Mol Biotechnol
DOI 10.1007/s12033-016-9957-7

ORIGINAL PAPER

T7 Endonuclease I Mediates Error Correction in Artificial Gene Synthesis

Ana Filipa Sequeira[1,2] · Catarina I. P. D. Guerreiro[2] · Renaud Vincentelli[3] · Carlos M. G. A. Fontes[1,2]

Abstract Efficacy of de novo gene synthesis largely depends on the quality of overlapping oligonucleotides used as template for PCR assembly. The error rate associated with current gene synthesis protocols limits the efficient and accurate production of synthetic genes, both in the small and large scales. Here, we analysed the ability of different endonuclease enzymes, which specifically recognize and cleave DNA mismatches resulting from incorrect impairments between DNA strands, to remove mutations accumulated in synthetic genes. The *gfp* gene, which encodes the green fluorescent protein, was artificially synthesized using an integrated protocol including an enzymatic mismatch cleavage step (EMC) following gene assembly. Functional and sequence analysis of resulting artificial genes revealed that number of deletions, insertions and substitutions was strongly reduced when T7 endonuclease I was used for mutation removal. This method diminished mutation frequency by eightfold relative to gene synthesis not incorporating an error correction step. Overall, EMC using T7 endonuclease I improved the population of error-free synthetic genes, resulting in an error frequency of 0.43 errors per 1 kb. Taken together, data presented here reveal that incorporation of a mutation-removal step including T7 endonuclease I can effectively improve the fidelity of artificial gene synthesis.

Keywords Gene synthesis · Error removal · Enzyme mismatch cleavage (EMC) · T7 endonuclease I

Introduction

The de novo assemblage of DNA molecules is rapidly emerging as a highly powerful molecular tool to generate any desired gene sequence [1]. Artificial gene synthesis technologies do not require pre-existing DNA templates which is becoming pivotal to explore accumulating genomic and metagenomics information for which natural sequences are difficult to access. Gene synthesis is also changing established paradigms within the recombinant DNA technology field, in particular for heterologous gene expression, vaccine development, gene therapy and molecular engineering. In recent years, improvements in artificial DNA production methodologies have originated more robust, simple and cost-effective gene assembly technologies [2, 3]. In addition, exploitation of the latest high-throughput molecular technologies supported the large-scale and low-cost production of DNA sequences [4]. However, current methods for de novo gene synthesis still display significant limitations. Prevailing DNA synthesis methodologies are based on the enzymatic assembly of chemically synthesized overlapping oligonucleotides, which span the entire length of the desired sequence.

Electronic supplementary material The online version of this article (doi:10.1007/s12033-016-9957-7) contains supplementary material, which is available to authorized users.

✉ Carlos M. G. A. Fontes
cafontes@fmv.ulisboa.pt

[1] Centro Interdisciplinar de Investigação em Sanidade Animal (CIISA), Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisbon, Portugal

[2] NZYTech Genes & Enzymes, Campus do Lumiar, Estrada do Paço do Lumiar, 1649-038 Lisbon, Portugal

[3] Unité Mixte de Recherche (UMR) 7257, Centre National de la Recherche Scientifique (CNRS), Aix-Marseille Université, Architecture et Fonction des Macromolécules Biologiques (AFMB), Campus de Luminy, 163 Avenue de Luminy, 13288 Marseille Cedex 09, France

Published online: 22 June 2016

Products of the assembly reaction often contain mutations that primarily result from errors accumulated during the chemical synthesis of oligonucleotides [5]. Thus, the major drawback to high-fidelity gene synthesis remains the quality of the single-stranded DNA oligonucleotides used for gene construction [6].

Various strategies have been developed to reduce the number of errors observed in synthetic genes. Despite recent progresses achieved in the production of high quality oligonucleotides, error removal from synthetic genes during or after gene assembly remains highly required. Current methods used for the correction of DNA sequences rely on the detection and correction of mismatches present in hetero-duplex DNA molecules using mismatch-binding proteins, mismatch cleavage by endonucleases or site-directed mutagenesis [5, 6]. Although only mutation correction with site-directed mutagenesis offers total fidelity for gene synthesis, approaches incorporating mismatch active enzymes are less laborious, cost-effective and easily adapted to the large scale. Hence, there is a considerable degree of evidence suggesting that enzymatic mismatch cleavage (EMC) using endonuclease enzymes is the most promising approach to effectively reduce the levels of inaccuracies within synthetic nucleic acids. Conceptually, this method is based on mismatch cleavage through the action of specific endonucleases, in particular those that recognize and cleave DNA at mismatches in hetero-duplex molecules, followed by a second DNA fragment assembly step. In combination with DNA polymerases presenting 3′–5′ exonuclease activity, also termed proofreading DNA polymerases, this approach was shown to be very effective [7]. However, it remains largely unknown which class of mismatch-specific endonucleases are most effective for error removal during gene synthesis.

The family mismatch-specific nucleases includes: (1) single-strand-specific nucleases, such as S1 and P1 nucleases, mung bean nuclease and CEL I nuclease [8, 9]; (2) mismatch repair endonucleases, such as MutH [10]; and (3) resolvases, such as phage T4 endonuclease VII, T7 endonuclease I and *Escherichia coli* endonuclease V [5]. All these enzymes have a specific activity towards DNA molecules [7, 11, 12]. T7 endonuclease I is a bacteriophage resolvase that has been extensively used in the detection of single-base pair mismatches and mutational screening. In addition, this endonuclease was suggested to recognize all types of mismatches, including those occurring in small hetero-duplex loops [13, 14]. However, Fuhrmann et al. [7] have reported that T7 endonuclease I failed to cleave selected mispairs. Thus, the specific cleavage activity of mismatch-specific nucleases, in general, and of T7 endonuclease I, in particular, is poorly understood. Here, an integrated gene synthesis protocol was used to synthesize the *gfp* gene, which encodes green fluorescent protein. Different endonucleases were used in an EMC assay to reduce error rates and improve gene synthesis fidelity. The data revealed that T7 endonuclease I is highly effective to remove mutations accumulating during artificial gene synthesis.

Methods

Synthesis, Cloning, Expression and Purification of Mismatch Cleavage Nucleases from Different Sources

Mismatch-specific nucleases have the ability to cleave single-base pair mismatches in hetero-duplex DNA templates. In this study, six endonucleases from different sources (Table 1) were selected to understand the influence of enzymatic mismatch cleavage as an error correction tool during in vitro gene synthesis. Four of the six endonucleases chosen belong to the P1/S1 nuclease family and display strong primary sequence conservation especially at the active site, where critical catalytic residues are identical between these enzymes. The other two nucleases selected, endonuclease V from *E. coli* and T7 endonuclease I from bacteriophage T7, were reported as mismatch cleavage enzymes in different studies involving either error removal or mutation detection [7, 15]. Genes encoding the six endonucleases were synthesized in vitro with a codon usage optimized for expression in *E. coli* and cloned into pHTP1 (6HIS tag) expression vector. DNA sequences of the six de novo designed genes are reported in supplementary Table 1S. The gene encoding the maltose-binding protein was also cloned in fusion with T7 endonuclease I to promote expression solubility and protein stability. The seven recombinant plasmids (T7 endonuclease I gene was present in the fused and the unfused form) were used to transform *E. coli* BL21(DE3) cells. Expression of each one of the 7 recombinant endonucleases was achieved by adding isopropyl β-D-thio-galactopyranoside (IPTG) (1 mM final concentration) to mid-exponential phase cultures and incubation for 16 h at 16 °C. The His_6-tagged recombinant proteins were purified from cell-free extracts by immobilized metal ion affinity chromatography (IMAC) using standard methodologies [16]. Fractions containing purified proteins were analysed through SDS-PAGE.

Design of a 967 nt *lac-gfp* Gene Using Overlapping Oligonucleotides

An artificial gene with 967 nt was designed by combining the coding sequence of the green fluorescence protein (GFP) with the lacZ promoter (Table2S). Additional

Table 1 Six endonucleases selected from different sources were used for error removal in gene synthesis protocols

Mismatch cleavage endonuclease	Family	Origin	Organism
Endonuclease V	Endonuclease	Bacteria	*Escherichia coli*
Endonuclease III-wt	S1/P1 nuclease	Eubacteria	eubacterium SCB49
Endonuclease III-mut	S1/P1 nuclease	Eubacteria	eubacterium SCB49
Endonuclease I	S1/P1 nuclease	Plant	*Apium graveolens*
Endonuclease II	S1/P1 nuclease	Fungi	*Tulasnella calospora*
T7 Endonuclease I	Resolvase	Enterobacteria	Bacteriophage T7

cloning sequences (16-bp sequence at 5′ and 3′ ends) were included in the artificial gene to facilitate ligation-independent cloning (LIC). The DNA sequence encoding GFP protein was designed to display a codon usage optimized for high expression levels in *E. coli*. The 967 DNA sequence was parsed into 24 oligonucleotides of 60-mer, including 20 nt overlap regions between complementary pairs and allowing gaps of 20 nt. Oligonucleotides were synthesized by Integrated DNA Technologies (IDT) using the smallest scale available, with no purification. The sequence of the gene construct and all oligonucleotides used for the gene-assembling protocol are presented in supplementary Tables2S and 3S, respectively.

PCR Assembly to Produce Synthetic Nucleic Acids

Synthetic genes were produced by assembly PCR. Internal oligonucleotides used in the assembling PCR reaction were grouped into a pool, termed the inner oligonucleotide mixture and diluted to 125 nM stock solution. The two outer (or external) forward and reverse primers were used at higher final concentrations (800 nM). The first PCR (PCR1) was performed in a final volume of 50 µL using 1 unit of KOD Hot Start DNA polymerase (EMD-Millipore), 1× reaction buffer provided by the enzyme manufacturer, 0.2 mM dNTPs and 1.5 mM $MgCl_2$. Outer and inner oligonucleotides were used at final concentrations of 800 and 20 nM, respectively. PCR1 cycling parameters were 95 °C for 2 min, followed by 15 cycles of denaturation at 95 °C for 20 s, annealing at 55 °C for 10 s and extension at 70 °C for 15 s. A 5-µL aliquot of resulting PCR1 product was used as template to perform a second PCR (PCR2). PCR2 was performed incorporating exclusively outer oligonucleotides to ensure the production of full-length variants of the gene of interest. PCR2 was carried out in a final volume of 50 µL containing 1 unit of KOD Hot Start DNA polymerase (EMD-Millipore), 0.2 mM dNTPs, 1.5 mM $MgCl_2$ and 250 nM of each outer primer. The PCR conditions were 1 cycle at 95 °C for 2 min, 30 cycles at 95 °C for 20 s, 60 °C for 10 s and 70 °C for 20 s. Amplified nucleic acids from PCR2 were visualized by agarose gel electrophoresis and purified using silica-based columns.

Endonuclease Activity Assay

The 967 nt synthetic PCR product obtained as described above was used to analyse the cleavage activity of the seven recombinant endonucleases. Enzymatic activity was tested by incubating 25 ng of the nucleic acid in a standard reaction buffer (50 mM NaCl, 10 mM Tris–HCl, 10 mM $MgCl_2$, 1 mM $ZnCl_2$, 1 mM DTT, pH 7.9) and 1 µL (5 pmol) of each recombinant endonuclease, at 25 °C for 1 h. The endonuclease activity of three commercial enzymes was also accessed following the suppliers recommendations. The enzymes used as controls were S1 nuclease (ThermoFisher Scientific), T7 endonuclease I (New England Biolabs) and CorrectASE (Invitrogen) and were incorporated in the reaction at a 1 µL volume. After incubation–digestion, reaction products were heated for 20 min at 65 °C to inactivate nuclease activity and resulting nucleic acids integrity analysed through agarose gel electrophoresis (1.5 % w/v). The efficacy of the different enzymes to degrade 50 ng of plasmid DNA (pUC18) in quantities ranging from 13.5 to 0.5 pmol was subsequently evaluated.

Error removal by Enzymatic Cleavage of DNA Mismatches

An enzyme treatment step involving the use of mismatch cleavage nucleases was designed to increase the percentage of error-free DNA fragments resulting from PCR assembly reactions. Thus, the products resulting from PCR2 were used in an enzymatic mismatch cleavage (EMC) assay. Resulting nucleic acids were employed as template in a final PCR reaction (PCR 3) to ensure that only DNA fragments with correct sequences were amplified. The complete workflow of the protocol employed for gene synthesis and enzymatic error removal is presented in Fig. 1. To produce incorrect impairment between DNA bases, also termed DNA mismatches, which act as substrates for mismatch endonuclease enzymes, PCR2 products were denatured and re-annealed (Fig. 1). Thus, PCR2 products were diluted to 25 ng/µL in standard reaction buffer (50 mM NaCl, 10 mM Tris–HCl, 10 mM $MgCl_2$, 1 mM $ZnCl_2$,1 mM DTT, pH 7.9), and DNA was denatured at 98 °C for 2 min and slowly re-hybridized by reducing the temperature down to 4 °C in order to obtain

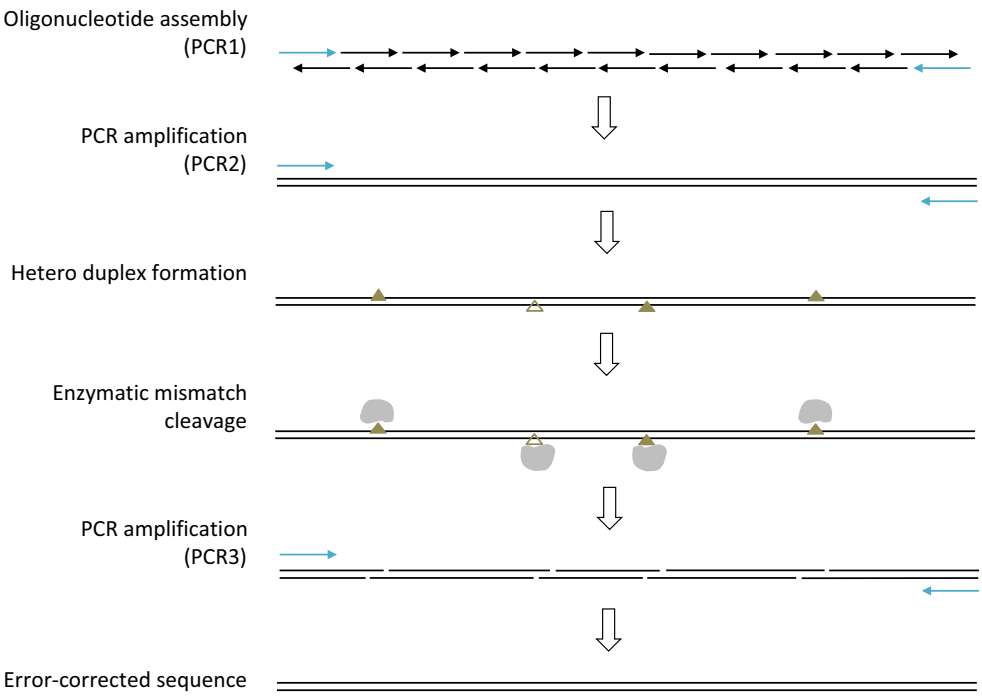


Fig. 1 Gene synthesis workflow including an endonuclease mismatch cleavage assay. Overlapping oligonucleotides are assembled (PCR1) and used as template for a second PCR reaction (PCR2) to build the full-length nucleic acid. A denaturation–renaturation step is used to form hetero-duplex DNA containing mismatches. DNA mismatches are recognized and cleaved by endonucleases. Using the digestion reaction as template, a third PCR reaction is used (PCR3) to recover the error-corrected DNA fragment

hetero-duplex nucleic acids. Samples remained at 4 °C for 5 min, followed by 5 min at 37 °C and a final step at 4 °C. To proceed with the EMC reaction, 10 µL of the re-annealed DNA was incubated with 1 µL of each endonuclease studied. Precisely 13.5 or 2.7 pmol, the two T7 endonuclease I recombinant derivatives were used in the cleavage reaction. Commercial enzymes used as controls, namely T7 endonuclease I-control 1 (New England Biolabs), CorrectASE-control 2 (Invitrogen), were used at a 1 µL, thus at an unknown concentration. A negative reaction that did not incorporate nucleases but contained exclusively hybridized DNA was also incubated with 1 µL of 1x reaction buffer. Reactions were allowed to proceed for 1 h at 25 °C. After incubation, reactions were heated for 20 min at 65 °C to inactivate the nucleases. A final PCR reaction (PCR3) was performed combining 2 µL of digestion reaction under similar conditions to PCR2 and as described above. PCR3 cycling conditions were of 1 cycle at 95 °C for 2 min and 30 cycles at 95 °C for 20 s, 60 °C for 10 s and 70 °C for 20 s. Synthesized genes were gel-purified following standard protocols.

Functional Analysis and Sequencing of Synthetic *gfp* Gene

The error-removal efficacy of recombinant endonucleases during de novo gene synthesis of the *gfp* gene was evaluated in a functional assay and by DNA sequencing. The fluorescence of GFP protein allows for a simple and expedite assessment of the success of gene synthesis, as it is likely that DNA fragments accumulating mutations will lead to the production of non-fluorescent GFP derivatives. Thus, fluorescence, visible under UV or blue light, may indicate that the resulting genes do not include mutations in the DNA sequence. After gene synthesis, full-length *gfp* gene constructs were cloned into pHTP0 cloning vector using the NZYEasy cloning kit (NZYTech, Ltd), according to the manufacturer's conditions. Recombinant plasmids were transformed into *E. coli* DH5α competent cells that were grown in LB agar plates containing 200 µg/mL ampicillin and IPTG (0.1 mM final concentration) to induce the expression of the GFP protein. After overnight incubation at 37 °C, a functional assay for the initial qualitative evaluation of the integrity of the artificial *gfp* gene sequence was performed by counting the number of fluorescent versus non-fluorescent colonies grown in the LB agar plates. In addition, 32 colonies of each endonuclease treatment were randomly selected for DNA sequencing, without regard to GFP expression. In total, 224 (32 × 7 treatments) plasmids were purified from the bacterial pellet using a silica-based protocol and the integrity of resulting nucleic acids quantified by Sanger sequencing.

Results

Cleavage Activity of Recombinant Endonucleases

In this study, six endonucleases known to present mismatch cleavage activity were selected to identify the most

efficient and accurate enzyme to incorporate in a gene synthesis protocol and contribute to error removal (Table 1). Some of the selected enzymes, such as T7 endonuclease I, display the capacity to cleave mismatch regions in hetero-duplex nucleic acids and have been applied in mutation screening [11, 15] and to a lesser degree in repair systems applied in artificial gene synthesis [2, 7]. The endonucleases were of prokaryotic or eukaryotic origins (Table 1). A mutant derivative of endonuclease III from *Eubacterium* SCB49, which has a redesigned active site presenting conserved residues observed in endonuclease I from plants, was also produced for these studies. Thus, the genes encoding the 6 different endonucleases were designed with a codon usage optimized for high expression in *E. coli* and synthesized artificially following conventional protocol [17]. The six artificial genes were cloned into pHTP1 expression vector. This vector incorporates a 6-His tag to the N-terminus of the recombinant protein. Recombinant T7 endonuclease I was also expressed in fusion with an N-terminal maltose-binding protein domain and an internal 6-His tag. The seven recombinant proteins were expressed in *E. coli* and purified (Fig. 2). All five prokaryotic enzymes were expressed in the soluble form by *E. coli*, which failed to produce the two eukaryotic proteins at significant levels. These enzymes presented a molecular mass in agreement with that deduced from primary sequence. The proteins from plant or fungal origin were found predominantly in the form of inclusion bodies. Inclusion bodies occur as a result of intracellular accumulation of partially folded expressed proteins which aggregate through non-covalent hydrophobic or ionic interactions and are commonly observed for several recombinant polypeptides expressed at high levels in *E. coli* [18]. Thus, considering that *E. coli* cells are unable to produce soluble forms of the two eukaryotic endonucleases analysed in this study, these enzymes were excluded from further analysis.

To determine the capacity of various endonucleases to cleave DNA molecules, the five recombinant prokaryotic proteins were incubated with a homogeneous *gfp* DNA fragment and integrity of resulting nucleic acid was evaluated through agarose gel electrophoresis. The efficacy of the recombinant nucleases was compared with those of three commercial endonucleases, T7 endonuclease I (New England Biolabs)-control 1, CorrectASE (Life Science technologies)-control 2 and S1 nuclease (ThermoFisher Scientific)-control 3. Nuclease activity was analysed through the digestion of 25 ng of a 967-nt PCR product with 5 pmol of each recombinant enzyme at 25 °C for 1 h. The data, presented in Fig. 3A, revealed that only the T7 endonucleases I recombinant derivatives (6HIS and MBP) displayed apparent nuclease activity, presenting identical cleavage patterns when compared with control enzymes T7 endonuclease I and CorrectASE. Activity displayed by endonuclease V is difficult to interpret and suggests an increase in the size of the nucleic acid. In addition, recombinant endonuclease V, endonuclease III-wt and endonuclease III-mut, present no nuclease activity under the reaction conditions established here. These enzymes were excluded from the titration studies presented below. In order to define more precisely, the optimal concentration of recombinant T7 endonuclease I required to efficiently cleave double-stranded DNA fragments resulting from artificial gene synthesis; six different quantities of the two recombinant forms of the enzyme (varying from 0.5 to 13.5 pmol) were used to cleave 50 ng of plasmid DNA. The data, presented in Fig. 3B, suggest that the optimal quantity of the two recombinant nucleases varied from 2.7 to 13.5 pmol. Interestingly, the nuclease activity of the two recombinant T7 endonuclease I derivatives when used at

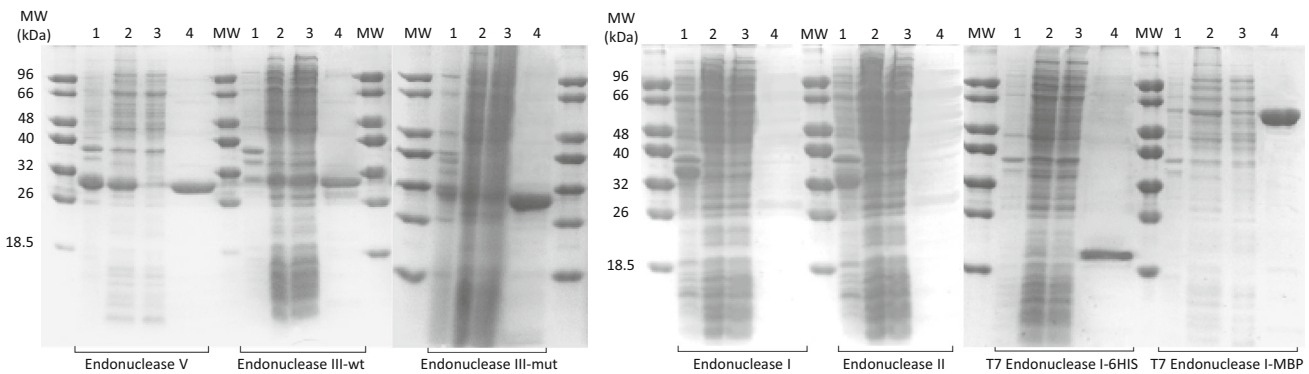


Fig. 2 Recombinant expression and purification of DNA endonucleases in *Escherichia coli*. Seven endonucleases were purified through IMAC from *E. coli* cells and protein homogeneity evaluated by SDS-PAGE. *Lanes 1* insoluble protein cellular extract; *Lanes 2* soluble protein cellular extract; *Lanes 3* protein fraction not retained by the affinity column; *Lanes 4* purified recombinant endonucleases. Sizes of molecular mass protein markers are shown. The names of the seven recombinant proteins are displayed below the corresponding SDS-PAGE figure

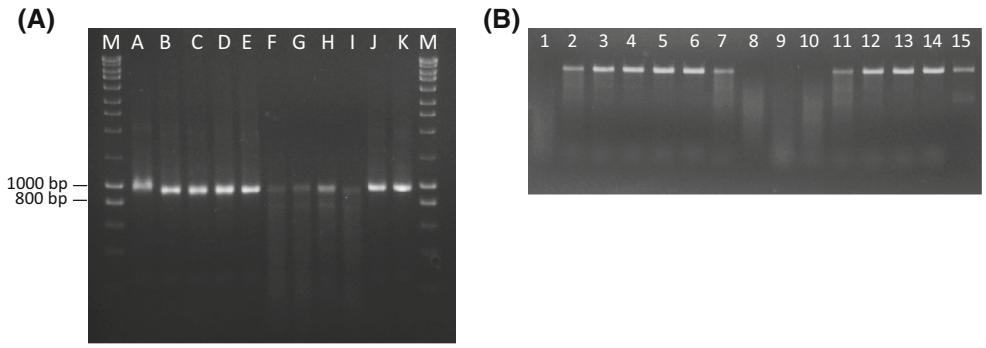


Fig. 3 Activity of recombinant endonucleases expressed in *E. coli*. **a** The capacity of different endonucleases to affect the integrity of a 967-nt PCR product was evaluated. *Lane A*, endonuclease V; *lane B*, endonuclease III-wt; *lane C*, endonuclease III-mut; *lane F*, T7 endonuclease I-6HIS; *lane G*, T7 endonuclease I-MBP; *lane H*, T7 endonuclease I-control 1; *lane I*, CorrectASE-control 2; *lane J*, S1 nuclease-control 3 and *lanes D, E, K* negative control reactions, which do not incorporate endonuclease enzymes. *Lanes M* NZYDNALadder III (NZYTech). **b** Effect of recombinant endonuclease concentration on enzyme activity (*lanes 1–6* T7 endonuclease I-6HIS; *lanes 9–14* T7 endonuclease I-MBP). Plasmid DNA (50 ng pUC18) was digested with the two T7 endonuclease I derivatives used in different quantities (*lanes 1, 9*—13.5 pmol; *lanes 2, 10*—2.7 pmol; *lanes 3, 11*—1.4 pmol; *lanes 4, 12*—1.1 pmol; *lanes 5, 13*—0.8 pmol; *lanes 6, 14*—0.5 pmol) for 1 h at 25 °C. Efficiency of recombinant endonucleases was compared with two commercial enzymes (*lane 7*—T7 endonuclease I-control 1; *lane 8*—CorrectASE-control 2). The negative reaction was performed without enzyme (*lane 15*). *Lane M* NZYDNALadder III (NZYTech)

the above quantities was similar to the two commercial enzymes.

Error Removal by Enzymatic Cleavage of DNA Mismatches

To investigate the capacity of the two T7 endonuclease I recombinant derivatives to remove DNA mismatches arising during artificial gene synthesis, the two enzymes were incorporated in the gene synthesis protocol developed here (Fig. 1). The protocol was used to synthesize *gfp*, which encodes a 27-kDa green fluorescent protein (GFP) that exhibits bright green fluorescence when exposed to light in the blue to ultraviolet range. The repairing step incorporated in the gene synthesis protocol occurs after PCR2 and is preceded by a denaturation–renaturation cycle required to produce hetero-duplex DNA molecules incorporating DNA mismatches (Fig. 1). The resulting nucleic acids were incubated with enzymes expressing nuclease activity in order to remove DNA mismatches and the integral gene recovered in a final PCR assembly phase (PCR3). The gene encoding the GFP protein was optimized for expression in *E. coli* and is controlled by an upstream lacZ promoter to ensure the expression of GFP protein in *E. coli* DH5α cells. To assemble the *gfp* gene, 24 overlapping oligonucleotides 60 nt length were designed with an overlap region of 20 nt and a gap of 20 nt and assembled through PCR. Efficacy of assembly reactions was analysed by agarose gel electrophoresis of nucleic acids resulting from PCR1 and PCR2. The data, presented in Fig. 4a, reveal that although no band is apparent for PCR1, a clear and specific band of the correct size is observed for PCR2.

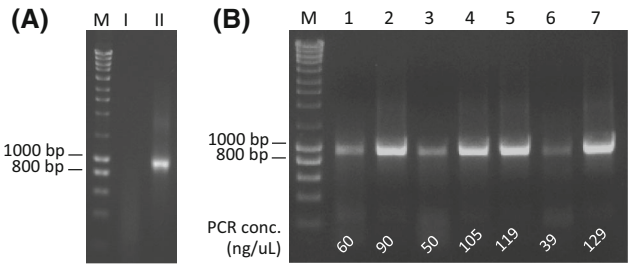
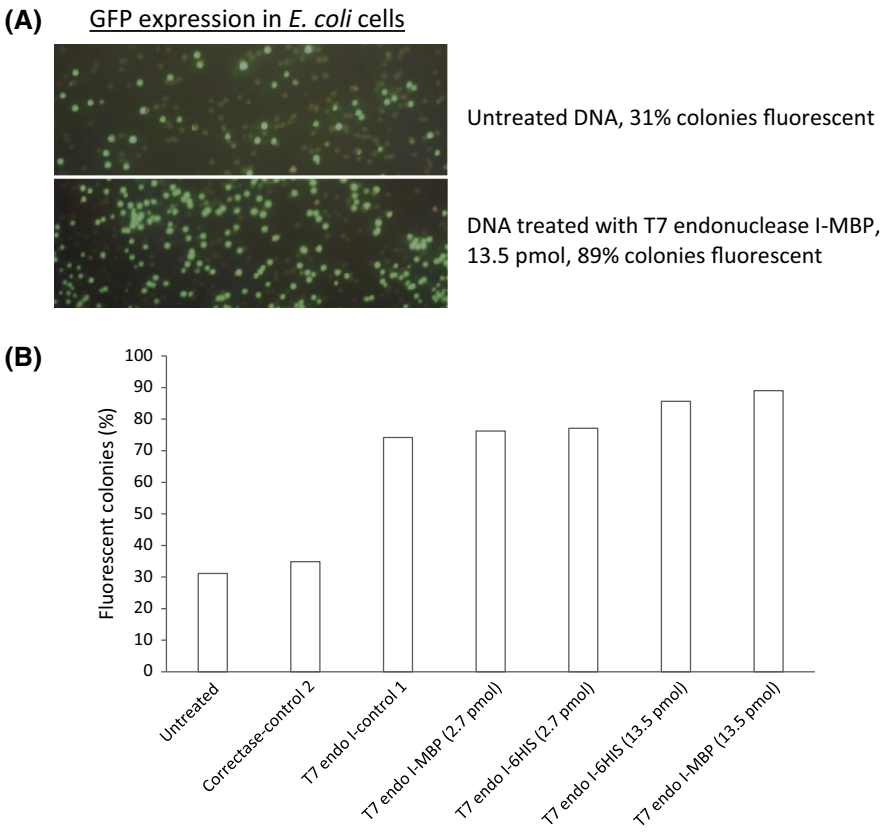


Fig. 4 Gene synthesis of *gfp* gene was performed using a set of four step reactions that include an additional error removal step. **a** Efficiency of oligonucleotide assembly, PCR1 (*lane I*), and PCR amplification, PCR2 (*lane II*), reactions as evaluated by agarose gel electrophoresis. **b** Integrity of PCR3 products obtained after mismatch cleavage and yield of each PCR product obtained. *Lane 1* T7 endonuclease I-6HIS (13.5 pmol); *lane 2* T7 endonuclease I-6HIS (2.7 pmol); *lane 3* T7 endonuclease I-MBP (13.5 pmol); *lane 4* T7 endonuclease I-MBP (2.7 pmol); *lane 5* T7 endonuclease I (NEB)-control 1; *lane 6* CorrectASE (Invitrogen)-control 2; and *lane 7* negative reaction with no enzyme. *Lanes M* NZYDNALadder III (NZYTech)

Subsequently, formation of hetero-duplex DNA was promoted through the method described by Carr, 2004 [19], by denaturation of the assembled DNA and slowly stimulating a hybridization reaction. To remove DNA mismatches from resulting nucleic acids, hybridized gene products were incubated with different mismatch cleavage endonucleases. The two recombinant T7 endonuclease I derivatives were tested at two enzyme quantities per reaction (13.5 and 2.7 pmol), and their cleavage activities compared with two control commercial proteins (T7 endonuclease I and CorrectASE). After endonuclease cleavage, reaction products were immediately used in PCR3 to recover the full length and corrected DNA sequence (Fig. 1). The



Enzymatic mismatch cleavage	no	yes	yes	yes	yes	yes	yes
Fluorescent colonies (%)	31	35	74	76	77	86	89
Fluorescent colonies	594	15	1528	1516	1558	955	819
Analysed clones	1906	43	2059	1988	2021	1115	920

Fig. 5 Analysis of GFP activity expressed by *E. coli* colonies derived from *gfp* genes artificially synthesized in the presence of different endonucleases. **a** Illustrative representation of GFP activity expressed by *E. coli* colonies resulting from expression of a synthetic *gfp* gene synthesized through a protocol not incorporating (with no error correction) or incorporating (with T7 endonuclease I-MBP, 13.5 pmol) and endonuclease treatment step. Colonies expressing GFP protein are *green* and *white* colonies correspond to the absence of GFP expression. **b** The graph above represents the percentage of colonies expressing GFP activity that were transformed with a gene that was subjected to a treatment with different endonucleases. The table below presents the raw data collected to produce the graph

results, presented in Fig. 4b, revealed that when compared with the control PCR3 products generated from a PCR reaction using a template hetero-duplex DNA not exposed to a nuclease enzyme, only three PCR3 reactions contain a significantly lower percentage of nucleic acids. These PCR reactions result from amplification of template generated after treatment with the recombinant T7 endonuclease I derivatives used at higher concentrations and the commercial enzyme CorrectASE. These results suggest that the gene products digested by these three enzymes may contain a reduced number of errors, when compared with untreated gene fragments.

Although gene products resulting from PCR3 after treatment with recombinant nucleases used at higher concentrations may contain fewer errors, it is possible that all enzymes may have contributed to improve the fidelity of the gene synthesis process. To test this, all 7 fragments resulting from PCR3 reaction were cloned into pHTP0 and the resulting plasmids were used to transform *E. coli* cells. Fidelity of gene synthesis was initially evaluated by detecting the activity of GFP in bacterial colonies derived from the transformation reaction. Expression of GFP protein was induced by adding IPTG into LB agar plates and activity detected under blue light (Fig. 5a). The data, presented in Fig. 5a, revealed an improvement in the number of fluorescent colonies that appeared in plates generated from treated gene products when compared with plates resulting from the transformation with untreated nucleic acids. As shown in Fig. 5b, only 31 % of colonies resulting from transformation with synthetic *gfp* gene not subjected to an error removal step exhibited fluorescence. In contrast, the proportion of fluorescent colonies was increased by 2.87-fold (from 31 to 89 %) when the synthetic gene was previously incubated with high concentrations of the recombinant T7 endonuclease I-MBP (Fig. 5b). Percentage of fluorescent colonies generated from enzyme-treated DNA was higher than the control reactions and ranged from 35 to 89 %. Thus, although all enzymes seem to function effectively to remove errors accumulated during gene synthesis, the data suggest that recombinant T7

Table 2 Error analysis of synthetic *gfp* gene with and without error correction

Error type	Untreated	T7 endo I (6HIS), 270 ng	T7 endo I (MBP), 270 ng	T7 endo I (control 1)	CorrectASE (control 2)
Deletion	68	16	4	18	11
A	12	1	1	4	2
T	21	4	1	1	3
C	9	9	2	5	2
G	26	2	0	8	4
% Deletions	75.6	61.5	33.3	81.8	55
Insertion	2	0	0	0	1
A	0	0	0	0	0
T	0	0	0	0	0
C	0	0	0	0	0
G	2	0	0	0	1
% Insertions	2.2	0	0	0	5
Substitution	20	10	8	4	8
Transition	5	0	0	1	8
G/T to A/T	3	0	0	1	3
A/T to G/C	2	0	0	0	0
Transversion	15	10	8	3	5
G/C to C/G	5	2	3	0	0
G/C to T/A	7	8	5	2	3
A/T to C/G	0	0	0	0	2
A/T to T/A	3	0	0	1	0
% Substitutions	22.2	38.5	66.7	18.2	40
Clones sequenced	27	27	29	27	17
% Clones with errors	88.9	51.9	31	55.6	70.6
Total errors	90	26	12	22	20
Bases sequenced	26109	26109	28043	26109	16439
Error frequency (per kb)	3.45	0.99	0.43	0.84	1.22

endonuclease I derivatives used at higher concentrations (13.5 pmol) constitute the most efficient enzyme treatments to reduce the percentage of mutations (Fig. 5). Diluted versions of the recombinant T7 endonuclease I (2.7 pmol) presented an identical but slightly reduced ratio of fluorescent clones ($\sim$76 %). The results suggest that detection of fluorescence constitutes a simple measurement of the success of *gfp* gene synthesis allowing to distinguish "putative error-free clones" (fluorescent colonies) from "error clones" (white colonies).

Error Frequency of Clones Treated with Mismatch Endonucleases

Data presented above suggest that treatment with mismatch cleavage enzymes improves the fidelity of gene synthesis. However, presence of conserved mutations within *gfp* is not detected using the qualitative functional assay that evaluates GFP activity in bacterial colonies per se. Thus, confirmation of the integrity of the nucleic acids resulting from all enzyme treatments was further performed by sequencing. Plasmid DNA from a total of 32 colonies of each one of the seven treatments, randomly selected irrespectively of presenting GFP expression, was isolated and sequenced. Together, 224 *gfp* genes were completely sequenced in both strands to identify DNA errors that surpassed mismatch cleavage treatments. The data, presented in Table 2 and Fig. 6 (data from the diluted versions of T7 endonuclease I derivatives is not shown), confirmed that treatment with mismatch cleavage nucleases dramatically reduced the error frequency observed in synthetic *gfp*. Overall, the error rate, expressed by number of errors per kb of synthetic DNA, was reduced from 3.45 to 0.43. This represents a eightfold reduction in the mutation frequency as a result of incubation with T7 endonuclease I-MBP. In general, almost all types of mutations were observed in synthetic genes, with exception of insertions with A, T and C nucleotides (Table 2). The type of errors identified in

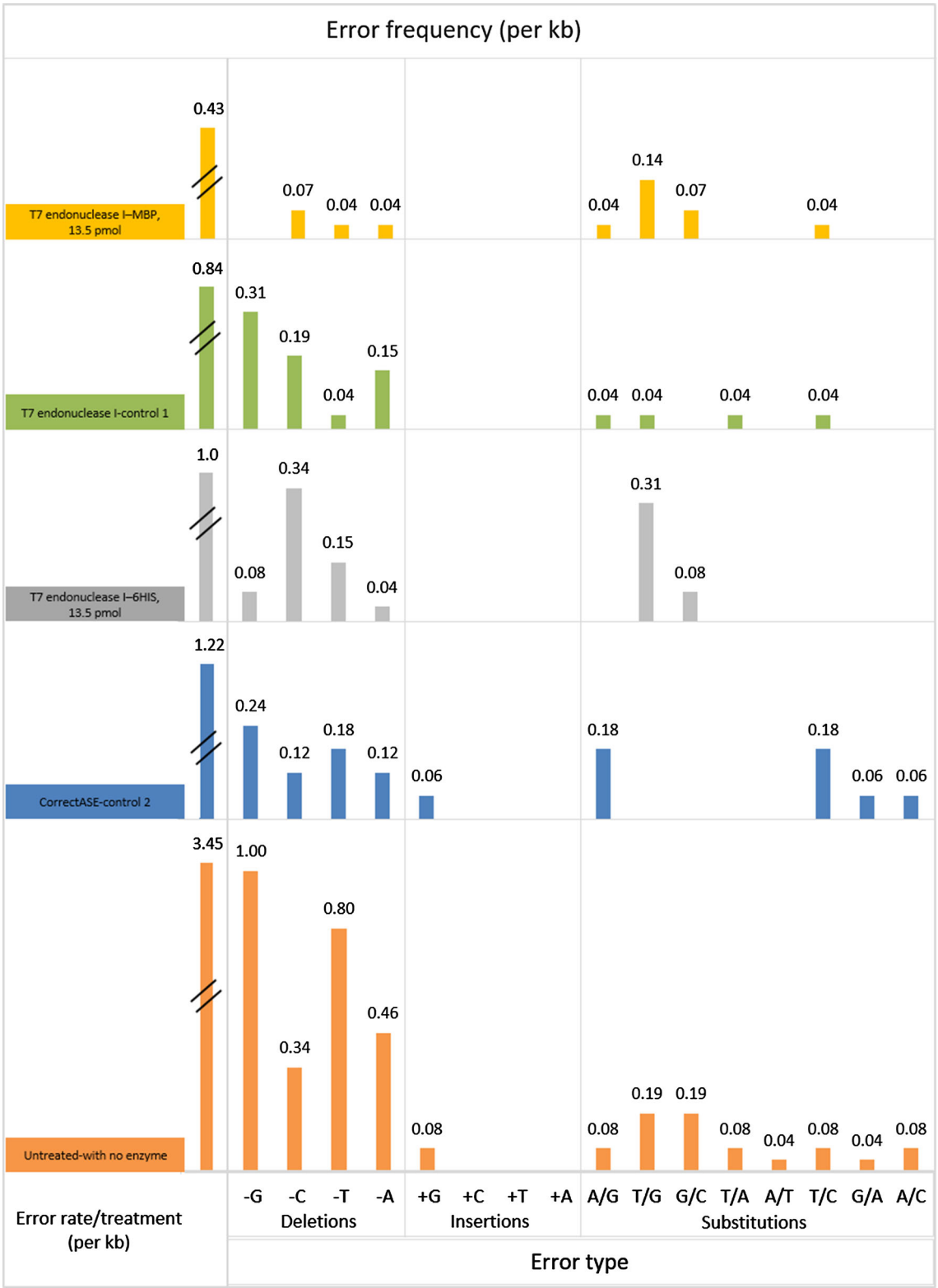


Fig. 6 Analysis of error-removal efficiency of T7 endonuclease I. Representation of error frequency per kb for untreated synthetic genes and nucleic acids artificially synthesized following a protocol incorporating an error correction step using endonucleases. The error frequency, expressed in number of mutations per kb, reflects the efficacy of each enzyme to remove errors accumulated during de novo gene synthesis. Error rate for each treatment was also calculated, and it is shown in the *left side* of the chart

Table 3 Localization of errors within *gfp* synthetic gene before and after treatment with endonucleases

Treatments	5′-end (first 60 nt)	core gene	3′-end (last 60 nt)
Untreated, with no enzyme	1	98	1
T7 endonuclease I-MBP, 270 ng	8	84	8
T7 endonuclease I-6HIS, 270 ng	31	65	4

synthetic genes was different when the nucleic acid was exposed to different enzymes, although deletions and substitutions generally predominated (Fig. 6). Significantly, untreated genes presented a higher frequency of deletions, while this type of mutations was mostly reduced in synthetic genes exposed to the enzyme treatments. For example, error correction using T7 endonuclease I-MBP reduced the presence of single deletions (per kb) by 17-fold. When the type of deletion was analysed, the reduction was of 4.5-fold for deletion of C, 21-fold for deletion of T, 12-fold for deletion of A and no deletions of G's were detected. Although not of this magnitude, similar levels of reduction in deletions were also verified for other enzyme treatments. To evaluate the location of the errors that survived the enzymatic mismatch cleavage, the distribution of mutations within *gfp* synthetic gene was analysed. The data, displayed in Table 3, suggest that the errors accumulated within untreated synthetic genes were mostly located in the core sequence. In contrast, errors which survived to error correction assays seem to be spread along the entire gene product.

Discussion

Gene synthesis is a powerful tool to create de novo DNA fragments irrespective of length and sequence. However, recurrent error rates resulting from different gene synthesis protocols have been a significant drawback to the efficient construction of DNA fragments [4]. Errors accumulated in synthetic genes can come from many sources. However, it is now well established that usage of imperfect synthetic oligonucleotides is the principal cause of low fidelity gene synthesis [5, 20]. Chemical synthesis of oligonucleotides is rarely 100 % efficient [6]. Deletions and insertions can occur in primers with a frequency of 0.5 and 0.4 % per position, respectively [6]. An improvement in the quality of oligonucleotides used for gene synthesis may result from including extra purification steps that reduce the percentage of truncated or extended molecules. However, besides the high cost of these oligonucleotides, current purifications available do not offer 100 % fidelity. In addition, enzymatic gene assembly can also cause the incorporation of mutations in synthetic genes. DNA polymerases can amplify gene products with mistakes, which are replicated during gene construction. Thus, gene assembly is also error-prone although these errors are less frequent than errors resulting from incorrect oligonucleotide synthesis [5]. Thus, there is a clear need to develop novel and effective methods to correct mutations resulting from artificial gene synthesis. Although different alternatives that minimize the number of mutations in artificial nucleic acids were previously reported, it is clear that more research is required to identify efficient enzymes to correct mutations occurring during artificial gene synthesis.

Here, the efficacy of mismatch cleavage endonucleases to remove incorrect impairments of DNA strands, thus providing a mechanism to reduce mutations in artificial genes, was evaluated. Phage T7 endonuclease I was shown to present a high capacity to cleave DNA fragments. This enzyme was expressed in *E. coli* in two variants containing or not an additional MBP fusion partner. Overall data presented here revealed a higher correctase activity for the T7 endonuclease I-MBP fusion protein, suggesting that the 45 kDa MBP tag may improve the folding and provide further stabilization to the associated nuclease domain. In addition, T7 endonuclease I-MBP reduced by eightfold error frequency in synthetic genes when compared with untreated samples. Deletion, insertions and substitutions were observed in all synthetic genes generated irrespective of the enzymatic treatment. For untreated samples, the most frequent mutations observed in synthetic genes were single deletions (75.6 %), followed by substitutions (22.2 %) and only 2.2 % of insertions. Deletions were also observed in high proportions for almost all treatments, except when T7 endonuclease I-MBP was used. Thus, this enzyme activity very effectively reduces the number of deletions in nucleic acids and in this case nucleotide substitution predominate (66.7 %). Overall, the data suggest that oligonucleotide truncation as a result of inefficient chemical synthesis leads to the accumulation of a significant proportion of deletions in the artificial genes. However, it seems that T7 endonuclease I-MBP is remarkably effective in removing single deletions in nucleic acids. Identical endonuclease activity of T7 endonuclease I was reported by Tsuji & Niida [14] in studies involving mutational screening.

Although mismatch-cleavage enzymes were effective in reducing the percentage of mutations observed in synthetic genes, data presented here confirm that inclusion of an enzymatic error removal step in gene synthesis protocols is unable to completely abolish the number of errors in

resulting artificial nucleic acids. Formation of hetero-duplex DNA after the PCR assembly steps of the gene synthesis protocol is a key to produce the required mismatches that will be targeted by the corrective nucleases. It is clear that a fraction of DNA sequences containing mutations will re-anneal and thus will not form mismatches that result from the re-annealing with sequences containing no mutations. Eventually, a cycling mismatch corrective step where hetero-duplex DNA and enzyme treatments would occur in several cycles could contribute to improve the efficacy. For this, thermal tolerant mismatch-cleavage nucleases would be required. In addition, it is possible that complete abolition of the number of mutations observed in synthetic genes may require more effective enzymes. Nevertheless, data presented here suggest that the beneficial effect of adding a mismatch removal step in gene synthesis protocols is highly dependent on enzyme concentration. Lower mutation frequencies were observed when synthetic genes were treated with 13.5 pmol of T7 endonuclease I. Thus, the lower efficacy revealed by the commercial T7 endonuclease I mixture could result from an inadequate amount of enzyme used in the treatment reaction; concentration of commercial enzymes is unknown, and a volume of 1 µl of enzyme was employed for mismatch cleavage. Significantly, a higher proportion of mutations that survived enzymatic mismatch cleavage were observed at the ends of the gene. These results suggest that some of the errors that accumulate in the final gene were introduced during the final PCR amplification and were carried by outer primers used in PCR3. These data suggest that errors present in the outer primers used for PCR assembly will be very difficult to remove from synthetic genes.

T7 endonuclease I primarily resolves four-way junctions generated by both homologous and site-specific recombination reactions by simultaneously introducing two nicks on the two non-crossing strands at 5′ sides of the junctions [21, 22]. However, it is well known that this enzyme presents a broad substrate specificity which allowed its use in a variety of biotechnological applications [22, 23]. Data presented here revealed that fidelity of *gfp* synthetic gene was strongly improved when an additional step of enzymatic cleavage with T7 endonuclease I-MBP was integrated in the gene synthesis protocol. Error frequencies (mutations/kb) were reduced from 3.45/kb (untreated) to 0.43/kb for samples treated with this mismatch cleavage enzyme. This improvement is related with the specific mismatch cleavage activity of T7 endonuclease I. This endonuclease most possibly recognizes the incorrect impairment of double DNA strands and cleaves DNA near to mismatches in both strands, resulting in the creation of 5′-end in DNA fragments. The 5′-end exposed phosphate groups are important substrates for 3′–5′ exonuclease digestion [7, 15]. The combination of T7 endonuclease I (DNA mismatch cleavage enzyme) with *Kod* DNA polymerase that has a strong 3′–5′ exonuclease activity, used to generate proofreading activity, strongly contributed to reduce mutations in artificial genes. These nucleic acids were re-assembled into a full-length gene sequence through a final PCR that enrich the error-free DNA fragments in the assembly mixture.

In conclusion, inclusion of an enzymatic treatment step during the production of synthetic nucleic acids leads to a dramatically increment in the fidelity of artificial DNA sequences generated using PCR assembly methods. Thus, the screening of integral genes from a pool of synthetic genes is facilitated by the incorporation of mismatch cleavage enzymes during gene synthesis. This approach reduces the dependence of gene synthesis fidelity on the quality of oligonucleotides used as initial templates for PCR assembly. Moreover, error removal using T7 endonuclease I derivatives leads to a more cost-effective gene synthesis and allows a simpler and quicker identification of error-free synthetic DNA products. In summary, by presenting novel evidence on the capacity of T7 endonuclease I to improve the fidelity of gene synthesis, this work opens novel avenues to explore the extraordinary potency of current gene synthesis technologies, an increasingly valuable source of nucleic acids for both fundamental and applied research.

Funding Ana Filipa Sequeira was supported by Fundação para a Ciência e a Tecnologia (Lisbon, Portugal) and NZYTech through the individual fellowship SFRH/BD/51602/2011.

Compliance with Ethical Standards

Conflict of interest The authors declare competing financial interests since NZYTech provides gene synthesis services. Renaud Vincentelli declares no competing interests.

References

1. Chao, R., Yuan, Y., & Zhao, H. (2014). Recent advances in DNA assembly technologies. *FEMS Yeast Research,*. doi:10.1111/1567-1364.12171.
2. Currin, A., Swainston, N., Day, P. J., & Kell, D. B. (2014). SpeedyGenes: An improved gene synthesis method for the efficient production of error-corrected, synthetic protein libraries for directed evolution. *Protein Engineering, Design & Selection, 27*(9), 273–280. doi:10.1093/protein/gzu029.
3. Zampini, M., Stevens, P. R., Pachebat, J. A., Kingston-Smith, A., Mur, L. A., & Hayes, F. (2015). RapGene: A fast and accurate strategy for synthetic gene assembly in *Escherichia coli*. *Scientific Reports, 5*, 11302. doi:10.1038/srep11302.
4. Kosuri, S., & Church, G. M. (2014). Large-scale de novo DNA synthesis: Technologies and applications. *Nature Methods, 11*(5), 499–507. doi:10.1038/nmeth.2918.

5. Ma, S., Saaem, I., & Tian, J. (2012). Error correction in gene synthesis technology. *Trends in Biotechnology, 30*(3), 147–154. doi:10.1016/j.tibtech.2011.10.002.
6. Tian, J., Ma, K., & Saaem, I. (2009). Advancing high-throughput gene synthesis technology. *Molecular BioSystems, 5*(7), 714. doi:10.1039/b822268c.
7. Fuhrmann, M., Oertel, W., Berthold, P., & Hegemann, P. (2005). Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. *Nucleic Acids Research, 33*(6), e58. doi:10.1093/nar/gni058.
8. Desai, N. A., & Shankar, V. (2003). Single-strand-specific nucleases. *FEMS Microbiology Reviews, 26*(5), 457–491. doi:10.1111/j.1574-6976.2003.tb00626.x.
9. Yang, B., Wen, X., Kodali, N. S., Oleykowski, C. A., Miller, C. G., Kulinski, J., et al. (2000). Purification, cloning, and characterization of the CEL I nuclease. *Biochemistry, 39*(13), 3533–3541. doi:10.1021/bi992376z.
10. Smith, J., & Modrich, P. (1997). Removal of polymerase-produced mutant sequences from PCR products. *Proceedings of the National Academy of Sciences of the United States of America, 94*(13), 6847–6850. doi:10.1073/pnas.94.13.6847.
11. Till, B. J., Burtner, C., Comai, L., & Henikoff, S. (2004). Mismatch cleavage by single-strand specific nucleases. *Nucleic Acids Research, 32*(8), 2632–2641. doi:10.1093/nar/gkh599.
12. Saaem, I., Ma, S., Quan, J., & Tian, J. (2012). Error correction of microchip synthesized genes using Surveyor nuclease. *Nucleic Acids Research, 40*(3), 1–8. doi:10.1093/nar/gkr887.
13. Babon, J. J., McKenzie, M., & Cotton, R. G. H. (2003). The use of resolvases T4 endonuclease VII and T7 endonuclease I in mutation detection. *Molecular Biotechnology, 23*(1), 73–81. doi:10.1385/MB:23:1:73.
14. Tsuji, T., & Niida, Y. (2008). Development of a simple and highly sensitive mutation screening system by enzyme mismatch cleavage with optimized conditions for standard laboratories. *Electrophoresis, 29*(7), 1473–1483. doi:10.1002/elps.200700729.
15. Huang, M. C., Cheong, W. C., Lim, L. S., & Li, M.-H. (2012). A simple, high sensitivity mutation screening using Ampligase mediated T7 endonuclease I and Surveyor nuclease with microfluidic capillary electrophoresis. *Electrophoresis, 33*(5), 788–796. doi:10.1002/elps.201100460.
16. Cheung, R. C. F., Wong, J. H., & Ng, T. B. (2012). Immobilized metal ion affinity chromatography: A review on its applications. *Applied Microbiology and Biotechnology, 96*(6), 1411–1420. doi:10.1007/s00253-012-4507-0.
17. Wu, G., Wolf, J. B., Ibrahim, A. F., Vadasz, S., Gunasinghe, M., & Freeland, S. J. (2006). Simplified gene synthesis: A one-step approach to PCR-based gene construction. *Journal of Biotechnology, 124*(3), 496–503. doi:10.1016/j.jbiotec.2006.01.015.
18. Rosano, G. L., & Ceccarelli, E. A. (2014). Recombinant protein expression in *Escherichia coli*: Advances and challenges. *Frontiers in Microbiology, 5*(April), 1–17. doi:10.3389/fmicb.2014.00172.
19. Carr, P. A. (2004). Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Research, 32*(20), e162. doi:10.1093/nar/gnh160.
20. Wan, W., Li, L., Xu, Q., Wang, Z., Yao, Y., Wang, R., et al. (2014). Error removal in microchip-synthesized DNA using immobilized MutS. *Nucleic Acids Research, 42*(12), e102. doi:10.1093/nar/gku405.
21. de Massy, B., Studier, F. W., Dorgai, L., Appelbaum, E., & Weisberg, R. A. (1984). Enzymes and sites of genetic recombination: Studies with gene-3 endonuclease of phage T7 and with site-affinity mutants of phage lambda. In *Cold Spring Harbor Symposia on Quantitative Biology* (Vol. 49, pp. 715–726).doi:10.1101/SQB.1984.049.01.081.
22. Aravind, L., Makarova, K. S., & Koonin, E. V. (2000). SURVEY AND SUMMARY: Holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Research, 28*(18), 3417–3432. doi:10.1093/nar/28.18.3417.
23. White, M. F., Giraud-Panis, M. J., Pöhler, J. R., & Lilley, D. M. (1997). Recognition and manipulation of branched DNA structure by junction-resolving enzymes. *Journal of Molecular Biology, 269*(5), 647–664. doi:10.1006/jmbi.1997.1097.