# Blocking uncertain errors of PCR

**Takumi Takahashi**[1], **Hiroyuki Aoyanagi**[1], **Simone Pigolotti**[2], **and Shoichi Toyabe**[1,*]

[1]Department of Applied Physics, Graduate School of Engineering, Tohoku University, Sendai, Japan
[2]Biological Complexity Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, Japan
[*]toyabe@tohoku.ac.jp

The polymerase chain reaction (PCR) plays a central role in genetic engineering and is routinely used in various applications, from biological and medical research to the diagnosis of viral infections. PCR is an extremely sensitive method for detecting target DNA sequences, but it is substantially error-prone. In particular, the mishybridization of primers to contaminating sequences can result in false positives for virus tests. The blocker method, also called the clamping method, has been developed to suppress mishybridization errors. However, its application is limited by the requirement that the contaminating template sequence must be known in advance. Here, we demonstrate that a mixture of multiple blocker sequences effectively suppresses the amplification of contaminating sequences even in the presence of uncertainty. This result is based on a prediction of the optimal blocker mixture that minimizes the error fraction. Our method extends the applicability of PCR and other hybridization-based techniques, including genome editing, RNA interference, and DNA nanotechnology, by improving their fidelity.

## 1 Introduction

Polymerase chain reaction (PCR) is a technique for amplifying DNA strands characterized by specific sequences[1]. In PCR, a pair of short strands, called primers, are mixed in the reaction solution. They specify the sequence region to be replicated and trigger DNA replication. By running several replication cycles, PCR is able to exponentially amplify a DNA sample. This exponential amplification makes PCR extremely sensitive, but also error-prone. On the one hand, the incorporation of incorrect nucleotides during polymerase extension can be significantly suppressed by using a high-fidelity polymerase and carefully tuning the reaction conditions. On the other hand, errors can be caused by mishybridization of primers onto contaminating strands[2] (Fig.1a). In practical applications, test samples are often far from being pure, and contain a variety of DNA sequences. Primers readily hybridize to contaminating sequences similar to the target sequence, resulting in false positives in viral testing.

The blocker or clamping method was developed to suppress mispriming in PCR[3–6]. (Fig.1b). In this method, a nucleic acid sequence called a blocker is added to the PCR mixture. The blocker sequence is designed to be complementary to the contaminating sequence. It suppresses errors by blocking mishybridization of the primer to the contaminating sequence (Fig. 1c). A chimeric strand of DNA and locked nucleic acids (LNA)[7] or DNA and peptide nucleic acids (PNA)[8] are often used as blockers due to their high specificity[3,9]. In our previous work, we established that blocking combines energetic and kinetic discrimination toward correct hybridization to the target sequence[10]. Because of this mechanism, error suppression remains effective even at relatively low temperatures, and therefore, it is expected to be applicable to thermally sensitive systems such as biological cells. Since the blocker sequence is designed to be complementary to the contaminating sequence, the blocker method requires to know the contaminating sequence in advance. This limitation can be quite severe in realistic situations where there are multiple possible contaminating sequences, but we do not know which sequences are contaminating each sample.

This work extends the blocker method to situations where the information about contaminating sequences is limited. We specifically consider the following scenario. A PCR mixture contains a target sequence to be detected, R, and a contaminating sequence, $\mathrm{W}_i$, randomly sampled from $N$ possible sequences with probability $p_i$ ($1 \leq i \leq N$). The primer P is complementary to the primer-binding region of R. The contaminating sequence has a corresponding sequence region with single or multiple mutations, which cause mismatches in the hybridization with P. The blockers are labeled as $\mathrm{B}_i$, such that $\mathrm{B}_i$ is perfectly complementary to $\mathrm{W}_i$ (Fig. 1c).

## 2 Model

Our method is based on adding a combination of blocker sequences $\{\mathrm{B}_i\}$ that are complementary to all the possible contaminating sequences $\{\mathrm{W}_i\}$. Throughout this paper, we use $\{\cdot\}$ to denote a set. The primer we use in this work is 20 bases long, which is typical for PCR virus tests. Therefore, there exist in principle $3^{20}$ possible W sequences. However, the binding strength of P to W with two or more mismatches is negligible in standard PCR assays. Therefore, in practice, we only need to worry about the contaminating sequences with a single mismatch with respect to R, which are limited to $3 \times 20 = 60$ sequences.

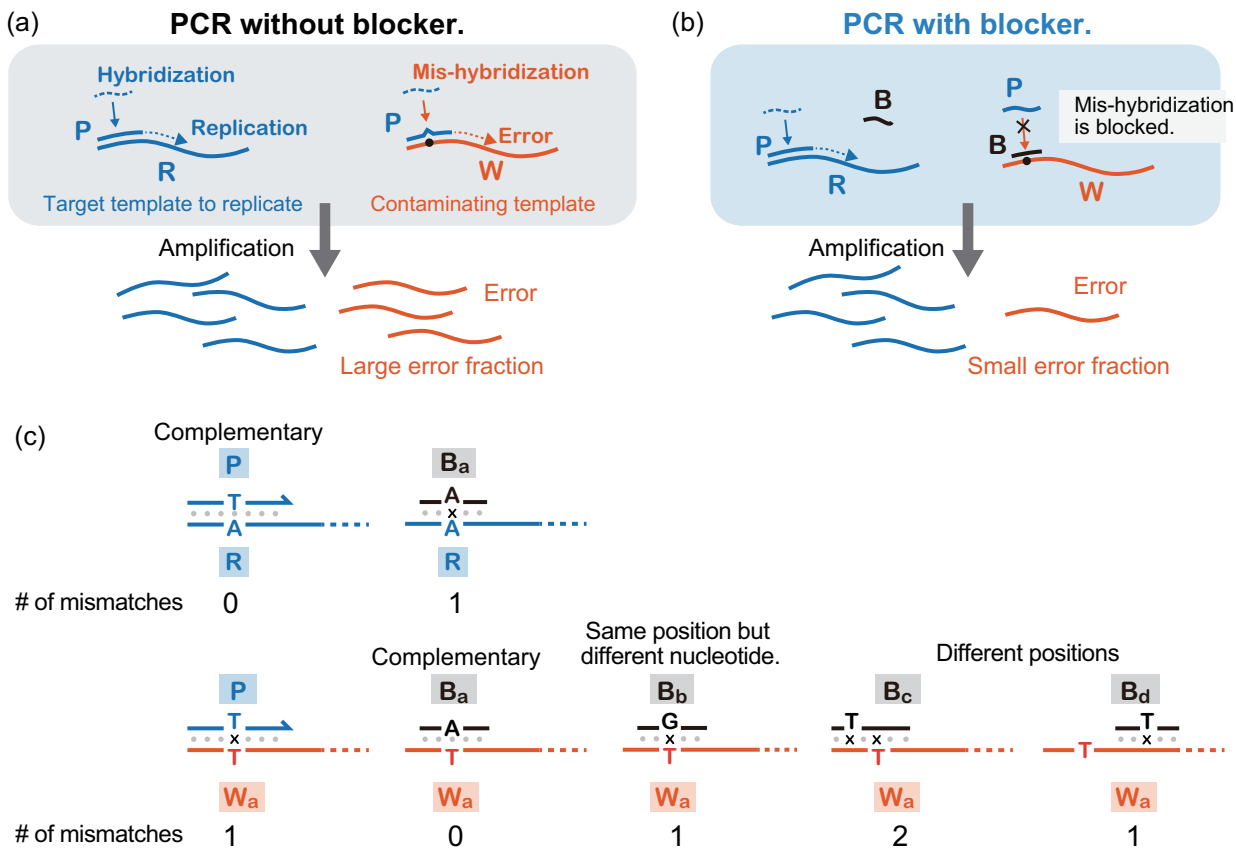


**Figure 1.** PCR and blocker method. (a) Mishybridization of the primer strands (P) to the contaminating template strands (W) results in errors. (b) Blocker method. The sequence of the blocker (B) is designed to be complementary to the sequence of W and blocks the primer binding to W. (c) Different sequences $W_i$ and $B_j$ are labeled so that $i = j$ when $W_i$ and $B_j$ are complementary. The number of mismatches changes depending on the relative positions of the mutation in W and the target mutation in B.

We define a growth rate $r$ as the number of strands copied in a PCR cycle divided by the number of template strands. With this definition, perfect PCR efficiency corresponds to $r = 2$. We call $r_{\mathrm{R}}$ and $r_{\mathrm{W}_i}$ the growth rates of R and $\mathrm{W}_i$, respectively. Then, the error fraction is defined as[11]

$$\varepsilon_i = \frac{r_{\mathrm{W}_i}}{r_{\mathrm{R}}}. \tag{1}$$

We aim at finding the optimal combination of blocker sequences that minimizes the error fraction, and the resulting error fraction. More formally, we call $b_i$ the concentration of $\mathrm{B}_i$ and $\varepsilon_i(\{b_k\})$ the error fraction associated with $\mathrm{W}_i$ in the presence of a blocker distribution $\{b_k\}$. We seek for the optimal blocker distribution $\{b_k^*\}$ that minimizes either the mean error fraction or the maximum error fraction, defined as

$$\varepsilon_{\mathrm{mean}}(\{b_k\}) = \sum_i p_i \varepsilon_i(\{b_k\}), \quad \varepsilon_{\mathrm{max}}(\{b_k\}) = \max_i \varepsilon_i(\{b_k\}), \tag{2}$$

respectively, under the constraint that the total concentration $b_{\mathrm{tot}} = \sum_i b_i$ is fixed. This constraint reflects the limitation that too many blocker strands also suppress the replication of R, as we will see (Fig. S3). The choice between minimizing $\varepsilon_{\mathrm{mean}}$ or $\varepsilon_{\mathrm{max}}$ depends on the purpose of the PCR assay. For example, if one is trying to prevent the worst case, it would be natural to minimize $\varepsilon_{\mathrm{max}}$.

We consider a biophysical model for evaluating $\varepsilon_i(\{b_k\})$. First, we consider the situation with a single B sequence and then extend the model to multiple B sequences. In experiments, the blocker concentration is relatively high ($\sim$ µM). We therefore assume that the blocker binding and dissociation are always equilibrated[10]. Then, the growth rates are given by

$$r_{\mathrm{W}_i}(b_j) = \frac{\tilde{r}_{\mathrm{W}_i}}{b_j/K_{i,j} + 1}, \qquad r_{\mathrm{R}}(b_j) \simeq \tilde{r}_{\mathrm{R}}. \tag{3}$$

Here, $K_{i,j}$ is the dissociation constant associated with the binding of $\mathrm{W}_i$ and $\mathrm{B}_j$. The factor $1/(b_j/K_{i,j} + 1)$ represents the fraction of the free $\mathrm{W}_i$ that is not bound by $\mathrm{B}_j$. We assumed that $b_j$ is not extremely large. $\tilde{r}_{\mathrm{W}_i}$ is the growth rate in the absence of the blocker sequences. We assume that the primer then binds to free $\mathrm{W}_i$, initiating the polymerase extension. Because the binding of $\mathrm{B}_j$ to R is negligible under the present condition (Fig. S3), the replication of $\mathrm{R}_i$ is not blocked by the blockers; $r_{\mathrm{R}}(b_j) \simeq \tilde{r}_{\mathrm{R}}$.

Combining Eqs. (1) and (3), we obtain an expression for the error fraction in the presence of a blocker $\mathrm{B}_j$:

$$\varepsilon_i(b_j) = \frac{\tilde{\varepsilon}_i}{b_j/K_{i,j} + 1}. \tag{4}$$

Here, $\tilde{\varepsilon}_i = \tilde{r}_{\mathrm{W}_i}/\tilde{r}_{\mathrm{R}_i}$ is the error fraction in the absence of blocker sequences. Thus, the model predicts that the blocking strength of $\mathrm{W}_i$ by $\mathrm{B}_j$ is essentially determined by $K_{i,j}$. In the presence of multiple blockers, Eq. (4) becomes

$$\varepsilon_i(\{b_k\}) = \frac{\tilde{\varepsilon}_i}{\sum_j b_j/K_{i,j} + 1}. \tag{5}$$

Since $b_j$ and $K_{i,j}$ are non-negative, Eq. (5) implies $\varepsilon_i(\{b_k\}) \leq \tilde{\varepsilon}_i$. As expected, the error is always reduced by the addition of blockers.

We first validate the models (4, 5) by experiments and then evaluate the optimal blocker distributions $\{b_i^*\}$ that minimize either $\varepsilon_{\mathrm{mean}}$ or $\varepsilon_{\mathrm{max}}$.

## 3 Experiments

We performed experiments to characterize the effect of the blockers on the error fraction; to validate Eqs. (4) and (5); and also to evaluate $\tilde{\varepsilon}_i$ and $K_{i,j}$. We performed a linear PCR using a single primer instead of a primer pair since the error fraction is directly evaluated from the product amounts in the linear PCR (Fig. S1). The reaction mixture contains 2.5 nM R template, 2.5 nM $\mathrm{W}_i$ template, 100 nM primer, thermostable (Taq) DNA polymerase enzyme, and standard buffer mixture. Blocker sequences are added depending on the experiments. See the Methods and SI for details.

We use W sequences with mutations at the positions where the original nucleotide of R is either A or T. Mutations at A and T cause more frequent errors than those at C and G positions due to their lower binding stability. The reason for this choice is to focus on the worst cases for practical assays.

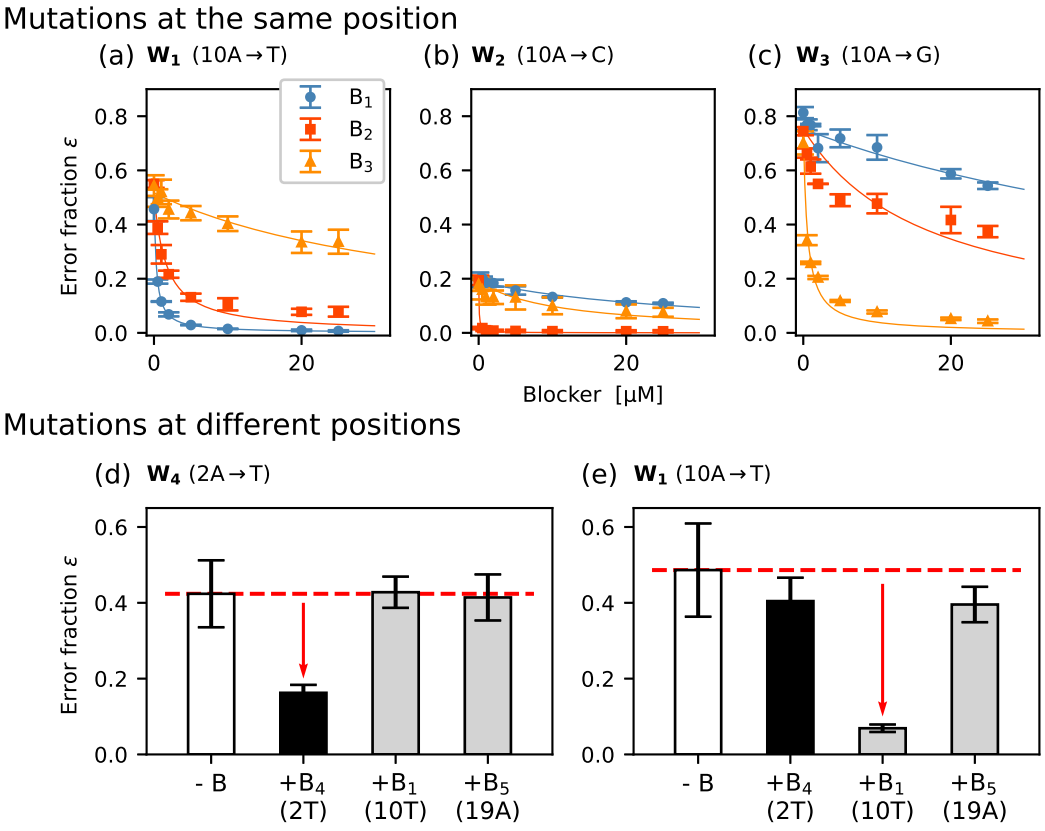


**Figure 2.** Error suppression by a single blocker sequence. **a**–**c**, Dependence of the error fractions $\varepsilon$ on blocker concentration for different combinations of W and B, where the target mutations of B are all at the same position in W. The subscript $i$ of $\mathrm{W}_i$ is used for labeling different W sequences. $\mathrm{B}_i$ is the blocker sequence complementary to $\mathrm{W}_i$. The expression, for example, $\mathrm{W}_1$ (10A $\rightarrow$ T) indicates that A, the tenth base from the 3' end in R, is mutated to T in $W_1$. See SI Fig. S3 for the production amounts of R and W. The solid curves are fitting curves by Eq. (4) with the fitting parameter $K_{i,j}$. **d**, **e**, Error suppression including the situations where the positions of the target mutations of B are different from the mutated position of W. The concentration of the blocker is 2 μM. Error bars indicate standard deviations. Each data point represents an average over three independent experiments.

### 3.1 Dependence of PCR error on blocker concentration.

We measured the dependence of the error fractions on the blocker concentrations for different combinations of $W_i$ and $B_j$ (Fig. 2). The error fraction $\varepsilon_i(b_j)$ decreased significantly with $b_j$ for the complementary pair $i = j$ (Fig. 2a–c). This dependence is well fitted by Eq. (4), supporting the validity of the model.

We distinguish two cases for non-complementary pairs of $W_i$ and $B_j$ ($i \neq j$). If the mismatch positions of $i$ and $j$ are the same, the blocker still suppresses the error (Fig. 2a–c). This is because the binding of $B_j$ to $W_i$ is not negligible in this case. In particular, purine-pyrimidine pairs (A and C) and (G and T) bind weakly even though they are not complementary. This crosstalk effect between B and W explains the tendency of the error fractions in Fig. 2a–c. For example, in Fig. 2a, $B_2$ suppresses errors more than $B_3$. This tendency is due to the stronger binding of $B_2$, which has G at the corresponding mutation position, to $W_1$ than $B_3$, which has C. The error fraction without blocker is the highest in Fig. 2c because G, the mutation in $W_3$, and T of the primer at the corresponding position bind to some extent. Equation (4) fits the results well for such pairs.

When the mismatch positions of $i$ and $j$ are different, error suppression is negligible (Fig. 2d, e). This is because the number of mismatched nucleotides is two (see, for example, $W_a$ and $B_c$ in Fig. 1c) and their hybridization is unstable. This may not be the case if the two mismatches are far apart (see $W_a$ and $B_d$ in Fig. 1c). In such cases, however, the primer binding region is not sufficiently covered by the blocker, and blocking is again not effective.

When the blocker concentration is higher than 10 µM, the replication of R is also suppressed (Fig. S3). Therefore, we used $b_{\mathrm{tot}} = 10\,\mu\mathrm{M}$ in the following experiments.

### 3.2 Estimation of dissociation constants $K_{i,j}$ from sequences.

Directly measuring $K_{i,j}$ for all the $60 \times 60$ pairs of $W_i$ and $B_j$ is unpractical. Based on the results in Fig. 2d, e, we set $K_{i,j} = \infty$ for non-complementary pairs with different mismatch positions. For other pairs, we estimated the $K_{i,j}$ values by the nearest neighbor method[9,12,13] with LNA parameters including mismatches (Fig. 3). We found that the values estimated by the nearest neighbor method are similar to those obtained by the fitting of Eq. (4) in Fig. 2(a–c).

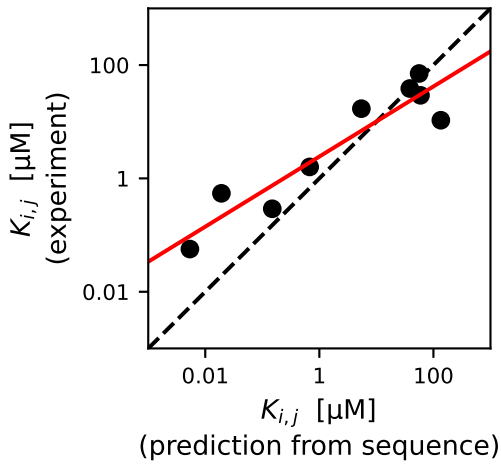


**Figure 3.** Experimental vs. predicted values of the dissociation constants $K_{i,j}$. The experimental values $K_{i,j}$ are obtained by fitting Eq. (4) to the experimental curves in Fig. 2(a–c). The predicted values are the estimation from the sequences by the nearest-neighbor method[9,12,13]. We used the BioPython package with modifications[14]. The dashed line indicates the equality between the experimental values and predicted values. The solid line is a fit by a curve linear in the log scale: $\ln y = a \ln x + b$ with $a = 0.619$ and $b = -4.38$. Here, $x$ and $y$ are in the units of M.

Broadly, this result supports the validity of the model given by Eq. (4) for the blocking effect. However, there are small deviations between the experiments and predictions. The deviations could be due to the fact that the parameters of the nearest neighbor method are not optimized for our experimental conditions. For better estimation, we used a linear fitting in the log scale (red solid curve in Fig. 3) to calibrate the $K_{i,j}$ values estimated by the nearest-neighbor method.

### 3.3 Estimation of error fraction without blocker $\tilde{\varepsilon}_i$.

We measured the dependence of $\tilde{\varepsilon}_i$ on the position and nucleotide species of the mismatch of W (Fig. 4). We observed a tendency for $\tilde{\varepsilon}_i$ to be small at both the 5' and 3' ends of the primer binding region of W. This is because a mismatch near the ends causes the ends to flap and reduces stability more than internal mismatches. In particular, rigid hybridization at the 5'-end of the primer binding region of W (3'-end of the primer) is necessary for the polymerase to initiate extension[15–17].

The value of the parameter $\tilde{\varepsilon}_i$ depends on the nucleotide species (Fig. 4). However, this dependence is hard to model accurately. For example, the sequence near the mismatch position should affect $\tilde{\varepsilon}_i$. Measuring all sixty values of $\tilde{\varepsilon}_i$ is not practical. We could calculate $K_{i,j}$ from the sequences because the blocker concentration is sufficiently large, and the equilibrium is expected for the blocker binding. However, equilibrium is not expected for primer binding because the primer concentration is not large enough[10]. Therefore, it is not straightforward to estimate $\tilde{\varepsilon}_i$ only from the sequences. Here, we assume a simple model where $\tilde{\varepsilon}_i$ depends only on the mismatch position and neglect the dependence on the nucleotide species. To account for the flat dependence on position in the internal region, we fit $\tilde{\varepsilon}_i$ with a combination of two hyperbolic tangent functions (Fig. 4). We, however, observed a significant variation around the fitted line. We will see how the variation affects the optimization of the blocker concentrations.

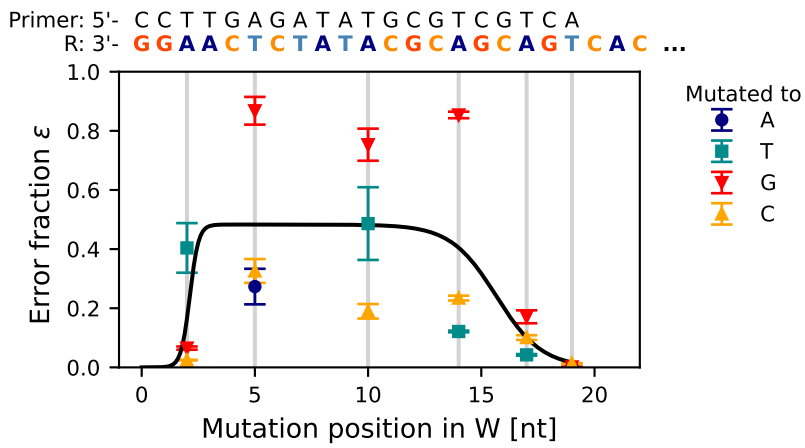


**Figure 4.** Dependence of error fraction on the mutation position and the nucleotide species of W in the absence of the blockers. Solid line is a fitting curve $f(x) = p[1+\tanh(q(x-r))][1-\tanh(s(x-u))]$ with the fitting parameters $p = 0.12$, $q = 2.6$, $r = 2.1$, $s = 0.50$, and $u = 16$. Each symbol corresponds to a different nucleotide type in the W sequence. Error bars indicate standard deviations. The number of independent experiments is three for each point. The sequences of R and P are shown on the top for the reference.

### 3.4 Error suppression by blocker combination.

We can now calculate $\varepsilon_i(b_j)$ by Eq. (4) using $K_{i,j}$ calculated by the nearest neighbor method with calibration (Fig. 3) and $\tilde{\varepsilon}_i$ modeled in Fig. 4. Next, we experimentally test the model for multiple blocker sequences, Eq. (5).

We contaminated one of three specific W sequences and evaluated error suppression by a combination of three blocker sequences targeting the three W sequences (Fig. 5a–c). The sum of blocker concentrations was kept at $\sum_j b_j = 10\,\mu\text{M}$. We only tested the W sequences with the same mutation positions since blocking is ineffective if the target position of B differs from the mutation position of W, see Fig. 2d, e. We found a gradient of the error fraction in a ternary diagram. The steepest gradient is obtained along the axis of the blocker complementary to the contaminating sequence ($i = j$), as expected. Eq. (5) qualitatively reproduces the experimental results under different conditions (Fig.5d–f). The absolute values show some discrepancy between the experimental results and the model calculations. The main source of the discrepancy seems to be the approximation by neglecting the dependence of $\tilde{\varepsilon}_i$ on the nucleotide species (Fig. 4).

### 3.5 Optimal blocker combinations

We theoretically derive the optimal blocker distribution $\{b_k^*\}$ that minimizes the mean or maximum error fraction (Eq. (2)). By substituting Eq. (5) to Eq. (2), we obtain

$$\varepsilon_{\text{mean}}(\{b_k\}) = \sum_i \frac{p_i \tilde{\varepsilon}_i}{\sum_j b_j / K_{i,j} + 1}, \qquad \varepsilon_{\max}(\{b_k\}) = \max_i \frac{\tilde{\varepsilon}_i}{\sum_j b_j / K_{i,j} + 1}. \tag{6}$$

We obtain $\{b_k^*\}$ numerically under the constraints $b_{\text{tot}} = \sum_j b_j = 10\,\mu\text{M}$ and $b_j \geq 0$ for $\forall j$. It can be shown that $\varepsilon_{\text{mean}}(\{b_k\})$ and $\varepsilon_{\max}(\{b_k\})$ are convex functions of $\{b_k\}$ for arbitrary $\{\tilde{\varepsilon}_i\}$ and $\{K_{i,j}\}$ (see Methods). This implies that we can efficiently find $\{b_k^*\}$ by a gradient descending method, without risking to be stuck in a local minimum (see Methods).

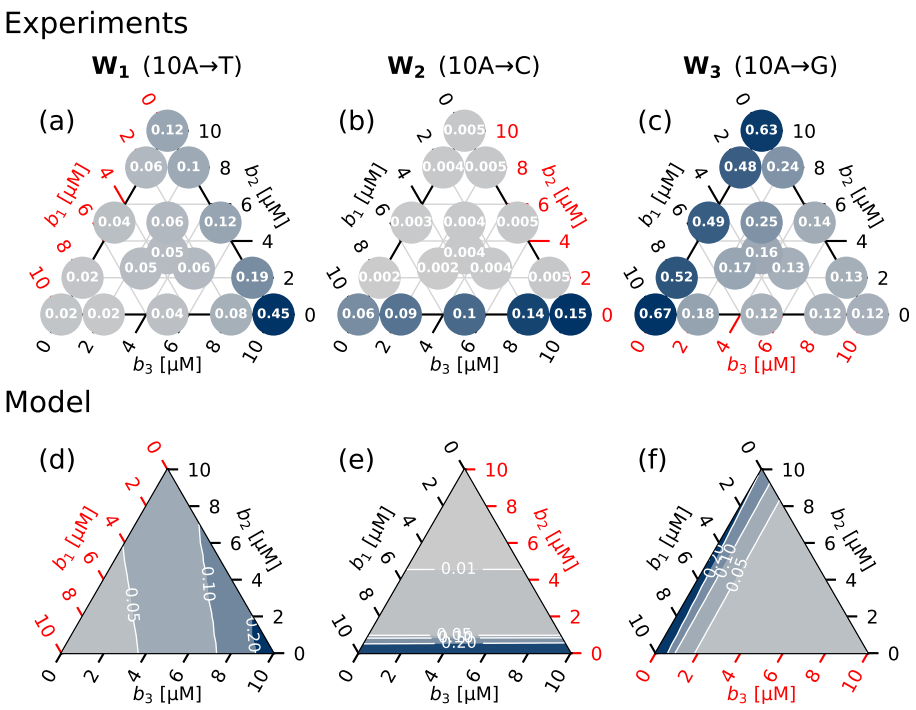


**Figure 5.** Error suppression by combinations of multiple blocker sequences. Experimental data (top) and corresponding model calculation by Eq. (5) (bottom). The error fractions without blockers $\tilde{\varepsilon}$ are 0.52 (a), 0.19 (b), 0.75 (c), and 0.48 (d–f). The numbers in the circles in (a–c) are the mean error fractions of three independent experiments. The blockers highlighted by red color are complementary to the contaminating sequences. $b_{\mathrm{tot}} = 10\,\mu\mathrm{M}$. The plots are colored according to the error fractions normalized by the maximum error fraction in each plot.

We apply this approach to three situations with different mismatch patterns (Fig. 6 and S4), where we have partial prior knowledge of the errors (i, ii) or no prior knowledge (iii). In all examples, we considered the worst case, where the contamination probability is uniform, $p_i = 1/N$. The results are summarized in Table 1.

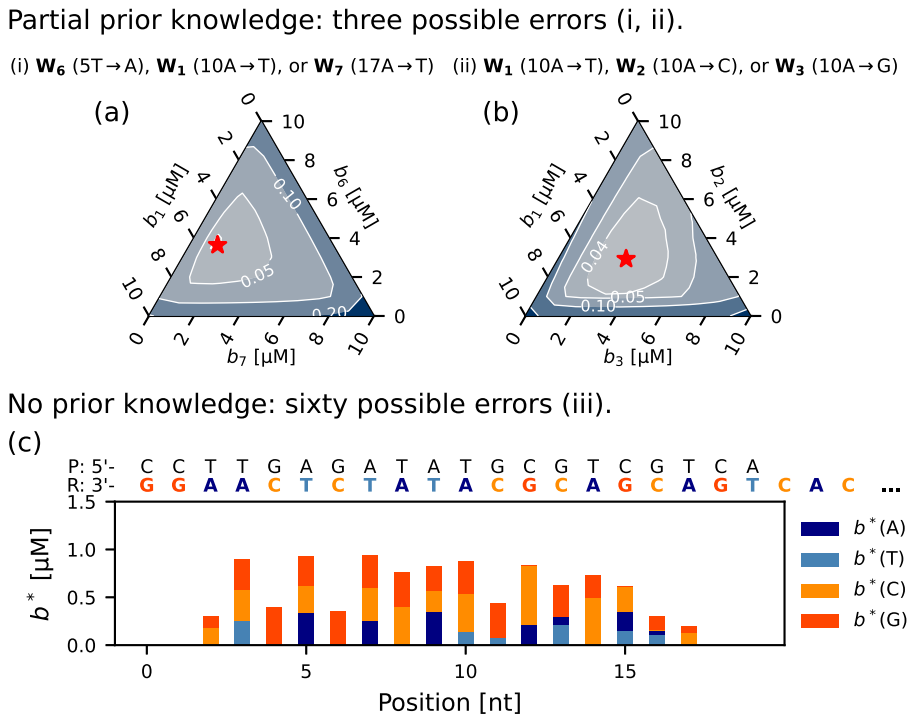


**Figure 6.** Optimal blocker concentrations minimizing $\varepsilon_{\mathrm{mean}}$ given by (6) under the constraint $b_{\mathrm{tot}} = 10\,\mu\mathrm{M}$. See SI Fig. S4 for the optimal blocker concentrations minimizing $\varepsilon_{\mathrm{max}}$. (a), (b) There are possible three mutations at different positions (a) or same position (b). The values of $\varepsilon_{\mathrm{mean}}$ are indicated by colors normalized by the maximum error rate in each plot. The optimal blocker concentrations are indicated by stars. (c) Optimal blocker concentrations for the case with 60 possible errors. The sequences of R and P are shown on the top for the reference. The expressions $b^*(\mathrm{N})$ (N is A, T, C, or G) indicate the optimal concentrations of the blockers targeting N mutation at each position. $\varepsilon_{\mathrm{mean}}$ without blockers are 0.36 (a), 0.48 (b), and 0.33 (c). The minimum $\varepsilon_{\mathrm{mean}}$ at the optimal blocker concentrations are 0.039 (a), 0.035 (b), and 0.18 (c).

***(i) The mutation is located in one of the three positions. The nucleotide species is specified (Fig.6a).***
In this scenario, $\mathrm{B}_j$ has a negligible error suppression effect on $\mathrm{W}_i$ that does not match ($i \neq j$) as seen in Fig. 2d, e. Hence, the effect of the blockers can be decomposed as $\varepsilon_i(\{b_k\}) = \varepsilon_i(b_i)$, leading to a simple superposition $\varepsilon_{\mathrm{mean}} = \sum_i p_i \varepsilon_i(b_i)$. The balance among the three complementary combinations of W and B determines the optimal blocker combinations. The model predicts that the optimal blocker combination indicated by a star in Fig.6a is able to reduce $\varepsilon_{\mathrm{mean}}$ by a factor of 9.0 (Table 1).

The reduction ratio was also calculated based on experimental data by interpolating the experimental error fractions to the optimal blocker concentrations (Fig. S5). The ratio was 7.1, similar to that predicted by the model (see Table 1 for the comparison). We obtained a similar reduction ratio for $\varepsilon_{\mathrm{max}}$.

Taking a uniform blocker concentration $b_i = b_{\mathrm{tot}}/N$ as a reference, we still have a significant error reduction, although the optimal concentrations give better error reduction (Table 1). This shows the effectiveness of the blocker method as we will discuss in the Discussion section.

Our results show that the $\varepsilon_{\mathrm{mean}}$ (Fig. 6a, b) and $\varepsilon_{\mathrm{max}}$ (Fig. S4a, b) have single minima, confirming our theoretical conclusion based on the convexity of $\varepsilon_{\mathrm{mean}}$ and $\varepsilon_{\mathrm{max}}$.

***(ii) The mutation is located in the specified position, but the nucleotide species is not identified (Fig.6b).***
In this case, the model predicted that the optimal blocker combination reduced $\varepsilon_{\mathrm{mean}}$ by a factor 14. This factor is larger than in case (i) due to crosstalk. As shown in Fig. 2a–c, $B_j$ can suppress the replication of $W_i$ even if $i \neq j$ when the mutations of $i$ and $j$ are located at the same position. The reduction ratio calculated for the experimental data deviated from the ratio calculated by

the model both for $\varepsilon_{\mathrm{mean}}$ and $\varepsilon_{\mathrm{max}}$ (Table 1). The discrepancy is possibly due to neglecting the nucleotide dependence of $\tilde{\varepsilon}$. In this particular scenario, this assumption could significantly affect the estimation.

However, the blocker addition still significantly reduced errors in the experiments, supporting the practical effectiveness of the method also in this case.

***(iii) No prior knowledge. (Fig.6c).***

In this scenario, we added a mixture of all sixty blocker sequences at the optimal concentrations, assuming that there are sixty error possibilities. We found that the blocker method is still effective, and the optimal blocker combination reduced $\varepsilon_{\mathrm{mean}}$ and $\varepsilon_{\mathrm{max}}$ with reduction ratios of 1.8 and 1.4, respectively. Although this case may not be relevant for practical virus testing, it illustrates that the blocker addition always reduces errors even in this worst-case scenario.

The optimal blocker distribution $\{b_j^*\}$ can be explained by the dependence of $\tilde{\varepsilon}_i$ on the position and the dependence of $K_{i,j}$ on the sequence combinations. In particular, $b_j^*$ are small at the edges because $\tilde{\varepsilon}_i$ are smaller at the edges than at the internal positions (as shown in Fig. 4). Therefore, the priority to target the edges with blockers is low.

The concentrations $\{b_j^*\}$ optimizing $\varepsilon_{\mathrm{mean}}$ are strongly biased toward targeting C and G mutations (Fig. 6c). Since C and G bind more strongly than A and T, corresponding $K_{i,j}$ values are typically small. Therefore, targeting C and G mutations is more effective in reducing $\varepsilon_{\mathrm{mean}}$ under the constraint of a fixed $b_{\mathrm{tot}}$. This tendency is quantified by the blocking efficacy $\alpha_j$ of the blockers, defined as $\alpha_j = \sum_i (1/K_{ij})$ (Fig. S6a). $b_j^*$ for $\varepsilon_{\mathrm{mean}}$ increases with $\alpha_j$ at small $\alpha_j$. The optimal concentrations $b_j^*$ for $\varepsilon_{\mathrm{mean}}$ decrease at large $\alpha_j$ because a small amount is enough for the blockers with large $\alpha_j$ to suppress errors.

In contrast, $\{b_j^*\}$ for $\varepsilon_{\mathrm{max}}$ is strongly biased toward targeting A and T mutations (Fig. S4c). Accordingly, $b_j^*$ for $\varepsilon_{\mathrm{max}}$ decreases monotonically with $\alpha_j$ (Fig. S6b). In particular, $b_j^*$ is roughly proportional to $1/\alpha_j$ (Fig. S6b, inset). This tendency is consistent with intuition as well. If $\varepsilon_i$ of a particular template $\mathrm{W}_i$ is larger than other templates, we need to increase $b_i$ to reduce $\varepsilon_{\mathrm{max}}$. The error $\varepsilon_i$ is large when $\mathrm{B}_i$ is not effective and, therefore, has a small $\alpha_i$ value. Hence, $b_j^*$ for the blockers with small $\alpha_j$ should be large to minimize $\varepsilon_{\mathrm{max}}$, which results in a uniform $\varepsilon_i$ distribution.

**Table 1.** Summary of error suppression. $\varepsilon_{\mathrm{mean}}$ and $\varepsilon_{\mathrm{max}}$ without blocker (None), with blocker combinations with the uniform concentrations (Uniform, $b_k = b_{\mathrm{tot}}/N$), or with the optimal concentrations (Optimal, $b_k^*$). The experimental values are obtained by interpolating the blocker concentrations in Fig. S5.

| | | $\varepsilon_{\mathrm{mean}}$ | | | | $\varepsilon_{\mathrm{max}}$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Blocker | None | Uniform | Optimal | Ratio (None / Optimal) | None | Uniform | Optimal | Ratio (None / Optimal) |
| (i) | Model | 0.36 | 0.047 | 0.039 | 9.0 | 0.48 | 0.088 | 0.048 | 10 |
| | Experiment | 0.27 | 0.040 | 0.038 | 7.1 | 0.49 | 0.073 | 0.069 | 7.1 |
| (ii) | Model | 0.48 | 0.036 | 0.035 | 14 | 0.48 | 0.066 | 0.044 | 11 |
| | Experiment | 0.48 | 0.071 | 0.072 | 6.7 | 0.75 | 0.16 | 0.20 | 3.7 |
| (iii) | Model | 0.33 | 0.20 | 0.18 | 1.8 | 0.48 | 0.46 | 0.34 | 1.4 |

## 4 Discussion

We demonstrated the effectiveness of the blocker method in PCR with limited information about the contaminating sequences. Our methodology is based on adding combinations of multiple blocker sequences informed by biophysical considerations. In particular, we successfully modeled the effect of the blockers and obtained the relevant parameters. With this approach, we were able to reduce the mean and maximum error fractions even in the worst-case scenario where all sixty possible contaminations are present (Table 1).

Adding blockers does not increase the error rate unless sufficiently large concentrations of blocker sequences inhibit the replication of R. This means that even though the perfect optimization achieves the minimum error fraction, an inaccurate optimization is still effective. In fact, the average error $\varepsilon_{\mathrm{mean}}$ is also reduced by uniform blocker concentrations $b_i = b_{\mathrm{tot}}/N$ (Table 1). This tolerance is an inherent advantage of the blocker method because we can reduce the experimental cost at the expense of reducing the error suppression effect. For example, we neglected the dependence of $\tilde{\varepsilon}_i$ on the nucleotide species and considered only the mismatch position (Fig. 4). The $K_{i,j}$ values are estimated from the sequences by the nearest neighbor method with a calibration (Fig. 3). While these approximations may reduce the error suppression effect, they significantly lower the burden of evaluating parameters, $\tilde{\varepsilon}$ and $K_{i,j}$. In addition, assuming that the model of $\tilde{\varepsilon}_i$ and calibration curve of $K_{i,j}$ obtained in this work are applicable to other primer sequences, we do not need additional calibration for each system.

Our methodology for reducing uncertain errors improves the precision of PCR and expands its applicability. The blocker method targets nucleic acid hybridization. Since most gene technologies, including genome editing[18], RNA interference[19], and DNA nanotechnology[20], rely on hybridization, the applicability of the present method is broad and not limited to PCR.

## 5 Methods

The experiments are performed using the same procedure as that in a previous work[10].

The experiments consist of two PCR steps (Fig. S1). In the first PCR step, R and W are amplified by linear PCR. In the second PCR step, the amount of the product in the first step, $\bar{\mathrm{R}}$ and $\bar{\mathrm{W}}$, is evaluated by a quantitative PCR (qPCR).

### Linear PCR experiment

In linear PCR, the product amount is amplified linearly because a single primer sequence is used, which simplifies the error analysis (SI Text S1).

DNA strands were synthesized by Eurofins Genomics (sequences are listed in SI Table S1). DNA/LNA chimeric strands were synthesized by Aji Bio-Pharma. The reaction mixture for polymerization contained 25 units/ml Hot-start Taq DNA polymerase (New England Biolabs), Taq standard reaction buffer, 2.5 nM R, 2.5 nM W, 100 nM P, and indicated amounts of blocker strands. We performed initial heating for 30 s at 95 °C and, then, 10 cycles of 15 s at 95 °C, 30 s at 60 °C, and 5 s at 68 °C using a PCR cycler (BioRad). Immediately after the cycles, the mixture was cooled down on the ice to stop the enzyme reaction and diluted for quantification.

### Quantification

The second quantitative PCR (qPCR) step was performed on a real-time PCR cycler (BioRad) after the linear PCR experiment for quantifying $\bar{\mathrm{R}}$ and $\bar{\mathrm{W}}$ (see SI Text S2). The reaction mixture contains Luna Universal qPCR Master Mix (New England Biolabs), 200 nM each of the primers, and the diluted sample. The dilution rate of the sample was 1/250 in the final concentration. The thermal cycle consisted of initial heating for 60 s at 95 °C, 40 cycles of 15 s at 95 °C, 30 s at 66 °C, and 5 s at 72 °C.

### Optimization of the blocker concentrations

We want to minimize $\varepsilon_{\mathrm{mean}}$ and $\varepsilon_{\mathrm{max}}$ given by Eq. (6) under the constraint that $b_i \geq 0$ for $\forall i$ and $\sum_j b_j = b_{\mathrm{tot}}$. It is difficult to obtain the analytical solution under the constraints given by inequalities. Instead, we numerically find the optimum by a gradient descent method. For validating the method's effectiveness, we verify that $\varepsilon_{\mathrm{mean}}$ and $\varepsilon_{\mathrm{max}}$ are convex functions of $\{b_i\}$.

#### *(i) Minimization of* $\varepsilon_{\mathrm{mean}}$.

To minimize $\varepsilon_{\mathrm{mean}}$, we first note that the domain $\{b_k \geq 0 \forall k; \sum_j b_j = b_{\mathrm{tot}}\}$ is convex, that is, a point between any two points in the domain lies in the same domain. Moreover, the Hessian matrix given by

$$H_{m,k} \equiv \frac{\partial^2 \varepsilon_{\mathrm{mean}}}{\partial b_m \partial b_k} = \sum_i \frac{p_i \tilde{\varepsilon}_i / (K_{i,m} K_{i,k})}{(\sum_j b_j / K_{i,j} + 1)^3} \tag{7}$$

is of the form $H = K^{\mathrm{T}} D K$, where $D$ is a diagonal matrix with positive entries. Here, $K$ is a matrix with the entries $K_{i,m}$. This implies that the Hessian matrix is nonnegative, hence the function $f$ is convex.

The fact that our minimization problem concerns a convex function in a convex domain ensures that the minimum is unique. We can, therefore, find this minimum very efficiently numerically using a gradient descent method. In particular, calling $F_m = -\frac{\partial \varepsilon_{\mathrm{mean}}}{\partial b_m}$, the optimal solution is the unique stable fixed point of the replicator dynamics:

$$\frac{db_m}{dt} = \gamma b_m (F_m - \bar{F}), \qquad m = 1, 2, \ldots, N, \tag{8}$$

where $\bar{F} = \sum_m b_m F_m / b_{\mathrm{tot}}$ is the average gradient. The coefficient $\gamma$ determines the update rate of $\{b_m\}$. The second term in the right-hand side of Eq. (8) ensures that the normalization condition $\sum_j b_j = b_{\mathrm{tot}}$ is preserved by the dynamics.

Substituting the explicit expression of the gradient into Eq. (8), we obtain

$$\frac{db_m}{dt} = \gamma b_m \sum_i \left[ \frac{p_i \tilde{\varepsilon}_i}{(\sum_j b_j / K_{i,j} + 1)^2} \left( \frac{1}{K_{i,m}} - \frac{1}{b_{\mathrm{tot}}} \sum_j \frac{b_j}{K_{i,j}} \right) \right]. \tag{9}$$

We used the replicator equation (9) to obtain the optimal blocker concentrations that minimize the average error fraction $\varepsilon_{\mathrm{mean}}$.

***(ii) Minimization of $\varepsilon_{\max}$.***
The minimization of $\varepsilon_{\max}$ (Eq. (6)) is also possible in a similar manner to $\varepsilon_{\mathrm{mean}}$. We note that $\varepsilon_{\max}(\{b_k\})$ is a convex function as well since the maximum of convex functions is a convex function. Let $i^*$ be the index with the maximum $\varepsilon_i$. Then, the replicator equation reads

$$\frac{db_m}{dt} = \gamma b_m \frac{\tilde{\varepsilon}_{i^*}}{(\sum_j b_j/K_{i^*,j}+1)^2}\left(\frac{1}{K_{i^*,m}} - \frac{1}{b_{\mathrm{tot}}}\sum_j \frac{b_j}{K_{i^*,j}}\right), \qquad i^* = \underset{i}{\operatorname{argmax}}\,\varepsilon_i. \tag{10}$$

We note that, in the numerical integration of Eq. (10), one has to dynamically update $i^*$, i.e., compute at each time step the value of $i$ that maximizes $\varepsilon_i$.

## Author contributions statement

TT did the experiments. TT, HA, SP, and ST developed the theory, designed the system, and wrote the paper.

## Acknowledgments

ST is supported by JSPS KAKENHI Grant Number JP23K17658 and JST ERATO Grant Number JPMJER2302, Japan. SP is supported by KAKENHI Grant Number JP23H01146.

## Conflict of interest statement.

None declared.

## Data availability

Data are available from the authors upon reasonable request.

## References

1. Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. In *Meth. Enzym.*, 335–350, DOI: 10.1016/0076-6879(87)55023-6 (Elsevier, 1987).
2. Whiley, D. M. & Sloots, T. P. Sequence variation in primer targets affects the accuracy of viral quantitative PCR. *J. Clin. Virol.* **34**, 104–107, DOI: 10.1016/j.jcv.2005.02.010 (2005).
3. Ørum, H. *et al.* Single base pair mutation analysis by PNA directed PCR clamping. *Nucleic Acids Res.* **21**, 5332–5336, DOI: 10.1093/nar/21.23.5332 (1993).
4. Li, M. *et al.* Nucleic acid tests for clinical translation. *Chem. Rev* **121**, 10469–10558, DOI: 10.1021/acs.chemrev.1c00241 (2021).
5. Yang, H. *et al.* A tailored LNA clamping design principle: Efficient, economized, specific and ultrasensitive for the detection of point mutations. *Biotech. J.* **16**, 2100233, DOI: 10.1002/biot.202100233 (2021).
6. Homma, C. *et al.* Effectiveness of blocking primers and a peptide nucleic acid (PNA) clamp for 18s metabarcoding dietary analysis of herbivorous fish. *PLoS one* **17**, e0266268, DOI: 10.1371/journal.pone.0266268 (2022).
7. Petersen, M. & Wengel, J. LNA: a versatile tool for therapeutics and genomics. *Trends Biotech.* **21**, 74–81, DOI: 10.1016/s0167-7799(02)00038-0 (2003).
8. Pellestor, F. & Paulasova, P. The peptide nucleic acids (PNAs), powerful tools for molecular genetics and cytogenetics. *Eur. J. Hum. Genet.* **12**, 694–700, DOI: 10.1038/sj.ejhg.5201226 (2004).
9. Owczarzy, R., You, Y., Groth, C. L. & Tataurov, A. V. Stability and mismatch discrimination of locked nucleic acid–DNA duplexes. *Biochem.* **50**, 9352–9367, DOI: 10.1021/bi200904e (2011).
10. Aoyanagi, H., Pigolotti, S., Ono, S. & Toyabe, S. Error-suppression mechanism of PCR by blocker strands. *Biophys. J.* **122**, 1334–1341, DOI: 10.1016/j.bpj.2023.02.028 (2023).
11. Hopfield, J. J. Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Nat. Acad. Sci.* **71**, 4135–4139, DOI: 10.1073/pnas.71.10.4135 (1974).
12. Allawi, H. T. & SantaLucia, J. Thermodynamics and NMR of internal G·T mismatches in DNA. *Biochem.* **36**, 10581–10594, DOI: 10.1021/bi962590c (1997).

**13.** SantaLucia, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Nat. Acad. Sci.* **95**, 1460–1465, DOI: 10.1073/pnas.95.4.1460 (1998).

**14.** Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423, DOI: 10.1093/bioinformatics/btp163 (2009).

**15.** Kwok, S. *et al.* Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies. *Nuc. Acid. Res.* **18**, 999–1005, DOI: 10.1093/nar/18.4.999 (1990).

**16.** Huang, M.-M., Arnheim, N. & Goodman, M. F. Extension of base mispairs by Taq DNA polymerase: implications for single nucleotide discrimination in PCR. *Nuc. Acid. Res.* **20**, 4567–4573, DOI: 10.1093/nar/20.17.4567 (1992).

**17.** Bru, D., Martin-Laurent, F. & Philippot, L. Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16s rRNA gene as an example. *Appl. Environ. Microbiol.* **74**, 1660–1663, DOI: 10.1128/aem.02403-07 (2008).

**18.** Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotech.* **38**, 824–844, DOI: 10.1038/s41587-020-0561-9 (2020).

**19.** Agrawal, N. *et al.* RNA interference: Biology, mechanism, and applications. *Microbiol. Mol. Biol. Rev.* **67**, 657–685, DOI: 10.1128/mmbr.67.4.657-685.2003 (2003).

**20.** Seeman, N. C. & Sleiman, H. F. DNA nanotechnology. *Nat. Rev. Mat.* **3**, 17068, DOI: 10.1038/natrevmats.2017.68 (2017).