally showed no appreciable repair.[9] Limited functionality of repair kits was also observed across most studies of forensic DNA. Some of the most promising results were shown when trying to increase the number of detectable alleles in impaired DNA samples.[10,37–42]

To our best knowledge, for the purpose of repairing synthetic DNA strands, there exists no previous work that has investigated the application of commercial enzymatic repair mixes. We have thus used the PreCR® repair enzymes and optimized the repair procedure to compare 1) costs, 2) repair time and 3) repair capacity to our self-composed repair mix. We found that optimal repair was achieved after 4 h repair time and costs that were more than one order of magnitude higher than full turnaround costs of the APL repair mix developed in this work. We subsequently quantified PreCR® repair using PCR (Fig. 4a) and observed that repair capacity of nicks in synthetic DNA is lower than APL repair across all file sizes and degrees of damage, thus performing worse in all three points of investigation.

Simulating damage and repair for information recovery

Whereas our experimental work has focused on files with high copy numbers of DNA strands, the future of DNA data storage aims at decreasing the volume required for storage, and thus increasing the data density. The physical limit of reliable DNA data retrieval was calculated to be up to 17 exabytes per gram.[43]

Potential future applications for DNA data storage would, for example, be the archiving of valuable information in national or state archives. Whereas on one hand the density for storage is desired to be high (to grasp the full potential of DNA as a medium for data storage), it is imperative that data can still be read subsequently to storage. As the effect of physical redundancy, error correction coding redundancy (logical redundancy) and damage/repair of sequences is complex, we have simulated damage and enzymatic repair of DNA libraries for low physical coverage (= high storage density) scenarios, to express the benefit information recovery can have.

The design of a DNA storage starts with defining the targeted information density and storage horizon, resulting in a targeted physical redundancy and expected DNA decay. Figure 5 shows the expected loss of unique sequences (i.e. unique sequences no longer present in the pool), as a function of these parameters, thereby accounting for expected effects of synthesis bias and PCR bias.[22,44] (see supporting info for details). An error correction code would then be tuned to compensate for these anticipated losses by adding appropriate logical redundancies to the original data. As an example, if data would have to be stored at a physical redundancy of 10 and for 2 DNA half lives (equivalent to ca. 1000 years at room temperature, if stored encapsulated in silica), the model predicts that an error correcting code needs to compensate for a loss of ca. 30% of the unique sequences in the absence of enzymatic repair. According to the Shannon limit this requires a logical redundancy of at least 30%, and consequently, a surplus of 43% of unique sequences would have to be synthesized (surplus = 1/(1-0.3)-1)*100), relative to the case where we do not lose any sequences. However, when applying a DNA repair strategy able to recover 50% of the sequences lost at these parameter choices, significantly less sequences would be missing from the pool (12%), requiring a drastically smaller logical redundancy overhead, and a lower number of surplus DNA synthesis required. As DNA synthesis cost and throughput is currently the main bottleneck of DNA data storage, the application of the rather straightforward APL enzymatic repair, therefore, would have a significant impact on the synthesis cost and read out reliability of the project. Being able to read information from low physical redundancy pools is also of importance, if the DNA pool is distributed, as the case for DNA of things[45] and enzymatic repair could also be used to recover data from DNA data storage systems, where DNA has been stored longer than anticipated.

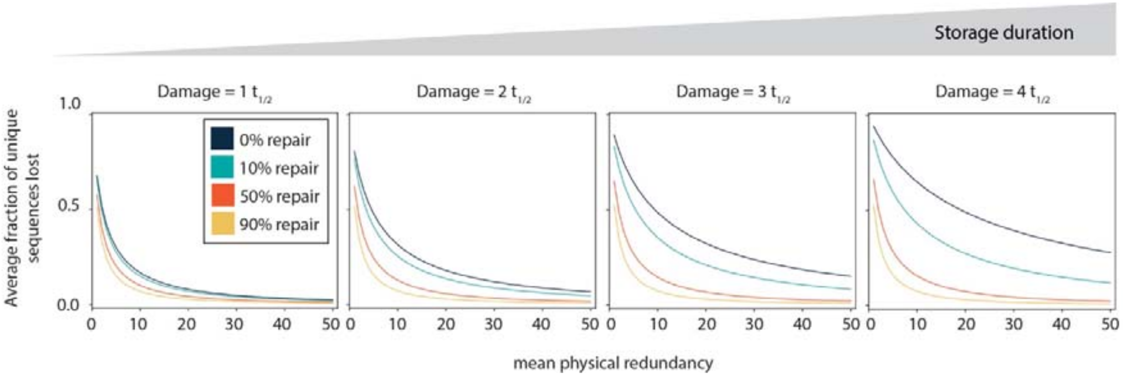

Figure 5: Simulating damage and repair for high-density storage regimes. *For a given physical redundancy and storage horizon (expressed in half lives of DNA), the data gives the fraction of unique sequences lost from the pool during storage (dark blue lines). This loss of sequences has to be compensated by the logical redundancy of an error correcting code. With DNA repair enabled, a fraction of the lost sequences can be recovered, and less logical redundancy is required. The benefits of DNA repair is largest in applications with a low physical redundancy and long storage durations, which are target domains for DNA data storage.*

SUMMARY

In this work, we analyze the decay and repair process of synthetic DNA for DNA data storage applications. We present a molecular analysis of the chemical process during nicking as a form of DNA decay and focus on nicking due to hydrolysis. Experimental conditions represent real-life scenarios during DNA handling steps and illustrate the threat of water (in the form of a medium for handling, and in the form of humidity during storage). We find that hydrolysis causes molecular obstruction at the 3' terminal when the strand is nicked, which results in strand loss during denaturation steps of PCR. As strand loss can only be compensated by redundancy of additional strands, nicking is a very expensive form of damage (compared to single-nucleotide mutations). We have thus devised an enzymatic solution containing three single enzymes (APE1, Bst polymerase, and Taq ligase) targeted at removing the obstructive aldehyde group after nicking and subsequently sealing the nick to repair DNA strands such that PCR amplification is possible again. We find that repair can recover information, with higher recovery in low-damage regimes and show more than 30%

recovery. Comparing our self-devised repair mix to commercial kits (i.e., PreCR® repair mix), our repair results in cheaper, faster, and more effective recovery of DNA.

At low physical coverage regimes, losing DNA strands can lead to data (contained in that strand) being lost and compromised read-out. We have thus simulated the requirement of physical and logical redundancy to predict and control the encoding process for DNA data storage such that read-out is still possible. This will be especially useful when moving towards high-density storage, for which DNA is especially well suited.

Enzymatic repair of DNA can, in the future, be especially interesting to data archives, when converting traditional archiving processes to include DNA as a storage medium becomes of greater interest. When storage times are unknown, a method for information recovery is beneficial, offering a recovery option if standard decoding and read-out shall fail.

BOX 1: Comparison of repair mixes

As referred to throughout this work, different enzymatic repair pathways exist (commercially or theoretically) that allow for repair of compromised DNA strands. The following table lists the components of the commercially available PreCR® repair mix, the enzymes as suggested by Lindahl et al. for the base-excision repair pathway[2], and our self-composed APL repair mix. Repairing the nick that occurred during hydrolysis does not require the full procedure of the base-excision repair pathway, as the repair of mutations is not necessary for data recovery for DNA data storage (the error-correction code can easily account for base mutations). This is why the APL repair mix requires only three enzymes that chemically target the obstructive damage caused by nicking.

Enzyme	**APL repair mix**	**PreCR ® repair mix**	**Lindahl et al.[2]**
UDG Glycosylase		x	x
FPG Glycosylase		x	
T4 PDG Glycosylase		x	
Endonuclease IV		x	x
Endonuclease VIII		x	
APE 1	x		
RecJ Protein			x
Polymerase I			x
Bst Polymerase	x	x	
T4 Ligase			x

Taq Ligase	x	x	

Method

PreCR enzyme kit – optimal repair conditions. The optimal ratios for repair are as follows. Master mix (ThermoPol buffer, dNTP, NAD+ (5:2:1)) was added to DNA (1 ng/µL) and water/PreCR® repair enzymes (ratio: 4:25:1, respectively). Enzymes were added last and reactions were kept on ice. After the addition of all components, reaction vials were shaken by hand and centrifuged. Repair was carried out in 30 °C incubator for 4 hours and subsequently centrifuged again. PCR quantification was performed by diluting 1 µL of incubated DNA in 499 µL water.

Self-composed enzyme repair mix – optimal repair conditions. Individual components were added on ice to make master mix: (ThermoPol buffer, dNTP, NAD+, APE1, Bst polymerase (full length), Taq ligase (ratio: 10:4:2:1:1:1). For repair reactions, 18.8 µL mili-Q water, 6.25 µL DNA (1 ng/µL), and 1.57 µL master mix were added (for control samples, the 1.57 µL master mix were exchanged for water). After the addition of all components, reaction vials were shaken by hand and centrifuged. Repair was carried out in 30 °C incubator for 15 minutes. Samples were subsequently centrifuged again. PCR quantification was performed by diluting 1 µL of incubated DNA in 499 µL water.

Sequencing preparation. Damaged, repaired and reference samples were denatured and prepared for sequencing using the Swift Biosciences Accel-NGS® 1S Plus DNA Library Kit and following the manufacturer's recommended protocol. To preserve small DNA fragments to the best extent possible, small fragment retention was performed according to the manufacturer's instructions. All steps were performed separately for all samples and no nicked samples were mixed.

Denaturing gel. DNA samples were denatured in the following way and then run on normal Agarose 2% SYBR Gold II gels: 200 mM sodium phosphate solution (pH 7.0) was prepared by adding 1.64 g of Na2HPO4 and 1.02 g of NaH2PO4 to 80 mL of deionized water. The pH was adjusted to 7 by using NaOH (this step may not be necessary). Water was added to bring the total volume to 100 mL. Denaturing buffer was made by adding 200 mM sodium phosphate solution (23.8 µL), 40% glyoxal (83.3 µL) and 99.7% DMSO (238 4 µL).

DATA AVAILABILITY

The sequencing data underlying Figure 3 will be made available on a figshare repository.

CODE AVAILABILITY

The code utilized to simulate the losses of sequences as a function of physical redundancy and storage time can be found in the supporting information.

ACKNOWLEDGEMENTS

The authors than Microsoft for funding and the Beat Christen Group for giving access to the iSeq 100 sequencer.

AUTHOR CONTRIBUTIONS

R.N.G. and K.S. initiated the project, R.N.G., K.S. B.H.N. and W.J.S. supervised the project. R.H., L.C.M and B.H.N. performed data analysis. L.C.M, G.L. and W.D.C. performed laboratory experiments, T.D. and A.L.G. coded, implemented and analyzed the model data. All figures prepared by L.C.M. L.C.M and R.N.G wrote the manuscript with input and approval from all authors.

COMPETING INTERSTS STATEMENT

R.N.G. and W.D.C. are authors of a patent application assigned to Microsoft. K.S. and B.H.N. are Microsoft employees.

References

1. Sancar, A. DNA Repair in Humans. *Annu. Rev. Genet.* **29**, 69–105 (1995).
2. Dianov, G. & Lindahl, T. Reconstitution of the DNA base excision-repair pathway. *Curr. Biol.* **4**, 1069–1076 (1994).
3. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **366**, 529–531 (1993).
4. Cleavera, J. E. Profile of Tomas Lindahl, Paul Modrich, and Aziz Sancar, 2015 nobel laureates in chemistry. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 242–245 (2016).

5. Lindahl, T. New classes of enzymes acting on damaged DNA. *Nature* 64–66 (1976).
6. Lindahl, T. DNA Repair Enzymes. *Ann. Rev. Biochem* **41**, 61–87 (1982).
7. Dianov, G., Price, A. & Lindahl, T. Generation of single-nucleotide repair patches following excision of uracil residues from DNA. *Mol. Cell. Biol.* **12**, 1605–1612 (1992).
8. Dovgerd, A. P. & Zharkov, D. O. Application of repair enzymes to improve the quality of degraded DNA templates for PCR amplification. *Appl. Biochem. Microbiol.* **50**, 235–242 (2014).
9. Mouttham, N., Klunk, J., Kuch, M., Fourney, R. & Poinar, H. Surveying the repair of ancient dna from bones via high-throughput sequencing. *Biotechniques* **59**, 19–25 (2015).
10. Diegoli, T. M., Farr, M., Cromartie, C., Coble, M. D. & Bille, T. W. An optimized protocol for forensic application of the PreCR^{TM} Repair Mix to multiplex STR amplification of UV-damaged DNA. *Forensic Sci. Int. Genet.* **6**, 498–503 (2012).
11. Galimberti, A. *et al.* DNA barcoding as a new tool for food traceability. *Food Res. Int.* **50**, 55–63 (2013).
12. Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
13. Church, G. M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* **337**, 1628 (2012).
14. 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. *Angew. Chem. Int. Ed.* **54**, 2552–2555 (2015).
15. Tabatabaei Yazdi, S. M. H., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. *Sci. Rep.* **5**, 1–10 (2015).
16. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
17. Organick, L. *et al.* Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–249 (2018).

18. Meiser, L. C. *et al.* Reading and writing digital data in DNA. *Nat. Protoc.* **15**, 86–101 (2020).

19. Yazdi, S. M. H. T., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A Rewritable, Random-Access DNA-Based Storage System. *Sci. Rep.* **5**, 1–10 (2015).

20. Matange, K., Tuck, J. M. & Keung, A. J. DNA stability: a central design consideration for DNA data storage systems. *Nat. Commun. 2021 121* **12**, 1–9 (2021).

21. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15**, 366–370 (2016).

22. Heckel, R., Mikutis, G. & Grass, R. N. A Characterization of the DNA Data Storage Channel. *Sci. Rep.* **9**, 9663 (2019).

23. Bossert, M. *Channel coding for telecommunications*. (New York□: Wiley, 1999).

24. Nash, H. M. *et al.* Cloning of a yeast 8-oxoguanine DNA glycosylase reveals the existence of a base-excision DNA-repair protein superfamily. *Curr. Biol.* **6**, 968–980 (1996).

25. Antkowiak, P. L. *et al.* Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nat. Commun. 2020 111* **11**, 1–10 (2020).

26. Tabatabaei, S. K. *et al.* DNA punch cards for storing data on native DNA sequences via enzymatic nicking. *Nat. Commun. 2020 111* **11**, 1–10 (2020).

27. Allentoft, M. E. *et al.* The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B Biol. Sci.* **279**, (2012).

28. Xu, Y., Ren, X., Wang, H., Wang, M. & Li, G. Evaluation of DNA degradation and establishment of a degradation analysis model for Lepidoptera specimens. *Biotechniques* **68**, 138–147 (2020).

29. Kistler, L., Ware, R., Smith, O., Collins, M. & Allaby, R. G. A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Res.* **45**, 6310–6320 (2017).

30. Mikutis, G., Schmid, L., Stark, W. J. & Grass, R. N. Length-dependent DNA degradation kinetic model: Decay compensation in DNA tracer concentration measurements. *AIChE J.* **65**, 40–48 (2019).

31. Bonnet, J. *et al.* Chain and conformation stability of solid-state DNA: Implications for room temperature storage. *Nucleic Acids Res.* **38**, 1531–1546 (2009).

32. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods 2009 65* **6**, 343–345 (2009).

33. Wiederhold, L. *et al. AP Endonuclease-Independent DNA Base Excision Repair in Human Cells*. *Molecular Cell* vol. 15 (2004).

34. Abbotts, R. & Wilson III, D. M. Coordination of DNA Single Strand Break Repair. *Free Radic Biol Med* **107**, 228–244 (2017).

35. Abbotts, R. & Madhusudan, S. Human AP endonuclease 1 (APE1): From mechanistic insights to druggable target in cancer. *Cancer Treat. Rev.* **36**, 425–435 (2010).

36. Grin, I. R., Khodyreva, S. N., Nevinsky, G. A. & Zharkov, D. O. Deoxyribophosphate lyase activity of mammalian endonuclease VIII-like proteins. *FEBS Lett.* **580**, 4916–4922 (2006).

37. Ohta, J., Noda, N., Minegishi, S. & Sakurada, K. Application of DNA repair for Streptococcus salivarius DNA-based identification of saliva from ultraviolet-exposed samples. *Forensic Sci. Int.* **306**, 110077 (2020).

38. Liu, Y. *et al.* Comparison of different methods for repairing damaged DNA from buffered and unbuffered formalin-fixed tissues. *Int. J. Legal Med.* **132**, 675–681 (2018).

39. Ambers, A., Turnbough, M., Benjamin, R., King, J. & Budowle, B. Assessment of the role of DNA repair in damaged forensic samples. *Int. J. Legal Med.* **128**, 913–921 (2014).

40. San Pietro, D., Tagliaro, F. & Adamowicz, M. S. A preliminary assessment of the effect of PreCR^{TM} DNA repair treatment on mixture ratios in two person mixtures. *Sci. Justice* **58**, 308–314 (2018).

41. Robertson, J. M. *et al.* Assessing PreCR^{TM} repair enzymes for restoration of STR profiles from artificially degraded DNA for human identification. *Forensic Sci. Int. Genet.* **12**, 168–180 (2014).

42. Westen, A. A. & Sijen, T. Degraded DNA sample analysis using DNA repair enzymes, mini-STRs and (tri-allelic) SNPs. *Forensic Sci. Int. Genet. Suppl. Ser.* **2**, 505–507

(2009).

43. Organick, L. *et al.* Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.* **11**, 1–7 (2020).

44. Chen, Y. J. *et al.* Quantifying molecular bias in DNA data storage. *Nat. Commun. 2020 111* **11**, 1–9 (2020).

45. Koch, J. *et al.* A DNA-of-things storage architecture to create materials with embedded memory. *Nat. Biotechnol.* **38**, 39–43 (2020).