Deep learning-assisted analysis of amplification bias in multi-template PCR

First Author[1,2*], Second Author[2,3†] and Third Author[1,2†]

[1*]Department, Organization, Street, City, 100190, State, Country.
[2]Department, Organization, Street, City, 10587, State, Country.
[3]Department, Organization, Street, City, 610101, State, Country.

*Corresponding author(s). E-mail(s): iauthor@gmail.com;
Contributing authors: iiauthor@gmail.com; iiiauthor@gmail.com;
†These authors contributed equally to this work.

Abstract

To Do

Keywords: keyword1, Keyword2, Keyword3, Keyword4

Introduction

Multi-template PCR is fundamental to many workflows involving high-throughput sequencing, and is used across numerous fields, ranging from metabarcoding to DNA data storage. [1–3] However, a key concern regarding its use is the introduction of bias caused by inhomogeneous amplification of these templates, leading to skewed abundance data. [1, 4–6] This concern has prompted the development of unique molecular identifiers [7–9] and PCR-free workflows [4, 10, 11] to mitigate or circumvent such PCR bias in high-throughput sequencing. Nevertheless, the quantitative understanding of the magnitude and inherent reasons for the inhomogeneous amplification during multi-template PCR is still incomplete.

While PCR with a single, well-defined target amplicon is commonly optimized to ensure high amplification efficiency (e.g. via primer design and choice of annealing temperature) [12, 13], multi-template PCR faces another separate problem: even a minor difference in amplification efficiency between two templates leads to vastly different product-to-template ratios after exponential amplification. [1, 5, 14] As a result, the abundance of the products in sequencing data is often not representative of the actual template concentration in the sample. Additionally, poorly amplifying templates may also be missing from the sequencing data altogether. Commonly reported reasons for these differences in amplification efficiency include degenerate primers, amplicon length, template-product inhibition, amplicon GC content, polymerase choice, temperature profile, and stochastic effects. [1, 6, 15–20] However, less severe manifestations of PCR bias are also commonly observed in the field of DNA data storage, where the quality, abundance, and sequence identity of the template DNA are better controlled. [15, 21–24] In these multi-template PCR experiments using synthetic oligonucleotide pools, all templates share identical amplification adapters, have the same length, and are often deliberately devoid of undesired sequence properties (i.e. extreme GC content, long homopolymers, or secondary structure). [3, 21, 22, 25, 26] The presence of inhomogeneous amplification under these optimal conditions therefore suggests the existence of a smaller, but innate sequence-specific PCR bias independent of most factors previously reported in studies on biological samples.

In recent years, deep learning has revolutionized the analysis of DNA sequences, elucidating their intricate characteristics and interactions. At the cutting edge, Convolutional Neural Networks (CNNs) have significantly improved the prediction of DNA-protein interactions and the impacts of non-coding variants, while also enhancing our understanding of chromatin accessibility. [27–29] Simultaneously, progress in Recurrent Neural Networks (RNNs)-based models, such as Long Short-Term Memory (LSTM) [30] and Gated Recurrent Units (GRU) [31], has effectively targeted the sequential characteristics of DNA, identifying subtle inter-dependency between the nucleotides. [32, 33] The integration of CNN and RNN technologies into composite models leverages the advantages of both, providing a powerful approach for the extraction of features and the analysis of sequence data such as DNA sequences. [34, 35]

While deep learning models are proficient at predicting the properties of DNA sequences, their predictive capabilities alone do not fully satisfy the scientific quest for understanding the foundational mechanisms at play. [36, 37] DNA sequences contain

motifs, defined as short and recurring patterns critical to influencing both the properties and functionalities of these sequences. [38, 39] Traditionally, motif discovery was primarily handled through techniques like Multiple EM for Motif Elicitation (MEME) and Gibbs Sampling. [38, 40] These techniques, based on robust statistical and probabilistic foundations, played a crucial role in detecting these patterns in DNA sequences. Additional methods such as Weeder and DREME have further enriched motif analysis, employing exhaustive search techniques and differential analysis to expand our understanding of these critical patterns. [41, 42] However, deep learning models offer a new paradigm for motif discovery that transcends these traditional methods due to their capabilities in interpreting complex patterns in massive data. DeepBind [36], a pioneer study using deep learning for motif discovery, has shown significant success in identifying DNA sequence motifs that are highly associated with DNA-protein binding sites. Further enhancements in techniques of this kind, such as dilated convolutions and kernel-based structures, increase the effectiveness of deep learning in motif discovery, even in scenarios with limited data. [43, 44] These advancements not only boost the accuracy of predictions but also enrich our understanding of the dynamic interactions and functional roles that motifs play within DNA sequences, thus bridging the gap between mere prediction and deeper mechanistic understanding.

In this work, we investigate the magnitude and reproducibility of sequence-dependent amplification bias in multi-template PCR. We aim to pinpoint the source of this bias using deep learning to identify sequence motifs associated with poor amplification. We employ a simple exponential model of the PCR process to quantify PCR bias from sequencing data generated at different cycle numbers. Unlike previous studies applying this approach to biological samples [16, 45], we use random synthetic oligonucleotide pools as samples to control template and workflow variability as best as possible. In addition, we train one-dimensional convolutional neural networks (1D-CNNs) to predict poorly amplifying DNA sequences solely based on their sequence data. Furthermore, we propose a novel method to identify significant sequence motifs linked to low amplification efficiency. Finally, we assess the model's generalizability using data from the DNA data storage research [21, 22, 25, 46, 47]. We then validate the discovered motifs through independent externally-performed experiments.

Results

PCR Amplification Progressively Skews Coverage Distributions

To systematically investigate the inherent bias in multi-template PCR, we first analyzed the change in sequencing coverage for 12,000 random sequences as they were simultaneously amplified for up to 90 cycles (see Fig. 1). These random sequences – each consisting of 108 randomly generated nucleotides flanked by primer adapters based on Illumina TruSeq sequencing adapters – were commercially synthesized (Twist Biosciences) as an oligonucleotide pool. To obtain data on the sequences' abundance at different cycle counts, a serial amplification protocol was used to consecutively amplify a sample of the pool six times for 15 cycles each, while generating a sample ready for Illumina sequencing at each iteration (see Methods).

During serial amplification, a progressive broadening of the coverage distribution was observed (see Fig. 2a), as previously reported[23]. Whereas the overall coverage distribution changed only marginally after deep amplification for 90 cycles, a considerable number of sequences were either severely depleted or even no longer present in the sequencing data at all. Specifically, a progressive rise in the fraction of sequences with low (i.e. present at <30% relative to the mean read count) and essentially zero coverage (i.e. <5%) with increasing PCR cycles was observed (see Fig. 2b). While the presence of a GC-bias in amplification and sequencing is known,[6, 15, 17, 48] an identical analysis using an oligonucleotide pool in which the random sequences were constrained to 50% GC content (called GCfix) showed identical results. The progressive skew of the coverage distribution and the increased fraction of sequences with low coverages were comparabale between the GCall and GCfix pools (see Supplementary Fig. S12). This result strongly suggests that the presence of sequences with extreme GC content cannot explain the observed bias during PCR amplification on its own.

Estimating Aamplification Efficiencies from Sequencing Data

In order to translate the observed changes in sequencing coverage to a quantifiable bias in PCR amplification, a simple model of exponential PCR amplification was fitted to the sequencing data (see Methods)[16, 45]. This model, using two parameters per sequence, separates the initial bias in oligonucleotide pools - caused by uneven coverage after synthesis - from the PCR-induced bias, caused by each sequence's individual amplification efficiency (ϵ_i, see Supplementary Fig. S1 for an illustration). The obtained estimates for the initial coverage bias were comparable to experimental data using PCR-free sequencing of oligonucleotide pools[24] (see Fig. 2c, dashed line) and the distributions of amplification efficiencies were comparable across both datasets (see Supplementary Fig. S12). Interestingly, a small subset of sequences (representing around 2% of the pool) with very poor amplification efficiency was present in both datasets (see Fig. 2c, inset). With estimated efficiencies as low as 80% relative to the population mean (equivalent to a halving in relative abundance every 3 cycles), these sequences were often no longer present in the sequencing data at higher cycle counts. These poorly amplifying sequences also caused the rising fraction of sequences with low coverages observed in Fig. 2b and described above, highlighting their relevance for further investigation.

Poor Amplification is Reproducible

A first verification experiment was undertaken to verify the estimates of amplification efficiency on a sequence level, using qPCR dilution curves of four selected sequences from the pool. The mean qPCR efficiencies differed significantly between samples (one-way ANOVA, $F(4, 10) = 252$, $p = 5 \times 10^{-10}$) and a post-hoc Tukey's range test revealed a significantly lower qPCR efficiency for the two sequences which had been previously identified as poorly amplifying ($\epsilon = 67.8 \pm 0.8\%$ and $89.2 \pm 0.4\%$) compared to the whole pool ($93.7 \pm 2.0\%$). Surprisingly, the qPCR efficiencies of the other two individual sequences – one with average and one with superior amplification efficiency – were also marginally lower than that of the GCall pool itself ($90.8 \pm 1.0\%$ and $92.5 \pm 1.0\%$, full results in Supplementary Table S4). This effect could be explained by

the increasing over-representation of well-amplifying sequences in the GCall pool after many PCR cycles, yielding a slightly inflated qPCR efficiency. Importantly however, the qPCR experiment showed that the sequences assigned with a low PCR efficiency by the PCR model also exhibited this low PCR efficiency in an orthogonal experiment.

To further support the estimates of amplification efficiency on a pool level, a second verification experiment performed an extended serial amplification workflow over 10x15 cycles, using a new oligonucleotide pool containing 1000 selected sequences from the GCall and GCfix pools. Fig. 2d shows the evolution of the sequence coverage as a function of PCR cycles for this pool, stratified by the assigned category of amplification efficiency. Evidently, the vast majority of sequences categorized as poorly-amplifying were drastically under-represented even after just 30 PCR cycles, and effectively drowned out completely by cycle number 60. In contrast, the sequences categorized as high performers were not considerably over-represented until around cycle 90, although many occurred more than twice as often as an average sequence by cycle 150 (see Supplementary Fig. S14 for all cycle counts).

The experimental data obtained in the two verification experiments support the parameter estimation by the PCR model, and verified that the observed effects are due to PCR amplification rather than just stochastic noise. Moreover, the observed effects of PCR bias were more severe for sequences with a low estimated amplification efficiency than for those with a high estimated amplification efficiency. Therefore, we focus on identifying contributors to a low amplification efficiency in our analysis, which is also more detrimental to PCR-based workflows in biological applications and DNA data storage.

Classification of DNA Sequences by Estimated Amplification Efficiency

The 1D-CNN model employs positional encoding to classify DNA sequences based on their amplification efficiency, categorizing them into either poorly amplified or not (see Panel **a** in Figure 3). When evaluated within the GCall and GCfix datasets, the model shows high predictive power, with a five-fold cross-validation average AUROC (Area Under the Receiver Operating Characteristic curve) of 0.88 and 0.87, and AUPRC (Area Under the Precision-Recall Curve) of 0.42 and 0.44, respectively (see Panel **b** in Figure 3). The evaluation across datasets shows the robustness of the model, achieving an AUROC score of 0.86 when applying the GCall-trained model to the GCfix dataset and 0.89 vice versa, while the corresponding AUPRC scores are 0.38 and 0.39, correspondingly (see Panel **c** in Figure 3). Despite the low prevalence rate of 2%, indicating a highly challenging and imbalanced classification task, the model manages to maintain AUPRC scores close to 40% across all evaluations.

Discovery and Validation of Motifs Associated with Low PCR Efficiency

In this study, significant motifs related to low PCR efficiency were identified using a comprehensive motif discovery approach (see Panel **a** in Figure 4 and Section 3). The analysis highlights a recurring 'CGTGT' subsequence in the most significant

motifs across both GCall and GCfix datasets (see Panel **b** in Figure 4). These motifs exhibited a marked propensity to occur at the start of poorly amplified sequences, a trend which is not observed in sequences with normal amplification efficiency (see Panel **c** in Figure 4).

Subsequent validation experiments involved iterative substitution of the identified motifs within the test datasets, which resulted in a pronounced decline in model performance, quantified by the change in AUROC (see Panels **d** and **e** in Figure 4). The initial substitution of the most statistically significant motif precipitated the largest drop in AUROC, approximating a 20% decrease. As the count of substituted motifs reaches 30, the model's performance deteriorated to that of a baseline random classifier. This degradation highlights the critical role these motifs play in the predictive capacity of the model. Furthermore, the potential of the 'CGTGT' motif to form inhibitory secondary structures, which could impede primer binding during PCR, was elucidated as a plausible biological mechanism underlying the observed PCR efficiency differences (see Panel **f** in Figure 4).

Generalization to Literature Datasets

Our evaluation strategy further includes datasets from a diverse set of literature to ascertain generalization capability of the proposed model (see Figure 5). The heatmap presents the AUROC scores, with the model trained on datasets listed by rows and tested on those by columns. High AUROC values above 80% are evident when models are tested within most datasets, excluding those from Choi et al. [21] and Koch et al. [46]. Notably, models trained on the GCall and GCfix datasets demonstrate remarkable generalizability to Koch et al.'s data, achieving AUROC scores comparable to which obtained during internal dataset evaluations. Furthermore, the model trained on the dataset from Gao et al. [22] shows high prediction performance when applied to Erlich et al.'s dataset [25] and vice versa, with AUROC scores being approximately 90%. In contrast, models trained on data from Choi et al. [21] and Koch et al. [46] generally underperform on other data sets, which can be attributed to differences in experimental protocols, including variations in polymerase and adapters, and data quality. Importantly, when a model shows generalization across different datasets, the motifs identified appear to be similar, evident in this trend between the GCall/GCfix and Koch et al.'s dataset, as well as between the datasets of Erlich et al. and Gao et al. We also perform the same cross-validation of the model performance using the baseline models mentioned in Section 3, showing the superior performance of the proposed model (see Supplementary Fig S16)

External Validation of Results

The performance of the 1D-CNN models, as well as the effects of the identified motifs, were validated using an externally-performed serial amplification experiment with a new oligonucleotide pool (see Fig. 6a). Specifically, the serial amplification was performed twice, once using the conditions of GCall/GCfix (KAPA SYBR FAST, 54 °C annealing), and once using the conditions of Erlich et al.[25] (Q5 High-Fidelity, 60 °C annealing). Strikingly, this change in polymerase and annealing temperature led to drastic changes in the sequences' relative amplification efficiency (Spearman rank

correlation: -0.23, see Supplementary Table S5). Nonetheless, the models trained on the GCall and GCfix data showed essentially identical performance as previously in the internal validation (GCall: AUROC 0.80, AUPRC 0.31; GCfix: AUROC 0.82, AUPRC 0.34) for the GCall/GCfix condition (see panel **b** of Fig. 6). However, the model trained on the literature data of Erlich et al. did not show an equivalent performance on the external validation dataset produced with the conditions of Erlich et al. (AUROC 0.43, AUPRC 0.02). Instead, the GCall/GCfix models retained a modest performance (AUROC 0.6, AUPRC 0.05).

Similarly, the introduction of motifs previously identified for the GCall/GCfix data (see Fig. 4b) led to a profound decrease in PCR efficiency in the GCall/GCfix condition, as shown in panel **c** of Fig. 6. At the extreme, inserting the motif *TCGTGT* at the 5' end of a sequence led to a decrease of 4.8 ± 2.4 percentage points in amplification efficiency on average. This change is equivalent to a halving of the relative abundance approximately every 14 cycles. More generally, negative effects were observed for all GCall/GCfix motifs, with a stronger effect upon insertion at either oligo end, and the 5' end in particular. Similar effects were not observed for the motifs identified in the data by Erlich et al. under the corresponding experimental condition (see panel **c** of Fig. 6, lower half). This is in line with the performances of the corresponding 1D-CNN models, as noted above, and hints at a combined dependence of motifs on polymerase and primer choice.

Further inspection of the estimated PCR efficiencies for both experimental conditions revealed a bimodal distribution for the Erlich et al. condition; unlike the long-tailed, but unimodal distributions observed before (see Supplementary Fig. S15b). Supporting this biased amplification in the Erlich et al. condition, a comparison of the coverage distributions of both conditions during serial amplification shows a considerable broadening only in the Erlich et al. condition (see Supplementary Fig. S15a). Both effects persisted when the serial amplification experiment using the Erlich et al. condition was repeated in-house, rather than externally (Spearman rank correlation of the estimated PCR efficiency: 0.91, see Supplementary Fig. S15 and Supplementary Table S5). These results suggest that an additional effect – possibly related to the interaction between the Q5 High-Fidelity polymerase and the TruSeq amplification primer – affects a larger proportion of the sequences as we considered before (up to 16% vs. 2%).

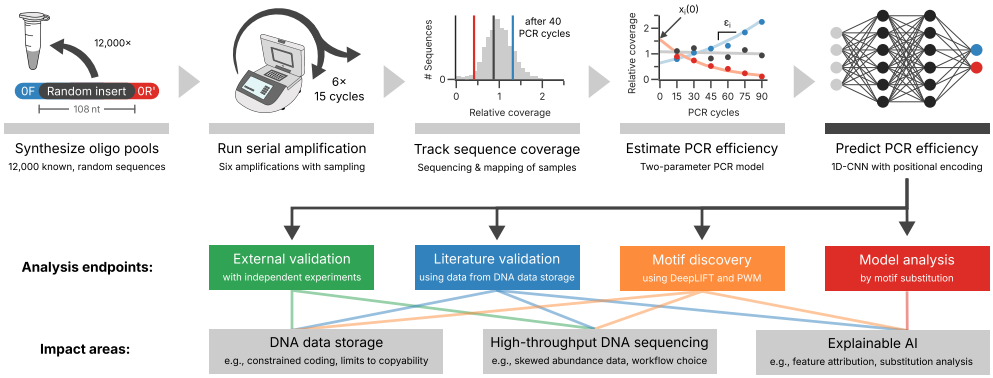

Fig. 1 Overview of the workflow and the analysis endpoints. The workflow starts with the synthesis of randomized oligonucleotide pools, which are consequently amplified serially to generate six samples with differing numbers of PCR cycles (from 15 to 90 cycles). After sequencing, the evolution of each sequence's coverage as a function of cycle number is used to estimate the PCR efficiency in the two-parameter PCR model (see Methods). These estimates of the PCR efficiencies are used in the training of an 1D-CNN model for the binary classification of PCR efficiency.

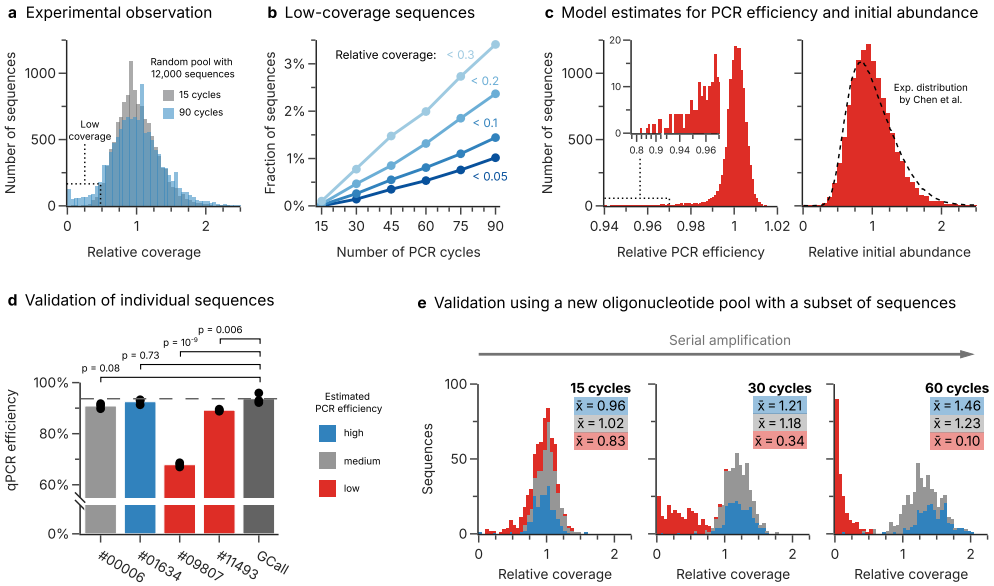

Fig. 2 Model-based estimation of PCR efficiency and its experimental validation. (a) Observed normalized coverage distributions of the GCall pool after the first (15 cycles, grey) and the sixth round of serial amplification (90 cycles, blue). **(b)** Observed fractions of underrepresented sequences in the GCall pool over the course of serial amplification. Low-coverage sequences are further grouped by their relative coverage, from occurring less frequently than 30% (light blue) to lower than 5% (dark blue). **(c)** Distributions of the relative PCR efficiency (left) and relative initial abundance (right) estimated from the experimental data using the two-parameter PCR model (see Methods). The inset for the PCR efficiency shows the subset of sequences with very low amplification efficiency. The dotted line superimposed onto the distribution of relative initial abundance shows the experimentally-determined distribution by Chen et al.[24], using a ready-to-sequence pool. **(d)** qPCR efficiencies of four individually synthesized sequences from the GCall pool (#00006 through #11493) and of the GCall pool itself, as measured with qPCR dilution curves. Two of the individual sequences had shown a low amplification efficiency during the serial amplification (#09807 and #11493, red bars). Samples #00006 and #01634 had shown average (grey) and good amplification performance (blue) respectively. Amplification efficiencies were significantly different by one-way ANOVA ($N = 3$ per sample, $F(4, 10) = 252$, $p = 5 \times 10^{-10}$), and the results of a post-hoc Tukey's range test are shown above the bars (see Supplementary Tables S3 and S4). **(e)** Observed normalized coverage distributions after 15, 30, and 60 cycles of serial amplification (iterations 1, 2, and 4 respectively) of a new pool containing a subset of the sequences present in the GCall and GCfix pools. Sequences were again selected by their PCR efficiency estimated from the PCR model, and grouped by a high (blue), medium (gray), or low (red) PCR efficiency. Insets show the mean coverage across all sequences in each category for that experiment.

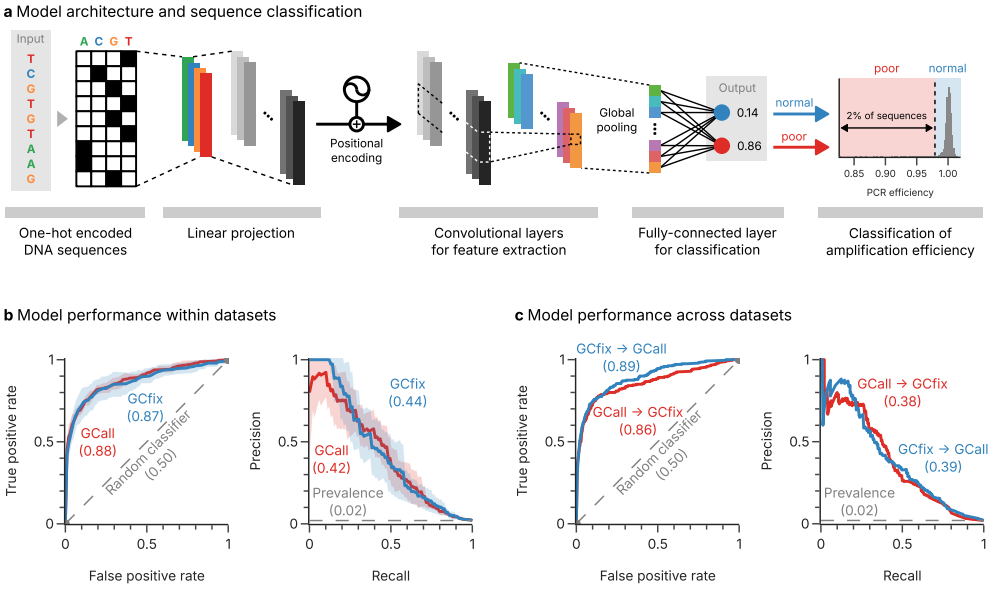

Fig. 3 Classification of DNA sequences by estimated amplification efficiency. (a) A depiction of a 1D-CNN model utilizing positional encoding to categorize sequences based on their structural attributes into poorly amplified or not. **(b)** Evaluation of the model's performance within the GCall (red) and GCfix (blue) datasets, presenting AUROC (left) and AUPRC (right) scores, with the shaded area showing the uncertainty in the results from the five-fold cross-validation. **(c)** Evaluation of the model's performance across datasets, from GCall to GCfix (red) and GCfix to GCall (blue), presenting AUROC (left) and AUPRC (right) scores.

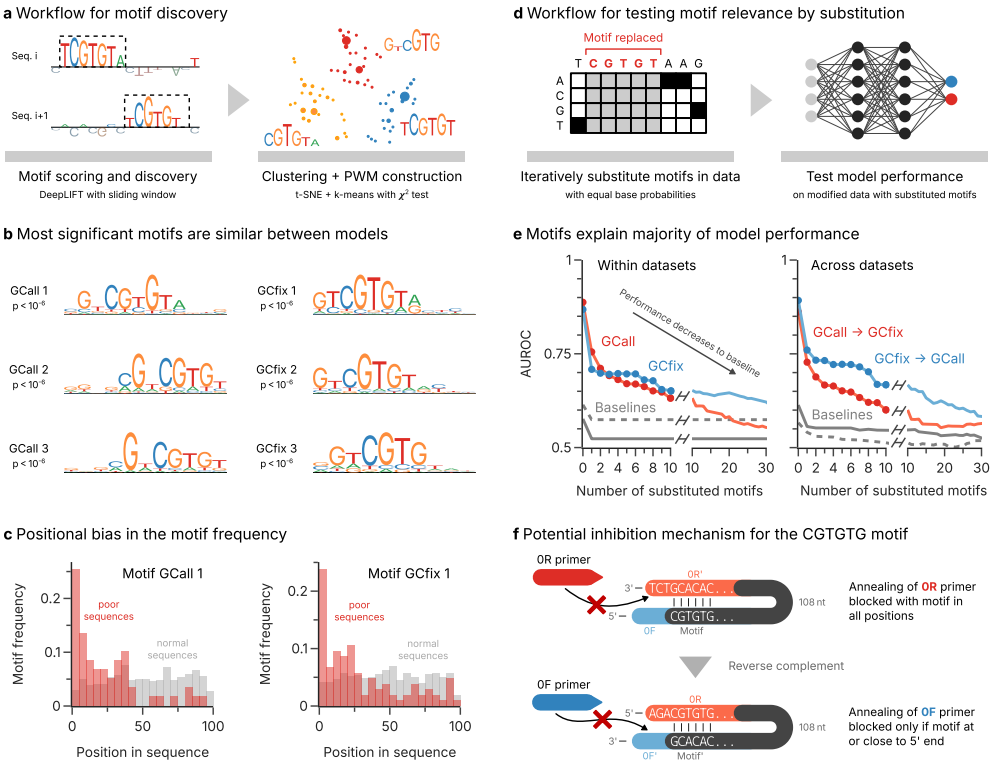

Fig. 4 Discovering motifs affecting amplification efficiency and testing their relevance. **(a)** Workflow for discovering motifs, based on extracting motifs using DeepLIFT, t-distributed stochastic neighbor embedding (t-SNE), and k-means clustering. The significance of the resulting position weight matrices (PWM) are assessed with a chi-squared test. **(b)** Most significant motifs identified for the GCall (left column) and GCfix datasets (right). The significance of the chi-squared test of each PWM is also shown. **(c)** Positional bias in the occurrence of the two motifs GCall 1 and GCfix 1 (see Panel b). The motifs in the poorly amplifying sequences (red) are more frequent at the beginning of the sequence, whereas there is no bias in normal sequences (gray). More data is shown in Supplementary Fig. S11. **(d)** Workflow for validation the discovered motifs in the trained model by iteratively replacing the motifs in the test data (ordered by p-values) without further retraining. The decrease in predictive power of the model upon motif replacement is correlated to that motif's relevance to the model output. **(e)** Evolution of model AUROC as a function of the number of substituted motifs in the test data, either using internal validation (left) or testing across datasets (right). The models trained on GCall (red) and GCfix (blue) approach the performance of the baseline LR model (gray; GCall solid, GCfix dashed) as the number of substituted motifs increases. An evaluation using AUPRC as model performance metric is shown in Supplementary Fig. S4. **(f)** Hypothesized inhibition mechanism to explain the CGTGTG motif discovered for both GCall and GCfix. The motif is complementary to the adapter present on all oligos, and thus potentially leads to hairpin structures which prevent adapter binding and subsequent extension.

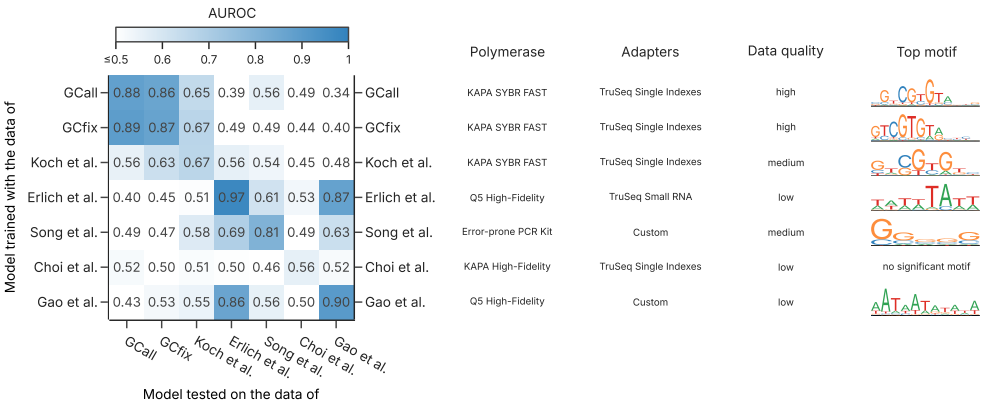

Fig. 5 Assessing model performance across literature datasets. Heatmap of the area under the receiver-operating characteristic (AUROC) metric for the models trained and tested on the different literature datasets. Additional information on the choice of polymerase and amplification adapters is provided, together with an assessment of the dataset quality and the highest-ranked motif. Additional information on the literature datasets is provided in Supplementary Table S2, and the corresponding heatmap using the area under the precision-recall curve (AUPRC) is given as Supplementary Fig. S5.

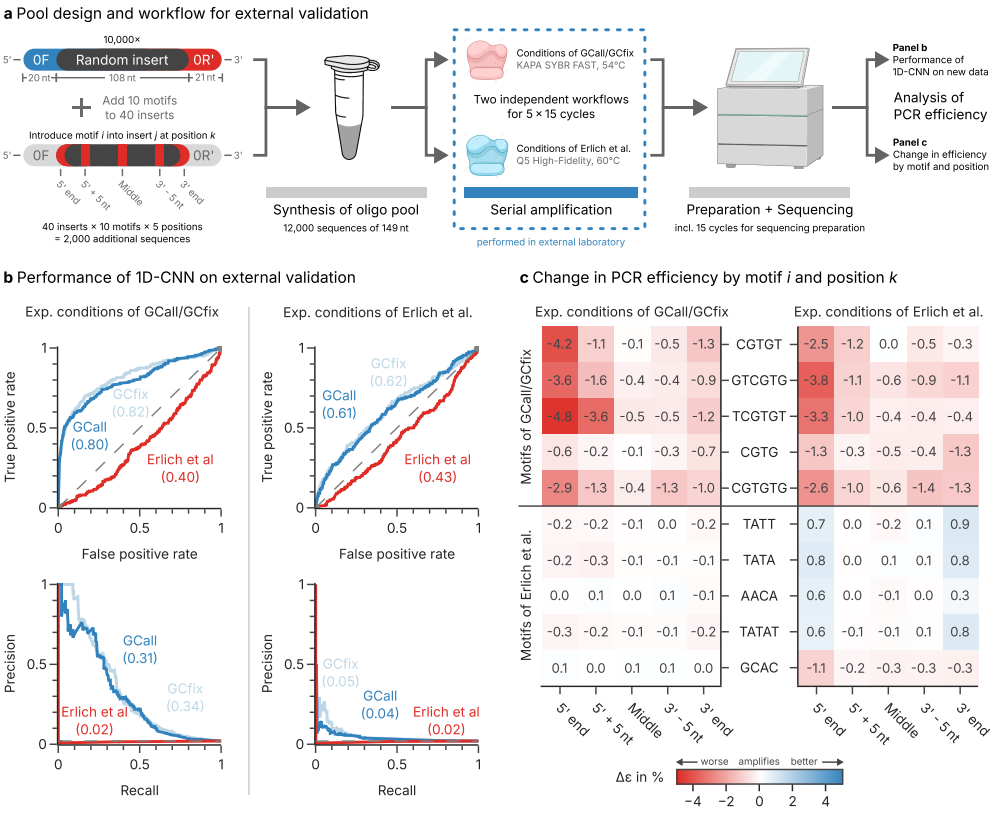

Fig. 6 Externally validating model performance and motif discovery with a motif-enriched oligonucleotide pool. (a) For external validation, a new oligonucleotide pool consisting of 10,000 random sequences and an additional 2,000 sequences with purposefully inserted motifs was used. This pool underwent serial amplification in two independent, parallel workflows which differed in the polymerase and the temperature profile used during PCR. The resulting sequencing data and estimated PCR efficiencies were used for external validation of model performance and motif discovery. **(b)** Performance metrics of the 1D-CNN models trained on the GCall (dark blue), GCfix (light blue), or Erlich et al. (red) datasets when tested on the data of the external validation experiments. The area under the receiver operating characteristic (AUROC, left) and the area under the precision-recall curve (AUPRC, right) are given in the legend. The performance metrics are shown both for the datasets created with the conditions of GCall/GCfix (solid lines) and Erlich et al. (dashed lines). **(c)** The mean change in amplification efficiency ($\Delta\epsilon$) across all sequences with a specific inserted motif (vertical axis) at a specific position (horizontal axis) compared to the sequences without inserted motif. In both workflows (left panel: conditions of GCall/GCfix, right panel: Erlich et al.), the presence of the motifs identified during motif discovery for the GCall/GCfix models lead to a considerable decrease in amplification efficiency if present at, or close to, the 5' end of the sequence.

Discussion

This study set out to characterize the amplification bias in multi-template PCR, and to identify motifs that are predictive of poor amplification using deep learning. By using multiple synthetic oligonucleotide pools with random sequences as templates, we established the presence and reproducibility of an amplification bias under well-controlled conditions. Poor amplification efficiency, affecting around 2% of randomly generated sequences, was correlated with the presence of short sequence motifs, as determined by motif discovery on deep learning-models trained on our data. Extension of our method to comparable sequencing datasets in the DNA data storage literature, [21, 22, 25, 46, 47] as well as validation with additional amplification experiments, revealed a strong dependence of the observed amplification bias on the choice of polymerase and primer. This result suggests that interactions between sequence motifs within the templates and primers or the polymerase induce an amplification disadvantage, leading to poor amplification efficiency. While such interactions – like the panhandle structure proposed in Fig. 4f – are generally avoided in quantitative PCR of single templates, [12, 13] their detrimental impact even at the short motif lengths considered in this work (approx. 6 nt) is unexpected. Similarly, the unforeseen, drastic change in sequences' amplification efficiency with a different polymerase (see Supplementary Fig.S15) highlights the need for further investigation into the apparent template-dependent inhibition mechanisms observed in this work.

A key limitation of this study is the narrow range of primer adapters and polymerases covered in our analysis. As evident from the limited generalizability of our 1D-CNN models across literature datasets (see Fig. 5), these experimental parameters appear critical to explaining the observed sub-populations of poorly amplifying sequences. However, our analysis extends the existing work that highlights and investigates PCR-induced bias [6, 16, 20, 45] by confirming its presence in random sequences and establishing a link with sequence motifs. Further work could include the integration of these experimental factors for the development of a more robust deep learning model for the prediction of the poor amplifcation efficiency, for example, to embed this information into the model architecture. Another limitation of this study concerns the reliability of the estimated amplification efficiency using the two-parameter PCR model. While our own experiments included a set of sequencing data at six cycle counts to estimate the amplification efficiency – whose reproducibility we confirmed both with experiments (see panel **c** and **d** in Fig.2) and in-silico (see Supplementary Fig. S2) – the literature datasets often included only two cycle counts. Therefore, the reliability of the results derived thereof must be corroborated prior to further investigation into the role of their sequence motifs. Nevertheless, the preliminary findings from the literature datasets yielded significant insights into the models' generalization capabilities and the method's effectiveness in identifying pertinent sequence motifs that correlate with the low amplification efficiency.

From the understanding of the PCR bias in multi-template PCR gained in this work, important considerations for its use in high-throughput sequencing can be deduced. For genetic studies of complex samples, such as in metabarcoding, the analysis presents a lower bound for the expected bias in multi-template PCR under optimal conditions (i.e., a halving of relative abundance for every 15 cycles). For

DNA data storage, the identified sequence features and the developed models present new avenues for optimal sequence design as part of constrained coding (i.e., avoiding primer motifs at the 5'-end). More generally, this work demonstrates a simple workflow to investigate amplification biases and their dependence on selected experimental choices (e.g. polymerase, primers, and adapters) in an isolated fashion using serial amplification. Moreover, the novel motif discovery apparoach developed for feature attribution and substitution analysis of 1D-CNN models for DNA sequences supports the interpretability of such models, which remains a critical challenge for their use as investigative tools.

Methods

Design of Oligonucleotide Pools

All oligonucleotide pools used in the experiments of this study were purchased from Twist Biosciences (Piscataway, NJ, United States) and designed with a fixed length of 149 nt. In all oligo pools, each design sequence contained a unique subsequence of 108 nt flanked by primer adapters (0F, 20 nt and 0R, 21 nt) for sequencing preparation, according to established protocols.[3] The two oligonucleotide pools comprised of 12,000 sequences either with (GCfix) or without (GCall) a fixed GC content of 50% contained fully randomly generated subsequences. These pools and their sequencing data have previously been used to estimate error rates and biases in the DNA data storage workflow.[23] The test pool used for assessing the reproducibility of the amplification bias comprised 2,000 sequences which were selected from the previous two pools on the basis of their estimated amplification efficiency. The validation pool used for external validation of the machine learning model and the effect of motifs comprised a total of 12,000 sequences. Of those, 10,000 were fully randomly generated without any constraint on GC content. For the remaining 2,000 sequences, we selected the five most significant motifs inferred from the models trained on the GCall/GCfix datasets or the literature dataset by Erlich et al.[25] and created additional sequences with these motifs. To do so, 40 sequences of the randomly generated subset that did not already contain any of the motifs were selected, and each motif was inserted into each sequence once at the start, the end, the middle, 5 nt from the start, and 5 nt from the end of the sequence, replacing the nucleotides present there. This resulted in 50 additional sequences for each of the 40 sequences selected from the subset, for a total of 2,000 additional non-random sequences. Due to the overlapping nature of some motifs (e.g. motif CGTG is contained in motif CGTGT), the resulting set of sequences contained duplicates (e.g. if the sequence randomly featured T at the position following the inserted motif CGTG). In the analysis of the sequencing data, these sequences were always associated with the longer motif (e.g. to CGTGT in the example) to isolate the effects of short motifs as best as possible.

Serial Amplification of Pools

All oligonucleotide pools were resuspended to 10 ng μL^{-1} in ultrapure water. All experiments except the second serial amplification of the validation pool used KAPA SYBR

FAST polymerase master mix from Sigma-Aldrich (St. Louis, MI, United States) for amplification, employing a temperature profile with an initial denaturation at 95 °C for 3 min, followed by 15 cycles at 95 °C for 15 s, 54 °C for 30 s, and 72 °C for 30 s. The second serial amplification of the validation pool used Q5 Hot Start High-Fidelity master mix (Catalog# M0494) from New England Biolabs (Ipswich, MA, United States) instead, employing a temperature profile with an initial denaturation at 95 °C for 30 s, followed by 15 cycles at 95 °C for 10 s, 60 °C for 30 s, and 72 °C for 30 s.[25] For each amplification, 5 µL of sample were mixed with 10 µL of 2x master mix, 3 µL of ultrapure water, and 1 µL each of the forward and reverse primer at 10 µM (Microsynth AG, Balgach, Switzerland). An overview of the primer sequences used in our experiments and during the generation of the literature datasets is given in Supplementary Table **??**.

The serial amplification of oligonucleotide pools followed an iterative protocol, as previously described,[23] to prevent resource exhaustion during PCR. In short, each iteration started by diluting 1 µL of the sample from the previous iteration by a factor of 3800x in ultrapure water (or 7600x, if the sample had approached the plateau phase after 15 cycles in the previous iteration). Then, the sample was amplified for 15 cycles in two wells: once using the standard primers (0F/0R), and once using primers with an overhang containing indexed sequencing adapters (2FUF/2RIF). The PCR product with sequencing adapters was then stored at −20 °C, whereas the PCR product with the standard primers was directly used for the next iteration.

In the case of the validation pool, the amplifications with the standard primers (0F/0R) were performed in the laboratory of Prof. Dr. Beat Christen at the Institute of Microbiology of the University of Stuttgart. The amplification with sequencing adapters (2FUF/2RIF) and the sequencing, both of which are common and identical for all samples, were then performed in-house.

For the first iteration of each serial amplification, the oligonucleotide pool at 10 ng µL^{-1} was diluted by a factor of 500x in ultrapure water and used without further dilution.

Sequencing and Data Preprocessing

The PCR product with indexed sequencing adapters was purified by excision of the appropriate band on an agarose gel (E-Gel EX Agarose Gels 2%, Invitrogen) with a 50 bp ladder (Invitrogen), and subsequent spin-column purification (ZymoClean Gel DNA Recovery Kit, ZymoResearch). All samples were quantified by fluorescence (Qubit dsDNA HS Kit, Invitrogen) prior to dilution to 1 nM with ultrapure water. Multiple samples were then pooled, further diluted to 50 pM, and finally sequenced on an Illumina iSeq 100 sequencer with 150 bp paired reads.

The demultiplexed sequencing data was post-processed by adapter trimming and read mapping using BBMap[49] (v39.01) against the pool's reference sequences. Reference sequences whose reads occurred in fewer than two sequencing runs across the dataset were removed from the data. The read counts for all remaining reference sequences, normalized by the mean number of reads per reference sequence in the dataset, were used as coverage distributions for further analysis.

Parameter Estimation from Coverage Distributions

To estimate the synthesis bias $x_i(0)$ and the relative amplification efficiency ϵ_i of each reference sequence i in a set of serial amplification experiments, we model the evolution of the relative sequence coverage $x_i(c_j)$ after c_j cycles as shown in Eq. 1.[16, 23, 45]

$$x_i(c_j) = x_i(0) \cdot \epsilon_i^{c_j} \tag{1}$$

Full estimation of the parameters of all N reference sequences across all M serial amplifications experiments is given by the solution to the least-squares problem of the sparse, log-linearized system of equations described by the PCR model in Eq. 2.

$$\log x_i(c_j) = \log x_i(0) + c_j \cdot \log \epsilon_i \quad \forall i \in [1, ..., N], \forall j \in [1, ..., M] \tag{2}$$

The estimated parameters were finally normalized to their mean. To validate the chosen approach, artificial sequencing data was generated *in-silico* using different defined distributions of initial synthesis bias and relative amplification efficiency. In a first test, a simple model based on Eq. 1 was used to generate sequencing data without the stochastic effects of PCR, dilution, and sequencing. In a second test, the full workflow was implemented in a digital twin of the DNA data storage process[23] to investigate the approach's robustness against stochastic noise. These validations of the model showed sufficient reliability of the parameter estimates under the expected experimental noise (e.g., stochastic sampling, or stochastic PCR effects) and no sensitivity to the underlying distribution of the parameters (see Supplementary Figure S2). However, accuracy of the parameter estimates decreased if a sequence was observed only in few sequencing runs (Supplementary Figure S2). Thus, we did not consider sequences which occurred less than two times across a set of sequencing data in the following analysis.

Efficiency Measurements by qPCR

Experimental quantification of the qPCR efficiency was performed for four selected sequences of the GCall oligonucleotide pool. Of the four selected sequences, #00006 had an estimated amplification efficiency of 0.999, #01634 of 1.014, #11493 of 0.854, and the efficiency of #09807 could not be estimated because it was filtered out due to occurring in only one sequencing run (relative coverage after first amplification: 0.21, thereafter 0). Sequence #09807 was included nonetheless to assess whether such sequences were missing in the sequencing data due to stochastic effects or because of an extremely poor amplification efficiency. These oligonucleotide sequences were synthesized individually by Microsynth (Balgach, Switzerland), dissolved with ultrapure water to 100 µM, serially diluted five times by factors of 10x, and each dilution measured by qPCR with the standard primers (0F/0R) in duplicates to create a calibration curve. A total of three independent dilutions and qPCR runs were performed for each oligonucleotide sequence, with the results shown in Supplementary Figure S3 and Supplementary Table S3. The qPCR efficiency of the GCall oligonucleotide pool itself was also measured in triplicate with the same range of dilutions. Comparison of qPCR-derived efficiencies was performed using one-way independent ANOVA after

testing for homoscedasticity with Levene's test, followed by post-hoc testing using Tukey's range test.

Selection of Literature Datasets

Multiple additional sequencing datasets from the literature were selected to test and train the classification model. For parameter estimation, datasets must include sequencing data for at least two different PCR cycle counts and their sequencing coverage must be sufficiently high to yield accurate coverage distributions. Moreover, to preclude any possible bias stemming from sequences with biological function or extreme sequence properties (such as GC content or long nucleotide repeats), we limited our search to datasets derived from synthetic oligonucleotide pools which contain close-to-random sequences. Multiple such datasets were identified in the DNA data storage literature, from Erlich et al.[25], Koch et al.[46], Song et al.[47], Choi et al.[21], and Gao et al.[22] and processed as described above. A detailed overview of the experimental parameters and sequencing endpoints of all literature datasets is given in Supplementary Table S2.

Deep Learning Model

In this work, the main model we propose to use is a 1D-CNN model to predict whether a DNA sequence is of low PCR amplification efficiency based on the sequence structure. Given the lack of prior knowledge about the likelihood of low PCR efficiency in randomly synthesized sequences, we empirically set a 2% threshold to categorize sequences into low efficiency sequences and normal efficiency sequences. Following this categorization, we formulate the prediction task as a binary classification task.

1D-CNN model has proven its superior efficacy over traditional methods in many DNA sequence property prediction tasks, as introduced in the introduction section. However, 1D-CNN does not have a notion of sequential order information of the nucleotides, because of its translation-invariant nature, while such information could potentially be important for the identification of the motifs. For example, in eukaryotic chromosomes, telomeres are repetitive nucleotide sequences at each end of a chromosome. The specific motif sequence, which often comprises repetitions of a short DNA sequence (like 'TTAGGG' in vertebrates). The positional specificity of these motifs at the very ends of the chromosomes is vital for their function in protecting the chromosome from deterioration or fusion with neighboring chromosomes. [**de2005shelterin blackburn2005telomeres**] Therefore, an additional positional encoding (PE) component, which was first introduced with the Transformer model [50], is added to the 1D-CNN model, to provide an unique positional information of each nucleotide in the sequence.

To be specific, positional encoding in neural networks is designed to embed sequence position information into models that do not inherently process sequential data, and the formula:

$$PE(p, 2i) = \sin\left(\frac{p}{10000^{\frac{2i}{d}}}\right) \tag{3}$$

$$PE(p, 2i+1) = \cos\left(\frac{p}{10000^{\frac{2i}{d}}}\right)$$

uses sinusoidal functions to encode each position p in a high-dimensional space, with d representing the encoding dimensionality, the factor 10000 in the denominator is to create a smooth and predictable change in the encoding values as the position increases, as used in [50]. This method allows the model to learn and take advantage of the positional information of elements in a given sequence.

Baseline Models

To demonstrate the efficacy of the 1D-CNN model in identifying sequences with lower PCR efficiency, we established baseline models for comparative evaluation of the proposed 1D-CNN model with PE. The baseline models include:

- 1D CNN model without PE.
- Recurrent neural network (RNN)-based model.
- Lasso regularized logistic regerssion (LR) model with the frequency of each nucleotide and the GC content in the sequence.

Experimental Setup

We use two metrics for our analysis: the Area Under the Receiver Operating Characteristic (AUROC) and the Area Under the Precision-Recall Curve (AUPRC). For evaluation within dataset, we split the data into three sets: training (70%), validation (10%), and test (20%). And we perform a 5-fold cross-validation (CV), and the CV was stratified to maintain the class imbalance in all folds, specifically for positive cases. For evaluation across datasets (training the model on one dataset and validating it on the other), the development dataset is split into a training set (80%) and a validation set (20%). As the task is a binary classification task, the model is trained to minimize the binary cross entropy loss in the training data. For each evaluation scheme, we perform a randomized search over 50 iterations (each corresponding to a hyperparameter configuration, including the learning rate, width, depth, batch size, weight decay, dropout, etc. The parameter search space is detailed in Table 1 and the best hyperparameter configuration is determined on the performance of the validation set, while we report the performance on the test set. In evaluation within dataset, the best hyperparameter configuration is used in all five training sets (which were the five repetitions of splitting) to create five repetition models to assess the robustness of the performance when different training data are used. Then the model of each repetition is applied to the test split to evaluate and report the final performance metrics. We also use class weights inversely proportional to the frequency of the positive class, allowing the loss function to assign the same weight to the few positively labeled sequences as to the many negatively labeled ones (2% positive vs 98% negative).

Motif Discovery and Analysis

As introduced in Section 3, DNA motifs are short and recurring subsequences found within DNA sequences, which are believed to play critical biological roles. In this

Hyperparameter	Search values
Number of convolutional layers	1, 2, 3
Number of convolutional filters	32, 64, 128
Length of convolutional filters	4, 8, 12, 16
Learning rate	10^{-3}, 10^{-4}, 10^{-5}
Dropout	0, 0.1, 0.3, 0.5
Batch size	64, 128, 256
Weight decay	0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}

Table 1 Hyperparameter grid and ranges for the hyperparameter search of the 1D-CNN model.

study, we specifically focus on motifs that negatively impact the PCR amplification efficiency (ϵ) of DNA sequences.

To this end, we propose an innovative and comprehensive motif discovery approach to interpret the predictions made by the 1D-CNN model. The approach is based on the DeepLIFT (Deep Learning Important FeaTures) [51] methodology. DeepLIFT is a technique for interpreting predictions made by deep learning models. It works by comparing the activation of each neuron to its reference activations and assigns an attribution score to each individual feature based on the difference. Attribution scores approximating the gradient of the model are obtained by back-propagating these differences in one pass from the output to input nodes. The attributions calculated using DeepLIFT are on the nucleotide level, meaning that each nucleotide in the test sequence is assigned to a score indicating their impacts toward the binary prediction target.

After obtaining the attribution score per nucleotide, we can evaluate how each nucleotide affects the prediction of the model on each sequence. And a straightforward approach for motif discovery using the attribution scores is to examine these attributions for each sequence individually. However, this method has notable limitations. Inspecting sequences one by one using DeepLIFT can lead to inconsistent interpretations, particularly with complex or subtle motifs, especially when dealing with intricate patterns that might be easily missed or misinterpreted. Additionally, this approach might not fully capture the diversity of the motifs, as it could overlook subtle variations or patterns that are more evident when analyzing the entire dataset collectively.

To address these challenges, we propose to blends unsupervised clustering techniques with convolutional kernel principles to analysis the attribution scores from DeepLIFT. This integrated approach allows for a holistic and comprehensive analysis of test sequences, capitalizing on DeepLIFT's feature extraction strengths while addressing the method's limitations in analyzing individual sequences. In the following paragraphs we detail how the analysis is performed:

We employ a sliding-window technique to cover all positions in the sequence with window sizes ranging from 4 to 12 nucleotides. Then we take the window with the highest cumulative attribution scores as the most significant k-mer in that sequence. These significant k-mers are collected across all test sequences. Then we compute pairwise Hamming distances for the subsequences under each window size. Hamming distance is defined as the number of positions at which the corresponding symbols

differ between two k-mers. This metric helps in quantifying the dissimilarity between k-mers.

Following this, we utilize the t-SNE (t-distributed Stochastic Neighbor Embedding) [52] algorithm to project the pairwise distance matrix onto a two-dimensional space. In this 2D space, we then apply weighted KMeans clustering. This advanced form of KMeans clustering considers the weight of each data point during the clustering process. In the weighted KMeans algorithm, the center of each cluster is recalculated as the weighted average of all data points assigned to that cluster. These weights are proportional to the frequency of occurrence of each pattern in the data set, allowing patterns that occur more frequently to have a greater influence on the clustering outcome. This method ensures that repetitive or prevalent patterns are adequately represented in the analysis, providing a more accurate and representative cluster of motifs. Mathematically, weighted KMeans differs from regular KMeans in how the positions of the new centers are updated: for cluster j, the new center C_j is updated as:

$$C_j = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{4}$$

where w_i is the weight of the data point x_i, and the sum is over all data points assigned to the cluster j. And the weight w_i is determined by the number of occurrences of the pattern i in all the sequences. This allows repetitive patterns to have more influence on the clustering process.

The number of clusters K is determined using the silhouette score of the clusters in the t-SNE embedded space, as we do not have any prior knowledge of the number of clusters in these subsequences. The silhouette score for a single data point is calculated as:

$$s = \frac{b - a}{\max(a, b)} \tag{5}$$

where a is the mean distance between the data point and all other points in the same cluster, and b is the smallest mean distance of the data point to all points in any other cluster, of which the data point is not a member. This score ranges from -1 to 1, where a high value indicates that the data point is well matched to its own cluster and poorly matched to neighboring clusters.

Once the clustering process is completed, we construct a Positional Weight Matrix (PWM) from each cluster. The PWM is formulated based on the frequency of occurrence of patterns within the cluster. These matrices serve as a detailed representation of the motifs corresponding to each cluster of subsequences. The PWM for each cluster is then used as a convolution kernel, traversing all sequences. The convolution operation assesses the similarity between the PWM and the subsequence at each position. The highest similarity score is indicative of the similarity between the PWM and the target sequence. Specifically, the similarity score, $\mathcal{F}$, is determined as follows:

$$\mathcal{F} = \max\left(\frac{\langle P, W_i\rangle}{k}\right) \quad \text{for } i = 1, 2, \ldots, L-k+1 \tag{6}$$

where P denotes the PWM and W_i represents the corresponding window in the sequence, and k s is the window length and L is the length of the target sequence.

If the score $\mathcal{F}$ is higher than the predefined threshold of 0.5, it suggests the presence of the motif in the given sequence, otherwise we consider absence. After that, we compute the contingency table for the presence and absence of all the motifs in the negative and positive sets of the test sequences. The $\tilde{\chi}^2$ test is performed to test whether the motif is statistically significantly enriched in the positive set of the sequences, with the null hypothesis being that there is no association between the presence of a motif and the sequence being in the low PCR efficiency group. The chi-squared statistic is calculated based on the observed and expected frequencies in the contingency table. A statistically significant chi-squared value suggests a rejection of the null hypothesis, indicating a non-random association between the motif and the positive set of sequences (i.e. the sequence considered to have low PCR amplification efficiency). We set the base α value as 0.05 and the Bonferroni correction is applied to adjust for multiple testing problem. The detailed procedure for implementing the proposed motif discovery approach is described in pseudo-code presented in Algorithm 1.

Lastly, to further evaluate the impact of the motifs we discovered, we substitute these identified motifs (arranged by the p-value from the lowest to the highest) in all sequences with an average subsequences (each entry is 0.25 in the one-hot encoding setting). This substitution process is carried out exclusively in the test set without the need to retrain the model for evaluation both within and across the datasets. And by comparing the performance from the original set of sequences and from the set of sequences after each substitution, the impact of each motif on the prediction can be quantified.

Algorithm 1 Motif Discovery and Analysis Method

Input: Lower ϵ sequences $\mathbf{S}$: $\{s_1, s_2, \ldots, s_n\}$ and the rest sequences $\mathbf{S}'$: $\{s'_1, s'_2, \ldots, s'_{n'}\}$ each of length l; Trained 1D-CNN model $\mathcal{M}$
Output: Motifs contribute to lower ϵ
Step 1: DeepLIFT feature attribution & Significant k-mer Discovery
for $k = 1$ to n **do**
  Compute attribution score z_k for sequence s_k using DeepLIFT and $\mathcal{M}$
  for Window lengths $w = 4$ to 12 **do**
    Extract most significant subsequence $subseq_{k,w}$ from z_k
  end for
end for
Step 2: k-mer Clustering & PWM Construction
for each window length $w = 4$ to 12 **do**
  Compute pairwise Hamming distance matrix $\mathbf{D}_w$ for all extracted subsequences of length w
  Project $\mathbf{D}_w$ to a lower-dimensional representation $\mathbf{D'}_w$ using t-SNE
  Cluster $\mathbf{D'}_w$ using weighted K-means to form clusters $\{C_{1,w}, C_{2,w}, \ldots, C_{m,w}\}$
  for each cluster $C_{i,w}$ **do**
    Compute PWM $P_{i,w}$ for all subsequences in cluster $C_{i,w}$
  end for
end for
Step 3: $\tilde{\chi}^2$ Test for PWMs
for each PWM $P_{i,w}$ **do**
  Assess similarity between $P_{i,w}$ and all sequences in $\mathbf{S}$ and $\mathbf{S}^{'}$ using convolution operation.
  Compute p-value for $P_{i,w}$ based on its enrichment in $\mathbf{S}$ compared to in $\mathbf{S}^{'}$ using $\tilde{\chi}^2$ test.
  Adjust p-values for multiple comparisons using Bonferroni correction
  if $P_{i,w}$ is statistically significantly enriched **then**
    return $P_{i,w}$ as a discovered motif
  end if
end for
Step 4: Motif Substitution Analysis
for each statistically significantly enriched PWM $P_{i,w}$, ordered by p-value **do**
  Substitute $P_{i,w}$ with an average one-hot encoded subsequence in both $\mathbf{S}$ and $\mathbf{S}'$
  Assess $\mathcal{M}$ performance post substitution
end for

Data availability

The experimental sequencing data generated in this study have been deposited in the European Nucleotide Archive under accession code XXXXXX. The literature sequencing data are available from Gimpel et al.[23] (PRJEB65931), Koch et al.[46] (PRJEB35217), Erlich et al.[25] (PRJEB19305 and PRJEB19307), Song et al.[47] (doi:10.6084/m9.figshare.16727122.v2, doi:10.6084/m9.figshare.17193128.v1, doi:10.6084/m9.figshare.18515045.v1), Gao et al.[22] (pers. communication), Choi et al.[21] (PRJNA555140). Source data are provided with this paper.

Code availability

Acknowledgements

This project was partially financed by the European Union's Horizon 2020 Program, FET-Open: DNA-FAIRYLIGHTS, grant agreement no. 964995, and also supported by European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Grant Agreement No. 813533 and core funding by the Max Planck Society (to K.B.). Data analysis and simulations were performed on the Euler cluster operated by the High-Performance Computing group at ETH Zürich. Figures were partially created with BioRender.com.

Competing interests

The authors declare no competing interests.

Authors' contributions

References

1. Kalle, E., Kubista, M. & Rensing, C. Multi-template polymerase chain reaction. *Biomolecular Detection and Quantification* **2,** 11–29. ISSN: 2214-7535. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5121205/ (2024) (Dec. 4, 2014).
2. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270.** Publisher: Royal Society, 313–321. https://royalsocietypublishing.org/doi/10.1098/rspb.2002.2218 (2024) (Feb. 7, 2003).
3. Meiser, L. C. *et al.* Reading and writing digital data in DNA. *Nature Protocols 2019 15:1* **15.** Publisher: Nature Publishing Group, 86–101. ISSN: 1750-2799. https://www.nature.com/articles/s41596-019-0244-5 (2021) (Nov. 29, 2019).
4. Van Dijk, E. L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research* **322,** 12–20. ISSN: 0014-4827. https://www.sciencedirect.com/science/article/pii/S0014482714000160 (2022) (Mar. 10, 2014).
5. Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology* **62.** Publisher: American Society for Microbiology, 625–630. ISSN: 00992240. https://journals.asm.org/journal/aem (2022) (1996).
6. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12,** R18. ISSN: 1474-760X. https://doi.org/10.1186/gb-2011-12-2-r18 (2024) (Feb. 21, 2011).
7. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9.** Publisher: Nature Publishing Group, 72–74. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.1778 (2024) (Jan. 2012).
8. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods* **11.** Publisher: Nature Publishing Group, 163–166. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.2772 (2024) (Feb. 2014).
9. Casbon, J. A., Osborne, R. J., Brenner, S. & Lichtenstein, C. P. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research* **39,** e81. ISSN: 0305-1048. https://doi.org/10.1093/nar/gkr217 (2024) (July 1, 2011).
10. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods 2009 6:4* **6.** Publisher: Nature Publishing Group, 291–295. ISSN: 1548-7105. https://www.nature.com/articles/nmeth.1311 (2022) (Mar. 15, 2009).
11. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461.** Publisher: Nature Publishing Group, 814–818. ISSN: 1476-4687. https://www.nature.com/articles/nature08390 (2024) (Oct. 2009).
12. Saiki, R. K. in *PCR Technology: Principles and Applications for DNA Amplification* (ed Erlich, H. A.) 7–16 (Palgrave Macmillan UK, London, 1989). ISBN: 978-1-349-20235-5. https://doi.org/10.1007/978-1-349-20235-5_1 (2024).
13. Rodríguez, A., Rodríguez, M., Córdoba, J. J. & Andrade, M. J. in *PCR Primer Design* (ed Basu, C.) 31–56 (Springer, New York, NY, 2015). ISBN: 978-1-4939-2365-6. https://doi.org/10.1007/978-1-4939-2365-6_3 (2024).

14. Korvigo, I., Igolkina, A. A., Kichko, A. A., Aksenova, T. & Andronov, E. E. Be aware of the allele-specific bias and compositional effects in multi-template PCR. *PeerJ* **10,** e13888. ISSN: 2167-8359. https://peerj.com/articles/13888 (2023) (Aug. 30, 2022).
15. Qiao, H. *et al.* Oligo replication advantage driven by GC content and Gibbs free energy. *Biotechnology Letters* **44,** 1189–1199. ISSN: 1573-6776. https://doi.org/10.1007/s10529-022-03295-2 (2023) (Oct. 1, 2022).
16. Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Research* **43.** Publisher: Oxford University Press, e143. ISSN: 13624962. /pmc/articles/PMC4666380/ (2022) (2015).
17. 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.** Publisher: Future Science Ltd London, UK. ISSN: 07366205. https://www.future-science.com/doi/abs/10.2144/000113809 (2022) (Feb. 3, 2012).
18. O'Donnell, J. L., Kelly, R. P., Lowell, N. C. & Port, J. A. Indexed PCR Primers Induce Template-Specific Bias in Large-Scale DNA Sequencing Studies. *PLOS ONE* **11.** Publisher: Public Library of Science, e0148698. ISSN: 1932-6203. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0148698 (2024) (Mar. 7, 2016).
19. Bonk, F., Popp, D., Harms, H. & Centler, F. PCR-based quantification of taxa-specific abundances in microbial communities: Quantifying and avoiding common pitfalls. *Journal of Microbiological Methods* **153.** Publisher: Elsevier, 139–147. ISSN: 0167-7012. (2022) (Oct. 1, 2018).
20. Mallona, I., Weiss, J. & Marcos, E. C. PcrEfficiency: A Web tool for PCR amplification efficiency prediction. *BMC Bioinformatics* **12.** Publisher: BioMed Central, 1–7. ISSN: 14712105. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-404 (2022) (Oct. 20, 2011).
21. Choi, Y. *et al.* DNA Micro-Disks for the Management of DNA-Based Data Storage with Index and Write-Once–Read-Many (WORM) Memory Features. *Advanced Materials* **32.** Publisher: Wiley-VCH Verlag, 2001249. ISSN: 0935-9648. https://onlinelibrary.wiley.com/doi/10.1002/adma.202001249 (2022) (Sept. 29, 2020).
22. Gao, Y., Chen, X., Qiao, H., Ke, Y. & Qi, H. Low-Bias Manipulation of DNA Oligo Pool for Robust Data Storage. *ACS Synthetic Biology* **9.** Publisher: American Chemical Society, 3344–3352. ISSN: 21615063. https://pubs.acs.org/doi/full/10.1021/acssynbio.0c00419 (2022) (Dec. 18, 2020).
23. Gimpel, A. L., Stark, W. J., Heckel, R. & Grass, R. N. A digital twin for DNA data storage based on comprehensive quantification of errors and biases. *Nature Communications* **14.** Number: 1 Publisher: Nature Publishing Group, 6026. ISSN: 2041-1723. https://www.nature.com/articles/s41467-023-41729-1 (2024) (Sept. 27, 2023).
24. Chen, Y.-J. *et al.* Quantifying molecular bias in DNA data storage. *Nature Communications 2020 11:1* **11.** Publisher: Nature Publishing Group, 1–9. ISSN:

2041-1723. https://www.nature.com/articles/s41467-020-16958-3 (2021) (June 29, 2020).
25. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355.** Publisher: American Association for the Advancement of Science, 950–954. https://www.science.org/doi/abs/10.1126/science.aaj2038 (2021) (Mar. 3, 2017).
26. Ping, Z. *et al.* Towards practical and robust DNA-based data archiving using the yin–yang codec system. *Nature Computational Science 2022 2:4* **2.** Publisher: Nature Publishing Group, 234–242. ISSN: 2662-8457. https://www.nature.com/articles/s43588-022-00231-2 (2022) (Apr. 25, 2022).
27. Emamjomeh, A., Choobineh, D., Hajieghrari, B., MahdiNezhad, N. & Khodavirdipour, A. DNA–protein interaction: identification, prediction and data analysis. *Molecular biology reports* **46,** 3571–3596 (2019).
28. Zeng, H. & Gifford, D. K. Predicting the impact of non-coding variants on DNA methylation. *Nucleic acids research* **45,** e99–e99 (2017).
29. Nair, S., Kim, D. S., Perricone, J. & Kundaje, A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinformatics* **35,** i108–i116 (2019).
30. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9,** 1735–1780 (1997).
31. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
32. Zhang, J. X. *et al.* A deep learning model for predicting next-generation sequencing depth from DNA sequence. *Nature communications* **12,** 4387 (2021).
33. Tang, X. *et al.* Deep6mAPred: A CNN and Bi-LSTM-based deep learning method for predicting DNA N6-methyladenosine sites across plant species. *Methods* **204,** 142–150 (2022).
34. Yang, B. *et al.* BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics* **33,** 1930–1936 (2017).
35. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic acids research* **44,** e107–e107 (2016).
36. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology* **33,** 831–838 (2015).
37. Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Briefings in bioinformatics* **18,** 851–869 (2017).
38. Bailey, T. L., Elkan, C., *et al.* Fitting a mixture model by expectation maximization to discover motifs in bipolymers (1994).
39. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16,** 16–23 (2000).
40. Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *science* **262,** 208–214 (1993).

41. Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research* **32,** W199–W203 (2004).
42. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27,** 1653–1659 (2011).
43. Gupta, A. & Rush, A. M. Dilated convolutions for modeling long-distance genomic dependencies. *arXiv preprint arXiv:1710.01278* (2017).
44. Chen, D., Jacob, L. & Mairal, J. Biological sequence modeling with convolutional kernel networks. *Bioinformatics* **35,** 3294–3302 (2019).
45. Silverman, J. D. *et al.* Measuring and mitigating PCR bias in microbiota datasets. *PLOS Computational Biology* **17** (ed McHardy, A. C.) e1009113. ISSN: 1553-7358. https://dx.plos.org/10.1371/journal.pcbi.1009113 (2023) (July 6, 2021).
46. Koch, J. *et al.* A DNA-of-things storage architecture to create materials with embedded memory. *Nature Biotechnology 2019 38:1* **38.** Publisher: Nature Publishing Group, 39–43. ISSN: 1546-1696. https://www.nature.com/articles/s41587-019-0356-z (2021) (Dec. 9, 2019).
47. Song, L. *et al.* Robust data storage in DNA by de Bruijn graph-based de novo strand assembly. *Nature Communications 2022 13:1* **13.** Publisher: Nature Publishing Group, 1–9. ISSN: 2041-1723. https://www.nature.com/articles/s41467-022-33046-w (2022) (Sept. 12, 2022).
48. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biology* **14.** Publisher: BioMed Central, 1–20. ISSN: 1474760X. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-5-r51 (2022) (May 29, 2013).
49. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner* LBNL-7065E (Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), Mar. 17, 2014). https://www.osti.gov/biblio/1241166 (2023).
50. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).
51. Shrikumar, A., Greenside, P. & Kundaje, A. *Learning important features through propagating activation differences* in *International conference on machine learning* (2017), 3145–3153.
52. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).

Extended Data

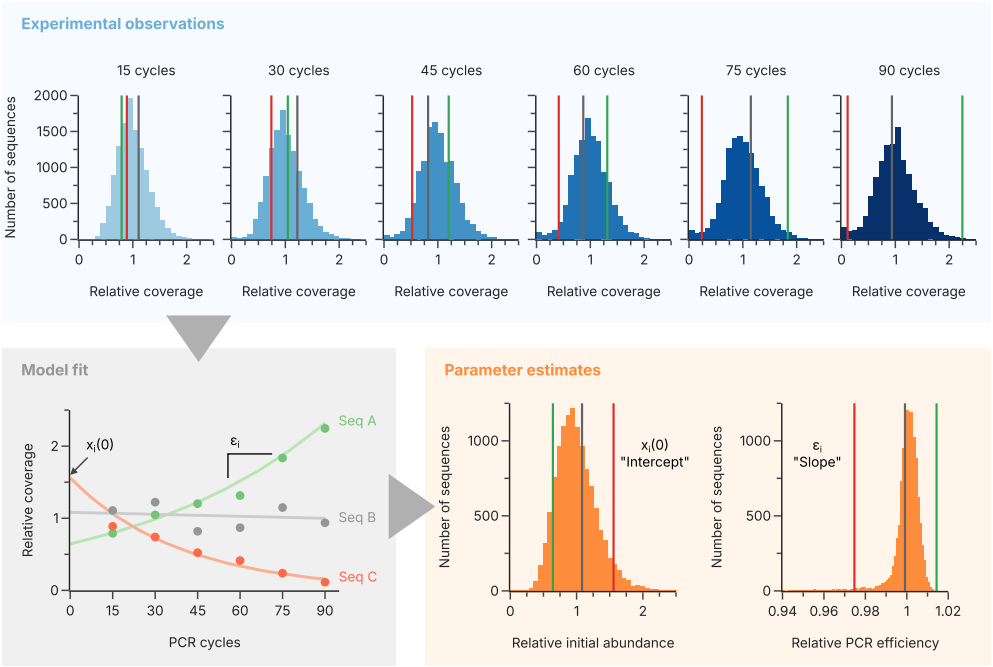

Fig. S1 Illustration of the model workflow.

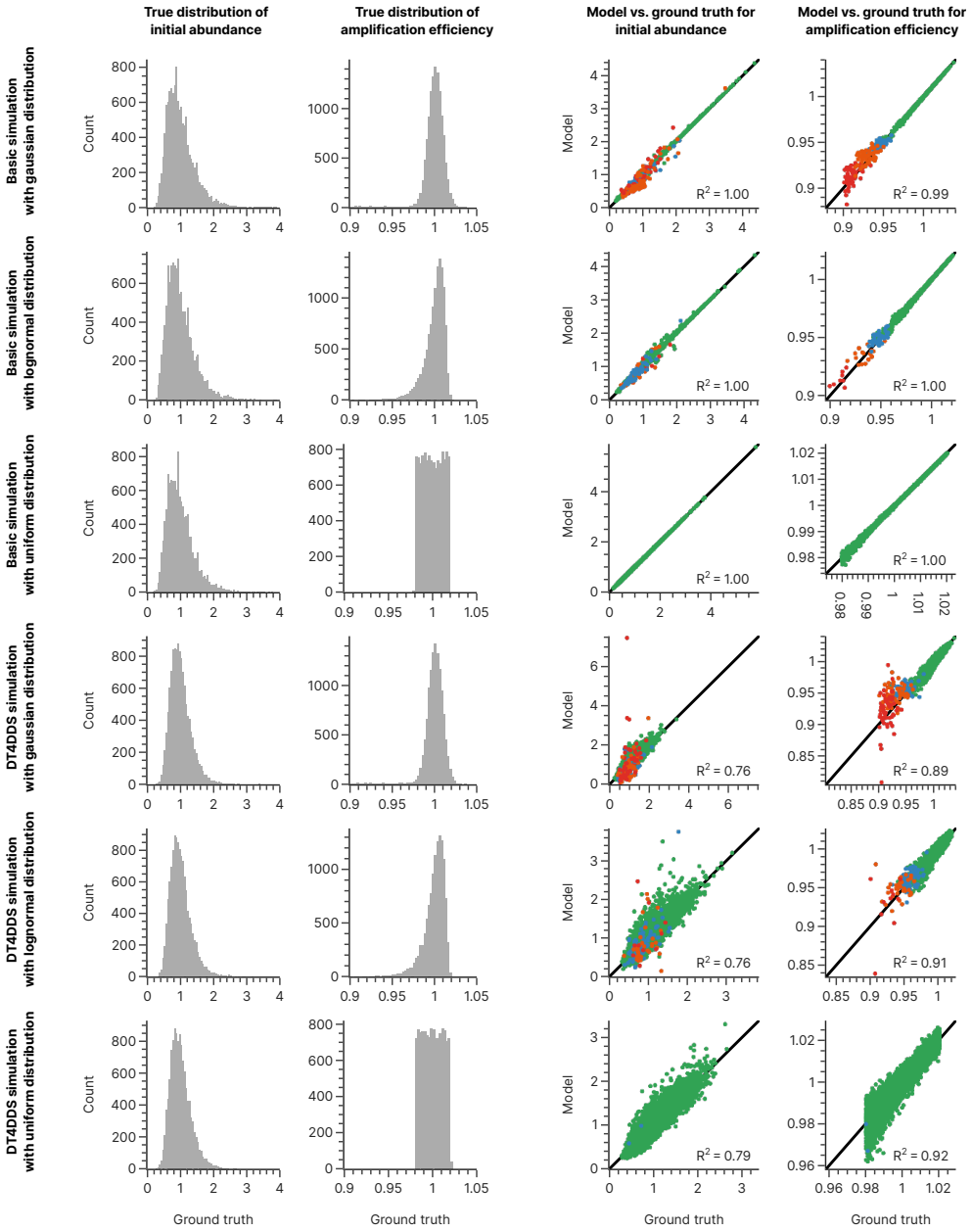

Fig. S2 Verification of the two-parameter PCR model using in-silico simulations of sequencing data with varying distributions of the true amplification efficiency and the true initial coverage bias. The simulated datasets also consist of six experimental endpoints at different cycle counts. The color of the points in the model vs. ground truth plot correspond to the frequency with which the sequence was not observed in the six experimental endpoints: never (i.e., always present, green), once (blue), twice (orange), and at least three times (red).

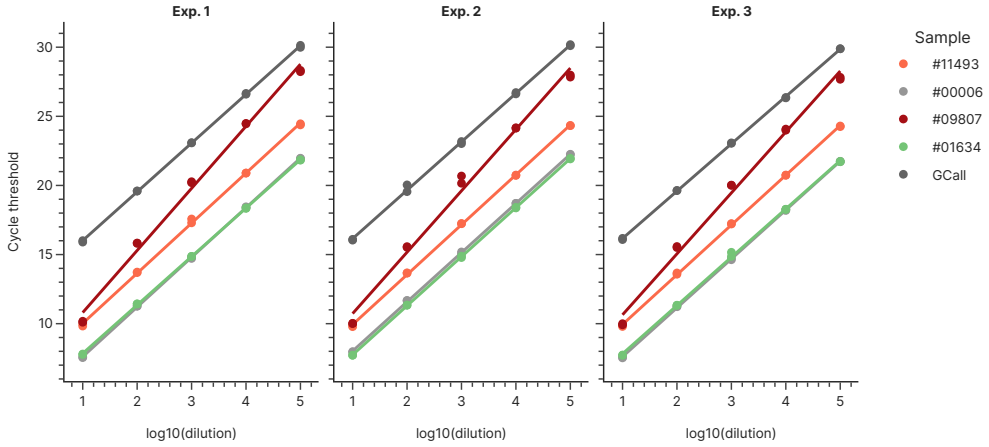

Fig. S3 Dilution curves for the qPCR-based efficiency verification of selected sequences and the full GCall pool.

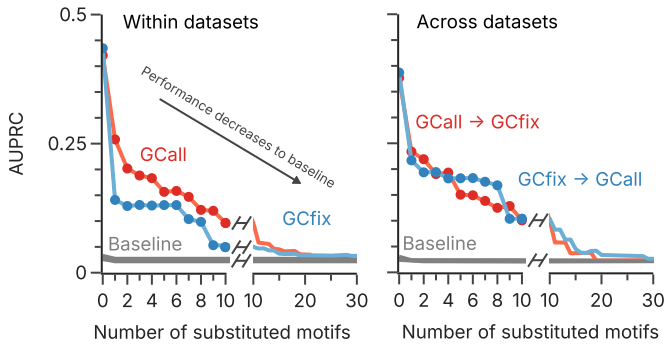

Fig. S4 AUPRC version of motif replacement figure.

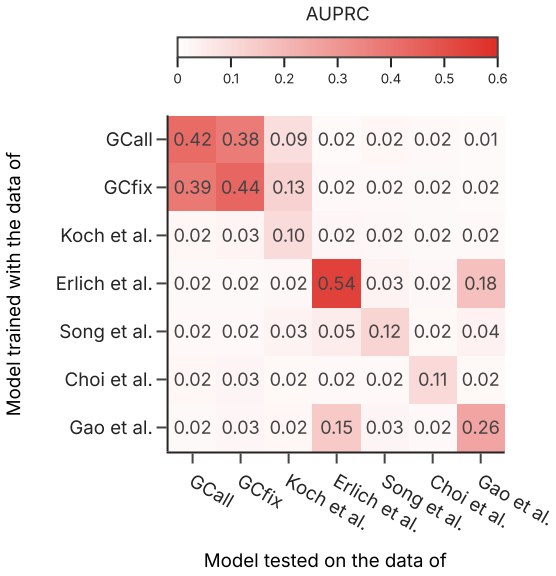

Fig. S5 AUPRC version of generalization figure.

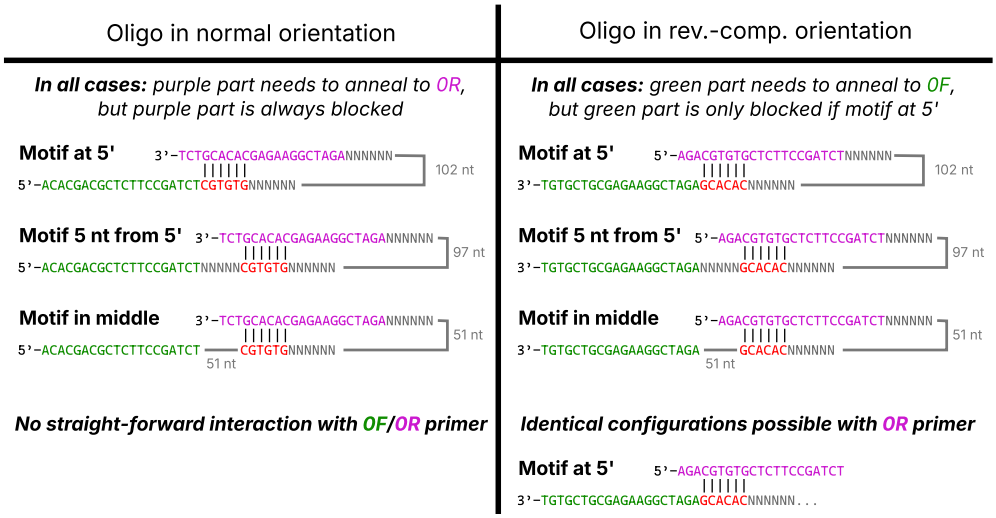

Fig. S6 Motif-adapter interactions for CGTGTG.

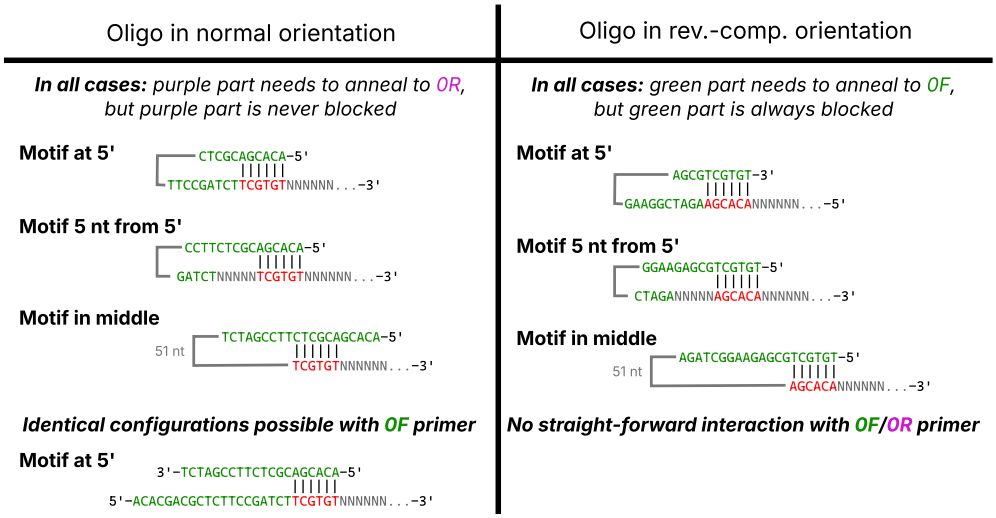

Fig. S7 Motif-adapter interactions for TCGTGT.

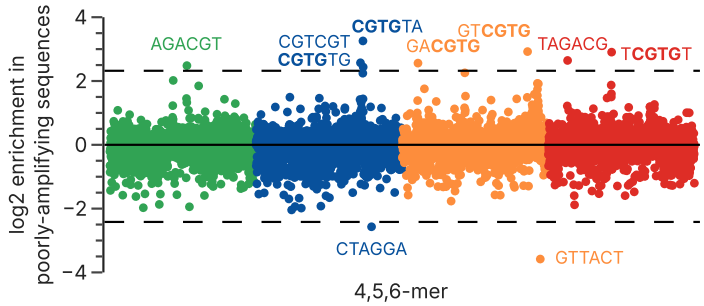

Fig. S8 K-mer frequency analysis for the combined GCall and GCfix datasets.

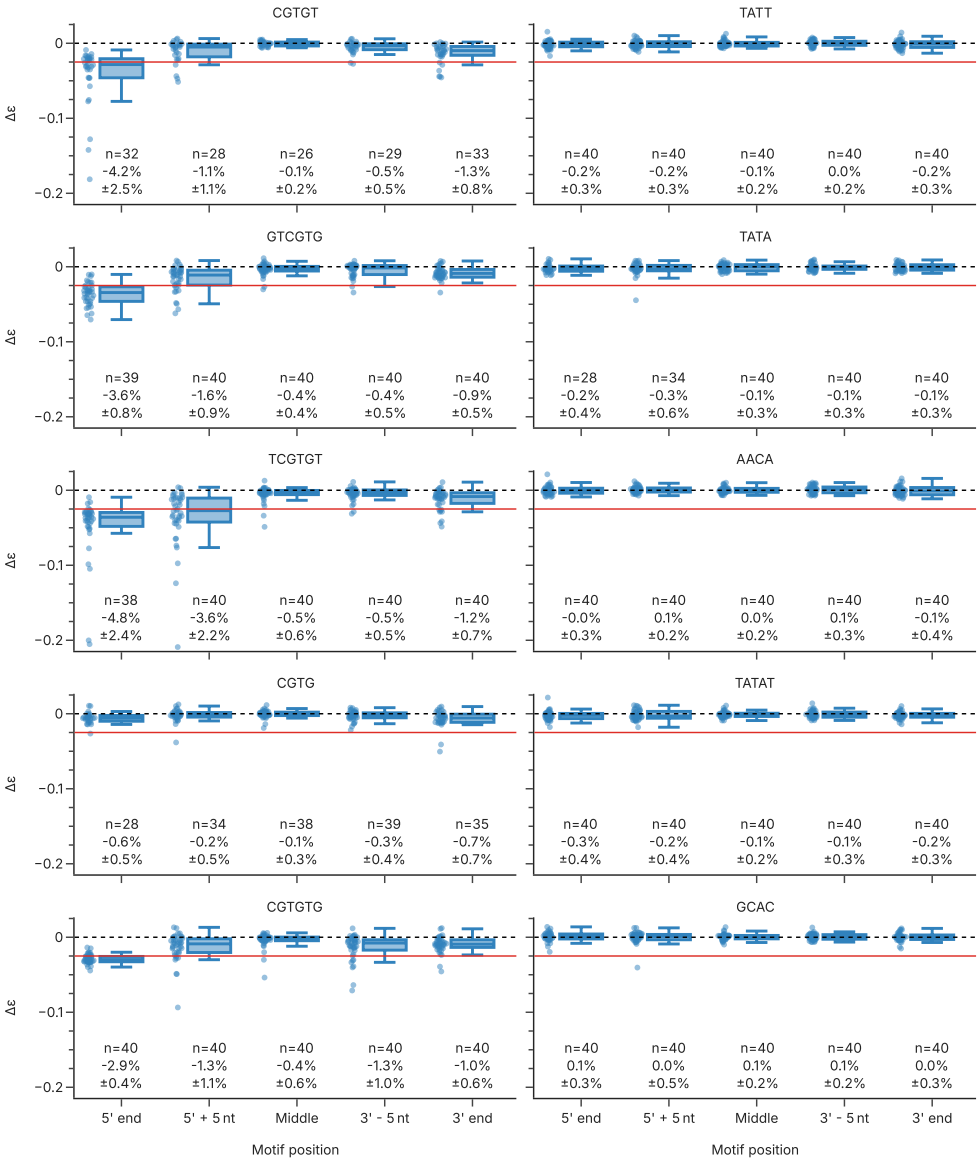

Fig. S9 GCall/fix.

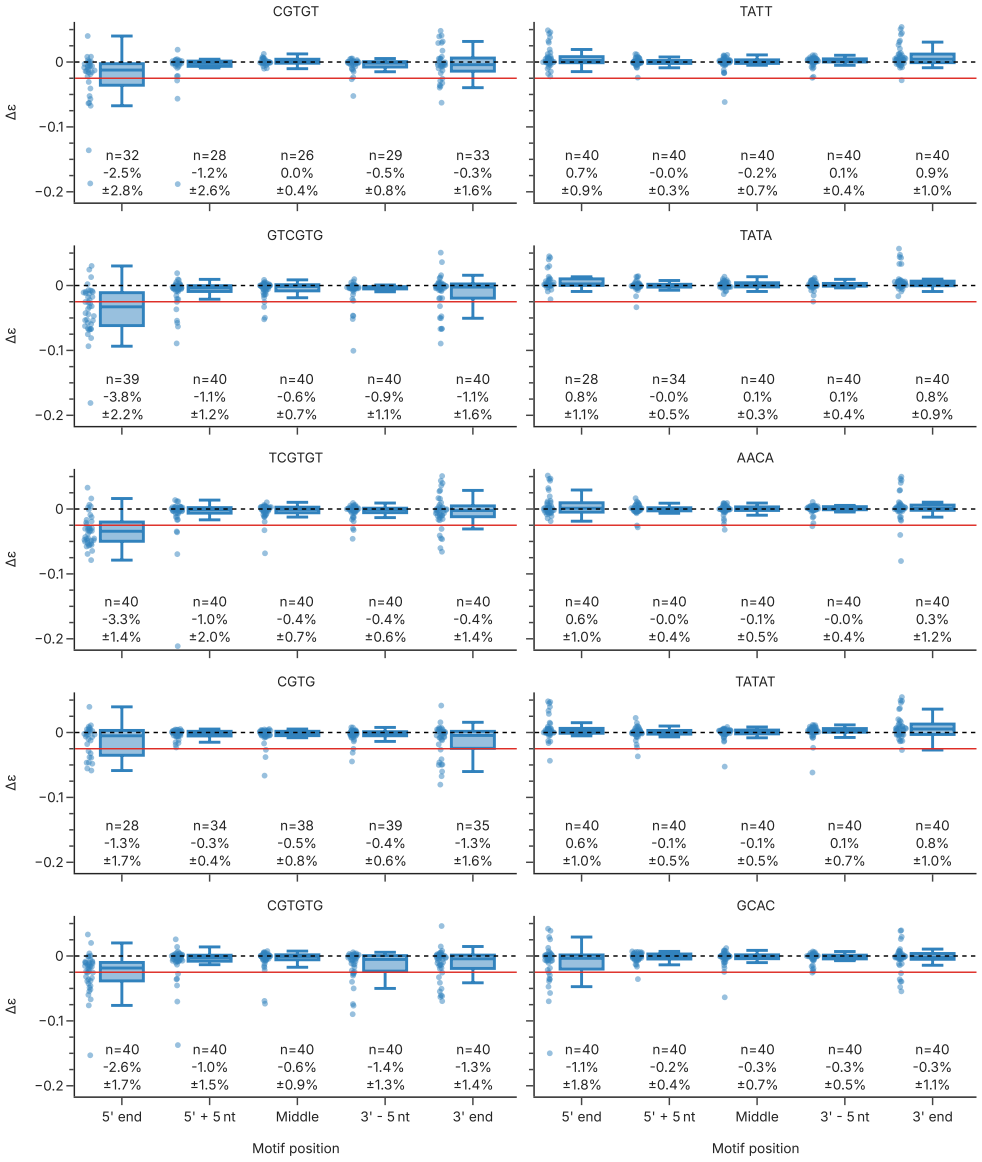

Fig. S10 Erlich.

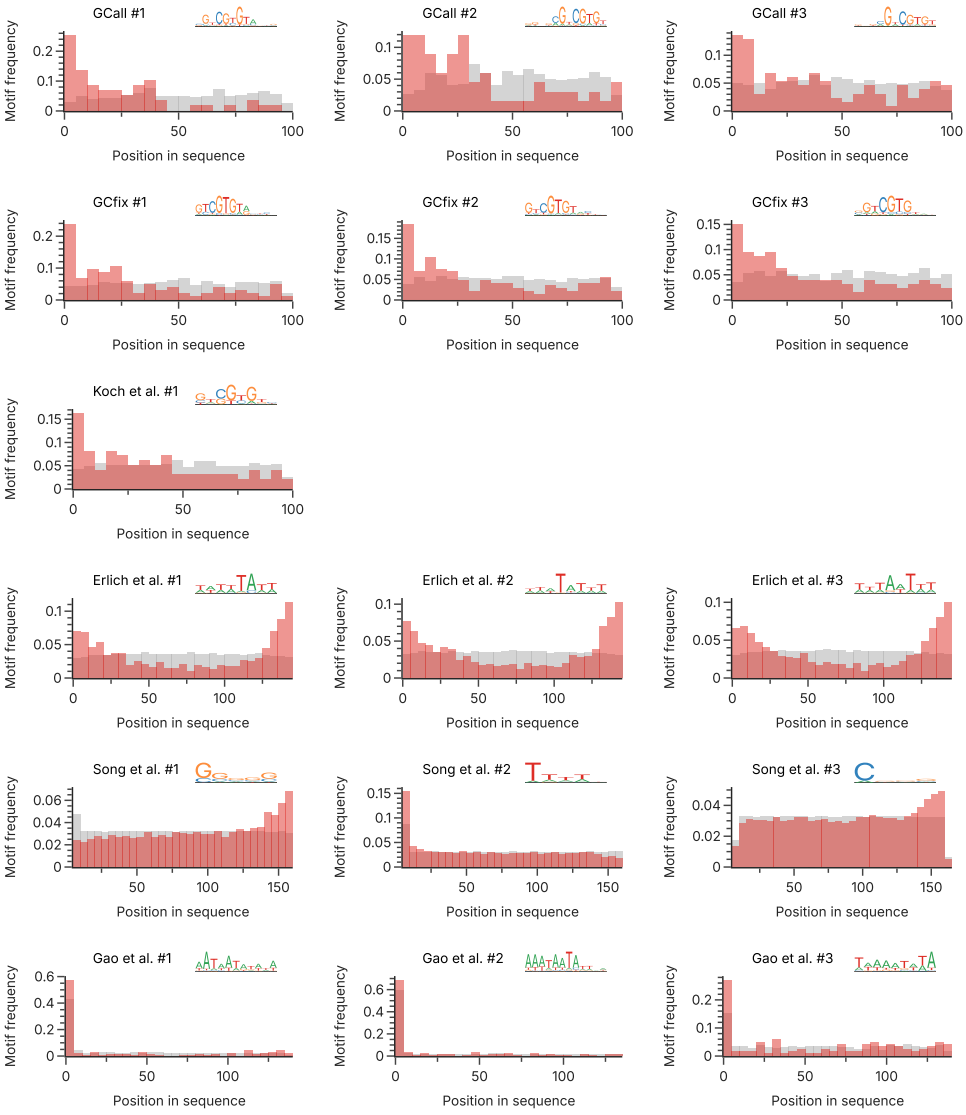

Fig. S11 Top 3 motifs and their position.

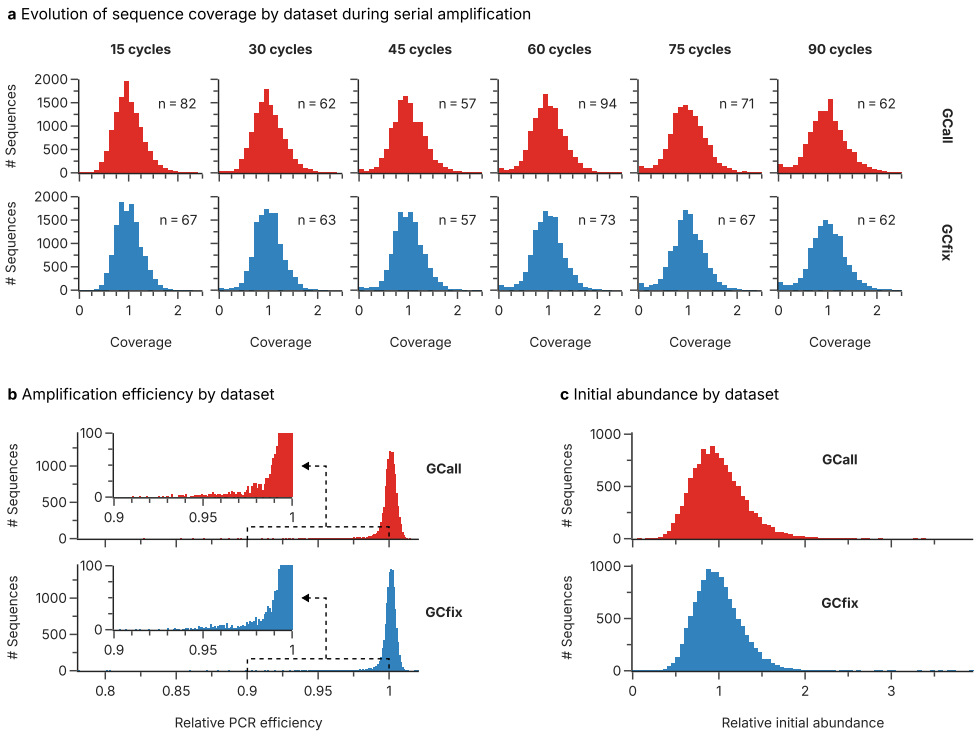

Fig. S12 GCall/GCfix pool full data.

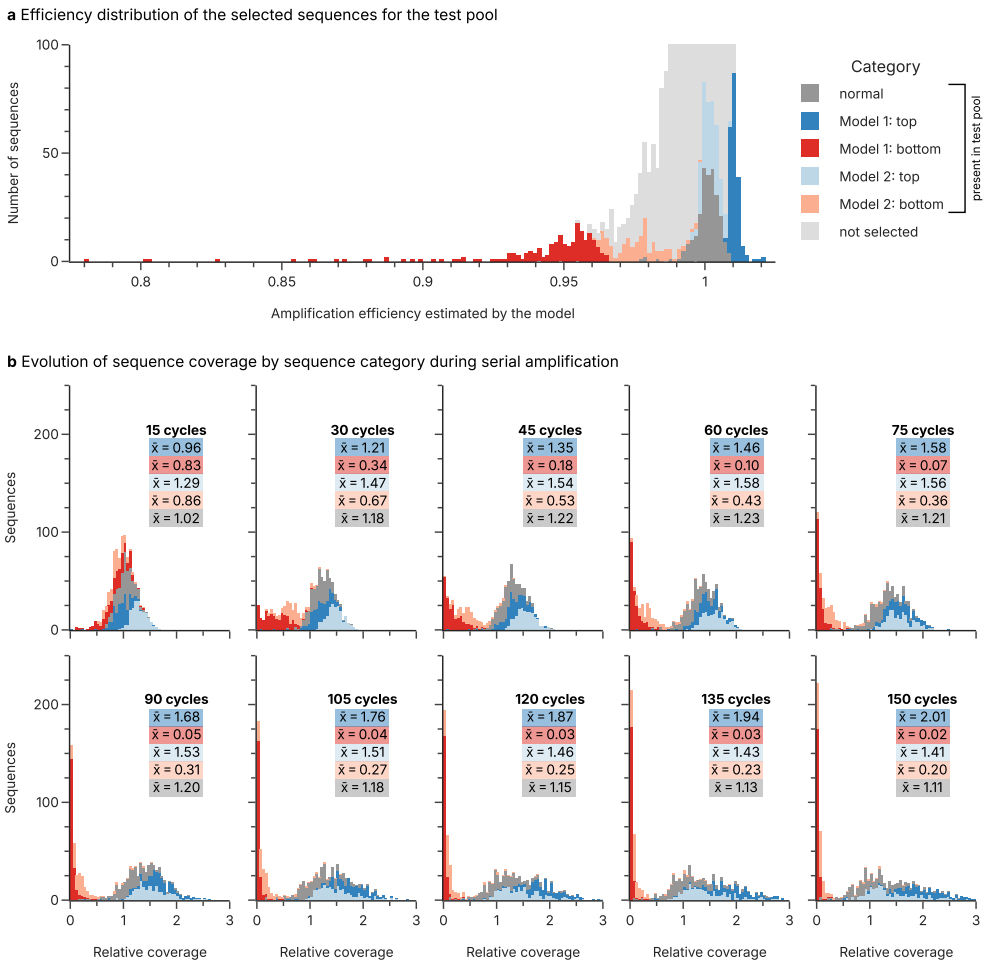

Fig. S13 Full composition and evolution of the test pool.

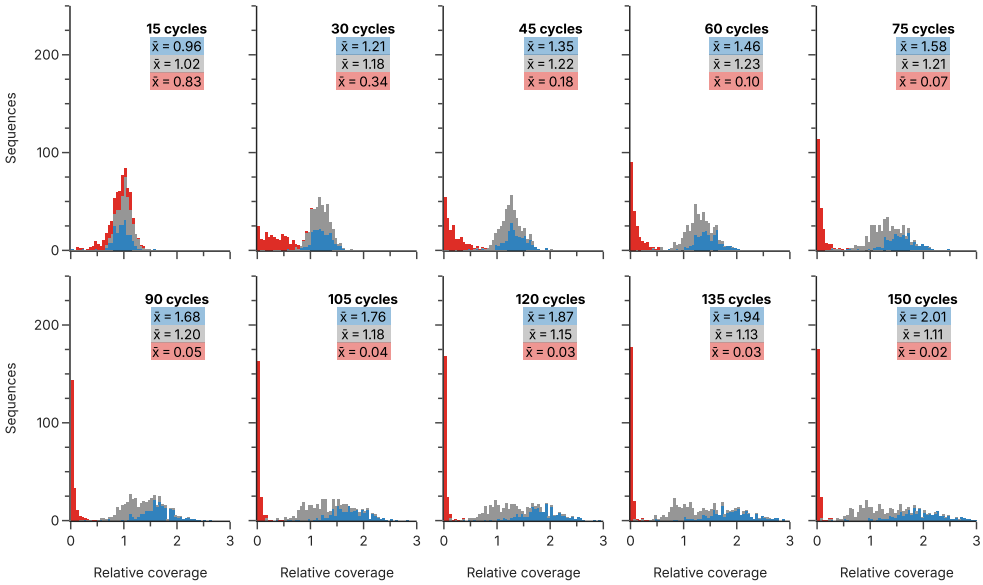

Fig. S14 Test pool full evolution.

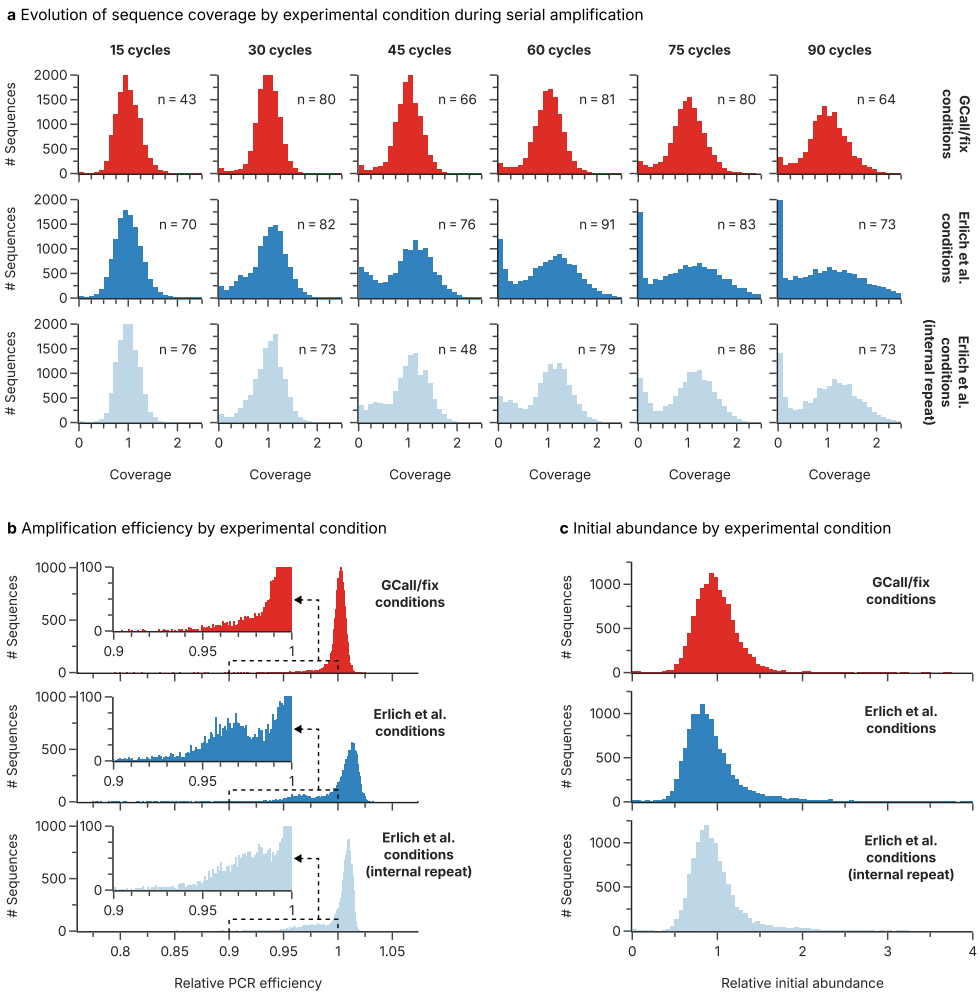

Fig. S15 Validation pool full data.

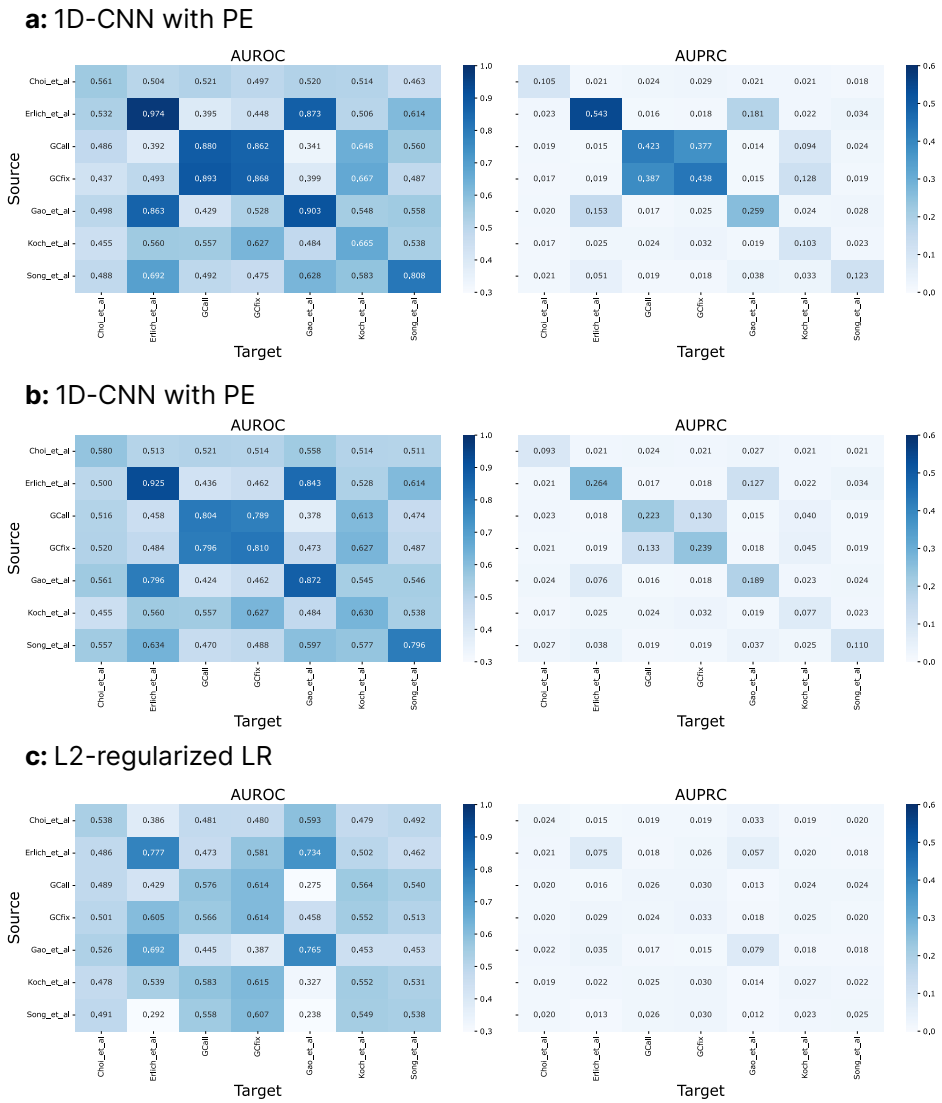

Fig. S16 Different ML models performance under the threshold of 2%.

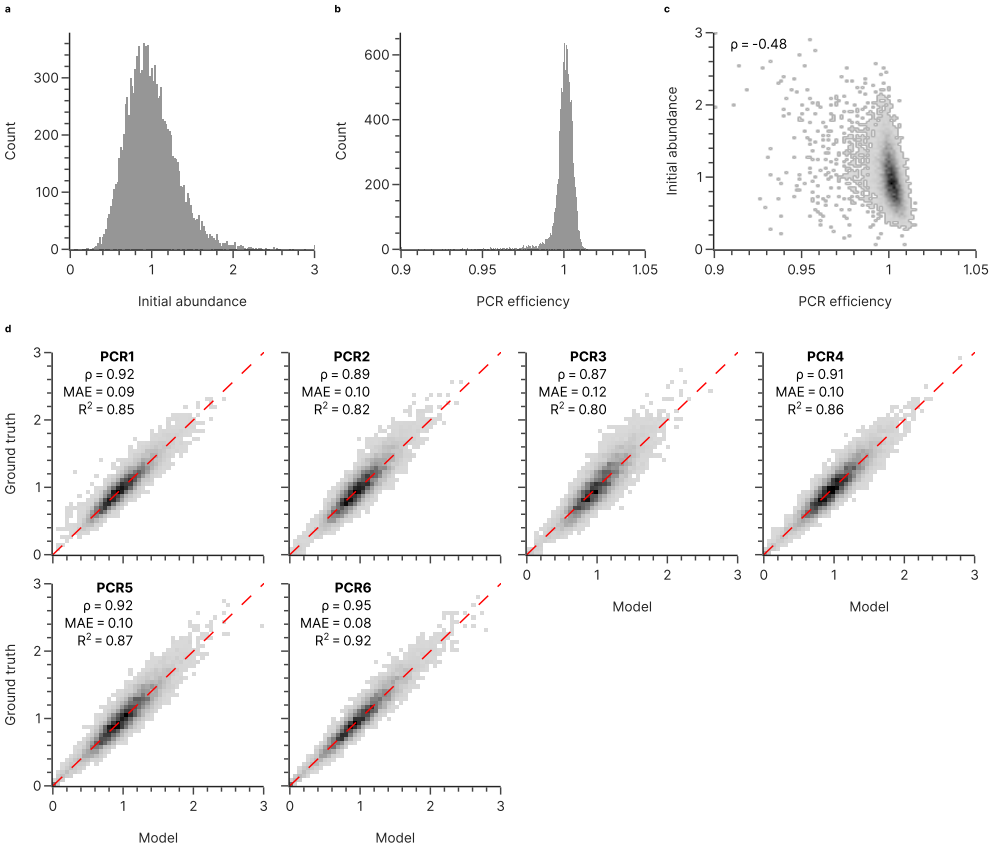

Fig. S17 GCall.

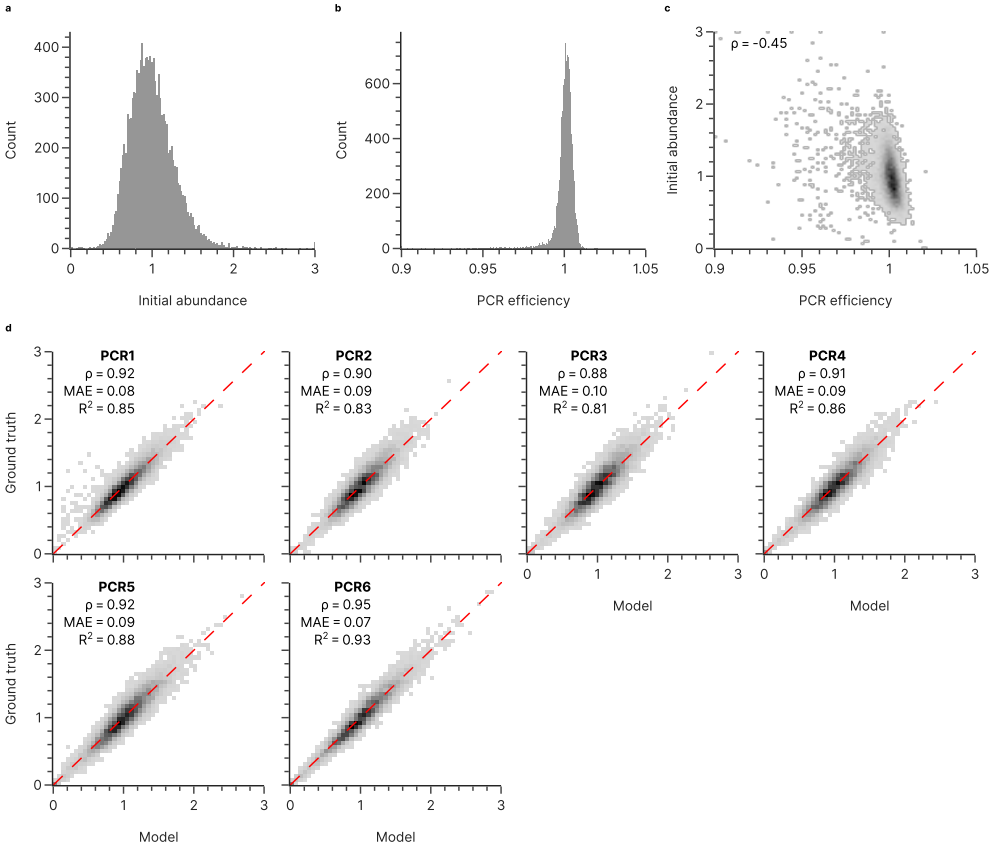

Fig. S18 GCfix.

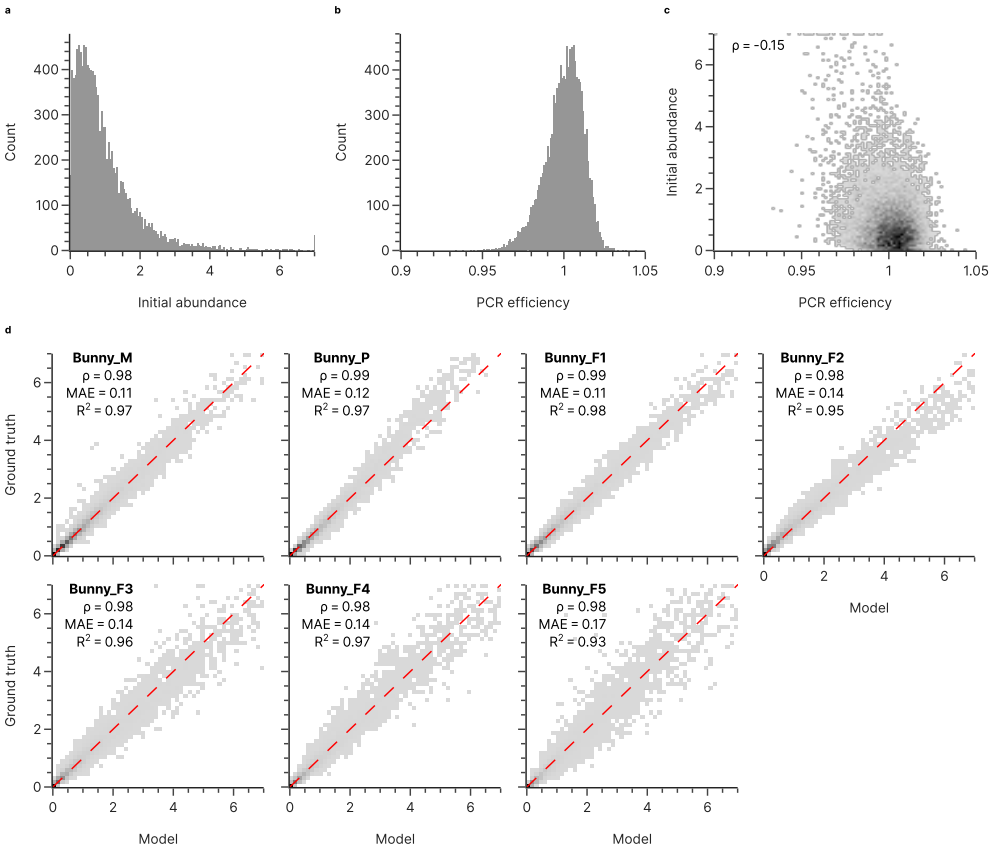

Fig. S19 Koch et al.

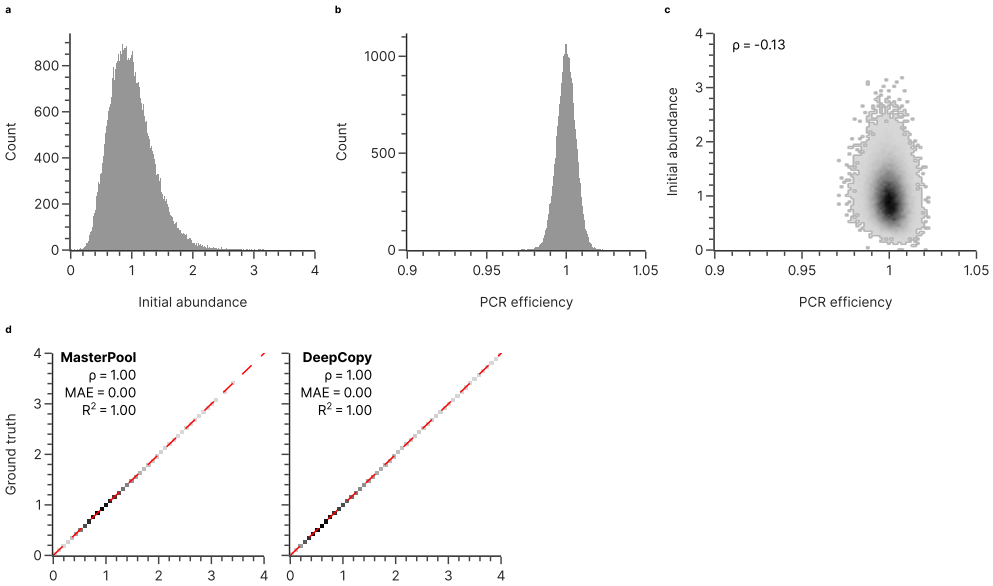

Fig. S20 Erlich et al.

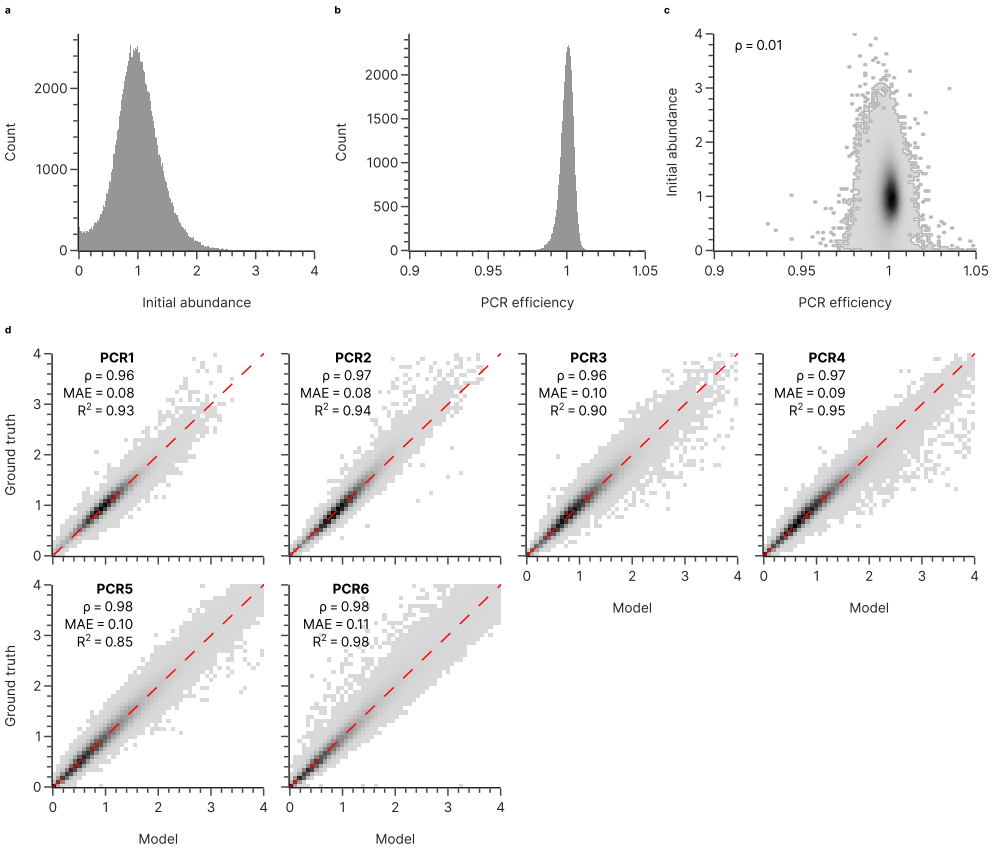

Fig. S21 Song et al.

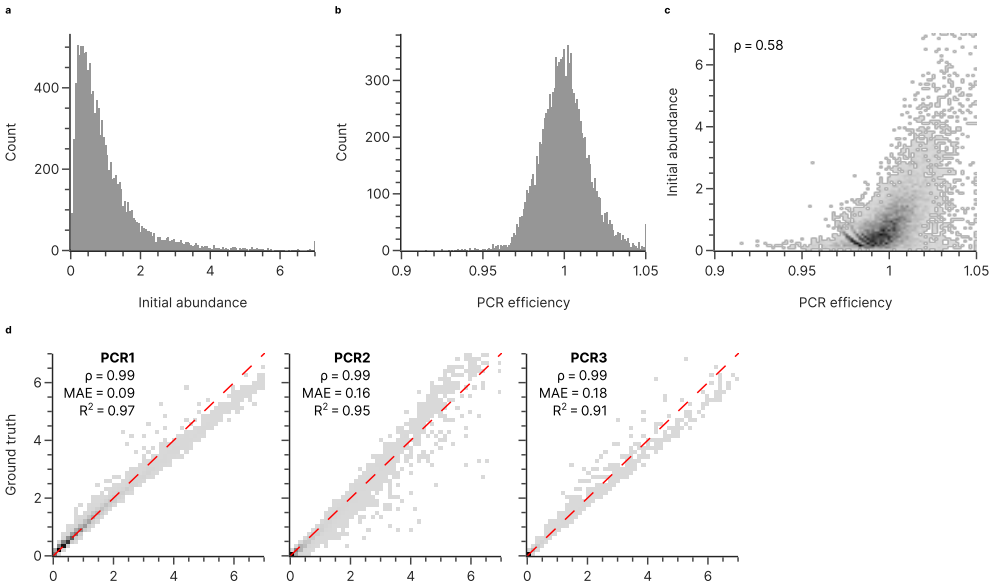

Fig. S22 Gao et al.

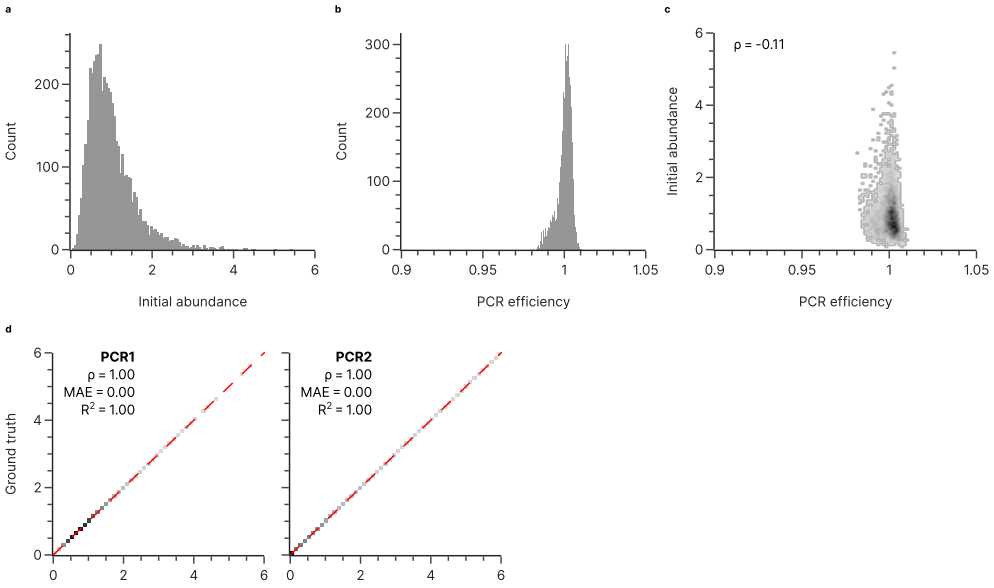

Fig. S23 Choi et al.

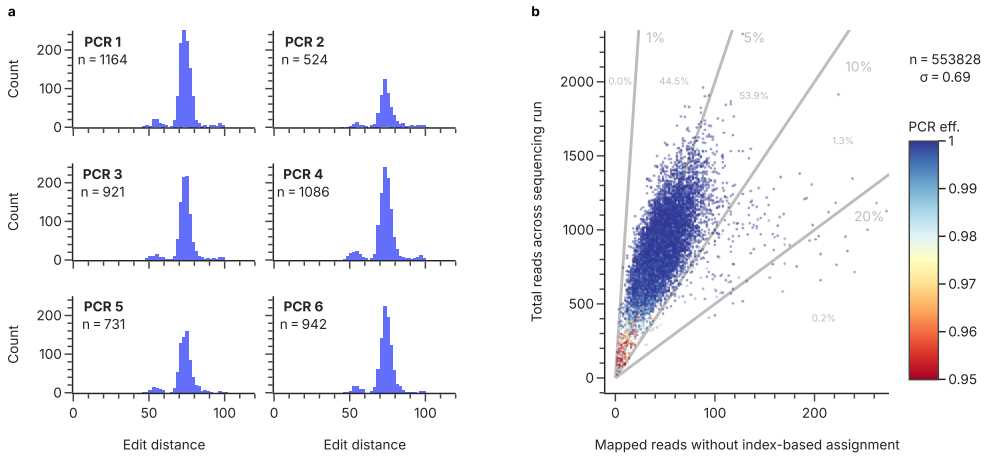

Fig. S24 Trash digging GCall.

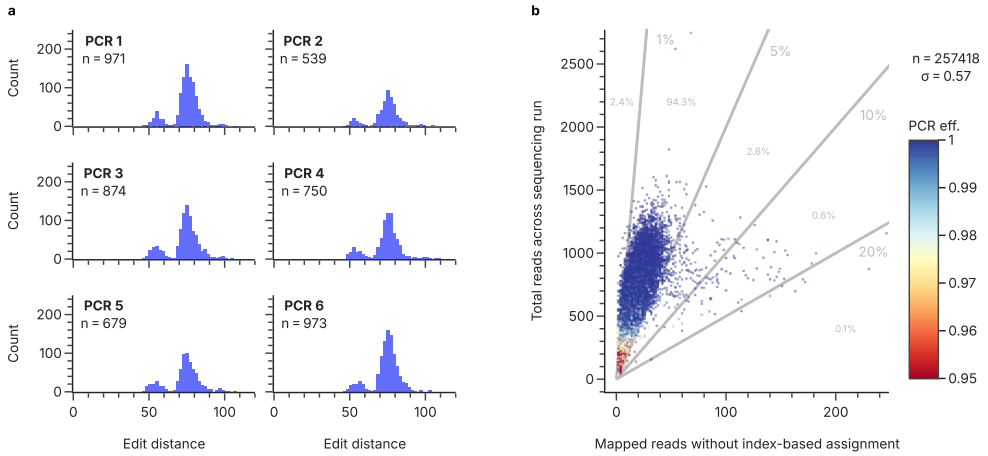

Fig. S25 Trash digging GCfix.

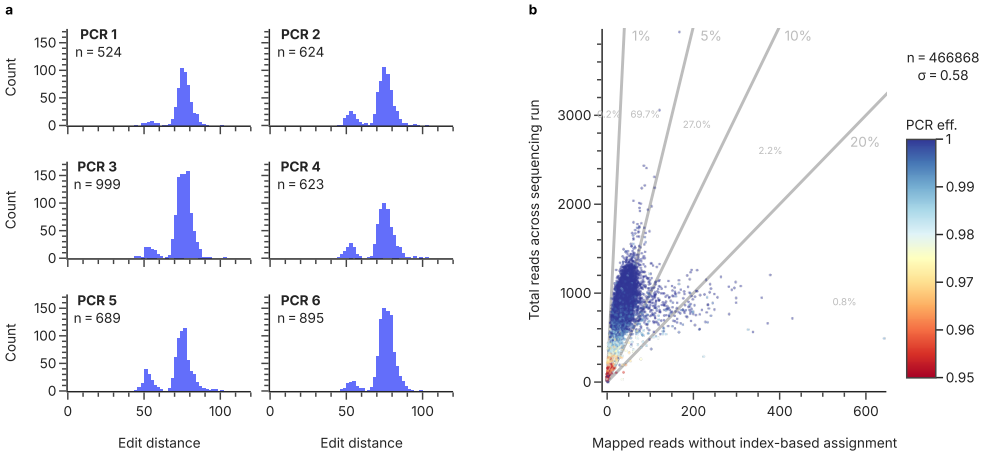

Fig. S26 Trash digging validation Taq.

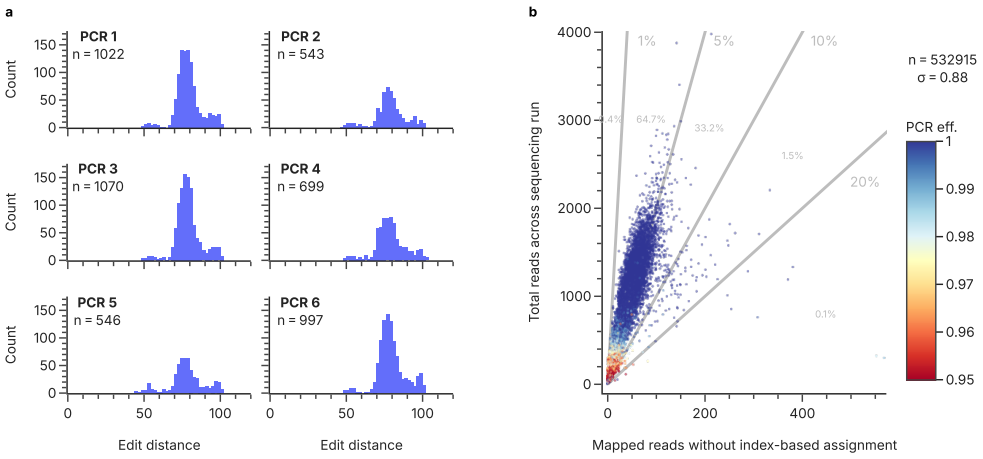

Fig. S27 Trash digging validation Q5.

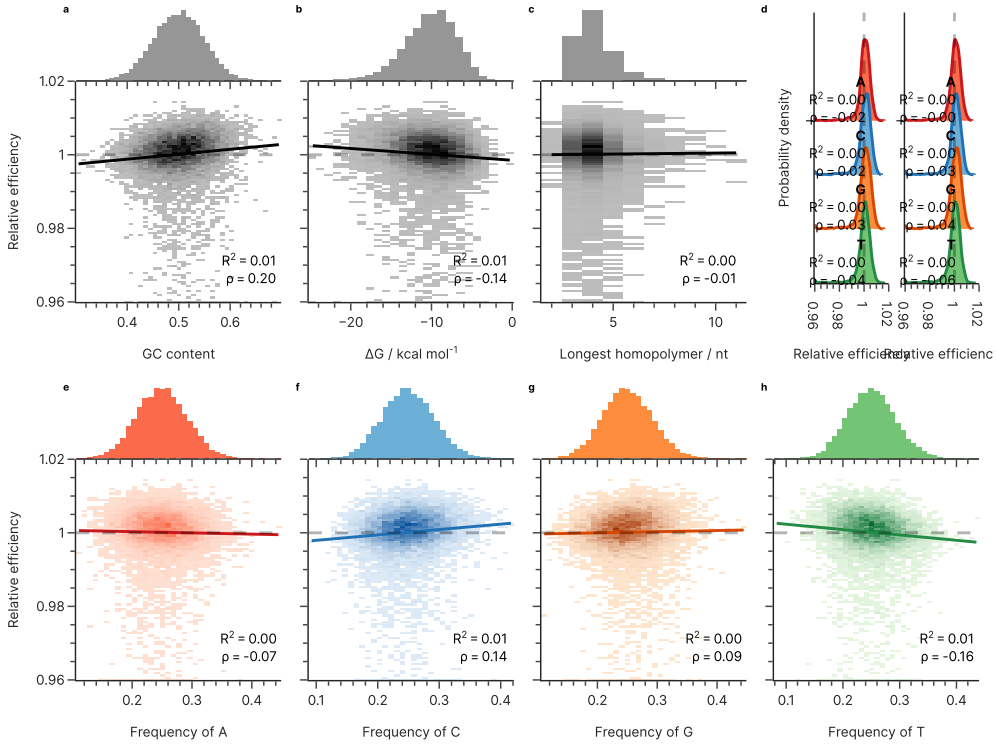

Fig. S28 GCall eff.

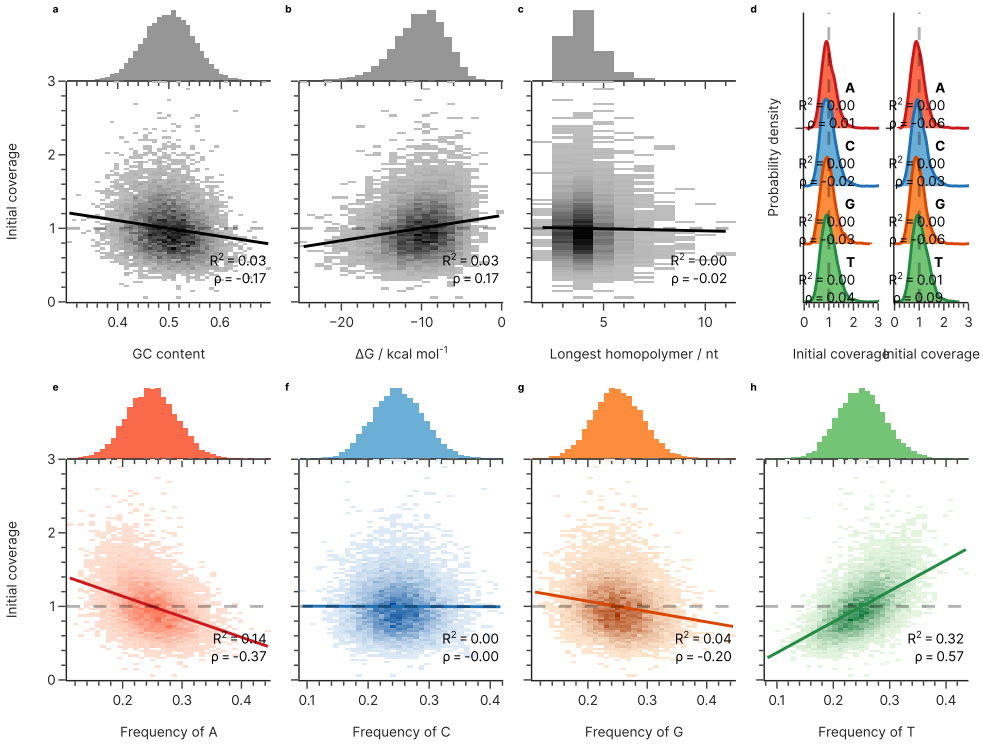

Fig. S29 GCall x0.

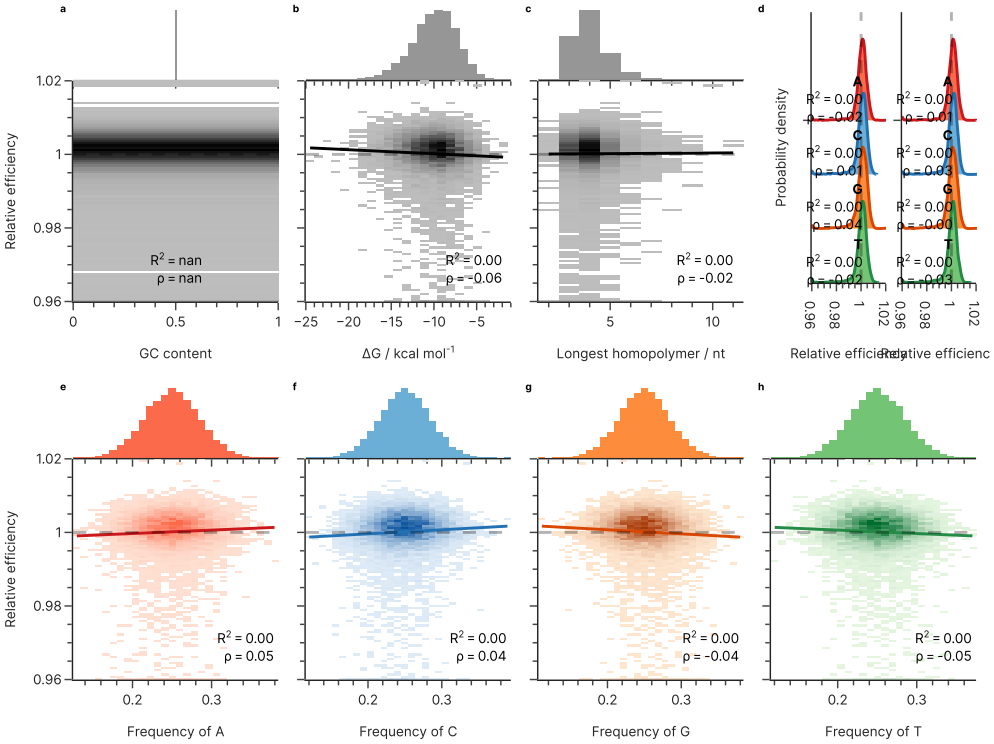

Fig. S30 GCfix eff.

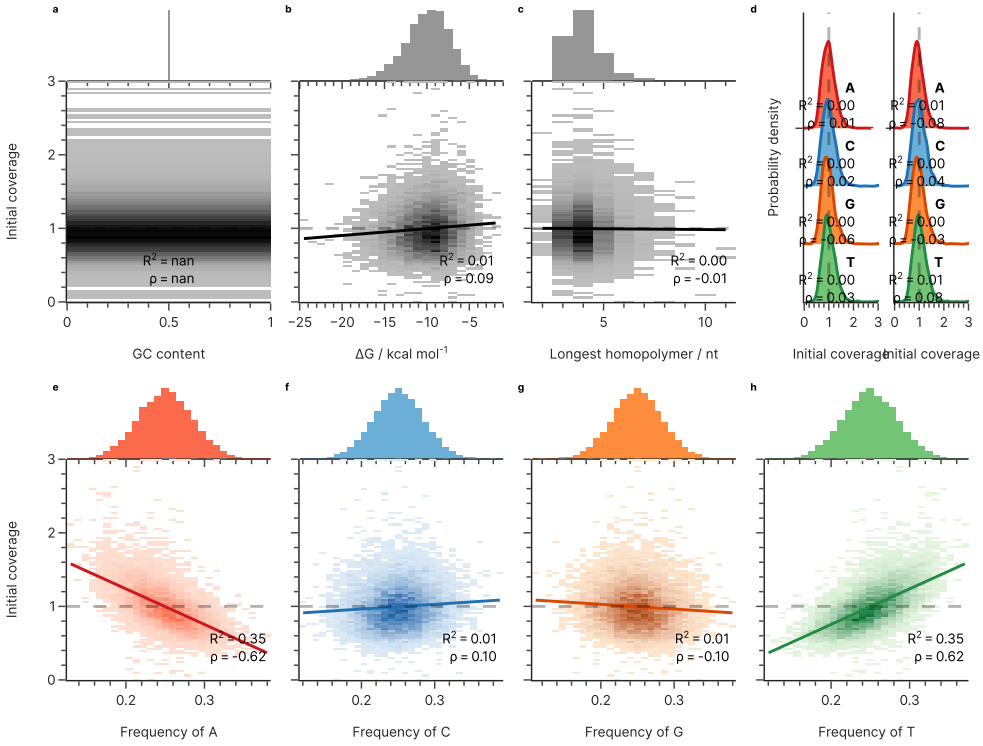

Fig. S31 GCfix x0.

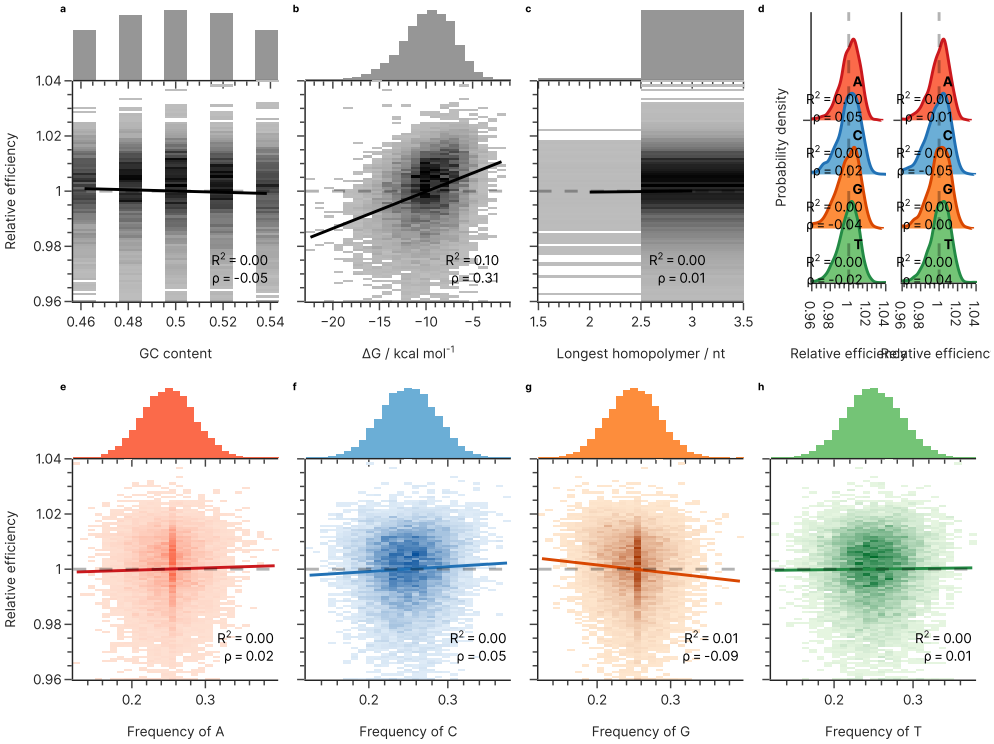

Fig. S32 Koch eff.

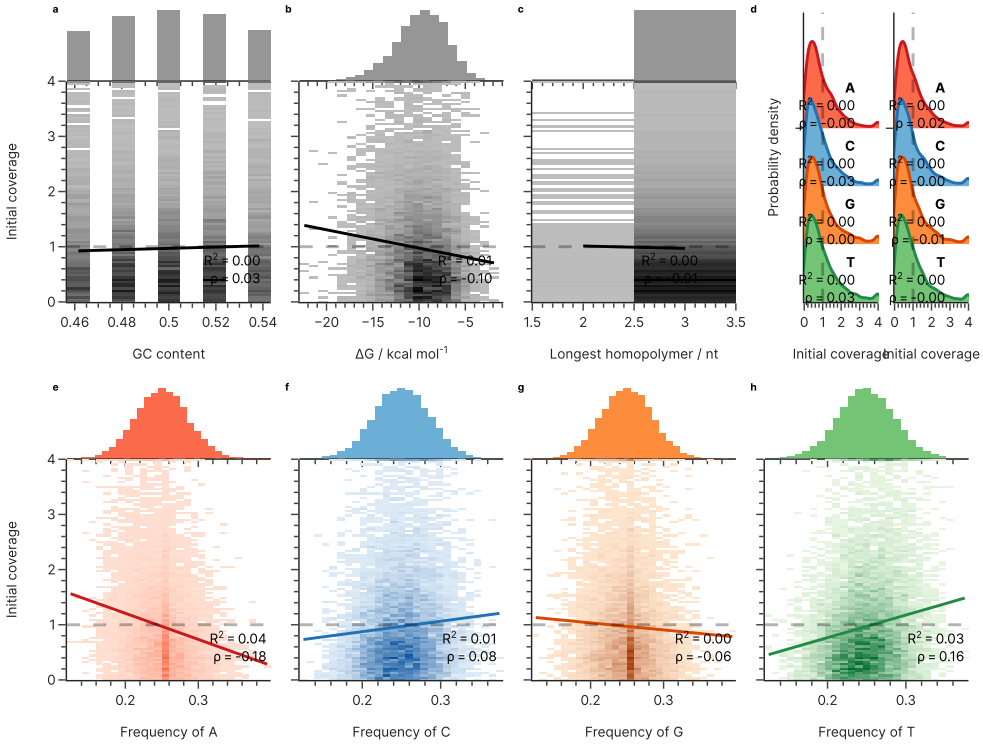

Fig. S33 Koch x0.

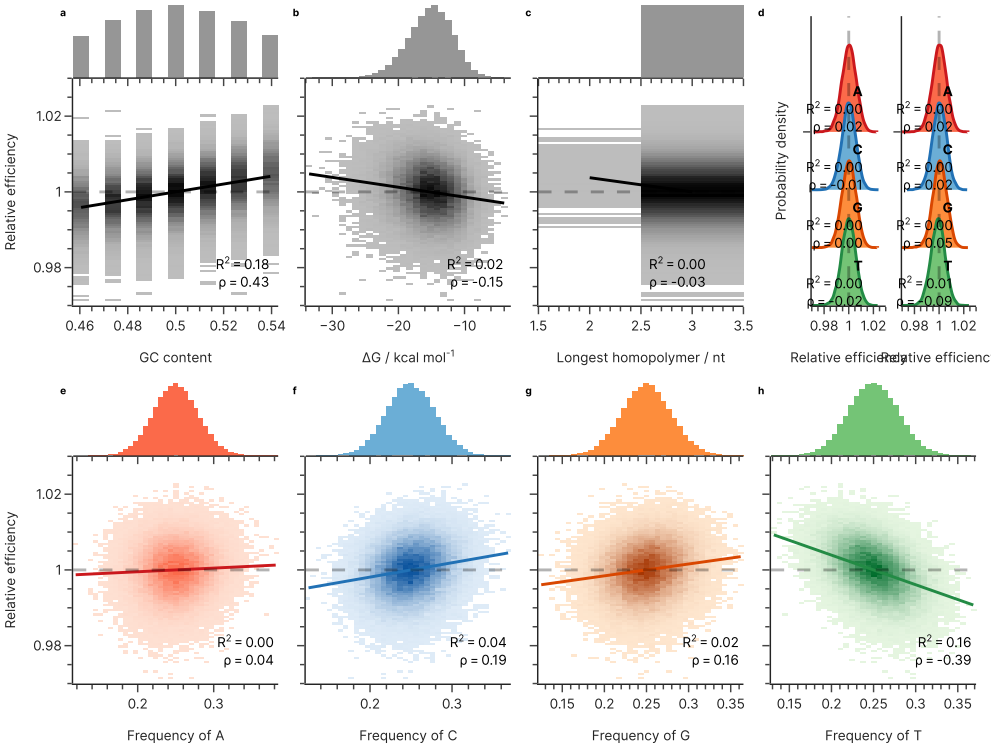

Fig. S34 Erlich eff.

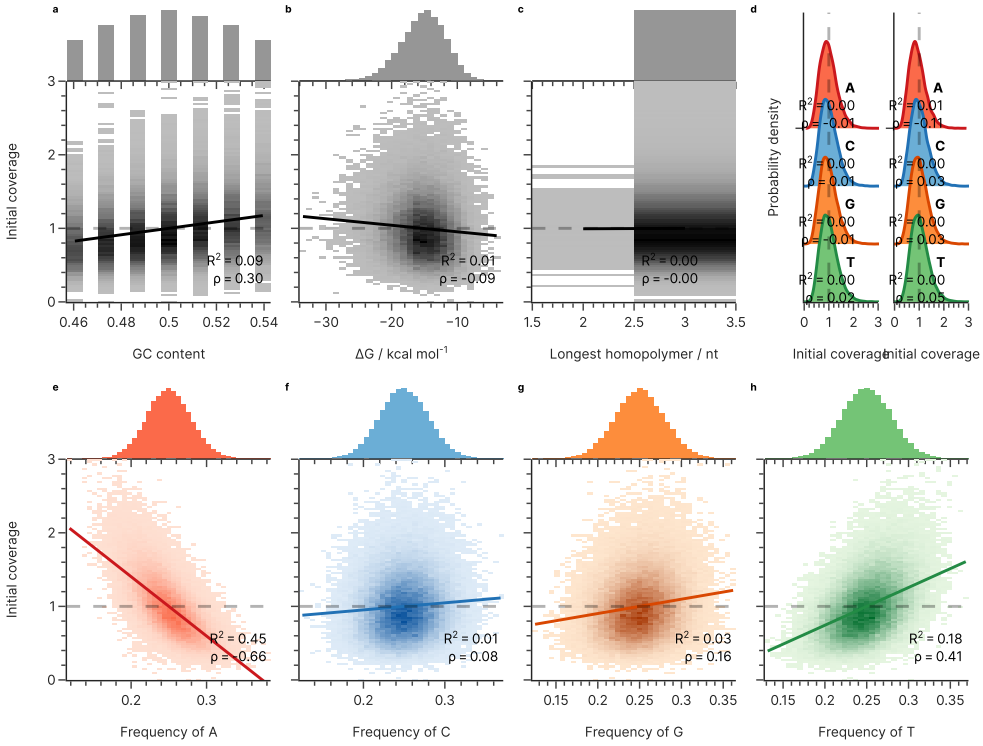

Fig. S35 Erlich x0.

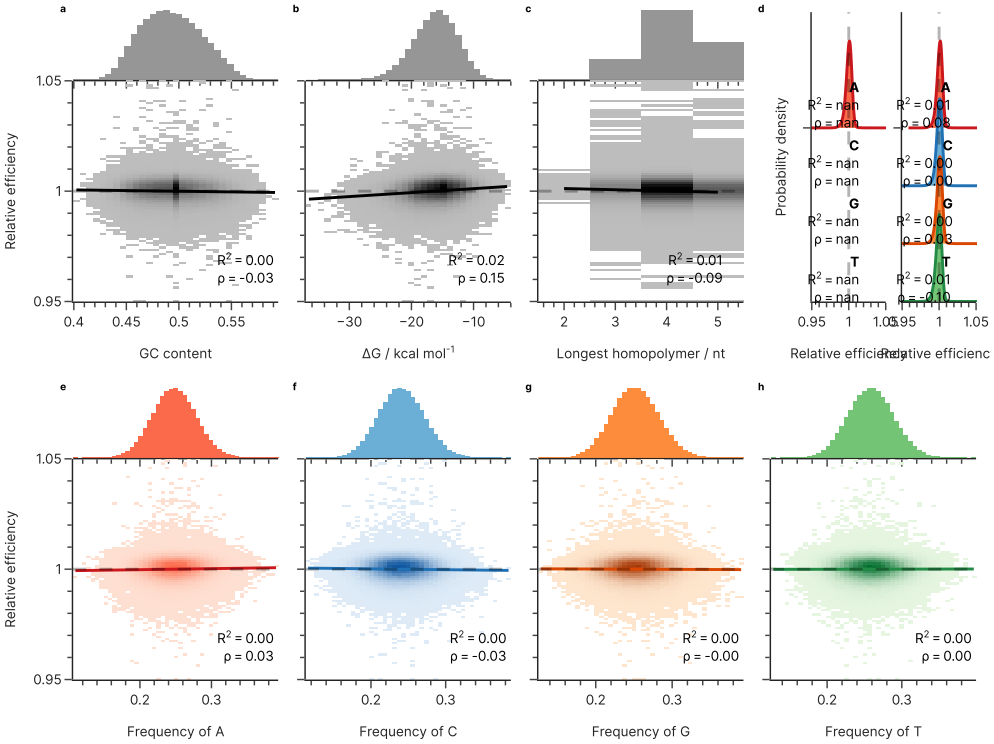

Fig. S36 Song eff.

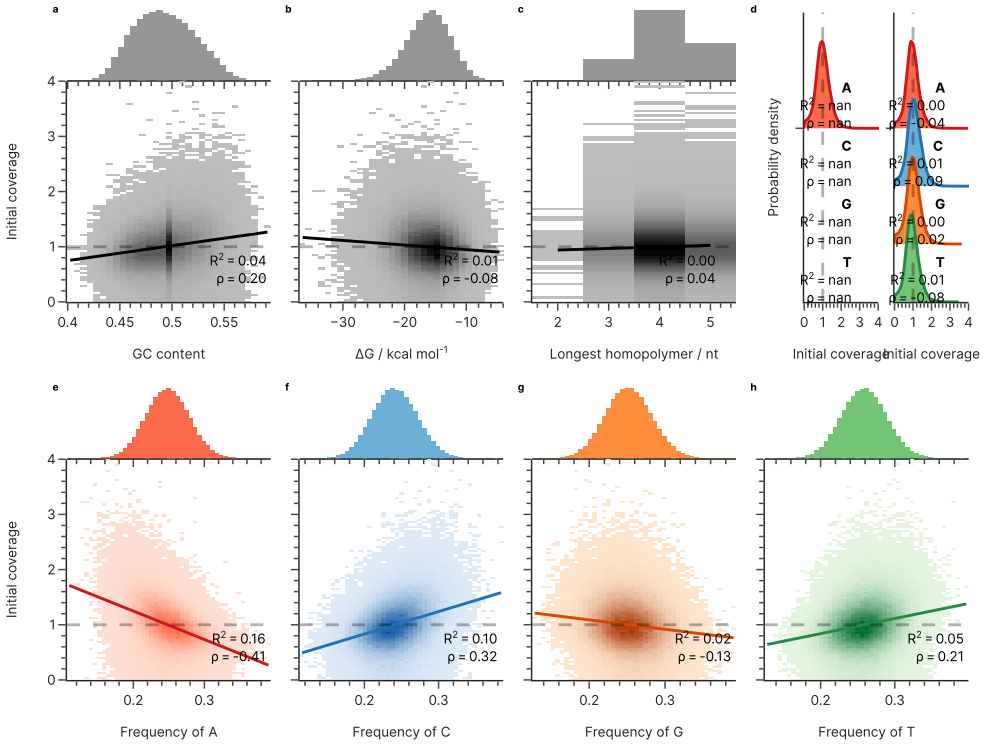

Fig. S37 Song x0.

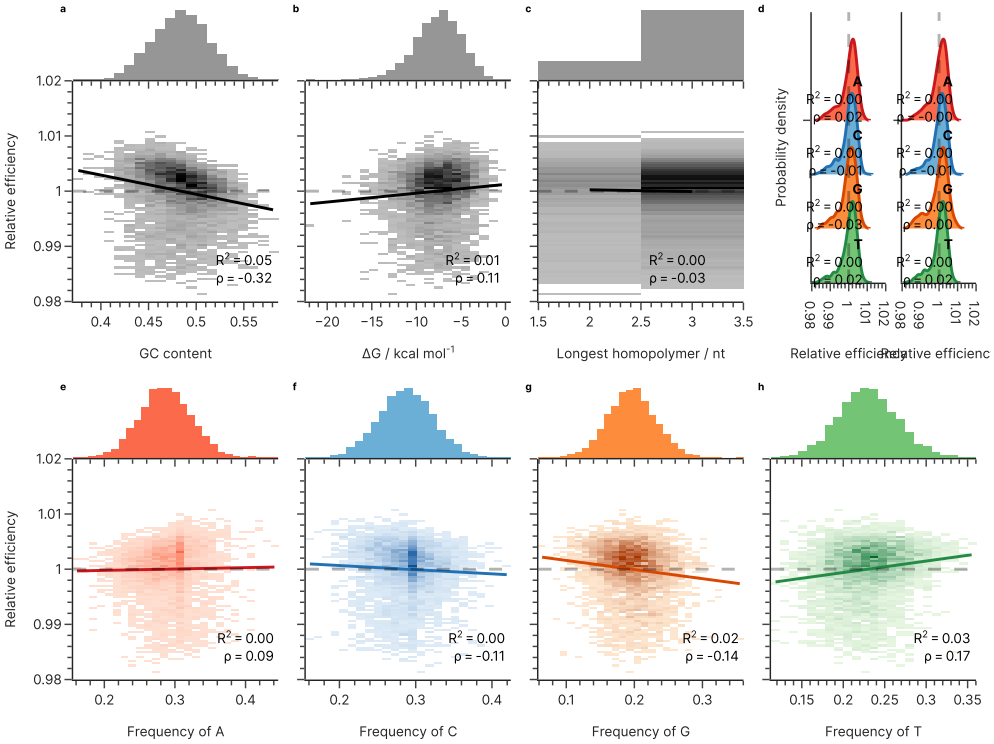

Fig. S38 Choi eff.

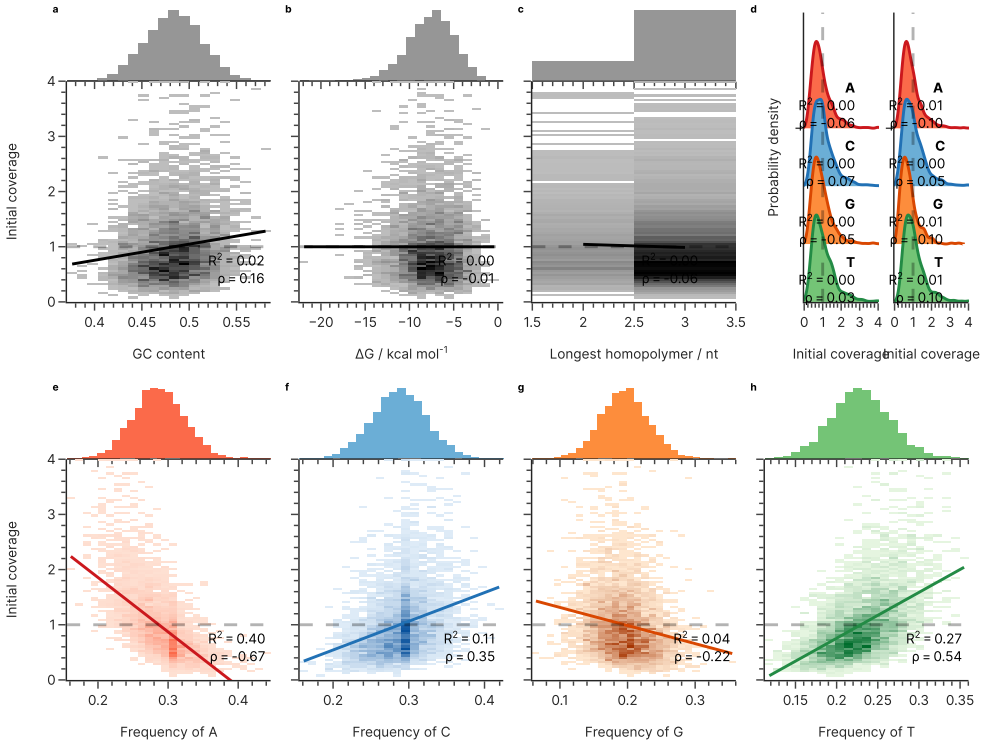

Fig. S39 Choi x0.

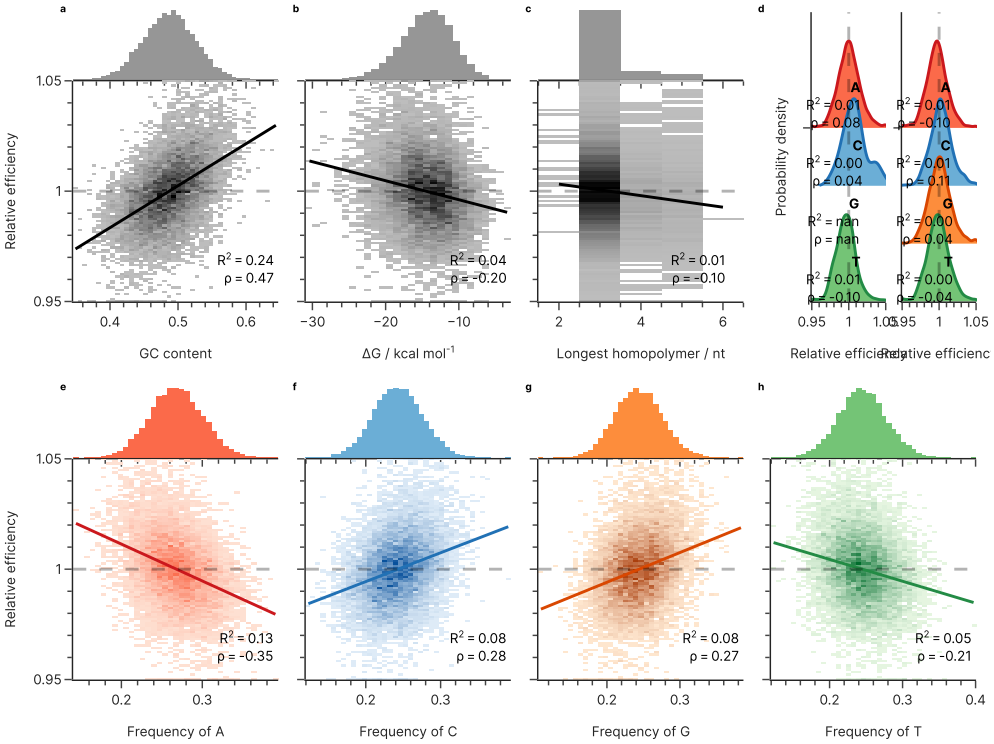

Fig. S40 Gao eff.

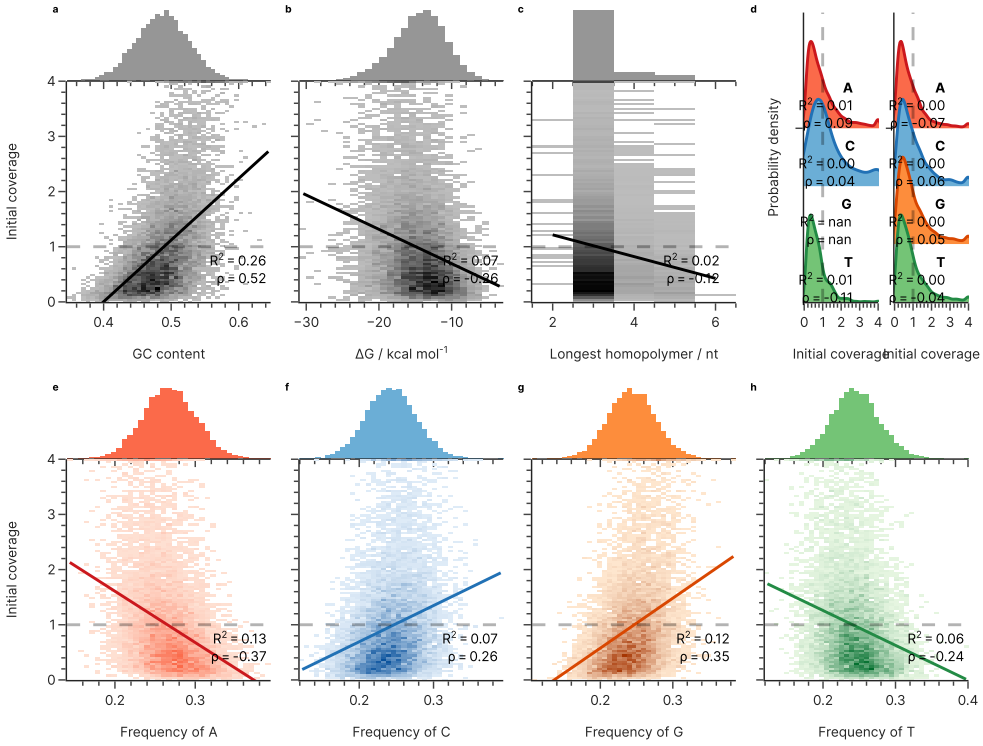

Fig. S41 Gao x0.

Supplementary Tables

Name	Description	Sequence (5'-3')
0F	Forward primer	ACACGACGCTCTTCCGATCT
0R	Reverse primer	AGACGTGTGCTCTTCCGATCT
2FUF	Forward sequencing primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT
2RIF-GM5	Indexed reverse sequencing primer	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGT TCAGACGTGTGCTCTTCCGATCT
2RIF-GM7	Indexed reverse sequencing primer	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGT TCAGACGTGTGCTCTTCCGATCT
2RIF-GM8	Indexed reverse sequencing primer	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGT TCAGACGTGTGCTCTTCCGATCT
2RIF-GM10	Indexed reverse sequencing primer	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGT TCAGACGTGTGCTCTTCCGATCT
2RIF-GM11	Indexed reverse sequencing primer	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGT TCAGACGTGTGCTCTTCCGATCT
2RIF-GM12	Indexed reverse sequencing primer	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGT TCAGACGTGTGCTCTTCCGATCT
2RIF-GM17	Indexed reverse sequencing primer	CAAGCAGAAGACGGCATACGAGATCTCTACGTGACTGGAGT TCAGACGTGTGCTCTTCCGATCT
#11493	qPCR test sequence	ACACGACGCTCTTCCGATCTCGTGTATAGGCTGACTGTTAT GTTCGTGCAGCACGCTGCATGGCTGATCGTATG-TACTTGGG CACTCTAAAGATTCAAGGCTAAGGAAGAAG-GATAACGCTAC CCAACAGATCGGAAGAGCACACGTCT
#00006	qPCR test sequence	ACACGACGCTCTTCCGATCTTATGTACTCTTGCACGCTTGG CGGGTGTTAAACTTCTGGTTCCGACCTAGCGCTCGTGC-CTT CCGGTGAATGTAGGTTGCCGAATGACACATCCATGCC-CTAT TCAGCAGATCGGAAGAGCACACGTCT
#09807	qPCR test sequence	ACACGACGCTCTTCCGATCTGAGTTCAAGCGCCGTGTAGCC TGATCTGGTCCTTATACTAGCCTGTTCACAAAGATATAGAT CTGTATAAACAGCGCCATTCAAGCTGAGAGAGAGGGCCCAG GGGCCAGATCGGAAGAGCACACGTCT
#01634	qPCR test sequence	ACACGACGCTCTTCCGATCTTTGCTTTGCTGTTGTGCGGTA CACCCCAACGTGATCTTTAGTCCTTGGAAAGTCCACCACAT CTCTTGTAGGGCTAGCTTATTCTAGAACAACAGCAAGGGCT GCCACAGATCGGAAGAGCACACGTCT

Table S1 Sequences used in the study.

Dataset	GCall	GCfix	Koch et al.	Erlich et al.	Song et al.	Choi et al.	Gao et al.
Source			PRJEB35217	PRJEB19305/7		PRJNA555140	pers. comm.
Experimental design							
#Experiments	6	6	7	2	6	2	3
Cycle counts	15-90	15-90	44-119	10, 100	30-180	17, 340	10, 50, 100
Synthesis provider	Twist	Twist	CustomArray	Twist	Twist	CustomArray	Twist
Polymerase	KAPA FAST	KAPA FAST	KAPA FAST	Q5 HiFi	-	KAPA HiFi	Q5 HiFi
Pool design							
#Sequences	12000	12000	12000	72000	210000	5173	11520
Seq. length	108	108	104	152	164	93	146
Seq. constraints	-	GC	GC, HP	GC, HP	GC, HP	-	motifs
Seq. randomized	Yes	Yes	No	No	No	No	No
Mapping							
#Reads mapped							
%Reads mapped							
Pre-processing							
Mean coverage							
#Seq. removed	2	6	35	0	53	408	15
%Seq. removed	0.017%	0.05%	0.29%	0%	0.025%	7.9%	0.13%
PCR model							
#observations							
#parameters							
R^2_{baseline}	0.750	0.758	0.868	0.755	0.424	0.684	0.434
R^2_{model}	0.859	0.868	0.957	1.000	0.967	1.000	0.911
df_1	11998	11994	11965	72000	209947	4765	11505
df_2	47992	47976	59825	0	839788	0	11505
$F(df_1, df_2)$	3.08	3.33	10.2	-	65.8	-	5.38

Table S2 Overview of datasets.

		GCall	#11493	#00006	#09807	#01634
	Category	reference	low ϵ	medium ϵ	low ϵ	high ϵ
Exp. 1	Slope	3.52	3.62	3.59	4.49	3.50
	Intercept	12.5	6.40	4.01	6.31	4.35
	R^2	>0.999	0.999	>0.999	0.994	>0.999
	ϵ	92.2%	88.8%	89.8%	67.0%	92.9%
Exp. 2	Slope	3.50	3.61	3.56	4.44	3.55
	Intercept	12.6	6.33	4.46	6.29	4.20
	R^2	0.999	>0.999	>0.999	0.991	>0.999
	ϵ	93.0%	89.2%	91.0%	68.0%	91.3%
Exp. 3	Slope	3.42	3.60	3.54	4.41	3.49
	Intercept	12.7	6.34	4.07	6.24	4.33
	R^2	>0.999	>0.999	>0.999	0.993	>0.999
	ϵ	96.0%	89.6%	91.8%	68.6%	93.3%
Overall	$\bar{\epsilon}$	93.7%	89.2%	90.8%	67.8%	92.5%
	SEM	1.14%	0.24%	0.58%	0.48%	0.61%

Table S3 Overview of qPCR results.

Sample 1	Sample 2	$\Delta\epsilon$	p-value	Confidence interval	
#00006	#01634	1.67%	0.4521	-1.48%	4.82%
#00006	#09807	-23.00%	3×10^{-9}	-26.14%	-19.85%
#00006	#11493	-1.64%	0.4671	-4.79%	1.51%
#00006	GCall	2.85%	0.0812	-0.3%	6.00%
#01634	#09807	-24.67%	1×10^{-9}	-27.81%	-21.52%
#01634	#11493	-3.31%	0.0385	-6.46%	-0.16%
#01634	GCall	1.18%	0.7341	-1.97%	4.33%
#09807	#11493	21.36%	6×10^{-9}	18.21%	24.50%
#09807	GCall	25.84%	9×10^{-10}	22.7%	28.99%
#11493	GCall	4.49%	0.0059	1.34%	7.64%

Table S4 Multiple comparisons of the qPCR results using Tukey's range test.

	Initial abundance			PCR efficiency		
Group	GCall/fix	Erlich (ext.)	Erlich (int.)	GCall/fix	Erlich (ext.)	Erlich (int.)
GCall/fix	1.00	0.30	0.28	1.00	-0.23	-0.22
Erlich (ext.)	0.30	1.00	0.63	-0.23	1.00	0.91
Erlich (int.)	0.28	0.63	1.00	-0.22	0.91	1.00

Table S5 Spearman rank correlation of the estimated paramters of the three external validation datasets.