SCIENTIFIC REPORTS

OPEN

A Characterization of the DNA Data Storage Channel

Reinhard Heckel[1], Gediminas Mikutis[2] & Robert N. Grass[2]

Received: 21 November 2018
Accepted: 4 June 2019
Published online: 04 July 2019

Owing to its longevity and enormous information density, DNA, the molecule encoding biological information, has emerged as a promising archival storage medium. However, due to technological constraints, data can only be written onto many short DNA molecules that are stored in an unordered way, and can only be read by sampling from this DNA pool. Moreover, imperfections in writing (synthesis), reading (sequencing), storage, and handling of the DNA, in particular amplification via PCR, lead to a loss of DNA molecules and induce errors within the molecules. In order to design DNA storage systems, a qualitative and quantitative understanding of the errors and the loss of molecules is crucial. In this paper, we characterize those error probabilities by analyzing data from our own experiments as well as from experiments of two different groups. We find that errors within molecules are mainly due to synthesis and sequencing, while imperfections in handling and storage lead to a significant loss of sequences. The aim of our study is to help guide the design of future DNA data storage systems by providing a quantitative and qualitative understanding of the DNA data storage channel.

Recent years have seen an explosion in the amount, variety, and importance of data created, and much of that data needs to be archived. As an example, CERN, the European particle research organization has spent billions of dollars to generate more than 100 petabytes of physical data which it archives for analysis by future generations of scientists. However, standard storage media such as optical discs, hard drives, and magnetic tapes only guarantee data lifetimes of a few years. This has spurred significant interest in new storage technologies. Fueled by the excitement about its longevity and enormous information density, Deoxyribonucleic acid (DNA), a molecule that carries the genetic instruction of living organisms, has emerged as a promising archival storage medium.

At least since the 60 s, computer scientists and engineers have dreamed of harnessing DNA's storage capabilities[1,2], but the field has only been developed in recent years: In 2012 and 2013 groups lead by Church[3] and Goldman[4] stored about a megabyte of data in DNA and in 2015 Grass *et al.*[5] demonstrated that millenia long storage times are possible by information theoretically and physically protecting the data. Later, in the same year, Yazdi *et al.*[6] showed how to selectively access files, and in 2017, Erlich and Zielinski[7] demonstrated that DNA achieves very high information densities. In 2018, Organick *et al.*[8] scaled up those techniques and successfully stored and retrieved more than 200 megabytes of data.

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 several technological constraints for writing (synthesizing), storing, and reading (sequencing) DNA. The perhaps most significant one is that in practice it is difficult to synthesize strands of DNA significantly longer than one-two hundred nucleotides. While there are approaches that generate significantly longer strands of DNA, those are based on writing short strands of DNA and stitching them together[9], which is currently not a scalable approach. As a consequence, all recently proposed systems[3–8,10] stored information on DNA molecules of one-two hundred nucleotides. The second technological constraint is that the DNA molecules are stored in a pool and cannot be spatially ordered. We do not have straightforward random access to the DNA fragments in the pool, and can therefore not choose which DNA fragments to read (however, one can design PCR hooks to selectively amplify and sequence part of the DNA[6,8]) Accessing the information is done via state-of-the-art sequencing technologies (including Illumina and third-generation sequencing technologies such as nanopore sequencing). This corresponds to (randomly) sampling and reading molecules from the DNA pool. In practice, sequencing is preceded by potentially several cycles of Polymerase Chain Reaction (PCR) amplification. In each cycle, each DNA molecule is replicated by a factor of about 1.6–1.8, but that number depends on the PCR method and varies by sequence. The proportions of molecules in the pool depend on the synthesis method, the PCR steps, and the decay of DNA during storage. In summary, in the process of synthesizing, storing, handling, and sequencing, the following errors occur:

[1]Rice University, Department of Electrical and Computer Engineering, Houston, 77005, Texas, USA. [2]ETH Zurich, Department of Chemistry and Applied Biosciences, Zurich, 8093, Switzerland. Correspondence and requests for materials should be addressed to R.H. (email: rh43@rice.edu)

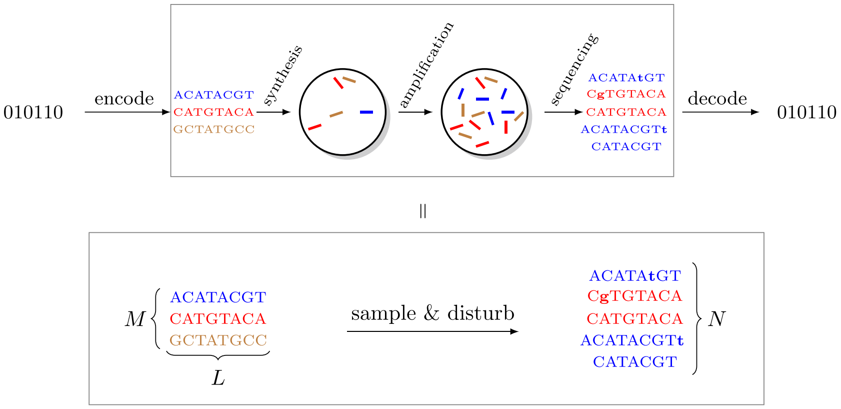


Figure 1. Channel model for DNA storage systems. Only short molecules can be synthesized, and of each molecule a large number of copies is generated in the synthesis process. For reading, the data is first amplified and then sequenced. Abstractly, the input to the channel is a multi-set of M length-L DNA molecules, while the output is a multi-set of N draws from the pool of DNA molecules that is disturbed by insertions, substitutions, and deletions (marked as lowercase and boldface letters). The sampling distribution as well as the process inducing errors in individual molecules account for errors in synthesis, storage, and sequencing.

i. Molecules might not be successfully synthesized, and some might be synthesized many more times than others. Current synthesis technologies generate not only one but several thousand to millions copies of a strand, which can all contain possibly different errors.
ii. During storage, DNA decays, which results in a loss of molecules.
iii. Since reading amounts to drawing from the pool of molecules, we only see a fraction of the molecules that are in the pool. This fraction depends on the distribution of the molecules in the pool and the number of draws (i.e., reads) we take.
iv. Sequencing of DNA, and in particular synthesis of DNA may lead to insertions, deletions, and substitutions of nucleotides in individual DNA molecules[11,12].

Given these constraints, a good abstraction of the DNA data storage channel that captures the essential parts of a DNA storage system is to view the input as a multiset of M DNA molecules of length L, and the output as sampling N times independently from the multiset, and then disturbing the sampled molecules with insertions, deletions, and substitutions, to account for errors within individual molecules. See Fig. 1 for an illustration. The sampling distribution and distribution of insertions, deletions, and substitutions determine the channel and thus how much data can be stored.

A statistical understanding of those distributions and the processes that lead to those distributions is crucial for designing good DNA data storage systems, and in particular for designing good encoder/decoder pairs. Encoder/decoder pairs use error correcting codes that add redundancy in order to correct errors. To decide which codes to use or to develop, and how to choose their parameters (specifying the amount of redundancy, for example), requires an statistical understanding of the error distributions. Providing such an understanding is the goal of this paper.

For a given channel—determined by the handling procedures, experimental designs, synthesis and sequencing tools, and storage time—the goal of an encoder/decoder pair is to maximize the number of bits per nucleotide, ML, while guaranteeing successful decoding. The goal when designing a DNA data storage system is to minimize the cost for storing data, or to maximize the number of bits per *total number of nucleotides* stored, which is typically significantly larger than the number of nucleotides in the multiset, ML. Therefore, it is important to understand how different parameters of the channel, such as storage times and physical density (i.e., number of copies of the DNA molecules in the pool) affect the error distributions. To determine the best trade-off for maximizing density or minimizing costs, an understanding of the error statistics and how they change with different experimental setups is important.

The aim of this paper is to obtain a both quantitative and qualitative understanding of the DNA data storage channel, which amounts to quantifying the sampling distribution as well as the distribution of errors within the molecules, for different experimental setups, and assign, when possible, error sources to processes such as reading or writing the DNA. As currently the cost of well designed experiments to analyze all those errors in a fully controlled manner is very high, we chose to use data for quantifying error statistics from the DNA data storage literature, as well as complement that data with data from some of our own experiments.

Error Sources and the Need for Error Correcting Codes

We start by giving an explanation of the basic chemical processes involved, together with a short theoretical discussion of the potential errors. Next, we identify differences between the various performed experiments and assign the observed differences in error probabilities to the individual steps. This should not only give theoreticians estimates of the overall expected errors, but also directions on how errors could be experimentally avoided.

Errors during DNA synthesis. The currently utilized synthesis strategies work on solid surfaces (chips), on which one end of the DNA molecule is attached, and nucleotides are added one by one by a chemical process[13]. These synthesis methods are currently able to generate up to a few millions of (typically distinct) sets of DNA strands per chip. Current technologies are not able to generate single DNA strands, they generate sets of DNA strands. Each set typically consists of millions copies (femtomoles)[14], and is generated on a geometrically limited 2-D surface of the chip (i.e., a spot). Nucleotides are directed to these spots by light via optically sensitive protection groups[15,16], are directed electrochemically with electrodes via pH sensitive protection groups[17], or are directed with printing technology by using a movable printing head[11]. During the chemical addition of new nucleotides on a growing DNA strand, various errors can occur: a nucleotide may not be positioned *where it should* (resulting in a deletion), a nucleotide might be positioned *where it should not* (resulting in an insertion), and a nucleotide other than the intended one is added (resulting in a substitution). Moreover, the growing strand may be terminated, meaning that the chemical reactivity at the end of the growing strand is lost, and nucleotides can no longer be added to this strand in later steps. The probability of this termination reaction—which varies by technology used and also potentially by the position within the sequence[11]—limits the length of DNA strands that can be generated, since the longer the target sequence, the more sequences at a spot do not reach the target length. Also note that typically not all individual growing strands at the same spot (corresponding to the same target sequence) undergo the same error, meaning that for every target sequence variations may be present.

Depending on the chosen synthesis method, the number of DNA copies generated per spot may be unevenly distributed. Spots on the chip-edge may have a different synthesis yield (number of complete DNA strands synthesized per sequence), so that some DNA sequences intrinsically have higher yields than others (e.g., synthesis limited by self-binding of sequences) and that physical imperfections on the chip surfaces limit the ideal synthesis on some chip locations.

All chemical synthesis methods generate single stranded DNA (ssDNA), which is chemically detached from the chip surface after the synthesis procedure to generate a DNA pool in aqueous solution. In order to clean up the synthesis pool, i.e., remove non-complete sequences, and to generate double stranded DNA, polymerase chain reaction (PCR) is performed. During this step utilizing biotechnological enzymes, only DNA sequences which have correct primers on both ends are amplified about 10,000 fold over about 15 cycles (different labs utilize slightly different procedures). This process dilutes out non-complete strands (DNA strands that do not have primers on both ends can not be amplified).

Although PCR by itself is understood as a high-fidelity process and thus has few errors[18,19]. PCR is known to have a preference for some sequences over others, which may further distort the copy number distribution of individual sequences[20–23]. Thus, each cycle of PCR in expectation multiplies each molecule by a certain number which lies typically slightly below two and is sequence dependent. For example, high GC-contents (i.e., a large number of Gs and Cs in the sequence) can lead to a smaller expected number and thus fewer corresponding sequences[24].

Errors during DNA storage. During DNA storage, as well as during any DNA processing step, such as removing DNA from the chips, and during heating intervals in PCR, the DNA strands are prone to chemical decay of DNA, especially hydrolytic damage. The main effects of hydrolytic damage are depurination[25], which eventually results in strands breaking[26], and deamination of cytosine (C), in the C-G basepair, resulting in uracil (an U-G basepair)[27].

Using current DNA reading technologies which involve DNA amplification stages, namely standard PCR in sample preparation and bridge amplification during Illumina sequencing, any DNA strand that has undergone strand-breakage following depurination is no longer read. This is due to the fact that broken strands do not have the required amplification primers on both ends of the DNA strand, are therefore not amplified, and as a consequence are diluted out during the amplification stages.

The effect of hydrolytic deamination is less extreme, and depends on the choice of enzymes utilized during subsequent PCR steps. Proof-reading enzymes (3′ to 5′ exonuclease, high-fidelity) have a significantly different effect on the errors introduced during deamination (i.e., U-G basepairs) than non-proof-reading enzymes. For most proof reading enzymes, the amplification reaction stalls at an U nucleotide in the first PCR cycle, and no complete copy of a sequence comprising a U nucleotide is formed. These incomplete DNA copies are then not further amplified in the next PCR cycle, due to a lack of primers on both ends of the sequence, and are therefore diluted out. On the complementary strand (having a G), the sequence still amplifies. In this context, the enzyme removes the errors which have occurred during DNA storage. However, if a DNA strand has at least one deamination (C to U reaction) on both strands, the whole information stored in the sequence is lost (diluted out), as the amplification reaction stalls at the U nucleotide at both strands and neither are amplified (see Fig. 2). If non-proof reading enzymes are utilized, the U-G basepair is in half amplified to the correct C-G basepair, and in half amplified incorrectly as a T-A basepair, formally resulting in a C2T error.

In summary, storage can lead to significant loss of whole DNA molecules, as well as to substitution errors in the DNA molecules.

Errors during DNA sequencing. The errors during DNA sequencing depend on the technology that is used. The currently most utilized sequencing platform is Illumina's, and in this paper we only consider datasets that have been sequenced with this technology, although recently DNA data storage readout has also been performed with nanopore sequencing[8,28]. Illumina errors—characterized as early as 2008[29]—are not random, but are strand specific[29,30], and the error rates are higher at the end of a read. According to a recent study[30], the substitution reading error rates are about 0.0015–0.0004 errors per base. We remark that the exact number, however, depends on the particular dataset. Insertions and deletions are significantly less likely, specifically they are on the order of 10^{-6}.

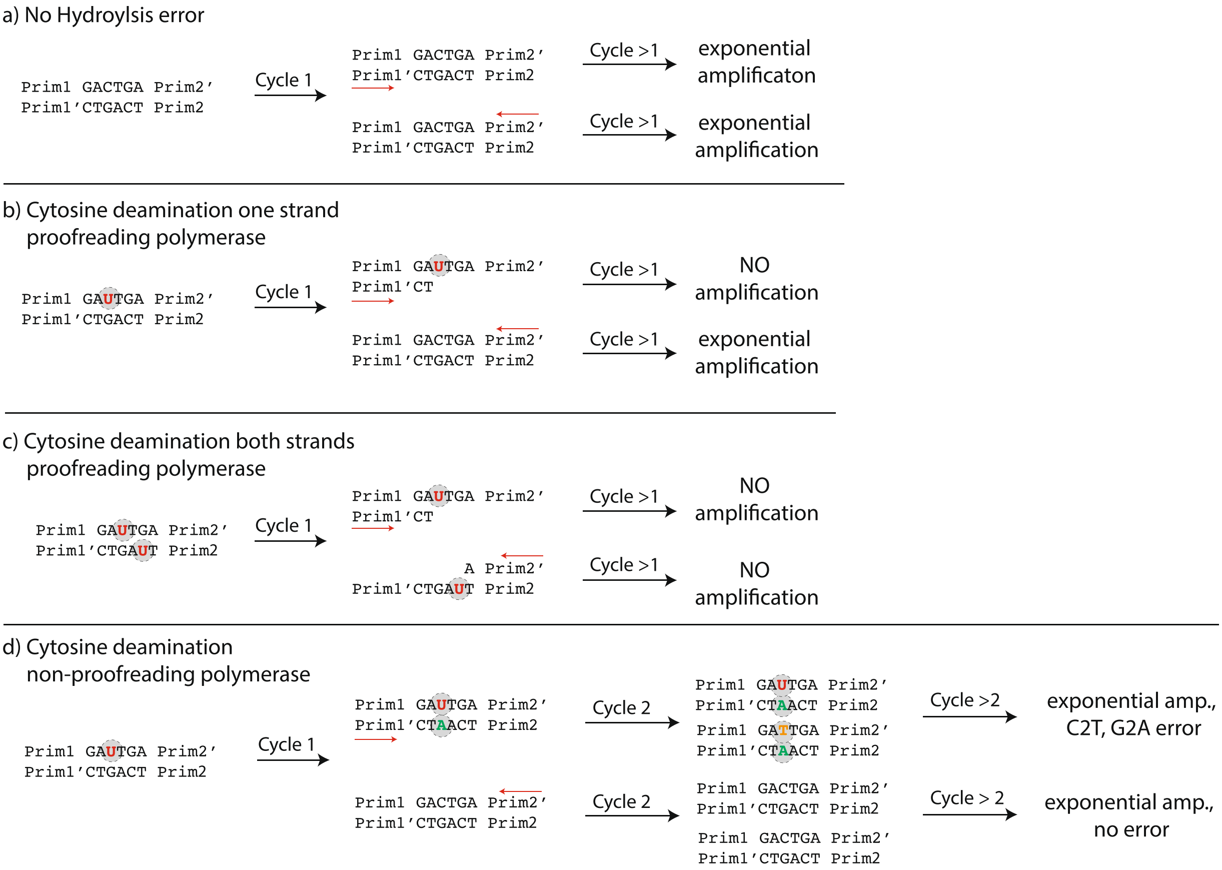


Figure 2. Deamination of cytosine results in the transformation of cytosine to uracyl. Most proof-reading PCR enzymes cannot copy past an uracyl. As a consequence any strand containing a de-aminated cytosine is not amplified (**b**). If both strands of a dsDNA molecule contain uracyl moieties, all information originally contained in the molecule is lost (diluted out) during PCR (**c**). If non-proof reading enzymes are used for PCR (e.g. taq polymerase), any uracyl is misinterpreted by the enzyme as a thymine, introducing an adenine on the growing complementary strand, resulting in C2T and G2A errors.

Two properties that raise the error probability for most sequencing (and synthesis) technologies are high GC content and long homopolymer stretches (e.g., GGGGGG)[24,31]. In more detail, the paper[24], [Fig. 5] reports that in particular the substitution and deletion error rates increase significantly for homopolymer stretches longer than six. Moreover, molecules with high GC content exhibit high dropout rates and PCR errors are significantly more likely in regions with a high GC content[7,31]. For example, the paper[24], [Fig. 3] reports that the coverage obtained by common sequencing technologies such as Illumina's HiSeq technology is significantly lower for fragments with GC content smaller than 20% and larger than about 75%. For reasons described in the previous section, any DNA strand that is broken, or for any other reason, does not have correct sequencing primers on both ends of the DNA strand can not be read by Illumina sequencing technologies, as the initial phase of the Illumina sequencing procedure involves bridge-amplification[32].

The need for error correcting code and implications of this work on their design. In view of the relatively low error levels encountered in DNA synthesis and sequencing (see Section 4.1), one may be tempted to store data in DNA without error correction coding, relying purely on the intrinsic data redundancy of the system: during synthesis every designed DNA strand is synthesized in many copies (often millions) and standard sequencing technologies read each designed DNA strand a multitude of times (often 100 fold read coverage). As DNA synthesis is orders of magnitude more expensive than DNA sequencing, it is not surprising that the first two attempts of storing 1 MB of data in DNA did not use a principled way of error correction[3,4] but as a consequence were unable to retrieve the stored data without manual intervention. The intrinsic data redundancy enables coping with statistical DNA read and write errors by data clustering and averaging if the number of reads and the number of redundancy copies is sufficiently large, but it fails at systematic errors: specific DNA sequences may be hard (or impossible) to synthesize or sequence, which include long base repetitions as well as hairpins/loops and other forms of higher-order structure[33]. As this paper shows, both statistical and systematic errors lead to a loss of sequences that make data redundancy a necessity for enabling perfect information retrieval. This is typically accomplished with error correcting codes[5].

In addition to the described problem of systematic and statistical errors, a main advantage of DNA data storage, namely its enormous data capacity (theoretically up to 455 EB/g[3]) is offset significantly if the system relies

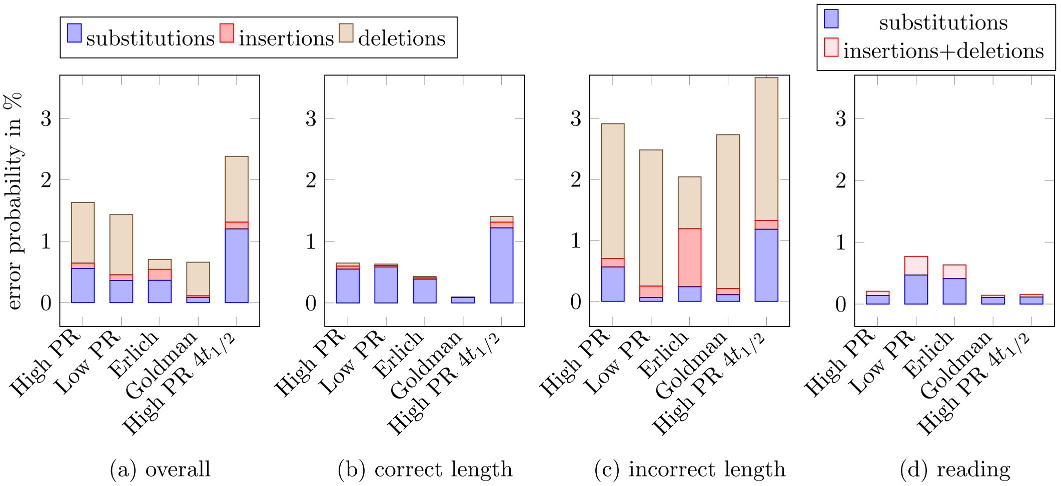


Figure 3. (**a**) Error probabilities of all molecules, (**b**) Error probabilities of the molecules that have the correct length (thus the number of insertions is equal to the number of deletions). (**c**) Error probabilities of molecules that *do not* have the correct length. In the five datasets, 56.6%, 57.7%, 56.7%, 83%, and 78.6% have the correct length. In the molecules that have the correct length, the substitution errors dominate by far. (**d**) Estimates of the reading errors.

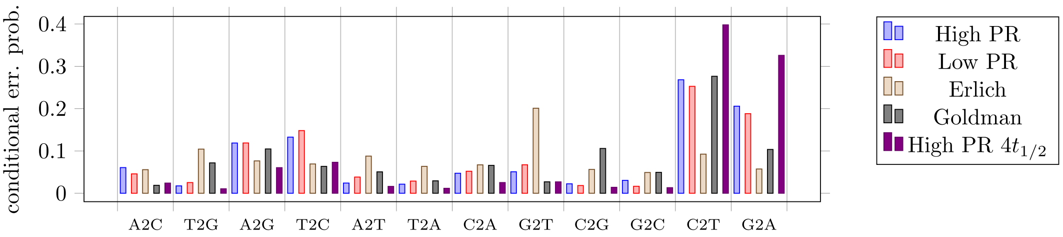


Figure 4. Conditional error probability for mistaking one nucleotides for another (e.g., A2C means mistaking A for C). Since those are conditional error probabilities (conditioned on a substitution error having occurred), the bars sum to one.

heavily on high physical redundancy of the sequences. This is not only true during the actual storage of the data as DNA, but also into how much DNA has to be fed into a sequencing experiment in order to allow sufficient data variability. Consequently, it is also advantageous to use error correction within individual DNA sequences. As Section 4.1 shows, errors within a sequence are in the 1–2% range per nucleotide. With common DNA strand lengths of 60–180 nucleotides, this means that statistically nearly every read strand has an error, and most strands have 2–3 errors. Relying on repetition alone, and assuming statistical errors, each designed strand would have to be read many times in order to precisely determine the correct sequence. This task is additionally complicated by the nature of DNA sequencing, which corresponds to randomly drawing sequences from a pool of sequences resulting in a coupon collector problem and requiring a significant increase in DNA read coverage. Currently applied DNA error correction schemes for DNA data storage[5] can correct any 2–3 substitution errors by adding 5–10 nucleotides to every sequence. This results in an increase of less than 10% of increased synthesis cost, which in most cases offsets the aforementioned problems of needing to read every sequence several times.

From the error rates estimated here (see Section 4.1) we see that the inner and outer code error correcting schemes proposed in literature, typically allowing for 1–3 error corrections per sequence, and a loss of 2–10 complete sequences, are able to perfectly recover the information, and are also within the economic specificities of the DNA read and write costs, using current DNA synthesis and sequencing technologies (i.e. high yield array synthesis and Illumina Dye Sequencing). However, if DNA data storage is to become a common DNA archiving technology, costs and accessibility of read and write technologies will have to advance significantly. This will most probably result in lower levels of precision, making error correcting schemes even more important to enable perfect data recovery.

Material and Methods

We estimate the error probabilities of the DNA data storage experiments performed by three different research groups and discuss the allocation of the errors to individual parts (writing, storage, reading) of the DNA storage model. In this section, we describe the data sets we consider, and the estimation of error probabilities.

Name	Strands synth.	target length	synthesis method	decay retained	physical redundancy	PCR cycles	sequencing method	read coverage
Goldman	153335	117	Agilent	100	22172	22	HighSeq 2000	519
Erlich	72000	152	Twist B.	100	1.2810^{7}	10	Miseq V4	281
Erlich D1–7	72000	152	Twist B.	100	$1.2810^{7}-1.28$	40+	Miseq V4	281–503
High PR	4991	117	Customa.	100	3.910^{3}	65	Miseq 2 × 150	372
High PR $4t_{1/2}$	4991	117	Customa.	6.25	3.910^{4}	65	Miseq 2 × 150	456
Low PR	4991	117	Customa.	100	1.2	68	Miseq 2 × 150	461
Low PR $4t_{1/2}$	4991	117	Customa.	5.75	17.9	68	Miseq 2 × 150	396

Table 1. Parameters of the datasets analyzed in this paper. Custom Array uses electrochemical synthesis, and Twist and Agilent use material deposition as processes for the synthesis.

Data sets. We analyze datasets from our group as well as datasets from Goldman's[4] and Erlich's groups[7]. Each dataset was obtained by synthesizing DNA molecules, storing the molecules, and sequencing them. For each dataset, we are given the channel input consisting of the DNA molecules to be synthesized, and the channel output, in the form of forward and backward reads from Illunmina sequencing technology. For each dataset, we obtained DNA molecules by stitching the one-sided reads together using the FLASH[34] algorithm. We set the maximal overlap of the reads to the target length, which is known in each experiment, the minimal overlap to 87% of the target length, and finally, the maximum allowed ratio between the number of mismatched based pairs and the overlap length to 20%.

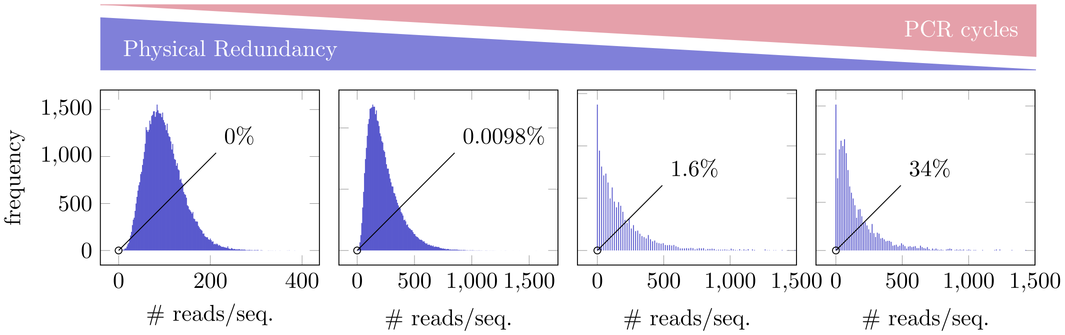
The datasets differ in the number of strands synthesized, the target length of each of the sequences, the synthesis method, the expected decay due to thermal treatment (modeling accelerated aging), measured in the percent of the original DNA that remains intact after heat treatment, the physical redundancy, which is the expected number of copies of each strand in the pool of DNA, the number of PCR cycles, which is the total number of PCR cycles conducted from synthesis to sequencing, and finally the read coverage which is number of reads per strand that has been synthesized. We briefly summarize how the datasets have been obtained below, and the key parameters in Table 1.

- Goldman *et al.*[4] synthesized 153,335 DNA molecules, each of length 117, containing no homopolymers. DNA was synthesized using the Agilent Oligo Library Synthesis (OLS) Sureprint technology process, and sequenced with Illumina HighSeq 2000 by taking forward and backward reads of length exactly 104.
- Erlich and Zielinski[7] synthesized 72,000 DNA molecules, each of length 152, containing no homopolymers longer than two. DNA was synthesized with Twist Bioscience technology, and sequenced with Illumina Miseq V4 techonlogy, taking forward and backward reads of length exactly 151. We also consider data from the dilution experiment in[7], which varies the physical redundancy by factors of ten starting from approximately 10^{7} by diluting the DNA (Erlich D1–7).
- Grass *et al.*[5] synthesized 4991 DNA molecules, each of length 117, containing no homopolymers longer than three. DNA was synthesized with Customarray, and the DNA was read after storing it for a negligible amount of time and read again after thermal treatment corresponding to four half-lifes. Both dataset have rather high physical redundancy and are therefore denoted by High PR and High PR $4t_{1/2}$. The DNA was sequenced with Illumina Miseq. 2 × 150 bp Truseq technology.
- In another dataset from our group, we synthesized again 4991 DNA molecules, each of length 117, containing no homopolymers longer than three. In contrast to High PR and High PR $4t_{1/2}$, we diluted the DNA so that the physical redundancies are low. We again sequenced the original DNA (High PR) as well as DNA after thermal treatment corresponding to four half-lifes; the resulting datasets are denoted by Low PR and Low PR $4t_{1/2}$.

Estimation of overall error probabilities. In order to estimate the errors within sequences, we consider the aligned reads (obtained from the two single sided reads as explained in the previous section "Data sets"), and for those reads, we estimate the substitution, insertion, and deletion probabilities. Towards this goal, for each aligned read m, we find a molecule $\mathrm{m}_{\mathrm{orig}}$ in the set of original sequences which minimizes the edit (Levenshtein) distance to, and take the substitution, insertion, and deletion probabilities as the average number of substitutions, insertions, and deletions per nucleotide required for aligning m and $\mathrm{m}_{\mathrm{orig}}$. We aggregate the corresponding statistics for all molecules, as well as for molecules that do have the correct target length and do not have the correct target length (Fig. 3(a–c)). Note that molecules that do not have the correct target length necessarily contain deletion or insertion errors.

Estimation of reading error probabilities. The overall error consists of errors due to synthesis, storage and sequencing. Out of those error sources, the sequencing error is the only error that can be estimated independently due to two-sided reads. The goal of this section is to estimate the sequencing error and to understand which proportion of errors reported in Fig. 3(a–c) can be attributed to synthesis and which to sequencing.

We estimate the sequencing error as follows. For each dataset, we consider the reads that have been successfully aligned, as those are the reads on which the overall error probability estimates (see Fig. 3a–c)) are based on. Out of those N reads, we consider the two single sided reads r_{1k} and r_{2k}, $k = 1, \ldots, N$, and find the best alignment that does



(a) Erlich
PR 1.2810^7, 10 PCRC

(b) Goldman
PR 22172, 22 PCRC

(c) High PR
PR 3898, 65 PCRC

(d) Low PR
PR 1.2, 68 PCRC

Figure 5. The distribution of the number of reads per each given sequences that has been synthesized, along with the physical redundancy (PR) and PCR cycles (PCRC). The percentage in the picture corresponds to the percentage of molecules that are not observed in the sequenced data. Erlich's and Goldman's data is approximately negative binomial distributed, whereas High PR and Low PR have a long tail and peak at zero. The reading coverage of all datasets is approximately the same. Likely, the difference in distribution of (**a**) and (**b**) to (**c**) and (**d**) is due to the significantly more cycles of PCR in (**a**) and (**b**), while the difference of (**a**) and (**b**) is due to low physical redundancy and dilution.

not penalize gaps at the beginning and at the end of the sequences. We then count the number of substitutions, s_k, and number of deletions, d_k, plus number of insertions, i_k, that are necessary to align the two reads. We then divide that number by the sum of the length of the two single sided reads $\mathrm{len}(r_{1k}) + \mathrm{len}(r_{2k})$, and average it over all reads to obtain an estimate of the substitution and insertion plus deletions error probabilities that occur during reading (for example, the substitution error probability estimate is $\frac{1}{N}\sum_{k=1}^{N} \frac{s_k}{\mathrm{len}(r_{1k}) + \mathrm{len}(r_{2k})}$). This estimate would be the error we obtain when choosing the nucleotide from one of the two single sided reads at random, in case there is a mismatch in the alignment. Since we ignore the quality information of the reads, this is a suboptimal way of obtaining a single sided read from the two sided reads, thus the corresponding error probabilities can be viewed as an upper bound on the reading error probability that is achieved by aligning the single sided reads. Note that our claim that this is an estimate of the read error probabilities also assumes that the single sided reads are independent.

Results

We characterize errors on a molecule level (substitutions, insertions, and deletions) as well as the distribution of molecules that we observe on the output of the channel.

Errors within molecules. We start with reporting the substitution, deletion, and insertion errors in individual molecules. As discussed in the introduction, those errors are mainly due to synthesis and sequencing. In Fig. 3 we report results for Goldman's and Erlich's dataset, and the High PR dataset, as those datasets differ in the synthesis method. Moreover, we report results for our heat treated DNA, High PR $4t_{1/2}$, as decay of DNA introduces substitution errors as well, (as discussed previously).

The majority of the reads in each dataset have the correct target length. As depicted in Fig. 3, out of the reads that have correct target length, in all datasets, the substitution errors dominate by far. In the overall error probabilities, the deletions dominate (at first sight, in Erlich's data, this seems to be different, as the number of insertions is similar to the number of deletions. However, this is only an artifact of how Erlich's reads are filtered: all single-sided reads have length 150, and the target length is 152. Therefore, reads that would align to shorter molecules are filtered out and are not accounted for in the estimates). Comparing this to the estimated reading error probabilities, we can conclude that likely most of the deletions (and insertions) that we see in the overall reads are due to synthesis, while the substitution errors are likely dominated by synthesis and sequencing and are also impacted by DNA decay and PCR. This is consistent with what we expect from the literature, which found that at sequencing, substitution errors are significantly more likely than deletions and insertions.

In Fig. 4 we examine the substitution errors in more detail by computing conditional error probabilities for mistaking a certain nucleotide for another. The results show that, for example, mistaking T for a C and A for a G is much more likely than other error probabilities. These two transitions can be explained by the vulnerability of cytosine to deamination and the consequent formation of an uracyl. During PCR amplification—either prior to sequencing, or as part of the sequencing process itself—and, if non-proofreading polymerases are used, uracyls are misinterpreted as thymines by the polymerase and adenines are introduced on the complimentary strand (see Fig. 2). While cytosine deamination is a hydrolysis reaction (involving water as reactant), it can proceed during many of the DNA processing steps, including chemical cleavage of the DNA molecules from the synthesis chip, PCR cycling itself (involving high temperatures), DNA handling and storage. As a result, DNA samples that have undergone decay due to storage show increased C2T and G2A errors. Interestingly, and considering the large experimental differences between the individual datasets (different synthesis, different PCR polymerases, different Illumina sequencing kits, different storage/handling), the observed error levels only display marginal variances, depicting the robustness of the technologies involved.

Distribution of molecules. In the coming subsections we discuss the distribution of the molecules that we observe at the output of the channel. The distribution of molecules is a function of the proportions of the molecules in the pool, as well as the sequencing process itself and the number of reads. The proportion of the molecules in the pool prior to sequencing depends on the original physical redundancy determined by the synthesis method and potential amplification steps, as well as the decay of the DNA which causes a loss of molecules, and on amplification and potential dilution steps during processing the DNA and for preparation of the sequencing step.

We are particularly interested in the fraction of sequences observed at least once, since this parameter is relevant for designing coding schemes. Specifically, as shown in the paper[35], the number of bits that can be stored on a given number of DNA sequences depends on the *fraction of molecules observed at least once*, or in more detail, the fraction out of the molecules that are given to the synthesizer and are read at least once when sequencing the DNA.

Distribution of the molecules in control sets and PCR bias. We start by analyzing Erlich's and Goldman's data as well as our High PR and Low PR datasets. The main difference in the experiments that generated those datasets is the number of PCR steps as well as the physical redundancy of the molecules. The synthesis method in some of the experiments differs as well, but the reading coverage and sequencing method in all the experiments are comparable. In Fig. 5, we depict the distribution of the number of reads per molecule that has been synthesized, along with the physical redundancy (PR) and the number of PCR cycles (PCRC). We find that Erlich's and Goldman's data follows approximately a negative binomial distribution, whereas our datasets (High PR and Low PR) have a long tail and peak at zero.

Since the reading process is essentially the same for all datasets (Illumina sequencing), the difference of Erlich's to Goldman's to our original DNA (High PR) can in principle be attributed to either a maldistribution of the number of molecules generated during synthesis and/or to PCR amplification bias.

However, a significant maldistribution of the number of molecules can be ruled by considering Fig. 5(d) and noting that only 34% of the sequences are lost, even though the physical density is around one, and thus the number of DNA fragments in the dataset is approximately equal to the number of distinct fragments that have synthesized. In more detail, the corresponding dataset has been obtained by taking a small sample from the original synthesized DNA, which contains many copies of each fragment. Suppose the number of copies in the original DNA are very far from being uniformly distributed. Then, if we draw a small subset corresponding to a physical redundancy of about one, we would expect to lose a significantly larger proportion of fragments. As a consequence, we can conclude that the synthesized molecule distribution is not very far from uniform, leaving PCR bias as an possible explanation, as explained next.

In each PCR cycle, every sequence is amplified with a sequence-specific factor that is slightly below two[20–23]. This leads to significantly different proportion of the sequences in the original pool of DNA in particular when the process is repeated many times, i.e., when the number of PCR cycles is large. For example, if a sequence has PCR efficiency 80%, whereas another one has efficiency 90%, then after 60 PCR cycles, the proportions change from 1/1 to $(1.8/1.9)^{60} = 0.039$. Thus, a large number of PCR cycles leads to a distribution with a long tail; later we report results from an computational experiment demonstrating this effect. This effect is further confirmed by the dilution experiment from Erlich (Fig. 6), where every dilution step consequently requires more PCR cycles until the PCR process saturates.

Finally, note that the Low PR dataset has a significantly larger fraction of lost sequences, see Fig. 5(d), compared to the High PR dataset in Fig. 5(c). That can be explained by the low physical redundancy in combination with the PCR amplication bias, as discussed in more detail in the next section.

Low physical redundancy and loss of sequences. Very low physical redundancy, obtained by taking a subset of a pool of DNA sequences with large physical redundancy, leads to a loss of sequences which in combination with PCR bias (as discussed in the previous section), results in a read distribution that has a long tail and many molecules with few or no reads. This can be seen by comparing Fig. 5(c) with Fig. 5(d). To see that the loss of sequences can be partly attributed to generating a pool with low physical redundancy, consider a pool of DNA sequences where each sequence has about the same amount of copies, and that number is large so that the physical redundancy is large. Suppose we obtain a pool of DNA with very low physical redundancy (say about one, as in Fig. 5(d)) by taking a random subset of the large pool, via pipetting. If the DNA sequences are mixed up well in the pool, that process corresponds to drawing a subset of the pool uniformly at random. Then, the expected fraction of distinct sequences in the subset is about $1-e^{-r}$, where r is the physical redundancy. In other words, at a physical redundancy of 1, about 36% of the sequences will be lost. Below, we present a computational experiment demonstrating this effect. Moreover, sequencing a pool of DNA with low physical redundancy requires more PCR steps, which, as discussed in the previous section, due to PCR bias, leads to a distribution that has a long tail and further contributes to a loss of sequences.

In order to investigate the effect of low physical redundancy further, we next consider the data from an experiment carried out by Erlich and Zielinski[7], in which the authors varied the physical redundancy from 1.2810^7–1.28 copies per molecule synthesized by factors of 10, by consecutively diluting the DNA. In Fig. 6, we depict the corresponding distribution of the molecules. As we can see, initially, when the physical redundancy is large, the molecules are approximately negatively binomial distributed. As the physical density becomes low (about 1000 and lower) the fraction of molecules that are not seen becomes increasingly large, and as it becomes very low, a large fraction of the molecules is not seen at the output, and the distribution has an extremely long tail, for example, at physical density 1.28, there is one sequence that is observed 98800 times at the output.

Again, this can be attributed to the effect of sampling a small set of sequences from a pool of DNA, as well as to the fact that an increasingly larger number of PCR steps is required for sequencing, thus the effect of PCR bias

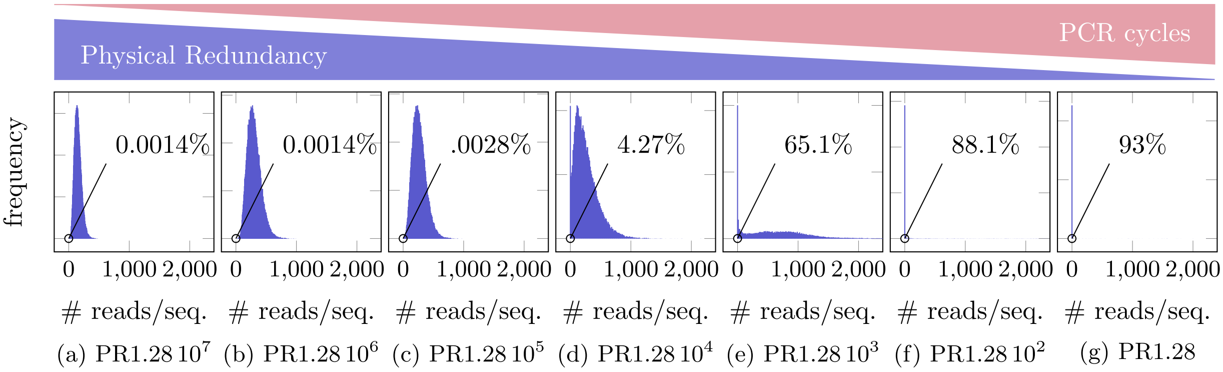


Figure 6. The distribution of the number of reads per each given sequences that has been synthesized for Erlich's dilution experiment, which varies the physical density from approximately 1.2810^7–1.28 copies per molecule synthesized. The percentage in the picture corresponds to the percentage of molecules that are not observed in the sequenced data. The distribution has an increasingly long tail (not depicted), e.g., at physical density 1.28, there is one sequence that has 98800 reads/sequence. While all samples went through the same amount of cycles, a less concentrated sample requires more cycles until the process reaches saturation in PCR, even if all samples go through the same number of cycles. Thus, the number of PCR cycles in the figure can be understood as PCR cycles in which the DNA is effectively amplified.

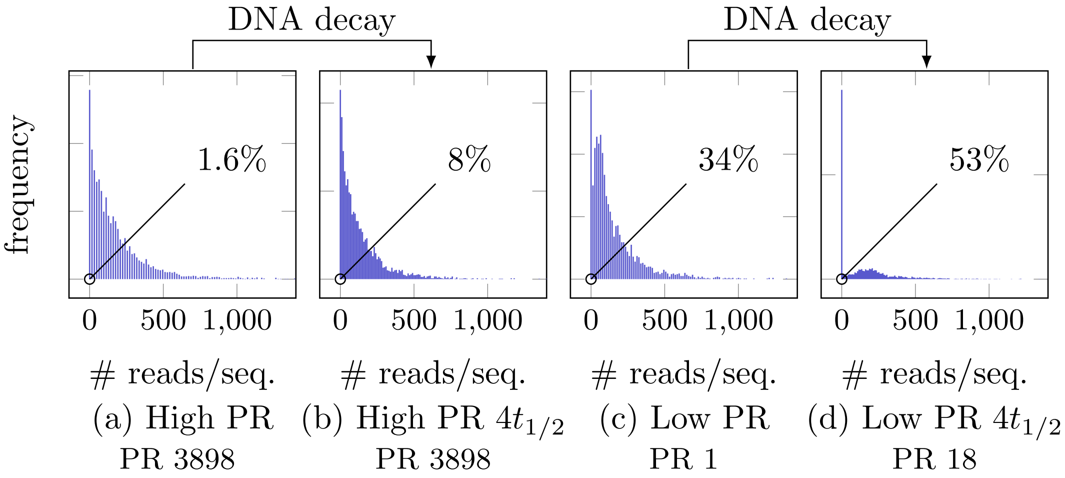


Figure 7. The distribution of the number of reads per sequence before and after storage. The percentage in the picture corresponds to the percentage of molecules that are not observed in the sequenced data. The number of PCR cycles is comparable (65–68 cycles) in all four experiments. In both experiments, the number of lost sequences increases significantly from the original DNA to the aged one.

becomes more pronounced (In the experiment by Erlich and Zielinski[7], all samples undergo 40 PCR cylces, however, it is to be expected that the PCR reaction saturates approximately 3.3 cycles later for every 10 fold dilution).

Errors due to storage. Finally, we examine the effect of storage on the errors by considering the data from our accelerated aging experiments. See Fig. 7 for the corresponding distributions of the molecules. We find that, in each experiment, the number of molecules that are not seen is significantly larger in the aged DNA in the respective dataset compared to the corresponding non-aged dataset. Specifically, as shown in Fig. 7, the number of molecules that are not seen at the output in the dataset High PR is 1.6% and increases to 8% after 4 half-lifes (High PR $4t_{1/2}$), and in the dataset Low PR it is 34% and increases to 53% after 4 half-lifes (Low PR $4t_{1/2}$). Note that the original DNA and the aged DNA were both diluted so that the decayed samples and non-decayed samples required a comparable number of PCR cycles for PCR saturation.

As the storage related DNA decay process can be best described as a first order reaction, the resulting DNA concentration is expected to decay exponentially[5]. At one half-life of DNA decay (corresponding to about 500 years for DNA encapsulated in bone[36]), half of the DNA has decayed and is no longer amplifiable via PCR. To generate DNA with an equivalent thermal history, we have utilized accelerated aging. As non-decayed DNA is diluted prior to amplification and reading to achieve equivalent PCR cycles for the decayed and non-decayed samples, the experiment Low PR versus Low PR $4t_{1/2}$ singles out the effect of aging, since other experimental factors, such as synthesis, amplification and sequencing are kept constant. Figure 7 shows that storage has an effect on the

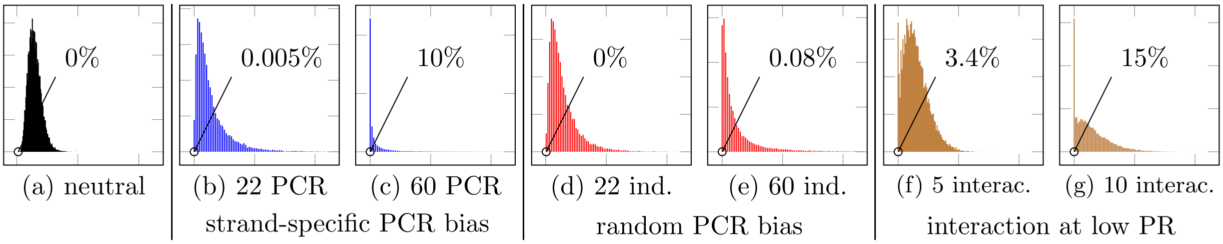


Figure 8. The effect of PCR bias and interaction at low physical redundancy on the distribution of the molecules; the y-axis are the number of molecules, and the x axis in each figure is the number of reads per sequence, and varies from 1–2500 in each figure; the longer tails of the distributions have been cut off. The specific frequency values on the y-axis are not important and are therefore intentionally not show. The percentage value in the figures correspond to the percentage of molecules that are not observed in the sequenced data. Panel (a) shows the distribution of copies obtained by "sequencing" the original pool, and panels (b) and (c) show the distribution after 22 and 60 cycles of PCR, where we took the PCR efficiencies as Gaussian distributed with mean 1.85 and standard deviation 0.07. Panels (d) and (e) depict results from the same experiment, but the efficiencies for each molecule are now different in each cycle and are Gaussian distributed with mean 1.85 and standard deviation 0.25. Finally, panels (f) and (g) show the distribution after repeated interaction with a pool of small (100) physical redundancy. The results show that both PCR bias as well as interaction (i.e., repeatedly taking copies of the DNA) at low physical densities leads to a significant loss of fragments.

distribution of sequences in the pool, most importantly resulting in an decreased number of sequences that are read at least once. Consequently it can be concluded that DNA decay due to storage alters the distribution of the molecules in the pool, which may be attributed to a sequence dependence of the decay process[37–40].

We also found that aging only has a marginal effect on the insertions and deletion probabilities, but increases the substitution error probabilities substantially, as discussed previously.

Impact of PCR bias and processing DNA at low physical redundancy: A computational experiment.

In this section, we perform a simple experiment demonstrating that PCR bias and processing DNA at low physical redundancies significantly impacts the distribution of the reads. The corresponding results contribute to interpreting and understanding the experiments discussed in the previous sections.

Towards this goal, we generate a random pool of DNA by taking $M = 20{,}000$ sequences, and generate copies of those fragments by drawing from a Gamma distributed with shape parameter 8 and scale parameter 16. Thus, the expected physical density in the pool is $8 \cdot 16 = 128$. We artificially sequence from this pool at physical redundancy 300 by taking $300M$ draws with replacement. See Fig. 8(a) for the corresponding distribution of molecules. Note that this implies that the reads are approximately Poisson-Gamma distributed. Specifically, they are approximately distributed as a mixture of Poisson distributions, with the mean parameter γ of the Poisson distribution being Gamma distributed). A Poisson-Gamma mixture distribution in turn is equivalent to a negative binomial distribution, which has been found to model empirical sequencing distributions well. While this justifies our model for generating a random pool of DNA, the particular distribution of the molecules in the pool is not crucial for illustrating how PCR bias impacts the distribution of the molecules, and the conclusions of the computational experiments discussed next are not sensitive to the particular distribution of the molecules in the pool; for example they continue to hold if the molecules in the pool are uniformly distributed.

PCR bias. We start by simulating the impact of PCR bias. In theory, a single PCR step generates exactly one copy per molecule. In reality, PCR amplifies each fragment on average by a factor that is tightly concentrated around a value slightly smaller than two[20–23]. In our first experiment, we model the efficiency E of PCR as a Gaussian random variable with mean 1.85 and standard deviation 0.07. We then amplify each molecule by a factor $E^{N_{\mathrm{PCR}}}$, where N_{PCR} is the number of PCR cycles. The corresponding distributions, obtained by artificially sequencing as explained above, after 22 and 60 such PCR steps, are depicted in Fig. 8(b,c). Note that after a moderate number of PCR steps (22), the distribution is still approximately negative binomial, albeit with a longer tail. After many PCR steps (60), however, a significant proportion of molecules is not seen and the distribution has a long tail, since the proportions of the molecules in the pool changes drastically, with some frequencies being much more frequent than others.

Note that this experiment assumes that the PCR efficiency is strand specific and constant at each cycle of PCR. While this has been found to be true in the literature[20–23], we point out that we would observe the same effect even if the PCR efficiency would not be strand specific or would slightly vary at each cycle. To see that, we perform the same experiment, but now, at each cycle of PCR, we draw a strand specific DNA efficiency from a Gaussian distribution with mean 1.85 and standard deviation 0.25. While this is chemically implausible, it shows that the findings from the previous experiment are not sensitive to the assumption of each molecule being duplicated with exactly the same strand-specific factor. The corresponding reading distributions, after 22 and 60 such "molecule independent" PCR steps is shown in Fig. 8(d,e).

DNA storage at low physical redundancy. A common viewpoint is that when storing data on DNA, we can interact with the DNA by reading the DNA many times, or taking copies of the DNA as we please, without significantly changing the pool. However, at low physical redundancy, such interactions with the DNA change the distribution of the sequences in the pool significantly, and in particular lead to a significant loss of sequences. To demonstrate this effect, we next consider a computational experiment that repeatedly takes a subset of the DNA, and then applies perfect PCR (without a bias) to re-obtain a pool of the original size. This corresponds to making copies of a pool of DNA or reading the pool at low physical redundancy. We generate a pool consisting of 100 copies of $M = 20,000$ distinct DNA molecules, corresponding to a physical redundancy of 100. We then take a small part of the pool and perfectly amplify that part. In particular, we choose one tenth of the pool uniformly at random, and then amplifying each sequence by a factor of 10. We then read at coverage 300 by taking 300 M draws with replacement from the pool, as before. The results after 5 and 10 such interaction steps are depicted in Fig. 8(f,g). The results show that repeated interaction with the pool at very low physical densities can lead to a significant loss of the fragments. In contrast, interaction at high physical redundancy does have very little effect on the distribution (not shown here).

Discussion

In this paper, we characterized the error probabilities of the DNA data storage systems, and assigned, when possible the error sources to processes such as reading and writing the DNA. The key findings are as follows.

Errors within sequences are mainly due to synthesis, sequencing, and to a smaller extent to decay of DNA. Synthesis introduces mainly deletion errors, and potentially a few insertions. If next generation sequencing technology is used for reading, then sequencing introduces additional substitution errors, but only very few insertions or deletions, if at all. Long term storage leads to additional substitution errors, due cytosine deamination. However, the majority of the reads has no substitution or deletions errors, thus the errors within sequences are relatively small.

PCR bias significantly changes the distribution of the reads, and a large number of PCR cycles leads to a distribution with a very long tail, which in turn increases the number of sequences that are never read. PCR bias is particularly apparent when storing DNA at low physical redundancy.

Storing DNA at low physical redundancy is desirable—amongst a variety of other important parameters such as longevity, cost, read-and write speeds—since it results in a large information density. However, interacting with DNA at low physical redundancy by copying the pool or sequencing parts of the pool changes the corresponding read distributions and results in a significant loss of sequences. Also, low physical redundancies ask for more PCR cycles, making the system more prone to PCR bias. Hence, if data is stored at low physical redundancies, the anticipated interactions have to be taken into account in the design of the system, in particular the parameters of the error-correcting scheme. Distortion at low physical redundancy may not be observed when working with high physical redundancies, and it is desirable to further understand the sequence dependency of these processes (especially polymerase efficiency and DNA decay) in order to improve the design of DNA storage systems.

The consequences of our key findings for designing DNA storage systems, in particular error correcting schemes for DNA data storage systems, are as follows. DNA storage systems typically use so called outer and inner codes to correct errors. The outer code adds redundancy in the form of additional sequences in order to be able to recover lost sequences, and the inner code adds redundancy within each sequence, in order to be able to correct errors on the sequence level (for example substitution errors). Our error characterization shows that an outer code is absolutely crucial in order to account for the loss of entire sequences since such loss is unavoidable. The parameters of the code, i.e., the number of erasures that it can correct should be chosen based on the expected number of lost sequences, which in turn depends on the storage time, physical redundancy, and the interactions anticipated with the DNA (such as PCR steps, number of copies). Since the error probability within sequences is generally very low (most reads are error-free), there are three sensible approaches to dealing with them: (i) Errors on a sequence level can be, at least partly, corrected with an inner code. (ii) The reads can algorithmically be combined, in order to obtain an estimate of each sequence. Due to redundant reads some of the errors within sequences can thus be potentially be corrected. (iii) The outer code can also correct the reads that are erroneous. Of course, the three approaches can be combined, for example one can use an inner code to correct some errors, and correct errors that the inner code cannot correct with the outer code, or one can only algorithmically combine the fragments, and again correct the remaining errors with the outer code. The design choice can be based on the error probabilities that are anticipated from the estimates reported in this paper.

References

1. Neiman, M. S. Some fundamental issues of microminiaturization. *Radiotekhnika* **1**, 3–12 (1964).
2. Baum, E. B. Building an associative memory vastly larger than the brain. *Sci.* **268**, 583–585 (1995).
3. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Sci.* **337**, 1628–1628 (2012).
4. Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nat.* **494**, 77–80 (2013).
5. Grass, R., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie Int. Ed.* **54**, 2552–2555 (2015).
6. Yazdi, H. T., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A rewritable, random-access DNA-based storage system. *Sci. Reports* **5** (2015).
7. Erlich, Y. & Zielinski, D. DNA fountain enables a robust and efficient storage architecture. *Sci* (2017).
8. Organick, L. *et al.* Random access in large-scale dna data storage. *Nat. Biotechnol* (2018).
9. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Sci.* **329**, 52–56 (2010).
10. Bornholt, J. *et al.* A DNA-Based Archival Storage System. In *Proc. of ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 637–649 (2016).
11. LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150 mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522–2540 (2010).
12. Agbavwe, C. *et al.* Efficiency, error and yield in light-directed maskless synthesis of dna microarrays. *J. Nanobiotechnology* **9** (2011).
13. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).

14. Schmidt, T. L. *et al*. Scalable amplification of strand subsets from chip-synthesized oligonucleotide libraries. *Nat. Commun.* **6**, 8634 (2015).
15. Sack, M., Kretschy, N., Rohm, B., Somoza, V. & Somoza, M. M. Simultaneous light-directed dynthesis of mirror-image microarrays in a photochemical reaction cell with flare suppression. *Anal. Chem.* **85**, 8513–8517 (2013).
16. Singh-Gasson, S. *et al*. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.* **17**, 974–978 (1999).
17. Maurer, K. *et al*. Electrochemically generated acid and its containment to 100 micron reaction areas for the production of DNA microarrays. *Plos One* **1**, e34 (2006).
18. Cline, J., Braman, J. C. & Hogrefe, H. H. PCR fidelity of PFU DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res.* **24**, 3546–3551 (1996).
19. Lubock, N. B., Zhang, D., Sidore, A. M., Church, G. M. & Kosuri, S. A systematic comparison of error correction enzymes by next-generation sequencing. *Nucleic Acids Res.* **45**, 9206–9217 (2017).
20. Ruijter, J. M. *et al*. Amplification efficiency: linking baseline and bias in the analysis of quantitative pcr data. *Nucleic Acids Res.* **37** (2009).
21. Pan, W. *et al*. DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnol.* **14**, 10 (2014).
22. Warnecke, P. M. *et al*. Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res.* **25**, 4422–4426 (1997).
23. Caldana, C., Scheible, W.-R., Mueller-Roeber, B. & Ruzicic, S. A quantitative RT-PCR platform for high-throughput expression profiling of 2500 rice transcription factors. *Plant Methods* **3** (2007).
24. Ross, M. G. *et al*. Characterizing and measuring bias in sequence data. *Genome Biol.* **14** (2013).
25. Lindahl, T. & Nyberg, B. Rate of depurination of native deoxyribonucleic acid. *Biochem.* **11**, 3610–3618 (1972).
26. Suzuki, T., Ohsumi, S. & Makino, K. Mechanistic studies on depurination and apurinic site chain breakage in oligodeoxyribonucleotides. *Nucleic Acids Res.* **22**, 4997–5003 (1994).
27. Lindahl, T. & Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochem.* **13**, 3405–3410 (1974).
28. Yazdi, H. T., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Sci. Reports* **7** (2017).
29. Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R. & Hannon, G. J. Alta-cyclic: a self-optimizing base caller for next-generation sequencing. *Nat. Methods* **5**, 679–682 (2008).
30. Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinforma.* **17**, 125 (2016).
31. Schwartz, J. J., Lee, C. & Shendure, J. Accurate gene synthesis with tag-directed retrieval of sequence-verified DNA molecules. *Nat. Methods* **9**, 913 (2012).
32. Bentley, D. R. *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nat.* **456**, 53 (2008).
33. Nelms, B. L. & Labosky, P. A. A predicted hairpin cluster correlates with barriers to PCR sequencing and possibly BAC recombineering. *Sci. Reports* **1** (2011).
34. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinforma.* **27**, 2957–2963 (2011).
35. Heckel, R., Shomorony, I., Ramchandran, K. & Tse, D. N. C. Fundamental limits of DNA storage systems. In *IEEE International Symposium on Information Theory (ISIT)*, 3130–3134 (2017).
36. Allentoft, M. E. *et al*. The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proc. Royal Soc. Lond. B: Biol. Sci* (2012).
37. Pedone, F. & Santoni, D. Sequence-dependent DNA helical rise and nucleosome stability. *BMC Mol. Biol.* **10**, 105 (2009).
38. Fujii, S., Kono, H., Takenaka, S., Go, N. & Sarai, A. Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.* **35**, 6063–6074 (2007).
39. Goddard, N. L., Bonnet, G., Krichevsky, O. & Libchaber, A. Sequence dependent rigidity of single stranded DNA. *Phys. Rev. Lett.* **85**, 2400–2403 (2000).
40. Hunter, C. A. Sequence-dependent DNA structure. the role of base stacking interactions. *J. Mol. Biol.* **230**, 1025–1054 (1993).

Acknowledgements

We would like to thank Yaniv Erlich for sharing the sequencing data of their paper[7], and for helpful comments on a previous version of this manuscript. We would also like to thank Goldman and his group for making the data from the paper[4] publicly available.

Author Contributions

R.H. conceived the analysis, G.M. and R.G. did all work related to the chemistry, R.H. analyzed the data and estimated the error probabilities, R.H. and R.G. wrote the paper. All authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.