*doi: 10.1093/nar/gku405*

# Error removal in microchip-synthesized DNA using immobilized MutS

**Wen Wan[1,2,†], Lulu LI[1,2,†], Qianqian Xu[1,2], Zhefan Wang[1,2], Yuan Yao[1,2], Rongliang Wang[1,2], Jia Zhang[1,2], Haiyan Liu[1,2], Xiaolian Gao[1,2,3,*] and Jiong Hong[1,2,*]**

[1]School of Life Science, University of Science and Technology of China, Hefei, Anhui, People's Republic of China,
[2]Hefei National Laboratory for Physical Science at the Microscale, Hefei, Anhui, People's Republic of China and
[3]Department of Biology and Biochemistry, University of Houston, Houston, TX, USA

Received March 14, 2014; Revised April 17, 2014; Accepted April 24, 2014

## ABSTRACT

**The development of economical *de novo* gene synthesis methods using microchip-synthesized oligonucleotides has been limited by their high error rates. In this study, a low-cost, effective and improved-throughput (up to 32 oligos per run) error-removal method using an immobilized cellulose column containing the mismatch binding protein MutS was produced to generate high-quality DNA from oligos, particularly microchip-synthesized oligonucleotides. Error-containing DNA in the initial material was specifically retained on the MutS-immobilized cellulose column (MICC), and error-depleted DNA in the eluate was collected for downstream gene assembly. Significantly, this method improved a population of synthetic enhanced green fluorescent protein (720 bp) clones from 0.93% to 83.22%, corresponding to a decrease in the error frequency of synthetic gene from 11.44/kb to 0.46/kb. In addition, a parallel multiplex MICC error-removal strategy was also evaluated in assembling 11 genes encoding ∼21 kb of DNA from 893 oligos. The error frequency was reduced by 21.59-fold (from 14.25/kb to 0.66/kb), resulting in a 24.48-fold increase in the percentage of error-free assembled fragments (from 3.23% to 79.07%). Furthermore, the standard MICC error-removal process could be completed within 1.5 h at a cost as low as $0.374 per MICC.**

## INTRODUCTION

*De novo* gene synthesis is playing an increasingly important role in synthetic biology, systems biology and general biomedical sciences (1–9). Currently, synthesis of gene-size fragments (500∼5000 bp) typically begins with oligonucleotides (oligos) as building blocks. Oligos are synthesized on controlled-pore glass (CPG) followed by an assembly step for producing long gene fragments (10,11). The technologies for assembling these gene-size fragments into much longer synthetic DNA constructs are now fairly mature (7,11–15). Thus, the relatively high cost and low-throughput oligos synthesis have become the limitation of scaling up DNA synthesis. Currently, CPG oligos (100∼200 nt) supplied by vendors generally cost about $0.40∼$1.00/bp (16,17) and the throughput is low (from one to 1534 oligos/batch) (7,18,19). Fortunately, compared to traditional CPG oligos, the price of oligos that are synthesized on microarrays can potentially be much cheaper with higher throughput (16,20–25). One million distinct oligos can be simultaneously synthesized on a single chip (26,27), and in some cases, microarray-based methods can provide oligo pools containing about one million 60-mers for $600 (16,21,28). However, the utilization of microchip-synthesized oligonucleotides (MCp-oligos) in *de novo* gene synthesis has been hindered by several technical bottlenecks. (i) Minute quantities of each MCp-oligo (1∼10 fmol per sequence) (21). Although a large number of distinct sequences can be simultaneously synthesized on one chip, the quantity of each sequence is too low to support subsequent applications. (ii) High complexity and high background (22–24,29). After cleavage from the microchip, the oligos form a large pool containing a huge number of different sequences. The great diversity of these sequences in the pool makes subsequent gene assembly more difficult. (iii) Low fidelity of the MCp-oligos (30). Fortunately, the problems of small oligo quantities and complex oligo pool composition can be partially resolved via high-fidelity polymerase chain reaction (PCR) amplification and separation of the MCp-oligo pool into subpools via addition of primers at both ends of the synthesized oligos (16,22). Consequently, the oligo quality has become the primary obstacle to *de novo* gene synthesis.

[*]To whom correspondence should be addressed. Tel: +86 551 63600705; Fax: +86 551 63601443; Email: hjiong@ustc.edu.cn
Correspondence may also be addressed to Xiaolian Gao. Tel: +1 713 743 2805; Fax: +1 713 743 2709; Email: xgao@uh.edu
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

The quality of synthetic oligos can be improved by enhancing the efficiency of each synthesis cycle during the synthesis process and reducing the errors in the synthesized oligos after synthesis (16,22,29,31–35) (Tables 1 and 2). A new chemical strategy that optimizes reagent flows and minimizes depurination during the oligo synthesis process can reduce the corresponding side reactions and consequently improve oligo yield and quality (31). However, the quality of the oligos still cannot meet the requirement of *de novo* DNA synthesis. High performance liquid chromatography and polyacrylamide gel electrophoresis (PAGE) can also be performed to improve the fidelity of synthetic DNA (Tables 1 and 2) (22,33–35). Using these methods, approximately 90% of the impurities of incorrect size can be removed (29). However, these steps often prove ineffective for error removal in oligo pools, especially when the size of the oligos does not differ (e.g. substitution errors) or when the oligos in the pool initially vary in size. Alternatively, a hybridization and selection method requiring two microchips successfully removed errors in oligos, providing a 8.74-fold reduction in the error frequency (22) (Table 2). Since this method was based on hybridization on microchips, a quality assessment microchip was necessary, which effectively doubles the cost of each synthesis. The combination of high-throughput pyrosequencing and oligo retrieval-mediated error correction appears to be the most efficient method to date (Table 2) (36). This highly parallel method uses a robotic system to image and directly recover beads containing sequence-verified oligos based on a next-generation pyrosequencing platform. This method improves the fidelity of MCp-oligos by 500-fold compared to the initial oligo pool. However, this process is expensive, as it depends on next-generation pyrosequencing reagents and instruments.

On the other hand, error correction of synthetic DNA can also be performed after assembly (16,35,37–41). Endonucleases, which can recognize the mismatch site of DNA, combined with exonucleases have been applied in error correction (37) (Tables 1 and 2). Surveyor nuclease, a commercially available CEL endonuclease that has also been successfully used for error correction of synthetic genes (40), reduced the error frequency from one error per 526 bp to one error per 8701 bp (40). Furthermore, the ErrASE kit, another commercially available CEL-based enzyme cocktail that corrects errors in a similar fashion, reduced the error rates of DNA assembled from MCp-oligos from one error per 1500 bp to one error per 7017 bp (16) (Table 2). However, these enzymatic mismatch cleavage (EMC) methods are expensive and time-consuming for scalable multi-gene treatments and may decrease the probability of generating assembled products due to over-digestion (37). Alternatively, mismatch binding proteins have been used to remove error-containing DNA. The mismatch binding protein MutS can specifically recognize and bind to all possible single-base mismatches, as well as 1∼5 bases insertion or deletion loops, with varying affinities and functions independently of other proteins or cofactors (42,43). A *Thermus aquaticus* MutS protein (*Taq*MutS)-mediated error-correction method corrected *de novo* synthesized gene using CPG synthesized oligos at a fidelity of one error per 10 000 bp (38) (Table 1). Another modified error-correction method through consensus shuffling with *Taq*MutS reduced the error frequency of synthetic green fluorescent protein (GFPuv) gene by 3.5- to 4.3-fold, reaching a final fidelity of one error per 3500 bp (39) (Table 1). However, the current MutS-mediated error-correction methods are only validated for assembled products. Additionally, these methods typically use *Taq*MutS because it is more stable and has a lower binding affinity to perfectly matched DNA compared to *Escherichia coli* MutS (*Eco*MutS), but its binding affinity to mismatch-containing DNA is lower than that of *Eco*MutS (44,45), reducing the efficiency of these MutS-mediated error-correction methods.

These existing approaches for error correction of DNA are not suitable for low-cost and high-throughput error removal from MCp-oligos that form a complex oligo pool consisting of oligos with high error rates and varying Tm values (22,36). In this study, a low-cost, effective and improved-throughput error-removal method using immobilized cellulose columns containing a combination of two homologs of the mismatch binding protein MutS (*Eco*MutS and *Taq*MutS) was produced to generate high-quality DNA from oligos, especially MCp-oligos. After optimization of the method using various MutS-immobilized cellulose columns (MICCs) to remove errors from the *de novo* synthesized EGFP gene assembled from MCp-oligos, the method was further validated for its ability to remove errors from a soluble methane monooxygenase (sMMO) gene cluster (containing sMMO X, Y, B, Z, D, C, H and G) and the epothilone (Epo) A, B and C genes that were synthesized *de novo* from MCp-oligos on a larger scale.

## MATERIALS AND METHODS

### Chemicals and strains

All chemicals were reagent grade or higher and were purchased from Sangon Biotech Co. (Shanghai, China) unless otherwise noted. All restriction enzymes and T4 DNA ligase were obtained from Thermo Fisher Scientific Inc. (MA, USA). PrimeStar HS DNA polymerase was from TaKaRa (TaKaRa Biotechnology (Dalian) Co. Ltd, Dalian, China). Pfu DNA polymerase was from Biocolor BioScience & Technology Company (BBST, Shanghai, China). KOD Plus DNA polymerase was from TOYOBO (Osaka, Japan). *T. aquaticus* NBRC 103206 was obtained from the NITE Biological Resource Center (NBRC, Japan). *E. coli* DH5α was used as a host cell for all DNA manipulations. *E. coli* BL21 Star (DE3) (Invitrogen, Carlsbad, CA, USA) containing a protein expression plasmid was used to express the recombinant proteins. Luria-Bertani (LB) medium containing 100 μg/ml ampicillin was used to cultivate *E. coli* and to produce the recombinant proteins.

### Expression, purification and functional evaluation of the recombinant *Eco*MutS-CBM3-EGFP (eMutS) and CBM3-*Taq*MutS-EGFP (tMutS) proteins

The MutS gene from *E. coli* or *T. aquaticus* (GenBank: *Eco*MutS, HG738867.1 and *Taq*MutS, U33117.1) was amplified from the pET32-muts plasmid (a gift from Dr Tianyin Zhong) or the *T. aquaticus* genome using corresponding primers (Supplementary Table S1), respectively,

**Table 1.** The effectiveness of different error-removal methods on CPG-oligos

| Method[a] | Tool[b] | Error correction stage (bp) | Analyzed DNA length (nts) | Before correction | | After correction | | Ref. |
|---|---|---|---|---|---|---|---|---|
| | | | | Error frequency (errors per kb) | Error-free DNA ratio (%) | Error frequency (errors per kb) (fold) | Error-free DNA ratio (%) (fold) | |
| Function selection | Synthetic ORF selection vector | Assembled DNA | 717 | NA[c] | 28%[d] | NA[c] | 82%[d] (2.93-fold) | (35) |
| PAGE | PAGE | Oligos | 717 | NA[c] | 28%[d] | NA[c] | 64%[d] (2.29-fold) | (35) |
| EMC | T4 endonuclease VII | Assembled DNA (616) | 616 | 6.52 | 4.10%[d] | 1.62 (4.02-fold) | 46.9%[d] (11.44-fold) | (37) |
| EMC | *E. coli* endonuclease V | Assembled DNA (616) | 616 | 6.52 | 4.10%[d] | 1.98 (3.29-fold) | 31%[d] (7.56-fold) | (37) |
| MMC | *Taq* MutS | Assembled DNA (993) | 993 | 1.8 | NA[c] | 0.1 (18-fold) | NA[c] | (38) |
| MMC | *Taq* MutS | Fragmented DNA[q] (∼150) | 760 | 1.30 | 54%[d] | 0.3 (4.33-fold) | 93%[d] (1.72-fold) | (39) |
| MMC | *Taq* MutS | Fragmented DNA[q] (∼150) | 760 | 0.98 | 67%[d] | 0.28 (3.50-fold) | 93%[d] (1.39-fold) | (39) |

[a] EMC: enzyme-mediated correction; MMC: MutS-mediated correction.
[b] The protein and technology used in the corresponding error correction method.
[c] Not available in the literature.
[d] Percentage of active clones (contains perfect clones).

and cloned into the pET-21c vector together with the CBM3 (GenBank: HF912725.1) and EGFP (GenBank: ACX42327.1) genes, which were amplified from the pCG plasmid (46) using the primers shown in Supplementary Table S1. The resulting plasmids were termed p*Eco*MutS-CBM3-EGFP (Supplementary Figure S1a), which expresses the *Eco*MutS fusion protein (eMutS) and pCBM3-EGFP-*Taq*MutS (Supplementary Figure S1b), which expresses the *Taq*MutS fusion protein (tMutS). The expression plasmids were transformed into *E. coli* BL21 Star (DE3). Expression of the MutS fusion protein was induced in LB medium containing 1 mM isopropyl-D-thiogalactoside (IPTG). Then, the expressed MutS fusion protein was purified using a Ni-NTA affinity column according to the manufacturer's protocol (Qiagen, The Netherlands). Both MutS fusion proteins contained CBM3, EGFP, MutS and a 6-His tag. The recombinant MutS fusion proteins could be immobilized on cellulose via CBM3, and the protein purification and immobilization of the constructed fusion proteins could be monitored via the fluorescence of EGFP. Detailed procedures regarding MutS expression vector construction and MutS expression and purification are described in the Supplementary Data.

Because regenerator amorphous cellulose (RAC) slurry was used to immobilize MutS, the binding ability of RAC to MutS was also determined. The maximum adsorption of MutS per gram RAC (*Amax*) and the binding constant (*Ka*) of eMutS and tMutS to RAC were calculated based on Langmuir equation and Hanes–Woolf method as described previously (47).

To evaluate the binding properties of *Eco*MutS-CBM3-EGFP (eMutS) and CBM3-*Taq*MutS-EGFP (tMutS), 10 oligos were synthesized by Sangon Biotech Co. (Shanghai, China) (Supplementary Table S2). These oligos (A1∼A5 and B1∼B5) can form all possible single-base mismatches through the annealing of two oligos to form heteroduplexes respectively (Supplementary Table S3). According to previous studies, the mismatches corresponding to deletion or insertion errors are preferred over both *Eco*MutS and *Taq*MutS (44,45). Therefore, the binding of MutS containing an unpaired T and perfectly matched DNA is initially used to determine the nonspecific binding of MutS to perfectly matched DNA. MutS-DNA binding reactions were performed using various molar ratios between MutS (eMutS or tMutS) and the DNA (58 bp homoduplexes and unpaired T duplexes) ranging from 0:1 to 40:1. Then, binding reactions of eMutS or tMutS to the various mismatched DNA sequences were performed to determine the binding affinities of MutS to these mismatches. All of the results were evaluated via a band-shift assay (48). The details of these processes are described in the Supplementary Data.

## Construction and functional evaluation of the MICCs

To prepare the MICCs, the MutS fusion protein was immobilized on RAC slurry (46) via CBM3 by mixing the eMutS or tMutS fusion protein with the RAC slurry (Supplementary Figure S2a) (600 pmol MutS in 500 μl of the RAC slurry (20 mg/ml)) and incubating at room temperature for 10 min. Then, the MutS-immobilized RAC slurry (1 ml) was

**Table 2.** The effectiveness of different error-removal methods on MCp-oligos

| Method[a] | Tool[b] | Error correction stage (bp) | Analyzed DNA length (nts) | Before correction | | After correction | | Ref. |
|---|---|---|---|---|---|---|---|---|
| | | | | Error frequency (errors/kb) | Error-free DNA ratio (%) | Error frequency (errors/kb) (fold) | Error-free DNA ratio (%) (fold) | |
| PAGE | PAGE | Oligos (70) | 1,755 | 6.29 | NA[c] | 2.20<br>(2.86-fold) | NA[c] | (22) |
| Hybridization | Microchip | Oligos (70) | 297~1,755 | 6.29 | NA[c] | 0.72<br>(8.74-fold) | NA[c] | (22) |
| NGS | Pyrosequencing platform | Oligos (40) | 137~255 | 25 | 3.07% | 0.05<br>(500-fold) | 84.28%<br>(27.45-fold) | (37) |
| EMC | ErrASE | Assembled DNA (779) | 779 | 0.67 | 69.8% | 0.14<br>(4.79-fold) | 90%<br>(1.29-fold) | (16) |
| EMC | ErrASE | Assembled DNA (779) | 779 | 0.88 | 60% | 0.20<br>(4.40-fold) | 85.7%<br>(1.43-fold) | (16) |
| EMC | ErrASE | Assembled DNA (708~777) | 708~777 | ~4.00 | NA[c] | ~3.17<br>(1.26-fold) | 12.50% | (16) |
| EMC | ErrASE | Assembled DNA (720~732) | 720~732 | NA[c] | 6.8%~7.5%[d] | NA[c] | 26%~49%[d]<br>(3.82-fold~6.53-fold) | (16) |
| EMC | Surveyor nuclease | Assembled DNA (678) | 678 | ~1.9 | 50.20%[d] | ~0.19<br>(10-fold) | 84%[d]<br>(1.67-fold) | (21) |
| EMC | Surveyor nuclease | Assembled DNA (1134) | 723 | 1.9 | 50.20% [d] | 0.11<br>(17.27-fold) | 94%[d]<br>(1.87-fold) | (40) |
| MMC | MICC | MCp-oligos (69~118) and assembled DNA (258~260) | 720 | 11.44 | 0.93%[d] | 0.46<br>(24.87-fold) | 83.22%[d]<br>(89.48-fold) | This study |
| MMC | MICC | MCp-oligos (63~129) and assembled DNA (286~456) | 286~456 | 14.25 | 3.23% | 0.66<br>(21.59-fold) | 79.07%<br>(24.48-fold) | This study |

[a] EMC: enzyme-mediated correction; MMC: MutS-mediated correction.
[b] The primary tools applied in the corresponding error correction technology.
[c] Not available in the literature.
[d] Percentage of active clones (contains perfect clones).

added to the chromatography column (diameter × length: 0.4 cm × 7 cm, Sangon Biotech Co.) up to a length of 2 cm. After the slurry settled in the column, 1 ml of binding buffer (5 mM $MgCl_2$, 100 mM KCl, 20 mM Tris-HCl (pH 7.6) and 1 mM DTT) was added to wash away the free proteins.

Due to the differing mismatch binding specificity and capacity of eMutS and tMutS, five types of MICCs that contained eMutS (eMICC), tMutS (tMICC) or a combination of both MutS proteins, termed etMICC (e/tMICC, t/eMICC or e+tMICC; differing in the packing mode of the two MutS proteins), were constructed to evaluate their error-removal ability, as shown in Supplementary Figure S2b. In the eMICC and the tMICC, only the corresponding MutS was immobilized on the RAC slurry. In the combined etMICC column, three types of MICCs, corresponding to the three packing modes, were constructed (Supplementary Figure S2b). For the e/tMICC, 0.5 ml of tMutS-cellulose slurry was packed on the bottom of the column, followed by another 0.5 ml of eMutS-cellulose slurry to form a length of 2 cm. For t/eMICC, the MutS-cellulose slurry was packed in the reverse order compared to e/tMICC. For e+tMICC, equivalent concentrations of eMutS and tMutS were mixed, immobilized on the RAC slurry and packed on the column. For the eMICC and the tMICC, the molar ratio of DNA to MutS was 1:10 and 1:20, respectively, based on the results of the analysis of MutS-DNA binding experiments. For the etMICC, the molar ratio of DNA to eMutS and tMutS was 1:10:10 unless otherwise noted. In these experiments, cellulose columns of equivalent length (20 mg of the RAC slurry, 2 cm column length) were used.

The error-removal ability of these MICCs (eMICC, tMICC and etMICC) was evaluated based on the binding specificity to heteroduplexes in mixtures of duplex DNA (60 pmol oligos containing 59 bp heteroduplex DNA (unpaired T: '+T') (45 pmol) and 54 bp homoduplex DNA (15 pmol) (Supplementary Table S2 and S3)) as described in the Supplementary Data. Because of the different sizes of the heteroduplex and homoduplex DNA, the 59 bp heteroduplex and the 54 bp homoduplex DNA could be separated, detected and semi-quantified via PAGE. After error removal, the error-depleted DNA (eluted fractions) was collected at 80 μl/tube and analyzed via PAGE.

### Design and synthesis of the microchip-synthesized oligonucleotides

The MCp-oligo pool (EGFP pool), encoding the *egfp* gene, was used to determine the error-removal ability of each MICC and establish the optimal error-removal protocol. The 60 oligos of lengths between 69 and 118 nt in the EGFP pool were separated into six separately amplifiable subpools (Supplementary Table S4). Each subpool was defined by unique primer binding sites at each terminal of each oligo (Supplementary Table S5 and S6 and Supplementary Figure S3). The primer binding sites, which contained a *Mly* I restriction site, could be removed by *Mly* I digestion. After the error-removal efficiency of MICC system was confirmed during the synthesis of EGFP gene described above, an additional two sets of 253 MCp-oligos (sMMO) and 640 MCp-oligos (Epo A, B, C) of lengths from 63 to 129 nt were separately designed and synthesized (sMMO and Epo pool) on microchips (Supplementary Table S4). The sMMO and Epo pools were separated into 57 subpools by adding various pairs of primer binding sites to each oligo subpool (Supplementary Table S5 and S6 and Supplementary Figure S4). The details of the design of these oligos are described in the Supplementary Data, and the sequence information regarding these designed oligos is supplied in the Supplementary Data of Sequences.

Each of the above MCp-oligos, encoding the EGFP gene, the sMMO gene cluster or the Epo A, B and C genes, were individually synthesized by LC Sciences (Houston, Texas) on 4-k microchips using light-directed synthesis methods (20).

### Primer removal before gene assembly

The PCR products of the MCp-oligos were digested using *Mly* I to remove the primer region. Then, the cleaved primer sequences were removed using the UNIQ-10 oligonucleotide purification kit (Sangon, Shanghai, China) according to the manufacturer's instructions.

### Amplification and assembly of oligo pools

The oligos were released from the microchip surface, forming an oligo pool that contained a total of ∼20 picomoles of oligos as previously described (22). This oligo pool was used as the PCR template without any additional purification. PCRs were performed using KOD Plus DNA polymerase and the appropriate primers (Supplementary Tables S5 and S6). After the primer regions of the oligos were removed as described above, polymerase cycling assembly (PCA) (10), ligase chain reaction (LCR) (22) or the combination of these two methods (PCA-LCR) was performed to assemble these oligos into target fragments. Then, the assembled fragments corresponding to each gene, containing overlaps of ∼30 bp between the neighboring fragments, were assembled into full-length genes via overlapping extension PCR (OE-PCR) (49). The details of oligo amplification and assembly and OE-PCR for full-length genes are described in the Supplementary Data section.

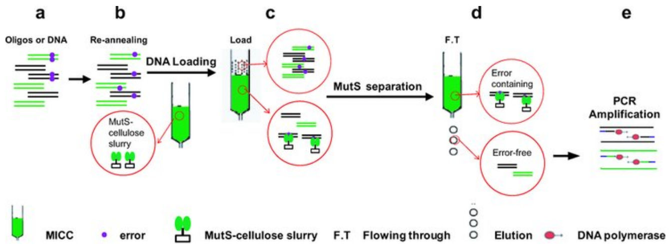


**Figure 1.** Schematic representation of removal of error-containing oligonucleotides or assembled DNA using a MICC. (**a**) Microchip-synthesized oligos or assembled DNA fragments are amplified via PCR. (**b**) Amplified oligos or DNA fragments are re-annealed to expose errors. (**c**) Re-annealed oligos or DNA fragments are loaded onto a MICC. (**d**) After elution, the error-containing oligos or DNA fragments are retained on the column, and the error-free oligos or DNA fragments elute through the column and are collected. (**e**) The collected error-free oligos or DNA fragments are amplified via PCR to generate additional material for subsequent applications.

### Error removal using a MICC

Removal of error-containing DNA was performed using a MICC as shown in Figure 1. First, the DNA was re-annealed to expose the errors as mismatches. For the re-annealing procedure, the DNA samples were diluted in 50 μl annealing buffer containing 10 mM Tris-HCl (pH 7.6), 50 mM NaCl and 1 mM EDTA to a final concentration of 50 ng/μl (approximately 1 μM for oligos and 0.3 μM for fragments). Then, the DNA samples were slowly cooled from 100°C to 25°C in a water bath. Next, 240 μl of the re-annealed DNA (diluted in binding buffer to 12.5 ng/μl) was loaded on the MICC. The error-depleted oligos were eluted in 1 ml of binding buffer and collected in fractions of 80 μl per 1.5 ml Eppendorf tube. The DNA concentration of each fraction was quantified via Nanodrop or detected via 6% PAGE when the concentration was too low to be quantified via Nanodrop. The first several collected fractions that contained error-depleted DNA served as the templates for the subsequent steps. In brief, 0.5 μl of the error-depleted DNA was used as the PCR template without any additional purification. KOD Plus DNA polymerase was used for oligo amplification, and Pfu DNA polymerase was used for fragment amplification using the appropriate primers (Supplementary Tables S5 and S6) as described in the Supplementary Data section.

### Evaluation of the efficiency of the MICC methods to remove errors during *de novo* EGFP gene synthesis

After the ability of MICCs to remove error-containing CPG oligos was confirmed, unpurified MCp-oligos that could be assembled into a 720 bp gene encoding EGFP were chosen to evaluate the ability of the MICC system to remove error-containing DNA during *de novo* gene synthesis. The error-removal process was performed on the EGFP oligos as shown in Figure 2. In brief, to improve the oligo quantity and reduce the complexity of the MCp-oligo pool, the original oligo pool cleaved from microchip was separated into subpools via differential amplifications using various primer pairs that were added at the terminals of oligos contained in each subpool (Figure 2a∼c). The oligos in each subpool were re-annealed to expose errors (Figure 2d). Then, the error-containing oligos in each subpool were removed using each individual MICC (Figure 2e). After error removal, these subpools were amplified via PCR (Figure 2f). These amplified products were used to assemble specific target DNA fragments or genes (Figure 2g and h). Each oligo subpool of EGFP was amplified using specific primer pairs (Supplementary Tables S5 and S6), re-annealed and corrected according to the methods described in the section 'Error removal using a MICC'.

To evaluate and optimize the error-removal ability of the MICC method, three MICCs (eMICC, tMICC and etMICC) were individually examined. Firstly, the error-removal efficacy of three types of combined etMICCs (e/tMICC, t/eMICC and e+tMICC), which differed in the packing mode as described above, were compared. Then, the error-removal ability of four t/eMICCs containing various molar ratios of eMutS and tMutS (1:0.5, 1:1, 1:2 and 1:3) in cellulose columns of the same length (20 mg cellulose slurry, 2 cm column length) were investigated (the molar ratios of DNA to eMutS and tMutS were 1:10:5, 1:10:10; 1:10:20 and 1:10:30, respectively). After error removal using the MICC, the error-depleted oligos were assembled into fragments via the LCR method as described in the Supplementary Data section.

To further improve the fidelity of the *de novo* synthesized EGFP genes, another round of error removal using a MICC was performed on the fragments (assembled from the error-depleted oligo subpool).

### Functional evaluation and sequencing of the synthetic EGFP gene

The error-removal efficiencies of MICCs during *de novo* synthesis of EGFP gene were evaluated via functional validation and sequencing. After these EGFP fragments (without error removal or with one or two rounds of error removal) were collected, they were fused to form full-length EGFP gene sequences as described above. Functional validation of the synthesized EGFP gene sequences was performed by counting the number of visible fluorescent colonies via a plating assay (50). In brief, the assembled EGFP full-length DNA was digested using *Nhe* I and *Xho* I and ligated to the pET-21c plasmid using T4 DNA ligase (NEB) and transformed into *E. coli* BL21 star (DE3) via electroporation (51). After the transformants were cultivated at 37°C for 10 h on a LB agar plate containing 100 μg/ml ampicillin, IPTG (0.1 mM) was sprayed on the surface of the plate to induce the expression of EGFP. The proportion of the clones with green fluorescence in the total clones (harboring synthesized EGFP genes) roughly indicated the error-removal efficiency of each MICC system.

To evaluate the error-removal efficiency via sequencing, the assembled EGFP fragments or genes were cloned into pMD18-T vectors (TaKaRa Bio, Dalian). Then, the clones were randomly selected for sequencing. The sequences of the selected clones were aligned with the EGFP DNA sequence using the BioEdit tool (http://www.mbio.ncsu.edu/bioedit/bioedit.html). The results were statistically analyzed according to previously reported methods (37,40) to quantitatively determine the error-removal efficiency.

### *De novo* gene synthesis of the sMMO gene cluster and the Epo A, B and C genes

Before the errors in the MCp-oligos for sMMO gene cluster and the Epo A, B and C genes were removed, the quality of these oligos was determined. After the MCp-oligos were amplified directly and cloned into pMD18-T vectors (TaKaRa Bio, Dalian), they were sequenced and analyzed as described above. Then, larger scale error removal was performed during the *de novo* gene synthesis of sMMO gene cluster and Epo A, B and C genes using MCp-oligo pools as described in Figure 2. The 57 oligo subpools of the sMMO gene cluster and the Epo A, B and C genes were individually amplified, re-annealed and error-removed using etMICC. The error-depleted oligos were assembled into fragments. The errors in some of the fragments were further removed using etMICCs. The above error-depleted oligos, assembled fragments and genes were randomly selected for sequencing and analyzed according to the methods described above.

### Gene expression of the sMMO gene cluster

The assembled genes of the sMMO gene cluster produced as described above were additionally validated via expression in *E. coli* BL21 star (DE3). The error-free sMMO X, Y, B, Z, D, C and H genes were individually inserted between the *Nde* I and *Xho* I sites of the pET21-c vector. Alternatively, the error-free sMMO G gene was cloned into the *Nde* I and *Xho* I sites of the pET28-a vector. The resulting sMMO gene expression vectors were transformed into BL21 (DE3) cells, cultivated at 37°C in LB medium (containing 100 μg/ml ampicillin or 50 μg/ml kanamycin) and induced by IPTG (final concentration: 1 mM) under 37°C for 4 h when $OD_{600}$ reached to 0.6. The cells were harvested and then analyzed via 12% SDS-PAGE (52). Because the Epo A, B and C genes only constitute a portion of the recombinant epothilone (Epo) synthesis pathway in *Streptomyces coelicolor* and were synthesized for another laboratory, these genes were only validated via sequencing.

## RESULTS

### Expression and functional evaluation of MutS and the MICCs

The constructed MutS fusion protein was easily expressed in *E. coli* BL21(DE3), and approximately 30 mg (∼0.21 μmol) of the purified MutS fusion protein (approximately 95% pure) (Supplementary Figure S5) was typically obtained from 1000 ml of culture. This amount of the purified proteins could support the production of 175 standard etMICCs.

The constructed fusion protein tMutS displayed lower nonspecific binding to perfectly matched DNA than eMutS.

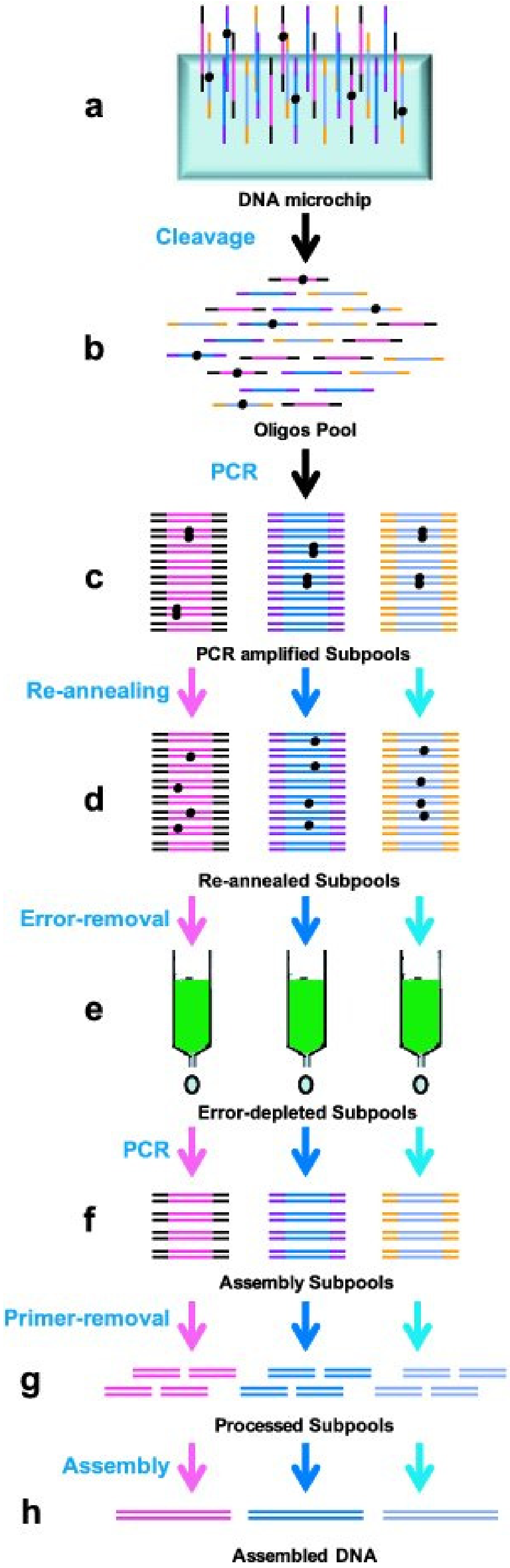


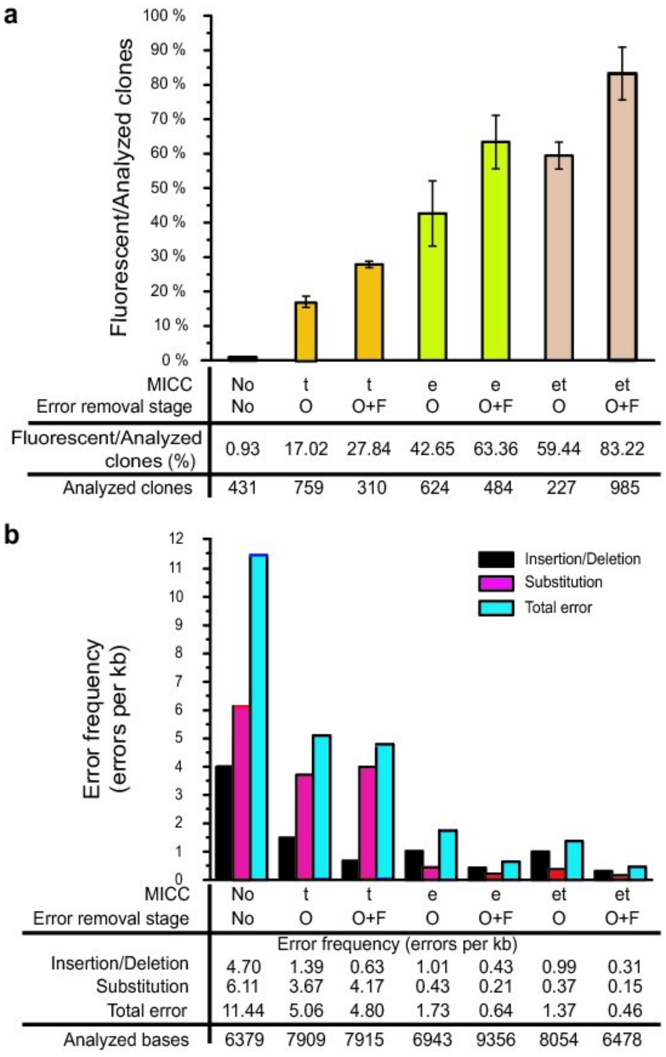
**Figure 2.** Schematic representation of error removal of microchip-synthesized oligonucleotides during *de novo* gene synthesis. (**a∼c**) Oligos are synthesized, cleaved and amplified. Specific primers (black, purple or yellow) are added to separate the oligo pool into subpools via PCR**.** (**d**) The oligos are re-annealed to expose synthetic errors, such as mismatches (black dot). (**e**) Errors are removed using a MICC. Each subpool is eluted through one MICC. (**f**) The error-depleted subpools are amplified separately. (**g**) Primers are removed. (**h**) The DNA is assembled.

As shown in Supplementary Figure S6, when the molar ratio of eMutS to DNA was more than 10:1, the remaining yield of perfectly matched DNA was significantly reduced, and when the ratio was raised to 20:1, almost all of the perfectly matched DNA was bound (Supplementary Figure S6a). In contrast, even when the molar ratio of tMutS to DNA was more than 20:1, only less than half of the perfectly matched DNA was bound (Supplementary Figure S6c). Therefore, the molar ratio of MutS to DNA was 10:1 and 20:1 for eMutS and tMutS, respectively, to avoid problematic nonspecific binding.

Furthermore, the MutS fusion proteins displayed distinct binding capacities to various mismatches, and eMutS bound to most mismatches more effectively than tMutS. As shown in Supplementary Figure S7, both MutS fusion proteins displayed high binding affinity to most deletion/insertion mismatches, and the binding affinities of eMutS to different substitutions were varied (G:T>G:G>C:A, C:C>A:A>G:A, T:C, T:T), whereas tMutS displayed similar binding affinity to all substitution mismatches. Furthermore, using the optimal molar ratios of MutS to DNA, eMutS bound to most mismatches more effectively than tMutS (Supplementary Figure S7).

The *Amax* of eMutS and tMutS to RAC slurry were 8.89 μmol/g and 11.93 μmol/g, respectively, and the *Ka* of eMutS and tMutS to RAC slurry were 7.71 μM and 4.58 μM, respectively. So during the construction of MICC, in which 1.2 nmol of MutS and 20 mg of RAC (1 ml of 20 mg/ml RAC) were mixed, theoretically, more than 99.9% MutS could be immobilized. Actually, when a standard MICC was constructed, less than 0.1% MutS in flow through and washout fraction could be detected, proving that almost all of the added MutS proteins were immobilized on RAC column.

The constructed MICCs could functionally retain mismatch-containing DNA from a DNA mixture. As shown in Supplementary Figure S8, when three-fourths of the oligos consisted of mismatch-containing heteroduplexes (59 bp), after elution through the tMICC, only the first elution (Elution 2 in Supplementary Figure S8b) contained no detectable 59 bp heteroduplexes. In contrast, after elution using the eMICC or the etMICC, only the perfectly matched 54 bp homoduplex (the error-free DNA) was detected. These results indicated that both the eMICC and etMICC could effectively retain mismatch-containing DNA (even when 75% of the DNA sample consisted of mismatch-containing DNA), and the ability of the tMICC to retain mismatches was not as effective as that of eMICC or etMICC. However, because 54 bp homoduplexes (error-free DNA) were visibly detected earlier during the elution (Elution 2, Supplementary Figure S8b) using the tMICC, the tMICC could still be used for correction processing. The recovery efficiency of error-free DNA (the eluted fractions which only contained the 54 bp homoduplex DNA band were considered as error-free DNA recovery) of these MICCs were 5.9%, 51.6% and 86.2% for the tMICC, the eMICC and the etMICC, respectively.



| MICC | No | t | t | e | e | et | et |
|---|---|---|---|---|---|---|---|
| Error removal stage | No | O | O+F | O | O+F | O | O+F |
| Fluorescent/Analyzed clones (%) | 0.93 | 17.02 | 27.84 | 42.65 | 63.36 | 59.44 | 83.22 |
| Analyzed clones | 431 | 759 | 310 | 624 | 484 | 227 | 985 |

| MICC | No | t | t | e | e | et | et |
|---|---|---|---|---|---|---|---|
| Error removal stage | No | O | O+F | O | O+F | O | O+F |
| | Error frequency (errors per kb) | | | | | | |
| Insertion/Deletion | 4.70 | 1.39 | 0.63 | 1.01 | 0.43 | 0.99 | 0.31 |
| Substitution | 6.11 | 3.67 | 4.17 | 0.43 | 0.21 | 0.37 | 0.15 |
| Total error | 11.44 | 5.06 | 4.80 | 1.73 | 0.64 | 1.37 | 0.46 |
| Analyzed bases | 6379 | 7909 | 7915 | 6943 | 9356 | 8054 | 6478 |

**Figure 3.** Evaluation of the error-removal ability of various MICCs. (**a**) Functional analysis of the synthesized *egfp* gene. The ratio of 'fluorescent clones' to 'analyzed clones' is calculated as described in the manuscript for a series of assays with or without error removal using a MICC. (**b**) Sequencing analysis of the synthesized *egfp* gene. The error frequencies of synthesized genes were analyzed as described in the text for the synthesized fragments and genes with or without error removal using a MICC. Then, the occurrence of different types of errors was counted, and the error frequency (errors per kb) of various error-removal protocols was calculated as the ratio of each error to the total bases analyzed. t, tMICC; e, eMICC; et, etMICC; O, one round of error removal at the oligo stage; O+F, two rounds of error removal at both the oligo and fragment stages.

## Error removal using a MICC during EGFP gene synthesis

The fidelity of the synthetic EGFP gene using MCp-oligos without error removal was very poor. As shown in Figure 3a, only 0.93% of the analyzed clones harboring synthetic EGFP full-length gene displayed fluorescence. Seventy-three errors were found among the 14 randomly selected EGFP fragments or full-length genes (a total of 6379 bp), and the error frequency was 11.44/kb (Table 3). Almost all types of mutations were detected, except for the A/T to C/G transition (Table 3).

**Table 3.** Error analysis of synthesized *egfp* gene sequences with or without MICC-mediated error removal

| Error type | Untreated | tMICC | | eMICC | | etMICC | |
|---|---|---|---|---|---|---|---|
| | | One-round[a] | Two-round[b] | One-round[a] | Two-round[b] | One-round[a] | Two-round[b] |
| Multi-error[c] | 4 | 0 | 0 | 2 | 0 | 0 | 0 |
| Deletion | 24 | 11 | 5 | 7 | 4 | 8 | 2 |
| A | 5 | 1 | 0 | 0 | 0 | 0 | 0 |
| C | 8 | 2 | 0 | 1 | 0 | 1 | 0 |
| T | 6 | 7 | 4 | 5 | 4 | 7 | 2 |
| G | 5 | 1 | 1 | 1 | 0 | 0 | 0 |
| Insertion | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Substitution | 39 | 29 | 33 | 3 | 2 | 3 | 1 |
| Transition | 27 | 13 | 9 | 1 | 2 | 0 | 0 |
| G/C to A/T | 26 | 9 | 8 | 0 | 2 | 0 | 0 |
| A/T to G/C | 1 | 4 | 1 | 1 | 0 | 0 | 0 |
| Transversion | 12 | 16 | 24 | 2 | 0 | 3 | 1 |
| G/C to C/G | 2 | 1 | 0 | 0 | 0 | 1 | 0 |
| G/C to T/A | 9 | 14 | 21 | 1 | 0 | 1 | 0 |
| A/T to C/G | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| A/T to T/A | 1 | 0 | 2 | 0 | 0 | 1 | 0 |
| Total errors | 73 | 40 | 38 | 12 | 6 | 11 | 3 |
| Bases sequenced | 6379 | 7909 | 7915 | 6943 | 9356 | 8054 | 6478 |
| Error frequency (errors per kb) | 11.44 | 5.06 | 4.80 | 1.73 | 0.64 | 1.37 | 0.46 |

[a] One round of error removal at the oligo stage.
[b] Two rounds error removal at both the oligo and fragment stages.
[c] Error site located in a sequence that contains more than three adjacent consecutive nucleotide errors.

Error removal using a MICC dramatically reduced the error frequency of synthetic EGFP gene. Firstly, the combination of the two MutS homologs resulted in a higher efficiency of error removal using MCp-oligos than that of using either MutS homolog alone. As shown in Figure 3a, the proportion of fluorescent clones was increased by 18.30-fold (from 0.93% to 17.02%), 45.86-fold (from 0.93% to 42.65%) and 63.91-fold (from 0.93% to 59.44%) after one round of error removal at the oligo stage using the tMICC, the eMICC and the etMICC, respectively. Moreover, the error frequency of the synthetic EGFP gene was reduced by 2.26-, 6.61- and 8.35-fold (from 11.44/kb to 5.06/kb, 1.73/kb and 1.37/kb) using the tMICC, the eMICC and the etMICC, respectively (Table 3). These results revealed that the one round of error correction at the oligo stage using MICCs containing a single MutS homolog significantly increased the proportion of 'fluorescent clones' among the 'analyzed clones', and using a combination of MutS homologs further improved the error-correction efficiency. The order of the error-correction efficiency was etMICC>eMICC>tMICC.

Secondly, the packing mode of etMICC displayed no significant effect on the error-removal efficiency, but the molar ratio between the two MutS homologs was found to influence the error-removal efficiency. As shown in Supplementary Table S7, after error removal using e/tMICC, t/eMICC or e+tMICC, which differed with respect to the packing mode, at the EGFP oligo stage, the error frequencies of the synthetic EGFP gene were decreased to 1.44/kb for the e/tMICC, 1.41/kb for the e+tMICC and 1.21/kb for the t/eMICC (Supplementary Table S7). These results indicated that all of these etMICCs, which were produced according to different packing modes, significantly improved the quality of the assembled gene at a similar efficiency level. On the other hand, as shown in Supplementary Table S8, after error correction of the etMICCs that contained different eMutS:tMutS ratios (1:0.5, 1:1, 1:2 or 1:3), the error rates were decreased to 2.31/kb, 1.21/kb, 1.93/kb and 1.73/kb, respectively. These results indicated that the t/eMICC containing a 1:1 eMutS:tMutS ratio showed the highest effectiveness.

Finally, iterated treatment using MICCs further reduced the DNA errors. Gene assembly processing is error-prone, introducing additional errors into the synthetic DNA. Moreover, after one round of MICC error removal at the oligo stage, some error-containing oligos might escape and appear in the produced DNA constructs. Thus, iterating the error-removal process at the assembled fragment stage was expected to further improve the fidelity of the synthetic DNA. To analyze whether repetition of this error-removal process at the assembled fragment stage could further reduce the error frequency of the synthetic EGFP gene, the error-removal process was performed at both the oligo and fragment stages. The errors in the DNA fragments assembled from the error-depleted oligos were removed using the corresponding MICCs once again. Both the functional assay (fluorescent clone proportion) and the sequencing results revealed that this multi-step error-removal process further improved the quality of the synthetic DNA.

Based on the functional assay, compared to one round of error removal at the oligo stage, the ‘fluorescent clones’ to ‘analyzed clones’ ratio of the two-round error-removal process was further increased by 1.64-fold (from 17.02% to 27.84%), 1.49-fold (from 42.65% to 63.36%) or 1.40-fold (from 59.44% to 83.22%) after error removal using the tMICC, the eMICC or the etMICC, respectively (Figure 3a). The sequencing results also indicated that the error frequency of the synthetic EGFP genes was further decreased by 1.05-fold (from 5.06/kb to 4.80/kb) using the tMICC, 2.70-fold (from 1.73/kb to 0.64/kb) using the eMICC and 2.98-fold (from 1.37/kb to 0.46/kb) using the etMICC (Table 3). After two rounds of error removal, the proportion of fluorescent clones was increased by 29.94-fold (from 0.93% to 27.84%), 68.13-fold (from 0.93% to 63.36%) and 89.48-fold (from 0.93% to 83.22%) using the tMICC, the eMICC and the etMICC, respectively (Figure 3a). In addition, the error frequencies of the synthetic EGFP gene were reduced by 2.38-, 17.88- and 24.87-fold (from 11.44/kb to 4.80/kb, 0.64/kb and 0.46/kb) using the tMICC, the eMICC and the etMICC, respectively (Table 3). Therefore, iterated error removal using MICCs at both the oligo and fragment stages significantly improved the fidelity of synthesized genes.

### Statistical analysis of DNA sequences from *de novo* EGFP gene synthesis

Error removal using the eMICC or the etMICC significantly improved the probability of obtaining an error-free synthetic gene. During *de novo* gene synthesis, the greatest concern is the number of clones that must be analyzed to identify at least one error-free sequence. As shown in Figure 4, two-round error removal using the eMICC or the etMICC at both the oligo and fragment stages significantly improved the probability of generating an error-free clone compared to gene assembly using untreated oligos or fragments. Especially, the number of clones required to be analyzed to identify one error-free sequence was dramatically reduced due to the increased percentage of correct clones among the total clones. For example, to obtain one 1 kb error-free synthetic gene (probability >90%), two to three clones must be screened using the two-round eMICC error-removal protocol, and only one to two clones must be analyzed using the two-round etMICC error-removal protocol. In contrast, without error removal, 47 to 48 clones must be analyzed to obtain a 1 kb error-free gene, which means a vast waste of materials, time and effort. However, using the two-round tMICC error-removal protocol, although the fidelity of the synthetic *egfp* gene was improved from 11.44/kb to 4.80/kb (Table 3), the probability of synthesizing an error-free 1 kb double-stranded product displayed only little improvement (Figure 4); i.e. 44 to 45 clones must be screened to obtain one 1 kb error-free synthetic gene at a probability of >90%. This result may be due to the low effectiveness of the tMICC in substitution error removal. Although the tMICC can remove most deletion/insertion errors (Figure 3b) and can improve the fidelity of synthetic DNA, the substitution errors remained in the DNA products, which resulted in the low probability of obtaining an error-free sequence.

### Error removal and gene synthesis of the sMMO gene cluster and the Epo A, B and C genes

The successful generation of a 720 bp *egfp* gene, along with the 24.87-fold reduction of error frequency in the synthetic gene, suggested that this method could be applied for error removal during *de novo* gene synthesis on a larger scale. This MICC system was further evaluated by the error removal for the MCp-oligos encoding the sMMO gene cluster or the Epo A, B and C genes. In this process, the errors in the oligos of each subpool (containing 11∼32 distinct oligos) were removed using one standard etMICC, and all of the error removals for each oligo subpool could be performed in parallel. Without error removal, the ratio of error-free oligos was 32.11%, and this ratio was significantly improved to 93.04% after one round of error removal using the etMICC, corresponding to a reduction in the error frequency from 12.24/kb to 0.88/kb (Supplementary Table S9). The binding abilities of MutS to the DNAs containing various mismatches were different (Supplementary Figure S7). Therefore, the amount of oligos bound to MutS present in the column is actually sample-dependent. However, the average binding ability of MutS to DNAs was obtained through the half-quantification of the DNA in eluates (unbound DNA) via PAGE (data not shown). About 6∼8 pmol of MCp-oligos could be recovered after error removal through an etMICC, which indicated that 1.2 nmol of MutS (immobilized on an etMICC) could effectively bind about 52∼54 pmol of oligos (about 60 pmol of oligos were loaded onto etMICC). So, the average amount of oligos bound to MutS was about 0.043 mol/mol.

Next, the 57 error-depleted oligo subpools were fully assembled into 78 target fragments as shown in Supplementary Figure S9. The fidelity of the assembled fragments using one round of error correction at the oligo stage was significantly improved by 4.50-fold (error frequency reduced from 14.25/kb to 3.17/kb). Furthermore, the error rate was further reduced to 0.66/kb after performing another round of error removal at the fragment stage. This fidelity improvement trend was also confirmed by the ratio of correct sequence to the total analyzed sequences. The ratio of error-free fragments to all analyzed fragments (∼335 bp) was increased by 11.93-fold (from 3.23% to 38.53%) (Table 4). The additional round of error removal at the fragment stage further improved the percentage of error-free fragments (from 38.53% to 79.07%) (Table 4).

### Gene expression of the sMMO gene cluster

The eight *de novo* synthesized genes of the sMMO gene cluster after MICC error removal, which were codon optimized based on *E. coli* codon usage, were strongly expressed in *E. coli* BL21 (DE3) under the control of the T7/lac promoter. There was no shift-frame error in the synthesized genes, and the expected size of the eight expressed proteins in the cluster was detected via PAGE (Supplementary Figure S10). However, these genes could not be expressed to form an active complex of sMMO due to the insoluble expression of sMMO X and Y.

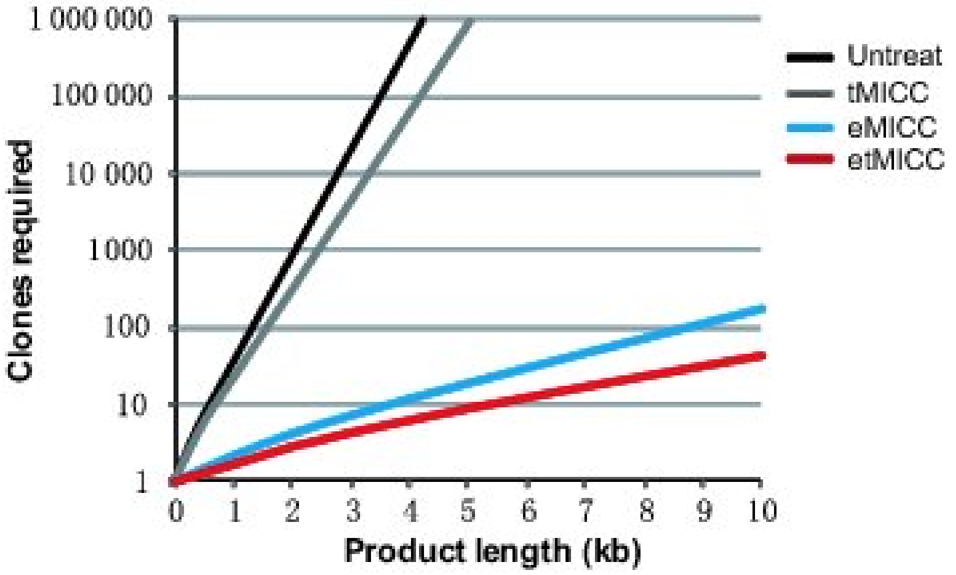


**Figure 4.** Influence of the error rates on *de novo* gene synthesis. The number of clones that must be sequenced to identify at least one error-free sequence with a high probability (90%) after two rounds error removal at both oligo and fragment stages using various MICCs.

**Table 4.** Error analysis of assembled fragment sequences of the sMMO gene cluster and the Epo A, B and C genes with or without error removal using the etMICC

| Error type | Untreated (%[a]) | One-round (%) | Two-round (%) |
|---|---|---|---|
| Multi-error[b] | 6 (1.34%) | 6 (2.63%) | 0 (0.00%) |
| Deletion | 175 (39.06%) | 61 (26.75%) | 3 (16.67%) |
| Insertion | 38 (8.48%) | 11 (4.82%) | 0 (0.00%) |
| Substitution | 229 (51.12%) | 150 (65.79%) | 15 (83.33%) |
| Total errors | 448 | 228 | 18 |
| Bases sequenced | 31 445 | 71 984 | 27 357 |
| Error frequency (error per kb) | 14.25 | 3.17 | 0.66 |
| Percentage of error-free synthetic fragments or genes[c] (%) | 3.23 | 38.53 | 79.07 |

[a] The ratio of each type of error to the total number of errors.
[b] Error site located in a sequence that contains more than three adjacent consecutive nucleotide errors.
[c] The length of the synthetic fragments or genes was ∼335 bp for the sMMO gene cluster and the Epo A, B and C genes.

## DISCUSSION

In this study, a high-throughput and cost-effective MICC error-removal method was developed, and this method could conveniently remove errors from MCp-oligo pools or assembled fragments.

This etMICC system, containing two homologs of immobilized MutS, was more efficient than the use of a single MutS for error removal. Due to the different binding affinities of *Eco*MutS and *Taq*MutS to various types of errors (44,45), a single MutS (*Eco*MutS or *Taq*MutS) immobilized MICC exhibited bias in binding to various types of errors. As shown in Figure 3b and Table 3, the tMICC effectively removed insertion/deletion errors but was less effective in removing substitution errors. The frequencies of insertion/deletion and substitution errors after two rounds of error removal were reduced from 4.70/kb to 0.63/kb and 6.11/kb to 4.17/kb, respectively. In contrast, the eMICC removed both insertion/deletion and substitution errors more effectively than the tMICC. The frequencies of insertion/deletion and substitution errors were reduced from 4.70/kb to 0.43/kb and from 6.11/kb to 0.21/kb, respectively. Combining these two MutS homologs further improved the efficiency of the MICC to remove both substitution and insertion/deletion errors, and also reduced the influence of biased binding. With two rounds of etMICC error removal, the error frequencies were reduced from 4.70/kb to 0.31/kb and 6.11/kb to 0.15/kb for the insertion/deletion and substitution errors, respectively. Therefore, the insertion/deletion error frequency was reduced by 7.46-, 10.93- and 15.16-fold, and the substitution error frequency was reduced by 1.47-, 29.10- and 40.73-fold for the tMICC, the eMICC and the etMICC, respectively (Figure 3b). Consequently, the etMICC was demonstrated to be the optimal type of MICC for error removal.

The MICC method was simpler and more cost-effective than EMC methods. Although the Surveyor nuclease-mediated EMC method reduced the error frequency of *de novo* synthesized genes using MCp-oligos from 1.9/kb to as low as 0.11/kb (21), which was the lowest error rate among EMC methods to date (Table 2), there were still several disadvantages compared to this MICC error removal method. In the EMC methods (16,37,40), after digestion of mismatch-containing DNA by the endonuclease, several steps, including mismatched nucleotide removal and re-assembly, are required to generate full-length error-free DNA sequences. In contrast, in the MICC method, MICC only bound and retained the mismatch-containing DNA in the column and did not destroy the perfectly matched DNA structure, so that the full-length error-free sequences were maintained. As a consequence, the unbound DNA (error-free DNA) could be utilized directly for subsequent applications. Compared to the EMC methods, the expensive exonuclease used for mismatched nucleotide removal and the DNA polymerase or ligase used for DNA re-assembly are not required in the MICC method. Furthermore, the MICC method avoids potential problems such as over-digestion, cross-hybridization and misassembly.

The MICC error-removal method was more convenient than previous MutS methods. CBM3, which exhibits high affinity to a cellulose (RAC) slurry, is a robust and easily accessible molecular tag for protein purification (53,54). In this study, after simple mixing, the MutS fusion protein was easily immobilized on cellulose via CBM3, forming a column that could specifically retain mismatch-containing DNA. There were several advantages of this process over previous methods. (i) The immobilization is simple and stable. The MutS fusion protein can be immobilized on cellulose by simply mixing them together, and the binding is very stable in a hydrophilic solution. (ii) The manipulation is easier. In most previously reported MutS-mediated error correction methods (38,39), error binding and removal are performed separately (errors were bound by MutS in solution, followed by centrifugation, electrophoresis or column separation to remove the DNA-MutS complexes). In contrast, the entire MICC error-removal process was performed on a column. Therefore, the binding and removal of the DNA-MutS complexes occured simultaneously. In this study, the MICC error-removal procedure, including error removal and amplification of error-depleted DNA, could be completed within 1.5 h (Supplementary Figure S11). (iii) The efficiency is enhanced (38). Because the MICC is similar to an affinity chromatography column, the chromatographic effect of the MICC renders it more effective at separating the unbound error-free DNA from the mismatched DNA-MutS complexes. In the previously reported MutS-mediated error correction methods (38,39), the error binding reactions of error-containing DNA by MutS were performed directly in solution. In this study, error removal by MutS was also performed on MCp-oligos directly in solution, but the result was unsatisfactory (data not shown). Because the error rates of the DNA sample were higher (for example, three-fourths of the oligos containing errors), the error-containing DNA from the escaped DNA-MutS complexes significantly reduced the fidelity of the synthetic DNA, causing poor repeatability. Moreover, as described previously, MutS-mediated error-removal methods are more effective for smaller DNA sizes due to the reduced incidence of errors per DNA duplex (38,39). In this study, the MICC containing MutS was suitable for performing error removal at the MCp-oligos stage, significantly reducing the oligo (63∼129 bp) error frequency, and the error-removal efficiency at the oligo stage was higher than that at the fragment stage after oligos assembly (258∼456 bp) (Table 3, Supplementary Table S10).

Aside from amplification, common primers can also improve the efficiency of error removal. In most cases, these primers cannot be avoided when using MCp-oligos for DNA synthesis as described in the 'Introduction' section. The primers were used for oligo amplification and subpool separation. In this study, there is another benefit of the primers: the common primers can partially avoid the low binding affinity of MutS to mismatched sites at the edges of DNA duplexes. In previous reports (38), after a MutS error-removal process, many errors within 15 bp of these edges were retained due to the low binding affinity of MutS to the edges of DNA duplexes. However, in this study, no bias in the error location was detected, which may be due to a benefit of using common primers (≥15 bp) (data not shown).

The MICC technology is cost-effective. To construct a MICC, the RAC slurry used for MICC production was less expensive and more stable than other matrices, such as chitin beads. Using 1 g of RAC ($5/g (53)) and 60 nmol of MutS fusion protein ($0.65; Supplementary Table S11), 50 standard etMICCs could be prepared. In this study, one etMICC could remove errors from one subpool in one batch, and the obtained error-free oligos could be used to assemble one DNA segment 300∼400 bp in length. For each etMICC, the cost (matrix and MutS protein) was about $0.374 (Supplementary Table S11). Furthermore, using two rounds of error removal, the cost of error removal for each oligo was as low as $0.0234/oligo ($0.374 × 2/32 oligos) or ∼$0.0016/bp ($0.374 × 2/456 bp) for a final synthesized DNA sequence. Especially, due to the high probability of obtaining correct sequence after etMICC error removal, this system could decrease the cost of cloning and sequencing to confirm the correct sequences after *de novo* DNA synthesis, consequently reducing the cost of DNA synthesis.

The throughput of the MICC method (throughput: up to 32 oligos in each error correction reaction) was also higher than previously reported protein-mediated methods (typically one fragment per error correction reaction) (22,37,39,40). This MICC system provided an improved-throughput error correction method for oligo pools. The throughput of the MICC system was 11∼32 distinct oligos per MICC treatment, which could be further improved via parallel error-removal processing. For example, the error removal of 57 oligo subpools containing 11∼32 distinct oligos per subpool could be performed in parallel.

In this study, although the MICC system was only applied for *de novo* gene synthesis based on MCp-oligos, the high efficiency and easy operability of this system renders method amenable to utilization for other applications. The next step in the examination of the MICC method will be focused on the scalability of this error correction system to a larger scale *de novo* gene synthesis using MCp-oligos.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGMENTS

We thank Dr Tianyin Zhong and Lijun Bi of the Chinese Academy of Sciences for their gift of the *E. coli* MutS gene. We also thank LC Science for providing the micro-fluid chips for this study.

## FUNDING

National Natural Science Foundation of China [31270149]; National High Technology Research and Development Program ['863' Program:2012AA02A708]; National Basic Research Program of China [2011CBA00801].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Wang,H.H., Isaacs,F.J., Carr,P.A., Sun,Z.Z., Xu,G., Forest,C.R. and Church,G.M. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460**, 894–898.
2. Carr,P.A. and Church,G.M. (2009) Genome engineering. *Nat. Biotechnol.*, **27**, 1151–1162.
3. Bayer,T.S. and Smolke,C.D. (2005) Programmable ligand-controlled riboregulators of eukaryotic gene expression. *Nat. Biotechnol.*, **23**, 337–343.
4. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
5. Gibson,D.G., Benders,G.A., Andrews-Pfannkoch,C., Denisova,E.A., Baden-Tillson,H., Zaveri,J., Stockwell,T.B., Brownley,A., Thomas,D.W., Algire,M.A. *et al.* (2008) Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science*, **319**, 1215–1220.
6. Kobayashi,H., Kaern,M., Araki,M., Chung,K., Gardner,T.S., Cantor,C.R. and Collins,J.J. (2004) Programmable cells: interfacing natural and engineered gene networks. *Proc. Natl. Acad Sci. U.S.A.*, **101**, 8414–8419.
7. Kodumal,S.J., Patel,K.G., Reid,R., Menzella,H.G., Welch,M. and Santi,D.V. (2004) Total synthesis of long DNA sequences: Synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 15573–15578.
8. McDaniel,R. and Weiss,R. (2005) Advances in synthetic biology: on the path from prototypes to applications. *Curr. Opin. Biotechnol.*, **16**, 476–483.
9. Cello,J., Paul,A.V. and Wimmer,E. (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science*, **297**, 1016–1018.
10. Stemmer,W.P.C., Crameri,A., Ha,K.D., Brennan,T.M. and Heyneker,H.L. (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, **164**, 49–53.
11. Gibson,D.G. (2009) Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res.*, **37**, 6984–6990.
12. Gibson,D.G., Glass,J.I., Lartigue,C., Noskov,V.N., Chuang,R.Y., Algire,M.A., Benders,G.A., Montague,M.G., Ma,L., Moodie,M.M. *et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, **329**, 52–56.
13. Li,M.Z. and Elledge,S.J. (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat. Methods*, **4**, 251–256.
14. Bang,D.H. and Church,G.M. (2008) Gene synthesis by circular assembly amplification. *Nat. Methods*, **5**, 37–39.
15. Shao,Z.Y., Zhao,H. and Zhao,H.M. (2009) DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.*, **37**, e16.
16. Kosuri,S., Eroshenko,N., LeProust,E.M., Super,M., Way,J., Li,J.B. and Church,G.M. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat. Biotechnol.*, **28**, 1295–1299.
17. Kong,D.S., Carr,P.A., Chen,L., Zhang,S. and Jacobson,J.M. (2007) Parallel gene synthesis in a microfluidic device. *Nucleic Acids Res.*, **35**, e61.
18. Horvath,S.J., Firca,J.R., Hunkapiller,T., Hunkapiller,M.W. and Hood,L. (1987) An automated DNA synthesizer employing deoxynucleoside 3'-phosphoramidites. *Method Enzymol.*, **154**, 314–326.
19. Sindelar,L.E. and Jaklevic,J.M. (1995) High-throughput DNA-synthesis in a multichannel format. *Nucleic Acids Res.*, **23**, 982–987.
20. Gao,X.L., LeProust,E., Zhang,H., Srivannavit,O., Gulari,E., Yu,P.L., Nishiguchi,C., Xiang,Q. and Zhou,X.C. (2001) A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res.*, **29**, 4744–4750.
21. Quan,J.Y., Saaem,I., Tang,N., Ma,S.M., Negre,N., Gong,H., White,K.P. and Tian,J.D. (2011) Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat. Biotechnol.*, **29**, 449–452.
22. Tian,J.D., Gong,H., Sheng,N.J., Zhou,X.C., Gulari,E., Gao,X.L. and Church,G. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.
23. Richmond,K.E., Li,M.H., Rodesch,M.J., Patel,M., Lowe,A.M., Kim,C., Chu,L.L., Venkataramaian,N., Flickinger,S.F., Kaysen,J. *et al.* (2004) Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.*, **32**, 5011–5018.
24. Zhou,X., Cai,S., Hong,A., You,Q., Yu,P., Sheng,N., Srivannavit,O., Muranjan,S., Rouillard,J.M., Xia,Y. *et al.* (2004) Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences. *Nucleic Acids Res.*, **32**, 5409–5417.
25. Kim,C., Kaysen,J., Richmond,K., Rodesch,M., Binkowski,B., Chu,L., Li,M., Heinrich,K., Blair,S., Belshaw,P. *et al.* (2006) Progress in gene assembly from a MAS-driven DNA microarray. *Microelectron. Eng.*, **83**, 1613–1616.
26. Cleary,M.A., Kilian,K., Wang,Y.Q., Bradshaw,J., Cavet,G., Ge,W., Kulkarni,A., Paddison,P.J., Chang,K., Sheth,N. *et al.* (2004) Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nat. Methods*, **1**, 241–248.
27. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.
28. Baker,M. (2011) Microarrays, megasynthesis. *Nat. Methods*, **8**, 457–460.
29. Tian,J., Ma,K. and Saaem,I. (2009) Advancing high-throughput gene synthesis technology. *Mol. Biosyst.*, **5**, 714–722.
30. Linshiz,G., Ben Yehezkel,T., Kaplan,S., Gronau,I., Ravid,S., Adar,R. and Shapiro,E. (2008) Recursive construction of perfect DNA molecules from imperfect oligonucleotides. *Mol. Syst. Biol.*, **4**, 191.
31. LeProust,E.M., Peck,B.J., Spirin,K., McCuen,H.B., Moore,B., Namsaraev,E. and Caruthers,M.H. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.*, **38**, 2522–2540.
32. Chan,L.Y., Kosuri,S. and Endy,D. (2005) Refactoring bacteriophage T7. *Mol. Syst. Biol.*, **1**, doi:10.1038/msb4100025.
33. Ellington,A. and Pollard,J.D. Jr (2001) Introduction to the synthesis and purification of oligonucleotides. *Curr. Protoc. Nucleic Acid Chem.*, Appendix 3, doi:10.1002/0471142700.nca03cs00.
34. Andrus,A. and Kuimelis,R.G. (2001) Analysis and purification of synthetic nucleic acids using HPLC. *Curr. Protoc. Nucleic Acid Chem.*, Chapter 10, Unit 10.5, doi:10.1002/0471142700.nc1005s01.
35. Cox,J.C., Lape,J., Sayed,M.A. and Hellinga,H.W. (2007) Protein fabrication automation. *Protein Sci.*, **16**, 379–390.
36. Matzas,M., Stahler,P.F., Kefer,N., Siebelt,N., Boisguerin,V., Leonard,J.T., Keller,A., Stahler,C.F., Haberle,P., Gharizadeh,B. *et al.* (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat. Biotechnol.*, **28**, 1291–1294.
37. Fuhrmann,M., Oertel,W., Berthold,P. and Hegemann,P. (2005) Removal of mismatched bases from synthetic genes by enzymatic mismatch cleavage. *Nucleic Acids Res.*, **33**, e58.

38. Carr,P.A., Park,J.S., Lee,Y.J., Yu,T., Zhang,S. and Jacobson,J.M. (2004) Protein-mediated error correction for de novo DNA synthesis. *Nucleic Acids Res.*, **32**, e162.
39. Binkowski,B.F., Richmond,K.E., Kaysen,J., Sussman,M.R. and Belshaw,P.J. (2005) Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Res.*, **33**, e55.
40. Saaem,I., Ma,S.Y., Quan,J.Y. and Tian,J.D. (2012) Error correction of microchip synthesized genes using Surveyor nuclease. *Nucleic Acids Res.*, **40**, e23.
41. Smith,H.O., Hutchison,C.A., Pfannkoch,C. and Venter,J.C. (2003) Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 15440–15445.
42. Whitehouse,A., Deeble,J., Parmar,R., Taylor,G.R., Markham,A.F. and Meredith,D.M. (1997) Analysis of the mismatch and insertion/deletion binding properties of Thermus thermophilus, HB8, MutS. *Biochem. Biophys. Res. Commun.*, **233**, 834–837.
43. Stanislawska-Sachadyn,A., Sachadyn,P., Jedrzejczak,R. and Kur,J. (2003) Construction and purification of his6-Thermus thermophilus MutS protein. *Protein Expr. Purif.*, **28**, 69–77.
44. Brown,J., Brown,T. and Fox,K.R. (2001) Affinity of mismatch-binding protein MutS for heteroduplexes containing different mismatches. *Biochem. J.*, **354**, 627–633.
45. Cho,M., Chung,S., Heo,S.D., Ku,J. and Ban,C. (2007) A simple fluorescent method for detecting mismatched DNAs using a MutS-fluorophore conjugate. *Biosens. Bioelectron.*, **22**, 1376–1381.
46. Hong,J., Ye,X., Wang,Y. and Zhang,Y.H. (2008) Bioseparation of recombinant cellulose-binding module-proteins by affinity adsorption on an ultra-high-capacity cellulosic adsorbent. *Anal. Chim. Acta*, **621**, 193–199.
47. Bothwell,M.K. and Walker,L.P. (1995) Evaluation of parameter-estimation methods for estimating cellulase binding constants. *Bioresour. Technol.*, **53**, 21–29.
48. Jiricny,J., Hughes,M., Corman,N. and Rudkin,B.B. (1988) A human 200-kDa protein binds selectively to DNA fragments containing G.T mismatches. *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 8860–8864.
49. Higuchi,R., Krummel,B. and Saiki,R.K. (1988) A general method of in vitro preparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. *Nucleic Acids Res.*, **16**, 7351–7367.
50. Cabantous,S. and Waldo,G.S. (2006) In vivo and in vitro protein solubility assays using split GFP. *Nat. Methods*, **3**, 845–854.
51. Miller,E.M. and Nickoloff,J.A. (1995) Escherichia coli electrotransformation. *Methods Mol. Biol.*, **47**, 105–113.
52. Laemmli,U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680–685.
53. Wan,W., Wang,D.M., Gao,X.L. and Hong,J. (2011) Expression of family 3 cellulose-binding module (CBM3) as an affinity tag for recombinant proteins in yeast. *Appl. Microbiol. Biot.*, **91**, 789–798.
54. Shoseyov,O., Shani,Z. and Levy,I. (2006) Carbohydrate binding modules: biochemical properties and novel applications. *Microbiol. Mol. Biol. Rev.*, **70**, 283–295.