DNA Stability Evaluation Method for DNA Data Storage Containment Systems

Version 0.9, Revision 5

ABSTRACT: This specification defines a standard procedure to measure and a standard metric to characterize the molecular stability of DNA in a DNA Data Containment System (DCS) so that a DCS being considered as a part of a DNA data storage solution can be objectively compared, in terms of how effectively the DCS protects the media, vs. other DCSs being so considered.

Publication of this Working Draft for review and comment has been approved by the DNA TA TWG Data Retention subgroup. This draft represents a “best effort” attempt by the DNA TA TWG Data Retention subgroup to reach preliminary consensus, and it may be updated, replaced, or made obsolete at any time. This document should not be used as reference material or cited as other than a “work in progress.” Suggestions for revisions should be directed to http://www.snia.org/feedback/.

Working Draft

August 28, 2024

USAGE

The SNIA hereby grants permission for individuals to use this document for personal use only, and for corporations and other business entities to use this document for internal use only (including internal copying, distribution, and display) provided that:

1. Any text, diagram, chart, table or definition reproduced shall be reproduced in its entirety with no alteration, and,
2. Any document, printed or electronic, in which material from this document (or any portion hereof) is reproduced, shall acknowledge the SNIA copyright on that material, and shall credit the SNIA for granting permission for its reuse.

Other than as explicitly provided above, you may not make any commercial use of this document or any portion thereof, or distribute this document to third parties. All rights not explicitly granted are expressly reserved to SNIA.

Permission to use this document for purposes other than those enumerated above may be requested by e-mailing tcmd@snia.org. Please include the identity of the requesting individual and/or company and a brief description of the purpose, nature, and scope of the requested use.

All code fragments, scripts, data tables, and sample code in this SNIA document are made available under the following license:

- BSD 3-Clause Software License
- Copyright (c) 2024, The Storage Networking Industry Association.
- Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:
- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of The Storage Networking Industry Association (SNIA) nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.
- THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

DISCLAIMER

The information contained in this publication is subject to change without notice. The SNIA makes no warranty of any kind with regard to this specification, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The SNIA shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance, or use of this specification.

Revision History

Revision	Date	Sections	Originator:	Comments
v0.7 r3	29-March-2024	All	Dave Landsman	All comments integrated.
0.7 r4	03-April-2024	• References • Appendix A	Dave Landsman	• Removed references 31-34, as we don't cite in document, and resolved the citations in text. • Updated version of Table 1 (removed fossil rows)
0.9 r0	16-July-2024	• Various	Dave Landsman Marthe Colotte Chris Takahashi	• Added variable definitions for equation 1 • Added "n", number of nt, in equation 1 • Modified figure 6 – Aging kinetics timing model • Various editorial revisions throughout
0.9 r2	13-Aug-2024	•	Dave Landsman Marthe Colotte Chris Takahashi	• Re-wrote sections 6.1.1. and 6.1.2 to improve degradation model and better explain the timing model and experiment flow. • Accepted other editorial changes throughout
0.9 r3	19-Aug-2024		Dave Landsman Marthe Colotte Chris Takahashi Lee Organick	• Updated notation in 6.1.1 and 6.1.2 to explicitly note temperature dependency of the data series from t_1 to t_q, including Figure 6 update. • Miscellaneous other editorial updates.
0.9 r4	21-Aug-2024			• Few editorial changes in review w/ Chris • Accepted all changes for clean r4
0.9 r5	28-Aug-2024			• Changed spec title – removed "data", as we are evaluating molecular stability, not data stability

Contents

1. Scope/Purpose

There is a need to begin clarifying how long end-users can count on storing their digital data in synthetic DNA. This specification defines a standard procedure to characterize the molecular stability properties of the Storage part of the DNA data storage pipeline, the DNA Containment System (DCS), independent of the other phases. In so doing, we create a means for end users to objectively compare the stability claims made by different DCS vendors and make the appropriate cost/benefit decisions for a particular DCS, and we begin to create trust in DNA as a media type.

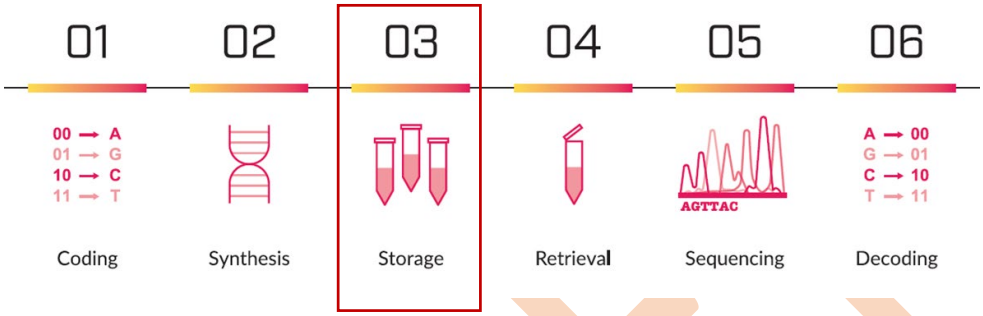

Figure 1 – DNA Data Storage Pipeline

Data retention for DNA data storage, which is currently a technology dependent and end-to-end system problem, is outside the scope of this standard. This standard does, however, begin to indirectly address data retention, because it is based on premises from the literature, and documented in Appendix A, supporting the conclusion that if DNA molecules stored in a DCS survive storage without incurring structural defects that halt polymerization (i.e., that halt PCR) the data encoded in the recovered intact molecules is recoverable.

2. References

1. Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ. Robust Chemical Peservation of Digital Information on DNA in Silica with Error-Correcting Codes. Angew Chem Int Ed Engl. 2015; 54(8): 2552-2555.
2. Lindahl T, Nyberg B. Rate of depurination of native deoxyribonucleic acid. Biochemistry. 1972; 11(19): 3610-3618
3. Bonnet J, Colotte M, Coudy D, Couallier V, Portier J, Morin B, et al. Chain and conformation stability of solid-state DNA: implications for room temperature storage. Nucleic Acids Res. 2010; 38(5): 1531-1546
4. Anchordoquy TJ, Molina MC. Frontiers in Clinical Research. Preservation of DNA. Cell Preservation Technology. 2007; 5(4): 180-188
5. Molina MC, Anchordoquy TJ. Degradation of lyophilized lipid/DNA complexes during storage: The role of lipid and reactive oxygen species. Biochim Biophys Acta-Biomembr. 2008; 1778(10): 2119-2126
6. Organick L, Nguyen BH, McAmis R, Chen WD, Kohll AX, Ang SD, et al. An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage. Small Methods. 2021; 5(5): e2001094
7. Coudy D, Colotte M, Luis A, Tuffet S, Bonnet J. Long term conservation of DNA at ambient temperature. Implications for DNA data storage. PLoS One. 2021; 16(11): e0259868

8. Wandeler, P. Patterns of nuclear DNA degeneration over time: a case study in historic teeth samples, Mol. Ecol. 12 (2003) 1087–1093.
9. Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc Biol Sci. 2012; 279(1748): 4724-4733.
10. Colotte M, Couallier V, Tuffet S, Bonnet J. Simultaneous assessment of average fragment size and amount in minute samples of degraded DNA. Anal Biochem. 2009; 388345–347
11. Kohll AX, Antkowiak PL, Chen WD, Nguyen BH, Stark WJ, Ceze L, et al. Stabilizing synthetic DNA for long-term data storage with earth alkaline salts. Chem Commun (Camb). 2020
12. Antkowiak PL, Koch J, Rzepka P, Nguyen BH, Strauss K, Stark WJ, et al. Anhydrous calcium phosphate crystals stabilize DNA for dry storage. Chem Commun (Camb). 2022.
13. Strauss, K et al 2021 US2021291134 (A1) Silica Encapsulated DNA on Magnetic Nanoparticles.
14. Jahanshahi-Anbuhi S et al. Pullulan encapsulation of labile biomolecules to give stable bioassay tablets. Angew Chem Int Ed Engl. 2014; 53(24): 6155–6158
15. Liu Y, et al. DNA preservation in silk. Biomater Sci. 2017.
16. Strauss, K. 2019 US2019390194 (A1) High-density DNA storage with salt.
17. Moon, W.-C. US2007254294 (A1) 2007 Method for Storing Dna by Using Chitosan, and Products Using the Methods.
18. Khavani, M. Application of amino acid ionic liquids for increasing the stability of DNA in long term storage, J Biomol Struct Dyn 1-15 2022
19. Hiroyuki O, et al US2007196826 A1 Solvent For Dissolving Nucleic Acid, Nucleic Acid-Containing Solution And Method Of Preserving Nucleic Acid
20. Mohanty, P.S. et al AU2020244494B2 Capillary assisted vitrification processes and devices.
21. Whitney, S. E., et al 2014 US 2014/0065627Al compositions and methods for biological sample storage.
22. Hogan, M. et alUS2017145477 (A1) Matrices and media for storage and stabilization of biomolecules.
23. Horton, J. K. et al US20170151545A1 Oligonucleotide data storage on solid supports
24. Fomovskaia, G. and et. Al WO0062023 (A1) FTA-coated media for use as a molecular diagnostic tool.
25. https://300k.bio/ accessed on 2024-03-27
26. Newman, S., et al High density DNA data storage library via dehydration with digital microfluidic retrieval. Nature communications 1706 10 (1) 2019
27. Provisional patent application (application number 62/812,521) filed on March 1, 2019 by the University of Washington
28. Trapmann, S. et al, Development of a novel approach for the production of dried genomic DNA for use as standards for qualitative PCR testing of food-borne pathogens. , Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement 695-699 9 (11) 2004
29. Wong, P.C. US2006024811 (A1), Storing data encoded DNA in living organisms.
30. Colotte, M. et al., Adverse Effect of Air Exposure on the Stability of DNA Stored at Room Temperature. Biopreserv Biobank. 2011; 9(1): 47–50
31. Molina, M et al, Degradation of lyophilized lipid/DNA complexes during storage: The role of lipid and reactive oxygen species, https://doi.org/10.1016/j.bbamem.2008.04.003

32. ACS Nano 63-9 1 (1) 2007 A general approach for DNA encapsulation in degradable polymer microcapsules. Zelikin, A. N., A. L. Becker, A. P. Johnston, K. L. Wark, F. Turatti and F. Caruso
33. Langmuir 34-42 21 (1) 2005 Encapsulation of DNA by cationic diblock copolymer vesicles. Korobko, A. V., W. Jesse and J. R. van der Maarel
34. Mater Today Bio 100900 24 2024 DNA data storage in electrospun and melt-electrowritten composite nucleic acid-polymer fibers. Soukarie, D., L. Nocete, A. M. Bittner and I. Santiago
35. ACS Synthetic Biology 2023 Magnetic Bead Spherical Nucleic Acid Microstructure for Reliable DNA Preservation and Repeated DNA Reading. Shen, P., X. Qu, Q. Ge, T. Huang, Q. Sun and Z. Lu
36. Adv Sci (Weinh) e2305921 2024 "Cell Disk" DNA Storage System Capable of Random Reading and Rewriting. Hou, Z., W. Qiang, X. Wang, X. Chen, X. Hu, X. Han, W. Shen, B. Zhang, P. Xing, W. Shi, J. Dai, X. Huang and G. Zhao
37. National Science Review 8 (5) 2021 An artificial chromosome for data storage. Chen, W., M. Han, J. Zhou, Q. Ge, P. Wang, X. Zhang, S. Zhu, L. Song and Y. Yuan
38. Nature Computational Science 2022 Towards practical and robust DNA-based data archiving using the yin–yang codec system. Ping, Z., S. Chen, G. Zhou, X. Huang, S. J. Zhu, H. Zhang, H. H. Lee, Z. Lan, J. Cui, T. Chen, W. Zhang, H. Yang, X. Xu, G. M. Church and Y. Shen

3. Terms and definitions

3.1. Reference Media Pool

The complete pool of DNA molecules that is apportioned into units of a DCS under test and formatted as specified in this standard.

3.2. Alternative Media Pool

A pool of DNA molecules that is apportioned into units of the DCS under test but, due to the requirements of the DCS, uses a different physical form of DNA (e.g., long oligos, plasmids), as discussed in section 8 (Alternative Methods).

3.3. DNA Containment System (DCS)

A sealed or unsealed container or substrate, including any associated Preservation Method, in which DNA is encoded with digital data and stored for later retrieval, reading, and decoding.

3.4. DCS Unit

The smallest physical unit of a DCS into which DNA is stored and preserved for later access to the media. A DCS Unit can comprise the entire DCS or can be a subsegment of the DCS. For example, all of the media in a DCS could be stored in a single storage vessel, which would mean the DCS has a single DCS Unit, or the media could be subdivided into small vials, in which case each vial, or a defined group of vials, is a DCS Unit.

3.5. Preservation Method

The totality of the steps (e.g., drying, addition of chemical additives, insertion of inert gasses, container sealing) by which DNA molecules are prepared for storage in a DCS Unit.

3.6. Stability

The property of a DNA molecule to maintain its molecular structure (e.g., no strand breaks) such that it will not halt polymerization (i.e., can be amplified using PCR) after storage.

3.7. Half-Life

For a given temperature, the time by which fifty percent of the DNA strands within a DCS Unit under test have incurred at minimum one strand break.

3.8. Accelerated Aging

In the context of DNA stability evaluation, accelerated aging is the application of exaggerated environmental conditions (e.g., high temperature, high humidity) to the DCS to cause the rate of molecular degradation in the DNA to be faster than would occur if the DCS were used to store data under normal use. (Note: It is known that the molecular degradation caused by the exaggerated conditions used in accelerated aging does not perfectly model molecular degradation during non-accelerated aging (i.e., storage under real-world use cases over long periods), but there is substantive evidence in the literature showing that accelerated aging of DNA reasonably models the non-accelerated case [3] [30].)

3.9. Degradation Rate Constant

Rate constant (k) for first-order or pseudo first-order kinetics which indicates the rate at which molecular degradation processes occur.

3.10. Arrhenius plot

Mathematical function that describes the approximate relationship between the rate constant of a chemical reaction and the temperature and energy of activation.

3.11. Quantitative PCR (qPCR)

A method used to detect and quantify the presence of DNA molecules in a sample. The use of qPCR requires the use of a Calibration Standard to calibrate the qPCR instrument.

3.12. Calibration Standard

A predefined sequence of DNA used to calibrate qPCR instruments, providing a frame of reference for reading the results of a qPCR run.

3.13. Digital PCR (dPCR)

A method used to detect and quantify the presence of DNA molecules in a sample. While using the same underlying chemistry (i.e., PCR) as qPCR, dPCR does not require

calibration, allowing absolute quantification of the number of amplifiable targets by applying the Poisson law.

3.14. Relative humidity (RH)

Ratio, defined as a percentage, of the existing partial vapor pressure of water to the vapor pressure at saturation. Usually, but not always, equal to the percentage of the amount of moisture in the air to that at saturation.

3.15. Strand Break

A degradation event in a DNA molecule characterized by a physical break in the sugar-phosphate chain.

3.16. Blocked Strand

A DNA strand which, due to a variety of possible root causes and molecular processes, has structural defects that will halt polymerization (i.e., will halt PCR). While multiple and complex molecular processes can lead to a strand becoming a Blocked Strand, the predominant modalities are processes leading to the loss of a base, which in turn leads to a Strand Break. (See Appendix A for background on DNA molecular breakdown)

4. General principles and spec overview

The outline of the DNA Stability Evaluation protocol is shown in figure 2. A Reference Media Pool is created per Section 5. Aliquots of the Reference Media Pool are then apportioned into the DCS and the stored media is aged in incubators according to the number of temperature and humidity conditions required to generate enough degradation events in a reasonable period of time to enable degradation analysis. At defined time points (t_1-t_q) during aging, qPCR (or equivalent) is used to measure the percentage of intact strands of original DNA remaining. Finally, an Arrhenius curve fit of the sampled data points are used to estimate the Half-Life metric (Figure 3) for the DCS at 25°C.

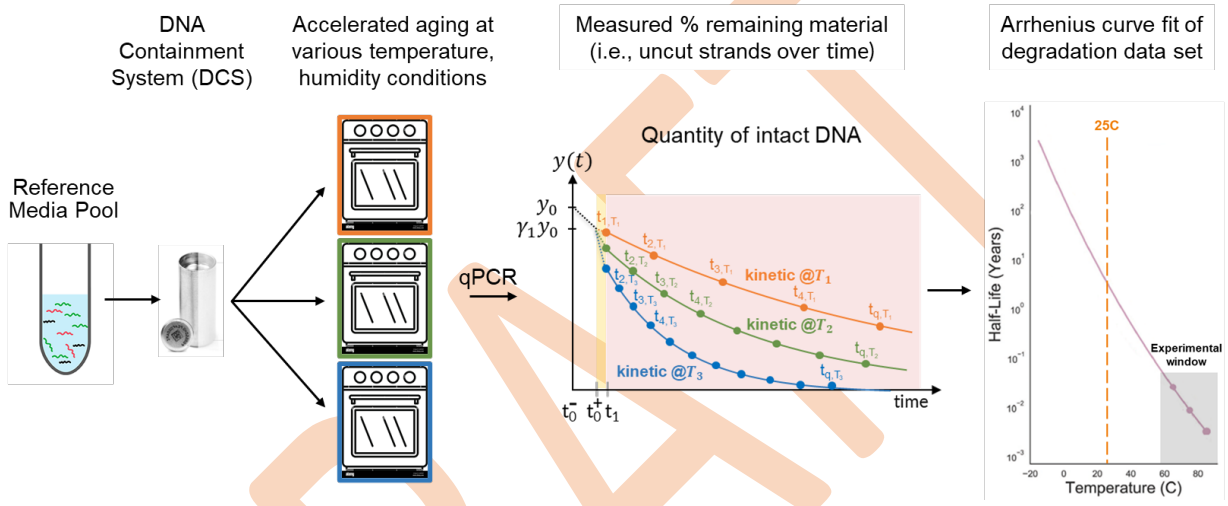

Figure 2 – Degradation Stability Evaluation Method

The DNA Stability Evaluation Method specification is based on the fundamental premise that we can establish the Half-Life of a DCS by counting the rate at which the DNA molecules stored in the DCS incur Strand Breaks during storage. (Figure 3)

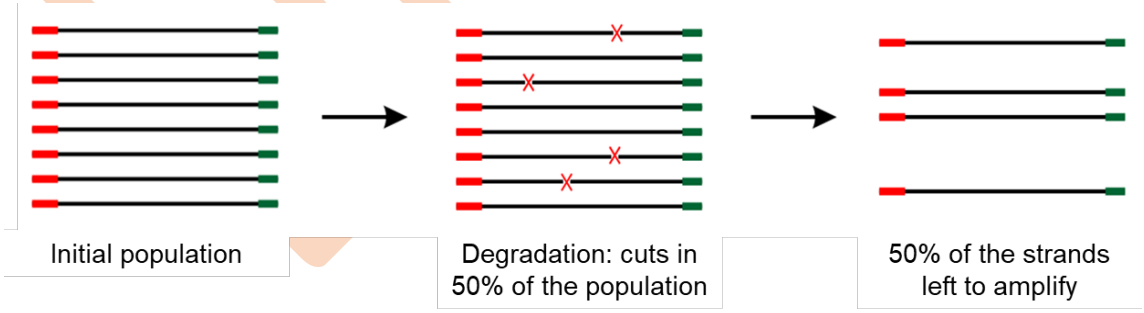

Figure 3 - DNA Half-Life in storage

A backgrounder on DNA molecular degradation is in Appendix A but the following are the key conclusions from the literature underlying the specification's premise:

1. The predominant form of DNA degradation events in storage is a strand break (break in the sugar-phosphate backbone of the molecule), which can be detected and quantified with qPCR or dPCR since a broken chain cannot be amplified.

2. Strand breaks during storage appear to be independent of the sequences in the DNA strands stored.
3. If a DNA strand survives storage with no strand breaks, the data stored in that strand appears to be recoverable. In other words, there is no observed sequence bias for the DNA storage methods tested to date.

While research in DNA degradation continues, and these conclusions could be revisited, this standard posits that competing Half-Life claims regarding DNA Data Containment Systems can be compared, independent of the surrounding end-to-end DNA data storage system.

The structure of the DNA Stability Evaluation Method specification is as follows:

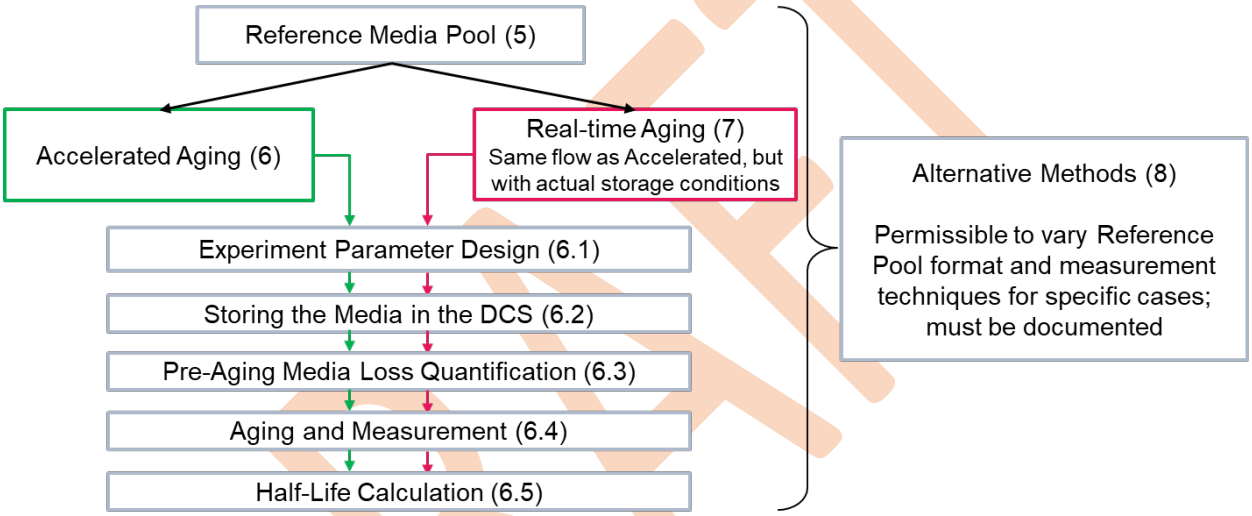

Figure 4 - Stability Evaluation Method Specification Flow

5. Reference Media Pool

Although, as noted in General Principles (section 4), the research indicates that there is no sequence bias associated with a DCS, this specification recommends the use of a standard formatted pool, if possible. The Reference Media Pool is defined as follows.

Each sequence in the Reference Media Pool shall contain a random 110 nucleotide (nt) long poly (N) sequence as the payload, where N is a random nucleotide incorporated during synthesis, flanked by a set of 20 nucleotide primers to be used when amplifying and quantifying sequences, as shown in Figure 5. Theoretically, there are a maximum of 4^{110} unique sequences.

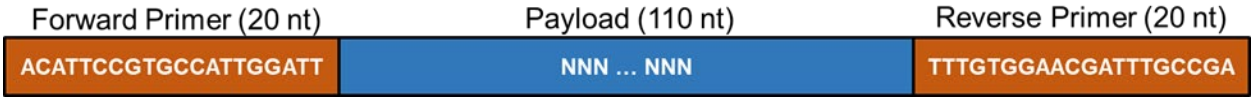

Figure 5 – Standard Reference Media Pool Oligo format

If different primer sequences are used, for example because the Preservation Method for a DCS requires a specified adapter sequence, they shall be identical throughout the pool, as inconsistencies could affect the degradation kinetics.

Note – Non-Standard Pool: In some cases, it may be necessary or desirable to use a pool which differs in form and format from the Reference Media Pool; this is permitted by the standard (see Section 8, Alternative Methods).

6. Accelerated Aging Protocol

The protocol for the accelerated aging media stability experiment is as follows:

1. **Experiment Parameter Design (Section 6.1):** The time points, temperatures, humidity levels to be tested for the given DCS are specified. It may be necessary to run some preliminary "calibration kinetics" (Section 6.1.5) to choose the selected parameters.
2. **Storing the media in the DCS (Section 6.2):** Aliquots of the Reference DNA Pool are stored in a population of DCS Units, prior to aging conditions being applied, per how many temperature/humidity points, replicants, etc. are needed to conduct the aging test.
3. **Pre-Aging Media Loss Quantification (Section 6.3):** A quantification of DNA material is done after steps 1 & 2, to account for possible material loss due to Media Preservation, prior to the application of aging conditions.
4. **Aging and Measurement (Section 6.4):** The DCS units are placed in incubators and the aging conditions are applied. Starting at time t_1, and at each subsequent time point through t_q, the fraction of intact DNA in each DCS unit is measured with qPCR or dPCR. This step includes calibration if qPCR is being used.
5. **Analysis and Reporting (Section 6.5):** After all data is collected, the half-life for the DCS is calculated with an Arrhenius curve fit.

6.1. Experiment Parameter Design

This section explains and specifies the degradation model underlying the aging protocol, and the parameter model for the aging experiment; that is, the timing, sampling, temperature, and humidity requirements for constructing and conducting the accelerated aging experiment.

6.1.1. Degradation Model

This standard assumes the degradation model shown in equation (1), expressed in terms of the Arrhenius equation modified by two constant terms ($\boldsymbol{\gamma_1}$ and $\boldsymbol{\gamma_2}$) to account for: (a) material loss prior the time that the aging conditions are applied; and (b) the early phase of aging, during which degradation rates have not settled into a steady state.

The degradation model is as follows:

$$\mathrm{y(t)} = \begin{cases} y_0 & \text{for } t = t_0^- \\ \gamma_1 y_0 & \text{for } t = t_0^+ \\ \gamma_1 \gamma_{2,T} y_0 \exp\left(-nA \exp\left(\frac{-E}{T}\right) t\right) & \text{for } t \geq t_{1,T} \end{cases} \qquad (1)$$

where:

$\boldsymbol{y_0 = y(t_0^-)}$	=	the initial quantity of DNA prepared for preservation in a DCS Unit (i.e., single sample), but before it is prepared and stored into the DCS.
$\boldsymbol{\gamma_1 = y(t_0^+)/y_0)}$	=	DCS Recovery Rate %; the fraction of DNA remaining after the sample has been processed and stored into a DCS Unit but prior to the degradation experiment starting (i.e., prior to aging conditions being applied). This accounts for loss due to material handling in preparation for preservation in the DCS (see also Pre-Aging Media Loss Quantification - Section 6.3).
$\boldsymbol{\gamma_{2,T} = y(t_{1,T})/(t_0^+)}$	=	the fraction of DNA remaining after short term losses early in the aging experiment that is not accounted for by a constant degradation rate between time t_0^+ and t_1, but not including the losses defined by γ_1. γ_2 depends on temperature and time in the early stages of aging and is included in the model to exclude early anomalies prior to reaching a steady state degradation rate.
$\boldsymbol{t_0^-}$	=	time by which the DNA is prepared for storage into the DCS, but before being stored in the DCS
$\boldsymbol{t_0^+}$	=	time at which aging conditions are applied to the DCS containing DNA.
$\boldsymbol{t_{1,T}}$	=	time by which aging conditions have been applied to the DNA in the DCS long enough for the degradation rate to have reached steady state for the kinetic being run at temperature T. $t_{1,T}$ shall be defined such that the interval between t_0^+ and $t_{1,T}$ is long enough to allow the degradation rate to stabilize to a constant rate.
$\boldsymbol{n}$	=	the number of nucleotides per strand in the tested sample
$\boldsymbol{A}$	=	the Arrhenius constant (frequency factor)
$\boldsymbol{E = E_a/R}$	=	the activation energy (E_a) divided by the universal gas constant (R)

Note: In this specification, for a time value, t, or an index for a time value (i.e., q), the use of the subscript T (e.g., $t_{i,T}$ or q_T) signifies values that are part of a data series which is dependent on temperature T.

6.1.2. Sampling time points

The DNA shall be sampled at a set of time points, $\{t_0^-, t_0^+, t_1, t_2, \ldots, t_q\}$, as follows (also see Figure 6), where:

- $t_{i,T}$, for $1 \leq i \leq q_T$, is the set of equally spaced time points at which sample data is to be collected for the aging part of the experiment, where Δt_T is a constant for the kinetic being run (i.e., $t_{q,T} = t_{1,T} + (q_T - 1)\Delta t_T$); and
- q_T shall be ≥ 4

An overview of the timing model is shown in Figure 6. In this example, DNA degradation kinetics are run at three temperatures: T_1, T_2, T_3. Dashed lines represent either discrete, or unmodeled kinetics, while solid lines represent the continuous degradation kinetics that are the basis of calculating the degradation rate constants (k_T). DNA is prepared at time t_0^- and inserted into the DCS by t_0^+ at which point the aging conditions are applied; i.e., the DNA spends zero time in the aging conditions prior to $t = t_0^+$. Between t_0^+ and

$t_{1,T}$(shaded yellow) empirical observations show a rapid change in the degradation rate before reaching the steady state modeled by the Arrhenius equation (shaded light red).

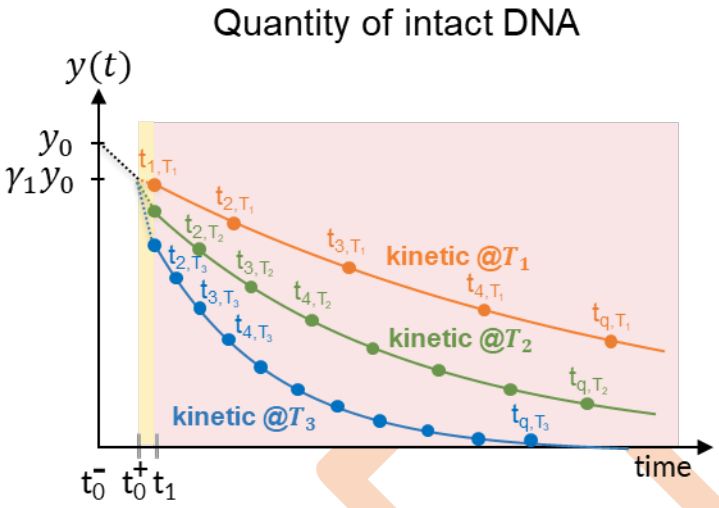

Figure 6 – DNA degradation kinetics at temperatures (T_1, T_2, T_3), fixed relative humidity

An example of the timing model being used is as follows:

Let us say we are running three kinetics as shown in Figure 6, where $T_1 = 65C$, $T_2 = 75C$, $and\ T_3 = 85C$. First, a solution of 10 µM DNA ($t = t_0^-$) is prepared with the intent of dispensing 10 µl per DCS Unit in the experiment (one DCS Unit per temperature point). This means that $y_0 = y(t_0^-) = 100$ pmols, for all three kinetics. Following this, the DNA is stored in each DCS Unit, the DCS Units are placed in ovens, and the aging conditions are applied ($t = t_0^+$) for each kinetic (65°C, 75°C, and 85°C).

Also, at time t_0^+, one additional sample (plus replicates) is held back for pre-aging quantification (the amount of material lost in preserving the sample in the DCS Units, but before aging conditions are applied). Let us say that the pre-aging measurement at t_0^+ is 75 pmol. This means that the constant $\gamma_1 = y(t_0^+)/y_0) = 75/100 = 0.75$. This applies to all three kinetics, as the pre-aging/preservation stage is considered non-temperature dependent.

Twenty-four hours later (i.e. for this example, $t_{1,T} - t_0^+ = 24$ hrs) the experimenter takes one sample per kinetic (plus replicates) and quantifies them. Let us say that the measurements for T_3 (85°C) yield an average of 50 pmol. This means that, for the kinetic @T_3, $\gamma_2 = y(t_{1,T})/(t_0^+) = 50/\gamma_1 100 = 0.666$. The constant γ_2 is calculated the same way for the other two kinetics (T_1, T_2).

Following the measurement at t = $t_{1,T}$, the experimenter takes samples (plus replicates) at regular intervals. The three data series generated by the measurements taken at $t_{1,T}, t_{2,T}, \ldots, t_{q,T}$ for $T = \{T_1, T_2, T_3\}$ are then used for fitting E and A, extrapolating the

degradation rate at 25°C from the Arrhenius equation, and finally calculating half-life at 25°C.

6.1.3. Temperature

Temperature is the acceleration parameter. Using high temperature (at constant humidity) to accelerate the reaction rate enables the extrapolation of degradation rate constants at any temperature from experiments with manageably short experimental windows.

For a constant controlled humidity level, a minimum of 3 temperatures shall be employed, and each temperature point shall be a minimum of 10 degrees Celsius apart from one to the next.

The actual observed temperature shall be continuously monitored, recommended at no more than 5 minute intervals, and shall not deviate more than +/- 2°C except for the 10 minutes after each time point when samples are removed for testing.

If more than 2% of the temperature readings between time points are out of the +/- 2°C specification, the tester shall use a heat sink (e.g., aluminum block or sand batch) to ensure samples stay at a constant temperature.

6.1.4. Relative Humidity (RH)

RH is a controlled parameter during the accelerated aging experiment and shall be held constant at each temperature. RH variance shall be no larger than that specified by the chamber manufacturer, except for the 10 minutes after each time point when samples are removed for testing.

The experiment shall be run at two mandatory RH points (50% and 75%). More RH points are optional. If added, optional RH points shall be at a minimum 20% from the two mandatory points, and from each other.

6.1.5. Calibration Kinetics

Establishing values for the aging conditions needed to get good results during accelerated aging (T, t_q, Δt, etc.) is dependent on the characteristics of a particular DCS (degradation rate constant, activation energy, drying or other physical storage requirements, etc.). If prior experimental knowledge of the DCS degradation characteristics is not available, an initial aging calibration run, or multiple calibration runs, may be needed to pick the accelerated aging parameter value ranges that ensure a good final curve fit and half-life results.

An example aging calibration run:

1. Setup DNA in storage conditions
2. Setting t_1=1 day to avoid any fast degradation dynamics, collect at least 3 time points (e.g., t_1=1 day, t_2=15 days, t_3= 30 days) over at least two temperatures (e.g., 70°C, 90°C). Temperatures and times should be based on prior performance expectations.
3. Fit data to the Arrhenius equation.

4. If uncertainty is too high and or excessive or insufficient degradation has occurred repeat steps 1-3 with updated performance expectations.

An aging calibration run is optional; however, the DCS provider should consider providing a description of the aging calibration or other data or studies to explain the aging parameter choices used.

6.2. Storing the media in the DCS

Prior to aging conditions being applied, aliquots of the Reference DNA Pool are stored in a population of DCS Units, according to how many temperature and humidity points are needed to conduct the aging experiment. The aliquots shall be identical in quantity and quality.

6.3. Pre-Aging media loss quantification (i.e., DCS Recovery Rate %)

To properly characterize the Half-Life of a DCS using this specification, it is important to ensure that any material loss prior to the beginning of the aging phase of the experiment is accounted for. It is for this reason that we defined the γ_1 (DCS Recovery Rate %) constant in the degradation model. In addition to its use as a constant in the degradation model, DCS Recovery Rate % is a valuable metric in and of itself to characterize the quality of a DCS.

The DCS Recovery Rate % shall be calculated and reported.

Two measurements are required to calculate DCS Recovery Rate % ($\gamma_1 = y(t_0^+)/y(t_0^-)$)

1) A measurement of the amount of intact DNA before the media is stored in the DCS, $y(t_0^-)$; and
2) A measurement of the amount of remaining intact DNA after the media is stored in the DCS but before experimental conditions are applied, $y(t_0^+)$.

For (1) and (2), a minimum of three measurements (experimental replicates) are required.

The quantification method used for these measurements shall be the same as that used throughout the rest of the aging experiment, regarding both instrumentation and minimum number of technical replicates used.

The data collected to calculate the DCS Recovery Rate % shall not be considered in degradation rate calculations.

6.4. Aging and Measurement

To run the aging experiment, the DCS Units are placed in incubators and the aging conditions are applied ($t = t_0^+$).

For sampling time points $t_{1,T}, t_{2,T}, \ldots, t_{q,T}$, for all temperature points T being run in the experiment (see Figure 6),, samples shall be retrieved and measured, or stored for later measurement at 4°C.

The procedure used to take the samples must be reproducible; such reproducibility is of paramount importance since the amount of retrieved DNA has a direct impact on the subsequent quantification of intact strands.

Measurement shall be done with qPCR or dPCR using state of the art protocols (e.g., MIQE guidelines), at least for technical duplicates.

The entire measurement and retrieval procedure (volumes, temperature, incubation time if any, number of mixing steps) shall be documented.

6.4.1. qPCR Assay Calibration

If using qPCR, calibration is required.

A minimum of five Calibration Standards shall be used, starting with a concentration equal to the maximum possible concentration to be measured and serially diluted by an order of magnitude. A sixth Negative Control (i.e., 0 ng/uL), shall be used to calibrate the qPCR readout and ensure accurate quantification.

The preparation process and source for the Calibration Standards shall be documented, and the following requirements apply:

1. Calibration Standards shall use the same primer sequences as those in the Reference Media Pool but their payload may consist of a fixed unique sequence (i.e., need not be random).
2. The overall length of the target/amplified sequence of the Calibration Standard (primer plus payload) shall have the same length as the target/amplified sequence of the strands in the Reference Media Pool.
3. The DNA in the Calibration Standards shall be of the same purity as the DNA in the Reference Media Pool, to ensure that qPCR efficiency is not affected by different purity levels, which could result in skewed measurement data. To ensure this, either:
    a. the Calibration Standards shall be treated using the DCS Preservation Method; or
    b. the samples taken from the DCS during aging shall be purified to remove any substances added by the Preservation Method that would affect the qPCR results.
4. The experimenter shall quantify the Calibration Standards using fluorometry prior to the experiment onset to validate that the measured concentration of the Calibration Standard matches the concentration as specified by the Calibration Standard's source.

The Calibration Standards are used to determine:

1. **Lower limit:** qPCR measurements of the Reference Media Pool with critical threshold (Ct) values equal to or greater than the Negative Control (i.e., 0ng/ul) shall be treated as beyond the lower limit of sensitivity.
2. **Upper limit:** The readouts shall not exceed the upper limits of detection as given by the instrument manufacturer and must have Ct values less than or equal to the minimum Calibration Standard Ct value.

3. **PCR efficiency:** The PCR efficiency of the standards shall be reported and should be kept as close to equal to the PCR efficiency of the Reference Media Pool as possible. If there is a chance that the Reference Media Pool contains agents that will have an impact on PCR efficiency, so that the calibration curve of the Calibration Standards is not giving an accurate number of copies in the Reference Pool, then the tester should change the Calibration Standards to reconcile this.

During calibration, a minimum of two readings (technical replicates) per sample shall be used in each assay measurement to ensure consistent assay readings.

6.5. Analysis & Reporting

6.5.1. This section specifies the analysis and reporting requirements for the standard. Estimation of degradation rates (k_T)

To calculate the degradation rates (k_T), the measured data series for each kinetic, y(t) for t = $t_{1,T}, t_{2,T}, \ldots, t_{q,T}$ for each value of T (i.e., the part of the experiment where the degradation rates are assumed to have become constant - see Figure 6), shall be fit to the following equation, for each temperature T:

$$ytT = \mathrm{y}(\mathrm{t}_{1,T})\exp(-nk_T(t_T - t_{1,T})) \qquad (2)$$

From each exponential equation, the k_T (expressed in units of per nucleotides per second) is estimated.

The tester shall report on the quality fit of the curve fit (Residuals) and the method used to calculate it (e.g., R-square)

6.5.2. Arrhenius Plot and Extrapolation of degradation rate

For each RH value, an Arrhenius plot is built as follows:

1) the -log10(k_T) are calculated and plotted versus 1/T (in °K).
2) the parameters of the Arrhenius equation are then estimated and with this equation, the k_T can be found by extrapolation to different temperatures, in particular 25°C.

The tester shall report on the quality fit of the curve fit (Residuals) and the method used to calculate it (e.g., R-square)

The tester should document an estimate of standard error for variables A, E in equation (1) and use an estimate of standard error to compute a confidence interval for a 25°C degradation rate estimate.

6.5.3. Half-Life Calculation

Once all of the k_T are known, the half-lives ($t_{1/2}$) of DNA fragments of a size of n nucleotides can be calculated as follows:

$$\frac{y}{y_0} = \exp\left(-nk_T t_{1/2}\right) = 0.5 \qquad (3)$$

and then

$$t_{1/2} = \frac{\ln 0.5}{-nk_T} \qquad (4)$$

with time $t_{1/2}$ in seconds and k_T in per nucleotides per second.

The tester should document an estimate of standard error of the half-life estimate at 25°C to compute a confidence interval.

7. Real Time Aging Protocol

It is highly recommended that, in addition to running an accelerated wear test, which is optimized to enable a high-confidence extrapolation of half-life with a relatively short experiment duration, the DCS vendor also run a real time aging experiment; that is, an aging experiment where the DCS is stored in conditions approximating the actual use case envisioned for the DCS.

This specification does not require real time testing, nor does it prescribe a temperature/humidity/timepoint regimen for real time testing, since the preservation methods used by different DCSs vary, as do the actual use cases envisioned by an end-user. The recommendation is to run the test as part of a general product quality and verification regime, even if the results are that no measurable degradation is detected. For example, for a DCS with a claimed half-life of 10 years at 25°C for a 150 nt-long fragment, the proportion of intact fragments should still be 87% after 2 years. A significant deviation from this result should lead to re-evaluation.

Real time testing can be conducted under non-controlled room temperature laboratory conditions, in controlled conditions under which the archive is planned to be stored, or any other conditions meaningful to confirming the survivability of DNA media in the DCS during actual use. If real time testing is conducted, the DCS vendor shall document the conditions used.

8. Alternative Methods

It might not be possible to use the Reference Media Pool as specified in section 5.

First, in the context of evaluating an end-to-end DNA data storage system, it might not be practical or even possible to test the DCS integrated within the end-to-end system using the standard Reference Media Pool.

Second, the properties of some DCSs might make the use of 150nt fragments cause the experiment duration to be very long:

- Some DCSs may not be able to withstand elevated temperatures; for example, some organic matrices may lose their protective properties above a certain temperature.
- For a DCS ensuring very robust preservation (e.g., one that completely isolates the media from the atmosphere), the kinetics with 150nt long fragments might also be very long, even at elevated temperatures.

These factors might make it necessary to consider using longer fragments or other forms of DNA that are more sensitive to degradation at lower temperatures. This, in turn, may require the use of measurement methods more sensitive than qPCR or dPCR, such as gel electrophoresis.

For example, a DCS could be tested using supercoiled plasmid DNA and measuring the rate of appearance of the first break by electrophoresis. Indeed, one cut on one of the 2 strands will lead to the relaxing of the supercoiled DNA into the open circle form, both distinguishable on a gel (Figure 7).

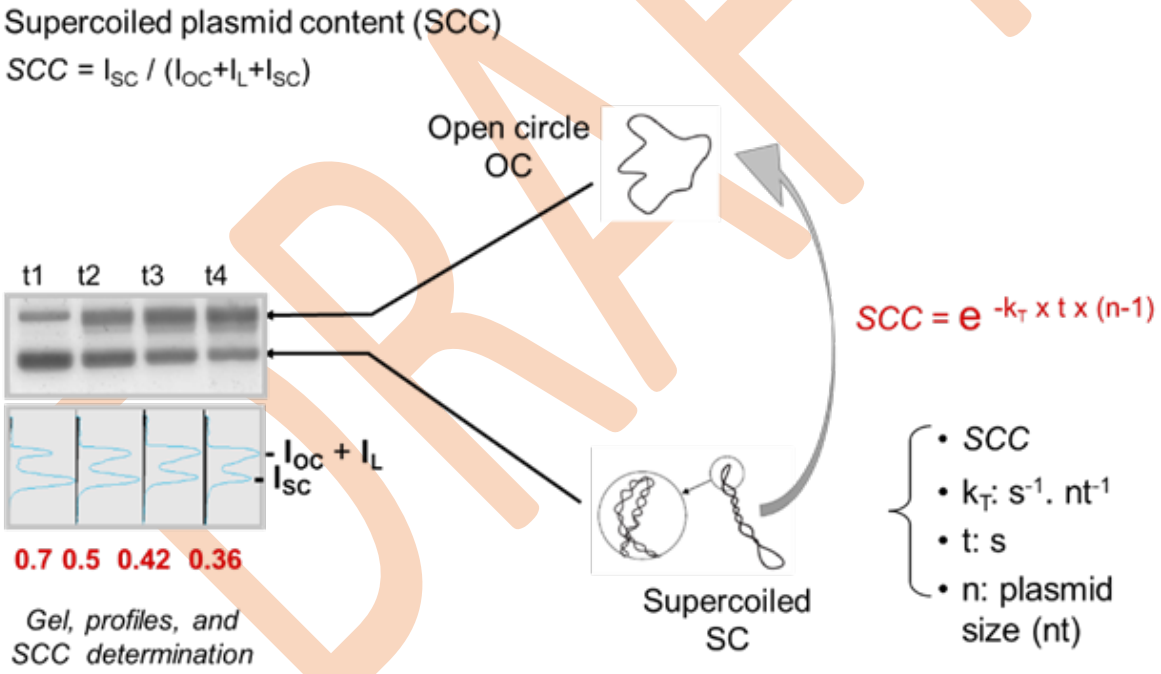

Figure 7 - Supercoiled plasmid aging example

Examples of the use of supercoiled plasmid DNA in degradation kinetics are found in Molina [31] (Fig 3) and Bonnet [3] (Fig 7).

For the reasons above, or any others which render the standard Reference Media Pool impractical, this standard allows alternatives to the Reference Media Pool (Alternative Media Pool), and alternative measurement methods as may be required for these alternatives. This is justified because, in general (see Section 9 and Section 4) it has been shown that the rate of “unrecoverable” errors (i.e., strand breaks) during long term

storage is not affected by the sequence of bases in the strands that were stored. It is thus possible to get valid aging results (i.e., ones that can yield a quality Arrhenius curve fit) with DNA sequences that deviate in form from that specified by the Reference Media Pool.

The aging protocol defined in Section 6 shall be followed when using Alternative Methods, except for the changes required to accommodate the form and format of the molecules in an Alternative Media Pool, and the measurement techniques, other than qPCR or dPCR, that may be required to read the DNA in an Alternative Media Pool.

The reported results using Alternative Methods should be normalized for strands of length 150 nucleotides, to enable comparison with the standard method.

9. Appendix A - DNA Stability Research Background

As described in section 4 (General Principles), this specification assumes that by counting DNA strand breaks during storage in a standard manner, we can objectively compare different storage/preservation methods, independent of the end-to-end DNA data storage system. This appendix describes some of the research on which this assumption was based.

9.1. DNA Molecular Degradation

There are a variety of DNA degradation events documented in the literature (Figure 8). In the context of preserving a pool of DNA molecules encoded with digital data, we are concerned with degradation events which ultimately lead to a break in the Sugar-Phosphate chain of a molecule, because DNA strands with broken chains cannot be read (sequenced) to recover encoded data, whereas strands with intact chains can be. To be more precise, strands with chain breaks will not polymerize (PCR block) and can thus not be sequenced; as regards DNA data storage such strands are, for all intents and purposes, lost.

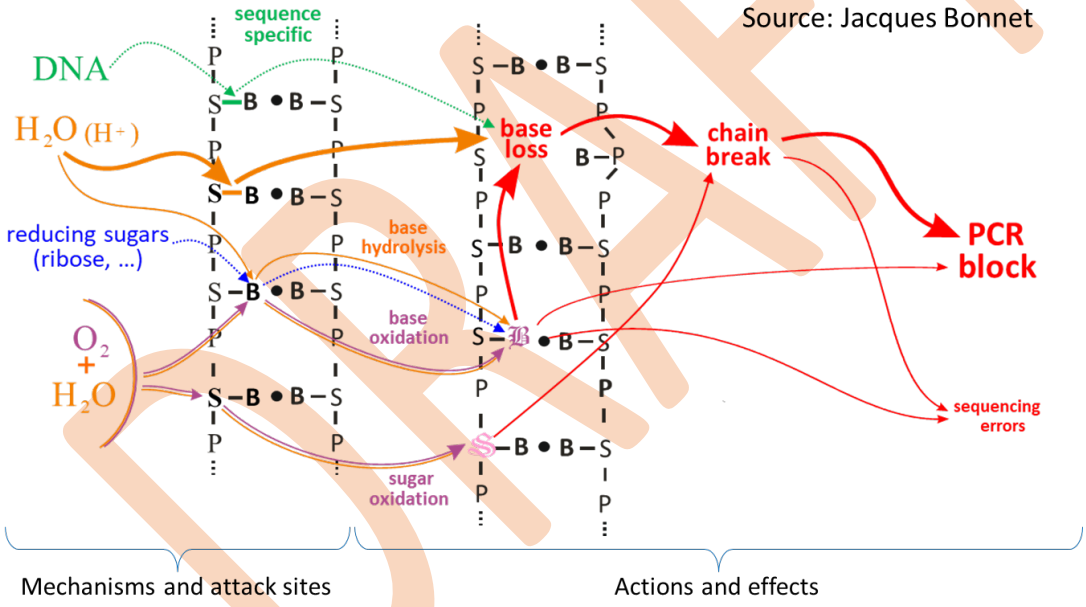

Figure 8 - Abstracted view of DNA molecular degradation mechanisms

As shown in Figure 8, while many precursor events can lead to chain breaks, by far the predominant types of events are those leading to base loss through depurination, and H_2O is involved in nearly all of these events, either directly or indirectly. It is for this reason that the predominant methods for preserving DNA involve either drying or methods which completely encapsulate the media from atmospheric moisture, thus avoiding any second order effects (Table 1).

Preservation Category	Preservation Substrate/Method		Drying	Protection from atmosphere	Stability estimation
Chemical encapsulation	Encapsulation in salts	[12, 16]	✓		Accelerated aging
	Degradable Polymer Microcapsules	[32]			
	Cationic Diblock Copolymer	[33]			
Physical encapsulation	Silica nanoparticles	[1]		✓	Arrhenius
	Stainless steel capsules	[3, 7, 30]	✓	✓	Arrhenius
	Magnetic silica nanoparticles	[13]		✓	Accelerated aging
Inclusion in a matrix	DNAstable	[1, 21]	✓		Arrhenius
	Gentegra DNA	[1, 22]	✓		
	Pullulan	[14]	✓		
	Silk	[15]	✓		
	composite nucleic acid-polymer fibers	[34]	✓		Accelerated aging
	300K matrix inclusion	[25]	✓		
Absorption on paper	FTA paper	[1, 23, 24]	✓		Arrhenius
	Chitosan treated paper	[17]	✓		
Dehydration on solid supports	Capillaries	[20]	✓		
	Glass	[26, 27]	✓		
	Tube walls	[28]			
Dissolution in liquid salts	Imidazolium cations	[18]			
	Imidazolium cations	[19]			
Living organism	yeast genome	[36]			
	Ecoli genome	[37]			
	yeast cells	[38]			
	Bacteria	[29]			
DNA beads	Magnetic Bead Spherical Nucleic Acid Microstructure	[35]	✓		Arrhenius

Table 1 - Overview of Preservation Methods in DCS Stability Studies (source: Bonnet, Colotte)

9.2. Digital Data Reliability in DNA

In developing the DNA Stability Evaluation Method for DNA Data Storage Containment systems, we examined the literature on DNA degradation during storage [1-38].

Figure 9, edited from Organick [6], summarizes data demonstrating that, at room temperature, with various preservation methods, it is possible to achieve a very long half-life for intact strands, on the order of decades or centuries.

In addition to the half-life conclusions for different DNA data storage methods, the following subsections detail a few conclusions on accelerated wear methods that underly the premises of this specification.

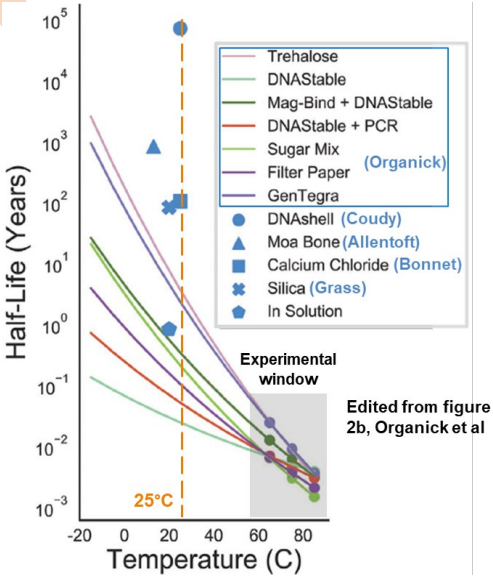

Figure 9 – Half-life plot for 150-nt long oligos in various preservation methods

9.2.1. Grass et al [1] - Can we recover data from the intact strands that survive accelerated wear? Answer: Yes

Grass et al conducted an accelerated wear experiment to show the durability of silica, a "synthetic fossil", for storing digital data in DNA. Figure 10 shows the qPCR results (% of DNA material present, i.e., able to be amplified) for the timepoints in the experiment.

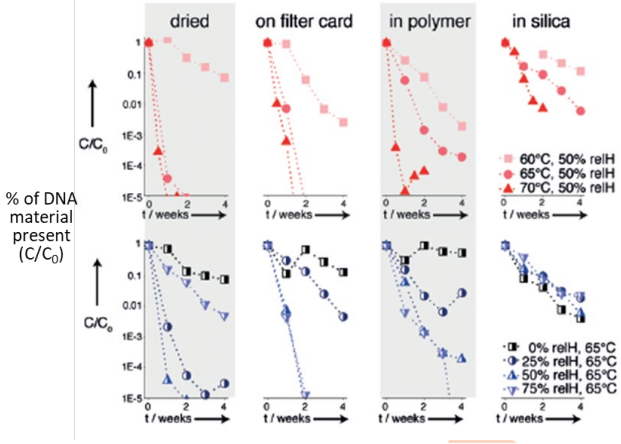

Figure 10 – Concentration over time, at temp/humidity [1]

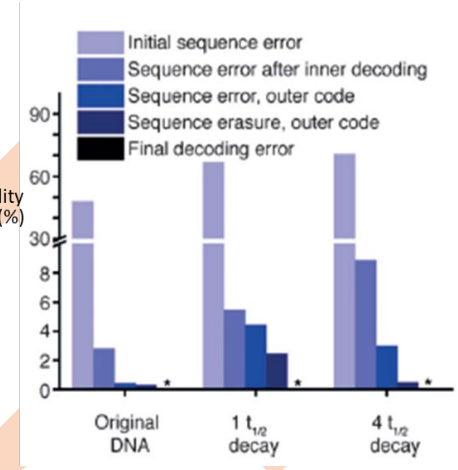

Figure 11 – Data retention after storage [1]

To verify that the intact molecules in the samples retained the original encoded data after accelerated aging, the media was sequenced and decoded. As shown in Figure 11, although there were significantly more errors detected in the media after thermal treatment, in the end, all data was able to be recovered. This result from the study reinforces two conclusions:

1) Enough strands survived the temperatures used in the accelerated aging protocol to validate the efficacy of using a high temperature stress method; and

2) Notwithstanding that we cannot conclusively correlate actual long term (i.e., non-accelerated) aging effects with accelerated aging effects, if strands survive long term storage intact in non-accelerated (i.e., less hostile than accelerated) aging conditions, we'll be able to recover encoded data from them.

9.2.2. Organick et al [6] - Do read errors vary by storage method? Answer: No

Organick et al conducted an accelerated wear experiment on multiple methods of preserving data in DNA (Figure 9). The analysis of error rates in the study is illuminating for the DNA Stability Evaluation Method.

Per figure 12, for all error categories (Insertion, Deletion, Substitution), there was minimal variation in error rates across storage methods. Moreover, per the authors (emphasis added), "Even when looking intently at the substitution rate, which has the most variation, we still observed a maximum difference of just over one percent, which

is not practically significant." Most importantly, despite very different strand breaking rates, "There was no particular storage method(s) that showed more or fewer errors than other methods across the different temperatures and time points, *which suggests that insertion, deletion, and substitution errors are independent from the storage method*".

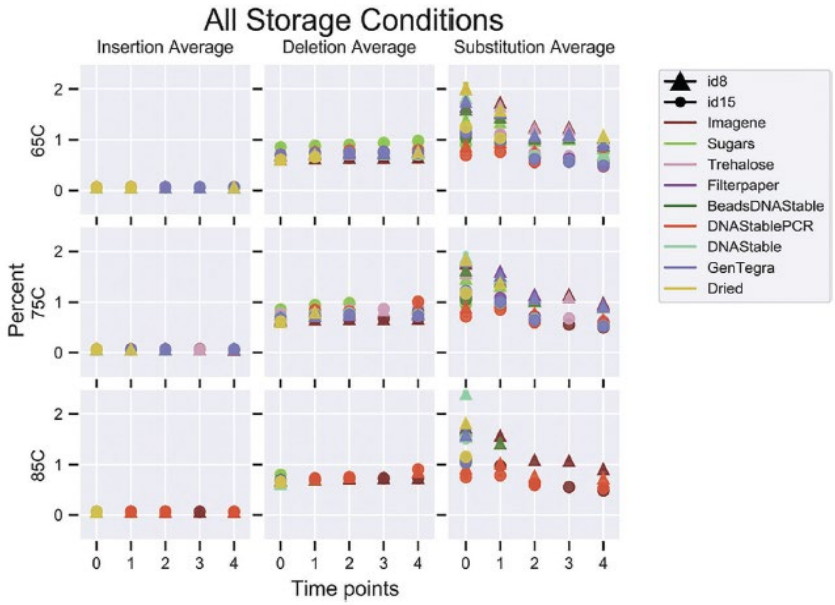

Figure 12 – Variation of error rates across different preservation methods

These results indicate that the errors created, and able to be corrected, by the synthesis, retrieval, and sequencing part of an end-to-end DNA data storage system can be ignored for the purposes of evaluating and comparing DCS options for their durability claims.

9.2.3. Organick et al [6] – Do certain sequences cause errors with specific storage methods? Answer: No

The authors also checked to see if the survivability of any particular sequence in the stored DNA strands was affected by the storage method.

The total # of sequences found missing during sequencing (across all methods, all time points, all temperatures) were analyzed for sequence loss. (Figure 13). As the authors concluded (emphasis added), "If the total number of sequences missing increased after the pre-aging time point 0, we could hypothesize that there was some sequence-dependent degradation as more-vulnerable sequences degraded. However, we observed no difference in the number of sequences missing across all time points. *This suggests that sequence loss is stochastic across all storage methods*."

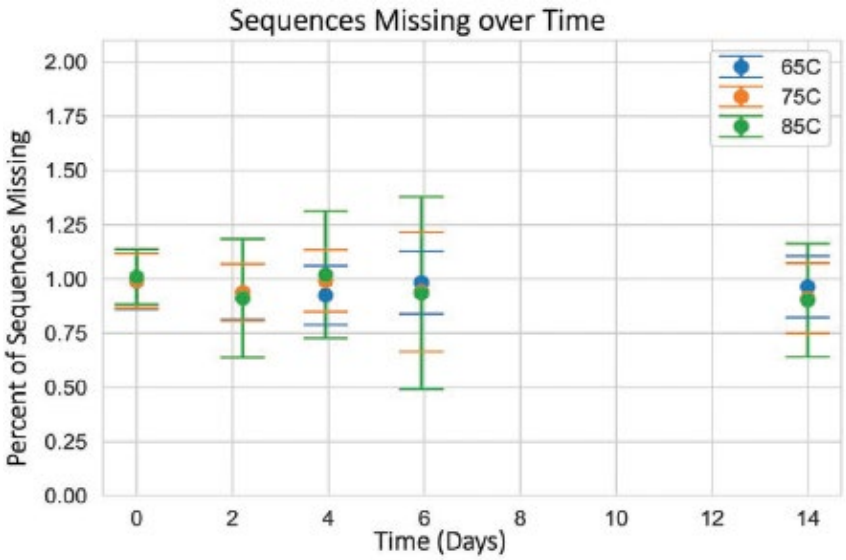

Figure 13 – Sequences loss (and found) over time [6]

This finding was reinforced by a separate finding in the study that individual sequences missing at a particular time point had a >90% probability of reappearing in other time points and being successfully sequenced, indicating that a sequence missing at one time point was missed simply due to sampling error.

These results, that there is no apparent sequence bias as related to storage method, further reinforce that one can define a standard stability evaluation methodology that is independent of the effects of synthesis, retrieval, and sequencing.

9.3. Summary

While new storage methods are always being developed, and while more research will continue to shed light on DNA degradation kinetics, the results from known studies indicate that counting strand breaks using qPCR at various time points during thermal treatment, extrapolating a degradation rate, and then extrapolating half-life based on an Arrhenius curve fit, is a valid way of enabling “apples-to-apples” data reliability comparison of DNA storage/containment systems.