






Quantifying molecular bias in DNA data storage

Yuan-Jyue Chen ¹✉, Christopher N. Takahashi², Lee Organick², Callista Bee ², Siena Dumas Ang¹, Patrick Weiss³, Bill Peck³, Georg Seelig ^{2,4}, Luis Ceze ²✉ & Karin Strauss ¹✉

DNA has recently emerged as an attractive medium for archival data storage. Recent work has demonstrated proof-of-principle prototype systems; however, very uneven (biased) sequencing coverage has been reported, which indicates inefficiencies in the storage process. Deviations from the average coverage in the sequence copy distribution can either cause wasteful provisioning in sequencing or excessive number of missing sequences. Here, we use millions of unique sequences from a DNA-based digital data archival system to study the oligonucleotide copy unevenness problem and show that the two paramount sources of bias are the synthesis and amplification (PCR) processes. Based on these findings, we develop a statistical model for each molecular process as well as the overall process. We further use our model to explore the trade-offs between synthesis bias, storage physical density, logical redundancy, and sequencing redundancy, providing insights for engineering efficient, robust DNA data storage systems.

¹Microsoft Research, Redmond, Washington 98052, USA. ²Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA. ³Twist Bioscience, San Francisco, California 94158, USA. ⁴Department of Electrical and Computer Engineering, University of Washington, Seattle, Washington 98195, USA. ✉email: yuanjc@microsoft.com; luisceze@cs.washington.edu; kstrauss@microsoft.com

Storing data in DNA is attractive due to its information density of petabytes of data per gram, and excellent durability¹. Relative to other forms of molecular-level or atomic-level data storage, DNA is unique because of the ease of copying DNA (i.e., using PCR) and its eternal relevance (people will always be interested in sequencing DNA)². High-throughput (HT) sequencing and synthesis technologies^{3,4} have evolved and made storing information in synthetic DNA an increasingly realistic alternative to traditional long-term storage methods^{5–8}. However, the sequencing coverage (number of read counts of a unique sequence) of an oligonucleotide (henceforth referred to simply as “oligo”) was found to be very uneven, requiring modern error correction codes capable of handling sequence dropout^{7–12}. Current methods typically require either trial-and-error of experimental protocols or brute-force use of hundreds to thousands of sequencing reads per sequence to capture under-represented sequences. This inefficiency stems from lack of understanding about bias in oligo copy distribution, as well as how it changes as the oligos are manipulated in DNA data storage systems.

In more recent work, errors and bias were studied using sequencing data from DNA storage systems¹³. However, direct PCR and sequencing in a DNA storage system cannot distinguish bias created by DNA synthesis from bias caused by PCR and sequencing. As our first foray in separating bias effects stemming from DNA synthesis versus PCR, we tag an arbitrarily chosen DNA archival file with over 400,000 sequences using unique molecular identifiers (UMI), random barcodes to label each molecule¹⁴. UMI labeling allow us to decouple synthesis bias from PCR bias, and we find significant bias from DNA synthesis. To corroborate this finding, we order from Twist Bioscience a carefully designed ready-to-sequence pool with 1,536,168 sequences, each of which unique, and already containing necessary segments of DNA to be sequenced. This ready-to-sequence pool can be sequenced using an Illumina sequencer directly, with no need for intermediate PCR or DNA ligation steps required for sequencing library preparation. Thus we can quantify synthesis oligo distribution without any interference from molecular processes. To the best of our knowledge, this is the first time an oligo pool from array-based synthesis is characterized in this way. We find that synthesis bias is highly related to spatial location of oligos on a synthesis chip.

After quantifying synthesis bias, we study PCR bias from two sources—guanine/cytosine (GC) content and PCR stochasticity. GC content of individual sequences were previously found to affect PCR amplification efficiency in biological DNA^{15–17}. In DNA storage, the GC content of each strand is determined by a data-to-DNA sequence encoder. We test GC bias using two different oligo pools: one pool is encoded to avoid all homopolymers (non-homopolymer pool); in contrast, the other is encoded without homopolymer avoidance steps (homopolymer pool). Even though these two encoding strategies lead to different GC distributions, we find no practically important association between GC content and PCR bias. Instead, we find that PCR stochasticity widens oligo copy distributions of our test DNA archival file and, based on our observations, seems to be a dominant factor in PCR bias. PCR is an exponential process, so small random variations early on in amplification can have a large impact on distribution^{18–24}.

Based on these observations, we construct a computational model for predicting molecular bias in a DNA data storage system (Fig. 1). We observe strong association between the bias predicted from this model and from our experimental data. Furthermore, we use our model to investigate the tradeoffs between synthesis bias, physical redundancy for storing DNA (i.e., oligo copy number), logical redundancy (additional information to aid error correction and mitigate missing sequences), and sequencing redundancy (i.e., sequencing coverage). A system model can be very useful to determine the best parameters for a given DNA storage system.

Results

DNA synthesis is a prominent source of sequence bias. Determining the source of bias in DNA data storage, and more generally in arbitrary DNA pools, is complicated because synthesis bias and PCR bias are typically coupled. To decouple them, we applied UMI, barcodes to individually identify each molecule of an initial pool, in our case an arbitrarily chosen DNA file with over 400,000 sequences (Fig. 2a and Supplementary Fig. 1). Synthetic DNA pools include multiple copies of each sequence, and UMI labeling ensures with high probability that each molecule will include a tag different from any other. The UMI-labeled oligos were sequenced, and the resulting reads were aligned to the

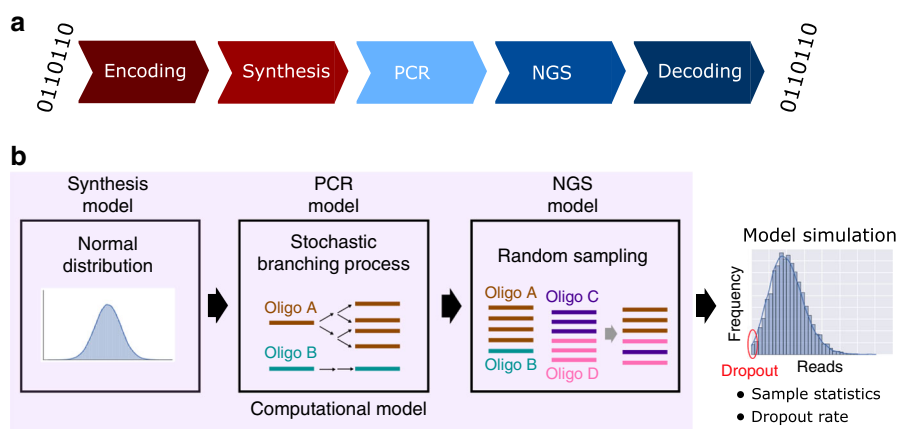


Fig. 1 | DNA storage system model. a Workflow of DNA data storage. Digital information is first encoded into oligonucleotide (oligo) sequences, resulting in multiple 150-base DNA fragments synthesized using array-based DNA synthesis technology, which are then stored. To read back the stored data, target DNA oligos can be selectively (random-) accessed using polymer chain reaction (PCR), then sequenced via next generation sequencing (NGS), and decoded back to digital information. **b** The computational model approximates each molecular process in the DNA storage system: it uses a normal distribution for modeling sequence copy numbers from synthesis, a stochastic branching process for PCR, and random sampling for sequencing. The computational model makes predictions for oligo copy distribution to help researchers estimate statistics such as sequence dropout rate.

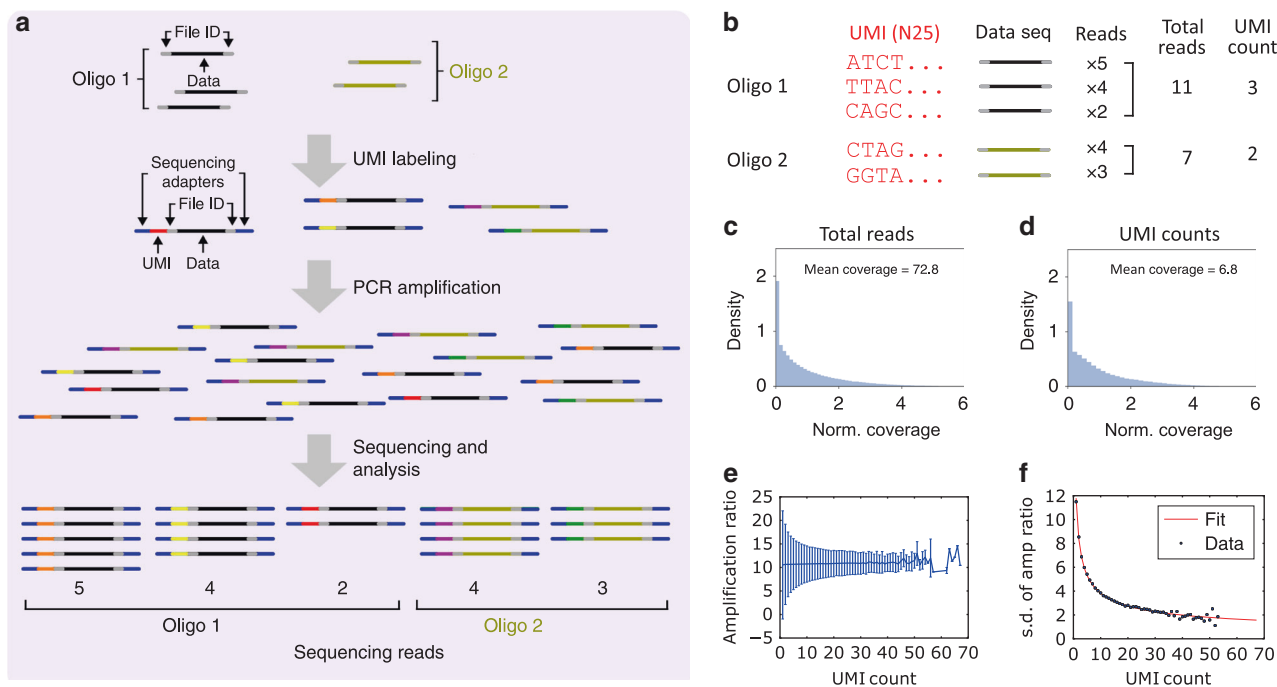


Fig. 2 Estimating oligonucleotide bias using unique molecule identifiers (UMIs). **a** Overview of tagging each single DNA molecule with UMIs. Each oligo sequence (e.g., represented in black, beige) in a pool has multiple copies and each copy is labeled with a UMI (represented in different colors) and universal Illumina sequencing adapters (represented in gray). After UMI labeling, oligos are PCR-amplified and sequenced. **b** Hypothetical examples of UMI counting. The UMI count of each sequence is a proxy for the oligo copy number from DNA synthesis. The total number of reads containing the same UMI is a proxy for the number of copies of a DNA molecule created by PCR. **c** The distribution of number of reads for each sequence, normalized to 83.0 mean coverage. Read counts are normalized to form a probability density (y-axis); the integral of the probability density is 1 (see "Methods" section). **d** The distribution of UMI counts for each sequence, normalized to 7.7 mean coverage. The biased UMI count distribution indicates that pools are already biased immediately after DNA synthesis, before any PCR is performed. **e** Amplification ratio versus UMI count. The average amplification ratio is roughly constant across UMI counts, but oligos with low initial copy numbers show higher variation. The error bars indicate standard deviation (s.d.) of amplification ratio. **f** Standard deviation (s.d.) of amplification ratio versus UMI count. The experimental data agree with Eq. (1). The number of unique sequences (sample size) is 457,772. Source data are available in the Source Data file.

file sequences in two manners. First, these reads were aligned to individual sequences in the file using Burrows-Wheeler Aligner (BWA)²⁵, independent from UMI, and their respective counts (coverage) are reported in Fig. 2c. Second, the same reads were aligned to sequences in the file, then further filtered by UMI label (Fig. 2b), and finally reported in Fig. 2d. The UMI-filtered results are a proxy for the oligo distribution after DNA synthesis, and the copy number is clearly variable, indicating that the old synthesis process is far more skewed (there have since been process improvements, discussed in the next section). This distribution is also very similar in shape to the distribution after PCR, indicating that PCR does not significantly increase bias overall. Nevertheless, PCR still has an impact on individual sequence counts, so we decided to study the amplification ratio of each sequence as a function of the number of initial molecules representing it. We define the amplification ratio to be the ratio of total reads after PCR to UMI count (i.e., oligo count before PCR) for each sequence. Figure 2e shows that regardless of the initial oligo copy number, the average amplification ratio remains constant. On the other hand, the amplification ratio was observed to have high variation when oligos had very low copy numbers, indicating that the amplification ratio was affected by stochastic effects at these low copy numbers. Indeed, since a PCR process is composed of successive rounds of binomially distributed copying (each molecule has some probability of being copied), we would expect the standard deviation (s.d.) of the amplification ratio to be inversely proportional to the square root of the initial number of strands. Additionally, since observation takes a sequencing reaction

(another binomial process) we would expect a constant amount of added deviation. These observations lead us to the model:

$$\sigma_{\alpha} = \frac{a}{\sqrt{\text{UMI count}}} + b \quad (1)$$

where σ_{α} is the s.d. of the amplification ratio, and a and b are constants. Our experimental data was fitted using Eq. (1) and shown in Fig. 2f.

Bias is related to the spatial location on the synthesis chip. To further understand the synthesis bias, we ordered a carefully designed a ready-to-sequence pool with 1,536,168 unique DNA sequences from Twist Bioscience. Oligos in this pool already contain universal Illumina adapters and Illumina sequencing primers on both the 5' and 3' ends, allowing us to sequence it without any sequencing library preparation such as PCR or ligation. By mapping the sequencing reads of each sequence back to its corresponding location on the synthesis chip, a distinct pattern can be observed (Fig. 3a), indicating that synthesis bias was related to the spatial location on the synthesis chip. After further discussion with Twist Bioscience, their synthesis process was improved, and the oligo counts on the synthesis chip became much more even (Fig. 3c). Interestingly, the oligo distribution before the synthesis process improvement did not follow a normal distribution, but the oligo distribution using the improved synthesis process is now well fitted to a normal distribution (Fig. 3b, d).

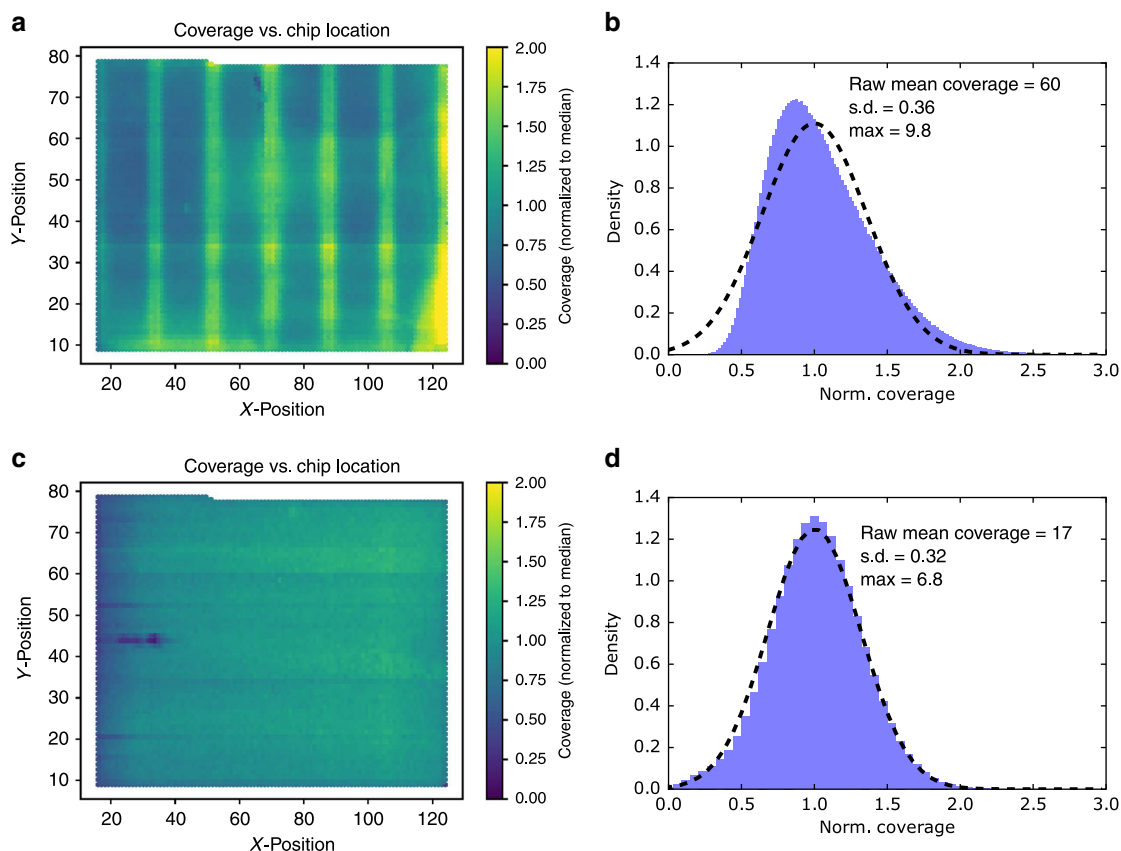


Fig. 3 Oligo copy distribution on the synthesis chip. **a** The sequencing coverage of each oligo from the first ready-to-sequence pool was mapped back to its corresponding location on the synthesis chip. Coverages are normalized to the median. **b** The histogram of normalized sequencing coverage of the first ready-to-sequence pool (blue). The distribution does not fit a normal distribution (dashed line). **c** The sequencing coverage of each oligo from the second ready-to-sequence pool, mapped back to its corresponding location on the synthesis chip. Coverages are normalized to the median. **d** The histogram of normalized sequencing coverage of the second ready-to-sequence pool (blue). The second ready-to-sequence pool showed much more even oligo copy distribution and fits a normal distribution (dashed line) much more closely than the first. The unique number of sequences (sample size) is 1,536,168. Source data are available in the Source Data file.

Oligo synthesis quality lives within a select set of parameters broadly defined as dosage, where we define dosage as: {time, temperature, and concentration}. The boundaries of this quality window define a dosage tolerance relevant to the particular application, in this case oligonucleotides used for data storage. Initial data used in this paper were produced with a process that allowed process excursions outside of the dosage tolerance window. To address these spatial gradients in quality, Twist Bioscience increased the dosage tolerance window with proprietary chemical modifications to the phosphoramidite chemistry, making the additive synthesis process less susceptible to process dosage excursions. Twist Bioscience also introduced engineering changes to the hardware and chemical process parameters to ensure more uniform evacuation of chemical reagents in the flowcell process with enhanced temporal control. The combination of these changes has resulted in dramatically decreased error rates and more robust molecules in subsequent processes.

Population fraction change for quantifying PCR bias. We now turn to studying the PCR bias by creating metrics to quantify it at the sequence level. We begin by defining the population fraction of a sequence i after $k \in \{Z_{\geq 0}\}$ cycles of PCR as

$$x_i^{(k)} := \frac{N_i^{(k)}}{\sum_j N_j^{(k)}} \quad (2)$$

where $N_i^{(k)}$ is the number of reads of sequence i after k PCR cycles. Here j is across all sequences. We then define the population fraction change for sequence i to be

$$Q_i = Q_i^{(k)} := \frac{x_i^{(k)}}{x_i^{(0)}}. \quad (3)$$

We consider a PCR process to be unbiased when $\mathbb{E}[Q_i | x_i^{(0)} > 0] = 1$ for all sequences, that is, no sequence becomes over or underrepresented after a PCR, whereas we consider a PCR process to be biased when $\mathbb{E}[Q_i | x_i^{(0)} > 0] \neq 1$ for any sequence i . We then can say that experiments with higher standard deviation over the population fraction change, $\text{s.d.}[Q := \{Q_i | x_i^{(0)} > 0\}]$, show more bias when all other conditions are equivalent. It is worth noting that even an unbiased process will have $\text{s.d.}[Q] > 0$ for finite sample sizes. Furthermore, $\text{s.d.}[Q]$ should asymptotically decrease with the total number of reads.

PCR bias is not correlated with GC content. Although previous studies observed PCR bias in genomic biological sample amplification^{15–17}, it remained unclear whether such bias is significant in DNA data storage. To assess this, we used the 1.5 million-sequence ready-to-sequence pool and compared its distribution before PCR and after PCR. The ready-to-sequence pool was sequenced in two ways: (1) directly from the synthesized pool and (2) after one 6-cycle plus five 5-cycle PCR processes, for a total of

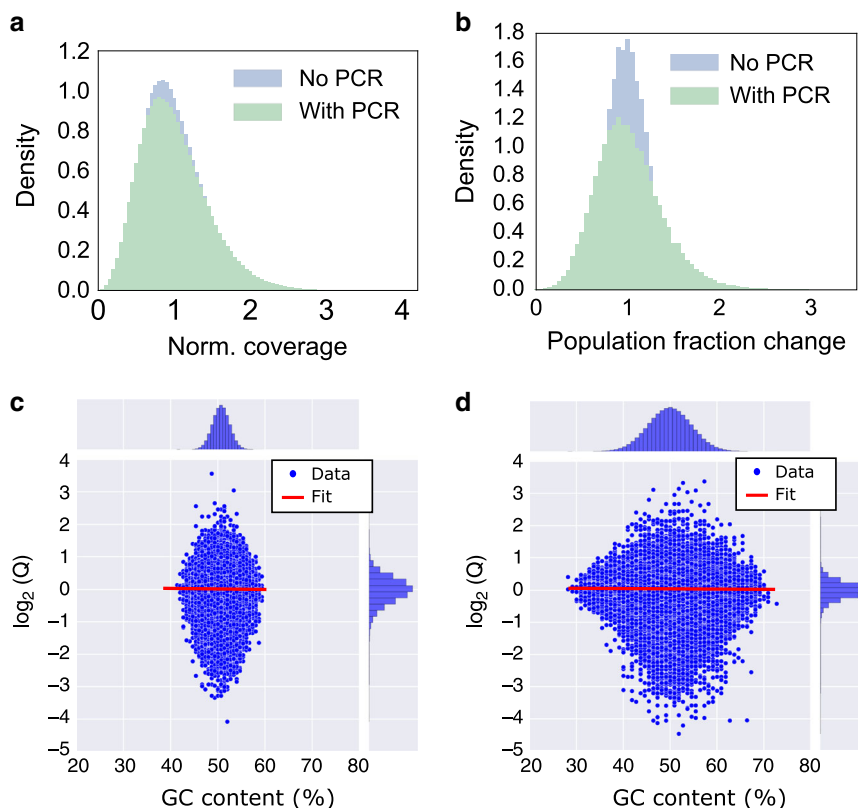


Fig. 4 PCR-induced Population Fraction Changes and Impact of GC content. **a** Histogram of sequencing coverages for optimized ready-to-sequence pool. The ready-to-sequence pool was sequenced directly, without PCR (blue). The same pool was amplified using PCR for 31 cycles (green) and sequenced separately. Both were randomly sampled to coverage of 20 \times for direct comparison, and they look quite similar (c.v. = 0.41 and c.v. = 0.45, respectively). **b** The blue histogram shows the population fraction change Q distribution of the ready-to-sequence pool, before PCR, after being subsampled to 20 \times coverage, with respect to its overall available coverage (60 \times). The green histogram shows the population fraction change Q distribution of the ready-to-sequence pool, after PCR and after being subsampled to 20 \times coverage, with respect to the pre-PCR pool distribution at 60 \times coverage. The blue distribution shows the effect of subsampling, while the further widening of the green curve with respect to the blue curve is attributed to stochastic bias in the PCR process. The sample size (number of unique sequences) in **a**, **b** is 1,536,168. **c**, **d** The GC content is plotted against the \log_2 of the population fraction change Q for the ready-to-sequence, non-homopolymer pool (**c**) and a homopolymer pool (**d**). The experimental data are shown as blue dots, and the linear fit is shown as a red line. The histograms of GC content and $\log_2(Q)$ are shown at the top and right, respectively. The sample size (number of unique sequences) in **c**, **d** is 1,536,168 and 1,358,998, respectively. Source data are available in the Source Data file.

31 cycles. Each PCR process was limited to no more than 6 cycles to prevent resource exhaustion (i.e., there was always an excess of primer and other reagents). Sequencing data (Fig. 4a) shows qualitatively little change in the coefficient of variation (c.v.) of oligo copy distribution before and after PCR (0.41 and 0.45, respectively, when both are subsampled to 20 \times coverage). The two datasets were then compared at a sequence level by observing population fraction changes with respect to the overall available pre-PCR pool coverage, 60 \times (Fig. 4b). The distribution before PCR shows the effect of subsampling on population fraction, and the distribution after PCR shows the effect of PCR itself. The latter showed much higher standard deviation. The standard deviations of population fraction changes were 0.24 and 0.37 before PCR and after PCR, respectively, and these two numbers were statistically different ($p < 0.005$, computed by bootstrapping $n = 1000$). This indicates that PCR increased bias relative to a random sampling process.

We then asked whether population fraction changes were caused by GC content. We first examined the ready-to-sequence pool, which was encoded to avoid homopolymers⁶ (Fig. 4c). Although the association between population fraction changes and GC content of this pool (between 40 and 60%) was found to be statistically significant (P value < 0.05), the association between the two was very small and practically unimportant (the slope of

the linear fit was < 0.01). Additionally, we tested another 9 different DNA archival files with a total of 1,358,998 unique sequences that allow random homopolymers (Fig. 4d; Supplementary Fig. 2 shows experimental workflow details). These homopolymer files had a wider range of GC content from 25 to 75%, but the association between GC content and the population fraction changes was still very small and not practically important (the slope of the linear fit was < 0.01). The negligible bias impact from GC content in our experimental data was likely because these oligos were relatively short (150-nt), and the use of KAPA HIFI polymerase also reduced the impact of GC bias²⁶. Having established that GC content was not the main effect being observed, we turned to hypothesizing that PCR stochasticity was the culprit.

PCR stochasticity can lead to significant bias. Because PCR is not perfect (i.e., replication of an individual molecule has a probability of less than one), even small random divergence in early phases of amplification can create significant bias, which is known as PCR stochastic bias. We have shown that PCR bias is related to oligo copy number in the UMI quantification experiment, especially for sequences with low copy numbers in the initial pool (from a previous PCR process or from a biased

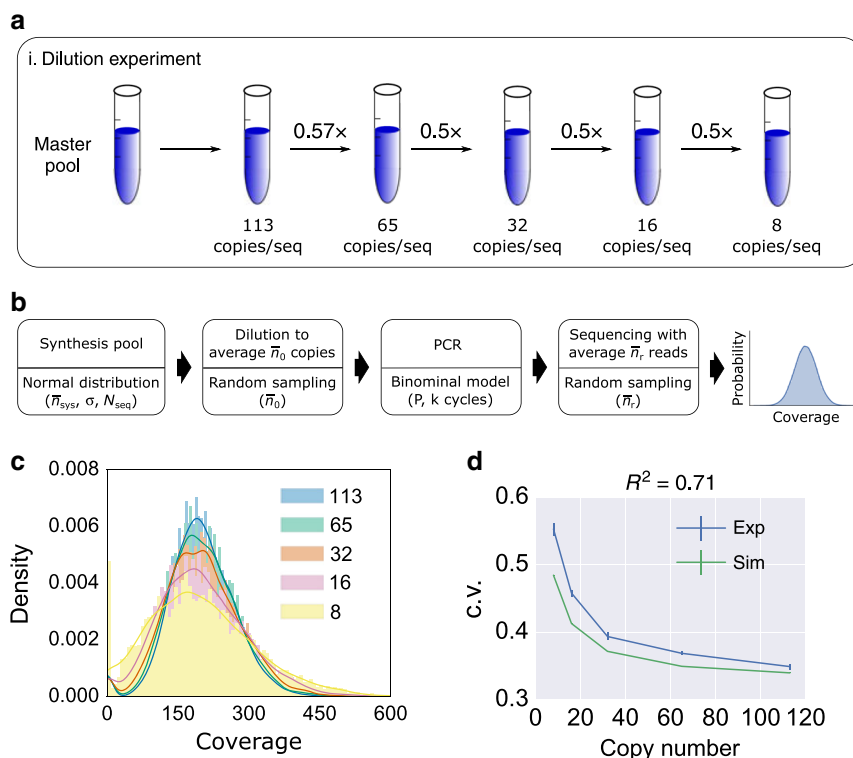


Fig. 5 Dilution-PCR experiment. **a** The experimental workflow. A master DNA pool was diluted to different average copy numbers as indicated in the drawing. Each dilution sample was PCR-amplified and sequenced using an Illumina NextSeq instrument, and the results sampled at 200x coverage. **b** A computational model for the dilution-PCR experiments. The synthesis pool model used $N_{seq} = 7,373$ number of sequences, and normally distributed copy numbers with mean $\bar{n}_{syn} = 10^8$, and standard deviation $\sigma = 3.2 \times 10^7$. The c.v. of the synthesis pool in this simulation ($\frac{\sigma}{\bar{n}_{syn}} = 0.32$) was set to be equal to the c.v. of our ready-to-sequence pool sequenced at mean coverage 17. The dilution process was simulated using random sampling with a mean copy number \bar{n}_0 , ranging from 8 to 113. PCR was simulated as a binomial process with a probability of successful amplification $P = 0.95$ and 18 PCR cycles. The simulated sequencing result was obtained using random sampling with an average coverage $\bar{n}_r = 200$. **c** Simulated post-PCR sequencing coverage histogram of each dilution-PCR sample. The initial (pre-PCR) average copy number of each histogram is shown in the legend, ranging from 8 to 113. Coverage counts are normalized to display a probability density. A Gaussian estimated density curve is added as an outline of each histogram to help with visualization. **d** Sequencing coverage c.v. of the post-PCR mix versus average copy number in the pre-PCR mix. The model prediction (green) shows good agreement with the experimental data (blue) with $R^2 = 0.71$. The error bars of experimental data indicate standard error of the mean calculated from triplicate experiments. The error bars of model outputs indicate standard error of the mean calculated from 100 repeated simulations. Source data are available in the Source data file.

synthesis pool). Now we want to understand better how PCR stochastic bias affects our DNA storage system.

To quantify PCR stochastic bias, we used an arbitrarily chosen DNA pool with 7,373 sequences to perform a serial dilution-PCR experiment (Fig. 5a). The master pool was diluted to different average copy numbers ranging from 8 to 113 (the copy numbers were quantified using qPCR). Then each sample was amplified with 18 cycles of PCR using primers with Illumina sequencing primer overhangs. Subsequently, a second step of PCR was carried out to include the Illumina adapters where we adjusted the number of cycles to equalize the final library concentration (Supplementary Fig. 3 shows workflow details). The second PCR was carried out at high copy number of the templates (over a million oligo copies per sequence) to avoid introducing additional bias. Our experimental results show that as average copy number decreased, oligo distribution skewed further away from its mean (Fig. 5c). We plot average copy number in a pre-PCR mix against the coefficient of variation (c.v.) of sequencing coverage (Fig. 5d) and standard deviation of population fraction change Q (Supplementary Fig. 4). Both plots show that the lower oligo copy numbers were, the greater the PCR stochastic bias was.

A computational model can predict molecular bias. After characterizing the bias caused by synthesis and PCR sequencing

retrieval, we construct a DNA storage model that encompasses the entire workflow of DNA storage, starting from synthesis \rightarrow aliquot into pre-PCR reaction \rightarrow PCR amplification with k cycles \rightarrow sequencing with mean \bar{n}_r reads (Fig. 5b). We model the oligo copy distribution of synthesis as a normal distribution with total number of sequences N_{seq} , mean copy number per sequence \bar{n}_{syn} , and standard deviation of oligo copy number σ . The PCR process is modeled as a stochastic branching process using the following recursive equation:

$$n_{j+1} = n_j + B(n_j, P) \quad (4)$$

where n_j is the number of molecules in the j -th cycle; $B(n_j, P)$ is a binomially distributed random variable with n_j molecules, and P is the probability of a successful amplification. Illumina sequencing was previously observed to have bias on GC-extreme sequences^{15,27,28}, but GC content in our files did not show practically significant bias in the PCR GC bias test. Therefore, high-throughput sequencing and sample dilution are modeled using random sampling. Note that for performance reasons our model does not perform stochastic simulation for high copy number PCR. PCR carried out at high copy number of templates should obey the law of mass action and therefore be effectively deterministic.

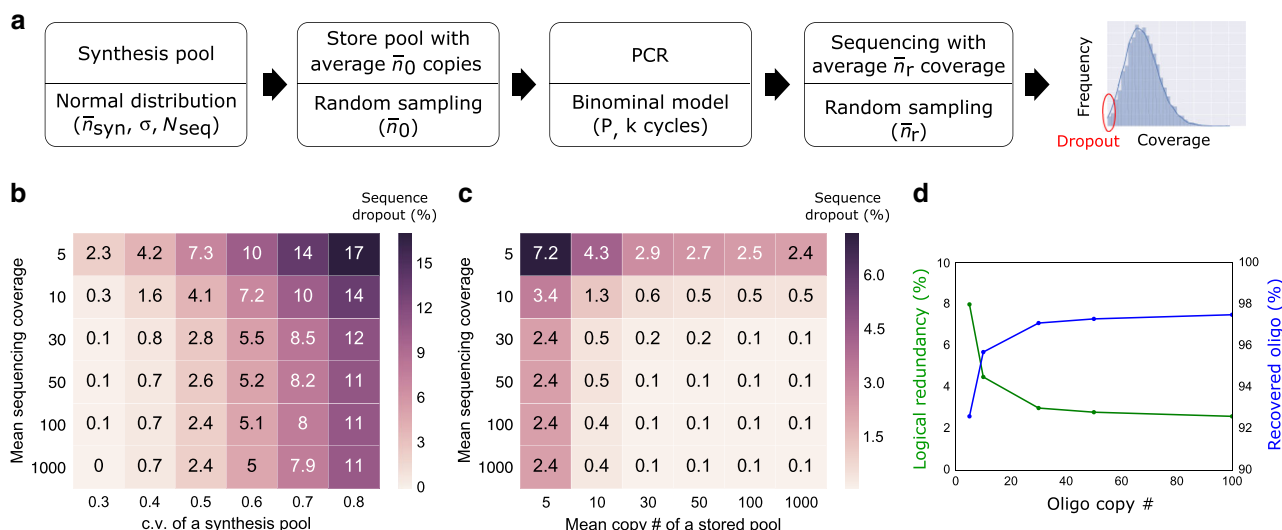


Fig. 6 A computational model can help determine system parameters for DNA data storage. **a** A synthesis pool was generated with $N_{seq} = 10,000$ total number of sequences, with normally distributed copy numbers with a mean of $\bar{n}_{syn} = 10^8$ and standard deviation $\sigma = 3.2 \times 10^7$. The pool was simulated to store an average copy number $\bar{n}_0 = 100$, followed by 20 cycles of PCR amplification with $P = 0.95$, and high-throughput sequencing with average sequencing coverage $\bar{n}_r = 200$. Sequence dropout (i.e., coverage of 0 for a given sequence) rates were quantified. **b** Sequence dropout percentage as a function of variable synthesis pool c.v. and variable mean sequencing coverage \bar{n}_r . **c** Sequence dropout percentage as a function of variable mean copy number n_0 and variable mean sequencing coverage \bar{n}_r . **d** Percentage of recovered oligo and logical redundancy as a function of average oligo copy number. The mean sequencing coverage is 5. When a very low copy number of oligo is stored, more sequences drop out, and thus higher logical redundancy is needed. The reported dropout percentage was the average of 100 repeated simulations.

We then interrogated our computational model to determine whether it can estimate the bias observed in the serial dilution-PCR experiment. Despite not being able to observe the oligo population directly after synthesis, our UMI experiment (Fig. 2) has provided evidence that its population distribution is quite similar to the distribution resulting from a PCR process that starts from a large average copy count sample coming from that synthesized pool. As such, the copy distribution of a synthesis pool is modeled as a normal distribution with the same c.v. as the experimental data from the (optimized) ready-to-sequence pool. Then we used our system model to simulate the dilution-PCR experiment. Figure 5d shows that our model prediction is in good agreement with the c.v. of the experimental data ($R^2 = 0.71$). The model also predicted the trend of standard deviation of population fraction change Q : the lower starting copy number in the PCR showed higher standard deviation ($R^2 = 0.84$; Supplementary Fig. 4).

A computational model can help determine system parameters.

Taking it one step further, we used our computational model to study a range of parameters associated with DNA storage: synthesis bias, physical redundancy for storing DNA, logical redundancy, and sequencing redundancy (Fig. 6a). In particular, we investigated the impact of these parameters on sequence dropout rate, which is critical for error-free decoding. Figure 6b plots sequence dropout rates as a function of the c.v. of a synthesis pool and sequencing reads. It shows that a biased synthesis pool (i.e., high c.v.) is the dominant factor in sequence dropout and cannot be proportionally compensated by additional sequencing reads. Sequence dropout is caused by physical storage with a limited number of oligo copies coupled with PCR stochastic bias. Figure 6c plots sequence dropout rates as a function of the copy number of stored DNA and sequencing reads. It shows that physical storage density is a more important factor than sequencing reads in modulating sequence dropout. Interestingly, our model estimates that it is possible to store as few as 10 copies per oligo sequence (physical density of 9.3 EB per g - EB: exabytes; 10^{18} bytes), while achieving less

than 2% sequence dropout. This estimated physical density is over 10-fold higher than prior work by Erlich and Zielinski¹⁰ and is aligned with what we have recently observed in practice²⁹. The next important question is how much logical redundancy is needed to handle missing sequences. Take Reed-Solomon code in our previous work⁸ as an example, the maximum percentage of missing strands that can be tolerated is $\frac{R}{100+R}\%$, where R is the percentage of logical redundancy. Figure 6d shows a simulation of how oligo copy number affects the recovered oligo percentage (100% minus oligo dropout) and the required logical redundancy to recover the data. Interestingly, at the low end, a modest increase in logical redundancy allows for a significant decrease in the required oligo copy number and enables an almost proportional increase in physical density. For example, at 30 copies, the required logical redundancy for data recovery is 3% whereas at 10 copies the logical redundancy grows to only 8%, nearly tripling physical density. It is worth pointing out the example here ignores all other errors such as insertions, deletions, and substitutions. These errors depend on synthesis and sequencing technologies, and they should be taken into account when determining the proper logical redundancy. Finally, we give an example in Supplementary Figure 5 to show how our system model can be used to optimize two important parameters in a given DNA storage system: physical redundancy (determining physical density) and sequencing redundancy (determining sequencing cost).

Discussion

In this work, we quantified molecular bias in a DNA storage system, and we identified two significant bias sources: synthesis bias and PCR stochastic bias. Synthesis bias was found to be related to the spatial location on the synthesis chip, and this observation was later used to inform and improve the synthesis process. PCR stochastic bias was identified as the second main driver of oligo copy variation. Indeed, prior work also found that PCR copy data from a deeply diluted oligo pool resulted in dramatic bias, which is less suitable for data recovery¹⁰.

Another important contribution of this manuscript is the construction of the first process-wide model that provides a quantitative understanding of how oligo copy distribution is skewed as it goes through a DNA storage system. Importantly, such system model helps researchers rationally optimize the use of DNA physical density, logical redundancy, and sequencing redundancy for reliable data decoding without conducting hundreds of experimental trials. We believe this is an important step towards engineering robust, efficient DNA storage systems.

In this study, we found that oligos from unbiased synthesis and sequencing processes can be well modeled as a normal distribution and random sampling, respectively. While the experiments were tested using Twist Bioscience and Illumina sequencing, the proposed system model can in principle be applied to other synthesis and sequencing technologies. It is worth noting that when applying our model to other technologies, additional quantification and modeling is likely needed. For example, Ion Torrent and Oxford Nanopore sequencing show limited ability to accurately sequence long homopolymers, which is less significant in Illumina sequencing. Different array-based synthesis technologies could also have their own unique dependent bias caused by processes specific to them, such as uneven fluidic operation, surface treatment, and other factors. Our system model was experimentally tested by PCR-amplifying a single file without any other non-targeted files in a pool. This experiment was designed to avoid complexity from other files for proper quantification of the impact of PCR stochastic bias. Next, we plan to investigate whether PCR random access of a file from a complex pool with additional files will lead to more bias. We suspect that amplifying a very small file from a complex pool with relatively large number of sequences will exhibit more copy number variation due to non-specific binding of primers. New methods will probably be needed for such system.

Methods

Reagents. All DNA pools were synthesized by Twist Bioscience (San Francisco, CA). All DNA pools were resuspended to 10 ng per μL in 1 \times TE buffer (pH 7.5). All primers were purchased as desalted, unpurified DNA from Integrated DNA Technologies (IDT; Coralville, IA). All primers were resuspended to 100 μM in 1 \times TE buffer (pH 7.5). KAPA HIFI polymerase was purchased from Kapa Biosystems. T4 ligase and T4 Polynucleotide Kinase (T4 PNK) were purchased from New England Lab.

PCR protocol. In a 20 μL PCR reaction, 1 μL of 1 ng per μL of ssDNA pool was mixed 1 μL of 10 μM of the forward primer and 1 μL of 10 μM of the reverse primer, 10 μL of 2 \times KAPA HIFI enzyme mix, and 7 μL of molecular biograde water. The reaction followed a thermal protocol: (1) 95 $^{\circ}\text{C}$ for 3 min, (2) 98 $^{\circ}\text{C}$ for 20 s, (3) 62 $^{\circ}\text{C}$ for 20 s, (4) 72 $^{\circ}\text{C}$ for 15 s. After PCR, the length of the PCR products was confirmed using a Qiaxcel fragment analyzer, and the sample concentration was measured using a Qubit 3.0 fluorometer. Primer sequences see Supplementary Table 2.

Sample preparation for sequencing. Before sequencing, the concentrations of all samples were quantified using qPCR. The final sample was then prepared for sequencing by following the NextSeq System Denature and Dilute Libraries Guide. The final concentration of the loaded sample for our Illumina NextSeq is 1.3 pM, and a 10–20% PhiX was spiked in as a control (PhiX is a genomic DNA sample provided by Illumina).

Protocols of UMI labeling. The general workflow for UMI labeling of a single-stranded DNA pool is divided into 5 steps (Supplementary Fig. 1; sequences see Supplementary Table 1): (1) phosphorylation of a ssDNA pool and Illumina P7 adapters, (2) assembly of a ssDNA pool with Illumina adapters with DNA staples by heat annealing, (3) ligation of Illumina adapters to the ssDNA pool, (4) extraction of the ligated sample using denaturing polyacrylamide gel electrophoresis (D-PAGE), and finally (5) PCR enrichment of the full length product.

The phosphorylation of ssDNA was performed using the following recipe: 5 pmole of the single-stranded DNA pool, 20 units of T4 Polynucleotide Kinase (T4 PNK), 1 μL of 10 \times T4 ligase buffer and 1 μL of 10 \times T4 PNK buffer were mixed in a 10 μL total volume reaction. 500 pmole of single-stranded Illumina P7 adapter, 200 units of T4 PNK, 5 μL of 10 \times T4 ligase buffer and 5 μL of 10 \times T4 PNK buffer were mixed in a 50 μL total volume reaction. The mixtures were incubated for 30 min at 37 $^{\circ}\text{C}$.

The assembly of the single-stranded DNA pool with adapters were performed with the following recipe: In a 25 μL reaction, 15 pmole of single-stranded DNA pool, 30 pmole of DNA staples and 45 pmole of Illumina P5 and P7 adapters were mixed. The mixture was heated up to 95 $^{\circ}\text{C}$ for 2 min, and then cooled down to 25 $^{\circ}\text{C}$ at a rate of 1 degree per minute.

Ligation of DNA was performed with a 15 μL reaction in which 10 μL of the assembled DNA mixture, 2 μL of the T4 ligase (10 units per μL), 1.5 μL of T4 ligase buffer and 1.5 μL of molecular water were mixed. The ligation mixture was incubated at room temperature for 30 min, followed by heat inactivation at 65 $^{\circ}\text{C}$ for 10 min.

A 10% D-PAGE gel was made by mixing 2.5 mL of 19:1 40% acrylamide/bus, 1.2 mL of 10 \times TBE, 5.04 g of urea and deionized water to 12 mL. Then 72 μL APS and 4.8 μL of TEMED were added to help polymerization. DNA sample was mixed with 2 \times TBE/Urea denaturing loading buffer (Bio-Rad). Gels were run at 200 V for 55 min at 55 $^{\circ}\text{C}$. The extracted band was incubated with 1 \times TE buffer overnight at room temperature for elution.

The eluted single-stranded DNA was PCR-amplified using the end primers of Illumina adapters. The PCR reaction used 1 μL of the eluted single-stranded DNA, 10 pmole of the forward and reverse primers, 10 μL of 2 \times KAPA HIFI polymerase and 8 μL of molecular water. The thermal protocol is as follows: (1) 95 $^{\circ}\text{C}$ for 3 min, (2) 98 $^{\circ}\text{C}$ for 20 s, (3) 60 $^{\circ}\text{C}$ for 20 s, (4) 72 $^{\circ}\text{C}$ for 15 s.

Sequence alignment using Burrows-Wheeler Aligner (BWA). We used BWA to align our expected, short references against reads from a sequencer. We then used the alignment counts for each reference sequence produced by BWA to generate distribution plots.

Density histogram plots. The y-axis of a density histogram shows probability density, and the area (or integral) under the histogram is 1. The probability density d_i is calculated by dividing the count by the sample size times its bin width (see the following equation).

$$d_i = \frac{N_i}{\left(\sum_j N_j\right) * W_i} \quad (5)$$

where N_i is the count of the i -th bar, and W_i is the bin width of the i -th bar. Displaying the y-axis as probability density makes it possible to compare distributions. In Fig. 5c, a Gaussian estimated curve is added to help visualize each histogram.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The sequencing coverage data underlying Figs. 2c–f, 3b, d, 4a, b, and 5 are available in the Source Data file and are on GitHub at this URL: <https://github.com/uwmisl/storage-biasing-ncmms20>. Any additional data will be made available upon reasonable request. Source data are provided with this paper.

Code availability

The analysis code that supports the findings of this study is available upon request/is on GitHub at this URL: <https://github.com/uwmisl/storage-biasing-ncmms20>. The code simulates the whole molecular process and was used to demonstrate how synthesis bias, physical redundancy, and sequencing redundancy affect the sequence dropout rate.

Received: 8 October 2019; Accepted: 19 May 2020;

Published online: 29 June 2020

References

- Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15**, 366–370 (2016).
- Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nat. Rev. Genet.* **20**, 456–466 (2019).
- Cox, J. P. Long-term data storage in DNA. *Trends Biotechnol.* **19**, 247–250 (2001).
- Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
- Church, G. M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* **337**, 1628–1628 (2012).
- Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
- Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem. Int. Ed.* **54**, 2552–2555 (2015).

8. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **36**, 242–248 (2018).
9. Yazdi, S. M. H. T., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A rewritable, random-access DNA-based storage system. *Sci. Rep.* **5**, 14138 (2015).
10. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
11. Bornholt, J. et al. A DNA-based archival storage system. *ACM SIGOPS Oper. Syst. Rev.* **50**, 637–649 (2016).
12. Yazdi, S. M. H. T., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Sci. Rep.* **7**, 5011 (2017).
13. Heckel, R., Mikutis, G. & Grass, R. N. A characterization of the DNA data storage channel. *Sci. Rep.* **9**, 9663 (2019).
14. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
15. Ross, M. G. et al. Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
16. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52**, 87–94 (2012).
17. Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
18. Jagers, P. & Klebaner, F. Random variation and concentration effects in PCR. *J. Theor. Biol.* **224**, 299–304 (2003).
19. Stolovitzky, G. & Cecchi, G. Efficiency of DNA replication in the polymerase chain reaction. *Proc. Natl Acad. Sci. USA* **93**, 12947–12952 (1996).
20. Hassibi, A., Kakavand, H. & Lee, T. A stochastic model and simulation algorithm for polymerase chain reaction (PCR) systems. In *Proc. of IEEE Workshop on Genomics Signal Processing and Statistics* (IEEE, 2004).
21. Piau, D. Confidence intervals for nonhomogeneous branching processes and polymerase chain reactions. *Ann. Probab.* **33**, 674–702 (2005).
22. Lalam, N., Jacob, C. & Jagers, P. Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency. *Adv. Appl. Probab.* **36**, 602–615 (2004).
23. Peccoud, J. & Jacob, C. Theoretical uncertainty of measurements using quantitative polymerase chain reaction. *Biophys. J.* **71**, 101–108 (1996).
24. Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* **43**, e143 (2015).
25. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
26. Quail, M. A. et al. Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* **9**, 10–11 (2012).
27. Chen, Y., Liu, T., Yu, C., Chiang, T. & Hwang, C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE* **8**, e62856 (2013).
28. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40**, e72 (2012).
29. Organick, L. et al. Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.* **11**, 1–7 (2020).

Acknowledgements

We would like to thank Patrick Finn, Siyuan Chen, Andrew Stewart, Bernadette Arias, and Emily Leproust from Twist Bioscience for supplying us with the DNA and for suggestions on the analysis of DNA synthesis data. We also thank Leila Zelnick and Jeff Nivala for discussion and comments on the paper.

Author contributions

Y.-J.C. designed, performed, and analyzed experiments. C.T. and C.B. designed experiments and analyzed data. L.O. performed experiments. S.D.A. ran the DNA sequence encoder and decoder, and analyzed data. P.W. and B.P. contributed improvements to the DNA synthesis process. G.S., K.S., and L.C. directed and supervised the work.

Competing interests

Y.-J.C., S.D.A., and K.S. are or were Microsoft employees while this project was conducted. P.W. and B.P. are Twist Bioscience employees. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-16958-3>.

Correspondence and requests for materials should be addressed to Y.-J.C., L.C. or K.S.

Peer review information *Nature Communications* thanks Stephen Yip and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020