Oligo replication advantage driven by GC content and Gibbs free energy

Hongyan Qiao · Yanmin Gao · Qian Liu · Yanan Wei · Jiaojiao Li · Zhaoguan Wang · Hao Qi

Abstract Large scale DNA oligo pools are emerging as a novel material in a variety of advanced applications. However, GC content and length cause significant bias in amplification of oligos. We systematically explored the amplification of one oligo pool comprising of over ten thousand distinct strands with moderate GC content in the range of 35–65%. Uniquality amplification of oligos result to the increased Gini index of the oligo distribution while a few oligos greatly increased their proportion after 60 cycles of PCR. However, the significantly enriched oligos all have relatively high GC content. Further thermodynamic analysis demonstrated that a high value of both GC content and Gibbs free energy could improve the replication of specific oligos during biased amplification. Therefore, this double-G (GC content and Gibbs free energy) driven replication advantage can be used as a guiding principle for the sequence design for a variety of applications, particularly for data storage.

Keywords Oligo pool · DNA storage · DNA amplification · DNA free-energy

Introduction DNA pools from array synthesis comprising thousands or even millions of distinct oligos have showed
great potential for various advanced bioengineering applications, such as CRISPR-based gene editing, probe blending, DNA origami, and DNA data storage (Kosuri and Church 2014). Especially for emerging DNA-based data storage, the oligo pool size is several orders of magnitude larger than for other research applications (Lipshutz et al. 1999; Kosuri et al. 2010; Bonde et al. 2015). However, the sheer copy number of oligos synthesized on a DNA chip is relatively small, roughly from $10^5$ to $10^{12}$ copies of each oligo, which is highly dependent on the synthesis platform (Tian et al. 2004; Klein et al. 2016). For example, hundreds of nanograms of DNA in micromolar concentrations are required to meet the demands of high-quality DNA sequencing on a commercial DNA sequencing platform, such as Illumina (Linnarsson 2010), in order to retrieve the complete information. Therefore, controllable and unbiased replication of the oligo pool is crucial for many applications, particularly for DNA data storage (Goldman et al. 2013; Organick et al. 2018).

Since its invention in the 1980s, the polymerase chain reaction (PCR) has become the most popular method for amplifying low-abundance DNA (Mullis et al. 1986). However, various PCR-based techniques are prone to generate biases during the parallel amplification of multiple DNA sequences due to its intrinsic mechanism, such as inefficient priming and product-as-template (Polz and Cavanaugh 1998). For a large-scale oligo pool for DNA data storage with enormous sequence complexity, amplification bias can greatly alter the copy numbers of individual oligos, potentially leading to data losses (Chen et al. 2020; Erlich and Zielinski 2017). In addition, researchers usually amplify the minor oligo strands via multiple thermal amplification cycles (deep replication, ≥ 40 cycles) in order to further increase the physical density and reduce the cost of DNA data storage (Erlich and Zielinski 2017). The oligo pool is therefore further skewed by a highly biased PCR with 40 or 60 cycles, and more sequencing reads are needed to recover the complete information, thus increasing decoding calculation and the cost of sequencing.

Amplification bias is currently considered to mainly be the result of two major groups of factors. The first is intrinsic to the DNA sequence, including GC content, secondary structures and the length of the DNA sequence (Aird et al. 2010; Benjamini and Speed 2012), while the other is related to the preferences of the used DNA polymerase (Dabney and Meyer 2012; Pan et al. 2014). The PCR amplification of DNA sequences with extremely low (< 11%) or high GC content (> 56%) is hindered by the formation of hairpins and various other secondary intramolecular structures, resulting in incomplete or non-specific products (Sahdev et al. 2007; Aird et al. 2011). A number of studies have attempted to minimize PCR amplification bias and improve the PCR yield of GC-rich templates by optimizing the thermocycling conditions (Don et al. 1991; Frey et al. 2008; Hecker and Roux 1996; Percze and Meszaros 2020) or adding molecular enhancers (Henke et al. 1997; Baskaran et al. 1996) such as betaine, dimethyl sulfoxide (DMSO) (Mihovilovic and Lee 1898) and bovine serum albumin (BSA) (Farell and Alexandre 2012). This body of research indicates that GC content is the crucial factor most significantly influencing the performance of molecular replication.

Here, we investigated an oligo pool comprising 11,520 distinct strands with a length of 201 nucleotides (nt). A moderate GC content in the range of 35 to 65% was designed for all oligos. The oligo pool was amplified using 10, 30 and 60 PCR cycles to explore the intrinsic features influencing the amplification efficiency and bias. In agreement with previous reports, deep statistical analysis revealed that the oligo distribution greatly shifted with increasing numbers of replication cycles. However, contrary to the current theory, we found that oligos with relatively high GC content (> 50%) exhibited an advantage in replication. After 60 cycles of PCR amplification, oligos with GC content > 50% were enriched up to about 14.7-fold, while oligos with a GC content < 44% were depleted to about one-fourteenth of the initial relative abundance. Thermodynamic analysis revealed that intrinsic features of both GC content and Gibbs free energy contribute to the biased replication of the oligo pool. We provide a new principle for the balanced sequence design of large numbers of distinct oligos that can help avoid amplification bias.

Materials and methods

Materials

Q5® High-Fidelity DNA Polymerase, Q5® Hot Start High-Fidelity DNA Polymerase and Taq DNA
Polymerase were purchased from New England Biolabs (NEB). Eastep Gel and PCR Cleanup Kit was purchased from Promega. Oligo pool was synthesized by Twist Bioscience. All primers we used were synthesized by GENEWIZ and in Table S1.

DNA master oligo pool

The synthesized oligo pool was resuspended in 1×TE buffer for final concentration of 2 ng/μL. The DNA master pool 1 containing 11,520 DNA strands used in this study was previously prepared in our laboratory. The details of preparation the DNA master pool were as follows: PCR was performed using 10 ng ssDNA pool, 4 μM of the forward primer F1/F2 and 4 μM of the reverse primer R1/R2, 10 μL 5×Q5 Reaction Buffer, 0.2 mM dNTPs, 0.5 μL Q5 High-Fidelity DNA Polymerase in a 50 μL reaction. Thermocycling conditions were as follows: 5 min at 98 °C; 10 cycles of: 30 s at 98 °C, 30 s at 54 °C, 30 s at 72 °C, followed by a 5 min extension at 72 °C. The reaction was then purified and eluted in 50 μL DNase/RNase-free water according to the instructions in the Eastep Gel and PCR Cleanup Kit. This library was considered the master pool for deep replication.

Deep replication reaction in PCR amplification

Deep replication was performed using 0.5 μL Q5 High-Fidelity DNA Polymerase/Q5 Hot Start High-Fidelity DNA Polymerase, 10 ng DNA master pool, 4 μM of forward primer F1-1/F2; 4 μM of reverse primer R1/R2, 10 μL 5×Q5 reaction buffer in a 50 μL reaction. Thermocycling conditions were as follows: 5 min at 98 °C; 10 cycles/30 cycles of: 30 s at 98 °C, 30 s at 54 °C, 30 s at 72 °C, followed by a 5 min extension at 72 °C. The product was then purified and eluted in 50 μL DNase/RNase-free water according to the instructions in the Eastep Gel and PCR Cleanup Kit. The details of qPCR were as follows: 1 ng of each strand (total 10 ng), 0.4 μM of the forward primer F1-1 and 0.4 μM of the reverse primer R1, 0.2 μM TaqMan probe T1 (T1-1) and 0.2 μM TaqMan probe T2, 5 μL 10×Taq Reaction Buffer, 2 μL Taq DNA Polymerase in a 50 μL reaction. The mixtures were incubated in a QuantStudio 6 qPCR System (Thermo Fisher Scientific) as follows procedure: 5 min at 95 °C; 40 cycles of: 30 s at 95 °C, 30 s at 54 °C, 20 s at 68 °C, with fluorescence measurements being taken at each cycle.

Preparation of the standard curve of the TaqMan probe: 1 pg, 10 pg, 100 pg, 1 ng and 10 ng of top 1/bottom 1, 0.4 μM of the forward primer F1-1 (Top1-F/Bottom1-F) and 0.4 μM of the reverse primer R1 (Top1-R/Bottom1-R), 0.2 μM TaqMan probe T1 (T1-1) and 0.2 μM TaqMan probe T2, 5 μL 10×Taq Reaction Buffer, 0.2 mM dNTPs, 2 μL Taq DNA Polymerase in a 50 μL reaction. The mixtures were incubated in a QuantStudio 6 qPCR System (Thermo Fisher Scientific) as follows procedure: 5 min at 95 °C; 40 cycles of: 30 s at 95 °C, 30 s at 54 °C, 20 s at 68 °C, with fluorescence measurements being taken at each cycle.
qPCR for the product of deep replication

The DNA mixtures of ten oligos at different concentration were amplified using 10, 30 and 60 PCR cycles. 0.25 ng, 0.45 ng, 0.5 ng, 0.6 ng, 0.7 ng of 1–5 and 2.5 ng, 2 ng, 1.5 ng, 1 ng, 0.5 ng of 6–10 DNA strand (total 10 ng, top 1:bottom 1 = 1:10), or 1 ng of each strand (total 10 ng, top 1:bottom 1 = 1:1), or 2.5 ng, 2 ng, 1.5 ng, 1 ng, 0.5 ng of 1–5 and 0.25 ng, 0.45 ng, 0.5 ng, 0.6 ng, 0.7 ng of 6–10 DNA strand (total 10 ng, top 1:bottom 1 = 10:1), 4 μM of forward primer F1-1, 4 μM of reverse primer R1, 10 μL 5 × Q5 reaction buffer in a 50 μL reaction. Thermocycling conditions were the same as deep replication reaction above mentioned. The 10, 30 and 60 cycles PCR product was purified by Eastep Gel and PCR Cleanup Kit and eluted in 50 μL DNase/RNase-free water. Then 10 ng of DNA out of PCR products were subjected to qPCR analysis, 0.4 μM of the forward primer Top1-F/Bottom1-F and 0.4 μM of the reverse primer Top1-R/Bottom1-R, 0.2 μM TaqMan probe T1-1/T2, 5 μL 10× Taq Reaction Buffer, 0.2 mM dNTPs, 2 μL Taq DNA Polymerase in a 50 μL reaction. The mixtures were incubated in a QuantStudio 6 and the procedures were the same as above mentioned.

Sequencing on an Illumina Hiseq 4000 platform

Sample collection and preparation

DNA degradation and contamination were monitored on 2% agarose gels. DNA purity was checked using the NanoPhotometer spectrophotometer (IMPLEN, CA, USA). DNA concentration was measured using Qubit DNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, CA, USA).

Library preparation for sequencing

A total amount of 700 ng DNA per sample was used as input material for the DNA sample preparations. Sequencing libraries were generated using NEB Next Ultra DNA Library Prep Kit for Illumina (NEB, USA) following manufacturer’s recommendations and index codes were added to attribute sequences to each sample. Briefly, the Chip DNA was purified using AMPure XP system (Beckman Coulter, Beverly, USA). After adenylation of 3’ ends of DNA fragments, the NEB Next Adaptor with hairpin loop structure were ligated to prepare for hybridization. Then electrophoresis was used to select DNA fragments specified in length. 3 μL USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated DNA at 37 °C for 15 min. At last, the products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system.

Clustering and sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using HiSeq 4000 PE Cluster Kit (Illumina) according to the manufacturer’s instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq 4000 platform and 150 bp paired-end reads were generated.

The bioinformatic statistical analysis

We combined the sequenced read pairs using PEAR to obtain the complete sequenced reads. The sequenced reads were aligned to the actual sequences using BLAST. The coverage and number per million reads were obtained using the Valid_Coverage_Number.pl, and the frequency per coverage was calculated by dividing the sum of these numbers (Fig. 2A). The Gini coefficient was calculated using R (Fig. 2B). The GC content was obtained using the get_GC.pl (Fig. 2C). The read number of each oligo was obtained using the Valid_Coverage_Number.pl, after which the depth per oligo was calculated by dividing the reads number of each oligo by the total reads number to display the distribution of depth as shown in Fig. 4A. The increment was calculated by dividing the depth in the PCR reaction with 60 cycles or 30 cycles by the corresponding depth in the PCR reaction with 10 cycles (Fig. 3A). The coverage of aligned sequences was sorted from small to large and numbered in sequence. The serial numbers were then used to select the top 1% (115 oligos) and bottom 1% (115 oligos) from the PCR reactions with 10, 30 and 60 cycles were selected and the average GC content of these sequences was calculated (Fig. 3B). The increment was sorted from small to large and numbered in sequence. The serial numbers were then used to select the top 1% (115 oligos) and bottom 1% (115 oligos) increments for 30 cycles/10 cycles and 60
cycles/10 cycles, after which the average increment of these sequences was calculated (Fig. 3C). The secondary structures and Gibbs free energy values of oligonucleotides were calculated using NUPACK (http://www.nupack.org) (Fig. 4B). The k-mers of the sequences were analyzed using the kmer.pl (Fig. S1). Based on these data, the distribution of GC content, Gibbs free energy and increment could be plotted as shown in Fig. 5.

Results

Oligo pool for deep replication

To assess the amplification bias, the DNA master pool 1 containing 11,520 DNA strands previously prepared in our laboratory was used in this study (Gao et al. 2020). To reduce or eliminate the impact of different primers on the priming efficiency, all 11,520 oligos used the same forward and reverse primers. The length of the oligos was 201 nt and the payload region with 167 nt had a GC content ranging from 35 to 65% (Fig. 1A). As reported, the original unevenness of the chip-synthesized oligo pool would be further enlarged through PCR with dozens of cycles, generating a skewed coverage distribution with a long tail that consists of high-copy-number oligos (Fig. 1B). We aimed to analyze the main characteristics of easily depleted oligos (low-copy-number) as well as the enriched oligos (high-copy-number), including the GC content, Gibbs free energy, and secondary structures. In addition, the oligo pool 2 containing 11,520 DNA oligos was designed using the same principles to rule out certain chance factors. The length of oligos was 200 nt and the payload region with 155 nt also had a GC content ranging from 35 to 65%.

Biased amplification of the oligo pool

First, 10 ng of oligos from the master pool was amplified using PCR with 10, 30 and 60 thermal cycles, respectively. The PCR products were sequenced on a commercial Illumina HiSeq 4000 platform with 150 paired-end reads. The coverage distributions of three samples were obtained via a set of statistical analyses developed based on the bioinformatic BLAST program (Fig. 2A). In addition, sequencing analysis revealed various errors, including base substitution and indel (Fig. S1A). With increasing number of
PCR cycles, the coverage distribution was increasingly skewed towards a high “head” and a long “tail”, which indicated that certain DNA sequences (located in the “tail” of the distribution) were preferred over others (situated in the “head” region) during the PCR amplification. It is worth noting that the same oligo showed the highest copy number, irrespective of the number of PCR cycles. Moreover, the relative abundance of this oligo increased from 0.21% after 10 cycles to 2.5% after 60 cycles, which indicated that PCR had a significant preference for the oligo. The inequality of the oligo pool was illustrated using the Gini coefficient, a measure of income inequality (Fig. 2B). The Gini index rose from 0.495 (10 cycles) to 0.579 (30 cycles) and 0.734 (60 cycles), which demonstrated that the oligo pool was increasingly skewed after deep replication.

For sake of convenience when analyzing the characteristics of DNA sequences located in “head” and “tail” of the statistical distribution, we firstly analyzed the GC content of the oligo pool, as shown in the histogram of the GC% distribution for the 11,520 oligos in Fig. 2C. The distribution of the GC content is a normal distribution and oligos with GC content of 43.84 to 53.42% accounted for 81.32% of the oligo pool, whereby a majority of oligos were concentrated around 48.63%.

The GC content is one of the most important factors leading to amplification bias in the process of PCR. To validate whether the amplification bias was
related to the GC content of DNA sequences, the increase in the relative abundance of each oligo was analyzed after 10, 30 and 60 PCR cycles. The frequency ratios for each amplicon produced by PCR with 30/60 cycles and the corresponding amplicon produced by PCR with 10 cycles were defined as the increment (Fig. 3A). The results showed that the increment rose as the GC content increased, and the increment distribution of 60 cycles/10 cycles was more skewed than 30 cycles/10 cycles, which illustrated that with the increase in the number of PCR cycles, the amplification bias also increased. Compared to 10 cycles, the DNA sequences preferred by PCR were enlarged up to 14.65 times, and the DNA sequences without amplification advantage were depleted to one-fourteenth.

Therefore, we further analyzed the GC content of DNA sequences in the top 1% of high coverage and the bottom 1% of low coverage in the samples of PCR reactions with 10, 30, and 60 cycles (Fig. 3B and Fig. S1B). It was obvious that the GC content of the top 1% (on average, 55%) was higher than that of the bottom 1% (on average, 45%). And the average GC content of top 1% slightly increased with the increasing number of PCR cycles. This further corroborated the finding that the sequences with high GC content were more easily amplified than those with relatively low GC content during the deep replication of the oligo pool. At the same time, the average increment of the top and bottom 1% for 30 cycles/10 cycles and 60 cycles/10 cycles was also analyzed. The increment value of the top 1% for 60 cycles/10 cycles (3.89) was higher than for 30 cycles/10 cycles (2.40). Conversely, the increment value of the bottom 1% for 60 cycles/10 cycles (0.19) was lower than for 30 cycles/10 cycles (0.35) (Fig. 3C). The mean increment values were 0.88 for 30 cycles/10 cycles and 0.62 for 60 cycles/10 cycles. These results reflected that the competitive advantage of DNA sequences with relatively high GC content increased with the increase of deep replication, leading to the decrease in the relative abundance or even extinction of certain DNA sequences with low GC content. Contrary to existing theory (Aird et al. 2011; Dabney and Meyer 2012), we found that the DNA sequences with relatively high GC content from 12 to 65% as well as a complex secondary structure exhibited a competitive advantage over sequences with low GC content in the process of deep replication, greatly increasing their relative abundance.
Double-G driven replication advantage

Next, we conducted a concrete analysis of the GC content of each amplicon produced by PCR with 60 cycles, and the frequency of each oligo was plotted against its GC content (Fig. 4A). Overall, the frequencies (copy number) of oligos with >50% GC was higher than the frequency of those with <50% GC. Additionally, the secondary structures of five oligos with the highest frequency and five oligos with the lowest frequency (the blue dotted box in A), C The percent of each oligo in the mini pool of 10 oligos after ten serial PCR amplifications were analyzed using NUPACK (Fig. 4B), a web version of the software for the design and analysis of nucleic acid structures. As shown in Fig. 4B, the GC content of the five most abundant (highly enriched) DNA sequences was almost 60%, significantly higher than the 40% value of the five least amplified DNA sequences. Compared with the low-copy-number DNA sequences, the secondary structures of the high-copy-number sequences were more complex. Moreover, the ten oligos were

Fig. 4 Features of the depleted and enriched DNA sequences. A The frequency (depth) distribution of each oligo after deep replication with 60 cycles plotted against its GC content. B The secondary structure and GC content of five DNA sequences with the highest frequency and five DNA sequences with the lowest frequency (the blue dotted box in A). C The percent of each oligo in the mini pool of 10 oligos after ten serial PCR amplifications

Fig. 5 The distribution of Gibbs free energy and increment for 60 cycles/10 cycles with corresponding GC content. The increment of an oligo over 1 indicates that the DNA sequence is prone to be enriched (amplified) by PCR. DNA sequences with increment less than 1, in turn, are most likely to be depleted or even go extinct during deep replication
individually synthesized, purified, and mixed at different concentration. amplified, and subjected to qPCR analysis (Fig. S2A). And we designed two TaqMan probes for the oligo with the maximum frequency (highly enriched) and the oligo with the lowest frequency (Fig. S2B).

First, the standard curves and regression equations of the two DNA sequences with the highest and lowest increment were established by qPCR (Fig. S2C, D). As you suggested, the DNA pool, containing ten oligos at different concentration, was subjected to qPCR analysis. Although the Ct value was a little different, the copy number of the two oligos was consistent when the initial concentration of them was the same (Fig. S2E, F). This is because qPCR is a technology used for measuring the initial copy number of DNA molecules in the sample using the standard curve method. Meanwhile, we also noted that the copy number of the oligo with low GC content was more than that with high GC content at the early stage of PCR (6.58 cycles), which suggested that the oligo with low GC content is more easily enriched at the abundant reaction components. However, PCR components, such as primers and dNTPs, will be exhausted with increasing PCR cycles, amplification bias is currently supposed to occur in last stage of the deep replication for the large-scale oligo pool (Fig. S3A). Here, the amplification of the oligos with high GC content and low GC content at the last stage of PCR was mainly investigated by qPCR. Detailedly, the DNA mixtures of ten oligos at different concentration were amplified using 10, 30 and 60 PCR cycles. Then, the amplicons were purified. And ten nanogram of DNA out of PCR products were subjected to qPCR analysis and the two oligos with the highest (top 1) and lowest increment (bottom 1) were analyzed through inner primer, which was specially designed for them (Fig. S2B). We discovered whether the initial copy number ratio of the top 1 and bottom 1 was 1:10, 1:1 or 10:1, the copy number of the top 1 increased with increasing PCR cycles, and the copy number of the bottom 1 decreased with increasing PCR cycles (Fig. S3B–D). Just as we expected, there was a golden point that the copy number of the top 1 were more than bottom 1 when the initial copy number ratio of both was 1:10. In addition, on development trend of copy number, we supposed that there might be a golden point when the initial copy number ratio of both was 1:1, which revealed that the oligo with low GC content is more easily enriched at the abundant reaction components, but the oligo with high GC content is more advantages over that with low GC content. Meanwhile, to further verify the amplification bias, these 10 oligos were mixed at an equal molar ratio and amplified through nine serial PCR amplifications. The amplicons were sequenced in manner of PCR-free and we found that the percent of top five oligos were higher than that of bottom five oligos and the percent of three of top five oligos exceeded 10% which is the initial percent of each oligo, proving once again that the sequences with high GC content were more easily amplified than those with relatively low GC content during the deep replication of the oligo pool (Fig. 4C). In addition, we further analyzed the 6-mer distribution of the top and bottom 1% sequences after 60 cycles (Fig. S4), but there was no obvious difference. This indicated that PCR amplification bias is not relevant to the partial features of the DNA sequence.

Finally, the Gibbs free energy of oligos and their increment distribution for 60 cycles/10 cycles was plotted against the corresponding GC content (Fig. 5). Notably, the increment for 60 cycles/10 cycles increased with the increase of the absolute value of the Gibbs free energy as well as the GC content of the DNA sequences. These results showed that the Gibbs free energy and GC content (double-G) of the DNA sequences had an influence on the increment. Meanwhile, we further plotted the Gibbs free energy against the increment and analyzed the Gibbs free energy of 10, 50, and 100 DNA sequences with the highest and lowest increment (Fig. S5). There was a trend that the DNA sequence with high increment had a high absolute value of the average Gibbs free energy. However, the higher absolute value of the Gibbs free energy was not always the high increment, and vice versa. Therefore, we supposed that the increased increment was not only caused by single factor, but by multi-factor, such as GC content and Gibbs free energy. It’s not a simple linear relationship between the Gibbs free energy and the increment. Therefore, the amplification preference is largely driven by the GC content and Gibbs free energy of the DNA sequences, and oligos with relatively high double-G values are more likely to replicate. We speculated that the reason for this phenomenon may
be the three hydrogen bonds between GC, which makes the structures more stable and results in rapid polymerization.

Conclusions

In summary, the analysis of PCR amplifications with 10, 30, and 60 cycles revealed bias in the replication of the oligo pool, whereby oligos with relatively high GC content (50–65%) were more prone to increase their abundance than those with 35–45% GC. Notably, this is not in agreement with the current theory, which stipulates that the replication efficiency of sequences with a high GC content is low. In the extreme case, oligos with 60% GC were enriched up to 14.65-fold, while oligos with 41% GC were reduced in their abundance to almost one-fourteenth as the worst case. Moreover, Taq DNA polymerase and Q5 Hot Start High-fidelity DNA polymerase were also used for deep replication and similar results were obtained. Systematical thermodynamic analysis revealed a double-G (GC content and Gibbs free energy) driven replication advantage, which to the greatest extent explained the biased replication of a large number oligos. As a fundamental principle, it could assist the construction of oligo pools comprising large numbers of distinct strands with more predictable molecular behavior to improve applications such as DNA data storage, CRISPR gene editing, and DNA self-assembly.

Supporting information Supplemental Information includes Transparent Methods, five figures and one table.

Supplementary Figure S1—The error distribution and amplification bias in deep replication, related to Figure 2A and Figure 3B.

Supplementary Figure S2—Features of the ten DNA sequences (five high GC content and five with low GC content, related to Figure 4).

Supplementary Figure S3—The amplification bias in the deep replication, related to Figure 4.

Supplementary Figure S4—Count distribution with category of 6-mer, related to Figure 4.

Supplementary Figure S5—The distribution of increment with Gibbs free energy, related to Figure 5.

Supplementary Table S1—Sequence information of primers, related to Figure S2B and transparent methods.

Author contributions HQ, YG, QL and YW: contributed equally to this work. HQ and YG: did the main experiments, statistical analysis and prepared the original draft. QL: performed most of the bioinformatics statistical analysis programs, analyzed the Gibbs free energy based on NUPACK and revised this manuscript. YW: did a part of the experiments and were involved in some statistical analyses. JL and ZW: assisted with a part of the experiments. HQ: designed the experiments, analyzed the data, prepared the manuscript and supervised the study.

Funding This work was supported by the National Key R&D Program of China (Grant No. 2020YFA0712104) and partly by the National Key R&D Program of China (Grant No. 2021YFF1200102).

Data and code availability This study did not generate new code. All the data supporting the conclusions of this article are included within the article and its supplemental information. Furthermore, the original sequencing FASTQ file and the designed sequence file can be obtained via: (http://pan.tju.edu.cn:80/link/06C85DE43C821C0B00413DBDD9575CD, code: oFzx).

Material availability All unique reagents generated in this study are available from the lead contact without restriction.

Declarations

Conflict of interest The authors declare no competing interests.

References

Aird D, Chen WS, Ross M, Connolly K et al (2010) Analyzing and minimizing bias in Illumina sequencing libraries. Genome Biol 11(1):P3. https://doi.org/10.1186/gb-2010-11-S1-P3

Aird D, Ross MG, Chen W-S, Danielsson M et al (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12(2):1–14. https://doi.org/10.1186/gb-2011-12-2-r18

Baskaran N, Kandpal RP, Bhargava AK, Glynn MW, Bale A, Weissman SM (1996) Uniform amplification of a mixture of deoxyribonucleic acids with varying GC content. Genome Res 6(7):633–638. https://doi.org/10.1101/gr.6.7.633

Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40(10):e72. https://doi.org/10.1093/nar/gks001

Bonde MT, Kosuri S, Genee HJ, Sarup-Lytzen K, Church GM, Sommer MO (2015) Direct mutagenesis of thousands of genomic targets using microarray-derived oligonucleotides. ACS Synth Biol 4(1):17–22. https://doi.org/10.1021/sb5001565

Chen YJ, Takahashi CN, Organick L, Bee C, Ang SD, Weiss P, Peck B, Seeleg G, Ceze L, Strauss K (2020) Quantifying molecular bias in DNA data storage. Nat Commun 11(1):3264. https://doi.org/10.1038/s41467-020-16958-3

Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA.
sequencing libraries. Biotechniques 52(2):87–94. https://doi.org/10.2144/000113809

Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS (1991) Touchdown PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res 19(14):4008–4008. https://doi.org/10.1093/nar/19.14.4008

Erlich Y, Zielinski D (2017) DNA Fountain enables a robust and efficient storage architecture. Science 355(6328):950–954. https://doi.org/10.1126/science.aaj2038

Farell EM, Alexandre G (2012) Bovine serum albumin further enhances the effects of organic solvents on increased yield of polymerase chain reaction of GC-rich templates. BMC Res Notes 5:257. https://doi.org/10.1186/1756-0500-5-257

Frey UH, Bachmann HS, Peters J, Siffert W (2008) PCR-amplification of GC-rich regions: ‘slowdown PCR.’ Nat Protoc 3(8):1312–1317. https://doi.org/10.1038/nprot.2008.112

Gao Y, Chen X, Qiao H, Ke Y, Qi H (2020) Low-bias manipulation of DNA oligo pool for robust data storage. ACS Synth Biol. https://doi.org/10.1021/acssynbio.0c00419

Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, Birney E (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494(7435):77–80. https://doi.org/10.1038/nature11875

Hecker KH, Roux KH (1996) High and low annealing temperatures increase both specificity and yield in touchdown and stepdown PCR. Biotechniques 20(3):478–485. https://doi.org/10.2144/19962003478

Henke W, Herdel K, Jung K, Schnorr D, Loening SA (1997) Betaine improves the PCR amplification of GC-rich DNA sequences. Nucleic Acids Res 25(19):3957–3958. https://doi.org/10.1093/nar/25.19.3957

Klein JC, Lajoie MJ, Schwartz JJ, Strauch EM, Nelson J, Baker D, Shendure J (2016) Multiplex pairwise assembly of array-derived DNA oligonucleotides. Nucleic Acids Res 44(5):e43. https://doi.org/10.1093/nar/gkv1177

Kosuri S, Church GM (2014a) Large-scale de novo DNA synthesis: technologies and applications. Nat Methods 11(5):499–507. https://doi.org/10.1038/nmeth.2918

Kosuri S, Eroshenko N, Leproust EM, Super M, Way J, Li JB, Church GM (2010) Scalable gene amplification by selective amplification of DNA pools from high-fidelity microchips. Nat Biotechnol 28(12):1295–1299. https://doi.org/10.1038/nbt.1716

Linnarsson S (2010) Recent advances in DNA sequencing methods—general principles of sample preparation. Exp Cell Res 316(8):1339–1343. https://doi.org/10.1016/j.yexcr.2010.02.036

Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nat Genet 21(1):20–24. https://doi.org/10.1038/4447

Mihovilovic M, Lee JE (1898) An efficient method for sequencing PCR amplified DNA. Biotechniques 7:14–16

Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA In Vitro: the polymerase chain reaction. Cold Spring Harbor Symp Quant Biol 51:263–273. https://doi.org/10.1101/sqb.1986.051.01.032

Organick L, Ang SD, Chen YJ, Lopez R et al (2018) Random access in large-scale DNA data storage. Nat Biotechnol 36(3):242–248. https://doi.org/10.1038/nbt.4079

Pan W, Byrne-Steele M, Wang C, Lu S, Clemmons S, Zohrach RJ, Han J (2014) DNA polymerase preference determines PCR priming efficiency. BMC Biotechnol 14:10. https://doi.org/10.1186/1472-6750-14-10

Percze K, Meszaros T (2020) Analysis of modified nucleotide aptamer library generated by thermophilic DNA polymerases. ChemBioChem 21(20):2939–2944. https://doi.org/10.1002/cbic.202000236

Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol 64(10):3724–3730. https://doi.org/10.1128/AEM.64.10.3724-3730.1998

Sahdev S, Saini S, Tiwari P, Saxena S, Singh Saini K (2007) Amplification of GC-rich genes by following a combination strategy of primer design, enhancers and modified PCR cycle conditions. Mol Cell Probes 21(4):303–307. https://doi.org/10.1016/j.mcp.2007.03.004

Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, Church G (2004) Accurate multiplex gene synthesis from programable DNA microchips. Nature 432(7020):1050–1054. https://doi.org/10.1038/nature03151

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

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.