# Information-Theoretic Foundations of DNA Data Storage

**Other titles in Foundations and Trends® in Communications and Information Theory**

*Asymptotic Frame Theory for Analog Coding*
Marina Haikin, Matan Gavish, Dustin G. Mixon and Ram Zamir
ISBN: 978-1-68083-908-1

*Modeling and Optimization of Latency in Erasure-coded Storage Systems*
Vaneet Aggarwal and Tian Lan
ISBN: 978-1-68083-842-8

*An Algebraic and Probabilistic Framework for Network Information Theory*
S. Sandeep Pradhan, Arun Padakandla and Farhad Shirani
ISBN: 978-1-68083-766-7

*Theoretical Foundations of Adversarial Binary Detection*
Mauro Barni and Benedetta Tondi
ISBN: 978-1-68083-764-3

*Polynomial Methods in Statistical Inference*
Yihong Wu and Pengkun Yang
ISBN: 978-1-68083-730-8

*Information-Theoretic Foundations of Mismatched Decoding*
Jonathan Scarlett, Albert Guillen i Fabregas, Anelia Somekh-Baruch and Alfonso Martinez
ISBN: 978-1-68083-712-4

# Information-Theoretic Foundations of DNA Data Storage

**Ilan Shomorony**
University of Illinois at Urbana-Champaign
USA
ilans@illinois.edu

**Reinhard Heckel**
Technical University of Munich
Germany
reinhard.heckel@tum.de

now
the essence of knowledge
Boston — Delft

**Foundations and Trends® in Communications and Information Theory**

*Published, sold and distributed by:*
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

*Outside North America:*
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

I. Shomorony and R. Heckel. *Information-Theoretic Foundations of DNA Data Storage*. Foundations and Trends® in Communications and Information Theory, vol. 19, no. 1, pp. 1–106, 2022.

ISBN: 978-1-68083-957-9

**Foundations and Trends® in Communications and Information Theory**
Volume 19, Issue 1, 2022

# Editorial Board

# Editorial Scope

## Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

## Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2022, Volume 19, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328 . Also available as a combined paper and online subscription.

# Contents

# Information-Theoretic Foundations of DNA Data Storage

Ilan Shomorony[1] and Reinhard Heckel[2]

[1] *University of Illinois at Urbana-Champaign, USA; ilans@illinois.edu*
[2] *Technical University of Munich, Germany, and Rice University, USA; reinhard.heckel@tum.de*

ABSTRACT

Due to its longevity and enormous information density, DNA is an attractive medium for archival data storage. Natural DNA more than 700.000 years old has been recovered, and about 5 grams of DNA can in principle hold a Zetabyte of digital information, orders of magnitude more than what is achieved on conventional storage media. Thanks to rapid technological advances, DNA storage is becoming practically feasible, as demonstrated by a number of experimental storage systems, making it a promising solution for our society's increasing need of data storage.

While in living things, DNA molecules can consist of millions of nucleotides, due to technological constraints, in practice, data is stored on many short DNA molecules, which are preserved in a DNA pool and cannot be spatially ordered. Moreover, imperfections in sequencing, synthesis, and handling, as well as DNA decay during storage, introduce random noise into the system, making the task of reliably storing and retrieving information in DNA challenging.

This unique setup raises a natural information-theoretic question: how much information can be reliably stored on

Ilan Shomorony and Reinhard Heckel (2022), "Information-Theoretic Foundations of DNA Data Storage", Foundations and Trends® in Communications and Information Theory: Vol. 19, No. 1, pp 1–106. DOI: 10.1561/0100000117.

and reconstructed from millions of short noisy sequences? The goal of this monograph is to address this question by discussing the fundamental limits of storing information on DNA. Motivated by current technological constraints on DNA synthesis and sequencing, we propose a probabilistic channel model that captures three key distinctive aspects of the DNA storage systems: (1) the data is written onto many short DNA molecules that are stored in an unordered fashion; (2) the molecules are corrupted by noise and (3) the data is read by randomly sampling from the DNA pool. Our goal is to investigate the impact of each of these key aspects on the capacity of the DNA storage system. Rather than focusing on coding-theoretic considerations and computationally efficient encoding and decoding, we aim to build an information-theoretic foundation for the analysis of these channels, developing tools for achievability and converse arguments.

# 1

# Introduction

In recent years, the number of applications that require or are enabled by digital data storage and the amount of data generated by a variety of technologies have increased dramatically. This has spurred significant interest in new storage technologies beyond hard drives, magnetic tapes, and memory chips. In this context, DNA—the molecule that carries the genetic instructions of all living organisms—emerged as a promising storage medium. DNA has two key advantages over conventional digital storage technologies: extreme longevity and information density. This makes DNA an interesting storage medium, particularly for archival storage.

Data on DNA can last very long, if stored appropriately, as nature itself proves. As demonstrated by recently sequenced DNA extracted from a mammoth tooth found in the Siberian permafrost [113], the information in a DNA molecule can be preserved for more than a million years. In contrast, information on memory chips lasts no more than a few years, and data on hard drives and magnetic tapes lasts no more than a few decades. While conventional storage media could be redesigned to preserve data longer, the longevity of DNA is currently unmatched, as illustrated in Figure 1.1(a).

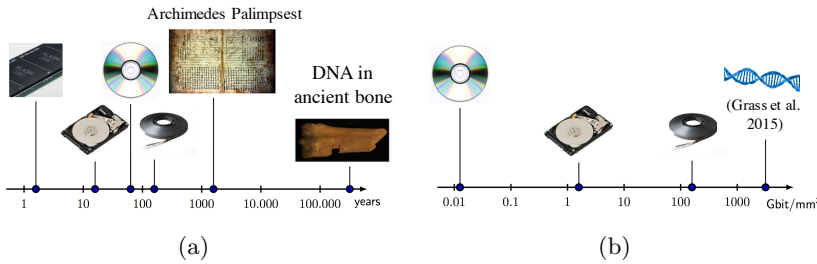


**Figure 1.1:** (a) Longevity of different data storage media. Archimedes Palimpsest containing "The Methods of Mechanical Theorems" survived more than 1000 years. Recently, 700,000 year old DNA from ancient horse bones has been successfully sequenced [91]. (b) Information density of different storage media. The proof-of-concept DNA-based storage system of [39] achieved an information density that is over one order of magnitude higher that of magnetic tape.

The information density of DNA is also extremely large. Just 5 grams of DNA contain about $4 \cdot 10^{21}$ nucleotides, which in principle could hold $8 \cdot 10^{21}$ bits, or one zettabyte. In a practical system, the redundancy required for error-correction coding required to build a reliable system reduces these numbers, but we can achieve information densities orders of magnitude larger than the highest information densities achieved on hard drives and tapes, as shown in Figure 1.1(b).

## 1.1 A brief history of DNA data storage

Computer scientists and engineers have dreamed of harnessing DNA's storage capabilities already in the 60s [9], [79], and in recent years DNA data storage, or more broadly, molecular information storage, developed into an active field of research. In 2012 and 2013 groups lead by Church [27] and Goldman [38] independently stored about a megabyte of data in DNA. Later, Grass, Heckel, Puddu, *et al.* [39] demonstrated that millennia-long storage times are possible by protecting the data both physically and information-theoretically, and designing a robust DNA data storage scheme using error-correcting codes. Yazdi, Yuan, Ma, *et al.* [119] showed how to selectively access files, and Erlich and Zielinski [33] demonstrated that a DNA storage system can achieve very high information densities, close to the absolute maximum of two bits

per nucleotide. In 2018, Organick, Ang, Chen, *et al.* [80] scaled up these techniques and stored about 200 megabytes of data. Together, these and other works demonstrated that writing, storing, and retrieving data using DNA as a medium is possible with today's technology, and achieves information densities and information lifetimes that are far beyond what state-of-the-art tapes and discs achieve.

DNA is a long molecule made up of four nucleotides (Adenine, Cytosine, Guanine, and Thymine) and, for storage purposes, can be viewed as a string over a four-letter alphabet. However, there are hard technological constraints for writing on DNA and for reading DNA, which need to be considered in the design of a practical DNA-based storage system. While in a living cell a DNA molecule may consist of millions of bases (the human chromosome 1, for example, is 250 million bases long), due to practical technological constraints, it is difficult and inefficient to synthesize long strands of DNA. For that reason, all recent works that have demonstrated working DNA storage systems stored information on molecules of no longer than 100-200 nucleotides [27], [33], [38], [39], [80], [119].

The process of determining the order of nucleotides in a DNA molecule, or DNA sequencing, suffers from similar length constraints. State-of-the-art sequencing platforms such as Illumina cannot sequence DNA segments longer than a few hundred nucleotides. While recently developed, so-called third-generation technologies such as Pacific Biosciences and Oxford Nanopore can provide reads that are several thousand bases long, their error rates and reading costs are significantly higher [30], [117].

Due to those constraints in the length of the DNA molecules that can be synthesized, stored and sequenced, a practical DNA-based storage system consists of many *short* DNA molecules stored in an *unordered fashion* in a solution.

Another technological limitation to DNA-based storage comes from the fact that state-of-the-art sequencing technologies rely on the *shotgun* sequencing paradigm. This corresponds to (randomly) sampling and reading sequences from the DNA pool. Furthermore, sequencing is usually preceded by several cycles of Polymerase Chain Reaction (PCR) amplification. In each cycle, the amount of each DNA molecule is

**Table 1.1:** Parameters of a few DNA storage systems using array-based synthesis, listed chronologically.

| | length of seqs. | number of seqs | data stored (in MB) | error correction |
|---|---|---|---|---|
| Church, Gao, and Kosuri [27] | 115 | 54,898 | 0.65 | None |
| Goldman, Bertone, Chen, *et al.* [38] | 117 | 153,335 | 0.75 | Repetition |
| Grass, Heckel, Puddu, *et al.* [39] | 117 | 4,991 | 0.08 | RS |
| Blawat, Gaedke, Hütter, *et al.* [12] | 190 | 900,000 | 22 | RS |
| Bornholt, Lopez, Carmean, *et al.* [14] | 120 | 45,652 | 0.15 | RS |
| Erlich and Zielinski [33] | 152 | 72,000 | 2.14 | Fountain |
| Organick, Ang, Chen, *et al.* [80] | 150 | $13.4 \cdot 10^9$ | 200.2 | RS |
| Chandak, Tatwawadi, Lau, *et al.* [17] | 150 | 13,716 | 0.192 | LDPC |
| Heckel and Grass [46] | 105 | $3.88 \cdot 10^9$ | 63.1 | RS |
| Antkowiak, Lietard, Darestani, *et al.* [7] | 60 | 16,383 | 0.1 | RS |

amplified by a factor between 1.6 and 1.8, and this factor can be sequence-dependent, thus leading to very different concentration of distinct DNA molecules. Last but not least, the DNA molecules in a DNA-based storage system are subject to errors such as insertions, deletions, and substitutions of nucleotides at the time of synthesis, during the storage period, and during sequencing.

## 1.2 Overview of existing DNA storage systems

In the last decade, several research groups have shown that using today's technologies, it is possible to store on the order of megabytes of data reliably. All systems that stored megabytes of data and demonstrated correct recovery relied on array-based synthesis, where data is stored on a set of many short sequences. In Table 1.1, we list several of these implementations with some of their key parameters in chronological order. All systems listed in the table stored a unique index on each sequence to deal with the shuffling character of the channel at decoding, and starting from Grass, Heckel, Puddu, *et al.* [39], all systems used outer error-correcting codes to deal with the loss of individual sequences.

There are alternatives to DNA data storage based on array-based synthesis. For example, Yazdi, Yuan, Ma, *et al.* [119] used column-based synthesis, where individual sequences are generated one-by-one. With this approach, the authors successfully stored 0.017 MB on 32

sequences of length 1000 each. Another example is the implementation by Lee, Kalhor, Goela, *et al.* [63], which stored 18 bytes using enzymatic synthesis. Tabatabaei, Wang, Athreya, *et al.* [109] avoided synthesis altogether, and stored data in form of nicks at certain positions on the backbone of existing DNA. Yet another example is the work by Yim, McBee, Song, *et al.* [121], which stored data (about 72 bits) in living cells via CRISPR arrays. All these approaches are conceptually interesting alternatives to array-based synthesis, but scaling them to store more than a few bytes of data currently looks challenging.

Array-based synthesis generates many copies of each sequence, by building up each sequence on a different spot of the array. It is also possible to modify array-based synthesis to grow sequences at one spot of the array so that a predefined fraction of the sequences contain, say, nucleotide A but others contain, say, nucleotide G at a given position. Anavy, Vaknin, Atar, *et al.* [5] explored this idea and proposed the notion of *composite DNA letters* to reduce the number of synthesis cycles. However, the use of composite letters comes at a higher sequencing cost to guarantee that composite letters can be identified from the sequenced DNA molecules, and also comes at a higher decoding complexity.

While the proof-of-concept implementations listed in Table 1.1 demonstrate that DNA storage can be practical, their overall cost is still an obstacle for them to become practically viable. For example, using the architecture proposed in Erlich and Zielinski [33], the estimated cost of synthesizing 1GB of data was $3.27 million [33, Supplementary Material], and the current cost of storing a Megabyte of DNA is around $500 [7]. However, until DNA storage becomes viable for commercial archival storage applications, there are already applications of DNA storage that are not possible with other storage media. For example, Koch, Gantenbein, Masania, *et al.* [54] demonstrated that data stored in DNA can be embedded into any product made of plastic, providing information about the product inside the product.

## 1.3 Information Theory and DNA storage

DNA-based storage is a fundamentally new way of storing data, due to the way DNA is written, stored, and read. The technology is still

under development, and details such as error profiles, the exact length of synthesized DNA molecules, and the sequencing throughput are likely to change. An information-theoretic perspective and an understanding of the fundamental limits of DNA storage will enable a system design based on key conceptual insights and tradeoffs.

**Joint design of physical system and coding schemes:** Conceptual advances in terms of how to optimally code for this new storage paradigm can inform biochemists in their development of new synthesis and sequencing technologies for building DNA storage systems. For example, is it worth developing very expensive technologies to allow long DNA molecules to be synthesized? Or is it possible to store data at high densities with very short molecules? An information-theoretic perspective may provide the foundations to answering these questions, enabling the design of an efficient "physical layer" for DNA storage systems.

**Emerging technologies beyond DNA:** There are many other interesting media for future storage. For example, synthetic polymers are also a potential substrate for data storage [83], in which case tandem mass spectrometry could be used instead of sequencing for data retrieval [62]. Futhermore, the idea of storing data in quartz glass [6] also promises to achieve incredibly high information densities.

As new technologies are proposed and compared, it is important to obtain a basic understanding of their capabilities. In this context, an information-theoretic perspective may allow these emerging approaches to be compared at a more fundamental level, rather than based on specific prototypes. Moreover, design principles may be transferable between them, and different technologies may be more suitable to specific applications depending on basic tradeoffs between cost, computational complexity, and reading and writing speeds. In Section 7.4 we briefly discuss additional storage capacity problems that are motivated by recent technological advances.

**Connections with classical information theory problems:** The growing interest on DNA data storage has sparked renewed interest in classical

problems in information theory. As several DNA sequencing technologies suffer from insertion and deletion errors, new attention has been given to the capacity of insertion/deletion channels and "sticky" channels [25]. Moreover, many of the topics explored in this monograph will be connected with permutation channels [3], [10] and, in Section 7.3, we discuss a connection between DNA storage in the short-molecule regime and discrete-time Poisson channels [60].

## 1.4 Organization of this monograph

In Section 2, we formalize and discuss a general class of channels to model DNA storage systems. This general model, called the noisy shuffling-sampling channel, will be the main object of our information-theoretic analysis of DNA-based data storage. In Section 3, we start the exploration of the capacity of DNA storage systems by studying a simple noiseless shuffling-sampling channel, where the input sequences are shuffled and sampled before being observed at the channel output. The capacity expression provides a precise understanding of the storage rate costs that the shuffling and sampling operations incur, and provide intuition on how to develop optimal codes for such channels. In the same section, we also consider channels that break the sequences at random points and shuffle the resulting pieces.

In Section 4, we study the impact of adding noise to the shuffling-sampling channels. Specifically we will study the capacity of channels where the sequences are not only shuffled and randomly sampled, but also contain errors, and we will discuss the impact of different noisy channel models on the overall capacity.

In Section 5, we study the multi-draw nature of DNA storage channels. Since multiple copies of each sequence are typically stored in DNA storage systems, they can be sequenced with potentially different noise patterns, which can help in error correction. This is captured in our information-theoretic framework by considering multi-draw channels, where each sequence in the DNA library may yield multiple copies with independent noise patterns at the output. In this setting, a natural approach is to first cluster the sequences and then solve several *trace reconstruction* problems. We will study whether such clustering-based

schemes are optimal, and we will provide results on the fundamental limits of DNA storage channels with multi-draws.

DNA storage is a relatively new field and many important questions remain unanswered. In Section 7, we end the monograph with a collection of important open problems and a discussion of connections to existing channel models and classical results in information theory.

# References

[1] M. Abroshan, R. Venkataramanan, L. Dolecek, and A. G. i Fàbregas, "Coding for deletion channels with multiple traces," in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 1372–1376, 2019.

[2] J. Acharya, H. Das, O. Milenkovic, A. Orlitsky, and S. Pan, "String reconstruction from substring compositions," *SIAM Journal on Discrete Mathematics*, vol. 29, no. 3, 2015, pp. 1340–1371.

[3] R. Ahlswede and A. Kaspi, "Optimal coding strategies for certain permuting channels," *IEEE transactions on information theory*, vol. 33, no. 3, 1987, pp. 310–314.

[4] D. Aird, M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke, "Analyzing and minimizing pcr amplification bias in illumina sequencing libraries," *Genome biology*, vol. 12, no. 2, 2011, R18.

[5] L. Anavy, I. Vaknin, O. Atar, R. Amit, and Z. Yakhini, "Data storage in dna with fewer synthesis cycles using composite dna letters," *Nature Biotechnology*, vol. 37, no. 10, Oct. 2019, pp. 1229–1236.

[6] P. Anderson, R. Black, A. Cerkauskaite, A. Chatzieleftheriou, J. Clegg, C. Dainty, R. Diaconu, R. Drevinskas, A. Donnelly, A. L. Gaunt, *et al.*, "Glass: A new media for a new era?" In *10th {USENIX} Workshop on Hot Topics in Storage and File Systems (HotStorage 18)*, 2018.

[7] P. L. Antkowiak, J. Lietard, M. Z. Darestani, M. Somoza, W. J. Stark, R. Heckel, and R. N. Grass, "Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction," *Nature Communications*, 2020.

[8] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," *roceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2004, pp. 910–918.

[9] E. B. Baum, "Building an associative memory vastly larger than the brain," *Science*, vol. 268, no. 5210, 1995, pp. 583–585.

[10] T. W. Benjamin, *Coding for a Noisy Channel with Permutation Errors*. Cornell University, 1975.

[11] V. Bhardwaj, P. A. Pevzner, C. Rashtchian, and Y. Safonova, "Trace reconstruction problems in computational biology," *IEEE Transactions on Information Theory*, 2020.

[12] M. Blawat, K. Gaedke, I. Hütter, X.-M. Chen, B. Turczyk, S. Inverso, B. W. Pruitt, and G. M. Church, "Forward Error Correction for DNA Data Storage," *Procedia Computer Science*, International Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA, vol. 80, no. Supplement C, 2016, pp. 1011–1022.

[13] N. A. Bokulich, S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso, "Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing," *Nature methods*, vol. 10, no. 1, 2013, pp. 57–59.

[14] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-Based Archival Storage System," in *Proc. of ASPLOS*, pp. 637–649, New York, NY, USA: ACM, 2016. (accessed on 01/14/2017).

[15] J. Brakensiek, V. Guruswami, and S. Zbarsky, "Efficient low-redundancy codes for correcting multiple deletions," *IEEE Transactions on Information Theory*, vol. 64, no. 5, May 2018, pp. 3403–3410.

[16] M. H. Caruthers, *A brief review of DNA and rna chemical synthesis*, 2011.

[17] S. Chandak, K. Tatwawadi, B. Lau, J. Mardia, M. Kubit, J. Neu, P. Griffin, M. Wootters, T. Weissman, and H. Ji, "Improved read/write cost tradeoff in dna-based data storage using ldpc codes," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, pp. 147–156, 2019.

[18] Z. Chase, "New lower bounds for trace reconstruction," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Institut Henri Poincaré, vol. 57, pp. 627–643, 2021.

[19] Y. M. Chee, H. M. Kiah, and T. T. Nguyen, "Linear-time encoders for codes correcting a single edit for DNA-based data storage," in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 772–776, 2019.

[20] Y. M. Chee, H. M. Kiah, and H. Wei, "Efficient and explicit balanced primer codes," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019.

[21] Y.-J. Chen, C. N. Takahashi, L. Organick, C. Bee, S. D. Ang, P. Weiss, B. Peck, G. Seelig, L. Ceze, and K. Strauss, "Quantifying molecular bias in dna data storage," *Nature Communications*, vol. 11, no. 1, 2020, p. 3264.

[22] K. Cheng, V. Guruswami, B. Haeupler, and X. Li, "Efficient linear and affine codes for correcting insertions/deletions," in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, ser. Proceedings, pp. 1–20, 2021.

[23] M. Cheraghchi, "Capacity upper bounds for deletion-type channels," *Journal of the ACM (JACM)*, vol. 66, no. 2, 2019, pp. 1–79.

[24] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Transactions on Information Theory*, 2020.

[25] M. Cheraghchi and J. Ribeiro, "Sharp analytical capacity upper bounds for sticky and related channels," *IEEE Transactions on Information Theory*, 2019.

[26] J. Chrisnata, H. M. Kiah, and E. Yaakobi, "Optimal reconstruction codes for deletion channels," in *2020 International Symposium on Information Theory and Its Applications (ISITA)*, IEEE, pp. 279–283, 2020.

[27] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, 2012, pp. 1628–1628.

[28] R. Con, A. Shpilka, and I. Tamo, "Linear and reed solomon codes against adversarial insertions and deletions," *arXiv:2107.05699 [cs, math]*, 2021.

[29] A. De, R. O'Donnell, and R. A. Servedio, "Optimal mean-based algorithms for trace reconstruction," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1047–1056, 2017.

[30] E. L. van Dijk, Y. Jaszczyszyn, D. Naquin, and C. Thermes, "The third revolution in sequencing technology," *Trends in Genetics*, vol. 34, no. 9, 2018, pp. 666–681.

[31] R. C. Edgar, "Muscle: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, 2004, pp. 1792–1797.

[32] P. Elias, "Coding for two noisy channels," in *Information Theory, 3rd London Symposium, London, England, Sept. 1955*, 1955.

[33] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, 2017, pp. 950–954.

[34] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes: New bounds and constructions," in *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

[35] R. Gabrys, S. Pattabiraman, and O. Milenkovic, "Mass error-correction codes for polymer-based data storage," in *2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 25–30, 2020.

[36] R. Gabrys and F. Sala, "Codes correcting two deletions," *IEEE Transactions on Information Theory*, vol. 65, no. 2, 2019, pp. 965–974.

[37] R. Gabrys and E. Yaakobi, "Sequence reconstruction over the deletion channel," *IEEE Transactions on Information Theory*, vol. 64, no. 4, 2018, pp. 2924–2931.

[38] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, 2013, pp. 77–80.

[39] R. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie International Edition*, vol. 54, no. 8, 2015, pp. 2552–2555.

[40] V. Guruswami and J. Hastad, "Explicit two-deletion codes with redundancy matching the existential bound," *IEEE Transactions on Information Theory*, vol. 67, no. 10, 2021, pp. 6384–6394.

[41] B. Haeupler and M. Mitzenmacher, "Repeated deletion channels," in *2014 IEEE Information Theory Workshop*, pp. 152–156, Nov. 2014.

[42] B. Haeupler and A. Shahrasbi, "Synchronization strings: Codes for insertions and deletions approaching the singleton bound," in *ACM SIGACT Symposium on Theory of Computing*, pp. 33–46, 2017.

[43] B. Haeupler and A. Shahrasbi, "Synchronization strings and codes for insertions and deletions—a survey," *IEEE Transactions on Information Theory*, vol. 67, no. 6, 2021, pp. 3190–3206.

[44] L. Hartung, N. Holden, and Y. Peres, "Trace reconstruction with varying deletion probabilities," in *2018 Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, SIAM, pp. 54–61, 2018.

[45] T. Haveliwala, A. Gionis, and P. Indyk, "Scalable techniques for clustering the web," 2000.

[46] R. Heckel and R. Grass, *Instructions to encode the first Biohackers episode from DNA*, Mar. 2021, URL: https://github.com/reinhardh/dna_rs_coding.

[47] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the dna data storage channel," *Scientific Reports*, vol. 9, no. 1, 2019, pp. 1–12.

[48] N. Holden, R. Pemantle, and Y. Peres, "Subpolynomial trace reconstruction for random strings and arbitrary deletion probability," in *Conference On Learning Theory*, 2018.

[49] T. Holenstein, M. Mitzenmacher, R. Panigrahy, and U. Wieder, "Trace reconstruction with constant deletion probability and related results," in *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, Citeseer, pp. 389–398, 2008.

[50] Y. Kanoria and A. Montanari, "Optimal coding for the binary deletion channel with small deletion probability," *IEEE Transactions on Information Theory*, vol. 59, no. 10, 2013, pp. 6192–6219.

[51] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *IEEE Trans. on Information Theory*, vol. 62, no. 6, 2016, pp. 3125–3146.

[52] H. M. Kiah, T. T. Nguyen, and E. Yaakobi, "Coding for sequence reconstruction for single edits," in *2020 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 676–681, 2020.

[53] H. M. Kiah, A. Vardy, and H. Yao, "Efficient bee identification," in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1943–1948, Jul. 2021.

[54] J. Koch, S. Gantenbein, K. Masania, W. J. Stark, Y. Erlich, and R. N. Grass, "A dna-of-things storage architecture to create materials with embedded memory," *Nature Biotechnology*, vol. 38, no. 11, Jan. 2020, pp. 39–43.

[55] S. Kosuri and G. M. Church, "Large-scale de novo DNA synthesis: Technologies and applications," *Nature methods*, vol. 11, no. 5, 2014, p. 499.

[56] M. Kovačević, "Runlength-limited sequences and shift-correcting codes: Asymptotic analysis," *IEEE Transactions on Information Theory*, 2019.

[57] M. Kovačević and V. Y. Tan, "Asymptotically optimal codes correcting fixed-length duplication errors in DNA storage systems," *IEEE Communications Letters*, vol. 22, no. 11, 2018, pp. 2194–2197.

[58] M. Kovačević and V. Y. Tan, "Codes in the space of multisets—coding for permutation channels with impairments," *IEEE Transactions on Information Theory*, vol. 64, no. 7, 2018, pp. 5156–5169.

[59] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: A mathematical analysis," *Genomics*, vol. 2, no. 3, 1988, pp. 231–239.

[60] A. Lapidoth and S. M. Moser, "On the capacity of the discrete-time poisson channel," *IEEE Transactions on Information Theory*, vol. 55, no. 1, 2008, pp. 303–322.

[61] A. Lapidoth, J. H. Shapiro, V. Venkatesan, and L. Wang, "The discrete-time poisson channel at low input powers," *IEEE Transactions on Information Theory*, vol. 57, no. 6, 2011, pp. 3260–3272.

[62] C. Laure, D. Karamessini, O. Milenkovic, L. Charles, and J.-F. Lutz, "Coding in 2d: Using intentional dispersity to enhance the information capacity of sequence-coded polymer barcodes," *Angewandte Chemie International Edition*, vol. 55, no. 36, 2016, pp. 10 722–10 725.

[63] H. H. Lee, R. Kalhor, N. Goela, J. Bolot, and G. M. Church, "Terminator-free template-independent enzymatic dna synthesis for digital information storage," *Nature Communications*, vol. 10, no. 1, Jun. 2019, p. 2383.

[64] A. Lenz, I. Maarouf, L. Welter, A. Wachter-Zeh, E. Rosnes, and A. Graell i Amat, "Concatenated codes for recovery from multiple reads of dna sequences," in *IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2021.

[65] A. Lenz, P. Siegel, A. Wachter-Zeh, and E. Yaakobi, "An upper bound on the capacity of the DNA storage channel," in *IEEE Information Theory Workshop*, 2019.

[66] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Anchor-based correction of substitutions in indexed sets," in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2019.

[67] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for dna storage," *IEEE Transactions on Information Theory*, vol. 66, no. 4, 2019, pp. 2331–2351.

[68] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakohi, "Achieving the capacity of the dna storage channel," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 8846–8850, 2020.

[69] V. Levenshtein, "Reconstruction of objects from a minimum number of distorted patterns," *Doklady Mathematics*, vol. 55, no. 3, 1997, pp. 417–420.

[70] K. Levick, R. Heckel, and I. Shomorony, "Achieving the capacity of a dna storage channel with linear coding schemes," *$56^{th}$ Annual Conference on Information Sciences and Systems (CISS), IEEE*, 2022.

[71] D. J. C. MacKay, "Fountain codes," *IEE Proceedings - Communications*, vol. 152, no. 6, 2005, pp. 1062–1068.

[72] A. Magner, J. Duda, W. Szpankowski, and A. Grama, "Fundamental bounds for sequence reconstruction from nanopore sequencers," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, 2016, pp. 92–106.

[73] A. Makur, "Information capacity of bsc and bec permutation channels," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, pp. 1112–1119, 2018.

[74] W. Mao, S. N. Diggavi, and S. Kannan, "Models and information-theoretic bounds for nanopore sequencing," *IEEE Transactions on Information Theory*, vol. 64, no. 4, 2018, pp. 3216–3236.

[75] M. Mitzenmacher and E. Drinea, "A simple lower bound for the capacity of the deletion channel," *IEEE Transactions on Information Theory*, vol. 52, no. 10, 2006, pp. 4657–4660.

[76] M. Mitzenmacher, "On the theory and practice of data recovery with multiple versions," in *2006 IEEE International Symposium on Information Theory*, IEEE, pp. 982–986, 2006.

[77] A. S. Motahari, G. Bresler, and N. David, "Information theory of DNA shotgun sequencing," *IEEE Transactions on Information Theory*, vol. 59, no. 10, 2013, pp. 6273–6289.

[78] F. Nazarov and Y. Peres, "Trace reconstruction with exp (o(n1/3)) samples," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1042–1046, 2017.

[79] M. S. Neiman, "Some fundamental issues of microminiaturization," *Radiotekhnika*, vol. 1, no. 1, 1964, pp. 3–12.

[80] L. Organick, S. D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, and et al., "Random access in large-scale dna data storage," *Nature Biotechnology*, 2018.

[81] L. Organick, B. H. Nguyen, R. McAmis, W. D. Chen, A. X. Kohll, S. D. Ang, R. N. Grass, L. Ceze, and K. Strauss, "An empirical comparison of preservation methods for synthetic dna data storage," *Small Methods*, vol. 5, no. 5, 2021, p. 2 001 094.

[82] S. Pabinger, S. Rödiger, A. Kriegner, K. Vierlinger, and A. Weinhäusel, "A survey of tools for the analysis of quantitative pcr (qpcr) data," *Biomolecular Detection and Quantification*, vol. 1, no. 1, 2014, pp. 23–33.

[83] S. Pattabiraman, R. Gabrys, and O. Milenkovic, "Reconstruction and error-correction codes for polymer-based data storage," in *2019 IEEE Information Theory Workshop (ITW)*, IEEE, pp. 1–5, 2019.

[84] K. R. Pomraning, K. M. Smith, E. L. Bredeweg, L. R. Connolly, P. A. Phatale, and M. Freitag, "Library preparation and data analysis packages for rapid genome sequencing," in *Fungal Secondary Metabolism*, Springer, 2012, pp. 1–22.

[85] C. Rashtchian, K. Makarychev, M. Racz, S. Ang, D. Jevdjic, S. Yekhanin, L. Ceze, and K. Strauss, "Clustering billions of reads for dna data storage," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[86] A. N. Ravi, A. Vahid, and I. Shomorony, "Capacity of the torn paper channel with lost pieces," in *IEEE International Symposium on Information Theory (ISIT)*, 2021.

[87] N. Roquet, S. P. Bhatia, S. A. Flickinger, S. Mihm, M. W. Norsworthy, D. Leake, and H. Park, "Dna-based data storage via combinatorial assembly," *bioRxiv*, 2021.

[88] O. Sabary, Y. Orlev, R. Shafir, L. Anavy, E. Yaakobi, and Z. Yakhini, "Solqc: Synthetic oligo library quality control tool," *Bioinformatics*, vol. 37, no. 5, 2021, pp. 720–722.

[89] O. Sabary, E. Yaakobi, and A. Yucovich, "The error probability of maximum-likelihood decoding over two deletion/insertion channels," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 763–768, 2020.

[90] O. Sabary, A. Yucovich, G. Shapira, and E. Yaakobi, "Reconstruction algorithms for dna-storage systems," *bioRxiv*, 2020.

[91] T. H. Saey, "Story one: Ancient horse's DNA fills in picture of equine evolution: A 700,000-year-old fossil proves astoundingly well preserved," *Science News*, vol. 184, no. 2, 2013, pp. 5–6.

[92] F. Sala, R. Gabrys, C. Schoeny, and L. Dolecek, "Exact reconstruction from insertions in synchronization codes," *IEEE Transactions on Information Theory*, 2017.

[93] J. Sayir, *Lecture notes for advanced communications and coding: Binary linear codes over the erasure channel*, 2014, URL: https://www-sigproc.eng.cam.ac.uk/foswiki/pub/Main/4F5/cod1.pdf.

[94] M. Schirmer, R. D'Amore, U. Z. Ijaz, N. Hall, and C. Quince, "Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data," *BMC Bioinformatics*, vol. 17, Mar. 2016, p. 125.

[95] S. Shamai and A. Lapidoth, "Bounds on the capacity of a spectrally constrained poisson channel," *IEEE Transactions on Information Theory*, vol. 39, no. 1, 1993, pp. 19–29.

[96] S. Shin, R. Heckel, and I. Shomorony, "Capacity of the erasure shuffling channel," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 8841–8845, 2020.

[97] T. Shinkar, E. Yaakobi, A. Lenz, and A. Wachter-Zeh, "Clustering-correcting codes," in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2019.

[98] I. Shomorony and R. Heckel, "Dna-based storage: Models and fundamental limits," *IEEE Transactions on Information Theory*, vol. 67, no. 6, 2021, pp. 3675–3689.

[99] I. Shomorony and A. Vahid, "Communicating over the torn-paper channel," in *IEEE Global Communications Conference (IEEE GLOBECOM)*, 2021.

[100] J. Sima and J. Bruck, "On optimal k-deletion correcting codes," *IEEE Transactions on Information Theory*, vol. 67, no. 6, Jun. 2021, pp. 3360–3375.

[101] J. Sima, N. Raviv, and J. Bruck, "On coding over sliced information," in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 767–771, 2019.

[102] J. Sima, N. Raviv, and J. Bruck, "Robust indexing - optimal codes for dna storage," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 717–722, Jun. 2020.

[103] S. Singh-Gasson, R. D. Green, Y. Yue, C. Nelson, F. Blattner, M. R. Sussman, and F. Cerrina, "Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array," *Nature Biotechnology*, vol. 17, no. 1010, Oct. 1999, pp. 974–978.

[104] W. Song, K. Cai, and K. A. S. Immink, "Sequence-subset distance and coding for error control in dna-based data storage," *IEEE Transactions on Information Theory*, vol. 66, no. 10, 2020, pp. 6048–6065.

[105] W. Song, K. Cai, and K. A. Schouhamer Immink, "Sequence-subset distance and coding for error control in dna-based data storage," *IEEE Transactions on Information Theory*, vol. 66, no. 10, Oct. 2020, pp. 6048–6065.

[106] S. R. Srinivasavaradhan, M. Du, S. Diggavi, and C. Fragouli, "Symbolwise map for multiple deletion channels," in *2019 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 181–185, 2019.

[107] S. R. Srinivasavaradhan, M. Du, S. Diggavi, and C. Fragouli, "On maximum likelihood reconstruction over multiple deletion channels," in *2018 IEEE International Symposium on Information Theory (ISIT)*, IEEE, pp. 436–440, 2018.

[108] S. R. Srinivasavaradhan, S. Gopi, H. D. Pfister, and S. Yekhanin, "Trellis bma: Coded trace reconstruction on ids channels for dna storage," 2021.

[109] S. K. Tabatabaei, B. Wang, N. B. M. Athreya, B. Enghiad, A. G. Hernandez, C. J. Fields, J.-P. Leburton, D. Soloveichik, H. Zhao, and O. Milenkovic, "Dna punch cards for storing data on native dna sequences via enzymatic nicking," *Nature Communications*, vol. 11, no. 1, Apr. 2020, p. 1742.

[110] R. Tamir and N. Merhav, "Error exponents in the bee identification problem," *arXiv preprint arXiv:2011.09799*, 2020.

[111] A. Tandon, V. Y. Tan, and L. R. Varshney, "The bee-identification problem: Bounds on the error exponent," *IEEE Transactions on Communications*, 2019.

[112] A. Tandon, V. Y. Tan, and L. R. Varshney, "The bee-identification error exponent with absentee bees," *IEEE Transactions on Information Theory*, vol. 66, no. 12, 2020, pp. 7602–7614.

[113] T. van der Valk, P. Pečnerová, D. Díez-del-Molino, A. Bergström, J. Oppenheimer, S. Hartmann, G. Xenikoudakis, J. A. Thomas, M. Dehasque, E. Sağlıcan, and et al., "Million-year-old dna sheds light on the genomic history of mammoths," *Nature*, vol. 591, no. 7849, Mar. 2021, pp. 265–269.

[114] S. Verdú, "Poisson communication theory," *International Technion Communication Day in Honor of Israel Bar-David*, vol. 66, 1999.

[115] H. Wei and M. Schwartz, "Improved coding over sets for dna-based data storage," 2021, pp. 1–1.

[116] N. Weinberger and N. Merhav, "The DNA storage channel: Capacity and error probability," *arXiv preprint arXiv:2109.12549*, 2021.

[117] J. L. Weirather, M. de Cesare, Y. Wang, P. Piazza, V. Sebastiano, X.-J. Wang, D. Buck, and K. F. Au, "Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis," *F1000Research*, vol. 6, 2017.

[118] G. Wong, I. Wong, K. Chan, Y. Hsieh, and S. Wong, "A rapid and low-cost pcr thermal cycler for low resource settings," *PLoS one*, vol. 10, no. 7, 2015, e0131701.

[119] H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific Reports*, vol. 5, 2015.

[120] H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific Reports*, vol. 7, no. 1, 2017.

[121] S. S. Yim, R. M. McBee, A. M. Song, Y. Huang, R. U. Sheth, and H. H. Wang, "Robust direct digital-to-biological data storage in living cells," *Nature Chemical Biology*, vol. 17, no. 3, Mar. 2021, pp. 246–253.